# Speech Intelligibility Prediction Based on Mutual Information

Jensen, Jesper; Taal, Cees H.

*Published in:*
I E E E Transactions on Audio, Speech and Language Processing

*DOI (link to publication from Publisher):*
10.1109/TASLP.2013.2295914

*Publication date:*
2014

*Document Version*
Early version, also known as pre-print

*Citation for published version (APA):*
Jensen, J., & Taal, C. H. (2014). Speech Intelligibility Prediction Based on Mutual Information. *I E E E Transactions on Audio, Speech and Language Processing*, *22*(2), 430-440.
https://doi.org/10.1109/TASLP.2013.2295914

# Speech Intelligibility Prediction Based on Mutual Information

Jesper Jensen and Cees H. Taal

***Abstract*—This paper deals with the problem of predicting the average intelligibility of noisy and potentially processed speech signals, as observed by a group of normal hearing listeners. We propose a model which performs this prediction based on the hypothesis that intelligibility is monotonically related to the mutual information between critical-band amplitude envelopes of the clean signal and the corresponding noisy/processed signal. The resulting intelligibility predictor turns out to be a simple function of the mean-square error (mse) that arises when estimating a clean critical-band amplitude using a minimum mean-square error (mmse) estimator based on the noisy/processed amplitude. The proposed model predicts that speech intelligibility cannot be improved by any processing of noisy critical-band amplitudes. Furthermore, the proposed intelligibility predictor performs well ($\rho > 0.95$) in predicting the intelligibility of speech signals contaminated by additive noise and potentially non-linearly processed using time-frequency weighting.**

***Index Terms*—Instrumental measures, noise reduction, objective distortion measures, speech enhancement, speech intelligibility prediction.**

## I. Introduction

MONAURAL speech intelligibility prediction methods aim at predicting the average intelligibility of noisy and/processed speech signals, as judged by a group of listeners. Our motivation for studying speech intelligibility predictors is twofold. Firstly, reliable intelligibility predictors are of great practical importance, e.g., in guiding the development process of speech processing algorithms, and replacing costly listening tests in early stages of the development phase. Secondly, the development and study of intelligibility predictors may lead to a better understanding of the mechanism behind human intelligibility capabilities.

Historically, two main branches of intelligibility predictors may be identified: methods based on the *Articulation Index (AI)* [1], proposed first by French and Steinberg [2] and later refined by Kryter [3], and the *Speech Transmission Index (STI)* [4] proposed by Steeneken and Houtgast [5].

The basic AI approach assumes that intelligibility is a function of the speech information available to the listener across several frequency bands, each of which carries an independent contribution to the total intelligibility. Assuming that speech and masker signals are available in isolation, effective signal-to-noise ratios (SNRs) are computed for each frequency band; the SNRs are then limited to a certain pre-specified SNR range, normalized to a value between 0 and 1, and combined as a perceptually weighted average. The AI approach has later been refined further and standardized as the *Speech Intelligibility Index (SII)* [6]. AI and SII are based on long-term spectra of speech and masker and therefore may be inaccurate for fluctuating maskers. To reduce this problem, Rhebergen proposed the *Extended SII* [7], which divides the speech and masker signal into short-time frames (9–20 ms), computes the instantaneous SII value for each frame, and then averages the per-frame SII values to find a final intelligibility prediction. Another extension of the SII is the *Coherence SII (CSII)* which was proposed to better take into account non-linear distortions such as peak- and center-clipping [8].

The AI based methods described above were originally formulated with the focus on simple linear degradations, e.g., linear filtering and additive, uncorrelated noise. The *Speech Transmission Index (STI)* [5], [9] extends the type of degradations to convolutive noise sources, such as reverberance and the effects of room acoustics. The STI is based on changes in the modulation transfer function. Specifically, STI relies on the observation that reverberation and additive noise tends to reduce the depth of the temporal amplitude/intensity modulations compared to the clean reference signal. Originally, STI used synthetic bandpass filtered probe signals with various acoustic center frequencies, intensity-modulated with a range of low-frequency sinusoidal modulators, whose frequencies were chosen in the range $f = 0.63$ Hz to $f = 12.7$ Hz to emulate the dominating modulation frequencies present in human speech. Later, in an attempt to better take into account the effects of various non-linear processing operations, such as dynamic amplitude compression [10], and envelope clipping [11], the class of *speech STI (sSTI)* methods were introduced [12] which replaced the artificial probe signals by actual speech signals. More recently, Jørgensen and Dau presented a speech intelligibility prediction model based on the envelope power signal-to-noise ratio $\mathrm{SNR}_{\mathrm{env}}$ at the output of a modulation filter bank [13]. This model showed promising results for noisy speech subjected to reverberation and spectral subtraction, but has only been evaluated for stationary speech-shaped noise.

The AI and STI based intelligibility predictors considered as a whole are suitable for a range of distortion types including additive noise, convolutive noise, filtering, and clipping, but they are less suited for speech signals distorted by non-stationary

Manuscript received April 02, 2013; revised September 09, 2013; accepted December 01, 2013. Date of publication January 02, 2014; date of current version January 08, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Brian Kingsbury.

J. Jensen is with Oticon A/S, 2765 Smørum, Denmark, and also with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: jsj@oticon.dk; jje@es.aau.dk).

C. H. Taal was with the ENT Department, Leiden University Medical Center, The Netherlands. He is now with Applied Sensor Technologies, Philips Research, 5656 AE Eindhoven, The Netherlands (e-mail: cees.taal@philips.com).

Digital Object Identifier 10.1109/TASLP.2013.2295914

noise sources and processed by time-varying and non-linear filtering systems such as those typically used in single-channel speech enhancement systems [14], [15]. To better take this type of distortions into account, new intelligibility predictors were proposed such as the method of Christiansen and Dau [16], and the *Short-Time Objective Intelligibility (STOI)* measure [17] by Taal *et al.* STOI shows similarities to the speech-based STI methods [12] in that speech envelopes extracted with band-pass filters are compared; however, unlike most variants of the speech-based STI methods which are based on long-term statistics, STOI compares the envelopes via *short-term* measures.

In this study we constrain ourselves to monaural intelligibility prediction, that is, only one realization of the noisy/processed signal and the clean reference is available. We further assume that the noise is additive but not necessarily stationary, and we consider processing methods, which can be described in a time-frequency analysis-modification-synthesis framework, e.g. [18]: in the analysis stage the signal is decomposed into time-frequency units, typically using a short-time band-pass filter bank, in the modification stage gain factors are multiplied onto the time-frequency units, and in the synthesis stage the modified time-frequency units are used to reconstruct processed time-domain signals. Since the gain factors are not necessarily constant across time and generally depend on short-term signal characteristics, the resulting processing may be time-varying and non-linear.

The proposed intelligibility prediction model makes use of basic information theoretic tools such as entropy and mutual information [19]. It appears natural to use tools developed to characterize information transmission. After all, the speech communication process can be viewed as the process of transmitting a speech signal across a time-varying, non-linear channel (the acoustic channel, the auditory periphery, and the higher stages of the auditory pathway) to reach the brain of the receiver; see also the work of Allen [20] who observed that the expression for the AI shows strong similarities to the expression for the capacity of a memoryless Gaussian channel, and the work of Leijon [21] who studied the potential relationship between AI (and SII) and the acoustic-to-auditory information rate. More specifically, the basic idea of the proposed method is to compare the critical-band amplitude envelopes of the clean and noisy/processed signal and estimate the intelligibility of the noisy/processed signal based on this comparison. In particular, we assume that the clean critical-band envelopes contain all information relevant for speech intelligibility, and consider the question: how much information (measured in bits) about the clean envelopes can be extracted, on average, by observing the envelopes of the noisy/processed signal? If the noisy/processed envelopes provide no information whatsoever, i.e., the mutual information between clean and noisy/processed envelopes is zero bits, then we expect the intelligibility of the noisy/processed signal to be zero. If, on the other hand, the noisy/processed envelopes provide much information about the clean envelopes, we expect the intelligibility of the noisy/processed input signal to be high.

