

**Beyond speech intelligibility and speech  
quality: measuring listening effort with an  
auditory flanker task**

*Mark Andrew Wibrow*

A thesis submitted in partial fulfilment  
of the requirements for the degree of

**Doctor of Philosophy**

of

**University College London**

Department of Speech, Hearing and Phonetic Sciences

University College London

13th July 2015

## **Declaration**

I, Mark Andrew Wibrow, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.


## Abstract

If listening to speech against a background of noise increases listening effort, then the effectiveness of a speech technology designed to reduce background noise could be measured by the reduction in listening effort it provides. Reports of increased listening effort in environments with greater background noise have been linked to accompanying decreases in performance (e.g., slower responses and more errors) which are commonly attributed to the increased demands placed on limited cognitive resources in these challenging listening environments, particularly when performing more than one task. As these cognitive resources are also implicated in maintaining attention and reducing distraction, the work reported here proposes to measure listening effort by measuring changes in distraction while listening to noisy and digitally-noise-reduced speech using an auditory flanker task designed to simulate an everyday situation: listening on the telephone. Over a series of experiments this novel listening effort measure is enhanced by the inclusion of a simultaneous memory task and contrasted with listening effort ratings and conventional speech technology evaluation measures (intelligibility and speech quality). However, while there are indications that increased background noise can increase listening effort and digital noise reduction fails to reverse this effect, the results are not consistent. These equivocal results are discussed in light of the recent surge of interest in listening effort research.

## Acknowledgements

I would like to thank my supervisors: [www.google.co.uk](http://www.google.co.uk) and <http://scholar.google.co.uk/>. I would also like to thank Jyrki Tuomainen for making the latter part of my Ph.D. tolerable, and my viva committee for making me realise that research can be fun. In addition, I would like to thank friends and family who provided nutritional and financial support, and the proof-readers whose advice I capriciously followed or ignored. Finally, I would like to thank Peter Wallis for his invaluable three point plan for completing a Ph.D.:

1. Start the Ph.D.
2. Do the Ph.D.
3. Most importantly of all, stop doing the Ph.D.

This Ph.D. was funded by the voice quality department at  **BlackBerry**. Although the department was disbanded in the middle of the Ph.D., I hope that Nikolai Kouznetsov and Leigh Thorpe would have (at least) been intrigued by the outcome of the research.

# Contents

<b>Contents</b>	<b>5</b>
<b>List of Figures</b>	<b>9</b>
<b>List of Tables</b>	<b>13</b>
<b>Glossary</b>	<b>15</b>
<b>Experimental conditions</b>	<b>17</b>
<b>1 Introduction</b>	<b>18</b>
1.1 Speech technologies and speech technology evaluation . . . . .	18
1.2 Intelligibility and speech quality . . . . .	20
1.2.1 Intelligibility . . . . .	21
1.2.2 Speech quality . . . . .	23
1.3 Beyond intelligibility and speech quality: listening effort . . . . .	24
1.4 The current work . . . . .	25
1.5 Summary and overview . . . . .	26
1.5.1 Overview . . . . .	27
1.5.2 Statistical analyses . . . . .	27

<b>2</b>	<b>Selective attention in an auditory flanker task</b>	<b>29</b>
2.1	Auditory attention and auditory flanker(s) tasks . . . . .	29
2.1.1	Selective attention and ‘Load theory’ . . . . .	31
2.1.2	The flanker(s) task . . . . .	33
2.2	The drift-diffusion model . . . . .	35
2.3	Experiment Ia . . . . .	40
2.3.1	P-Centres . . . . .	41
2.3.2	Materials . . . . .	42
2.3.3	Methods . . . . .	47
2.3.4	Results . . . . .	49
2.3.5	Discussion . . . . .	66
2.4	Experiment Ib . . . . .	70
2.4.1	Materials . . . . .	70
2.4.2	Methods . . . . .	71
2.4.3	Results . . . . .	71
2.4.4	Discussion . . . . .	73
2.5	General discussion . . . . .	75
<b>3</b>	<b>The auditory flanker task with noisy and ‘de-noised’ targets</b>	<b>79</b>
3.1	Selective attention and noise . . . . .	79
3.1.1	Digital noise reduction . . . . .	81
3.2	Experiment II . . . . .	82
3.2.1	Materials . . . . .	84
3.2.2	Methods . . . . .	85
3.2.3	Results . . . . .	85
3.2.4	Discussion . . . . .	97

3.3	Experiment III . . . . .	100
3.3.1	Materials . . . . .	102
3.3.2	Methods . . . . .	102
3.3.3	Results . . . . .	104
3.4	Discussion . . . . .	112
3.5	General discussion . . . . .	113
<b>4</b>	<b>The flanker task and speech technology evaluation</b>	<b>116</b>
4.1	Comparing speech technology evaluations . . . . .	116
4.2	Experiment IV . . . . .	116
4.2.1	Methods . . . . .	118
4.2.2	Tasks . . . . .	118
4.2.3	Results . . . . .	126
4.2.4	Discussion . . . . .	154
4.3	General discussion . . . . .	154
<b>5</b>	<b>Conclusions</b>	<b>159</b>
5.1	Experimental summary . . . . .	159
5.2	Research themes . . . . .	164
5.2.1	The auditory flanker task . . . . .	164
5.2.2	Empirical Bayesian analysis . . . . .	166
5.2.3	Listening effort . . . . .	167
5.3	Conclusions . . . . .	169
	<b>Appendices</b>	<b>172</b>

<b>A Bayesian inference</b>	<b>172</b>
A.1 Bayesian statistics . . . . .	172
A.2 Bayesian model specification . . . . .	173
A.3 Parameter estimation: Markov-Chain Monte Carlo . . . . .	174
A.3.1 MCMC convergence . . . . .	175
A.3.2 Model fit . . . . .	177
A.4 Hypothesis testing and comparisons . . . . .	178
A.5 Reporting Bayesian analysis . . . . .	183
<b>B Graphs for Bayesian estimation</b>	<b>185</b>
B.1 Model specifications . . . . .	185
B.2 HDDM . . . . .	185
B.3 Correlation . . . . .	186
B.4 Repeated measures BANOVA . . . . .	187
B.5 Mixed effects linear regression . . . . .	188
B.6 Mixed effects logistic regression . . . . .	189
<b>C Simulating telephone use in the auditory flanker task</b>	<b>190</b>
C.1 Simulating the distractor speaker using head-related transfer functions . . . . .	190
C.2 Simulating telephone use with a monaural speaker . . . . .	192
<b>D Cognitive failures questionnaire</b>	<b>194</b>
<b>Bibliography</b>	<b>196</b>



## List of Figures

2.1	Energetic and informational masking . . . . .	30
2.2	The Drift-Diffusion Model . . . . .	37
2.3	P-centre alignment . . . . .	43
2.4	Simulated listening environment . . . . .	44
2.5	Flanker trial . . . . .	48
2.6	RT violin plot . . . . .	51
2.7	Posterior distributions for RT regression parameters . . . . .	52
2.8	Posterior distributions for logistic regression parameters . . . . .	54
2.9	Observed and predicted RTs . . . . .	55
2.10	Means, modes and 95% HDIs for drift-rate . . . . .	56
2.11	Drift-rate comparison in congruent and incongruent trials . . . . .	57
2.12	Means, modes and 95% HDIs for non-decision time . . . . .	59
2.13	Non-decision time comparison between congruent and incongruent trials . . . . .	59
2.14	Means, modes and 95% HDIs for threshold separation . . . . .	60
2.15	Threshold separation comparison between congruent and incongruent trials . . . . .	61
2.16	RTs and SOAs in congruent and incongruent trials . . . . .	62
2.17	Effect of SOA on drift-rate in congruent and incongruent trials . . . . .	64
2.18	Congruency $\times$ SOA regression parameters for drift-rate . . . . .	65
2.19	Effect size of SOA on drift-rate in congruent trials . . . . .	65

2.20 Drift-rate in all SOA conditions . . . . . 67

2.21 Drift-rate flanker effects for each participant . . . . . 68

2.22 DDM model fit for experiment Ib . . . . . 72

2.23 Drift-rate means, modes and 95% HDIs . . . . . 74

2.24 Drift-interference with binaural and dichotic distractors . . . . . 74

2.25 Comparison of drift-interference between binaural and dichotic distractors . . . . . 75

3.2 Model fit by congruency, SNR and DNRs for experiment II . . . . . 87

3.3 Drift-rate means, modes and HDIs for experiment II . . . . . 88

3.4 Change in drift-rate across SNRs in experiment II . . . . . 89

3.5 Average drift-interference in experiment II . . . . . 90

3.6 SNR comparisons for drift-interference in experiment II . . . . . 90

3.7 Effects of DNR on congruent and incongruent drift-rate in experiment II . . . . . 92

3.8 Posterior non-decision times and HDIs for all experimental conditions . . . . . 94

3.9 Non-decision time between congruent and incongruent trials . . . . . 94

3.10 SNR comparisons for non-decision time . . . . . 95

3.11 DNR comparisons for non-decision time . . . . . 95

3.12 Posterior threshold separations and HDIs for all experimental conditions . . . . . 97

3.13 Comparison of threshold-separation in congruent and incongruent trials . . . . . 98

3.14 SNR comparisons for threshold-separation . . . . . 98

3.15 DNR comparisons for non-decision time . . . . . 99

3.16 Auditory flanker task embedded in a memory probe trial . . . . . 103

3.17 Selected observed and predicted RT plots for the flanker task . . . . . 105

3.18 Drift-rate means, modes and HDIs for experiment III . . . . . 106

3.19 Drift-interference under high and low cognitive load . . . . . 108

3.20 Selected observed and predicted RT plots for the probe task . . . . . 110

3.21	Probe drift-rate means, modes and HDIs . . . . .	111
3.22	Probe drift-rate comparison between SNRs . . . . .	112
3.23	Probe drift-rate comparison between DNRs . . . . .	113
4.1	Format of a single trial in the probe/flanker task . . . . .	121
4.2	AOSpan trial . . . . .	123
4.3	P.835 user interface . . . . .	125
4.4	Intelligibility and listening effort trial format . . . . .	127
4.5	Selected observed and predicted RT plots for the flanker task . . . . .	128
4.6	Means, modes and HDIs for the flanker drift-rate . . . . .	129
4.7	Drift-interference between high and low cognitive load . . . . .	131
4.8	Selected observed and predicted RT plots for the probe task . . . . .	132
4.9	Posterior means, modes and HDIs for probe drift-rate . . . . .	133
4.10	Posterior means, modes and HDIs for recall . . . . .	136
4.11	AOSpan and CFQ correlation . . . . .	139
4.12	Box-plots of P.835 speech quality ratings . . . . .	140
4.13	Box-plots of intelligibility scores . . . . .	147
4.14	Box-plots of perceived effort ratings . . . . .	150
4.15	Correlation between intelligibility and effort ratings . . . . .	153
4.16	Changes in credible drift-interference with reduced sample sizes . . . . .	157
A.1	Model dependencies for RT linear regression . . . . .	174
A.2	Unconverged and converged MCMC traces . . . . .	176
A.3	Anatomy of a posterior plot . . . . .	179
A.4	Examples of posterior plots for hypothesis testing . . . . .	182
B.1	Bayesian graph for correlation . . . . .	186

B.2	Bayesian graph for BANOVA . . . . .	187
B.3	Bayesian graph for mixed effects regression . . . . .	188
B.4	Bayesian graph for mixed effects logistic regression . . . . .	189

## List of Tables

2.1	The parameters of the DDM . . . . .	38
2.2	Fifty CVC singular noun stimuli . . . . .	44
2.3	Participant and group RTs . . . . .	51
2.4	Participant and group accuracy . . . . .	53
2.5	Drift-rate in congruent and incongruent trials . . . . .	56
2.6	Non-decision time in congruent and incongruent trials . . . . .	58
2.7	Threshold separation means and standard deviations . . . . .	60
2.8	Fitted DDM parameters in experiment Ib . . . . .	73
3.1	Drift-rate means and standard deviations for experiment II . . . . .	86
3.2	Posterior means and standard deviations for non-decision time . . . . .	93
3.3	Posterior means and standard deviations for threshold separation . . . . .	96
3.4	Drift-rate means and standard deviations for experiment III . . . . .	106
3.5	Means and standard deviations for probe-drift rate . . . . .	110
4.1	Flanker drift-rate means and standard deviations . . . . .	129
4.2	Planned comparisons for flanker drift-interference . . . . .	130
4.3	Posterior means and standard deviations for probe drift-rate . . . . .	133
4.4	Selected comparison for the change in probe drift-rate . . . . .	134
4.5	Prior means and standard deviations for recall accuracy . . . . .	135

4.6	Selected posterior comparisons for recall . . . . .	137
4.7	AOSpan and CFQ scores . . . . .	138
4.8	Means and standard deviations for the P.835 evaluation task . . . . .	139
4.9	Posterior comparisons for speech distortion ratings . . . . .	143
4.10	Posterior comparisons for background ratings . . . . .	144
4.11	Posterior comparisons for speech quality ratings . . . . .	145
4.12	Means and standard deviations for intelligibility and listening effort . . . . .	147
4.13	Posterior comparisons for intelligibility . . . . .	148
4.14	Posterior comparisons for listening effort . . . . .	152
B.1	Hyperprior parameters for correlation . . . . .	186
B.2	Hyperprior settings for BANOVA . . . . .	187
B.3	Hyperprior parameters for mixed effects regression . . . . .	188
B.4	Hyperprior parameters for mixed effects logistic regression . . . . .	189

# Glossary

**2AFC** 2-alternative forced choice

**AOSpan** Automated operating span. A measure of attentional control

**Auto-correlation** MCMC samples which tend to clump together (see section [A.3.1](#))

**BANOVA** ANOVA using empirical Bayesian methods

**Burn-in** Samples at the beginning of MCMC sampling which are very inaccurate and need to be discarded (see section [A.3.1](#))

**CFQ** Cognitive failures questionnaire

**DDM** Drift-diffusion model (see section [2.2](#))

**DNR** Digital noise reduction

**Geweke statistic** Test to see if MCMC samples are from the same distribution (see section [A.3.1](#))

**HDDM** Hierarchical drift-diffusion model. DDM fitted using MCMC methods

**HDI** Highest density interval (see figure [A.3](#) and section [A.4](#))

**HRTF** Head-related transfer function (see appendix [C](#))

**ITU** International Telecommunication Union

**MAD** Median absolute deviation

- MCMC** Markov-chain Monte Carlo. A procedure for approximating the parameters of a distribution using random sampling (see section [A.3](#))
- MMSE** Minimum mean-square error DNR algorithm (Ephraim & Malah, 1985)
- Mode** The most credible 1% of the posterior distribution
- MSE** Mean squared error
- NHST** Null hypothesis significance testing
- P.835** Speech quality ratings standardised by the International Telecommunications Union
- Raftery-Lewis procedure** A procedure for obtaining a recommended set of parameters for an MCMC process (see section [A.3.1](#))
- RMS** Root mean square
- ROPE** Region of practical equivalence (see figure [A.3](#) and section [A.4](#))
- SNR** Signal to noise ratio
- SOA** Stimulus onset asynchronicity
- Thinning** Discarding every  $n$ th sample during MCMC sampling to reduce auto-correlation (see section [A.3.1](#))



## Experimental conditions

<b>CL</b>	+60 dB SNR	<b>NG</b>	Negative SOA
<b>CON</b>	Congruent	<b>NN</b>	No DNR (i.e., noisy)
<b>HCG</b>	High cognitive load	<b>PS</b>	Positive SOA
<b>HG</b>	+4 dB SNR	<b>SB</b>	Spectral-subtraction algorithm
<b>INC</b>	Incongruent	<b>SP</b>	Sub-space algorithm
<b>LCG</b>	Low cognitive load	<b>VH</b>	+8 dB SNR
<b>LO</b>	-4 dB SNR	<b>VN</b>	Very negative SOA
<b>MD</b>	0 dB SNR	<b>VP</b>	Very positive SOA
<b>MM</b>	MMSE algorithm	<b>ZO</b>	Zero SOA

## Chapter 1

# Introduction

### 1.1 Speech technologies and speech technology evaluation

“...[S]peech technologies are now everywhere...” asserts the preface to the Springer *Handbook of Speech Processing*, and will soon be “...impossible to miss in our daily life” (Benesty, Sondhi & Huang, 2008, p. vii). Even if speech technologies were not literally ‘everywhere’, then with almost 83 million mobile phone subscriptions (Ofcom, 2013) and 2 million hearing aid users (Action on Hearing Loss, 2011) in the United Kingdom alone, it is certainly the case that at least some speech technologies are difficult, if not impossible, to miss. Furthermore, the development of telephones and hearing aids, and other speech technologies such as voice messaging, real-time translation, cochlear implants and voice input-output communication aids, enable speech based communication between increasing numbers of humans throughout their daily lives, regardless of distance, time, linguistic skill and disability (cf. Whittaker, 2002).

A speech technology is any technology that facilitates communication by “partially replacing the natural air-path” (Wältermann, 2013, p. 23) from the speakers vocal tract to listener’s inner ear. This includes entire telecommunication networks as well as telecommunication and assistive communication devices (labelled “end products” by Benesty et al., 2008, p. vii). But the term also applies to the hardware and software components that they contain. For example, a mobile telephone handset contains an analogue-to-digital converter to transform the continuous voltage representing the air pressure fluctuations received via the handset microphone into a digital signal, and a digital-to-analogue converter to transform digital speech data received over the telephone network into air pressure fluctuations via the handset speaker. In addition, one of several microprocessors on the handset circuit board would be used to compress the

digital speech data for more efficient transmission over networks and also to decode received signals. Some handsets may also contain processors to apply some form of noise reduction to try to reduce interfering noise from the speaker's environment (a technology also increasingly found in hearing aids). Finally, the operating system of a handset may provide automatic speech recognition for voice dialling (or as a voice interface to internet searching), and text-to-speech synthesis for reading menus (for partially sighted users).

With the development of any new technology, regardless of whether it is a component or an end product, there is a need for some form of evaluation. Evaluation is a “vital component” (Polkosky & Lewis, 2003, p. 161) of speech technology development, and can be used as a guide for prototypes or on going development (Swaffield & Richards, 1959), to set performance targets at each stage of development and establish standards (Thorpe, 1998), to select between competing technologies (Schmidt-Nielsen, 1993), or to gauge customer satisfaction with the technology (Chateau, Gros, Durin & Macé, 2006). But perhaps the most compelling reason to evaluate a technology is to measure how the technology fulfils its purpose (Moller, Engelbrecht, Kühnel, Wechsung & Weiss, 2009) or, in other words, to “demonstrate the *utility* of the emerging technology” (Hirschman, 1998, p. 282): as Gaizauskas (1998, p. 251) points out “. . . it is no use having a brilliant piece of technology if it cannot help you to do what you want to do in the actual context of use in which you will deploy it”.

The ‘context of use’ for speech technologies is invariably some form of spoken communication. Real world spoken communication takes place against a complex and challenging background of noise, distraction and multi-tasking (Cherry, 1953; Hughes & Jones, 2003; Baldwin, 2012; Konig, Buhner & Murling, 2005), and given these challenges posed by real world communication, the principle argument of the current work is that to demonstrate the ‘utility’ of a speech technology designed to aid communication, it must be demonstrated that by using the technology, communication becomes less challenging for listeners (or speakers) so that the individual would find it easier to divide their attention between communication and any simultaneous activity (Sörqvist, 2010).

So, the focus of the current work is to investigate a novel evaluation of a speech technology by measuring how the technology reduces the challenge of communication. However, focus will be given only to one side of the communication exchange: listening, and to one particular communication challenge faced by listeners: the division of attention between listening

to speech and performing other tasks. This particular kind of ‘challenging listening’ is also known as *ease of listening* (Feuerstein, 1992; Mackersie, Boothroyd & Prida, 2000) or *listening effort* (e.g., Downs, 1982; Gosselin & Gagné, 2010), and although listening always takes place against a background of noise (Baldwin, 2012), as the noise increases it is commonplace for listeners to report increases in perceived effort (Larsby, Hällgren, Lyxell & Arlinger, 2005). In addition, as background noise increases, performance on concomitant tasks tends to worsen (e.g., Rabbitt, 1966, 1968; Howard, Munro & Plack, 2010; Sarampalis, Kalluri, Edwards & Hafter, 2009; Baldwin, 2012) and as decrements in performance correspond to increments in perceived effort (Larsby et al., 2005), it is common to link decrements in performance to increases in listening effort (Mackersie et al., 2000).

Therefore, the focus of the current work is narrowed further to consider only one class of speech technology: digital noise reduction (DNR), a technology designed to reduce background noise without distorting the signal of interest (Loizou, 2007). It is increasingly common for DNR to be included as a component in end-products such as telephones (Goulding & Bird, 1990; Hu & Lee, 2009) and hearing aids (where only about half of users are satisfied with the results — Edwards, 2007). There is little evidence that DNR processing provides any consistent benefits (compared to unprocessed noisy speech) when using conventional evaluation methods which either measure the percentage accuracy or words correctly reported when listening to DNR processed speech (intelligibility) or collect user opinions regarding the DNR processed speech (quality or preference ratings) (Hu & Loizou, 2007; Stelmachowicz et al., 2010). Consequently, the evaluation of systems with DNR components is a matter of on-going research in both telecommunications (Benesty, Makino & Chen, 2005; Hu & Lee, 2009; Pourmand, Parsa & Weaver, 2013) and audiology (Brons, Houben & Dreschler, 2012; Chung, 2004). According to the arguments presented above, the evaluation of a DNR system should include a demonstration of the ‘utility’ of the system, so if listening to noisy speech increases listening effort, a speech technology designed to reduce background noise should reduce listening effort.

## **1.2 Intelligibility and speech quality**

The idea of using listening effort as a criterion for speech technology evaluation is not new (e.g., Swaffield & Richards, 1959). Nor is the claim that listening to noisy speech can have an impact on performance in other tasks (Rabbitt, 1968, 1966). Nevertheless, over the last ninety

years, speech technology evaluations have fallen into two general classes (Schmidt-Nielsen, 1993; Egan, 1948), neither of them involving listening effort: speech intelligibility and speech quality.<sup>1</sup>

### 1.2.1 Intelligibility

Speech intelligibility tests were developed to characterise the ‘information capacity’ of a technology (cf. Shannon, 1948), and initially focused on quantifying the ‘amount’ of speech information correctly reported when listening via a speech technology (Egan, 1948), although measurements could focus on the amount of speech material that was *not* perceived correctly (Allen, 2005), for example, by analysing patterns of errors (Miller & Nicely, 1955).

In the general form of the intelligibility test (which has remained broadly unchanged since the first half of the twentieth century), speech stimuli (typically nonsense syllables, meaningful words or sentences)<sup>2</sup>, are presented to participants having been processed by the speech technology (or technologies) under evaluation. The stimuli are either presented in isolation or in a ‘carrier phrase’ and participants either repeat back what they hear (open-set responses) or select from a number of given alternatives (closed-set responses or forced-choice). An intelligibility score is then derived from these responses — usually by counting the number of correctly reported units of interest (e.g., phonemes or words). For example, an intelligibility score may be based on the number of correctly identified words presented in isolation (e.g., Pollack, Rubenstein & Decker, 1959), the number of correctly identified phonemes presented in a nonsense-syllables (cf. Fletcher & Steinberg, 1929), or correctly identified keywords presented as part of a sentence (Nilsson, Soil & Sullivan, 1993; Kalikow, Stevens & Elliott, 1977).

A common criticism of traditional intelligibility tests is that in addition to being “tedious and time-consuming” (House, Williams, Hecker & Kryter, 1965, p. 158), they are liable to show ceiling effects (Nilsson et al., 1993; Levitt & Rabiner, 1967), with intelligibility for all

---

<sup>1</sup>Some researchers argue that, in fact, intelligibility is one dimension of speech quality (Grancharov & Kleijn, 2008), whereas others argue that intelligibility is a precondition for speech quality measurements: if intelligibility is low then speech quality measurements are unnecessary (Thorpe, 1998). Others argue that intelligibility and speech quality are only weakly related as correlations between intelligibility scores and quality ratings decrease as the SNR increases (Studebaker & Sherbecoe, 1988), or take a much stronger position that intelligibility and speech quality are independent of each other (Loizou, 2007).

<sup>2</sup>When the speech material involves part-words (e.g., syllables) or nonsense words the term ‘articulation’ is usually applied (Allen, 2005; Möller, 2000; Fletcher & Steinberg, 1929). For the purposes of this discussion no distinction will be made between the two and ‘intelligibility’ will be used, regardless of the meaningfulness of the speech.

the speech technologies under consideration uniformly high, particularly when technologies output speech with low distortion, and intelligibility is above 90% (Hecker, Stevens & Willaims, 1966). Furthermore, intelligibility is dependent on the context provided by the speech material used with higher intelligibility scores when more linguistic context is available (Miller, Heise & Lichten, 1951; Rubenstein, Decker & Pollack, 1959; Francis & Nusbaum, 1999; Hirsh, Reynolds & Joseph, 1954). Intelligibility is also positively correlated with the relative frequency of speech material (e.g., words) in the language (Howes, 1957) although this ‘frequency-effect’ may be more evident in intelligibility tests with open response sets (Pollack et al., 1959). In addition, intelligibility tests using closed response sets typically result in higher intelligibility scores (Pisoni, Manous & Dedina, 1987) as guessing is less likely to result in an incorrect response (Mackersie, Neuman & Levitt, 1999).

Initial alternatives to the basic intelligibility test aimed to provide more information than intelligibility scores. The so called ‘rhyme tests’ (Fairbanks, 1958; House et al., 1965; Voiers, 1967), presented participants with the ends of words (e.g., *-ip*, *-ack*, *-oon*) and required the identification of the initial consonant (e.g., *r*, *b*, *m*). These tests aimed to establish which sounds were more confusable or examine response errors along particular features such as voicing, place of articulation, duration. Focus was generally only on consonants as vowel intelligibility was “normally a problem of relatively minor consequence” (Voiers, 1967, p. 2). However, like the simpler intelligibility tests, the rhyme tests have not proved to be particularly sensitive tests for speech that is highly intelligible, and the success of analysing response errors according to particular phonetic features is entirely dependent on the choice of features (Greenspan, Bennet & Syrdal, 1998).

Further attempts to extend intelligibility tests included measures of response time, usually only being applied when systems produced highly intelligible speech (Mackie, Dermody & Katsch, 1987). For example, Hecker et al. (1966) presented suggestive (albeit not rigorously analysed) data indicating that response times were slower with when very low levels of background noise increased (i.e., from +30 and +20 dB) even though recognition accuracy in a rhyme test was at ceiling (i.e., 100%). Pratt (1981) replicated this effect to evaluate microphone performance. Stimuli were recorded through various microphones and played to participants who were provided with response alternatives on a computer screen. Using an auditory threshold detector, vocal response times were measured from the onset of the presented

word to the onset of the (spoken) recognised word and Pratt found that different microphones yielded different response times when intelligibility was constant, and argued that augmenting intelligibility test with response times would provide more sensitive discriminations than basic intelligibility scores. However, in subsequent research using speech in noise, Mackersie et al. (1999), found contrary to Pratt (1981), intelligibility scores were more sensitive to changes in levels of background noise than response times.

### 1.2.2 **Speech quality**

The measurement of ‘speech quality’ (often referred to as ‘voice quality’ or ‘listening quality’, Rix, 2004) arose out of a need to assess signal distortions that did not affect intelligibility (Egan, 1948), but subsequently have become linked to customer satisfaction (Chateau et al., 2006). What constitutes a ‘quality’ speech technology is left up to the opinion of listeners who make the evaluations (Kondo, 2012) but speech quality is generally considered a multidimensional phenomenon (Grancharov & Kleijn, 2008) which can include subjective judgements of loudness, intelligibility, background noise intrusiveness, and the perception of circuit noise or transmission delay (Möller, 2000). As these judgements are internal to the listener, it is impossible to know how much weight listeners attach to each perceptual dimension (Hu & Loizou, 2007; Nagle & Eadie, 2012).

Speech quality evaluations require listeners to provide ratings of speech samples on a discrete five-point Likert scale (e.g., 1 – Good, 2 – Fair, 3 – Average, 4 – Poor, 5 – Bad) which are averaged to create a ‘Mean Opinion Score’ (MOS). The MOS test is ubiquitous in telecommunication engineering (Thorpe, 1998; Kondo, 2012) and has become standardised by the International Telecommunications Union (e.g., ITU-T Rec. P.835, 2003). Despite this popularity, MOS rating scales are acknowledged as not being generally reliable, producing different results when the tests are repeated (Jekosch, 2002), and also lacking sensitivity, as listeners tend to avoid the extremes of scales (Jekosch, 2010), even though MOS is only reliable for distinguishing systems that lie at these extremes.<sup>3</sup> There may also be cross-cultural differences in ratings as the text labels (or ‘anchors’) for points on the rating scales may not be semantically identical in different languages, and speech technologies may affect languages differently due to the differing acoustic-phonetic structure of these languages (Goodman & Nash, 1982). Further-

---

<sup>3</sup>Leigh Thorpe (formerly head of voice quality research at Blackberry), personal comment.

more, despite high intelligibility being a prerequisite for speech quality evaluations, judgements of speech quality do not necessarily correlate with intelligibility (Punch & Beck, 1986; Boike & Souza, 2000)

### **1.3 Beyond intelligibility and speech quality: listening effort**

Although intelligibility and speech quality have become the dominant forms of speech technology evaluation, it was suggested over fifty years ago that that a key outcome in designing communication systems is a system which “never requires users to exert unreasonable amounts of mental or vocal effort in conversation” (Swaffield & Richards, 1959, p. 65). Subsequent research showed that listening to noisy speech can affect listeners’ recall of the speech material (Rabbitt, 1968, 1966), leading to what has been called the “effortfulness hypothesis” (Tun, Benichov & Wingfield, 2010, p. 731), where the ‘effort’ required to hear noisy or degraded speech uses mental resources that could otherwise be deployed in concomitant tasks. Underlying the effortfulness hypothesis is the concept that individuals have a limit set of attentional resources to deploy in everyday tasks (Kahneman, 1973) representing their ‘capacity’ for attention. Difficult tasks require additional attentional resources and if tasks of sufficient difficulty are attempted, then the attentional capacity is used up, and failures of attention will occur. These failures of attention are shown by slower responses or more mistakes when trying to complete the tasks. It is the degree to which attentional resources are depleted in relation to the individuals attentional capacity that constitutes ‘effort’ and listening effort is believed to be increased when listening to noisy speech as extra attentional resources are deployed to supplement the degraded sensory input, perhaps by coordinating lexical or semantic knowledge (Boothroyd & Nittrouer, 1988; Wild et al., 2012; Mattys, Brooks & Cooke, 2009).

The effortfulness hypothesis has received a great deal of attention in audiological research (Downs and Crum, 1978; Downs, 1982; Edwards, 2007; Rönnberg, 2003, Rönnberg, Rudner and Zekfeld, 2009; Finkelman and Glass, 1970, Rönnberg, Rudner, Foo and Lunner, 2008, Rönnberg et al., 1998, Rönnberg, Rudner, Lunner and Zekveld, 2010), although, ‘listening effort’ still has no agreed definition (McGarrigle et al., 2014). Commonly cited definitions of listening effort include “[t]he cognitive resources allocated for speech recognition” (Picou, Ricketts & Hornsby, 2011, p. 1416), “the attention and cognitive resources required to understand speech” (Gosselin & Gagné, 2010, p. 45), “the attentional requirements necessary to



understand speech” (Bourland-Hicks & Tharpe, 2002, p. 572–573) or “the mental exertion required to attend to, and understand, an auditory message” (McGarrigle et al., 2014, p. 434), and a considerable amount of research has been carried out under the rubric of ‘listening effort’ providing more evidence that the processing of noisy or distorted speech does have a measurable impact on the individual (e.g., Bertoli and Bodmer, 2014; Tun et al., 2010; Picou, 2011; McCreery and Stelmachowicz, 2013; Bourland-Hicks and Tharpe, 2002; Mackersie et al., 2000; Howard et al., 2010; Hornsby, 2013; Downs, 1982; Sarampalis et al., 2009; Larsby et al., 2005; Houben, van Doorn-Bierman and Dreschler, 2013; Zekveld, Kramer and Festen, 2010, see McGarrigle et al., 2014 for a recent summary).

Taken together, the existing research suggests that ‘listening effort’ can be operationalised as any perceived or measured changes in perceptual, cognitive or physiological function that results from listening to speech that has been corrupted during its production, transmission, reception, perception and/or comprehension. All the proposed methods for measuring listening effort involve an intelligibility component (so that participants demonstrate that they have attended to the noisy speech), but differ in the measure used to indicate listening effort. Behavioural approaches require listeners to perform a task in addition to identifying to the noisy speech, and measure listening effort according to performance on this secondary task which could involve memory for all or part of the noisy stimulus (McCreery & Stelmachowicz, 2013; Howard et al., 2010; Sarampalis et al., 2009; Hornsby, 2013) and reaction times to the noisy stimulus (Houben et al., 2013; James, Cheesman, Cornelisse and Miller, 1994), or to an unrelated external cue (i.e., ‘vigilance tasks’, Downs, 1982; Hornsby, 2013; Picou, 2011). Physiological approaches attempt to gauge listening effort by measuring physical changes in the listener while attending to noisy speech including pupil dilation (Zekveld et al., 2010), cortisol levels (Bourland-Hicks & Tharpe, 2002), ERP (Bertoli & Bodmer, 2014), GSR (skin conductance) and heart rate (Mackersie & Cones, 2011) or fMRI (Wild et al., 2012). Finally, subjective approaches require listeners to provide ratings on their perceived effort typically on discrete Likert scales (Larsby et al., 2005; Brons et al., 2012).

