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

Yan Tang[a,*], Qingju Liu[b], Wenwu Wang[b], Trevor J. Cox[a]

[a] *Acoustics Research Centre, University of Salford, UK*
[b] *Centre for Vision, Speech and Signal Processing, University of Surrey, UK*

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.

Keywords: objective intelligibility measure, non-intrusive, binaural intelligibility, noise, glimpsing, neural network, blind source separation, 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-

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

The ASR-based OIMs normally comprise the feature extraction and ASR components. Indeed, they can provide detailed modelling of speech perception in noise and make phoneme-level intelligibility predictions compared to word- and sentence-level predictions offered by normal intrusive OIMs. This permits, for example, more transparent and profound analyses to be performed on the model's errors. Therefore, they are also known as *microscopic* OIMs. However, knowing exactly what constants and vowels a listener may misperceive is unnecessary in many practical situations where a simple intelligibility estimate is sufficient. In addition, except for the glimpsing model (Cooke, 2006), all the microscopic OIMs mentioned above were only evaluated in speech-shaped noise (SSN). Their performance in more commonly-occurring noise conditions (e.g. fluctuating noise) was not investigated. Although an ASR can be trained for any target noise masker, deploying an ASR is onerous, especially for a robust ASR system.

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

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

One solution to overcome the lack of subjective training data is to use objective intelligibility score provided by an established OIM as the target output. Usually the performance of an established OIM was rigorously evaluated in previous studies by comparing its predictions to subjective data, it is expected to be able to provide reasonable estimation on subjective intelligibility. Li and Cox (2003) trained a neural network on the Speech Transmission Index (STI, IEC, 2011) from the low frequency envelope spectrum of running speech, to predict intelligibility. Sharma et al. (2016) further improved LCIA and extended it to both speech quality and intelligibility predictions. In terms of intelligibility, the GMM used in the enhanced version of LCIA, renamed as the Non-Intrusive Speech Assessment (NISA), was trained on the predictive scores of the short-time objective intelligibility (STOI, Taal et al., 2010), which was validated to show good match to the subjective data measured in Hilkhuysen et al. (2012). Despite extensive objective evaluations performed, the NISA was regretfully not further evaluated using subjective data. This leaves the question of whether the high correlation with the objective scores can be translated to a good match with subjective intelligibility unanswered. There is some evidence (Tang and Cooke, 2012; Tang et al., 2016b) suggesting that STOI lacks predictive accuracy when making predictions for algorithmically-modified speech or across different types of maskers.

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

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

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

1.2. Overview of this work

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

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

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

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

2. Proposed system

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

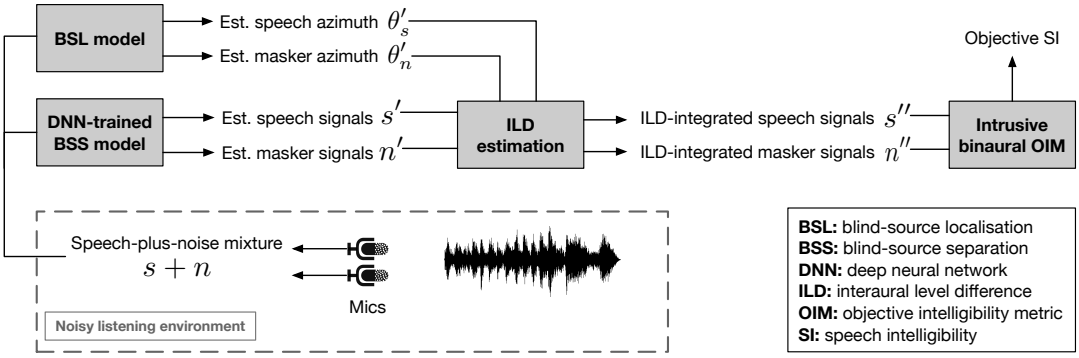

Figure 1: Schematic of the proposed system

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

2.1. Blind source separation using deep neural network

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

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

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

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

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

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

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

2.2. Blind source location estimation

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

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

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

where τ and F_s are the candidate delay and the sampling frequency, respectively. $*$ denotes the complex conjugate. Assuming the mixing process is

