Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 82 (2016) 26–37

www.elsevier.com/locate/specom

Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms

Yan Tang*, Richard J. Hughes, Bruno M. Fazenda, Trevor J. Cox

Acoustics Research Centre, University of Salford, Salford M5 4WT, UK

Received 20 December 2015; received in revised form 5 April 2016; accepted 14 April 2016
Available online 28 May 2016

Abstract

A distortion-weighted glimpse proportion metric (BiDWGP) for predicting binaural speech intelligibility were evaluated in simulated anechoic and reverberant conditions, with and without a noise masker. The predictive performance of BiDWGP was compared to four reference binaural intelligibility metrics, which were extended from the Speech Intelligibility Index (SII) and the Speech Transmission Index (STI). In the anechoic sound field, BiDWGP demonstrated high accuracy in predicting binaural intelligibility for individual maskers ($\rho \geq 0.95$) and across maskers ($\rho \geq 0.94$). The reference metrics however performed less well in across-masker prediction ($0.54 \leq \rho \leq 0.86$) despite reasonable accuracy for individual maskers. In reverberant rooms, BiDWGP was more stable in all test conditions ($\rho \geq 0.87$) than the reference metrics, which showed different predictive patterns: the binaural STIs were more robust for the stationary than for the fluctuating noise masker, whilst the binaural SII displayed the opposite behaviour. The study shows that the new BiDWGP metric can provide similar or even more robust predictive power than the current standard metrics.

Keywords: Speech; Binaural; Objective intelligibility metric; Glimpsing; Noise; Reverberation.

1. Introduction

Objective intelligibility measures (OIMs) provide a fast and robust approach to estimating the intelligibility of speech. They have therefore been widely adopted in place of subjective tests for interim intelligibility evaluation in fields in which speech intelligibility is a concern – such as in telephony quality assessment (ANSI S3.5, 1997; Fletcher, 1921), audiology for hearing impairment (Holube and Kollmeier, 1996; Santos et al., 2013), acoustics design (Houtgast and Steeneken, 1985; IEC, 2011) and algorithm development for speech enhancement and modification (Gomez et al., 2012; Taal et al., 2010). As the majority of the OIMs estimate intelligibility based on purely *monaural* listening, their usability may be limited in more practical situations in which listeners hear *binaurally*. Therefore, an added advantage of developing binaural OIMs is that the effects of room acoustics (e.g. reverberation) on how listeners hear sounds in realistic environments may be more accurately taken into account.

Nearly all existing binaural intelligibility metrics (e.g. Andersen et al., 2015; Beutelmann et al., 2010; Jelfs et al., 2011; Schlesinger et al., 2010; van Wijngaarden and Drullman, 2008; Zurek, 1993) extend their monaural counterpart such as the Speech Intelligibility Index (SII, ANSI S3.5, 1997), the Speech Transmission Index (STI, IEC, 2011) and the short-time objective intelligibility measure (Taal et al., 2010), by taking the head shadow effect and binaural interaction into account. As this study demonstrates, existing binaural metrics do not work reasonably well in all test conditions. More recently, Tang et al. (2015) proposed a method for predicting binaural speech intelligibility by extending the distortion-weighted glimpse proportion (DWGP, Tang, 2014). Originally developed as a monaural method, the DWGP metric provides an objective assessment of speech modification algorithms that aim to boost speech intelligibility in noise. In the binaural version of the DWGP metric (BiDWGP), the better ear effect resulted from the head-shadow effect is

* Corresponding author.
E-mail address: y.tang@salford.ac.uk (Y. Tang).

http://dx.doi.org/10.1016/j.specom.2016.04.003

modelled with better-ear glimpses, which are essentially the time-frequency regions of speech with energy exceeding the noise by a certain threshold. The binaural interaction is quantified using the binaural masking level difference (BMLD, Levitt and Rabiner, 1967). In Tang et al. (2015), evaluation using subjective listening tests in a simulated anechoic sound field demonstrated that the intelligibility predicted by BiDWGP is highly correlated with listener performance in a word identification task in both a single stationary or fluctuating noise masker (Pearson correlation coefficients $\rho_p = 0.98$), and two or three of the same types of masker ($\rho_p \geq 0.94$).

The monaural DWGP metric incorporates a distortion weighting factor with the glimpse proportion metric (GP, Cooke, 2006; Tang, 2014). This weighting factor was initially introduced in Tang (2014) to increase the consistency of predictions by the GP metric across different noise maskers, especially between stationary (e.g. speech-shaped noise) and fluctuating (e.g. single-talker competing speech) maskers (Tang et al., 2016). The calculation of the distortion weighting factor was inspired by a STI-based metric, the normalise-covariance metric (Holube and Kollmeier, 1996), which uses the cross-correlation coefficient of the reference clean and noise-corrupted speech envelopes within each frequency band to determine the speech-to-distortion level. The DWGP metric adopts this approach and uses the cross-correlation coefficient directly to weight the number of the glimpses in a frequency band. This enables DWGP to take into account the impact of the masker on the speech envelope, in addition to the masked-audibility that is accounted for by the original idea of glimpse detection. STI metrics are reported to perform well when predicting speech intelligibility in reverberation (e.g. Houtgast and Steeneken, 1985; Houtgast et al., 1980; Plomp et al., 1980). Thus, it may be hypothesised that the BiDWGP metric may also preserve its predictive power for reverberation conditions, as it contains a STI-inspired component which operates in the modulation domain. However, despite its accurate predictions in anechoic conditions (Tang et al., 2015), the performance of BiDWGP in reverberant conditions has never been assessed. Therefore, the main aim of this study is to explore whether and how well the BiDWGP metric can predict intelligibility in reverberation.

In Section 2 of this paper the BiDWGP metric and four other reference intelligibility metrics with their binaural extensions are introduced. To evaluate their performance, the model predictions are compared with subjective data obtained from two listening experiments conducted in a simulated anechoic sound field and three rooms varying in size and reverberation time, both with and without a noise masker (Section 3). In addition to speech-shaped noise, which has the long term average spectrum of the chosen speech corpus and is widely used in evaluations of objective intelligibility metrics, competing speech uttered by a female speaker was also tested as a masker. Predicting intelligibility in the presence of competing speech is challenging due to the large temporal fluctuations present in the competing speech, and the possibility of it introducing informational masking – thus, compared to speech-shaped noise, it is used less often as a masker in relevant studies. As listening to speech in the presence of other talkers is a common realistic scenario, examining the performance of predictors for competing speech maskers has practical implications. Section 4 focuses on discussing the aspects that affect the performance of the BiDWGP metric; its limitations and further work are also explored. Finally, we draw conclusions from the study in Section 5.

2. The distortion-weighted glimpse proportion metric and reference metrics

In this section a technical overview of the BiDWGP metric will be presented first, followed by introductions to four state-of-the-art metrics with their binaural extensions, including the binaural Speech Intelligibility Index, and the three binaural Speech Transmission Index metrics with different implementations. As each metric may take different inputs for analysis, for the sake of clarity, six variables are defined first, which will be further referred to in this section:

- s, s': clean speech in anechoic and reverberant conditions
- n, n': noise masker in anechoic and reverberant conditions
- m, m': noise-corrupted speech (i.e. speech+noise mixture) in anechoic and reverberant conditions

2.1. An overview of the distortion-weighted glimpse proportion metric (BiDWGP)

For anechoic conditions, Zurek (1993) suggested a method to estimate the effective binaural signals from a single channel signal using a free-field to eardrum transformation of the sound pressure level (Shaw and Vaillancourt, 1985). With this approach as the first stage of the BiDWGP metric, Tang et al. (2015) demonstrated that BiDWGP can predict binaural intelligibility well from just a set of single channel signals (s, n and m), provided that the azimuth angle and distances for speech and masker sources relative to the listener are known. Further analyses have confirmed that intelligibility predictions by BiDWGP using single channel signals with the location information, and direct binaural signals are highly consistent ($\rho_p = 0.998$, and the Euclidean distance of 0.091 for indices falling between 0 and 1). However, the estimated binaural signals do not carry room acoustic information, and therefore cannot be used by the metric to account for the effects of room acoustics such as reverberation. In the current study, we assume binaural signals are available such that the estimation stage in Tang et al. (2015) is unnecessary.

