# An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen

***Abstract*—In the development process of noise-reduction algorithms, an objective machine-driven intelligibility measure which shows high correlation with speech intelligibility is of great interest. Besides reducing time and costs compared to real listening experiments, an objective intelligibility measure could also help provide answers on how to improve the intelligibility of noisy unprocessed speech. In this paper, a short-time objective intelligibility measure (STOI) is presented, which shows high correlation with the intelligibility of noisy and time–frequency weighted noisy speech (e.g., resulting from noise reduction) of three different listening experiments. In general, STOI showed better correlation with speech intelligibility compared to five other reference objective intelligibility models. In contrast to other conventional intelligibility models which tend to rely on global statistics across entire sentences, STOI is based on shorter time segments (386 ms). Experiments indeed show that it is beneficial to take segment lengths of this order into account. In addition, a free Matlab implementation is provided.**

***Index Terms*—Noise reduction, objective measure, speech enhancement, speech intelligibility prediction.**

## I. Introduction

SPEECH processing systems often introduce degradations and modifications to clean or noisy 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 costly and time consuming. Therefore, accurate and reliable objective evaluation methods are of interest since they might replace listening tests, at least in some stages of the algorithm development process. Although it is not straightforward to completely characterize a noisy or processed speech signal, people tend to divide the evaluation into the attributes speech quality, (i.e., pleasantness/naturalness of speech) and speech intelligibility. In this paper, we focus 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 [1]. Kryter [2] 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 as S3.5-1997 [3]. Since AI is mainly meant for simple linear degradations, e.g., additive noise, Steeneken and Houtgast [4] proposed the speech transmission index (STI), which is also able to predict the intelligibility of reverberated speech and nonlinear 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. The majority of recently published models are still based on the fundamentals of AI, e.g., [5], [6] and STI (see the work from Goldsworthy and Greenberg [7] for an overview).

Although the just mentioned objective intelligibility measures are suitable for several types of degradation (e.g., additive noise, reverberation, filtering and clipping), it turns out that they are less appropriate for methods where noisy speech is processed by some type of time–frequency (TF) varying gain function. This includes single-channel noise-reduction algorithms (see the work from Loizou [8] for an overview), but also speech separation techniques like ideal time frequency segregation (ITFS) [9], where typically a binary TF-weighting is used. For example, STI and various STI-based measures predict an intelligibility improvement when spectral subtraction is applied [7], [10], [11]. This is not in line with the results of listening experiments in literature, where it is reported that single-channel noise-reduction algorithms generally are not able to improve the intelligibility of noisy speech, e.g., [12]. Furthermore, measures like the coherence SII (CSII) [6] and a normalized covariance-based STI procedure (CSTI) [7], both show low correlation with the intelligibility of ITFS-processed speech [13].

In a recent study, Ma *et al.* [14] showed that several intelligibility measures could benefit from the use of new (signal-dependent) band-importance functions (BIF). For example, the correlation of CSII and CSTI with the speech intelligibility of single-channel noise-reduced speech increased significantly by the use of these new BIFs [14]. It is of interest to see if these methods would also work for other types of TF-weighted noisy speech, e.g., ITFS-processed speech. Also two different methods have been proposed lately, which indicate promising results for ITFS-

Manuscript received August 24, 2010; revised November 30, 2010; accepted February 04, 2011. Date of publication February 14, 2011; date of current version July 22, 2011. This work was supported by the Oticon foundation and the Dutch Technology Foundation STW. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

C. H. Taal, R. C. Hendriks, and R. Heusdens are with the Delft University of Technology, Signal Information and Processing Lab, 2628 CD Delft, The Netherlands (e-mail: c.h.taal@tudelft.nl).

J. Jensen is with Oticon A/S, 2765 Smørum, Denmark.

Digital Object Identifier 10.1109/TASL.2011.2114881

1558-7916/$26.00 © 2011 IEEE

processed speech [15], [16]. These methods have not been evaluated yet for intelligibility prediction of single-channel noise reduced-speech.

Therefore, a reliable objective intelligibility measure which has high correlation with the speech intelligibility of noisy and various types of TF-weighted noisy speech is of great interest. Such a measure could be used for the analysis of algorithms that process noisy speech. In addition, new algorithms could be developed, which optimize for such an objective measure. To analyze the effect of certain signal degradations on the speech intelligibility in more detail, an objective measure must be of a simple structure. Nevertheless, some measures are based on a large amount of parameters which are extensively trained for a certain dataset. This makes these measures less transparent, and therefore less appropriate for these evaluative purposes.

In this paper, a simple objective intelligibility measure is proposed which has a strong monotonic relation with the intelligibility scores of various listening tests where noisy speech is processed by some type of TF-weighting.[1] The model has a simple structure in the sense that it is based on only two free parameters. Moreover, it shows better performance than five other reference objective intelligibility measures for these listening tests.

### A. Rationale of Proposed Intelligibility Measure

A general approach in the field of objective intelligibility assessment is to make some type of correlation-based comparison between the spectro-temporal internal representations of the clean and degraded speech signal. For example, CSTI [7] determines a correlation coefficient between octave-band temporal envelopes and CSII [6] is based on the coherence function, which is a measure of correlation between complex Fourier-coefficients, over time, as a function of frequency. Another example of a correlation-based measure is the normalized subband envelope correlation (NSEC) proposed by Boldt and Ellis [16]. In contrast to SNR-based measures (e.g., [3], [5]), the benefit of such a correlation-based approach is the fact that the introduced degradation (i.e., "the noise") is not needed as a separate signal in isolation from the clean speech. Hence, in addition to speech corrupted by background noise, a correlation-based comparison can also be used for other (nonlinear) types of distortions, e.g., noise-reduced speech, where it is not that straightforward how to separate the clean speech from its introduced distortion.

Several correlation-based measures estimate correlation values for the complete signal of interest at once (e.g., [6], [7], [16]). Typically, these signals have a length in the order of tens of seconds. A problem which occurs with an analysis length of this order is the fact that a few signal regions with high amplitudes (either from the clean or the degraded speech) may dominate the eventual estimated correlation. There are also measures based on a very short segment size (20–30 ms), e.g., [15]. However, as a consequence of their poor modulation frequency resolution, certain low temporal modulations are excluded which are important for speech intelligibility. According the results from Drullman *et al.* [17] temporal modulations below 2–3 Hz can be removed without affecting intelligibility. Therefore, an analysis window with a length around 333–500 ms would be more appropriate. This is also more in line with the results from van den Brink [18] which suggest that the temporal integration time of the auditory system has an upper bound of a few hundreds of milliseconds.

Motivated by this we propose a short-time objective intelligibility (STOI) measure, based on a correlation coefficient between the temporal envelopes of the clean and degraded speech, in short-time (384 ms), overlapping segments. Indeed, by experimenting with this segment-length we will show that one actually benefits using segments of this duration.

### B. Further Outline

The remaining part of this paper is organized as follows: first more details are given about STOI in Section II. Then, in Section III, three different intelligibility listening experiments are described for different types of processed noisy speech. These results are used to evaluate the intelligibility prediction performance of STOI. Next, more details are given in Section IV about the general evaluation procedure. Finally, the evaluation results are presented together with a discussion in Section V after which conclusions are drawn.

## II. STOI

The basic structure of STOI is illustrated in Fig. 1. It is a function of the clean and degraded speech, denoted by $x$ and $y$, respectively. The output of STOI is a scalar value which is expected to have a monotonic relation with the average intelligibility of $y$ (e.g., the percentage of correctly understood words averaged across a group of users). A sample-rate of 10 kHz is used, in order to capture a relevant frequency range for speech intelligibility [1].[2]

First, both signals are TF-decomposed in order to obtain a simplified internal representation resembling the transform properties of the auditory system. This is obtained by segmenting both signals into 50% overlapping, Hann-windowed frames with a length of 256 samples, where each frame is zero-padded up to 512 samples. Before evaluation, silent regions which do not contribute to speech intelligibility are removed. This is done by first finding the frame with maximum energy of the clean speech signal. Both signals are then reconstructed, excluding all the frames where the clean speech energy is lower than 40 dB with respect to this maximum clean speech energy frame. Then, a one-third octave band analysis is performed by grouping DFT-bins. In total 15 one-third octave bands are used, where the lowest center frequency is set equal to 150 Hz and the highest one-third octave band has a center-frequency equal to approximately 4.3 kHz.

