Predicting the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech

Heidemann Andersen, Asger; de Haan, Jan Mark; Tan, Zheng-Hua; Jensen, Jesper

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

DOI (link to publication from Publisher):
10.1109/TASLP.2016.2588002

Publication date:
2016

Document Version
Accepted author manuscript, peer reviewed version

Predicting the Intelligibility of Noisy and Non-Linearly Processed Binaural Speech

Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen

***Abstract*—Objective speech intelligibility measures are gaining popularity in the development of speech enhancement algorithms and speech processing devices such as hearing aids. Such devices may process the input signals non-linearly and modify the binaural cues presented to the user. We propose a method for predicting the intelligibility of noisy and non-linearly processed binaural speech. This prediction is based on the noisy and processed signal as well as a clean speech reference signal. The method is obtained by extending a modified version of the short-time objective intelligibility (STOI) measure with a modified equalization-cancellation (EC) stage. We evaluate the performance of the method by comparing the predictions with measured intelligibility from four listening experiments. These comparisons indicate that the proposed measure can provide accurate predictions of 1) the intelligibility of diotic speech with an accuracy similar to that of the original STOI measure, 2) speech reception thresholds (SRTs) in conditions with a frontal target speaker and a single interferer in the horizontal plane, 3) SRTs in conditions with a frontal target and a single interferer when ideal time frequency segregation (ITFS) is applied to the left and right ears separately, and 4) the advantage of two-microphone beamforming as applied in state-of-the-art hearing aids. A MATLAB implementation of the proposed measure is available online[1].**

***Index Terms*—binaural speech intelligibility prediction, speech enhancement, speech transmission, binaural advantage**

I. Introduction

THE speech intelligibility prediction problem consists of predicting the intelligibility of a particular noisy/processed/distorted speech signal to an average listener. The problem was initially studied with the purpose of improving telephone systems [2], [3]. Since then, it has been applied as a development tool in related fields such as telecommunication [4], architectural acoustics [5], [6] and speech processing [7]–[12]. Many endeavours in these fields focus on improving speech understanding in particular conditions. This introduces the need for measuring speech intelligibility through listening experiments, which is a time consuming and expensive task. Objective (computational) measures of intelligibility can provide estimates of the results of such experiments faster and at a lower cost, while being easily reproducible.

An early Speech Intelligibility Prediction (SIP) method is the Articulation Index (AI) [3], [13], which can be seen as a common ancestor for most of the methods which have been proposed since then. The AI considers the condition in which a listener is presented with monaural speech contaminated by additive, stationary noise. It is assumed that speech and noise at the ear of the listener are available as separate signals. The AI estimates intelligibility as a weighted sum of normalized Signal to Noise Ratios (SNRs) across a range of third octave bands. It has later been shown that, under certain assumptions, this is in fact an estimate of the channel capacity from Shannon's information theory [14]. A refined and standardized version of the AI is known as the Speech Intelligibility Index (SII) [15]. Notably, the AI and the SII are unsuitable for conditions involving fluctuating noise interferers, binaural conditions and conditions where speech and noise are not combined linearly (due to e.g. distorting transmission systems or non-linear speech processing algorithms).

Many SIP methods have been proposed since the introduction of the AI, mainly focussing on extending the domain in which accurate predictions can be made. For example, the Speech Transmission Index (STI) estimates the impact of a transmission channel (e.g. the acoustics of a room or a noisy and distorting transmission system) on intelligibility by measuring the change in modulation depth across the system [16], [17]. It has, however, been shown that the STI does not perform well at predicting the impact of speech enhancement algorithms, on speech intelligibility [9], [18]–[22]. A more recent modulation-based and physiologically motivated method, the speech-based Envelope Power Spectrum Model (sEPSM), has been shown to perform well at predicting the impact of spectral subtraction [22]. Another notable method is the Extended SII (ESII), which is a variation of the SII that provides more accurate predictions in conditions with fluctuating noise interferers [23], [24]. The Coherence SII (CSII) is yet another variation of the SII which aims to predict the influence on intelligibility of non-linear distortion from clipping [25]. The CSII and several other intelligibility measures are evaluated with speech processed by noise reduction algorithms in [26]. The recent Hearing-Aid Speech Perception Index (HASPI) is closely related to the CSII, but involves a more sophisticated auditory model and aims to predict the intelligibility of processed speech for hearing impaired listers [27]. Recently, the Short-Time Objective Intelligibility

A. H. Andersen is with Oticon A/S, 2765 Smørum, Denmark, and also with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: aand@oticon.com; aha@es.aau.dk).

J. M. de Haan is with Oticon A/S, 2765 Smørum, Denmark (e-mail: janh@oticon.com).

Z.-H. Tan is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: zt@es.aau.dk).

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: jesj@oticon.com; jje@es.aau.dk).

Parts of this work has been presented in [1]. The present work extends that of [1] by more rigorously covering theoretical derivations, and by providing additional evaluation of the proposed intelligibility measure.

[1]See `http://kom.aau.dk/project/Intelligibility/`.

(STOI) measure [7] has become very popular for evaluation of noisy and processed speech. The STOI measure has shown to compare favorably to several other SIP methods, with respect to predicting the impact of various single microphone enhancement schemes as well as Ideal Time Frequency Segregation (ITFS) [7]. This observation is confirmed for hearing impaired listeners in [28], which shows that the CSII and the STOI measure perform favorably to other measures at predicting the effect of noise reduction algorithms. The STOI measure has later been shown to compare well with other measures with respect to predicting the impact of a number of detrimental effects and processing schemes relevant to users of hearing aids and cochlear implants [8] and for predicting the intelligibility of noisy speech transmitted by telephone [4]. Finally, we mention the Speech Intelligibility prediction based on Mutual Information (SIMI) measure [12], which is very similar to the STOI measure in structure and performance, but which is based on information theoretical considerations.

The methods discussed up to this point all assume that speech is presented monaurally or diotically to the listener. However, in many real world scenarios, humans obtain an advantage from listening with two ears. This is partly because one can, to some extent, choose to listen to the ear in which the speech is more intelligible, and partly because the brain can combine information from the two ears [29]. The Equalization-Cancellation (EC) stage is an early simple model which predicts Binaural Masking Level Differences (BMLDs) accurately in a range of conditions [30], [31] (i.e. the binaural advantage obtained in tasks such as detecting a tone in noise). Several attempts have been made at developing SIP methods which account for binaural advantage [32]–[40], i.e. the advantage in intelligibility obtained through the presence of interaural, source-dependent, phase and level differences. Notably, the Binaural Speech Intelligibility Measure (BSIM) [33] uses the EC stage as a preprocessor to the SII to predict the intelligibility of binaural signals. The same paper proposes another binaural method, the short-time Binaural Speech Intelligibility Measure (stBSIM), with properties similar to the ESII (i.e. the ability to handle fluctuating noise interferers) [33]. A number of ways to extend the BSIM, such as to predict the detrimental effect of reverberation, are investigated in [39], [40]. A different approach for combining the SII with an EC stage is proposed in [36]. The method estimates the Speech Reception Threshold (SRT) of the better ear and subtracts an estimate of binaural advantage obtained by an EC based method proposed in [41], [42]. This method has later been expanded further to account for aspects such as to multiple interferers and reverberation [37], [38], [43].

It should be realized that none of the above-mentioned methods are able to predict the simultaneous impact of both non-linear processing and binaural advantage. This is in spite of the fact that both effects are important in the context of modern audio processing devices that present signals dichotically to a user, e.g. hearing aids. In [44] we introduced an early version of the proposed method, that has shown promising results in predicting both the effects of processing and binaural advantage. The method is obtained by extending the STOI measure such as to predict binaural advantage, and is therefore referred to as the Binaural STOI (BSTOI) measure. Taking inspiration from [33], this measure is obtained by using a modified EC stage to combine the left and right ear signals, prior to predicting intelligibility with the STOI measure. Because the EC stage includes internal noise sources, which model inaccuracies in the human auditory system, computationally expensive Monte Carlo simulation is used to obtain an estimate of the expected STOI measure across these noise sources [44]. Results presented in [44] indicate that the BSTOI measure can predict both binaural advantage and the effect of non-linear processing with ITFS. It was not investigated whether the BSTOI measure can account for both effects simultaneously (i.e. they were investigated separately).

In the present study, we introduce a refined version of the BSTOI measure, which we refer to as the Deterministic BSTOI (DBSTOI) measure. In order to avoid Monte Carlo simulation, the DBSTOI measure introduces some minor changes in the STOI measure which allow us to derive an analytical expression for the expectation of the output measure across the internal noise sources in the EC stage. The DBSTOI measure is therefore much less computationally demanding to evaluate than the BSTOI measure. Furthermore, the DBSTOI measure produces fully deterministic outputs. Except for the mentioned advantages of the DBSTOI measure, no noteworthy performance differences between the DBSTOI and BSTOI measures have been found. Furthermore, we provide a thorough evaluation of the prediction performance of the measure by comparing to the results of four different listening experiments, including one with both non-linear speech enhancement and binaural advantage combined. The ability to handle such conditions allows the measure to predict intelligibility of e.g. users of assistive listening devices in complex real-world scenarios.

The remainder of the paper is organsized as follows: In Sec. II, the proposed intelligibility measure is described in detail. In Sec. III, four sets of experimental data are described. In Sec. IV, the procedure used for evaluating the measure is described. In Sec. V, the results are presented. Sec. VI concludes upon the proposed measure and its performance.

