A Differentiable Approximation to Speech Intelligibility Index with Applications to Listening Enhancement

Richard Stanton¹, Nikolay D. Gaubitch², Patrick Naylor³, and Mike Brookes¹
¹Imperial College London, London, United Kingdom
²Delft University of Technology, The Netherlands
Correspondence should be addressed to Richard Stanton (richard.stanton08@imperial.ac.uk)

ABSTRACT
The Speech Intelligibility Index (SII) [1] is a standardised objective measure for estimating the intelligibility of speech in noise. It is, however, difficult to use it in the iterative optimisation of speech enhancement algorithms because it is a discontinuous function of its input parameters. In this paper, we derive an approximation for the Speech Intelligibility Index that is both continuous and differentiable, which allows for more efficient optimisation procedures. The use of the approximation is demonstrated in an application to near-end speech enhancement.

1. INTRODUCTION
The Speech Intelligibility Index (SII) [1] is a standardized objective measure which is correlated with the intelligibility of speech under a variety of adverse listening conditions. It is an intrusive measure which is a function of the long term average spectra of the speech and noise signals. It is based on the Articulation Index defined by French and Steinberg [2] and on work from Kryter [3]. It can be used to predict the performance of speech enhancements systems that involve linear processing [4].

Acoustic forensics is frequently concerned with speech at low Signal-to-Noise Ratios (SNRs) that are at the margin of intelligibility. When used in court, speech can only be dependable when it is above a threshold of intelligibility [5, 6, 7]. Thus it would be useful to use the SII in a speech processing algorithm as a metric to maximise intelligibility.

Several speech processing techniques, for example near-end listening enhancement, use optimisation routines that seek to maximise the SII. Near-end listening enhancement focuses on the presentation of far-end speech in near-end noise that cannot be controlled. The far-end speech is linearly filtered in order to maximise the intelligibility, [8, 9, 10, 4, 11, 12].

In its present state, the SII has no continuous derivative with respect to the speech power spectrum. Because of this, approximations are commonly made in order for the SII to be used in optimisation. In the contribution from Taal et al. [13], a convex approximation to the SII was made, which gives a single stable optimisation point. It was an extension of the work by Sauert et al. [14] in which an alternative approximation was used for the SII.

In this paper, we derive a formula that closely approximates the SII. The proposed formula has several advantages. It is both continuous and differentiable with respect to the speech spectral levels, which can lead to more efficient optimisation. We demonstrate the use of this approximation in the application of near-end speech enhancement as an extension to [13], in which, we see further positive gains in SII over [13].

The remainder of the paper is organised as follows. The SII is summarised in Section 2. The formula that closely approximates the SII calculation based on the ANSI standard is derived in Section 3. Simulation results are provided to confirm the correspondence between the two methods of calculation. Subsequently, this formula is used in Section 4 to set up a cost function for the calculation of the equaliser settings for a near-end speech enhancement system and simulation results are provided to demonstrate its performance. Conclusions form this work are drawn in Section 5.

2. DESCRIPTION OF SII
The SII, \( \Psi \), is defined in [1]. It is a function of the average speech spectrum power levels, \( E_i \), and the average noise spectrum power levels, \( N_i \), both given in dB, where \( i \) represents the \( i \)th subband. The SII is computed as the weighted sum over \( l_{\text{max}} \) subbands of the band audibility function, \( A_i \):

\[
\Psi(E, N) = \sum_{i=1}^{l_{\text{max}}} I_i A_i
\]

\[
E = [E_1 \ldots E_{l_{\text{max}}}] \nonumber
\]

\[
N = [N_1 \ldots N_{l_{\text{max}}}] \nonumber
\]

where \( I_i \) is the predefined band importance function, which represents the relative contribution of each band to the overall intelligibility. Typically, third-octave or critical bands are used [15].

The band audibility function is a measure of the effective dynamic range in each subband that contributes to the intelligibility of the speech. It can be described as a function of the level distortion factor, \( L_i \), and \( K_i \), a parameter that models masking effects and that varies from 0 to 1:

\[
A_i = L_i K_i. \tag{1}
\]
The level distortion factor considers the decrease in intelligibility when speech is presented at high levels, thus causing distortions. It is given by:

\[ L_i = \min \left\{ 1 - \frac{E_i - U_i - 10}{160}, 1 \right\}, \quad (2) \]

where \( U_i \) is the standard speech spectrum level at normal vocal effort.