Let $\hat{x}(k,m)$ denote the $k^{th}$ DFT-bin of the $m^{th}$ frame of the clean speech. The norm of the $j^{th}$ one-third octave band, referred to as a TF-unit, is then defined as

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2} \tag{1}$$

[1] An intelligibility model can also predict absolute intelligibility scores (e.g., a percentage of correctly understood words), however, for analysis and/or optimization monotonicity with speech intelligibility is already of great interest.

[2] Note, that the sample-rate of 10 kHz is not critical. When the window length (in ms) and the frequency-range of the critical bands is preserved the method can be extended to other sample-rates.

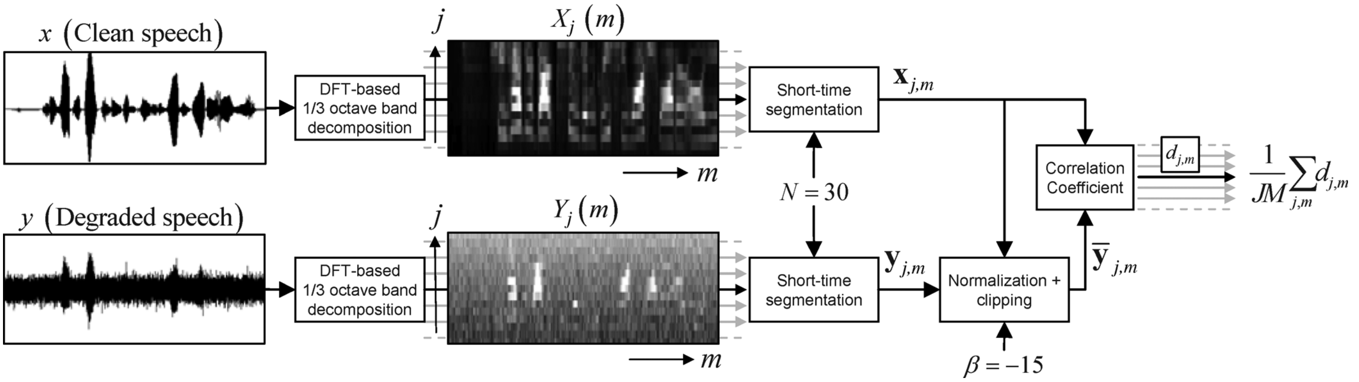


Fig. 1. STOI is a function of the clean and degraded speech, which are first decomposed into DFT-based, one-third octave bands. Next, short-time (384 ms) temporal envelope segments of the clean and degraded speech are compared by means of a correlation coefficient. Before comparison, the short-time degraded speech temporal envelopes are first normalized and clipped (see text for more details). These short-time intermediate intelligibility measures $(d_{j,m})$ are then averaged to one scalar value, which is expected to have a monotonic increasing relation with the speech intelligibility.

where $k_1$ and $k_2$ denote the one-third octave band edges, which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly, and is denoted by $Y_j(m)$.

STOI is a function of a TF-dependent intermediate intelligibility measure, which compares the temporal envelopes of the clean and degraded speech in short-time regions by means of a correlation coefficient. The following vector notation is used to denote the short-time temporal envelope of the clean speech:

$$\mathbf{x}_{j,m} = \left[X_j(m-N+1), X_j(m-N+2), \ldots, X_j(m)\right]^T . \tag{2}$$

where $N = 30$ which equals an analysis length of 384 ms (see Section V-C for details on this particular choice). Similarly, $\mathbf{y}_{j,m}$ denotes the short-time temporal envelope of the degraded speech. As illustrated in Fig. 1, $\mathbf{y}_{j,m}$ is first normalized and clipped before comparison. The rationale behind the normalization procedure is to compensate for global level differences which should not have a strong effect on the speech intelligibility (e.g., due to different playback levels of $x$ and $y$). The clipping procedure makes sure that the sensitivity of the model towards one TF-unit which is severely degraded is upper bounded. As a consequence, further degradation of a speech TF-unit which is already completely degraded (i.e., "unintelligible") does not lead to a lower intelligibility prediction by the model.

Let $\mathbf{x}(n)$ denote the $n^{th}$ element of $\mathbf{x}$, where $n \in \{1, \ldots, N\}$ and $\|\cdot\|$ represent the $\ell_2$ norm. The normalized and clipped version of $\mathbf{y}$, say $\bar{\mathbf{y}}$, is then given by

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{\|\mathbf{x}_{j,m}\|}{\|\mathbf{y}_{j,m}\|}\mathbf{y}_{j,m}(n), (1+10^{-\beta/20})\mathbf{x}_{j,m}(n)\right). \tag{3}$$

where $\beta = -15$ dB refers to the lower signal-to-distortion (SDR) bound. Indeed, for this case we have that

$$SDR = 10\log_{10}\left(\frac{\mathbf{x}_{j,m}(n)^2}{\left(\bar{\mathbf{y}}_{j,m}(n) - \mathbf{x}_{j,m}(n)\right)^2}\right) \geq \beta. \tag{4}$$

The intermediate intelligibility measure is defined as the sample correlation coefficient between the two vectors

$$d_{j,m} = \frac{\left(\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}\right)^T \left(\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m}}\right)}{\left\|\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}\right\| \left\|\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m}}\right\|} \tag{5}$$

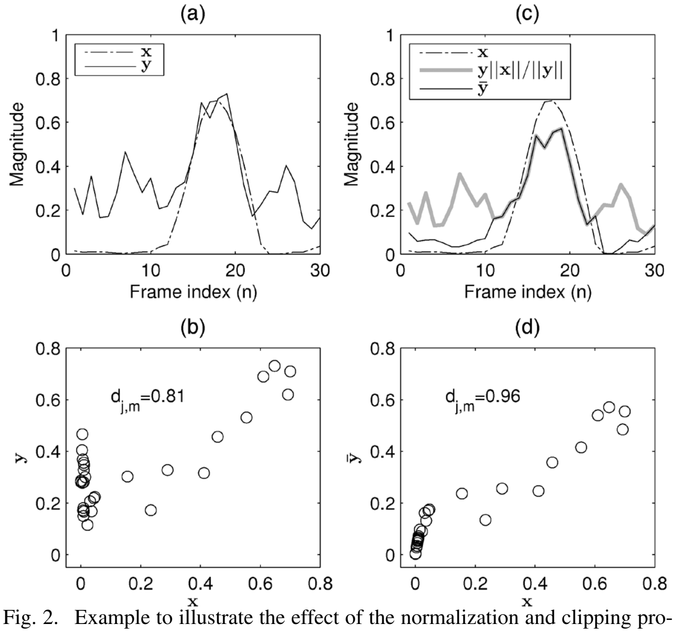


Fig. 2. Example to illustrate the effect of the normalization and clipping procedure. A clean $(\mathbf{x})$ and noisy $(\mathbf{y})$ speech vector of 30 time-frames (386 ms) is shown in (a) together with a corresponding scatter-plot in (b). Similarly, (c) and (d) show the results for the normalized $(\mathbf{y}\|\mathbf{x}\|/\|\mathbf{y}\|)$ and clipped+normalized $(\bar{\mathbf{y}})$ degraded vector (See text for more details). Notice that the clipping procedure reduces the effect of the noise in noise-only regions.

where $\mu_{(\cdot)}$ refers to the sample average of the corresponding vector. Finally, the average of the intermediate intelligibility measure over all bands and frames is calculated:

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m} \tag{6}$$

where $M$ represents the total number of frames and $J$ the number of one-third octave bands.