II. The DBSTOI Measure

In this section we present the proposed intelligibility measure in detail. The measure applies to conditions in which a human subject is listening to a well defined target speaker in the presence of some form of interference. Furthermore, the combination of speech and interferer may have been non-linearly transformed by e.g. a speech enhancement algorithm or a distorting transmission system. Intelligibility is predicted on the basis of four input signals: the left and right *clean* signals, $x_l(t)$ and $x_r(t)$, and the left and right *noisy and processed* signals, $y_l(t)$ and $y_r(t)$. The clean signals are measured at the ear of the listener but in the presence of only the target speaker (and in the absence of both interferer and processing). An example of this is illustrated in Fig. 1a. It is assumed that the clean signals are fully intelligible. The aim is to predict the intelligibility of the noisy and processed signals. These are given by the processed mixture of target and

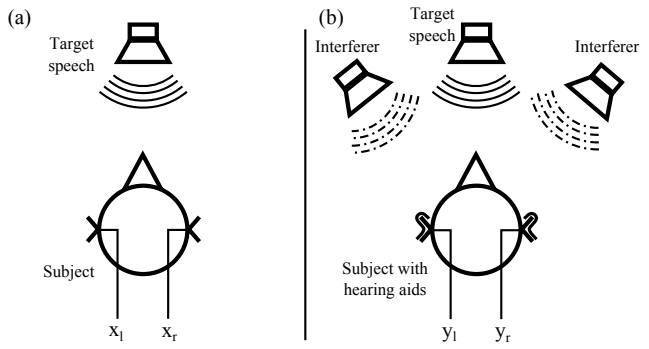

Fig. 1. The four input signals needed by the proposed measure in the exemplifying application where it is used to predict the intelligibility of speech which has been processed by a hearing aid system. a) The left and right clean signals, $x_l(t)$ and $x_r(t)$, are obtained by measuring the acoustic signal in the ear canal of the left and right ear of the subject when listening only to the unprocessed target speaker. b) The left and right noisy/processed signals, $y_l(t)$ and $y_r(t)$, are measured in the ear canals while the subject is wearing hearing aids and is listening to the combination of target and interferer.

interferer as measured at the ear of the listener. An example of this is illustrated in Fig. 1b. The clean and degraded signals are assumed to be time aligned for each ear. E.g. if the degraded signals include a substantial processing delay, the clean signals should be delayed correspondingly to compensate for this difference. It should be stressed that the use of the measure for predicting the impact of hearing aid processed speech, as illustrated in Fig. 1, is merely an example. The measure is applicable in virtually any condition in which noisy and processed speech is presented binaurally to a listener. The clean and noisy/processed signals may be either recorded or simulated by use of Head Related Transfer Functions (HRTFs). A block diagram of the computational structure of the measure is shown in Fig. 2. The block diagram is separated in three steps: 1) a Time-Frequency (TF) decomposition based on the Discrete Fourier Transformation (DFT), 2) a modified EC stage which models binaural advantage and 3) a STOI based stage which rates intelligibility on a scale from -1 to $+1$. The three steps are described in detail in sections II-A, II-B and II-C, respectively.

A. Step 1: Time-Frequency Decomposition

The first step of computing the DBSTOI measure is adopted from the STOI measure [7] with no significant changes. The four input signals are first resampled to 10 kHz. Then, regions in which the target speaker is silent are removed with a simple frame-based Voice Activity Detector (VAD). This is done by 1) segmenting the four input signals into 256-sample Hann-windowed segments with an overlap of 50%, 2) finding the frame with the highest energy for each of the two clean signals, respectively, 3) locating all frame indices where the energy of both clean signal frames are more than 40 dB below their respective maximum and 4) resynthesising the four signals, but excluding the frame numbers which where found in 3. This produces four signals which are time aligned, because the same frames are removed in all the signals.

A TF decomposition of the signals is then obtained in the same manner as for the STOI measure [7]. This is done by segmenting the signals into 256-sample frames with an overlap of 50%, followed by zero-padding each frame to 512 samples and applying the DFT. We refer to the k'th frequency bin of the m'th frame of the left clean signal as $\hat{x}^{(l)}_{k,m}$. Similarly, the same TF units of the right clean signal and the left and right noisy/processed signals are denoted by $\hat{x}^{(r)}_{k,m}$, $\hat{y}^{(l)}_{k,m}$ and $\hat{y}^{(r)}_{k,m}$, respectively.

B. Step 2: Equalization-Cancellation Stage

The second step of computing the measure consists of combining the left and right signals into a single clean signal and a single noisy/processed signal while accounting for any potential binaural advantage. This is done by use of a modified EC stage.

The originally proposed EC stage models binaural advantage under the assumption that the left and right speech and interferer signals are known in separation [30], [31]. The stage introduces relative time shifts and amplitude adjustments between the left and right signals (equalization) and subtracts the two from each other (cancellation) to obtain one signal. This is done separately for the left and right clean signals and the left and right interferer signals such as to obtain a single clean signal and a single interferer signal. The same time shifts and amplitude adjustments are applied for both clean and noisy/processed signals, and these are chosen such as to maximize the SNR of the output. For wideband signals, such as speech, the EC stage is typically applied independently in auditory bands.

The original EC stage cannot be applied in the present case, because the interferer signal is not available in separation. Instead, the processed combination of speech and interferer is available. We propose the following changes in order to adapt the EC stage to work with the available signals:

1) The left and right clean signals and the left and right noisy/processed signals are combined using the same procedure as that of the original EC stage.
2) The time shifts and amplitude adjustment factors are determined such as to maximize the STOI measure of the output, rather than the SNR.

This essentially correspond to assuming that the human brain combines the signals from the two ears such as to maximize intelligibility rather than SNR. The combination of the left and right signals by the modified EC stage, is carried out in the frequency domain as follows:

$$\hat{x}_{k,m} = \lambda_{k,m}\,\hat{x}^{(l)}_{k,m} - \lambda^{-1}_{k,m}\hat{x}^{(r)}_{k,m}, \tag{1}$$

$$\hat{y}_{k,m} = \lambda_{k,m}\,\hat{y}^{(l)}_{k,m} - \lambda^{-1}_{k,m}\hat{y}^{(r)}_{k,m}, \tag{2}$$

where the time and frequency dependent complex-valued factor $\lambda_{k,m}$ represents the time shift and the amplitude adjustment. Specifically, this factor is given by:

$$\lambda_{k,m} = 10^{(\gamma_{k,m}+\Delta\gamma_{k,m})/40} e^{j\omega(\tau_{k,m}+\Delta\tau_{k,m})/2}, \tag{3}$$

where $\gamma_{k,m}$ is the relative amplitude adjustment (in dB), $\tau_{k,m}$ is the relative time shift (in seconds), and $\Delta\gamma_{k,m}$ and $\Delta\tau_{k,m}$

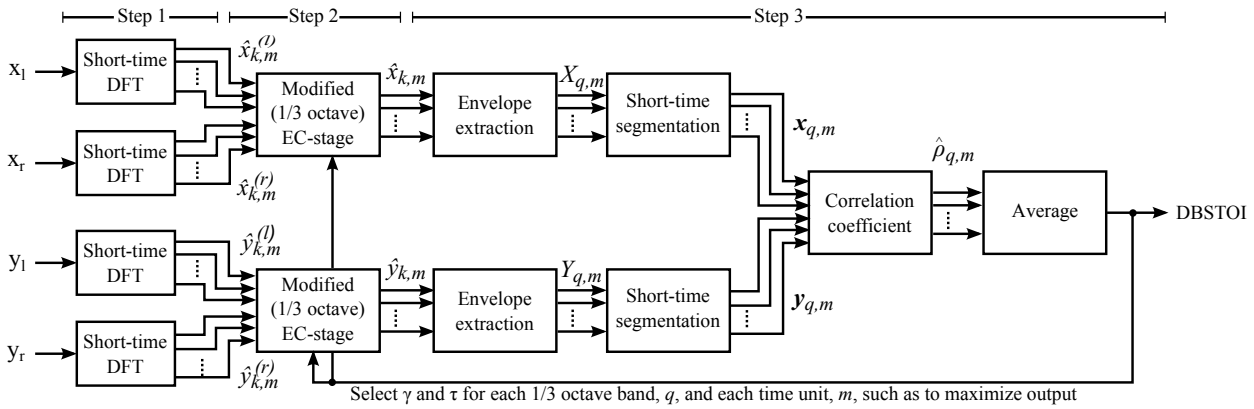

Fig. 2. A block diagram illustrating the computation of the proposed measure.

are uncorrelated random variables which serve to model the suboptimal performance of the human auditory system [30], [45]. These are normally distributed with zero mean and variance (adapted from [45] in the same manner as is done in [32], [33]) given by[2]:

$$\sigma_{\Delta\gamma}(\gamma_{k,m}) = \sqrt{2}\cdot 1.5 \text{ dB}\cdot\left(1+\left(\frac{|\gamma_{k,m}|}{13 \text{ dB}}\right)^{1.6}\right) \quad \text{[dB]}, \quad (4)$$

$$\sigma_{\Delta\tau}(\tau_{k,m}) = \sqrt{2}\cdot 65\ \mu\text{s}\cdot\left(1+\frac{|\tau_{k,m}|}{1.6 \text{ ms}}\right) \quad \text{[s]}. \quad (5)$$

The values of $\gamma_{k,m}$ and $\tau_{k,m}$ are determined independently for each time unit and third octave band such as to maximize the STOI measure of the combined signals (i.e. $\gamma_{k,m}$ and $\tau_{k,m}$ have the same value for all k belonging to one third octave band). The details of this are covered in Sec. II-D. Henceforth, for notational convenience, we discard time and frequency indices such as to denote $\lambda_{k,m}$, $\gamma_{k,m}$ and $\tau_{k,m}$ as λ, γ and τ, respectively. The same is done for the noise sources $\Delta\gamma$ and $\Delta\tau$.

