A glimpse-based approach for predicting binaural intelligibility with single and multiple maskers in anechoic conditions

Yan Tang[1], *Martin Cooke*[2,3], *Bruno M. Fazenda*[1], *Trevor J. Cox*[1]

[1]Acoustic Research Centre, University of Salford, UK
[2]Ikerbasque (Basque Science Foundation), Bilbao, Spain
[3]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
y.tang@salford.ac.uk, m.cooke@ikerbasque.org, b.m.fazenda@salford.ac.uk, t.j.cox@salford.ac.uk

Abstract

A distortion-weighted glimpsing metric developed for estimating monaural speech intelligibility is extended to predict binaural speech intelligibility in noise. Two aspects of binaural listening, the better ear effect and the binaural advantage, are taken into account in the new metric, which predicts intelligibility using monaural target and masker signals and their location, and is therefore able to provide intelligibility estimates in situations where binaural signals are not readily available. Perceptual listening experiments were conducted to evaluate the predictive power of the proposed metric for speech in the presence of single and multiple maskers in anechoic conditions, for a range of source/masker azimuth combinations. The binaural metric is highly correlated ($\rho > 0.9$) with listeners' performance in all conditions tested, but overestimates intelligibility somewhat in conditions where multiple maskers are present and the target speech source location is unknown.

Index Terms: speech intelligibility, binaural listening, noise

1. Introduction

Objective intelligibility measures (OIMs) have been used in place of subjective perceptual listening tests for the interim evaluation of speech intelligibility prior to final subjective tests. For example, OIMs have been used in the development of speech modification algorithms [1, 2, 3]. Estimating speech intelligibility using OIMs is fast and cheap. They are therefore extremely useful for providing a first approximation of speech intelligibility when the subjective data is not directly available, or perceptual listening tests are impractical, especially when a large set of conditions need to be continuously evaluated.

Many OIMs have been proposed for predicting *monaural* speech intelligibility in various additive noise or reverberant conditions, such as the Speech Intelligibility Index (SII) [4], the Speech Transmission Index (STI) [5] and the Christiansen-Pederson-Dau metric [6]. Fewer measures have been designed for the prediction of *binaural* speech intelligibility. Some approaches work by adapting existing monaural OIMs to binaural signals. For instance, Zurek modified the SII metric to predict intelligibility for situations where the target speech and maskers are situated at different locations on a horizontal plane [7]. In [8] the STI was extended to deal with binaural listening. Jelfs et al. [9] proposed an approach to predict binaural intelligibility which adapts the SII-concept. The SII is based on frequency-dependent signal-to-noise ratios (SNRs), with contributions from each frequency region weighted by a band importance function (BIF). In [9], the final output is the speech reception threshold rather than the conventional real-valued index in the range $[0, 1]$. In general, these binaural models extend their monaural counterparts by taking two prominent aspects in binaural listening into account: (i) the better ear advantage for lateralised sources i.e., the more favourable SNR at one ear resulted from the head shadow effect; and (ii) binaural unmasking due to interaural time differences at the two ears.

To make an intelligibility prediction, existing methods such as [8, 9] require the binaural signals or corresponding binaural room impulse responses of the target and masking sources. However, such signals or impulse responses are not always available. For example, a sound designer in a radio drama may wish to know the approximate speech intelligibility of a sound scene before sound rendering. Therefore, it is convenient to develop approaches which are able to estimate intelligibility using solely the monaural signals and locations of the sources. Such an approach would permit a flexible sound design process through manipulation of SNR and locations for sound sources to meet some intelligibility criterion. Indeed, the binaural SII [7] meets such requirements, but has been evaluated only in the presence of single masking sources.

The current study proposes a glimpse-based approach for the prediction of binaural speech intelligibility in anechoic conditions using knowledge of the monaural speech and masker signals and their location only (section 2). The proposed approach is evaluated in two listening experiments with both noise and speech maskers. In the first experiment (section 3) subjective intelligibility was tested with a single masker at different locations, while in a second experiment (section 4) multiple maskers were employed.

