An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers

Jesper Jensen and Cees H. Taal

Abstract—Intelligibility listening tests are necessary during development and evaluation of speech processing algorithms, despite the fact that they are expensive and time-consuming. In this paper, we propose a monaural intelligibility prediction algorithm, which has the potential of replacing some of these listening tests. The proposed algorithm shows similarities to the Short-Time Objective Intelligibility (STOI) algorithm but works for a larger range of input signals. In contrast to STOI, Extended STOI (ESTOI) does not assume mutual independence between frequency bands. ESTOI also incorporates spectral correlation by comparing complete 400-ms length spectrograms of the noisy/processed speech and the clean speech signals. As a consequence, ESTOI is also able to accurately predict the intelligibility of speech contaminated by temporally highly modulated noise sources in addition to noisy signals processed with time-frequency weighting. We show that ESTOI can be interpreted in terms of an orthogonal decomposition of short-time spectrograms into intelligibility subspaces, i.e., a ranking of spectrogram features according to their importance to intelligibility. A free Matlab implementation of the algorithm is available for non-commercial use at http://kom.aau.dk/~jje/.

I. INTRODUCTION

When developing speech communication systems for human receivers, listening tests play a major role both for monitoring progress in the development phase and for verifying the performance of the final system. Often, listening tests are used to quantify aspects of speech quality and speech intelligibility. Although listening tests constitute the only tool available for measuring ground-truth end-user impact, they are time-consuming, they may require special auditory stimuli data and test equipment, and they require the availability of a group of typical end-users. For these reasons, listening tests are costly and can typically not be employed many times during the development phase of a speech communication system. Hence, cheaper alternatives or supplements are of interest.

In this paper, we focus on intrusive, monaural intelligibility prediction models, i.e., algorithms which — rather than conducting an actual listening test — predict the outcome of the listening test based on the auditory stimuli of the test. Historically, two lines of research serve as the foundation for existing intelligibility prediction models: i) the Articulation Index (AI) [1] by French and Steinberg [2], which was later refined and standardized as the Speech Intelligibility Index (SII) [3], and ii) the Speech Transmission Index (STI) [4] by Steeneken and Houtgast [5].

AI and SII were developed with simple linear signal degradations, e.g., additive noise, in mind. To estimate intelligibility, the methods divide the signal under analysis into frequency subbands and assume that each subband contributes independently to intelligibility. The contribution of a subband is found by estimating the long-term speech and noise power within the subband to arrive at the long-term subband signal-to-noise ratio (SNR). Then, subband SNRs are limited to the range from -15 to +15 dB, normalized to a value between 0 and 1, and combined as a perceptually weighted average.

STI extends the range of distortions to include convolutive noise, e.g., reverberant speech and effects of room acoustics. STI is based on the observation that reverberation and/or additive noise tend to reduce the depth of temporal signal modulations compared to the clean, undistorted reference signal. To measure changes in the modulation transfer function, STI generates bandpass filtered noise probe signals at different center-frequencies, and amplitude-modulates each such signal at different modulation frequencies relevant to speech intelligibility. Each modulated probe signal is then passed through the communication channel in question (e.g. characterised by a room impulse response), and the reduction in modulation depth is finally translated into an intelligibility index.
Despite the importance of AI and SII, the methods have a number of limitations. First, they require the long-term spectrum of the additive noise signal to be known in advance. Secondly, since the methods rely on long-term statistics, they cannot discern modulated noise signals from un-modulated ones, when their long-term spectra are identical. In other words, the intelligibility of speech contaminated by modulated and un-modulated noise is judged to be identical, although it is well-known that this is generally not the case, e.g. [6], [7]. Finally, AI and SII are not directly applicable to signals which have been passed through some non-linear processing stage before presentation to the listener, because in this case it is no longer clear which noise spectrum to use.

Various methods have been proposed to reduce the limitations mentioned above and extend the range of acoustic situations for which intelligibility prediction can be made. In [6], Rhebergen et al. proposed the Extended SII (ESII), which avoids the use of long-term noise spectra in SII. Specifically, ESII divides the masker signal into short time frames (9–20 ms) and averages the SII computed for each frame individually, to predict intelligibility for fluctuating noise sources. In a somewhat similar manner, the Glimpse Model by Cooke [8] uses realizations of speech and additive noise signals to estimate the glimpse percentage, i.e., the fraction of time-frequency tiles whose SNR exceeds a certain threshold, which is then translated to an intelligibility estimate. The Coherence SII (CSII) by Kates et al. [9] extends SII to better take into account various non-linear distortions, including center- and peak-clipping. Similarly, Goldsworthy et al. [10] proposed a modified STI approach, speech STI (sSTI), which replaces the traditional noise probe signals with actual speech signals. This was done to better take into account the effect of non-linear distortions such as envelope clipping [11], and dynamic amplitude compression [12].

More recently, methods have emerged which are inspired by both original lines of research. For example, Jørgensen et al. [13] decompose the speech signal and an additive noise masker in a modulation filter bank. Intelligibility prediction is then based on the envelope/ power SNR at the output of this filter bank. Taal et al. [14] proposed the Short-Time Objective Intelligibility (STOI) measure, which extracts temporal envelopes of undistorted and noisy/processed speech signals in frequency subbands. The envelopes are then subject to a clipping procedure, compared using short-term linear correlation coefficients, and a final intelligibility prediction is constructed simply as an average of the correlation coefficients. STOI has proven to be able to predict quite accurately the intelligibility of speech in many acoustic situations, including the speech output of mobile phones [15], noisy speech processed by ideal time-frequency masking and single-channel speech enhancement algorithms [14], speech processed by cochlear implants [16], and STOI appears robust to different language types incl. Danish [14], Dutch [17], and Mandarin [18]. Although STOI performs well in many cases, some of the algorithmic choices, e.g. the use of a linear correlation coefficient as a basic distance measure, are less well motivated from a theoretical point of view. However, in [17] Jensen et al. proposed the Speech Intelligibility prediction based on Mutual Information (SIMI) method, which suggests that characteristics of STOI may be explained using information theoretic arguments.

STOI, and several of the methods described above, show only weak links to the properties of the auditory system, and much more elaborate models have been proposed, e.g. [19], [20]. For example, the Hearing-Aid Speech Perception Index (HASPI) by Kates et al. [20] employs level- and hearing-profile dependent auditory filterbanks to compute quantities resembling mel-frequency cepstral coefficients (MFCCs) for the clean reference signal and the noisy/processed signal, respectively. Then, long-term correlations between clean and noisy/processed MFCCs are computed, before the average across the cepstral dimension is found. Finally, this cepstral correlation average is combined with estimates of the auditory signal coherence for low-, mid-, and high-intensity signal regions [9], to form an intelligibility index.