\( K_i \) represents the decrease in intelligibility caused by masking from noise and can be further expanded as:

\[ K_i = \min (1, \max (0, \sigma_i)), \quad (3) \]

where \( \sigma_i \) is a function of the equivalent speech spectrum level and the disturbance level, \( D_i \):

\[ \sigma_i = \frac{E_i - D_i + 15}{30}. \quad (4) \]

The disturbance level is defined as:

\[ D_i = \max \{ Z_i, X_i + T_i \}, \quad (5) \]

where \( X_i \) represents the internal noise spectrum and \( Z_i \) is the equivalent masking spectrum. If the listener is impaired we can introduce an equivalent hearing threshold level, \( T_i \), which accounts for various impairments and is added to the internal noise spectrum. The equivalent masking spectrum, \( Z_i \) is expanded as follows for third octave bands:

\[ Z_i = \begin{cases} B_i, & i = 0 \\ 10 \log_{10} \left( 10^{\frac{N_i}{10}} + \sum_{k=1}^{i-1} 10^{\frac{f_k+3.5f_k \log (0.89 f_k)}{10}} \right), & i \neq 0 \end{cases} \]

\( i \neq 0 \),

where \( N_i \) is the noise spectrum level and \( f_i \) is the nominal midband frequency of the \( i \)th sub-band. \( C_k \) is the slope per octave of the spread of masking which, for third-octave bands, is given as:

\[ C_i = -80 + 0.6 (B_i + 10 \log_{10} f_i) - 6.353). \quad (7) \]

Finally, \( B_i \) is the effective noise in the \( i \)th band:

\[ B_i = \max \{ N_i, V_i \}, \quad (8) \]

where \( V_i \) is the self-masking spectrum level of the speech:

\[ V_i = E_i - 24. \quad (9) \]

### 3. APPROXIMATION FOR SII

The above formulation is not very easy to use in optimisation procedures. In order to fully utilise the SII we need to find an approximation that is better behaved. Several of the expressions above use max and min functions which are not differentiable. Thus we approximate these terms using the Jacobian maximum:

\[ \max (a, b) \approx \ln \left( e^a + e^b \right). \]

The corresponding minimum can be found using the following identity:

\[ \min (a, b) = -\max (-a, -b). \]

The above functions are used to approximate (2), (5) and (8) as follows:

\[ L_i \approx -\ln \left( e^{-\frac{U_i-E_i}{100}} + e^{-1} \right) \]

\[ D_i \approx \ln (e^{N_i} + e^{X_i + T_i}) \]

\[ B_i \approx \ln (e^{N_i} + e^{V_i}). \]

We can approximate (3) using both the maximum and minimum approximations from above:

\[ K_i \approx \frac{-\ln \left( e^{-\lambda} + e^{-\ln (1+e^{\lambda \sigma_i})} \right)}{\lambda}, \quad (10) \]

where \( \lambda \) determines the accuracy of the estimate. As (3) uses both the minimum and maximum functions close together, \( \lambda \) is introduced to force the approximation closer to the turning point of each function. Higher values for \( \lambda \) result in a more accurate approximation but a higher peak derivative. As (3) is itself an approximation based on the original experiments in [2], by considering this work a better value for \( \lambda \) can be chosen. In our experiments we have set \( \lambda = 100. \)

#### 3.1. Differential

The above approximations allow us to form the partial derivative of the SII with respect to each spectral level, \( E_i \), whilst holding all other spectral levels constant, as follows:

\[ \frac{\partial \Psi}{\partial E_j} = \frac{\partial}{\partial E_j} \sum_{i=1}^{\max_i} L_i K_i \]

\[ = \sum_i^n \left( \frac{\partial L_i}{\partial E_j} K_i + \frac{\partial K_i}{\partial E_j} L_i \right) \]

The level distortion factor is differentially as follows:

\[ \frac{\partial L_i}{\partial E_j} = \begin{cases} -\frac{1}{100} e^{-\ln \left( U_i-E_i+170 \right)} & i = j \\ 0 & i \neq j \end{cases} \]

Noise and masking differentiation:

\[ \frac{\partial K_i}{\partial E_j} = \begin{cases} \frac{1}{30} \left( \frac{1+e^{\lambda \sigma_i}}{e^{-\lambda} + (1+e^{\lambda \sigma_i})} \right)^2 & i = j \\ 0 & i < j \end{cases} \]

\[ = \begin{cases} \frac{1}{30} \left( \frac{1+e^{\lambda \sigma_i}}{e^{-\lambda} + (1+e^{\lambda \sigma_i})} \right)^2 & i < j \end{cases} \]

\[ \frac{\partial D_i}{\partial E_j} = \frac{e^{E_i}}{e^{X_i} + e^{E_i + T_i}} \frac{\partial Z_i}{\partial E_j} \]