2. A binaural distortion-based glimpse proportion metric

The proposed measure extends to the binaural domain an objective intelligibility measure operating on monaural signals – the distortion-based weighted glimpse proportion (DWGP) metric [10]. DWGP was originally developed to improve the predictive power of OIMs for modified natural and synthetic speech. In [10], DWGP was extensively evaluated in a range of maskers/SNR combinations, demonstrating high correlations with subjective data.

DWGP consists of two main processes operating on the output of a model of the auditory periphery. First, distortion-weighting correlates the temporal envelopes of speech and speech-plus-masker in each frequency band in an attempt to quantify the effect of speech envelope distortions induced by the masker. Second, energetic masking is assessed by measuring the proportion of 'glimpses' i.e., spectro-temporal regions

where the speech is more energetic than the masker. Glimpse counts in each frequency band are multiplied by the distortion-weightings to compute the DWGP metric.

The binaural DWGP metric – BiDWGP – introduces two new ideas. First, a measure of binaural unmasking, obtained through the binaural masking level difference, is incorporated into the decision of what constitutes a glimpse. Second, the better ear effect is simulated by assuming the existence of binaural glimpses whenever either or both of left and right channels possess a glimpse. The BiDWGP model is schematised in figures 1-3 and its components are described further below.

2.1. Peripheral filtering and binaural unmasking

Each source on a horizontal plane is specified by its location in polar coordinates (r, θ) whose origin is at the centre of the listeners head, where r is the source distance in metres and θ is the azimuth angle subtended by the source relative to the 0° baseline straight ahead of the listener. Signal attenuation with distance is approximated using an inverse-square law. The elevation of source is not considered in this modelling.

The initial stage of the model simulates peripheral auditory filtering and binaural unmasking (Fig. 1). The target speech source s at location (r^s, θ^s) and one or multiple maskers $n1 \dots ni$ at locations $(r^{n1}, \theta^{n1}) \dots (r^{ni}, \theta^{ni})$ are processed independently by a bank of 34 gammatone filters [11] with centre frequencies in the range 100-7500 Hz spaced equally on the ERB-rate scale [12], using an implementation described in [13]. To compute spectro-temporal excitation patterns (STEPs) for the two ears, following [7] a transformation of sound pressure level from the free field to the eardrum [14] results in azimuth- and frequency-dependent gains $d_f(\theta)$, which are then used to weight the outputs of the gammatone filters. The Hilbert envelope of each weighted filter output is smoothed by a leaky integrator with a 8 ms time constant [15], downsampled to 100 Hz and log-compressed. Filterbank outputs from different maskers are summed prior to computation of the STEP.

As proposed in [16] the gain due to binaural unmasking can be measured as the binaural masking level difference (BMLD). This is calculated for each frequency f using a simplified form from [7] (originally Eq. 17 in [17])

$$BMLD_f = 5 \log_{10}\{1 + \frac{C_f[1 + \beta_f^2 - 2\beta_f \cos(\phi_f^s - \phi_f^n)]^2}{\max(1, \beta_f^4)}\} \quad (1)$$

$$\beta_f = \frac{d_f(\theta^s)/d_f(-\theta^s)}{d_f(\theta^n)/d_f(-\theta^n)} \quad (2)$$

where C_f is a frequency-dependent gain [7], $d_f(\theta^s)$ and $d_f(\theta^n)$ are transformation functions for speech and masking sources with separations in azimuth of θ^s and θ^n respectively relative to 0°, and ϕ_f^s and ϕ_f^n denote the interaural phase shifts of the speech and masker at this frequency. In the presence of multiple maskers, the BMLD between the speech and each masker is calculated separately and the overall BMLD at each frequency $\overline{BMLD_f}$ is the mean over the individual masker BMLDs, as shown in the lower part of Fig. 1.

