A Binaural Short Time Objective Intelligibility Measure for Noisy and Enhanced Speech

Asger Heidemann Andersen\textsuperscript{1,2}, Jan Mark de Haan\textsuperscript{2}, Zheng-Hua Tan\textsuperscript{1}, Jesper Jensen\textsuperscript{1,2}  
\textsuperscript{1} Dept. of Electronic Systems, Aalborg University, 9220 Aalborg Øst, Denmark  
\textsuperscript{2} Oticon A/S, 2765 Smørum, Denmark  
aha@es.aau.dk/aand@oticon.com, jmh@oticon.dk, zt@es.aau.dk, jje@es.aau.dk/jsj@oticon.dk

Abstract

Objective intelligibility measures are increasingly being used to assess the performance of speech processing algorithms, e.g., for hearing aids. It has been shown that the short time objective intelligibility (STOI) measure yields good results in this respect. In this paper we propose a binaural extension of the STOI measure, which predicts binaural advantage using a modified equalization cancellation (EC) stage. The proposed method is evaluated for a range of acoustic conditions. Firstly, the method is able to predict the advantage of spatial separation between a speech target and a speech shaped noise (SSN) interferer. Secondly, the method yields results comparable to the monaural STOI measure when presented with noisy speech processed by ideal time-frequency segregation (ITFS). Finally, the method also performs well when presented with a selection of different acoustic conditions combined with beamforming as used in hearing aids.

Index Terms: binaural speech intelligibility prediction, enhanced speech, speech in noise

1. Introduction

The interest in predicting the intelligibility of noisy speech was initially sparked by the telephone industry where speech intelligibility is of key importance [1]. It is cumbersome and expensive to guide the development in such an industry purely through human listening tests, thus motivating the need for objective means of Speech Intelligibility Prediction (SIP).

The subject of 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]. The SII predicts the intelligibility of monaural speech in additive stationary noise. Another early method is the Speech Transmission Index (STI), which focuses on noisy and distorting transmission systems (e.g., telephone lines) [3, 4]. While the SII and STI have been shown to correlate well with human speech intelligibility in some conditions, they work poorly in others. E.g., the methods do not work well when non-linear signal processing is applied to the noisy speech [5, 6]. This has led to the development of a multitude of SIP methods with different characteristics.

One such characteristic is the ability to predict the intelligibility of noise-reduced or enhanced speech [6–12]. The Short Time Objective Intelligibility (STOI) measure has been successful in predicting the intelligibility of noisy speech which has been processed by Time Frequency (TF) weighting methods [6], e.g., as used in noise reduction systems.

Another characteristic is the ability to predict the advantage obtained through binaural listening. It is well known that humans may obtain a great advantage in conditions where the speech and noise sources are spatially separated [13]. This effect can be predicted by the Equalization Cancellation (EC) model [14, 15], which has been used in the development of binaural SIP methods. The Binaural Speech Intelligibility Measure (BSIM) is essentially the combination of the EC model with the SII [16, 17]. Another EC-based binaural SIP method is proposed in [18–20], which focuses on predicting binaural advantage for different room acoustical conditions. Both of these methods, however, perform SIP for the case of additive noise and no processing.

These characteristics are both highly relevant in many areas, e.g., for the development of hearing aids and cochlear implants [12, 21]. Modern hearing aids apply non-linear speech enhancement algorithms and may modify the binaural cues presented to the user. Hence, binaural SIP of enhanced speech could be used to guide developments in this field. Unfortunately, to the knowledge of the authors, no existing SIP method can handle binaural processed signals (e.g., predict the effect that a speech enhancement algorithm has on binaural advantage).

In this paper, we propose a binaural SIP method which is able to predict the intelligibility of noisy, potentially processed, speech. Specifically, we combine a modified EC stage with the STOI measure to obtain such a method. We refer to the proposed method as the Binaural STOI (BSTOI) measure. The method is presented and evaluated for a range of conditions including ideal Time Frequency Segregation (ITFS) and spatial separation between target and interferer.

2. The BSTOI Method

The objective of the BSTOI measure is to predict the intelligibility of binaural speech which may be corrupted by additive noise and processed by a speech enhancement algorithm. The method is intrusive in the sense that it assumes knowledge of the clean speech signal as a reference. The method therefore takes four inputs: the clean speech as recorded at the left and right ears of the listener, and the corresponding noisy/processed speech at the left and right ears. As output, the method produces an index in the range of 0 to 1, which should correlate well with the fraction of the spoken words which are understandable to a normal hearing listener. The BSTOI method is outlined by the block diagram in Figure 1.