C. Step 3: Intelligibility Prediction

At this point, the left and right ear signals have been combined into one clean signal, $\hat{x}_{k,m}$, and one noisy/processed signal, $\hat{y}_{k,m}$, cf. Fig. 2. This allows us to estimate intelligibility using the STOI measure. However, the signals $\hat{x}_{k,m}$ and $\hat{y}_{k,m}$ are stochastic due to the noise sources $\Delta\gamma$ and $\Delta\tau$ in the EC stage, and the resulting STOI measure is therefore also a stochastic variable. For the BSTOI measure this problem was solved by averaging the output across many realizations of $\Delta\gamma$ and $\Delta\tau$ [44]. This solution is computationally expensive and does not lead to entirely deterministic results. The DBSTOI measure instead applies a slight variation of the originally proposed STOI measure[3], which allows us to derive a closed form expression of the expectation of the final measure across $\Delta\gamma$ and $\Delta\tau$. The remainder of this section describes these matters in detail.

The clean signal "power envelopes" (envelopes squared) are first determined in $Q = 15$ third octave bands with center frequencies starting from 150 Hz. These bands are obtained by grouping DFT coefficients, exactly as in the original STOI measure [7]. The border between two adjacent bands are given by the geometric mean of their respective center frequencies. The upper and lower frequency bin indices of the q'th band are denoted, respectively, by $k_1(q)$ and $k_2(q)$. The resulting expression for the clean signal power envelope is given by:

$$\begin{aligned} X_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}|^2 = \sum_{k=k_1(q)}^{k_2(q)} \left|\lambda\,\hat{x}^{(l)}_{k,m} - \lambda^{-1}\hat{x}^{(r)}_{k,m}\right|^2 \\ &= 10^{\frac{\gamma+\Delta\gamma}{20}} \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}^{(l)}_{k,m}\right|^2 + 10^{-\frac{\gamma+\Delta\gamma}{20}} \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}^{(r)}_{k,m}\right|^2 \\ &\quad - 2\text{Re}\left[\sum_{k=k_1(q)}^{k_2(q)} \hat{x}^{(l)*}_{k,m}\hat{x}^{(r)}_{k,m} e^{-j\omega_k(\tau+\Delta\tau)}\right] \\ &\approx 10^{\frac{\gamma+\Delta\gamma}{20}} X^{(l)}_{q,m} + 10^{-\frac{\gamma+\Delta\gamma}{20}} X^{(r)}_{q,m} \\ &\quad - 2\,\text{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} X^{(c)}_{q,m}\right], \end{aligned} \quad (6)$$

where:

$$\begin{aligned} X^{(l)}_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}^{(l)}_{k,m}|^2, \\ X^{(r)}_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}^{(r)}_{k,m}|^2, \\ X^{(c)}_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} \hat{x}^{(l)*}_{k,m}\hat{x}^{(r)}_{k,m}, \end{aligned} \quad (7)$$

and where ω_k is the angular frequency of the k'th frequency bin and ω_q is the center angular frequency of the q'th third octave band. The last step in (6) assumes that the signal energy is located at the center of each third octave band. The

[2]In [45], noise sources are added independently to the left and right ear signals. Here, one noise source is applied symmetrically. This leads to a multiplicative factor of $\sqrt{2}$ in (4) and (5) compared to [45].

[3]For mathematical tractability, we use "power envelopes" (envelopes squared) rather than magnitude envelopes as originally proposed for the STOI measure [7]. This is also done in [46] and appears to have no significant effect on predictions [46], [47]. Furthermore, we discard the clipping mechanism used in the original STOI measure. The same variation is applied in [46] and the changes do not appear to significantly impair the prediction performance of the measure.

same procedure is applied for the noisy/processed signal to obtain $Y_{q,m}$ as well as $Y_{q,m}^{(l)}$, $Y_{q,m}^{(r)}$ and $Y_{q,m}^{(c)}$.

The obtained power envelope samples are then arranged temporally in zero-mean vectors of $N = 30$ samples, in the same manner as is done in the STOI measure [7]:

$$\mathbf{x}_{q,m} = [X_{q,m-N+1}, \ldots, X_{q,m}]^\intercal - \mathbf{1} \sum_{m'=m-N+1}^{m} \frac{X_{q,m'}}{N}, \quad (8)$$

where $\mathbf{1}$ is a column vector of all ones. Similar vectors are defined from the other power envelope signals: $\mathbf{x}_{q,m}^{(l)}$, $\mathbf{x}_{q,m}^{(r)}$, $\mathbf{x}_{q,m}^{(c)}$, $\mathbf{y}_{q,m}^{(l)}$, $\mathbf{y}_{q,m}^{(r)}$ and $\mathbf{y}_{q,m}^{(c)}$. From (6) we then have:

$$\mathbf{x}_{q,m} \approx 10^{\frac{\gamma+\Delta\gamma}{20}} \mathbf{x}_{q,m}^{(l)} + 10^{-\frac{\gamma+\Delta\gamma}{20}} \mathbf{x}_{q,m}^{(r)} - 2\,\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} \mathbf{x}_{q,m}^{(c)}\right]. \quad (9)$$

A similar expression holds for $\mathbf{y}_{q,m}$.

In order to compute the expectation across the final measure, we assume that the input signals are wide sense stationary stochastic processes across the duration of one segment (i.e. 386 ms). It follows that the third octave band envelope samples, $X_{q,m}$ and $Y_{q,m}$ are also samples of a stochastic process, due to the stochastic nature of the input signals, but also due to the random variables, $\Delta\gamma$ and $\Delta\tau$, introduced in the EC stage. A basic assumption of the original STOI measure is that speech intelligibility is related to the average sample correlation between the vectors $\mathbf{x}_{q,m}$ and $\mathbf{y}_{q,m}$ [7]. This, however, may be interpreted simply as an estimate of the correlation between the processes $X_{q,m}$ and $Y_{q,m}$:

$$\rho_{q,m} = \frac{E\left[(X_{q,m} - E\left[X_{q,m}\right])(Y_{q,m} - E\left[Y_{q,m}\right])\right]}{\sqrt{E\left[(X_{q,m} - E\left[X_{q,m}\right])^2\right] E\left[(Y_{q,m} - E\left[Y_{q,m}\right])^2\right]}}, \quad (10)$$

where the expectation is taken across both input signals, $\Delta\gamma$ and $\Delta\tau$. An estimate of this expectation across $N = 30$ envelope samples is given by:

$$\bar{\rho}_{q,m} = \frac{E_\Delta\left[\mathbf{x}_{q,m}^\intercal \mathbf{y}_{q,m}\right]}{\sqrt{E_\Delta\left[||\mathbf{x}_{q,m}||^2\right] E_\Delta\left[||\mathbf{y}_{q,m}||^2\right]}}, \quad (11)$$

where $E_\Delta[\cdot]$ denotes the expectation across $\Delta\gamma$ and $\Delta\tau$. A closed form approximation of this expectation is derived in Appendix A, and is given by:

$$\begin{aligned} &E_\Delta\left[\mathbf{x}_{q,m}^\intercal \mathbf{y}_{q,m}\right] \approx \\ &\quad (e^{2\beta} \mathbf{x}_{q,m}^{(l)\intercal} \mathbf{y}_{q,m}^{(l)} + e^{-2\beta} \mathbf{x}_{q,m}^{(r)\intercal} \mathbf{y}_{q,m}^{(r)}) e^{2\sigma_{\Delta\beta}^2} \\ &+ \mathbf{x}_{q,m}^{(r)\intercal} \mathbf{y}_{q,m}^{(l)} + \mathbf{x}_{q,m}^{(l)\intercal} \mathbf{y}_{q,m}^{(r)} - 2e^{\sigma_{\Delta\beta}^2/2} e^{-\omega^2\sigma_{\Delta\tau}^2/2} \times \\ &\quad \left\{ \left(e^{\beta} \mathbf{x}_{q,m}^{(l)\intercal} + e^{-\beta} \mathbf{x}_{q,m}^{(r)\intercal}\right) Re\left[\mathbf{y}_{q,m}^{(c)} e^{-j\omega\tau}\right] \right. \\ &+ \left. Re\left[e^{-j\omega\tau} \mathbf{x}_{q,m}^{(c)}\right]^\intercal \left(e^{\beta} \mathbf{y}_{q,m}^{(l)} + e^{-\beta} \mathbf{y}_{q,m}^{(r)}\right) \right\} \\ &+ 2\left(Re\left[\mathbf{x}_{q,m}^{(c)H} \mathbf{y}_{q,m}^{(c)}\right] + e^{-2\omega^2\sigma_{\Delta\tau}^2} Re\left[\mathbf{x}_{q,m}^{(c)\intercal} \mathbf{y}_{q,m}^{(c)} e^{-j2\omega\tau}\right]\right), \end{aligned} \quad (12)$$

where:

$$\beta = \frac{\ln(10)}{20}\gamma, \qquad \sigma_{\Delta\beta}^2 = \left(\frac{\ln(10)}{20}\right)^2 \sigma_{\Delta\gamma}^2. \quad (13)$$

The approximation in (12) stems from the approximation introduced in (6). A similar expression can be used to compute $E_\Delta\left[||\mathbf{x}_{q,m}||^2\right] = E_\Delta\left[\mathbf{x}_{q,m}^\intercal \mathbf{x}_{q,m}\right]$. This is obtained simply by replacing all occurrences of $\mathbf{y}$ in (12) with $\mathbf{x}$. In a similar manner, an expression for $E_\Delta\left[||\mathbf{y}_{q,m}||^2\right]$ can be obtained. This makes it possible to evaluate (11) in closed form.

