A METHOD FOR PREDICTING THE INTELLIGIBILITY OF NOISY AND NON-LINEARLY ENHANCED BINAURAL SPEECH

Asger Heidemann Andersen[⋆†], *Jan Mark de Haan*[†], *Zheng-Hua Tan*[⋆], *Jesper Jensen*[⋆†]

[⋆] Dept. of Electronic Systems, Aalborg University, 9220 Aalborg Øst, Denmark
[†] Oticon A/S, 2765 Smørum, Denmark

aha@es.aau.dk/aand@oticon.com, janh@oticon.com, zt@es.aau.dk, jje@es.aau.dk/jesj@oticon.com

ABSTRACT

We propose and evaluate a binaural speech intelligibility measure. The measure is a binaural extension of the Short-Time Objective Intelligibility (STOI) measure and focuses on predicting the intelligibility of noisy speech which has been enhanced by a speech processing algorithm (e.g. in a hearing aid). We show that the measure can accurately predict 1) the Speech Reception Threshold (SRT) for a frontal speaker masked by a point noise source in the horizontal plane, 2) the improvement in SRT obtained by independently processing the left and right ear signals with Ideal Time Frequency Segregation (ITFS), and 3) the intelligibility of speech in the presence of multiple interferers as well as the effect of processing the noisy signals with 2-microphone MVDR beamforming as used in hearing aids. Finally, we show that the computational demands associated with the measure are favourable in comparison with those of a previously proposed measure with similar properties.

***Index Terms*—** binaural speech intelligibility prediction, enhanced speech, speech in noise

1. INTRODUCTION

The topic of Speech Intelligibility Prediction (SIP) has been widely investigated since the introduction of the Articulation Index (AI) [1], which was later refined and standardized as the Speech Intelligibility Index (SII) [2]. While the research interest initially came from the telephone industry [1], the possible application to hearing aids and cochlear implants has recently gained attention [3, 4].

The SII predicts monaural intelligibility in conditions with additive, stationary noise. Another early and highly popular method is the Speech Transmission Index (STI), which predicts the intelligibility of speech, which has been transmitted through a noisy and distorting transmission system (e.g. a reverberant room) [5, 6]. Many additional SIP methods have been proposed, mainly with the purpose of extending the range of conditions under which predictions can be made (e.g. [7–17]).

For SIP methods to be applicable in relation to binaural communication devices such as hearing aids, the operating range of the classical methods must be expanded in two ways. Firstly, they must be able to take into account the non-linear processing that typically happens in such devices. This task is complicated by the fact that many SIP methods assume knowledge of the clean speech and interferer in separation; an assumption which is not meaningful when the combination of speech and noise has been processed non-linearly. One example of a method which does not make this assumption, is the STOI measure [7] which predicts intelligibility from a noisy/processed signal and a clean speech signal. The STOI measure has been shown to predict well the influence on intelligibility of multiple enhancement algorithms [7]. Secondly, SIP methods must take into account the fact that signals are commonly presented binaurally to the user. Binaural auditory perception provides the user with different degrees of advantage, depending on the acoustical conditions and the applied processing [18]. Several SIP methods have focused on predicting this advantage [11–17]. Existing binaural methods, however, can generally not provide predictions for non-linearly processed signals.

In [19] we proposed a binaural extension of the STOI measure: the Binaural STOI (BSTOI) measure. The BSTOI measure was shown to predict well the intelligibility (including binaural advantage) obtained in conditions with a frontal target and a single point noise source in the horizontal plane. The BSTOI measure was also shown to predict the intelligibility of diotic speech which had been processed by ITFS.

In this paper we present an improved version of the BSTOI measure which is computationally less demanding and, unlike BSTOI, produces deterministic results. We furthermore show that the proposed measure is able to predict intelligibility in conditions where *both* binaural advantage *and* non-linear processing simultaneously influence intelligibility. To the knowledge of the authors, no other SIP method is capable of producing predictions in conditions where intelligibility is affected by both. We refer to the improved binaural speech intelligibility measure as the Deterministic BSTOI (DBSTOI) measure.