The proposed intelligibility prediction model shares characteristics with the method proposed in [22], although the motivation for the proposed model is quite different. Specifically, the proposed model arises as a consequence of describing speech information transmission in a simple model of the auditory periphery, whereas the method in [22] has a more heuristic foundation in that it replaces the linear correlation operation used in the STOI model with a generalization, namely mutual information. Furthermore, the proposed model employs a short-term stationary signal model, whereas the method in [22] assumes the clean and noisy/processed speech signals to be realizations of *strictly* (long-term) stationary stochastic processes. Finally, the proposed model relies on lower bounds of mutual information, leading to simple equations in terms of second-order statistics, whereas the method in [22] estimates mutual information, which generally involves higher-order statistics.

The proposed intelligibility prediction model also bears some similarities to the STOI model [17], as it compares critical-band amplitude envelopes in terms of the linear correlation coefficient. However, whereas the use of linear correlation in STOI has a heuristic foundation, it follows in the proposed model as a consequence of the assumed signal model and a crude model of the auditory periphery; in this sense, the proposed model might be seen as a better motivated model.

With the proposed model, we have aimed at simplicity. For example, the proposed model does not make use of band importance functions to emphasize certain critical bands over others. Instead, each critical band contributes equally to intelligibility. In fact, from the information theoretical path followed in this paper, band importance functions are hard to justify. Furthermore, the proposed model appears to work well without (see Section V). If one would introduce band-importance functions or other additional free parameters, which model aspects not taken into account by the model, we expect performance to increase.

The article is organized as follows. In the following section we introduce the basic auditory model used in the proposed method. Section III derives the proposed mutual information lower bound. Section IV presents implementational details and discusses the numerical values of the few free parameters of the proposed method. In Section V the proposed intelligibility predictor is compared to intelligibility predictors from the literature for several noise sources and processing conditions. Finally, Section VI concludes the work.

## II. Auditory Front-End and Notation

We consider a crude signal processing model of the auditory periphery, which is similar in structure to front-ends used in speech enhancement [23], automatic speech recognition [24], and intelligibility predictors [17]. The model consists of a band-pass filter bank simulating the bandpass filtering characteristics of the cochlea, and a full-wave rectification, which simulates coarsely the mechanism of the hair cell transduction in the inner ear. The resulting "inner representations" are rough abstractions of the signal transmitted via the auditory nerve to the higher stages of the auditory system.

The model is shown in more detail in Fig. 1. We use capital letters to denote random processes and variables and lower-case letters to denote the corresponding realizations. Let $S(n)$ and $X(n)$ denote random processes modeling a clean speech input

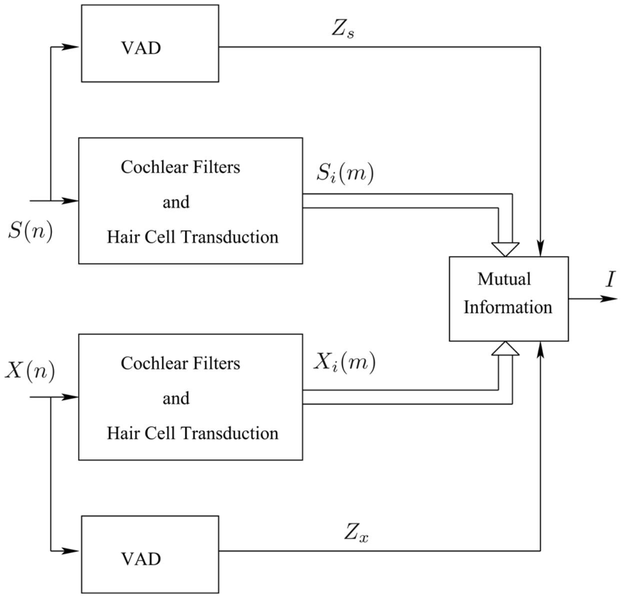

VAD
$Z_s$
Cochlear Filters and Hair Cell Transduction
$S_i(m)$
$S(n)$
Mutual Information
$I$
$X(n)$
Cochlear Filters and Hair Cell Transduction
$X_i(m)$
$Z_x$
VAD

Fig. 1. Proposed intelligibility prediction scheme. It is assumed that the critical-band amplitude envelopes $S_i(m)$ of the clean speech signal contain all information relevant for speech intelligibility. The intelligibility of the noisy/processed signal $X(n)$ is estimated as the average information in its critical-band envelopes $X_i(m)$ about the clean amplitude envelopes $S_i(m)$.

signal, and the corresponding noisy/processed signal, respectively. Band pass filtered signals are obtained by dividing the time-domain input signals into successive, overlapping analysis frames, applying an analysis window, and transforming these time-domain frames to the frequency domain using a Discrete Fourier Transform. The resulting DFT coefficients are given by

$$\tilde{S}(k,m) = \sum_{n=0}^{N-1} S(mD+n)w(n)e^{-j2\pi kn/N},$$

and

$$\tilde{X}(k,m) = \sum_{n=0}^{N-1} X(mD+n)w(n)e^{-j2\pi kn/N},$$

respectively, where $k$ and $m$ denote the frequency bin index and the frame index, respectively, $D$ is the frame shift in samples, $N$ is the frame length in samples, and $w(n)$ is the analysis window. Then, a one-third octave band analysis is performed by grouping DFT bins, resulting in critical-band amplitudes

$$S_i(m) = \sqrt{\sum_{k\in CB_i} |\tilde{S}(k,m)|^2}, \tag{1}$$

and

$$X_i(m) = \sqrt{\sum_{k\in CB_i} |\tilde{X}(k,m)|^2},$$

respectively, where $CB_i$ is the frequency index set representing the $i$th one-third octave band, $i = 1,\ldots,L$. The amplitude envelope signals $S_i(m)$ and $X_i(m)$ are also random processes.

The Voice Activity Detection (VAD) blocks are used to identify and exclude low-energy frames from the computation. Signal $S(n)$ typically contains low-energy frames, e.g. silence regions, which certainly do not contribute to speech intelligibility and therefore can be excluded from the mutual information computation. For this reason a simple energy-based per-frame VAD (details are given in Section IV) is applied to $S(n)$ resulting in the frame index set $Z_s$ of speech active frames. In an identical manner, the VAD in the lower branch identifies low- and high-energy frames in $X(n)$. The low-energy frames are typically i) noise-only (i.e., silence) frames, or ii) they occur due to certain types of aggressive processing which essentially suppress entire signal frames, which *do* carry speech information. The high-energy frames are represented by the frame index set $Z_x$.

Let $M$ denote the number of frames in a given speech sentence, and let

$$\mathcal{X} = [X_1(1)X_2(1)\cdots X_L(1)X_1(2)\cdots X_L(M)]^T$$

and

$$\mathcal{S} = [S_1(1)S_2(1)\cdots S_L(1)S_1(2)\cdots S_L(M)]^T$$

denote random super vectors, formed by stacking critical-band amplitude spectra for successive frames. We are interested in the average mutual information (to be defined exactly below) between clean and noisy/processed critical-band amplitudes, i.e., $\frac{1}{L|Z_s|}I(\mathcal{S};\mathcal{X})$, where $|\cdot|$ denotes set cardinality, and $L|Z_s|$ estimates the number of speech-active critical-band amplitudes in the clean signal. Assuming that the entries in each super vector are statistically independent, an assumption which is routinely made in the area of speech enhancement[1], it is easy to verify that the mutual information $I(\mathcal{S};\mathcal{X})$ decomposes into a summation of mutual information $I(S_i(m);X_i(m))$ terms,

$$\begin{aligned}\frac{1}{L|Z_s|}I(\mathcal{S};\mathcal{X}) &= \frac{1}{L|Z_s|}\sum_m \sum_{i=1}^{L} I(S_i(m);X_i(m)) \\ &= \frac{1}{L|Z_s|}\sum_{m\in Z_s\cap Z_x} \sum_{i=1}^{L} I(S_i(m);X_i(m)).\end{aligned}$$

The second equation follows because summation over the frame index set $m \in Z_s \cap Z_x$, where both signals $S(n)$ and $X(n)$ are speech active, excludes $I(S_i(m);X_i(m))$ terms which are all zero. Specifically, silence frames in $S(n)$ are excluded, and speech information loss due to over-suppressed frames in $X(n)$ is taken into account (that is, all $I(S_i(m);X_i(m))$ terms in such frames are set to zero). An alternative, and perhaps physiologically more plausible, implementation of the described VAD function is to replace the VAD blocks by additive, uncorrelated internal noise sources, see e.g. [27].