In the same manner as in the STOI measure, we define the final measure to be the average of these correlation estimates:

$$\mathrm{DBSTOI} = \frac{1}{QM} \sum_{m=1}^{M} \sum_{q=1}^{Q} \bar{\rho}_{q,m}. \quad (14)$$

It should be noted that $\bar{\rho}_{q,m}$ is dependent on the parameters of the EC stage, γ and τ.

D. Determination of γ and τ

As stated, the parameters γ and τ are determined such as to maximize predicted intelligibility, i.e. (14). These parameters are determined independently for each estimated correlation coefficient, $\bar{\rho}_{q,m}$, i.e. for each time unit m and third octave band q. The values of $\gamma = \gamma_{k,m}$ and $\tau = \tau_{k,m}$ are therefore held constant for all frequency bins, k, within one third octave band, q, and for all N envelope samples, m, within one set of envelope vectors, $\{\mathbf{x}_{q,m},\ \mathbf{y}_{q,m}\}$.

The values of γ and τ are found separately for each estimated correlation coefficient, by maximizing the correlation:

$$\bar{\rho}_{q,m} = \max_{\gamma,\tau} \ \bar{\rho}_{q,m}(\gamma, \tau). \quad (15)$$

It has not been possible to find a simple analytical procedure for solving this optimization problem. Instead, an approximately optimal solution is found by evaluating (15) for a range of combinations of γ and τ. In practice, we search all combinations of an evenly spaced range of 100 τ-values from -1 ms to $+1$ ms and an evenly spaced range of 40 γ values from -20 dB to $+20$ dB. On top of the mentioned (γ,τ)-combinations, the correlation is also estimated for each ear individually (a "better-ear option"), corresponding to $\gamma = \pm\infty$. These estimates are as well included in the search of the maximum in (15). Preliminary experiments have indicated that the quality of the output is not highly sensitive to the searched range of (γ,τ)-combinations. For applications with scarce computational resources, the cost of computing the measure can be lowered by more coarsely searching these variables. The computational cost of the DBSTOI measure is compared to that of the STOI measure in Table I. It should, however, be noted that the cost of both methods depends on the choice of parameters and can most likely be decreased significantly by implementing them a low-level language.

A noteworthy special case of the DBSTOI measure arises for diotic stimuli, where $\mathbf{x}_{q,m}^{(l)} = \mathbf{x}_{q,m}^{(r)} = \mathbf{x}_{q,m}^{(c)}$ and $\mathbf{y}_{q,m}^{(l)} =$

TABLE I
TIME SPENT PRODUCING A SCORING OF 100 SECONDS OF WHITE NOISE ON A LENOVO W530 WITH AN INTEL CORE I7-3820QM, 2.7 GHZ. THE AUTHORS' OWN MATLAB IMPLEMENTATION OF THE DBSTOI MEASURE WAS USED, WHILE A STOI MEASURE IMPLEMENTATION WAS PROVIDED BY THE AUTHORS OF [7].

Algorithm	STOI	DBSTOI
Time	5.3 s	62.2 s

$\mathbf{y}_{q,m}^{(r)} = \mathbf{y}_{q,m}^{(c)}$. This implies that $\mathbf{x}_{q,m}^{(c)}$ and $\mathbf{y}_{q,m}^{(c)}$ are real-valued. Therefore, (12) simplifies to:

$$E_\Delta\left[\mathbf{x}_{q,m}^\intercal \mathbf{y}_{q,m}\right] \approx \Big(\left(e^{2\beta}+e^{-2\beta}\right)e^{\sigma_{\Delta\beta}^2}+2 -42e^{\sigma_{\Delta\beta}^2/2}e^{-\omega^2\sigma_{\Delta\tau}^2/2}\left(e^{\beta}+e^{-\beta}\right)Re\left[e^{-j\omega\tau}\right]+2 +e^{-2\omega^2\sigma_{\Delta\tau}^2}Re\left[e^{-j2\omega\tau}\right]\Big)\mathbf{x}_{q,m}^{(l)\intercal}\mathbf{y}_{q,m}^{(l)}. \quad (16)$$

Inserting this in (11), it can be verified that the entire τ- and β-dependent factor cancels, because the same factor appears in the denominator. This implies that the DBSTOI measure simplifies to the monaural (modified) STOI measure for diotic signals.

III. EXPERIMENTAL DATA

We evaluate the performance of the proposed measure by comparing it to the results of four listening experiments. In this section we describe these experiments.

The first two experiments make it possible to investigate the performance of the proposed measure in comparison with other measures of intelligibility. The third and fourth experiments make it possible to investigate the performance of the DBSTOI measure in conditions with both binaural advantage and processing. The conditions of experiments 2–4 are summarized in Table II (Experiment 1 is excluded due to the large number of conditions and the fact that it is thoroughly documented in [48]).

A. Experiment 1: Diotic Presentation and Ideal Time Frequency Segregation

This data set was collected as part of a study on ITFS [48], but has kindly been made available for evaluation of the present work. Subjects were presented with noisy and processed sentences from the Dantale II corpus [49]. After each sentence, the subjects were requested to repeat as many words as possible. The experimenter marked the correctly repeated words. Sentences were presented diotically via headphones together with one of four different interferers: Speech Shaped Noise (SSN), café noise, bottling factory noise and car interior noise. Sentences mixed with each noise type were ITFS processed with Ideal Binary Masks (IBMs) at 8 different threshold values. Furthermore, sentences mixed with each noise type, excluding SSN, were ITFS processed with Target Binary Masks (TBMs) at 8 different threshold values. Each combination of noise and processing was presented at 3 SNRs and with two sentences for each. The experiment was carried out with 15 normal hearing Danish speaking subjects. This resulted in the collection of results for 15 subjects $\times$ 7 noise/mask combinations $\times$ 8 thresholds $\times$ 3 SNRs $\times$ 2 repetitions $=$ 5040 sentences. See [48] for a detailed description of the experimental procedure.

TABLE II
SUMMARY OF EXPERIMENTS 2–4. FOR DETAILS SEE THE TEXT.

Cond.	Interferer type	Interferer location	Proc.
2.1	SSN	$-160°$	–
2.2	SSN	$-115°$	–
2.3	SSN	$-80°$	–
2.4	SSN	$-40°$	–
2.5	SSN	$20°$	–
2.6	SSN	$0°$	–
2.7	SSN	$40°$	–
2.8	SSN	$80°$	–
2.9	SSN	$140°$	–
2.10	SSN	$180°$	–
3.1	SSN	$-115°$	ITFS
3.2	SSN	$0°$	ITFS
3.3	SSN	$20°$	ITFS
3.4	Bottling factory	$-115°$	–
3.5	Bottling factory	$0°$	–
3.6	Bottling factory	$20°$	–
3.7	Bottling factory	$-115°$	ITFS
3.8	Bottling factory	$0°$	ITFS
3.9	Bottling factory	$20°$	ITFS
4.1	SSN	isotropic	–
4.2	SSN	$\{-115°,180°,115°\}$	–
4.3	ISTS	$\{-115°,180°,115°\}$	–
4.4	SSN	$\{30°,180°\}$	–
4.5	ISTS	$\{30°,180°\}$	–
4.6	SSN	isotropic	Beamforming
4.7	SSN	$\{-115°,180°,115°\}$	Beamforming
4.8	ISTS	$\{-115°,180°,115°\}$	Beamforming
4.9	SSN	$\{30°,180°\}$	Beamforming
4.10	ISTS	$\{30°,180°\}$	Beamforming

The original STOI measure has been shown to correlate well with the results of this experiment [7]. We include it in this study in order to investigate the impact of the differences between the STOI and DBSTOI measures for diotic stimuli.

B. Experiment 2: A Single Source of SSN in the Horizontal Plane

Speech intelligibility was measured in the condition of a frontal speaker masked by a single SSN interferer in the horizontal plane [44]. An anechoic environment was simulated binaurally by use of the CIPIC HRTFs [50] and the result was presented via Sennheiser HDA200 headphones at a comfortable level. Sentences from the Danish Dantale II material were used as target signals while SSN was generated by filtering Gaussian noise to have the same long time spectrum as these sentences. Speech intelligibility was measured for 10 interferer angles, each for 6 SNRs. The SNRs were equally spaced by 3 dB, centred around a rough estimate of the SRT for each condition. Sentences were presented one at a time and the subject was requested to repeat as many words of each sentence as possible. The experimenter marked the correctly repeated words. Three sentences were presented for each combination of interferer position and SNR. The experiment was carried out for 10 normal hearing Danish speaking subjects. In total, results were collected from the presentation of 10 subjects $\times$ 10 interferer positions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1800 sentences. The conditions of Experiment 2 are summarized in Table II.