2. THE DBSTOI MEASURE

The DBSTOI measure scores intelligibility based on four signals: The noisy/processed signal as presented to the left and right ears of the listener and a clean speech signal, also at both ears. The clean signal should be the same as the noisy/processed one, but with neither noise or processing. The DBSTOI measure produces a score in the range 0 to 1. The aim is to have a monotonic correspondence between the DBSTOI measure and measured intelligibility, such that a higher DBSTOI measure corresponds to a higher intelligibility (e.g. percentage of words heard correctly).

The DBSTOI measure is based on combining a modified Equalization Cancellation (EC) stage with the STOI measure as proposed in [19]. Here, we introduce further structural changes in the STOI measure to allow for better integration with the EC-stage. This allows for computing the measure deterministically and in closed form, contrary to the BSTOI measure [19], which is computed using Monte Carlo simulation.

The structure of the DBSTOI measure is shown in Fig. 1. The procedure is separated in three main steps: 1) a time-frequency-decomposition based on the Discrete Fourier Transformation (DFT), 2) a modified EC stage which extracts binaural advantage and 3) a modified version of the monaural STOI measure. The three steps are described in Secs. 2.1, 2.2 and 2.3, respectively.

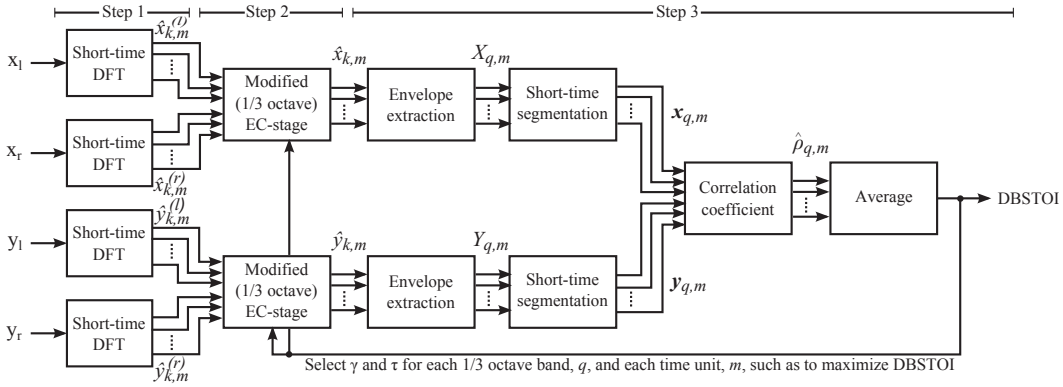

Fig. 1. A block diagram which illustrates the computation of the DBSTOI measure.

2.1. Step 1: TF Decomposition

The first step resamples the four input signals to 10 kHz, removes segments with no speech (via an ideal frame based voice activity detector) and performs a short-time DFT-based Time Frequency (TF) decomposition. This is done in exactly the same manner as for the STOI measure [7]. Let $\hat{x}^{(l)}_{k,m} \in \mathbb{C}$ be the TF unit corresponding to the clean signal at the left ear at the m'th time frame and the k'th frequency bin. Similarly, let $\hat{x}^{(r)}_{k,m}$, $\hat{y}^{(l)}_{k,m}$ and $\hat{y}^{(r)}_{k,m}$ denote the right ear clean signal and the left and right ear processed signal TF units, respectively.

2.2. Step 2: EC Processing

The second step of computing the measure combines the left and right ear signals using a modified EC stage to model binaural advantage [20,21].

A combined clean signal is obtained by relatively time shifting and amplitude adjusting the left and right clean signals and thereafter subtracting one from the other. The same is done for the noisy/processed signals to obtain a single noisy/processed signal. The relative time shift of τ (seconds) and amplitude adjustment of γ (dB) is given by the factor:

$$\lambda = 10^{(\gamma+\Delta\gamma)/40} e^{j\omega(\tau+\Delta\tau)/2}, \tag{1}$$

where $\Delta\tau$ and $\Delta\gamma$ are uncorrelated noise sources which model imperfections of the human auditory system [20–22]. The resulting combined clean signal is given by:

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

A combined noisy/processed TF-unit, $\hat{y}_{k,m}$, is obtained in a similar manner (using the same value of λ).

