A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers[a)]

Yan Tang[b)]
Acoustics Research Centre, University of Salford, Salford M5 4WT, United Kingdom

Martin Cooke[c)]
Ikerbasque (Basque Science Foundation), Bilbao, Spain

Bruno M. Fazenda and Trevor J. Cox
Acoustics Research Centre, University of Salford, Salford M5 4WT, United Kingdom

(Received 7 October 2015; revised 10 August 2016; accepted 25 August 2016; published online 21 September 2016)

One criterion in the design of binaural sound scenes in audio production is the extent to which the intended speech message is correctly understood. Object-based audio broadcasting systems have permitted sound editors to gain more access to the metadata (e.g., intensity and location) of each sound source, providing better control over speech intelligibility. The current study describes and evaluates a binaural distortion-weighted glimpse proportion metric—BiDWGP—which is motivated by better-ear glimpsing and binaural masking level differences. BiDWGP predicts intelligibility from two alternative input forms: either binaural recordings or monophonic recordings from each sound source along with their locations. Two listening experiments were performed with stationary noise and competing speech, one in the presence of a single masker, the other with multiple maskers, for a variety of spatial configurations. Overall, BiDWGP with both input forms predicts listener keyword scores with correlations of 0.95 and 0.91 for single- and multi-masker conditions, respectively. When considering masker type separately, correlations rise to 0.95 and above for both types of maskers. Predictions using the two input forms are very similar, suggesting that BiDWGP can be applied to the design of sound scenes where only individual sound sources and their locations are available.

[http://dx.doi.org/10.1121/1.4962484]

[MSS] Pages: 1858–1870

I. INTRODUCTION

Speech output, both natural and synthetic, is increasingly used in applications such as spoken dialogue systems, broadcast audio, and in public address systems. A key goal in the deployment of generated speech is to ensure that the information conveyed by speech is correctly received by the target audience. Traditional channel-based broadcasting systems are being gradually challenged by object-based systems, which have greater flexibility for sound production (customisation of properties of individual sounds) and can provide better transplantability to audio products (adaptation to suit reproduction devices). Within many object-based audio systems, information about the spatial configuration of the target speech source and potential maskers is available as a parameter of the design process. For example, in broadcast audio applications where dialogue is involved (e.g., Sonnenscheinn, 2001; Mapp, 2008), a sound editor may wish to know the approximate speech intelligibility of the "sound scene" that results after mixing of acoustic sources. A similar design problem has to be solved to guarantee minimum intelligibility levels as a function of the locations of the target sound and listeners, as well as the speech-to-background ratio, in a multi-loudspeaker announcement system. Clearly, the measurement or estimation of intelligibility is a critical component of the sound scene design process.

The traditional approach to measuring intelligibility involves the use of listener panels. However, reliance on subjective evaluation is slow and expensive and, consequently, limits the use of intelligibility scores as a key part of the design process. An alternative is to use objective intelligibility metrics (OIMs), which make quantitative predictions of the proportion of words likely to be heard correctly based on access to the speech signal and other contextual information such as masking noise or listening configuration. For example, OIMs have been applied recently to the problem of how to modify speech to render it more intelligible in noise via closed-loop optimisation of an intelligibility metric (Sauert and Vary, 2010; Tang and Cooke, 2010; Taal and Heusdens, 2014). The current study describes an OIM designed to estimate the intelligibility of speech sources in binaurally presented sound scenes.

Many OIMs are based on modeling the masked audibility of speech. The Speech Intelligibility Index (SII; ANSI S3.5, 1997) and its descendants [e.g., extended SII (ESII);

[a)]A preliminary version of part of this work was presented in "A glimpse-based approach for predicting binaural intelligibility with single and multiple maskers in anechoic conditions," Proceedings of INTERSPEECH, Dresden, Germany, September 2015.
[b)]Electronic mail: y.tang@salford.ac.uk
[c)]Also at Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain.

Rhebergen and Versfeld, 2005], as well as the glimpse proportion (GP; Cooke, 2006) fall into this category. Other OIMs operate by estimating the distortion induced by the masker or reverberation to the speech in a modulation domain. The latter class of OIMs include the Speech Transmission Index (STI; Steeneken and Houtgast, 1980), the normalised-covariance measure (NCM; Ma *et al*., 2009), and the short-term objective intelligibility index (STOI; Taal *et al*., 2010). While earlier metrics such as the SII and the STI operate on the long-term speech signal, more recent OIMs (e.g., ESII, STOI, GP) integrate short-term information, influenced by the ability of listeners to attend to dips in the masker (Miller and Licklider, 1950; Howard-Jones and Rosen, 1993).

In order to deal with more common listening situations, Zurek (1993) adapted SII to enable intelligibility predictions for a speech target and a noise masker separated in azimuth. A model based on combining equalisation-cancellation theory (Durlach, 1963, 1972) with SII was introduced by Beutelmann and Brand (2006) to predict binaural intelligibility in spatial-separated noise and reverberation conditions. van Wijngaarden and Drullman (2008) extended the STI to deal with binaural listening. Lavandier and Culling (2010) proposed an approach that augments the SII concept with components to account for binaural unmasking and better-ear listening, leading to predicted speech reception thresholds (SRTs). More recently, Jelfs *et al*. (2011) revised the model of Lavandier and Culling (2010) to enable the direct use of binaural room impulse responses (BRIRs) when predicting spatial release from masking. Further versions of these models have also been proposed to take into account short-term information with modulated maskers (e.g., Beutelmann *et al*., 2010; Collin and Lavandier, 2013). Cosentino *et al*. (2014) extended a monaural measure—speech to reverberation modulation energy ratio (Falk and Chan, 2008)—by integrating two additional components accounting for the better-ear effect and binaural unmasking. As a non-intrusive measure, this measure allows predicting binaural intelligibility from speech + noise mixture without separate access to the speech and noise signals.

While some of the aforementioned binaural algorithms (e.g., Lavandier and Culling, 2010; Beutelmann *et al*., 2010) aim to *model* the detailed binaural processes involved in human speech perception, and which output intelligibility estimates as a side-effect, our aim is more modest. The goal of the current study is to develop and evaluate an easily-computed *metric*, like the standard intelligibility measures SII and STI, capable of making robust predictions of overall intelligibility of a speech target in a spatial configuration alongside one or more sources of masking noise.

One important aspect of binaural listening is the better-ear advantage for spatially separated sources, based on the notion that listeners can exploit whichever ear has the more favourable signal-to-noise ratio (SNR) for the speech target due to head shadow effects. Previous application of the better-ear effect in binaural OIMs (e.g., Zurek, 1993; Lavandier and Culling, 2010) makes use of the long-term frequency-dependent SNR at the better ear. Studies (e.g., Shannon *et al*., 1995; Drullman, 1995) have suggested that listeners are able to decode speech without having access to all the spectro-temporal information in the speech, implying that rather than an entire frequency, individual spectro-temporal regions may contribute to listeners' speech understanding. This leads to the notion that glimpsing those spectro-temporal regions where speech is released from masking could be adequate for speech to be understood in the presence of noise (Cooke, 2006). Brungart and Iyer (2012) studied the efficiency of better-ear glimpsing with symmetrically placed competing talkers, and suggested that listeners are capable of extracting information from better-ear glimpses that fluctuate rapidly across frequency between the two ears. Collin and Lavandier (2013) observed lower SRTs when a unique one-voice modulated masker was used than a different masker for each target sentence. The authors ascribed this to listeners' ability of making use of predictable dips of the masker while listening, but they further suggested that the benefit from "listening-in-dip" could be reduced if the dips' positions within the masker are less predictable to listeners.

The notion of better-ear glimpsing motivates the design of the binaural OIM introduced in this paper. Specifically, the metric assumes that listeners have access to a glimpse of the speech target whenever the target is deemed to be glimpsed in either of the two ears. However, there is evidence that better-ear glimpsing alone cannot fully account for spatial release from masking (Glyde *et al*., 2013). Consequently, the proposed metric also incorporates a component which reflects binaural unmasking due to interaural time differences (ITDs) at the two ears, using an estimate of the binaural masking level difference (BMLD; Durlach, 1963; Levitt and Rabiner, 1967). The BMLD is applied at the point of deciding which spectro-temporal regions contain glimpses of the speech target.