For notational convenience, we skip in the following the subband and frame index where possible, and simply replace $S_i(m)$ and $X_i(m)$ by $S$ and $X$, respectively. The mutual information $I(S;X)$ between clean and noisy/processed critical-band amplitudes is given by [19]

$$I(S;X) = h(S) - h(S|X),$$

[1] This assumption is approximately valid if the frame size $N$ is large compared to the correlation time of the signals in question [25], [26].

where the differential entropy[2] of $S$ is

$$h(S) = \int_S f_S(s) \ln f_S(s) ds,$$

and the conditional differential entropy $h(S|X)$ is

$$h(S|X) = \int_X \int_S f_{S,X}(s,x) \ln f_{S|x}(s|x) ds dx. \tag{2}$$

For certain simple situations, the joint probability density function (pdf) $f_{S,X}(s,x)$ may be given and the conditional differential entropy $h(S|X)$ might be derived analytically. However, in general, since the exact processing leading to $X$ may be complicated or even unknown, deriving or estimating from limited data the joint pdf $f_{S,X}(s,x)$ needed to compute $h(S|X)$ is difficult at best. Instead, to circumvent this difficulty, we propose to lower bound the mutual information $I(S;X)$; as we show in the following, this requires only second-order statistics of $f_{S,X}(s,x)$.

## III. Lower Bounds on $I(S;X)$

We derive lower bounds on the mutual information $I(S;X)$ by upper bounding the conditional entropy $h(S|X)$, see e.g. the work of Bialek *et al.* [28] for another application of this procedure.

### A. Upper Bounds on $h(S|X)$

From the expression in Eq. (2) for the conditional entropy $h(S|X)$, it follows that

$$\begin{aligned} h(S|X) &= -\int_{x \geq 0} f_X(x) \int_{s \geq 0} f_{S|x}(s|x) \ln f_{S|x}(s|x) ds dx \\ &\leq \int_{x \geq 0} f_X(x) \frac{1}{2} \ln 2\pi e \sigma^2_{S|x} dx \\ &\triangleq E_X\left(\frac{1}{2} \ln 2\pi e \sigma^2_{S|x}\right) \\ &\leq \frac{1}{2} \ln 2\pi e E_X\left(\sigma^2_{S|x}\right). \end{aligned} \tag{3}$$

The first inequality holds because the maximum entropy pdf for a non-negative random variable $Y$ with a given variance $\sigma^2_Y$ approaches a Gaussian pdf for large means, which has differential entropy $h(Y) = \frac{1}{2} \ln 2\pi e \sigma^2_Y$.[3] The second inequality follows from Jensen's inequality [19, Thm. 2.6.2] and the fact that $\ln(\cdot)$ is concave.

The quantity $\sigma^2_{S|x}$ is the variance of a non-negative random variable distributed according to the pdf $f_{S|x}(s|x)$. Let $\mu_{S|x} = \int_y y f_{S|x}(y|x) dy$ denote the mean of this variable. Then,

$$\begin{aligned} \sigma^2_{S|x} &= \int_{y \geq 0} (y - \mu_{S|x})^2 f_{S|x}(y|x) dy \\ &= \int_{y \geq 0} (y - \hat{s}_{mmse}(x))^2 f_{S|x}(y|x) dy \\ &\triangleq D_{mmse}(x). \end{aligned} \tag{4}$$

The second equation follows by recalling that the conditional mean $\mu_{S|x}$ is identical to the minimum mean-square error (mmse) estimator $\hat{s}_{mmse}(x)$ of the clean random variable $S$ upon observing the noisy and/or processed realization $x$. So, it is clear that $\sigma^2_{S|x} \triangleq D_{mmse}(x)$ is nothing more than the mean-square error (mse) resulting from estimating $S$ upon observing $x$, using an mmse estimator.

Let $D_{mmse}$ denote $D_{mmse}(x)$ averaged across all realizations of the noisy/processed critical-band amplitude $x$, that is

$$D_{mmse} = \int_{x \geq 0} f_X(x) D_{mmse}(x) dx. \tag{5}$$

Inserting Eq. (4) in Eq. (3) and using Eq. (5), we arrive at

$$\begin{aligned} h_{mmse}(S|X) &\triangleq \frac{1}{2} \ln 2\pi e D_{mmse} \\ &\geq h(S|X). \end{aligned}$$

To find $D_{mmse}$ we must form the mmse estimator $\mu_{S|x}$ and average the resulting mse across realizations $x$ of the noisy and/or processed critical-band amplitudes. Finding closed-form expressions for $E(S|x)$ generally requires knowledge (or assumption) of the joint pdf $f_{S,X}(s,x)$. This has been a central topic in the area of single-channel speech enhancement over the last decades, for the case where a clean speech signal is contaminated by additive and independent noise; so, in this special case, it might be possible to derive closed-form expressions for $D_{mmse}$, and $h_{mmse}(S|X)$ could be evaluated. However, for the more general situation considered in this paper, the observations $x$ may be a result of some, potentially unknown, processing applied to the noisy observations, so that the joint pdf $f_{S,X}(s,x)$ would certainly be unknown, and estimating $D_{mmse}$ reliably from limited observations would be difficult.

To circumvent this practical difficulty, observe that replacing the conditional mean estimator $\hat{s}_{mmse}(x) = E(S|x)$ with the *linear* mmse estimator $\hat{s}_{lmmse}(x)$, leads to an mse of

$$\begin{aligned} D_{lmmse}(x) &= \int_{y \geq 0} (y - \hat{s}_{lmmse}(x))^2 f_{S|x}(y|x) dy \\ &\geq D_{mmse}(x), \end{aligned}$$

with equality for jointly Gaussian $(S, X)$, and

$$\begin{aligned} D_{lmmse} &\triangleq \int_{x \geq 0} f_X(x) D_{lmmse}(x) dx \\ &\geq D_{mmse}. \end{aligned} \tag{6}$$

It therefore follows that a looser upper bound on the conditional differential entropy $h(S|X)$ based on linear mmse estimators is given by

$$\begin{aligned} h_{lmmse}(S|X) &\triangleq \frac{1}{2} \ln 2\pi e D_{lmmse} \\ &\geq h_{mmse}(S|X). \end{aligned} \tag{7}$$

The quantity $D_{lmmse}$ is a function of second-order statistics, rather than the joint pdf $f_{S,X}(s,x)$.

To derive an expression for $D_{lmmse}$, let $\mu_S = E_S(S)$ and $\mu_X = E_X(X)$ denote expected values of $S$ and $X$, respectively, and let $r_{SX} = E_{S,X}(SX)$, $\sigma^2_S = E_S(S^2) - \mu^2_S$ and $\sigma^2_X = E_X(X^2) - \mu^2_X$ denote the cross-correlation between $S$ and $X$,

[2] For notational convenience, we use the natural logarithm and express all entropies in the units of nats, throughout. One nat corresponds to 1/ ln 2 =1.44 bits [19].