## **1.4 The current work**

The research described below aims to investigate whether listening effort can be used as a criterion to evaluate a particular speech technology (digital noise reduction, DNR) using beha-

vioural measures. A simple computer based behavioural measure of listening effort would be a quick, efficient and economically viable means of evaluating speech technologies. This is not to downplay the research into listening effort that involves EEG (Mackersie & Cones, 2011), fMRI (Wild et al., 2012) and or eye-tracking measures (Zekveld et al., 2010). It simply reflects the fact that no matter how reliable or scientifically interesting a test may be, if it requires more expensive equipment, more experienced operators, or considerably more participants and testing time compared to an existing evaluation, it is unlikely to gain any wide acceptance (Houben et al., 2013; Greenspan et al., 1998).

The vast majority of the listening effort research has its origins in audiological research, and is targeted at special populations such as children (Howard et al., 2010), the elderly (Rönnerberg et al., 2010; Bertoli & Bodmer, 2014) or the hearing impaired (Downs & Crum, 1978; Downs, 1982; Sarampalis et al., 2009). However, the current work will focus on using adult listeners with unimpaired hearing and unimpaired cognitive function. In doing so there is always the possibility that adult listeners with unimpaired hearing and unimpaired cognitive function are able to complete any listening effort task without difficulty, except in the presence of extreme noise or distortion which renders the speech unintelligible (at which point a simple intelligibility test would be sufficient rather than a listening effort test). Thus, throughout the research described below, it will be borne in mind that an attempt to find a behavioural measure of listening effort to evaluate DNR using ‘normal’ listeners maybe *a priori* impossible. Nevertheless, even in the event of a failure to find a consistent, reliable and practical behavioural measure in listening effort, the research is considered useful (and particularly apposite given the surge in interest in measuring listening effort) as it may help to delineate future research directions.

## 1.5 Summary and overview

Speech technologies that are designed to be used for spoken communication should make communication easier and evaluating speech technologies is an important activity to ensure (among other things) that they do what they are intended (or claimed) to do (Gaizauskas, 1998). Conventional evaluations such as intelligibility and speech quality overlook the fact that there are effects of listening to noisy or distorted speech through speech technologies that go beyond the ability to identify the words spoken or a preference for listening to one technology over another.

Listening effort is broadly defined as changes in the deployment, maintenance and control of attention when listening to noisy or distorted speech (Picou et al., 2011; Gosselin & Gagné, 2010; Bourland-Hicks & Tharpe, 2002; McGarrigle et al., 2014). Research suggests that listening effort can be measured using (i) behavioural measures such as reaction times (Houben et al., 2013) and accuracy (Sarampalis et al., 2009) in primary or secondary tasks, (ii) physiological measures, such as pupil dilation (Zekveld et al., 2010) or brain activity (Wild et al., 2012), or (iii) subjective listener ratings of listening effort (e.g., Larsby et al., 2005). The current work will focus on developing a behavioural measure of listening effort due to the relative simplicity and cost-effectiveness of behavioural methods in comparison to physiological methods (Houben et al., 2013), and their increased objectivity in comparison to subjective methods (Sarampalis et al., 2009). Although the majority of listening effort research has its origins in audiology and uses ‘special’ populations (such as children or the elderly) or clinical populations, the current work aims to investigate the use of listening effort to evaluate speech technologies with unimpaired adult listeners.

As increased listening effort is linked to increased levels of background noise, the work described below will also focus on a technology designed to remove background noise, digital noise reduction (DNR). If listening to noisy speech increases listening effort, then DNR systems should be evaluated in terms of the reduction of listening effort that they provide, in order to ensure that they do make communication easier (Downs, 1982).

### 1.5.1 Overview

The remainder of this thesis is as follows: In chapter 2, the behavioural task used throughout this thesis (an auditory ‘flanker task’) is introduced and some important properties of the task are discussed. In chapter 3, the effects of noise and digital noise reduction on performance in the flanker task are investigated along with the effects of adding a secondary task to increase the effort required to successfully complete the flanker task. Chapter 4 compares the auditory flanker task with some conventional speech technology evaluations, and finally chapter 5 summarises the work presented throughout this thesis and suggests future directions for research.

### 1.5.2 Statistical analyses

All the analyses of the experiments below use empirical Bayesian methods to test hypotheses and make inferences (e.g., Kruschke, 2010a, 2010b, 2011, 2013, 2015) rather than traditional

null hypothesis significance testing (NHST). Although Bayesian inference is a matter of active research (Wiecki, Sofer & Frank, 2013) the basic ideas (e.g., Bayes, 1763) predate the establishment of the techniques used in NHST (e.g., Fisher, 1934, Neyman and Pearson, 1933). Consequently, one of the aims of the current work is to provide practical examples of Bayesian analysis applied to psychological research, including ANOVA, mixed-effects linear regression, mixed-effects logistic regression and correlations. Appendix A provides a (non-exhaustive) account of some the key details of Bayesian analysis and it may necessary to consult appendix A first before reading the analyses of the experiments.

## Chapter 2

# Selective attention in an auditory flanker task

## 2.1 Auditory attention and auditory flanker(s) tasks

In order to communicate successfully, listeners must attend to the incoming speech of their conversation partner in order to respond appropriately. However, in typical listening environments the incoming speech is just one of many sources of sound arriving at the listener's ears, and in particular, the other sources of sound could be irrelevant speech from other conversations. Thus, the listener's task is not only to attend to their conversation partner's speech but also to ignore other speech in the environment.

The ability to focus on one sound in the auditory environment (or 'auditory scene') depends on perceptual processes which divide the auditory scene into perceptually distinct 'streams' of auditory information or *auditory objects* (Alain & Arnott, 2000; Nudds, 2007; Shamma, 2008; Shinn-Cunningham, 2008; Kubovy & Valkenburg, 2001; Griffiths & Warren, 2004), which are perceived as coming from the same source. The formation of auditory objects involves a bottom-up analysis of the acoustic scene in terms of its temporal and spectral properties (Alain & Izenberg, 2003), generic grouping processes that group the components according to similarities in structure, intensity, timing or location (Bregman, 1990; Best, Gallun, Carlile & Shinn-Cunningham, 2007; Darwin & Hukin, 1999; Kubovy & Valkenburg, 2001) and the use of 'top-down' information (e.g., linguistic knowledge — Shinn-Cunningham and Wang, 2008) to bias the assignment of components to perceptually distinct auditory objects (Shamma, 2008).

In challenging listening situations these processes can be hindered by the overlap in frequency and time of spectral components from distracting auditory sources (e.g., other conversations) which obscure (i.e., mask) the spectral components from the target auditory source (e.g.,

the speech of a conversation partner) making them less intelligible. This *energetic masking* is usually contrasting with *informational masking* (Durlach et al., 2003; Brungart, Simpson, Ericson & Scott, 2001; Cooke, Lecumberri & Barker, 2008; Ihlefeld & Shinn-Cunningham, 2008; Lutfi, 1990) which has been defined as “everything that reduces intelligibility once energetic masking has been accounted for” (Cooke et al., 2008, p. 414) and is associated with the degree of uncertainty that exists regarding whether particular spectral components should be assigned to one auditory object or another (Barker & Shao, 2009). Figure 2.1 illustrates this contrast between informational and energetic masking.

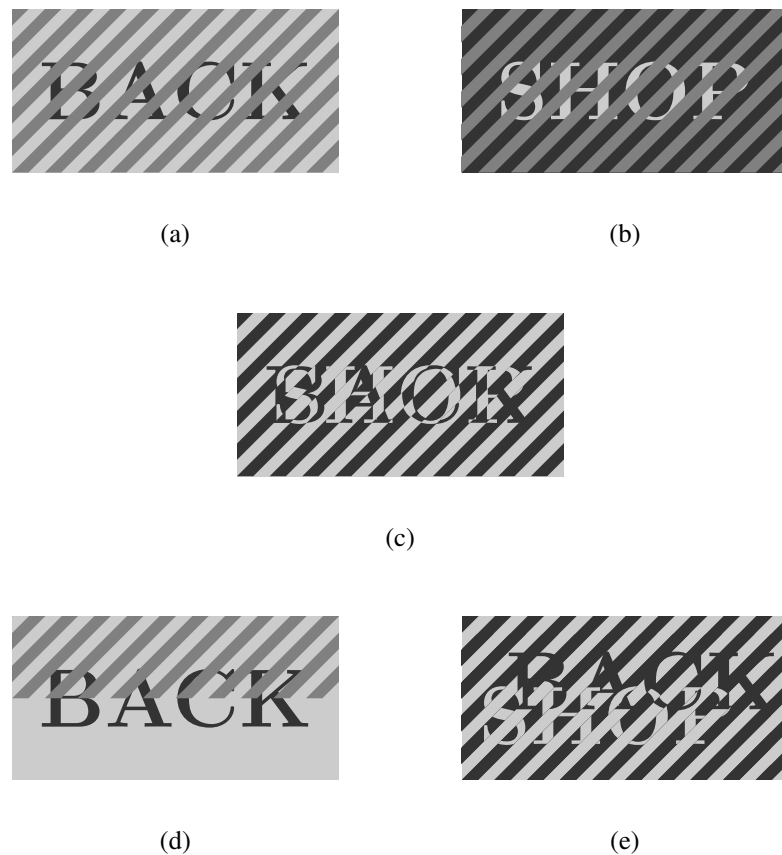


Figure 2.1: Visual analogies of energetic and informational masking in the auditory domain. In (a) and (b) parts of the object are obscured by another object (energetic masking), but the words are still visible so there is no uncertainty as to which parts belong to which object. In (c) the same parts of each of the objects from (a) and (b) are obscured, so there is the same energetic masking but there is more uncertainty regarding which parts belongs to which objects (informational masking). Spatial separation of targets and distractors (maskers) reduces the effects of both energetic masking (d) and informational masking (e). Based on Barker and Shao (2009) and Mattys, Brooks and Cooke (2009).

One consequence of informational masking is the “increased difficulty of attending to one [auditory object] among many” (Shamma, 2008, p. 1143) leading to increased failures of attention (Gutschalk, Micheyl & Oxenham, 2008), particularly when the other auditory objects are from speech sources (Brungart et al., 2001; Kidd Jr., Richards, Mason, Gallun and Huang, 2008; Wightman and Kistler, 2005). These failures of attention typically involve reporting part or all of the distracting speech instead of the target speech (Brungart & Simpson, 2007), or making slower responses to target speech in the presence of distracting speech (Chan, Merrifield & Spence, 2005; Francis, 2010).

### 2.1.1 Selective attention and ‘Load theory’

‘Selective attention’ refers to the perceptual and cognitive processes that control the selection of target (or task-relevant) information and inhibition of distracting (or task-irrelevant) information and the conditions under which failures of attention can occur (Driver, 2001; Guerreiro, Murphy & Gerven, 2010). Two of the key issues in selective attention research have involved (i) whether selective attention processes focus on task-relevant information at a relatively early or relatively late stage in processing, and as a consequence, (ii) the extent to which task-irrelevant information is processed (Driver & Tipper, 1989; Lachter, Forster & Ruthruff, 2004, 2008).

In the auditory domain, early research suggested that the focus of attention was established fairly early and that very little processing of the irrelevant information took place (Broadbent, 1958). For example, Cherry (1953) presented two speech samples dichotically and found that participants shadowing the target speech in one ear could not remember any of the distracting speech in the other. In addition, although participants were aware that the irrelevant speech was, in fact, speech, they were unable to identify the language, or unanimously identify if the distracting speech was reversed; they were, however, able to identify changes in the gender of the speaker producing the irrelevant speech. However, participants were required to shadow the target speech and this additional task demand beyond simple speech identification may have increased engagement with the task reducing the interference from the irrelevant speech (Hughes, Hurlstone, Marsh, Vachon & Jones, 2012; Murphy, Fraenkel & Dalton, 2013). Subsequently, Moray (1959) found that around a third of participants recognised their own name in supposedly unattended speech, a finding which was replicated by Wood and Cowan (1995) who also suggested that participant’s attention had slipped to the unattended speech (see also Lachter et al.,

2004). Further research suggested that the susceptibility to irrelevant speech can be assessed by using measures of attentional control (Conway et al., 2005) and individuals who score highly on these measures are less likely to be distracted (Conway, Cowan & Bunting, 2001) and less likely to fail to attend to targets in demanding tasks (Colflesh & Conway, 2007).

'Load theory' (Lavie, 1995, 2000, 2005, 2010) is a recent model of selective attention, developed specifically to account for the circumstances under which task-irrelevant information is processed and the degree to which it is processed. Load theory was originally developed as an account of selective attention in the visual domain. However, recent research has begun to consider its relevance in the auditory domain either implicitly (Mattys et al., 2009; Mattys & Wiget, 2011) or explicitly (Gomes, Barrett, Duff, Barnhardt & Ritter, 2008; Francis, 2010; Murphy et al., 2013). Load theory proposes that selective attention has two levels of focus represented by two limited-capacity sub-systems whose interaction can result in task-irrelevant information being rejected at a relatively early stage (resulting in very little processing of the irrelevant information), or at a relatively late stage (resulting in the irrelevant information being processed to a considerable degree). The first sub-system is a passive perceptual system which automatically processes all incoming sensory information from both the relevant and irrelevant stimuli. However, as the perceptual system is subject to capacity limits, more complex task-relevant stimuli require more perceptual capacity, leaving less spare capacity for processing irrelevant information (Lavie, 2000; Francis, 2010; Lavie, 2010; Dyson & Quinlan, 2003). This reduced availability of perceptual capacity is referred to as 'perceptual load' and results in a reduction or elimination of the processing of irrelevant information.

The second sub-system is an active cognitive system which maintains current task goals by controlling the focus of attention (i.e., what constitutes task-relevant information) and inhibiting responses to irrelevant information that is processed by the perceptual system (Conway et al., 2001; Ahmed & de Fockert, 2012; Dalton, Santangelo & Spence, 2009; Lavie, 2005). The cognitive system also has a limited capacity, and more complex tasks require more cognitive capacity leading to failures in maintaining the focus of attention, or failures in inhibiting irrelevant information (Lavie, 2000; Lavie & De Fockert, 2005; Francis, 2010; Lavie, 2010). The reduced availability of cognitive capacity is referred to as 'cognitive load' and results in increased processing of any irrelevant information that has passed through the perceptual system.



### 2.1.2 The flanker(s) task

The ‘flanker(s) task’ (Eriksen & Eriksen, 1974; Eriksen, 1995) has been used to evaluate load theory and the impact of perceptual and cognitive load on the degree to which irrelevant information is processed in both the visual domain (Lavie, de Fockert & Viding, 2004) and the auditory domain (Francis, 2010; Murphy et al., 2013). In addition, the flankers task has been used to investigate the control of attention in infants (Smith & Trainor, 2011), children (McDermott, Pérez-Edgar & Fox, 2007; Mullane, Corkum, Klein & McLaughlin, 2009), and the elderly (Guerreiro et al., 2010), in addition to clinical populations involving attention deficit hyperactivity disorder (Mullane et al., 2009) and Parkinson’s disease (Praagstra, Stegeman, Cools & Horstink, 1998).

In the original flankers experiments, Eriksen and Eriksen (1974) presented a target letter (task-relevant information) surrounded (i.e., ‘flanked’) by distractor letters (task-irrelevant information). For example, the target letter *K* was flanked by the distractor *S* to produce a flanker stimulus *S S S K S S S*. Each letter was associated with one of two responses (which involved pressing a lever positioned to the participants’ left or right) or no response (i.e., ‘neutral’ letters). Response times (RTs) for correct responses were fastest when the distractors were ‘response compatible’ or *congruent* with the target (i.e., associated with the same response) and slowest when the distractors were ‘response incompatible’ or *incongruent* with the target (i.e., associated with opposite response). The difference in RTs between trials where the distractor response was congruent with the target response and trials where the distractor response was incongruent with the target response was labelled the ‘flanker effect’, and is taken to indicate the degree to which a distractor is perceived (e.g., Chan et al., 2005; Murphy et al., 2013; Paquet, 2001; Lavie et al., 2004).

In what is claimed to be the first demonstration of flanker effects in the auditory domain, Chan et al. (2005, experiment 1) presented target words from a loudspeaker 1.11 m in front of the participant (i.e., at 0° azimuth and elevation) and distractor words from speakers either side of the target loudspeaker (i.e., at ±30° azimuth and 0° elevation). Participants were required to determine if the target word was *bat* or *bet*. Chan et al. used ‘inverse efficiency’ as their principle performance measure, which was calculated as the mean response time for a particular condition divided by the accuracy for that condition (but still measured in seconds). Inverse efficiency scores were lower in congruent trials than in incongruent trials, indicating that listeners

were faster and more accurate in congruent trials compared to incongruent trials. Chan et al. (2005, experiment 2) also showed that the flanker effect was still evident with increased spatial separation between the target and the distractors (at  $\pm 60^\circ$  and  $\pm 90^\circ$  azimuth) suggesting that the effect could not be explained by masking of the target by the distractor at the auditory periphery.

The flanker effect is explained “at least in large part” (Eriksen, 1995, p. 101) by ‘response competition’ (Hazeltine, Poldrack & Gabrieli, 2000) where the distractor is processed sufficiently to influence the decision process that determines the response that is finally made. In order to complete the task correctly the participant must inhibit the ‘competing’ response. Crucially, the flanker effect (and interference from distractors in general) can be reduced by making the target more complex and perceptually challenging to process (e.g., Lavie & De Fockert, 2003; Hughes et al., 2012) and increased by making the task more challenging and more demanding to complete (e.g., Lavie et al., 2004; Lavie & De Fockert, 2005; Francis, 2010). According to load theory, stimulus complexity constitutes a ‘perceptual load’ which exhausts the capacity of the passive perceptual system resulting in less processing of the distractor and a reduction in the flanker effect. By contrast, increases in task complexity constitute a ‘cognitive load’ (Lavie, 2000) which exhausts the capacity of the active cognitive system resulting in a reduction in the control of distractor rejection and an increase in the flanker effect.

Francis (2010) illustrated effects of both perceptual and cognitive load in a speech-based auditory flanker task where participants had to attend to a target presented to both ears (in stereo) while ignoring a monaural distractor presented to either the left or right ear. Stimulus complexity (i.e., perceptual load) was introduced by adding a tone to the target word and requiring participants to respond only if the tone matched certain criteria. In low perceptual load conditions the response criterion was based on one perceptual feature (pitch). In high perceptual load conditions the response criterion was based on two perceptual features (pitch and modulation). Task complexity (i.e., cognitive load) was introduced by requiring participants to divide their attention between remembering one digit (low cognitive load) or six digits (high cognitive load) and successfully completing a flanker trial. As predicted by load theory, Francis found increased perceptual load resulted in a reduction in the flanker effect and increased cognitive load resulted in an increase in the flanker effect (although this latter effect was enhanced by increasing the spectral overlap between the target and distractor).

The results from Francis (2010) (and also Lavie et al., 2004; Dalton et al., 2009; Ahmed and de Fockert, 2012) show that increased cognitive load can lead to increases in interference from distractors. While this shows that attending to targets and resisting interference from distractors are dependent on the level of cognitive load it also implies that the degree of interference from distractors could be taken as a measure of cognitive load. Furthermore, if increased listening effort is linked to increases in cognitive load and these increases are qualitatively and quantitatively similar to the experimental manipulations used by Francis (2010) which led to increased interference from distractors (and there are indications that they are — see Howard et al., 2010; Sarampalis et al., 2009) then the degree of interference from distractors could be used as a measure of listening effort (cf. Dhamani, Leung, Carlile & Sharma, 2013). No existing research appears to have used a flanker task in this way, so using an auditory flanker task to measure listening effort could not only represent an advancement in the understanding of how listening effort in challenging listening situations affects attention, but could also provide a novel measure of listening effort. Furthermore, by examining how interference from distractors changes when targets are either contaminated by background noise or processed with a speech technology to remove background noise (DNR), the flanker task could then be used to evaluate DNR in terms of the reduction in listening effort it provides (cf. Sarampalis et al., 2009).

## **2.2 The drift-diffusion model**

Distraction in the flanker task is assessed by contrasting RTs and accuracies from congruent and incongruent trials, typically by comparing some measure of central tendency (e.g., the mean or median RT) between congruent and incongruent trials (e.g., Francis, 2010; Murphy et al., 2013). RTs are often further analysed by calculating an ‘interference’ measure using the difference in mean RT between congruent and incongruent trials (e.g., Wyatt & Machado, 2013; Lavie et al., 2004), and sometimes normalising the difference by dividing by the mean of the RTs in both congruent and incongruent trials (Francis, 2010). Another measure involves dividing mean RT by accuracy to create a composite ‘inverse-efficiency’ measure (Chan et al., 2005) which is calculated for congruent and incongruent trials separately.

However, RTs are typically positively skewed making inference with traditional statistical analyses (which assume a normal distribution) inadequate (Van Zandt, 2000; Ratcliff, 1979). This can clearly be seen in the RTs collected in experiment Ia (see figure 2.6 on page 51)

where the density of the RTs is clearly positively skewed in both congruent and incongruent conditions. Nevertheless, typical analyses of RTs in flanker tasks (e.g., Francis, 2010; Chan et al., 2005; Murphy et al., 2013) perform an ANOVA on the mean RTs for each participant in each condition and impose normality on the dependent variables by discarding outliers or applying transformations to the data (Ratcliff, 1993) although this runs the risk of introducing bias into the models (Ulrich & Miller, 1994). Other approaches to modelling RTs involve using different distributions which include parameters to model the positive skew such as the log-normal (Rouder, Province, Morey, Gomez & Heathcote, 2014), the inverse Gaussian (Lachaud & Renaud, 2011), or the ex-Gaussian (McVay & Kane, 2012; Shahar, Teodorescu, Pereg & Meiran, 2014; Hervey et al., 2006).

RT and accuracy represent indices of a number of underlying processes, including stimulus encoding, response selection, speed-accuracy trade-off and response initiation (Merkt et al., 2013; Wagenmakers, 2009). Despite attempts to relate experimental manipulations to changes in parameters of distributions such as the ex-Gaussian (Shahar et al., 2014; Hervey et al., 2006), it is not always clear that researchers modelling skewed RTs have given much consideration to how the psychological processes underlying the observed RTs are related to the parameters of any proposed distribution (although see Rouder et al., 2014). With this in mind, the auditory flanker task developed below models responses in the flanker task using the drift-diffusion model (DDM —Ratcliff, 1978; Ratcliff and Rouder, 1998; Ratcliff and Tuerlinckx, 2002; Vandekerckhove, Tuerlinckx and Lee, 2011, see Voss, Nagler and Lerche, 2013 and Wagenmakers, 2009 for reviews) which can be used to model the observed RTs and accuracies in a speeded '2-alternative forced choice' task (2AFC, i.e., a task with two possible responses, such as the flanker task). The DDM models 2AFC responses as the result of a noisy decision process that accumulates information over time for each of two responses. The decision process commences after an initial (sensory) processing of a stimulus and continues until sufficient information for one of the responses has been accumulated and a response is made.

More precisely, RTs are modelled as a Wiener process (e.g., Vandekerckhove et al., 2011): a random walk from an initial state of rest towards one of two thresholds. The thresholds are upper and lower boundaries representing different responses (e.g., correct or incorrect responses). The random walk represents the process of information accumulation (i.e., the evaluation of incoming sensory information) subject to random fluctuations (i.e., noise). As information for the

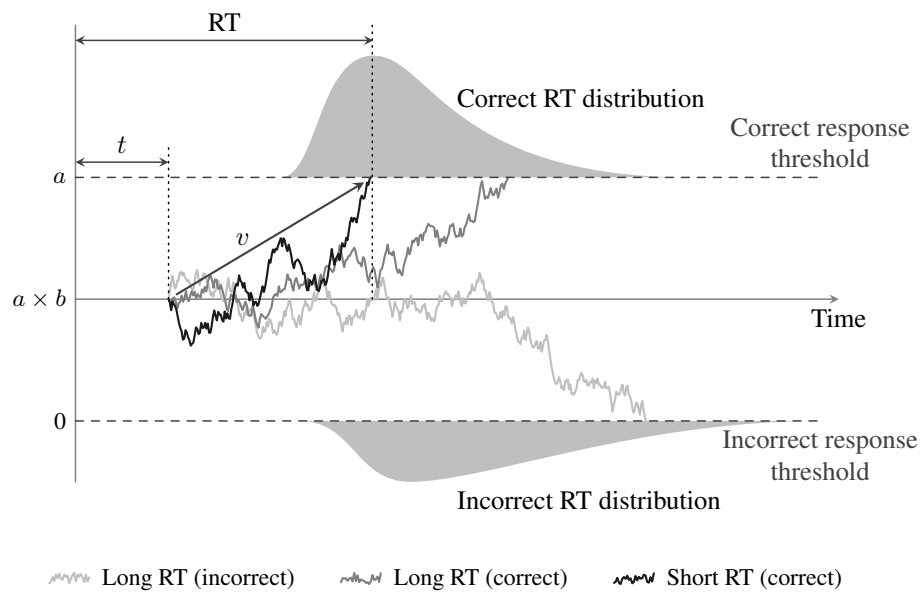


Figure 2.2: The parameters of the Drift-Diffusion Model (DDM), which models RTs as the noisy accumulation of information. Examples of three decision processes are shown, with one process annotated with the parameters of the DDM. The DDM models RTs using the average rate of information accumulation  $v$  from a start point  $b \times a$  to one of two thresholds (0 or  $a$ ), with an initial delay  $t$  representing non-decision processes.

response associated with the upper threshold is accumulated the trajectory of the random-walk ‘drifts’ towards the upper threshold and as evidence for the response associated with the lower threshold is processed the trajectory of the random-walk drifts towards the lower threshold. When either of the thresholds is crossed the decision process terminates, a response is made and the RT recorded. The drift process is illustrated in figure 2.2 which shows the four basic parameters of the DDM:  $a$ ,  $b$ ,  $t$ , and  $v$  which are summarised in table 2.1 (p. 38).

The  $a$  parameter represents the separation between the response thresholds and is often interpreted as an indicator of the speed-accuracy trade off (Krajbich, Armel & Rangel, 2010) and response caution or ‘conservatism’ (Voss, Rothermund & Voss, 2004). Larger threshold separation indicates that more evidence was accumulated in favour of a particular response (assuming no bias for one response over another) before a response was made. Although there is no obligatory mapping between experimental responses and the thresholds in the DDM, in the flanker tasks reported below the upper threshold represents a correct response and the lower threshold represents an incorrect response. The  $b$  parameter represents any bias that participants may have

Parameter	Description
$v$	<i>Drift-rate</i> The average rate of information accumulation in the drift process.
$a$	<i>Threshold separation</i> Response ‘caution’. The amount of information that is accumulated before a response is initiated.
$t$	<i>Non-decision time</i> The time taken for processes not associated with information accumulation (sensory encoding and response initiation).
$b$	<i>Bias</i> The <i>a priori</i> bias for making a particular response; the relative weight given to information associated with the lower threshold compared to information associated with the upper threshold

Table 2.1: The parameters of the drift-diffusion model.

for making one response over another. Bias can be specifically manipulated by increasing the probability that one response is will be made, or by assigning a reward for a particular response (e.g., Mulder, Wagenmakers, Ratcliff, Boekel & Forstmann, 2012). However, in experiments without these manipulations (as in the experiments reported below) this parameter is fixed at 0.5 to reflect the fact that participants have no prior knowledge which physical response (i.e., left or right) corresponds to a correct or incorrect response on a trial-by-trial basis, so cannot have any bias to respond in one way or the other (Merkt et al., 2013, p. 324, footnote 1).

The  $v$  parameter represents the average rate of information accumulation for the two responses. For the response associated with the upper threshold  $v$  is positive and for the response associate with the lower threshold  $v$  is negative. For the flanker task, the DDM makes specific predictions regarding the drift-rate during the decision process. For congruent trials, information from the distractor is consistent with information for the target (correct) response so slips or leaks of attention to the distractor would not alter the trajectory of the diffusion process (i.e., the average rate of information accumulation). In addition, with both the target and the distractor associated with the same response, errors are less likely. Assuming that on average errors are relatively low, in both of these cases the diffusion process will drift towards the upper (correct) threshold more quickly, resulting in a higher drift-rate. For incongruent trials, information from

the distractor is inconsistent with information for the target response and consistent with the incorrect response and slips or leaks of attention to the distractor would alter the trajectory of the diffusion process towards the lower (incorrect) boundary. In addition, with the distractor associated with the opposite response to the target, failures of attention are more likely to result in an incorrect response. Assuming that on average errors are relatively low, then in both of these cases the diffusion process drifts towards the upper threshold less quickly, resulting in a lower drift-rate.

The  $t$  parameter represents the contribution of any ‘non-decision’ processes to the overall RT. It includes the initial processing of stimuli required to initiate the decision process and time to physically initiate the response. Logically, these processes occur at opposite ends of the decision process but for the purpose of the modelling (and in particular, parameter estimation) they are combined together into a single parameter. It is also important to note that sensory processing takes place throughout the decision process until one of the thresholds is crossed (Krajbich, Lu, Camerer & Rangel, 2012) but that the  $t$  parameter includes the portion of sensory processing that occurs prior to the onset of the decision process.

The DDM has become increasingly popular for modelling RTs in a wide variety of psychological research (Voss et al., 2013), including purchasing decisions (Krajbich et al., 2012), decisions under high and low pressure (Milosavljevic, Malmaud, Huth, Koch & Rangel, 2010), speed-accuracy trade-offs in visual perception (Zhang & Rowe, 2014), memory retrieval (Ratcliff, 1978; Pearson, Raškevičius, Bays, Pertzov & Husain, 2014), semantic categorisation (Klauer, Voss, Schmitz & Teige-Mocigemba, 2007), visual word-recognition in dyslexia (Zeguers et al., 2011), and inference in ageing populations (McKoon & Ratcliff, 2013). But although the DDM has also been used to model RTs in visual flanker tasks (e.g., Merkt et al., 2013), at the time of writing the work presented below appears to be the first attempt to model RTs in an auditory flanker task.

However, the application of the DDM in the auditory flanker tasks presented below is not simply motivated by analytic novelty. It is hoped that using the DDM will provide insight into the processes underlying performance in the flanker task. In addition, as the experiments reported below modify the flanker task in different ways to vary the amount of listening effort that participants require to successfully complete the flanker task, it is hoped that the DDM will provide more insight into exactly how these experimental manipulations affect the under-

lying decision process, and whether the DDM parameters relate in any way to the evaluation of listening effort.

In addition, the DDM models trials where correct responses *and* incorrect responses are given, so the relationship between RT and accuracy (e.g., Swensson, 1972) is automatically included in the model. This contrasts with approaches which analyse RT and accuracy separately with only RTs being analysed from trials in which a correct response is given (although see Davidson and Martin, 2013). Although some researchers combine RT and accuracy into a composite score such as inverse efficiency (e.g., Chan et al., 2005) these combinations can increase the variation in the derived statistic or even introduce false effects from the data (Bruyer & Brysbaert, 2011).

## 2.3 Experiment Ia

The aims of experiment Ia were threefold. The first aim was to replicate flanker effects using a novel 3D auditory display of a monaural target and a binaural distractor to simulate telephone use in the listening environment illustrated in figure 2.4 (p. 44).<sup>1</sup> The second aim was to confirm the suitability of the drift diffusion model for analysing responses in an auditory flanker task. Despite the fact that the flanker task was established over forty years ago (Eriksen & Eriksen, 1974) and the DDM first proposed over thirty years ago (albeit for modelling memory retrieval Ratcliff, 1978), it appears that it is only recently that a few studies have applied the DDM to flanker tasks (e.g., Merkt et al., 2013; White, Ratcliff & Starns, 2011). However, none of these studies used the DDM to model performance in *auditory* flanker tasks. Specifically, it was expected that in congruent trials, the drift-rate (i.e., the  $v$  parameter) would be higher (i.e., larger in magnitude and more positive) than in incongruent trials, reflecting the fact that in incongruent trials slips or leaks of attention to the distractor would alter the trajectory of the decision process towards the incorrect (i.e., distractor) response. There were no specific predictions regarding the threshold-separation ( $a$ ) and non-decision time ( $t$ ), but these parameters would still be estimated from the data to establish if there was any relationship between threshold-separation and non-decision time and flanker performance.

The third aim of experiment Ia was to generalise the approaches of Chan et al. (2005), Francis (2010) and Murphy et al. (2013) and confirm flanker effects when using a much larger

---

<sup>1</sup>See also appendix C for details of how the listening environment was simulated.



selection of stimulus words. This was considered necessary, as speech technology evaluation (the intended application of the flanker task) typically draws from large inventories of speech material at their disposal (although they usually only use a subset of them — Hu and Loizou, 2007). These inventories are designed to increase ‘coverage’ and be in some way representative of the language from which they are derived. Without a representative sample of the speech sounds in the target language the evaluation of a speech technology may be invalid (Egan, 1948), as it would only reflect performance on this small number of sounds. However, previous speech-based auditory flanker tasks have used only a small number of words with Chan et al. (2005) using only four words (*bat*, *bed*, *rod*, and *red*), Francis (2010) using only two (*bead* and *bad*), while Murphy et al. (2013) used eight ‘words’ (the spoken letters A, C, H, G, J, L, X and T). So, in order to make the stimuli more representative, more words (and more speakers) needed to be used. This would significantly increase the variation in the stimuli, which might have an impact on the flanker effect.

The use of a wider variety of stimulus words presented the problem of how to align the words for simultaneous presentation. Although Murphy et al. (2013) replicated the flanker effect using asynchronous auditory flankers in a sequence of targets with an average of 96 ms stimulus onset asynchronicity (SOA) between the flanker and one of the targets, in the majority of flanker tasks the target and distractor are presented simultaneously, and Chan et al. (2005) and (Francis, 2010) aligned the target and distractor words by their acoustic onset. Given the wider variety of words used in the experiment reported here target and distractor word pairs were aligned according to their ‘perceptual centre’ (or *p-centre*).

