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 shows 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, 7]) 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 peripheral not being further processed by the brain. 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

(task-independent) attention, as well as slower top-down cognitive (task-dependent) attention. First, the bottom-up processing occurs and attracts attention towards conspicuous or salient locations of the scene in an unconscious manner. Then, the top-down processing shifts the attention voluntarily towards locations of cognitive interest. Only the information selectively attended to is allowed to progress through the cortical hierarchy for high-level processing and detailed analysis [20, 21]. Therefore, saliency detection is considered to be a key attentional mechanism used to economically allocate and efficiently use the brain's limited processing capacity [22, 23].

The saliency of an object is the state or quality by which it stands out relative to its neighbours or background. In a complex auditory scene, a salient sound object may stand a bigger chance relative to other competing sources to gain a listener's attention. Saliency-based approaches were initially proposed as a major component in modelling bottom-up visual attention [24, 25, 26, 18]. The way in which the auditory cortex responds to sound stimuli is similar in terms of feature analyses on spectral or temporal modulation for instance [27, 28, 29, 30]. Many studies (e.g. [31, 32, 33, 34, 35]) on auditory saliency adopt the same analytical feature extraction mechanism to model auditory attention. The features used mainly include intensity, temporal contrast, spectral contrast and orientation which simulates the dynamics of the auditory neuron responses to moving ripples [36, 37]. In general, the modelling of auditory attention closely resembles that of visual attention, in which features essentially approximate the receptive field sensitivity profile of orientation-selective neurons in the primary visual cortex [38].

The output of the saliency analysis is usually a spectro-temporal representation called a saliency map. Kayser et al. generated the saliency map from intensity, temporal and spectral contrasts using a standard Fourier analysis [31]. By comparing the model prediction to the results from behavioural studies on human listeners and macaque monkeys, it was confirmed that different primate sensory systems rely on common principles for extracting relevant sensory events. In more recent studies [32, 33, 34, 35], the features used for composing the saliency map were extracted from the output of auditory peripheral analysis instead of via Fourier. This in principle provided a more physiologically-valid representation for the saliency analysis. Besides the same features used in [31], Kalinli and Narayanan included the orientation information in the saliency map [32]. A saliency score, which was a function of time, was further computed by collapsing the saliency map across frequencies followed by normalisation. This was used to predict the 'prominent' syllables and words in sentences drawn from a speech corpus. Their model achieved a better accuracy than when orientation information was excluded. However, further adding pitch information did not improve the model accuracy. Some other features were also used for generating a saliency map. Kaya and Elhilali added temporal envelope, rate and bandwidth as features to further emphasise the impact of the spectro-temporal modulations [35].

As both contemporary auditory saliency and glimpse analyses are performed on T-F representations, it is therefore possible to use a common representation at the early stage of the models for the purposes. This study aims to integrate saliency analysis into the ext. GP measure, in an attempt of quantifying the IM effect in a physiologically-plausible approach. The performance of the proposed method were evaluated along with another three reference OIMs, by comparing the model predictions to measured subjective intelligibility in noise maskers, some of which introduce IM.

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 SM, 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[1]. The centre frequencies of the 64 filters are evenly distributed on the equivalent rectangle 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 (Δ=3 dB), and be above the hearing level (HL, set to 25 dB),

$$G(t,f) = STEP_s(t,f) > max(STEP_n(t,f) + \Delta, HL) \quad (1)$$

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 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.