The proposed metric uses as its base a glimpse-based estimate of single-channel intelligibility known as distortion-weighted glimpse proportion (DWGP; Tang, 2014). DWGP was developed in response to the observation that many OIMs show poor predictive accuracy when considered across different types of maskers. An evaluation of seven published OIMs using common datasets and subjective scores reported in Tang *et al*. (2016) suggested that those OIMs motivated by masked audibility (e.g., GP, ESII) tend to overestimate intelligibility in fluctuating maskers such as competing speech (CS) relative to stationary maskers such as speech-shaped noise (SSN), while those OIMs inspired by measuring the distorting effect of noise (e.g., STOI, NCM) exhibit the converse behaviour. Tang (2014) demonstrated that adding distortion-weighting to a measure of GP leads to better predictions across stationary and fluctuating maskers.

A further consideration in developing a binaural intelligibility metric concerns the form of the input. For an anechoic environment, one approach (e.g., van Wijngaarden and Drullman, 2008; Jelfs *et al*., 2011) requires explicit left and right ear signals or head-related transfer functions (HRTFs) for speech and masker(s). Given the fact that the metadata of each sound source, such as the intensity, the relative distance from the listener, and the azimuth on a horizontal plane, are available in an object-based audio system, an alternative is to start with anechoic monophonic

recordings of speech and masker along with information about their locations. This approach is particularly relevant in sound scene design if HRTFs are not available or are costly to collect. In principle, with the aid of a binaural OIM, sound scene designers can then more easily optimise intelligibility by manipulating properties such as speech-to-background ratio and source locations, while meeting a specified intelligibility criterion. Here, we show how the DWGP metric with its binaural extensions can be used with both forms of input, i.e., binaural recordings or anechoic monophonic recordings along with information about their spatial location relative to the listener.

Section II summarises the DWGP metric on which the proposed metric is based. The binaural metric, which we call BiDWGP, is defined in Sec. III. By taking the binaural Speech Transmission Index (BiSTI; van Wijngaarden and Drullman, 2008) as a reference metric, the predictive capacity of both metrics is evaluated with respect to subjective scores from two listening experiments involving the identification of keywords in sentences in the presence of one masker (experiment I, Sec. IV) or several maskers (experiment II, Sec. V). Performance is evaluated in both stationary noise and CS maskers in a variety of spatial configurations involving separation in azimuth and distance to the listener.

II. MONAURAL DWGP

The DWGP metric, which forms the basis for the current binaural extension, was introduced by Tang (2014) and is briefly reviewed here. The initial stage of the metric simulates peripheral auditory filtering. Target speech source s, masker n, and their sum y are processed independently by a bank of 34 gammatone filters (Patterson *et al.*, 1988) using an implementation described in Cooke (1993). Filter centre frequencies lie in the range 100–7500 Hz spaced equally on the scale of equivalent rectangle band (Moore and Glasberg, 1983). To model audibility in quiet, the output of each filter is adjusted by a frequency-dependent gain converted from the hearing threshold, interpolated from ISO 389-7 (2006). This approach permits the use of a constant hearing level value (HL = 25 dB for normal hearing cohort) to subsequently define which glimpses are supra-threshold [see Eq. (2) below]. Spectro-temporal excitation patterns (STEPs) of the speech S, masker N, and their mixture Y are computed by extracting the Hilbert envelope of each filter output, smoothing with a leaky integrator with an 8 ms time constant (Moore *et al.*, 1988), and downsampling to 100 Hz.

Following temporal envelope extraction, a weighting designed to model the effect of masker-induced fluctuations on speech envelope is applied. The frequency-dependent distortion weighting W_f is defined as the normalised temporal cross-correlation of the STEP temporal envelopes of the clean $S_f(t)$ and noise-corrupted speech signals $Y_f(t)$:

$$W_f = \frac{\sum_{t=1}^{T}\left(Y_f(t)-\bar{Y}_f\right)\left(S_f(t)-\bar{S}_f\right)}{\sqrt{\sum_{t=1}^{T}\left(Y_f(t)-\bar{Y}_f\right)^2\sum_{t=1}^{T}\left(S_f(t)-\bar{S}_f\right)^2}}, \quad (1)$$

where $\bar{S}_f$ and $\bar{Y}_f$ represent across-time means of $S_f(t)$ and $Y_f(t)$.

Glimpses are defined as spectro-temporal regions where the speech is above the HL and exceeds the masker by 3 dB ($\alpha = 3$ dB), with which the DWGP metric was found to provide the best match to listener performance (Tang, 2014). That is, glimpses in frequency channel f must meet criterion G_f,

$$G_f: \quad S_f(t) > \max(N_f(t)+\alpha, \mathrm{HL}), \quad (2)$$

where S and N are expressed in decibels. Finally, as summarised in Eq. (3), the DWGP is computed by weighting the GP in each frequency band by the distortion weighting W_f, multiplying by a SII band importance function (BIF) K_f interpolated from the values provided in Table III of ANSI S3.5 (1997), summing across frequency and finally compressing the output by a quasi-logarithmic function v, which models the finding (e.g., Barker and Cooke, 2007) that ceiling intelligibility occurs for GPs substantially lower than unity

$$\mathrm{DWGP} = v\left(\frac{1}{T}\sum_{f=1}^{F}\left(K_f W_f \sum_{t=1}^{T}\mathcal{H}(G_f)\right)\right), \quad (3)$$

where

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

and

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

T and F are the number of time frames and frequency channels and $\mathcal{H}(.)$ is the Heaviside unit step function, which counts the time frames meeting the glimpsing criterion G_f in channel f.

III. BINAURAL DISTORTION-WEIGHTED GLIMPSE PROPORTION (BiDWGP)

The DWGP metric is extended to binaural signals via components that model the better-ear advantage and binaural unmasking. The former involves a combination of left and right ear glimpses (Sec. III C below) while the latter is based on an estimate of the BMLD.

Two forms of input are handled by the BiDWGP metric. One, denoted the "binaural input" condition, assumes the availability of binaural recordings for target speech and maskers. The other, the "source + location" case, assumes that the input consists of anechoic monophonic recordings for speech and masker(s) together with their azimuth and distance relative to the listener. In practice, the key difference between these scenarios lies in an extra stage—estimating binaural signals—required by source + location input.

The metric in the current study is to predict intelligibility in anechoic conditions for spatial configurations of

speech and one or more maskers defined by their location on a horizontal plane.

A. BMLD for binaural inputs

As proposed in Levitt and Rabiner (1967) the gain due to binaural unmasking—BMLD—can be computed for each frequency f using an approach described in Culling *et al.* (2005), which was later adopted by Lavandier and Culling (2010) in their predictions of binaural intelligibility

$$\mathrm{BMLD}_f = 10\log_{10}\left[\frac{k-\cos\left(\phi_f^s-\phi_f^n\right)}{k-\rho_f^n}\right], \tag{4}$$

where

$$k = (1+0.25^2)\exp((2\pi f)^2\cdot 0.000\,105^2)$$

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. If more than one masker is present, ϕ_f^n and ρ_f^n in Eq. (4) are computed after summing the gammatone outputs to all maskers at frequency f.

B. BMLD for source + location inputs

In this scenario, a binaural signal corresponding to each source has to be estimated from the anechoic monophonic recording and location on a horizontal plane. Location is specified in polar coordinates (r,θ) with reference to an origin at the centre of the listener's head, for source distance r in metres and azimuth angle θ subtended by the source relative to the 0° baseline in front of the listener.

First, source amplitude is adjusted to simulate signal attenuation due to distance, relative to a reference distance at which a sound pressure level (SPL) is measured. The signal is then processed by an auditory filterbank as described in Sec. II. Following Zurek (1993), to construct a binaural signal the difference in SPL between each ear and the listener's frontal position is interpolated using a transformation of SPL from the free field to the eardrum (Shaw and Vaillancourt, 1985). The azimuth- and frequency-dependent gains $d_f(\theta)$ converted from the SPL differences are then used to weight the outputs of the gammatone filters, resulting in the signals for each ear with estimated interaural level difference (ILD).

In order to calculate the frequency-dependent BMLD using Eq. (4) without access to binaural signals, the ITD for each frequency needs to be estimated for interaural phase shifts of the speech ϕ_f^s and the masker ϕ_f^n, and the interaural coherence of the noise masker ρ_f. Based on a model of ITD as a function of azimuth (Kuhn, 1977), Eq. (11) in Zurek (1993) provides an approach to do so using assumed nominal head radius. With the estimated ITD for frequency f, the ILD-adjusted signals for each ear from the early step are shifted forward or backward for certain sample points. The BMLD_f is then calculated as described in Sec. III A for both single- and multi-masker conditions.