2.2. Distortion weighting

BiDWGP inherits from DWGP [10] a stage of distortion weighting designed to measure the effect of masker-induced envelope perturbations. Distortion weighting involves computing

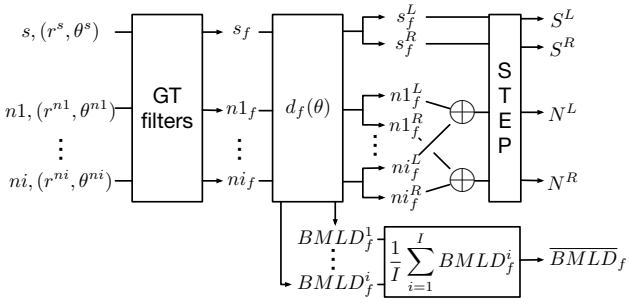

Figure 1: *Stage 1 of the BiDWGP model: from speech and masker signals to excitation patterns and binaural masking level differences. GT filters: gammatone filters.*

the normalised cross-correlation in time of the STEP temporal envelope in each frequency band of the speech signal and the speech-plus-noise mixture. As shown in Fig. 2, distortion-weighting is extended to the binaural case by averaging cross-correlations for the left and right ear STEPs, resulting in a frequency-dependent binaural weighting W_f^{bi}.

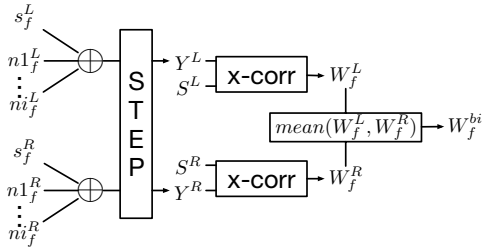

Figure 2: *Stage 2 of the BiDWGP model: computation of a binaural distortion-weighting W_f^{bi} between the envelopes of speech S^L, S^R and speech-plus-noise Y^L, Y^R in each frequency band f.*

2.3. Binaural glimpsing

In the glimpsing model [18] STEPs of speech and masker are compared and time-frequency regions where the speech excitation pattern exceeds that of the masker by a certain amount, known as the local criterion ($LC = 3\,dB$), are treated as speech glimpses. The DWGP [10] modifies the glimpse definition to ensure that glimpses exceed the hearing level ($HL = 25\,dB$):

$$G = S_f(t) > max(N_f(t) + LC, HL) \quad (3)$$

For the BiDWGP, the glimpsing criterion is further extended to incorporate the BMLD and the better ear effect. The frequency-dependent $BMLD_f$ is applied at the stage of glimpse definition, replacing Eq. 3 by

$$G' = (S_f(t) > HL) \wedge (S_f(t) + \overline{BMLD}_f > N_f(t) + LC) \quad (4)$$

Inspired by the binaural SII [7], which models the better ear effect by treating the ear with the largest SNR in each frequency band as the effective SNR, the better ear effect is simulated in the BiDWGP by combining glimpses from the two ears. Glimpses defined by Eq. 4 are computed separately for left and right ear models, G^L and G^R, and combined to produce binaural glimpses G^{bi} in all time-frequency regions where either or both individual ears produce a glimpse, i.e., the inclusive 'or' of

glimpsed spectro-temporal locations for the left and right ears, G^L and G^R (Fig. 3).

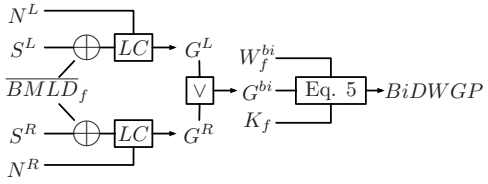

Figure 3: *Stage 3 of the BiDWGP model: computation of binaural glimpses G^{bi} and final objective score.*

2.4. The BiDWGP metric

The final binaural distortion-weight glimpse proportion metric is computed as shown in Eq. 5. The output of the BiDWGP metric lies in the range 0-1, with higher scores indicating better intelligibility.