The conditions of Experiment 2 contain no processing and are therefore applicable to a range of existing binaural SIP methods. We include it in this study to allow for a comparison of the DBSTOI measure with existing measures.

C. Experiment 3: A Single Interferer in the Horizontal Plane and Ideal Time Frequency Segregation

This experiment measured intelligibility in 9 conditions with a frontal speaker masked by a single interferer. Conditions 1–3 included an SSN masker at some position in the horizontal plane (as in Experiment 2) but with ITFS applied independently to the signals of each ear. This was done in as manner similar to that described in [51]: 1) a short-time DFT was applied to the speech and interferer signals in separation, prior to mixing, 2) DFT coefficients for which the SNR of the mixed signal was below 0 dB (i.e. where the magnitude of the interferer coefficient was larger than that of the target coefficient) were attenuated by 10 dB, 3) the signals were reconstructed. The finite attenuation of 10 dB was chosen to restrict the improvement in intelligibility (as ITFS can, otherwise, make speech fully intelligible regardless of SNR [48]). The interferer positions were chosen as a representative subset of those used in Experiment 2. The same interferer positions were used for conditions 4–6 and 7–9. In conditions 4–6, a "bottling factory noise" was used as interferer in place of SSN. This is a fluctuating noise type with more energy at higher frequencies than speech [52]. Conditions 7–9 were the same as 4–6 but with ITFS. The Dantale II corpus was also used in this experiment. The environment was simulated using the CIPIC HRTFs [50] and the signals presented via Sennheiser HDA200 headphones. The subjects were presented with one sentence at a time. After each sentence the subjects were requested to select the words they heard on a screen. For each of the five words in each sentence the subjects were offered a choice between 10 possible words and the option to pass (if the word had not been heard at all). In [53], [54], this procedure is shown to yield results almost identical to the verbal procedure used for collecting the results of Experiment 2. The experiment was run with 14 normal hearing Danish speaking subjects. In total, results were collected from the presentation of 14 subjects $\times$ 9 conditions $\times$ 6 SNRs $\times$ 3 repetitions $= 2268$ sentences. The conditions of Experiment 3 are summarized in Table II.

While experiments 1 and 2 investigate conditions with *either* processing or binaural advantage, Experiment 3 includes conditions with *both* processing and binaural advantage. Therefore, none of the mentioned existing SIP methods can be applied.

D. Experiment 4: Multiple Interferers and Beamforming

This experiment measured speech intelligibility in 10 somewhat more complex conditions relevant to the evaluation of hearing aids [44]. Conditions were again simulated binaurally by use of HRTFs and presented via Sennheiser HD 280 Pro headphones at a comfortable level. The Dantale II speech material was used as target speech material and the target speaker was placed in front of the subject in all conditions. Responses were given in the same way as in Experiment 2. In each condition, the subject was presented with speech at 6 different SNRs. In condition 1, the target was masked by cylindrically isotropic SSN. In condition 2 the target was masked by three sources of SSN positioned in the horizontal plane at azimuths of $110°$, $180°$ and $-110°$. Condition 3 was the same as condition 2, but used randomly selected segments of the International Speech Test Signal (ISTS) [55] instead of SSN as interferer. In condition 4 the target was masked by two sources of SSN positioned in the horizontal plane at azimuths of $30°$ and $180°$. Condition 5 was the same as condition 4, but again used segments of the ISTS instead of SSN as interferer. Conditions 6–10 were the same as conditions 1–5 but included 2-microphone beamforming as used in hearing aids. This was accomplished by using HRTFs measured from far field and to the two microphones of a behind-the-ear hearing aid, and combining the obtained signals with a time-invariant linear MVDR beamformer. The experiment was carried out with 10 normal hearing Danish speaking subjects. In total, results were collected from the presentation of 10 subjects$\times 10$ conditions$\times$ 6 SNRs $\times$ 3 repetitions $= 1800$ sentences. The conditions of Experiment 4 are summarized in Table II.

Experiment 4 is included to provide insights into the performance of the DBSTOI measure in acoustically varied scenes. Furthermore, beamforming is an increasingly important type of processing in e.g. hearing devices.

IV. Evaluation Procedure

Sec. III presents a substantial quantity of data. In each of the conditions, considered in the experiments, we can rate the intelligibility on an arbitrary scale (i.e. one that has an unknown relationship with speech intelligibility), using the proposed DBSTOI measure. In this section we present a range of tools which are used to compare the experimental results to these objective ratings of intelligibility, and thereby to quantify the performance of the DBSTOI measure.

A. Representation of Experimental Data

We represent the results of the described listening experiments either in terms of the average fraction of correctly repeated words or in terms of SRTs. We define the SRT as the SNR at which a subject is able to correctly repeat 50% of words. We determine this point from the measured data by maximum-likelihood-fitting a logistic function [56]:

$$p(\text{SNR}) = \frac{1}{1 + e^{4 \cdot s_0 \cdot (\text{SRT} - \text{SNR})}}, \tag{17}$$

with respect to the parameters SRT and s_0, where s_0 is the slope of the function at SNR $=$ SRT.

B. Predicting the Fraction of Correct Words

The DBSTOI measure provides an output on an arbitrary scale. We assume that a monotonic relationship exists between the DBSTOI measure and actual intelligibility (i.e. the fraction

of words repeated correctly). In the proposal of the STOI measure, a logistic function was used to model this relationship [7]. The same procedure is followed here:

$$f(d) = \frac{100\%}{1 + e^{ad+b}}, \tag{18}$$

where d is the DBSTOI measure, $f(d)$ is the estimated fraction of correctly repeated words and a and b are free parameters which we fit by maximum likelihood, such as to provide the best possible predictions.

C. Prediction of SRTs

By calibrating the proposed measure to a reference condition with known SRT, we may directly predict SRTs for other conditions. First, the proposed measure is evaluated for the reference condition at SRT, in order to output a reference value. Assuming that the measure correlates well with intelligibility, this reference value can be assumed to correspond to the SRT in other conditions as well. We may therefore predict the SRT for another condition by evaluating the proposed measure for a sequence of different input SNRs which are chosen adaptively such that the output approaches the reference value (e.g. using bisection). The SNR at which this procedure converges is taken to be an estimate of the SRT.

D. Measures of Prediction Accuracy

Whenever comparing listening test results and corresponding predictions, we rely on the following three performance statistics. Let x_i be the experimentally measured intelligibility (either fraction of correctly repeated words or the SRT) and y_i be corresponding predicted intelligibility, for conditions $i = 1, \ldots, I$. The performance statistics are then given by:

1) Sample standard deviation:

$$\sigma = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(y_i - x_i)^2}. \tag{19}$$

2) Pearson correlation:

$$\rho = \frac{\sum_{i=1}^{I}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{I}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{I}(y_i - \mu_y)^2}}, \tag{20}$$

where $\mu_x = \frac{1}{I}\sum_{i=1}^{I} x_i$ and $\mu_y = \frac{1}{I}\sum_{i=1}^{I} y_i$.

3) Kendall rank correlation [57][4]:

$$\phi = \frac{c_\mathrm{c} - c_\mathrm{d}}{\frac{1}{2}I(I-1)}, \tag{21}$$

where c_c is the number of concordant pairs, i.e. the number of unique tuples, (i, j), such that $(x_i > x_j) \wedge (y_i > y_j) \vee (x_i < x_j) \wedge (y_i < y_j)$, and c_d is the number of discordant pairs, i.e. the number of unique tuples, (i, j), such that $(x_i > x_j) \wedge (y_i < y_j) \vee (x_i < x_j) \wedge (y_i > y_j)$.

[4]Conventionally, τ is used for the Kendall rank correlation. We do not follow this convention because τ is used for a different purpose throughout the paper.

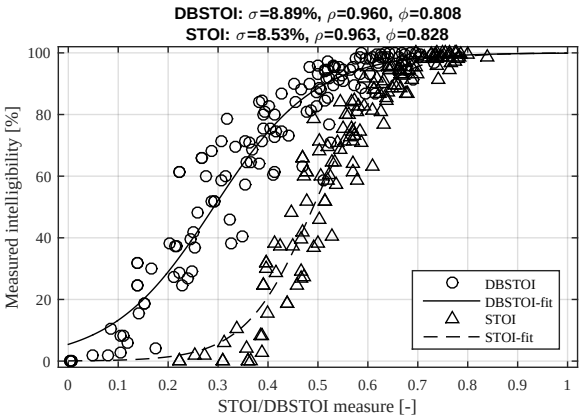

Fig. 3. Measured intelligibility for the conditions/SNRs of Experiment 1 compared with the DBSTOI measure and the STOI measure. A logistic function was maximum likelihood fitted for each method. The statistics σ and ρ were computed by comparing the shown data points with the predictions made by the fitted logistic curves.

V. Results

In this section we apply the proposed measure to yield predictions of the results of the listening experiments described in Sec. III.