To include the BMLD component into the metric, the frequency-dependent BMLD_f is applied at the stage of glimpse definition, replacing Eq. (2) by

$$\begin{aligned} G_f : {} & S_f(t) > HL \ \wedge \\ & S_f(t) + \mathrm{BMLD}_f > N_f(t) + \alpha \end{aligned} \tag{5}$$

C. The better-ear effect

Inspired by findings from Zurek (1993) and Brungart and Iyer (2012), the better-ear effect is modeled in BiDWGP by combining glimpses from the two ears. Glimpses are computed separately for left and right ear models and combined to produce binaural glimpses in all time-frequency regions where either or both individual ears produce a glimpse, i.e., the inclusive "or" of glimpsed spectro-temporal locations for the left and right ears

$$G_f^{\mathrm{bi}} = G_f^L \vee G_f^R, \tag{6}$$

where G_f^L and G_f^R indicate glimpses in channel f defined by Eq. (5) occurring in the left or right ear.

D. BiDWGP

The BiDWGP metric is defined in Eq. (7). Distortion-weighting is extended to the binaural case by averaging cross-correlations for the left and right ear STEPs, resulting in a frequency-dependent binaural weighting W_f^{bi}. The glimpsing criterion is extended to incorporate the BMLD and the better-ear effect

$$\mathrm{BiDWGP} = v\left(\frac{1}{T}\sum_{f=1}^{F}\left(K_f W_f^{\mathrm{bi}}\sum_{t=1}^{T}\mathcal{H}\left(G_f^{\mathrm{bi}}\right)\right)\right). \tag{7}$$

IV. EXPERIMENT I: INTELLIGIBILITY FOR SINGLE MASKERS

The first experiment was designed to evaluate the predictive accuracy of the BiDWGP metric for target speech in the presence of a single masker, and consists of conditions in which source azimuth and distance are varied. Binaural unmasking varies as a function of the separation in azimuth between target speech and masker, hence, subjective intelligibility is expected to change with masker location for a fixed speech target. Likewise, the distances from the listener to speech and masker will also affect intelligibility due to increasing signal attenuation with increasing source-listener distance.

A. Design

Target speech material was drawn from recordings of the Harvard sentences (Rothauser *et al.*, 1969) spoken by a British English male talker. Each sentence contains five or six keywords (e.g., "**take** the **winding path** to **reach** the **lake**" or "**many hands help get** the **job done**"), which are used for scoring purposes. A stationary noise (SSN, with spectrum matching the long-term corpus average) and a

fluctuating masker (CS) were used as maskers. CS was generated by concatenating sentences uttered by a British English female talker from the SCRIBE corpus (University College London *et al.*, 1992). In order to minimise the informational masking, listeners were explicitly instructed to always focus on the target male voice when the competing female voice was present. Speech-in-noise stimuli were generated by mixing the Harvard sentences with each masker at two SNRs: −9 and −6 dB for SSN and −18 and −15 dB for CS, values based on pilot tests aimed at producing keyword recognition rates of approximately 25% and 50% when the target and masker are co-located in front of the listener (i.e., $\theta^s = \theta^n = 0°$).

The target speech source was fixed at 0° relative to the listener (i.e., $\theta^s = 0$), while the azimuth of the masker θ^n varied across conditions. Three source-listener distances r_s and r_n for speech and masker sources, respectively, were also tested: (i) target and masker equally distant from the listeners ($r_s = r_n = 2$ m, 10 azimuths); (ii) speech closer to listener than masker ($r_s = 1.5$ m, $r_n = 2.5$ m, 4 azimuths); (iii) masker closer to listener than speech ($r_s = 2.5$ m, $r_n = 1.5$ m, 4 azimuths). This design leads to a total of 72 conditions (2 masker types × 2 SNR levels × 18 masker locations). Figure 1 shows the locations of the noise masker in the horizontal plane.

B. Listeners

Fourteen native British English speakers from the University of Salford with ages ranging from 24 to 40 yr (mean age 30 yr) were recruited as paid participants in experiment I. Audiological screening suggested that all participants had a hearing threshold below 20 dB HL at frequencies 500–4000 Hz.

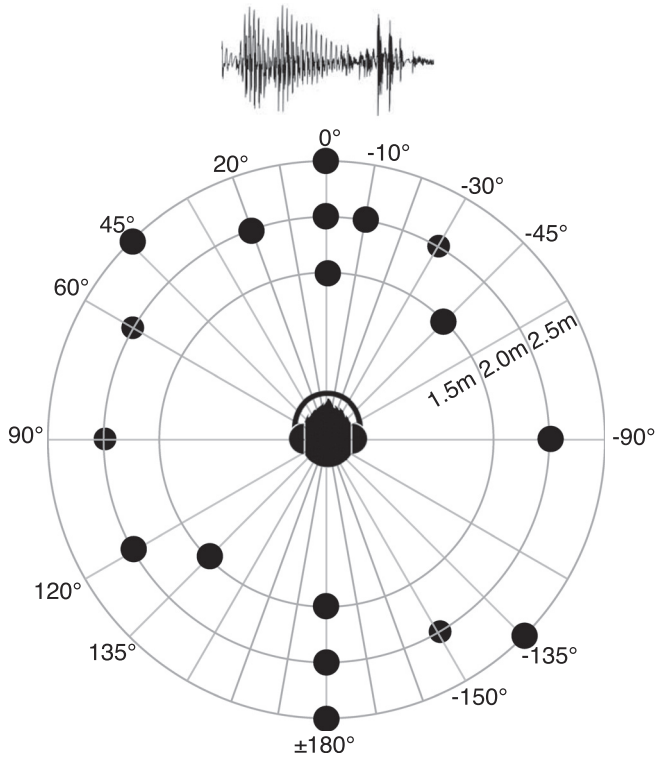

FIG. 1. Masker locations used in experiment I. The speech target is located at an azimuth of 0° at a distance of 1.5, 2.0, or 2.5 m.

C. Procedure

The experiment was conducted in a semi-anechoic room at the Acoustics Research Centre at the University of Salford. A virtual sound field was simulated by convolving anechoic monophonic speech and noise recordings with BRIRs recorded in an anechoic chamber (Wierstorf *et al.*, 2011). Stimuli were presented to listeners via Sennheiser HD650 headphones following pre-amplification by a Focusrite Scarlett 2i4 USB audio interface. The presentation level of speech was calibrated and fixed at 63 dB(A) at the listener's ears when the source was 2 m in front of the listener; the masker presentation level was adjusted to achieve the required SNR. For unequal speech/masker distances, presentation levels were adjusted using an inverse-square law, taking the level at 2 m as a reference. This was performed after the SNR level had been adjusted as described above. In each speech + noise mixture, the masker preceded and followed the speech by 300 ms; each mixture was further ramped in and out for 10 ms.

Participants listened to 3 sentences in each of the 72 conditions; no sentence was repeated for any listener. Sentences were arranged into four blocks according to masker type and SNR level. To minimise the bias due to the difference in intrinsic intelligibility among the sentences, over the entire experiment, each sentence was only presented once in each condition and was heard by each listener only once using a balanced design. In each block, the order of sentences was randomised for each listener. Stimuli were delivered to listeners using a MATLAB interface. Each stimulus was only played once. After each stimulus, listeners typed any words they heard from the sentence using a physical computer keyboard; the following stimulus was played as soon as the "enter" button was pressed. All listeners completed the experiment within one hour.

D. Postprocessing

Subjective performance in each condition is computed as the keyword identification rate. In order to deal with words that may have multiple correct spellings, a predefined homophone dictionary was used during scoring.

The performance of the metric was evaluated in terms of the Pearson correlation coefficient ρ between the metric outputs and the mean subjective scores, transformed to rationalised arcsine units (RAU; Studebaker, 1985), along with the possible lowest root-mean-square error RMSE′ after a linear fit to raw metric outputs: $\text{RMSE}' = \sigma_d\sqrt{1-\rho^2}$, where σ_d is the standard deviation of subjective scores in a given condition.