[3] This inequality is often used for random variables with double-infinite support (∞;∞) [19]. However, it also holds for non-negative random variables $Y \in [0; \infty)$; in this case the maximum entropy pdf is a truncated Gaussian whose entropy approaches from below that of a Gaussian for large means [29].

the variance of $S$, and the variance of $X$, respectively. The linear mmse (lmmse) estimator is then given by (e.g., [30]),

$$\hat{s}_{lmmse}(x) = gx + b, \quad g, b \in \mathfrak{R}, \tag{8}$$

with

$$\begin{aligned} g &= \frac{r_{SX}(SX) - \mu_S \mu_X}{\sigma_X^2}, \\ b &= \mu_S - g\mu_X. \end{aligned}$$

Inserting Eq. (8) in Eq. (6) we get

$$D_{lmmse} = \sigma_S^2 \left( 1 - \frac{(r_{SX} - \mu_S \mu_X)^2}{\sigma_S^2 \sigma_X^2} \right). \tag{9}$$

With the derived upper bounds on $h(S|X)$ we have the following lower bounds on the mutual information $I(S;X)$,

$$\begin{aligned} I_{LB,mmse}(S;X) &\triangleq \max\{h(S) - h_{mmse}(S|X), 0\}, \\ I_{LB,lmmse}(S;X) &\triangleq \max\{h(S) - h_{lmmse}(S|X), 0\}, \end{aligned} \tag{10}$$

and

$$I_{LB,lmmse}(S;X) \leq I_{LB,mmse}(S;X) \leq I(S;X).$$

### B. *Differential Entropy $h(S)$*

The bounds discussed in this paper are functions of the entropy $h(S)$ of the clean speech critical-band amplitudes. To derive an expression for this quantity, we note that when the frame size $N$ is large compared to the correlation time of the clean signals $s(n)$, then the real and imaginary parts of the DFT coefficients $\tilde{S}(k,m)$ can be considered independent and can be modeled as zero-mean Gaussian variables [25], and e.g. [26]. Assuming further that the DFT coefficients within the same critical band $\tilde{S}(k,m)$, $k \in CB_i$ are identically distributed (that is, the speech power spectral density is constant), then $S_i(m)$ given in Eq. (1) is a (scaled) chi-distributed random variable with $k' = 2|CB_i|$ degrees of freedom.

To derive an expression for $h(S)$, note first that in the special case when the real and imaginary parts of $\tilde{S}(k,m)$, $k \in CB_i$ are zero-mean, *unit-variance* Gaussians, then the corresponding critical-band amplitude, $Z$, has an expected value of

$$E(Z) = \sqrt{2}\frac{\Gamma((k'+1)/2)}{\Gamma(k'/2)},$$

a variance of

$$\sigma_Z^2 = k' - E(Z)^2,$$

and a differential entropy given by [19, Table 16.1]

$$h(Z) = \ln \Gamma(k'/2) + \frac{1}{2}(k' - \ln 2 - (k'-1)\Psi(k'/2)),$$

where $\Gamma(\cdot)$ and $\Psi(\cdot)$ denote the Gamma and the digamma function, respectively.

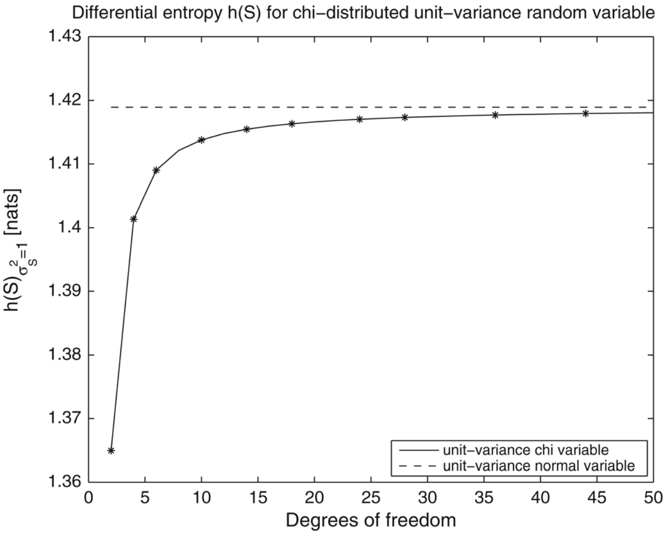


Fig. 2. Differential entropy of unit-variance chi-distributed variable as a function of degrees of freedom. The degrees of freedom corresponding to $k' = 2|CB_i|$ are marked with $*$. The dashed line indicates differential entropy of a unit-variance Gaussian.

In the general case, where the real and imaginary parts of $\tilde{S}(k,m)$, $k \in CB_i$ are not unit-variance, the differential entropy of the corresponding critical-band amplitude is

$$h(S) = h(Z) - \frac{1}{2}\ln \sigma_Z^2 + \frac{1}{2}\ln \sigma_S^2, \tag{11}$$

where we used the fact that [19]

$$h(cY) = h(Y) + \ln |c|,$$

for any random variable $Y$ and constant $c$. Thus, the differential entropy $h(S)$ is a simple function of the variance $\sigma_S^2$ of the critical-band amplitudes, because the two first terms in Eq. (11) are functions only of the number of degrees of freedom $k'$ and can therefore be computed offline. Fig. 2 plots $h(S)$ as a function of the number of degrees of freedom for the case where the critical-band amplitude variance is unity, $\sigma_S^2 = 1$. For comparison, the differential entropy of a unit-variance Gaussian, $\frac{1}{2}\ln 2\pi e \approx 1.4189$, is included. Clearly, $h(S)$ is close to the upper bound except for the lowest frequency critical bands where $k' = 2$.

It can be argued that estimation of $h(S)$ in the present context is not crucial. Specifically, if the main interest is to determine how a given type of processing leading to the processed signal $x_1$ compares to another type of processing leading to a processed signal $x_2$ in terms of intelligibility, then we are interested in determining whether $I(S;X_1) < I(S;X_2)$. As $h(S)$ appears on both sides of this inequality sign, the exact value of $h(S)$ becomes irrelevant.

Inserting Eqs. (11) and (7) in the middle line of Eq. (10), we find the following lower bound on mutual information

$$\begin{aligned} I_{LB,lmmse}(S;X) = \max \Bigg\{ & h(Z) - \frac{1}{2}\ln \sigma_Z^2 - \frac{1}{2}\ln 2\pi e \\ & + \frac{1}{2}\ln \frac{\sigma_S^2}{D_{lmmse}}, 0 \Bigg\} \text{ [nats]}. \end{aligned} \tag{12}$$

### C. Observations

*Remark 1:* In the special case when $X$ and $S$ are statistically independent, then $\hat{s}_{lmmse}(x) = E_S(S)$, $D_{lmmse} = \sigma_S^2$, and the last term in Eq. (12) is zero. In this case, the sum of the first three terms is negative (the difference between the solid and the dashed curve in Fig. 2), but the max-operator ensures $I_{LB,lmmse}(S;X) = 0$ as expected.

*Remark 2:* The expression is scaling-invariant, that is,

$$\begin{aligned} I_{LB,lmmse}(S;X) &= I_{LB,lmmse}(cS;X) \\ &= I_{LB,lmmse}(S;cX), \forall c > 0. \end{aligned}$$

This is typically a desirable property, e.g. when the processing leading to $X(n)$ (which may be unknown) reduces the general signal level significantly. However, the proposed model does not take masking effects into account: if a given spectral component is suppressed below the masking threshold, either due to spread of masking or the threshold in quiet, it would still (erroneously) contribute positively to speech intelligibility according to the proposed model. For the processing types considered in the simulation experiment, however, this potential weakness does not appear to play a big role.

*Remark 3:* From Eq. (12) it follows that

$$\begin{aligned} I_{LB,lmmse}(S;X) &\approx \frac{1}{2} \ln \frac{\sigma_S^2}{D_{lmmse}} \\ &= \frac{1}{2} \ln \frac{E_S(S^2) - \mu_S^2}{E_{S,X}(S - \hat{S}_{lmmse}(X))^2} \geq 0 \end{aligned} \tag{13}$$

because the sum of the first three terms is close to 0. Since the denominator is a mse arising from estimating a clean quantity $S$ based on a noisy/processed observation $X$, $I_{LB,lmmse}$ may be recognized as an "SNR-type" of measure [18]. In fact, it resembles closely the frequency-based segmental SNR *fwSNRseg* with constant band-importance functions ($B_j = 1/K$)[4], see [18, p. 504] [31]. However, whereas *fwSNRseg* in this case would be interpreted as a predictor of the *quality* of a noisy signal enhanced by an lmmse-estimator $s_{lmmse}(x)$ [31], the developed theory suggests that it can be interpreted differently: it characterizes the *intelligibility* of the noisy/processed signal $x$, *not* the quality of the signal $\hat{s}_{lmmse}(x)$.