Existing intelligibility prediction methods may be divided into two classes: 1) methods which require that the target speech signal and the distorting component (e.g., the additive noise) are available in separation, e.g., [3], [4], [6], [8], [13] and 2) methods which do not impose this requirement, e.g., [9], [10], [14], [17]. Class-1 methods have the advantage that they can use the access to speech and noise realizations to compute SNR realizations in different time-frequency regions and find an intelligibility index based on these. Class-2 methods, on the other hand, cannot observe SNRs directly, but must rely on features estimated from, generally, limited data, e.g., short-time correlations estimated from the noisy/processed and clean speech signal. The disadvantage of Class-1 methods is that they are not applicable to non-linearly processed noisy speech signals, because in this situation noise and speech signals are not readily available in separation. Class-2 methods are more generally applicable, i.e., also to noisy signals, which have been non-linearly processed.

As reported in [21], STOI – and as we show in Sec. IV – other Class-2 methods have limitations for target speech signals in additive noise sources with strong temporal modulations, as e.g. a single competing speaker. To demonstrate this point, Fig. 1 shows an example of intelligibility predicted by STOI vs. actual intelligibility, measured in listening tests with speech signals degraded by 10 highly modulated masker signals at different SNRs (details are given in Table I and will be discussed later). Clearly, STOI performs less well in this situation: the linear correlation coefficient between predicted and measured intelligibility is as low as $\rho = 0.47$.

In this paper, we propose a new Class-2 intelligibility predictor – ESTOI (Extended Short-Time Objective Intelligibility) – which, unlike many existing Class-2 methods, works well for highly modulated noise sources as the example above\(^1\). Importantly, ESTOI also works well in situations, where existing Class-2 methods work well.

As the name suggests, ESTOI is inspired by STOI [14]. As with STOI, ESTOI operates within a 384 ms analysis window on amplitude envelopes of subband signals. This analysis window is used in order to include important temporal

\(^1\)The algorithm name ESTOI follows the terminology introduced by Rhebergen et al., who used the name Extended SII (ESII) for their algorithm to improve the performance of SII for modulated noise maskers [6].
modulation frequencies relevant for speech intelligibility [14].

To understand the differences between \textit{STOI} and \textit{ESTOI}, let us first interpret, how \textit{STOI} computes a correlation coefficient for a 384 ms analysis window. \textit{STOI} first computes linear correlation coefficients for each subband between undistorted and noisy/processed signals\(^2\); this is equivalent to computing inner products between mean- and variance- normalized envelope signals. To find a correlation coefficient for the 384 ms analysis window, \textit{STOI} then averages these temporal correlation coefficients across frequency, an operation which implies independent frequency band contributions to intelligibility, and which is not in line with literature, e.g., [22]. \textit{ESTOI} shares the first step with \textit{STOI}: mean- and variance-normalization is applied to subband envelopes. However, rather than computing the average inner products between these normalized envelopes, and relying on the additive-intelligibility-across-frequency assumption as is done in \textit{STOI}, \textit{ESTOI} instead computes \textit{spectral} correlation coefficients, which are finally averaged across time within the 384 ms analysis segment. This allows \textit{ESTOI} to better capture the effect of time-modulated noise maskers, where spectral correlation ‘in the dips’ is often preserved better capture the effect of time-modulated noise maskers, and in linking its performance to perceptual studies.

To model crudely the signal transduction in the cochlear inner hair cells, a one-third octave band analysis is approximated by summing STFT coefficient energies,

\[
S_j(m) = \left\| \sum_{k \in CB_j} |S(k,m)|^2 \right\|,
\]

where \(j\) is the one-third octave band index, \(CB_j\) denotes the index set of STFT coefficients related to the \(j\)th one-third octave frequency band, and \(J\) denotes the number of subbands.

Let us collect spectral values \(S_j(m)\) for each frequency band \(j = 1, \ldots, J\), and across a \textit{time segment} of \(N\) spectral samples, and arrange these in a short-time spectrogram matrix

\[
S_m = \begin{bmatrix}
S_1(m - N + 1) & \cdots & S_1(m) \\
\vdots & & \vdots \\
S_J(m - N + 1) & \cdots & S_J(m)
\end{bmatrix}.
\]

Hence, the \(j\)th row of \(S_m\) represents the temporal envelope of the signal in subband \(j\). Typical parameter choices are \(J = 15\) and \(N = 30\) (corresponding to 384 ms) [14]. The noisy/processed short-time spectrogram matrix \(X_m\) is defined analogously.

\textit{ESTOI} operates on mean- and variance- normalized rows and columns of \(S_m\) and \(X_m\) as follows. Let

\[
\bar{s}_{j,m} = [S_j(m - N + 1) S_j(m - N + 2) \cdots S_j(m)]^T
\]

denote the \(j\)th row of the spectrogram matrix \(S_m\). The \(j\)th mean- and variance- normalized row of \(S_m\) is given by

\[
\bar{s}_{j,m} = \frac{1}{\|s_{j,m} - \mu s_{j,m}\|} (s_{j,m} - \mu s_{j,m})_1,
\]

where \(\|y\| = \sqrt{y^Ty}\) is the vector 2-norm, \(\bot\) is an all-one vector, and \(\mu s_{j,m}\) is the sample mean given by

\[
\mu s_{j,m} = \frac{1}{N} \sum_{m'=0}^{N-1} S_j(m - m').
\]

\footnote{We ignore the clipping procedure used in \textit{STOI} in this description.}
Figure 2. The proposed intelligibility predictor, ESTOI, is a function of the noisy/processed signal \( x(n) \) and clean speech signal \( s(n) \). First, the signals are passed through a one-third octave filter bank, and the temporal envelopes of each subband signal are extracted. The resulting clean and noisy/processed short-time envelope spectrograms are time- and frequency-normalized before the "distance" between them is computed, resulting in intermediate, short-time intelligibility indices \( d_m \). Finally, the intermediate indices are averaged to form the final intelligibility index \( d \). More details are given in Sec. II. Signal examples of the various stages are shown in Fig. 3.

Note that the sample mean and variance of the elements in vector \( \hat{s}_{j,m} \) is zero and one, respectively. The mean and variance-normalized rows \( \hat{x}_{j,m} \) of the noisy/processed signal are defined similarly.

As mentioned, this row-normalization procedure is similar to the one used in STOI. Specifically, STOI uses an intermediate temporal correlation coefficient for the \( j \)th subband in the \( m \)th time segment, which can be expressed as the inner product of normalized vectors,

\[
\hat{s}_{j,m}^T \hat{x}_{j,m}.
\]

However, as mentioned, in ESTOI we do not use Eq. (3) directly, but introduce a spectral normalization as follows. Let us first define the row-normalized spectrogram matrix

\[
\hat{S}_m = \begin{bmatrix}
\hat{s}_{1,m} \\
\vdots \\
\hat{s}_{N,m}
\end{bmatrix}.
\]

Then, let \( \hat{s}_{n,m} \) denote the mean- and variance-normalized \( n \)th column, \( n = 1, \ldots, N \) of matrix \( \hat{S}_m \), where the normalization is carried out analogously to Eqs. (1) and (2). We finally define the row- and column-normalized matrix \( \hat{S}_{m} \) as

\[
\hat{S}_m = [\hat{s}_{1,m}, \ldots, \hat{s}_{N,m}].
\]

Hence, the columns of \( \hat{S}_m \) represent unit-norm, zero-mean normalized spectra (which themselves are computed from normalized temporal envelopes). The row- and column-normalized matrix \( \hat{X}_m \) of the noisy/processed signal in time segment \( m \) is defined in a similar manner. Fig. 3 demonstrates the effect of the various normalizations on example clean and noisy spectrograms.