E. Results

The upper panels of Fig. 2 show keyword identification rates as a function of target-masker separation in azimuth, for conditions where both target and masker are at the same distance from the listener. These results confirm the

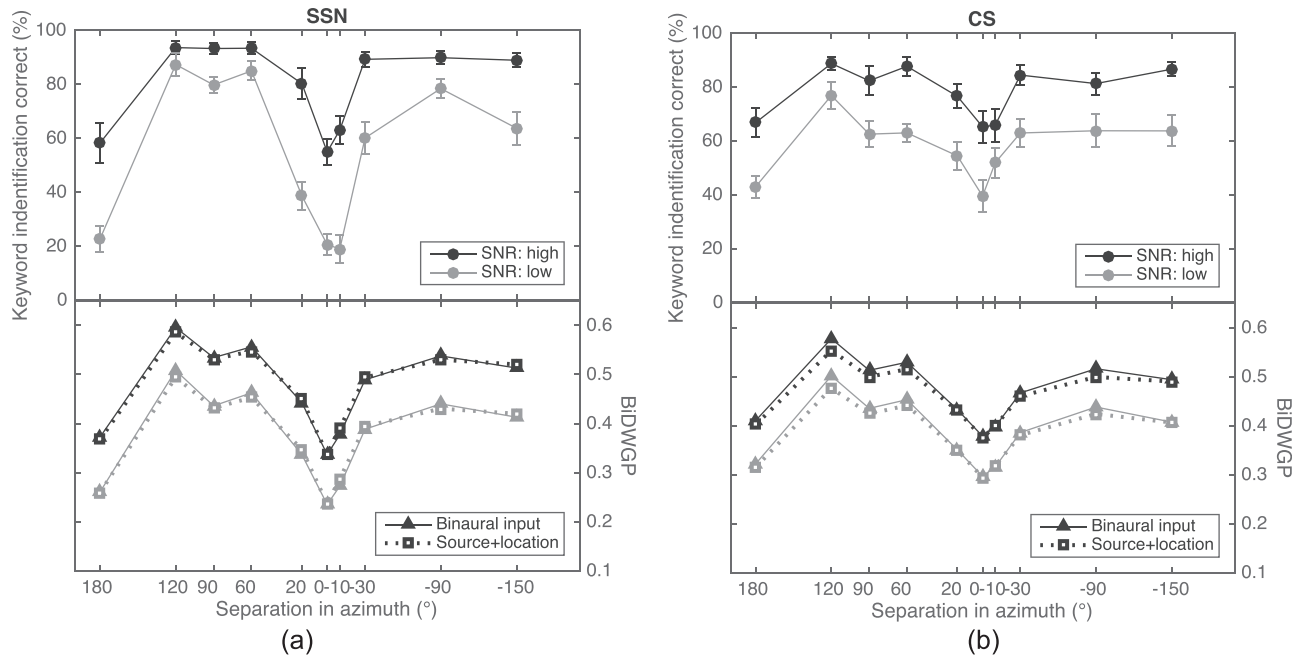

FIG. 2. Listeners' keyword identification rates (upper) and BiDWGP predictions (lower) for different speech-masker separations when the speech and masker were at the same distance from the listener ($r_s = r_n = 2m$), in SSN (left) and CS (right). Error bars indicate ±1 standard error.

well-documented listener benefit from binaural unmasking when target and maskers are spatially separated (Hirsh, 1950; Dirks and Wilson, 1969; Hawley *et al.*, 1999; Hawley *et al.*, 2004): identification rates are lowest when target and masker are co-located ahead of the listener, but increase rapidly with increasing separation up to a maximum for separations in the range of 60–120° before falling in the 180° condition where ILD and ITD cues are similar for target speech and masker. A similar pattern is apparent for both maskers and SNRs. In SSN, the maximum benefits from separation in azimuth are 68.3 and 38.7 percentage points at low and high SNRs, respectively; for CS, more modest gains of 37.3 and 23.5 percentage points are observed.

The lower panels of Fig. 2 depict predictions of intelligibility from the BiDWGP metric for both binaural and source + location inputs. The pattern exhibited in the subjective results is repeated in the predicted scores, with $\rho = 0.95$ for both forms of input; see Table I.

The effect of azimuthal separation for a speech target and masker which differ in distance to the listener is shown in Fig. 3. For the cases where the target is closer than the masker ($r_s < r_n$), speech is substantially more intelligible than for the reverse case ($r_s > r_n$), especially when speech and masker have the same azimuth or are separated by 180° ($p < 0.001$). When the speech source is further from the listener than the masker, the BiDWGP metric shows a similar pattern ($\rho = 0.97$). For the converse case, some differences are apparent ($\rho = 0.82$), especially when the subjective intelligibility almost converges despite different SNR levels, the model predictions still show a large departure (~0.1 BiDWGP) between the two SNR levels.

As is evident from Figs. 2 and 3, almost identical BiDWGP predictions result from binaural and source + location forms of input to the metric, with an overall correlation of $\rho = 0.998$ and RMSE = 0.01 between the two. Table I further presents the performance of BiSTI, which is compared with BiDWGP using chi-squared tests on *Z*-transformed scores. Except for $r_s < r_n$ ($Z = 1.933, \chi^2 = 3.739, p = 0.05$), the BiDWGP metric with both input forms outperforms the standard intelligibility metric with its binaural extensions in all other sub-conditions ($Z \geq 3.543, \chi^2 \geq 12.642, p < 0.001$).

TABLE I. Listener-metric Pearson correlation coefficients ρ (with RMSE′ in RAU in parentheses) for the BiDWGP metric with two forms of input and the BiSTI metric in a number of sub-conditions of experiment I (*N* indicates the number of data points in each sub-condition). For all ρ, $p < 0.001$.

	N	Binaural	Source + location	BiSTI
SSN	36	0.95 (10.1)	0.95 (10.7)	0.87 (15.5)
CS	36	0.96 (5.6)	0.96 (5.5)	0.88 (9.7)
$r_s < r_n$	16	0.82 (5.4)	0.82 (5.3)	0.60 (7.5)
$r_s > r_n$	16	0.97 (6.6)	0.97 (6.6)	0.62 (20.3)
$r_s = r_n$	40	0.95 (6.8)	0.95 (6.4)	0.73 (14.7)
Overall	72	0.95 (8.6)	0.95 (8.9)	0.78 (17.1)

F. Discussion

The BiDWGP metric predicts the pattern of listeners' keyword identification rates for target speech in the presence of a stationary or fluctuating masker varying in azimuth or distance with an overall correlation of 0.95. Encouragingly, almost identical predictions result from binaural and source + location forms of input, demonstrating the applicability of the metric for a range of application scenarios. There is some evidence of a ceiling effect for listeners in conditions where the target speech is closer than the masker. The ceiling effect is clearly seen in Fig. 4, which plots predictions of the BiDWGP metric (source + location input) and the BiSTI against RAU-transformed subjective scores for each condition of experiment I.

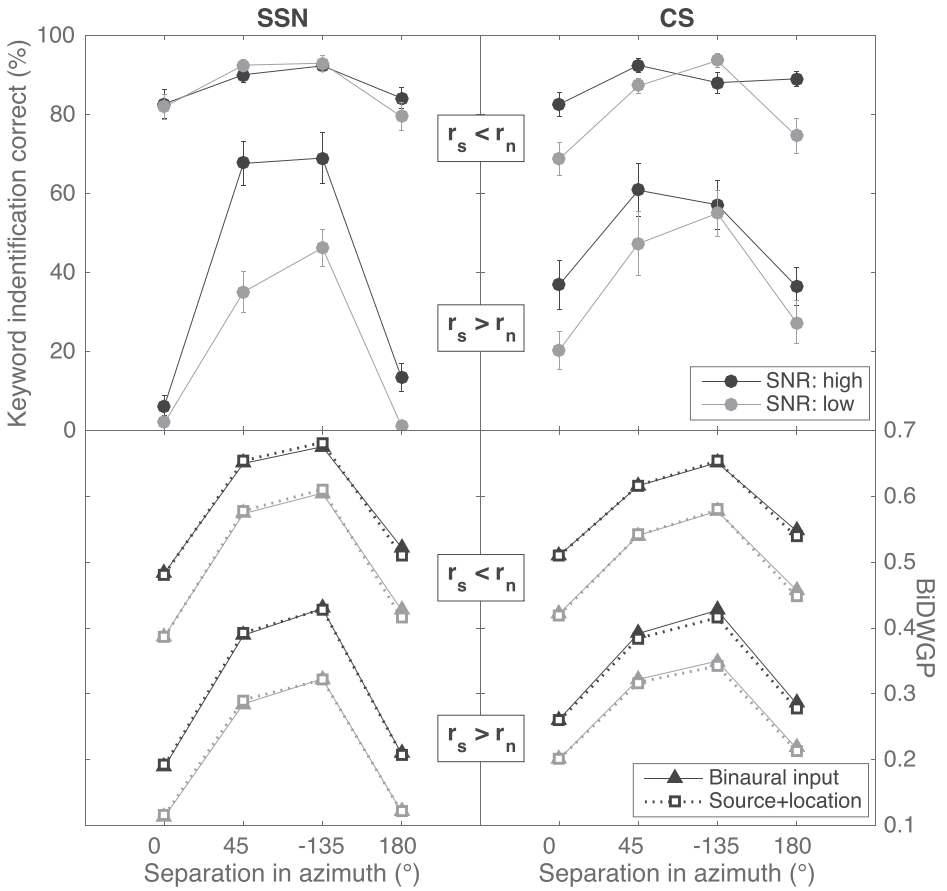

FIG. 3. Listeners' keyword identification rates (top) and BiDWGP predictions (bottom) for conditions where the target speech and masking sources are at different distances from the listener. Error bars indicate ±1 standard error.