The uncorrelated noise sources, $\Delta\tau$ and $\Delta\gamma$, are normally distributed with zero mean and standard deviation (adapted from [22] in the same manner as is done in [11,12])[1]:

$$\sigma_{\Delta\gamma}(\gamma) = \sqrt{2} \cdot 1.5 \text{ dB} \cdot \left(1 + \left(\frac{|\gamma|}{13 \text{ dB}}\right)^{1.6}\right), \qquad \text{[dB]} \tag{3}$$

$$\sigma_{\Delta\tau}(\tau) = \sqrt{2} \cdot 65 \cdot 10^{-6} \text{ s} \cdot \left(1 + \frac{|\tau|}{0.0016 \text{ s}}\right). \qquad \text{[s]} \tag{4}$$

Following the principle introduced in [19], the values γ and τ are determined such as to maximize the scoring of intelligibility. This is covered in Sec. 2.4.

[1] In [22], noise is added separately to the left and right ear signals. Here, one noise source is applied symmetrically. This leads to a multiplicative factor of $\sqrt{2}$ in (3) and (4) compared to [22].

2.3. Step 3: Intelligibility Prediction

At this point the four input signals have been condensed to two signals: a clean signal, $\hat{x}_{k,m}$, and a noisy/processed signal, $\hat{y}_{k,m}$. We compute an intelligibility score for these signals by use of a variation of the STOI measure[2].

The clean and processed signal power envelope is determined in $Q = 15$ third octave bands:

$$\begin{aligned} X_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}|^2 \\ &\approx \alpha X^{(l)}_{q,m} + \alpha^{-1} X^{(r)}_{q,m} - 2\text{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} X^{(c)}_{q,m}\right], \end{aligned} \tag{5}$$

where $\alpha = 10^{\frac{\gamma+\Delta\gamma}{20}}$ and:

$$X^{(l)/(r)}_{q,m} = \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}^{(l)/(r)}_{k,m}|^2, \qquad X^{(c)}_{q,m} = \sum_{k=k_1(q)}^{k_2(q)} \hat{x}^{(l)*}_{k,m} \hat{x}^{(r)}_{k,m}, \tag{6}$$

and where $k_1(q)$ and $k_2(q)$ denote the lower and upper DFT bins for the q'th third octave band, respectively, and ω_q is the center frequency of the q'th frequency band. The approximate equality is obtained by inserting (1) and (2) and assuming that the energy in each third octave band is contained at the center frequency. A similar procedure for the processed signal yields third octave power envelopes, $Y_{q,m}$.

If we assume that the input signals are wide sense stationary stochastic processes, the power envelopes, $X_{q,m}$ and $Y_{q,m}$ are also stochastic processes, due to the stochastic nature of the input signals as well as the noise sources, $\Delta\tau$ and $\Delta\gamma$, in the EC stage. An underlying assumption of STOI is that intelligibility is related to the correlation between clean and noisy/processed envelopes [7]:

$$\rho_q = \frac{E[(X_{q,m} - E[X_{q,m}])(Y_{q,m} - E[Y_{q,m}])]}{\sqrt{E[(X_{q,m} - E[X_{q,m}])^2] E[(Y_{q,m} - E[Y_{q,m}])^2]}}, \tag{7}$$

where the expectation is taken across both input signals and the noise sources in the EC stage. To estimate ρ_q, the power envelopes are arranged into vectors of $N = 30$ samples [7]:

$$\mathbf{x}_{q,m} = [X_{q,m-N+1}, X_{q,m-N+2}, ..., X_{q,m}]^\intercal. \tag{8}$$

[2] For mathematical tractability, we use power envelopes rather than magnitude envelopes as originally proposed in STOI [7]. This is also done in [3] and appears not to have a significant effect on predictions [3, 23]. Furthermore, we discard the clipping mechanism contained in the original STOI, as also done in [3]. We have seen no indication that this negatively influences results.

