Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions

Asger Heidemann Andersen\textsuperscript{a,b,*}, Jan Mark de Haan\textsuperscript{a}, Zheng-Hua Tan\textsuperscript{b}, Jesper Jensen\textsuperscript{a,b}

\textsuperscript{a}Oticon A/S, Kongebakken 9, Smørum, Denmark  
\textsuperscript{b}Aalborg University, Department of Electronic Systems, Fredrik Bajers Vej 7B, Aalborg, Denmark

Abstract

Speech intelligibility prediction methods have recently gained popularity in the speech processing community as supplements to time consuming and costly listening experiments. Such methods can be used to objectively quantify and compare the advantage of different speech enhancement algorithms, in a way that correlates well with actual speech intelligibility. One such method is the short-time objective intelligibility (STOI) measure. In a recent publication, we proposed a binaural version of the STOI measure, based on a modified version of the equalization cancellation (EC) model. This measure was shown to retain many of the advantageous properties of the STOI measure, while at the same time being able to predict intelligibility correctly in conditions involving both binaural advantage and non-linear signal processing. The biggest prediction errors were found for conditions involving multiple spatially distributed interferers. In this paper, we report results for a new listening experiment including different mixtures of isotropic and point source noise. This exposes that the binaural STOI measure has a tendency to overestimate the intelligibility in conditions with spatially distributed interferers at low signal to noise ratios (SNRs). This condition-dependent error can make it difficult to compare intelligibility across different acoustical conditions. We investigate the cause of

*Corresponding author

Preprint submitted to Speech Communication June 3, 2018
this upward bias, and propose a correction which alleviates the problem. The modified method is evaluated with five datasets of measured intelligibility, spanning a wide range of realistic acoustic conditions. Within the tested conditions, the modified method yields very accurate predictions, and entirely alleviates the aforementioned tendency to overestimate intelligibility in conditions with spatially distributed interferers.

**Keywords:** speech intelligibility prediction, binaural hearing, speech enhancement

---

1. **Introduction**

Speech Intelligibility Prediction (SIP) algorithms aim to predict the intelligibility of noisy or degraded recordings of speech. Such algorithms were first studied early in the twentieth century as a means to quantify the intelligibility of speech transmitted via telephone [1, 2]. Recently, SIP has become an increasingly popular tool for objectively assessing intelligibility within applications such as speech enhancement [3, 4, 5, 6, 7, 8], architectural acoustics [9, 10], and telecommunications [11].

SIP algorithms can be separated in two groups: intrusive ones and non-intrusive ones. Intrusive SIP algorithms predict intelligibility using some representation of the degraded signal as well as a clean reference signal, while non-intrusive SIP algorithms require only the degraded signal. The focus of this paper is on intrusive SIP algorithms.

The first widely known SIP algorithm was the Articulation Index (AI), which assumes speech intelligibility to be proportional to a weighted sum of long-term Signal-to-Noise Ratios (SNRs) across different frequency bands [2]. A theoretical justification of the AI is provided in [12], which shows it to be an estimate of the channel capacity from Shannon’s information theory. A refined and standardized version of the AI is known as the Speech Intelligibility Index (SII) [13]. While the SII has proven successful in many respects, conditions exist, where it is unable to provide accurate predictions. Therefore, many al-
ternative SIP algorithms have been proposed, often inspired by the SII, but aiming to extend the domain of signals for which accurate predictions can be made (some overviews are provided by [14, 15, 9, 16, 17, 4, 11]). Here, we provide a brief overview of SIP algorithms which extend the domain of the SII in three different ways: by taking into account fluctuating interferers, non-linear processing and binaural advantage, respectively:

- **Fluctuating interferers**: Speech intelligibility may be substantially higher in the presence of modulated or interrupted noise, than in the presence of stationary noise at a similar level [18]. This suggests that humans are capable of effectively collecting information from short instants with high SNR. The SII relies on long-time signal statistics, and is therefore unable to take such an effect into account [19]. This ability is, however, important to consider, as many real-world noise signals are highly non-stationary (e.g. competing talkers). One solution to this problem is given by the Extended SII (ESII) [19, 20], which computes the SII in short time windows, and averages contributions across time. The ESII has shown high prediction accuracy in conditions with interrupted and sinusoidally modulated Speech Shaped Noise (SSN) at various frequencies [20]. The principle of averaging contributions across short time segments has repeatedly been used for the same purpose in other predictors [14, 8, 21]. An alternative model, proposed in [22], predicts intelligibility from the fraction of the time-frequency plane for which the SNR exceeds a fixed threshold.

- **Non-linear processing**: Since the SII bases predictions on SNRs, it is not applicable in conditions, where the noisy speech signal has been non-linearly processed or distorted, because the SNR is not well defined in this case. Since SIP is often used for evaluating the performance of non-linear speech enhancement algorithms, it is desirable to have SIP algorithms which can handle such processing. One such approach is the Coherence SII (CSII), which computes the SII based on coherence rather than SNR [23].
This measure has shown good prediction performance for distorted speech \cite{23} and noise reduction processing \cite{17}. Another approach is the frequency-weighted segmental SNR (fwSNRseg) \cite{14,15}, which uses the difference between clean and degraded signal sub-band envelopes as an estimate of the noise envelope, and further uses this to compute an SII-like measure. A somewhat different model is the speech-based Envelope Power Spectrum Model (sEPSM) \cite{24}, which is based on a modulation domain model of human hearing \cite{25}. The sEPSM performs well at predicting the impact of spectral subtraction algorithms \cite{21} and noisy speech transmitted via telephone \cite{11}. It has also been proposed in a short-time version which can handle fluctuating interferers \cite{21}. Recently, the Short-Time Objective Intelligibility (STOI) measure has gained considerable popularity, due to its simplicity and high performance in a range of key scenarios such as different noise types \cite{3}, noise reduction processing \cite{3,17}, hearing aid processing \cite{4} and noisy speech transmitted via telephone \cite{11}. The Extended STOI (ESTOI) measure \cite{26} has later been proposed with the aim of improving prediction performance in conditions with fluctuating interferers.

- **Binaural advantage:** The SII considers speech presented to one ear only or, similarly, identical speech signals presented to both ears. In real-world situations, a large binaural advantage may be obtained, because the acoustics of the head causes different sound signals to reach the left and right ears. Binaural advantage can have a large impact on speech intelligibility \cite{27}, and is increasingly considered in speech enhancement applications \cite{28,29}. It is therefore a highly desirable property of a SIP algorithm to predict this effect. The binaural advantage can be seen as originating from two separate sources: 1) the human head casts an acoustical shadow, that leads to Interaural Level Differences (ILDs), dependent on the source location with respect to the ears \cite{30}. In some acoustical conditions, this causes the SNR to be higher at one ear than at the other. The potential ability of the brain to listen to the ear with the higher SNR...
may lead to a considerable advantage, which we refer to as the "better-ear" advantage. 2) The distance between the ears causes Interaural Time Differences (ITDs) which depend on the location of the source \[30\]. Experiments have indicated that the brain can exploit ITDs to effectively segregate sound signals from different spatial locations \[31, 27\]. We refer to this as an advantage due to "binaural unmasking". The Equalization-Cancellation (EC) stage \[32, 33\] models binaural advantage, by assuming that the brain uses delays and gain adjustments to align any interferer components in the left and right ears, and thereafter subtracts the two signals from one-another to cancel the interferers. The Binaural Speech Intelligibility Measure (BSIM) \[34, 35\] combines the SII and the EC stage in order to predict intelligibility including binaural advantage. This is done by optimizing the delay and gain parameters in the EC stage to maximize the SII. A short-time version, the short-time BSIM (stBSIM), is proposed in \[35\]. An approach similar to that of the BSIM is followed in \[36\], with a short-time version following in \[37\]. A different approach comes from \[38, 39, 40\], which proposes to compute the SII from the better ear, and adding binaural advantage as estimated by an expression from \[41\]. A binaural version of the STOI measure is proposed in \[42\] and further improved in \[43, 44\]. The improved version is referred to as the Deterministic Binaural STOI (DBSTOI) measure. This is based on extending the STOI measure with an EC stage, similar to how the BSIM extends the SII. The DBSTOI measure retains many of the favorable properties of the STOI measure while being able to predict binaural advantage \[44\]. A binaural version of the multiresolution sEPSM (mr-sEPSM), the Binaural sEPSM (B-sEPSM), is proposed in \[45\]. This too, is based on an EC stage. Both the DBSTOI measure and the B-sEPSM are short-time methods, and both can handle non-linearly processed speech.

