A Spectro-Temporal Glimpsing Index (STGI) for Speech Intelligibility Prediction

Amin Edraki1, Wai-Yip Chan1, Jesper Jensen2,3, and Daniel Fogerty4

1Department of Electrical and Computer Engineering, Queen’s University, Kingston, Canada
2Department of Electronic Systems, Aalborg University, Aalborg, Denmark
3Oticon A/S, Smørum, Denmark
4Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

a.edraki@queensu.ca, chan@queensu.ca, jesj@demant.com, dfogerty@illinois.edu

Abstract

We propose a monaural intrusive speech intelligibility prediction (SIP) algorithm called STGI based on detecting glimpses in short-time segments in a spectro-temporal modulation decomposition of the input speech signals. Unlike existing glimpse-based SIP methods, the application of STGI is not limited to additive uncorrelated noise; STGI can be employed in a broad range of degradation conditions. Our results show that STGI performs consistently well across 15 datasets covering degradation conditions including modulated noise, noise reduction processing, reverberation, near-end listening enhancement, checkerboard noise, and gated noise.

Index Terms: speech intelligibility, speech quality model, spectro-temporal modulation, glimpsing

1. Introduction

In human-human communications mediated by machines and networks, speech signals are often distorted and not perfectly perceived by the listener. Machine algorithms can facilitate predicting the perceptual effect of different types of distortion-processing on speech signals. An ideal speech intelligibility prediction (SIP) algorithm would accurately predict a degraded/processed signal’s average intelligibility as perceived by a group of listeners. This study aims to develop an intrusive or reference-based SIP algorithm for normal-hearing listeners that relies on inputting a clean reference signal.

Different approaches to computational modelling of speech perception in adverse listening conditions have been proposed [1–4]. In this regard, the glimpsing model of speech perception [4] was proposed based on two observations about the spectro-temporal energy distribution of the speech signal: the redundancy of the carried information and the sparse distribution of the regions with high energy. Based on these observations, the glimpsing model of speech perception hypothesizes that listeners process degraded speech by taking advantage of “glimpses”-time-frequency (TF) regions where the signal is least distorted. Using different glimpse detection criteria, Cooke [4] suggested that the proportion of the TF regions glimpsed -glimpse proportion (GP)- is a good predictor of intelligibility. GP has shown strong positive correlation with the intelligibility of speech degraded by additive uncorrelated noise (AUN) [4–7]. The glimpsing model proposed in [4] uses the local SNR as the glimpse detection criterion: the TF units whose local SNR exceeds 3 dB are detected as glimpses. While providing a simple and effective criterion for glimpse detection, using SNR limits the application of the method to speech degraded by AUN. In many common scenarios, the speech degradation cannot be treated as caused by AUN, e.g., speech processed by non-linear noise reduction algorithms and reverberated speech. To remove this limitation, here, we modify GP to extend its applicability to distortions beyond AUN. To this end, we use the normalized cross-correlation (NCC) between the spectro-temporal modulation (STM) envelopes of the clean and degraded/processed speech signals [6–10] as a measure of local signal quality and craft a glimpse detection criterion tailored to the measure. The NCC is calculated between the STM envelopes of the clean speech signal and their distorted versions and does not require a separate noise signal to be available. Therefore, the resulting algorithm can be applied to any pair of time-aligned clean and degraded speech signals.

2. Glimpse Proportion

To motivate the proposed algorithm, we synopsize the GP introduced in [4, 5] and defined as the proportion of the regions in the TF representation of the speech signal where the local SNR exceeds a certain threshold:

\[ GP = \frac{1}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} \frac{C[f, t]}{SNR_{\alpha}} \] (1)

where \( 1 \leq t \leq T \) indexes time frames, \( 1 \leq f \leq F \) indexes acoustic frequency channels, \( T \) and \( F \) are the number of time frames and acoustic frequency bins, respectively, and:

\[ C[f, t] = \mathbb{I}_+ (\text{SNR}[f, t] - \alpha) \] (2)

where SNR\([f, t]\) denotes the local SNR, \( \alpha \in \mathbb{R} \) denotes an SNR threshold, and \( \mathbb{I}_+: \mathbb{R} \rightarrow \{0, 1\} \) is the indicator function:

\[ \mathbb{I}_+(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases} \] (3)

GP takes the time-aligned clean and AUN signals as input and computes the local SNR for each TF unit using the spectro-temporal excitation patterns (STEP) of the inputs, where STEP is a representation of the envelope of the basilar membrane response to a sound signal. GP has shown high correlation with the intelligibility of speech corrupted by AUN [4–7]. However, in many common scenarios, the difference between the clean and corrupted speech cannot be treated as AUN. E.g., for noise suppression processed speech, the difference signal is some non-linear mixture of the noise and speech signals.
3. Proposed Algorithm

To overcome the AUN limitation, we modify GP in two ways. First, we use a Gabor STM filter-bank [7, 11, 12] to decompose the input speech signals into STM frequency sub-bands. Perception of speech is crucially dependant on preserving the fidelity of salient STMs [9, 13–15], and STM analysis has previously shown promising results in SIP [6, 7, 10] and automatic speech recognition [11, 12]. Second, we modify the glimpse detection function as opposed to SNR, as an estimator of speech intelligibility [6–8, 10]. The NCC is computed between the STM envelopes of the clean and degraded speech signals and, thus, does not require an additive noise signal to be separately available.