The BiDWGP consists of two main components that account for the factors that negatively affect intelligibility: (1) masked-audibility due to energetic masking, and (2) distortion of the speech envelope due to temporal fluctuation and smearing.

2.1.1. Binaural glimpse detection

Glimpse detection quantifies the regions of the speech with a local speech-to-noise ratio (SNR) above certain threshold *LT* in dB (Cooke, 2006); it is intended to reflect the local audibility of speech in noise. The binaural advantage measured as the binaural masking level difference *BMLD* is included at this stage by applying the gain to the glimpse definition. Implementation involves generation of the spectro-temporal excitation patterns (STEP) *S*, N' and M' for *s*, n' and m' from the outputs of 34 gammatone filters with centre frequencies in the range of 100–7500 Hz. Since larger number of filters results in similar model performance, the choice of 34 filters is for computational efficiency (Tang, 2014). The output of each filter is applied to a frequency-dependent gain interpolated from ISO 389-7 (2006), accounting for the hearing threshold. The Hilbert envelope of each filter output is then extracted and smoothed using a leaky integrator with a 8 ms time constant (Moore et al., 1988), followed by downsampling to 100 Hz. The glimpse G_f at frequency f is defined as,

$$G_f = (S_f(t) > HL) \wedge (S_f(t) + BMLD_f > N_f(t) + LT) \quad (1)$$

where *S* and *N* are logarithmically compressed into decibel and *HL* is a constant hearing level (set at 25 dB HL). The value of *LT* here is set to 0 dB, which is different from that of 3 dB used in Tang et al. (2015). This choice enables BiDWGP to operate in reverberant conditions in which no masking sources are present, i.e. using the reverberant clean speech s' instead of n' and m' to account for the effect of reverberation only. The original 3 dB *LT* leads to significantly lower scores in detecting glimpses when comparing *s* to s'. Nevertheless, a 0 dB *LT* does not significantly affect the model performance in noisy conditions. This is further studied and discussed in Section 4.

BMLD is computed for each frequency f here using an approach described in Culling et al. (2004, 2005) as,

$$BMLD_f = 10\log_{10}\left[\frac{k - \cos(\phi_f^s - \phi_f^n)}{k - \rho_f}\right] \quad (2)$$

where

$$k = (1 + 0.25^2)\exp\left((2\pi f)^2 \cdot 0.000105^2\right)$$

and ϕ_f^s and ϕ_f^n denote the interaural phase shifts of the speech and masker at this frequency. ρ_f is the interaural coherence of the noise masker, defined as the maximum value of the interaural cross-correlation at frequency *f*. To model the better-ear effect, G_f is computed separately for the left ear as G_f^L and the right ear G_f^R. The binaural glimpses G_f^{bi} are essentially all time-frequency regions where either or both individual ears produce a glimpse, defined as,

$$G_f^{bi} = G_f^L \vee G_f^R \quad (3)$$

2.1.2. Binaural distortion weighting

To account for disturbance due to masker to speech envelope, the number of glimpses in each frequency band f is weighted by a distortion factor W_f, which is the cross-correlation coefficient between the uncompressed STEP of *S* and M',

$$W_f = \frac{\sum_{t=1}^{T}(S_f(t) - \bar{S}_f)\cdot(M'_f(t) - \bar{M}'_f)}{\sqrt{\sum_{t=1}^{T}(S_f(t) - \bar{S}_f)^2 \cdot \sum_{t=1}^{T}(M'_f(t) - \bar{M}'_f)^2}} \quad (4)$$

where T is the number of time frames. $\bar{S}_f$ and $\bar{M}_f$ represent across-time means of $S_f(t)$ and $M_f(t)$. W_f is also calculated for each ear separately; the binaural distortion weighting W_f^{bi} is the mean W_f across the two ears for each frequency band.

The final BiDWGP score is a sum of the glimpse proportion in each frequency band weighted by the distortion weighting W_f^{bi} and a band importance function (BIF) K_f interpolated from the values provided in Table 3 of ANSI S3.5 (1997). A quasi-logarithmic function v is then applied which models that ceiling intelligibility occurs for glimpse proportions substantially lower than unity:

$$\text{BiDWGP} = v\left[\frac{1}{T}\sum_{f=1}^{34}(K_f W_f^{bi}\sum_{t=1}^{T}\mathcal{H}(G_f^{bi}))\right] \quad (5)$$

where

$$\sum_{f=1}^{34} K_f = 1$$

and

$$v(x) = \frac{\log(1 + x/\delta)}{\log(1 + 1/\delta)},\ \delta = 0.01$$

$\mathcal{H}(.)$ is the Heaviside unit step function which counts the time frames meeting the glimpsing criterion G_f in channel *f*.

2.2. The state-of-the-art objective metrics with binaural extensions

2.2.1. The binaural Speech Intelligibility Index (BiSII)

Zurek (1993) revised the standard intelligibility measure – the Speech Intelligibility Index (SII ANSI S3.5, 1997) – to enable binaural intelligibility predictions for a single masker. The frequency-dependent SNR at each ear is calculated following the standard procedure. The better ear effect is then accounted for by taking the maximal SNR between the left SNR_f^L and right ear SNR_f^R. With an additional frequency-dependent binaural interaction gain $BMLD_f$, the effective binaural SNR at frequency band f is defined as,

$$SNR_f^{bi} = max(SNR_f^L, SNR_f^R) + BMLD_f \quad (6)$$

where the result is converted to the binaural articulation index AI_f^{bi} after being limited to ± 15 dB,

$$AI_f^{bi} = \frac{min(15, max(-15, SNR_f^{bi})) + 15}{30} \quad (7)$$

The final binaural SII calculation is after (ANSI S3.5, 1997):

$$BiSII = \sum_{f=1}^{18} K_f \cdot AI_f^{bi} \quad (8)$$

In this study, BiSII is calculated from anechoic clean speech s and reverberant noise n' using 18 1/3-Octave bands centred at from 160 to 8000 Hz. The corresponding BIFs are read from Table 3 of ANSI S3.5 (1997). Note that with reverberation only (noise-free), the second input n' is replaced by the reverberant speech s'.

It is worth noting that there are two differences in terms of the BiSII implementation from the original approach described in Zurek (1993). First, the input signals are binaural in this study, and consequently the true binaural SNRs are calculated from the directly obtained binaural signals instead of from estimated inputs as in the original study. Second, the BMLD in Zurek (1993) is estimated using an approach based on the estimated interaural time difference (Colburn, 1977), which is a function of the azimuths of the sources relative to the listener. Since we use binaural signals as model inputs here, the BMLD therefore can be readily computed using Eq. (2). As all the calculations are based on the true binaural signal, the model accuracy may be expected to be somewhat better than the original approach in Zurek (1993) (see further Section 3.1.2), which has already shown good predictive power for stationary noise ($\rho_p = 0.92$) and fluctuating maskers ($\rho_p = 0.89$) as studied in Tang et al. (2015).