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

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

2.3. Integration of head-induced binaural level difference

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

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

$$s''(f) = k(f, \theta'_s) \cdot s'(f) \tag{4}$$

where

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

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

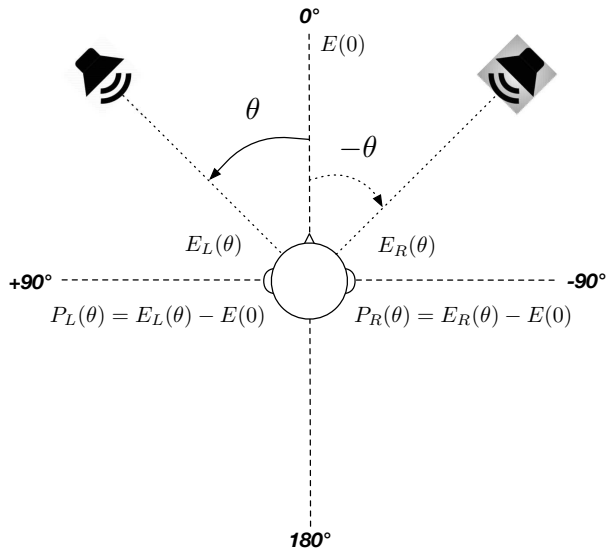

Figure 2: Difference ($P_L(\theta)$, $P_R(\theta)$) in sound pressure level between the left ear $E_L(\theta)$ and the listener's frontal position $E(0)$, and between the right ear $E_R(\theta)$ and $E(0)$ respectively, when the source is at an azimuthal position θ on a horizontal plane. The left-right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

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

$$P_R(f, \theta) = P_L(f, -\theta) = P_L(f, 360 - \theta) \tag{5}$$

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

2.4. Back-end binaural intelligibility predictor

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

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

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

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

3. Experiments

3.1. Preparation

The proposed system was evaluated in two rooms (referred to as Room 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

3.1.1. Binaural signal generation and test materials

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

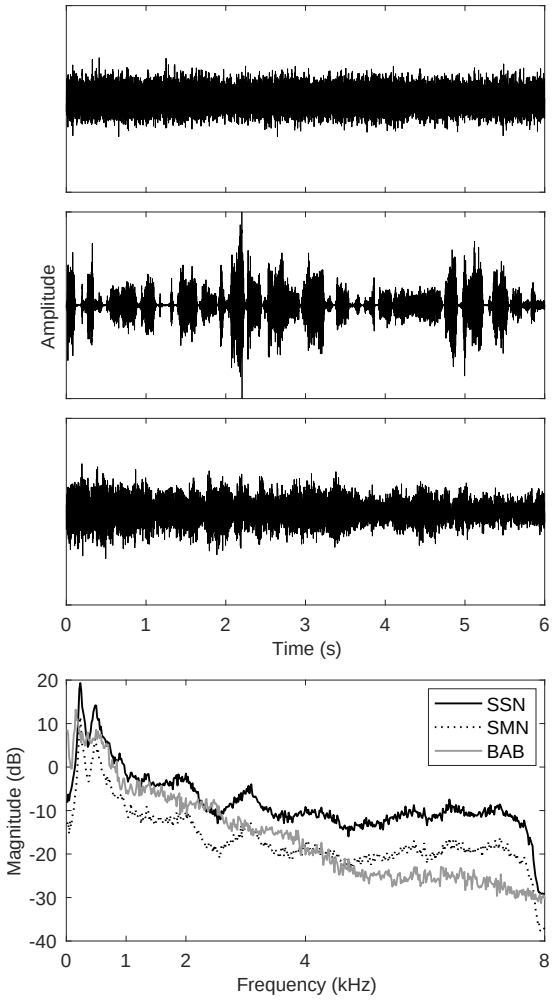

Figure 3: Sample waveform of SSN, SMN and BAB and their long-term average spectra. For illustration, the spectra of SSN and SMN are offset at ±3 dB, respectively.