### *A. Example of Normalization and Clipping Procedure*

To illustrate the effect of the normalization and clipping procedure an example is given in Fig. 2, where subplot (a) shows a short-time temporal envelope of a clean speech vector together with a noise corrupted version (one frequency band is shown). A corresponding scatter plot is given in Fig. 2(b), where $d_{j,m} = 0.81$ denotes the outcome of the intermediate intelligibility measure when clipping would be discarded, i.e., $\bar{\mathbf{y}}$ is replaced with

$\mathbf{y}$ in (5) (note, that the applied scaling due to the normalization does not directly affect the correlation coefficient). The normalized and clipped+normalized vectors of the degraded speech are shown in Fig. 2(c) together with a scatter-plot in Fig. 2(d). From Fig. 2(c) it can be observed that the clipping procedure is mainly effective in the noise-only regions (i.e., $n < 11$ and $n > 23$). As a consequence, a higher correlation is obtained ($d_{j,m} = 0.96$) compared to the situation when clipping would be discarded ($d_{j,m} = 0.81$). This is desired, since it is expected that degrading these regions (where speech is absent within a sentence) will only have a minor impact on speech intelligibility.

## III. Listening Experiments

In order to evaluate the performance of STOI, its output as described in (6) is compared with the intelligibility scores from three different intelligibility listening experiments. In each of these listening tests, noisy speech is processed with different types of TF-weightings. While the first experiment comprises a method where noisy speech signals are ITFS-processed [19], the second listening test evaluates the effect on speech intelligibility due to two conventional single-channel noise-reduction schemes. The last experiment evaluates the effect of modifying the applied TF-weighting based on ITFS with artificially introduced errors [20]. Next, more details will be given about these three listening tests.

### A. Ideal Time–Frequency Segregation

The intelligibility data from the first experiment is obtained from a listening test conducted by Kjems *et al.* [19], where noisy speech signals are ITFS-processed. ITFS is a technique which can improve the intelligibility of noisy speech significantly by applying a binary modulation pattern in a TF-representation.[3] This binary modulation pattern has a value equal to one, when the SNR within a certain TF-component exceeds a user-defined local criterion (LC), and is commonly referred to as the ideal binary mask (IBM). The IBM is given as follows:

$$IBM(t,f) = \begin{cases} 1, & \text{if } T(t,f) - M(t,f) > LC \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

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 based on a 64-channel gammatone filterbank linearly spaced on an ERB scale between 55 and 7500 Hz. The filterbank is followed by a time segmentation of 20-ms windowed frames with an overlap of 10 ms.

Lowering the LC-parameter in (7) will increase the number of ones in the IBM, where LC $= -\infty$ will result in an IBM with ones only (i.e., the noisy speech is unprocessed). High values for LC will result in sparse IBMs. Kjems *et al.* showed that for certain settings of the LC-parameter as a function of the global SNR, noisy speech can be made fully intelligible. This even holds for the situation that essentially pure noise is modulated with the IBM [19]. An alternative IBM is also included which is only based on the clean speech. This so-called target binary mask (TBM) [19] 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-component. Therefore, the noise itself is not needed in order to determine the TBM. For more details on the algorithm (e.g., signal reconstruction) the reader is referred to Kjems *et al.* [19].

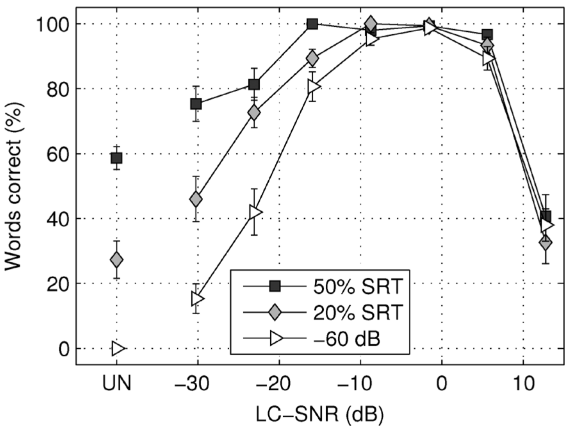


Fig. 3. Ideal time–frequency segregation: average-user intelligibility scores with standard errors of clean speech degraded with speech shaped noise (SSN) at three different SNRs (20%SRT, 50%SRT, $-60$ dB), followed by ITFS processing (replotted from Kjems *et al.* [19]). The percentage of correct words is plotted as a function of the ITFS algorithm's LC-parameter corrected with the global SNR (see text for more details). The leftmost point refers to an IBM with only ones, i.e., $LC = -\infty$, which equals the noisy unprocessed speech (UN).

The test signals are taken from the Dantale II corpus [21], where each excerpt consists of five words, all spoken by the same Danish female speaker. These sentences are degraded by four different types of additive noise: speech shaped noise (SSN), cafeteria noise, noise from a bottling factory hall and car interior noise at three different SNRs: 20% and 50% speech reception threshold (SRT)[4] and an SNR of $-60$ dB, which represents essentially pure noise. Eight different LC-values are chosen, including an unprocessed condition where only the noisy speech is presented, i.e., LC $= -\infty$ (see the work from Kjems *et al.* [19] for more details on the SNR values and LC-parameters).

For the listening experiment, 15 normal-hearing native Danish speaking subjects participated, where the correctly recognized words are recorded by an operator without providing any form of feedback. Each subject listened to two five-word sentences for each condition. The average score for all users for one condition is then obtained by the average percentage of correct words. In total, this gives $(4 * \text{IBM} + 3 * \text{TBM}) * (3 * \text{SNR}) * (8 * \text{LC}) = 168$ conditions to be tested in the listening experiment. Only three TBM conditions are included since the TBM equals the IBM for the case that SSN is used, by definition.

As an example, the results for all SSN conditions processed with an IBM are plotted in Fig. 3. Here, the percentage of correct words is plotted as a function of the LC-parameter corrected with the global SNR. By subtracting the global SNR from the LC-parameter one can observe from the figure that the noisy

[3] Note, that here the clean speech is needed separately from the noise source, therefore, large intelligibility improvements are possible. Although this may not seem practical in real-life noisy conditions, this type of processing will deliver a wide variety of processed signals with largely varying intelligibility scores. This is of interest for evaluating STOI.

[4] The $x\%$ SRT is the SNR at which the average listener achieves $x\%$ intelligibility.

speech becomes fully intelligible when the corrected SNR is close to 0 dB. Note, that the leftmost point refers to an IBM with only ones, which equals the condition where the noisy speech is unprocessed (indicated by UN in Fig. 3.

### B. Single-Channel Noise Reduction

The second 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-Malah (EM) [22] which was developed under the assumption that speech and noise DFT coefficients are Gaussian, and 2) an improved version by Erkelens *et al.* (SG) [23], 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 [22] with a smoothing factor of $\alpha = 0.98$. The noise PSD in EM and SG is estimated using minimum statistics [24] and the noise-tracker by Hendriks *et al.* [25], 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 $\nu = 0.6$ [23].

As with the previous listening experiment from Section III-A the speech signals are from the Dantale II corpus [21], which are degraded by 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 [19]) 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) and five SNR values make up 3*5 = 15 conditions. For each of the 15 conditions, each listener is presented with ten five-word sentences. The average score for all users and for one condition was consequently obtained by the average percentage of correct words.

The results from the listening experiment are shown in Fig. 4. 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 I). This result is in line with the conclusions from Hu and Loizou [26] where, in general, no noise-reduction scheme could improve the intelligibility of noisy speech.