Figure 1 shows a block diagram of the proposed algorithm. The algorithm takes the time-aligned clean \( x[n] \) and degraded \( y[n] \) speech signals as input. First, the STM envelopes of the input signals are extracted using a Gabor STM filter-bank. Next, the envelopes are divided into short-time segments, and each segment of the degraded signal is compared to that of the clean signal via NCC to produce a similarity measure. Next, similar to Eq. 2, a thresholding is applied to the similarity measures to detect STM glimpses. Finally, the overall intelligibility is estimated as the proportion of the glimpsed segments over the entire input. The building blocks of the algorithm are described in the following subsections.

3.1. Auditory-Modulation Analysis

First, an ideal voice activity detector (VAD) based on the cleanc reference signal [6, 8] is used to remove silent frames from the clean and degraded signals. Next, a Mel-frequency spectrogram is calculated for both input signals following the ETSI standard [7, 11, 16]. The Mel-spectrograms for the clean and degraded speech signals are denoted by \( X \) and \( Y \), respectively; and consist of \( F \) acoustic frequency channels equally spaced on the Mel-frequency scale. Next, the Mel-spectrograms are decomposed using an STM filter bank [7, 11, 12]. The filter bank consists of \( S \) spectral and \( R \) temporal Gabor modulation filters covering the range of modulation frequencies that are known to be important for speech perception [11] and SIP [7].

The STM filter-bank is implemented as a two-step procedure: spectral modulation filtering followed by temporal modulation filtering. The filter bank’s output consists of \( S \times R \) filtered spectrograms that exhibit modulation patterns characteristic of the associated filters. We denote the resultant TF decompositions by \( \hat{X}[f, t; s, r] \) and \( \hat{Y}[f, t; s, r] \), where \( 1 \leq s \leq S \) and \( 1 \leq r \leq R \) index the spectral and temporal modulation center frequencies, respectively, of the associated filters. For a detailed description of the modulation filter-bank, we refer to [7, 12].

3.2. Envelope Comparison

As discussed in Sec. 2, GP uses the local SNR as a measure of speech degradation in each TF unit. Here, we develop a similarity measure between the clean and degraded speech signals by comparing their STM envelopes in short-time segments [6, 8]. We present the processing steps for a single STM channel, i.e., one filtered spectrogram; the steps are identical for all filtered spectrograms. Let \( \hat{X}^k, 1 \leq k \leq T - N + 1 \) denote the \( k \)-th short-time segment of length \( N \):

\[
\hat{X}^k = \begin{bmatrix}
\hat{X}_1[k] & \hat{X}[1, k + 1] & \cdots & \hat{X}[1, k + N - 1] \\
\hat{X}[2, k] & \hat{X}[2, k + 1] & \cdots & \hat{X}[2, k + N - 1] \\
\vdots & \vdots & \ddots & \vdots \\
\hat{X}[F, k] & \hat{X}[F, k + 1] & \cdots & \hat{X}[F, k + N - 1]
\end{bmatrix}.
\]

(4)

Note that the dependencies on \( s \) and \( r \) are dropped for simplicity. First, mean and variance normalization is applied to each row and each column of the short-time segments [6]. We denote the clean and degraded row- and column-normalized short-time segments by \( \hat{X}^k = [\hat{x}_{s, r}^k] \) and \( Y^k = [\hat{y}_{s, r}^k] \), respectively. Next, the columns of \( \hat{X}^k \) and \( Y^k \) are compared using NCC, and the NCCs are averaged across columns to produce the STM channel’s similarity measure for the \( k \)-th short-time segment:

\[
d^k[s, r] = \frac{1}{N} \sum_{j=1}^{F} \sum_{r=1}^{R} \hat{x}_{s, r}^k \hat{y}_{s, r}^k.
\]

(5)

Each of the STM channels \([s, r]\) contributes a similarity measure \( -1 \leq d^k[s, r] \leq 1 \) for the \( k \)-th time segment.

3.3. Glimpse Detection

As described in Eq. 2, GP compares the local SNR with a fixed threshold to determine each STM channel’s glimpse estimate for the \( k \)-th short-time segment:

\[
g^k[s, r] = 1 + (d^k[s, r] - \beta_{sr}),
\]

(6)

where \( -1 \leq \beta_{sr} \leq 1 \) is the glimpsing threshold for the \([s, r]\) STM channel to be determined with the aid of a small dataset in Sec. 3.5. Note that each STM channel contributes a glimpse estimate for the \( k \)-th time segment. Lastly, the glimpsing index for the \( k \)-th short-time segment is computed as the average of glimpse estimates over all STM channels:

\[
g[k] = \frac{1}{SR} \sum_{s=1}^{S} \sum_{r=1}^{R} g^k[s, r].
\]

(7)

Fig. 2 shows the glimpsing index \( g[k] \) as a function of time for a speech signal sample degraded by AUN.