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

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

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

3.1.2. DNN training of the BSS model

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

As the DNN-trained BSS algorithm employed in the current study operates on a monophonic signal, the separation does not rely on any binaural features such as ILD and ITD. Unlike in the previous study (Liu et al., 2016), where both ILD and ITD cues were used as features, and consequently several individual azimuth-dependent models were required when source location changed, the advantage here is that only one universal BSS model was 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

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

3.2. System prediction

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

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

3.3. Subjective data collection

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

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

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

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

4. Results

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

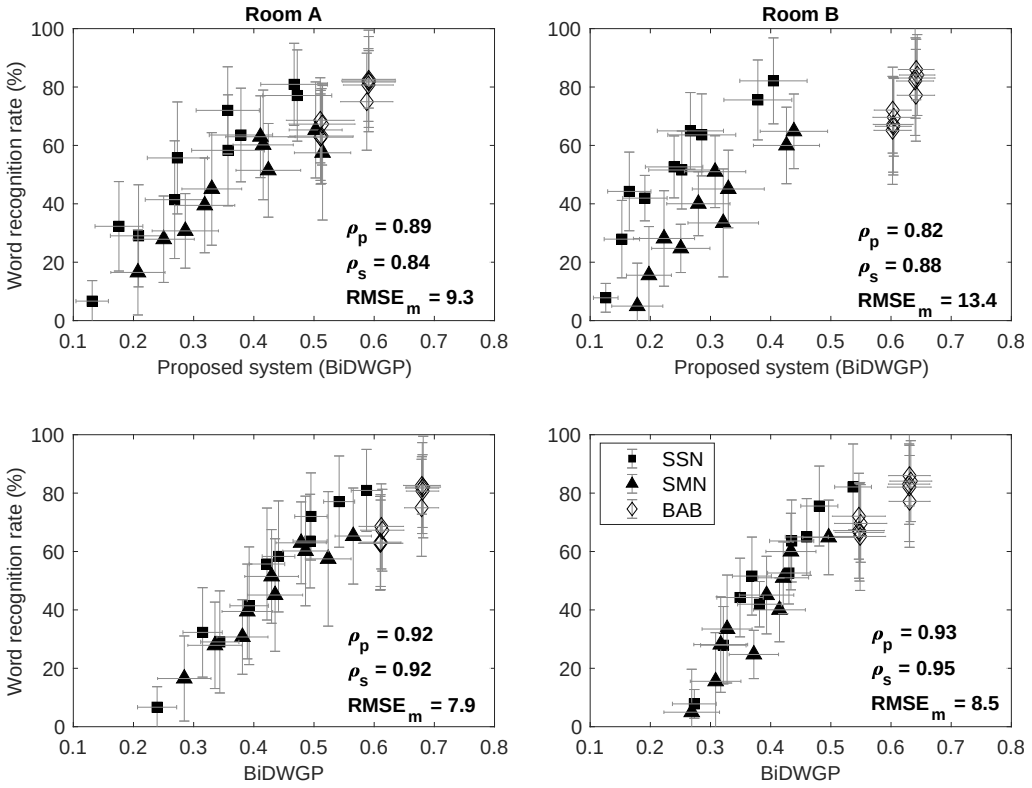

Figure 4: Objective-subjective correlation in Room A (left column) and B (right column), with reference performance provided in the second row. ρ_p, ρ_s and RMSE_m are displayed for each subplot. Error bars indicate standard deviations of subjective intelligibility (vertical) and BiDWGP scores (horizontal)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