### 2.3.1 P-Centres

The *p-centre* of a word is its “psychological moment of occurrence” (Morton, Marcus & Prankish, 1976, p. 405) and corresponds to the point in time when an acoustic stimulus is subjectively judged to occur or when two acoustic stimuli are judged to be in synchrony (Scott, 1998). This is not necessarily the same as the word’s acoustic onset (Whalen, Cooper & Fowlert, 1991) which can be seen by analysing the acoustic onsets of sequences of words spoken at regular intervals (Patel, Löfqvist and Naito, 1999, and in particular Morton et al., 1976, figure 1). Some authors maintain that *p-centres* can only be established using behavioural experiments (Villing, Repp, Ward & Timoney, 2011) where, for example, participants adjust the alignment of two

presented words until they judge them to occur simultaneously (e.g., Marcus, 1981). However, others have tried to link acoustic properties of the stimulus to the location of the p-centre (e.g., Scott, 1994; Marcus, 1981; Harsin, 1997; see also Howell, 1988), although some of these models have only been verified for digits (Villing, Ward & Timoney, 2003).

To automatically align words by their p-centres, Scott (1994) filtered the acoustic input with a seven-band gammatone filter-bank with centre frequencies spaced using the ‘equivalent-rectangular bandwidth’ (ERB) rate function (see Moore & Glasberg, 1983, figure 2), corresponding to 109, 299, 578, 997, 1638, 2651, and 4342 Hz, and bandwidths of 4 ERB. Each filter-band was full-wave rectified and high-pass filtered with a cutoff of 25 Hz. Scott found that the first point where the energy in the frequency-band centred at 578 Hz reached 50% of the maximum energy for the entire word significantly correlated with the p-centre of the word obtained using human participants.

So, the recordings of the target and distractor words were normalised to the same level and a 4th order<sup>2</sup> gammatone filter (as implemented by Brookes, 2003) was applied (using only the single critical filter-band centred at 578 Hz with a bandwidth corresponding to 4 ERB). The rest of the procedure was the same as Scott (1994), described above. Aligning words by their p-centres frequently results in words that are not aligned by the acoustic onsets (Morton et al., 1976). This is most obvious when one word starts with a stop (e.g., *back*) and another starts with a fricative (e.g., *shop*) as shown in figure 2.3. In these cases it could be argued that participants could use the relatively early distractor information to exclude one of the responses in the flanker task before the relatively late onset of the target, rather than responding to the target. In particular, this effect of ‘stimulus-onset asynchronicity’ (SOA) may increase the variance in any measure of distraction (Wyatt & Machado, 2013). The effects of SOA are discussed below in section 2.3.4.3.

### 2.3.2 Materials

The words that formed the basis for the visual and auditory stimuli are shown in table 2.2, and were the 50 highest-frequency English CVC singular common nouns consisting of three to five

---

<sup>2</sup>Scott (1994) does not appear to specify the order (i.e., the steepness of the frequency attenuation outside the critical bandwidth) of the filter-bands that she used. Previous research has suggested use of 1, 2, 4, or 8 order filters (Slaney, 1993), so the choice of 4th order filters was partially arbitrary, but motivated by practical rather than theoretical considerations: increasing the order filter order resulted in errors from the software used to generate the filter-bank.

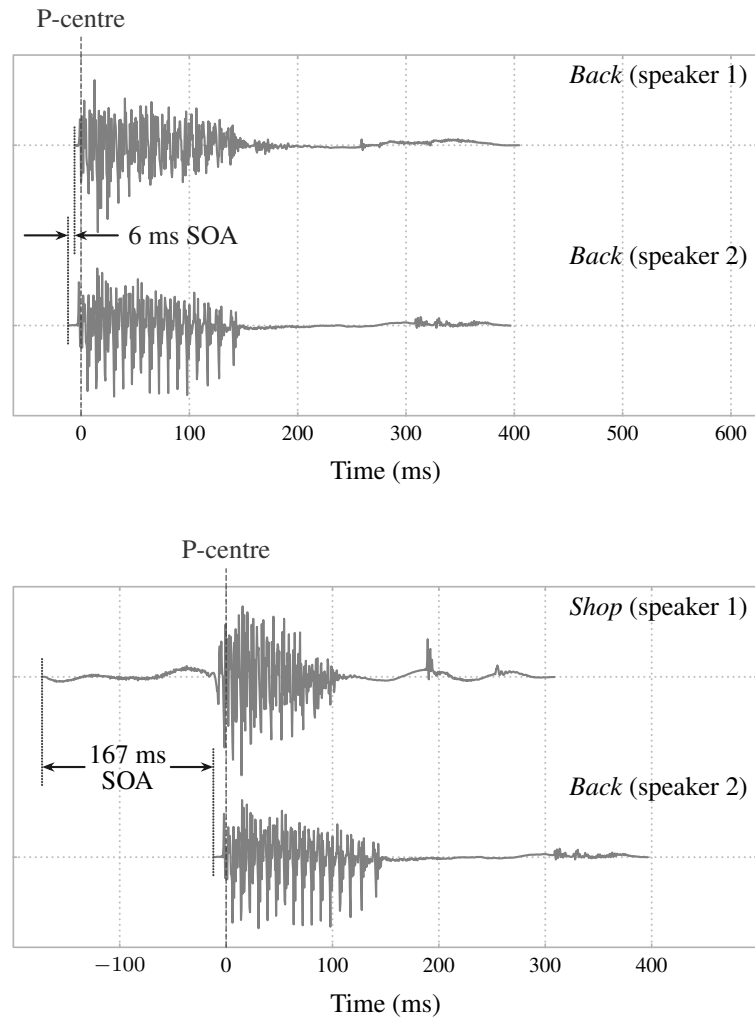


Figure 2.3: Two examples of p-centre alignment showing possible differences in the onset of p-centre aligned words. In particular, it should be noted that when one word begins with a stop /b/ and the other begins with a fricative /ʃ/, the time between the acoustic onsets of the words, or stimulus onset asynchronicity (SOA), is relatively large.

letters, selected automatically using a computer script which cross-referenced the frequency lists derived from the spoken transcriptions of the British National Corpus (Leech, Rayson & Wilson, 2011) and the phoneme transcriptions from the British English Example Pronunciation corpus (BEEP — Robinson, 1996). The words pronounced /fʌk/, /ʃɪt/, /hæk/ and /bɪtʃ/ were excluded, as sexual or taboo distractor words have been shown to alter accuracy in attention based tasks (e.g., Mathewson, Arnell & Mansfield, 2008).

The words were selected from a single syntactic category as Borowsky et al. (2013, experiment 2) found that response times to nouns were faster than reaction times to verbs, and it was

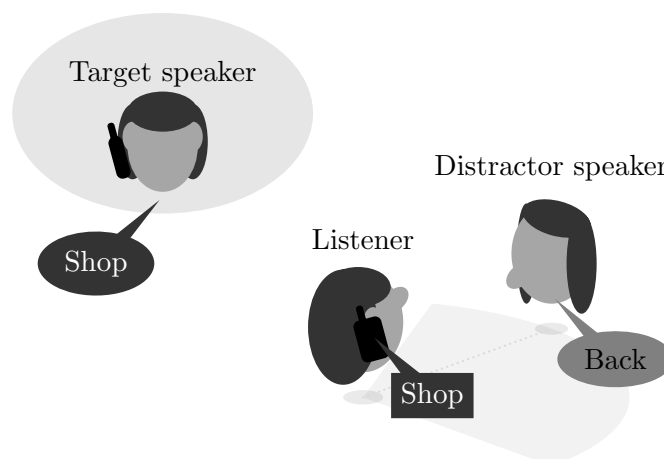


Figure 2.4: The listening environment simulation simulated over headphones in experiment Ia. The listener receives a telephone call from a friend who utters a single word ‘Shop’ (the target). At the same time a speaker in front the listener also speaks a single word ‘Back’ (the distractor).

back	court	form	head	life	man	part	shop	tape	type
bed	cup	game	hell	line	mum	phone	side	thing	week
book	dad	girl	home	look	name	piece	size	time	wife
case	dog	god	house	lord	night	road	sort	top	word
cause	food	half	job	love	page	room	south	town	work

Table 2.2: The 50 high frequency monosyllabic CVC singular noun word list. Lexical frequency was determined from the frequency lists taken from Leech, Rayson and Wilson (2011).

assumed that comparable differences may be found with other syntactic categories. Selecting words from the same syntactic category would therefore ensure that any changes in flanker performance were due to the experimental manipulation of congruent and incongruent trials rather than changes in syntactic category.

In addition, words were selected to be of similar lexical frequency because Boulenger, Hoen, Ferragne, Pellegrino and Meunier (2010) found that response times can be longer to targets with high frequency distractors compared to targets with low frequency distractors. The minimum frequency resulting from the automatic selection was 90 instances per million word

tokens which was higher than the criterion for high frequency of 45 instances per million tokens used by Boulenger et al. (2010). Selecting words with similar lexical frequencies would provide support that any changes in flanker performance were due to the experimental manipulation of congruent and incongruent trials rather than changes in lexical frequency of the stimuli.

The selection of words did not result in a list that was phonetically balanced (i.e., containing a distribution of phonemes that approximates the distribution found in a suitably large corpus of the language — Egan, 1948), and an analysis of the CVC list, showed that two consonants /ð/ and /ʒ/ and three diphthong vowels /ɛə/, /ɪə/, /ɔɪ/ and /ʊə/ were missing according to the phoneme inventory derived from the BEEP lexicon (Robinson, 1996). However, words starting or ending with the consonants /ʒ/ (e.g., *gigue*) or /ð/ (e.g., *mouth*, *tithe*) were either not nouns, common nouns or were not in the frequency lists. Furthermore, the diphthongs /ɛə/ and /ʊə/ are typically pronounced as /ɛ:/ and /ɔ:/, respectively, in modern variants of Standard Southern British English (Hughes & Trudgill, 1997). Of these, /ɔ:/ was present in the word list and there were no instances of CVC singular nouns in the BNC frequency lists with a vowel /ɛ:/. Thus, the only real omission was the diphthong /ɔɪ/ (e.g., *noise*, *voice*). However, given the intended application of the auditory flanker task was speech technology evaluation where vowel intelligibility is “normally a problem of relatively minor consequence” Voiers (1967, p. 2), it was felt that this omission was not significant.

### 2.3.2.1 Audio recordings

The 50 CVC words were recorded in sound-proof recording booths using a Røde NT-1A microphone via a Roland UA25-EX USB sound card connected to a laptop running custom software written in Python on the Xubuntu (12.10) operating system. Three male and three female monolingual Standard Southern British English speakers, aged 21 to 38 years old (mean 29.57, sd. 5.50 years) recorded the words and were instructed to use a normal conversational voice, and repeat each word three times, with a significant gap between them to avoid word boundary co-articulations; each recording was saved as an uncompressed 44100 Hz, 16 bit PCM file.

For each word, each of the three repetitions was examined for acoustic artifacts and the repetition that contained minimal or no acoustic artefacts was excised from the recording and manually trimmed to within 5 ms of its perceptual onset and 10 ms of its perceptual offset. Each word was then amplitude normalised and a 5 ms linear ramp was applied to the onset and

offset of the recording to minimise artefacts introduced excising the word from the recording (cf. Mattys & Wiget, 2011).

### 2.3.2.2 Stimulus generation

#### **Target and distractor selection**

For each trial, two words with the same number of letters and no common phonemes were selected at random to appear on the left and the right of the screen. The words were selected so that they did not share any phonemes in order to reduce the chance that participants would fail to attend to the target if the distractor was similar to the target (Dishon-Berkovits & Algom, 2000).<sup>3</sup> In addition, the words were selected so that neither word had appeared in the immediately previous trial, to reduce the effects of *negative priming* (e.g., Lavie & Fox, 2000; Driver & Tipper, 1989) where responses to targets are slower when the target has been a distractor in the previous trial. The target and distractor were selected from this pair of words: in congruent trials (CON), the target and the distractor were the same word; for incongruent trials (INC), the target and the distractor were different words.

The position of the target (i.e., the side of the screen on which the target word appeared) and the position of the distractor was counter-balanced across trials. The words recorded by four of the speakers (two male and two female) were used for the main experimental trials, and the other two speakers (one male and one female) were used to create the practice trials. Target-distractor speaker pairs were selected so that target and the distractor were never spoken by the same speaker in the same trial.<sup>4</sup>

#### **Target and distractor alignment**

Using the estimated p-centres, the target and the distractor were padded with silence at both ends so that they were both the same length, and when summed would be aligned by their respective p-centres. The target was then mixed to the left or right channel of a stereo signal to simulate a telephone speaker, and the distractor was projected to 0° azimuth and elevation using the Kemar head-related transfer functions (HRTFs, Gardner and Martin, 1995, see appendix C)

---

<sup>3</sup>Although each word having the same number of letters clearly makes them similar along at least one dimension, it was assumed that this would be irrelevant for auditory attention.

<sup>4</sup>It was not feasible to test participants on all combinations of four speakers and 50 words. For example, for congruent trials there were 600 possible combinations of words and speakers and for incongruent trials there were 3536 possible combinations of words and speakers.

to simulate a speaker from in front of the participant. The target and distractor signals were summed and prefixed with an ‘auditory fixation’ tone, consisting of a 500 ms, 500 Hz tone mixed to the same channel as the target ear. This was done to ensure that participants were attending to the correct ear in each trial.

### 2.3.3 Methods

Eight participants (four males, four female), aged 18–38 years (mean 24.70, sd. 6.49 years) were recruited to take part in the experiment from the University College London ‘Psychology subject [sic] pool’. All participants were paid 10 GBP for their participation. All participants reported being monolingual native British English speakers from birth, with no known speaking, hearing or reading disorders and with normal (or corrected to normal) vision. Participants’ hearing thresholds were tested using a Kamplex KD 29 diagnostic audiometer; the inclusion criteria for normal hearing was thresholds of 20 dB HL or better at 125, 250, 500, 750, 1000, 2000, 3000, 4000, 6000 and 8000 Hz (BSA, 2011) and all participants met these criteria.

The experiment was run on an HP desktop computer running the Arch Linux operating system (with real time Linux kernel 3.2) using a customised version of PsyToolkit (Stoet, 2010). The stimuli were presented using AKG272 mkII headphones controlled by an Asus Xonar PCI sound card. Participants sat in a sound-proof booth, approximately 50 cm in front of a 17 in (43.18 cm) 1280 × 1024 VGA Dell monitor and used a USB keyboard to make responses. Visual stimuli were presented in white on a black background. A visual fixation (a white dot) was presented in the centre of a black screen for 500 ms. The words were presented in the centre of the screen using a mono-spaced font (GNU FreeMono bold, pixel size 72). The words were positioned on the screen so that the distance between the last letter of the left word and the first letter of the right word was constant. With the participants sat approximately 50 cm from the screen, the two words subtended a visual angle of 12.72° (for five letter words) to 9.20° (for three letter words).

The format for each trial is illustrated in figure 2.5. The participants had 1000 ms to read both words, before the onset of the auditory stimulus; the words remained on the screen until the end of the trial. The time of 1000 ms was established as a reasonable time to read both words during piloting, and this was well within the suggested average reading time for reading two 3–5 letter words (cf. Legge, Mansfield & Chung, 2001; Legge & Bigelow, 2011). After

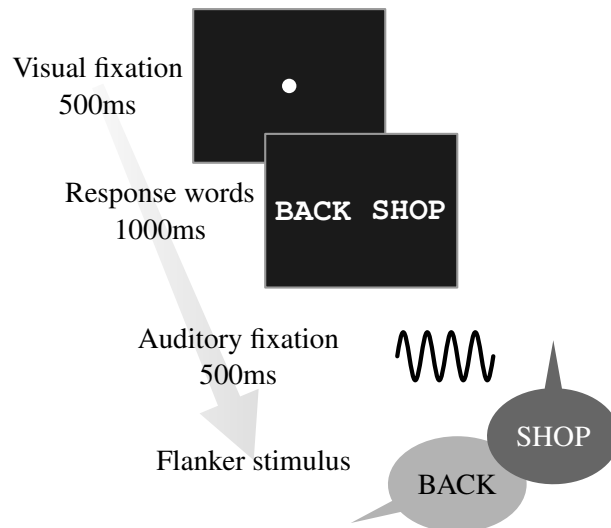


Figure 2.5: An example trial of the flanker task used in experiment Ia. As the words in the flanker stimulus are different, this represents an incongruent trial.

1000 ms, the auditory stimulus was presented. The level of auditory stimulus presentation was set so that the target was presented at 65 dB SPL (determined by a Sono OKKI CF-3502 FFT analyser connected to a Brüel & Kjær artificial ear type 4513). Participants were told they had 2000 ms from the end of the target stimulus to indicate which of the two words they had heard in the target ear using the computer keyboard;<sup>5</sup> this time of ‘2000 ms plus target offset’ was similar to value used by Francis (2010). Participants responded with their left hand (using the left-control key) if the word they heard in the target ear appeared on the left of the screen, and their right hand (using the right-control key) if the word they heard in the target ear appeared on the right of the screen. If participants responded incorrectly, a red cross appeared in the centre of the screen for 500 ms; nothing was shown for a correct response. For both correct and incorrect responses, the next trial was presented after a 250 ms delay.

Participants completed 576 trials in 6 blocks of 96 trials. Prior to the main trials, they completed a practice session consisting of 32 trials, which used the words spoken by the two speakers that were not used in the main trials. Between each block they were given the opportunity to take a short break. Within each block the target was presented to the same ear (the ‘target ear’) and the target ear alternated between blocks. Prior to each block, an instruction was

<sup>5</sup>Due to a technical error, participants actually had 2006 ms from the *onset* of the stimulus. This is discussed below in section 2.3.4.



presented on screen informing participants which ear to concentrate on (e.g., “Pay attention to words in your RIGHT ear. Ignore any other words”) which (along with the auditory fixation in the target ear) was designed to eliminate asymmetries in left-ear and right-ear responses when the target ear is unknown (Simon, 1967). Although this instruction was technically misleading as the binaural distractor was also partly presented in the target ear, the experimenter confirmed during and after the practice session that each participant understood the instruction as it related to the task. In addition, participants were encouraged not to sacrifice accuracy for speed. Target ear (left or right), target position (left or right of the screen — which determined the response hand), distractor position (left or right of the screen — determining the congruency condition), and target and distractor speaker pairs were fully crossed between trials.<sup>6</sup>

### 2.3.4 Results

All data processing and analyses were carried using R (R Core Team, 2013), PyMC (Patil, Huard & Fonnesbeck, 2010), HDDM (Wiecki et al., 2013), NumPy (van der Walt, Colbert & Varoquaux, 2011), and IPython (Pérez & Granger, 2007).

#### 2.3.4.1 Response times

Although Linux has been demonstrated (by setting appropriate system parameters) to provide superior playback latency to other operating systems given the same hardware and software (e.g., Wang, Stables & Reiss, 2010), the playback latency of the sound card was estimated at 11 ms using the procedure described in Wright, Cassidy and Zbyszyński (2004). This value was subtracted from all response time (RT) measurements to auditory stimuli in this and subsequent experiments. A preliminary examination of participants’ accuracy scores suggested that accuracy was very high across all conditions (mean 94.66%, sd. 6.55%), and participants’ accuracy scores were within three ‘median absolute deviations’ (MAD) of the median (Leys, Ley, Klein, Bernard & Licata, 2013), so no participants’ data were discarded.

RTs were measured from the onset of the target. Due to a technical error, participants in this experiment had only 2006 ms from the *onset* of the stimulus rather than 2000 ms from the *offset* of the target (as they were instructed). This meant that participants had less time (mean

---

<sup>6</sup>A legitimate issue with this experimental design is that the use of targets and responses that could be oriented to either left or right could mean that participants responses would be faster when the required response was in the same direction same as the stimulus (i.e., when the response hand and the target ear were the same). This ‘Simon effect’ (e.g., Simon, 1969) was presumed to be ‘averaged out’ by balancing the trials according to target ear and response hand.

712 ms, sd. 117 ms) to respond than was intended. The expectation was that this would lead to a large number of ‘timeouts’, where participants did not have time to respond. However, only 40 timeout errors occurred in total (less than 1% of the data), 7 in congruent trials and 33 in incongruent trials, which corresponded to 1.4% and 0.3% of the data in congruent and incongruent trials, respectively. Given that (i) these were very low numbers (and accounted for only 21% and 15% of all the errors made in congruent and incongruent trials, respectively) and (ii) this was approximately the same amount of data (1%) that Francis (2010) discarded as outliers, it was concluded that this procedural error would not lead to significant bias in the results.

#### 2.3.4.2 RT and accuracy

To demonstrate that the ‘traditional’ flanker effect based on RTs was replicated in this experiment a regression analysis was carried out but within a Bayesian framework (e.g., Kruschke, 2010a, 2010b; Feinberg and Gonzalez, 2012 see appendix A.1 for more details).

#### **Response times**

Figure 2.6 illustrates the distribution of RTs in congruent and incongruent trials and suggests that despite the considerable skew in the distributions, RTs for CON trials may be faster than RTs in INC trials, which is reinforced by the individual and group mean RTs shown in table 2.3. A Bayesian multi-level linear regression model (e.g., Kruschke, 2010a) was fit to the participants’ mean RTs obtained from trials in which a correct response was made (4362 trials, 94.66% of the data) using the Bayesian graph and parameter settings shown in section B.5.

The posterior distributions of the parameters for the RT model were estimated using the Markov Chain Monte-Carlo sampling (MCMC, e.g., Bishop, 2006; Andrieu, De Freitas, Doucet and Jordan, 2003, see section A.3 in appendix A.1 for details) with the Metropolis-Hastings algorithm implemented in PyMC (Patil et al., 2010). The MCMC sampling was run for 2994916 iterations with 4908 trials discarded as burn-in and no thinning; these numbers were determined using the Raftery-Lewis diagnostic (Raftery & Lewis, 1992). Convergence was confirmed with a visual inspection of the traces and the Geweke statistic (Geweke, 1992) which revealed no concerns. In addition, model fit was acceptable ( $MSE \ll 0.01$ ).

Figure 2.7 shows the posterior distributions for the regression parameters averaged across participants, with  $\beta_0$  the intercept (posterior mean RT in INC trials averaged across participants)

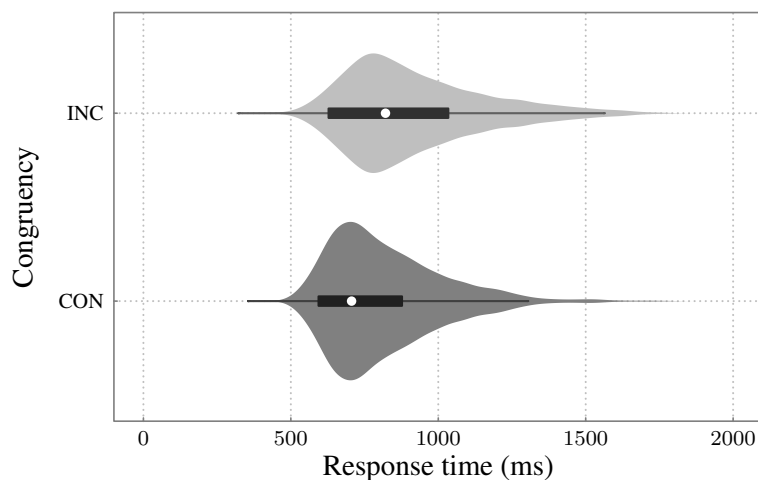


Figure 2.6: Violin plot showing response time (RT) distributions split across congruent (CON) and incongruent (INC) trials for all participants. The boxplot indicates the median, the first and third quartiles and 1.5 times the interquartile range outside the first and third quartiles (Frigge, Hoaglin & Iglewicz, 1989). Outliers are not indicated; the box-plot is augmented with a density curve derived from the observed data (Hintze & Nelson, 1998).

Participant	Response time (ms)					
	Congruency condition		CON		INC	
	mean	sd.	mean	sd.	mean	sd.
1	899	196	1052	265		
2	1021	249	1238	245		
3	632	136	737	186		
4	654	156	710	161		
5	599	110	744	224		
6	765	193	841	231		
7	582	105	697	171		
8	839	203	1042	277		
Group	749	169	883	220		

Table 2.3: Participant and group mean RTs for congruent (CON) and incongruent (INC) trials in experiment Ia.

and  $\beta_1$  the ‘slope’ (the change in posterior mean RT from INC to CON trials, averaged across participants). Although that the mode of posteriors is the preferred measure of central tendency as it represents the most credible 1% of the posterior, the means of the parameters ( $\beta_1 = 883.02$  and  $\beta_2 = -134.91$ ) are very close to the group means that can be obtained from table 2.3 (note that the  $\beta_1$  parameter represents the change in RTs between incongruent and congruent trials, and table 2.3 gives  $749 - 883 = -136$  ms). In order to establish if there is a credible difference between congruent and incongruent trials it is sufficient to demonstrate that the null value 0 is not one of the credible values of the posterior distribution for the  $\beta_1$  parameter (Kruschke, 2010a). Figure 2.7 demonstrates that the null value and its corresponding ROPE are outside the 95% HDI for the  $\beta_1$  parameter (in fact, they are so far outside the HDI they are not on the posterior plot) providing strong evidence that mean RTs for congruent trials are credibly faster than the mean RTs for incongruent trials. In addition, with no overlap between the ROPE and the 95% HDI there is considerable certainty in this difference. The posterior effect size for the  $\beta_1$  parameter is also shown in figure 2.7 (on the right) and with the null value and ROPE outside the 95% HDI the 133 ms difference between CON and INC trials is a credible moderately large effect size ( $\hat{d} = 0.48$ ).<sup>7</sup>

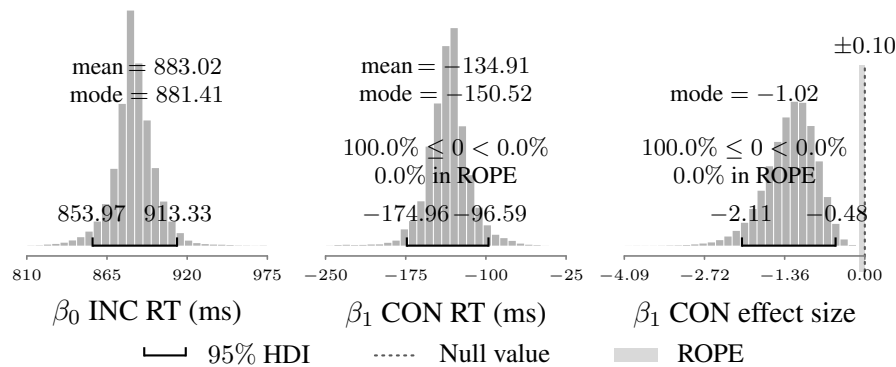


Figure 2.7: Posterior densities of the parameters for a Bayesian multilevel regression showing the intercept  $\beta_0$  for incongruent (INC) trials on the left and the ‘slope’  $\beta_1$  for congruent (CON) trials in the middle. The posterior for  $\beta_1$  represents the change in RTs from INC to CON trials and is used to test the hypothesis that RTs in CON trials are faster than RTs in INC trials. The effect size for this hypothesis test is shown on the right.

<sup>7</sup>The approximate effect size  $\hat{d}$  (analogous to Cohen’s  $d$ ) is defined here as the difference between the null value and the nearest HDI limit when the null is outside the HDI (if it is inside the HDI the difference is essentially meaningless).

Participant	Accuracy	
	Congruency condition	
	CON	INC
1	0.99	0.97
2	0.99	0.82
3	0.99	0.95
4	1.00	0.98
5	1.00	0.92
6	0.98	0.94
7	0.97	0.82
8	1.00	0.97
Group	0.99	0.92

Table 2.4: Participant and group accuracy scores for congruent (CON) and incongruent (INC) trials in experiment Ia.

### Accuracy

A Bayesian multi-level logistic regression model was fit to the participants accuracy scores shown in table 2.4 with the intercept representing mean RT in INC trials and the ‘slope’ representing the difference in mean RT between INC and CON trials. The Bayesian network and hyperprior parameters for the logistic regression are shown in section B.6. The posterior distributions of the parameters for the accuracy model were estimated using MCMC sampling with 1263747 steps and 462 samples discarded as burn-in with no thinning. Model convergence was assessed visually and with the Geweke statistic and revealed no concerns. Model fit was adequate ( $MSE < 0.01$ ).

Figure 2.8 shows the posterior distributions for the regression parameters averaged across participants, with  $\beta_0$  the intercept (posterior log-odds in INC trials averaged across participants) and  $\beta_1$  the ‘slope’ (the change in posterior log-odds from INC to CON trials, average across participants). Converting the means of the posteriors from log-odds to accuracy scores (i.e., probabilities) for each condition gives  $(1 + e^{-\beta_1})^{-1} = 0.93$  (INC trials) and  $(1 + e^{-(\beta_1 + \beta_2)})^{-1} = 0.99$  (CON trials) which can be seen to be similar to the group means that are given in table 2.4. In order to establish if there is a credible difference in accuracy between CON and INC trials it is sufficient to demonstrate that the null value 0 is not one of the most credible values of the posterior distribution for the  $\beta_1$  parameter and as 2.8 (middle

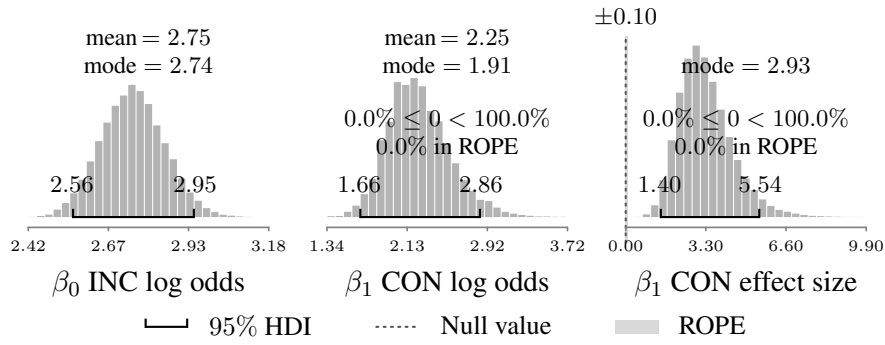


Figure 2.8: Posterior densities of the parameters for a Bayesian multilevel logistic regression for accuracy (log-odds) showing the intercept  $\beta_0$  for incongruent (INC) trials on the left and the ‘slope’  $\beta_1$  for congruent (CON) trials in the middle. The posterior for  $\beta_1$  represents the change in log-odds from INC to CON trials and is used to test the hypothesis that the log-odds of a correct response in CON trials is greater than the log-odds of a correct response in INC trials. The effect size for this hypothesis test is shown on the right.

panel) clearly shows the null value and its corresponding ROPE are outside the 95% HDI for the  $\beta_1$  parameter (in fact, they are so far outside the HDI they are not on the posterior plot) demonstrating credibly higher accuracy for CON trials than INC trials. In addition with none of the HDI overlapping with the ROPE there is considerable certainty in this difference in accuracy. The right panel of 2.8 shows the effect size posterior for the  $\beta_1$  parameter and as the null value and its corresponding ROPE are outside the 95% HDI, the difference between accuracy in CON and INC trials constitutes a credibly large effect size ( $\hat{d} = 1.40$ ).

#### 2.3.4.3 DDM parameters

The parameters of the DDM were estimated via Bayesian estimation using HDDM (Wiecki et al., 2013). Distractor congruency was set as the conditional dependency for the  $v$ ,  $t$  and  $a$  parameters and the  $b$  parameter was fixed at 0.5 to reflect the fact that participants had no *a priori* knowledge which left or right response would be correct (Merkt et al., 2013). Prior and hyperprior distributions for each parameter were set automatically according to the recommendations of Wiecki et al. (2013, p. 3), and parameters were estimated for each participant in congruent and incongruent conditions with each participants’ variance constrained at the group level. The parameters of the MCMC sampling process were determined using the Raftery-Lewis diagnostic (Raftery & Lewis, 1992). The MCMC was run for 45480 iterations and samples from

the first 45 iterations were discarded as ‘burn-in’ but with no thinning to reduce auto-correlation in the MCMC samples (as recommended by Wiecki et al., 2013).<sup>8</sup>

Convergence was assessed visually and with the Geweke statistic revealing no concerns. Model fit was adequate ( $MSE < 0.01$ ). A visual examination of the fit (formed by averaging 500 posterior samples from the fitted model and overlaying the resulting density over the observed RTs obtained in the experiments (Zhang & Rowe, 2014)) showed no serious concerns and is shown in figure 2.9.

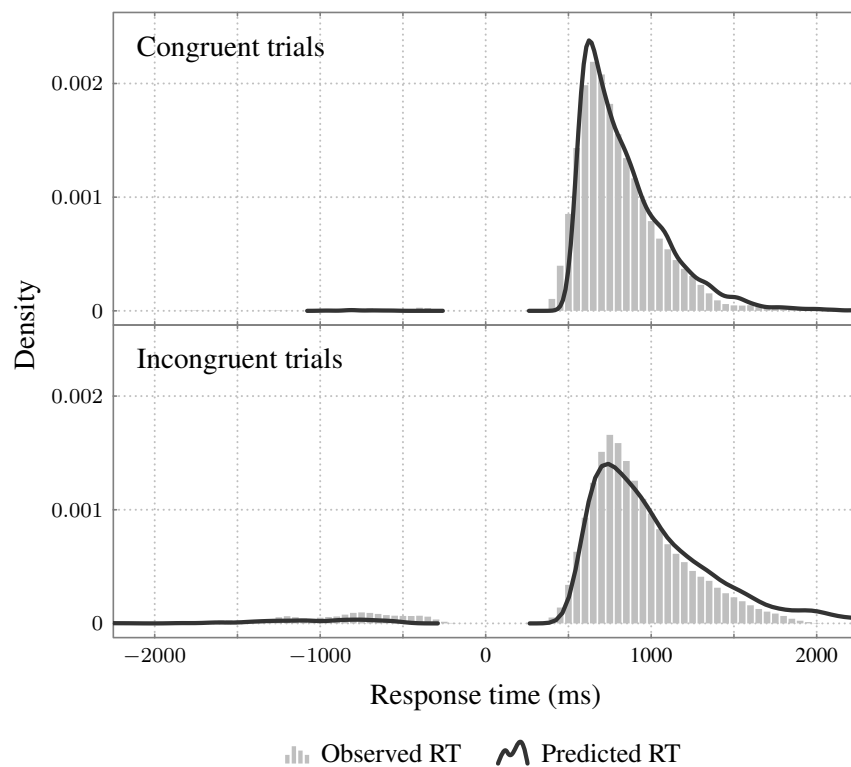


Figure 2.9: Observed and predicted RTs for congruent and incongruent trials from experiment Ia. Predicted RTs were generated using the parameters of the DDM estimated by Monte-Carlo simulation. Negative response times indicate response times for incorrect responses.