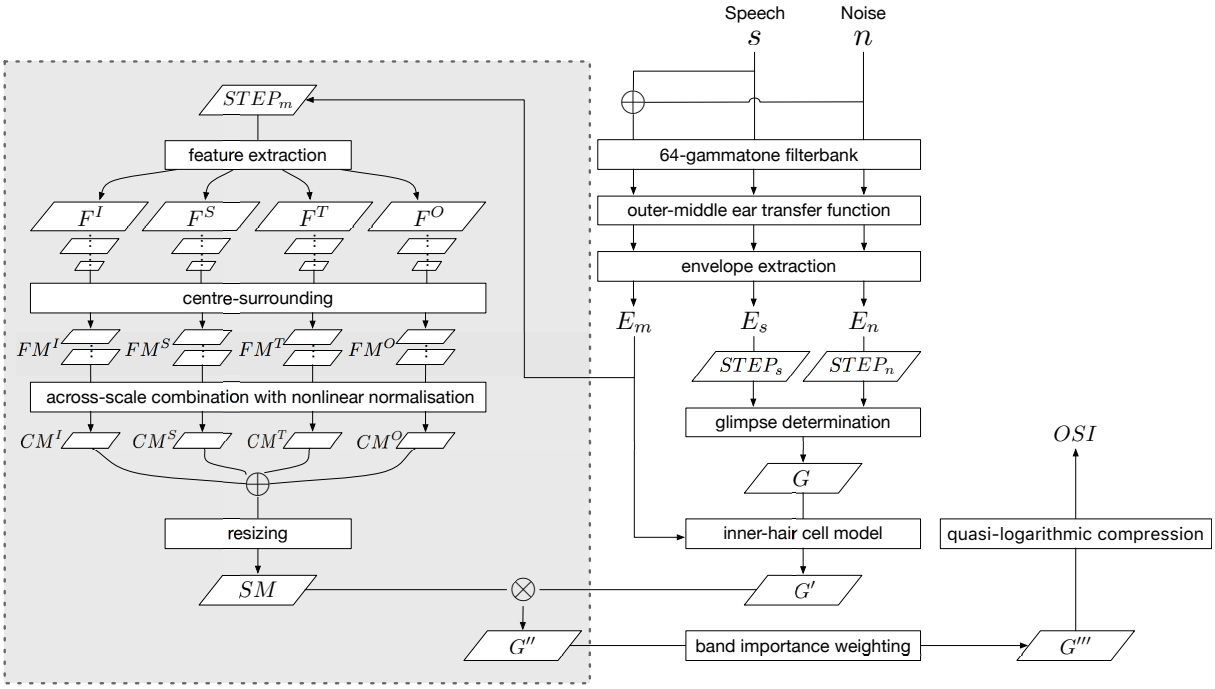

Figure 1: *Diagram of the proposed system. The shaded part on the left performs saliency analysis, partly accounting for the effect of informational masking. The unshaded part on the right describes the mechanism of the extended GP [11].*

malisation, combination and resizing. After [42, 32, 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 [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 θ, which represents one of the four target features, as listed in Table 1:

Table 1: *Parameters of the Gabor filters for each feature*

Feature	θ	α
Intensity	$\pi/2$	0
Spectral contrast	0	1
Temporal contrast	$\pi/2$	1
Orientation	$\{\pi/4, 3\pi/4\}$	1

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:

$$FM(c, s) = |FM(c) - FM(s)| \tag{3}$$

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 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 6×5 feature maps[2], 6 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 [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:

$$CM \longleftarrow |FM + FM * DoG - 0.02| \geq 0 \tag{4}$$

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 recovery the map size (currently at scale 4) back to the original size (at scale 1, same to the $STEP_m$).

[2]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.

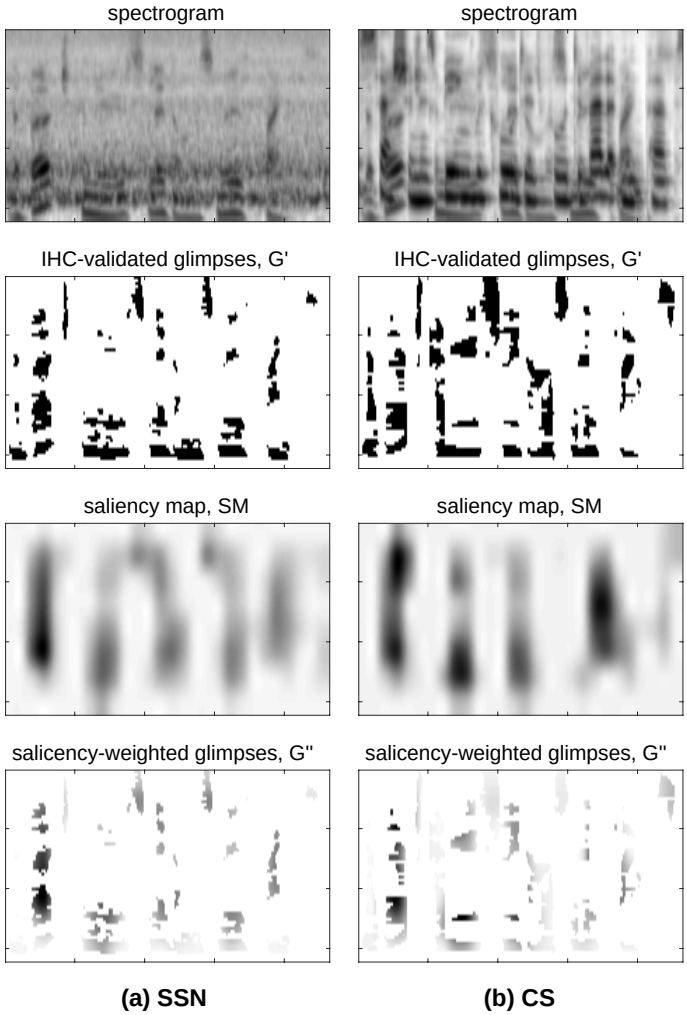

Figure 2: *Spectrograms, IHC-validated glimpses (G'), saliency maps (SM) and saliency-weighted glimpses (G'') of the sentence 'the birch canoe slid on the smooth planks' in SSN (left column) and CS (right column) at 1 and -7 dB SNR, respectively.*

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 [5] , followed by a quasi-logarithmic compression in a form of

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

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

where F=64 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.