*Remark 4:* As a consequence of Remark 2, linear processing of noisy critical-band amplitudes cannot improve intelligibility beyond that of the underlying noisy signal. To extend the range of this statement further, recall that we introduced the linear estimator $\hat{s}_{lmmse}(x)$ in Eq. (8) only for ease of estimation. The argument can be repeated with the generally non-linear estimator $\hat{s}_{mmse}(x)$ from Eq. (4), and the conclusion is that no processing of noisy critical-bands, linear or otherwise, allows intelligibility improvements. This prediction is in line with the results by Loizou [32] and Taal *et al.* [15], who showed that single-channel noise reduction systems in general provide no or very modest intelligibility improvements.

*Remark 5:* The proposed intelligibility predictor shows similarities to the STOI measure [17]. Let

$$\rho = \frac{r_{SX} - \mu_S \mu_X}{\sqrt{\sigma_S^2 \sigma_X^2}}$$

denote the linear correlation coefficient. Inserting Eq. (9) in Eq. (13) and using this expression for $\rho$, we find

$$I_{LB,lmmse}(S;X) \approx \frac{1}{2} \ln (1 - \rho^2)^{-1}. \tag{14}$$

The proposed intelligibility predictor averages these values of $I_{LB,mmse}(S;X)$ across speech-active time-frequency units (see Eq. (15) in the next Section) and uses this average as a predictor of intelligibility.

STOI, on the other hand, computes the average of $\rho$ across speech-active frequency units[5]. In this way, the proposed intelligibility predictor and STOI show strong similarities; the main difference to STOI appears to be the non-linearity $\ln\,(1-\rho^2)^{-1}$ applied to $\rho$ before averaging.

STOI is mainly heuristically motivated (for example, computing the linear correlation coefficient for speech-active time-frequency units with a subsequent averaging operation is not linked to any underlying theoretical reasoning). However, Eq. (14) is a *consequence* of the assumed signal model and auditory model and can in this sense be seen to offer some theoretical justification of the less well-motivated choices made in STOI.

*Remark 6:* The proposed intelligibility predictor does not make use of band-importance functions. As above, this is not a choice, but rather a consequence of the model. Although many existing predictors do make use of band-importance functions, e.g. [6], [33], both the proposed predictor and STOI appear to work quite well without.

## IV. Implementation

Our implementation shows similarities to the STOI model described in [17]. Signals are resampled to a sampling frequency of 10 kHz, to ensure that the frequency region relevant for speech intelligibility is covered [2]. Signals are divided into frames of length $N = 256$ samples, and a Hann analysis window $w(n)$ is applied; we use a frame shift of $D = N/2 = 128$ samples. A DFT order of $N = 256$ is used. DFT coefficients are grouped into a total of $L = 15$ third-order octave bands, with a center frequency of the lowest band set to 150 Hz, and the center frequency of the highest band set to approximately 4.3 kHz.

The VAD blocks in Fig. 1 are implemented by identifying signal frames with energy no less than $\Delta_E$ dB of the signal frame with maximum energy. The indices of these signal frames are collected in the index sets $Z_s$ and $Z_x$ for the clean and noisy/processed signals, respectively.

Let $\bar{S}_i(m)$ and $\bar{X}_i(m)$ denote the critical-band amplitudes with frame indices $m \in Z_s \cap Z_x$. The first and second moments needed to evaluate $I(\bar{S}_i(m), \bar{X}_i(m))$ via Eqs. (12) and (9) are estimated using first-order recursive smoothing, i.e.,

$$\hat{r}_{S_i X_i}(m+1) = \alpha \hat{r}_{S_i X_i}(m) + (1-\alpha) \bar{S}_i(m+1) \bar{X}_i(m+1),$$

and similarly for the other moments.

[4] We note that in the original proposal [31], band-importance functions were used, which were equal to the short-term magnitude spectrum of the clean signal, raised to a power.

[5] Note, though, that STOI applies a clipping procedure to the noisy/processed time-frequency units before computing $\rho$.

TABLE I
PARAMETER VALUES IN PROPOSED MODEL.

| Parameter | $\alpha$ | $\Delta_E$ [dB] | $I_{max}$ [nats] |
|---|---|---|---|
| Value | 0.95 | 30 | 0.2 |

Let $\hat{I}(S_i(m); X_i(m))$ denote the estimate of $I_{LB,lmmse}(S_i(m); X_i(m))$ obtained by replacing expected values by recursively estimated moments. The average per sentence mutual information is finally computed as

$$\tilde{I}(\mathcal{S};\mathcal{X}) = \frac{1}{L|Z_s|} \times \sum_{m \in Z_x \cap Z_s} \sum_{i=1}^{L} \min(\hat{I}(S_i(m); X_i(m)), I_{max}). \tag{15}$$

We have introduced here an upper limit $I_{\max}$ on the information content per critical-band amplitude to avoid that the final information score is dominated by a single high-information time-frequency unit. The idea of upper limiting the impact of a single time-frequency unit is not new, but has e.g. been used in the SII intelligibility measure, where the estimated critical-band SNR is upper limited to 15 dB [6]. It can be motivated by the observation that at a sufficiently high SNR, a signal is perfectly intelligible, and increasing the SNR beyond this point cannot increase intelligibility further.

The values of the three parameters, $\alpha$, $\Delta_E$, and $I_{\max}$ are summarized in Table I. The value of $\alpha = 0.95$ corresponds to a time constant of 250 ms. The idea of averaging signal statistics over longer time spans such as 250 ms rather than 20-40 ms which are often used in speech processing applications, e.g. [18], but much shorter than the typical long-term statistics, e.g. as used in SII [6], is not new. For example, in the STOI model [17], statistics were computed across time spans of roughly 400 ms, and it was suggested that this time span could be linked to temporal integration processes taking place in the auditory system. Performance with the proposed model is not sensitive to the exact value of $\alpha$. The choice of $\Delta_E = 30$ dB is not controversial. For clean speech, most speech frames have an energy content larger than this threshold. The value of $I_{\max} = 0.2$ nats was determined heuristically, but prediction performance does not seem to be very sensitive with respect to the exact value of this parameter either.

## V. SIMULATION RESULTS

In the following we evaluate the proposed intelligibility predictor using noisy speech signals processed with different time-frequency weighting strategies. We compare the performance of the proposed method with that of algorithms proposed in the literature. The sample rates mentioned below are used in the listening experiments. When applying the proposed method, the signals are down- or upsampled to 10 kHz.

### A. Signals and Processing Conditions

*1) Additive Noise:* The first set of signals is from the study described by Kjems in [34]. In this study, speech signals from the Dantale II sentence test [35] are contaminated by four different additive noise sources. The speech sentences consists of 150 5-word sentences spoken by a female Danish speaker. The noise sources are i) stationary speech shaped noise created by filtering white noise through a shaping filter with a frequency response corresponding to the long-term spectrum of the speech sentences, ii) car cabin noise recorded in a car driving on the highway, iii) bottle hall noise, and iv) cafeteria noise, which is a recording of a conversation in Danish between a male and a female speaker, i.e. two-talker babble, equalized to have the same long-term spectrum as the test sentences [34]. The sample rate is 20 kHz.

Kjems conducted a listening test to establish the 20% and 80% speech reception threshold (SRT)[6] for each noise source, and a logistic function was fitted to the SRTs to estimate the underlying psychometric function. Finally, we generated noisy test signals with SNRs from $-20$ dB to 5 dB in steps of 2.5 dB, and the corresponding intelligibility scores were established by sampling the psychometric functions at these input SNR values. The total number of conditions therefore amounts to 4 noise types x 11 SNRs = 44 conditions. A number of 15 normal-hearing subjects in the age range 25–52 years participated in the test.