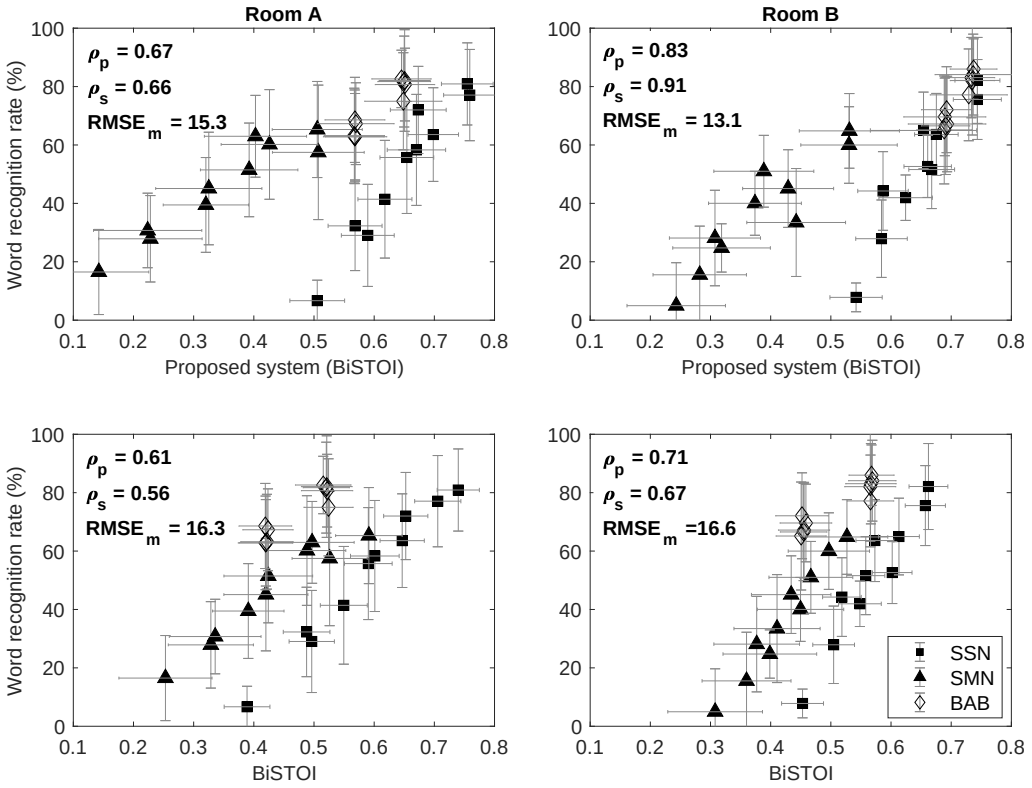

Figure 5: As for Fig. 4 but when BiSTOI is used as the intelligibility predictor.

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

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

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

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

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

Table 3 further details the performance of the proposed system with BiDWGP or BiSTOI for individual maskers in each target room, along with the reference counterparts. When BiDWGP was used, despite the declined overall predictive accuracy when making predictions across different types of maskers in Room B as observed above, the proposed system achieved similar performance to the reference method for individual maskers [all $\chi^2 \leq 2.907,\ p \geq 0.09$], except for the ranking ability for SMN in Room A [$\chi^2 = 8.865,\ p < 0.05$]. When BiSTOI was used and the overall performance is less good, the system also provided predictive accuracy for individual maskers that is similar to the reference performance in most of conditions [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 subcondtions 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

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

5. Discussion

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

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

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

5.1. Error in SNR between BSS-estimated signals

The robustness of the BSS algorithm may considerably affect the predictive accuracy because it determines the quality of the estimated source signals that an intrusive OIM uses to make intelligibility prediction. In order to separate the target speech and masker signals from the mixture, the DNN-trained BSS model essentially estimates the IRM of the target speech. If the IRM contains too much information about the masker signal, the estimated speech signal will still be noisy, while the separated masker signal will be missing parts of its original constituents. This potentially leads to higher SNR between the estimated signals than the original SNR, and hence an overestimation of intelligibility when the back-end intelligibility predictor makes predictions using the estimated signals. The opposite case on the other hand is caused by the IRM missing too much information from the target speech signal. As SNR is one of the most dominant effects affecting speech intelligibility in noise, its errors in the BSS-estimated signals may lead to inaccuracy in ultimate intelligibility prediction. Liu et al. (2016) investigated the error in SNR preservation of a binaural BSS algorithm, which uses both ILD and ITD as cues for separation. They found that while the interaural SNR can be well maintained by the algorithm, the overall SNR between estimated speech and masker signals tended to be underestimated. Consequently, decreased predictive performance was observed for all tested intrusive binaural OIMs which made predictions from the BSS outputs.