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

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 2: *Values of parameters a and b used in the sigmoid transformation for the OIMs*

	proposed	ext. GP	HEGP	SII
a	-2.201	-2.864	-4.007	-1.009
b	8.284	5.837	8.024	5.339

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

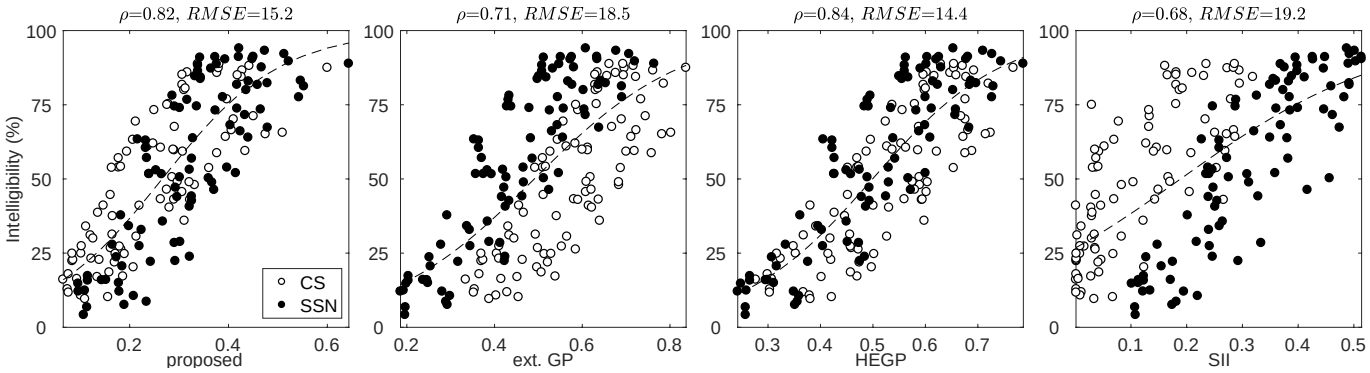

Figure 3: *Listener intelligibility versus model predictions in all 180 conditions. The dashed line in each plot is the sigmoid fitting for the OIM.*

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.*

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

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 ($\rho = 0.82$) over ext. GP ($\rho = 0.71$) [$Z = 3.216$, $p < 0.01$]. SII performance was similar to ext. GP ($\rho = 0.48$) [$Z = 0.781$, $p = 0.535$]. With such overall listener-model correlation, the proposed method performed as almost the best as reported in [12] ($\rho = 0.83$). The proposed is also comparable to HEGP [$Z = 0.970$, $p = 0.332$], despite the latter leading to the highest correlation ($\rho = 0.84$).

The performances of the OIMs were also examined in a series of sub-conditions, as displayed in Table 3. For individual maskers, all the OIMs achieved similar performance [all $\chi^2(3) \leq 3.923$, $p \geq 0.270$]. When making predictions separately for natural and synthetic speech, the proposed was equivalent to the HEGP [all $Z \leq 0.797$, $p \geq 0.426$], however was clearly more robust than the other two OIMs [all $Z \geq 2.927$, $p < 0.01$], especially for synthetic speech.

4. DISCUSSION

The current study aimed to improve the predictive power of the ext. GP metric [11] 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 [11], 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 overestimation 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 [12], which is a more dominant factor to speech intelligibility in noise than IM [45].

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 [48] 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

[1] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Computer Speech and Language*, vol. 49, pp. 1–16, 2018.

[2] Q. Liu, W. Wang, P. J. B. Jackson, and Y. Tang, "A Perceptually-Weighted Deep Neural Network for Monaural Speech Enhancement in Adverse Background noise," in *2017 European Signal Processing Conference*, Kos island, Greece, 2017.

[3] Y. Tang, B. M. Fazenda, and T. J. Cox, "Automatic speech-to-background ratio selection for maintaining speech intelligibility in broadcasts using an objective intelligibility metric," *Appl. Sci.*, vol. 8, no. 1, pp. 59, 2018.

[4] Inga Holube and Birger Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.*, vol. 100, pp. 1703–1716, 1996.

[5] ANSI S3.5, "ANSI S3.5-1997 Methods for the calculation of the Speech Intelligibility Index," 1997.

[6] IEC, ""Part 16: Objective rating of speech intelligibility by speech transmission index (4th edition)," in IEC 60268 Sound System Equipment (Int. Electrotech. Commiss., Geneva, Switzerland)," 2011.