2.2.2. The binaural Speech Transmission Index (BiSTI)

Unlike the SII, which predicts intelligibility in noise by quantifying masked audibility in frequencies, the STI measures the reduction of the temporal modulation of speech, which has been found to be important and correlated with speech intelligibility (Houtgast and Steeneken, 1973, 1985). More importantly, STI has been more directly applied to the measurement of speech intelligibility degradation caused by room effects such as reverberation. The reduction of modulation in each frequency is calculated as the modulation transfer function (MTF). Further procedures are then taken to convert the MTF to a final intelligibility index (e.g. Goldsworthy and Greenberg, 2004; IEC, 2011). Many approaches (e.g. Drullman et al., 1994; Holube and Kollmeier, 1996; Payton and Shrestha, 2013) have been proposed to compute the MTF other than the original (Houtgast and Steeneken, 1973; Steeneken and Houtgast, 1980), from using artificial test signals to using real time running speech directly.

In order to predict intelligibility with binaural listening using the traditional STI, van Wijngaarden and Drullman (2008) introduced a general extension to account for binaural interaction based on an interaural correlogram. As binaural interaction is most prominent between 500 and 1500 Hz, the extension is only applied for frequency bands with central frequencies falling within this range. van Wijngaarden and Drullman (2008) further included a 2 kHz band, although the impact to the final performance is subtle. The brief implementation is as follows: signals are resampled at 23 kHz, so that the analysis can be done on seven Octave bands centred at 125, 500, 1000, 2000, 4000 and 8000 Hz. The envelope of each band is then extracted by squaring the output of the filter followed by low-pass filtering with a cut-off frequency of 50 Hz. For the three bands of 500, 1000 and 2000 Hz, the envelopes of the left and right ears are segmented into 30-ms time frames without overlap. For each frame, an interaural correlogram is generated using the cross-correlation between the two ear signals with any offset removed (ensuring that the smallest value is 0). Only the part where the interaural delay is less than 2 ms on the correlogram is kept for further interaural MTF calculation. The frame-based correlograms are generated for both anechoic clean speech s and reverberant noise-corrupted speech m' signals. The MTFs are then calculated for each frame within an interaural time delay from −0.8 to 0.8 ms (Schlesinger et al., 2010); the largest value is chosen as the MTF of this frame. The final MTF value for this frequency band f is the average across all the frames in that band. This extension theoretically can be applied to any STI-based metrics in which the MTF is calculated using different approaches mentioned above. For all other four bands under 500 Hz and over 2 kHz, the MTFs are calculated as in the monaural version for both the left and right ears. The better-ear apparent SNR for frequency f is then calculated from the larger MTF between the two ears as,

$$SNR_f^{bi} = 10\log_{10}\left(\frac{MTF_f}{1 - MTF_f}\right) \tag{9}$$

The apparent SNR is then converted to the binaural transmission index TI_f^{bi} using Eq. (7). Finally, the overall STI is calculated by summing up the weighted TIs across all frequencies, further taking upwards spread of masking into account,

$$BiSTI = \sum_{f=1}^{7} \alpha_f \cdot TI_f^{bi} - \sum_{f=1}^{6} \beta_f \sqrt{TI_f^{bi} \times TI_{f+1}^{bi}} \tag{10}$$

where α and β are the STI weighting and redundancy factors specified in Table A.3 of IEC (2011), respectively.

Within the same framework of binaural STI, two different approaches were chosen in this study to calculate MTF. The first one is a phase-locked MTF introduced in Drullman et al. (1994), but with revised normalisation term k proposed by Goldsworthy and Greenberg (2004), defined as,

$$k_f = \frac{\bar{S}_f}{\bar{S}_f + \bar{N}'_f} \tag{11}$$

where $\bar{S}_f$ and $\bar{N}_f$ denote the mean intensities of anechoic clean speech envelope $S_f(t)$ and the estimated reverberant noise envelope $N_f(t)$ at frequency band f. Given that the STI normally takes anechoic clean speech s and reverberant noise-corrupted speech m' as inputs, the reverberant noise envelope $N'_f(t)$ is estimated by

$$N'_f(t) = |M'_f(t) - S_f(t)| \tag{12}$$

where $M'_f(t)$ is the envelope of m'. By defining k this way, normalisation singularities resulting from reduction in the overall amplitude of the envelope of the received signal during processing may be avoided (Goldsworthy and Greenberg, 2004). For each frequency band f, the $MTF_f(i)$ is calculated for 14 one-third octave modulation frequencies covering from 0.63 to 12.7 Hz. The mean of the 14 MTFs is then taken as

Table 1
Summary of input signals required, number and type of analysis filters, measurement for the effects of noise and reverberation, and modelling of binaural listening in the five binaural intelligibility metrics introduced in Section 2.

	Input	Analysis filters	Effects of noise and reverberation	Better ear	Binaural interaction
BiDWGP	s, n', m'	34 gammatone	Short-term $SNR_f(t)$ between s and n'; long-term W_f between envelopes of s and m'	$G_f^L \vee G_f^R$	$BMLD_f$
BiSII	s, n'	18 1/3-Octave	Long-term SNR_f (ANSI S3.5, 1997)	$max(SNR_f^L, SNR_f^R)$	$BMLD_f$
BiNCM	s, m'	18 1/3-Octave	Long-term r_f between envelopes of s and m' (Holube and Kollmeier, 1996)	$max(r_f^L, r_f^R)$	Interaural correlogram-based r_f for seven bands between 500 and 2000 Hz
BiSTI1	s, m'	7 Octave	Long-term MTF_f, using the real cross-power spectrum method (Drullman et al., 1994)	$max(MTF_f^L, MTF_f^R)$	Interaural correlogram-based MTF_f for 500, 1000 and 2000 Hz
BiSTI2	s, m'	7 Octave	Long-term MTF_f, using the envelope regression method (Goldsworthy and Greenberg, 2004)	$max(MTF_f^L, MTF_f^R)$	Interaural correlogram-based MTF_f for 500, 1000 and 2000 Hz

the MTF for this band,

$$MTF_f = \frac{1}{14}\sum_{i=1}^{14}\left(k_f \cdot Re\left\{\frac{P_{SM}(i)}{P_{SS}(i)}\right\}\right) \tag{13}$$

where $Re\{\cdot\}$ indicates taking the real part of the complex numbers. P_{SS} and P_{SM} are the power spectra of S, and the cross-power spectrum of S and M, respectively.

The second approach is the Envelope Regression method described in Goldsworthy and Greenberg (2004). The MTF of frequency f is calculated directly from the intensity envelopes as,

$$MTF_f = k_f \cdot \frac{\frac{1}{T}\sum_{t=1}^{T}(S_f(t)\cdot M_f(t)) - \bar{S}_f \cdot \bar{M}_f}{\frac{1}{T}\sum_{t=1}^{T} S_f(t)^2 - \bar{S}_f^2} \tag{14}$$

where $\bar{M}_f$ is the mean intensity envelope of M_f and T is the number of samples.

2.2.3. The binaural normalised covariance metric (BiNCM)

Another variant of the STI, the normalised covariance metric (NCM), was initially proposed to estimate speech intelligibility for hearing-impaired listeners (Holube and Kollmeier, 1996). Similar to the SII calculation (Section 2.2.1), the analysis here is performed on 18 1/3-Octave bands. Instead of the MTF, NCM measures the distortion on the speech envelope in each band caused by masker and reverberation using the cross-correlation coefficient r_f between S_f and M'_f, which is computed using Eq. (4).