A. Diotic Presentation and Ideal Time Frequency Segregation

We first consider the data of Experiment 1 (Sec. III-A), which investigated the intelligibility of diotic noisy speech processed with ITFS. Due to the diotic nature of the signals, we can obtain predictions by use of the original STOI measure. The STOI measure was designed with the main focus of predicting the impact of TF weighting such as ITFS, and has been shown to perform very well at doing so [7]. We may also obtain predictions with the DBSTOI measure, by simply using the same signals as inputs to the left and right channels of the measure (corresponding to presenting the same signals on the left and right ears of a subject). This allows us to investigate whether the desirable performance of the STOI measure is retained in the DBSTOI measure in spite of the introduced modifications.

Predictions with both the STOI and DBSTOI measures were based on sequences of 30 Dantale sentences. Measured intelligibility was obtained for each condition/SNR by averaging the fraction of correct words across subjects and repetitions. The results are shown in Fig. 3. It is evident that there is a strong relationship between measured intelligibility and both STOI and DBSTOI measures. The three statistics, shown at the top of Fig. 3, indicate that the STOI measure performs marginally better than the proposed measure. In Sec. II-D, it was shown that the effect of the EC stage cancels for diotic stimuli. Therefore, the differences between the STOI measure and the DBSTOI measure, in these conditions, stem only from the modifications introduced in the STOI measure, and not from the extension with an EC stage[5]. The similar performance

[5]The slight decrease in performance may stem from either the use of "power envelopes" rather than conventional envelopes, or from the fact that the DBSTOI measure does not include a clipping mechanism such as the one used in the original STOI [7].

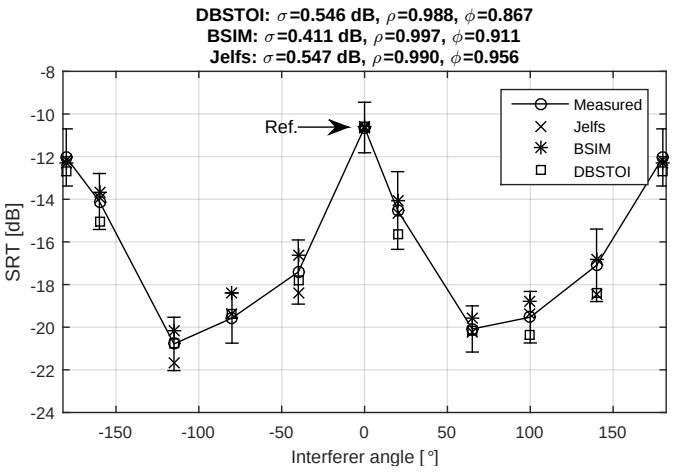

Fig. 4. SRTs estimated from the results of Experiment 2 along with the predictions of three different measures of speech intelligibility. The measures are all calibrated to the S_0N_0-condition (where both speech and interference comes from the front). The error bars show the standard deviation of SRTs among subjects for the measured results. The reference condition is marked by an arrow.

of the two methods indicate that the modifications applied to the STOI measure to obtain the proposed measure do not strongly impair the performance in a task which is central to the original STOI measure.

B. A Single Source of SSN in the Horizontal Plane

We now consider the results of Experiment 2 (Sec. III-B), which involved frontal speech masked by a single source of SSN in the horizontal plane. The conditions of this experiment allow subjects to obtain a binaural advantage but include no processing. The STOI measure is unsuitable for predicting these results as it relies on monaural/diotic signals. Instead we compare the predictions of the DBSTOI measure to predictions of two existing methods which do consider binaural advantage (but which do not allow for non-linearly processed signals).

Firstly, we compare to the BSIM [33], which is a binaural measure obtained by combining the EC stage with the SII. The BSIM requires knowledge of binaural speech and interference in separation and outputs a number between 0 and 1. It can be used to predict SRTs following the procedure described in Sec. IV-C. This was done by calibrating to the S_0N_0-condition (the condition in which speech and interference sources are co-located in front of the listener). When carrying out predictions with the BSIM, SSN was used for both target and interferer signals (as the method is also evaluated in this manner in [33]). The BSIM was implemented following the description given in [33].

Secondly, we compare to a method described by Jelfs et. al. in [37]. This method uses an SII-like scheme to predict the SRT for the better ear, and a correlation-based model for predicting the additional binaural advantage. The method outputs SRTs but with a significant offset [37], and was therefore calibrated by shifting all outputs by an additive constant, chosen such as to yield correct predictions in the S_0N_0-condition. For this method, SSN was also used as both target and masker (as the method is also evaluated in this manner in [37]). An implementation of this method was kindly provided by the authors of [37].

The DBSTOI measure was used to carry out SRT predictions as described in Sec. IV-C after being calibrated to the S_0N_0-condition. The predictions were based on a clean signal composed of 30 concatenated Dantale II sentences and an SSN interferer of the same length, both convolved with appropriate HRTFs. Signals of the same length were used for the methods of comparison.

Fig. 4 shows the results of measurements along with the predictions of the three methods. It is evident that all methods produce very accurate predictions in all conditions, especially considering the standard deviations on the measurements, and the fact that measurements of binaural advantage can vary by several dB from one study to another [29]. The statistics on top of Fig. 4 indicate that the DBSTOI measure produces predictions which are slightly less accurate than those of the BSIM. This conclusion, however, should be viewed in the light of the facts that 1) the DBSTOI measure does not have access to the interferer in separation (i.e. it does not assume that the speech signal is merely degraded by an additive interferer), while the two existing measures require access to speech and interference in separation, and 2) the DBSTOI measure uses actual speech signals as input while the two existing measures use SSN as both speech and interferer signals.

The DBSTOI measure predictions in isolation, indicate that the measure can indeed predict binaural advantage. This suggests that the modifications introduced in the EC stage, to make it handle non-linearly processed signals, have not severely degraded its performance.

C. A Single Interferer in the Horizontal Plane and Ideal Time Frequency Segregation

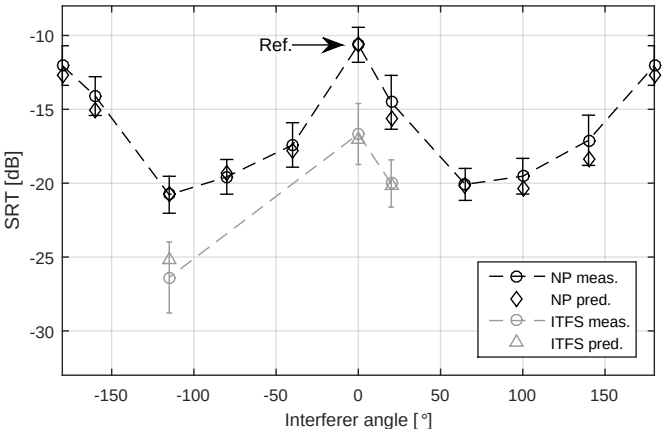

Fig. 5. Comparison between measured and predicted SRTs for all 10 conditions of Experiment 2 (denoted NP) and conditions 1–3 from Experiment 3 (denoted ITFS). The conditions are shown together because they all use SSN for masking. SRTs were estimated from the measured results for each subject individually for each condition. The results, averaged across subjects, are shown as dotted lines with standard deviations. The SRT of one reference condition was used to make SRT predictions for the other conditions. The reference condition is marked by an arrow.

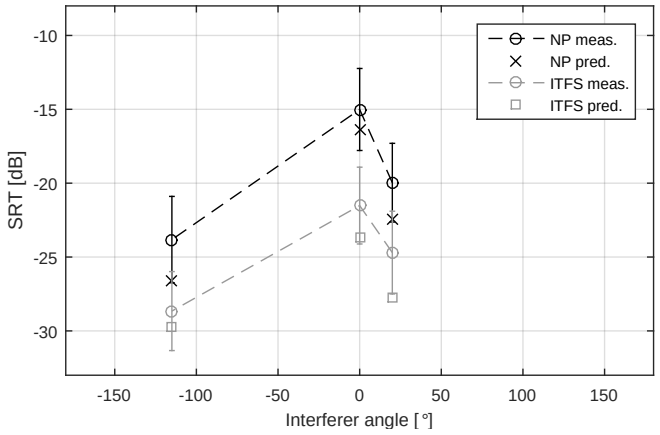

Fig. 6. The results of conditions 4–6 (NP) and 7–9 (ITFS) of Experiment 3. These conditions are shown together because they all use bottling factory noise for masking. SRTs were estimated from the measured results for each subject individually for each condition. The results, averaged across subjects, are shown as dotted lines with standard deviations. The SRT were predicted with the proposed measure using the reference condition shown in Fig. 5.

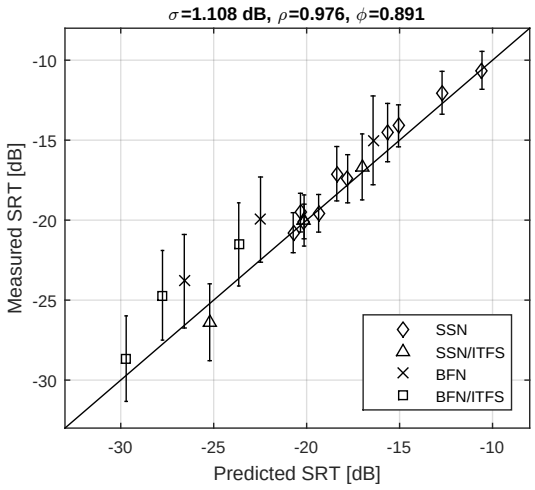

Fig. 7. Comparison of measured and predicted SRTs from all conditions of Experiment 2 and Experiment 3 with both SSN and bottling factory noise (BFN) interferers. SRTs were estimated separately for each subject for each conditions. The shown SRT values are averaged across subjects with standard deviations shown as error bars. The diagonal line represents perfect predictions. Measures of accuracy in the top of the plot are computed by comparing the measured and predicted SRTs.