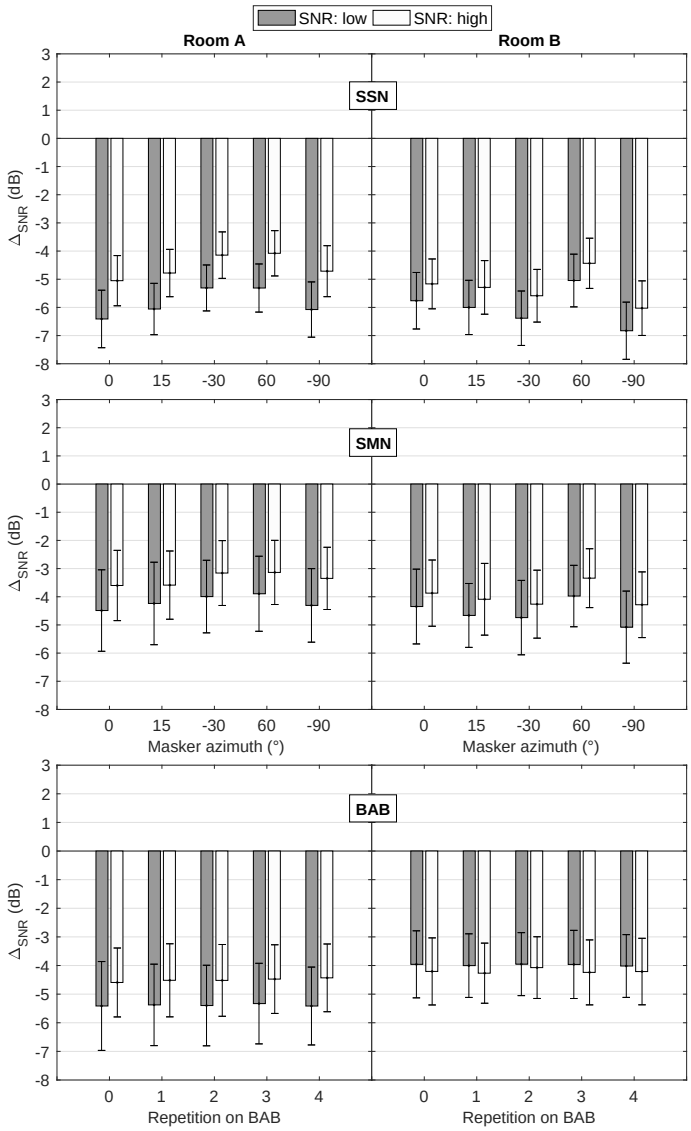

Figure 6: Difference between the target SNR and that calculated from the BSS-separated signals when the masker (SSN or SMN) is at different locations. The results for BAB are calculated from the corresponding five repeated conditions. Columns display the results for individual rooms while rows for mask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

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

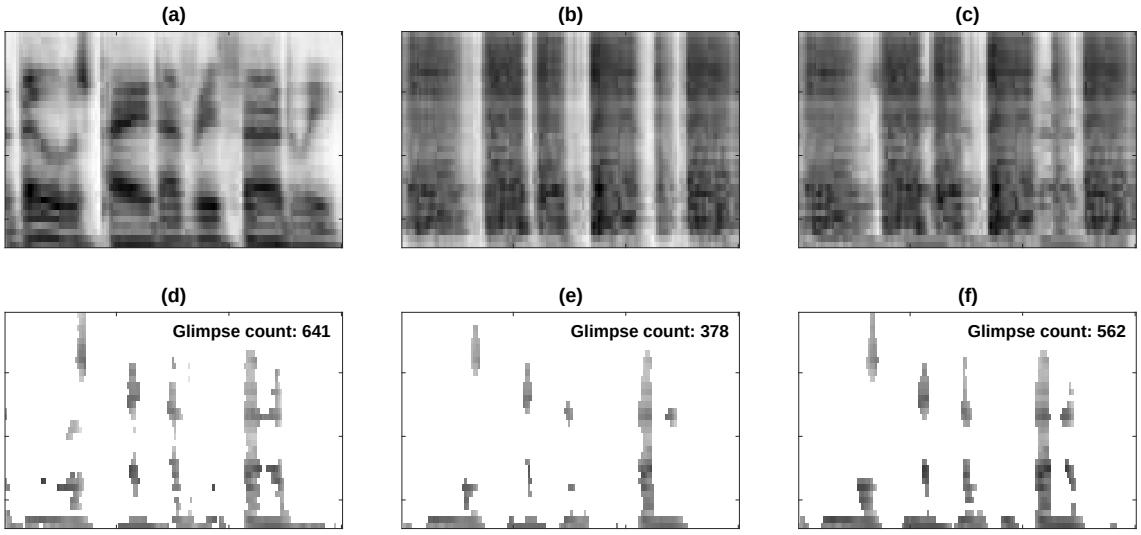

Figure 7: Spectrograms and glimpse analyses of the sentence 'the bill was paid every third week' in SMN at -12 dB SNR in Room A. (a): spectrogram of the clean speech signal; (b): spectrogram of the SMN signal; (c): spectrogram of the speech-plus-noise mixture; (d): glimpses calculated from the direct 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).

