

1 A non-intrusive method for estimating binaural speech
2 intelligibility from noise-corrupted signals captured by a
3 pair of microphones

4 Yan Tang^{a,*}, Qingju Liu^b, Wenwu Wang^b, Trevor J. Cox^a

5 ^a*Acoustics Research Centre, University of Salford, UK*

6 ^b*Centre for Vision, Speech and Signal Processing, University of Surrey, UK*

7 **Abstract**

A non-intrusive method is introduced to predict binaural speech intelligibility in noise directly from signals captured using a pair of microphones. The approach combines signal processing techniques in blind source separation and localisation, with an intrusive objective intelligibility measure (OIM). Therefore, unlike classic intrusive OIMs, this method does not require a clean reference speech signal and knowing the location of the sources to operate. The proposed approach is able to estimate intelligibility in stationary and fluctuating noises, when the noise masker is presented as a point or diffused source, and is spatially separated from the target speech source on a horizontal plane. The performance of the proposed method was evaluated in two rooms. When predicting subjective intelligibility measured as word recognition rate, this method showed reasonable predictive accuracy with correlation coefficients above 0.82, which is comparable to that of a reference intrusive OIM in most of the conditions. The proposed approach offers a solution for fast binaural intelligibility prediction, and therefore has practical potential to be deployed in situations where on-site speech intelligibility is a concern.

8 *Keywords:* objective intelligibility measure, non-intrusive, binaural
9 intelligibility, noise, glimpsing, neural network, blind source separation,
10 blind source localisation, microphone

*Corresponding author at: Acoustics Research Centre, University of Salford, UK
Email address: y.tang@salford.ac.uk (Yan Tang)

1. Introduction

Objective intelligibility measures (OIMs) have been widely used in the place of subjective listening tests for speech intelligibility evaluation, due to their fast but cheap operation and the reliable feedback they provide. In fields such as telephony quality assessment (Fletcher, 1921; ANSI S3.5, 1997), acoustics design (Houtgast and Steeneken, 1985; IEC, 2011), audiology for hearing impairment (Holube and Kollmeier, 1996; Santos et al., 2013) and algorithm development for speech enhancement and modification (Taal et al., 2010; Gomez et al., 2012), OIMs have been playing an important role for nearly a century. More recently, in order to promote their usability in more realistic listening situations, work on OIM development has focused on improving their predictive performance in conditions such as additive noise (Rhebergen and Versfeld, 2005; Jørgensen et al., 2013; Tang and Cooke, 2016) and reverberation (Rennies et al., 2011; Tang et al., 2016c). Other work has enabled them to predict intelligibility from binaural listening (van Wijngaarden and Drullman, 2008; Jelfs et al., 2011; Andersen et al., 2015; Tang et al., 2016a).

To predict speech intelligibility in noise, the clean speech signal is an essential input required by the OIMs for detailed analyses and comparisons against the noise-corrupted speech signal. Some OIMs alternatively use a separate noise signal to operate (e.g. ANSI S3.5, 1997; Tang and Cooke, 2016). This class of OIMs therefore are referred to as *intrusive* OIMs, and all the aforementioned OIMs fall into this category. In strictly controlled or experimental conditions, the clean speech signal is usually known and accessible, hence intelligibility estimation can be readily performed using an intrusive OIM. However, in situations such as live broadcasting in public crowds, where the speech signal has already been contaminated by any non-target background sounds or the clean speech reference is not available, predicting intelligibility consequently becomes problematic. This therefore greatly limits the use of this class of OIMs. In contrast to intrusive OIMs, those which operate directly on noise-corrupted speech signals are known as *non-intrusive* OIMs.

1.1. A review of non-intrusive OIMs

In early studies, non-intrusive OIMs were based on automatic speech recognition (ASR) techniques. Holube and Kollmeier (1996) proposed an approach to predict hearing-impaired listeners' recognition rate on consonant-

1 vowel-consonant (VCV) words corrupted by continuous speech-shaped noise
2 (SSN). The dynamic-time-warping (DTW) ASR recogniser (Sakoe and Chiba,
3 1978) used in their system was trained using the outputs of an auditory model
4 (Dau et al., 1996) as the features. During prediction, the DTW recogniser
5 made a decision based on the similarity between all possible responses and
6 the test word. Jurgens and Brand (2009) further adopted this approach with
7 a modulation filter bank (Dau et al., 1997) added at the stage of feature
8 extraction for better modelling of human auditory processing. Based on a
9 different theory, Cooke (2006) proposed a glimpsing model to simulate hu-
10 man speech perception in noise. The model consists of two parts: the front-
11 end glimpse detector and a back-end Hidden Markov model (HMM)-based
12 missing-data ASR recogniser. Because the missing-data recogniser requires
13 a glimpse mask computed from separate speech and masker signals, strictly
14 speaking the glimpsing model is not a non-intrusive OIM. More recently, Ger-
15 avanchizadeh and Fallah (2015) extended the system of Holube and Kollmeier
16 (1996) by introducing a unit that accounts for the better-ear (BE) advantage
17 and binaural unmasking (BU) in binaural listening. They used the system to
18 predict listeners’ speech reception threshold (SRT) when the target speech
19 and masking sources were spatially separated on a horizontal plane.

20 The ASR-based OIMs normally comprise the feature extraction and ASR
21 components. Indeed, they can provide detailed modelling of speech percep-
22 tion in noise and make phoneme-level intelligibility predictions compared to
23 word- and sentence-level predictions offered by normal intrusive OIMs. This
24 permits, for example, more transparent and profound analyses to be per-
25 formed on the model’s errors. Therefore, they are also known as *microscopic*
26 OIMs. However, knowing exactly what constants and vowels a listener may
27 misperceive is unnecessary in many practical situations where a simple intel-
28 ligibility estimate is sufficient. In addition, except for the glimpsing model
29 (Cooke, 2006), all the microscopic OIMs mentioned above were only evalu-
30 ated in speech-shaped noise (SSN). Their performance in more commonly-
31 occurring noise conditions (e.g. fluctuating noise) was not investigated. Al-
32 though an ASR can be trained for any target noise masker, deploying an
33 ASR is onerous, especially for a robust ASR system.

34 With the facilitation of machine learning techniques, other non-intrusive
35 OIMs were also proposed. Inspired by the Low Complexity Speech Quality
36 Assessment method (Grancharov et al., 2006), Sharma et al. (2010) suggested
37 an algorithm, the Low Cost Intelligibility Assessment (LCIA), for predict-
38 ing intelligibility from noise-corrupted speech signal. LCIA uses a Gaussian

1 mixture model (GMM) to generate the predictive score from frame-based
2 features, such as spectral flatness, spectral centroid, excitation variance and
3 spectral dynamics. As the GMM model is trained using a supervised ap-
4 proach with the measured subjective intelligibility score as the desired out-
5 put, which is expensive and time-consuming to collect, it is difficult for this
6 approach to be generalised for a wider range of conditions, in spite of the
7 high correlation with the subjective data in the testing conditions.

8 One solution to overcome the lack of subjective training data is to use
9 objective intelligibility score provided by an established OIM as the target
10 output. Usually the performance of an established OIM was rigorously evalu-
11 ated in previous studies by comparing its predictions to subjective data, it is
12 expected to be able to provide reasonable estimation on subjective intelligibil-
13 ity. Li and Cox (2003) trained a neural network on the Speech Transmission
14 Index (STI, IEC, 2011) from the low frequency envelope spectrum of run-
15 ning speech, to predict intelligibility. Sharma et al. (2016) further improved
16 LCIA and extended it to both speech quality and intelligibility predictions.
17 In terms of intelligibility, the GMM used in the enhanced version of LCIA,
18 renamed as the Non-Intrusive Speech Assessment (NISA), was trained on the
19 predictive scores of the short-time objective intelligibility (STOI, Taal et al.,
20 2010), which was validated to show good match to the subjective data mea-
21 sured in Hilkhuisen et al. (2012). Despite extensive objective evaluations
22 performed, the NISA was regrettably not further evaluated using subjective
23 data. This leaves the question of whether the high correlation with the objec-
24 tive scores can be translated to a good match with subjective intelligibility
25 unanswered. There is some evidence (Tang and Cooke, 2012; Tang et al.,
26 2016b) suggesting that STOI lacks predictive accuracy when making predic-
27 tions for algorithmically-modified speech or across different types of maskers.

28 Based on full-band clarity index C50 (Naylor and Gaubitch, 2010), a
29 data-driven non-intrusive room acoustic estimation method for predicting
30 ASR performance in reverberant conditions was introduced (Peso Parada
31 et al., 2016). On the other hand, rather than a direct feature-score mapping,
32 Karbasi et al. (2016) sought to cater for intrusive OIMs by reconstructing the
33 clean speech signal from the noise-corrupted signal, using a speech synthesiser
34 based on a twin HMMs. With STOI as the back-end intelligibility predictor,
35 the proposed system can achieve comparable performance to STOI, when
36 used in its ordinary intrusive manner. Indeed, this approach permits almost
37 all intrusive OIMs to serve for the purpose of blind intelligibility prediction.
38 However, it also faces a similar issue that the ASR-based OIMs encounter: it

1 is difficult to build a synthesiser without access to a large amount of resources
2 including speech corpora accompanied by transcriptions.

3 A non-machine learning-based metric was proposed by Falk et al. (2010).
4 It can predict speech intelligibility in conditions including noisy, reverberant
5 and the combination of the former two based on speech-to-reverberation mod-
6 ulation energy ratio (SRMR). Santos and Falk (2014) extended this method
7 to predict intelligibility for hearing-impaired listeners by limiting the range
8 of modulation frequencies and applying a threshold to the modulation en-
9 ergy. Furthermore, the binaural extensions were also introduced to SRMR
10 by Cosentino et al. (2014), so that SRMR can be further used to predict
11 SRT when a listener listens binaurally. While SRMR has been reported to
12 deal well with conditions where stationary noise (e.g. SSN) was mostly used,
13 its predictive power may be limited in fluctuating maskers such as modu-
14 lated and babble noises. These fluctuating maskers can not only reduce the
15 modulation depth of the speech signal, but also introduce stochastic distur-
16 bance to speech modulation (Dubbelboer and Houtgast, 2007). The latter
17 effect does not necessarily always lead to increased energy at high modulation
18 frequencies.

19 *1.2. Overview of this work*

20 In this study, a framework for predicting binaural speech intelligibility
21 from noise-corrupted signals captured by a pair of closely-spaced microphones
22 is proposed. In practice, all the aforementioned non-intrusive OIMs assume
23 that the binaural signals are directly accessible from a head and torso sim-
24 ulator, or can be simulated using existing head-related transfer functions
25 (HRTFs) or binaural room impulse responses (BRIRs). For the latter case,
26 the source locations must be known to be able to choose correct HRTFs or
27 BRIRs. Therefore, this approach further intends to deal with conditions in
28 which the source locations are unknown, and consequently the binaural sig-
29 nals that a human listener perceives can not be easily simulated; the method
30 is also suitable for situations in which HRTFs and BRIR are not available
31 at all. The system also aims to overcome some of the problems that the
32 state-of-the-art non-intrusive approaches encounter as reviewed above, such
33 as lacking predictive power in fluctuating noise.

34 The novelty of the proposed system is to bring together techniques in-
35 cluding blind-source separation (BSS), blind-source localisation (BSL), and
36 intrusive binaural intelligibility prediction. The BSS and BSL provide an
37 estimation of the binaural signals of both the speech and the masker signal,

1 and hence allows the intrusive OIM to calculate the speech intelligibility.
 2 Therefore, similar to the approach of Karbasi et al. (2016), the framework
 3 allows any component in the proposed system to be replaced by counter-
 4 parts if that is desired. As a proof of concept, the components adopted in
 5 the current study were optimised for their best performance.