For predicting binaural intelligibility, NCM was here extended with the same procedures applied to the STI metrics as described in Section 2.2.2. The frequency-dependent binaural apparent SNR can be computed using the larger r_f from the two ears, which is defined as,

$$SNR_f^{bi} = 10\log_{10}\left(\frac{r_f^2}{1-r_f^2}\right) \tag{15}$$

After converting SNR to a transmission index (Eq. (7)), unlike the conventional STI calculation (Eq. (10)), NCM is calculated using the SII approach formulated in Eq. (8).

Table 1 lists the input signals required by each metric and the number and type of filters on which the analysis is performed. It further summarises the measurement used by each metric to account for the effects of noise and reverberation, as well as the modelling of the better-ear effect and binaural interaction in binaural listening. For all the five metrics, the output representing predicted intelligibility is a number falling between 0 and 1, with larger values indicating better intelligibility.

3. Evaluation

The performance of all the metrics introduced above are evaluated using the Pearson correlation (ρ_p) and Spearman's rank correlation coefficients (ρ_s) between measured listener performance and model predictions, along with the error of the standard deviation (σ_e) for each type of correlation, defined as,

$$\sigma_e = \sigma_d\sqrt{1-\rho^2} \tag{16}$$

where σ_d is the standard deviation of the subjective scores in a given condition. While the Pearson correlation reflects the linear relationship between the measured and predicted intelligibility, the Spearman correlation assesses the ranking capacity of the model prediction with respect to the measured intelligibility. Before computing correlation coefficients, listener performance is arcsine-transformed into rationalised arcsine units (RAU, Studebaker, 1985), in order to enable more accurate linear tests on the subjective data which may not be strictly Gaussian when listener performance is close to 0 or 1.

3.1. Metric performance in simulated anechoic sound field

3.1.1. Subjective intelligibility

The subjective intelligibility data in anechoic conditions was reported in Tang et al. (2015). Within simulated anechoic conditions over headphones, fourteen native British English speakers (mean 30.0 years, s.d. 4.9 years) with normal

Table 2
Listener-model Pearson (ρ_p) and Spearman (ρ_s) correlation coefficients (with σ_e in parentheses) for the five evaluated binaural OIMs in individual masker and overall conditions in an anechoic environment. The number following the condition name indicates the number of the data points in each condition. Darker and lighter grey codings highlight the highest ρ_p and ρ_s, respectively for each condition. For all ρ, $p < .001$.

	SSN [36]		CS [36]		overall [72]	
	ρ_p	ρ_s	ρ_p	ρ_s	ρ_p	ρ_s
BiDWGP	0.96 (0.09)	0.97 (0.08)	0.96 (0.06)	0.95 (0.06)	0.94 (0.09)	0.96 (0.07)
BiSII	0.93 (0.12)	0.96 (0.10)	0.91 (0.08)	0.87 (0.10)	0.74 (0.19)	0.80 (0.16)
BiNCM	0.93 (0.13)	0.96 (0.09)	0.84 (0.11)	0.87 (0.10)	0.53 (0.23)	0.60 (0.22)
BiSTI1	0.89 (0.15)	0.93 (0.12)	0.88 (0.10)	0.92 (0.08)	0.79 (0.17)	0.86 (0.14)
BiSTI2	0.84 (0.18)	0.91 (0.14)	0.80 (0.12)	0.86 (0.10)	0.74 (0.18)	0.83 (0.15)

hearing identified keywords from the Harvard sentences (e.g. 'the **birch canoe slid** on the **smooth planks**', Rothauser et al., 1969) uttered by a British male talker. 216 non-repetitive sentences were mixed with speech-shaped noise (SSN) or female competing speech (CS) at two SNR levels: −9 and −6 dB for SSN and −18 and −15 dB for CS. While the target speech source was always fixed straight ahead (i.e. $\theta_s = 0^\circ$) of the listener, the azimuth of the masker θ_n relative to the listener varied across conditions. For different source-listener distances r_s and r_n for speech and masker respectively, the locations of the masker were:

- $r_s = r_n = 2m$: $\theta_n \in$[0 −10 20 −30 60 −90 90 −150 120 180]°.
- $r_s = 1.5m$, $r_n = 2.5m$: $\theta_n \in$[0 −45 135 180]°.
- $r_s = 2.5m$, $r_n = 1.5m$: $\theta_n \in$[0 45 −135 180]°.

More experimental details are described in Tang et al. (2015). In total, this dataset consists of 72 conditions. The subjective intelligibility for each condition was taken as the mean keyword identification rate across all the listeners.

3.1.2. Results

Table 2 presents the performance of all the metrics for each type of masker and their overall performance across the entire dataset. In general, all metrics made better predictions for a stationary masker (SSN) than for a fluctuating masker (CS). Initial Chi-square tests on dependent correlations suggest that all metrics performed differently in all conditions measured by both the Pearson [$\chi(4)^2 > 29.777$, $p < .001$] and Spearman correlation coefficients [$\chi(4)^2 > 21.008$, $p < .001$]. Post-hoc statistical comparisons using Z tests were further performed. In terms of linear relationships, while BiDWGP, BiSII and BiNCM achieved similarly good Pearson correlations for SSN ($\rho_p \geq 0.93$) [$Z = 1.409$, $p = .159$], only BiDWGP and BiSII provided comparable results for CS ($\rho_p \geq 0.91$) [$Z = 1.858$, $p = .063$]. BiDWGP demonstrated significantly better predictive power when compared to the two STI metrics for each of the sub-conditions [$Z > 4.568$, $p < .001$]. All metrics showed good ranking capacity for SSN ($\rho_s > 0.90$). In CS, BiDWGP and STI1 maintained their high ranking capacity

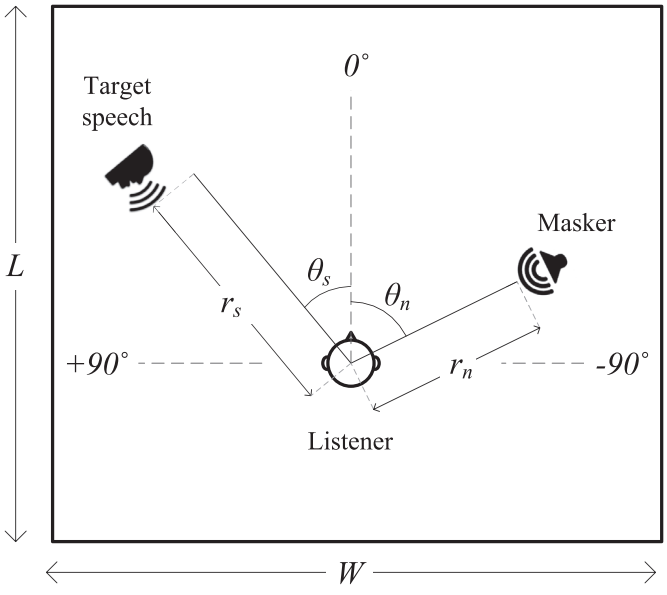

Fig. 1. 2-D layout in the simulated room. The target speech and noise masker are situated r_s and r_n metres away from the listener, with an azimuth of θ_s° and θ_n° off the straight ahead of the listener ($\theta = 0^\circ$), respectively. L and W indicate the length and width of the room in metre.