The evaluation of the DBSTOI measure in \[44\] generally shows high prediction accuracy in the tested conditions. The largest prediction errors are seen for
conditions with spatially distributed interferers such as isotropic noise or multiple point-like interferers in different spatial locations. It is important that a binaural SIP algorithm works well in conditions with spatially distributed interferers which are common in practice (e.g. at a cocktail party or in a reverberant room). Furthermore, many speech enhancement algorithms modify the binaural cues of the processed signals, and thereby modify the conveyed acoustical scene. An extreme example of this is binaural beamforming which may transform the binaural signal into a diotic one, thus collapsing all sound sources towards the front of the listener [28, 29]. It is important that a binaural SIP algorithm is able to correctly predict the impact of such processing.

In this paper we carry out a thorough investigation of the DBSTOI measure for distributed interferers. In Section 2 we introduce five datasets of measured intelligibility, three of which include different combinations of spatially distributed interferers. We introduce the datasets at this early point, as we refer to them throughout the paper. In Section 3 we 1) provide an outline of the DBSTOI measure, 2) investigate the problems arising with the DBSTOI measure in conditions with spatially distributed interferers as well as the cause of these, and 3) propose a modified measure, the Modified BSTOI (MBSTOI) measure, which solves the observed problems. We also propose a computationally simpler better-ear version, the better-ear MBSTOI (beMBSTOI), which does not model binaural unmasking. We do so, mainly to investigate the relative contributions of better-ear advantage and binaural unmasking to predicted intelligibility. In Section 4 we compare the DBSTOI, MBSTOI, and beMBSTOI measures, as well as the B-sEPSM, using the five previously mentioned datasets. Section 5 concludes upon our findings.

2. Experimental Data

In this section we describe five datasets of measured intelligibility, used throughout the paper for evaluating SIP algorithms. We use existing datasets from [42], [43], and [46], as well as a new dataset, that has been collected.
Datasets are listed in Table 1, but serve well to demonstrate binaural advantage. The experiment included 10 frontal interferer, masked by a single point-like source of SSN located at different azimuths in the horizontal plane. Such conditions are acoustically simple but serve well to demonstrate binaural advantage. The experiment included 10 point noise sources.

Table 1: Summary of conditions. Abbreviations are as follows: SSN (speech shaped noise), BFHN (bottling factory hall noise), ISTS (international speech test signal), ISO (cylindrically isotropic SSN), IBM (ideal binary mask processing), un. (unprocessed), bil. (bilateral), bin. (binaural), MVDR (minimum variance distortionless response beamforming), MWF (multichannel Wiener filtering), Kukl. (algorithms proposed in [46]), Braun (algorithms proposed in [47]).

<table>
<thead>
<tr>
<th>Cond.</th>
<th>Interferer</th>
<th>Interferer location(s)</th>
<th>Proc.</th>
<th>Cond.</th>
<th>Interferer</th>
<th>Interferer location(s)</th>
<th>Proc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1-1</td>
<td>SSN</td>
<td>−160°</td>
<td>–</td>
<td>D3-4</td>
<td>SSN</td>
<td>{30°,180°}</td>
<td>–</td>
</tr>
<tr>
<td>D1-2</td>
<td>SSN</td>
<td>−115°</td>
<td>–</td>
<td>D3-5</td>
<td>ISTS</td>
<td>{30°,180°}</td>
<td>–</td>
</tr>
<tr>
<td>D1-3</td>
<td>SSN</td>
<td>−80°</td>
<td>–</td>
<td>D3-6</td>
<td>SSN</td>
<td>ISO</td>
<td>Beamforming</td>
</tr>
<tr>
<td>D1-4</td>
<td>SSN</td>
<td>−40°</td>
<td>–</td>
<td>D3-7</td>
<td>SSN</td>
<td>{±115°,180°}</td>
<td>Beamforming</td>
</tr>
<tr>
<td>D1-5</td>
<td>SSN</td>
<td>20°</td>
<td>–</td>
<td>D3-8</td>
<td>ISTS</td>
<td>{±115°,180°}</td>
<td>Beamforming</td>
</tr>
<tr>
<td>D1-6</td>
<td>SSN</td>
<td>0°</td>
<td>–</td>
<td>D3-9</td>
<td>SSN</td>
<td>{30°,180°}</td>
<td>Beamforming</td>
</tr>
<tr>
<td>D1-7</td>
<td>SSN</td>
<td>40°</td>
<td>–</td>
<td>D3-10</td>
<td>ISTS</td>
<td>{30°,180°}</td>
<td>Beamforming</td>
</tr>
<tr>
<td>D1-8</td>
<td>SSN</td>
<td>80°</td>
<td>–</td>
<td>D4-1</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Un. front mics</td>
</tr>
<tr>
<td>D1-9</td>
<td>SSN</td>
<td>140°</td>
<td>–</td>
<td>D4-2</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Un. left front mic</td>
</tr>
<tr>
<td>D1-10</td>
<td>SSN</td>
<td>180°</td>
<td>–</td>
<td>D4-3</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Bil. MVDR Kukl.</td>
</tr>
<tr>
<td>D2-1</td>
<td>SSN</td>
<td>−115°</td>
<td>IBM</td>
<td>D4-4</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Bil. MWF Kukl.</td>
</tr>
<tr>
<td>D2-2</td>
<td>SSN</td>
<td>0°</td>
<td>IBM</td>
<td>D4-5</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Bin. MVDR Braun</td>
</tr>
<tr>
<td>D2-3</td>
<td>SSN</td>
<td>20°</td>
<td>IBM</td>
<td>D4-6</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Bin. MWF Braun</td>
</tr>
<tr>
<td>D2-4</td>
<td>BFHN</td>
<td>−115°</td>
<td>–</td>
<td>D4-7</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Bin. MVDR Kukl.</td>
</tr>
<tr>
<td>D2-5</td>
<td>BFHN</td>
<td>0°</td>
<td>–</td>
<td>D4-8</td>
<td>ISTS</td>
<td>{±90°,±135°,180°}</td>
<td>Bin. MWF Kukl.</td>
</tr>
<tr>
<td>D2-6</td>
<td>BFHN</td>
<td>20°</td>
<td>–</td>
<td>D5-1</td>
<td>SSN</td>
<td>0°</td>
<td>–</td>
</tr>
<tr>
<td>D2-7</td>
<td>BFHN</td>
<td>−115°</td>
<td>IBM</td>
<td>D5-2</td>
<td>SSN</td>
<td>½(0°)+½(ISO)</td>
<td>–</td>
</tr>
<tr>
<td>D2-8</td>
<td>BFHN</td>
<td>0°</td>
<td>IBM</td>
<td>D5-3</td>
<td>SSN</td>
<td>½(0°)+½(ISO)</td>
<td>–</td>
</tr>
<tr>
<td>D2-9</td>
<td>BFHN</td>
<td>20°</td>
<td>IBM</td>
<td>D5-4</td>
<td>SSN</td>
<td>ISO</td>
<td>–</td>
</tr>
<tr>
<td>D3-1</td>
<td>SSN</td>
<td>ISO</td>
<td>–</td>
<td>D5-5</td>
<td>SSN</td>
<td>½(ISO)+½(−115°)</td>
<td>–</td>
</tr>
<tr>
<td>D3-2</td>
<td>SSN</td>
<td>{±115°,180°}</td>
<td>–</td>
<td>D5-6</td>
<td>SSN</td>
<td>½(ISO)+½(−115°)</td>
<td>–</td>
</tr>
<tr>
<td>D3-3</td>
<td>ISTS</td>
<td>{±115°,180°}</td>
<td>–</td>
<td>D5-7</td>
<td>SSN</td>
<td>−115°</td>
<td>–</td>
</tr>
</tbody>
</table>

for evaluating the DBSTOI measure in isotropic noise. In combination, these datasets span a wide range of realistic conditions. We denote the datasets as D1, D2, D3, D4 and D5, respectively. The acoustical conditions of the five datasets are listed in Table 1.

