Improving intelligibility prediction under informational masking using an auditory saliency model

Tang, Y and Cox, TJ

<table>
<thead>
<tr>
<th>Title</th>
<th>Improving intelligibility prediction under informational masking using an auditory saliency model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Tang, Y and Cox, TJ</td>
</tr>
<tr>
<td>Type</td>
<td>Conference or Workshop Item</td>
</tr>
<tr>
<td>URL</td>
<td>This version is available at: <a href="http://usir.salford.ac.uk/47113/">http://usir.salford.ac.uk/47113/</a></td>
</tr>
<tr>
<td>Published Date</td>
<td>2018</td>
</tr>
</tbody>
</table>

USIR is a digital collection of the research output of the University of Salford. Where copyright permits, full text material held in the repository is made freely available online and can be read, downloaded and copied for non-commercial private study or research purposes. Please check the manuscript for any further copyright restrictions.

For more information, including our policy and submission procedure, please contact the Repository Team at: usir@salford.ac.uk.
IMPROVING INTELLIGIBILITY PREDICTION UNDER INFORMATIONAL MASKING USING AN AUDITORY SALIENCY MODEL

Yan Tang
Acoustics Research Centre,
University of Salford
Salford, UK
Y.Tang@salford.ac.uk

Trevor J. Cox
Acoustics Research Centre,
University of Salford
Salford, UK
T.J.Cox@salford.ac.uk

ABSTRACT

The reduction of speech intelligibility in noise is usually dominated by energetic masking (EM) and informational masking (IM). Most state-of-the-art objective intelligibility measures (OIM) estimate intelligibility by quantifying EM. Few measures model the effect of IM in detail. In this study, an auditory saliency model, which intends to measure the probability of the sources obtaining auditory attention in a bottom-up process, was integrated into an OIM for improving the performance of intelligibility prediction under IM. While EM is accounted for by the original OIM, IM is assumed to arise from the listener’s attention switching between the target and competing sounds existing in the auditory scene. The performance of the proposed method was evaluated along with three reference OIMs by comparing the model predictions to the listener word recognition rates, for different noise maskers, some of which introduce IM. The results show that the predictive accuracy of the proposed method is as good as the best reported in the literature. The proposed method, however, provides a physiologically-plausible possibility for both IM and EM modelling.

1. INTRODUCTION

Speech communication often takes place in non-ideal listening environments. Speech intelligibility is often negatively affected by background noise, leading to the potential failure of information transmission. In order to efficiently quantify the extent to which the background noise harms intelligibility, a great number of objective intelligibility measures (OIM) have been proposed in the last decades. They have been used as a perceptual guide in activities such as development of modification algorithms for highly-intelligible speech [1], speech enhancement [2], production of TV or radio broadcast [3] and research in hearing impairment [4]. OIMs have an important role in developing speech and noise processing algorithms for an inclusion.

Standard measures, such as the Speech Intelligibility Index (SII, [5]) and the Speech Transmission Index [6], and early OIMs (e.g. [4, 2]) make intelligibility predictions based on long-term masked audibility (e.g. SII) or modulation reduction (e.g. STI) of the target speech signal. More recent methods [8, 9, 10, 11] operate on short windows (10-300 ms), in order to improve the predictive accuracy in temporally-fluctuating noise maskers. In addition, some of the measures [10, 11] were developed on the basis of sophisticated auditory models, and have demonstrated more robust predictive power in a wide range of conditions [12]. In the light of the fact that listener do not need all time-frequency (T-F) information to successfully decode the speech [13], Cooke proposed a glimpsing model of speech perception in noise [14]. In [14], the percentage of the T-F regions of speech with a local speech-to-noise ratio (SNR) meeting a given criteria was calculated as the intelligibility proxy, known as the glimpse proportion (GP). It can be thought of the overall contribution from the local audibility of all the T-F regions to intelligibility in noise. Tang and Cooke further extended the GP to a complete intelligibility measure – the extend GP (ext. GP) – which performs detail modelling of the masking effect taking place in the auditory peripheral from the outer-middle ear, through the cochlea to the inner-hair cells. The predictions by ext. GP are well correlated with listener word recognition performance in various noise maskers, with correlation coefficients greater than 0.85 [11].