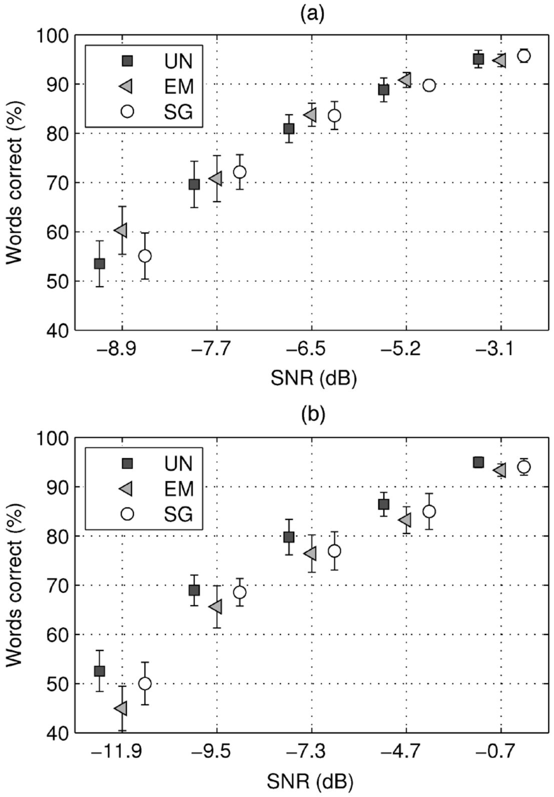


Fig. 4. Single-channel noise reduction: average-user intelligibility scores with standard errors for unprocessed noisy (UN) speech, and two noise-reduction schemes (EM, SG) for (a) speech shaped noise and (b) cafe noise. (a) SSN. (b) Cafe.

TABLE I
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 |

### C. ITFS With Artificially Introduced Errors

As with Section III-A, the last listening experiment is also based on ITFS. Since the clean speech is needed in (7), the high intelligibility improvements illustrated in Fig. 3 are generally not obtained in real-life noisy conditions. In practice, one has to estimate the IBM from the noisy speech, which will typically lead to errors [27]. In order to find implications for noise reduction, Li and Loizou [20] investigated the effect of artificially introduced errors in the IBM for the case that $\mathrm{LC} = 0$ dB. We regenerated these processed signals as described by Li and Loizou [20], which are used for the evaluation of STOI.

Three types of errors in the IBM were considered by Li and Loizou. 1) A general error, which refers to the procedure where the value of a random selection of TF-units (FFT-based) per time-frame (20 ms) is changed, i.e., a zero in the IBM becomes a one and vice-versa. 2) A type-I error, where a certain percentage of TF-units in the IBM, originally labeled as zero, is changed into a one. (3) A type-II error, where a random selection of TF-units, originally labeled as one, are changed into a zero. For the general errors, five amounts of error in terms of percentage are used (5%–40%) and three noise types are considered (SSN, 2-talker babble noise and 20-talker babble noise) all mixed at $-5$-dB SNR. For the type-I and type-II errors only the 20-talker babble noise is used (also $-5$-dB SNR) and eight percentages are considered (20%–95%). Moreover, the unprocessed noisy speech for all three noise types is also included. This gives us a total of 31 conditions: $(3 \text{ noise types} * 5 \text{ error values}) + (2 \text{ error types}) * (8 \text{ error values}) + 3 * \text{unprocessed}$.

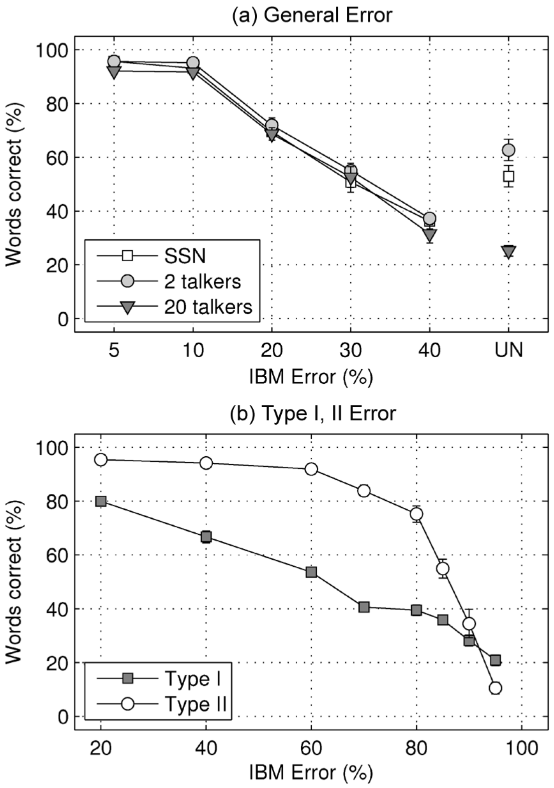


Fig. 5. ITFS with artificially introduced errors: average-user intelligibility scores with standard errors for artificially introduced errors in the IBM (replotted from the work of Li and Loizou [20]). (a) The effect on speech intelligibility due to a general error for three different noise types at $-5$-dB SNR. (b) The effect on speech intelligibility due to a type-I or type-II error for 20-speaker babble noise at $-5$-dB SNR (see text for details). UN indicates the unprocessed noisy speech.

Seven normal-hearing listeners participated in the listening experiment from Li and Loizou, where all subjects were native American English speakers. The speech material consisted of sentences taken from the IEEE database, see, e.g., [8], all produced by the same male speaker, where 20 sentences were used per condition.

The results from Li and Loizou are replotted in Fig. 5 [20]. Fig. 5(a) illustrates that a general error in the IBM has a similar impact on intelligibility for all noise types. That is, the gain in intelligibility due to the applied IBM drops fast when the percentage of incorrectly TF-units is larger then 10%. In Fig. 5(b), it can be clearly observed that a Type-I error has a stronger effect on intelligibility compared to a Type-II error.

## IV. Evaluation Procedure

In order to evaluate STOI, 30 sentences are taken from the relevant corpus for each condition of the three listening experiments. That is, the Dantale sentences [21] for listening experiment Sections III-A and III-B, and the IEEE sentences (see, e.g., [8]) for listening experiment III-C. These 30 clean and processed sentences are then concatenated and resampled to 10 kHz. We experimented with different values of $N \in [10, 20, 30, 50, 100, 500]$ and $\beta \in [-\infty, -35, -25, -15, -10]$ only for the intelligibility data originating from the ITFS listening experiment from Section III-A. Note, that $\beta = -\infty$ equals the condition without clipping and therefore without normalization (without clipping the correlation coefficient from (5) is independent of the applied normalization procedure).

Best performance was obtained with $N = 30$ and $\beta = -15$ dB. These settings were used for evaluating STOI with respect to the remaining two listening tests.

### A. Mapping

We are interested in measuring the monotonic relation between the outcomes of STOI and the actual intelligibility scores. First, a mapping is used in order to account for a nonlinear relation between the STOI outcomes and the intelligibility scores. The main reason for this mapping procedure is to linearize the data such that we can use merits like a linear correlation coefficient. Second, with this procedure the STOI scores are mapped to an absolute intelligibility prediction which makes it possible to reveal the distribution of prediction errors amongst all the listening test conditions. For this a logistic function is used:

$$f(d) = \frac{100}{1 + \exp(ad + b)} \qquad (8)$$

where $a$ and $b$ are free parameters, which are fitted to the data with a nonlinear least squares procedure. Note, that a logistic function is also monotonic and will therefore not influence the monotonicity between STOI and the intelligibility scores.

While experiments Section III-A and III-B use the Danish Dantale sentences, listening experiment III-C uses the English IEEE database. In contrast to the IEEE database, the Dantale sentences are taken from a closed set of words. As a consequence, the Dantale sentences are easier to understand for equal adverse listening conditions compared to the IEEE sentences. Objective measures, in general, do not exploit this *a priori* knowledge and will therefore need a different mapping function for each corpus. Motivated by this, the mapping procedure is applied independently for both corpora denoted by $f_{\mathrm{Dantale}}$ and $f_{\mathrm{IEEE}}$. Moreover, for the Dantale corpus only the data from the ITFS listening test is used to fit the mapping, which is then reused for the single-channel noise reduction conditions.

### B. Reference Objective Measures

The results of STOI are compared with five other reference objective measures which are all promising candidates for intelligibility prediction of TF-weighted noisy speech. In this section, some details will be given for each model.