**B. Intelligibility index**

The row- and column-normalized matrices \( \hat{S}_m \) and \( \hat{X}_m \) serve as the basis for the proposed intelligibility predictor. In particular, we define an intermediate intelligibility index, related to time segment \( m \), simply as

\[
d_m = \frac{1}{N} \sum_{n=1}^{N} \hat{s}_{n,m}^T \hat{x}_{n,m}.
\]

Since \( \hat{s}_{n,m} \) and \( \hat{x}_{n,m}, n = 1, \ldots, N \) are unit-norm vectors, each term in the sum may be recognized as the (signed) length of the orthogonal projection of the noisy/processed vector \( \hat{x}_{n,m} \) onto the clean vector \( \hat{s}_{n,m} \) or vice versa. It follows that \(-1 \leq \hat{s}_{n,m}^T \hat{x}_{n,m} \leq 1\). Similarly, \( d_m \) may be interpreted as the (signed) length of these projections, averaged across time within a time segment. In low-noise situations where \( \hat{x}_{n,m} \approx \hat{s}_{n,m} \), then \( d_m \) will be close to its maximum average projection length of 1, whereas if the elements of \( \hat{x}_{n,m} \) and \( \hat{s}_{n,m} \) are uncorrelated, then \( d_m \approx 0 \), i.e., the vectors are approximately orthogonal. Also, from the definitions of \( \hat{s}_{n,m} \) and \( \hat{x}_{n,m} \), \( d_m \) may be interpreted simply as sample correlation coefficients of the columns of \( \hat{S}_m \) and \( \hat{X}_m \) (i.e., spectra which have been normalized according to their subband envelopes), averaged across the \( N \) frames within a segment.

For simplicity, the intelligibility index related to the entire noisy/processed signal of interest is then defined as the temporal average of the intermediate intelligibility indices,

\[
d = \frac{1}{M} \sum_{m=1}^{M} d_m,
\]

where \( M \) is the number of time segments in the signal of interest. Since \(-1 \leq d_m \leq 1\), it follows that \(-1 \leq d \leq 1\).

**C. Implementation**

ESTOI operates at a sampling frequency of 10 kHz to ensure that the frequency region relevant for speech intelligibility is covered [2]; all signals are resampled to this frequency before applying the method. Then, signals are divided into frames of 256 samples, using a frame shift of \( D = 128 \), the frames are windowed with a Hann window, and an FFT of order \( N' = 512 \) is applied. Before computing the intelligibility index, frames with no speech content are discarded. These are identified as the frames of the reference speech signal \( s(n) \) with energy less than 40 dB than the signal frame with maximum energy. DFT coefficients of speech active frames are grouped into \( J = 15 \) one-third octave bands, with center frequencies of 150 Hz and approximately 4.3 kHz, for the lowest and highest band, respectively. Finally, time segments of length \( N = 30 \) (corresponding to 384 ms) are used (for further details on this choice, we refer to Sec. IV-C).
In this section we present interpretations of Eqs. (4) and (5) which provide insights into ESTOI. Specifically, we show that ESTOI can be interpreted in terms of a decomposition of (row- and column- normalized) noisy/processed short-time spectrograms into orthogonal one-dimensional subspaces. The decomposition assigns an intelligibility score to each such subspace, so that the sum of all the subspace intelligibilities equals the total intermediate intelligibility $d_m$ of the noisy/processed short-time spectrogram. The decomposition therefore allows us to rank each subspace according to their (predicted) contribution to intelligibility, revealing which spectro-temporal features are predicted to be important to intelligibility.

A. Preliminaries

To focus our exposition, we re-write the expression for $d_m$ (Eq. (4)) using the columns of the row- and column-normalized matrices $\hat{X}_m$ and $\hat{S}_m$. Let us concatenate the $N$ columns of $\hat{S}_m$ into a supervector

$$\hat{s}_m = [\hat{s}_{1,m}^T \ldots \hat{s}_{N,m}^T]^T,$$

where $\hat{s}_m \in \mathbb{R}^{NJ \times 1}$. A similar definition holds for the noisy/processed supervector $\hat{x}_m$. Furthermore, let us collect supervectors for each segment $m$ as columns in super matrices. The clean speech super matrix $\hat{S} \in \mathbb{R}^{NJ \times M}$ is given by

$$\hat{S} = [\hat{s}_1, \ldots, \hat{s}_M].$$

The noisy/processed matrix $\hat{X}$ is defined similarly.

The intermediate intelligibility index $d_m$ (Eq. (4)) may then be written as

$$d_m = \frac{1}{N} \hat{X}_m^T \hat{X}_m,$$

and inserting this into Eq. (5) leads to

$$d = \frac{1}{MN} \text{Tr} (\hat{S}^T \hat{X}),$$

where $\text{Tr}(\cdot)$ denotes the matrix trace operator.

Let us introduce the following orthogonal decomposition of the noisy/processed supervectors $\hat{x}_m$,

$$\hat{x}_m = \sum_{l=1}^{NJ} e_l e_l^T \hat{x}_m,$$

where $e_l e_j = \delta(i,j)$ are orthonormal vectors, and $P_l = e_l e_l^T$ is an orthogonal projection matrix onto the one-dimensional subspace spanned by $e_l$. Since each noisy/processed supervector $\hat{x}_m$ describes a time-frequency region of $N \times J$ one-third octave spectral values, Eq. (8) provides - when the basis vectors $e_l$ are specified - a decomposition of a noisy/processed time-frequency region into mutually orthogonal one-dimensional subspaces.

B. Intelligibility Subspace Decomposition

Our goal is to determine the orthogonal basis vectors $e_l$ in Eq. (8), ordered according to their (estimated) impact on intelligibility: first, we find the basis vector $e_1$, which carries most intelligibility on average across noisy/processed spectrograms. Next, we find the basis vector $e_2$, orthogonal to $e_1$, which carries most intelligibility. This procedure is repeated for the remaining dimensions, leading to an orthogonal subspace decomposition in terms of intelligibility.