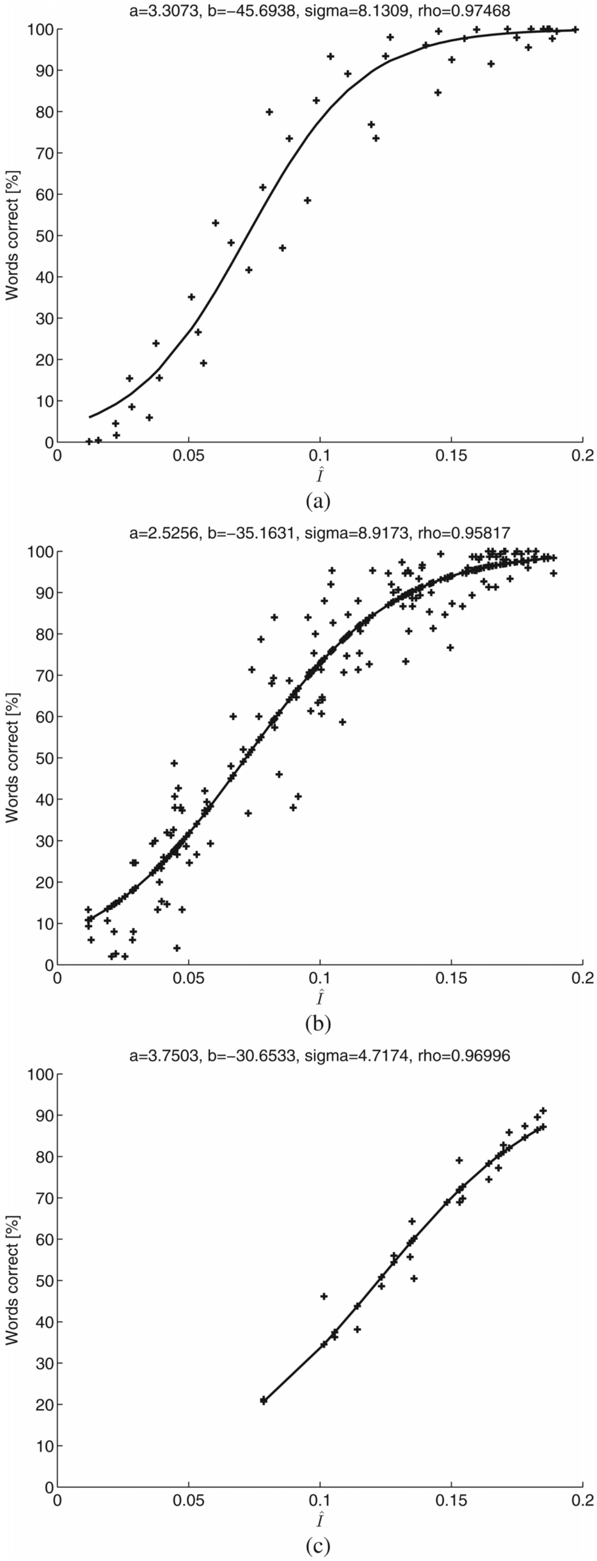
*2) Ideal Binary Mask Signals:* In a second experiment, Kjems [34] processed the noisy signals, using the technique of ideal time-frequency segregation (ITFS) [36], and measured the intelligibility of the resulting signals for different processing conditions. More specifically, the IFTS processing decomposes the clean signal $s(n)$, the noise signal $w(n)$, and the noisy signal $x(n) = s(n) + w(n)$ in time-frequency tiles. In the implementation of Kjems, the time domain signals were analyzed in the short-term spectral domain using a gamma-tone filter bank with 64 channels, each with a bandwidth of 1 ERB, and channel center frequencies linearly spaced on the ERB scale with center frequencies between 55 and 7500 Hz. The filterbank signals were segmented into 20-ms windowed frames with an overlap of 50%, and the energy $\epsilon_{s,i}(m)$, $\epsilon_{w,i}(m)$, and $\epsilon_{x,i}(m)$ of the $i$th subband and $m$th frame was computed for the clean, noise and noisy signal, respectively. Then, a binary-mask value $g_i(m) \in \{0, 1\}$ was computed for each time-frequency unit. Finally, the resulting binary mask signal $g_i(m)$ was upsampled to the signal sample rate of 20 kHz, and point-wise multiplied with the noisy filterbank output, and the result was passed through a gamma-tone synthesis filter bank to synthesize the corresponding processed time domain signal.

Two methods for deriving the binary mask signal $g_i(m)$ were compared. In the *ideal binary mask* (IBM) method the binary mask signal $g_i(m)$ was found by comparing the local target-to-noise ratio $10\log_{10}(\epsilon_{s,i}(m)/\epsilon_{w,i}(m))$ to a threshold LC according to

$$g_i(m) = \begin{cases} 1 & 10\log_{10}(\epsilon_{s,i}(m)/\epsilon_{w,i}(m)) > \mathrm{LC} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

In the *target binary mask* (TBM) method, the binary mask signal was found by replacing the local time-frequency noise energy $\epsilon_{w,i}(m)$ with the value of the long-term speech spectrum, evaluated in the $i$‘th gammatone filter. For both the IBM and TBM methods, the sparsity of the binary mask $g_i(m)$ is a function of the threshold LC: the higher the value of LC, the fewer 1’s in the

[6] The $x\%$ SRT is defined as the SNR at which the average listener correctly identifies $x$ percent of the test words.



(a)

(b)

(c)

Fig. 3. Per-sentence mutual information versus measured intelligibility for various processing conditions. (3(a)): additive noise,(3(b)), and (3(c)) noisy signals processed by single-channel noise reduction algorithms. See text for further details on test signals and conditions.

mask. Eight different values of the threshold LC were used, one of which was LC $= -\infty$ dB, leading to $g_i(m) = 1$ and therefore noisy, unprocessed test signals. For each of the four noise types, noisy signals were generated at three different SNRs: two SNRs were selected corresponding to the 20% and 50% SRT. The third SNR was fixed at $-60$ dB. The total number of test conditions was (4 noise types (IBM) + 3 noise types (TBM)[7]) x 8 LCs x 3 SNRs = 168. As above, 15 normal hearing listeners participated in the intelligibility test, and the sampling rate was 20 kHz.

*3) Single-Channel Enhancement Signals:* The last set of signals consists of noisy signals processed by three single-channel noise reduction algorithms [37]. In contrast to the ITFS processing, which cannot be realized in practice because it requires knowledge of the clean speech signal and the noise signal realizations separately, the study in [37] considered single-channel noise reduction methods which *can* be realized in practice. We included this data set, because it is important to understand if a given intelligibility predictor is limited to synthetic signals, or if it can actually be applied to signals generated by practical applications.

Noisy signals were processed using a DFT-based analysis-modification-synthesis framework. Three processing methods were considered: two methods for finding a binary mask as in Eq. (16) but using only the noisy speech signal, and, in addition, a state-of-the-art single-channel noise reduction method, where the gain values applied to time-frequency units are not constrained to be binary (but are non-negative). Finally, the noisy unprocessed signals were included in the test as well, leading to four processing conditions in total.

The listening test was a closed Dutch speech-in-noise intelligibility test proposed in [38]. As in the previous section, the test sentences consisted of 5 words, spoken by a female speaker. The signals were sampled at a sampling-rate of 8 kHz, and degraded by speech-shaped noise at five different SNRs, namely $-8$, $-6$, $-4$, $-2$ and 0 dB. Thirteen native Dutch speaking subjects participated in the test. Each processing condition was presented five times, and each sentence was used only once. The order of presenting the different algorithms and SNRs was randomized. The signals were presented diotically through head-phones (Sennheiser HD 600). For each processing method and SNR pair, the intelligibility scores were averaged across the 13 listeners and the 5 repetitions, leading to 4 conditions x 5 SNRs = 20 processing conditions for which average intelligibility scores are computed.

## *B. Per Sentence Mutual Information vs Intelligibility*

Fig. 3 plots the per-sentence mutual information lower bound as calculated by Eq. (15) versus the intelligibility measured in listening tests, and averaged across test subjects. The free parameters, $\alpha$, $\Delta_E$, and $I_{\max}$ were chosen according to Table I for all subfigures.