*1) Dau Auditory Model:* The perceptual model developed by Dau *et al.* [28] (DAU) acts as an artificial observer and is originally used for accurately predicting masking thresholds for various masking conditions [29]. More recently, it is also shown that the model can be used as a good intelligibility predictor for ITFS-processed speech [13], [15]. We compare STOI with the intelligibility-model based on the Dau auditory model as proposed by Christiansen *et al.* [15]. This model is already evaluated with the ITFS intelligibility data from Section III-A [15] where good prediction results were obtained. It is of interest to see its performance compared to STOI. First, the spectro-temporal internal representations of $x$ and $y$ are determined as described in [28], followed by a segmentation in 20-ms frames within each auditory channel. Subsequently, the internal presentations within each frame of the clean and degraded speech are compared by means of a correlation coefficient jointly over time and frequency. As proposed by Christiansen *et al.* only a subset of frames with high speech energy were considered from which an average correlation coefficient is obtained.

*2) Coherence Speech-Intelligibility Index:* The coherence speech-intelligibility index (CSII) [6] is based on the coherence function which equals the normalized cross-spectral density between the clean and degraded speech. The coherence function is then translated to several frequency-band dependent SDRs, which are combined to one score as in the conventional speech intelligibility index (SII) [3]. That is, the SDRs are limited and normalized and are combined by computing a weighted average based on perceptual band-important functions (BIFs). It is shown that CSII can successfully predict the effect on speech intelligibility due to nonlinear types of speech distortions like peak-clipping and center-clipping [6]. Recent results show that by using signal-dependent BIFs instead [14], the performance of CSII with respect to single-channel noise-reduced speech signals will increase significantly. This CSII variant is also used for the comparison with STOI (referred to as $\mathrm{CSII}_{\mathrm{mid}}$, $W_4$, $p = 1$ by Ma *et al.* [14]).

*3) Normalized Covariance Based Speech Transmission Index:* The normalized covariance based speech transmission index CSTI is based on the correlation coefficient between the band magnitude envelopes within 8 octave bands [7], [30]. The measure shows good results with respect to various types of signal distortions, e.g., clipping and spectral subtraction [7]. The correlation coefficients per band are translated to an SNR and combined in a similar way as with the CSII. Also for this measure new BIFs were recently proposed in order to improve its performance with respect to single-channel noise reduced speech [14]. These BIFs are also used in this paper for comparison with STOI (referred to as NCM, $W_i^{(1)}$, $p = 1.5$ by Ma *et al.* [14]).

*4) Frequency-Weighted Segmental SNR:* The frequency-weighted segmental SNR (FWS) is included for comparison as proposed by Hu and Loizou [26]. The measure determines the SNR within several frequency bands in short-time frames (20 ms) which are limited and normalized. Here, the clean and processed speech frames are first normalized to have unit area. This normalization procedure was found to be of critical importance in order to predict the speech quality of enhanced speech [26]. In addition, FWS also showed promising results with respect to predicting the speech-intelligibility of single-channel noise-reduced speech [14]. Again, new BIFs are used as proposed by Ma *et al.* (referred to as fwSNRseg, $p = 1$ by Ma *et al.* [14]) in order to combine the clipped and normalized SNRs.

*5) Normalized Subband Envelope Correlation:* The final intelligibility measure is based on the normalized subband envelope correlation (NSEC) [16]. This model is already evaluated with the ITFS intelligibility data from Section III-A) [16] where good prediction results were obtained. Hence, it is of interest to compare its performance with STOI. First, a 64-channel gammatone filterbank is applied on the clean and processed speech, after which the normalized, compressed and high-pass filtered intensity envelopes are extracted. The eventual distance between the clean and processed speech is then defined by the normalized correlation over all time and frequency points. Similarly as with DAU, the correlation is determined jointly over time and frequency.

## V. Results

First the performance of STOI in terms of several correlation measures will be reported for each listening test after which more details are given about how the STOI intelligibility prediction errors are distributed over the various listening tests and processing conditions. Then the effect of the clipping parameter $\beta$ and the analysis length $N$ is analyzed followed by a comparison with several other intelligibility models.

### A. Correlation Between STOI and Intelligibility Scores

The performance of STOI is evaluated by means of the correlation coefficient $(\rho)$ and the standard deviation of the prediction error $(\sigma)$. A higher $\rho$ denotes better performance while for $\sigma$, lower values represent better results. Both merits are applied on the mapped objective scores, i.e., $f(d)$. The scatter-plots for all three listening tests are shown in Fig. 6, where their corresponding figures of merit are indicated at the top of each plot. In addition, the applied mapping function $f(d)$is shown. Table II summarizes the obtained values for the free parameters of the applied mappings.

The plots clearly show good performance by means of a strong monotonic relation between STOI and the speech-intelligibility scores, for all three listening tests. This is reflected in the correlation coefficients which are all above 0.9 and the obtained standard deviations of the predictions errors, which are below 9%. It can be observed from the plots in Fig. 6 that the logistic function for the IEEE sentences is shifted more to the right compared to the mapping function for the Dantale sentences: given a STOI score, the actual intelligibility score for the IEEE sentences is slightly lower compared to the Dantale sentences. As hypothesized in Section IV-A, this is probably due to the fact that the Dantale sentences are generated from a closed set of words, which makes them more intelligible than IEEE sentences under equal adverse conditions.

### B. Analysis of Absolute Intelligibility Predictions

As already mentioned in the introduction, the aim for STOI is to have a monotonic relation with speech intelligibility and not necessarily to predict absolute intelligibility scores. However, by mapping the STOI outcomes using the logistic function $f(d)$, some insight will be gained in the distribution of the prediction errors of STOI. The results are shown in Figs. 7–9 for the three listening tests from Sections III-A–III-C, respectively.

Fig. 7 shows the results for the first listening experiment for all noise-types, SNRs and other specific ITFS-settings (similar to Fig. 3). The plots reveal that STOI correctly predicts the effect of the LC-parameter on the speech intelligibility, for almost all cases. This includes the extreme cases where essentially a noise-only signal (−60-dB SNR) is ITFS-processed, resulting in almost 100% intelligible speech for specific LC-values. Note, that for this case all fine-structure of the clean speech is lost and the signals sound rather artificial; a challenging condition.

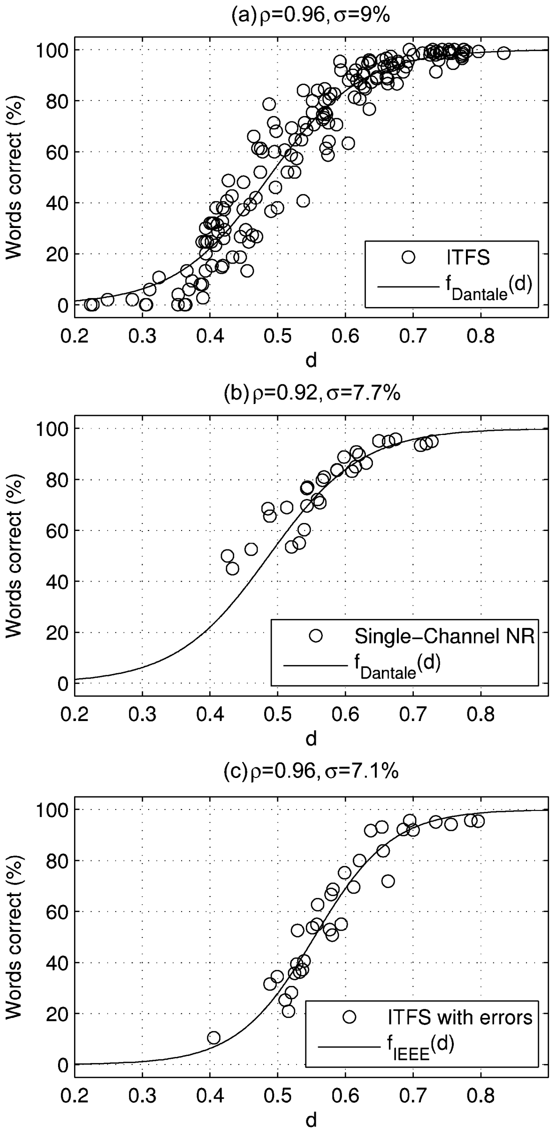


Fig. 6. Scatter plots between STOI and the speech-intelligibility scores from three different types of TF-weighted noisy speech: (a) ITFS-processed noisy speech (see Section III-A), (b) single-channel noise-reduced speech (see Section III-B), and (c) ITFS-processed noisy speech with artificially introduced errors (see Section III-C). At the top of each plot the correlation coefficient $(\rho)$ and the standard deviation of the prediction error $(\sigma)$ is denoted. (a) $\rho = 0.96$, $\sigma = 9\%$; (b) $\rho = 0.92$, $\sigma = 7.7\%$; (c) $\rho = 0.96$, $\sigma = 7.1\%$.