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

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

Here

- $\mathcal{H}(.)$ is the Heaviside unit step function which counts the number of glimpses in frequency channel f
- K_f is a band importance function interpolated from the values provided in Table 3 of [4].
- T and F are the number of time frames and frequency channels
- $v(.)$ is a quasi-logarithmic function which models the fact that ceiling intelligibility occurs for glimpse proportions substantially lower than 1:

$$v(x) = \frac{\log(1+x/\delta)}{\log(1+1/\delta)},\ \delta = 0.01 \qquad (6)$$

3. Experiment I: Single masker

Harvard sentences [19] uttered by a British English male talker were mixed with two noise maskers, a stationary noise masker (SSN: speech-shaped noise with spectrum matching the long-term corpus average) and a fluctuating noise masker (CS: competing speech) at two SNR levels. CS was generated by concatenating sentences uttered by a British English female talker from the SCRIBE corpus [20]. The SNR levels led to recognition scores of approximately 25% and 50% in each noise masker (-18 and -15 dB for CS; -9 and -6 dB for SSN) when the target and masker were co-located ahead of the listener. The azimuth of the speech source was fixed at 0° relative to the listener (i.e., $\theta^s = 0$), while the azimuth of the masker θ^n varied across conditions. Different source-listener distances r_s and r_n for speech and masker sources respectively were also tested, leading to the 18 conditions shown in Table 1.

A virtual sound field was simulated by convolving monaural speech and noise signals with head-related impulse responses recorded in an anechoic chamber [21]. Speech presentation level was fixed at 63 dB(A) at a distance of 2m from the listener, while the masker presentation level was adjusted to achieve the required SNR. Further presentation level change due to the change of distance from the listener is relative to the reference level at 2m.

Table 1: *Locations of the target speech and masker.*

r_s (m)	r_n (m)	θ^n (°)
2	2	[0 -10 20 -30 60 -90 90 -150 120 180]
1.5	2.5	[0 -45 135 180]
2.5	1.5	[0 45 -135 180]

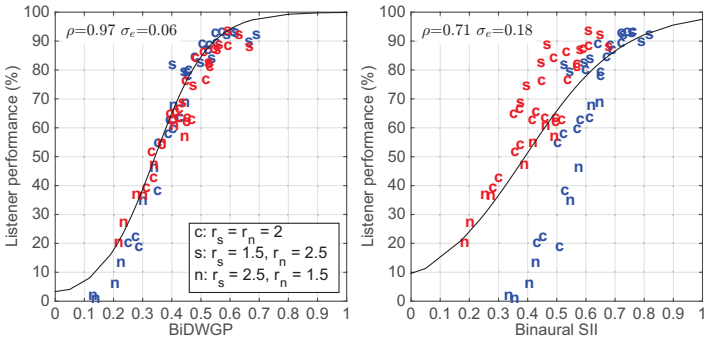

Figure 4: *Listener scores and model predictions by BiDWGP (left) and the binaural SII (right) in single masker conditions (blue: SSN; red: CS).*

In total, there were 72 conditions (2 masker types × 2 SNR levels × 18 locations) in the experiment. Each participant listened to three sentences in each condition. Sentences were blocked by masker type and SNR level. Fourteen native British English speakers listened to the stimuli in a semi-anechoic room over headphones. All participants reported normal hearing.

A 2-parameter logistic fit (Eq. 7) was applied to the raw output o of the model, in order to convert o to the predicted listener performance $f(o)$:

$$f(o) = \frac{1}{1+exp(-(a+b\cdot o))} \qquad (7)$$

The best fit, determined with the nonlinear least squares MATLAB procedure `glmfit`, was achieved at $a = -3.77$ and $b = 11.15$. The correlation ρ between the transformed objective prediction and subjective performance was computed together with the error of the standard deviation σ_d of subject scores in test condition d, defined by $\sigma_e = \sigma_d \cdot \sqrt{1-\rho^2}$.