For any good intelligibility predictor, the intelligibility-vs-prediction curve is monotonically increasing. Clearly, Fig. 3 shows a strong monotonic relation between speech intelligibility and mutual information, for all three test conditions. With the proposed intelligibility predictor, one is able to predict *relative* intelligibility: if $\tilde{I}$ is larger for one noisy/processed signal than for another (where the underlying clean speech signal is

[7] For speech shaped noise, IBM and TBM are identical.

the same), we would expect the intelligibility of the first to be larger than the latter.

However, in order to estimate *absolute* intelligibility, a mapping is needed between the outcome of the intelligibility predictor, and the true underlying intelligibility. This mapping is a function of many factors, including the noise type, the test type, the processing applied to the noisy signal, the redundancy of the speech material, and, obviously, the intelligibility predictor.

For additive noise experiments, the psychometric curve is often modeled as a logistic function of the input SNR. In [17], [39] it was proposed to extend the use of this function to model the relationship between the outcome of the intelligibility prediction $\tilde{I}(\mathcal{S};\mathcal{X})$ and the true underlying intelligibility $I$, i.e.,

$$f(\tilde{I}) = \frac{1}{1+\exp(a\tilde{I}+b)},$$

where $a, b \in \Re$ are test specific model parameters, which are estimated to fit the intelligibility data. In Fig. 3, the parameters $a, b$ were estimated to fit the data in each subfigure in a least-square sense; the resulting logistic function is overlaid each subfigure.

To evaluate numerically the performance of intelligibility predictors, we use two figures of merit, namely the normalized linear correlation coefficient $\rho$ between average intelligibility scores obtained through listening tests, and the outcomes of the intelligibility predictors, and the root mean-square prediction error $\sigma$. Let $\tilde{I}_k$ denote the intelligibility prediction for the $k$th processing condition, and let $SI_k$ denote the average across listeners in the corresponding intelligibility test. Furthermore, let $\mu_{f(\tilde{I})}$, and $\mu_{SI}$ denote the averages of $f(\tilde{I}_k)$ and $SI_k$, respectively, across listening test conditions $k$, and let $K$ denote the number of test conditions. The normalized linear correlation coefficient is then defined as

$$\rho = \frac{\sum_k (f(\tilde{I}_k) - \mu_{f(\tilde{I})})(SI_k - \mu_{SI})}{\sqrt{\sum_k (f(\tilde{I}_k) - \mu_{f(\tilde{I})})^2 \sum_k (SI_k - \mu_{SI})^2}}, \qquad (17)$$

and the root mean-square prediction error $\sigma$ is defined as

$$\sigma = \sqrt{\frac{1}{K}\sum_k (f(\tilde{I}_k) - SI_k)^2}. \qquad (18)$$

### C. Comparison to Other Intelligibility Predictors

In this section we compare the performance of the proposed intelligibility predictor, which will be abbreviated as *SIMI* (Speech Intelligibility prediction based on Mutual Information) to several methods from the literature, see Table II. Specifically, we consider *STOI* [17], *CSII-MID* (the Mid-level coherence SII as proposed in [8]), *CSII-BIF* (the coherence SII with signal-dependent band importance functions, referred to as $\mathrm{CSII}_{\mathrm{mid}}$, $W_4$, $p = 1$ in [33]), *STI-NCM* (the normalized covariance speech transmission index as proposed in [12]), *STI-NCM-BIF* (the normalized covariance speech transmission index with signal-dependent band-importance functions, referred to as NCM, $W_i^{(1)}$, $p = 1.5$ in [33]), and *NSEC* (the Normalized Subband Envelope Correlation method as proposed in [40]).

TABLE II
INTELLIGIBILITY PREDICTORS FOR COMPARISON.

| Method Name | Remarks |
|---|---|
| *STOI* [17] | The short-time objective intelligibility measure. |
| *CSII-MID* [8] | The mid-level coherence SII. |
| *CSII-BIF* [33] | The coherence SII with signal-dependent band importance functions (referred to as $\mathrm{CSII}_{\mathrm{mid}}$, $W_4, p = 1$ in [33]). |
| *STI-NCM* [12] | The normalized covariance speech transmission index. |
| *STI-NCM-BIF* [33] | The normalized covariance speech transmission index with signal-dependent band-importance functions (referred to as NCM, $W_i^{(1)}, p = 1.5$ in [33]). |
| *NSEC* [40] | The normalized subband envelope correlation method. |

We evaluate the performance of these intelligibility predictors in terms of $\rho$ and $\sigma$ computed from an $n$-fold cross-validation procedure. Specifically, for each data set, the set is randomly divided into $n = 4$ equal size subsets, the free parameters $a, b$ in the logistic function are fitted to the $n - 1$ subsets, after which $\rho$ and $\sigma$ are computed based on prediction of the remaining subset. This procedure is repeated for each subset, and the averages of the resulting $\rho$ and $\sigma$ values are computed. Tables III and IV summarize these results. The values in brackets are found from estimating $a, b$ using the entire data set, and estimating $\rho$ and $\sigma$ from the *same* set; these are included here for comparison with values reported in literature, which are often computed in this way, e.g. [17].

From Tables III and IV most intelligibility predictors work well in the case of additive noise, leading to correlation coefficients of $\rho > 0.93$; *STI-NCM-BIF* is an exception, but it should be noted that this method was proposed in [33] for single-channel noise reduced speech. For the ITFS processed signals, only *SIMI* and *STOI* work well, resulting in linear correlation coefficients of $\rho > 0.95$ and $\sigma < 9.0$. Most other methods essentially fail in this situation. Similar results were reported in [41]. For the single-channel enhanced speech signals, most intelligibility predictors perform well. It is worth noting that *STI-NCM-BIF* works particularly well in this situation; this result is in quantitative agreement with the results reported in [33]. The results of Tables III and IV are also in general agreement with the results reported in [17], [41]. Note finally that *SIMI* and *STOI* are the only methods which work well for all conditions.

## VI. CONCLUSION

Algorithms for estimating the outcome of intelligibility listening tests are of interest both for reducing the number of costly listening tests during algorithm development, but also have the potential to lead to new insights into the auditory system.

Historically, a wide range of intelligibility predictors have been proposed with varying validity domains including additive (stationary or non-stationary) or convolutive noise types, and several types of signal processing, including filtering, clipping, etc. In this work we consider the situation of additive, but not necessarily stationary, noise sources, and non-linear processing which can generally be referred to as *time-frequency weighting*.

TABLE III
PERFORMANCE OF INTELLIGIBILITY PREDICTORS IN TERMS OF LINEAR CORRELATION COEFFICIENT $\rho$, EQ. (17), FOR DIFFERENT NOISE/PROCESSING CONDITIONS. THE PERFORMANCE SCORES ARE ESTIMATED USING $n$-FOLD CROSS-VALIDATION ($n = 4$). PERFORMANCE SCORES IN BRACKETS ARE COMPUTED BY FITTING THE LOGISTIC FUNCTION TO THE ENTIRE DATA SET, AND COMPUTING THE RESULTING $\rho$ ACROSS THE *SAME* DATA SET.

| | *SIMI* | *STOI* | *CSII-MID* | *CSII-BIF* | *STI-NCM* | *STI-NCM-BIF* | *NSEC* |
|---|---|---|---|---|---|---|---|
| Additive Noise | 0.975 (0.975) | 0.969 (0.967) | 0.943 (0.949) | 0.978 (0.978) | 0.934 (0.935) | 0.808 (0.813) | 0.951 (0.953) |
| ITFS-Processing | 0.957 (0.958) | 0.966 (0.961) | 0.352 (0.455) | 0.517 (0.539) | 0.613 (0.727) | 0.481 (0.586) | 0.834 (0.868) |
| SC-Enhancement | 0.976 (0.970) | 0.985 (0.987) | 0.776 (0.625) | 0.918 (0.850) | 0.910 (0.844) | 0.971 (0.976) | 0.896 (0.799) |