<sup>8</sup>See appendix A section A.3.1 for further information regarding the MCMC sampling process.

Participant	Drift-rate			
	Congruency condition			
	CON		INC	
	mean	sd.	mean	sd.
1	2.81	0.15	1.96	0.11
2	2.47	0.15	0.98	0.09
3	3.69	0.21	2.14	0.12
4	3.71	0.26	2.78	0.16
5	4.35	0.26	1.73	0.11
6	2.55	0.14	1.69	0.10
7	3.01	0.17	1.22	0.10
8	3.26	0.22	1.90	0.11
Group	3.22	0.81	1.80	0.68

Table 2.5: Individual and group posterior means and standard deviations (sd.) for the drift rate parameter  $v$  in congruent and incongruent trials.

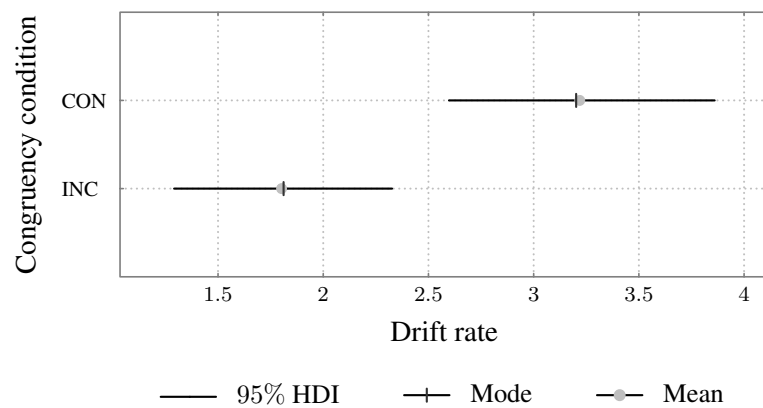


Figure 2.10: Means, modes and 95% HDIs for congruent CON and incongruent INC trials.

### Drift rate

The means and standard deviations of posterior drift rates for each participant and the overall group in congruent and incongruent trials are shown in table 2.5. Figure 2.10 illustrates the group posterior means, modes and 95% HDIs showing that credible values from 95% of the posterior for congruent (CON) samples would not be credible values for the incongruent (INC) sample, which suggests that drift-rate is credibly higher in CON trials compared to INC trials.



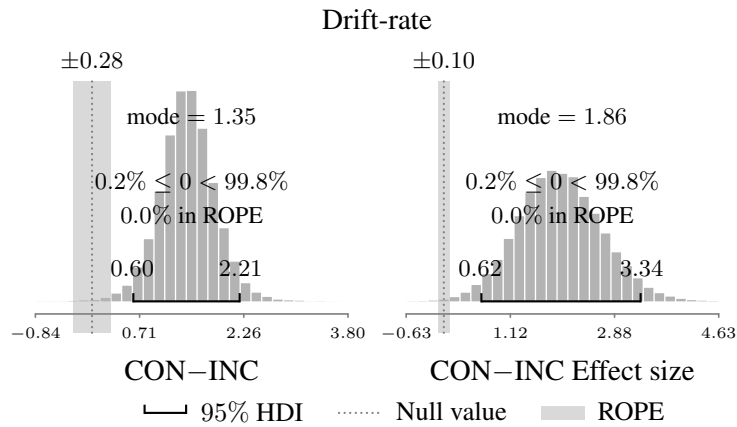


Figure 2.11: The posterior distribution for the CON–INC comparison formed by subtracting the trace for incongruent (INC) trials from the trace for congruent (CON) trials. The fact that the null value 0 (representing no difference between drift-rate in CON and INC trials) and its corresponding ROPE do not fall inside the 95% HDI provide strong evidence that the drift-rate is credibly different between congruent and incongruent trials.

This can be further illustrated by plotting the posterior density for the comparison between CON and INC trials, formed by subtracting the MCMC trace for INC trials from the MCMC trace for CON trials. Figure 2.11 (left) shows the posterior density for this CON–INC comparison, and it can be clearly seen that the CON–INC posterior is positive almost 100% of the time. More importantly, the entire ROPE containing the null value 0 is outside the 95% HDI and less than the lower limit of the HDI, providing strong evidence that drift-rate in CON trials is credibly higher than the drift-rate in INC trials. In addition, with 0% of the HDI overlapping with the ROPE, there is considerable certainty in this effect and the posterior effect size for this comparison (figure 2.11, right) constitutes a large effect size.

### Non-decision time

Table 2.6 shows the individual and group posterior means and standard deviations for non-decision times in CON and INC trials and figure 2.12 illustrates the group posterior means, modes and 95% HDIs. Although the mode is greater for incongruent trials than for congruent trials, there is greater variance in the posterior distribution which is reflected in the overlapping 95% HDIs. Although non-decision time represents both the time taken for sensory encoding before the decision process is started and time taken by the physical initiation of the response,

Participant	Non-decision time (ms)			
	Congruency condition		INC	
	CON			
	mean	sd.	mean	sd.
1	403	15	505	12
2	426	24	659	9
3	349	11	376	7
4	262	26	312	11
5	312	14	333	6
6	278	13	259	9
7	348	6	291	5
8	394	31	405	15
Group	352	31	406	56

Table 2.6: Individual and group means and standard deviations (sd.) for non-decision time in congruent (CON) and incongruent (INC) trials.

it is reasonable to assume that the time taken to initiate the response is constant for a given participant (or at least is drawn from a constant distribution). Thus, changes in non-decision time could be viewed as an indicator of changes in the amount of sensory encoding required to initiate the decision process. However, the almost completely overlapping 95% HDIs indicate that credible values for non-decision times from CON trials are also credible values from INC trials. This suggests that there is no credible difference in non-decision time between CON and INC trials.

This is further illustrated in figure 2.13, which shows the posterior density for the CON–INC comparison (left) and the posterior effect size for the comparison. It can be seen that while 80.0% of the INC distribution is greater than the CON distribution (i.e., where  $\text{CON} - \text{INC} < 0$ ) the null value 0 — representing no difference between the two distributions — and the entire ROPE is inside the 95% HDI showing that no difference in non-decision time is a credible interpretation of the evidence. In addition, with approximately 50% the HDI in the ROPE there is little certainty regarding the credibility of the difference. If non-decision time difference was sampled at random from the 95% HDI then approximately 50% of the samples would indicate a credible difference and other 50% would indicate no practical difference from the null. Thus, there is insufficient evidence to conclude with any certainty that there is a cred-

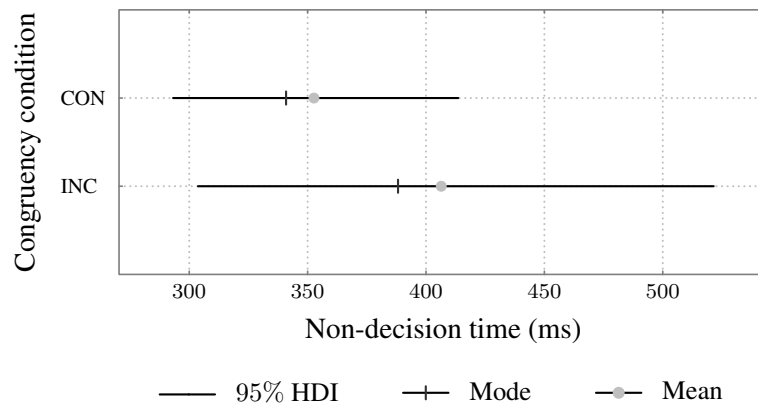


Figure 2.12: Means, modes and 95% HDIs for non-decision time in congruent (CON) and incongruent (INC) trials.

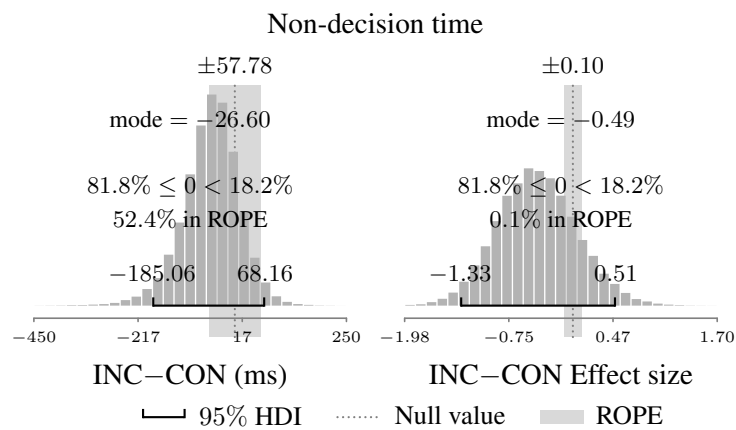


Figure 2.13: The posterior distribution for the non-decision time CON-INC comparison.

ible difference in non-decision processes between congruent and incongruent trials.

### Threshold separation

The means and standard deviations for the posterior threshold separations for each participant and averaged over all participants in CON and INC trials are shown in table 2.7, and figure 2.14 shows the group posterior means, modes and 95% HDIs for threshold separation in congruent and incongruent trials. The mode threshold separation for INC trials is less the mean threshold separation for CON trials suggesting that participants made responses on the

Participant	Threshold separation			
	Congruency condition			
	CON	INC		
	mean	sd.	mean	sd.
1	2.86	0.19	2.29	0.12
2	3.00	0.25	1.76	0.06
3	2.12	0.17	1.73	0.08
4	2.91	0.37	2.30	0.14
5	2.54	0.24	1.71	0.07
6	2.59	0.16	2.25	0.10
7	1.51	0.09	1.51	0.05
8	2.92	0.36	2.57	0.14
Group	2.53	0.71	2.02	0.48

Table 2.7: Individual and group means and standard deviations (sd.) for the posterior threshold separation in congruent (CON) and incongruent (INC) trials.

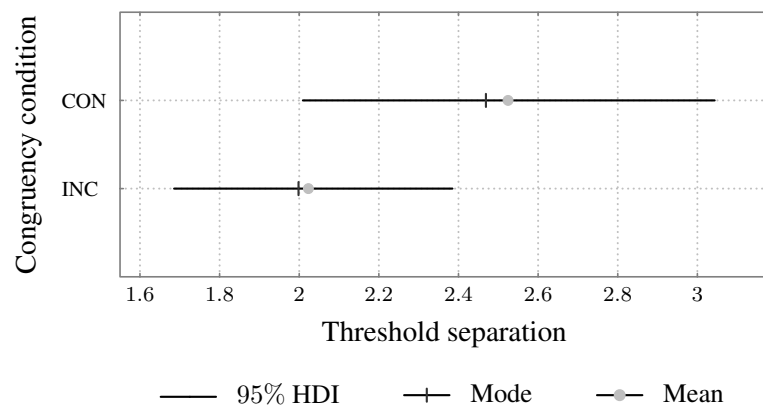


Figure 2.14: Means, modes and 95% HDIs for threshold separation in congruent (CON) and incongruent (INC) trials.

basis of less information in incongruent trials. However, the overlapping 95% HDIs indicate that credible values from 95% of the CON posterior could be credible values for the INC posterior, suggesting that any difference will not be credible.

Figure 2.15 shows the difference of the posterior density for threshold separation in CON and INC trials (left) along with the effect size for this comparison. It can be seen that 94.4% of the posterior is below the null suggesting that threshold separation is lower in CON trials compared to INC trials. However, the null value is one of the 95% most credible values, but

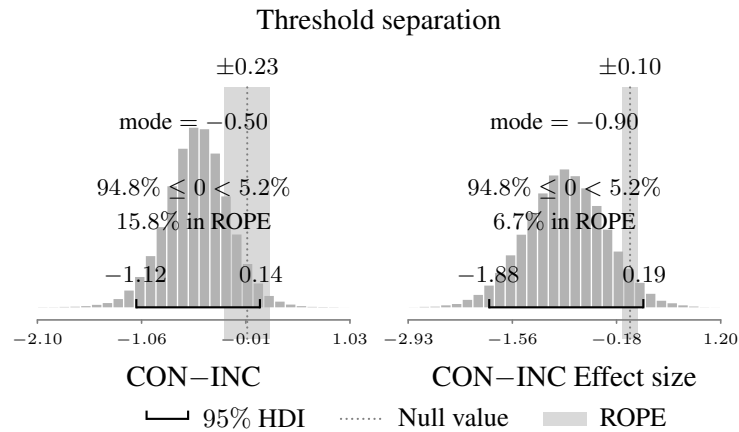


Figure 2.15: The posterior distribution for the threshold separation CON-INC comparison.

only 15% of the HDI is overlapping with the ROPE. This suggests that there is some uncertainty regarding whether this a credible difference. Around 15% of a random sample of the credible differences in threshold separation would practically indistinguishable from the null so if there is a difference in threshold-separation, it is impossible to state with absolute certainty if is credible given the current data.

### Effects of SOA on drift-interference

As mentioned in section 2.3.1, with the target and distractor words aligned by their P-centres the acoustic onsets of the targets and distractors resulting in substantial variations in SOA which are illustrated in relation to RTs in figure 2.16. The difference in SOA is most striking when one word starts with a stop (e.g., *back*) with a relatively early p-centre and another starts with a fricative (e.g., *shop*) with a relatively late p-centre as shown in figure 2.3 (p. 43). SOA was calculated by subtracting the target p-centre from the distractor p-centre. This meant that if a distractor had a relatively later p-centre compared to the target, this would result in a positive SOA and if the target had a relatively later p-centre compared to the distractor, this would result in a negative SOA.

Although SOA was not specifically manipulated but was a ‘side-effect’ of the random selection of stimulus words and speakers, and the flanker effect has been demonstrated with asynchronous auditory targets and distractors (Murphy et al., 2013), it was considered possible that the flanker effect may be modified by SOAs, as has been shown in visual flanker tasks (Wyatt

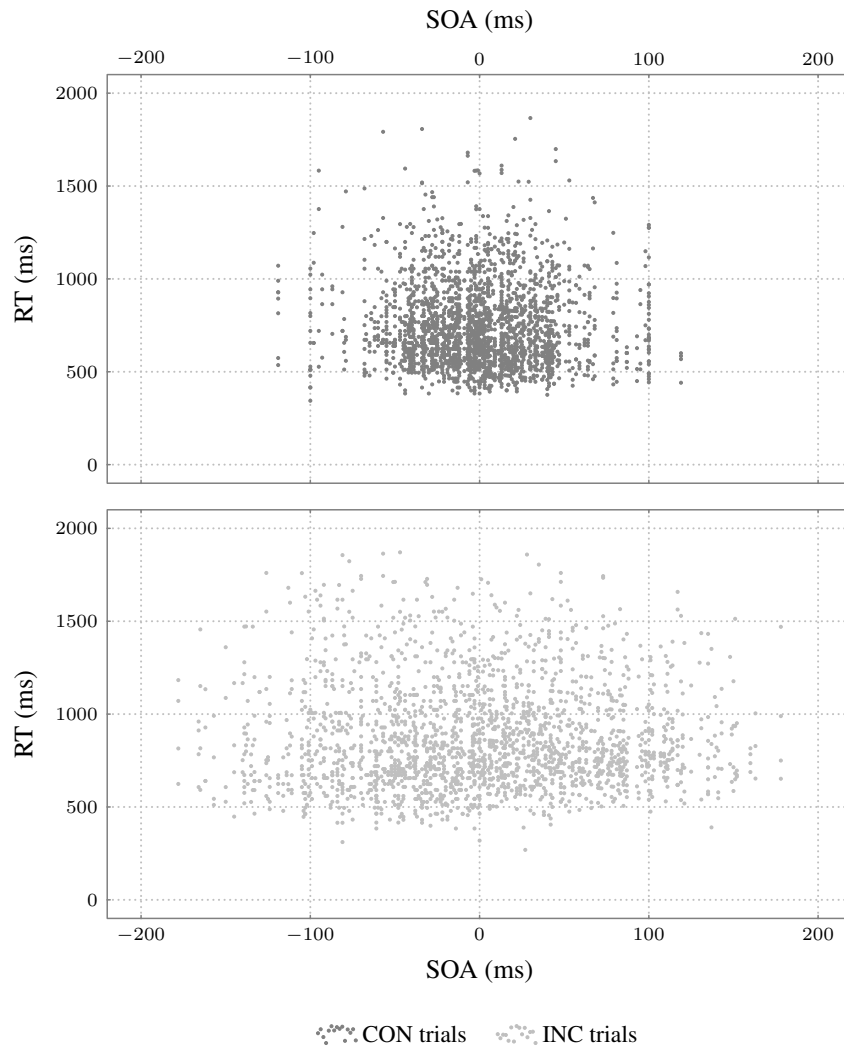


Figure 2.16: Scatter plot of response times (RT) against stimulus onset asynchronicity (SOA) for congruent (CON) trials (top) and incongruent (INC) trials (bottom).

& Machado, 2013). In addition, in an experiment where less flanker trials are run due to other experimental manipulations (e.g., target distortion or task difficulty), it was considered possible that the random selection of stimulus words and speakers could result in extreme SOAs represented unequally in CON and INC trials within these other experimental manipulations. If the flanker effect (i.e., the change in drift-interference between CON and INC trials) disappeared when there were unequal distributions of extreme SOAs within the experimental conditions then an experiment with relatively few flanker trials may report the absence of flanker effect resulting from the distribution of SOAs rather than other experimental manipulations.

The SOAs for each trial were standardised by dividing by twice the standard deviation

(Gelman, 2008) for each congruency condition separately (as the distribution of SOAs in CON and INC trials was substantially different). A DDM model was fit to the RT data with the drift-rate  $v$  for each participant dependent on congruency with the standardised SOAs as a covariate (i.e., similar to a ANCOVA) and the threshold-separation  $a$  and non-decision time  $t$  dependent on congruency. The MCMC was run for 10784 iterations with 44 samples discarded as burn-in and no thinning. Convergence was assessed visually and with the Geweke statistic which revealed no concerns. Model fit was adequate ( $MSE < 0.01$ ).

Figure 2.17 plots the regression (with the SOA back-transformed onto the millisecond scale) and figure 2.18 shows the posterior plots for the regression parameters. The modes for the intercepts  $\beta_0$  and  $\beta_1$  are the same as the group means those shown in table 2.5 (i.e.,  $INC = \beta_0 = 1.90$  and  $CON = \beta_0 + \beta_1 = 3.22$ ). The  $\beta_2$  ‘slope’ parameter is the posterior effect of SOA in INC trials showing an change in drift-rate of  $-0.01$  per 100 ms (the scale for the SOAs was back-transformed from the standardised values used in the regression before creating the posterior plots). It can be seen that that the null value 0 (representing a slope of 0, i.e., no effect of SOA) and its ROPE overlap with 100% of the 95% HDI demonstrating not only that SOA has no credible effect in INC trials but also that there is considerable certainty in this conclusion. The  $\beta_3$  slope parameter is the posterior change in SOA slope from CON to INC trials showing a small change in the drift-rate slope per 100 ms, and while the null value 0 (representing the change in the SOA slope from INC trials to CON trials) is just outside the 95% HDI, the ROPE overlaps with the HDI suggesting that this may not be a credible change in slope. However, with almost 50% of the HDI overlapping with the ROPE this conclusion is very uncertain. Figure 2.19 (right) shows the absolute SOA slope for CON trials formed by summing the  $\beta_2$  and  $\beta_3$  posteriors, and although the null is outside the 95% HDI, there is a considerable overlap between the HDI and the ROPE, showing that the effect of SOA slope may not be credible, with a reasonable amount of certainty. The effect size for this slope (Figure 2.19, left) shows that although the HDI overlaps with the ROPE, this is a very small overlap and the posterior effect size for this slope is approaching the margins of credibility for a small effect size.

Although no credible effects of SOA on drift-rate in INC trials were found, it was possible that drift rate did vary with SOA in CON trials, although it was difficult to say with much certainty if this was a credible relationship. If the interaction was credible this might

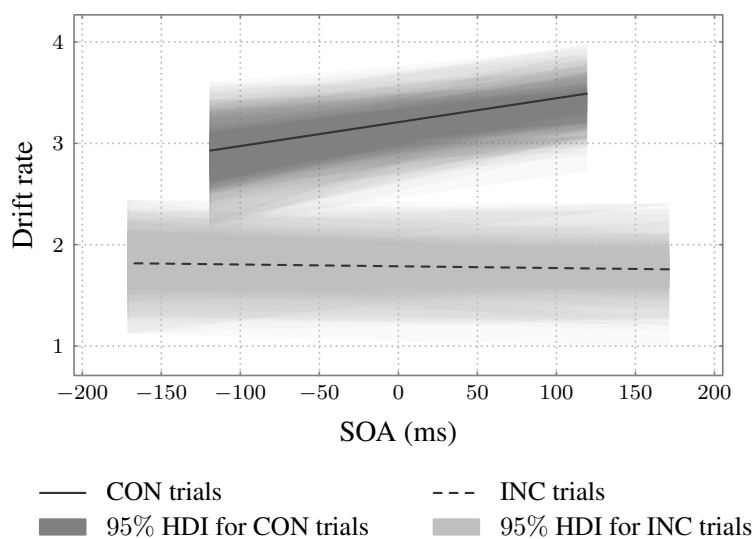


Figure 2.17: Plots of the regression parameters, with the SOAs transformed back from the standardised forms used in the regression on to a millisecond scale. The 95% HDIs were simulated using 200 samples from the posterior parameters.

affect a measure of distraction derived from the difference between drift-rates in CON and INC trials (cf. Wyatt & Machado, 2013). Although figure 2.17 appears to suggest that the drift-interference will always be present except at extremely negative SOAs, the use of the auditory flanker task with substantially reduced trials ran the risk of producing unequal and possibly extreme distributions of SOAs in other experimental conditions so it was considered necessary to demonstrate that there would be a difference between the drift-rate in CON and INC trials (drift-interference) regardless of the SOAs in these conditions resulting from the random selection of stimulus words and speakers.

The (standardised) SOAs were partitioned in five ‘bins’ to create five SOA conditions: very negative (VN), negative (NG) zero (ZO) positive (PS) and very positive (VP). These SOA conditions represented a range of SOAs from where the target onset was maximally before the distractor onset (very negative, VN) to where the distractor onset was maximally before the target onset (very positive, VP). The bin width was set automatically so that equal numbers of RTs were in each bin. A DDM model was fit to the RTs with participants’ drift-rate  $v$  dependent on the congruency  $\times$  SOA interaction and the threshold-separation  $a$  and non-decision time  $t$  dependent on congruency. The MCMC was run for 85934 iterations with 174 samples discarded



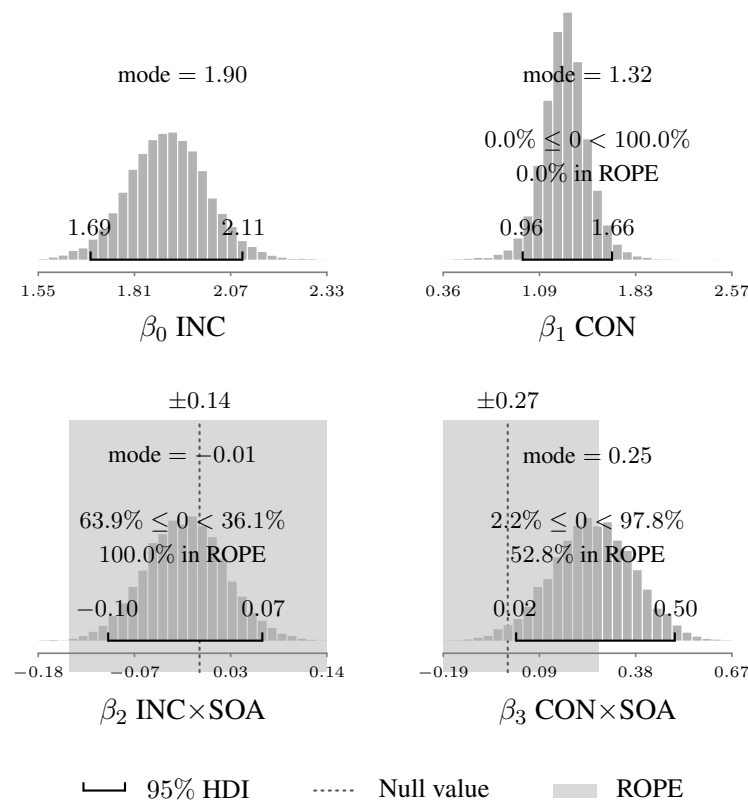


Figure 2.18: Drift-rate regression parameters for a Congruency  $\times$  SOA interaction. The  $\beta_0$  and  $\beta_1$  parameters represent the intercepts for congruent (INC) and incongruent (CON) trials, respectively. The  $\beta_2$  parameter represents the SOA slope in INC trials and the  $\beta_3$  parameter represents the change in SOA between INC and CON trials. Note, scale the x-axis represents the change in drift-rate per 100 ms SOA.

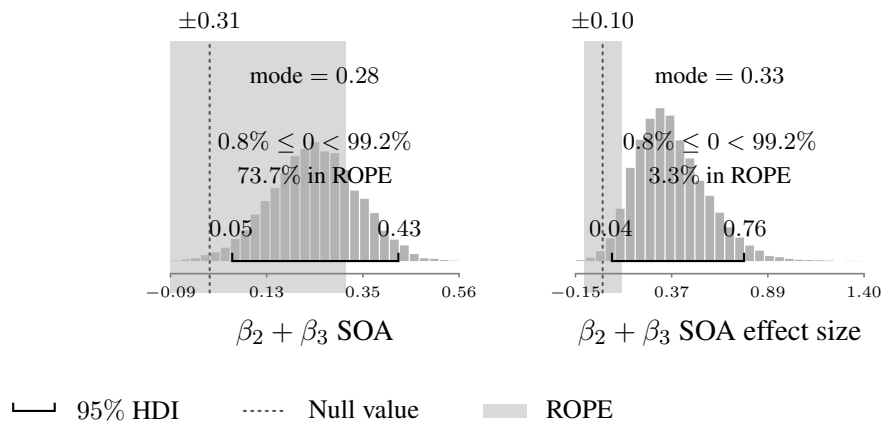


Figure 2.19: Effect size for the  $\beta_2 + \beta_3$  parameters representing the absolute SOA slope in congruent (CON) trials.

as burn-in with no thinning. Convergence was assessed visually and with the Geweke statistic which revealed no concerns. Model fit was acceptable ( $MSE < 0.02$ ).

Figure 2.20 (p. 67) shows the drift-interference resulting from all combinations of CON and INC trials and SOA conditions, revealing that drift-interference is positive (i.e., the drift-rate in CON trials is higher than the drift-rate in INC trials) for all SOA combinations and credibly different for all SOA combinations except the CON-VN–INC-VP comparison where the ROPE very slightly overlaps the 95% HDI. However, with only 0.4% of the HDI overlapping with the ROPE (which arguably, could be down to MCMC sampling error), a reasonable conclusion was that despite the interaction between SOA and congruency, that there was a credible difference between drift-rate in CON and INC trials (i.e., ‘drift-interference’) at all SOAs used in the experiment, and that there was considerable certainty in the conclusion.

### 2.3.5 Discussion

The first aim of experiment Ia was to replicate the flanker-effect (i.e., shorter RTs and greater accuracy in congruent trials compared to incongruent trials) using a novel auditory display and a wider variety of source materials for the targets and distractors than has been used in previous research (e.g., Chan et al., 2005; Francis, 2010; Murphy et al., 2013). Using a Bayesian analysis of RTs and accuracy the data provided strong evidence that flanker-effect was indeed present with RTs approximately 134 ms faster in congruent trials compared to incongruent trials and the odds of a correct response increasing from incongruent trials ( $\approx 12$ ) to congruent trials ( $\approx 109$ ).

The second aim was to narrow down the locus of the flanker-effect to specific processes that contribute to the changes in the RTs and accuracy found in flanker tasks that are taken to indicate interference from distractors. Modelling the flanker performance as a noisy decision process (the DDM) the results suggest that interference in the flanker task is accounted for by in changes in the average rate that information for target (correct) responses is accumulated during decision phase of the drift-diffusion process. Specifically, the average rate of target-information accumulation is higher for congruent trials than for incongruent trials and this particular flanker effect is consistent across all participants, as is shown in figure 2.21, although there are considerable differences in changes which presumably reflect individual differences in susceptibility to distraction (cf. Forster & Lavie, 2007).

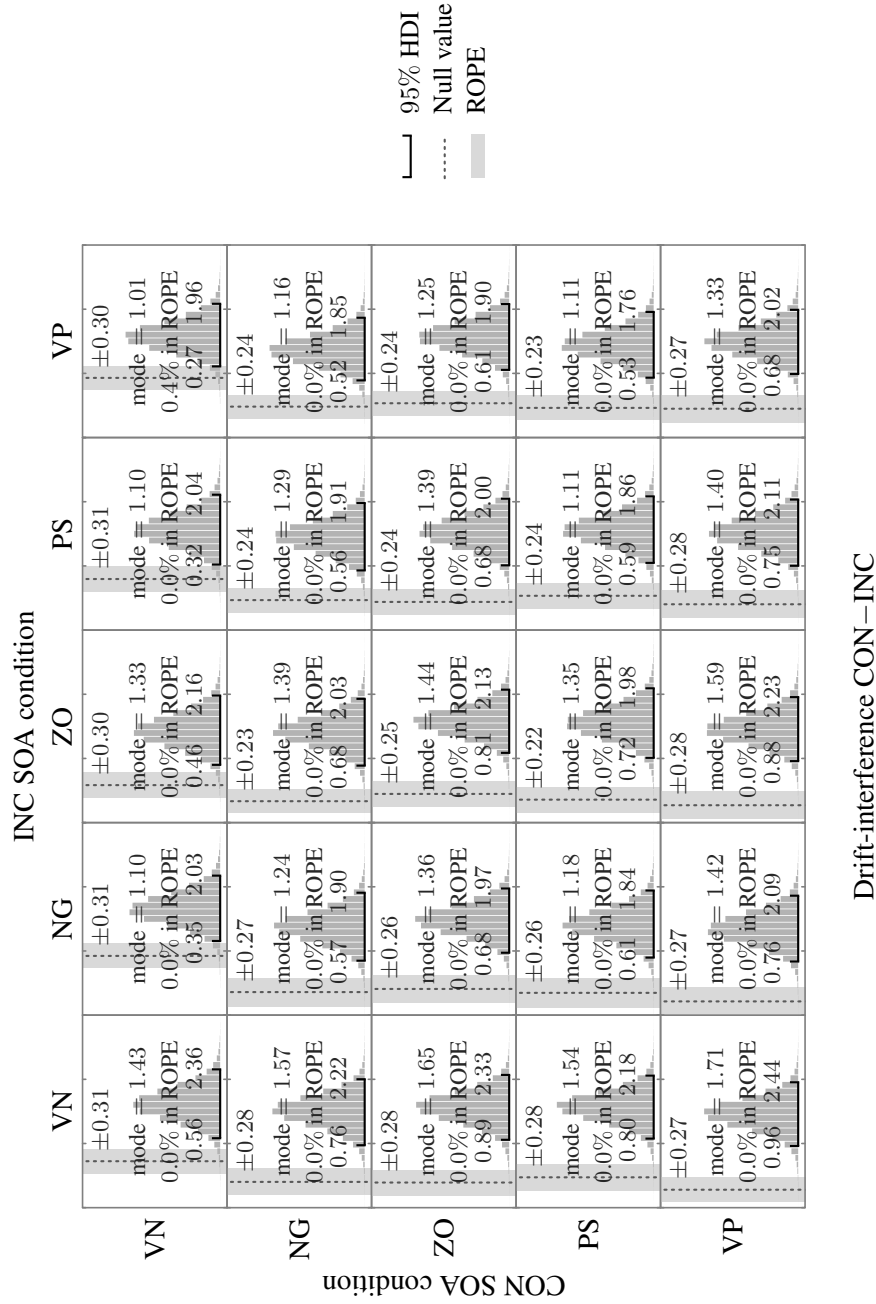


Figure 2.20: Drift-interference in all combinations of congruent (CON) and incongruent (INC) SOA conditions.

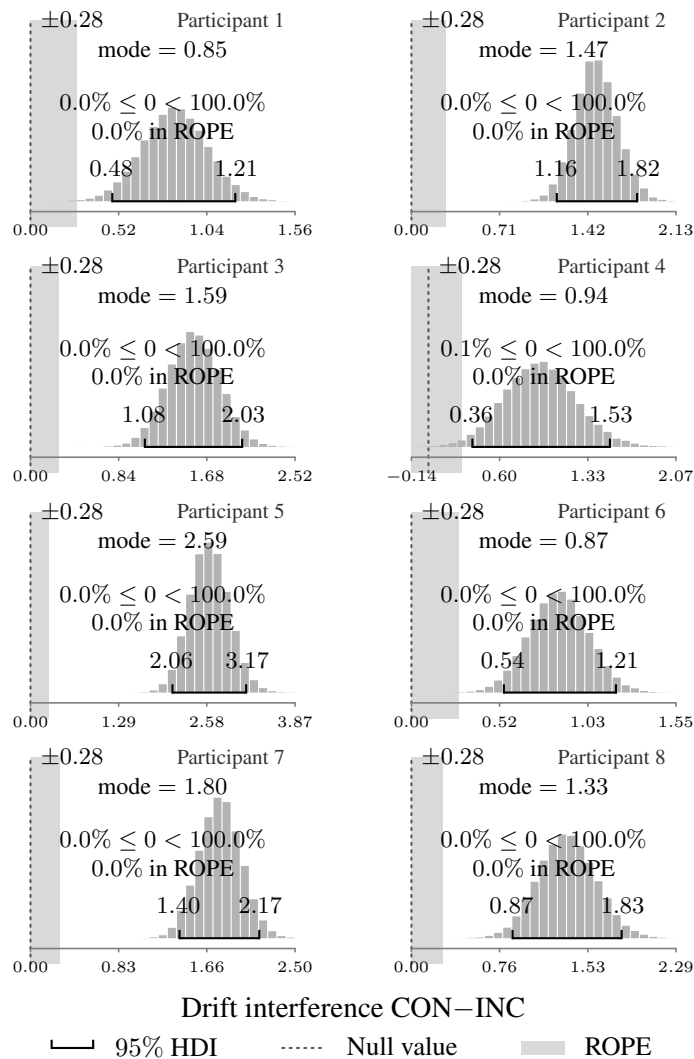


Figure 2.21: Drift-interference — the change in drift-rate between congruent (CON) and incongruent (INC) trials — for each of the eight participants in experiment Ia. All participants demonstrate the flanker effect, with the credible differences being outside (and above) the null region (the null value 0 and its corresponding ROPE).