[$Z = 1.253$, $p = .210$], but a significant drop in performance was observed for the other metrics compared to BiDWGP [$Z = -4.210$, $p < .001$]. As an overall predictor, BiDWGP exhibited the best linear relationship [$Z = 9.461$, $p < .001$] and rank capacity [$Z = 8.071$, $p < .001$] of all the metrics tested.

3.2. Metric performance in virtual rooms

3.2.1. Simulation of rooms

In order to simulate more realistic listening conditions, three rooms varying in size and acoustic properties were modelled. A room model approach was used to allow both flexibility of the presented acoustic conditions (room size, surface absorption etc.) and source/receiver positions. Simulations were carried out using an Image Source Model (ISM)

Table 3
Experiment settings for rooms, speech and masker locations, and SNR levels. Room dimension is specified as length × width × height in metre.

Room spec.	Speech		Masker		SNR
	θ_s (°)	r_s (m)	θ_n (°)	r_n (m)	(dB)
$RT_{60} \approx 0.4$ s	0	2	CS: 0	2	−12, −9
	0	2	CS: 60	2	−12, −9
Dim.: 5.8 × 6.6 × 2.8	0	2	SSN: −30	2	−6, −3
	0	2	SSN: 90	2	−6, −3
	30	2	SSN: 60	2	−6, −3
	0	2	–	–	In quiet
$RT_{60} \approx 1.2$ s	0	2	CS: −90	3	−12, −9
	0	2	CS: −60	5	−12, −9
Dim.: 17.4 × 19.8 × 7.5	30	4	CS: 0	2	−12, −9
	−60	4	CS: −90	4	−12, −9
	0	2	SSN: 30	4	−6, −3
	0	2	SSN: 0	6	−6, −3
	−60	4	SSN: 0	2	−6, −3
	0	2	–	–	In quiet
	0	4	–	–	In quiet
$RT_{60} \approx 3.0$ s	0	8	CS: 0	12	−12, −9
	30	14	CS: 30	14	−12, −9
Dim.: 43.5 × 49.5 × 15.0	0	8	SSN: 0	8	−6, −3
	30	6	SSN: −60	20	−6, −3
	−60	20	SSN: 6	30	−6, −3
	0	2	–	–	In quiet
	−30	8	–	–	In quiet
	0	20	–	–	In quiet

(Allen and Berkley, 1979) for a simple box-shaped room, following the principles of geometric acoustics (Savioja and Svensson, 2015), which was implemented in the frequency domain (Peterson, 1986) and extended to produce a binaural output (Wendt et al., 2014). The approximate reverberation time (RT_{60}) for the three spaces, taken as the average value in the 250 Hz–4 kHz octave band range, were given as 0.4 s, 1.2 s and 3 s respectively. Fig. 1 illustrates the layout of the source and listener positions used within the three rooms, for which the details including the dimensions of each room are summarised in Table 3.

The Binaural Room Impulse Response (BRIR) signals used for simulating the reverberant rooms were generated as follows. For each image source, representing a discrete separate reflection, the attenuation due to spherical spreading and wall reflections were calculated, and both time and angle of arrival (azimuth and elevation) at the listener were obtained. The appropriate delay and amplitude alteration were then applied to each. For simplicity, the source was assumed to be omnidirectional and uniform surface absorption was chosen to give a more even decay and to allow simpler estimation of the presented RT_{60}. In order to generate a binaural output, a bank of Head Related Impulse Responses (HRIRs) (Jin et al., 2014) was used (subject 7, selected arbitrarily), which included elevation data to account for reflections arriving from outside the horizontal plane. Each image was then convolved with the HRIR in the direction corresponding to the angle of arrival. For angles of arrival between available HRIR data points, a frequency domain HRIR interpolation method was used (Hartung et al., 1999). The final BRIR was then obtained by summing the individual reflection contributions along with the direct sound component. The method was then repeated for each source-receiver pair and room type.

3.2.2. Material and maskers

As in Tang et al. (2015), different sentences were drawn from the same Harvard corpus as described in Section 3.1.1. The same two types of noise maskers (SSN and CS) were adopted to generate speech+noise mixture, with the SNR levels being readjusted. By considering that the effect of reverberation in addition to the masking effect may potentially increase task difficulty, a pilot test was conducted to find the appropriate SNR levels for this study. The results suggested $SNR = -6$ and -3 dB for SSN; $SNR = -12$ and -9 dB for CS, leading to listeners' keyword recognition rates being spread between 0% to 100% in reverberant noisy conditions. Note that all the reported SNR levels were measured in anechoic conditions when the speech and masker were co-located. In addition, we tested reverberation effects on intelligibility on their own without the effects of noise. Table 3 further displays all the settings in terms of speech and masker locations. In total, this design led to 40 conditions being tested.

3.2.3. Listeners and procedures

Ten native British English speakers (mean 32.6 years, s.d. 5.6 years) from the University of Salford participated in this experiment. They were all undergraduates, graduates or staff working in the Acoustics Research Centre. All participants reported normal hearing. Student participants were paid for their participation.

The binaural stimuli were generated by convolving the speech and noise samples with BRIRs generated as described in Section 3.2.1. The experiment took place in a semi-anechoic listening room with a background noise level of 3.8 dB(A). The speech+noise mixtures were presented to listeners over Sennheiser HD650 headphones after being pre-amplified by a Focusrite Scarlett 2i4 USB audio interface. The presentation level of speech over headphones was calibrated using an artificial ear and fixed to 63 dB(A); the noise level was then adjusted to meet the SNR requirements.

For each of the 40 conditions, listeners heard 5 different sentences, resulting in 200 sentences in total. No sentence was presented twice to the same listener. Sentences were blocked by masker/SNR combination in addition to a quiet condition (reverberation only). Listeners always listened to the quiet block first, then the other four noisy blocks in a random order. All the sentences within a block were also randomised. Therefore, each condition (data point) was heard 50 times (5 sentences × 10 listeners).

The task for listeners was to identify the keywords in each sentence. Listeners used a physical computer keyboard to record their response via a MATLAB programme. The entire experiment lasted about 45–60 minutes in one session. The Research Ethics Panel at the College of Science and Technology, University of Salford, granted ethical approval for the experiment reported in this paper.