Figure 4 further reveals a similar predictive pattern which has been observed in Tang *et al.* (2016) for those envelope distortion-based monaural OIMs (e.g., NCM and STOI)—the BiSTI tends to underestimate intelligibility in fluctuating masker (CS). Tang *et al.* (2016) further discussed that such predictive bias to CS is possibly because distortion-based OIMs only consider distortion over time but not over frequencies; quantifying intelligibility only based on distortion in the time domain may overestimate the negative impact of CS on intelligibility. By using separate linear fits, while the BiDWGP predictions from source + location input at the 50% RAU score are 0.35 and 0.34 in SSN and CS, respectively, the BiSTI values are 0.38 and 0.32. This discrepancy may account for the decreased predictive power of the BiSTI ($\rho = 0.78$) when making predictions across the two types of maskers despite its reasonable accuracy for each type of masker alone ($\rho > 0.87$).

Experiment I explored a relatively simple listening situation. In more complex listening conditions the target talker may not be directly in front of the listener, or there may be more than a single masker. Experiment II evaluates the performance of BiDWGP under these conditions.

V. EXPERIMENT II: INTELLIGIBILITY WITH MULTIPLE MASKERS

Listeners' keyword identification performance was tested for a range of azimuthal locations of the speech

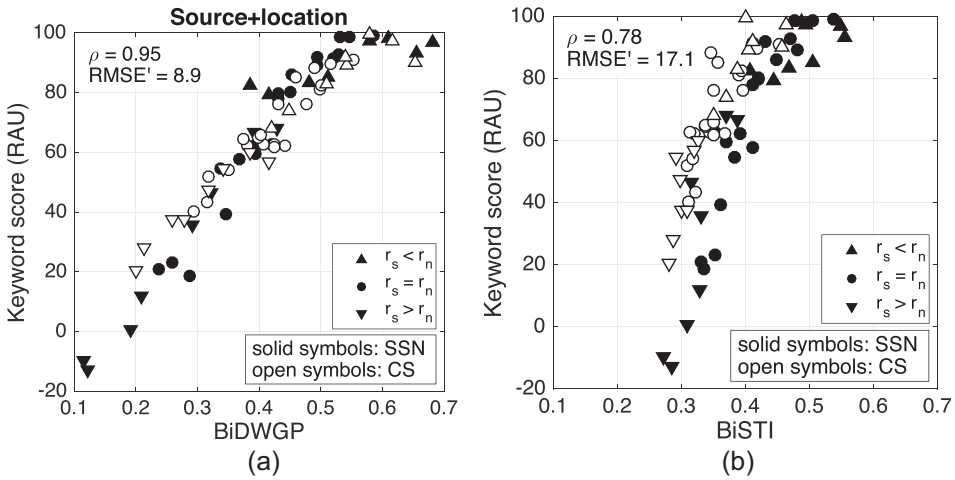

FIG. 4. Listener scores vs metric predictions for the conditions of experiment I.

TABLE II. Azimuth settings of the target speech and the masker. N_m is the number of maskers; Ψ_θ is the separation between speech and the closest masker.

N_m	$\theta^s(^\circ)$	$\theta^n(^\circ)$	$\Psi_\theta(^\circ)$
2	0	[30 −60], [30 90]	30, 30
	45	[0 90]	45
	−45	[45 90]	90
	90	[0 −90]	90
	−90	[0 −45]	45
3	0	[30 −60 90], [30 60 90], [−30 60 90]	30, 30, 30
	30	[0 60 90], [−60 −90 −120]	30, 90
	−60	[0 90 −120]	60

source in the presence of either two or three competing sources.

A. Design

Table II lists azimuth settings of the target speech and maskers. A total of 12 configurations were tested, 6 for each of the two- and three-masker cases. Target and maskers were located at a fixed distance of 2 m from the listener in all conditions. As in experiment I, target sentences were drawn from the Harvard corpus, but were different from those used in the first experiment. Sentences were presented in SSN or CS maskers, and in any condition all maskers were of the same type, i.e., SSN or CS. Due to the presence of additional maskers, individual masker SNRs were higher than those used in experiment I: −8 and −5 dB for SSN; −12 and −9 dB for CS. Note that *each* masker was adjusted to produce the specified SNR with respect to the target speech as described in Sec. IV A, and all maskers of the same type were uncorrelated. The true overall SNR is therefore approximately 3 dB and 4.7 dB lower than the quoted values in the two-masker and three-masker conditions, respectively. In total, experiment II consists of 48 conditions (2 masker types × 2 SNR levels × 12 source-masker location configurations).

B. Listeners and procedure

Fourteen native British English speakers (ages: 18–40 yr, mean age 27 yr) with a hearing threshold below 20 dB HL were recruited from the same source as in experiment I. Four had participated in the earlier experiment. Stimuli from the 48 conditions were presented to listeners in the same listening environment as used in experiment I. Listeners heard 5 different sentences in each condition, leading to a total of 240 sentences for the entire experiment.

C. Results

Figure 5 shows keyword identification rates (upper panels) along with predictions of the BiDWGP metric (lower panels). Intelligibility varies as a function of the configuration of the speech and maskers. For both masker types and SNRs, highest scores were obtained in the conditions where the speech has the largest azimuthal separation from the closest masker (90°). However, intelligibility was not monotonically related to the largest azimuthal separation. For example, a speech source directly ahead of the listener with maskers at 30° and 90° was significantly more intelligible than a target at 45° with maskers at 0° and 90° ($p < 0.001$). This finding suggests that the SNR at a listener's better ear is a more important determinant of speech intelligibility than the degree to which the speech and maskers are separated (Hawley *et al.*, 1999). This is particularly the case when all maskers are present on the lateral side of the listener while

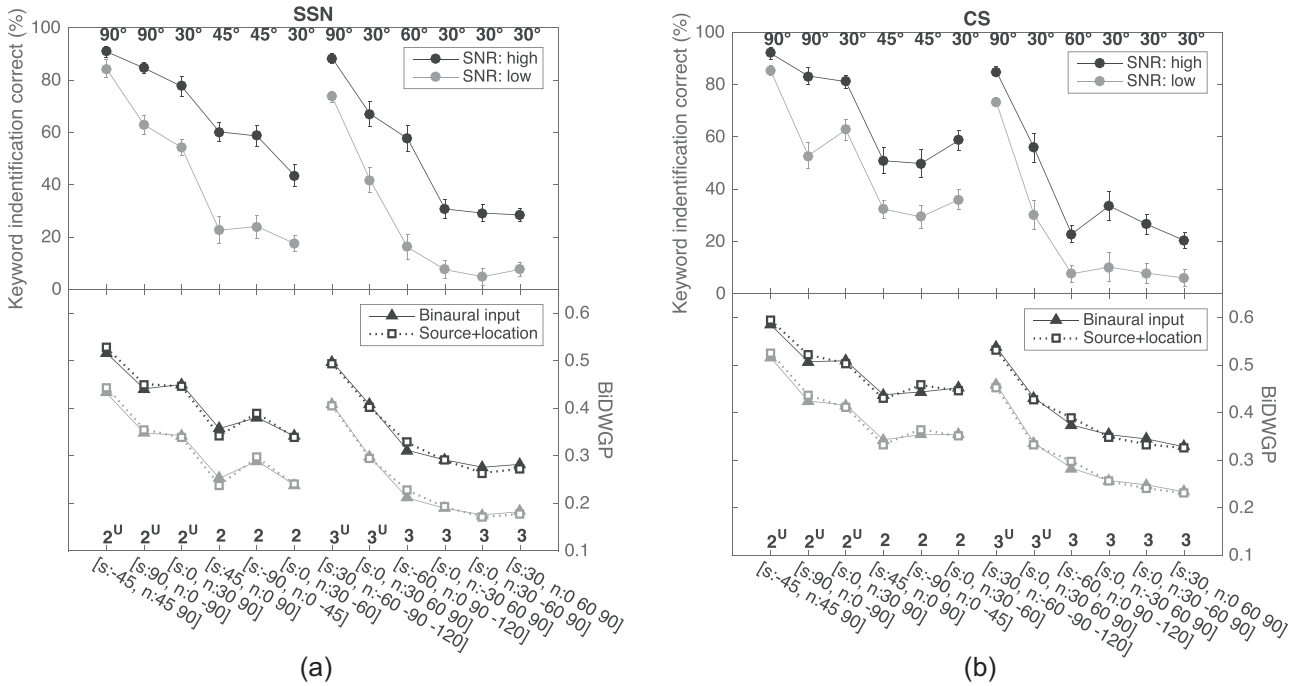

FIG. 5. Listeners' keyword identification rates (upper) and BiDWGP predictions (lower) in multiple SSN or CS maskers, grouped by number of maskers. Numbers above the x axis indicate the number of maskers (2 or 3) and the presence of superscript "U" denotes a unilateral distribution of maskers. Numbers in the upper part of the figure indicate the separations between speech and the closest masker in each setting Ψ_θ. Error bars indicate ±1 standard error.