Using (13) we can simplify the summation part of the derivative:

\[ \frac{\partial \Psi}{\partial E_j} = L_i \left( \frac{\partial L_i}{\partial E_j} K_i + \frac{\partial K_i}{\partial E_j} L_i \right) + \sum_{i=j+1}^{\max_i} L_i \frac{\partial K_i}{\partial E_j} \]

\[ - \sum_{i=1}^{\max_i} L_i \frac{\partial L_i}{\partial E_j} \]
We can simplify the derivative expressions by assuming that the effective noise level, \(N_e\), is greater than the self-speech masking level, \(V_s\), so that (8) becomes \(B_i = N_e\). This assumption has been made in several previous approaches [13, 12, 14] and holds well for SNRs up to 35 dB, where the SII is already high. As \(B_i\) is no longer a function of \(E_i\), \(\frac{d\xi_i}{dE_j} = 0\), and therefore \(\frac{dE_i}{dE_j} = 0\), which gives \(\frac{dE_i}{dE_j} = 0\). Thus our derivative can be further simplified using the following:

\[
\frac{\partial K_i}{\partial E_j} = 0, \quad i \neq j.
\]

Therefore (14) becomes:

\[
\frac{\partial \Psi}{\partial E_j} = I_j \left( \frac{\partial L_i}{\partial E_j} K_j + \frac{\partial K_j}{\partial E_j} I_j \right),
\]

where (12) becomes:

\[
\frac{\partial L_i}{\partial E_j} = -\frac{1}{160} \exp \left( -\frac{1}{100} (U_j - E_j + 170) \right)
\]

and from (13) we obtain:

\[
\frac{\partial K_j}{\partial E_j} = \frac{1}{30} \left( 1 + e^{\lambda \sigma_i} \right)^{-1} e^{\lambda \sigma_j}.
\]

3.2. Approximation Validity

To evaluate the accuracy of the estimate we tested a range of speech and noise signals. The speech signals comprised the TIMIT core set, [16], and we used 8 different noise types from NOISEX-92, [17]. The SNR was varied from −30 to 30 dB, and the speech presentation level was set to 62.5 dB SPL when measured with P.56 [18], which is specified as the normal speech presentation level from Table 2 in [1].

The signals were split into separate bands using a third-octave filter bank, where \(l_{\text{max}} = 18\), with a sampling frequency of 16 kHz: \(x_i(t)\) and \(n_i(t)\). The equivalent spectral levels are computed as follows:

\[
E_i = 10 \log_{10} \left( \hat{E}_i \right) - B_i,
\]

where \(\hat{E}_i = \frac{1}{T} \sum_{t=0}^{T-1} x^2_i(t)\) is the average power in band \(i\) and \(B_i\) is the filter bandwidth in dB, from Table 3 in [1]. Similarly we have \(N_i = 10 \log_{10} \left( \hat{N}_i \right) - B_i\) for the noise equivalent spectral levels.

For each pair of speech and noise segments the SII was computed using several methods. The mean average was taken over all the 240 speech segments and noise types. The results were compared to that of the original ANSI standard and to those proposed by [13]. The approximation from [13] is as follows:

\[
(\mathcal{T} - \tilde{N}_i) \approx \sum_i I_i \frac{\xi_i}{\xi_i + 1},
\]

where \(\xi_i = \frac{\hat{E}_i}{\hat{N}_i}\) is the SNR in each band.

The upper graph in Fig. 1 plots the SII versus SNR averaged over 240 speech utterances and 8 noise types together with our proposed approximation and that from [13]. We see that while [13] slightly overestimates the true SII, our proposed approximation is extremely close. The lower graph in Fig. 1 shows the standard deviation of the error in the two approximations and we see that this too has been greatly reduced and is extremely small for our proposed approximation.

3.3. Hearing Impaired Listeners

In [13], (5): \(D_l = \max \{ Z_i, X_i + T_i \} \), is approximated using \(D_l = Z_i\). However, for impaired listeners the internal noise spectrum, \(X_i + T_i\), can be much larger than \(Z_i\). Thus this approximation does not hold well in this case. Fig. 2 shows the case of a moderately-severe presbycusis listener with hearing loss of \(T_i \leq 60\) dB, [19, 20, 21]. The SII is poorly approximated by [13] above 0 dB, whereas the proposed method still remains accurate. The proposed approximation is therefore suitable for use in hearing aid applications where the accurate assessment of intelligibility is essential.
4. APPLICATION TO NEAR-END LISTENING ENHANCEMENT

The proposed continuous approximation allows us to utilise the derivative which can assist applications in which the aim is to maximise the SII. Near-end listening enhancement is one such application which attempts to find the optimal filter to improve the SII of a speech segment presented in a given noisy environment.