6 This paper is organised as follows. In Section 2, the proposed framework
 7 and each component are introduced. Section 3 focuses on evaluating the per-
 8 formance of the proposed system by comparing its intelligibility predictions
 9 to listener performance measured from two listening experiments. The as-
 10 pects which potentially influence the system performance are then analysed
 11 and discussed in Section 5. Conclusions are drawn in Section 6.

12 2. Proposed system

13 Fig. 1 illustrates the pipeline of the proposed system. In order to capture
 14 the signals heard by the listener, a pair of microphones are placed at the
 15 listener’s position. The speech-plus-noise mixture, $s + n$, is then processed
 16 by a BSS model, which is trained using a deep neural network (DNN), to
 17 estimate the signals of the speech s' and masker n' sources separately (Section
 18 2.1). The two-channel mixtures are also fed as the inputs into a BSL model
 19 (Section 2.2) to calculate the approximate locations of the speech θ'_s and the
 20 masker θ'_n , which are then used to estimate the head-induced interaural level
 21 differences (ILD) of the binaural signals. Early studies (Hawley et al., 2004;
 22 Culling et al., 2004) have suggested that head-shadowing plays an important
 23 role in binaural speech intelligibility in noise (Hawley et al., 2004; Culling
 24 et al., 2004). Because the signals captured by the microphones do not contain

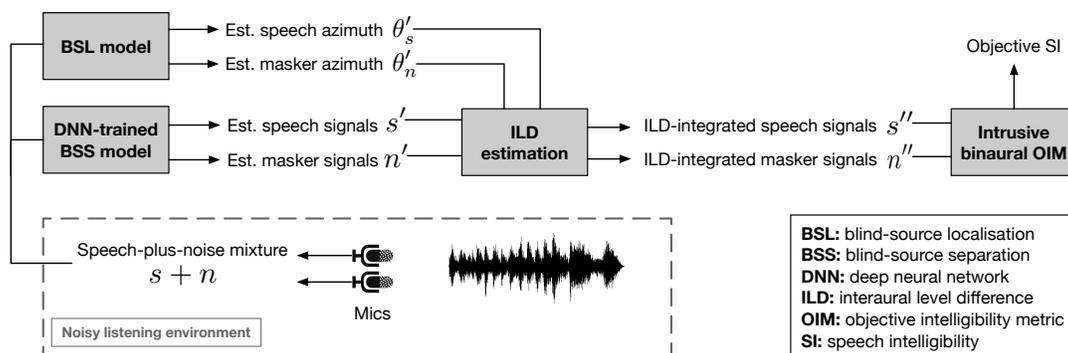


Figure 1: Schematic of the proposed system

1 head shadowing, it needs to be modelled in the binaural signals using the
 2 estimated ILD (Section 2.3) before they are passed to the intrusive OIM for
 3 intelligibility prediction. Finally, the chosen intrusive binaural OIM (Section
 4 2.4) makes predictions from the ILD-rectified speech and masker signals, s''
 5 and n'' .

6 2.1. Blind source separation using deep neural network

7 The BSS component extracts both the underlying speech and the noise
 8 signals from their mixtures $s + n$, as illustrated in Fig. 1. Traditional BSS
 9 methods have been carried out in the field of sensor array signal processing
 10 (Jutten and Herault, 1991; Comon, 1994; Mandel et al., 2010; Alinaghi et al.,
 11 2014; Virtanen, 2007). Recently, DNNs have achieved state-of-the-art per-
 12 formance in speech source separation (Grais et al., 2014; Huang et al., 2015;
 13 Nugraha et al., 2016; Yu et al., 2016) and enhancement/denoising (Xu et al.,
 14 2014; Liu et al., 2014; Weninger et al., 2015), and thus are exploited in the
 15 proposed system.

16 We employed the classic multilayer perceptron structure with three hid-
 17 den layers, each of which consists of 3000 rectified linear units. The DNN
 18 performs in the time-frequency (T-F) domain after short time Fourier trans-
 19 forming (STFT), whose input $\mathbf{x}(t)$ is a super vector consisting of the con-
 20 catenated log-power (LP) spectra from 11 neighbouring frames centred at
 21 the t -th frame, and the output vector $\hat{\mathbf{y}}(t)$ is the ideal ratio mask (IRM) as-
 22 sociated with the target speech. Denoting the LP of the ground-truth target
 23 and the estimated target as $S^{\text{LP}}(t, f)$ and $\hat{S}^{\text{LP}}(t, f)$ respectively, the weighted
 24 square error was used as the cost function during the DNN training:

$$\sum_{t,f} w \left(\hat{S}^{\text{LP}}(t, f), S^{\text{LP}}(t, f) \right) \left(\hat{S}^{\text{LP}}(t, f) - S^{\text{LP}}(t, f) \right)^2. \quad (1)$$

25 Motivated by mechanisms of existing perceptual evaluation metrics (Rix
 26 et al., 2001; Huber and Kollmeier, 2006), the adopted perceptual weight
 27 w is a balance between suppressing low energy components and boosting
 28 high energy components of the original speech signal, as well as suppressing
 29 distortions introduced in the estimated signal s' ,

$$w \left(\hat{S}^{\text{LP}}(t, f), S^{\text{LP}}(t, f) \right) = \psi(S^{\text{LP}}(t, f)) + (1 - \psi(S^{\text{LP}}(t, f)))\psi(\hat{S}^{\text{LP}}(t, f)). \quad (2)$$

30 In the above equation, $\psi(\cdot)$ is a sigmoid function $\psi(S) = \frac{1}{1 + \exp(-(S - \mu)/\sigma)}$, with
 31 the translation parameter $\mu = -7$ and scaling parameter $\sigma = 0.5$.

1 Standard back-propagation was performed during the DNN training with
 2 root mean square propagation optimisation (Tieleman and Hinton, 2012).
 3 The dropout was set to 0.5 in order to avoid over-fitting (Srivastava et al.,
 4 2014). The DNN output $\hat{y}(t, f)$ – the IRM associated with the target speech
 5 – can be applied to the mixture spectrum directly, followed by the inverse
 6 STFT to recover the waveform of the target speech source s' in the time
 7 domain. Similarly, the estimated masker signal n' can be obtained using the
 8 separation mask $1 - \hat{y}(t, f)$.

9 2.2. Blind source location estimation

10 The spatial locations of both target and masking sources affect the lis-
 11 tener’s binaural intelligibility, due to different head-shadow effects. In order
 12 to recover the ILD to account for this (Section. 2.3), the locations of the
 13 sources need to be estimated from the captured mixture $s + n$. To localise
 14 the sources from stereophonic recordings, some binaural acoustic features
 15 have proved to be useful. Three groups of audio localisation cues are of-
 16 ten used: high-resolution spectral covariance, time delay of arrival (TDOA)
 17 at microphone pairs, and steered response power (Asaei et al., 2014). The
 18 first group is sensitive to outliers, e.g. the multiple signal classification algo-
 19 rithm (Schmidt, 1986), while the third group often requires a large number
 20 of spatially-distributed microphones. TDOA cues have been widely used
 21 in speaker tracking (Vermaak and Blake, 2001; Lehmann and Williamson,
 22 2006; Ma et al., 2006; Fallon and Godsill, 2012) and are applicable for bin-
 23 aural recordings. Therefore, a BSL method based on TDOA (Blandin et al.,
 24 2012) is employed in the proposed system.

25 TDOA cues can be obtained by comparing the difference between the
 26 stereophonic recordings captured by a pair of microphones. This can be
 27 performed by identifying the peak positions from the angular spectra, us-
 28 ing generalised cross correlation (GCC) (Knapp and Carter, 1976) function.
 29 Blandin et al. (2012) demonstrated that a phase-transform GCC (PHAT-
 30 GCC) function is able to provide more robust estimation on TDOA against
 31 noise. Let $X_L(t, f)$ and $X_R(t, f)$ denote the STFTs of a pair of stereophonic
 32 signals at T-F location (t, f) . The PHAT-GCC can be calculated,

$$C_t(\tau) = \sum_f \frac{X_L(t, f)X_R^*(t, f)}{|X_L(t, f)X_R^*(t, f)|} e^{j2\pi \frac{fF_s}{\Omega} \tau} \quad (3)$$

33 where τ and F_s are the candidate delay and the sampling frequency, respec-
 34 tively. * denotes the complex conjugate. Assuming the mixing process is

1 time-invariant, a pooling process can be applied over all the frames via the
 2 direct summation $C(\tau) = \sum_t C_t(\tau)$. The peak positions in $C(\tau)$ indicate the
 3 TDOA cues.

4 The maximum TDOAs between the two microphones are then calculated
 5 based on sound velocity and distance between the two microphones. Using a
 6 linear interpolation between the two maximum delays (positive and negative),
 7 the candidate delays can be set with a linear grid, which can be further
 8 mapped to the estimated input angles θ' in the range of $[-90^\circ, 90^\circ]$.

9 *2.3. Integration of head-induced binaural level difference*

10 Before making intelligibility prediction from the BSS-estimated speech s'
 11 and masker n' signals, the head-induced ILD needs to be recovered for both s'
 12 and n' using their corresponding locations θ'_s and θ'_n determined by the BSL
 13 component (Section 2.2). Many studies (e.g. Hirsh, 1950; Durlach, 1963a,
 14 1972; Hawley et al., 2004; Culling et al., 2004) have revealed that ILD and
 15 interaural time difference (ITD) are the two prominent factors that affect
 16 intelligibility in binaural listening. As noted before, each of the originally
 17 captured mixture signals, $s + n$, lacks the effect of head-shadowing that gives
 18 ILD cues. Despite preserved ITD cues in $s + n$, studies (e.g. Lavandier and
 19 Culling, 2010) have suggested that binaural unmasking due to ITD alone
 20 cannot fully account for the spatial release from masking when the target
 21 and masking sources are spatially separated. In their binaural intelligibility
 22 modelling, Tang et al. (2016a) found that ILD plays an even more important
 23 role than ITD. This will be further discussed in Section 5.4.

24 Similar to the approach in Zurek (1993), the left s'^L and right s'^R channel
 25 of the estimated speech signal s' is processed by a bank of 55 gammatone
 26 filters, whose centre frequencies lie in the range between 100 to 7500 Hz on the
 27 scale of equivalent rectangle band (Moore and Glasberg, 1983). As expressed
 28 by Equation 4, the output of each filter $s'(f)$ is scaled by an azimuth- and
 29 frequency-dependent gain $k(f, \theta'_s)$, which is converted from the difference in
 30 sound pressure level between each ear and the listener’s frontal position, P ,
 31 in decibels.

$$s''(f) = k(f, \theta'_s) \cdot s'(f) \quad (4)$$

32 where

$$k(f, \theta'_s) = 10^{P(f, \theta'_s)/20}$$

33 Given a frequency f and a source location θ , $P_L(f, \theta)$ for the left ear
 34 of the listener can be directly interpolated using a transformation of sound

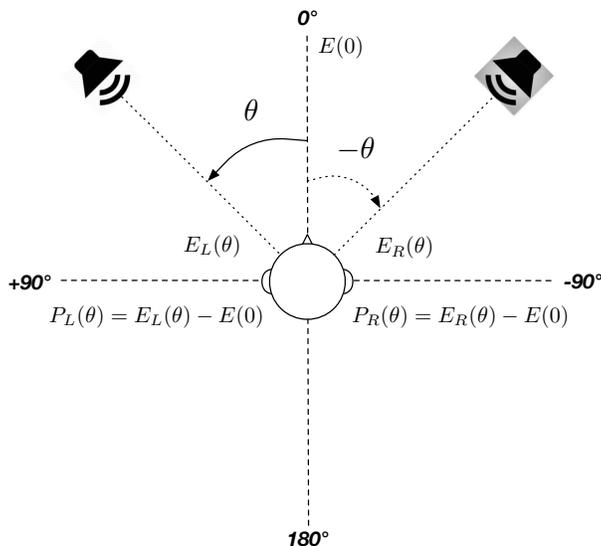


Figure 2: Difference ($P_L(\theta)$, $P_R(\theta)$) in sound pressure level between the left ear $E_L(\theta)$ and the listener’s frontal position $E(0)$, and between the right ear $E_R(\theta)$ and $E(0)$ respectively, when the source is at an azimuthal position θ on a horizontal plane. The left-right image source of the target is also shown at $-\theta$ in the grey square.