2.1. D1: Point Noise Sources

This dataset was initially described in [12], and features conditions with a frontal interferer, masked by a single point-like source of SSN located at different azimuths in the horizontal plane. Such conditions are acoustically simple but serve well to demonstrate binaural advantage. The experiment included 10...
Danish test subjects who reported normal hearing. They were presented with
noisy speech stimuli via headphones in quiet. An anechoic listening environment
was simulated using the CIPIC Head Related Transfer Functions (HRTFs) \[48\].
Sentences from the Dantale II corpus \[49\] were presented from the front, masked
by an SSN interferer placed at some azimuth in the horizontal plane. Subjects
were presented with one sentence at a time and asked to repeat it as accu-
rately as possible. The number of correctly repeated words was noted by the
experimenter. Speech intelligibility was measured for ten different interferer lo-
cations (see Table 1), each at six SNRs, uniformly spaced by 3 dB and centered
around a rough estimate of the Speech Reception Threshold (SRT) \[1\] (manually
obtained by the experimenter, by listening to samples of the stimuli). Three
sentences were scored for each subject/location/SNR, resulting in the scoring
of (10 subjects) × (10 interferer locations) × (6 SNRs) × (3 repetitions) = (1800
sentences).

2.2. D$_{2}$: Point Sources and Binary Mask Processing

This dataset was initially described in \[43\], and contains conditions similar
to dataset D$_{1}$, but adds Ideal Binary Mask (IBM) processing \[50\] and realistic
environmental noise. This dataset is included in order to investigate whether SIP
algorithms can predict the combined impact of binaural advantage and non-
linear processing. Measurements were carried out for 14 subjects who reported
normal hearing. Three representative interferer azimuths were chosen from the
conditions of dataset D$_{1}$ (0°, 20°, and −115°). Measurements were made for
three different variations of each of these three conditions:

- **Conditions 1–3:** use an SSN interferer as in dataset D$_{1}$. However,
the stimuli were IBM processed before being presented to the subjects:
the speech and noise signals were decomposed with a short-time Discrete

---

\[1\] We define the SRT as the SNR at which intelligibility is 50%. The SNR is computed as the ratio of average target power to average interferer power, both measured at the respective sources.
Fourier Transformation (DFT) and noisy DFT-units with an SNR of less than 0 dB were attenuated by 10 dB \[44\]. This processing was carried out independently on the left and right ear signals, e.g. to simulate a hearing aid with a powerful noise reduction system.

- **Conditions 4–6**: use an environmental noise recording in place of the SSN interferer: Bottling Factory Hall Noise (BFHN) \[51\]. BFHN is a recording of heavy machinery and bottles rattling on a conveyor belt. No IBM processing was applied.

- **Conditions 7–9**: are identical to conditions 4–6, but use IBM processing as in conditions 1–3.

Sentence material and scoring was carried out as for D\(_1\), except for the fact that subjects did not verbally repeat the sentences they heard, but rather entered them on a computer screen by selecting each word from a list of 10 possibilities\(^2\). This resulted in the scoring of \((14 \text{ subjects}) \times (9 \text{ conditions}) \times (6 \text{ SNRs}) \times (3 \text{ repetitions}) = (2268 \text{ sentences})\).

2.3. \(D_3\): Multiple Interferers and Beamforming

This dataset was initially described in \[42\], and includes different combinations of multiple interferers, isotropic noise and beamforming. It is therefore well suited for evaluating, how SIP algorithms cope with changing acoustical conditions and processing. Intelligibility was measured for 10 subjects, who reported normal hearing, using the same sentence material and collection procedure as dataset D\(_1\). In all the 10 conditions, speech was presented from the front. Otherwise, the conditions are as follows:

- **Condition 1**: speech is presented together with cylindrically isotropic SSN.

\(^2\)The collection procedure is described in more detail in \[41\]. A comparison of the two approaches for measuring intelligibility is conducted in \[52\], showing them to yield highly similar results.
• **Condition 2:** SSN interferers are placed in 115°, 180° and −115° azimuth.

• **Condition 3:** is acoustically identical to condition 2, but uses the International Speech Test Signal (ISTS) instead of SSN. ISTS is a speech-like, but unintelligible, signal, obtained by concatenating short segments of speech in different languages [53].

• **Condition 4:** SSN interferers are placed in, 30° and −180° azimuth.

• **Condition 5:** is acoustically identical to condition 4, but uses the ISTS instead of SSN.

• **Conditions 6–10:** are identical to conditions 1–5, but present subjects with the output of a time-invariant, two-microphone Minimum Variance Distortionless Response (MVDR) beamformer. This was accomplished by using HRTFs measured for the two microphones of a behind-the-ear hearing aid.

This resulted in the scoring of (10 subjects) × (10 conditions) × (6 SNRs) × (3 repetitions) = (1800 sentences).

2.4. D4: Reverberation and Beamforming

This dataset was initially described in [46], and consists of conditions with multiple spatially separated interferers and different types of beamforming. Some types of beamforming alter the binaural cues of the input signals, and this dataset allows for evaluating how SIP algorithms cope with such changes. Measurements were made for 20 subjects, using the same speech material and procedure as in dataset D2. Speech was presented from the front with reverberation simulated by use of a Binaural Room Impulse Response (BRIR) measured in a large cellar. Five ISTS interferers were placed at azimuths of ±90°, ±135° and 180°, respectively, using BRIRs measured in those locations in the same cellar. The interferers were fixed at the same power as the target speaker. Intelligibility was modified by synthetically altering the Direct to Reverberant
Ratio (DRR) of the target speaker. Intelligibility was measured for four different DRRs and for eight different combinations of hearing aid beamforming and noise reduction. Details on the applied processing are given in [46]. In total, intelligibility was scored for \( (20 \text{ subjects}) \times (8 \text{ conditions}) \times (4 \text{ DRRs}) \times (5 \text{ repetitions}) = (3200 \text{ sentences}) \).

2.5. \( D_5 \): Isotropic and Point Source Interferers

To gain even more knowledge about the performance of the DBSTOI measure in environments with distributed interferers, we collected an additional dataset with different combinations of cylindrically isotropic and point source SSN interferers. Specifically, we measured intelligibility for frontal speech, using the same speech material and procedure as that of dataset \( D_2 \). The speech was masked by interferer signals of the following form:

\[
\mathbf{n}(t; \alpha) = \begin{cases} 
\sqrt{1-\alpha} \, \mathbf{n}_0(t) + \sqrt{\alpha} \, \mathbf{n}_{\text{iso}}(t), & \text{if } 0 < \alpha < 1, \\
\sqrt{2-\alpha} \, \mathbf{n}_{\text{iso}}(t) + \sqrt{\alpha - 1} \, \mathbf{n}_{-115}(t), & \text{if } 1 < \alpha < 2,
\end{cases}
\]

(1)

where \( \mathbf{n}_0(t) : \mathbb{R} \to \mathbb{R}^{2 \times 1} \) is the binaural signal arising from a point source of SSN directly in front of the listener, while \( \mathbf{n}_{\text{iso}}(t) \) is the binaural signal arising from cylindrically isotropic SSN, and \( \mathbf{n}_{-115}(t) \) is the binaural signal arising from a point source of SSN at \(-115^\circ\) from the front. The noise source \( \mathbf{n}(t; \alpha) \) allows for sampling a continuum of mixtures of point source noise and isotropic noise.