TABLE II
USED VALUES FOR THE FREE PARAMETERS OF THE NONLINEAR MAPPINGS, FOR THE DANTALE AND IEEE CORPUS

| | $a$ | $b$ |
|---|---|---|
| $f_{Dantale}(d)$ | -14.5435 | 7.0792 |
| $f_{IEEE}(d)$ | -17.4906 | 9.6921 |

Small problems are observed for both the bottles-noise and the car-noise mixed at 50% SRT for the unprocessed noisy speech (UN) and low SNR-corrected LC values (first, third, fourth, and seventh plot of top-row in Fig. 7). For these conditions, STOI underestimates the speech intelligibility. An explanation for this could be the fact that these noise types have a significantly different average spectrum compared to the clean speech. Therefore, the errors are distributed in different frequency channels compared to the SSN and cafe-noise conditions. Perhaps these problems can be solved by introducing band-importance functions, see, e.g., [14]. Nevertheless, these problems are rather modest and generally STOI shows good agreement with the data. Note, that STOI was developed with simplicity in mind: the goal was to develop a model with very few parameters. For this reason we did not include any band-importance functions.

The absolute predicted intelligibility scores for the single-channel noise-reduced speech are shown in Fig. 8. From this plot we observe that for low SNRs the intelligibility scores for the SSN-conditions are slightly overestimated. However, these small overestimations are approximately equal for both the unprocessed noisy condition and the noise reduction algorithms EM and SG. By comparing the relative difference in predicted intelligibility scores before and after noise reduction, it can be concluded that STOI correctly predicts no significant effect on the intelligibility. This is in line with the results from the listening test. Similarly, for the cafe-noise no significant change in intelligibility is reported. Note, that several STI-based speech-intelligibility measures report an incorrect intelligibility improvement after noise reduction, e.g., [7], [10], [11].

Fig. 9 shows the STOI predictions for the ITFS-processed speech with artificially introduced errors from Section III-C. From the plot we can observe that STOI correctly predicts the effect of the introduced errors in the IBM. Specifically, the Type-I and Type-II error predictions are in strong correspondence with the actual intelligibility scores. For the general error introduced in the IBM, the plots reveal small deviations between the different noise types, i.e., the 2-talker noise conditions are slightly overestimated and the 20-talkers noise is slightly underestimated. However, these errors turn out be small.

### C. Effect of Parameters $N$ and $\beta$

The correlation coefficients obtained for the different values of $N \in [10, 20, 30, 50, 100, 500]$ and $\beta \in [-\infty, -35, -25, -15, -10]$, with respect to the ITFS listening experiment from Section III-A, are shown in Fig. 10. From the plot it can be observed that maximum correlation is obtained with $N = 30$ and $\beta = -15$ dB. The same conclusion holds for observing the standard deviations of the prediction errors (not shown). In general, the segment length $N$ has a bigger impact on the results compared to the clipping procedure.

The results with respect to $N$ are in line with the rationale behind STOI which was explained in Section I-A. While an estimated correlation coefficient based on very long segments (tens of seconds) may be dominated by outliers, an analysis length which is too short (20–30 ms) may exclude important temporal modulation frequencies. Several listening experiments show that temporal modulations above 2–3 Hz are important for intelligibility [17], [31]. For $N = 30$, STOI will be sensitive for temporal modulations of 2.6 Hz and higher which is roughly in accordance with the results of these listening tests. Moreover, the analysis length of $N = 30$ (384 ms) is also more in line with the maximum temporal integration properties of the auditory system, which is in the order of hundreds of milliseconds, e.g., [18].

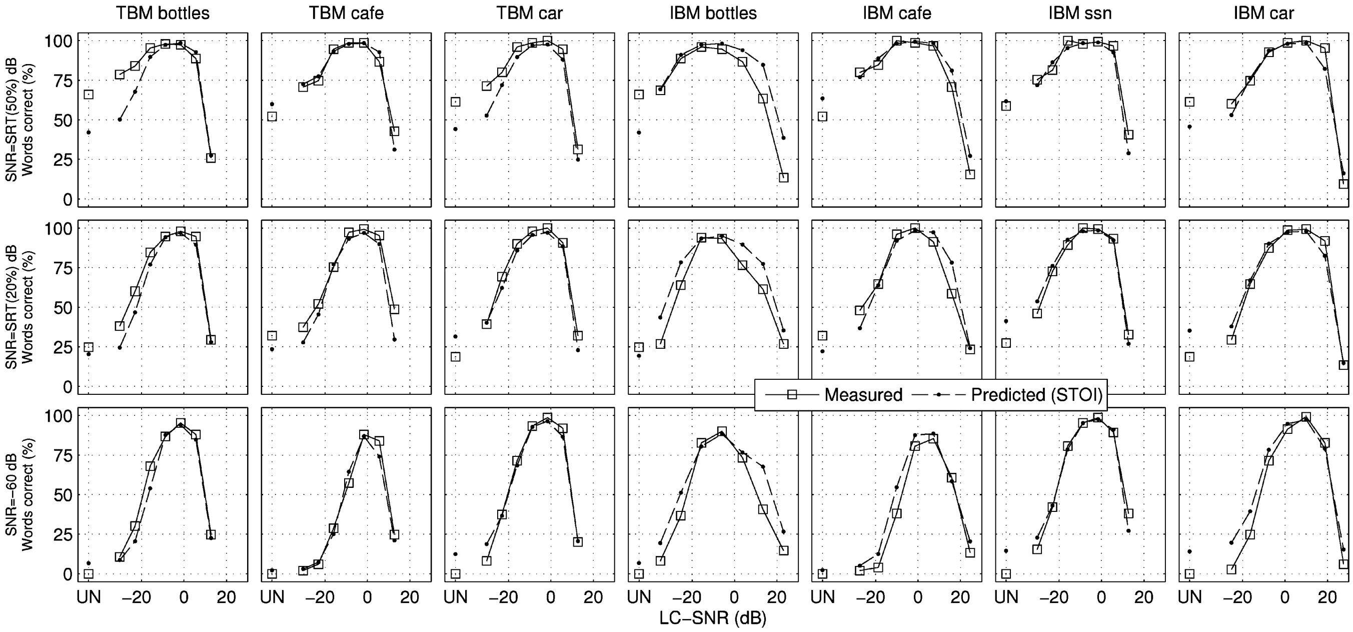


Fig. 7. Results of STOI predictions for the ITFS-processed speech and the actual measured intelligibility scores from Kjems *et al.* [19] as described in Section III-A. Each row of subplots refers to a certain SNR-value (50% SRT, 20% SRT, and −60 dB) where a column is related to a specific noise-type and ITFS-processing setting (TBM, IBM, see Section III-A for more details).