1 pressure level from the free field to the eardrum (see Table I in Shaw and
 2 Vaillancourt, 1985). As illustrated in Fig. 2, for the right ear P_R can be
 3 derived by assuming that the hearing abilities of the two ears of a normal
 4 hearing listener are symmetric, such that

$$P_R(f, \theta) = P_L(f, -\theta) = P_L(f, 360 - \theta) \quad (5)$$

5 The final ILD-rectified speech signal s'' is the sum of the scaled outputs
 6 of all the 55 filters. The RMS energy of $[s''_L, s''_R]$ is renormalised to that of
 7 $[s'_L, s'_R]$ to eliminate any changes in energy caused by the signal processing.
 8 The estimated noise signal n' is processed by the same procedure to generate
 9 the ILD-rectified masker signal n'' .

10 2.4. Back-end binaural intelligibility predictor

11 In principle, any binaural OIM may be used at the end of the pipeline
 12 to predict the intelligibility from the outputs of the ILD-rectification stage
 13 (Section 2.3). Liu et al. (2016) investigated three binaural OIMs: binaural
 14 STI (van Wijngaarden and Drullman, 2008), binaural Speech Intelligibility

1 Index (Zurek, 1993) and the binaural distortion-weighted glimpse proportion
2 (BiDWGP, Tang et al., 2016a), examining the correlation between the met-
3 rics and perceptual measurements of speech intelligibility. When the error in
4 speech-to-noise ratio (SNR) estimation due to the BSS processing was com-
5 pensated for, BiDWGP showed the least difference from its corresponding
6 benchmark performance, which was calculated from the known direct speech
7 and masker signals.

8 BiDWGP predicts intelligibility by quantifying the local audibility of T-F
9 regions, as ‘glimpses’ (Cooke, 2006), on the speech signal, and the effect of
10 masker- or reverberation-induced perturbations on the speech envelope. To
11 model binaural listening, glimpses and the frequency-dependent distortion
12 factors are computed for both ears. The binaural masking level difference
13 (Levitt and Rabiner, 1967) accounting for the BU effect is integrated at the
14 stage where the glimpses are calculated. The BE effect is then simulated by
15 combining glimpses from the two ears. The final intelligibility index is the
16 sum of the numbers of glimpses in each frequency band, weighted by the dis-
17 tortion factor and band importance function. As BiDWGP has demonstrated
18 more robust intelligibility predictions (correlation coefficients $\rho > 0.88$) than
19 the binaural counterparts of the standard intelligibility measures (e.g. SII:
20 $\rho > 0.69$ and STI: $\rho > 0.78$) in both anechoic (Tang et al., 2015, 2016a) and
21 reverberant noisy conditions (Tang et al., 2016c), the system performance
22 with BiDWGP as the intelligibility predictor was primarily examined in this
23 paper.

24 The binaural Short-Time Objective Intelligibility (BiSTOI, Andersen et al.,
25 2016) was also examined as the intelligibility predictor in the proposed sys-
26 tem to demonstrate the flexibility of the framework. BiSTOI extends its
27 monaural counterpart, STOI (Taal et al., 2010), which computes the pre-
28 dictive score by comparing the similarity between the clean reference speech
29 signal and the corrupted signal from T-F representations in every approxi-
30 mately 400 ms. STOI has been widely used for estimating intelligibility of
31 noisy speech and speech signals processed by speech enhancement algorithms
32 (e.g. ideal time frequency segregation). The binaural extension is essentially
33 to account for the binaural advantages using a modified model based on the
34 Equalisation-Cancellation theory (Durlach, 1963b). When estimating lis-
35 tener’s word recognition rate and SRT in conditions where a single masking
36 source was presented in the horizontal plane, BiSTOI has demonstrated good
37 predictive accuracy ($\rho > 0.95$) (Andersen et al., 2016).

1 3. Experiments

2 3.1. Preparation

3 The proposed system was evaluated in two rooms (referred to as Room
4 A and B). The dimensions and the reverberation time (RT_{60}) of the rooms
are described in Table 1.

Table 1: Dimension (*length* \times *width* \times *height*) and RT_{60} of each experimental room, and the relative distance between listener and each speech/masker source

	Dimension (m)	RT_{60} (s)	Listener-source distance (m)
Room A	$3.5 \times 3.0 \times 2.3$	0.10	1.2
Room B	$6.6 \times 5.8 \times 2.8$	0.27	2.2

5

6 3.1.1. Binaural signal generation and test materials

7 Two sets of room impulse responses (RIRs) were measured in each room.
8 The first set was recorded using a Brüel & Kjær head and torso simulator
9 (HATS) Type 4100 from a sine sweep as the excitation signal, which was
10 played back from a single GENELEC 8030B loudspeaker placed at different
11 target azimuths (0° , 15° , -30° , 60° and -90°) relative to 0° of the HATS. The
12 loudspeaker was mounted on top of a loudspeaker stand. The centre of the
13 main driver of the loudspeaker was at the same level as the ear height on
14 the HATS at approximately 1.5 m above floor-height. The distance between
15 the loudspeaker and the HATS was fixed, as shown in Table 1, regardless of
16 the azimuthal position of the loudspeaker. The target RIR at each azimuth
17 was then acquired by linearly convolving the recording from the HATS with
18 an analytical inverse filter preprocessed from the excitation signal (Farina,
19 2000). As this set of RIRs include complete binaural cues (for ITD and ILD),
20 it is further referred to as binaural RIR (BRIR), and was used to generate
21 binaural signals that a listener hears when the source is at different locations.
22 The second set of RIRs were recorded by replacing the HATS with a pair
23 of Behringer B-5 condenser microphones fixed on a dual microphone holder,
24 while all the other settings remained the same. The distance between the two
25 microphones was 18.0 cm, which was consistent with the distance between
26 the two ears on the HATS. In contrast to the BRIRs, this set of RIRs allowed

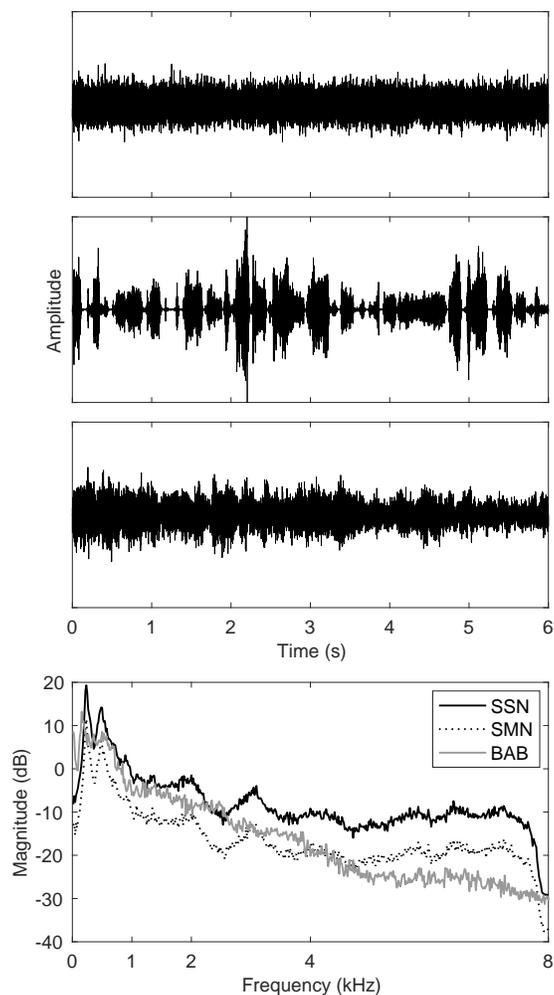


Figure 3: Sample waveform of SSN, SMN and BAB and their long-term average spectra. For illustration, the spectra of SSN and SMN are offset at ± 3 dB, respectively.

1 the creation of signals that were captured by the pair of microphones in the
 2 room. In total, four sets of RIRs were recorded and used in the subsequent
 3 work.

4 To generate binaural signals to allow the system to be assessed and also
 5 perceptual testing of intelligibility, monophonic recordings were convolved
 6 with the corresponding RIR at every target azimuthal location. The target
 7 source were speech sentences drawn from the Harvard corpus (Rothausen
 8 et al., 1969), which consists of 720 phonetically-balanced utterances produced

1 by a male British English speaker. The noise maskers included speech-shaped
2 noise (SSN), speech-modulated noise (SMN) and babble noise recorded in a
3 cafeteria (BAB), covering both stationary and fluctuating types of maskers.
4 SSN has the long-term spectrum of the speech corpus. SMN was generated
5 by applying the envelope of a speech signal randomly concatenated from
6 utterances of a female talker to the SSN. As a consequence, SMN has large
7 amplitude modulations in its waveform. Fig. 3 exemplifies the waveform of
8 each type of masker, along with their long-term average spectra displayed.
9 Both point and diffused sources were considered: while SSN and SMN were
10 treated as point sources, the diffused BAB condition was created by summing
11 the point BAB sources at all the five positions.

12 3.1.2. DNN training of the BSS model

13 From the Harvard corpus, the first 208 sequences were reserved for sub-
14 sequent objective and subjective evaluation of the system. The DNN model
15 was hence trained on the binaural signals produced from the remaining 512
16 sentences. In order to avoid the trained BSS model over-representing charac-
17 teristics of the maskers, similar to May and Dau (2014), the masker signals
18 used for training and testing were randomly drawn from two uncorrelated
19 9-min long signals for each masker. For each masker type, two different SNR
20 levels (referred to as *low* and *high*) were considered as shown in Table 2. The
21 chosen SNRs led to approximately 25% and 50% speech recognition rate for
22 listeners in a pilot test when the stimuli were presented to listeners monau-
23 rally. Note that, although the global SNRs used in model training were
24 limited (i.e. only two levels), the local SNR at each time frame or several
25 consecutive frames covered a much wider range due to the non-stationarity of
26 both the target and masker. In total, about five hours of training data were
27 generated for Room A. In order to inspect the robustness of the BSS model
28 to small changes in microphone and HATS placement, as well as to different
29 acoustics, in further evaluation no separate new BSS model was trained for
30 Room B.

31 As the DNN-trained BSS algorithm employed in the current study oper-
32 ates on a monophonic signal, the separation does not rely on any binaural
33 features such as ILD and ITD. Unlike in the previous study (Liu et al., 2016),
34 where both ILD and ITD cues were used as features, and consequently sev-
35 eral individual azimuth-dependent models were required when source loca-
36 tion changed, the advantage here is that only one universal BSS model was
37 trained regardless of the source location. While generating the input features

Table 2: SNR (dB) settings for each noise masker used in the experiments

	SSN	SMN	BAB
SNR: high	-6	-9	-4
SNR: low	-9	-12	-7

1 from the simulated binaural recordings sampled at 16 kHz, the two channels
 2 were treated independently. Each channel was first normalised, followed by
 3 512-point STFT with half-overlapped Hamming windows. After feature ex-
 4 traction, these LP features were then further normalised at each frequency
 5 bin, using frequency-dependent mean and variance calculated from all the
 6 training data. The five-hour training data was divided using a ratio of 80:20
 7 for training and validation, respectively. Both the training and validation
 8 data were randomised after each of 200 epochs.

9 3.2. System prediction

10 The proposed system made predictions from the speech-plus-noise mix-
 11 tures. As illustrated in Fig. 1, the mixture signals traverse the system pipeline
 12 from the BSS and BSL components until the back-end binaural OIM, where
 13 the objective intelligibility score is generated. The impact of each main com-
 14 ponents will be analysed and discussed in Section 5.