In this section we consider Experiment 3 (Sec. III-C), which is similar to Experiment 2, but involves conditions with ITFS and masking by either SSN or bottling factory noise (see Table II). We remark that the STOI measure is not applicable to predicting the intelligibility of the conditions in Experiment 3, as binaural advantage is involved. At the same time, the BSIM or the method by Jelfs et. al. are not applicable, as non-linear processing in the form of ITFS is involved. We therefore present results from the DBSTOI measure alone. This serves as a study of the prediction performance of the DBSTOI measure in conditions with *both* binaural advantage *and* processing. Predictions of SRTs were carried out as described in Sec. IV-C, with calibration to the same condition as in Sec. V-B. The predictions were based on 30 concatenated Dantale II sentences and interferer signals of the same length.

TABLE III
PREDICTED AND MEASURED BINAURAL ADVANTAGE.

	Meas. bin. adv.	Pred. bin.adv.
SSN-NP	10.2 dB	10.2 dB
SSN-ITFS	9.7 dB	8.2 dB
BFN-NP	8.8 dB	10.2 dB
BFN-ITFS	7.1 dB	6.1 dB

Fig. 5 shows the DBSTOI predictions and the average measured SRTs for the conditions with an SSN interferer (conditions 1–3) together with the measured and predicted SRTs from Experiment 2. It is apparent that ITFS leads to a large advantage in terms of speech intelligibility, as expected. Furthermore, the SRTs in the ITFS-conditions are predicted with an error of less than a standard deviation of the measurements. This indicates that the DBSTOI measure can account for the joint effect of binaural advantage and processing by ITFS. The results of the conditions with bottling factory noise masking are shown in Fig. 6. The large standard deviations of the measurements indicate that there are large differences between subjects for this masker type. Furthermore, predictions are biased downwards by 2–3 dB. This may be caused by the fact that predictions were made with a reference condition were another type of masker (SSN) was used. However, the relative differences in SRTs between the conditions appear to be rather accurately predicted.

A noteworthy feature of the results relates to binaural advantage. We define binaural advantage in this experiment as the difference in SRT between the S_0N_0-condition and the S_0N_{-115}-condition. Predicted and measured values of binaural advantage for SSN and bottling factory noise with and without ITFS are shown in Table III. From this table it can be seen that for both interferer types, the binaural advantage decreases, when the signals are processed with ITFS. A possible explanation of this is that ITFS improves the spectral features of the signal but fails to restore the phase, which is important for binaural advantage. It can be noted that this decrease in binaural advantage is indeed predicted by the DBSTOI measure. The decrease is, however, predicted to be larger than the actual values.

Fig. 7 compares predictions and measurements for both masker types. An average prediction error of slightly over 1 dB is obtained: an error dominated by the bias of predictions with bottling factory masking.

D. Multiple Interferers and Beamforming

Lastly, we consider the results of Experiment 4 (Sec. III-D), which involves multiple interferers and beamforming. Predictions were made with the DBSTOI measure on the basis of 30 concatenated Dantale II sentences. Fig. 8 compares measured intelligibility and outputs of DBSTOI. In most of the conditions there appears to be a very strong relationship between the DBSTOI score and the measured results. Especially, the impact of beamforming is well predicted. At low SNRs, there is a discrepancy between predictions made

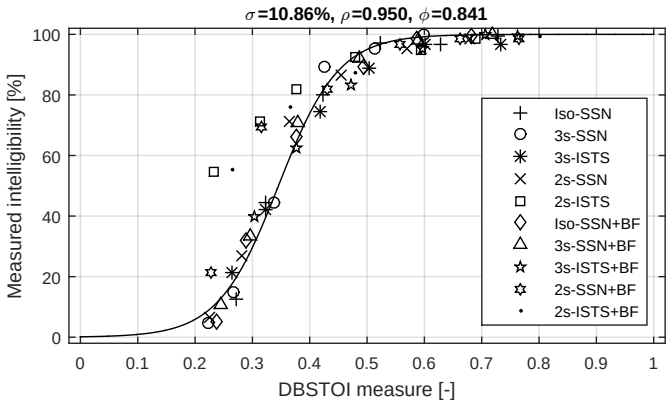

Fig. 8. Measured intelligibility, averaged across subjects, for the conditions/SNRs of Experiment 4 compared with the predictions of the proposed measure. The shown logistic curve was maximum likelihood fitted to the data. The statistics σ and ρ were computed by comparing the shown data points with the predictions made by the fitted logistic curve. The legend displays the layout of interferers: isotropic (`Iso`), 3 sources in 110°, 180° and −110° (`3s`), 2 sources in 30° and 180° (`2s`). Furthermore it shows the interferer type and whether or not beamforming was included (`BF`).

for some of the conditions with two interferers, and the remaining conditions. This may indicate that the DBSTOI measure does not provide fully consistent predictions when making comparisons across different complicated acoustical scenes. If this is the case, the measure should be separately calibrated to acoustically significantly different conditions (e.g. conditions with different numbers of maskers). However, this topic is outside the scope of the present work and has yet to be investigated in depth.

VI. Conclusions

We have proposed a binaural speech intelligibility measure based on combining the Short-Time Objective Intelligibility (STOI) measure with an Equalization-Cancellation (EC) stage. The proposed measure excels by being capable of predicting the impact of both binaural advantage and non-linear signal processing simultaneously. This makes the measure a potentially powerful tool for the development of signal processing devices which present speech binaurally to a user. The measure outputs ratings of intelligibility on an arbitrary scale, which is useful for comparing e.g. different speech processing algorithms. The measure can be calibrated such as to make direct predictions of Speech Reception Thresholds (SRTs) or of the percentage of correctly understood words. This is useful for predicting the outcome of listening experiments. The accuracy of the measure was investigated by comparing predictions with the results of four listening experiments. The measure was shown to predict the effect of Ideal Time Frequency Segregation (ITFS), with an accuracy similar to that of the original STOI measure. The measure was also shown to predict the effect of binaural advantage, in case of masking by a single point noise source, with an accuracy similar to that of two existing binaural methods. Furthermore, the measure was shown to accurately predict the effect of simultaneous binaural advantage and ITFS. Lastly, the measure was shown to predict well the effect of beamforming, in conditions with multiple interferers. The measure, however, showed some discrepancies when comparing between different conditions with multiple interferers. A detailed investigation of this issue is left for future work. The broad domain in which the measure is applicable calls for further investigation of performance in different conditions, e.g. different types of processing/distortion, different types of interference and different acoustical conditions. Finally, with respect to the particular application of hearing aid signal processing, future work could be directed towards incorporating a hearing loss model into the DBSTOI measure.

Appendix A

We derive an expression for the expectation of $\mathbf{x}_{q,m}^{\intercal}\mathbf{y}_{q,m}$ under the gaussian random variables, $\Delta\gamma$ and $\Delta\tau$, introduced in the EC stage. We make the assumption that all energy in each third octave band is contained at the center frequency. From (9), we obtain:

$$\begin{aligned}
&E_{\Delta}\left[\mathbf{x}_{q,m}^{\intercal}\mathbf{y}_{q,m}\right] \approx \\
&\quad E_{\Delta}\Big[e^{2\beta+2\Delta\beta}\mathbf{x}_{q,m}^{(l)\intercal}\mathbf{y}_{q,m}^{(l)} + e^{-2\beta-2\Delta\beta}\mathbf{x}_{q,m}^{(r)\intercal}\mathbf{y}_{q,m}^{(r)} \\
&\quad + \mathbf{x}_{q,m}^{(l)\intercal}\mathbf{y}_{q,m}^{(r)} + \mathbf{x}_{q,m}^{(r)\intercal}\mathbf{y}_{q,m}^{(l)} \\
&\quad - 2(e^{\beta+\Delta\beta}\mathbf{x}_{q,m}^{(l)\intercal} + e^{-\beta-\Delta\beta}\mathbf{x}_{q,m}^{(r)\intercal})\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{y}_{q,m}^{(c)}\right] \\
&\quad - 2\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)\intercal}\right](e^{\beta+\Delta\beta}\mathbf{y}_{q,m}^{(l)} + e^{-\beta-\Delta\beta}\mathbf{y}_{q,m}^{(r)}) \\
&\quad + 4\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)\intercal}\right]\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{y}_{q,m}^{(c)}\right]\Big], \quad (22)
\end{aligned}$$

where β is given by (13) and:

$$\Delta\beta = \frac{\ln(10)}{20}\Delta\gamma. \quad (23)$$

Because the expectation operator is linear we may evaluate the terms of (22) independently. Furthermore, since $\Delta\beta$ is zero-mean normally distributed with variance $\sigma_{\Delta\beta}^2$, we have:

$$\begin{aligned}
E_{\Delta}[e^{\beta+\Delta\beta}] &= e^{\beta}E_{\Delta}[e^{\Delta\beta}] = e^{\beta}e^{\sigma_{\Delta\beta}^2/2}, \\
E_{\Delta}[e^{-\beta-\Delta\beta}] &= e^{-\beta}E_{\Delta}[e^{-\Delta\beta}] = e^{-\beta}e^{\sigma_{\Delta\beta}^2/2}, \\
E_{\Delta}[e^{2\beta+2\Delta\beta}] &= e^{2\beta}E_{\Delta}[e^{2\Delta\beta}] = e^{2\beta}e^{2\sigma_{\Delta\beta}^2}, \\
E_{\Delta}[e^{-2\beta-2\Delta\beta}] &= e^{-2\beta}E_{\Delta}[e^{-2\Delta\beta}] = e^{-2\beta}e^{2\sigma_{\Delta\beta}^2}. \quad (24)
\end{aligned}$$