TABLE III. Listener-metric Pearson correlation coefficients ρ (RMSE′ in RAU in parentheses) overall and for various sub-conditions of experiment II (N indicates the number of the data points in each sub-condition). For all ρ, $p < 0.001$.

	N	Binaural	Source + location	STI
SSN	24	0.98 (6.1)	0.98 (5.9)	0.95 (8.8)
CS	24	0.98 (6.1)	0.96 (7.6)	0.93 (10.8)
Two-masker	24	0.88 (11.1)	0.88 (11.2)	0.95 (7.4)
Three-masker	24	0.90 (12.5)	0.90 (12.7)	0.95 (9.3)
Overall	48	0.91 (11.9)	0.91 (12.1)	0.93 (10.3)

the speech is at 0° or on the opposite side of the listener. For other conditions in which $\Psi_\theta < 60$ and the speech lies between maskers, no significant difference in listeners' performance was found ($p \geq 0.072$), regardless of the lateral distribution of speech and maskers, except for $\Psi_\theta = 30$ [($s: 0°; n: 30°, -60°$) in SSN/high ($p < 0.001$) and ($s: 30°; n: 0°, 60°, 90°$) in CS/high] ($p < 0.01$), in which intelligibility is worse than in other conditions. We speculate that this outcome is a consequence of the head shadow effect. When all the maskers are unilaterally distributed, the energetic masking effect at the opposite ear may be largely attenuated by the head, leading to better intelligibility.

Predictions from the BiDWGP metric with two forms of input (lower panels of Fig. 5) show a similar pattern to listeners' scores, but with a less dramatic differences between tested conditions in the predictions compared to listeners' scores. Nevertheless, the overall predictive patterns between the two input forms are highly consistent ($\rho = 0.996$ and RMSE $= 0.01$). Table III shows listener-metric correlations, indicating that overall both input types lead to a correlation of $\rho = 0.91$, with a higher value ($\rho \geq 0.96$) for each individual masker type. Compared to the BiSTI, the BiDWGP with both input forms shows more robust predictive accuracy for individual maskers ($Z \geq 2.260$, $\chi^2 \geq 5.128$, $p < 0.05$), but less in two- or three-masker sub-conditions ($Z \leq -2.260$, $\chi^2 \geq 5.134$, $p < 0.05$). Overall, all the metrics have demonstrated statistically similar performance ($Z = -1.328$, $\chi^2 = 1.786$, $p = 0.184$).

D. Discussion

For each of the two maskers considered independently, the BiDWGP metric is highly correlated with listener scores ($\rho \geq 0.96$) in a background of two or three maskers, exceeding the correlations seen in the single-masker case. However, at $\rho = 0.91$, the across-masker prediction is somewhat lower than for the single-masker case of experiment I ($\rho = 0.95$). Figure 6 depicts predictions from the BiDWGP metric (with source + location input) and the BiSTI against RAU-transformed listener scores. There is some evidence in this plot of a masker-specific effect to BiDWGP: relative to the SSN case, the metric predicts a higher subjective performance in the presence of a CS masker. While the BiDWGP prediction at the 50% RAU score for SSN hardly changes from experiment I to experiment II (0.35 vs 0.34), it has increased from 0.34 to 0.41 for CS, leading to the decreased overall correlation between the BiDWGP predictions and listener scores.

The same overestimation for CS can be also seen for the BiSTI metric. Compared to that of 0.32 for single-masker conditions, a larger prediction of 50% RAU for CS (0.37) has been received in experiment II. The BiSTI prediction of 50% RAU for SSN only has a trivial change from 0.38 for experiment I to 0.37. Having observed the underestimation of BiSTI for CS in experiment I, the shift of predictions of CS leads to a reduction of the discrepancy between SSN and CS predictions by BiSTI as illustrated in Fig. 6, *coincidentally* resulting in a largely improved overall correlation ($\rho = 0.93$) between the BiSTI predictions and listener scores.

The overestimation for CS by both BiDWGP and BiSTI might be due to a form of informational masking not taken into account in the metrics. Unlike in experiment I, in which listeners were aware of the target speech being located straight ahead, the location of the target speech relative to the listener changed randomly from trial to trial. Many participants reported that the CS masker conditions were more difficult than those involving the stationary masker due to the need to identify the location of the target speech. We speculate that attention-switching due to source localisation and segregation might have had a negative impact on performance here. Although Hawley *et al.* (1999) suggested that intelligibility is not significantly associated with a listener's localisation ability, if listeners can locate the target source they appear to be able to understand speech better in the presence of other background noise or CS sources than if the

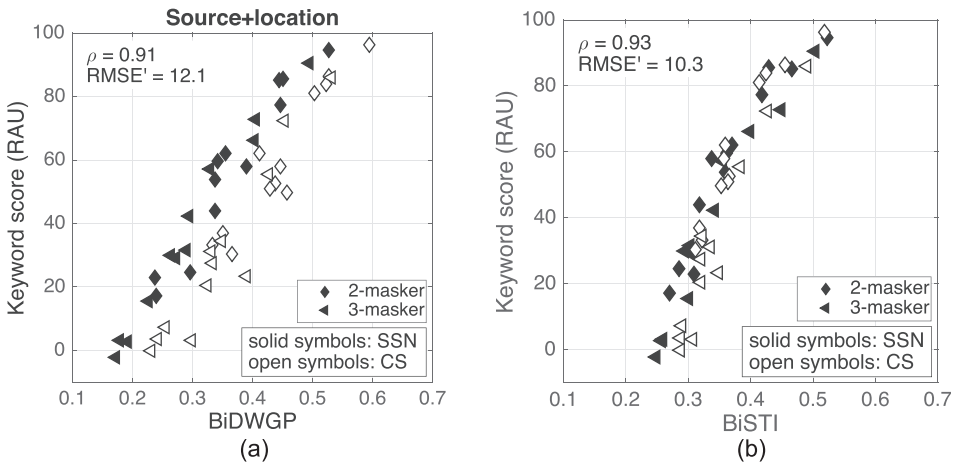

FIG. 6. Listener scores vs metric predictions for the conditions of experiment II.

location of the target source is unclear (Hirsh, 1950, 1971; Litovsky *et al.*, 2009)—a form of selective attention (Kock, 1950; Cherry, 1953; Litovsky *et al.*, 1999). Further experiments are needed to test the possibility of a cost of attention-switching hypothesis, perhaps using a visual cue to identify the location of the target source. Therefore, in order to improve the predictive accuracy of the metric, the loss of intelligibility due to any forms of attentional activity needs to be accounted for by an additional component in the metric if listeners have no prior knowledge of the location of the target speech.

VI. GENERAL DISCUSSION

BiDWGP, a binaural extension to a glimpse-based intelligibility metric, predicted subjective keyword identification rates for a speech target in the presence of 1–3 spatially separated stationary noise or CS maskers with correlations in the range of 0.91–0.95 across masker types, rising to 0.95–0.98 when masker types are considered separately.

The BiDWGP metric handles two forms of input, one consisting of binaural recordings, the other composed of the anechoic monophonic recordings of sources together with their locations in the horizontal plane. For the case of a single masker, both forms of input lead to very similar predictions. There are two key differences in the way the input types are handled in the BiDWGP metric. For the source + location input case, in order to construct binaural signals, ILD and ITD are estimated using source-to-eardrum transfer functions (Shaw and Vaillancourt, 1985) and an ITD model (Kuhn, 1977) in the azimuthal plane, respectively. Both approximations could conceivably produce potential errors in estimation. Here, we examine more closely possible differences arising from these computations.