A block diagram of a general listening enhancement system is shown in Fig. 3. Initially we compute the equivalent spectral levels of the speech, $E_i$, and the noise, $N_i$, in each band, as in (15). The filtering processing multiplies each speech subband spectral level by a constant, $W_i$, as in (15). The filtering processing multiplies each speech subband spectral level by a constant, $W_i$. Thus the equivalent spectral levels of the output signal are equal to $\tilde{E}_i = E_i + W_i$. The resulting SII is a function of $\tilde{E}_i$ and $N_i$, therefore, using a vector notation for a set of spectral levels $E = \{E_i\}$ etc., the optimisation problem is:

$$W_A = \arg \max_W \Psi(E + W, N)$$

subject to: $\sum_i 10^{F_i + W_i + R_i / 10} \leq \sum_i 10^{F_i + R_i / 10}$. (17)

The constraint imposes a limitation on the power output of the filtered signal. Due to limitations of the hardware, in real applications there is usually an upper bound on the signal power. For example, applications with small speakers such as a mobile phone or hearing aid. The speaker or amplifier is likely driven at its limit to maximise the output and further amplification will introduce distortions. Thus we do not want to increase the output power beyond this level.

The final output signal can be composed by adding the sub-band signals together, to give $y(t) = \sum_{i=1}^{\text{bands}} x(t) 10^{W_i / 10}$.

To solve the optimisation problem, previous approaches have used approximations of the SII to make the optimisation convex. For example, (16) is convex, which guarantees a unique solution. Let us denote the optimised filter weights from [13], (16), as $W_T$:

$$W_T = \arg \max_W \Psi_T(E + W, N)$$

subject to: $\sum_i 10^{F_i + W_i + R_i / 10} \leq \sum_i 10^{F_i + R_i / 10}$. (18)

In our proposed solution the derivative is defined everywhere, thus the proposed approximation can be effectively used in optimisation. However as there exist multiple local maxima for (17), it is possible for the optimiser to converge to a sub-optimal set of weights. For this reason we use the solution from (18), $W_T$, which is globally convex, as the initial filter weights and then optimise from there to converge on an improved solution, $W_P$:

$$W_P = \arg \max_W \Psi_P(E + W_T + W, N)$$

subject to: $\sum_i 10^{E_i + W_i + W_T_i + R_i / 10} \leq \sum_i 10^{E_i + R_i / 10}$. (19)

4.1. Results

The SII before and after listening enhancement was evaluated using the ANSI SII standard, on the output equivalent speech spectral levels: $\Psi(E + W, N)$. The results were averaged over all speech segments and all noise types as presented in Section 3.2. Fig. 4 shows the SII improvement from the weighted speech produced using $W_A$, $W_T$ and $W_P$. The weights were optimized using the fmincon non-linear optimiser within MATLAB [22].

Using our proposed method we see small performance gains over (18) and large gains over the un-enhanced speech as shown in Fig. 4. The increase in SII, over the unprocessed signal, can correspond to an increase in intelligibility of 5-20% for sentences not previously heard by the listener [23]. At SNRs above 15dB, the performance gains over (18) increase. At all SNRs the performance is positive as expected. Furthermore by using an inequality constraint on the instantaneous power, we are able to achieve gains in SII whilst reducing the power output of the enhanced signal. This is particularly noticeable at high SNRs, as shown in Fig. 5.

5. CONCLUSIONS

A new approximation to the SII has been proposed. It has a globally defined derivative which is continuous. Experiments show that the estimate closely matches the original ANSI standard. The proposed approximation was applied to the application of near-end listening enhancement. Over a varied set of speech and noise segments, results show that utilising the approximation gives large positive gains in SII.

![Fig. 3: Block diagram of a listening enhancement system](http://example.com/block_diagram.png)

![Fig. 4: SII improvement for different algorithms averaged over the TIMIT core test set and 8 noise types, top - mean average, bottom - standard deviation](http://example.com/sii_improvement.png)
Fig. 5: The relative power difference before and after listening enhancement.

compared with an unprocessed signal as well as an existing method. It also allows a reduction in signal power without a loss of intelligibility.

6. REFERENCES