3.4. Intelligibility Prediction

The spectro-temporal glimpsing index (STGI) is defined as the temporal average of the short-time glimpsing index over the entire input:

\[
\text{STGI} = \frac{1}{K} \sum_{k=1}^{K} g[k].
\]

(8)

3.5. Glimpse Thresholds

Here, we describe the procedure to select the glimpsing threshold \( \beta_{sr} \) for each STM channel. For simplicity, we employ a greedy approach and aim to maximize each individual STM
clean and degraded speech signals have different statistics and result in different distributions of the NCCs across the channels. Fig. 4 shows the resulting thresholds for the STM channels.

4. Implementation and Evaluation

Tables 1 and 2 summarize the model parameters used in this study. The glimpsing thresholds \( \beta_{sr} \) were calculated following the procedure described in Sec. 3.5. We compare the SIP performance of STGI with a suite of state-of-the-art SIP algorithms in terms of i) Pearson correlation coefficient and ii) Spearman rank correlation coefficient between the ground truth intelligibility and the algorithm outputs.

We evaluate the SIP performance of STGI across 15 listening test datasets, which together cover degradation conditions including modulated noise (ModN), checkerboard noise (CheckB), gated noise (GatedN), noise reduction processing (NR), reverberation (Rev), and near-end listening enhancement (NELE). Table 5 summarizes the data sets, while the accompanying references provide additional details. The first 6 datasets are for AUN, dataset 7 includes temporally filtered speech, datasets 8 and 9 include noisy speech processed by non-linear noise-reduction algorithms, dataset 10 covers ideal TF masking, datasets 11 and 12 consist of reverberated speech, and datasets 13-15 include NELE processed speech degraded by noise and reverberation.

Table 1: Model parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_s )</td>
<td>10 kHz</td>
<td>( F )</td>
<td>130</td>
</tr>
<tr>
<td>( S )</td>
<td>11</td>
<td>( R )</td>
<td>4</td>
</tr>
<tr>
<td>( N )</td>
<td>40 (512 ms)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Spectral and temporal modulation center frequencies.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral Modulation Filter Center Frequencies (cyc/kMel)</td>
<td>0, 0.28, 0.44, 0.72, 1.10, 1.77, 2.88, 4.54, 7.26, 11.6, 18.3</td>
</tr>
<tr>
<td>Temporal Modulation Filter Center Frequencies (Hz)</td>
<td>0, 4.9, 7.8, 12.4</td>
</tr>
</tbody>
</table>

5. Results and Discussion

Tables 3 and 4 show the Pearson and Spearman rank correlation coefficients between the ground truth intelligibility and each SIP algorithm’s output over the datasets introduced in Sec. 4. The results indicate excellent performance of STGI over a wide range of degradations and datasets.

Comparing STGI and GP, a few observations can be made. Firstly, different from GP, the application of STGI is not limited
to AUN and it can be used for SIP in acoustic conditions where the speech and noise are not separately available. Secondly, STGI consistently outperforms GP in the presence of AUN.

Another interesting observation is the failure of several state-of-the-art SIP algorithms in the presence of checkerboard and gated noises. These conditions consist of artificial maskers that provide a controlled environment for studying the influence of the spectro-temporal properties of the glimpses on speech perception. A SIP algorithm’s ability to successfully capture and model such properties can lead to performance gains in other degradation conditions. We note that several successful SIP algorithms (including wSTMI, eSTOI, and SIIB) that have excellent performance across many acoustic conditions, struggle to predict intelligibility in the presence of checkerboard and gated noises. On the other hand, STGI consistently shows high correlation with speech intelligibility in such degradation conditions (CheckB1-Fogerty, CheckB2-Fogerty, and GatedN-Fogerty datasets). Like STGI, wSTMI also decomposes the input speech signals using a Gabor STM filter bank and compares the filtered spectrograms using NCC. However, different from STGI, wSTMI uses a sparse linear model to combine information from only the most salient 8 STM channels. While the approach has shown excellent performance in a wide range of degradations, it fails to predict intelligibility in the presence of checkerboard noise. Unlike conventional distortions, the maskers’ energy in the checkerboard noise datasets is distributed over a small number of STM channels, thus preserving much of the speech information in the other STM channels. While the human auditory system can take advantage of the information from less affected STM channels, the sparse representation of wSTMI cannot capture this opportunistic behaviour. STGI emulates this behaviour by pooling the information from all the STM channels in the range that is usable for speech perception.

### 6. Conclusion

We presented an intrusive speech intelligibility prediction (SIP) algorithm based on detecting glimpses in a spectro-temporal modulation decomposition of the input speech signal. The algorithm -STGI- was developed by overhauling a pre-existing speech intelligibility prediction scheme -GP- and thereby expanding its narrow applicability to a wide range of distortions, including speech processed by non-linear noise suppression algorithms and reverberated speech. The results suggested that the proposed algorithm outperforms established SIP methods in the presence of temporally and spectro-temporally gated maskers, while also consistently provides high performance for the other tested degradation conditions.
7. References