15 The test mixtures as the system input were generated by convolving the
 16 monophonic recording of the reserved speech sentences (i.e. not used for DNN
 17 training) and corresponding masker signals with the RIRs recorded using the
 18 pair of microphones. In the experiments the speech source was always fixed
 19 at 0° of the listener, while the location of the masking source (SSN and SMN)
 20 varied in the five target azimuths as described in Section 3.1.1. Since diffused
 21 BAB was not location-specific, it hence was considered as one azimuthal
 22 condition. In order to yield the same number of conditions as for other
 23 maskers, the BAB condition was repeated four times with different sentences.
 24 This facilitated using a balanced design in the following perceptual listening
 25 experiments (Section 3.3). The SNRs at which the speech and masker were
 26 mixed are as shown in Table 2. In total, this design led to 30 conditions (3
 27 masker types \times 2 SNRs \times 5 masker locations as described in Section 3.1.1
 28 and 3.1.2) in each room.

1 *3.3. Subjective data collection*

2 Subjective intelligibility tests were undertaken as an independent eval-
3 uation of the performance of the system. Intelligibility was measured as
4 listener’s word recognition rate. The listening tests were conducted in the
5 same 30 conditions as described in 3.2. In contrast to the speech-plus-noise
6 mixtures from which the proposed system made predictions, the stimuli for
7 the listening tests were generated using the HATS-recorded BRIRs. Experi-
8 ments took place in Room A and B with background noise levels lower than
9 15 dBA. The listener was seated at the position where the HATS and the mi-
10 crophones were placed during the RIR recording. The stimuli were presented
11 to the listener over a pair of Sennheiser HD650 headphones after being pre-
12 amplified by a Focusrite Scarlett 2i4 USB audio interface. The presentation
13 level of speech over the headphones was calibrated using an artificial ear and
14 fixed to 72 dBA; the level of the masker was consequently adjusted to meet
15 the target SNR requirement in each condition.

16 Each Harvard sentence has five or six keywords (e.g. ‘GLUE the SHEET
17 to the DARK BLUE BACKGROUND’ with keywords being capitalised).
18 Each listener heard 5 sentences in each of the 30 conditions, leading to 150
19 sentences being presented through each experiment. All the 150 sentences
20 were unique and the listener heard no sentence twice. The same 150 sen-
21 tences were used in both experiments in Room A and B. In order to min-
22 imise the effect due to the intrinsic difference on intelligibility, a balanced
23 design was used to ensure that each sentence appeared and was heard in dif-
24 ferent conditions by different listeners. The 150 sentences were blocked into
25 6 masker/SNR sessions, which were presented in a random order. The 25
26 sentences in each session were also randomised. Listeners were not allowed to
27 re-listen to each sentence. The listener was asked to type down all the words
28 that s/he could hear after each sentence was played, in a MATLAB graphic
29 programme using a physical computer keyboard. The word recognition rate
30 was finally computed only from the predefined keywords using a computer
31 script. In order to reduce counting errors, the script checked the responses
32 against a homophone dictionary and a dictionary including common typos
33 during scoring.

34 A total of 30 native British English speakers (mean 28.2 years, s.d. 3.3
35 years) from the University of Salford participated in the experiments. The
36 participants were equally divided into two groups of 15, separately taking
37 part in the experiment in Room A and B. All participants reported normal
38 hearing. Student participants were paid for their participation. The Research

1 Ethics Panel at the College of Science and Technology, University of Salford,
 2 granted ethical approval for the experiment reported in this paper.

3 4. Results

4 The system predictions are compared against the mean subjective intelli-
 5 gibility over all subjects in the 30 testing conditions in the first row of Fig. 4
 6 and 5. The performance of the proposed system was evaluated as the Pear-
 7 son and Spearman correlation coefficients, ρ_p and ρ_s , between the system
 8 outputs (as BiDWGP in Fig. 4 or BiSTOI scores in Fig. 5) and subjective
 9 intelligibility. The possible minimum root-mean square error, $RMSE_m$, be-
 10 tween subjective data and predictions converted from raw objective scores

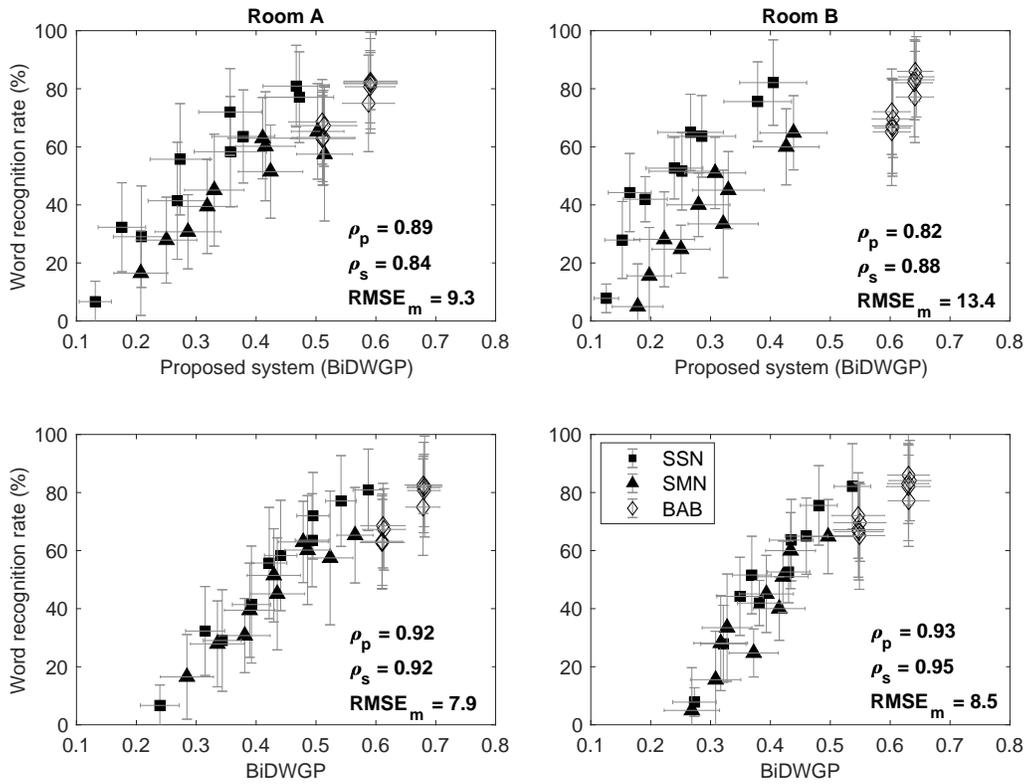


Figure 4: Objective-subjective correlation in Room A (left column) and B (right column), with reference performance provided in the second row. ρ_p , ρ_s and $RMSE_m$ are displayed for each subplot. Error bars indicate standard deviations of subjective intelligibility (vertical) and BiDWGP scores (horizontal) for each condition/data point.

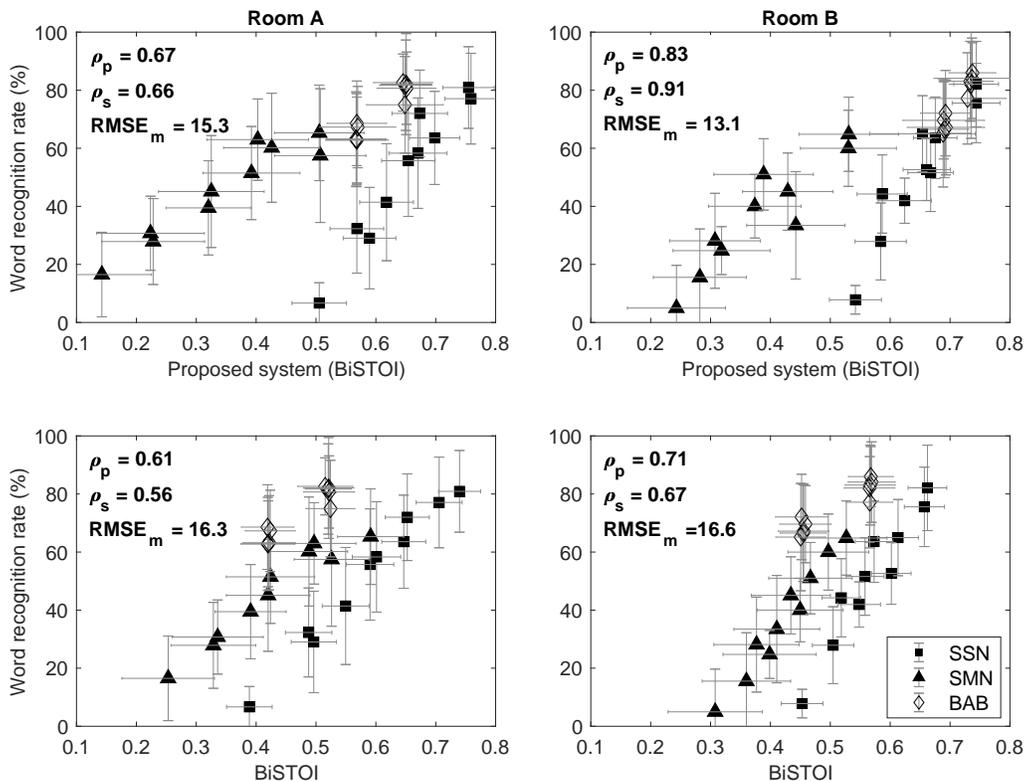


Figure 5: As for Fig. 4 but when BiSTOI is used as the intelligibility predictor.

1 using a linear fit is also computed as, $RMSE_m = \sigma_e \sqrt{1 - \rho_p^2}$, where σ_e is the
 2 standard deviation of the subjective data in a given condition.

3 As references, the performance of the BiDWGP and BiSTOI when pre-
 4 dicting from the *true* binaural speech and noise signals is also presented in
 5 the second row of Fig. 4 and 5. The input signals for the two OIMs here
 6 were the original signals used to make the speech-plus-noise mixtures for the
 7 listener tests (i.e. generated using the HATS-recorded BRIRs). As opposed
 8 to operating on the *estimated* signals (the outputs of the ILD-estimation
 9 component) in the proposed system, the reference performance is considered
 10 as the best possible performance of the OIMs. Therefore, ρ_p and ρ_s of the
 11 proposed system which are significantly higher or lower than the references,
 12 are caused by the errors in the estimated signals.

13 In Room A for which the BSS model was trained, the proposed system
 14 with BiDWGP as the predictor (Fig. 4) is able to provide similar predictive

1 accuracy ($\rho_p = 0.89$) compared to the corresponding reference performance
 2 ($\rho_p = 0.92$) [$\chi^2 = 1.219$, $p = 0.270$] in terms of the linear relationship with
 3 the subjective data. However, the reference method indeed shows better
 4 ranking ability measured as Spearman correlation ($\rho_s = 0.92$) to the subjective
 5 data than the proposed system ($\rho_s = 0.84$) [$\chi^2 = 5.507$, $p < 0.05$]. For
 6 Room B, where the BSS model trained for Room A was used, the decrease
 7 in the performance of the proposed system with BiDWGP as the predictor
 8 is evident compared to the reference [all $\chi^2 \geq 6.694$, $p < 0.05$].