The binaural SII [7] as the reference measure is also evaluated, after the same procedure of the logistic fit with the parameters $a = -2.36$ and $b = 6.03$. Fig. 4 plots mean listener performance and model scores in each condition. Compared to the binaural SII ($\rho = 0.71$), prediction by BiDWGP are strongly-correlated with listener scores ($\rho = 0.97$). The BiDWGP model also makes good predictions for individual maskers: $\rho = 0.98$, $\sigma_e = 0.06$ for SSN and $\rho = 0.98$, $\sigma_e = 0.04$ for CS. It is interesting to observe that the model achieves a similar level of predictive accuracy for the two maskers, since monaural OIMs typically exhibit decreased predictive power in fluctuating noise maskers [22, 10]. Although the binaural SII also exhibited reasonable predictions for individual maskers ($\rho = 0.92$, $\sigma_e = 0.11$ for SSN and $\rho = 0.89$, $\sigma_e = 0.90$ for CS), the overall correlation is poor due to the model score falling into discrepant numeric range for different maskers.

4. Experiment II: Multiple maskers

The second experiment simulated conditions involving two or three maskers. Moreover, the position of the speech source was no longer fixed solely at 0°. Speech and maskers were at the same distance from the listener ($r_s = r_n = 2$). Azimuth settings for speech (θ^s) and maskers (θ^n) are shown in Table 2.

Table 2: *Azimuth settings of the target speech and the masker.*

Num. maskers	θ^s (°)	θ^n (°)
2	0	[30 -60], [30 90]
	45	[0 90]
	-45	[45 90]
	90	[0 -90]
	-90	[0 -45]
3	0	[30 -60 90], [30 60 90], [-30 60 90]
	30	[0 60 90], [-60 -90 -120]
	-60	[0 90 -120]

Since increasing the number of maskers leads to additional masking, the SNR levels for Exp. II were increased to -12 and -9 dB for CS and -8 and -5 dB for SSN. Within each condition all maskers had the same SNR with respect to the target speech. Some 48 conditions (2 masker types × 2 SNR levels × 12 location settings) resulted from this design. Participants listened to five sentences in each condition. Fourteen native British English speakers with normal hearing participated in Exp. II of whom four had participated in Exp. I. Sentences and competing speech maskers were different from those used Exp. I.

The same logistic fitting procedure used in Exp. I resulted in parameter values of $a = -5.21$ and $b = 13.20$. Fig. 5 plots objective vs. subjective intelligibility in all conditions. In the presence of two or three maskers, the BiDWGP model predicted listeners' scores well ($\rho = 0.91$), though not at the level of the single masker conditions. High correlations for the individual maskers are also seen: (SSN, $\rho = 0.96$, $\sigma_e = 0.07$; CS, $\rho = 0.94$, $\sigma_e = 0.09$). For conditions split by the number of maskers, while the model achieved a correlation of ($\rho = 0.88$, $\sigma_e = 0.11$) when two maskers in present, a similar performance ($\rho = 0.90$, $\sigma_e = 0.12$) was with three maskers. The model displays a tendency to over-estimate listener performance in many CS masking conditions, particularly in the region where listeners' scores are below 60%.

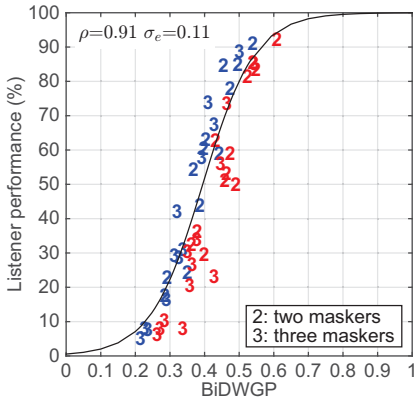

Figure 5: *Listener scores and model predictions in multiple masker conditions (blue: SSN; red: CS).*

5. Conclusion and discussion

Extending a monaural intelligibility predictor to the binaural context resulted in high correlations with listeners' speech identification scores. For single maskers, correlations in excess of 0.97 were observed with both overall and for individual stationary and fluctuating competing speech maskers. In the multiple masker scenarios overall correlation dropped to around 0.91, with better predictions for the individual maskers of 0.96 (SSN) and 0.94 (CS) respectively.