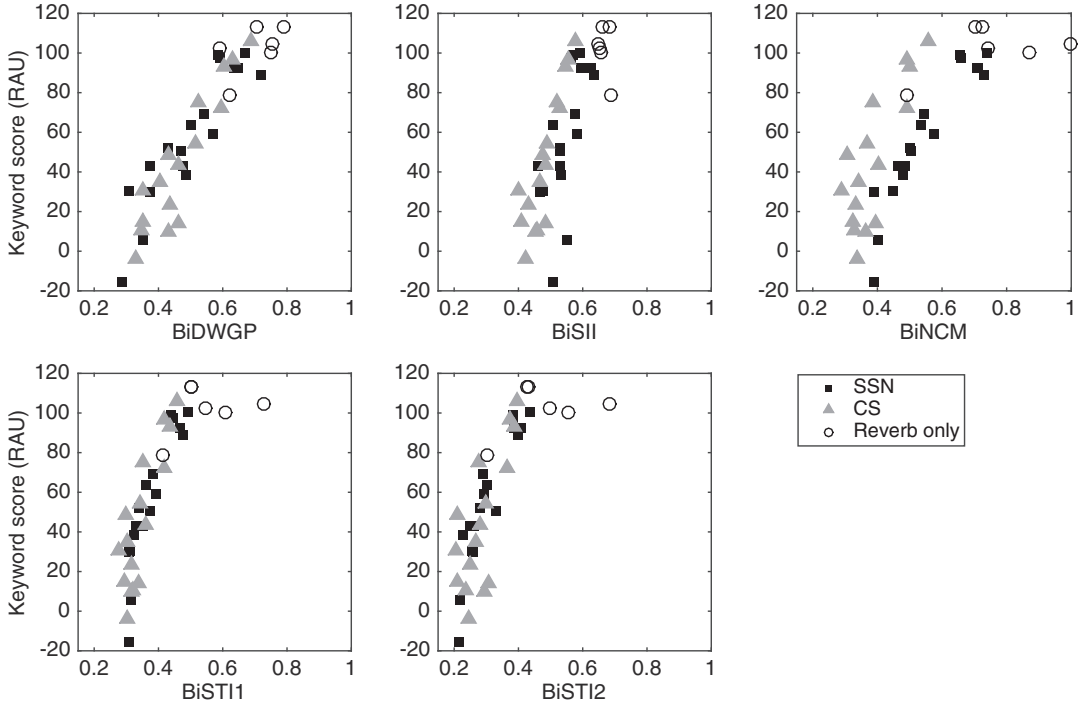

Fig. 2. Subjective vs predicted intelligibility in reverberant and noisy reverberant conditions, coded by masker type or reverberation.

3.2.4. Results

Fig. 2 displays the 40 data points representing listener-model correspondences for each metric for all reverberant and reverberant noisy conditions; the marker type distinguishes different sub-conditions. Similar to the anechoic conditions (Table 2), the evaluated metrics significantly varied in their performance in terms of a linear relationship [$\chi(4)^2 > 16.974$, $p < .01$] and in ranking capacity [$\chi(4)^2 > 10.473$, $p < .05$], for all conditions. Post-hoc Z tests confirm that overall, the four reference metrics show a similar linear relationship with listeners' performance, as indicated by the Pearson correlation coefficients ($0.79 \leq \rho_p \leq 0.81$) [$Z = .556$, $p = .578$]. However, the BiDWGP metric provides a closer match ($\rho_p = .92$) to the subjective data than do the four reference metrics [$Z = 4.067$, $p < .001$]. Nevertheless, the two STIs showed a similar rank capacity to BiDWGP [$Z = 1.071$, $p = .284$], despite the higher Spearman correlations of the latter ($\rho_s = 0.92$). The correlation coefficients for the overall performance are further detailed in the last two columns of Table 4.

The individual masker performance for each metric is listed in Table 4. For the stationary masker (SSN), all the metrics, except BiSII, demonstrated comparably good predictive power both in terms of a linear relationship with subjective intelligibility ($\rho_p \geq 0.91$) [$Z = .356$, $p = .722$] and the rank capacity ($\rho_p \geq 0.90$) [$Z = .525$, $p = .599$], with BiNCM exhibiting highest correlations. For the fluctuating masker (CS), BiDWGP, BiSII and BiSTI1 had a similar linear relationship with listeners' performance [$Z = 1.023$, $p = .307$]. The rank capacity of all the metrics, however, declined drastically in the presence of CS compared to that of SSN, particularly for the STI-based metrics. Interestingly, the results further reveal that while BiSII performed less well in SSN but considerably better in CS, the other three STI-based metrics demonstrated the opposite tendency. In contrast, the BiDWGP metric exhibited more balanced performance for the two noise maskers in this dataset. Due to the limited number of data points (six conditions), the performance of the metrics in the purely reverberant conditions was not assessed separately.

4. Discussion

4.1. Choosing local threshold for glimpse definition

As a free parameter in the BiDWGP metric, it is important to consider how the predictive performance is affected by the local threshold *LT* value used. A lower *LT* means more relaxed glimpse criteria (i.e. a lower local SNR), which allows for a greater number of time-frequency regions to be included for intelligibility prediction; a higher *LT* leads to fewer but more robust glimpses, which more securely ensure that the speech escapes from masking. Fig. 3 depicts the listener-model Pearson correlation coefficient ρ_p (left panel) and the Spearman correlation coefficient ρ_s (right panel) as a function of *LT*. The results are consistent to those found for the monaural DWGP metric in Tang (2014). Both measurements display similar patterns, although the Spearman correlation appears more sensitive to varying *LT* than the Pearson correlation. For SSN, better accuracy is achieved with a lower *LT*, as a high *LT* may eliminate too many glimpses that are required to make more accurate predictions. For CS, a higher *LT* is clearly favourable. This is likely because listeners may need more distinct glimpses to perform source separation for

Table 4
Listener-model Pearson (ρ_p) and Spearman (ρ_s) correlation coefficients (with σ_e in parentheses) for the five evaluated binaural OIMs in individual masker and overall conditions in simulated rooms. The number following the condition name indicates the number of the data points in each conditions. Darker and lighter grey codings highlight the largest ρ_p and ρ_s, respectively, for each condition. For all ρ, $p < .001$.

	SSN [18]		CS [16]		overall [40]	
	ρ_p	ρ_s	ρ_p	ρ_s	ρ_p	ρ_s
BiDWGP	0.91 (0.14)	0.90 (0.14)	0.92 (0.14)	0.87 (0.17)	0.92 (0.14)	0.92 (0.15)
BiSII	0.69 (0.24)	0.72 (0.23)	0.90 (0.15)	0.85 (0.18)	0.80 (0.22)	0.82 (0.21)
BiNCM	0.93 (0.12)	0.94 (0.11)	0.81 (0.20)	0.64 (0.26)	0.80 (0.22)	0.83 (0.20)
BiSTI1	0.91 (0.14)	0.93 (0.12)	0.85 (0.18)	0.67 (0.26)	0.81 (0.21)	0.90 (0.16)
BiSTI2	0.92 (0.13)	0.92 (0.13)	0.76 (0.23)	0.59 (0.28)	0.79 (0.22)	0.86 (0.19)

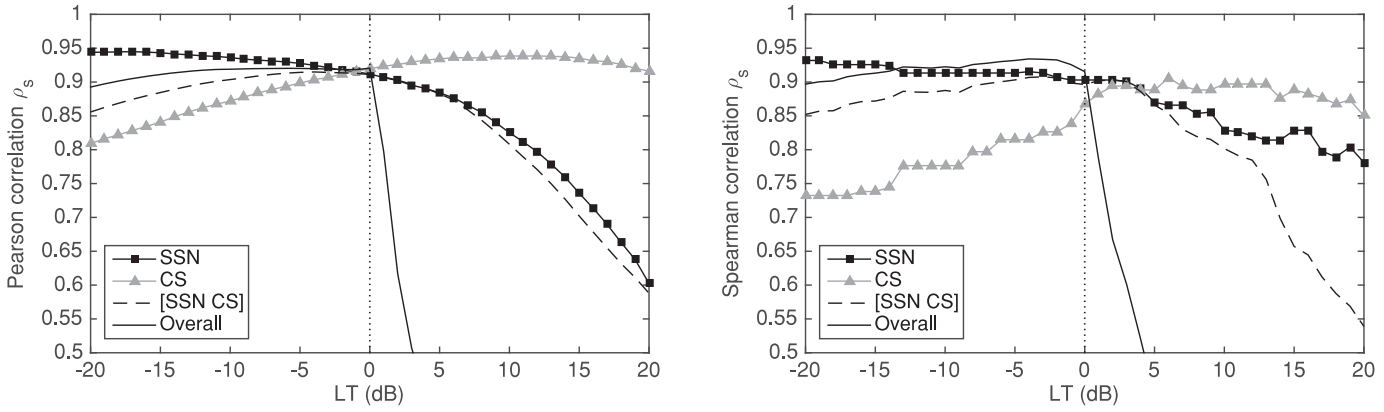

Fig. 3. Listener-model Pearson (ρ_p, left panel) and Spearman (ρ_s, right panel) correlation coefficient as a function of the local threshold (*LT*) for defining glimpses in sub-conditions: individual masker (SSN and CS), noisy reverberant conditions with the six reverberation-only conditions excluded ([SSN CS]) and all conditions with the six reverberation-only conditions included (Overall). The vertical dotted line indicates the chosen *LT* value in this study.

competing sources in high level auditory processing, which potentially introduces a large amount of informational masking (Brungart, 2001). Nevertheless, a value of *LT* falling into the range of −5 to 5 dB can broadly ensure reasonable model performance for each masker, as well as for across-masker predictions (black dash line in Fig. 3).