9 When BiSTOI is used as the predictor (Fig. 5), both the linear rela-
 10 tionship with the subjective data ($\rho_p = 0.67$) [$\chi^2 = 0.250$, $p = 0.618$] and
 11 the ranking ability of the system ($\rho_s = 0.66$) [$\chi^2 = 0.588$, $p = 0.444$] are
 12 comparable to the reference performance in Room A. However, the reference
 13 performance of BiSTOI appears to suffer considerably from underestimating
 14 in BAB (i.e. diffused) conditions relative to the other conditions – both ρ_p
 15 and ρ_s dramatically increase to 0.84 and 0.88 respectively, with the BAB
 16 data being excluded. In addition, it can be seen from the plots in the second
 17 row of Fig. 5 that BiSTOI has a tendency of underestimating in fluctuating
 18 masker (SMN) or overestimating in stationary masker (SSN). This finding
 19 is compatible with that on STOI, which is its monaural counterpart (Tang
 20 et al., 2016b). Such masker-specific bias of BiSTOI is worsened when making
 21 predictions from the estimated binaural signals in this system. Consequently,
 22 the corresponding system performance with BiSTOI under the same situa-
 23 tion is $\rho_p = 0.61$ and $\rho_s = 0.66$. In Room B, the system performance with
 24 BAB being excluded is $\rho_p = 0.71$ and $\rho_s = 0.67$, compared to $\rho_p = 0.85$ and
 25 $\rho_s = 0.85$ as the reference performance of BiSTOI. Similar to in Room A,
 26 the predictive bias of BiSTOI becomes greater with the estimated binaural
 27 signals, resulting in the decreased overall performance.

28 Table 3 further details the performance of the proposed system with
 29 BiDWGP or BiSTOI for individual maskers in each target room, along
 30 with the reference counterparts. When BiDWGP was used, despite the
 31 declined overall predictive accuracy when making predictions across differ-
 32 ent types of maskers in Room B as observed above, the proposed system
 33 achieved similar performance to the reference method for individual maskers
 34 [all $\chi^2 \leq 2.907$, $p \geq 0.09$], except for the ranking ability for SMN in Room
 35 A [$\chi^2 = 8.865$, $p < 0.05$]. When BiSTOI was used and the overall perfor-
 36 mance is less good, the system also provided predictive accuracy for individ-
 37 ual maskers that is similar to the reference performance in most of conditions
 38 [all $\chi^2 \leq 3.851$, $p \geq 0.05$], except for both ρ_p [$\chi^2 = 3.947$, $p < 0.05$] and ρ_s

Table 3: System performance for subconditions in the target rooms evaluated as ρ_p , ρ_s and RMSE_m in percentage points (pps). For all ρ , $p < 0.001$.

	Room A				Room B			
	SSN	SMN	BAB	overall	SSN	SMN	BAB	overall
Proposed system (BiDWGP as OIM):								
ρ_p	0.95	0.93	0.95	0.89	0.93	0.94	0.94	0.82
ρ_s	0.94	0.85	0.83	0.84	0.96	0.93	0.87	0.88
RMSE_m (pps)	7.2	6.2	2.7	9.3	7.9	6.5	2.8	13.4
BiDWGP:								
ρ_p	0.98	0.95	0.95	0.92	0.96	0.93	0.93	0.93
ρ_s	0.99	0.94	0.89	0.92	0.96	0.95	0.78	0.95
RMSE_m (pps)	4.5	5.0	2.6	7.9	6.0	6.9	2.9	8.5
Proposed system (BiSTOI as OIM):								
ρ_p	0.97	0.95	0.95	0.67	0.94	0.92	0.96	0.83
ρ_s	0.96	0.90	0.76	0.66	0.90	0.90	0.93	0.91
RMSE_m (pps)	5.7	5.0	2.8	15.3	7.4	7.3	2.3	13.1
BiSTOI:								
ρ_p	0.99	0.97	0.94	0.61	0.96	0.98	0.94	0.71
ρ_s	0.99	0.96	0.65	0.56	0.98	0.98	0.90	0.67
RMSE_m (pps)	3.3	4.3	3.0	16.3	6.5	3.6	2.8	16.6

- 1 $[\chi^2 = 4.839, p < 0.05]$ for SSN in Room A, and ρ_s $[\chi^2 = 5.487, p < 0.05]$ for
- 2 SSN in Room B. Overall, for masker-specific predictions the proposed system
- 3 with both binaural predictors can provide reasonable predictive accuracy.

4 5. Discussion

- 5 In this study we proposed an approach to predict binaural speech in-
- 6 telligibility from noise-corrupted signals captured by a pair of microphones.

1 Listeners’ word recognition rate in both stationary and fluctuating noise con-
2 ditions were measured in two target rooms which differ in dimension and
3 room acoustics. In Room A, which has smaller RT than the other room and
4 which the BSS model was trained for, the proposed method with BiDWGP
5 as the intelligibility predictor can provide predictions that match the objec-
6 tive performance as close as those estimated by a reference intrusive OIM in
7 most of the conditions. In Room B, a decrease in the predictive performance
8 in some testing conditions was observed when using the same BSS model
9 that was trained for Room A. Nevertheless, the performance for individual
10 maskers still remained robust ($\rho_p \geq 0.93$, $\rho_s \geq 0.87$ and $\text{RMSE}_m \leq 7.9\%$)
11 relative to the reference performance.

12 As the proposed system consists of several components, each of which
13 may potentially influence the final predictive performance, in this section
14 further analyses on the main components of the system are performed along
15 with a discussion of their contributions.

16 *5.1. Error in SNR between BSS-estimated signals*

17 The robustness of the BSS algorithm may considerably affect the pre-
18 dictive accuracy because it determines the quality of the estimated source
19 signals that an intrusive OIM uses to make intelligibility prediction. In or-
20 der to separate the target speech and masker signals from the mixture, the
21 DNN-trained BSS model essentially estimates the IRM of the target speech.
22 If the IRM contains too much information about the masker signal, the es-
23 timated speech signal will still be noisy, while the separated masker signal
24 will be missing parts of its original constituents. This potentially leads to
25 higher SNR between the estimated signals than the original SNR, and hence
26 an overestimation of intelligibility when the back-end intelligibility predic-
27 tor makes predictions using the estimated signals. The opposite case on the
28 other hand is caused by the IRM missing too much information from the
29 target speech signal. As SNR is one of the most dominant effects affecting
30 speech intelligibility in noise, its errors in the BSS-estimated signals may lead
31 to inaccuracy in ultimate intelligibility prediction. Liu et al. (2016) inves-
32 tigated the error in SNR preservation of a binaural BSS algorithm, which
33 uses both ILD and ITD as cues for separation. They found that while the
34 interaural SNR can be well maintained by the algorithm, the overall SNR
35 between estimated speech and masker signals tended to be underestimated.
36 Consequently, decreased predictive performance was observed for all tested
37 intrusive binaural OIMs which made predictions from the BSS outputs.

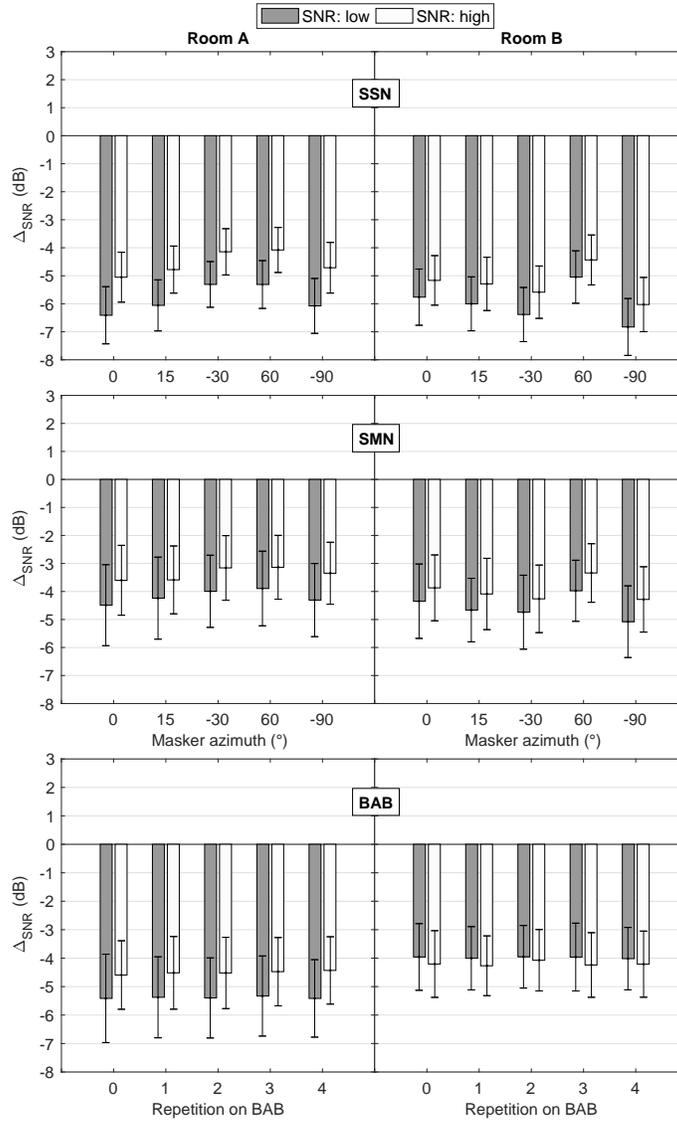


Figure 6: Difference between the target SNR and that calculated from the BSS-separated signals when the masker (SSN or SMN) is at different locations. The results for BAB are calculated from the corresponding five repeated conditions. Columns display the results for individual rooms while rows for mask types. $\Delta_{SNR} = SNR_{estimated} - SNR_{target}$. Error bars show standard deviation.

- 1 Fig. 6 displays the mean SNR error calculated as the difference between
- 2 the SNR of the BSS-estimated signals and the original target SNR over all

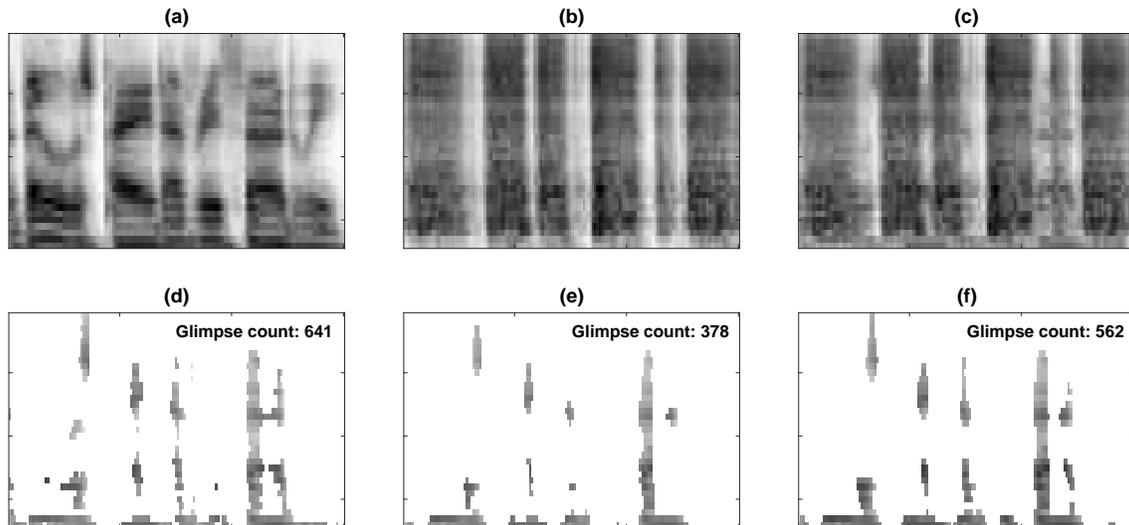


Figure 7: Spectrograms and glimpse analyses of the sentence ‘the bill was paid every third week’ in SMN at -12 dB SNR in Room A. (a): spectrogram of the clean speech signal; (b): spectrogram of the SMN signal; (c): spectrogram of the speech-plus-noise mixture; (d): glimpses calculated from the direct known speech and masker signals; (e): glimpses calculated from the BSS-estimated speech and masker signals; and (f): glimpses calculated from the BSS-estimated speech and masker signals with a gain of 4.7 dB applied to the speech signal. Glimpse count is also supplied for (d), (e) and (f).