To do this, insert Eq. (8) into Eq. (7),

$$d = \frac{1}{NM} \text{Tr} (\hat{S}^T \hat{X})$$

$$= \frac{1}{2NM} \text{Tr} (\hat{S}^T \hat{X} + \hat{X}^T \hat{S})$$

$$= \frac{1}{2NM} \text{Tr} \left( \hat{S}^T \sum_i P_i \hat{X} + \hat{X}^T (\sum_i P_i)^T S \right)$$

$$= \frac{1}{2NM} \text{Tr} \left( \hat{X} \hat{S}^T + \hat{S} \hat{X}^T \sum_i P_i \right)$$

$$= \frac{1}{2NM} \sum_{i=1}^{NJ} (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T) e_i,$$

where we used that $\text{Tr} A = \text{Tr} A^T$, that summation and trace are linear operators whose order can be interchanged, that
\[ \sum_i P_i = (\sum_i P_i)^T \] is a symmetric matrix by definition, and that \( \text{Tr} \ ABC = \text{Tr} \ CAB = \text{Tr} \ BCA \).

We can now perform the orthogonal intelligibility subspace decomposition described above by solving the following sequence of problems,

**Step 1:**

\[
\max_{e_1} \frac{1}{2NM} e_1^T (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T) e_1, \text{ such that } e_1^T e_1 = 1.
\]

Steps \((l = 2, \ldots, NJ)\):

\[
\max_{e_l} \frac{1}{2NM} e_l^T (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T) e_{l} \text{ such that } e_l^T e_l = 1,
\]

\[
\quad \quad \text{and } e_l \perp e_1, \ldots, e_{l-1}.
\]

It may be recognized that the solution vectors \(e_l\) are the eigenvectors of the symmetric matrix \(\frac{1}{2NM} (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T)\). The symmetry of this matrix ensures that a) the eigenvectors are mutually orthogonal, and b) the eigenvalues are real-valued, which allows a simple ranking of subspaces according to their contribution to intelligibility.

Note that inserting Eq. (8) with the found vectors \(e_l\) in Eq. (6) allows us to express the intermediate intelligibility index \(d_m\) in terms of a sum of orthogonal intelligibility subspaces. Note also that since the \(l\)th eigenvalue \(\lambda_l\) of the matrix \(\frac{1}{2NM} (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T)\) satisfies

\[
\frac{1}{2NM} (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T) e_l = \lambda_l e_l,
\]

then it follows that

\[
\lambda_l = \frac{1}{2NM} e_l^T (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T) e_l.
\]

Comparison to Eq. (9) shows that

\[
d = \sum_{i=1}^{NJ} \lambda_i.
\]

In other words, the total estimated intelligibility of the signal in question is completely determined by the eigenvalues of the sample cross-correlation matrix \(\frac{1}{2NM} (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T)\). Specifically, the contribution to intelligibility by the \(l\)th subspace is given by the corresponding eigenvalue \(\lambda_l\), and the sum of all eigenvalues equals the total intelligibility index.

**C. Intelligibility Subspaces - Example**

To demonstrate the intelligibility subspace decomposition we construct noisy (but in this example unprocessed) speech signals by adding noise to clean speech signals. Specifically, we study the impact of adding a 100% intensity-modulated, lowpass filtered noise sequence to 1680 signals from the TIMIT [25] database. The noise is a Gaussian white noise sequence (sampling rate of \(f_s = 16000\) Hz), filtered through a first-order IIR low-pass filter with a 3dB cut-off frequency at approximately 80 Hz (pole location at \(p = 0.97\)). Then, this lowpass filtered noise is amplitude modulated by the sequence

\[
a(n) = 1 + \sin(2\pi f_{\text{mod}} / f_s n + \phi), \quad n = 0, \ldots, N_s,
\]

with a modulation frequency of \(f_{\text{mod}} = 5\) Hz. \(N_s\) is the sequence length corresponding to the duration of the speech signal in question, and \(\phi\) is a uniformly distributed random phase value, drawn independently for each sentence. The noise is scaled to form an SNR of -10 dB for each sentence.

Based on this set of noisy signals, we apply the intelligibility subspace decomposition described above. Fig. 4 (lower-right) shows the \(NJ = 450\) eigenvalues \(\lambda_i\) in descending order for the decomposition of \(d\). In this example, the 11 dominant subspaces carry 46% of the total estimated intelligibility, while 115 dimensions carry 90%. Fig. 4 shows the basis vectors of the 11 dominant subspaces. The subspaces are characterized by regular spectro-temporal patterns, apparently with low spectro-temporal modulation frequencies.

Frequency analysis across the temporal dimension of each of the subfigures reveal temporal modulation frequencies - averaged across the acoustic frequency axis - ranging from 2.1 Hz to 5.8 Hz. This modulation frequency range is well-known to be particularly important for intelligibility. Specifically, Drullman et al. showed that intelligibility can be degraded significantly, if modulations in the frequency range of approximately 3-8 Hz are not preserved [26], [27]. Similarly, Elliot and Theunissen found temporal modulation frequencies in the range 1-7 Hz to be most important for speech intelligibility [24], while Kates and Arehart found frequencies less than 12.5 Hz to carry most information about intelligibility [28].

Applying Fourier transforms to the columns of each subspace spectrogram in Fig. 4 and computing the average magnitude spectrum shows maximum spectral modulations in the range 0.2-0.7 cycles/kHz. As for the temporal modulation content, these numbers are quantitatively well in line with the results in [24] who report spectral modulation frequencies < 1 cycle/kHz to be most important for speech intelligibility.

While the eigenvalues \(\lambda_i\) of the sample correlation matrix \(\frac{1}{2NM} (\hat{X} \hat{S}^T + \hat{S} \hat{X}^T)\) tend to be positive, a small subset of the lowest eigenvalues can be negative. In other words, the signal components represented by the corresponding subspaces degrade intelligibility as estimated by the model. For many practical situations, however, the impact of these negative intelligibility subspaces is small. For example, in Fig. 4, where the global SNR is \(-10\) dB, the 15 smallest eigenvalues are negative - their sum is approximately \(-0.001\), which is 0.3% of the total estimated intelligibility index. Generally speaking, the number and the impact of these negative intelligibility subspaces increases with decreasing SNR. For simple additive and stationary noise maskers, e.g., a single constant-frequency masker tone which occupy the same time-frequency region for all time segments, it can be verified that the time-frequency pattern of this masker may be represented well using the negative subspaces as basis functions. For non-stationary noise sources, on the other hand, e.g., the modulated low-pass noise used in Fig. 4, the negative subspaces do, generally, not represent the spectro-temporal noise pattern in a particular segment. Rather, the negative subspaces represent the average spectro-

\[\text{Note that an accurate comparison is difficult: the results in [24] are based on spectro-temporal analyses of log-magnitude spectra computed in a uniform-frequency filter bank, whereas the proposed method operates on linear, but energy-normalized one-third octave band magnitude spectra.}\]
temporal pattern of the noise within many time segments, across which the noise does not necessarily occupy the same time-frequency region in each time segment.