A. ILDs

Estimating ILD from source + location input is crucial to accurately model the better-ear effect. The upper panels of Fig. 7 show differences between estimated (from source + location input) and measured (from binaural input) ILDs, i.e., $\Delta_X = X_{\text{estimated}} - X_{\text{measured}}$, where X denotes measurement, at different azimuths for the SSN signals used in experiment I and II. Note that, instead of calculating the ILD at a certain frequency band, the ILD here is computed as an overall effect of all frequencies from binaural signals; the same applies to the later ITD comparisons. The mean absolute Δ_{ILD} across masker azimuths are 1.0 dB in both experiments I and II, with a maximum difference of 2.3 dB when the masker is at 60° in both experiments. Similar patterns were also observed for CS signals used in both experiments. Note that there is a small difference at azimuths of 0° and 180° where no ILD should exist in principle, which may be due to measurement errors when the head-related impulse responses were recorded. The same reason may explain that the estimated ILDs are asymmetrical for two symmetrical positions (e.g., 90°/–90° and 45°/–45°), as the measured values for the SPL transformation (Table I in Shaw and Vaillancourt, 1985) used for estimating ILD in this study are not strictly consistent given two symmetrical positions

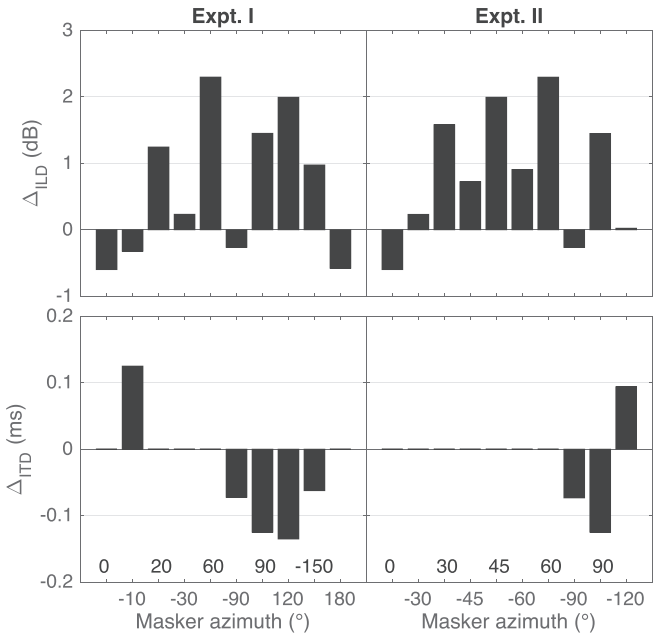

FIG. 7. Difference between the estimated and measured ILDs Δ_{ILD} and ITDs Δ_{ITD} of the SSN masker as a function of the azimuth relative to 0°.

relative to 0°. Although the estimated ILD is close to the measured ILD, there is a tendency to over-estimate the level difference. Nevertheless, the close correspondence in estimates ($\rho = 0.997$, RMSE = 0.01) made by BiDWGP with the two input forms without the BMLD component being integrated suggests that any deficiency in estimating the ILD in this study has little effect on the outcome of the metric.

B. ITDs

If binaural signals are available, the BMLD can be computed directly from phase differences between signal and masker, and the coherence of the masker based on equalisation-cancellation theory (Durlach, 1963, 1972) as described in Culling *et al.* (2004, 2005). For the BiDWGP metric with source + location input, ITD needs to be estimated in order to construct the binaural signals for both speech and masking sources. This enables the use of Eq. (4) for both input forms of BiDWGP for BMLD calculation. The lower panels of Fig. 7 show Δ_{ITD} at different azimuths for the SSN signals used in experiments I and II. The mean absolute Δ_{ITD} across masker azimuths are 0.05 and 0.03 ms with a maximum difference of 0.14 ms when the masker is at 120° in experiment I. Given the sampling frequency of 16 000 Hz for signals used in this study, the 0.14 ms difference is led by the inadequate shift of a mere 2.2 sample points between the estimated signals of the left and right ears.

Figure 8 compares the BMLDs calculated using measured or estimated ITDs for the SSN maskers in experiment II. Despite some errors existing in ITD estimation, the pattern of BMLD as a function of frequency in each condition for the two approaches is broadly consistent. In order to quantify the differences, the BIF-weighted BMLD for each approach is defined as $\sum_{f=1}^{F=34} K_f \text{BMLD}_f$ (Lavandier and Culling, 2010). The difference Δ_{BMLD} between using

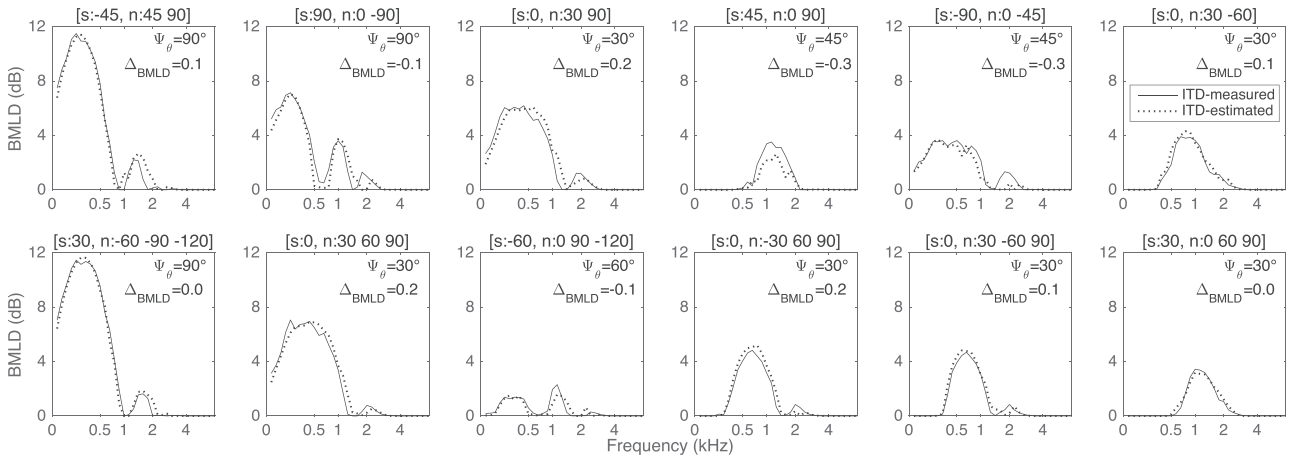

FIG. 8. BMLDs calculated using measured or estimated ITDs for the SSN maskers in experiment II. Speech and masker positions are indicted above each subplot; the separations between speech and the closest masker in each setting is displayed as Ψ_θ, with Δ_{BMLD} in dB showing the difference between BIF-weighted BMLDs of estimated and measured ITDs.

estimated and measured ITDs are also displayed in Fig. 8. Overall, the absolute Δ_{BMLD} is under 0.3 dB, with a mean of 0 dB across all 12 conditions. These findings are commensurate with predictions made by BiDWGP to the two forms of input.

C. The role of BMLD in the BiDWGP metric

To further clarify the role of BMLDs, the BiDWGP metric was recomputed after excluding the BMLD term from Eq. (5). For the single-masker conditions of experiment I, both forms of input produced overall correlations of $\rho = 0.92$ compared to 0.95 with the BMLD component. In the face of multiple maskers, the exclusion of BMLD led to falls to 0.87 from a value of 0.91 with BMLD. The reduction in predictive power echoes the finding that better-ear glimpsing alone appears not to fully account for spatial release from masking (Glyde *et al.*, 2013). In contrast, Lavandier and Culling (2010) demonstrated that in their model the BMLD component alone was not able to fully account for listeners' SRTs in reverberant conditions. However, the effect of the BMLD components to the correlation between listeners' keyword identification scores and BiDWGP predictions in this study is relatively small.

D. Limitation and extension

Tang *et al.* (2016) evaluated the BiDWGP metric with binaural input in reverberant only and reverberant noisy (SSN or CS) environments in which three rooms with different reverberation times (RTs) were simulated. Compared to the standard intelligibility metrics with their binaural extensions, such as the binaural SII (Zurek, 1993) and the BiSTI (van Wijngaarden and Drullman, 2008), the BiDWGP metric has demonstrated more robust predictive power across all tested conditions. However, in its current form, BiDWGP with source + location input is not able to account for the reverberation effect to intelligibility as anechoic monophonic recordings used as inputs by the metric do not carry any room acoustic information.