The current data did not support credible changes in the other parameters of the DDM (non-decision time and threshold separation) between congruent and incongruent trials with the same level of certainty that was demonstrated for the change in drift-rate. Any differences in non-decision time were highly uncertain, with the posterior indicating that a credible difference might be found but only around 50% of the time. Similarly, although credible differences in threshold separation were possible between congruent and incongruent trials, given the current data, it was also possible that no difference would be a plausible interpretation of the data at

least 15% of the time. Neither of these differences were as certain as the difference in drift-rate.

The higher drift-rate in congruent trials is perhaps unsurprising given that information for the distractor response in congruent trials is the same as information for the correct (target) response, so information for the target response is accumulated much more quickly than in incongruent trials where any information accumulated for the distractor response is information for the incorrect response. But assuming that the same amount of information from the distractor response is accumulated in both congruent and incongruent trials by each participant (determined — in part — by their individual susceptibility to distraction), then the difference in the rate of information accumulation for the target response provides an indicator of the degree to which information from the distractor was processed, and hence, provides the measure of distraction. This is similar to the arguments made for changes in RT in congruent and incongruent trials in flanker tasks as an indicator of distraction (e.g., Eriksen, 1995), but this experiment has provided evidence that it is a specific process underlying the observed RTs and accuracies that results in the flanker effect.

Lachter et al. (2004) suggest that the processing of distractor information is caused either by obligatory processing of the distractor by the perceptual system even though attention is focused on the target (see Lachter et al., 2008 for examples in a visual stroop task), or by unintentional redirection of attention to the distractor, perhaps by possible lack of attentional control (Lamey, Leber & Egeth, 2012; Lavie, 2000; Driver, 2001). However, in this experiment it is not possible to rule the fact that energetic masking at the auditory periphery may have accounted for the interference effect. Although the different signal processing applied to the target and the distractor undoubtedly gave the clear impression that the target and distractor came from different (albeit virtual) locations (at least according to participants reports), it is the case that the target ear would be receiving spectral energy from both the monaural target and one channel of the binaural distractor.

Energetic masking may induce the flanker effect by masking target information so that more of the target must be processed in order to accumulate sufficient information to initiate a correct response. In this case the information from the distractor is not being processed, it is merely the absence of target information that reduces the drift-rate in incongruent trials. Alternatively, in temporo-spectral regions where the distractor masks the target, the distractor information may be processed instead of the target information resulting in the accumulation

of evidence in favour of the incorrect (distractor) response, again resulting in a reduction in the average drift-rate. Regardless of which of these two possibilities is the case (or if some combination of the two possibilities operates simultaneously), an energetic masking account of the flanker effect would be ‘less interesting’ as performance in the flanker task could be explained entirely by masking at the auditory periphery and interference could then be predicted entirely by signal properties, specifically by the degree of spectral overlap at the auditory periphery (cf. Broadbent, 1958). For the purposes of the current work, this the auditory flanker task would not be a suitable task for measuring listening effort as the concept of listening effort is predicated on the assumption that there are effects on the listener which cannot be explained by masking at the auditory periphery (e.g., Gosselin & Gagné, 2010; McGarrigle et al., 2014).

## 2.4 Experiment Ib

Experiment Ib aimed to establish if energetic masking accounted for any of the effects of drift-interference found in experiment Ia, by contrasting binaural and dichotic stimuli in the flanker task. The target ear would again be the left or right ear, but the distractor would either be projected to 0 degrees azimuth and elevation as in experiment Ia (a ‘binaural’ condition), or mixed to the opposite channel of the stereo signal as the target (a ‘dichotic’ condition)

In this way, it would be possible to contrast interference in binaural and dichotic conditions. If interference was found in the binaural condition but not in the dichotic condition then it would be reasonable to conclude interference in this auditory flanker task was caused entirely by energetic masking from the overlapping of spectral information from the projected distractor in the target ear. In terms of the DDM, if the difference in drift-rate between congruent and incongruent trials was not found in dichotic trials then this would be strong evidence for an energetic masking account of interference in the flanker task.

### 2.4.1 Materials

The source materials were the same 50 CVC singular nouns spoken by the same six speakers used in experiment Ia (see section 2.3.2). Stimuli were constructed in the same way as before except that in half the trials the distractor was projected to 0 degrees azimuth and elevation (i.e., the binaural condition, BN) and the in the other half the distractor was mixed to the left stereo channel if the target ear was the right ear or the right stereo channel if the target ear was the left

ear (i.e., the dichotic condition, DC).

Target-distractor congruency (CON and INC), distractor-type (BN, DC), target-ear, and target-position were fully crossed between trials. Target-word distractor-word, target speaker and distractor speaker selection was randomised, so the stimulus word-pairs for each participant were different.

### 2.4.2 Methods

Ten new participants (six female, four male) aged 18–34 (mean 23.90, sd. 5.58) were recruited to take part in the experiment from the University College London ‘Psychology Subject [sic] Pool’. All participants were paid 10 GBP for their participation. All participants reported being monolingual native British English speakers from birth, with no known speaking, hearing or reading disorders and with normal (or corrected to normal) vision. Participants’ hearing thresholds were tested using a Kamplex KD 29 diagnostic audiometer; the inclusion criteria was the same as experiment 1a (see section 2.3.3), and all participants met these criteria for normal hearing.

The experiment was carried out in the same way using the same software and equipment as the previous experiment (see section 2.3.3). Participants completed 576 trials in eight blocks of 72 trials. Within each block the target ear (left or right) and distractor type (binaural BN or dichotic DC) remained the same. The target ear alternated between each block, and the stimulus type alternated every two blocks; the order of target-ear  $\times$  distractor-type was counter balanced across participants. Within each block, congruency (i.e., congruent and incongruent trials) was randomised.

### 2.4.3 Results

The results from two participants were excluded as one participant was observed attempting to complete the experiment using only one hand, and the other missed a number trials due to a technical error. Trials for the remaining eight participants were excluded where a response was not given (40 trials, less than 1% of the data). The remaining RTs were in the range 200–1500 ms. Thus, no RTs were considered to be short-outliers due to ‘fast-guesses’ (Swensson, 1972) or long-outliers (Ratcliff, 1993).

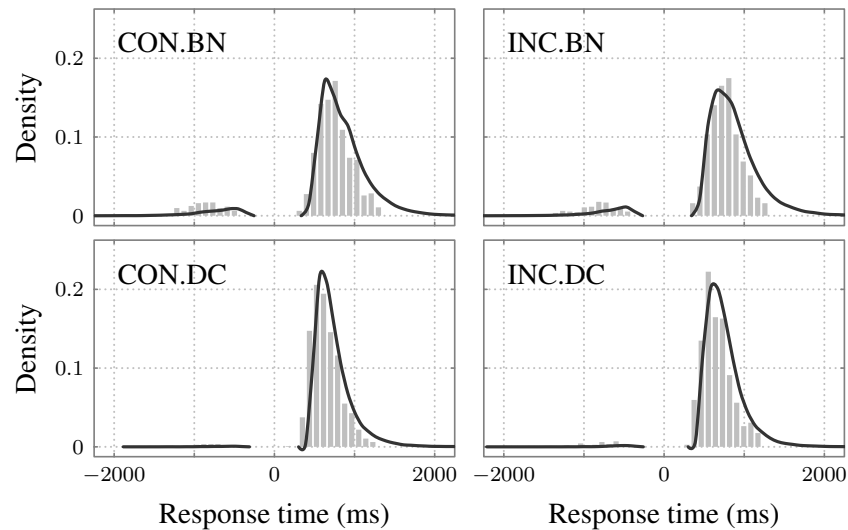


Figure 2.22: Posterior predictive RT samples generated from the DDM overlaid onto the observed RTs from experiment Ib.

#### 2.4.3.1 DDM parameters

The DDM was fitted to the RT data with the parameters  $v$ ,  $t$  and  $a$  for each participant dependent on the interaction of congruency (CON, INC) and stimulus-type (BN, DC). The MCMC chains were run for 14003 steps with 51 samples discarded as burn-in and no thinning. Convergence was assessed by visually inspecting the chains and using the Geweke statistic, which revealed no concerns. Model fit was adequate ( $MSE < 0.01$ ). Figure 2.22 illustrates the model fit for the experimental conditions revealing a reasonable fit to the data.

With only drift-rate yielding a consistent effect of congruency in experiment Ia with the highest certainty the analysis focuses only on the effects of distractor type on ‘drift-rate interference’, the difference in drift-rate between CON and INC trials which constitutes the ‘flanker-effect’ in the DDM. The posterior group means and standard deviations for drift-rate are shown in table 2.8 for each of the conditions. Figure 2.23 shows the group posterior densities for congruent and incongruent trials in the binaural and dichotic conditions. There are clear differences in the modes for congruent and incongruent trials in both the BN and DC distractor type condition. However there is a slight overlap in the 95% HDIs in the BN condition suggesting that credible values for CON trials could be credible values for INC trials in this condition. The posteriors for the comparison between CON and INC drift-rates (i.e., drift-interference) in BN and



Drift-rate							
Distractor type							
BN				DC			
Congruency		Congruency		Congruency		Congruency	
CON	INC	CON	INC	CON	INC	CON	INC
mean	sd.	mean	sd.	mean	sd.	mean	sd.
2.83	0.28	1.85	0.27	3.23	0.35	1.81	0.31

Table 2.8: Group means and standard deviations for the  $v$  parameter of the DDM model for experiment Ib

DC trials are shown in figure 2.23. In the left column (showing the CON–INC comparisons), the null region (i.e., the null and its ROPE) are outside the 95% HDI in both BN and DC conditions, and the effect sizes in the right column show credible effect sizes for both BN ( $\hat{d} = 0.19$ ) and DC ( $\hat{d} = 0.40$ ) conditions suggesting that there is a credible effect of drift-interference in both DC and BN conditions. Although the effect sizes suggest that drift-interference (i.e., the difference in drift-rate between CON and INC trials) is higher in DC trials than BN trials, figure 2.24 shows the comparison for ‘drift-interference’ between BN and DC trials and demonstrates that the null region (i.e., the null value and its ROPE) are credible values of the posterior. In particular, the mode (representing the most credible 1% values of the posterior) is inside the ROPE and almost 30% of the HDI is overlapping with the ROPE which was felt indicated sufficient uncertainty in the posterior to reject a credible difference in drift-interference between BN and DC conditions.

#### 2.4.4 Discussion

The aim of this experiment was to establish whether interference from a distracting auditory flanker could be explained entirely by energetic masking. This was considered important, because if interference from the distractor was entirely due to energetic masking at the auditory periphery, then interference in the auditory flanker task used in experiment Ia could be predicted by examining signal properties, perhaps by examining the signal-to-noise ratio in different frequency bands. Furthermore, the auditory flanker task would then not be appropriate for investigating listening effort, which assumes there are measurable effects of challenging listening that are not predictable from energetic masking (e.g., McGarrigle et al., 2014).

By contrasting interference in a binaural condition where the distractor was projected using

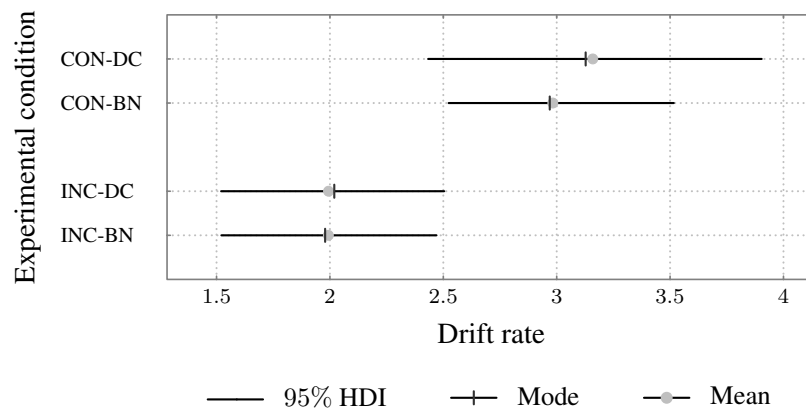


Figure 2.23: Means, modes and 95% HDIs for the drift-rate parameter in congruent (CON) and incongruent trials (INC) with binaural (BN) and dichotic (DC) distractors.

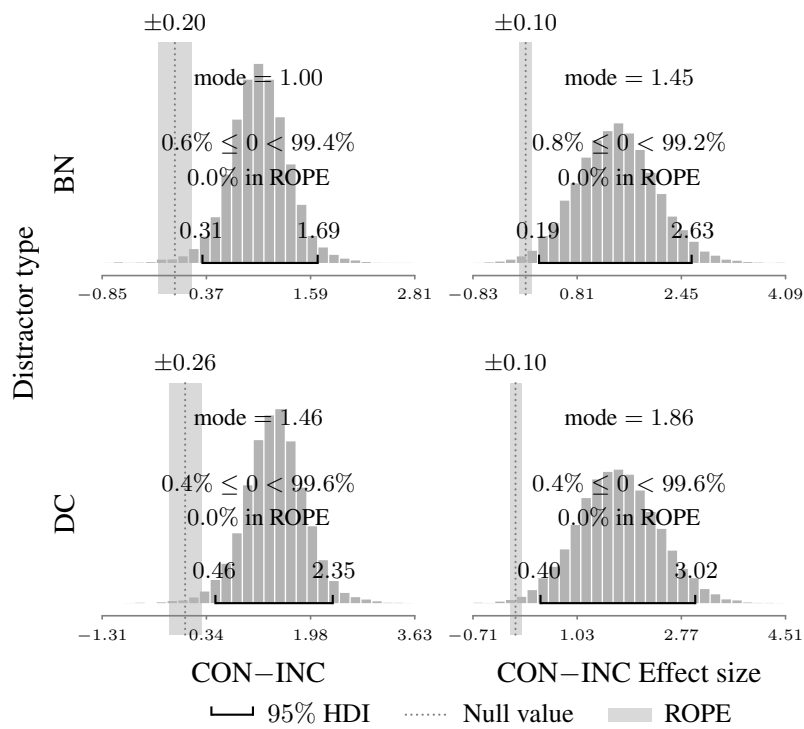


Figure 2.24: The posterior densities (left column) and effect sizes (right column) for ‘drift-interference’: the difference in drift-rate between congruent (CON) and incongruent (INC) trials with binaural (BN) and dichotic (DC) distractors.

an HRTF (i.e., some energetic masking) and a dichotic condition where the distractor was mixed to the opposite ear to the target (i.e., no energetic masking), no credible difference in the flanker

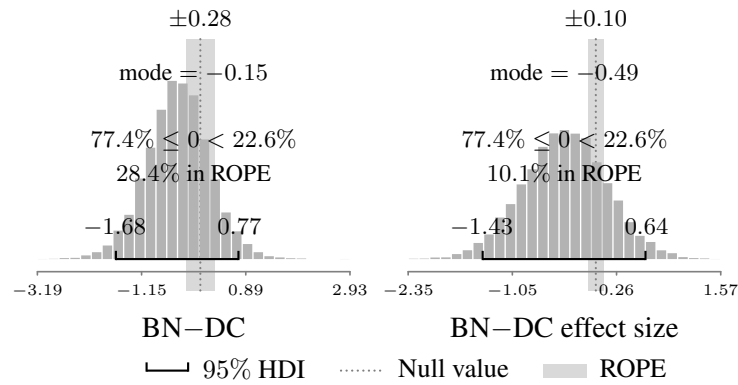


Figure 2.25: The posterior density for the comparison between drift-interference with binaural (BN) distractors and dichotic (DC) distractors.

effect was found between the binaural and dichotic conditions. This strongly suggests that the distractor interference effect in the auditory flanker task used in experiment Ia is *not* due to energetic masking at the auditory periphery and must be due to more ‘central’ or ‘high-level’ (Shamma, 2008, p. 1141) processes, possibly at the response selection stage (Eriksen, 1995).

## 2.5 General discussion

The measurement of listening effort requires an auditory task that provides a measure that is susceptible to changes in the psychological factors that are suggested to underlie listening effort. Mechanisms which control the focus and maintenance of attention are factors that have been implicated in listening effort, often under the term ‘working memory’ (e.g., Gosselin and Gagné, 2010; Larsby et al., 2005; Sarampalis et al., 2009) a concept which is also associated with susceptibility to distraction (Cowan et al., 2005; Whitney, Arnett, Driver and Budd, 2001, although see Macken, Phelps and Jones, 2009). So, auditory tasks that are sensitive to changes in attentional control and resistance to distraction could be used to measure listening effort.

### The flanker task and the drift diffusion model

In this chapter, an auditory flanker task was proposed as the auditory task to use to investigate listening effort. In a typical flanker task, performance is contrasted in trials where attention to targets is modulated by the presence of distractors which are either associated with the same response as the target (congruent) or with a different response (incongruent). The ‘flanker effect’ is shown by faster RTs and less errors in congruent trials compared to incongruent trials

and is an indicator of ‘interference’: the extent to which a distractor is processed (Lavie, 2010). In both the visual and the auditory domain, the flanker effect has been shown to be susceptible to increases in cognitive load (by increasing task complexity to influence the control of attention) and also to perceptual load (relating to stimulus complexity) (e.g., Lavie & De Fockert, 2005; Francis, 2010). Thus, if increases in listening effort correspond to increases in cognitive load or perceptual load the degree of interference from distractors in the flanker task may prove to be a useful measure of listening effort.

Experiment Ia replicated the flanker effect by analysing response time and response accuracy separately. In addition, the responses were modelled as a noisy decision processes which accumulates information for one of two responses over time, using the drift-diffusion model (DDM — e.g., Voss et al., 2013), which can model RT and accuracy simultaneously. One specific parameter of this model, the drift-rate parameter  $v$  was shown to account for performance in the flanker task. The drift-rate represented the average rate of target-information accumulation for a correct response with large positive drift-rates representing faster correct responses, small positive drift-rates representing slow correct responses, large negative drift-rates representing fast incorrect responses, and small negative drift-rates representing slow incorrect response.

The average drift-rate was shown to be higher (i.e., more positive) in congruent trials compared to incongruent trials, and represented the ‘flanker effect’ in the DDM. In congruent trials, any processing of the distractor would lead to higher drift-rates because information from the distractor was also information for the target response so the average information accumulation rate (i.e., drift-rate) for the target response would be relatively high. In incongruent trials, any processing of the distractor would lead to lower drift-rates as information from the distractor was information for the distractor (associated with the incorrect response) so the average information accumulation rate for the target response would be relatively low. Assuming participants processed distractor information to the same degree on congruent trials as well as incongruent trials, then the difference in drift-rate between the congruent trial and incongruent trials represented a measure of interference from distractors or ‘drift-interference’.

### **Effects of stimulus onset asynchronicity**

In order to use an auditory flanker task to assess listening effort it was considered important to enlarge the inventory of speech materials that were used to construct the stimuli, as traditional listening tests such as intelligibility and speech quality use large inventories of speech materials. Previous speech-based auditory flanker tasks have used only a small number of words and speakers: two words, two speakers (Francis, 2010), four words and four speakers (Chan et al., 2005) and eight words and one speaker (Murphy et al., 2013). The auditory flanker task presented in experiment 1a used six speakers and fifty words selected at random for each trial (although under some restrictions — see section 2.3.2.2). Although each participant did not hear all combinations of speakers and words (which with no restrictions on the selection of words and speakers would have required over 70000 incongruent trials), this represents a substantial extension to the auditory flanker tasks with speech materials.

Flanker tasks typically involve a simultaneous presentation of targets and distractors (e.g., Eriksen & Eriksen, 1974), which is less problematic in the visual domain compared to the auditory domain as visual objects are ‘distributed’ in space across the retina whereas auditory objects are distributed in time across the cochlea filters (cf. Chan et al., 2005). Although it is possible to present targets and distractors sequentially and still observe the flanker effect (Murphy et al., 2013), this chapter proposed using the ‘perceptual-centre’ (or p-centre — Morton et al., 1976; Fowler, 1979; Scott, 1998) as the means of aligning stimuli, where the p-centre for each word is the first point at which energy in a critical frequency band reached half the maximum energy for that band (Scott, 1994). Although alignment by p-centres meant that targets and distractors were perceived as occurring simultaneously, the variation in the pronunciation of the words by different speakers and the alignment of words with significantly different p-centres resulted in a considerable amount of variability in the onset of the targets and distractors (i.e., stimulus onset asynchronicity — SOA, see figure 2.3). Subsequent analysis, however showed that despite this variability, even extreme SOAs failed to completely eliminate drift-interference.

### **Effects of energetic masking**

The use of a monaural target and binaural distractor meant that there could be some spectral overlap in the target ear. This was considered to be more problematic in incongruent trials as despite the use of different speakers for target and distractor words in congruent trials there

would be less conflicting spectral information (e.g., if the target and the distractor word were both *shop*, the acoustic realisations of the target and the distractor would begin with begin with random noise at relatively high frequencies). If spectral (i.e., energetic) masking at the auditory periphery was the cause of slower RTs in incongruent trials, then not only would the explanation of the flanker effect be quite straightforward and could be predicted by signal properties, but the auditory flanker task would be less useful for measuring listening effort, which is related to the effects of challenging listening beyond the auditory periphery (McGarrigle et al., 2014). Experiment Ib addressed this issue by contrasting interference effects with binaural distractors (some energetic masking) with dichotic distractors (no energetic masking), and showed that there was no significant difference in interference for trials with binaural or dichotic distractors. This strongly suggested that there the flanker effect measure in experiment Ia was not due to energetic masking.

### **Summary**

In summary, an auditory flanker task has been introduced which demonstrates the flanker effect with a considerable variety in auditory stimuli and simulating ‘every-day’ challenging listening situation: listening to speech on the telephone while ignoring other, distracting speech. It is now possible to to apply the auditory flanker task to a more challenging listening-situation and examine how interference from distractors changes when the target is made harder to perceive by adding noise, and whether any changes are reversed by speech technologies which are designed to reduce noise.

## Chapter 3

# The auditory flanker task with noisy and ‘de-noised’ targets

### 3.1 Selective attention and noise

In the previous chapter, an imaginary listening situation was presented to participants, in which they were asked to imagine that they were in a quiet room and they received a telephone call from a friend who was also in a quiet room. As their friend spoke, somebody in front of the participant also spoke, and participants had to pay attention to their friend's speech (the target) while ignoring the speech from in front of them (the distractor). For the experiments in this chapter, the listening situation was made a little more complex. This time, the caller was in a noisy café, so the caller's speech was mixed with background noise. Sometimes the caller was using a modern handset equipped with different forms of ‘digital noise reduction’ (DNR) software to try to reduce the background noise, at other times the caller was using an old handset with no noise reduction at all. Nevertheless, the task remained the same and is illustrated in figure 3.1: as the caller spoke, somebody in front of the participant also spoke, and participants were required to pay attention to the caller's speech (the target) while ignoring the speech from in front of them (the distractor) and respond appropriately. So, the focus of the current chapter is whether attending to noisy targets affects the ability to resist distraction and whether processing the noisy targets with a technology designed to reduce noise will change the effects.

Listening to speech degraded with noise requires attention processes which are not required for clear speech (Wild et al., 2012) and given that almost all spoken communication

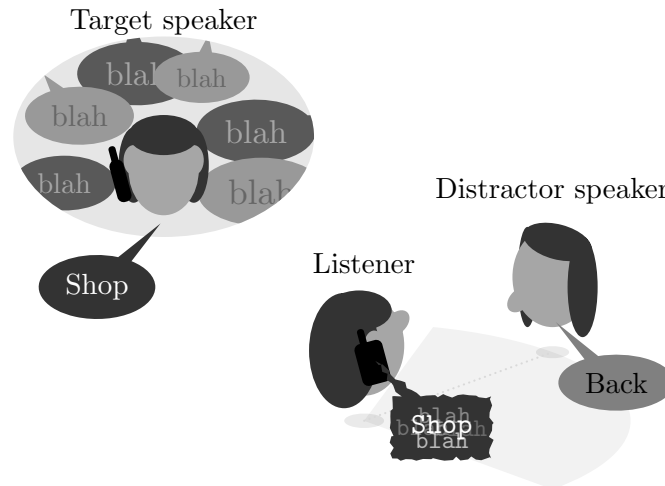


Figure 3.1: The imaginary listening situation used in the experiments in this and the next chapter. Participants must attend to a noisy monaural target (e.g., ‘shop’) while ignoring a binaural distractor (e.g., ‘back’).

takes place against a background of some level of noise (Baldwin, 2012), it follows that attention is requisite for listening to speech in any realistic environment (Shinn-Cunningham, 2008; Shamma, 2008). Furthermore, if increasing the background noise increases the attentional requirements for processing speech then increasing the background noise in the flanker targets should increase the attentional resources required to process targets in the flanker task.

If these attentional processes are related to capacity limits assumed to exist in the perceptual system responsible for the perceptual organisation of incoming sensory information (Lavie, 1995, 2000), then increased background noise would constitute an increased ‘perceptual load’, and should result in a decrease in distraction, as the capacity of the perceptual system would be depleted by processing the noisy target to the extent that there was no ‘spare’ capacity left to process the distractor (Francis, 2010; Murphy et al., 2013). If, on the other hand, the attentional processes were related to capacity limits assumed to exist in the cognitive system responsible for control of attentional focus, then the increased background noise would constitute an increased ‘cognitive’ load and would result in an increase in distraction, as the capacity of the cognitive system would be depleted by processing the noisy target to the extent that there was no ‘spare’ capacity to control the inhibition of the distractor (Dalton et al., 2009; Francis, 2010)

However, there is very little research on whether noisy auditory targets constitute a perceptual or cognitive load in auditory attention. Although Mattys et al. (2009) suggest that stimuli



degraded with noise can constitute a perceptual load, Lavie and De Fockert (2003) specifically argue that degraded stimuli do not constitute perceptual load. By contrasting the concept of sensory limits (i.e., degraded stimuli) with capacity limits (perceptual load) they argue that degraded stimuli, whilst increasing task difficulty, do not constitute a perceptual load, merely requiring more time to process and so providing more opportunities for interference from distractors (see Norman and Bobrow, 1975 for an in-depth discussion on the contrast between sensory and capacity limits). However, the experiments reported in Lavie and De Fockert (see also Yeshurun and Marciano, 2013) involved visual attention, and the ‘degraded’ targets were merely presented with less intensity (i.e., in dark grey rather than white) which is arguably analogous to reducing the volume of an auditory target rather than degradation by the addition of background noise. Furthermore, when visual stimuli are degraded by adding noise, distraction is reduced (Hughes et al., 2012).

It isn’t clear if the concept of perceptual load has any relation to the concept of listening effort, as the effects of listening effort are generally assumed to have a cognitive basis, so are presumably associated with cognitive load. But with the very broad definitions of listening effort (see section 1.3) including almost any aspect of attention, it is conceivable that the perceptual load might also come under the rubric of listening effort. Nevertheless, regardless of whether background noise in the flanker target induces a perceptual load or a cognitive load, then processing the noisy targets with an ideal DNR system should relieve the burden on attentional resources so that performance in the attentional system affected should return to a level that would be found with less noisy targets. If noisy targets constitute a perceptual load, and increased perceptual load results in decreases in interference from distractors, then DNR processed targets should result in a relative reduction in perceptual load (compared to unprocessed targets), leading to an increase in interference. Conversely, if the noisy targets constitute a cognitive load, and increased cognitive load increases interference then DNR processed targets should result in a relative reduction in cognitive load, leading to a decrease in interference.

### 3.1.1 Digital noise reduction

Digital noise reduction (DNR) also referred to as ‘speech enhancement’ aims to reduce background noise without distorting the target speech (Loizou, 2007; Bentler & Chiou, 2006) and DNR is a common component in mobile phones (Goulding & Bird, 1990) and modern hear-

ing aids (Brons et al., 2012; Levitt, 2001). The DNR algorithms used below are representative of three classes of noise reduction algorithms: spectral-subtraction (SB), statistics-based (MM after the specific algorithm, MMSE Ephraim and Malah, 1985), and sub-space (SP) (Hu & Loizou, 2007; Loizou, 2007). All three algorithms (i) break the noisy speech signal into small overlapping segments, (ii) analyse the signal in each segment into spectral components, (iii) estimate which components are related to speech and which are related to noise (iv) remove or attenuate components which are assumed to be related to the noise, and (v) recombine the components and stitch the segments back together to create the noise-reduced speech. The principle differences in the algorithms involve the assumptions about the relationship between the speech and the noise, and the transformations applied to the spectral components to enable them to be classified as speech or noise.

The specifics of each algorithm are given in Loizou (2007) (see also Benesty et al., 2005), and are not dealt with here as the focus of the work is not about the specifics of DNR implementations *per se* but whether DNR systems can be evaluated by changes in (measures of) listening effort. The algorithms used in the experiments below were based on the MATLAB code provided by Loizou (2007). Each algorithm used a ‘voice activity detector’ (VAD) which updated the estimation of the noise spectrum during periods in which speech is assumed not to be occurring, although a different VAD was used for the sub-space DNR which was appropriate for the type of algorithm (see Hu & Loizou, 2007, p. 591).

## 3.2 Experiment II

The aim of experiment II was to take a broad overview of the effects of listening to noisy and DNR-processed speech on interference from distracting speech. The auditory flanker task established in the previous chapter was used again, but with the speech of the targets mixed with background noise at different SNRs and processed with or without three kinds of DNR algorithm. In particular, the aim was to note any changes in drift-interference when listening to increasingly noisy targets, as this determined the criteria for what constitutes a ‘good’ DNR. If listening to speech at lower SNRs changes the level of drift-interference (compared to higher SNRs) and a (possibly ideal) DNR is supposed to ‘undo’ the effects of background noise, then a good DNR system should be one in which changes the level of drift-interference to a level similar to that found when listening to unprocessed noisy speech at a higher SNR. Specifically,

if attending to higher SNR targets *increases* interference from distractors (relative to lower SNR targets) then attending to lower SNR targets processed with DNR should *decrease* interference. Conversely, if listening to higher SNR targets *decreases* interference (relative to lower SNR targets), then attending to lower SNR targets processed with DNR should *increase* interference.

In terms of the DDM, increasing the noise in the targets would increase the energetic masking in the target, resulting in reduced availability of target information, so it is possible that more initial sensory processing would be required to initiate the decision process (i.e., longer non-decision times) and decisions would have to be made using less information (i.e., lower threshold separation). Reduced availability of target information would (everything else begin equal) result in target-information being accumulated at a lower rate in the decision process as more of the target may need to be processed to obtain sufficient information to initiate a response. This would be shown by reduced drift-rates. However, if attending to the noisy targets utilises attentional resources (Wild et al., 2012) and those attentional resources are located in the limited capacity perceptual system, then there will be less spare capacity to process the distractor, and less distractor information being accumulated in the decision process, resulting in relatively increased drift-rates compared to conditions where the targets are less noisy. Conversely, if the attending to the noisy speech utilises attentional resources that are located in the limited capacity cognitive system, then there will less spare capacity to control attention and inhibit the processing of the distractor, and more distractor information being accumulated in the decision process, resulting in relatively reduced drift-rates compared to conditions where the targets are less noisy.

So, if attending to noisy targets constitutes a perceptual load, then at lower SNRs there will be higher perceptual load, less interference from distractors, and higher drift-rates. At higher SNRs there will be lower perceptual load, more interference from distractors, and lower drift-rates. If, however, attending to noisy targets constitutes a cognitive load, then at lower SNRs there will be higher cognitive load, more interference from distractors, and lower drift-rates. At higher SNRs there will be lower cognitive load, less interference from distractors (more control over inhibition), and higher drift-rates.

### 3.2.1 Materials

The recordings used to construct the stimuli were the same 50 words and six speakers used in the experiments described in the previous chapter (see section 2.3.2). As before, two words and two speakers were selected at random, subject to the same constraints used in the experiments reported in chapter 2 (both words and speakers were different, and the words had same orthographic length, no phonemes in common, and neither word appeared in the previous trial). The target and distractor were selected from these two words and the distractor projected to 0 degrees azimuth and elevation using the KEMAR HRTFs (see appendix C).

The background noise consisted of babble noise from the NOISEX-90 database (Noisex, 1990) which consisted of a 235 s recording of 100 people talking in a canteen. The babble noise was up-sampled from the original 19.98 kHz to 44.1 kHz and root-mean square (RMS) smoothed using successive 16, 8, 4, 2, and 1 s sliding windows, to ensure that the amplitude was consistent throughout the whole recording. The target was padded at both ends with 250 ms of silence and a random segment of the same duration as the padded target was extracted from the babble noise and scaled to the appropriate level to obtain the required SNR to create three SNR conditions (LO  $-4$  dB, MD, 0 dB and HG  $+4$  dB). The scaling factor for the noise was calculated using the unpadded target and the segment of the babble noise that aligned with the onset and offset of the unpadded target. The noisy target was either left unprocessed (NN DNR condition) or with one three DNR algorithms implemented in the PYTHON programming language (based on the original MATLAB code from Loizou, 2007): MMSE (MM), Sub-space (SP), or Spectral-subtraction (SB). The target was then mixed to the left or right channel of a stereo signal and summed with the distractor. Finally, an ‘audio-fixation’ consisting of a 500 ms, 500 Hz tone mixed to the same stereo channel as the target ear was prepended to the stimulus.