TABLE IV
PERFORMANCE OF INTELLIGIBILITY PREDICTORS IN TERMS OF ROOT MEAN-SQUARE PREDICTION ERROR $\sigma$, EQ. (18), FOR DIFFERENT NOISE/PROCESSING CONDITIONS. THE PERFORMANCE SCORES ARE ESTIMATED USING $n$-FOLD CROSS-VALIDATION ($n = 4$). PERFORMANCE SCORES IN BRACKETS ARE COMPUTED BY FITTING THE LOGISTIC FUNCTION TO THE ENTIRE DATA SET, AND COMPUTING THE RESULTING $\sigma$ ACROSS THE *SAME* DATA SET.

| | *SIMI* | *STOI* | *CSII-MID* | *CSII-BIF* | *STI-NCM* | *STI-NCM-BIF* | *NSEC* |
|---|---|---|---|---|---|---|---|
| Additive Noise | 8.95 (8.13) | 9.45 (9.24) | 12.72 (11.47) | 7.95 (7.59) | 13.41 (12.74) | 21.11 (21.08) | 11.48 (10.89) |
| ITFS-Processing | 8.49 (8.92) | 8.20 (8.61) | 27.45 (27.72) | 25.73 (26.23) | 20.56 (21.38) | 24.43 (25.23) | 14.59 (15.51) |
| SC-Enhancement | 5.34 (4.72) | 3.41 (3.09) | 16.00 (15.12) | 11.45 (10.20) | 11.69 (10.38) | 4.53 (4.22) | 13.05 (11.65) |

This class of processing method is quite broad and is for example used in single-channel noise reduction algorithms.

In this context, we pursue the hypothesis that intelligibility could be monotonically related to the Shannon information about the (unknown) clean critical-band envelopes, which can be learned by observing their noisy and potentially processed counterparts. We derive lower bounds for this mutual information, which turn out to be analytically tractable. Specifically, the information lower bound can be computed as a function of the minimum mean-square error (mmse) arising from estimating the clean critical-band amplitude from its noisy/processed counterpart.

The proposed model has a number of surprising consequences. Traditionally, in speech signal processing the mse between a clean time-frequency unit and an estimate thereof has been linked to the speech quality resulting from the estimator in question. According to this paradigm, using an mmse estimator would lead to highest speech quality. However, the proposed model suggests that it could be interpreted differently: the mmse could be viewed as an indicator of the intelligibility of the underlying noisy (and potentially processed) signal. Furthermore, it is interesting to note that whereas several of the intelligibility predictors proposed in the literature are heuristically motivated, the proposed mmse based predictor is a consequence of a simple auditory model and signal model, and the assumption that mutual information can be used as a principle for comparing inner representations. Finally, the proposed model predicts that processing of noisy critical-band amplitudes (based on the noisy critical-band amplitudes only) cannot lead to intelligibility improvements, a prediction which is in line with several existing intelligibility test results, e.g., [15], [32].

Simulation experiments with the proposed method shows that it is able to reliably estimate the average intelligibility of speech signals contaminated by stationary and non-stationary noise sources as well time-frequency processed noisy speech. In a comparison with other intelligibility predictors from literature, this performance was only equalled by the STOI intelligibility predictor [17].

It is of interest to study in the future if the proposed principle is valid in a more general context than covered in this article. For example, the auditory model presented in this article is very simple; it would be of interest to study prediction performance if the proposed mutual information principle were combined with a more physiologically plausible model, e.g. the modulation filter based model of Dau *et al.* [42] and the intelligibility prediction model by Jørgensen *et al.* [13].

Another topic for future research includes the extension of the proposed principle to situations which are more natural to human listeners, e.g. a binaural listening setup. It appears possible that phenomena such as the spatial unmasking effect, e.g. [43] may be predicted well by such extended model.

## REFERENCES

[1] *American National Standard Methods for the Calculation of the Articulation Index*, ANSI S3.5, American National Standards Institute, New York, NY, USA, 1969.

[2] 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.

[3] K. D. Kryter, “Methods for the calculation and use of the articulation index,” *J. Acoust. Soc. Amer.*, pp. 1689–1697, 1962.

[4] *Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index*, IEC60268-16, Int. Electrotechnical Commission, Geneva, Switzerland, 2003.

[5] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech-transmission quality,” *J. Acoust. Soc. Amer.*, vol. 67, pp. 318–326, 1980.

[6] *Methods for the Calculation of the Speech Intelligibility Index*, ANSI S3.5, American National Standards Institute, New York, NY, USA, 1995.

[7] 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.

[8] 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.

[9] T. Houtgast and H. J. M. Steeneken, “Evaluation of speech transmission channels by using artificial signals,” *Acustica*, vol. 25, pp. 355–367, 1971.

[10] V. Hohmann and B. Kollmeier, “The effect of multichannel dynamic compression on speech intelligibility,” *J. Acoust. Soc. Amer.*, vol. 97, pp. 1191–1195, 1995.

[11] R. Drullmann, “Temporal envelope and fine structure cues for speech intelligibility,” *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 585–592, Jan. 1995.

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

[13] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Amer.*, vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

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

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Proc. Int. Workshop, Acoust. Echo Noise Control*, 2010.

[16] 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.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

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

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley, 1991.

[20] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing*, D. Pressnitzer, A. Cheveigné, S. McAdams, and L. Collet, Eds. New York, NY, USA: Springer, 2005, pp. 313–319.

[21] A. Leijon, "Articulation index and shannon mutual information," in *Hearing - From Sensory Processing to Perception*, B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, Eds. Berlin/Heidelberg, Germany: Springer, 2007, pp. 525–532.

[22] J. Taghia, R. Martin, and R. C. Hendriks, "On mutual information as a measure of speech intelligibility," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 65–68.

[23] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2001, pp. 167–170.

[24] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[25] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Philadelphia, PA, USA: SIAM, 2001.

[26] 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.

[27] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. model structure," *J. Acous. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996 [Online]. Available: http://link.aip.org/link/?JAS/99/3615/1

[28] W. Bialek, F. DeWeese, and D. Warland, "Bits and brains: Information flow in the nervous system," *Physica A*, vol. 200, no. 1-4, pp. 581–593, 1993.

[29] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*, 1st ed. ed. New York, NY, USA: Wiley - Interscience, Jul. 1990.

[30] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1992.

[31] 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.

[32] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 561–564.

[33] J. Ma, Y. Hu, and P. C. 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.

[34] U. Kjems *et al.*, "Role of mask pattern in intelligibility of ideal binary-masked nosy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

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

[36] D. Brungart *et al.*, "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, Dec. 2006.

[37] J. Jensen and R. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.

[38] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in noise test (matrix)," in *Proc. 8th EFAS Congr., 10th DGA Congr.*, Heidelberg, Germany, Jun. 2007.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech itelligibility prediction," in *Proc. Interspeech.*, Brighton, U.K., Sep. 6-10, 2009, pp. 1947–1950, ISCA.

[40] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proc. 17th Eur. Signal Process. Conf. (EUSIPCO)*, 2009, pp. 1849–1853.

[41] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Amer.*, vol. 130, pp. 3013–3027, 2011.

[42] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. detection and masking with narrow-band carriers," *J. Acoust. Soc. Amer.*, vol. 102, no. 5, pp. 2892–2905, 1997.

[43] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica*, vol. 86, pp. 117–128, 2000.

**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 a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Multimedia Information and Signal Processing (MISP), Department of Electronic Systems, at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

**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 computer science from the Delft University of Technology (DUT), Delft, The Netherlands, in 2007. From 2008 to 2012, he was a Ph.D. Researcher in the Multimedia Signal Processing Group, DUT, under the supervision of R. Heusdens and R.C.Hendriks in collaboration with Oticon A/S.

From 2012 to 2013 he held Postdoc positions at the Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden and the Leiden University Medical Center (LUMC), Leiden, the Netherlands. Currently, he is with Philips Research, Eindhoven, the Netherlands.

His main research interests are in the field of audio, speech and biomedical digital signal processing.