Finally, Fig. 5 shows the intelligibility subspace decomposition for speech in natural noise, namely a noise recorded in a busy office cafeteria [29]. While details obviously differ from the decomposition in Fig. 4, the main features of the dominant subspaces are the same: temporal modulation frequencies are in the range 2.0–5.7 Hz, while spectral modulations are in the range 0.2–0.8 cycles/kHz.

IV. SIMULATION RESULTS

In this section, we present a number of intelligibility listening tests for evaluating the proposed method. Furthermore, we study the performance as a function of the segment length \( N \) and the test signal duration. Finally, we compare the performance of ESTOI to a range of existing speech intelligibility predictors.

A. Signals and Processing Conditions

We study the performance of ESTOI using the results of five intelligibility tests with speech signals subjected to various noise sources and processing conditions. The first two tests used various additive noise sources with strong temporal modulations; we include these in the study to verify the ability of ESTOI to operate in this domain, and to verify the results reported in [21] that established intelligibility predictors work less well here. The third test used stationary and non-stationary additive noise sources, with less temporal modulations; quite some existing methods work well for this common class of noise sources, and it is important to establish that ESTOI does so too.

The fourth and fifth intelligibility test used processed noisy speech signals for which STOI works exceptionally well, while many other methods fail. As before, it is important to establish the performance of ESTOI in this situation.

1) Additive Noise Set I: The first set of signals consist of ten mainly non-stationary noise sources with significant modulation content, cf. Table I. The \( \text{icra1} \) signals are synthetic speech signals constructed by filtering Gaussian noise sequences through bandpass filters with time-varying gain to construct signals with speech-like spectro-temporal properties [30]. The \( \text{snam} \) signals are 100% sinusoidally modulated speech-shaped noise signals. The signals are constructed by point-wise multiplication of the unmodulated speech-shaped noise signal \( \text{icra1} \) with the modulation sequence

\[
a(n) = 1 + \sin(2\pi \omega n + \phi),
\]

where \( \omega \) denotes the angular modulation frequency, and \( \phi \in [-\pi; \pi] \) is a random phase-value, drawn independently for each signal generation. To construct machine gun noise \( \text{macgun} \) with sufficient masking power, the original machine gun noise signal from the Noisex database [31] was divided into successive 20 ms frames, and frames with energy less than 40 dB of the maximum frame energy were removed.

Speech signals from the Dantale II sentence test [32] were added to randomly selected sections of each of the ten noise sources at six different SNRs (cf. Table I). The SNRs were chosen so that, for each noise source, some noisy signals were almost perfectly intelligible, whereas others were essentially unintelligible. The total number of conditions was therefore 10 noise types x 6 SNRs = 60 conditions. Each condition was repeated 3 times (with different speech and noise realizations) leading to a total of 180 sentences to be judged per subject. The presentation order of noise types, SNRs, and repetitions was randomized. The sample rate was 20 kHz.

We conducted a closed Danish speech-in-noise intelligibility test, cf. [33]. The Dantale II sentences consist of five words with a correct grammatical structure. Candidate words were arranged in an 10-by-5 matrix on a computer screen, such that each of the five columns encompassed exactly the 10 possible alternatives for the corresponding word. Each column was extended with one entry, which allowed the subject to answer “Don’t know”. For each 5-word sentence, the subject must select via a graphical user interface the words that she heard. Subjects were seated in a sound treated room, where signals were presented diotically through headphones (Sennheiser HD 280 Pro). The \( \text{icra1} \) noise at the SNR of -8 dB was used to calibrate the presentation level to 65 dB (A). The subjects were allowed to adjust this level during a training session prior to the actual test. Twelve native Danish speaking subjects (normal-hearing, age range 26–44 years, 2 females, 10 males) participated in the test. The subjects volunteered for the experiments and were not paid for their participation.

2) Additive Noise Set II: The second data set, consisting of speech in additive fluctuating noise sources, is described in [7]. We use Maskers 1–13 [7], which include low-pass filtered unmodulated Gaussian noise, and various amplitude modulated Gaussian noise signals, including sinusoidally amplitude-modulated signals (modulation frequencies: 2.1, 4.9, 10.2, 19.9 Hz, and modulation depths of ±6 dB, ±12 dB, and ±100%), and three irregularly modulated noise signals found by adding the sinus-modulators with random initial phases. The speech material used was the Swedish version of the Hagerman material [34], which is similar in structure to the Dantale 2 set used above. For each noise source, noisy speech signals were generated with an SNR of -15 dB, and the corresponding speech intelligibility was recorded ([7, Fig. 5]). Hence, the number of conditions equalled 11 noise sources x 1 SNR = 11 conditions. Intelligibility tests were conducted with i) eleven young (17–33 years), normal hearing listeners, and ii) twenty elderly (54–69 years), normal hearing listeners (the study also included elderly, hearing-impaired listeners, but results of these tests are not used in this paper). The sample rate used was 20 kHz. For more details, we refer to [7].

3) Additive Noise Set III: We include a third additive noise set, for which many existing intelligibility predictors work well, see e.g. [14], [17] and the references therein. The data set encompasses Dantale 2 speech sentences contaminated by four additive noise sources: i) speech-shaped Gaussian noise, ii) car cabin noise recorded when driving on the highway, iii) bottling hall noise, and iv) cafeteria noise consisting of a conversation between a female and a male speaker, i.e., two-talker speech babble [35]. The noisy signals were generated with SNRs from -20 dB to 5 dB in steps of 2.5 dB, so the total number of conditions equals 4 noise types x 11 SNRs.
Intelligibility Subspace 1
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 2
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 3
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 4
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 5
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 6
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 7
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 8
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 9
Freq. [Hz]
0 100 200 300
300 756 1905
Intelligibility Subspace 10
Freq. [Hz]
0 100 200 300
300 756 1905

Figure 4. Decomposition of $d$ for speech in additive, speech-shaped, sinusoidally amplitude-modulated Gaussian noise ($f_{\text{mod}} = 5$ Hz, SNR = -10 dB). Basis functions $e_i$ of the 11 dominant intelligibility subspaces and decomposition of $d$ in terms of eigenvalues $\lambda_i$ (lower right).