Twelve audio conditions were created in this way with SNR crossed with DNR. A thirteenth audio condition was created as a control condition with no noise (actually  $+60$  dB SNR) and no DNR (CL-NN). Target ear, target position, distractor congruency and audio condition were fully crossed between trials.

### 3.2.2 Methods

Sixteen participants (eight male, eight female), aged 18 to 35 years (mean 23.23 years, *sd.* 4.26 years) were recruited from the University College London ‘Psychology Subject [sic] Pool’ and paid 10 GBP for their participation. All participants reported being native British English speakers with no hearing, reading or speaking difficulties, and normal (or corrected to normal) vision. As before, participants’ hearing thresholds were tested using a Kamplex KD 29 diagnostic audiometer with the inclusion criteria for normal hearing being hearing thresholds of 20 dB HL or better at 125, 250, 500, 750, 1000, 2000, 3000, 4000, 6000 and 8000 Hz. All participants met the criteria for normal hearing.

Prior to starting the experiment, the simulated listening environment was explained to the participants using the image shown in figure 3.1 (see page 80) and participants also completed 32 practice trials with clear-speech targets. Participants then completed 576 trials in 8 blocks of 72 trials. Within each block the target ear was the same and alternated between blocks. Half the participants started with the left target ear and the other half started with the right target ear. As before, prior to each block participants were shown a visual instruction (e.g., “Pay attention to your LEFT ear, ignore any other sounds”). During each block, the order of congruency, SNR and DNR was randomised.

### 3.2.3 Results

Trials where a response was not given were discarded (76 trials, < 1% of the data). The rest of the participant’s data were screened for accuracy performance below 3 MAD below the median accuracy for all participants (Leys et al., 2013), as excessively low accuracy could bias the estimates of drift-rate parameters. Although the software used to estimate the parameters of the DDM uses prior distributions which are robust to RT outliers (Wiecki et al., 2013), an informal visual check of RT histograms for each participant in each condition was made and revealed no concerns.

The HDDM model was fitted to the RT data with the  $v$ ,  $a$  and  $t$  parameters dependent on a three way Congruency  $\times$  SNR  $\times$  DNR interaction. Parameters were fitted for each participant constrained under a common variance within each condition (Wiecki et al., 2013). The MCMC sampling process was run for 43927 iterations, with a burn-in of 188 iterations and no thinning. Convergence was assessed informally by visually inspecting plots of the traces and formally

		Drift-rate			
DNR	SNR	Congruency			
		Congruent (CON)		Incongruent (INC)	
		Mean	sd.	Mean	sd.
None (NN)	-4 (LO)	1.78	0.40	0.46	0.57
	0 (MD)	1.96	0.28	1.69	0.51
	+4 (HG)	2.38	0.56	1.53	0.50
	+60 (CL)	2.28	0.38	0.42	0.60
MMSE (MM)	-4 (LO)	1.45	0.41	1.17	0.48
	0 (MD)	1.85	0.31	1.81	0.28
	+4 (HG)	2.63	0.53	1.13	0.50
Spec-sub. (SB)	-4 (LO)	1.09	0.35	1.73	0.24
	0 (MD)	1.95	0.63	1.61	0.52
	+4 (HG)	2.10	0.48	1.12	0.57
Sub-space (SP)	-4 (LO)	1.35	0.28	0.83	0.41
	0 (MD)	1.79	0.48	1.16	0.36
	+4 (HG)	2.51	0.47	0.74	0.62

Table 3.1: Means and standard deviation drift-rates for all experimental conditions.

using the Geweke statistic. Both assessments revealed no concerns. Model fit was assessed visually by overlaying the posterior predictive sample for each audio condition on the observed RTs (see figure 3.2), and was considered adequate ( $MSE < 0.01$ ).

### Drift-rate

Table 3.1 shows the group means and standard deviations for the posterior drift-rate in all experimental conditions, which are also illustrated (excluding the CL-NN condition) in figure 3.3 along with the 95% HDIs and the mode (the highest 1% credible values of the posterior). Some distinct trends are apparent, with increased SNR resulting in an increased drift-rate in all conditions. It also appears that drift-rate is higher in general in congruent (CON) trials than in incongruent (INC) trials as would be expected. But it is not clear if the difference in drift-rate between CON and INC trials (i.e., drift-interference) varies with change in SNR or DNR.

Figure 3.4 illustrates the planned comparisons carried out to examine the effects of SNR on CON and INC trials separately with SNR differences of 4 dB (the MD-LO and HG-MD comparisons) 8 dB (the HG-LO comparison) and also difference between the presence and absence of different levels of noise (the CL-HG, CL-MD and CL-LO comparisons). The changes are in the expected direction with drift-rate at lower SNRs being less than drift-rate at the higher SNRs (shown by the positive modes), consistent with the idea that energetic masking

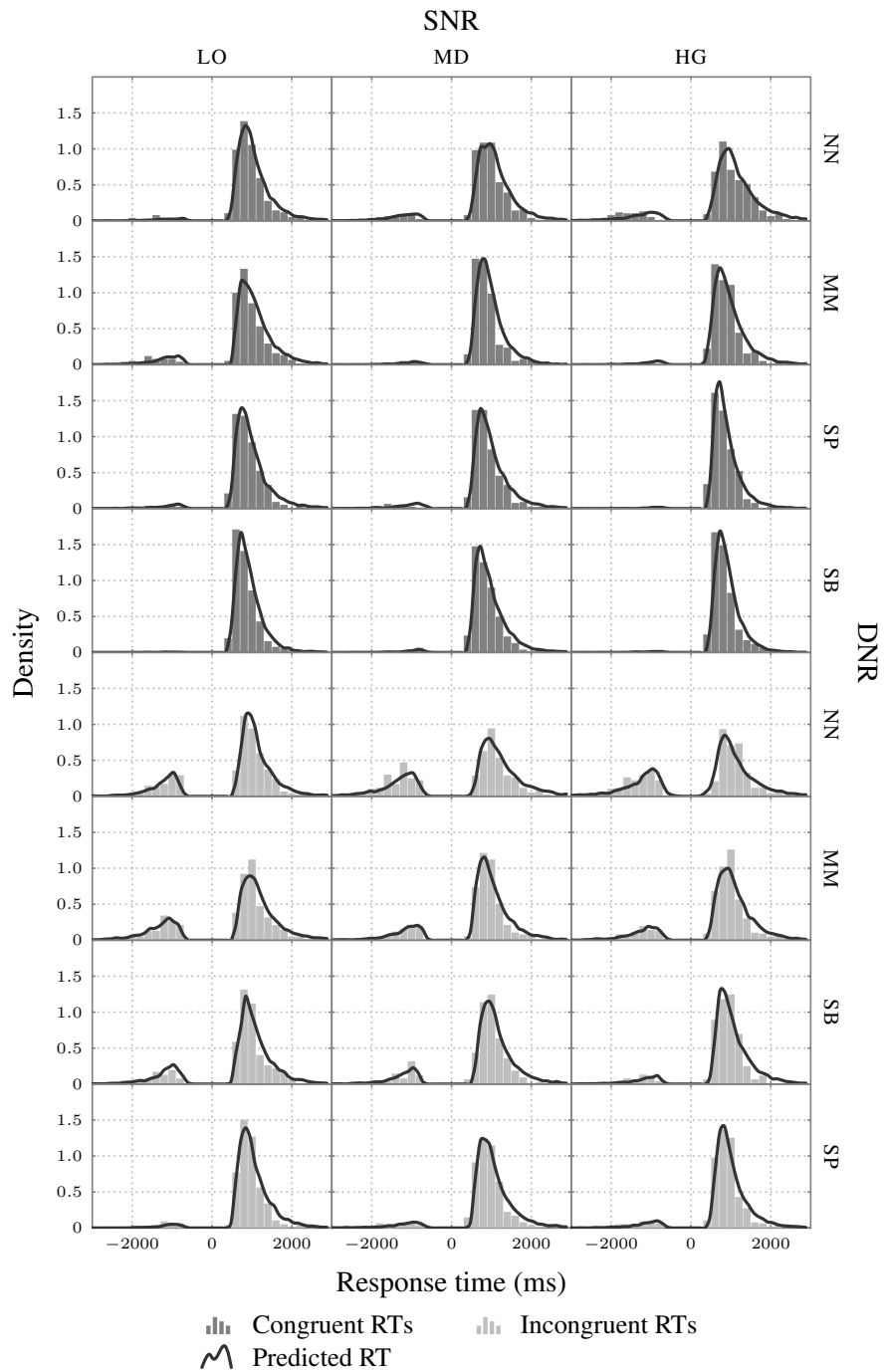


Figure 3.2: Plots of posterior predicted response time (RT) densities overlaid on to (normalised) histograms of the observed RTs for congruent trials (top) and incongruent trials (bottom). Fifty sets of posterior predictive samples were generated from the fitted model and averaged to produce the densities. Negative RTs are RTs for trials where an incorrect response was made.

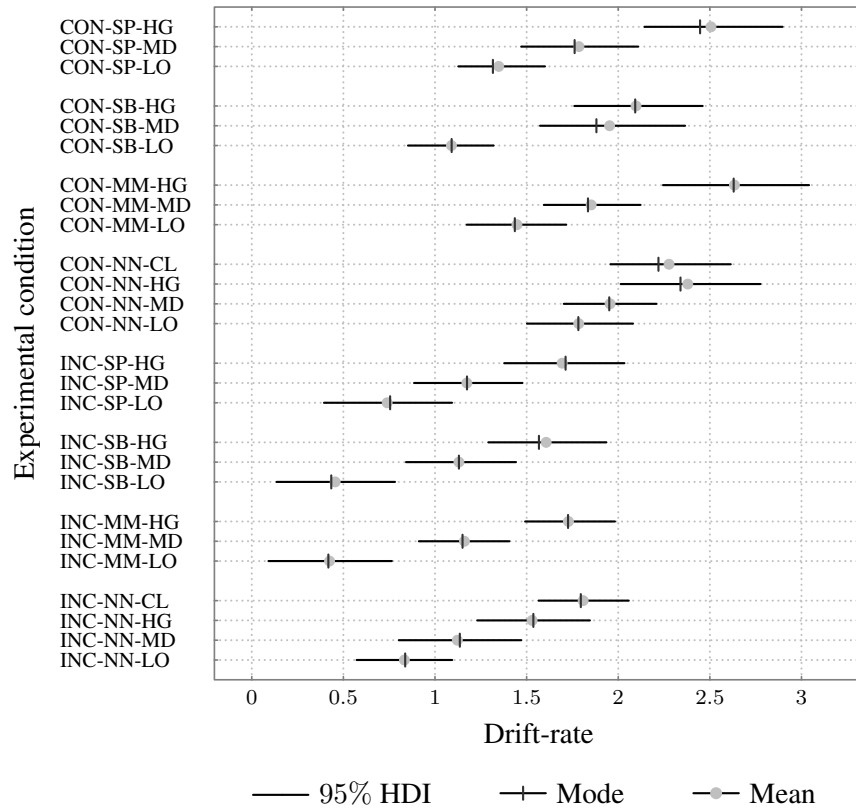


Figure 3.3: Posterior drift-rate means, modes and 95% HDIs for all experimental conditions.

reduces the availability of information so the decision process either takes longer or is more prone to error. However, it can be seen that the most credible changes in drift-rate occur only with relatively wide changes in SNR particularly the HG–LO and CL–LO comparisons in INC trials and that these differences constitute moderately large effects size. The other comparisons show varying levels of uncertainty in the differences, in particular the CL–HG arguably shows no difference as the null region is not only inside the 95% HDI, but is also very close to the most credible values of the posterior (i.e., the mode).

With little in the way of credible differences in CON and INC trials due to SNR, there was no expectation that the difference in drift-rate between to the two congruency conditions (i.e., drift-interference) would be credibly affected by increasing the SNR. Figure 3.5 shows that drift-interference averaged across conditions, is credibly replicated in the current experiment, but constitutes a fairly small effect size ( $\hat{d} = 0.14$ ). Figure 3.6 shows comparisons between drift-interference at various levels of SNR in the NN DNR condition and it can be seen that in



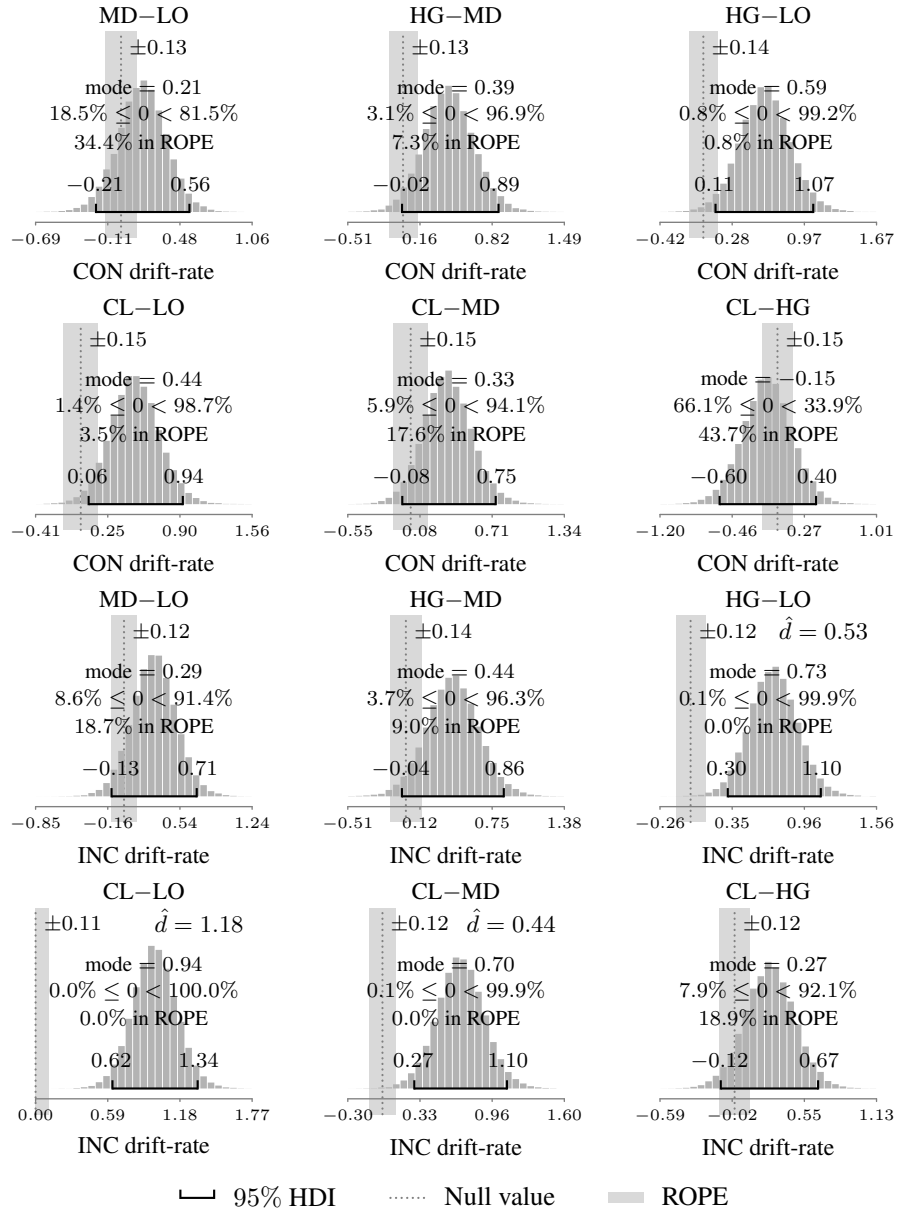


Figure 3.4: Histograms for the planned comparisons of drift-rate between selected SNRs for congruent (CON) trials (top two rows) and incongruent (INC) trials (bottom rows), comparing between various levels of noise (top and second from bottom rows) and the presence and absence of noise (bottom and second from top rows). Credible effect sizes ( $\hat{d}$ ) are shown only for the credible differences.

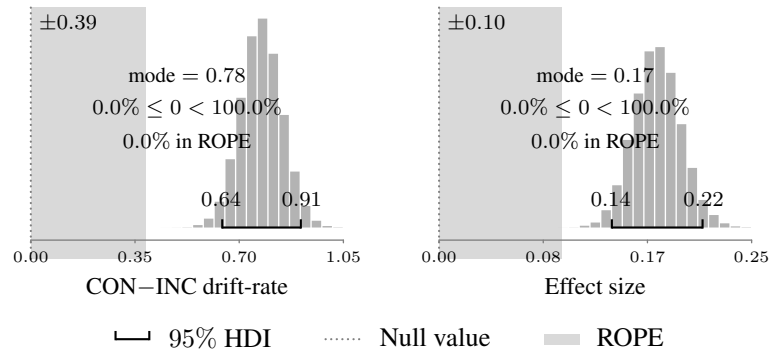


Figure 3.5: Histogram for the ‘drift-interference’ planned comparison between drift-rate in congruent (CON) and incongruent (INC) trials averaged across all DNRs and SNRs.

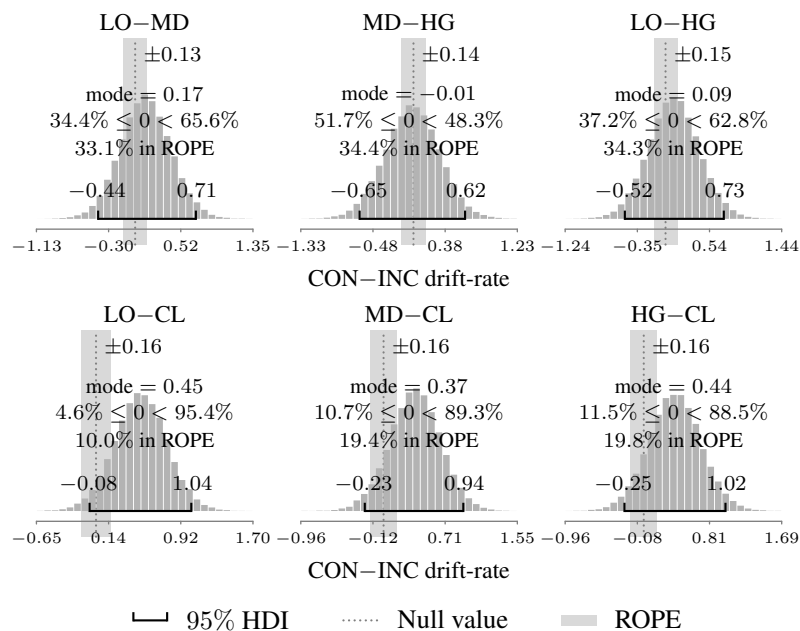


Figure 3.6: Comparison for the change in drift-rate (between various SNR levels without DNR processing (i.e., the NN condition)).

the comparisons with different levels of noise (figure 3.6, top row), no credible difference in drift-interference can be concluded. The null region is not only inside the 95% HDI, it either covers or is very close to the most credible differences of the posterior (i.e., the mode). Although almost 60% of the credible values in the HDI are outside the ROPE, the most credible changes in drift-interference are practically equivalent to the null. In the drift-interference comparisons for the presence and absence of noise (figure 3.6, bottom row), there are larger changes in drift-interference but these do not reach credibility with a large degree of certainty as the null region is either fully inside the HDI or overlapping with the HDI, showing that no difference is a credible interpretation.

With no credible effects of lowering SNR on interference, it is not possible to evaluate if any of the DNR algorithms reverses the effect of SNR. Nevertheless, with suggestions that some DNR algorithms can make performance *worse* (Hu & Loizou, 2007) due to increased distortion introduced into the signal, it was considered worth examining if the MM, SP or SB DNR conditions show credible performance differences in comparison with the NN DNR condition. Figure 3.7 shows SNR comparisons for drift-rate between NN–MM, NN–SB and NN–SP in congruent (CON) trials (top two rows) and incongruent (INC) trials (bottom two rows) for the LO and HG conditions. The presence of credible differences is only indicated with considerable certainty for the NN–SB comparison in the LO SNR condition for CON trials only, with the drift-rate in the SB condition lower than in the NN condition suggesting a longer decision process or more errors; this difference constitutes a moderately large effect size ( $\hat{d} = 0.62$ ). Other comparisons come very close to the margins of credibility but, in general, there are few credible differences that can be inferred with any certainty.

### Non-decision time

Table 3.2 shows the group means and standard deviations for the posterior non-decision times from experimental conditions and figure 3.8 illustrates the means, along with the modes and the 95% HDIs. Some trends are apparent, with increased SNR resulting in the expected decrease in the non-decision times as less initial sensory processing of the target would be required with reduced energetic masking. However, there is considerable overlap in the HDIs and in particular the modes (the most credible non-decision times of the posterior) are not particularly far apart suggesting that this SNR trend may not turn out to be credible. There is no impression, however, that non-decision time is lower in congruent (CON) trials compared to incongruent

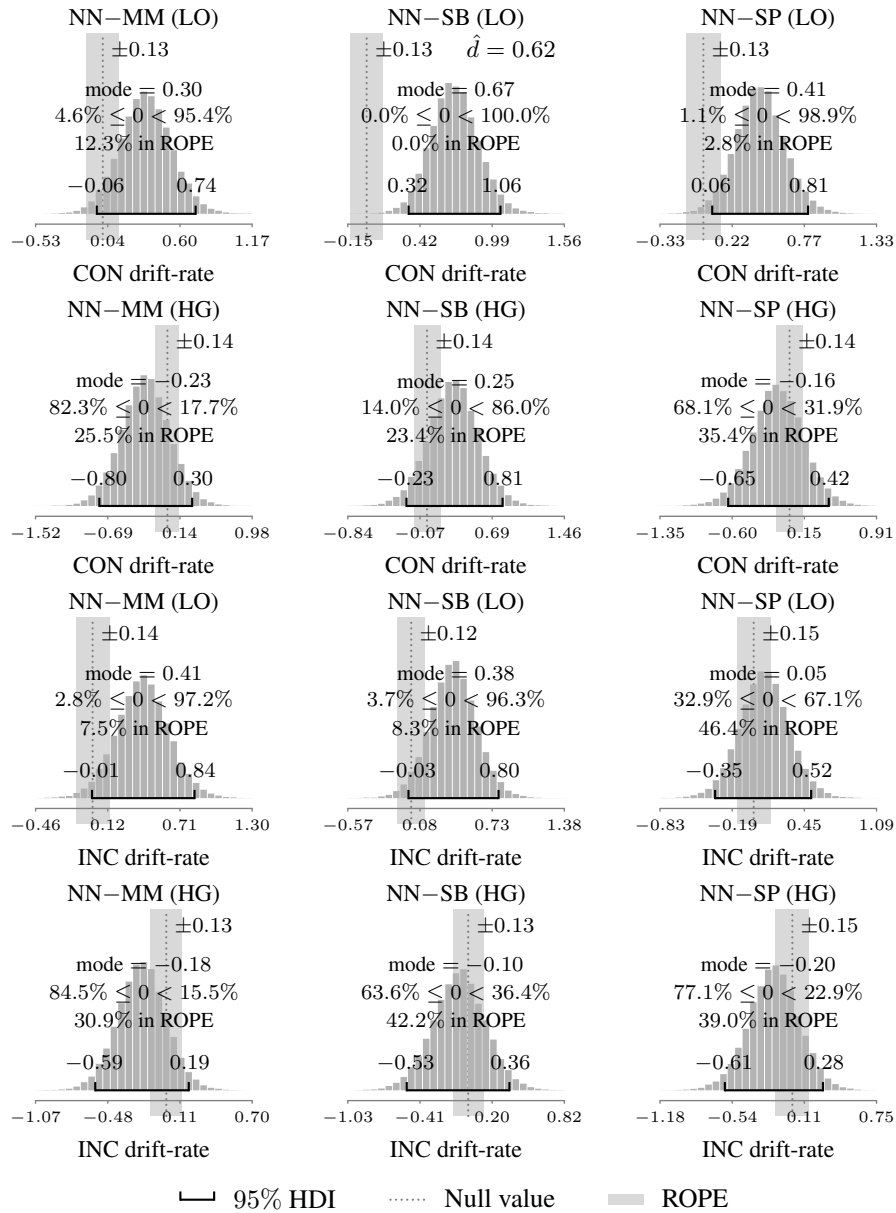


Figure 3.7: Posterior distributions for the change in drift-rate between the NN DNR condition and the MM, SB and SP DNR conditions, in the LO, HG SNR conditions for congruent (CON) trials (top two rows) and incongruent (INC) trials (bottom two rows). Credible effect sizes ( $\hat{d}$ ) are shown only for comparisons which reveal a credible change in drift-rate.

		Non-decision time (ms)			
DNR	SNR	Congruency			
		Congruent (CON)		Incongruent (INC)	
		Mean	sd.	Mean	sd.
None (NN)	+4 (HG)	535	21	470	20
	0 (MD)	580	24	501	28
	-4 (LO)	631	38	504	31
	+60 (CL)	501	27	484	24
MMSE (MM)	+4 (HG)	532	34	451	25
	0 (MD)	589	31	466	38
	-4 (LO)	598	30	578	33
Spec. Sub. (SB)	+4 (HG)	528	31	433	24
	0 (MD)	620	25	475	31
	-4 (LO)	572	43	559	30
Sub-space (SP)	+4 (HG)	548	26	489	24
	0 (MD)	610	29	485	21
	-4 (LO)	643	29	590	37

Table 3.2: Non-decision time means and standard deviations for all experimental conditions.

(INC) trials and the overlapping HDIs and similarity in modes suggest any differences would also be non-credible.

As experiment I found no credible difference in non-decision times between congruent (CON) and incongruent (INC) trials, and figure 3.9 shows no credible difference in non-decision times between CON and INC trial (averaged over all conditions), the traces for CON and INC conditions were averaged for making further inferences. Figure 3.10 shows SNR comparisons for the non-decision times in the NN DNR condition (i.e., without DNR processing). Although some of the comparisons (i.e., LO–HG and LO–CL comparisons) reach the margins of credibility, these are very small differences which fail to constitute even a small effect size ( $\hat{d} < 0.1$  in both cases).

But if decreasing the SNR increases the non-decision time (albeit with very little in the way of credible differences) then an effective DNR should reverse the effects of SNR and decrease non-decision time relative to the SNR condition without DNR processing. Figure 3.11 shows DNR comparisons between non-decision times at the LO and HG SNR levels showing that there are no credible differences in non-decision times between conditions with noisy (NN) targets and conditions with DNR processed targets (MM, SB and SP).

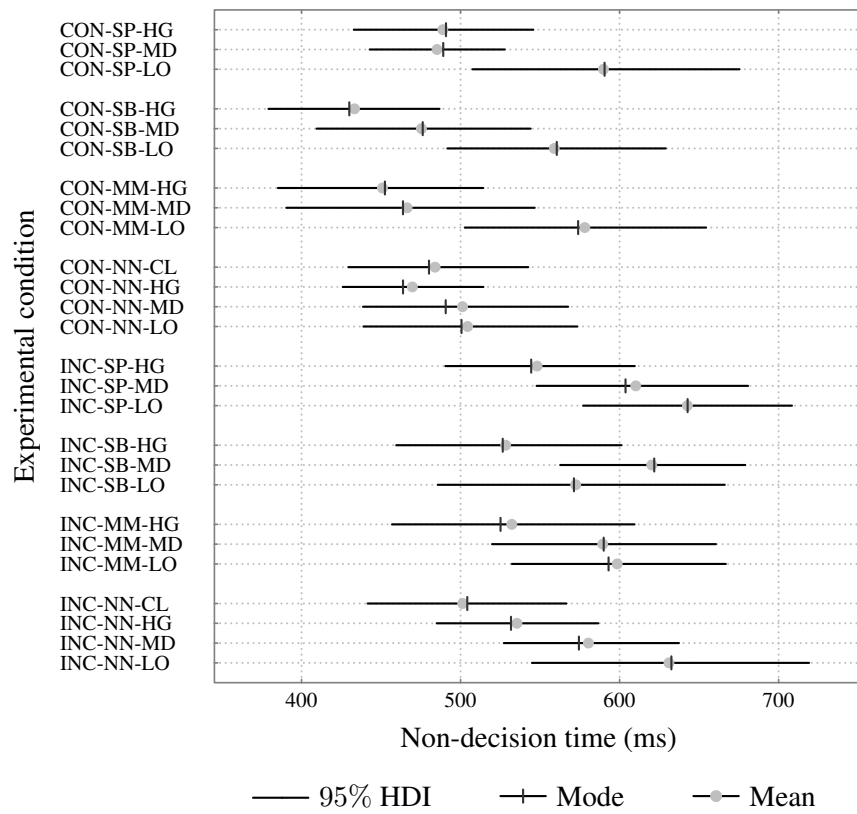


Figure 3.8: Posterior means, modes and 95% HDIs for non-decision times in all experimental conditions.

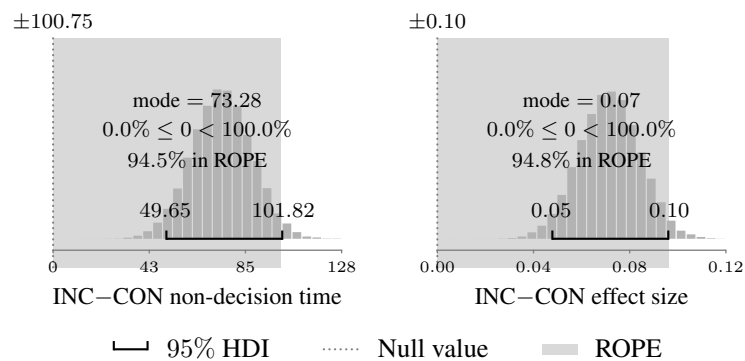


Figure 3.9: Comparison of non-decision time in congruent (CON) and incongruent (INC) trials, averaged over the other experimental conditions. With almost 100% of the HDI overlapping the ROPE it can be concluded with considerable certainty that there is no credible difference in drift-rate between CON and INC trials.

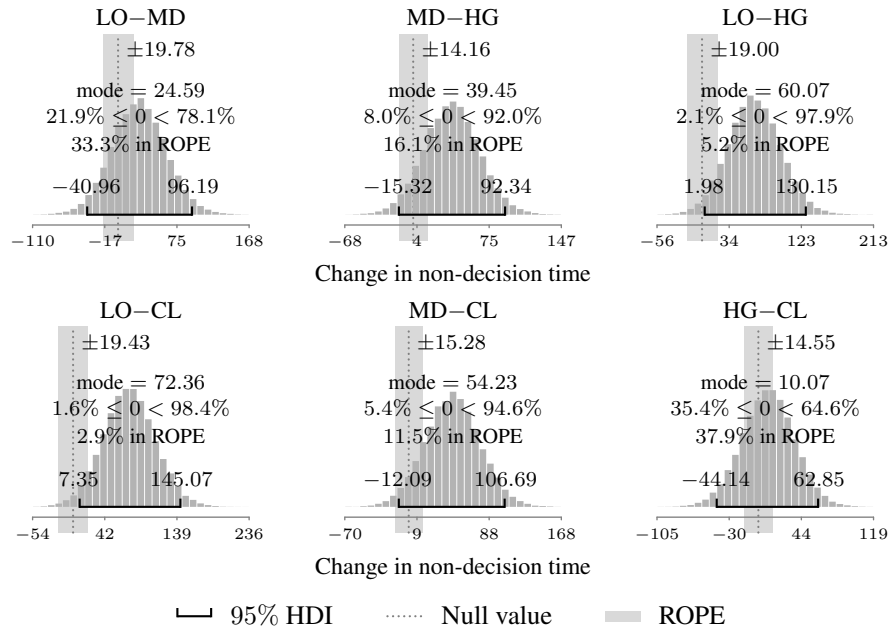


Figure 3.10: SNR comparisons for non-decision times with comparisons for different levels in SNR (top) and the presence and absence of various levels of noise (bottom).

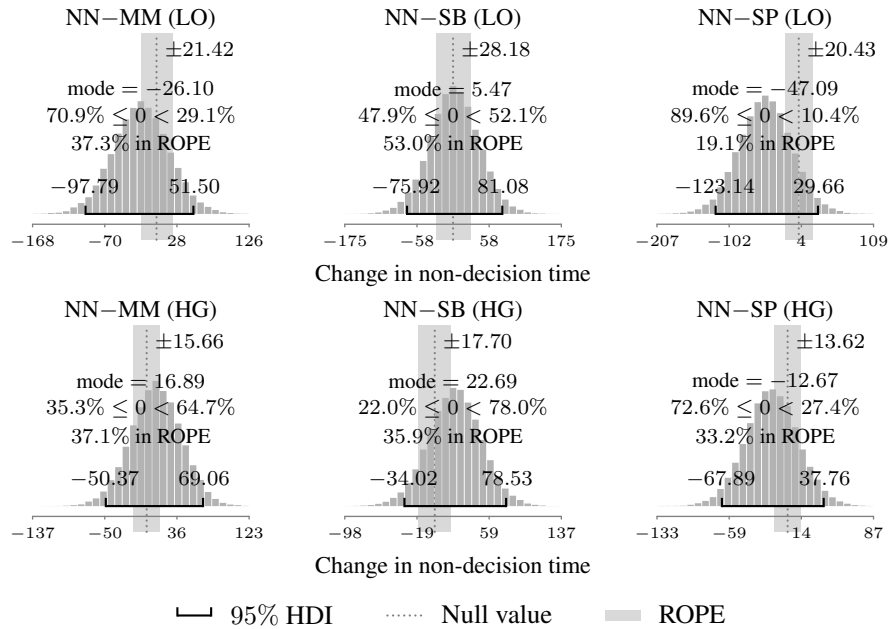


Figure 3.11: DNR comparisons non-decision times showing comparisons for the individual DNR algorithms MM, SB, SP compared to the NN (i.e. no) DNR condition.