2.1. Step 1: TF Decomposition

The first step of the method is to perform a short time DFT-based TF decomposition of the four input signals. This is done in exactly the same manner as in the STOI-method [6]. Let $\hat{x}_{k,m}^{(l)} \in \mathbb{C}$ 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 $\hat{y}_{k,m}^{(r)}$ and $\hat{y}_{k,m}^{(r)}$ denote the right ear clean signal and the processed left and right ear signal TF units.
2.2. Step 2: EC Processing

The second step of the method combines the left and right ear signals using a modified EC stage to model binaural advantage. The original EC stage assumes four inputs: left/right ear clean signal (e.g. speech) and left/right ear noise [14, 15]. The method time shifts and amplitude adjusts the left and right ear signals relative to one-another and are thereafter subtracted. The processed signals are treated similarly, to obtain a combined TF equalization and then proceeds to subtract the left ear signals from the right ear signals (cancellation) to obtain two output signals: one clean difference signal and one noise difference signal. The level of time shift and amplitude adjustment between the left and right ears is determined such that the Signal to Noise Ratio (SNR) at the output is maximized. The BSTOI method assumes access to the clean input signals but the noise is not assumed separately available. Instead, a potentially processed combination of speech and noise is available. The SNR is not directly computable from these signals, and therefore a conventional EC stage cannot be applied. Instead, we modify the EC stage by introducing two particular changes:

1. The left/right clean signals and the left/right processed/noisy signals are combined in exactly the same manner as the left/right clean signals and the left/right noise signals are combined in the conventional EC stage.

2. Instead of relying on maximizing the SNR for determining the relative time and level adjustments between the ears, the BSTOI measure of the resulting clean and processed signals is maximized.

The second modification is inspired by [16, 17] who found time and level adjustments in the EC stage, that maximized the SI. The linear combination of left and right ear clean signals in the EC stage is described by:

\[ x_{l/m} = \lambda \hat{x}_{l/m}^{(l)} - \lambda^{-1} \hat{x}_{l/m}, \]

where:

\[ \lambda = 10^{(\gamma + \Delta \gamma)/40} e^{j\omega(\gamma + \Delta \gamma)/2}. \]

The factor \( \lambda \) implements the equalization step in the EC stage. The left and right ear signals are level adjusted by \( \gamma \) (in dB) and time shifted by \( \tau \) relative to one-another and are thereafter subtracted. The processed signals are treated similarly, to obtain a combined TF signal. To obtain performance similar to that of humans, the EC stage adds noise sources, \( \Delta \gamma \) and \( \Delta \tau \), to the values of \( \gamma \) and \( \tau \) [14, 15, 22]. These noise sources are normally distributed with zero mean and standard deviation (adapted from [22] in the same manner as is done in [16, 17]):

\[ \sigma_{\Delta \gamma} = \sqrt{2} \sigma_{\Delta \gamma_0} \left(1 + \frac{|r|}{\alpha_0} \right)^p, \] (3)

\[ \sigma_{\Delta \tau} = \sqrt{2} \sigma_{\Delta \tau_0} \left(1 + \frac{|\tau|}{\tau_0} \right), \] (4)

with \( \sigma_{\Delta \gamma_0} = 1.5 \) dB, \( \alpha_0 = 13 \) dB, \( p = 1.6 \), \( \sigma_{\Delta \tau_0} = 65 \) ms and \( \tau_0 = 1.6 \) ms. The determination of the values \( \gamma \) and \( \tau \) is covered in Section 2.4.

2.3. Step 3: Intelligibility Prediction

At this point, the method progresses like the STOI method. The clean and processed signal envelopes are determined in \( Q = 15 \) third octave bands [6]:

\[ X_{q,m} = \frac{1}{k_1(q)-1} \sum_{k=k_1(q)}^{k_2(q)} \left| \hat{x}_{q,m} \right|^2, \] (5)

where \( k_1(q) \) and \( k_2(q) \) denote the lower and upper DFT bins for the \( q \)’th third octave band. The same is done for the processed signal to obtain third octave envelopes, \( Y_{q,m} \). These envelopes are arranged into vectors of \( N = 30 \) samples [6]:

\[ x_{q,m} = [X_{q,m} - N + 1, X_{q,m} - N + 2, \ldots, X_{q,m}]^\top. \] (6)

Similar vectors, \( y_{q,m} \in \mathbb{R}^{N \times 1} \) are defined for the processed signal. The processed envelope is subjected to a clipping procedure which bounds the sensitivity of the model to severely degraded frames (see [6] for details):

\[ \hat{y}(n) = \min\left( \frac{\|X_{q,m}\|}{\|\tilde{y}_{q,m}(n)\|}, \left(1 + 10^{-\beta/20}\right)\|X_{q,m}(n)\| \right), \] (7)

with \( \beta = -15 \) (dB) and \( n = 1, 2, \ldots, N \). Each clean signal envelope is then correlated with the corresponding clipped processed signal envelope [6]:

\[ d_{q,m} = \frac{\|X_{q,m} - \mu_{q,m}\|}{\|\tilde{y}_{q,m} - \mu_{q,m}\|} \frac{\|\tilde{y}_{q,m} - \mu_{q,m}\|}{\|Y_{q,m} - \mu_{q,m}\|}, \] (8)