<table>
<thead>
<tr>
<th>Noise Name</th>
<th>Description</th>
<th>SNR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>icra1</td>
<td>Unmodulated speech-shaped (male) Gaussian noise from the ICRA corpus (Track 1) [30].</td>
<td>-17:3:-2.</td>
</tr>
<tr>
<td>icra4</td>
<td>1-person babble (female) from the ICRA corpus (Track 4).</td>
<td>-29:3:-14.</td>
</tr>
<tr>
<td>icra6</td>
<td>2-persons babble (1 male and 1 female) from the ICRA corpus (Track 6)</td>
<td>-24:3:-9.</td>
</tr>
<tr>
<td>icra7</td>
<td>6-persons babble from the ICRA corpus (Track 7)</td>
<td>-19:3:-4.</td>
</tr>
<tr>
<td>snam2</td>
<td>100% intensity-modulated versions of icra1. Modulation frequency 2 Hz.</td>
<td>-27:3:-12.</td>
</tr>
<tr>
<td>snam4</td>
<td>As above with modulation frequency 4 Hz.</td>
<td>-23:3:-8.</td>
</tr>
<tr>
<td>snam8</td>
<td>As above with modulation frequency 8 Hz.</td>
<td>-25:3:-10.</td>
</tr>
<tr>
<td>snam16</td>
<td>As above with modulation frequency 16 Hz.</td>
<td>-22:3:-7.</td>
</tr>
<tr>
<td>macgun</td>
<td>(Modified) machine gun noise from the Noisex corpus [31].</td>
<td>-37:3:-22.</td>
</tr>
<tr>
<td>destop</td>
<td>Destroyers operation room noise from the Noisex corpus [31].</td>
<td>-14:3:1.</td>
</tr>
</tbody>
</table>

Table 1

Noise sources and SNR ranges used for intelligibility test with Additive Noise Set I. Notation $x:y:z$ indicates SNRs from $x$ to $z$ (both included) in steps of $y$ dB.

4) Ideal Time-Frequency Segregation: The fourth data set consists of the noisy signals from Additive Noise Set III, processed using the ideal time-frequency segregation (ITFS) technique [36]. Kjems [35] processed noisy signals with two different ITFS algorithms called ideal binary mask (IBM) and target binary mask (TBM), and used eight different variants of each algorithm (reflected by the $LC$ parameter, i.e., the threshold for which the algorithm suppresses a given time-frequency tile or not). Three different SNRs were used, leading to a total number of (4 noise types (IBM) + 3 noise types (TBM)) x 8 LC x 3 SNRs = 168 test conditions. Fifteen normal hearing subjects participated in the test. The sample rate was 20 kHz. More details are available in [35].

5) Single-Channel Noise Reduction: The last data set consists of noisy speech signals processed with three single-microphone noise reduction algorithms [37]. The three algorithms are all non-linear and aim at finding binary or soft minimum mean-square error (MMSE) estimates of the short-time spectral amplitude (STSA). We include this data set,

---

4The IBM and TBM algorithms are identical for speech-shaped noise.
because an obvious use of the proposed algorithm is for development/tuning of noise reduction algorithms. Speech-shaped (unmodulated) noise signals were added to speech signals (female speaker) from the Dutch version of the Hagerman test [33], [34] at fixed SNRs of -8, -6, -4, -2, and 0 dB. The noisy and processed speech signals were presented diotically via headphones, and the order of presenting the different algorithms and SNRs was randomized. The signals were evaluated in a closed Dutch speech-in-noise intelligibility test [33]. Each processing condition was repeated five times, leading to 4 conditions x 5 SNRs = 20 test conditions. Thirteen subjects participated in the test. The sample rate was 8 kHz.

B. Prediction of Absolute Intelligibility and Figures of Merit

Most intelligibility prediction methods (including ESTOI) do not predict intelligibility, i.e., the fraction of words understood, per se. Instead, they output a scalar \( \tilde{I} \) which, ideally, is monotonically related to absolute intelligibility \( I \). The monotonic mapping between predictor output and absolute intelligibility is generally hard to derive analytically, but in [14], [38] it was proposed to use the following logistic map,

\[
\tilde{I} = \frac{100}{1 + \exp(a\tilde{I} + b)},
\]

where \( a, b \in \mathbb{R} \) are constants that depend on the test material, test paradigm, etc., and which are estimated to fit the intelligibility data at hand.

To quantify the performance of intelligibility predictors, we use four figures of merit (see [17] for exact definitions): i) the linear correlation coefficient \( \rho_{\text{pre}} \) between average intelligibility scores obtained in listening tests, and the outcomes \( \tilde{I} \) of the intelligibility predictors before applying the logistic map (Eq. (10)), ii) the linear correlation coefficient \( \rho \) between average intelligibility scores and the outcomes \( \hat{I} \) of the intelligibility predictors, i.e., after the logistic map, iii) the root mean-square prediction error \( \sigma \) between measured and predicted intelligibility, and iv) Kendalls rank correlation coefficient (\( \tau \)).

C. Impact of Segment Length and Signal Duration

1) Sensitivity to segment length \( N \): ESTOI was developed with simplicity in mind and has few free parameters. This
section studies the performance of ESTOI as a function of the segment length $N$ for the various noise/processing situations in the five data sets described above. In particular, for a given data set, and for a given choice of the segment length $N$, the proposed method was applied to compute an intelligibility index for each test condition in that data set. Then, the free parameters $a$, $b$, of the logistic function, Eq. (10), were fitted to map the predicted intelligibility indices to absolute intelligibility as measured in a listening test. Finally, the performance in terms of $\rho$, $\sigma$, and $\tau$ was computed.

Fig. 6 shows the performance in terms of $\rho$ as a function of segment length $N$ (in the range $N = 5, \ldots, 60$ corresponding to durations of 64–768 ms). Clearly, the proposed method is fairly insensitive to the exact choice of the segment length $N$. In fact, for $20 \leq N \leq 50$ (corresponding to durations of 256–640 ms), the proposed method gives excellent performance with values of $\rho > 0.9$. The lower performance for $N < 20$ may be explained by the fact that with these short segment lengths, the method is less able to capture low-frequency temporal modulations, which are important for speech intelligibility [14]. While the discussion above focused on prediction performance in terms of $\rho$, similar conclusions may be drawn from performance analysis based on $\sigma$ and $\tau$ (not shown). Based on these observations, a value of $N = 30$ (384 ms) is used in the remainder of the simulation experiments.

![Figure 6. Speech intelligibility prediction performance (in terms of $\rho$) as a function of segment length $N$ for various noise/processing conditions.](image)

**2) Sensitivity to duration of test signals:** For highly modulated, additive noise sources, the instantaneous SNR can vary significantly across a short time span. For example, a sinusoidally amplitude-modulated noise source could completely mask the target signal at one instant, while leaving it essentially un-masked half a period later (cf. Fig. 3). Hence, the speech intelligibility for a particular short speech signal is highly dependent on the (random) location of the high SNR region with respect to the speech signal. Since our goal is to estimate the average speech intelligibility, we would therefore expect it to be necessary to average across many noise realizations, or, equivalently, to use longer test speech signals, than would e.g. be necessary for unmodulated noise sources. In this section we therefore study the sensitivity of the proposed method with respect to test signal duration $t_{\text{sig}}$.