Similar vectors, $\mathbf{y}_{q,m} \in \mathbb{R}^{N\times 1}$ are defined for the processed signal. An N-sample estimate of ρ_q across the input signals is then given by:

$$\hat{\rho}_{q,m} = \frac{E_\Delta\left[(\mathbf{x}_{q,m} - \mathbf{1}\mu_{\mathbf{x}_{q,m}})^\intercal (\mathbf{y}_{q,m} - \mathbf{1}\mu_{\mathbf{y}_{q,m}})\right]}{\sqrt{E_\Delta\left[||\mathbf{x}_{q,m} - \mathbf{1}\mu_{\mathbf{x}_{q,m}}||^2\right] E_\Delta\left[||\mathbf{y}_{q,m} - \mathbf{1}\mu_{\mathbf{y}_{q,m}}||^2\right]}}, \quad (9)$$

where $\mu_{(\cdot)}$ denotes the mean of the entries in the given vector, E_Δ is the expectation across the noise in the EC stage and $\mathbf{1}$ is the vector of all ones. A closed form expression for this expectation can be derived, and is given by (derivation omitted):

$$\begin{aligned}
&E_\Delta\left[(\mathbf{x}_{q,m} - \mathbf{1}\mu_{\mathbf{x}_{q,m}})^\intercal (\mathbf{y}_{q,m} - \mathbf{1}\mu_{\mathbf{y}_{q,m}})\right] = \\
&\quad (e^{2\beta}\mathbf{l}_{x_{q,m}}^\intercal \mathbf{l}_{y_{q,m}} + e^{-2\beta}\mathbf{r}_{x_{q,m}}^\intercal \mathbf{r}_{y_{q,m}}) e^{2\sigma_{\Delta\beta}^2} \\
&+ \mathbf{r}_{x_{q,m}}^\intercal \mathbf{l}_{y_{q,m}} + \mathbf{l}_{x_{q,m}}^\intercal \mathbf{r}_{y_{q,m}} - 2e^{\sigma_{\Delta\beta}^2/2} e^{-\omega^2\sigma_{\Delta\tau}^2/2} \times \\
&\quad \left\{ \left(e^{\beta}\mathbf{l}_{x_{q,m}}^\intercal + e^{-\beta}\mathbf{r}_{x_{q,m}}^\intercal\right) Re\left[\mathbf{c}_{y_{q,m}} e^{-j\omega\tau}\right] \right. \\
&\quad \left. + Re\left[e^{-j\omega\tau}\mathbf{c}_{x_{q,m}}^\intercal\right]\left(e^{\beta}\mathbf{l}_{y_{q,m}} + e^{-\beta}\mathbf{r}_{y_{q,m}}\right) \right\} \\
&+ 2\Big(Re\left[\mathbf{c}_{x_{q,m}}^H \mathbf{c}_{y_{q,m}}\right] + e^{-2\omega^2\sigma_{\Delta\tau}^2} Re\left[\mathbf{c}_{x_{q,m}}^\intercal \mathbf{c}_{y_{q,m}} e^{-j2\omega\tau}\right] \Big), \quad (10)
\end{aligned}$$

where:

$$\mathbf{l}_{x_{q,m}} = [X_{q,m-N+1}^{(l)}, ..., X_{q,m}^{(l)}]^\intercal - \mathbf{1}\sum_{k=m-N+1}^{m} \frac{X_{q,k}^{(l)}}{N}, \quad (11)$$

$$\mathbf{r}_{x_{q,m}} = [X_{q,m-N+1}^{(r)}, ..., X_{q,m}^{(r)}]^\intercal - \mathbf{1}\sum_{k=m-N+1}^{m} \frac{X_{q,k}^{(r)}}{N}, \quad (12)$$

$$\mathbf{c}_{x_{q,m}} = [X_{q,m-N+1}^{(c)}, ..., X_{q,m}^{(c)}]^\intercal - \mathbf{1}\sum_{k=m-N+1}^{m} \frac{X_{q,k}^{(c)}}{N}, \quad (13)$$