1 speech samples when the SSN or SMN masker is at each azimuth in the target
 2 rooms. Note that for BAB the results from the five repeated conditions are
 3 presented. Similar to the findings in Liu et al. (2016), the BSS algorithm
 4 tends to underestimate the SNR with larger errors in the low SNR conditions
 5 compared to that in the high SNR for all three maskers, despite the BSS
 6 techniques used in the two studies being different. Nevertheless, the errors
 7 appear to be fluctuating around -5 dB across all the conditions and rooms,
 8 with a mean of -4.7 dB (s.d.: 0.7). This is, however, different to what has
 9 been observed in Liu et al. (2016); the extent of the overestimation in SNR
 10 varied in the source azimuthal location, presumably due to the BSS algorithm
 11 employed in the early study performed on binaural features such as ILD and
 12 ITD cues, which are functions of azimuth.

13 An example of speech corrupted by SMN masker at -12 dB SNR in Room
 14 A is shown in Fig. 7, in order to compare the glimpse constitution when
 15 the glimpses are calculated from the direct known speech and masker signals
 16 (subplot d) and from the BSS-estimated speech and masker signals (subplot

Table 4: System performance with SNR compensation. For all ρ , $p < 0.001$.

	Room A				Room B			
	SSN	SMN	BAB	overall	SSN	SMN	BAB	overall
Proposed system (BiDWGP as OIM)								
ρ_p	0.97	0.95	0.95	0.91	0.96	0.95	0.94	0.83
ρ_s	0.96	0.88	0.89	0.86	0.96	0.93	0.93	0.88
RMSE _m (pps)	6.1	5.4	2.6	8.8	6.5	5.7	2.7	13.0
Proposed system (BiSTOI as OIM)								
ρ_p	0.97	0.95	0.95	0.72	0.94	0.94	0.96	0.86
ρ_s	0.96	0.92	0.83	0.75	0.90	0.90	0.95	0.92
RMSE _m (pps)	6.1	5.0	2.6	14.3	7.3	6.6	2.3	11.8

1 e). It is worth noting that since the BSS component in fact processes the
2 signals for each ear independently, the graphs are plotted using only the left
3 channel of the chosen binaural signal. In both cases, it is clearly illustrated
4 that in the time domain glimpses are largely produced in the gaps where
5 the energy of the masker is low, reflecting listeners' ability to listen in the
6 modulation dips of the masker (Howard-Jones and Rosen, 1993). Despite
7 the consistent locations of the glimpses in subplot d and e, the size of the
8 glimpses that are calculated from the BSS-estimated signals is substantially
9 smaller than the true number, which is obtained by comparing the known
10 speech signal against the masker signal. Consequently, the glimpse count –
11 what the BiDWGP metric relies on to make intelligibility prediction – in the
12 former case (378 in subplot e) is much smaller than in the latter case (641 in
13 subplot d). This demonstrates the effect due to the SNR underestimation.

14 To empirically compensate for the error in SNR, a gain of 4.7 dB was
15 applied to the estimated speech signal, leading to an increase in both glimpse
16 size and number (562 in subplot f) in the estimated speech. When applying
17 the constant 4.7 dB gain to all BSS-estimated samples, the performance of the
18 proposed system with either BiDWGP or BiSTOI as the predictor appears

1 to be improved over that without the gain as presented in Table 4.

2 For the reference performance, it is unclear why BiSTOI underestimated
3 intelligibility in the diffused BAB conditions relative to the other noises in
4 this study, resulting in the poor overall performance. Inheriting from STOI,
5 BiSTOI assumes that the supplied reference speech signal leads to perfect
6 intelligibility, hence the comparison is conducted between the reference and
7 the tested signals. When BiSTOI was used in the proposed system, the
8 exacerbated masker-specific bias between stationary and fluctuating maskers
9 is likely due to the use of the BSS-estimated speech signal as the reference,
10 which probably does not yield the same intelligibility and quality as the clean
11 unprocessed speech. Furthermore, the performance of the BSS probably
12 varies with masker type, leading to different intelligibility and quality of the
13 output signals. Therefore, the discrepancy on the BiSTOI outputs for the
14 same intelligibility in SSN and SMN becomes noticeably evident as seen in
15 Fig. 5. This warrants further investigation in how masker type affects BBS
16 performance.

17 5.2. Impact of room acoustics on system performance

18 With the BSS model trained for Room A, the system made less accurate
19 intelligibility predictions in Room B. The longer RT in room B was expected
20 to make separation more challenging (e.g. Mandel et al., 2010; Alinaghi et al.,
21 2014); this would lead to different distributions of the audio features for
22 the DNN input and output. Take the SSN condition at -9 dB SNR for
23 example, with the same mixing process using RIRs from Room A and Room
24 B separately, the frequency-independent mixture mean shifts from -0.62 to
25 -0.76. As a result, this mismatch between the training data and testing
26 data could have led to the decreased separation performance, and thus the
27 resulting reduction in the predictive accuracy of the OIMs.

28 To investigate this possibility, the BSS model was also trained for Room
29 B to replace the original model trained for Room A. The performance of the
30 system in different conditions is shown in Table 5. The overall performance,
31 ρ_p and ρ_s , with BiDWGP as the predictor in Room B indeed increase to
32 0.88 and 0.91 respectively, from 0.82 and 0.88 when the Room A model
33 was used. These results are comparable to the reference performance in
34 Room B ($\rho_p = 0.93$ and $\rho_s = 0.95$) [$\chi^2 \leq 3.727$, $p \geq 0.054$]. Although
35 the overall performance in Room A ($\rho_p = 0.89$ and $\rho_s = 0.84$) was not
36 significantly decreased by using the Room B BSS model, the accuracy for
37 individual maskers does tend to decline, especially for SSN and BAB [$\chi^2 \geq$

Table 5: System performance with BSS model trained for Room B. For all ρ , $p < 0.001$.

	Room A				Room B			
	SSN	SMN	BAB	overall	SSN	SMN	BAB	overall
Proposed system (BiDWGP as OIM)								
ρ_p	0.90	0.87	0.80	0.88	0.95	0.95	0.94	0.88
ρ_s	0.94	0.87	0.84	0.83	0.93	0.93	0.93	0.91
RMSE _m (pps)	10.2	8.3	5.1	9.0	6.7	5.7	2.7	10.5
Proposed system (BiSTOI as OIM)								
ρ_p	0.95	0.88	0.94	0.70	0.91	0.90	0.95	0.82
ρ_s	0.94	0.85	0.67	0.72	0.90	0.90	0.90	0.91
RMSE _m (pps)	7.8	7.8	2.9	14.7	8.2	6.6	2.8	13.3

1 4.741, $p \leq 0.032$]. Therefore, for the best predictive accuracy when using
2 BiDWGP in the system, ideally the BSS model is trained for the target space.
3 With BiSTOI as the predictor, using different BSS models however does not
4 substantially change the overall system performance, nor that for individual
5 maskers [$\chi^2 \leq 1.812$, $p \geq 0.093$]. As discussed above, using an *imperfect*
6 reference signal in BiSTOI seems to be an explanation for its low overall
7 performance.

8 5.3. Error in BSL-estimated source location

9 The motivation for employing a BSL model is to detect the source loca-
10 tions in the horizontal plane so that ILD cues can be estimated and integrated
11 into the binaural signals. As ILD is a function of azimuth (Fig. 2), the per-
12 formance of ILD estimation is therefore dependent on the accuracy of the
13 azimuth detection. The errors in the estimated azimuths compared to the
14 target azimuths for the SSN and SMN masker were computed. Since the
15 results for SSN and SMN are highly consistent, only those for SMN are pre-
16 sented in Fig. 8. The absolute errors fall into the range from 2.6° to 16.2°,
17 with smaller errors when the source is at 5° and 90° and bigger errors in be-
18 tween at -30° and 60°. In each target room, the errors are also similar. The

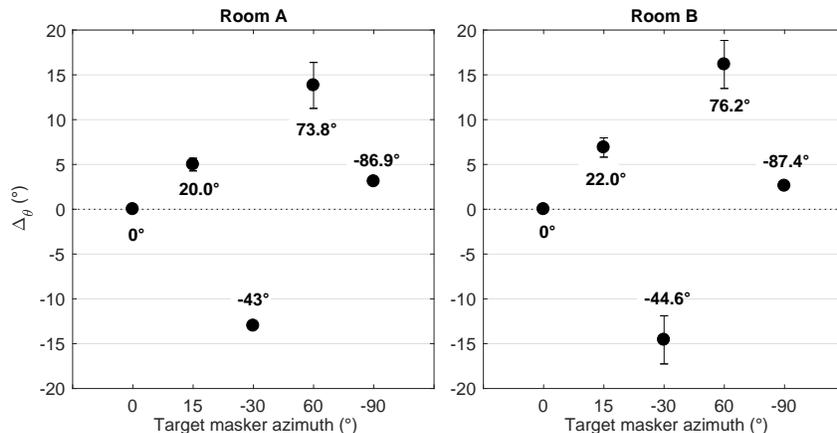


Figure 8: Difference between estimated and target azimuth. $\Delta\theta = \theta' - \theta$, where θ' and θ denote BSL-estimated azimuth and target azimuth respectively. Value of θ' is also supplied next to each data point. Error bars indicate standard deviation of $\Delta\theta$.

1 direct linear mapping from the TDOA to azimuth is used in the proposed
 2 system. However, their relationship is more complicated and may be non-
 3 linear. Since two sound sources are present in the mixture, the interference
 4 from the competing source may reduce the accuracy in localisation.

5 To further quantify the impact on the ILD estimation due to the error
 6 in azimuth detection, the estimated ILDs are computed on all SMN signals
 7 for the target azimuths (i.e. -30° and 60°) where the largest errors occurred
 8 and for the corresponding estimated azimuths (i.e. -43° and 76.2°). It
 9 found that the mean absolute ILD differences are 1.2 and 0.1 dB between the
 10 target -30° and estimated -43° , and between the target 60° and estimated
 11 76.2° , respectively. These small errors in ILD estimation probably do not
 12 significantly affect the predictive performance of the system.

13 5.4. The role of head-induced ILD integration

14 From the signals captured by the pair of microphones to those processed
 15 by the BSS separation, in principle there should be very limited ILD exist-
 16 ing between the two channel signals. Early analyses have verified that the
 17 BSS separation does not noticeably alter the ILD. With proper microphone
 18 calibration, the only possible ILD measured on the microphones comes from
 19 source-to-microphone distances being different for sources at 0° and 180° .
 20 But this is trivial compared to the ILD induced by the head-shadow effect.
 21 Fig. 9 compares the ILD of BSS output before or after ILD rectification, to

1 the head-induced ILD (measured from the signals recorded using the HATS).
2 Consequently, a Δ_{ILD} of 0 dB is desirable in theory because it shows the head-
3 shadow effect has been correctly estimated. Similar to Δ_{θ} in Fig. 8, only the
4 results of SMN are displayed for demonstration purpose since similar results
5 were observed for SSN.

6 The head-induced ILD increases with the increase of separation from 0°
7 up to 90° (Shaw and Vaillancourt, 1985). The mean ILD before ILD integra-
8 tion is up to 5.0 and 3.7 dB lower than the head-induced ILD in Room A and
9 Room B, respectively. After ILD correction, on the other hand, there is a ten-
10 dency to overestimation of up to 2.9 dB with a maximum when the source is at
11 -90° . This estimation error is however comparable to that of 2.3 dB reported
12 in Tang et al. (2016a). To identify the importance of the ILD integration
13 component in the proposed system, the performance of the proposed system
14 without the ILD estimation component is calculated. When BiDWGP was
15 used as the intelligibility predictor, compared to that with ILD integration
16 ($\rho_p = 0.89, 0.82$ and 0.85 for Room A, B, and A+B together, respectively),
17 the exclusion of ILD integration leads to the Pearson correlations with the
18 subjective data decreasing to $\rho_p = 0.71, 0.69$ and 0.69 . When BiSTOI was
19 used, the system performance dropped from $\rho_p = 0.67, 0.83$ and 0.74 to
20 $\rho_p = 0.62, 0.70$ and 0.69 , respectively. This finding echoes that of previous
21 studies (e.g. Lavandier and Culling, 2010; Tang et al., 2016a) on ILD con-
22 tribution to binaural speech intelligibility in noise, and confirms that ILD
23 integration plays a crucial role in the proposed system for robust predictive
24 power.