To do so, we generated speech signals contaminated by two additive noise sources: speech-shaped stationary noise, and synthetic 1-person babble (the icr4d noise source from [30]). The noise sources were scaled to achieve SNRs of -10 dB and -23 dB, respectively, corresponding approximately to the 50% speech reception threshold (SRT) (the SNR needed to achieve a recognition rate of 50%) for these noise sources. Noise and clean signals were generated in corresponding pairs with various durations in the range $1 \sim 80$ secs. Noisy signals were generated by adding the clean and noise signals. Clean and noisy signals were then passed through ESTOI. For comparison, the clean and noise signals were passed through the Extended Speech Intelligibility Index (ESII) algorithm [3] (see Sec. IV-D for implementational details). For each signal duration, $n_{\text{real}} = 100$ different realizations of the clean/noisy signal pairs were evaluated.

It is of interest to study to which extent an intelligibility prediction $\hat{I}_n(t_{\text{sig}})$ based on a single (the $n$th) test signal realization of duration $t_{\text{sig}}$ lies in the neighborhood of the ensemble-average, i.e., the average predictor value $\mu_{\hat{I}(t_{\text{sig}})} = \frac{1}{n_{\text{real}}} \sum_{n=1}^{n_{\text{real}}} \hat{I}_n(t_{\text{sig}})$ across many realizations $n_{\text{real}}$ of test signal pairs. To do so, let us define the sample standard deviation

$$
\sigma_{\hat{I}(t_{\text{sig}})} = \sqrt{\frac{1}{n_{\text{real}}} \sum_{n=1}^{n_{\text{real}}} (\hat{I}_n(t_{\text{sig}}) - \mu_{\hat{I}(t_{\text{sig}})})^2},
$$

and let us define the relative standard deviation as

$$
e(t_{\text{sig}}) = \sigma_{\hat{I}(t_{\text{sig}})}/\mu_{\hat{I}(t_{\text{sig}})} \times 100 \text{ [%].}
$$

Figs. 7a) shows $e(t_{\text{sig}})$ for ESII and ESTOI for unmodulated speech-shaped noise, while Fig. 7b) shows the results for babble noise. From Fig. 7 three conclusions can be drawn. First, as expected, the relative standard deviation declines with signal duration. Secondly, for a given test signal duration, ESII has lower relative standard deviation than ESTOI. This is because ESII makes explicit use of its access to the clean speech signal and the noise signal in separation to accurately compute SNRs in different time-frequency regions, and subsequently compute an intelligibility index based on these (i.e., ESII is a Class-1 method as discussed in the Introduction). ESTOI, on the other hand, does not make use of the access to the separated clean and noise components and is therefore more generally applicable, e.g., to non-linearly processed signals (i.e., it is a Class-2 method). This generality comes with the price of an increase in the estimation standard deviation. Thirdly, as expected, for a given test signal duration, the estimation standard deviation is higher for modulated than for non-modulated noise, both for ESII and ESTOI.

It is hard to decide a priori on a sufficient test signal duration, because a) this depends on the noise signal statistics in a non-trivial manner, and b) the noise statistics are unavailable to the proposed method. Hence, the test signal duration should simply be chosen as long as practically possible, and generally no less than some tens of seconds. Note that long test signals...
can be generated by concatenating several of the, potentially short, speech sentences used in the intelligibility test.

![Graph](image)

Figure 7. Relative standard deviation \(\varepsilon(t_{\text{sig}})\) of speech intelligibility predictors ESII and ESTOI, as a function of test signal duration \(t_{\text{sig}}\). a) Speech-shaped stationary noise (SNR = -10 dB), b) icna-l-noise (synthetic 1-person babble) [30] (SNR = -23 dB).

D. Comparison to Existing Methods

We compare the proposed intelligibility prediction method to reference methods from the literature. The methods are outlined in Table II. The methods CSII-BIF and STI-NCM-BIF are referred to as CSII\(_{\text{mid}}\), \(W_4, \rho = 1\) and NCM, \(W_4^{(3)}, p = 1.5\), respectively, in [39, Table IV]. We implemented the GLIMPSE method [8] using the one-third-octave filter bank used in ESTOI. The speech glimpse percentage was defined here as the percentage of time-frequency units with a local SNR exceeding -8 dB (this threshold was chosen because it lead to best performance in terms of \(\rho, \sigma, \) and \(\tau\)). Our implementation of the ESII algorithm computes the per-frame SII based on one-third octave filtering, and outputs the average of the per-frame SII. The implementation uses stationary speech-shaped Gaussian noise instead of undistorted real speech signals as input (as specified in [3], [6]), but excludes the upward-spread-of-masking functionality as defined in [3], because this appears to degrade performance. As in [6], we use the band importance functions derived for the test stimuli in the Speech in the Presence of Noise (SPIN) test ( [3, Table B.1]) and [40].

Tables III–VI summarize performance in terms of \(\rho_{\text{pre}}, \rho, \sigma, \) and \(\tau\), respectively, for the intelligibility predictors for the various additive noise and processing conditions. The \(\rho, \sigma, \) and \(\tau\), values are found by fitting \(a, b\) to the data sets in question. To identify statistically significant differences between \(\rho\)-values (Table IV), pairwise comparisons using the Williams \(t\)-test [41]–[43] were performed within each data set between the predictor with the largest \(\rho\) and the others (Bonferroni correction for multiple comparisons). Methods which do not perform statistically significantly worse than the method with highest \(\rho\) \((p < 0.05)\) are indicated with (*) in Table IV. In addition, a statistical analysis of significance was applied to the root mean-square prediction errors \(\sigma\) (Table V) as follows (see [44] for a brief outline of this approach). For each prediction method and for each of the five listening tests, the free parameters \(a, b\) in the logistic function were fitted to \(n - 1\) data points, where \(n\) denotes the number of conditions for a specific data set. Then this logistic function was applied to the left-out data point \(I_i\), where \(i\) is the index of the left-out data point, to find a prediction \(\hat{I}_i\) of the left-out subject result \(I_i\). The procedure was repeated for all data points, resulting in prediction errors \(e_i = I_i - \hat{I}_i, i = 1, \cdots, n\). Our goal is to compare the magnitude of these prediction errors across prediction methods. The data \(e_i\) for each intelligibility predictor and for each data set did not pass a chi-square goodness-of-fit test for normality \((p < 0.05)\). Hence, a Kruskal-Wallis test was performed, rejecting for each data set the hypothesis that the median of \(e_i\) is identical for all prediction methods \((p < 10^{-5})\). A multiple pairwise comparison test (Tukey HSD) was applied to identify prediction methods which, for a particular data set, performed statistically significantly worse than the method with lowest \(\sigma\) \((p < 0.05)\). The result of this comparison is indicated with (*) in Table V.

From these tables, a number of observations can be made. First, focusing on the highly non-stationary noise conditions, i.e., Additive Data Sets I and II, it is clear that ESTOI, GLIMPSE, and ESII appear to work quite well. The fact that GLIMPSE and ESII can work well in these conditions is well in line with results reported in [8] and [6], respectively. On the other hand, existing methods such as STOI and SIMI, which are known to work well for other less non-stationary noise sources and for various processing conditions, do not work well for the highly fluctuating noise sources: SIMI shows correlation values \(\rho \approx 0\), while STOI shows large but negative correlations for these data sets in Table III. This indicates that the STOI output \(\hat{I}\) decreases for increasing measured intelligibility (the fact that the same entry in Table IV is positive is because the logistic map from \(\hat{I}\) to \(I\) in this situation maps low \(\hat{I}\) values to high \(I\) values, and vice versa).