Intelligibility was measured for seven linearly spaced values of \( \alpha \): 0, \( 1/3 \), \( 2/3 \), 1, \( 4/3 \), \( 5/3 \) and 2, each for eight SNRs linearly spaced by 3 dB, centered on a rough estimate of the SRT. Measurements were carried out with eight subjects, who reported normal hearing, using the same procedure as that described for dataset \( D_2 \). This resulted in the scoring of \( (8 \text{ subjects}) \times (7 \text{ conditions}) \times (8 \text{ SNRs}) \times (3 \text{ repetitions}) = (1344 \text{ sentences}) \).

3. Methods

The purpose of this section is threefold: 1) to briefly outline the DBSTOI measure, which was initially proposed in [43, 44], 2) to highlight important dis-
Select $\gamma$ and $\tau$ for each 1/3 octave band, $q$, and each time unit, $m$, such as to maximize output.

Figure 1: A block diagram of the DBSTOI measure. Figure adapted from [44].
parities between measured intelligibility and predictions of the DBSTOI measure which arise in conditions with spatially distributed noise sources, and 3) to propose a modification of the DBSTOI measure, which greatly diminishes these disparities.

3.1. The DBSTOI measure

The DBSTOI measure, introduced in [44], extends the STOI measure to enable prediction of binaural advantage. At the same time, it was shown that the DBSTOI measure mostly retains the favorable monaural prediction performance of the STOI measure [44]. An overview of the DBSTOI measure is given in Fig. 1. To facilitate the following discussion, we give a compressed overview of the DBSTOI measure. A more thorough treatment can be found in [44].

As input signals, the DBSTOI measure assumes access to degraded signals, $y_l(t)$ and $y_r(t)$, recorded at the left and right ears of the listener, as well as the corresponding clean reference signals $x_l(t)$ and $x_r(t)$. These are analyzed with a short-time DFT, yielding coefficients, $\hat{y}_{l,k,m}$, $\hat{y}_{r,k,m}$, $\hat{x}_{l,k,m}$ and $\hat{x}_{r,k,m}$, where $k$ is the frequency bin index and $m$ is the frame index. The DFT is computed with the same parameters as for the STOI measure [3].

The DFT coefficients from the left ear and the right ear are combined using an EC stage [32, 33, 54], which models the binaural advantage obtained by having two ears, in situations with spatial separation between target and interferer [44]:

\[
\hat{x}_{k,m} = \lambda_{k,m} \hat{x}_{l,k,m} - \lambda_{k,m}^{-1} \hat{x}_{r,k,m},
\]

\[
\hat{y}_{k,m} = \lambda_{k,m} \hat{y}_{l,k,m} - \lambda_{k,m}^{-1} \hat{y}_{r,k,m},
\]

where:

\[
\lambda_{k,m} = 10^{(\gamma_{k,m} + \Delta \gamma_{k,m})/40} e^{j\omega (\tau_{k,m} + \Delta \tau_{k,m})/2}.
\]

Here, $\gamma_{k,m}$ is a relative amplitude change between the left and right ear signals, and $\tau_{k,m}$ is a relative time shift. These are used to compute the complex scalar $\lambda_{k,m}$, which represents both time and amplitude shift. Both are introduced before subtracting the right ear coefficient from the left ear coefficient.
to obtain the EC stage output. The parameters $\gamma_{k,m}$ and $\tau_{k,m}$ are determined in order to maximize predicted intelligibility of the output. Informally, the purpose of $\gamma_{k,m}$ and $\tau_{k,m}$ can be interpreted as that of aligning any interferer or distortion components in the left and right ear signals, such that these are canceled when the signals are subtracted from one another [32]. In signal processing terms, the EC stage can be viewed as a beamformer. Independent noise sources, $\Delta \gamma_{k,m}$ and $\Delta \tau_{k,m}$, referred to as "jitter" [32], are added to $\gamma_{k,m}$ and $\tau_{k,m}$, to limit performance to be consistent with human performance. The jitters are zero mean Gaussian with variances [54]:

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

(5)

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

(6)

The combined DFT coefficients are used to compute power envelopes, $X_{q,m}$ and $Y_{q,m}$, in $Q = 15$ one-third octave bands [44]:

$$X_{q,m} = \sum_{k=k_{1}(q)}^{k_{2}(q)} \left| \hat{x}_{k,m} \right|^2 = \sum_{k=k_{1}(q)}^{k_{2}(q)} \left| \lambda \hat{x}_{k,m}^{(l)} - \lambda^{-1} \hat{x}_{k,m}^{(r)} \right|^2$$

$$= 10^\frac{\gamma_{k,m}}{20} \sum_{k=k_{1}(q)}^{k_{2}(q)} \left| \hat{x}_{k,m}^{(l)} \right|^2 + 10^{-\frac{\gamma_{k,m}}{20}} \sum_{k=k_{1}(q)}^{k_{2}(q)} \left| \hat{x}_{k,m}^{(r)} \right|^2$$

$$- 2 \text{Re} \left[ \sum_{k=k_{1}(q)}^{k_{2}(q)} \hat{x}_{k,m}^{(l)}* \hat{x}_{k,m}^{(r)} e^{-j\omega_k (\tau + \Delta \tau)} \right]$$

$$\approx 10^\frac{\gamma_{k,m}}{20} X_{q,m}^{(l)} + 10^{-\frac{\gamma_{k,m}}{20}} X_{q,m}^{(r)}$$

$$- 2 \text{Re} \left[ e^{-j\omega_k (\tau + \Delta \tau)} X_{q,m}^{(c)} \right],$$

(7)
where:

\[
X_{q,m}^{(l)} = \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(l)}|^2,
\]

\[
X_{q,m}^{(r)} = \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(r)}|^2,
\]

\[
X_{q,m}^{(c)} = \sum_{k=k_1(q)}^{k_2(q)} \hat{x}_{k,m}^{(l)} \ast \hat{x}_{k,m}^{(r)},
\]

(8)

and where \(\omega_k\) is the angular frequency of the \(k\)’th frequency bin, \(\omega_q\) is the center angular frequency of the \(q\)’th third octave band, and \(k_1(q)\) and \(k_2(q)\) are, respectively, the lower and upper limits of the \(q\)’th one-third octave band.

Similar power envelope samples, \(Y_{q,m}\), are defined for the degraded signal. These envelopes are arranged in zero-mean vectors of \(N = 30\) samples:

\[
x_{q,m} = [X_{q,m-N+1}, \ldots, X_{q,m}]^T - 1 \sum_{m'=m-N+1}^{m} \frac{X_{q,m'}}{N},
\]

(9)

where \(1 \in \mathbb{R}^{N \times 1}\) is a vector of ones. Similar vectors, \(y_{q,m}\), are defined for the degraded signal, as well as for the other power envelope signals defined by (8): \(x_{q,m}^{(l)}, x_{q,m}^{(r)}, x_{q,m}^{(c)}, y_{q,m}^{(l)}, y_{q,m}^{(r)}\) and \(y_{q,m}^{(c)}\). Using (7), we then have [44]:

\[
x_{q,m} \approx 10^{-\frac{\Delta \gamma}{10}} x_{q,m}^{(l)} + 10^{-\frac{\Delta \gamma}{10}} x_{q,m}^{(r)} - 2 \text{Re} \left[ e^{-j\omega_q(\tau+\Delta \tau)} x_{q,m}^{(c)} \right],
\]

(10)

The STOI measure uses the average correlation between clean and degraded envelopes to predict intelligibility [3]. Assuming that \(X_{q,m'}\) and \(Y_{q,m'}\), for \(m' = m - N + 1, \ldots, m\), are samples of a wide sense stationary process, this can be interpreted as an estimate of the linear correlation coefficient:

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

(11)

where the expectation is computed with respect to both the input signals, and the jitter in the EC stage. By estimating the correlation in the same manner as
in the STOI measure, we arrive at [44]:

\[ \hat{\rho}_{q,m} = \frac{E_\Delta [x_{q,m}^T y_{q,m}]}{\sqrt{E_\Delta [x_{q,m}^T x_{q,m}] E_\Delta [y_{q,m}^T y_{q,m}]}} \]  

(12)

where \( E_\Delta [\cdot] \) is the expectation with respect to the jitter. As shown in [44], the numerator of this expression may be approximated by:

\[ E_\Delta [x_{q,m}^T y_{q,m}] \approx \left(e^{2\beta x_{q,m}^T y_{q,m}} + e^{-2\beta x_{q,m}^T y_{q,m}}e^{2\sigma^2_\Delta}\right) \]

\[ + \left(x_{q,m}^T y_{q,m} + x_{q,m}^T y_{q,m} - 2e^{\sigma^2_\Delta/2}e^{-\omega^2\sigma^2_\Delta/2}x_{q,m}^T y_{q,m}\right) \]

\[ \left\{e^{2\beta x_{q,m}^T y_{q,m}} + e^{-\beta y_{q,m}}\right\} Re \left[y_{q,m}^T e^{-j\omega \tau}\right] \]

\[ + \left(Re \left(x_{q,m}^T y_{q,m}\right) + e^{-\beta x_{q,m}^T y_{q,m}}\right) \]

\[ + 2 \left(Re \left(x_{q,m}^T y_{q,m}\right) + e^{-2\omega^2\sigma^2_\Delta}Re \left[y_{q,m}^T e^{-2\omega \tau}\right]\right) \]

(13)

where:

\[ \beta = \frac{\ln(10)}{20} \gamma, \]

\[ \sigma^2_\Delta = \left(\frac{\ln(10)}{20}\right)^2 \sigma^2_\gamma. \]  

(14)

The values of \( E_\Delta [x_{q,m}^T x_{q,m}] \) and \( E_\Delta [y_{q,m}^T y_{q,m}] \) can be obtained from similar expressions (by exchanging all occurrences of \( y \) with \( x \), or vice versa, in (13)). The final DBSTOI measure is obtained by averaging the intermediate correlation coefficients across time and one-third octave bands:

\[ \text{DBSTOI} = \frac{1}{QM} \sum_{m=1}^{M} \sum_{q=1}^{Q} \hat{\rho}_{q,m}. \]  

(15)

This value is influenced by the choice of \( \gamma_{k,m} \) and \( \tau_{k,m} \). These values are fixed within each one third octave band, so that \( \gamma_{k,m} = \gamma_{q,m} \) for \( k_1(q) \leq k \leq k_2(q) \). At the same time, the values are fixed across one vector of envelope samples, \( \gamma_{q,m} \). The values are chosen such as to maximize the intermediate correlation, i.e. the predicted intelligibility:

\[ \hat{\rho}_{q,m} = \max_{\gamma_{q,m},\tau_{q,m}} \hat{\rho}_{q,m}(\gamma_{q,m},\tau_{q,m}). \]  

(16)
In practice, this optimization is carried out as a simple grid search across all combinations of $-1 \text{ ms} < \tau_{q,m} < +1 \text{ ms}$ in 100 uniform steps and $-20 \text{ dB} < \gamma_{q,m} < +20 \text{ dB}$ in 40 steps. In addition, we also search the two "better-ear"-options given by the $(\gamma_{q,m}, \tau_{q,m})$-pairs $(+\infty, 0)$ and $(-\infty, 0)$. This corresponds to listening with only one ear and ignoring the other.

The DBSTOI measure outputs a rating of intelligibility on a somewhat arbitrary scale from zero to one. To obtain a prediction of measured intelligibility as a percentage of correctly repeated words, we transform this output with a logistic function [3, 44]:

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

(17)

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, to provide the best possible predictions [44].

3.2. The DBSTOI measure in spatially diverse conditions

In [44] the DBSTOI measure is evaluated in a range of conditions including different noise types, different configurations of interferers, as well as IBM processing and beamforming. Generally, the measure was found to correlate
well with intelligibility in the investigated conditions. However, some indication was also seen of the measure providing inaccurate predictions, when making comparisons between conditions with a point source interferer and conditions with multiple or distributed interferers (e.g. isotropic noise). In this section, we investigate this effect further.

Fig. 2 shows measurements from datasets D$_3$ and D$_5$, which contain both distributed and point-like interferers, plotted against the DBSTOI measure of each condition. The left plot shows the discrepancy which was noted in [10]. It appears that the conditions cluster along two distinct lines. In particular, it appears that conditions D$_3 - 5$, D$_3 - 9$, and D$_3 - 10$ cluster along a path to the left of the remaining data points. This left cluster, which has relatively high measured intelligibility compared with predictions, thereby consists mainly of the conditions with two ISTS interferers. The two conditions with two SSN interferers fall somewhere between the two clusters. The right cluster, containing the remaining conditions, terminates at a DBSTOI measure just above 0.2 and a measured intelligibility of about 0%. This cluster consists of the conditions with
Figure 4: Examples of the distribution of intermediate correlation coefficients, $\rho$, across the values sampled by the grid search, for a random time frame and for band $q = 4$ (with center frequency 300 Hz). Left: all four input signals are independent white noise signals. Right: clean and degraded inputs are independent white noise signals, but the left and right ear signals are identical for both pairs.

more spatially distributed interferers (i.e those with either three point sources or isotropic noise). The right plot of Fig. 2 shows predictions for dataset D$_5$. Here, we see the same tendency towards clustering even more clearly. A left cluster contains the two conditions which exclusively consist of point interferers (D$_5 - 1$ and D$_5 - 7$), while a right cluster contains the conditions where isotropic noise is included. For both datasets, the clusters join together at high levels of intelligibility, and the discrepancy is most pronounced at low intelligibility levels.

To shed further light on this issue, we investigate the behaviour of the DBSTOI measure at very low SNRs. For this purpose, we consider the predictions made for frontal speech masked by the same type of interferer signals used in dataset D$_5$, as given by (1). Fig. 3 shows the DBSTOI measure for different SNRs, for a frontal target, masked by noise signals corresponding to different $\alpha$-values. On the left, the DBSTOI measure is plotted against SNR, for three different $\alpha$-values. It is immediately apparent that the DBSTOI measure
does not converge exactly to 0, at low SNRs, for any of the conditions. The right plot shows the DBSTOI measure versus $\alpha$ at an SNR of $-100$ dB, i.e. a situation where the output of the DBSTOI measure should ideally be zero. This shows that the DBSTOI measure almost reaches 0 for the conditions with a point source interferer ($\alpha = 0$ and $\alpha = 2$), but settles at a higher value, whenever isotropic noise is present. This is a significant insight into the discrepancies seen in Fig. 2 as these can be characterized by the DBSTOI measure predicting too high intelligibility at low SNRs, whenever isotropic noise or multiple distributed interferers are present.

To understand this issue further, we look at the internals of the DBSTOI measure. When computing each intermediate correlation coefficient, as described by (16), the DBSTOI measure does a grid search of $(100$ values of $\tau) \times (40$ values of $\gamma) + (2$ better ear options) = $4002$ parameter combinations, with the aim of finding the combination that leads to the highest predicted intelligibility. However, this optimization process is not only influenced by intelligibility, but also by random fluctuations in the signals. To illustrate this point, the left plot of Fig. 4 shows a histogram of the intermediate correlation, $\bar{\rho}_{q,m}$, for all $4002$ possible $(\tau, \gamma)$-combinations for one frame, when the four input signals are uncorrelated white Gaussian noise. Since the clean and degraded signals are entirely uncorrelated in this scenario, the intermediate correlation should be 0 on average. However, the figure shows that a considerable variation occurs for the different EC parameters. When always picking the highest correlation from such distributions (i.e. what is done in (16)), a considerable upwards bias is introduced. The right plot of Fig. 4 shows the distribution arising from a similar procedure, but when the left and right signals are identical white Gaussian noise signals, while the clean and degraded signals are still uncorrelated. Here, it is seen that the intermediate correlation is the same for all sampled EC-parameters. In [44] it was shown that the DBSTOI measure reduces to the monaural STOI measure under such conditions. Thereby, no variation can be introduced in the EC stage. This explains why the upwards bias is not present, when highly correlated signals are presented to the left and the right ear, e.g.
when a single interferer is located in front of the listener together with the target.