25 5.5. Limitations and extensions

26 A robust system should be able to offer reasonable performance in any
27 unknown conditions. For reverberation, one solution could be to introduce a
28 de-reverberation component (e.g. Nakatani et al., 2008; Naylor and Gaubitch,
29 2010) to the system sitting in the pipeline before the BSS component, whose
30 separation model may even be trained in an anechoic condition. On the other
31 hand, to exploit the longer temporal relationship within each signal sequence,
32 recurrent neural networks such as long short term memory (Hochreiter and
33 Schmidhuber, 1997) could be considered in the future. In addition, since the
34 DNN is a data-driven machine learning approach, the training of the BSS
35 model could be performed on a larger database and using more sophisticated
36 DNN structures, for more robust performance in various conditions.

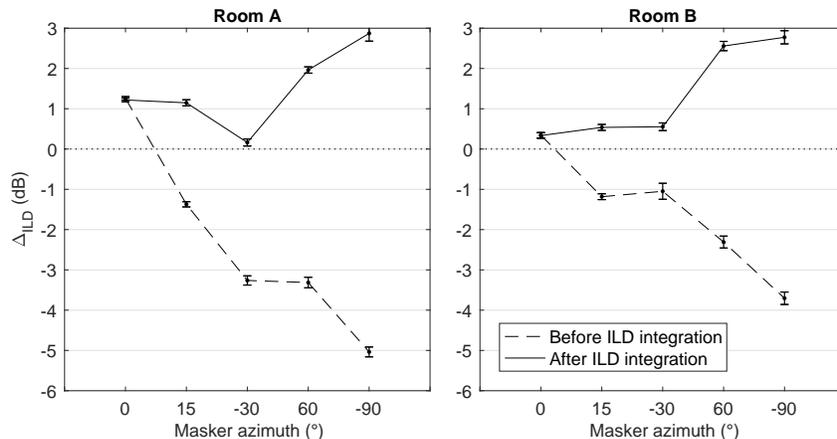


Figure 9: Difference between ILD of BSS output before or after ILD correction and head-induced ILD on SMN signals. $\Delta_{ILD} = ILD_X - ILD_{head-induced}$, where ILD_X is the ILD either before or after integration. Error bars indicate standard deviation of Δ_{ILD} .

1 The ILD estimation component may be further integrated within the BiD-
2 WGP metric. Because they both reconstruct the signal or generate auditory
3 representations for analysis using gammatone filters, signal processing here
4 can be done only once in order to save the computational time for online in-
5 stantaneous operation. Since the system is proposed as a general framework,
6 in order to facilitate any possible OIM serving as the back-end intelligibility
7 predictor, 55 filters are used by the ILD estimation component in the current
8 study for minimising the impact on the quality of the reconstructed signal
9 (Strahl and Mertins, 2009). Nevertheless, the number of filters can be re-
10 duced to 34, matching the number of frequencies that the BiDWGP metric
11 analyses.

12 6. Conclusions

13 A non-intrusive system for predicting binaural speech intelligibility in
14 noise is introduced. By placing a pair of closely-spaced microphones in the
15 target room, the system is able to make intelligibility estimations directly
16 from the captured signals, based on assumptions that the speech source is
17 straight ahead of the microphone pair and only one point or diffused source
18 exists in the target space. When compared to measured subjective intelligi-
19 bility, the system with the BiDWGP metric as the intelligibility predictor can
20 provide a reasonable match to listener’s word recognition rates in both sta-

1 tionary and fluctuating maskers, with correlation coefficients above 0.82 for
2 all testing conditions. Although it is still short in predictive power compared
3 to the state-of-the-art intrusive OIM, it could open the door for robust and
4 easy-to-deploy implementations for on-site speech intelligibility prediction in
5 practice. The study is mainly concluded as follows:

- 6 1. The proposed system provides a solution for fast binaural intelligibil-
7 ity prediction, when the reference speech signal is unavailable and the
8 location of the masking source is unknown.
- 9 2. The predictive performance of the system is dependent on the SNR
10 preservation of the BSS algorithm. An empirical gain may be applied
11 to the BSS-estimated signal to compensate for errors in SNR preser-
12 vation. Integrating head-induced ILD into the signals captured by the
13 microphones is also crucial for accurate binaural intelligibility predic-
14 tion. Errors in localisation appear to have less impact than the former
15 two factors.
- 16 3. The proposed system can deal with a single stationary or fluctuating
17 noise masker when it is presented as a point or diffused source on a
18 horizontal plane. However, the robustness needs to be enhanced to
19 enable handling of more than one spatially-separated masker.
- 20 4. The components (e.g. the back-end intelligibility predictor) in the
21 pipeline are not limited to those tested in the current study; other
22 techniques can be used in each place to serve for the same functions.
23 However, the predictive accuracy of the system may vary depending on
24 the *de facto* performance of chosen components and the mutual influ-
25 ences between elements in the processing chain. The entire framework
26 is also extensible for better predictive performance, such as including
27 a dereverberation component in reverberant conditions.
- 28 5. Since the DNN-trained BSS model operates on individual channels, the
29 proposed system can also be used to predict monaural speech intelli-
30 gibility using a monaural OIM as the back-end predictor. The BSL and
31 ILD estimation components should be excluded from the system for
32 this purpose.

1 **Acknowledgements**

2 This work was supported by the EPSRC Programme Grant S3A: Future
3 Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1)
4 and the BBC as part of the BBC Audio Research Partnership. The authors
5 would like to thank Huw Swanborough for conducting the listening experi-
6 ments. The MATLAB implementation of the BiSTOI metric was acquired
7 from <http://kom.aau.dk/project/Intelligibility/>. Data underlying
8 the findings are fully available without restriction, details are available from
9 <https://dx.doi.org/10.17866/rd.salford.5306746>.

10 Alinaghi, A., Jackson, P., Liu, Q., Wang, W., 2014. Joint mixing vector and
11 binaural model based stereo source separation. *IEEE/ACM Trans. Audio,
12 Speech, Language Process.* 22 (9), 1434–1448.

13 Andersen, A. H., de Haan, J. M., Tan, Z.-H., Jensen, J., 2015. A Binau-
14 ral Short Time Objective Intelligibility Measure for Noisy and Enhanced
15 Speech. In: *Proc. Interspeech*. pp. 2563–2567.

16 Andersen, A. H., de Haan, J. M., Tan, Z.-H., Jensen, J., 2016. Predicting
17 the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech.
18 *IEEE/ACM Transactions on Audio, Speech, and Language Processing*
19 24 (11), 1908–1920.

20 ANSI S3.5, 1997. ANSI S3.5-1997 Methods for the calculation of the Speech
21 Intelligibility Index.

22 Asaei, A., Boursard, H., Taghizadeh, M. J., Cevher, V., 2014. Model-based
23 sparse component analysis for reverberant speech localization. In: *Proc.*
24 *ICASSP*. pp. 1439–1443.

25 Blandin, C., Ozerov, A., Vincent, E., 2012. Multi-source TDOA estimation in
26 reverberant audio using angular spectra and clustering. *Signal Processing*
27 92 (8), 1950–1960.

28 Comon, P., April 1994. Independent component analysis, a new concept?
29 *Signal Processing* 36 (3), 287–314.

- 1 Cooke, M., 2006. A glimpsing model of speech perception in noise. *J. Acoust.*
2 *Soc. Am.* 119 (3), 1562–1573.
- 3 Cosentino, S., Marquardt, T., McAlpine, D., Culling, J. F., Falk, T. H., 2014.
4 A model that predicts the binaural advantage to speech intelligibility from
5 the mixed target and interferer signals. *J. Acoust. Soc. Am.* 135 (2), 796–
6 807.
- 7 Culling, J. F., Hawley, M. L., Litovsky, R. Y., 2004. The role of head-induced
8 interaural time and level differences in the speech reception threshold for
9 multiple interfering sound sources. *J. Acoust. Soc. Am.* 116 (2), 1057–1065.
- 10 Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modeling auditory process-
11 ing of amplitude modulation. I. Detection and masking with narrow-band
12 carriers. *J. Acoust. Soc. Am.* 102, 2892–2905.
- 13 Dau, T., Püschel, D., Kohlrausch, A., 1996. A quantitative model of the
14 “effective” signal processing in the auditory system. I. Model structure. *J.*
15 *Acoust. Soc. Am.* 99, 3615–3622.
- 16 Dubbelboer, F., Houtgast, T., 2007. A detailed study on the effects of noise
17 on speech intelligibility. *J. Acoust. Soc. Am.* 122 (5), 2865–2871.
- 18 Durlach, N. I., 1963a. Equalization and cancellation theory of binaural
19 masking-level differences. *J. Acoust. Soc. Am.* 35, 1206–1218.
- 20 Durlach, N. I., 1963b. Equalization and cancellation theory of binaural
21 masking-level differences. *J. Acoust. Soc. Am.* 35, 1206–1218.
- 22 Durlach, N. I., 1972. *Foundations of Modern Auditory Theory Vol. II.* Aca-
23 *ademic, New York, Ch. Binaural signal detection: Equalization and cancel-*
24 *lation theory.*
- 25 Falk, T. H., Zheng, C., Chan, W.-Y., 2010. A non-intrusive quality and intel-
26 ligibility measure of reverberant and dereverberated speech. *IEEE Trans.*
27 *Audio, Speech, Language Process.* 18 (7), 1766–1774.
- 28 Fallon, M. F., Godsill, S. J., May 2012. Acoustic source localization and
29 tracking of a time-varying number of speakers. *IEEE Trans. Audio, Speech,*
30 *Language Process.* 20 (4), 1409–1415.

- 1 Farina, A., Feb 2000. Simultaneous measurement of impulse response and
2 distortion with a swept-sine technique. In: Audio Engineering Society Con-
3 vention 108.
- 4 Fletcher, H., 1921. An empirical theory of telephone quality. AT&T Internal
5 Memorandum 101 (6).
- 6 Geravanchizadeh, M., Fallah, A., 2015. Microscopic prediction of speech in-
7 telligibility in spatially distributed speech-shaped noise for normal-hearing
8 listeners. *J. Acoust. Soc. Am.* 138 (6), 4004–4015.
- 9 Gomez, A. M., Schwerin, B., Paliwal, K., Mar. 2012. Improving objective in-
10 telligibility prediction by combining correlation and coherence based meth-
11 ods with a measure based on the negative distortion ratio. *Speech Com-*
12 *munication* 54 (3), 503–515.
- 13 Grais, E. M., Sen, M. U., Erdogan, H., May 2014. Deep neural networks for
14 single channel source separation. In: *Proc. ICASSP*. pp. 3734–3738.
- 15 Grancharov, V., Zhao, D., Lindblom, J., Kleijn, W., 2006. Low-complexity,
16 nonintrusive speech quality assessment. *IEEE Trans. Audio, Speech, Lan-*
17 *guage Process.* 14 (6), 1948–1956.
- 18 Hawley, M. L., Litovsky, R. Y., Culling, J. F., 2004. The benefit of binaural
19 hearing in a cocktail party: Effect of location and type of interferer. *J.*
20 *Acoust. Soc. Am.* 115 (2), 833–843.
- 21 Hilkhuisen, G., Gaubitch, N., Brookes, M., Huckvale, M., 2012. Effects of
22 noise suppression on intelligibility: Dependency on signal-to-noise ratios.
23 *Speech Communication* 131 (1), 531–539.
- 24 Hirsh, I. J., 1950. The relation between localization and intelligibility. *J.*
25 *Acoust. Soc. Am.* 22, 196–120.
- 26 Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural*
27 *Computation* 9 (8), 1735–1780.
- 28 Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-
29 impaired listeners based on a psychoacoustically motivated perception
30 model. *J. Acoust. Soc. Am.* 100 (3), 1703–1716.