It is interesting to note that ESTOI (and ESII and GLIMPSE), perform well for Additive Noise II, both for young and for elderly normal-hearing subjects. While the basic intelligibility predictors are unchanged, each intelligibility predictor employs a different logistic map (i.e., constants \(a\) and \(b\) in Eq. (10)) for the different subject groups, because the 30% speech reception threshold was 6 dB higher for the elderly compared to the young subjects. It appears that the SRT differences between these normal-hearing subject groups (e.g., differences in higher auditory stages, which are not captured by a standard listening test used to establish whether a subject is normal hearing or not) are well-modeled simply by changing the logistic map.

Secondly, for the less fluctuating (but still non-stationary) noise sources in Additive Noise Set III, most methods work well. In fact, for this data set, several intelligibility predictors, including ESTOI, show values of \(\rho > 0.95\), and \(\sigma\) values at or below 10%. Note that SII, which relies on long-term noise spectra, also works well in this situation.

For noisy signals processed with ideal time-frequency segregation and single-channel noise reduction, ESTOI, SIMI, and STOI work well with \(\rho > 0.94\). It is interesting to note that for single-channel noise reduced signals, STI-NCM-BIF works exceptionally well \((\rho > 0.97)\); an explanation is that STI-NCM-BIF was developed with this particular processing type in mind; also note that STI-NCM-BIF does not show this level of performance for any other noise/processing condition.

In summary, for highly fluctuating, additive noise sources, where STOI and SIMI fail, ESTOI performs at the level of
established methods such as GLIMPSE and ESII, without requiring access to the speech and noise signals in isolation. For less fluctuating noise sources, ESTOI performs as well as the best existing methods, such as SII, STOI, and SIMI. For non-linearly processed noisy signals, where methods such as SII, ESII, and GLIMPSE cannot operate, the proposed method still performs at the level of STOI and SIMI.

V. CONCLUSION

We presented an algorithm for monaural, intrusive intelligibility prediction: given undistorted reference speech signals and their noisy, and potentially non-linearly processed, counterparts, the algorithm estimates the average intelligibility of the latter, across a group of normal-hearing listeners. The proposed algorithm, which is called ESTOI (Extended Short-Time Objective Intelligibility), may be interpreted in terms of an orthogonal decomposition of energy-normalized short-time spectrograms into "intelligibility subspaces", i.e., one-dimensional subspaces which are ranked according to their importance with respect to intelligibility. This intelligibility subspace decomposition indicates that the proposed algorithm favors spectro-temporal modulation patterns, which are known from literature to be important for intelligibility. The proposed intelligibility predictor has only one free parameter, the segment length \( N \), i.e., the duration across which the short-time spectrograms are computed. We show, via simulation experiments, that performance is fairly insensitive to the exact choice of this parameter, and that durations in the range 256–640 ms lead to best performance – this allows the algorithm to capture relatively low-frequency modulation content, while still being able to adapt to changing signal characteristics. We study the performance of ESTOI in predicting the results of five different intelligibility listening tests: two with temporally highly modulated additive noise sources, one with more moderately modulated, additive noise sources, and two with noisy signals processed by ideal time-frequency masking and single-channel non-linear noise reduction algorithms, respectively. Compared to a range of existing speech intelligibility prediction algorithms, ESTOI performs well across all listening tests.

The present study has focused on speech intelligibility prediction performance within data sets, that each contain signals with similar distortions or processing types (e.g., additive, modulated noise or noisy speech processed by ideal time-frequency segregation (ITFS) algorithms, etc.). It is a topic for future research to study the performance of the proposed intelligibility predictor across data sets with different distortion and processing types. Compared to the present study, this would require conduction of larger intelligibility tests, where these different distortion or processing types are included in the same listening test.

A Matlab implementation of the proposed algorithm is available for non-commercial use at http://kom.au.dk/~jjel/.

ACKNOWLEDGEMENT

The authors would like to thank four anonymous reviewers whose constructive comments helped improve the presentation of this work. Discussions with Dr. Gaston Hilkuysen, Dr. Thomas Ulrich Christiansen, and Asger Heidemann Andersen are greatly acknowledged. Finally, the authors wish to thank Dr. Jalal Taglia for making the Matlab code of his intelligibility predictor publicly available.

REFERENCES


Jesper Jensen Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Information Processing (SIP), Department of Electronic Systems, at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

Gees H. Taal Cornelis (Cees) H. Taal received the B.S. and M.A. degrees in arts and technology from the Utrecht School of Arts, Utrecht, The Netherlands, in 2004 and his M.Sc. and Ph.D. degree in computer science from the Delft University of Technology, Delft, The Netherlands, in 2007 and 2012, respectively. From 2012 to 2013 he held Postdoc positions at the Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden and the Leiden University Medical Center, Leiden, the Netherlands. From 2013 he worked in industry with Philips Research, Eindhoven, the Netherlands as a research scientist in the field of biomedical signal processing. Currently, he is at Quby, Amsterdam, the Netherlands performing R&D as a DSP expert applying signal processing to smart-home applications, e.g., thermostats and power monitoring.


| Table VI | PERFORMANCE OF INTELLIGIBILITY PREDICTORS IN TERMS OF $\hat{\tau}$ KENDALL'S RANK CORRELATION COEFFICIENT. |
| ESTOI | 0.748 | 0.590 | 0.667 | 0.816 | 0.775 | 0.842 |
| SIMI | 0.354 | 0.077 | 0.256 | 0.825 | 0.767 | 0.884 |
| STOI | 0.376 | 0.436 | 0.436 | 0.818 | 0.798 | 0.905 |
| CSII-MID | 0.609 | 0.026 | 0.026 | 0.863 | 0.335 | 0.600 |
| CSII-BIF | 0.580 | 0.410 | 0.487 | 0.846 | 0.408 | 0.684 |
| STI-NCM | 0.328 | 0.180 | 0.590 | 0.818 | 0.538 | 0.684 |
| STI-NCM-BIF | 0.276 | 0.085 | 0.077 | 0.643 | 0.385 | 0.853 |
| NSEC | 0.514 | 0.128 | 0.128 | 0.854 | 0.595 | 0.505 |
| MIKNN | 0.554 | 0.436 | 0.410 | 0.751 | 0.695 | 0.684 |
| GLIMPSE | 0.673 | 0.461 | 0.385 | 0.742 | – | – |
| SH | 0.260 | 0.330 | 0.128 | 0.822 | – | – |
| ESH | 0.653 | 0.564 | 0.359 | 0.683 | – | – |