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

Figure 1: A block diagram which illustrates the computation of the BSTOI measure. The measure has the option of computing the intermediate correlation coefficients, \( d_{q,m} \), in a better-ear fashion whenever favourable. This is omitted from the figure for simplicity.
where \( \mu(\cdot) \) denotes the mean of the entries in the corresponding vector. The final BSTOI measure is obtained by averaging these intermediate correlation coefficients [6]:

\[
\text{BSTOI} = \frac{1}{QM} \sum_{q=1}^{Q} \sum_{m=1}^{M} d_{q,m},
\]

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

### 2.4. Determination of \( \gamma \) and \( \tau \)

Finally, we consider the determination of the parameters \( \gamma \) and \( \tau \). As stated, these are determined such as to maximize the final BSTOI measure. The parameters are determined individually for each time unit, \( m \), and third octave band, \( q \). Thus, each intermediate correlation coefficient is a function of its own set of parameters, \( d_{q,m} = d_{q,m}(\gamma, \tau) \). The BSTOI measure, (9), can therefore be maximized by maximizing each of the intermediate correlation coefficients individually:

\[
d_{q,m} = \max_{\gamma,\tau} d_{q,m}(\gamma, \tau).
\]

In practice, \( d_{q,m} \) is evaluated for a discrete set of \( \gamma \) and \( \tau \) values (a “grid”) and the highest value is chosen.

It should be noted that the BSTOI measure is stochastic due to the noise sources, \( \Delta \gamma \) and \( \Delta \tau \), in (2). The approximately optimal values of \( \gamma \) and \( \tau \) are therefore first determined without the noise. Using these \( \gamma \) and \( \tau \) values, the measure is averaged over ten noise realizations. In addition, the intermediate correlation coefficients are computed for the left and right ears separately, with the monaural STOI method [6]. Whenever a single ear provides a higher correlation than the corresponding EC processed alternative, the better-ear correlation is used.

### 3. Evaluation

Since the BSTOI measure accepts the combination of binaural and processed input signals, it is applicable to a wide range of acoustical conditions. Here, we evaluate the method against results from three different listening tests spanning different types of conditions.

#### 3.1. \( S_0N_0 \) with Dantale Target and SSN Interferer

We first investigate predictions of the binaural advantage in the case of a frontal speech target and a single Speech Shaped Noise (SSN) interferer from different directions, \( \theta \), in the horizontal plane under anechoic conditions (denoted \( S_0N_0 \)) [13, 18]. Sentences from the Danish Dantale II corpus [23] were used as target material, while SSN was generated by filtering Gaussian noise to obtain the same long-time average spectrum as that of the speech material. Both speech and noise was filtered with large pinnae KEMAR Head Related Transfer Functions (HRTFs) from the CIPIC database [24] to simulate the required binaural conditions. The signals were presented to the subjects by headphones. Ten normal hearing subjects participated in the listening experiment. Each subject was presented with sentences in noise from ten different angles and six fixed SNRs. The subjects were instructed to repeat any words they heard, and the number of correct words were recorded for each presented sentence. This resulted in the scoring of (10 subjects)\( \times \) (10 interferer angles)\( \times \) (6 SNRs)\( \times \) (3 repetitions) = 1800 sentences. The resulting data was averaged across all subjects and repetitions to obtain a total of 60 average scores. Logistic functions were fitted to data from each interferer angle, and 50% Speech Reception Thresholds (SRTs)\(^2\) were estimated from these.

![Figure 2: Measured and predicted SRTs for a frontal target and ten different interferer angles. Measured data points are connected by lines for ease of viewing.](image)

Figure 2 compares the measured SRTs with the predictions of three different binaural SIP methods, including the BSTOI measure. The method denoted as Jelfs was introduced in [19] and an implementation from the Auditory Modelling Toolbox [25] was used. The method predicts only relative SRTs and the predictions have therefore been offset such that the predictions match the measured results exactly in the \( S_0N_0 \)-condition. The stBSIM method was implemented as best as possible according to the description given in [17]. The Jelfs and stBSIM methods were both supplied with frontal SSN as target and SSN from different directions as interferer (to most closely resemble the conditions under which these methods have previously been validated [17, 19]).