Energetic masking (EM) and informational masking (IM) introduced by the noise masker mainly account for the reduced intelligibility. EM is the consequence of interactions of physical signals acting in the auditory peripheral. IM is different as it obstructs auditory identification and discrimination at the late stage of auditory pathway, when a sound is perceived in the presence of other similar sounds [15, 16, 17]. However, this is overlooked by the aforementioned OIMs, which can only quantify the impact of EM on intelligibility from the physical attributes of the speech and noise signals. When comparing the GP in speech-shaped noise (SSN) and competing speech (CS), Tang and Cooke found that to achieve the same intelligibility much fewer glimpses are required in SSN than in CS, which introduces large IM [11]. They postulated that IM made some glimpses ineffective.

With a classification of the glimpsed T-F regions based on energy, it was further found that the regions with energy above the average are more robust in noise. Those with energy under the average are more susceptible to both EM and IM [11]. Importantly, the amount of high-energy glimpses is broadly consistent for the same speech signal in SSN and CS under SNRs leading to the similar listener performance. Although this method, know as high-energy glimpse proportion (HEGP), is a crude approach for making consistent predictions when the masker is in presence or absence of IM, it confirmed an early hypothesis that IM may affect the effectiveness of the glimpsed T-F regions available for speech.

One possible explanation is that the listener switches attention between the target and the competing sources, leading to some of the target components that have triggered activities at the auditory pathway, when a sound is perceived in the presence of other similar sounds [15, 16, 17]. The perceptual and cognitive resources that a human’s nervous system can use to process the input sensory stimuli received in a short time window is limited. Consequently, the brain temporarily and selectively stores only a subset of available sensory information in short-term working memory for further processing [18, 19, 20]. This selection is a combination of rapid bottom-up signal-driven
2. PROPOSED METHOD

The proposed method consists of two main parts, as illustrated in Fig. 1. The first part is the ext. GP [11], which models the energetic masking taking place at the auditory peripheral. The second part (shaded and on the left of Fig. 1) performs saliency analysis on the given auditory scene as a whole, quantifying the probability of the T-F regions on the scene gaining processing in the later stage of the auditory pathway in a bottom-up process. The output of this part, the saliency map $S_M$, is subsequently combined with glimpse representation $G'$ from ext. GP, in order to adjust the contribution of the glimpses to the final intelligibility.

2.1. Quantifying energetic masking

Energetic masking is modelled using ext. GP [11]. To generate the auditory representations -- the spectro-temporal excitation patterns (STEP) -- for the signals, the clean speech signal $s$ and noise signal $n$ are passed through 64-gammatone filterbanks. The frequency locations of the 64 filters are evenly distributed on the equivalent rectangular bandwidth (ERB) scale, ranging from 100 to 7500 Hz, with a spectral resolution of 0.51 ERB. An outer-middle ear transfer function [39] is applied to the filter outputs, in order to account for the auditory sensitivity (i.e. hearing threshold) to the level of the signal at different frequencies. The Hilbert envelopes of each frequency band, $E(f)$, is then extracted, smoothed by a leaky integrator with an 8 ms time constant and downsampled to 100 Hz. A log-compression is imposed on the final output.

The glimpses are determined by comparing the STEP of the speech signal $STEP_s$, against that of the noise signal $STEP_n$. A glimpsed T-F region must possess a local SNR above a given threshold $(\Delta=3\, \text{dB})$, and be above the hearing level (HL, set to 25 dB).

\[
G(t, f) = \text{STEP}_s(t, f) > \max(\text{STEP}_n(t, f) + \Delta, \text{HL}) \quad (1)
\]

To account for forward masking, the raw glimpses $G$ are further validated using an inner-hair cell model (IHC, [40]), which also takes the envelopes of speech-plus-noise mixture $E_m$ as the input. The glimpsed T-F regions surviving from simultaneous masking are considered valid only when their corresponding IHC outputs are not masked during the IHC depleting and replenishing process. Hence, the IHC-validated glimpse $G'$ is defined as,

\[
G'(t, f) = G(t, f) \wedge \neg g(t, f) \quad (2)
\]

where $\wedge$ indicates logic ‘and’, and $g$ denotes the masked glimpses due to forward masking. For the rules for IHC validation, see [11] for details.

The plots in the second row of Fig. 2 exemplify the valid glimpsed T-F regions on a speech signal in SSN and CS at 1 and $-7\, \text{dB}$ SNR, respectively. The chosen SNRs led to a similar intelligibility in the two maskers [41].