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

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

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

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

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

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

5.2. Impact of room acoustics on system performance

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

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

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

5.3. Error in BSL-estimated source location

The motivation for employing a BSL model is to detect the source locations in the horizontal plane so that ILD cues can be estimated and integrated into the binaural signals. As ILD is a function of azimuth (Fig. 2), the performance of ILD estimation is therefore dependent on the accuracy of the azimuth detection. The errors in the estimated azimuths compared to the target azimuths for the SSN and SMN masker were computed. Since the results for SSN and SMN are highly consistent, only those for SMN are presented in Fig. 8. The absolute errors fall into the range from 2.6° to 16.2°, with smaller errors when the source is at 5° and 90° and bigger errors in between at -30° and 60°. In each target room, the errors are also similar. The

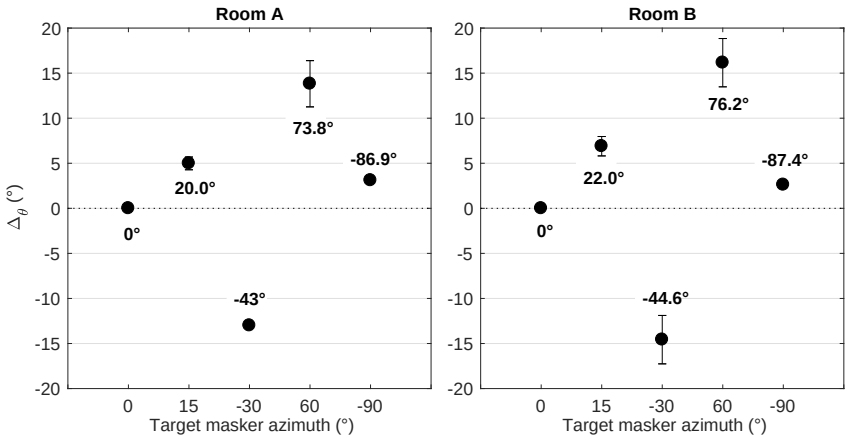

Figure 8: Difference between estimated and target azimuth. $\Delta_\theta = \theta' - \theta$, where θ' and θ denote BSL-estimated azimuth and target azimuth respectively. Value of θ' is also supplied next to each data point. Error bars indicate standard deviation of Δ_θ.

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

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

5.4. The role of head-induced ILD integration

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

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

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

5.5. Limitations and extensions

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

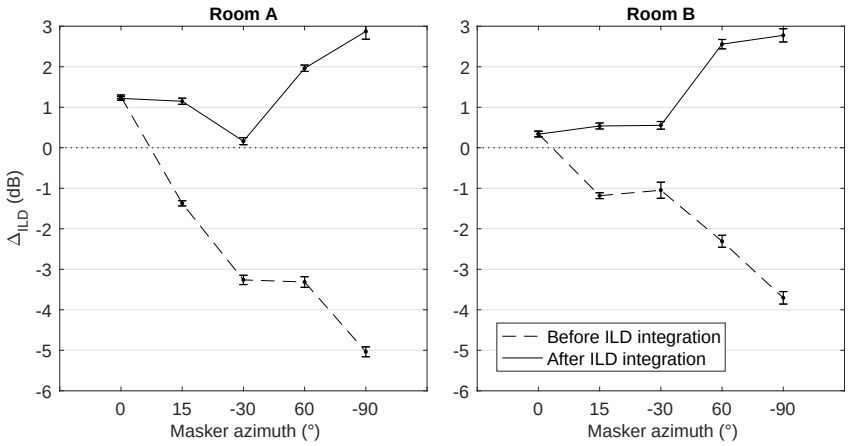

Figure 9: Difference between ILD of BSS output before or after ILD correction and head-induced ILD on SMN signals. $\Delta_{\mathrm{ILD}} = ILD_{\mathrm{X}} - ILD_{\text{head-induced}}$, where ILD_{X} is the ILD either before or after integration. Error bars indicate standard deviation of Δ_{ILD}.

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

6. Conclusions

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

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

1. The proposed system provides a solution for fast binaural intelligibility prediction, when the reference speech signal is unavailable and the location of the masking source is unknown.

2. The predictive performance of the system is dependent on the SNR preservation of the BSS algorithm. An empirical gain may be applied to the BSS-estimated signal to compensate for errors in SNR preservation. Integrating head-induced ILD into the signals captured by the microphones is also crucial for accurate binaural intelligibility prediction. Errors in localisation appear to have less impact than the former two factors.

3. The proposed system can deal with a single stationary or fluctuating noise masker when it is presented as a point or diffused source on a horizontal plane. However, the robustness needs to be enhanced to enable handling of more than one spatially-separated masker.