$$\beta = \frac{\ln(10)}{20}\gamma, \qquad \sigma_{\Delta\beta}^2 = \left(\frac{\ln(10)}{20}\right)^2 \sigma_{\Delta\gamma}^2, \quad (14)$$

and similarly for the noisy/processed signal. An expression for $E_\Delta\left[||\mathbf{x}_{q,m} - \mu_{\mathbf{x}_{q,m}}||^2\right]$ may be obtained from (10) by replacing all instances of $y_{q,m}$ by $x_{q,m}$ and vice versa for $E_\Delta\left[||\mathbf{y}_{q,m} - \mu_{\mathbf{y}_{q,m}}||^2\right]$.

The final DBSTOI measure is obtained by estimating the correlation coefficients, $\hat{\rho}_{q,m}$, for all frames, m, and frequency bands, q, in the signal and averaging across these [7]:

$$\text{DBSTOI} = \frac{1}{QM}\sum_{q=1}^{Q}\sum_{m=1}^{M} \hat{\rho}_{q,m}, \quad (15)$$

where Q and M is the number of frequency bands and the number of frames, respectively.

It can be shown that whenever the left and right ear inputs are identical, the DBSTOI measure produces scores which are identical those of the monaural STOI (that is, the modified monaural STOI measure based on (5) and without clipping).

2.4. Determination of γ and τ

Finally, we consider the parameters γ and τ. These parameters are determined individually for each time unit, m, and third octave band, q, such as to maximize the final DBSTOI measure. Thus, each correlation coefficient estimate is a function of its own set of parameters, $\hat{\rho}_{q,m}(\gamma,\tau)$. The DBSTOI measure, (15), can therefore be maximized by maximizing each of the estimated correlation coefficients individually:

$$\hat{\rho}_{q,m} = \max_{\gamma,\tau} \hat{\rho}_{q,m}(\gamma,\tau). \quad (16)$$

In practice, $\hat{\rho}_{q,m}$ is evaluated for a discrete set of γ and τ values and the highest value is chosen.

3. RESULTS

The DBSTOI measure accepts binaural input signals which have been non-linearly processed, and is therefore applicable to a large range of acoustical conditions. We investigate the prediction performance of the measure in a selection of conditions: 1) speech masked by a single additive point noise source in the horizontal plane (a condition often used for evaluation of binaural intelligibility predictors [11, 12, 14–16]), 2) speech masked by a single point source with separate non-linear enhancement for each ear (to the knowledge of the authors, no other method is capable of providing predictions under such a condition), and 3) speech masked by multiple interferers and linearly processed with a beamformer.

3.1. Frontal Target and Point Interferer With and Without ITFS

First, we investigate the ability of the DBSTOI measure to predict SRTs[3]. Predictions are compared to the results of two experiments where SRTs were measured in normal hearing Danish subjects. In both experiments, we simulated a binaural anechoic environment by use of Head Related Transfer Functions (HRTFs) [24] and presented the resulting binaural signals through Sennheiser HDA200 headphones at a comfortable level. The target signal consisted of sentences from the Dantale II corpus [25], always emanating from a point source directly in front of the subject. The target signal was masked by a single point noise source, located in the horizontal plane. The further specifics of the two experiments are given by:

Experiment 1: Speech reception was measured in 10 conditions, differing only by the location of a single Speech Shaped Noise (SSN) interferer in the horizontal plane. The subjects listened to one five-word sentence at a time, repeating whatever words were heard. The experimenter marked the correctly identified words. The experiment was carried out for 10 normal hearing adult subjects. For each condition three repeated measurements were taken at 6 different Signal to Noise Ratios (SNRs). This resulted in the scoring of 10 subjects $\times$ 10 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1800 sentences. ■