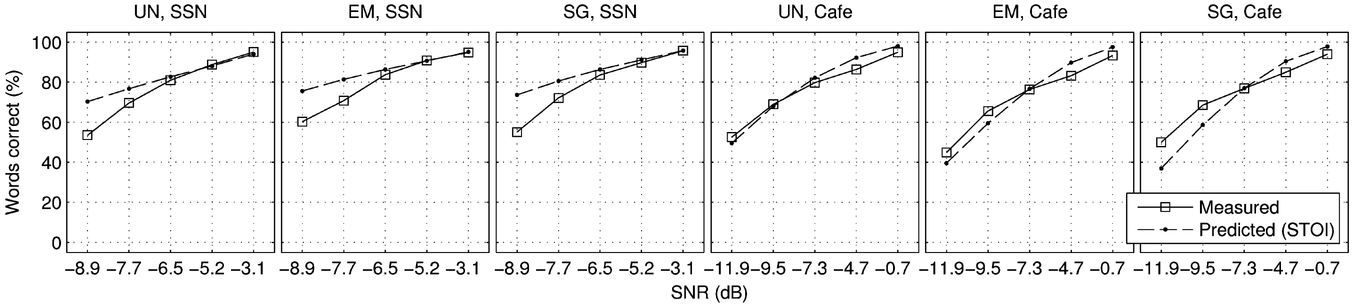


Fig. 8. STOI predictions for the single-channel noise-reduced speech conditions from Section III-B. Predicted and measured scores are shown for unprocessed noisy (UN) speech, and two noise-reduction schemes (EM, SG) for speech-shaped noise (SSN) and cafe noise.

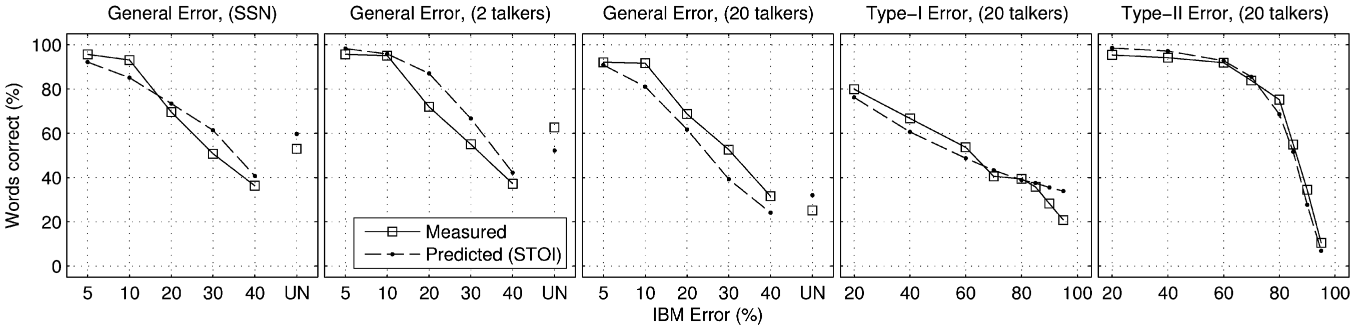


Fig. 9. Intelligibility scores of ITFS-processed speech with artificially introduced errors (replotted from Li and Loizou [20]) as described in Section III-C together with the predicted intelligibility scores from STOI.

### D. Comparison With Other Intelligibility Models

For the five reference objective measures the same evaluation procedure is used as with STOI (as described in Section IV). An additional figure of merit is included: the Kendall's Tau ($\tau$), e.g., [32], where a higher $\tau$ implies better performance. The Kendall's Tau is only based on the ranking and therefore independent of the applied mapping from model output to predicted intelligibility scores, as long as the mapping is monotonic. It is included to make the results more transparent, since the mapping procedure may show a better fit with the data for certain intelligibility models.

The results are shown in Fig. 11, where each column of subplots represents one of the three listening tests and each row represents a figure of merit. The average of the three

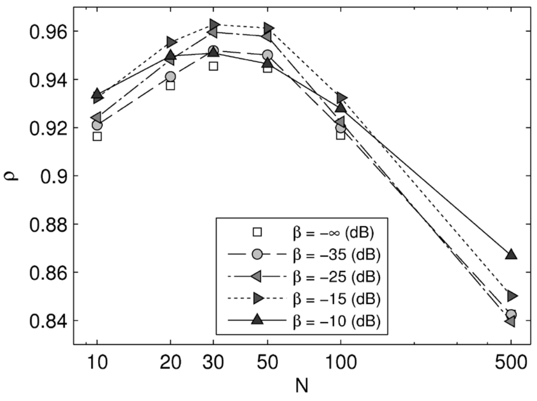


Fig. 10. Influence of the clipping parameter $\beta$ and the segment length $N$ for the intelligibility data originating from the ITFS listening experiment from Section III-A.

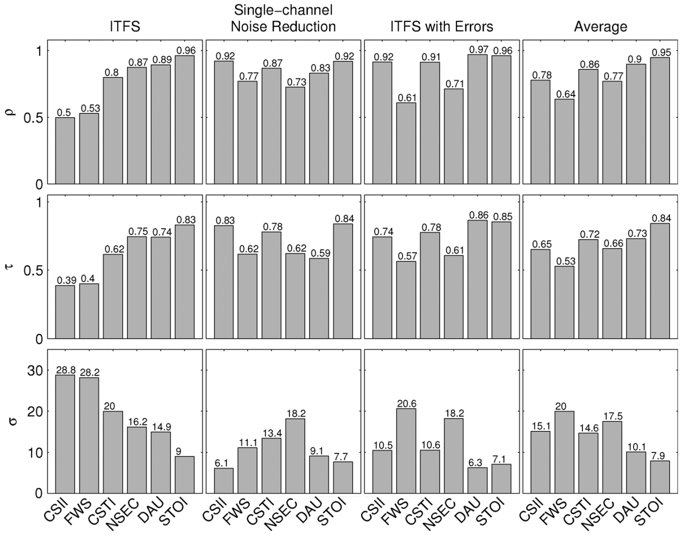


Fig. 11. Performance of STOI compared with five other reference objective intelligibility models. Each column of all subplots denotes one of the three listening tests as described in Section III, where the last row indicates the average performance measure for all three listening tests. The rows represent the correlation coefficient $(\rho)$, Kendall's Tau $(\tau)$ and the standard deviation of the prediction error $(\sigma)$.

outcomes for each merit is shown in the last column. From these results it can be concluded that STOI has the best average performance for all three listening tests with respect to all figures of merit. Also for the results with respect to each listening test independently, STOI has better performance compared to almost all other measures. Only CSII has similar performance for the "single-channel noise reduction" listening experiment and DAU shows slightly better results for the "ITFS with errors" data. Less good results were obtained with FWS which ended up lowest in ranking for the average results for all listening tests. In general, the rankings based on the correlation coefficient are roughly in accordance with the remaining two figures of merit, except for CSTI and NSEC when evaluated for the single-channel noise reduced speech. It turns out that for these two measures, the mapping function $f_{\mathrm{Dantale}}$, which was only trained on the ITFS-processed data, did not fit the noise-reduction dataset.

The good results obtained with DAU and NSEC for the first listening experiment (ITFS) are in accordance with the fact that these two measures were also designed and optimized for the ITFS listening experiment by Kjems *et al.* [19]. Furthermore, the performance and ranking of CSII, FWS and CSTI for the single-channel noise-reduction intelligibility data is in agreement with the results from Ma *et al.* [14].

## VI. Discussion

One may argue that STOI has better performance compared to the reference objective measures due to the fact that the parameters $\beta$ and $N$ have been optimized for. However, instead of extensively tuning these parameters, a limited amount of settings have been tested only with respect to the first listening test. Other settings than $\beta = -15$ dB and $N = 30$ for the last two listening experiments have not been considered. Note, that also NSEC and DAU were designed and optimized for the intelligibility data from Kjems *et al.* [19]. Furthermore, the output signals of the single-channel noise-reduction algorithms from the second listening experiment have different types of signal artifacts compared to the ITFS-processed speech. Also the ITFS-processed speech with artificially introduced errors (the third listening experiment) is based on a different speaker and different noise types. The latter two listening experiments contain a significant amount of "musical noise" due to their DFT-based approach, in contrast to the gammatone-based approach of the first listening experiment. In summary, STOI has not been optimized for listening tests 2 and 3.