2.2. Generating saliency map

A saliency map is also a T-F representation produced from STEP. Generating a saliency map often involves feature extraction, nor-

---

1 Saliency analysis requires a greater number of filters to maintain the T-F resolution of its output than previously used for ext. GP. Instead of 34-channel described in [11], 64-channel STEPs were used here for ext. GP, in order to keep the representations consistent. Tests have shown that filter numbers above 34 have little impact to the performance of ext. GP.
malisation, combination and resizing. After \cite{22, 52, 33}, the features $F(\sigma, \theta, \alpha)$ including intensity $F^I$, spectral contrast $F^S$, temporal contrast $F^T$ and orientation $F^O$ are extracted from the STEP of the speech-plus-noise mixture $STEP_m$. This is performed in a multi-scale manner \cite{18}; eight scales $\sigma = \{1, ..., 8\}$ are used, and the input $STEP_m$ is filtered and decimated by a factor of two iteratively for seven times; the output of the last iteration is the input of the next. This results in size reduction factors ranging from 1:1 to 1:128. The resized STEPs are then convolved by the Gabor filters (which are the product of a cosine grating and a 2D Gaussian envelope) with different $\theta$, which represents one of the four target features, as listed in Table 1.

Table 1: Parameters of the Gabor filters for each feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>$\theta$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>$\pi/2$</td>
<td>0</td>
</tr>
<tr>
<td>Spectral contrast</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Temporal contrast</td>
<td>$\pi/2$</td>
<td>1</td>
</tr>
<tr>
<td>Orientation</td>
<td>${\pi/4, 3\pi/4}$</td>
<td>1</td>
</tr>
</tbody>
</table>

In order to mimic the properties of local cortical inhibition, the ‘centre-surrounding’ differences are calculated after extracting features at multiple scales, yielding a set of feature maps $FM(c, s)$. This is done by across-scale subtraction between a centre finer scale $c \in \{2, 3, 4\}$ and a surrounding coarser scale $s$:

$$FM(c, s) = |FM(c) - FM(s)|$$  \hspace{1cm} (3)

where $s = c + \delta, \delta \in \{3, 4\}$. As the size of the feature representation varies across scales, it (from scale 1 to 8) needs to be normalised prior to the across-scale subtraction. Here the representations for each feature are resized to that of scale 4. The centre-surrounding step finally results in $6 \times 5$ feature maps\cite{42, 32, 33}, of which represent each of the features in different scales.

Across-scale combination aims to generate a so-called ‘conspicuity map’, $CM$, for each feature from the feature maps at different scales, using across scale addition. Due to different dynamic ranges resulting from the extraction process for each feature, the feature maps must be first handled by a nonlinear normalisation procedure in order to bring them into a comparable scale. Another purpose of the normalisation is to simulate competition between neighbouring salient locations \cite{43}. This nonlinear normalisation consists of certain number of iterations (three times is used here), each of which consists of self-excitation and inhibition induced by neighbours. To implement, a 2-D difference of Gaussians (DoG) filter is convolved with each feature map, followed by clamping the negative values to zero. A feature map $FM$ is then transformed in each iteration as follow:

$$CM \leftarrow |FM + FM \ast \text{DoG} - 0.02| \geq 0$$  \hspace{1cm} (4)

After normalisation, the normalised feature maps of different scales can then be summed up to a single conspicuity map. This is repeated for all the four features, resulting in four maps. The final saliency map, $SM$, is a linear combination of all the four normalised maps. A further resizing is required to recover the map size (currently at scale 4) back to the original size (at scale 1, same to the $STEP_m$).

\[ Due to orientation having two sub-conditions (i.e., $\theta \in \{\pi/4, 3\pi/4\}$), there are therefore 12 feature maps for orientation in total.\]
Figure 2: Spectrograms, IHC-validated glimpses \( G' \), saliency maps \( SM \) and saliency-weighted glimpses \( G'' \) of the sentence ‘the birch canoe slid on the smooth planks’ in SSN (left column) and CS (right column) at 1 and -7 dB SNR, respectively.

The plots in the third row of Fig. 2 show saliency maps of the same speech signal corrupted by SSN and CS. While the T-F regions where the glimpses occur are mostly salient in SSN, this is not always the case in CS. For CS, the glimpsed fricative components of speech that have energy concentrated at mid-high frequencies are scarcely salient in the example. These glimpses might have very limited contribution to intelligibility due to IM.