Experiment 2: Speech reception was measured in 9 conditions. Conditions 1–3 used SSN interferers at different positions in the horizontal plane. The left and right ear signals were independently subjected to ITFS with an Ideal Binary Mask (IBM) [26] as follows. The target and interferer signals were TF-decomposed with a short-time DFT, and TF-units with an SNR of less than 0 dB were attenuated by 10 dB. This finite attenuation was chosen to limit the improvement in intelligibility. Conditions 4–6 used the same interferer positions, but used bottle factory hall noise [27], rather than SSN. This noise type, which is a recording of bottles on a conveyor belt, has more energy at higher frequencies compared to SSN and is highly non-stationary. Conditions 4–6 did not include ITFS. Conditions 7–9 were the same as conditions 4–6 but with ITFS. The subjects were asked to select the words they heard on a screen. For each of the five words the subjects were shown 10 possible words, as well a pass-button to indicate that the given word had not been heard. A similar procedure is investigated in [28] and is shown to give results almost identical to those of the procedure used in experiment 1. Each condition was tested for 13 subjects, each at 6 SNRs and repeated 3 times. This resulted in the scoring of 13 subjects $\times$ 9 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 2106 sentences. ■

SRTs were determined individually for each subject for each condition above. This was done by performing a maximum likelihood fit of a logistic function to the measured data as described in [29]. The SRTs were

[3] The 50% SRT is the SNR where the subject scores 50% correct words.

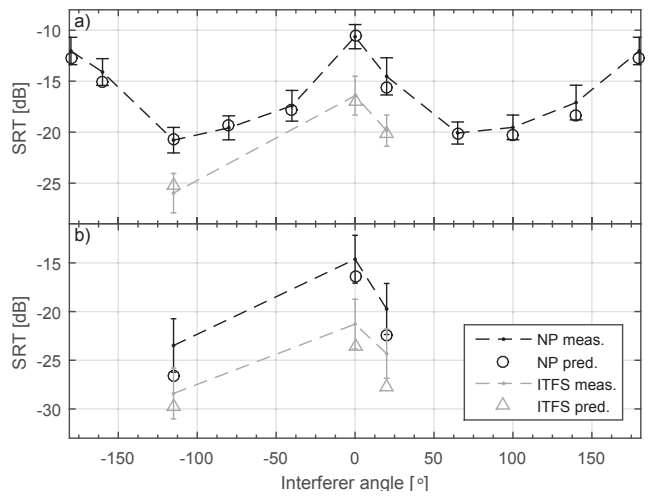

Fig. 2. Measured ("meas.") and predicted ("pred.") SRTs with ("ITFS") and without ITFS ("NP"). a) Frontal speech masked by a single SSN interferer and b) frontal speech masked by a single bottling factory hall noise interferer. The error bars indicate the standard deviation of the measured SRTs across subjects.

averaged across all subjects to obtain one mean SRT for each condition. To predict SRTs with the DBSTOI measure, a calibration constant was determined by scoring the condition with both target and SSN interferer in front, at an SNR equal to the SRT measured for this condition. SRT predictions in all the conditions were then made by adaptively varying the input SNR until the DBSTOI score was close to the calibration score.

The results of both measurements and predictions are shown in Fig. 2. The upper plot, showing conditions with SSN masking, indicates that by calibrating the DBSTOI to a single condition, it is possible to predict the results of all the other combinations of ITFS and interferer positions to within less than one standard deviation of the measurements. In the bottom plot, showing conditions with bottle factory masking, a downward bias of 1-3 dB is seen in the predictions. This is most likely due to the method being calibrated to SSN masking. Note, though, that the relative effects of interferer position and ITFS are still predicted accurately.

3.2. Frontal Target and Multiple Interferers With/Without Beamforming

In this section, we evaluate the DBSTOI measure for a range of conditions with multiple interferers. An experiment similar to experiment 1 discussed in Section 3.1 was carried out. Ten normal hearing subjects were presented with Dantale II sentences in six different conditions for a scoring of 10 subjects $\times$ 6 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1080 sentences in total. These were averaged across subjects and repetitions to produce a total of 36 data points. All six conditions were anechoic with speech originating from the front. The first condition was contaminated by isotropic ("Iso") SSN. The second condition was contaminated by uncorrelated SSN from point sources at 110°, 180° and -110° in the horizontal plane ("3s"). The third condition considered the same layout of noise sources as condition two, but used three different segments of the International Speech Test Signal (ISTS) as noise [30]. Conditions four to six were the same as one to three, but include monaural 2-microphone Minimum Variance Distortionless Response (MVDR) beamforming as used in a behind-the-ear hearing aid ("BF"). DBSTOI scorings were made for each of the 6 conditions and 6 SNRs. Fig. 3 shows the results.