The mechanism described above presents a fundamental issue with the DBSTOI measure: a signal dependent upward bias is introduced at low SNRs. This bias is larger when the left and right ear signals are uncorrelated (e.g. for dichotic stimuli) and smaller when they are correlated (e.g. for diotic stimuli). This can be more or less disruptive, depending on the application of the measure. When using it to compare predicted intelligibility between acoustically similar conditions, the issue is not severe, as the bias will be present to the same extent in all predictions. However, if using the measure to compare widely different acoustical situations, the bias may be of different magnitude in the different conditions and thus corrupt the comparability of predictions. The issue may also be disruptive when comparing different types of processing (e.g. speech enhancement or beamforming) as these may modify the binaural cues of the signal differently, thus effectively changing the acoustical scene conveyed by the binaural signal. Since such comparisons of different acoustical environments and different processing types are central use-cases of the DBSTOI measure, this issue is highly undesirable. We therefore propose a slight modification to the DBSTOI measure, which greatly diminishes the bias.

### 3.3. The MBSTOI measure

To approach the problematic bias, we consider a subset of the domain in which the DBSTOI measure is meant to be applied. We derive analytically a bias compensation strategy for this domain. In Section 4 we demonstrate that the strategy works well outside of this domain. Let us, therefore, consider the situations where the degraded signal can be written as a sum of the clean signal and a noise/distortion component:

\[
y_l(t) = x_l(t) + n_l(t),
\]

\[
y_r(t) = x_r(t) + n_r(t).
\]
Here, \( n_l(t) \) and \( n_r(t) \) may be additive noise or distortion, but are assumed to be uncorrelated with the clean reference signals \( x_l(t) \) and \( x_r(t) \). In this situation, the degraded speech power envelope (defined by (7)) can be written as:

\[
Y_{q,m} = \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m} + \hat{n}_{k,m}|^2
= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}|^2 + \sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2 + 2\Re(\hat{x}_{k,m}^* \hat{n}_{k,m}), \tag{19}
\]

where the first term is recognized as the clean signal power envelope, \( X_{q,m} \), and two following terms are collectively referred to as \( R_{q,m} \). We now define an envelope SNR:

\[
\text{SNR}_{q,m} \triangleq \frac{\sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}|^2}{\sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2}. \tag{20}
\]

Considering a situation with very low envelope SNR, i.e. \( \text{SNR}_{q,m} \to 0 \), we have:

\[
Y_{q,m} \bigg|_{\text{SNR}_{q,m} \ll 1} \approx R_{q,m} \bigg|_{\text{SNR}_{q,m} \ll 1} \approx \sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2. \tag{22}
\]

The above approximations become exact for SNRs tending toward negative infinity.

Now, define:

\[
r_{q,m} = [R_{q,m} - N + 1, \ldots, R_{q,m}]^\top - 1 \sum_{m' = m - N + 1}^{m} \frac{R_{q,m'}}{N}. \tag{23}
\]

\(^3\)To see the result from (22), decompose the clean signals, \( x_l(t) \) and \( x_r(t) \), as \( x_l(t) = c\phi_l(t) \) and \( x_r(t) = c\phi_r(t) \), where \( c > 0 \) and \( \phi_l(t) \) and \( \phi_r(t) \) are arbitrary real signals. Inserting this in (10) yields:

\[
Y_{q,m} = c^2 \sum_{k=k_1(q)}^{k_2(q)} |\hat{\phi}_{k,m}|^2 + \sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2 + c \sum_{k=k_1(q)}^{k_2(q)} 2\Re(\hat{\phi}_{k,m}^* \hat{n}_{k,m}). \tag{21}
\]

It is evident that the value of \( c \) directly controls \( \text{SNR}_{q,m} \). The first term diminishes for small \( c \), i.e. for low \( \text{SNR}_{q,m} \), leaving only \( R_{q,m} \). As \( c \to 0 \), only the second term remains.
From this point, we drop the frame and frequency band indexes, \( q \) and \( m \), from \( x_{q,m}, y_{q,m} \) and \( r_{q,m} \) for notational convenience. Then (19) implies \( y = x + r \). Inserting this into (12), we obtain:

\[
\bar{\rho}_{q,m} = \frac{E_{\Delta} [x^\top x] + E_{\Delta} [x^\top r]}{\sqrt{E_{\Delta} [x^\top x] E_{\Delta} [y^\top y]}} = \frac{E_{\Delta} [x^\top r]}{\sqrt{E_{\Delta} [y^\top y] E_{\Delta} [r^\top r]}}.
\]

(24)

Here, \( E_{\Delta} [\cdot] \) is the expectation across the jitter from the EC stage. We remind the reader that (19), (22) and (24) are functions of the EC parameters, \( \tau \) and \( \gamma \). Clearly, (22) holds independently of these values, except for the special case where the EC stage would provide an infinite improvement in SNR (i.e. in the hypothetical situation where the EC stage is able to completely cancel the noise). The jitter effectively prevents this from happening in practice [32].

For low SNRs, we thus have \( E_{\Delta} [y^\top y] \approx E_{\Delta} [r^\top r] \) and \( E_{\Delta} [y^\top y] \gg E_{\Delta} [x^\top x] \). Inserting this into (24), we get:

\[
\bar{\rho}_{q,m} \bigg|_{\text{SNR}_{q,m} \ll 1} \approx 0 + \frac{E_{\Delta} [x^\top r]}{\sqrt{E_{\Delta} [x^\top x] E_{\Delta} [r^\top r]}}.
\]

(25)

The first term approaches zero, and the second term is a short-time estimate of the correlation between the signals \( X_{q,m'} \) and \( R_{q,m'} \) for \( m' = m - N + 1, \ldots, m \). This second term is not in general zero, and is furthermore dependent on the EC parameters. We can conclude that the bias at low SNRs must stem from the fact that the EC stage parameters, \( \gamma \) and \( \tau \), are found such as to maximize the second term in (25). We therefore propose a modified objective function for finding EC stage parameters, in which only the first term of (24) is included. In other words, rather than finding the EC parameters, \( (\gamma_{q,m}, \tau_{q,m}) \), which maximize (24), we find instead the EC parameters, \( (\gamma_{q,m}, \tau_{q,m}) \), which maximize:

\[
g = \sqrt{\frac{E_{\Delta} [x^\top x]}{E_{\Delta} [y^\top y]}}.
\]

(26)

Note that by doing so, we ensure that in situations where the true intelligibility is certainly going to be 0% (e.g. additive uncorrelated noise at sufficiently
low SNR), the predictor output is also going to be zero. In other words, by
collection the function $g$ does not include the described upwards bias. Fur-
thermore, it should be noted that in high SNR situations, i.e. where $E_{\Delta} [x^T x] \approx
E_{\Delta} [y^T y]$, we have $g = 1$, corresponding to maximum predicted intelligibility.
We only use (26) to find the values of $\gamma$ and $\tau$. The found parameters are then
used to evaluate the intermediate correlation, (12), as in the DBSTOI measure.
We refer to the modified measure as the MBSTOI measure.