- 1 Houtgast, T., Steeneken, H. J. M., 1985. A review of the MTF concept in
2 room acoustics and its use for estimating speech intelligibility in auditoria.
3 *J. Acoust. Soc. Am.* 77 (3), 1069–1077.
- 4 Howard-Jones, P. A., Rosen, S., 1993. Uncomodulated glimpsing in “checker-
5 board” noise. *J. Acoust. Soc. Am.* 93, 2915–2922.
- 6 Huang, P. S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2015. Joint
7 optimization of masks and deep recurrent neural networks for monaural
8 source separation. *IEEE/ACM Trans. Audio, Speech, Language Process.*
9 23 (12), 2136–2147.
- 10 Huber, R., Kollmeier, B., November 2006. PEMO-Q –a new method for
11 objective audio quality assessment using a model of auditory perception.
12 *IEEE Trans. Audio, Speech, Language Process.* 14 (6), 1902–1911.
- 13 IEC, 2011. “Part 16: Objective rating of speech intelligibility by speech trans-
14 mission index (4th edition),” in IEC 60268 Sound System Equipment (Int.
15 Electrotech. Commiss., Geneva, Switzerland).
- 16 Jelfs, S., Culling, J. F., Lavandier, M., May 2011. Revision and validation of
17 a binaural model for speech intelligibility in noise. *Hear. Res.* 275 (1–2),
18 96–104.
- 19 Jørgensen, S., Ewert, S. D., Dau, T., 2013. A multi-resolution envelope-power
20 based model for speech intelligibility. *J. Acoust. Soc. Am.* 134 (1), 436–446.
- 21 Jurgens, T., Brand, T., 2009. Microscopic prediction of speech recognition for
22 listeners with normal hearing in noise using an auditory model. *J. Acoust.*
23 *Soc. Am.* 126 (5), 2635–2648.
- 24 Jutten, C., Herault, J., July 1991. Blind separation of sources, part I: An
25 adaptive algorithm based on neuromimetic architecture. *Signal Processing*
26 24 (1), 1–10.
- 27 Karbasi, M., Abdelaziz, A. H., Kolossa, D., 2016. Twin-HMM-based non-
28 intrusive speech intelligibility prediction. In: *Proc. ICASSP*. pp. 624–628.
- 29 Knapp, C., Carter, G. C., August 1976. The generalized correlation method
30 for estimation of time delay. *IEEE Trans. Audio, Speech, Language Pro-*
31 *cess.* 24 (4), 320–327.

- 1 Lavandier, M., Culling, J. F., 2010. Prediction of binaural speech intelligibil-
2 ity against noise in rooms. *J. Acoust. Soc. Am.* 127, 387–399.
- 3 Lehmann, E. A., Williamson, R. C., 2006. Particle filter design using im-
4 portance sampling for acoustic source localisation and tracking in rever-
5 berant environments. *EURASIP Journal on Advances in Signal Processing*
6 2006 (1), 1–9.
- 7 Levitt, H., Rabiner, L. R., Oct. 1967. Predicting binaural gain in intelligibility
8 and release from masking for speech. *J. Acoust. Soc. Am.* 42 (4), 820–829.
- 9 Li, F. F., Cox, T. J., 2003. Speech transmission index from running speech:
10 A neural network approach. *J. Acoust. Soc. Am.* 113 (4), 1999–2008.
- 11 Liu, D., Smaragdis, P., Kim, M., 2014. Experiments on deep learning for
12 speech denoising. In: *Proc. Interspeech*. pp. 2685–2689.
- 13 Liu, Q., Tang, Y., Jackson, P. J. B., Wang, W., 2016. Predicting binaural
14 speech intelligibility from signals estimated by a blind source separation
15 algorithm. In: *Proc. Interspeech*. pp. 140–144.
- 16 Ma, W.-K., Vo, B.-N., Singh, S. S., Baddeley, A., 2006. Tracking an unknown
17 time-varying number of speakers using TDOA measurements: a random
18 finite set approach. *IEEE Trans. Signal Process.* 54 (9), 3291–3304.
- 19 Mandel, M. I., Weiss, R. J., Ellis, D., Feb. 2010. Model-based expectation-
20 maximization source separation and localization. *IEEE Trans. Audio,*
21 *Speech, Language Process.* 18 (2), 382–394.
- 22 May, T., Dau, T., 2014. Requirements for the evaluation of computational
23 speech segregation systems. *J. Acoust. Soc. Am.* 136 (6), EL398–EL404.
- 24 Moore, B. C. J., Glasberg, B. R., 1983. Suggested formulae for calculating
25 auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.*
26 74, 750–753.
- 27 Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B. H., 2008.
28 Blind speech dereverberation with multi-channel linear prediction based on
29 short time fourier transform representation. In: *Proc. ICASSP*. pp. 85–88.
- 30 Naylor, P. A., Gaubitch, N. D. (Eds.), 2010. *Speech Dereverberation*.
31 Springer, New York, NY, USA.

- 1 Nugraha, A. A., Liutkus, A., Vincent, E., September 2016. Multichannel
2 audio source separation with deep neural networks. *IEEE/ACM Trans.*
3 *Audio, Speech, Language Process.* 24 (9), 1652–1664.
- 4 Peso Parada, P., Sharma, D., Lainez, J., Barreda, D., Waterschoot, T. v.,
5 Naylor, P. A., 2016. A single-channel non-intrusive C50 estimator cor-
6 related with speech recognition performance. *IEEE/ACM Trans. Audio,*
7 *Speech, Language Process.* 24 (4), 719–732.
- 8 Rennies, J., Brand, T., Kollmeier, B., 2011. Prediction of the influence of
9 reverberation on binaural speech intelligibility in noise and in quiet. *J.*
10 *Acoust. Soc. Am.* 130 (5), 2999–3012.
- 11 Rhebergen, K. S., Versfeld, N. J., 2005. A Speech Intelligibility Index-based
12 approach to predict the speech reception threshold for sentences in fluc-
13 tuating noise for normal-hearing listeners. *J. Acoust. Soc. Am.* 117 (4),
14 2181–2192.
- 15 Rix, A. W., Beerends, J. G., Hollier, M. P., Hekstra, A. P., 2001. Perceptual
16 evaluation of speech quality (PESQ)-a new method for speech quality as-
17 sessment of telephone networks and codecs. In: *Proc. ICASSP. Vol. 2.* pp.
18 749–752.
- 19 Rothausler, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker,
20 M. H. L., Urbanek, G. E., Nordby, K. S., Weinstock, M., 1969. *IEEE Rec-*
21 *ommended practice for speech quality measurements.* *IEEE Trans. Audio*
22 *Electroacoust* 17, 225–246.
- 23 Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization
24 for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*
25 26 (1), 43–49.
- 26 Santos, J. F., Cosentino, S., Hazrati, O., Loizou, P. C., Falk, T. H., 2013.
27 Objective speech intelligibility measurement for cochlear implant users in
28 complex listening environments. *Speech Communication* 55 (7–8), 815–824.
- 29 Santos, J. F., Falk, T. H., 2014. Updating the SRMR-CI metric for improved
30 intelligibility prediction for cochlear implant users. *IEEE/ACM Trans. Au-*
31 *dio, Speech, Language Process.* 22 (12), 2197–2206.

- 1 Schmidt, R. O., 04 1986. Multiple emitter location and signal parameter
2 estimation. *IEEE Trans. Antennas Propag.* 34, 276–280.
- 3 Sharma, D., Hilkhuyzen, G., Gaubitch, N. D., Naylor, P. A., Brookes, M.,
4 Huckvale, M., 2010. Data driven method for non-intrusive speech intelligi-
5 bility estimation. In: *Proc. EUSIPCO*. pp. 1899–1903.
- 6 Sharma, D., Wang, Y., Naylor, P. A., Brookes, M., 2016. A data-driven
7 non-intrusive measure of speech quality and intelligibility. *Speech Com-
8 munication* 80, 84–94.
- 9 Shaw, E., Vaillancourt, M. M., 1985. Transformation of soundpressure level
10 from the free field to the eardrum presented in numerical form. *J. Acoust.
11 Soc. Am.* 78 (3), 1120–1123.
- 12 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.,
13 2014. Dropout: A simple way to prevent neural networks from overfitting.
14 *J. Mach. Learn. Res.* 15 (1), 1929–1958.
- 15 Strahl, S., Mertins, A., 2009. Analysis and design of gammatone signal mod-
16 els. *J. Acoust. Soc. Am.* 126 (5), 2379–2389.
- 17 Taal, C. H., Hendriks, R. C., Heusdens, R., Jensen, J., 2010. A short time
18 objective intelligibility measure for time-frequency weighted noisy speech.
19 In: *Proc. ICASSP*. pp. 4214–4217.
- 20 Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-
21 dependent speech intelligibility enhancement. In: *Proc. Interspeech*. pp.
22 955–958.
- 23 Tang, Y., Cooke, M., 2016. Glimpse-based metrics for predicting speech intel-
24 ligibility in additive noise conditions. In: *Proc. Interspeech*. pp. 2488–2492.
- 25 Tang, Y., Cooke, M., Fazenda, B. M., Cox, T. J., 2015. A glimpse-based
26 approach for predicting binaural intelligibility with single and multiple
27 maskers in anechoic conditions. In: *Proc. Interspeech*. pp. 2568–2572.
- 28 Tang, Y., Cooke, M. P., Fazenda, B. M., Cox, T. J., Sep. 2016a. A metric for
29 predicting binaural speech intelligibility in stationary noise and competing
30 speech maskers. *J. Acoust. Soc. Am.* 140 (3), 1858–1870.

- 1 Tang, Y., Cooke, M. P., Valentini-Botinhao, C., 2016b. Evaluating the predic-
2 tions of objective intelligibility metrics for modified and synthetic speech.
3 *Computer Speech and Language* 35, 73–92.
- 4 Tang, Y., Hughes, R. J., Fazenda, B. M., Cox, T. J., Sep. 2016c. Evaluat-
5 ing a distortion-weighted glimpsing metric for predicting binaural speech
6 intelligibility in rooms. *Speech Communication* 82 (C), 26–37.
- 7 Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by
8 a running average of its recent magnitude. COURSERA: Neural Networks
9 for Machine Learning.
- 10 van Wijngaarden, S. J., Drullman, R., 2008. Binaural intelligibility prediction
11 based on the speech transmission index. *J. Acoust. Soc. Am.* 123 (6), 4514–
12 4523.
- 13 Vermaak, J., Blake, A., 2001. Nonlinear filtering for speaker tracking in noisy
14 and reverberant environments. In: *Proc. ICASSP*. Vol. 5. pp. 3021–3024.
- 15 Virtanen, T., Mar. 2007. Monaural sound source separation by nonnega-
16 tive matrix factorization with temporal continuity and sparseness criteria.
17 *IEEE Trans. Audio, Speech, Language Process.* 15 (3), 1066–1074.
- 18 Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J., Hershey,
19 J. R., Schuller, B., 2015. Speech enhancement with LSTM recurrent neu-
20 ral networks and its application to noise-robust ASR. In: *International*
21 *Conference on Latent Variable Analysis and Signal Separation*. pp. 91–99.
- 22 Xu, Y., Du, J., Dai, L. R., Lee, C. H., January 2014. An experimental
23 study on speech enhancement based on deep neural networks. *IEEE Signal*
24 *Process. Lett.* 21 (1), 65–68.
- 25 Yu, Y., Wang, W., Han, P., 2016. Localization based stereo speech source
26 separation using probabilistic time-frequency masking and deep neural net-
27 works. *EURASIP Journal on Audio Speech and Music Processing* 7.
- 28 Zurek, P. M., 1993. *Acoustical Factors Affecting Hearing Aid Performance*.
29 Allyn and Bacon, Needham Heights, MA, Ch. Binaural advantages and
30 directional effects in speech intelligibility, pp. 255–276.