Next to STOI, DAU also showed good performance for all listening tests. However, in contrast to STOI, DAU determines a correlation coefficient in segments of only 20 ms. In line with the results from Section V-C this could be a reason for their difference in performance with respect to the first two listening tests where STOI shows better performance. However, this is not in agreement with the results for the last listening experiment, where DAU shows slightly better performance than STOI. Maybe this can be explained by the use of the so-called adaptation loops in the DAU-model, which simulate the adaptation properties of the auditory nerve [28]. This stage shows a log-compressive behavior for stationary input signals while fast fluctuations are linearly transformed. As a consequence, DAU is more sensitive to transient regions which are of importance for speech intelligibility. This unique property of DAU is not represented in any of the other intelligibility models contained in this research. It would be of interest to investigate the contribution of these adaptation loops with respect to intelligibility prediction (e.g., by excluding them or replacing this stage with a simple log-transform).

Although not as good as STOI and DAU, CSTI also showed good performance with respect to all three listening tests. Note, that without the clipping procedure CSTI and STOI are similar measures in the sense that they are both based on a correlation coefficient per band. However, CSTI determines a correlation coefficient for the complete signal at once instead of the short-time segments used by STOI. In line with the earlier results shown in Fig. 10, this difference in analysis window length probably explains their difference in performance. The same holds for NSEC which also considers the correlation for the complete signal at once.

CSII showed good results for the second and third listening experiment; however, poor results were obtained with respect to predicting the intelligibility scores for the ITFS processed speech data. It was observed that CSII predicted incorrectly that all the ITFS-processed noisy speech signals mixed at $-60$-dB SNR were unintelligible. An explanation for this is the fact that CSII is sensitive for degradations in the temporal fine structure of the clean speech (in contrast to STOI). This is a direct consequence of the coherence function, which takes into account the phase component of the complex DFT coefficients. Note, that for the ITFS-processed noisy signals mixed at $-60$-dB SNR, the temporal fine structure is completely lost.

FWS is the only measure in this evaluation which is not based on a correlation-based comparison between the clean and degraded speech. Instead it uses a conventional SNR per frequency band. This property and the relatively short analysis window of 20 ms probably explains its low ranking compared to all other intelligibility models.

Although STOI is meant for predicting the intelligibility of TF-weighted noisy speech, it would be of interest to investigate its performance with respect to other types of degradations. A recent evaluative study [33] showed promising results for STOI with respect to envelope thresholding: a nonlinear operation that consists of setting to zero any samples of the original envelope that are below a threshold [7]. Also CSTI showed good results with respect to envelope thresholding [7]. As already explained, CSTI and STOI are similar in the sense that they are both based on the correlation coefficient between the temporal envelopes of the clean and degraded speech per frequency band. Goldsworthy and Greenberg concluded that with this correlation-based approach, CSTI was not capable to predict the intelligibility of reverberated speech in quiet and low noise environments [7]. It could be the case that this conclusion also holds for STOI. However, more research is needed to investigate the effect of the clipping procedure and shorter analysis window length of STOI compared to CSTI. Note, that STOI does work well for additive noise since each of the three different listening tests contain unprocessed noisy speech for different noise types and SNRs.

STOI does not take into account some type of absolute threshold in quiet. Therefore, its predictions may not be accurate for operations which significantly reduce the level per band and do not have a strong impact on its temporal envelope (e.g., as with low-pass or high-pass filtering).

## VII. Conclusion

A short-time objective intelligibility measure (STOI) is presented based on the correlation between temporal envelopes of the clean and degraded speech in short-time (382 ms) segments. This is different from other measures, which typically consider the complete signal at once, or use a very short analysis length (20–30 ms). Experiments with different segment lengths indeed show the benefit by using segment-lengths in the order of hundreds of milliseconds. Further extensive evaluation shows that STOI has high correlation with the speech intelligibility for three different listening tests ($\rho \geq 0.92$ for all listening tests). For each of these three listening tests, noisy speech is processed by some type of TF-varying gain function, including a signal processing technique called "ideal time frequency segregation" and conventional single-channel noise reduction algorithms. In general, STOI showed better correlation with speech intelligibility compared to five other reference objective intelligibility models. A free Matlab implementation is provided at http://siplab.tudelft.nl/.

## Acknowledgment

The authors would like to thank Dr. U. Kjems for providing the data from Section III-A and Prof. P. C. Loizou for additional help on generating the signals from Section III-C.

## References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[2] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[3] *Methods for Calculation of the Speech Intelligibility Index*, S3.5-1997, ANSI, New York, 1997.

[4] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, 1980.

[5] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[6] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[7] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[8] P. C. Loizou*, Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[9] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[10] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method—Comparison between observed scores and scores predicted from STI," *Scand. Audiol. Supplement.*, vol. 38, pp. 50–55, 1993.

[11] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3937–3946, 2008.

[12] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

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

[14] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, 2009.

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

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

[17] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2670–2680, 1994.

[18] G. van den Brink, "Detection of tone pulse of various durations in noise of various bandwidths," *J. Acoust. Soc. Amer.*, vol. 36, no. 6, pp. 1206–1211, 1964.

[19] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[20] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, p. 1673, 2008.

[21] K. Wagener, J. L. Josvassen, and R. Ardenkjaer, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, 2003.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[23] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[25] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4266–4269.

[26] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[27] Y. Hu and P. C. Loizou, "Techniques for estimating the ideal binary mask," in *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, 2008.

[28] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[29] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3623–3631, 1996.

[30] R. Koch, "Auditory sound analysis for the prediction and improvement of speech intelligibility (in German)," Ph.D. dissertation, Universität Göttingen, Göttingen, Germany, 1992.

[31] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2783–2791, 1999.

[32] D. J. Sheskin*, Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition*. Boca Raton, FL: Chapman & Hall/CRC, 2004.

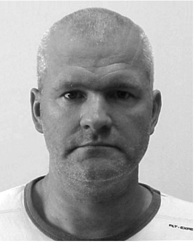
[33] A. Schlesinger and M. M. Boone, "The characterization of the relative information content by spectral features for the objective intelligibility assessment of nonlinearly processed speech," in *Proc. Interspeech*, 2010, pp. 1309–1312.

**Cees H. Taal** received the B.S. and M.A. degrees in arts and technology from the Utrecht School of Arts, Utrecht, The Netherlands, in 2004 and the M.Sc. degree in media and knowledge engineering from Delft University of Technology, Delft, The Netherlands, in 2007. He is currently pursuing the Ph.D. degree at the Signal and Information Processing Lab, Delft University of Technology, under the supervision of R. Heusdens and R. Hendriks in collaboration with Oticon A/S.

His main research topic is intelligibility enhancement of single-channel noisy speech. Other research interests include auditory modeling and applications thereof in the field of digital audio and speech processing.

**Richard C. Hendriks** received the B.Sc., M.Sc., (*cum laude*), and Ph. D. (cum laude) degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively.

From 2003 to 2007, he was a Ph.D. Researcher at Delft University of Technology. From 2007 to 2010, he was a Postdoctoral Researcher at Delft University of Technology. Since 2010, he has been an Assistant Professor in the Multimedia Signal Processing Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. From March 2008 to March 2009, he was a Visiting Researcher at Oticon A/S, Copenhagen, Denmark. His main research interests are digital speech and audio processing, including single-channel and multi-channel acoustical noise reduction, speech enhancement, and intelligibility improvement.



**Richard Heusdens** received the M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively.

Since 2002, he has been an Associate Professor in the Department of Mediamatics, Delft University of Technology. In the spring of 1992, he joined the digital signal processing group at Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio and speech processing activities within the ICT group. He held visiting positions at KTH (Royal Institute of Technology, Stockholm, Sweden) in 2002 and 2008. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, signal processing for digital hearing aids, distributed signal processing, and sensor networks.

**Jesper Jensen** received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively.

From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is with Oticon A/S, Smørum, Denmark. His main research interests are in the areas of acoustical signal processing, including signal retrieval from noisy observations, coding, speech and audio synthesis, signal processing for hearing aids, intelligibility enhancement of speech signals, and perceptual aspects of signal processing.