In order to make predictions for reverberation-only conditions, the largest value for *LT* ought to be 0 dB. Reverberation energy being added to the direct sound with delay may smear speech elements in the time domain, resulting in a certain level of masking to adjacent phonemes. The reverberation effect may also reduce the temporal resolution of speech. To quantify the masking effect due to reverberation, glimpses are detected in BiDWGP by comparing the anechoic clean speech to its reverberant version. Compared to the masking effect of noise masker, the level of the masking due to reverberation is low. When the reverberation time is short, it is almost equivalent to comparing the anechoic clean speech to itself. A value above 0 dB can therefore excessively reduce the number of the valid glimpses from the clean speech itself, hence a substantial drop in performance occurs, as illustrated by the black solid line in Fig. 3.

Table 5
Performance assessed as Pearson correlation coefficient ρ_p (with σ_e in parentheses) of solo components in the BiDWGP metric, and the complete metric. The tick and circle indicate components being included and excluded, respectively.

	Glimpsing *G*	Distortion *W*	SSN ρ_p	CS ρ_p	Overall ρ_p
	√	○	0.72 (0.23)	0.94 (0.12)	0.67 (0.27)
	○	√	0.95 (0.11)	0.79 (0.21)	0.84 (0.20)
BiDWGP	√	√	0.91 (0.14)	0.92 (0.14)	0.92 (0.14)

4.2. Role of the glimpsing and distortion components

The contribution of the glimpsing (Section 2.1.1) and distortion (Section 2.1.2) components, referred to as '*G*' and '*W*' in Table 5, was considered by comparing the correlations between the listener performance and the predictions made using each component in isolation. Since the two types of correlation show similar results, only the Pearson correlation coefficient ρ_p with the error of the standard deviation σ_e is presented here. As demonstrated in Table 5, an evident predictive tendency is observed: while the glimpsing component

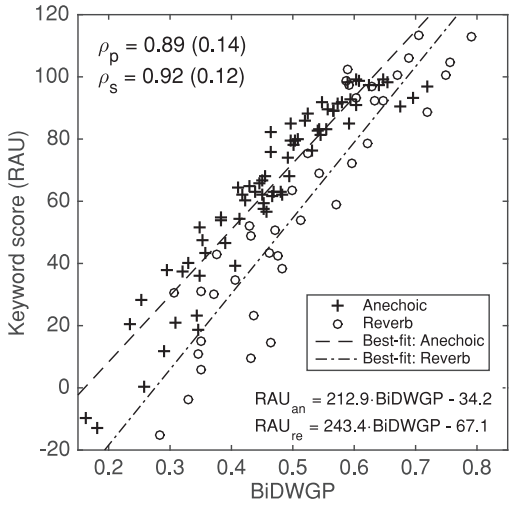

Fig. 4. Scatter plot of subjective-BiDWGP pairs in the simulated anechoic and room conditions. The overall Pearson (ρ_p) and Spearman (ρ_s) correlation coefficients are displayed (with σ_e in parentheses). The dashed and dashed-dotted lines show the best linear fitting for the anechoic and reverberant conditions respectively, with the equations provided.

alone can deal with CS very well, it performs dramatically worse in the face of SSN. Interestingly, the solo distortion component displays a totally opposite pattern. These findings are consistent with the observed tendency between BiSII and the STI-based metrics in Section 3.2.4, implying that neither of the two approaches alone can cope well with both types of maskers. For the glimpsing component, the poor performance in SSN may be due to the difficulty in dealing with the impact of reverberation (e.g. smeared speech envelope) (Rennies et al., 2011), especially when the RT_{60} is large. For the distortion component, calculating distortion based on the long-term envelope may help explain the low correlation in fluctuating noise. To account for the fluctuation due to the masker, an average of scores computed from several short windows may result in improved correlation (Payton and Shrestha, 2013; Rhebergen and Versfeld, 2005). Nevertheless, the synergetic effect of the two components is remarkable: BiDWGP seems to inherit the advantages of both approaches which account for the masking effects from different aspects.

4.3. Comparing predictions for anechoic and reverberant rooms

Having observed good overall predictive power for BiDWGP in both anechoic ($\rho \geq 0.94$ in Section 3.1.2) and reverberant conditions ($\rho = 0.92$ in Section 3.2.4), we further combined the data from the two experiments to investigate the overall BiDWGP performance, as illustrated in Fig. 4. While the rank capacity ($\rho_s = 0.92$) was maintained, a decrease in the Pearson correlation ($\rho_p = 0.89$) relative to that in early assessments is evident. It is noticeable that BiDWGP tends to somewhat overpredict intelligibility in the reverberation conditions compared to in the anechoic case, especially when listeners' performance is under approximately 60 RAU, i.e. in more adverse conditions. A separate linear fitting for each condition confirms the visual impression: a RAU score of 50 corresponds to 0.40 BiDWGP and 0.48 BiDWGP in anechoic and reverberant conditions, respectively. We speculate that this may be because the existence of reverberation increased the listeners' difficulties in noise by compromising the benefits of binaural listening to listeners. Culling et al. (1994) found that the effect of reverberation can cancel out the binaural advantage received when target vowels and maskers were spatially separated, compared to when they were co-located in anechoic conditions. This is also consistent with the findings in Rychtarikova et al. (2011). The effect of reverberation on binaural unmasking is mediated by reduced interaural coherence of the masker (Lavandier and Culling, 2007, 2008). When the RT_{60} is long enough, reverberation may also affect the intrinsic intelligibility of the target (Lavandier and Culling, 2008). However, this version of BiDWGP is unable to account for all of the negative impact of reverberation, resulting in overestimated intelligibility.

4.4. Limitations and further work

STI-based metrics typically rely on anechoic clean and reverberant noise-corrupted speech signals to make intelligibility predictions. Further to this, the BiDWGP metric also requires the noise signal in reverberant conditions to carry out more detailed audibility analysis. Although this may provide BiDWGP with more robust predictive accuracy, it also limits the use of the BiDWGP metric in situations where accessing separate noise signals is impossible. One alternative here could be to estimate the envelope of the reverberant noise by subtracting the clean speech envelope from that of the reverberant noise-corrupted speech signal. Further work is required to test to what extent the model performance would be affected by using an estimated noise envelope. Section 4.3 also discussed the improvements that could be made to the BiDWGP metric if the decreased binaural advantage due to reverberation were taken into account.

5. Conclusions

The BiDWGP metric for predicting binaural speech intelligibility was evaluated, along with the binaural version of four state-of-the-art metrics, for simulations of both anechoic conditions and reverberant rooms of different reverberation times. This was carried out in the presence of both stationary and fluctuating noise maskers. The BiDWGP metric demonstrated increased stability ($\rho > 0.87$) across all tested conditions compared to the reference metrics, implying potential for practical purposes in which speech intelligibility is a concern. Further work may focus on refining the model usability and revising it to allow more detailed modelling of noise and reverberation effects.