In these derivations, we assumed a very simple signal model, (18), which is
not strictly valid for all situations, in which the MBSTOI measure may find use,
i.e. when the degraded signal is a processed version of the noisy speech signal,
and this processing cannot be modeled as an uncorrelated additive component.
However, an argument, similar to the one carried out above, leads to the same
method, if the noisy speech has been processed, and this processing is 1) identi-
cal on both ears, 2) linear, 3) slowly time varying (with respect to the duration
of $N$ DFT windows), and 4) approximately constant within each one-third oc-
tave band. We hypothesize, furthermore, that the MBSTOI measure works
well in an even broader context. This is experimentally verified in Section 4.

3.4. The beMBSTOI measure

We also define a better ear version of the MBSTOI measure. This is eas-
ily obtained by performing the optimization across a search grid of only two
$(\gamma_{q,m}, \tau_{q,m})$-pairs: $(+\infty, 0)$ and $(-\infty, 0)$. Mathematically, this corresponds com-
pletely to disabling the EC stage, computing the (modified) STOI measure for
each ear separately, and maintaining only the larger result. This measure allows
us to investigate the contribution of binaural unmasking to predicted speech in-
telligibility. We refer to it as the beMBSTOI measure.

---

*In this case, the processing amounts approximately to a constant factor, $c_{q,m}^2$, multiplied onto (19).*
4. Results and Discussion

We investigate the performance of the proposed MBSTOI measure and other measures in terms of 1) bias characteristics, 2) general prediction performance, and 3) prediction of binaural advantage. We furthermore, discuss the relationship between the different evaluated SIP algorithms, as well as the extent to which the results generalize to conditions beyond those tested here.

4.1. Bias in the MBSTOI measure

The bias problem, encountered in the DBSTOI measure, is most directly apparent from Fig. 3. Fig. 5 therefore shows the same plot, using the same input signals, but for the MBSTOI measure. From this figure, it appears that almost no upwards bias is present in the MBSTOI measure. The right plot on the figure shows that a slight upwards bias is still present for \( \alpha \)-values just under 1. However, the magnitude of this effect is so small that it has no practical impact on prediction performance.
Fig. 5 indicates that upwards bias, in conditions with spatially distributed interferers, is not a pronounced issue in the MBSTOI measure, but it does not say anything about how the proposed modification impacts prediction performance. This is investigated in the following two sections.

4.2. General Prediction Performance

In this section we investigate prediction performance for each of the five datasets and for four different SIP algorithms: the DBSTOI measure [44], the MBSTOI measure, the beMBSTOI measure, and the B-sEPSM [45].

The B-sEPSM, proposed in [45], is a binaural extension of the mr-sEPSM, which handles binaural and non-linearly processed stimuli. The B-sEPSM bases predictions on a degraded speech signal and a clean noise signal, i.e., noise in the absence of speech and processing. An implementation of the measure has kindly been made available by the authors [55]. The B-sEPSM outputs a number, B-SNR_{env}, which is transformed to a sensitivity index [45]:

\[ d' = k (\text{B-SNR}_{env})^q, \]

and further to a prediction of the probability that a subject answers correctly [45]:

\[ P_{\text{correct}} = \Phi \left( \frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}} \right), \]

where \( \Phi(\cdot) \) is a cumulative normal distribution, \( k, q, \mu_N, \) and \( \sigma_N^2 \) are constants determined by the speech material and test setup, while \( \sigma_S^2 \) is fitted to the data at hand. In this work, we use the B-sEPSM slightly differently to make it more comparable with the other methods. We first transform the output, B-SNR_{env}, by (27) using \( k = \sqrt{1.2} \) and \( q = 0.5 \). These values were suggested in [24] for making predictions together with the Dantale II speech material, which was used in collecting the five datasets considered in this study. We then transform the resulting sensitivity index, \( d' \), by a logistic function, (17), and fit the

---

5The implementation is part of the Python Auditory Modelling Toolbox. See http://pambox.org/
Figure 6: The outputs of the four studied intelligibility predictors (rows) plotted against the measured intelligibility (averaged across all subjects and repetitions) of the five studied datasets (columns). Details about the conditions are given in Table 1. For the B-sEPSM, the plot shows $d'$, divided by a factor of 50, such as to make it fit on the same axis as the other predictors. Predictions for the D₄-data are missing for the B-sEPSM, because we were unable to obtain clean noise signals for this dataset. The black curves are logistic functions, fitted for each predictor (row).
parameters, $a$, and $b$ to the available data, exactly as is done for the other measures discussed in this paper. It should be noted that the logistic function is very similar to the cumulative normal distribution, and that the main difference therefore lies in the fact that we fit two parameters ($a, b$) to the available data rather than one ($\sigma_2^2$).

Fig. 6 shows the predictions of the four measures for each of the five datasets. It was not possible to obtain clean noise signals for dataset D$_4$, and B-sEPSM predictions for this dataset are therefore missing. A single logistic function was fitted for each measure (i.e. $a$ and $b$ are constant across each row in the figure).

The DBSTOI measure (Fig. 6, row 1) makes accurate predictions for the datasets with point source interferers, D$_1$ and D$_2$, while for the remaining datasets, D$_3$, D$_4$, and D$_5$, predictions fall in two clusters depending on whether the interferers are point-like or spatially distributed. The MBSTOI measure (Fig. 6, row 2) makes almost identical predictions for datasets D$_1$ and D$_2$. For the three remaining datasets, D$_3$, D$_4$, and D$_5$, the bias problem appears to be alleviated. Some deviation is, however, still seen for conditions $D_3-5$ and $D_3-10$, which contain two ISTS interferers. This corresponds to unintelligible two-talker babble, which is strongly fluctuating. The STOI measure is known to underestimate intelligibility for such interferers [26], and this is most likely the cause of observed deviations. It therefore appears that the proposed modification to the DBSTOI measure serves its intended purpose, without otherwise impairing prediction performance considerably. The tested conditions include a substantial range of combinations of different acoustical setups, different interferer types, noise reduction and beamforming. It therefore spans many of the types of conditions in which the MBSTOI measure could realistically be applied.

$^6$The B-sEPSM only processes frequency channels which exceed the absolute threshold of hearing [45]. The datasets used in this study were collected at normal conversational level, and thus not near the absolute threshold of hearing. We therefore ensured that the signals provided to the B-sEPSM were always powerful enough for all frequency channels to be processed.
The third row in Fig. 6 shows predictions by the beMBSTOI measure, which does not have the EC stage, but works purely in a better-ear fashion. For datasets D_1 and D_2, it is very evident that the beMBSTOI measure is not accounting for binaural unmasking, i.e. predictions are too low in conditions where an advantage can be gained from binaural unmasking. This is especially evident for dataset D_1, where intelligibility is underestimated in all conditions except the two where the interferer is located directly in front or directly behind. In conditions such as those in datasets D_1 and D_2, a simple better-ear approach is thus not sufficient to predict binaural advantage. A similar effect can be seen for dataset D_5. However, for datasets D_3 and D_4, predictions do not appear to be significantly impaired. These datasets cover conditions which are complicated in terms of both spatially distributed interferers and processing. At the same time they do not contain conditions with a single point-like interferer located away from the frontal position. The absence of such “pure” conditions, with large potential for binaural unmasking, may explain why predictions are reasonably good without the EC stage.

The last row of Fig. 6 shows predictions of the B-sEPSM. These follow the trend of the measurements, but are less accurate than the predictions of the MBSTOI measure. From dataset D_1, it appears that binaural advantage is overestimated. At the same time, from dataset D_2, it appears that the effect of IBM processing may also be overestimated. The B-sEPSM uses an EC stage similar to that of the MBSTOI measure, but the underlying SIP algorithm, the mr-sEPSM [21], is very different from the STOI measure. It is designed to take modulation masking into account, and therefore performs well in conditions with modulated interferers [56, 21]. The datasets considered in this paper lack such conditions.