4. The components (e.g. the back-end intelligibility predictor) in the pipeline are not limited to those tested in the current study; other techniques can be used in each place to serve for the same functions. However, the predictive accuracy of the system may vary depending on the *de facto* performance of chosen components and the mutual influences between elements in the processing chain. The entire framework is also extensible for better predictive performance, such as including a dereverberation component in reverberant conditions.

5. Since the DNN-trained BSS model operates on individual channels, the proposed system can also be used to predict monaural speech intelligibility using a monaural OIM as the back-end predictor. The BSL and ILD estimation components should be excluded from the system for this purpose.

Acknowledgements

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

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

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

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

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

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

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

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

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

Cosentino, S., Marquardt, T., McAlpine, D., Culling, J. F., 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 (2), 796–807.

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

Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102, 2892–2905.

Dau, T., Püschel, D., Kohlrausch, A., 1996. A quantitative model of the "effective"signal processing in the auditory system. I. Model structure. J. Acoust. Soc. Am. 99, 3615–3622.

Dubbelboer, F., Houtgast, T., 2007. A detailed study on the effects of noise on speech intelligibility. J. Acoust. Soc. Am. 122 (5), 2865–2871.

Durlach, N. I., 1963a. Equalization and cancellation theory of binaural masking-level differences. J. Acoust. Soc. Am. 35, 1206–1218.

Durlach, N. I., 1963b. Equalization and cancellation theory of binaural masking-level differences. J. Acoust. Soc. Am. 35, 1206–1218.

Durlach, N. I., 1972. Foundations of Modern Auditory Theory Vol. II. Academic, New York, Ch. Binaural signal detection: Equalization and cancellation theory.

Falk, T. H., Zheng, C., Chan, W.-Y., 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio, Speech, Language Process. 18 (7), 1766–1774.

Fallon, M. F., Godsill, S. J., May 2012. Acoustic source localization and tracking of a time-varying number of speakers. IEEE Trans. Audio, Speech, Language Process. 20 (4), 1409–1415.

Farina, A., Feb 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Audio Engineering Society Convention 108.

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

Geravanchizadeh, M., Fallah, A., 2015. Microscopic prediction of speech intelligibility in spatially distributed speech-shaped noise for normal-hearing listeners. J. Acoust. Soc. Am. 138 (6), 4004–4015.

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

Grais, E. M., Sen, M. U., Erdogan, H., May 2014. Deep neural networks for single channel source separation. In: Proc. ICASSP. pp. 3734–3738.

Grancharov, V., Zhao, D., Lindblom, J., Kleijn, W., 2006. Low-complexity, nonintrusive speech quality assessment. IEEE Trans. Audio, Speech, Language Process. 14 (6), 1948–1956.

Hawley, M. L., Litovsky, R. Y., 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 (2), 833–843.

Hilkhuysen, G., Gaubitch, N., Brookes, M., Huckvale, M., 2012. Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios. Speech Communication 131 (1), 531–539.

Hirsh, I. J., 1950. The relation between localization and intelligibility. J. Acoust. Soc. Am. 22, 196–120.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9 (8), 1735–1780.

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

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

Howard-Jones, P. A., Rosen, S., 1993. Uncomodulated glimpsing in "checkerboard" noise. J. Acoust. Soc. Am. 93, 2915–2922.

Huang, P. S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio, Speech, Language Process. 23 (12), 2136–2147.

Huber, R., Kollmeier, B., November 2006. PEMO-Q –a new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio, Speech, Language Process. 14 (6), 1902–1911.

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

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

Jørgensen, S., Ewert, S. D., Dau, T., 2013. A multi-resolution envelope-power based model for speech intelligibility. J. Acoust. Soc. Am. 134 (1), 436–446.

Jurgens, T., Brand, T., 2009. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. J. Acoust. Soc. Am. 126 (5), 2635–2648.

Jutten, C., Herault, J., July 1991. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24 (1), 1–10.

Karbasi, M., Abdelaziz, A. H., Kolossa, D., 2016. Twin-HMM-based non-intrusive speech intelligibility prediction. In: Proc. ICASSP. pp. 624–628.

Knapp, C., Carter, G. C., August 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Audio, Speech, Language Process. 24 (4), 320–327.