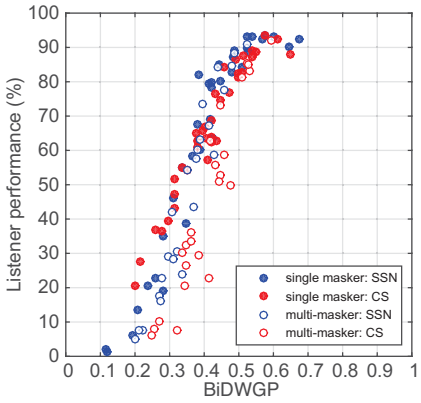

Figure 6: *Scatter plot of objective-subjective intelligibility pairs in Exp. I and Exp. II.*

Some overestimation of intelligibility in the face of multiple competing speech maskers was observed (Exp. II). Fig. 6 combines the results of the two experiments. Based on logistic fits, model predictions at the 50% correct point averaged 0.37 across the two experiments, with values of 0.34, 0.36 and 0.33 for the SSN/single masker, SSN/multiple maskers and CS/single masker combinations. For the CS/multiple masker case the model prediction of 0.43 was significantly higher. Most participants reported that the CS masker conditions were more difficult than those involving the stationary masker due to the need to identify the location of the target speech. We speculate that listeners' attention switching due to source localisation and segregation might negatively impact performance in keyword identification in the presence of multiple competing speech sources. This effect is more evident when noise is more dominant (e.g., listener scores below 40%), suggesting that it may take longer for listeners to locate the target speech in adverse listening conditions. Further experiments are needed to test the cost of attention-switching hypothesis, perhaps using a visual cue to identify the location of the target source. Future work will also address the possibility that informational masking by the competing speech contributes to the greater difficulty of this masker.

The BiDWGP method operates with monaural target and masker signals and knowledge of their locations in distance and azimuth. As such, it is applicable to a wide range of sound generation scenarios where assessment of intelligibility is a concern (e.g., design of radio programmes for narrative or dialogue; loudspeaker placement for public announcements). However, the BiDWGP metric does not take account of reverberant energy and is therefore limited to simulations of free-field conditions: future work will extend the BiDWGP metric to these acoustic conditions.

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.

6. References

[1] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *Proc. ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, 2010.

[2] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, 2010, pp. 1636–1639.

[3] ——, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, 2012, pp. 955–958.

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

[5] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

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

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

[8] S. J. van Wijngaarden and R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4514–4523, 2008.

[9] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hear. Res.*, vol. 275, no. 1-2, pp. 96–104, May 2011.

[10] Y. Tang, "Speech intelligibility enhancement and glimpse-based intelligibility models for known noise conditions," Ph.D. dissertation, Universidad del País Vasco, 2014.

[11] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS Final Report: The Auditory Filterbank," Technical Report 2341, 1988, MRC Applied Psychology Unit.

[12] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

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

[14] E. Shaw and M. M. Vaillancourt, "Transformation of soundpressure level from the free field to the eardrum presented in numerical form," *Journal of the Acoustical Society of America*, vol. 78, no. 3, pp. 1120–1123, 1985.

[15] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The shape of the ear's temporal window," *J. Acoust. Soc. Am.*, vol. 83, no. 7-8, pp. 1102–1116, 1988.

[16] H. Levitt and L. R. Rabiner, "Predicting binaural gain in intelligibility and release from masking for speech," *J. Acoust. Soc. Am.*, vol. 42, no. 4, pp. 820–829, Oct. 1967.

[17] H. S. Colburn, "Theory of Binaural Interaction Based on Auditory Nerve Data. II. Detection of tones in noise. Supplementary material," Tech. Rep., 1977, AIP document No. PAPS-JASMA-61-525-98.

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

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

[20] University College London, Cambridge University, Edinburgh University, the Speech Research Unit and the National Physical Laboratory, "SCRIBE – Spoken Corpus of British English," 1992, online, http://www.phon.ucl.ac.uk/resource/scribe, accessed on 19 Oct 2009.

[21] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *130th Convention of the Audio Engineering Society*, May 2011.

[22] K. S. Rhebergen and N. J. Versfeld, "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.*, vol. 117, no. 4, pp. 2181–2192, 2005.