2.3. Intelligibility prediction

The final saliency-adjusted glimpses \( G'' \) is the product of the IHC-validated glimpse \( G' \) and the saliency map \( SM \). The effect of this operation on the glimpses is visualised in the plots at the bottom of Fig. 2. The remaining procedure follows the calculation of ext. GP: \( G'' \) is subsequently weighted by the band importance function \( K \left[ 5 \right] \), followed by a quasi-logarithmic compression in a form of

\[
v(x) = \log(1 + x/0.01)/\log(1 + 1/0.01),
\]

\[
OSI = v \left[ \frac{1}{F} \sum_{f=1}^{F} (K(f) \sum_{t=1}^{T} G'(t, f) \cdot SM(t, f)) \right]
\] (5)

where \( F \) and \( T \) are the number of frequency bands and time frames, respectively. The final predictive index falls between 0 and 1, with the greater number indicating the better intelligibility.

3. EVALUATION

For the reference performance, ext. GP, HEGP and SII were evaluated along with the proposed method.

3.1. Subjective data

The subjective data was drawn from \[41, 44\]. In the two studies, the listener intelligibility was measured as the sentence-level word recognition rate in SSN and CS at three SNR levels for each masker, i.e. -9, -4 and 1 dB for SSN and -21, -14 and -7 dB for CS. The chosen SNRs led to the intelligibility of approximately 25%, 50% and 75% in each masker. While the target sentences were uttered by a male native English speakers, the CS was produced by a female speaker. In contrast to SSN, CS is able to cause strong IM \[45\]. In total, this corpus offers 180 conditions, covering the intelligibility range from 5% to 95%. As this corpus consists of 30 types of speech including those algorithmically-modified for better intelligibility and synthetic speech, it is rather challenging for OIMs to predict from. Tang et al. evaluated up to seven state-of-the-art OIMs using this corpus, the average overall performance – the correlation between the listener performance and the model predictions – across all the OIMs was merely 0.67, with 0.83 being the best \[12\]. Nevertheless, the use of SSN and CS maskers in the corpus provided this study with an ideal experiment protocol (i.e. inclusion of maskers which do or do not introduce IM) for evaluation the proposed method.

3.2. Procedure

The raw model output, \( O \), was transformed to the estimated listener performance using a two-parameter sigmoid function (Eqn. 6), in order to make a direct comparison with the subjective data.

\[
W = \frac{1}{1 + \exp(- (a + b \cdot O))}
\] (6)

where \( a \) and \( b \) are the two open parameters, the values of which are chosen to give a best fit to the subjective data for each OIM; values are presented in Table 2.

<table>
<thead>
<tr>
<th>proposed</th>
<th>ext. GP</th>
<th>HEGP</th>
<th>SII</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>-2.201</td>
<td>-2.864</td>
<td>-4.007</td>
</tr>
<tr>
<td>( b )</td>
<td>8.284</td>
<td>5.837</td>
<td>8.024</td>
</tr>
</tbody>
</table>

The main performance of the OIM was evaluated as the Pearson correlation coefficient \( \rho \) between the measured and estimated intelligibility, as well as the root-mean-square error RMSE.


3.3. Results

Fig. 3 compares the model predictions against the measured intelligibility in the 180 conditions. Overall, ext. GP and SII exhibited visually poorer performance than the other two due to the discrepancy between the predictions in SSN (solid circles) and in CS (open circles). While ext GP overestimated in CS or underestimated in SSN, SII displays opposite behaviour. This will be discussed later. By accounting for the effect of IM using the saliency map to further weight the contribution of glimpses, the proposed method decreases the discrepancy observed for ext. GP in Fig. 3. This led to a significant improvement in accuracy for the proposed method compared to the other two due to the similar fundamental mechanism and predictive performance, even if the proposed method presents a finer and more transparent modelling of speech intelligibility in noise than HEGP. As it quantifies the speaker’s attention during the bottom-up processing. Despite the similar fundamental mechanism and predictive performance, the proposed method presents a finer and more transparent modelling of speech intelligibility in noise than HEGP.

Table 3: Subjective-model Pearson correlation correlations ρ and RMSEs (in parentheses) as the model performance in each sub-conditions. Figure in squared brackets indicates the number of data points from which ρ and RMSE were calculated.