The BSTOI was supplied with an input signal consisting of 30 (randomly chosen) concatenated Dantale II sentences as speech and binaural SSN as interferer. The processed signal input to the BSTOI measure was generated simply as the sum of clean speech and interferer (as no processing was carried out). SRTs were determined with the BSTOI measure in a manner similar to that used in [17, 26, 27]: The BSTOI output at the measured SRT in the \( S_0N_0 \)-condition was computed and used as a reference. The SRTs of other conditions were then predicted by adaptively varying the input SNR until this reference value was found as the output of the BSTOI measure.

Figure 2 shows that BSTOI is able to predict the binaural advantage rather accurately. The largest prediction error is just under 0.7 dB. The performance of BSTOI is similar to that of Jelfs and stBSIM. The results suggest that the changes introduced to the EC model, to merge it with the monaural STOI measure, do not significantly impair its performance under the studied conditions. Furthermore, they indicate that the use of STOI in conjunction with the EC model is a viable approach to predicting binaural speech intelligibility.

#### 3.2. Ideal Time Frequency Segregation

The STOI measure has shown to perform well at predicting the intelligibility of speech which has been processed by ITFS [6]. It is obviously desired that this property is retained in the binaural extension of the measure. We therefore applied the BSTOI measure to the same collection of monaural ITFS data as used in [6]. The

\(^2\)The 50% SRT is the SNR where the subject scores 50% correct words.
data presented in [28] consist of clean and noisy/processed speech signals from the Dantale II corpus as well as the results from a listening test with 15 normal hearing subjects. The signals are contaminated by four noise types: SSN as well as cafeteria, bottling factory and car interior noise [28]. The processing consist of ITFS with two different mask types: Ideal Binary Masks (IBMs) and Target Binary Masks (TBMs) [28]. The listening test has been carried out in a manner quite similar to that described in the previous section (see [28] for details). The listening test results were averaged across all subjects and repetitions for each condition. Each condition was scored by the STOI and BSTOI measures using 30 concatenated Dantale II sentences as input. The signals were presented diotically to the BSTOI measure (the same signals were averaged across all subjects and repetitions for each condition. The listening test has been carried out in a manner quite similar to that described in the previous section (see [28] for details). The listening test results were averaged across all subjects and repetitions for each condition. Each condition was scored by the STOI and BSTOI measures using 30 concatenated Dantale II sentences as input. The signals were presented diotically to the BSTOI measure (the same signals were given for the left and right ear inputs).

Figure 3 shows the resulting scores vs. the average listening test scores. The results for the STOI measure are close to those presented in [6] as should be expected. Slight variations occur from the use of different Dantale II sentences as input. The BSTOI measure shows predictions which are similar to those of STOI albeit slightly higher. This relative difference is not important as none of the measures predict intelligibility relative to any fixed reference. A number of key measures are shown on the figure as well.

Figure 3: Measured intelligibility (averaged over 15 test subjects and multiple sentence scores) vs. STOI and BSTOI measures, and fitted logistic curves. The values above the plot show sample standard deviation from the respective logistic curves ($\sigma$), correlation coefficient ($\rho$), and Kendall’s tau ($\tau$).

3.3. Various Conditions with and without Beamforming

Lastly, we evaluate the BSTOI measure for a selection of different conditions. A listening experiment was carried out in a manner almost identical to that discussed in Section 3.1. Ten normal hearing subjects were presented with Dantale II sentences in six different conditions for a scoring of (10 subjects)×(6 conditions)×(6 SNRs)×(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. The ISTS is a speech-like noise signal created from recorded speech, but which is largely non-intelligible [29]. Conditions four to six are 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”). The signals which were presented to the subjects were saved. BSTOI predictions were made for each of these signals and averaged in the same way as the measured data.

Figure 4: The averaged results of the third discussed listening test vs. the averaged BSTOI predictions. A fitted logistic function is shown as well.

4. Conclusions

In this paper we present a binaural extension, BSTOI, of the Short Time Objective Intelligibility (STOI) measure. The method is based on extending the STOI measure using a modified Equalization Cancellation (EC) stage to model binaural advantage. Unlike existing methods, BSTOI is able to predict intelligibility for signals which are binaural and have been processed by speech enhancement algorithms. Initial experiments show promising results. Firstly, they indicate that BSTOI can predict both 1) the binaural advantage of spatial separation between a target and an interferer and 2) the intelligibility of ITFS processed speech diotically. Secondly, BSTOI predicts intelligibility well in a variety of different conditions with and without linear processing (beamforming). The investigated conditions, however, cover only a small part of the domain of input signals, processing methods and acoustical conditions to which the measure is theoretically applicable, and the evaluation can therefore only be considered as an initial validation. Future work includes further investigation of the proposed method for more complicated combinations of spatial layouts and binaural non-linear processing.

5. Acknowledgements

This work was funded by the Oticon Foundation and the Danish Innovation Foundation.
6. References