In contemporary sound design, artificial reverberation has been frequently added to audio scenes in order to increase realistic and immersive listening experience to listeners (see Välimäki *et al.*, 2012, for review). With dialogue intelligibility as a concern, it would be very useful to also take the reverberation effect to intelligibility into account at the stage of sound design, at least as the first approximation. It is worth noting that this is to only consider the intelligibility issue within an audio scene, given that the source signals and metadata (e.g., intended distance, azimuth, and RT, etc.) of all sounds are available. Intelligibility affected by actual listening environments is unlikely to be precisely predicted in practice at this stage, as the space in which a given audio scene is played to listeners may vary largely. Rennies *et al.* (2011) presented three different approaches to enhance the predictive accuracy of the model proposed by Beutelmann *et al.* (2010) in reverberation. Two of the approaches are to effectively use modulation transfer function (MTF; IEC 60268-16:2011, 2011) or the early-to-overall reflection energy ratio of BRIR (ISO 3382-1, 2009) as a correction factor to weight the apparent SNR. The third is to primarily modify the speech signal by convolving it to the early part of the BRIR, and the noise signal by adding the speech signal convolved by the late reflection to it, before feeding into the model. The evaluation suggested that all three methods produced closer matches to measured SRT than the original model. By adding a weighting function to the calculation of the early-to-overall reflection energy ratio, Rennies *et al.* (2014) further improved the model performance, especially with the third method introduced above.

Providing a parametrised approach offered by BiDWGP, i.e., using source location (distance and azimuth) relative to the listener, the notion of predicting intelligibility from source + location could also be further extended by including some room acoustic information, such as RT, as parameters. For instance, IEC 60268-16:2011 (2011) provides an equation for the MTF as a function of RT for a given modulation frequency. If the MTF for each frequency band could be integrated with the distortion weighting W in Eq. (7), it may be possible for BiDWGP with source + location input to further

include a component accounting for the effect of reverberation. This needs further investigation.

VII. CONCLUSIONS

This paper proposes an OIM based on weighting and combining glimpses at the output of a simulation of binaural processing. The metric predicts binaural intelligibility for a range of speech target and masker combinations in stationary and fluctuating maskers with listener-metric correlations in excess of 0.91. The metric operates with either binaural signals or single-channel source signals together with their locations, and is applicable to a range of sound generation scenarios in which the intelligibility of speech in a background of spatially located maskers is required.

ACKNOWLEDGMENTS

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (Grant No. EP/L000539/1) and the British Broadcasting Company (BBC) as part of the BBC Audio Research Partnership. Data underlying the findings are fully available without restriction from https://dx.doi.org/10.17866/rd.salford.3549774.

ANSI (**1997**). S3.5, "Methods for the calculation of the Speech Intelligibility Index" (Acoustical Society of America, New York).
Barker, J., and Cooke, M. (**2007**). "Modelling speaker intelligibility in noise," Speech Commun. **49**, 402–417.
Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331–342.
Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model," J. Acoust. Soc. Am. **127**, 2479–2497.
Brungart, D. S., and Iyer, N. (**2012**). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers," J. Acoust. Soc. Am. **132**, 2545–2556.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.
Collin, B., and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers," J. Acoust. Soc. Am. **134**, 1146–1159.
Cooke, M. (**1993**). *Modelling Auditory Processing and Organisation* (Cambridge University Press, Cambridge, UK).
Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.
Cosentino, S., Marquardt, T., McAlpine, D., Culling, J. F., and Falk, T. H. (**2014**). "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," J. Acoust. Soc. Am. **135**, 796–807.
Culling, J. F., Hawley, M. L., and 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**, 1057–1065.
Culling, J. F., Hawley, M. L., and 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**, 552.
Dirks, D. D., and Wilson, R. H. (**1969**). "The effect of spatially separated sound sources on speech intelligibility," J. Speech Hear. Res. **12**, 5–38.
Drullman, R. (**1995**). "Speech intelligibility in noise: Relative contributions of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.
Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**, 1206–1218.
Durlach, N. I. (**1972**). "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory Vol. II*, edited by J. V. Tobias (Academic, New York).
Falk, T., and Chan, W.-Y. (**2008**). "A non-intrusive quality measure of dereverberated speech," in *IEEE Proceedings of the International Workshop on Acoustic Echo and Noise Control*, pp. 978–989.
Glyde, H., Buchholz, J., Dillon, H., Best, V., Hickson, L., and Cameron, S. (**2013**). "The effect of better-ear glimpsing on spatial release from masking," J. Acoust. Soc. Am. **134**, 2937–2945.
Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). "Speech intelligibility and localization in a multi-source environment," J. Acoust. Soc. Am. **105**, 3436–3448.
Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.
Hirsh, I. J. (**1950**). "The relation between localization and intelligibility," J. Acoust. Soc. Am. **22**, 196–200.
Hirsh, I. J. (**1971**). "Masking of speech and auditory localization," Audiology **10**(2), 110–114.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.
IEC 60268-16:2011 (**2011**). *Part 16: Objective rating of speech intelligibility by speech transmission index* (International Electrotechnical Commission, Geneva, Switzerland), Sound System Equipment (fourth ed.).
ISO 3382-1 (**2009**). "Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces" (International Organization for Standardization, Geneva, Switzerland).
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" (International Organization for Standardization, Geneva, Switzerland).
Jelfs, S., Culling, J. F., and Lavandier, M. (**2011**). "Revision and validation of a binaural model for speech intelligibility in noise," Hear. Res. **275**, 96–104.
Kock, W. E. (**1950**). "Binaural localization and masking," J. Acoust. Soc. Am. **22**, 801–804.
Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157–167.
Lavandier, M., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms," J. Acoust. Soc. Am. **127**(1), 387–399.
Levitt, H., and Rabiner, L. R. (**1967**). "Predicting binaural gain in intelligibility and release from masking for speech," J. Acoust. Soc. Am. **42**, 820–829.
Litovsky, R., Colburn, H., and Yost, W. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**, 1633–1654.
Litovsky, R., Parkinson, A., and Arcaroli, J. (**2009**). "Spatial hearing and speech intelligibility in bilateral cochlear implant users," Ear Hear. **30**(4), 419–431.
Ma, J., Hu, Y., and Loizou, P. C. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.
Mapp, P. (**2008**). "Designing for speech intelligibility," in *Handbook for Sound Engineers*, 4th ed. (Focal, Oxford), pp. 1385–1414.
Miller, G., and Licklider, J. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulas for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (**1988**). "The shape of the ear's temporal window," J. Acoust. Soc. Am. **83**, 1102–1116.
Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS Final Report: The Auditory Filterbank," Technical Report **2341**, Medical Research Council (MRC) Applied Psychology Unit.
Rennies, J., Brand, T., and Kollmeier, B. (**2011**). "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," J. Acoust. Soc. Am. **130**, 2999–3012.
Rennies, J., Warzybok, A., Brand, T., and Kollmeier, B. (**2014**). "Modeling the effects of a single reflection on binaural speech intelligibility," J. Acoust. Soc. Am. **135**, 1556–1567.

Rhebergen, K. S., and 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**, 2181–2192.
Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Sauert, B., and Vary, P. (**2010**). "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *Proc. ITG-Fachtagung Sprachkommunikation* (Bochum, Germany).
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shaw, E., and Vaillancourt, M. M. (**1985**). "Transformation of sound pressure level from the free field to the eardrum presented in numerical form," J. Acoust. Soc. Am. **78**, 1120–1123.
Sonnenscheinn, D. (**2001**). *Sound Design: The Expressive Power of Music, Voice and Sound Effects in Cinema* (Michael Wiese Productions, CA).
Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Taal, C., Hendriks, R. C., and Heusdens, R. (**2014**). "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," Comput. Speech Lang. **28**, 858–872.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2010**). "A short time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. 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., and Cooke, M. (**2010**). "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, pp. 1636–1639.
Tang, Y., Cooke, M., and Valentini-Botinhao, C. (**2016**). "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," Comput. Speech Lang. **35**, 73–92.
Tang, Y., Hughes, R. J., Fazenda, B. M., and Cox, T. J. (**2016**). "Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms," Speech Commun. **82**, 26–37.
University College London, Cambridge University, Edinburgh University, the Speech Research Unit and the National Physical Laboratory (**1992**). "SCRIBE—Spoken corpus of British English," available at http://www.phon.ucl.ac.uk/resource/scribe (Last viewed October 19, 2009).
Välimäki, V., Parker, J. D., Savioja, L., Smith, J. O., and Abel, J. S. (**2012**). "Fifty years of artificial reverberation," IEEE Trans. Audio Speech Lang. Process. **20**, 1421–1448.
van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J. Acoust. Soc. Am. **123**, 4514–4523.
Wierstorf, H., Geier, M., Raake, A., and Spors, S. (**2011**). "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *130th Convention of the Audio Engineering Society*.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance* (Allyn and Bacon, Needham Heights, MA), pp. 255–276.