Lavandier, M., Culling, J. F., 2010. Prediction of binaural speech intelligibility against noise in rooms. J. Acoust. Soc. Am. 127, 387–399.

Lehmann, E. A., Williamson, R. C., 2006. Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments. EURASIP Journal on Advances in Signal Processing 2006 (1), 1–9.

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

Li, F. F., Cox, T. J., 2003. Speech transmission index from running speech: A neural network approach. J. Acoust. Soc. Am. 113 (4), 1999–2008.

Liu, D., Smaragdis, P., Kim, M., 2014. Experiments on deep learning for speech denoising. In: Proc. Interspeech. pp. 2685–2689.

Liu, Q., Tang, Y., Jackson, P. J. B., Wang, W., 2016. Predicting binaural speech intelligibility from signals estimated by a blind source separation algorithm. In: Proc. Interspeech. pp. 140–144.

Ma, W.-K., Vo, B.-N., Singh, S. S., Baddeley, A., 2006. Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. IEEE Trans. Signal Process. 54 (9), 3291–3304.

Mandel, M. I., Weiss, R. J., Ellis, D., Feb. 2010. Model-based expectation-maximization source separation and localization. IEEE Trans. Audio, Speech, Language Process. 18 (2), 382–394.

May, T., Dau, T., 2014. Requirements for the evaluation of computational speech segregation systems. J. Acoust. Soc. Am. 136 (6), EL398–EL404.

Moore, B. C. J., Glasberg, B. R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74, 750–753.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B. H., 2008. Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. In: Proc. ICASSP. pp. 85–88.

Naylor, P. A., Gaubitch, N. D. (Eds.), 2010. Speech Dereverberation. Springer, New York, NY, USA.

Nugraha, A. A., Liutkus, A., Vincent, E., September 2016. Multichannel audio source separation with deep neural networks. IEEE/ACM Trans. Audio, Speech, Language Process. 24 (9), 1652–1664.

Peso Parada, P., Sharma, D., Lainez, J., Barreda, D., Waterschoot, T. v., Naylor, P. A., 2016. A single-channel non-intrusive C50 estimator correlated with speech recognition performance. IEEE/ACM Trans. Audio, Speech, Language Process. 24 (4), 719–732.

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

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

Rix, A. W., Beerends, J. G., Hollier, M. P., Hekstra, A. P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. ICASSP. Vol. 2. pp. 749–752.

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

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust., Speech, Signal Process. 26 (1), 43–49.

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

Santos, J. F., Falk, T. H., 2014. Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users. IEEE/ACM Trans. Audio, Speech, Language Process. 22 (12), 2197–2206.

Schmidt, R. O., 04 1986. Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. 34, 276–280.

Sharma, D., Hilkhuysen, G., Gaubitch, N. D., Naylor, P. A., Brookes, M., Huckvale, M., 2010. Data driven method for non-intrusive speech intelligibility estimation. In: Proc. EUSIPCO. pp. 1899–1903.

Sharma, D., Wang, Y., Naylor, P. A., Brookes, M., 2016. A data-driven non-intrusive measure of speech quality and intelligibility. Speech Communication 80, 84–94.

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Strahl, S., Mertins, A., 2009. Analysis and design of gammatone signal models. J. Acoust. Soc. Am. 126 (5), 2379–2389.

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

Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In: Proc. Interspeech. pp. 955–958.

Tang, Y., Cooke, M., 2016. Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. In: Proc. Interspeech. pp. 2488–2492.

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

Tang, Y., Cooke, M. P., Fazenda, B. M., Cox, T. J., Sep. 2016a. A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers. J. Acoust. Soc. Am. 140 (3), 1858–1870.

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

Tang, Y., Hughes, R. J., Fazenda, B. M., Cox, T. J., Sep. 2016c. Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms. Speech Communication 82 (C), 26–37.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

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

Vermaak, J., Blake, A., 2001. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In: Proc. ICASSP. Vol. 5. pp. 3021–3024.

Virtanen, T., Mar. 2007. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech, Language Process. 15 (3), 1066–1074.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J., Hershey, J. R., Schuller, B., 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: International Conference on Latent Variable Analysis and Signal Separation. pp. 91–99.

Xu, Y., Du, J., Dai, L. R., Lee, C. H., January 2014. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21 (1), 65–68.

Yu, Y., Wang, W., Han, P., 2016. Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. EURASIP Journal on Audio Speech and Music Processing 7.

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