Using the above allows for computing the expectation of terms 1–4 in (22). For terms 5–6, we may make use of the fact that $E_{\Delta}[f(\Delta\beta)g(\Delta\tau)] = E_{\Delta}[f(\Delta\beta)]E_{\Delta}[g(\Delta\tau)]$ because $\Delta\beta$ and $\Delta\tau$ are statistically independent (where $f, g : \mathbb{C} \to \mathbb{C}$ are any functions). Furthermore, we note that $\mathrm{Re}[ab] = \mathrm{Re}[a]\mathrm{Re}[b] - \mathrm{Im}[a]\mathrm{Im}[b]$ for any $a, b \in \mathbb{C}$. This allows us to write:

$$\begin{aligned}
\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)}\right] &= \mathrm{Re}\left[e^{-j\omega_q\Delta\tau}\right]\mathrm{Re}\left[e^{-j\omega_q\tau}\mathbf{x}_{q,m}^{(c)}\right] \\
&\quad - \mathrm{Im}\left[e^{-j\omega_q\Delta\tau}\right]\mathrm{Im}\left[e^{-j\omega_q\tau}\mathbf{x}_{q,m}^{(c)}\right]. \quad (25)
\end{aligned}$$

By (24) and by symmetry, respectively, we have:

$$\begin{aligned}
E_{\Delta}\left[\mathrm{Re}\left[e^{-j\omega_q\Delta\tau}\right]\right] &= e^{-\omega_q^2\sigma_{\Delta\tau}^2/2}, \\
E_{\Delta}\left[\mathrm{Im}\left[e^{-j\omega_q\Delta\tau}\right]\right] &= 0, \quad (26)
\end{aligned}$$

which in turn leads to:

$$E_\Delta\left[\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)}\right]\right] = e^{-\omega_q^2\sigma_{\Delta\tau}^2/2}\mathrm{Re}\left[e^{-j\omega_q\tau}\mathbf{x}_{q,m}^{(c)}\right]. \tag{27}$$

To evaluate the last term in (22), we note that $\mathrm{Re}\left[a\right]\mathrm{Re}\left[b\right] = \frac{1}{2}\left(\mathrm{Re}\left[ab\right] + \mathrm{Re}\left[a^*b\right]\right)$, where $a, b \in \mathbb{C}$ and * represents complex conjugation:

$$\begin{aligned} &E_\Delta\left[4\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)\mathsf{T}}\right]\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{y}_{q,m}^{(c)}\right]\right] = \\ &2\left(E_\Delta\left[\mathrm{Re}\left[e^{-j2\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)\mathsf{T}}\mathbf{y}_{q,m}^{(c)}\right]\right] + \mathrm{Re}\left[\mathbf{x}_{q,m}^{(c)\mathrm{H}}\mathbf{y}_{q,m}^{(c)}\right]\right) = \\ &2\left(e^{-2\omega_q^2\sigma_{\Delta\tau}^2}\mathrm{Re}\left[e^{-j2\omega_q\tau}\mathbf{x}_{q,m}^{(c)\mathsf{T}}\mathbf{y}_{q,m}^{(c)}\right] + \mathrm{Re}\left[\mathbf{x}_{q,m}^{(c)\mathrm{H}}\mathbf{y}_{q,m}^{(c)}\right]\right), \end{aligned} \tag{28}$$

where $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose. By inserting (24), (27) and (28) into (22) (with appropriate substitution of variables), one reaches (12) as desired.

ACKNOWLEDGMENT

This work was funded by the Oticon Foundation and the Danish Innovation Foundation. We thank the authors of [37] for making an implementation of their method available.

REFERENCES

[1] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016.

[2] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, Oct. 1929.

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

[4] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[5] S. van Wijngaarden, "The speech transmission index after four decades of development," *Acoustics Australia*, vol. 40, no. 2, pp. 134–138, Aug. 2012.

[6] *Sound System Equipment. Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index*, IEC Std.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[8] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[9] 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.

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

[11] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO-2009)*. Glasgow, Scotland: EURASIP, Aug. 2009.

[12] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[13] K. D. Kryter, "Methods for the Calculation of and Use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[14] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing: Physiology, Psychoacoustics and Models*, D. Pressnitzer, A. de Cheveigne, S. McAdams, and L. Collet, Eds. Springer Verlag, 2004, pp. 314–320.

[15] *Methods for Calculation of the Speech Intelligibility Index*, ANSI Std. S3.5-1997, 1997.

[16] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

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

[18] I. M. Noordhoek and R. Drullmann, "Effect of reducing temporal intensity modulations on sentence intelligibility," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 498–502, Jan. 1997.

[19] V. Hohmann, "The effect of multichannel dynamic compression on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1191–1195, Feb. 1995.

[20] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interupted noise," *J. Acoust. Soc. Am.*, vol. 126, no. 6, pp. 3236–3245, Dec. 2009.

[21] F. Dubbelboer and T. Houtgast, "A detailed study on the effects of noise on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 122, no. 5, pp. 2865–2871, Nov. 2007.

[22] 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. Am.*, vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

[23] 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. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[24] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

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

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

[27] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Jul. 2014.

[28] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363–1374, Sep. 2014.

[29] A. W. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[30] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[31] ——, "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory Volume II*, J. V. Tobias, Ed. New York: Academic Press, 1972, pp. 371–462.

[32] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[33] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[34] S. J. van Wijngaarden and R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4514–4523, Jun. 2008.

[35] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sep. 2010.

[36] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[37] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[38] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 218–231, Jan. 2012.

[39] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the intelligibility of reverberation on binaural speech intelligibility in noise and in quiet," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2999–3012, Nov. 2011.

[40] J. Rennies, A. Warzybok, T. Brand, and B. Kollmeier, "Modelling the effects of a single reflection on binaural speech intelligibility," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1556–1567, Mar. 2014.

[41] J. F. Culling, M. L. Hawley, and R. Y. Litovsky, "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 1057–1065, Aug. 2004.

[42] ——, "Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]," *J. Acoust. Soc. Am.*, vol. 118, no. 1, p. 552, Jul. 2005.

[43] T. Leclère, M. Lavandier, and J. F. Culling, "Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. 3335–3345, Jun. 2015.

[44] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015.

[45] H. vom Hövel, "Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[46] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in coclear implants based on an intelligibility metric," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO-2012)*. Bucharest, Romania: EURASIP, Aug. 2012, pp. 504–508.

[47] 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. Am.*, vol. 130, no. 5, pp. 3013–3027, Nov. 2011.

[48] 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. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

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

[50] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proccedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2001.

[51] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[52] M. Vestergaard, "The eriksholm cd 01: Speech signals in various acoustical environments," 1998.

[53] E. R. Pedersen and P. M. Juhl, "Speech in noise test based on a ten-alternative forced choice procedure," in *Joint Baltic-Nordic Acoustics Meeting (BNAM)*, Odense, Denmark, Jun. 2012.

[54] ——, "User-operated speech in noise test: Implementation and comparison with a traditional test," *International Journal of Audiology*, vol. 53, no. 5, pp. 336–344, May 2014.

[55] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, Dec. 2010.

[56] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, Jun. 2002.

[57] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.

Asger Heidemann Andersen received the B.Sc. degree in electronics & IT and the M.Sc. degree in wireless communication (*cum laude*) from Aalborg University, Aalborg, Denmark, in 2012 and 2014, respectively. He is currently employed at Oticon A/S while pursuing the Ph.D. degree at the speech and information processing section at Aalborg University under the supervision of J. Jensen, J. M. de Haan and Z.-H. Tan. His main research interests are prediction and measurement of speech intelligibility and speech enhancement with applications to hearing aids.

Jan Mark de Haan Jan Mark de Haan received the M.Sc. degree in Electrical Engineering and the Ph.D. degree in Applied Signal Processing from Blekinge Institute of Technology, Karlskrona, Sweden, in 1998 and 2004, respectively. From 1999 to 2004 he was a Ph.D. student with the Department of Applied Signal Processing, Blekinge Institute of Technology. In 2003 he was a visiting researcher at the Western Australia Telecommunication Research Institute, Perth, Australia. Since 2004 he is employed at Oticon A/S, Copenhagen, Danmark. His main interest are in acoustic signal processing and signal processing applications in hearing aids.

Zheng-Hua Tan (M'00-SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999.

He is an Associate Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark, which he joined in May 2001. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human-robot interaction, and machine learning. He has published extensively in these areas in refereed journals and conference proceedings. He has served as an Editorial Board Member/Associate Editor for Elsevier Computer Speech and Language, Elsevier Digital Signal Processing and Elsevier Computers and Electrical Engineering. He was a Lead Guest Editor for the IEEE Journal of Selected Topics in Signal Processing. He has served as a program co-chair, area and session chair, tutorial speaker and committee member in many major international conferences.

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 Signal and Information Processing (SIP), 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.