Threshold separation

DNR	SNR		Congruency			
			Congruent (CON)		Incongruent (INC)	
			Mean	sd.	Mean	sd.
None (NN)	+4	(HG)	1.91	0.10	2.20	0.17
	0	(MD)	1.66	0.09	1.99	0.15
	-4	(LO)	1.57	0.08	2.19	0.18
	+60	(CL)	2.04	0.12	1.96	0.16
MMSE (MM)	+4	(HG)	2.01	0.15	2.57	0.19
	0	(MD)	1.69	0.08	2.13	0.17
	-4	(LO)	1.73	0.09	1.79	0.10
Spec. sub. (SB)	+4	(HG)	1.95	0.14	2.37	0.23
	0	(MD)	1.63	0.06	2.15	0.19
	-4	(LO)	1.69	0.08	1.91	0.11
Sub-space (SP)	+4	(HG)	1.89	0.14	2.24	0.21
	0	(MD)	1.70	0.10	2.03	0.14
	-4	(LO)	1.61	0.09	1.71	0.10

Table 3.3: Means and standard deviation threshold separations for all experimental conditions.

### Threshold separation

Table 3.3 shows the means and standard deviations of the posterior threshold separations in all conditions, and figure 3.12 shows the means, modes and 95% HDIs for all conditions. Some of the expected trends are suggested by the data, particularly that as SNR increases, the threshold-separation increases suggesting that the reduction in target energetic masking at higher SNRs enabled decisions to be made with more information, although it should be pointed out that this trend is not consistently represented across congruency and DNR conditions. As experiment Ia and experiment Ib found no credible differences in threshold separation in CON and INC trials, and figure 3.13 shows no credible difference between CON and INC conditions (averaged over all other conditions), these posteriors were averaged in each SNR and DNR condition.

Figure 3.14 shows the 4 dB and 8 dB SNR comparisons (top row), and the comparisons with absence and presence of noise (bottom row) in the NN DNR condition. No credible differences were found for any of the comparisons (the HG–MD comparison reached the margins of credibility but the effect size was not credible), suggesting that threshold separation was not credibly altered by the SNRs used in this experiment. In addition, figure 3.15 shows separ-



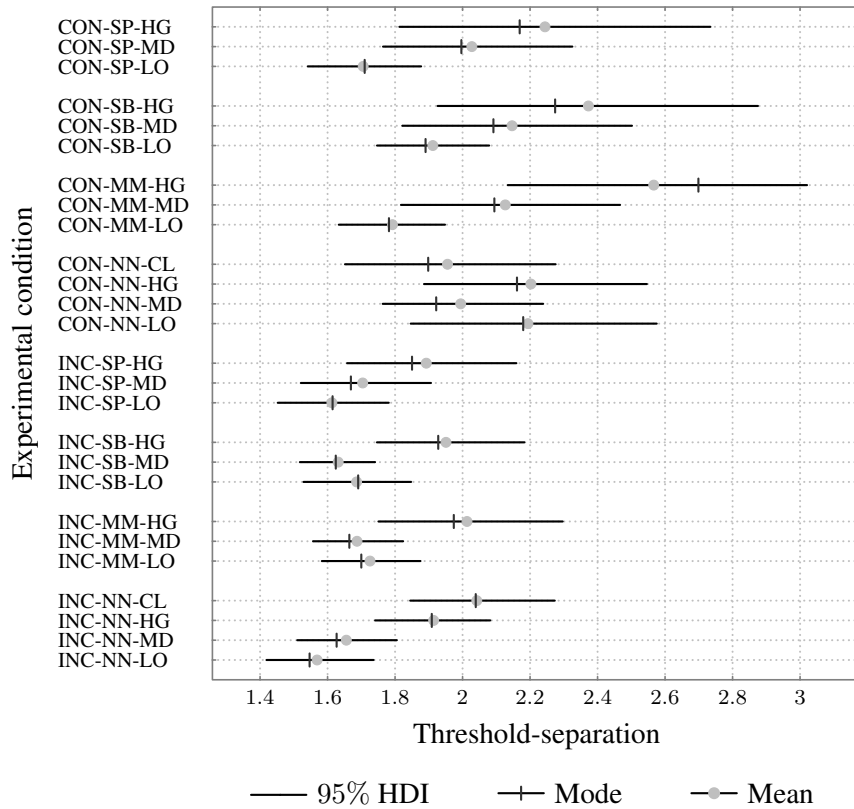


Figure 3.12: Posterior means and 95% HDIs for threshold separation in all experimental conditions.

ate comparisons between the NN condition and the three MM, SB and SP conditions together at LO and HG SNR. No credible differences between the presence or absence of DNR processing were found, demonstrating that the current experiment provided no evidence that DNR processing has a credible effect on threshold-separation.

### 3.2.4 Discussion

Having established a measure of interference using the DDM in experiment Ia and experiment Ib, experiment II aimed to take a ‘first look’ at how the DDM parameters and the measure of ‘drift-interference’ (the difference in the drift-rate between congruent and incongruent trials) were affected by noisy targets and whether any effects observed were reversed by the application of DNR. Background noise at three SNRs  $-4$ ,  $0$  and  $+4$  dB (the LO, MD and HG conditions, respectively) was mixed with flanker targets and processed with three representative DNR algorithms or left unprocessed (the MM, SP, SB and NN conditions, with an additional

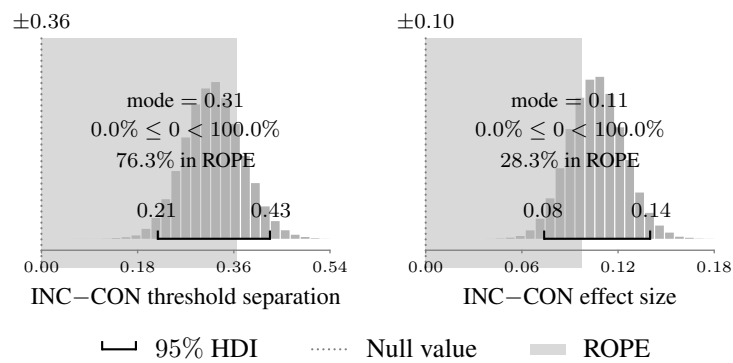


Figure 3.13: Comparison of threshold-separation in congruent (CON) and incongruent (INC) trials, averaged over the other experimental conditions. With over 75% of the HDI overlapping the ROPE there is a reasonable level of confidence that on average there is no credible difference in threshold-separation between CON and INC trials.

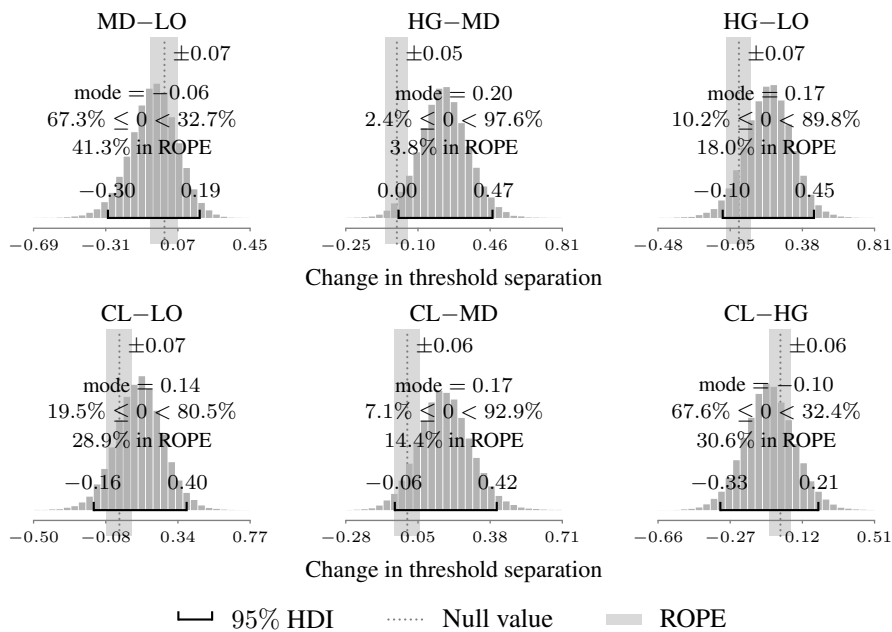


Figure 3.14: SNR comparisons for threshold-separation with comparisons for increases in SNR (top), and the presence of various levels of SNR and the absence of noise (bottom).

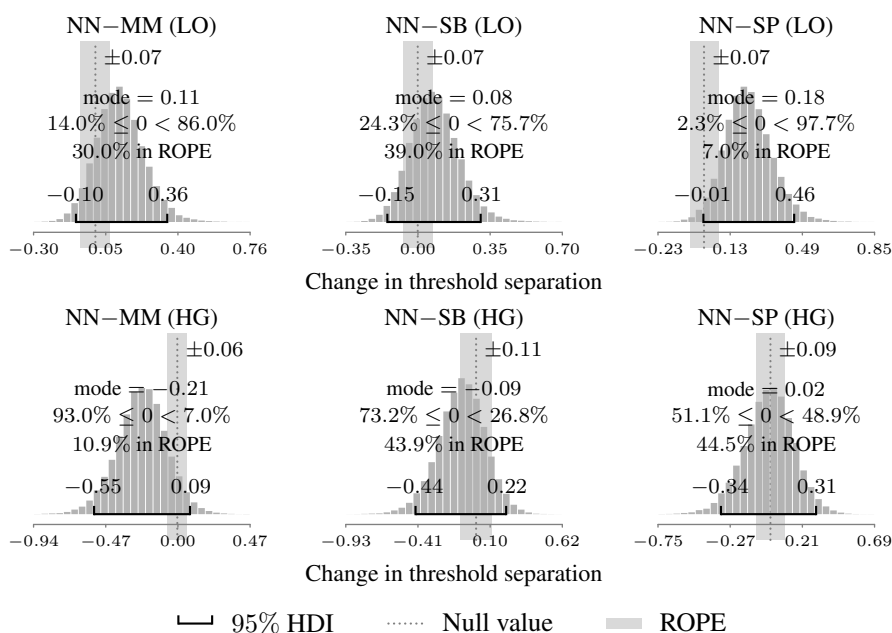


Figure 3.15: DNR comparisons for threshold separation showing comparisons for the individual DNR algorithms MM, SB, SP compared to the NN (i.e. no) DNR condition.

CL control condition). The results showed that although decreasing the SNR also decreased the drift-rate in both congruent and incongruent trials in all DNR conditions these differences were only credible for large differences in SNR. Furthermore, drift-interference (the change in drift-rate between congruent and incongruent trials) was not credibly different between any of the SNR or DNR levels. These results suggest that in its current form, the auditory flanker was not sensitive to small changes in SNR.

Although the sensitivity to SNR was non-credible, it is worth noting that in the drift-interference comparisons between noisy and noise-free targets at least 88% of the posterior was above the the null value, suggesting that 88% of the time interference would be greater with noisy-targets than with noise-free targets. While the current data did not provide evidence that this difference was credible, this implies that that noisy or distorted targets could increase distraction, but the measurement of distraction in the current experiment was not sensitive to the increase.

As increased distraction in the flankers task is associated with increased cognitive load, these results suggest that the noisy targets constitute a cognitive load, although no definitive

conclusions can be drawn on the basis of the current results. The current experiment may have found no credible evidence that any changes in distraction occurred because any increases in distraction were too small (or too variable). If the experimental task could be altered increase the likelihood that participants would be distracted, then increases in distraction due to energetic masking may be magnified, providing a measurable change in distraction in low SNR conditions. One such experimental manipulation involves explicitly increasing cognitive load and contrasting attentional performance in conditions of low and high cognitive load. The introduction of an explicit cognitive load has been shown to increase distraction in both the visual and auditory domains (Lavie et al., 2004; Dalton et al., 2009; Francis, 2010). In particular, Francis (2010) embedded each auditory flanker trial in a memory probe task, requiring participants to remember one or six digits (representing low and high cognitive load respectively) and found that RT-interference (measured as the difference of mean RT in congruent and incongruent trials divided by the mean of the RT means) increased under high cognitive load. So, experiment II was modified to introduce a cognitive load in a similar manner to Francis (2010).

### 3.3 Experiment III

The aim of experiment III was to explicitly increase cognitive load, and force participants to divide their attention between performance in the auditory flanker task and memory for digits in a probe task. Each flanker trial would be embedded between the presentation and recall phases of the probe task (see figure 3.16 on page 103) in which listeners would be presented with a recall set consisting of one digit (low cognitive load) or six digits (high cognitive load) and respond according to whether a subsequent ‘probe’ digit was present or absent from the recall set.

Maintenance of performance in the high and low cognitive load conditions was expected to vary the available capacity to control attention (and inhibit distraction) during the flanker task. As increases in cognitive load are associated with decreases in attentional control (Lavie, 2010), this manipulation of cognitive load was anticipated to lead to specific changes in performance in the flanker task. As increased noise would increase the energetic masking, this would reduce the target information, leading to a lower drift-rate. However, the increased cognitive load was expected to decrease the control of attention resulting in increased processing of distractor information, which in turn would result in further decreases in the drift-rate.

So, energetic masking would reduce the amount of target information available and cog-

nitive load would reduce the control of attention for processing the target information. At high SNR and low cognitive load relatively more target information would be available (less energetic masking) and attention would be more focused on that information (greater attentional control) resulting in more target information and less distractor information being accumulated in the decision process (i.e., higher drift-rates) with a relatively smaller difference between congruent and incongruent trials (i.e., lower drift-interference). By contrast, in low SNR and high cognitive load, relatively less target information would be available (increased energetic masking), and attention would be less focused on that information (reduced attentional control), so less target information and more distractor information would be accumulated in the decision process (i.e., lower drift-rates), but there would be a relatively larger difference in drift-rate between congruent and incongruent trials (i.e., higher drift interference).

In this way, it was anticipated that that the drift-rate would be reduced as the SNR was reduced (although as in experiment II this might not be a credible reduction in drift-rate), but that under conditions of high cognitive load the drift-rate would be reduced further in incongruent trials compared to congruent trials, producing credible increases in drift-interference compared to conditions of low cognitive load. It was not anticipated that threshold-separation would be affected by cognitive load as threshold-separation reflects the amount of information required to make a decision, and no prior research indicated a mechanism by which cognitive load would influence the amount of information used in making a decision (although see Mattys and Wiget, 2011 for suggestions that cognitive load may influence the *type* of information used in perceptual decision making). Nor was it anticipated that non-decision times would be increased under higher cognitive load. While failures to control attention may result in attention being initially directed to the distractor, the decision process is presumably initiated after sufficient sensory processing of *either* the target or the distractor, and the recovery from the failure of attention (assuming a correct response is given) would occur during the decision process and be reflected in lower drift-rates.

With predictions regarding how SNR would influence flanker performance, it was possible to specify the criteria for DNR evaluation. If decreasing SNR under high cognitive load credibly increased drift-interference then the criterion for DNR evaluation would be that that low SNR targets processed with DNR should show credible decreases in drift-interference compared to the same level of SNR (without DNR), with the effect being greater under high cognitive load.

### 3.3.1 Materials

The recordings used to form the flanker targets and distractors were identical to those used in previous experiments (see section 2.3.2), and procedure used to create the flanker stimuli was the same procedure used in experiment II (see section 3.2.1). The same three SNRs were used –4, 0 and +4 dB (LO, MD and HG SNR conditions, respectively), but there were only two DNR conditions MM (the MMSE DNR algorithm) and NN (no DNR).

Two cognitive load conditions were used. In the high cognitive load condition (HCG) the recall set consisted of six digits selected from the set 1–9 under the constraint that there were no repeated digits and no more than two consecutive digits formed an ascending or descending sequence (so 851249 was an acceptable sequence, but 851239 was not). In the low load condition (LCG) the recall set was a single randomly selected digit. Recall sets were chosen so that in consecutive trials the recall sets could contain at most three of the same digits, which had to be in a different order.

For each, trial a recall set was created and a probe selected. The probe was selected at random, subject to the constraints that (i) it was either in the recall set or not in the recall set with equal probability, (ii) the probe could not be the same on consecutive trials, and (iii) the probe could not be the same as the recall set on the previous trial in the LCG condition.

### 3.3.2 Methods

15 participants (4 male, 11 female), aged 17 to 46 years (mean 25.00, sd. 8.29 years) were recruited from the University College London ‘Psychology Subject [sic] Pool’ and paid 10 GBP for their participation. All participants reported being native British English speakers with no hearing, reading or speaking difficulties and with normal (or corrected to normal) vision. Participants’ hearing thresholds were tested using a Kamplex KD 29 diagnostic audiometer with the inclusion criteria for normal hearing being thresholds of 20 dB HL or better at 125, 250, 500, 750, 1000, 2000, 3000, 4000, 6000 and 8000 Hz. All participants met the criteria for normal hearing.

The experiment was run using the same equipment used in experiment Ia and in the same environment (section 2.3.3). Stimuli (recall set, mask, flanker words, and probe) were displayed in white on a black screen using the same font (mono-spaced GNU FreeMono bold, pixel size 72). The simulated listening environment was explained to participants with reference to an

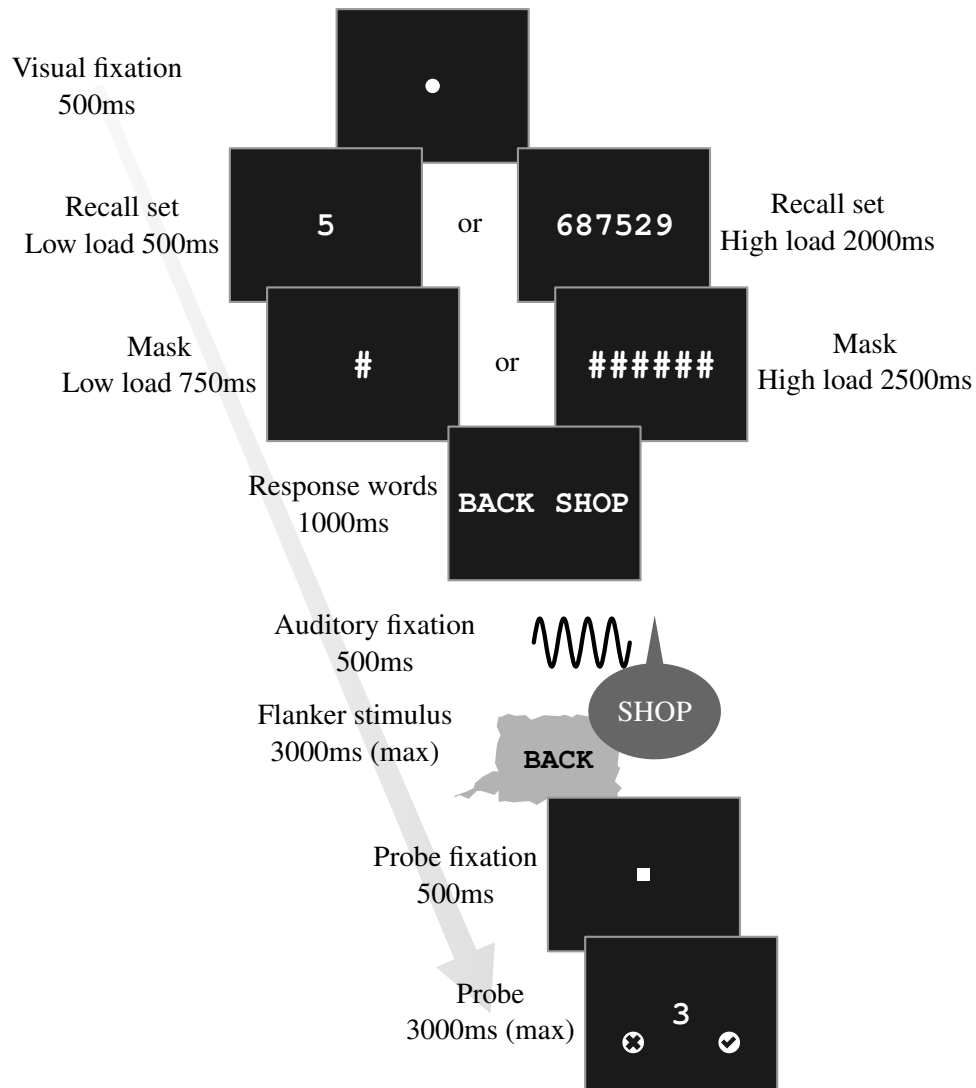


Figure 3.16: Format of a single trial for experiment III.

accompanying picture (reproduced in figure 3.1 on page 80). Participants were then shown an untimed version of a single trial with each stage of the trial explained verbally. Each trial consisted of a flanker trial embedded inside the memory probe task (similar to Francis, 2010 and Lavie et al., 2004) and is illustrated in figure 3.16. After a 500 ms visual fixation, the recall set was shown in the centre of the screen for 500 ms in the LCG condition, and 2000 ms in the HCG condition, followed by a mask (750 ms for LCG and 2500 ms for HCG). The timings for the recall set and mask were based on (Lavie et al., 2004). Following the mask, the response words were shown on the screen for 1000 ms, and then the flanker stimulus was played over headphones, consisting of a 500 ms auditory fixation in the target-ear followed by the stimulus

words. Participants had a maximum of 3000 ms to respond to the flanker, responding with their left hand if the target word was on the left of the screen, or their right hand if the target word was on the right of the screen (using the left- and right-control keys of the computer keyboard, respectively). A red-cross appeared on the screen for 250 ms if an incorrect response was made, but no feedback was given for a correct response. The trial continued to the recall stage of the memory probe task as soon as they had responded (or after 3000 ms if no response was made). Another visual fixation was displayed for 500 ms followed by the probe digit. Participants again responded using the computer keyboard with the right-control key if the probe was in the recall set and the left-control key if the probe was not in the recall set (two icons were shown on the screen at the same time as the probe as a reminder: a cross on the left of the screen and a tick on the right of the screen). A red-cross appeared on the screen for 250 ms if an incorrect response was made, and no feedback was given for a correct response. In both cases the next trial followed after a 250 ms delay.

Participants completed 32 practice trials using targets with no background noise or DNR processing, 16 trials at low load and 16 trials at high load. After finishing the practice trials participants then completed 288 trials in 4 blocks of 72 trials. Within each block the target ear and cognitive load was the same. Cognitive load alternated between blocks, and target ear alternated every two blocks. Between each block participants were given the opportunity to take a break, and before each block participants were given an on-screen instruction regarding the target ear (e.g., “Pay attention to your LEFT ear. Ignore any other sounds”). The order of cognitive load and target ear was balanced across participants. During each block, congruency, and audio condition (i.e., SNR×DNR) were randomised.

### 3.3.3 Results

One participant’s results were discarded for performing with very low accuracy in all conditions compared to the other participants (more than 3 median absolute deviations below the median accuracy for all participants cf. Leys et al., 2013).

For the analysis of the flanker task, data in which no response was given in the flanker task or an incorrect response was given in the probe task were excluded (686 trials, 17.01% of the data). The trials where an incorrect response was given in the probe task were excluded in case participants had ‘traded off’ performance in the flanker trial for performance in the probe trial.



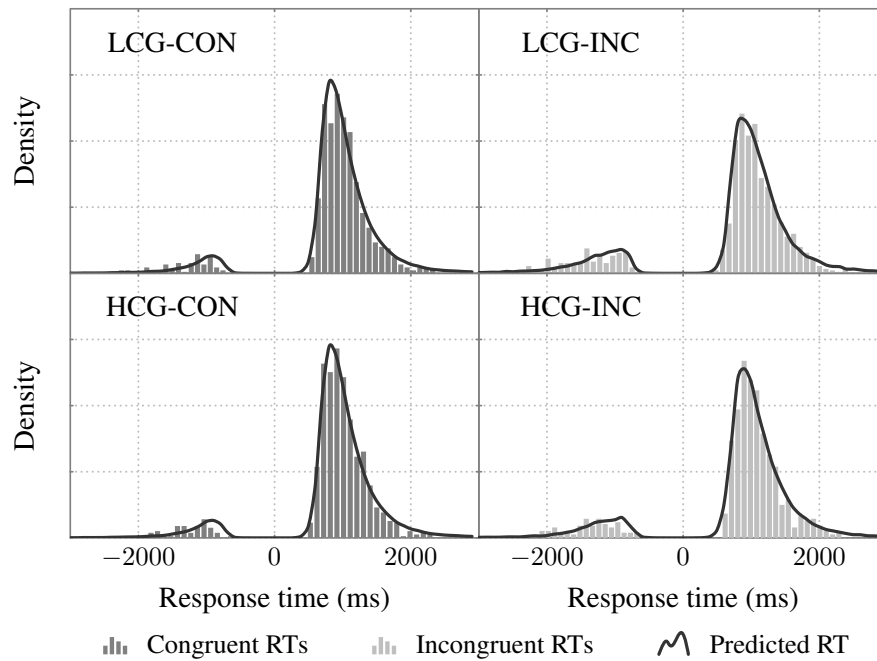


Figure 3.17: Predicted response time (RT) densities overlaid onto the observed RT histograms for the flanker task in congruent (CON) and incongruent (INC) trials under high (HCG) and low (LCG) load. Negative RTs indicate incorrect responses.

The DDM model was fitted to the data with the  $v$ ,  $t$  and  $a$  parameters dependent on the full interaction between cognitive load  $\times$  congruency  $\times$  SNR  $\times$  DNR. Parameters were estimated for each participant constrained by a common variance within each condition (Wiecki et al., 2013). The MCMC sampling ran for 24150 samples with 217 samples discarded as burn in and no thinning. Convergence was assessed by visually inspecting the posterior traces and by the Geweke statistic which revealed no concerns. Model fit was assessed visually (see figure 3.17 for some representative examples) and was considered adequate (MSE = 0.03).

As the only anticipated changes were in drift-interference, analyses are confined to the drift-rate parameter. Table 3.4 shows the drift-rate means and standard deviations for each experimental condition and figure 3.18 illustrates the means and modes (the highest 1% of the distribution where the most credible values are located) along with the 95% HDIs. There appears to be very little in the way of observable trends in the data, although there is the suggestion that drift-rate increases with SNR, although this is by no means a consistent trend, except between the LO and HG SNR conditions (a difference of 8 dB).

		Cognitive load							
		LCG				HCG			
		Congruency				Congruency			
		CON		INC		CON		INC	
DNR	SNR	mean	sd.	mean	sd.	mean	sd.	mean	sd.
NN	LO	1.22	1.53	2.04	1.94	0.95	1.56	1.94	1.77
NN	MD	1.75	2.00	2.78	1.80	2.51	1.90	2.70	2.01
NN	HG	3.30	2.65	2.90	1.84	3.10	2.74	3.57	3.50
MM	LO	0.84	1.10	1.53	1.13	2.02	1.72	2.15	1.88
MM	MD	2.58	2.47	2.41	2.10	1.75	1.49	3.81	2.77
MM	HG	2.46	1.88	2.96	2.03	3.00	2.22	2.87	1.60

Table 3.4: Drift-rate means and standard deviations for all experimental conditions

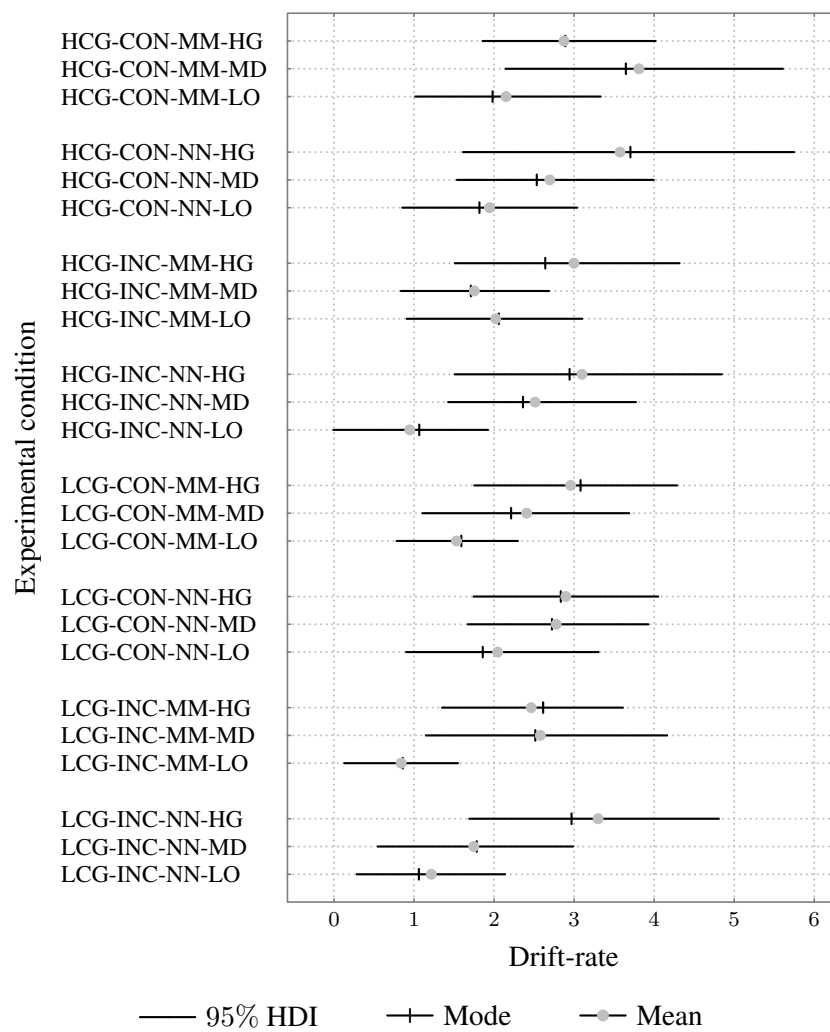


Figure 3.18: The drift-rate means, modes and 95% HDIs for all the experimental conditions.

As with experiment II, consideration was first given to the effects of the key experimental manipulations on distraction for unprocessed noisy speech (i.e., the NN condition) to establish what (if any) changes in distraction DNR must induce to reverse the effects of background noise. The posterior distributions for drift-interference (i.e., the difference in drift-rate between CON and INC trials) are illustrated in figure 3.19 (top two rows) and demonstrate that no difference (i.e., the null value 0) is a credible value for each comparison and in most cases the null region (i.e., the null value and its ROPE) are close to the most credible values of the posterior (i.e., the mode).

The elimination of drift-interference in both LCG and HCG load conditions is contrary to the idea that increasing cognitive load increases interference (Lavie et al., 2004; Francis, 2010). In order to see if the elimination of the drift-interference effect was confined to the NN condition figure 3.19 (bottom two rows) shows the same comparisons (drift-interference in LCG and HCG load conditions at each SNR) for the MM DNR condition. It can be seen that in general there is not a credible drift-interference effect, although in the HCG-MD condition, the comparison approaches the margins of credibility, but the difference does not constitute a credible effect size ( $\hat{d} < 0.10$ ). With no credible or consistent pattern in these effects, it is reasonable to conclude that the addition of the probe task eliminated the consistent drift-interference effect across the SNR and DNR conditions found in experiment II. As drift-interference was the measure of distraction, it is not possible to examine variations in distraction across the levels of SNR, DNR or cognitive load.

In typical dual-task experiments the impact of maintaining performance on the primary task is measured by decrements in the secondary task. In the current experiment, maintenance of performance in the probe task was expected to have an impact on the flanker task leading to increases in distraction. But, if the addition of the probe task had resulted in participants prioritising the tasks in a way that had not been intended, then it may be that maintenance of performance in the flanker task would have an impact on performance in the probe task. If this was the case then there may be patterns of performance in the probe task that reflect the demands of maintaining performance in the flanker task with the noisy and de-noised targets. In particular, the demands of attending to noisy or distorted speech has been demonstrated to have an impact on recall (Rabbitt, 1968, 1966; Howard et al., 2010; Sarampalis et al., 2009; Luce, Feustel & Pisoni, 1983) so an analysis of the probe task may be revealing.

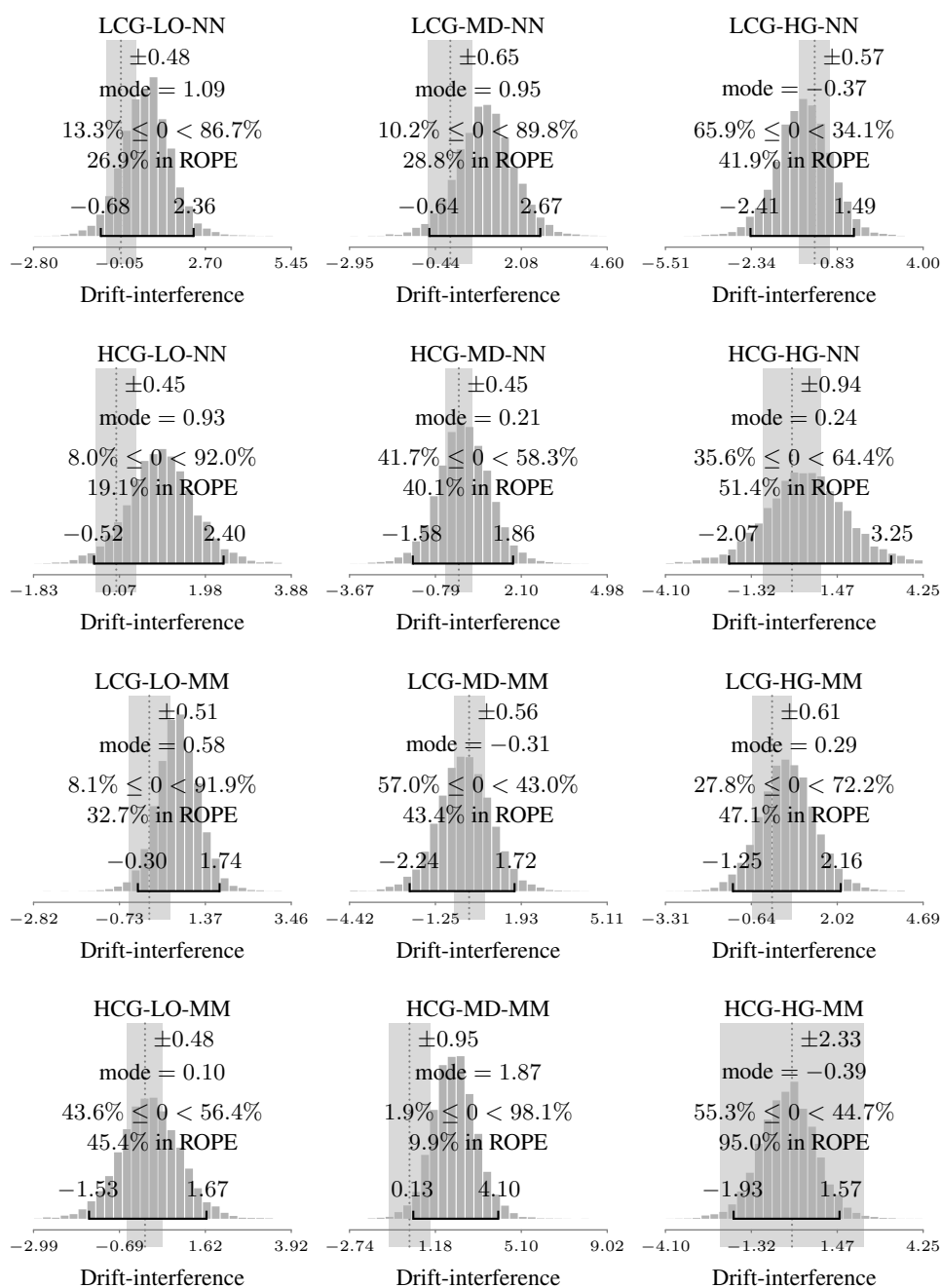


Figure 3.19: Drift-interference (the comparison of drift-rate between congruent and incongruent trials) in low (LCG) and high (HCG) cognitive load, at three levels of SNR low (LO) medium (MD) and high (HG) with no DNR (NN) processing in the target (top two rows) and MMSE DNR (MM) processing in the target (bottom two rows).