Acknowledgements

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. Data underlying the findings are fully available without restriction, details are available from https://dx.doi.org/10.17866/rd.salford.3172921.

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943–950.

Andersen, A.H., de Haan, J.M., Tan, Z.-H., Jensen, J., 2015. A binaural short time objective intelligibility measure for noisy and enhanced Speech. In: Proceedings of the Interspeech, pp. 2563–2567.

ANSI S3.5, 1997. ANSI S3.5–1997 Methods for the calculation of the Speech Intelligibility Index.

Beutelmann, R., Brand, T., Kollmeier, B., 2010. Revision, extension, and evaluation of a binaural speech intelligibility model. J. Acoust. Soc. Am. 127 (4), 2479–2497.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109 (3), 1101–1109.

Colburn, H.S., 1977. Theory of Binaural Interaction Based on Auditory Nerve Data. II. Detection of tones in noise. Supplementary material. Technical Report. AIP. document No. PAPS-JASMA-61-525-98.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119 (3), 1562–1573.

Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2004. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. J. Acoust. Soc. Am. 116 (2), 1057–1065.

Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2005. Erratum: The role head--induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J Acoust Soc Am 116, 1057 (2004)]. J. Acoust. Soc. Am. 118 (4), 552.

Culling, J.F., Summerfield, Q., Marshall, D.H., 1994. Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. Speech Commun. 14 (1), 71–95.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95 (5), 2670–2680.

Fletcher, H., 1921. An empirical theory of telephone quality. AT&T Intern. Memo. 101 (6).

Goldsworthy, R.L., Greenberg, J.E., 2004. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J. Acoust. Soc. Am. 116 (6), 3679–3689.

Gomez, A.M., Schwerin, B., Paliwal, K., 2012. Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio. Speech Commun. 54 (3), 503–515.

Hartung, K., Braasch, J., Sterbing, S.J., 1999. Comparison of different methods for the interpolation of head-related transfer functions. In: Proceedings of the 16th International Conference of the Audio Engineering Society, pp. 319–329.

Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100 (3), 1703–1716.

Houtgast, T., Steeneken, H.J.M., 1973. The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acta Acust United Ac 28 (1), 66–73.

Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77 (3), 1069–1077.

Houtgast, T., Steeneken, H.J.M., Plomp, R., 1980. Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. Acta Acust United Ac 46 (1), 60–72.

IEC, 2011. "Part 16: Objective rating of speech intelligibility by speech transmission index," in Proceedings of the International Electrotechnical Commission, Geneva, SwitzerlandIEC 60268 Sound System Equipment (fourth edition).

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.

Jelfs, S., Culling, J.F., Lavandier, M., 2011. Revision and validation of a binaural model for speech intelligibility in noise. Hear. Res. 275 (1–2), 96–104.

Jin, C.T., Member, S., Guillon, P., Epain, N., Zolfaghari, R., Van Schaik, A., Tew, A.I., Hetherington, C., Thorpe, J., 2014. Creating the Sydney York morphological and acoustic recordings of ears database. IEEE Trans. Multimed. 16 (1), 37–46.

Lavandier, M.N., Culling, J.F., 2007. Speech segregation in rooms: effects of reverberation on both target and interferer. J. Acoust. Soc. Am. 122 (3), 1713–1723.

Lavandier, M.N., Culling, J.F., 2008. Speech segregation in rooms: monaural, binaural, and interacting effects of reverberation on target and interferer. J. Acoust. Soc. Am. 123 (4), 2237–2248.

Levitt, H., Rabiner, L.R., 1967. Predicting binaural gain in intelligibility and release from masking for speech. J. Acoust. Soc. Am. 42 (4), 820–829.

Moore, B.C.J., Glasberg, B.R., Plack, C.J., Biswas, A.K., 1988. The shape of the ear's temporal window. J. Acoust. Soc. Am. 83 (7–8), 1102–1116.

Payton, K.L., Shrestha, M., 2013. Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data. J. Acoust. Soc. Am. 134 (5), 3818–3827.

Peterson, P.M., 1986. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. J. Acoust. Soc. Am. 80 (5), 1527–1529.

Plomp, R., Steeneken, H.J.M., Houtgast, T., 1980. Predicting speech intelligibility in rooms from the modulation transfer function. II. Mirror image computer model applied to rectangular rooms. Acta Acust United Ac 46 (1), 73–81.

Rennies, J., Brand, T., Kollmeier, B., 2011. Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. J. Acoust. Soc. Am. 130 (5), 2999–3012.

Rhebergen, K.S., Versfeld, N.J., 2005. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. J. Acoust. Soc. Am. 117 (4), 2181–2192.

Rothauser, E.H., Chapman, W.D., Guttman, N., Silbiger, H.R., Hecker, M.H.L., Urbanek, G.E., Nordby, K.S., Weinstock, M., 1969. IEEE Recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust 17, 225–246.

Rychtarikova, M., Van Den Bogaert, T., Vermeir, G., Wouters, J., 2011. Sound source localisation and speech intelligibility in virtual rooms. In: Proceedings of the Jaarvergadering ABAV.

Santos, J.F., Cosentino, S., Hazrati, O., Loizou, P.C., Falk, T.H., 2013. Objective speech intelligibility measurement for cochlear implant users in complex listening environments. Speech Commun. 55 (7-8), 815–824.

Savioja, L., Svensson, U.P., 2015. Overview of geometrical room acoustic modeling techniques. J. Acoust. Soc. Am. 138 (2), 708–730.

Schlesinger, A., Ramirez, J.-P., Boone, M.M., 2010. Evaluation of a speech-based and binaural speech transmission index. In: Proceedings of the 40th International Conference: Spatial Audio: Sense the Sound of Space (Audio Engineering Society Conference).

Shaw, E., Vaillancourt, M.M., 1985. Transformation of soundpressure level from the free field to the eardrum presented in numerical form. J. Acoust. Soc. Am. 78 (3), 1120–1123.

Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. 67 (1), 318–326.

Studebaker, G.A., 1985. A 'rationalized' arcsine transform. J. Speech Hear. Res. 28, 455–462.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short time objective intelligibility measure for time-frequency weighted noisy speech. In: Proceedings of the ICASSP, pp. 4214–4217.

Tang, Y., 2014. Speech Intelligibility Enhancement and Glimpse-based Intelligibility Models for Known Noise Conditions, Ph.D. thesis. Universidad del País Vasco.

Tang, Y., Cooke, M., Fazenda, B.M., Cox, T.J., 2015. A glimpse-based approach for predicting binaural intelligibility with single and multiple maskers in anechoic conditions. In: Proceedings of the Interspeech, pp. 2568–2572.

Tang, Y., Cooke, M.P., Valentini-Botinhao, C., 2016. Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech. Computer Speech and Language 35, 73–92.

Wendt, T., van de Par, S., Ewert, S.D., 2014. A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. J. Audio Eng. Soc. 62 (11), 748–766.

van Wijngaarden, S.J., Drullman, R., 2008. Binaural intelligibility prediction based on the speech transmission index. J. Acoust. Soc. Am. 123 (6), 4514–4523.

Zurek, P.M., 1993. Acoustical Factors Affecting Hearing Aid Performance. Allyn and Bacon, Needham Heights, MA, pp. 255–276. chapter Binaural advantages and directional effects in speech intelligibility.