<table>
<thead>
<tr>
<th></th>
<th>proposed</th>
<th>ext. GP</th>
<th>HEGP</th>
<th>SII</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN [90]</td>
<td>0.89 (13.0)</td>
<td>0.88 (13.0)</td>
<td>0.88 (13.1)</td>
<td>0.87 (13.5)</td>
</tr>
<tr>
<td>CS [90]</td>
<td>0.82 (14.0)</td>
<td>0.81 (14.4)</td>
<td>0.83 (13.7)</td>
<td>0.77 (15.7)</td>
</tr>
<tr>
<td>natural [132]</td>
<td>0.86 (13.5)</td>
<td>0.73 (17.8)</td>
<td>0.87 (13.0)</td>
<td>0.70 (18.7)</td>
</tr>
<tr>
<td>synthetic [48]</td>
<td>0.92 (9.0)</td>
<td>0.79 (13.5)</td>
<td>0.93 (8.0)</td>
<td>0.74 (14.8)</td>
</tr>
<tr>
<td>overall [180]</td>
<td>0.82 (15.2)</td>
<td>0.71 (18.5)</td>
<td>0.84 (14.4)</td>
<td>0.68 (19.2)</td>
</tr>
</tbody>
</table>

4. DISCUSSION

The current study aimed to improve the predictive power of the ext. GP metric under informational masking by incorporating an auditory saliency model into the OIM. Having observed that the speech-dominant T-F regions contribute to intelligibility differently in the face of different maskers, a weighting based on the likelihood of a region being selected for further auditory processing in a bottom-up procedure, can help account for the IM effect. Hence, improved performance over the original ext. GP metric is seen, especially when performing across maskers which do or do not introduce IM.

The overall performance of both ext. GP and SII suffers from the separation of their outputs in the two maskers, as seen in Fig. 3. Since speech is more tolerant of EM in CS (i.e. fluctuating masker) than in SSN (stationary masker) at the same SNR level, speech in CS must be presented at a lower global SNR to obtain the same intelligibility level as in SSN. However, due to its large envelope modulations, CS provides more opportunities for glimpsing T-F regions on the target signal than in SSN. Even so, the additional glimpses in CS are not translated to intelligibility gain. The over-estimation of ext. GP in CS is thus attributed to the IM effect not being accounted for. On the other hand, Tang et al. explained that SII scores lower in CS than in SSN is due to its long-term spectral SNR-based calculation being sensitive to any change in global SNR, which is a more dominant factor to speech intelligibility in noise than IM.

The proposed method achieved the same performance as the HEGP metric, which assumes that the amount of the high-energy T-F regions on the speech signal is determinant for intelligibility prediction in noise. In terms of EM, more energy offers bigger chance of surviving from the masking to this group of T-F regions, it is therefore more likely for them to be glimpsed by the listener. In the meantime, relatively high intensity in these regions may cause large spectral and temporal contrasts across both the time and frequency at the boundaries when intensity dramatically increases or decreases, e.g. at the transition between a consonant and a vowel. Consequently, these T-F regions are likely to be more salient than others, and hence more probable to win the completion of the auditory attention during the bottom-up processing. Despite the similar fundamental mechanism and predictive performance, the proposed method presents a finer and more transparent modelling of speech intelligibility in noise than HEGP. As it quantifies...
the EM and IM effects in different components, modelling of each effect could be further improved and extended separately. There is some evidence suggesting that in English the glimpses taking place on vowels are more important to the intelligibility than those on consonants [46, 47], implying that the contribution of the glimpsed T-F regions could be further re-weighted for voicing and invoicing segments. In addition, a top-down auditory spotlight searching could be also considered in the metric for better modelling of IM.

5. CONCLUSIONS

An auditory saliency model was used in conjunction with a state-of-the-art OIM to improve the accuracy for intelligibility prediction under IM. The evaluation confirmed the validity of this approach, whose performance for the given dataset was comparable to the best reported in the literature. This study presents a detailed and yet physiologically-plausible approach for modelling both EM and IM to speech intelligibility. The proposed method could be thus used as a perceptual guide in audio production and reproduction, where speech intelligibility is a concern. However, the complexity of IM occurring at the later stage of the auditory pathway warrants investigations in future.

6. ACKNOWLEDGMENTS

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

7. REFERENCES