The overall performance of the four methods is summarized in Table 2 in terms of Root Mean Squared Error (RMSE), Kendall’s Tau [57], and Pearson correlation. These metrics support the previous observations. Specifically, the minimum RMSE is obtained by the MBSTOI measure for all five datasets. We remark that Kendall’s Tau is independent of one-to-one mappings, and that
Table 2: Summary of prediction performance. The best performance in each column is written in bold. Pearson correlation values which are not statistically significantly lower than the highest value of the column are marked with an asterisk.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>Kendall’s Tau</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D$_1$</td>
<td>D$_2$</td>
<td>D$_3$</td>
</tr>
<tr>
<td>DBSTOI</td>
<td>10.80%</td>
<td>8.73%</td>
<td>12.87%</td>
</tr>
<tr>
<td>MBSTOI</td>
<td>7.34%</td>
<td>6.92%</td>
<td>7.07%</td>
</tr>
<tr>
<td>beMBSTOI</td>
<td>16.57%</td>
<td>13.06%</td>
<td>8.49%</td>
</tr>
<tr>
<td>B-seEPSM</td>
<td>15.12%</td>
<td>17.82%</td>
<td>16.12%</td>
</tr>
</tbody>
</table>
the fitted logistic mapping therefore has no effect on this metric. This also means that performance is independent of whether the B-sEPSM framework is used as described in [45], or in the slightly different manner in which we use it here. Alongside the Pearson correlation values, Table 2 shows the results of a statistical significance test. For each dataset, we compared each predictor to the one with the highest Pearson correlation using the statistic by Williams for comparing dependent correlations [58, 59, 60]. The results were compensated for multiple comparisons by use of Bonferroni correction. In Table 2, the Pearson correlation scores which are not significantly lower than the highest correlation of the corresponding dataset are marked with an asterisk. This shows that the Pearson correlation for the MBSTOI predictions are not significantly lower than the highest one in any case. Furthermore, for three out of five datasets, the MBSTOI measure achieves the highest Pearson correlation while all of the other methods achieve statistically significantly lower correlation.

### 4.3. Prediction of Binaural Advantage

In this section we evaluate the performance of the four methods in terms of predicting binaural advantage. We therefore focus specifically on the results of dataset D1, which contains conditions with a frontal target and a single interferer at different azimuths, i.e. conditions where binaural advantage is expected to be high.

The SRT was estimated for each of these conditions, by maximum-likelihood-fitting a logistic function to measured intelligibility as a function of SNR, and computing the intersection with 50% intelligibility [44]. We define the binaural advantage of each condition of dataset D1 to be the difference in SRT between the particular condition and the condition with an interferer azimuth of 0°, i.e. condition D1-6 (see Table 1). The binaural advantage, as computed in this manner, is plotted in Fig. 7 ("Andersen et al."). Similar measurements were carried out by [31] and these are also included ("Bronkhorst & Plomp, FF"). This study also measured binaural advantage with artificially processed stimuli, where no binaural unmasking is possible (see [31] for details). For this type of
Figure 7: Binaural advantage as measured in two independent studies, and predicted by four SIP algorithms. The error bars on the "Andersen et al."-data show quartiles (the binaural advantage was computed separately for the 10 subjects).

stimuli, no additional performance is predicted by using an EC stage relative to better-ear processing as used in the beMBSTOI measure. These results are also shown in Fig. 7 ("Bronkhorst & Plomp, dL"). The studied conditions for "Andersen et al." and "Bronkhorst & Plomp, FF" are identical except for differences in the sampled interferer angles. Correspondingly, the results are seen to be similar. Since the conditions in "Bronkhorst & Plomp, dL" do not allow for binaural unmasking, the measured binaural advantages are smaller by 0-2 dB, depending on the interferer angle.

To predict binaural advantage using the studied predictors, calibration values were generated by evaluating each predictor at the SRT in condition $D_1$-6 (which has the target and interferer co-located at the front). The calibration
value for each predictor was then assumed to always correspond to the SRT. Hence, the SRT in another condition can be predicted by adaptively adjusting the input SNR to a predictor, until the predictor output is equal to the calibration value. The binaural advantage, predicted with the four studied predictors in this way, is shown in Fig. 7.

From Fig. 7, we first note that the DBSTOI measure predicts binaural advantage accurately in these conditions, as was also concluded in [44]. The DBSTOI measure predicts binaural advantage to increase as the interferer is moved away from 0°. Peaks are located both below and above 90°. This well-known phenomenon, which is most likely caused by the acoustics of the human head [27], is also seen in the "Andersen et al."-data, but not in the "Bronkhorst & Plomp, FF"-data. The binaural advantage measured in 90° varies considerably between studies [27], so this difference may be caused by slight differences in the experimental setup or the measured interferer angles.

The MBSTOI measure predicts binaural advantage to be up to 1.5 dB smaller than the DBSTOI measure does. This could be due to the MBSTOI measure also removing some bias in these conditions. Fig. 3 indicates a slight bias also in the S0N−115-condition (the DBSTOI measure does not converge to 0 for low SNRs). While this happens to decrease prediction performance of the MBSTOI measure relative to the DBSTOI measure, the MBSTOI-predictions are still well in line with the measurements. This, furthermore, should be seen in the light of the considerable variability inherent to such measurements (as seen from the error bars on Fig. 7).

As expected, the beMBSTOI measure predicts binaural advantage to be much smaller than the other methods, because it does not include the prediction of binaural unmasking. It, however, reveals the predicted better ear advantage in isolation. Similarly, the difference in predictions by the MBSTOI and beMBSTOI measures show the amount of binaural unmasking included in predictions by the MBSTOI measure. This is seen to be around 2 dB when the interferer is not near 0° or 180°. Fig. 7 also shows results from [31], of measured binaural advantage ("Bronkhorst & Plomp, FF") and of binaural advantage,
measured using modified stimuli that does not allow for binaural unmasking ("Bronkhorst & Plomp, dL"). The vertical distance between each corresponding pair of points from "Bronkhorst & Plomp, dL" and "Bronkhorst & Plomp, FF" amounts to an empirical observation of binaural unmasking. It is interesting to note that the distance between these measurements correspond well with the distance between prediction by the MBSTOI and beMBSTOI measures. Moreover, the beMBSTOI measure predicts the data points of "Bronkhorst & Plomp, dL" quite well. This indicates that the MBSTOI measure correctly estimates the separate contributions of better-ear advantage and binaural unmasking.

The B-sEPSM overestimates binaural advantage considerably. This result is somewhat surprising, as the B-sEPSM also uses an EC stage, similar to that of the MBSTOI measure, to predict binaural advantage. However, the fact that it uses a degraded signal and a clean noise signal, could possibly alter the operation of the EC stage. Furthermore, as mentioned, the underlying SIP algorithm, the mr-sEPSM, is significantly different from the STOI measure. The result is consistent with observations of [45], which, however, also indicate that the B-sEPSM performs well in more complicated binaural conditions with reverberation and fluctuating interferers. In such conditions, the performance of the MBSTOI measure is unclear. While dataset D4 shows the MBSTOI measure to cope well with some reverberation, the STOI measure, on which the MBSTOI measure is based, is known to fail in conditions with strongly fluctuating interferers [26].

5. Conclusions

In this paper we have carried out a detailed investigation of the Deterministic Binaural Short-Time Objective Intelligibility (DBSTOI) measure in acoustical conditions with distributed or multiple interferers. We showed that, in such conditions at low Signal-to-Noise Ratios (SNRs), the DBSTOI measure suffers from an upward bias. The reason for this bias was studied in detail, resulting in a modified algorithm which does not suffer from this issue: the Modified
BSTOI (MBSTOI) measure. Across five different datasets, the MBSTOI measure performed similarly to the DBSTOI measure in conditions with point-like interferers, and outperformed the DBSTOI measure in conditions with multiple or spatially distributed interferers. Within the scope of the five datasets used for testing, the MBSTOI measure essentially removed the issues relating to bias, with no apparent side effects.

6. Acknowledgments

This work was supported by the Oticon Foundation and the Danish Innovation Foundation. We thank Adam Kuklasiński for providing dataset D4.

References