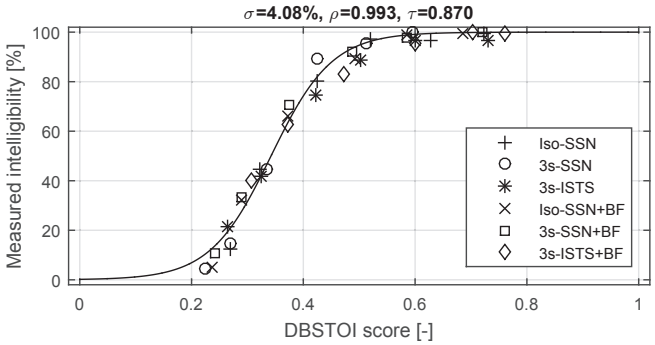

Fig. 3. The results of the multiple-interferer experiment vs. computed DBSTOI scores. A fitted logistic function is shown too. Standard deviation (σ) and Pearson correlation (ρ) are computed relative to the fitted logistic function. The Kendall rank correlation (τ) is also shown.

STOI	BSTOI	DBSTOI
5.3 s	1086.3 s	62.2 s

Table 1. 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 implementations of the BSTOI and DBSTOI measures were used, while a STOI measure implementation was provided by the authors of [7].

Although the investigated conditions are highly diverse, the DBSTOI predictions appear to be very well in line with the measured intelligibility.

3.3. Computational Cost

A key motivation for the DBSTOI measure is to avoid the necessity of Monte Carlo simulation, as was the case for the BSTOI measure proposed in [19]. It is therefore expected that the DBSTOI measure is computationally less demanding than the BSTOI measure. To verify this, the measures were each used to score 100 seconds of white noise (the computational demand of computing the measures is independent of signal type). This was done simply with the `timeit`-function in MATLAB. The results are shown in Table 3.3. Evaluating the DBSTOI measure is approximately 12 times more time consuming than evaluating the monaural STOI measure. Evaluating the Monte-Carlo-based BSTOI measure is, however, more than 17 times as time consuming as evaluating the proposed DBSTOI measure.

4. CONCLUSIONS

In this paper we present and investigate a binaural speech intelligibility measure, DBSTOI, which accepts input signals that have been processed by e.g. a speech enhancement algorithm. The measure is obtained by combining the Short-Time Objective Intelligibility (STOI) measure with a modified Equalization Cancellation (EC) stage. The presented measure improves upon the previously proposed BSTOI measure by providing deterministic results at a lower computational cost. We demonstrate that the measure is able to predict accurately the effect on intelligibility of simultaneous non-linear signal enhancement and binaural advantage, in addition to simpler conditions with non-linear enhancement or binaural advantage separately. We furthermore show that the computational costs associated with the proposed DBSTOI measure is more than 17 times lower than those of the previously proposed BSTOI measure.

5. REFERENCES

[1] 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.

[2] American National Standards Institute, "S3.5-1997: Methods for calculation of the speech intelligibility index," 1997.

[3] 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)*, Bucharest, Romania, Aug. 2012, pp. 504–508.

[4] 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.

[5] 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.

[6] 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.

[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, Sept. 2011.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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, Sept. 2010.

[18] 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.

[19] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *INTERSPEECH*, Dresden, Germany, Sept. 2015, pp. 2563–2567.

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] M. Vestergaard, "The eriksholm cd 01: Speech signals in various acoustical environments," Tech. Rep. 050-08-01, Oticon Research Centre, Eriksholm, Snekkersten, 1998.

[28] E. R. Pedersen and P. M. Juhl, "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.

[29] 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, June 2002.

[30] 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.