[7] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[8] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[10] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Comm.*, vol. 52, no. 7-8, pp. 678–692, 2010.

[11] Y. Tang and M. Cooke, "Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions," in *Proc. Interspeech*, San Francisco, US, 2016, pp. 2488–2492.

[12] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Computer Speech and Language*, vol. 35, pp. 73–92, 2016.

[13] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5, pp. 303–304, 1995.

[14] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[15] G. Miller, "The masking of speech," *Psychol. Bull.*, vol. 44, pp. 105–129, 1947.

[16] I. Pollack, "Auditory informational masking," *J. Acoust. Soc. Am.*, vol. 57, no. S1, pp. S5–S5, 1975.

[17] G. Kidd Jr and H. S. Colburn, "Informational masking in speech recognition," in *The Auditory System at the Cocktail Party*, pp. 75–109. Springer, 2017.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[19] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sounds.*, The MIT Press, 1990.

[20] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Front. Biosci.*, vol. 55, pp. 202–212, 2000.

[21] S. Harding, M. Cooke, and P. Koenig, "Auditory gist perception: An alternative to attentional selection of auditory streams," in *WAPCV2007*, 2007.

[22] W. Schneider and R. M. Shiffrin, "Controlled and automatic human information processing: I. detection, search, and attention," *Psychological Review*, vol. 84, no. 1, pp. 1–66, 1977.

[23] R. M. Shiffrin and W. Schneider, "Controlled and automatic human information processing: II perceptual learning, automatic attending and a general theory," *Psychological Review*, vol. 84, no. 2, pp. 127–190, 1977.

[24] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[25] R. Milanese, S. Gil, and T. Pun, "Attentive Mechanisms for Dynamic and Static Scene Analysis," *Optical Eng.*, vol. 34, no. 8, pp. 2428–2434, 1995.

[26] S. Baluja and D. A. Pomerleau, "Expectation-Based Selective Attention for Visual Monitoring and Control of a Robot Vehicle," *Robotics and Autonomous Systems*, vol. 22, no. 3–4, pp. 329–344, 1997.

[27] J. P. Rauschecker, B. Tian, and M. Hauser, "Processing of complex sounds in the macaque nonprimary auditory cortex," *Science*, vol. 268, pp. 111–114, 1995.

[28] C. E. Schreiner, H. L. Read, and M. L. Sutter, "Modular organization of frequency integration in primary auditory cortex," *Rev. Neurosci.*, vol. 23, pp. 501–529, 2000.

[29] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," *J. Neurophysiol.*, vol. 87, pp. 516–527, 2002.

[30] S. Kaur, R. Lazar, and R. Metherate, "Intracortical pathways determine breadth of subthreshold frequency receptive fields in primary auditory cortex," *J. Neurophysiol.*, vol. 91, pp. 2551–2567, 2004.

[31] C. Kayser, C. I Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[32] O. Kalinli and S. S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech.," in *Proc. Interspeech*, 2007, pp. 194–1944.

[33] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *Proc. 20th International Congress on Acoustics (ICA 2010)*, 2010, pp. 1–8.

[34] T. Tsuchida and G. W. Cottrell, "Auditory saliency using natural statistics," in *Proc. Annual Meeting of the Cognitive Science (CogSci)*, 2012, pp. 1048–1053.

[35] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *46th Annual Conference on Information Sciences and Systems (CISS*, 2012.

[36] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 1998.

[37] S. Shamma, "On the role of space and time in auditory processing," *Trends in cognitive sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[38] A. G. Leventhal, *The Neural Basis of Visual Function: Vision and Visual Dysfunction, vol. 4.*, Boca Raton, Fla.: CRC Press, 1991.

[39] ISO 389-7:2006, "Acoustics – Reference Zero For The Calibration Of Audiometric Equipment – Part 7: Reference Threshold Of Hearing Under Free-field And Diffuse-field Listening Conditions," 2006.

[40] M. Cooke, *Modelling Auditory Processing and Organisation*, Cambridge University Press, 1993.

[41] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. 55, pp. 572–585, 2013.

[42] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[43] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electron. Imaging*, vol. 10, no. 161–169, pp. 1102–1116, 2001.

[44] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[45] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–1109, 2001.

[46] R. Cole, Y. Yan, B. Mak, and M. Fanty, "The contribution of consonants versus vowels to word recognition in fluent speech," *J. Acoust. Soc. Am.*, vol. 100, pp. 2689, 1996.

[47] D. Fogerty and D. Kewley-Port, "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *J. Acoust. Soc. Am.*, vol. 126, no. 2, pp. 847–857, 2009.

[48] S. Treue, "Directing the auditory spotlight," *Nature Neuroscience*, vol. 100, pp. 2689, 2006.