### 3.3.3.1 Probe task

Like the flanker task, the probe-task is an example of a 2AFC task and could be analysed using the DDM.<sup>1</sup> For modelling memory processes, the drift-rate represents the accumulation of information from the memory trace, and lower drift-rates indicate a less intact memory trace (Ratcliff, Thapar & McKoon, 2004). It was anticipated that increased SNR in the flanker targets would reduce memory performance (Rabbitt, 1968, 1966; Howard et al., 2010; Sarampalis et al., 2009), resulting in lower drift rates under high cognitive load in trials with low SNR flanker targets, and higher drift-rates under low cognitive load in trials with high SNR flanker targets. Threshold-separation and non-decision time were not expected to be credibly different under the different experimental conditions, as degradation of the memory trace was expected to be reflected in reduced drift rate (Ratcliff et al., 2004).

If probe drift-rate decreased as the SNR decreased then the application of DNR to the flanker targets should result in increased probe drift-rate in trials where the flanker target was not processed with a DNR. It was anticipated that this effect would be more credible in high cognitive load as the effects of SNR with unprocessed flanker targets were expected to be greater with high cognitive load.

Data where responses were not given, or an incorrect response was given in the flanker trial were discarded (801 trials, 18.54% of the data). The DDM model was fitted to the probe RT data with the  $v$ ,  $t$  and  $a$  parameters dependent on the full interaction between cognitive load  $\times$  congruency  $\times$  SNR  $\times$  DNR. Parameters were estimated for individual participants constrained by group level variance (Wiecki et al., 2013) and subsequently averaged for posterior inferences. The MCMC sampling ran for 24364 iterations with 70 samples discarded as burn in (as determined by the Raftery-Lewis procedure) with no thinning. Convergence was assessed by visually inspecting the posterior traces and by use of the Geweke statistic, which both revealed no concerns. Model fit was assessed visually (see figure 3.20 for representative examples) and was considered acceptable (MSE = 0.02).

With only drift-rate implicated as an index of performance in memory tasks (Ratcliff et al., 2004), the analysis is confined to the drift-rate parameter. Furthermore, the distinction between congruent and incongruent trials was considered not relevant to the probe task, as

---

<sup>1</sup>In fact, the DDM was originally intended as a model of memory recall (Ratcliff, 1978)

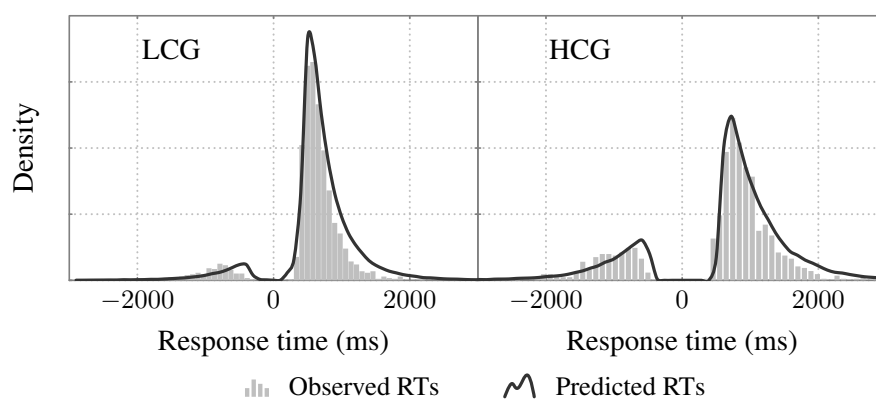


Figure 3.20: Predicted response time (RT) densities overlaid onto the observed RT histograms for the probe task under high (HCG) and low (LCG) load. Negative RTs indicate incorrect responses.

		Cognitive load							
		LCG				HCG			
		Congruency		Congruency		Congruency		Congruency	
		CON		INC		CON		INC	
DNR	SNR	mean	sd.	mean	sd.	mean	sd.	mean	sd.
NN	LO	1.35	0.53	0.86	0.52	1.29	0.24	0.76	0.63
NN	MD	1.80	0.36	1.14	0.66	1.75	0.21	1.52	0.41
NN	HG	1.92	0.27	1.94	0.80	1.83	0.84	1.90	0.65
MM	LO	1.26	0.20	0.68	0.27	1.44	0.30	1.22	0.21
MM	MD	1.66	0.28	1.52	0.52	2.13	0.32	1.31	0.40
MM	HG	1.88	0.37	1.65	0.68	2.01	0.21	1.85	0.43

Table 3.5: Means and standard deviations for probe-drift rate average across congruent and incongruent trials.

the intention was to investigate the impact of listening to noisy targets on recall, regardless of the congruency of the distractor in the flanker trials. So, drift-rates were averaged across congruency, and table 3.5 summarises the posterior group means and standard deviations for drift-rate in the remaining experimental conditions. Figure 3.21 illustrates the means, modes and 95% HDIs. The expected SNR trends are apparent with drift-rate increasing as the SNR increases, and this is a consistent trend, although the 95% HDIs overlap somewhat, there is a clear distinction between HG and LO SNRs. However, there is no obvious distinction between LCG and HCG load trials.

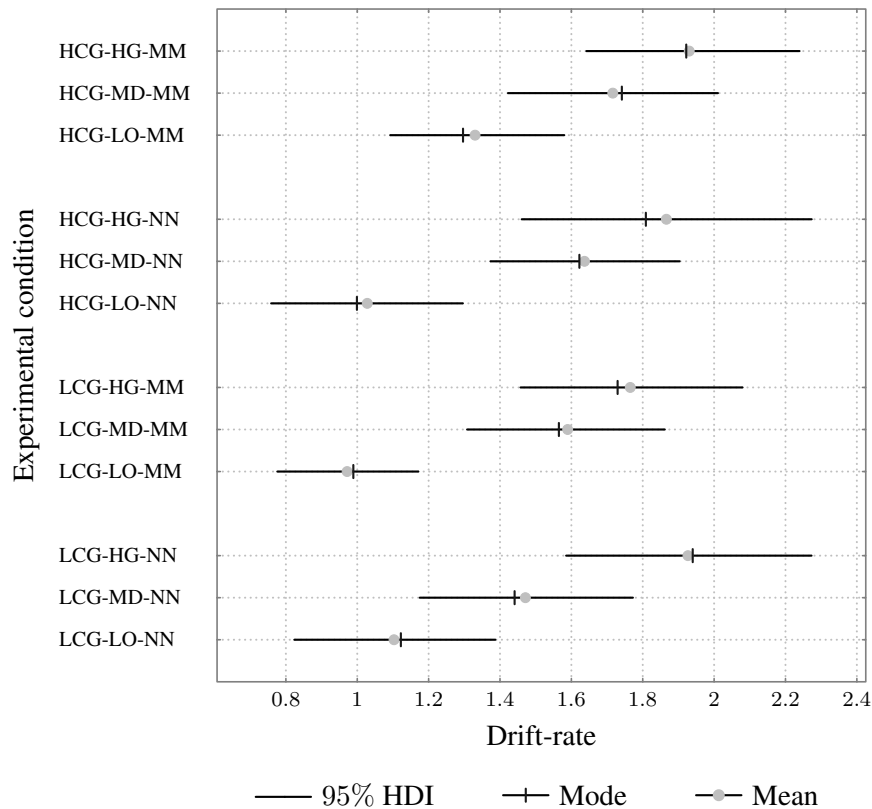


Figure 3.21: Means, modes and 95% HDIs for the memory probe drift-rates.

Planned comparisons between different levels of SNR (MD–LO, HG–MD and HG–LO) for each level of cognitive load in the NN DNR condition are illustrated in figure 3.22. There are credible differences in drift-rate for the HG–LO comparison in both the LCG and HCG load conditions and also for the MD–LO comparison in the HCG load condition. In each, case the drift-rate increases at the higher SNR relative to the lower SNR (i.e., the modes of the comparisons are positive) demonstrating that the memory trace for the recall set was more intact in trials when the target in the flanker trial was presented with a higher SNR. In addition, there is a larger effect size ( $\hat{d} = 0.80$ ) between the 8 dB HG–LO comparison compared to the 4 dB MD–LO comparison ( $\hat{d} = 0.69$ ) suggesting the effect is related to the difference in SNR. The credible reduction in probe drift-rate when the SNR is lowered in the flanker targets indicates that the criterion for DNR performance should be that probe drift-rate should be increased in trials where the flanker target was processed with a DNR (MM) compared to trials where the flanker task was not processed with DNR (NN), at the same SNR. However, figure 3.23 (p. 113) shows that as the null region (i.e., the null value and its corresponding ROPE) is either partially

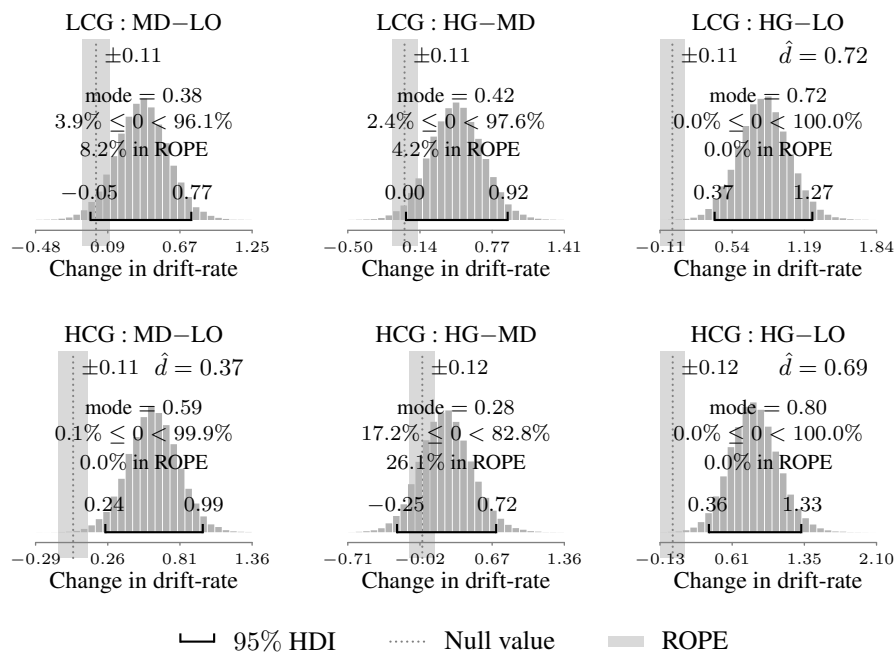


Figure 3.22: Comparison for memory probe drift-rates between trials flanker targets were mixed with different levels of background noise (i.e., NN DNR condition) in high (HCG) and low (HCG) load conditions.

or completely overlapping with the 95% HDI there are no credible differences between the NN and MM DNR conditions at any level of SNR, under either LCG or HCG load.

### 3.4 Discussion

The analysis of performance in the combined memory probe auditory flanker task produced the unexpected result that (drift) interference from flankers was eliminated under cognitive load, at every level of SNR, with or without DNR processing, regardless of the level of cognitive load. However, an examination of the probe task revealed that probe drift-rate (the rate of information accumulation from the memory trace for the probe recall set) was credibly less at lower SNRs, suggesting that the memory trace had become less distinct while completing a flanker trial with low SNR targets. The effects were presumably only due to the addition of noise, as in other respects (i.e., the control of attention in the flanker trial) the trials were largely identical and were found both at high and low cognitive load for differences of 8 dB SNR and at under high load for one of the 4 dB comparisons; other comparisons for differences of 4 dB SNR



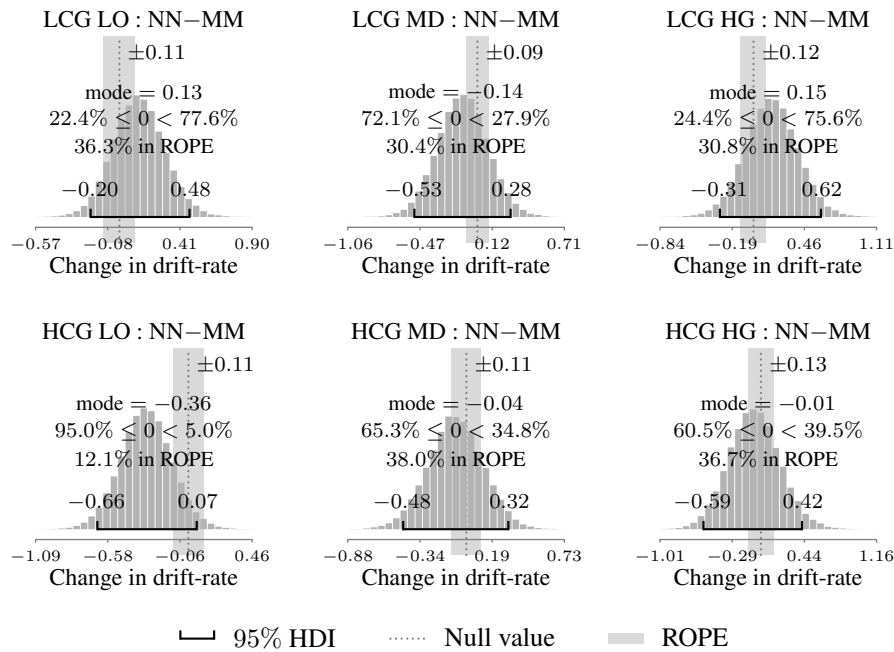


Figure 3.23: Comparison for memory probe drift-rates between trials where flanker targets were DNR-processed (MM) or unprocessed (MM) at three SNRs in high (HCG) and low (HCG) load conditions.

approached the margins of credibility. However, no benefit was found for DNR.

### 3.5 General discussion

In this chapter, the effects of attending to noisy targets in the auditory flanker task was investigated. Previous research had suggested that attending to noisy targets required the utilisation of attentional processes that were not required for clear speech (Wild et al., 2012). So it was hoped that by using noisy targets in the flanker task, more light could be shed on the attentional processes that were utilised for noisy speech. In addition, by examining how attention to noisy targets altered performance in the auditory flanker task, it was hoped that this would provide the criterion by which a digital noise reduction system should be evaluated.

Load theory (Lavie, 2005) proposes a hybrid attentional system which framed the formation of hypotheses regarding how noisy flanker targets would affect interference. According to load theory, attention comprises two limited capacity sub-systems, a passive perceptual system responsible which processes and organises all sensory input subject to available capacity

limits, and a cognitive system for controlling the maintenance of task priorities (e.g., resistance from distraction). Increased load on the perceptual system (due to more complex stimuli) leaves less spare capacity for processing distractors resulting in less interference from distractors. Increased load on the cognitive system (due to increased task demands) leaves less spare capacity to control the focus of attention resulting in more interference from distractors. If the attentional processes required for processing noisy targets were located in the passive limited capacity perceptual sub-system responsible for organising incoming sensory information, then attending to noisy targets would leave no spare capacity to process the distractor, leading to decreased interference. Alternatively, if the attentional processes required for processing noisy targets were located in the active cognitive sub-system responsible for controlling the focus of attention and inhibiting distraction, then attending to noisy targets would result in less control over the inhibition of distractors, leading to increased interference.

Experiment II demonstrated no credible changes in drift-interference due to SNR, so it was not possible to unambiguously conclude that the introduction of background noise constituted a perceptual load (resulting in decreased interference from distractors) or a cognitive load (resulting in increased interference from distractors). However, there were suggestive (but non-credible) indications that drift-interference was higher with noisy targets than with clear targets, implying that the background noise used in experiment II induced a cognitive load, but that it was too small to be credible.

Experiment III aimed to explicitly increase cognitive load in the flanker task in anticipation of inducing a credible change in drift-interference when the SNR was varied. Previous research had indicated that an increased cognitive load would result in increases in interference based on RTs (Francis, 2010) and it was anticipated that similar effects would be noted with the addition of a similar cognitive load in the auditory flanker task developed above. However, under conditions of both high and low cognitive load, drift-interference was eliminated across all other experimental conditions, making it impossible to evaluate if cognitive load influenced distraction in the auditory flanker task. Despite suggestions that cognitive load only increases interference when the stimuli used to increase cognitive load share common properties with the targets (Dittrich & Stahl, 2012) and can even facilitate attention if the material shares features with the distractor (Park, Kim & Chin, 2007), this result was curious given that the cognitive load was operationalised in similar manner to previous research (e.g., Lavie et al., 2004; Francis,

2010). Although Francis (2010) only demonstrated conclusive effects of cognitive load when the target and the distractor were fully overlapping (whereas the targets and distractors in experiment III were only partially overlapping) even in the experiments of Francis where cognitive load did not produce a conclusive change in interference, there was still *some* interference, as opposed to the complete elimination of interference found in experiment III.

However, while it was possible that the addition of the memory probe task may have increased the participant's engagement with the flanker task resulting in a reduction in interference from the distracting speech (Hughes et al., 2012; Murphy et al., 2013), it was equally possible that participants may have shifted their priorities in the combined probe/flanker task, and instead of maintaining performance in the probe task so that the effects of cognitive load would be evident in the flanker task, participants had instead maintained performance in the flanker task (cf. Ahmed & de Fockert, 2012). As a consequence, performance in the probe task was analysed to examine if attending to noisy targets had had an effect on the probe recall (cf., Rabbitt, 1968). Like the flanker task, the probe task could also be modelled as a drift-diffusion process with the probe drift-rate (averaged across congruent and incongruent trials) acting as an indicator of the degradation in the memory trace (Ratcliff et al., 2004). Credible effects of SNR were found, with probe drift-rate lower in trials where the corresponding flanker trial had lower SNR targets compared to trials where the flanker trial had higher SNR targets, which suggested that the memory trace was degraded to a greater degree in trials where targets in the flanker trial had a greater level of background noise. These effects were found under both high and low cognitive load but mainly for large differences in SNR (8 dB).

Considering the application of the auditory flanker task to evaluating speech technologies in terms of their impact on listening effort, the results of the current chapter are not promising. With the only credible results an unintended 'side effect' of the probe/flanker task, and then only for reasonably large differences in SNR, the use of the auditory flanker task as viable alternative or complement to existing speech technology evaluations looks less likely. Nevertheless, if the auditory flanker task is not appropriate for evaluating speech technologies, it remains to be seen if existing speech technology evaluations can really do any better. This issue forms the basis of the next chapter.

## Chapter 4

# The flanker task and speech technology evaluation

### 4.1 Comparing speech technology evaluations

Previous chapters have (i) developed an auditory flanker task, (ii) examined how performance in the flanker task changes with either noisy targets or noisy targets processed with a DNR algorithm to remove the noise, and (iii) examined how performance in the flanker task with noisy targets changes when the flanker trial is made more complex by the addition a memory task. The results have been inconclusive and largely disappointing, suggesting not only that the auditory flanker task may not be a valid method for measuring listening effort, but also that the flanker task is not, therefore, an appropriate task for evaluating speech technologies.

In order to confirm this, the auditory flanker from experiment III is repeated in this chapter along with some ‘traditional’ speech technology evaluations: intelligibility, subjective listening effort, and a version of a standardised test for evaluating communication systems with a DNR component, the ITU-T P.835 test (see section 4.2.2.4). If any of these evaluations provide measurable distinctions between speech at different levels of SNR, or differences between speech with and without DNR processing, then the case for rejecting the flanker task (at least in its current form) as a means of evaluating speech technologies will be stronger.

### 4.2 Experiment IV

The aims of experiment IV were threefold. Firstly, the experiment aimed to replicate experiment III and demonstrate a decrease in the probe drift-rate (indicating poorer recall) for lower

SNRs. Although experiment III had found no difference in probe drift-rate between trials where the flanker target was mixed with background noise and trials where the noisy target was processed with the MMSE DNR algorithm, experiment IV would reintroduce the other DNR algorithms from experiment II (the spectral subtraction and sub-space algorithms) to see if similar differences could be found.

The second aim of experiment IV was to contrast performance in the flanker/probe task with three ‘traditional’ methods for evaluating the impact of noise and distortion on listeners: (i) speech intelligibility, (ii) subjective listening effort ratings, and (iii) and a version of the International Telecommunication Union (ITU) standardised procedure for measuring speech signal quality in DNR systems. Although the flanker/probe task has so far failed to provide a useful measure of listening effort and, the only measurable change in performance was not the expected one, it would be worth while to consider if traditional methods can do better. If the traditional methods provide measurable distinctions between DNR processed and unprocessed speech then this could demonstrate that the flanker/probe task, at least in its current form was not an appropriate task for speech technology evaluation.

Thirdly, participants would complete a task to measure the extent of their attentional control, the ‘Operation span’ (Turner & Engle, 1989; Unsworth, Heitz, Schrock & Engle, 2005), and complete a subjective measure of susceptibility to distraction, the Cognitive Failures Questionnaire (Broadbent, Cooper, FitzGerald & Parkes, 1982). This was to rule out the possibility that the participants were ‘unusually’ resistant to distraction so that if the flanker effect was eliminated in current experiment it could not be attributed to the participants’ uniform ability to resist distraction. In the previous experiments, variations in susceptibility to distraction have been acknowledged but largely ignored. It may be possible that the very large flanker effects observed in experiment I were the result of recruiting participants who were unusually susceptible to distraction. Conversely, the elimination of the flanker effect in experiment III may be the result of recruiting some participants who were unusually resistant to distraction. Before asserting that the flanker task is an inappropriate task for measuring listening effort and, therefore, inappropriate for evaluating speech technologies, it is necessary to ensure that effects (or lack of effects) cannot be attributed to the fact that the individual differences in the participants’ susceptibility to distraction were not that different at all, and unrepresentative of the susceptibility to distraction in general populations.

### 4.2.1 Methods

Twelve right-handed participants (two male, ten female) aged 19–30 years (mean 21.75, *sd.* 3.48 years) were recruited from the University College London ‘Psychology subject [sic] pool’ and paid 30 GBP for their participation. All participants reported being monolingual native British English speakers from birth, with no known speaking, hearing or reading disorders, and with normal (or corrected to normal) vision. Participants’ hearing thresholds were tested using a Kamplex KD 29 diagnostic audiometer; the inclusion criteria for normal hearing were thresholds of 20 dB HL or better at 125, 250, 500, 750, 1000, 2000, 3000, 4000, 6000 and 8000 Hz (BSA, 2011) and all participants met these criteria for normal hearing.

### 4.2.2 Tasks

Experiment IV consisted of five tasks, and all except two of the participants completed the five tasks in a single 2.5 hour session. The remaining two participants required two sessions to complete all of the tasks, due to the first session over-running; the order of tasks was ‘quasi-completely’ balanced across participants.<sup>1</sup> All tasks were completed in a sound-proof booth, and all tasks except the CFQ task were run on an HP desktop computer running the Arch Linux operating system (with real time Linux kernel 3.2) with the participants sat approximately 50 cm in front of a 17 in (43.18 cm) 1280 × 1024 VGA Dell monitor and used a USB keyboard or USB mouse to make responses (depending on the task). The CFQ task was run on a Blackberry playbook tablet, and responses were collected using the touch interface. All auditory stimuli were presented using AKG272 mkII headphones controlled by an Asus Xonar PCI sound card and calibrated so that audio targets (or stimulus words) were presented at 65 dB SPL (determined by a Sono OKKI CF-3502 FFT analyser connected to a Bruel & Kjør artificial ear type 4513).

#### 4.2.2.1 Auditory Flanker and memory probe task

The auditory flanker and memory probe task was the same as experiment III, but with some minor adjustments. Firstly, the contrast between low cognitive (LCG) load and high cognitive (HCG) load was changed to six digits in ascending order (low load) and six digits in random order (high load). This was done to ‘standardise’ the presentation of LCG and HCG recall

---

<sup>1</sup> Although not used in subsequent analysis participants also completed revised version of the Edinburgh handedness inventory (Oldfield, 1971).

lists, as some participants in experiment III had commented that they felt that under LCG load, the shorter display time for the single-digit recall list and its mask made the task more difficult. The timings were based on previous research (Lavie et al., 2004) which had argued that shorter presentation times in LCG load conditions were warranted to prevent the creation of more intact memory trace in comparison to the HCG condition (so better performance in LCG conditions would be due to greater opportunity for rehearsal rather than the level of cognitive load itself). Nevertheless, if the participants' sense of 'difficulty' had some genuine underlying performance decrements, this may have biased the results in experiment III, perhaps by reducing the difference in the performance between LCG and HCG conditions to the extent that no credible distinction could be made between them. The use of the ascending and random order digit lists of the same length to contrast LCG and HCG load conditions had some motivation in the literature (Ahmed & de Fockert, 2012; Dalton et al., 2009), so in addition to allaying participants' sense of difficulty in LCG tasks, the change would still constitute a valid manipulation of cognitive load.

Secondly, the SNRs were reduced to two levels: 0 and +8 dB (MD and VH SNR conditions). This was done to enable an increase in the number of DNRs without increasing the duration of the entire probe/flanker task to a point where participant fatigue or boredom became an issue (although this did lead to a reduction in the number probe/flanker trials in each condition) and to keep the task within the recommended time (i.e., under an hour) for speech technology evaluations (Thorpe, 1998). In addition, DNR algorithms are not often optimised to perform below 0 dB (Loizou, 2007) so any comparison between DNRs may not be valid at very low SNRs as they would not be performing optimally.

The final difference between the probe/flanker tasks in experiment III and experiment IV was to require participants to recall the entire probe recall set at the end of each trial. This was to add an 'extra dimension' to the memory component of the task which required participants to remember the recall set in order (a requirement which is not imposed in the probe task) and provide an additional measure of the impact of listening to noise on recall (cf. Rabbitt, 1964; Howard et al., 2010; Sarampalis et al., 2009).

The format of a single trial is shown in figure 4.1. After a 500 ms visual fixation (a white dot in the centre of the screen), the recall set was shown for 2500 ms in both LCG load and HCG load conditions, in order to give participants sufficient time to read the six digits in both

conditions (Lavie et al., 2004). This was followed by a mask for 2500 ms (in both LCG and HCG conditions). The response words for the flanker trial were presented for 1000 ms followed by the 500 ms auditory fixation in the target ear, and the flanker stimulus. Participants had a maximum of 3000 ms to respond and proceeded to the probe recall phase immediately after responding (or after 3000 ms if no response was made). In the probe recall phase a 500 ms probe fixation (a white square in the middle of the screen) was displayed followed by the probe, and participants were given 3000 ms to decide if the probe was present or absent from the recall set. Following this, participants were required to recall all the digits in the recall set in the correct order and input them using the USB keyboard. This last stage was untimed.

Prior to completing this task, the task was explained verbally to the participants using an untimed version of the probe/flanker trial. Following 16 practice trials using clear-speech targets, participants completed 256 trials in four blocks of 64 trials with cognitive load (LCG, HCG), congruency (CON, INC), SNR (MD, VH), DNR (NN, MM, SB, SP), target ear (left, right) and response hand (left, right) fully crossed between trials. Target and distractor words and speakers were selected at random. Within each block, cognitive load and target ear were kept constant. Cognitive load alternated between blocks and the target ear alternated every two blocks. The order of cognitive load and target ear was balanced across participants and the same order was also used the practice trials.

#### 4.2.2.2 Automated operation span

A custom version of the Automated Operation Span (AOSpan Unsworth et al., 2005), written in the PYTHON programming language, was administered to participants on the same desktop computer used in for all tasks (except the cognitive failures questionnaire) and responses were collected using the computer's USB mouse. The procedure for the AOSpan is illustrated in figure 4.2. Participants were required to remember lists of letters (from the set F,H,J,K,L,N,P,Q,R,S,T,Y) in the correct order. For each item in the list, participants were asked to solve a simple mathematical problem (the equation stage) such as  $(1 + 5) \div 3$ , and then indicate whether a candidate answer was correct or not (the decision stage), after which the recall item was displayed. The problems were always of the form  $(X o_1 Y) o_2 Z$ , where  $X$ ,  $Y$  and  $Z$  were all non-identical positive single digit integers (i.e., in the range 1–9 inclusive), and  $o_1$  was either addition  $+$  or subtraction  $-$  and  $o_2$  was either multiplication  $\times$  or division  $\div$ . The



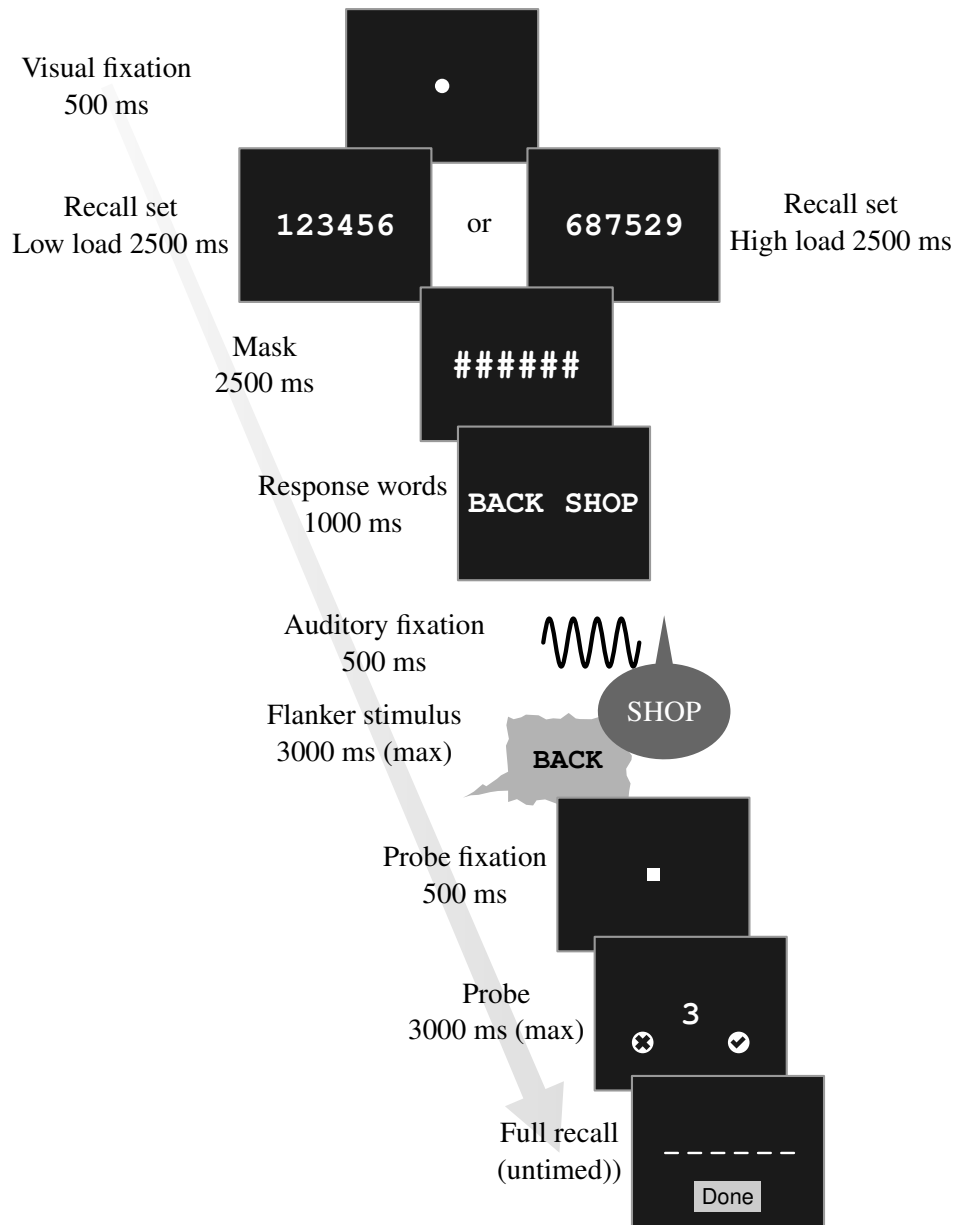


Figure 4.1: Illustration of a single trial in the probe/flanker task.

solution to the problem was always a single digit positive integer, and problems were sampled at random from a pre-generated set of all possible problems, so that no problem was repeated during the AOSpan procedure for a given participant. The candidate answer was correct or incorrect with equal probability.

After all items in the list had been displayed, participants were asked to recall the list in order (recall stage). Participants, were presented with fifteen lists, three lists each of three, four, five, six and seven items, and after recalling each list participants were given feedback on their

recall and math performance. Although, the recall items were always displayed for 500 ms and the recall and feedback stages were untimed, the timing of the other stages (i.e., the equation and decision stages) were set automatically for each participant individually during a training session, to take into account differences in mathematical ability and facility with the computer mouse (cf. Unsworth et al., 2005). Each participants' operation span was measured according to the recommendations of Unsworth et al., p. 501 as the total number of correctly recalled items (i.e., the list length) for lists where serial recall was correct. This provided an integer scale of 0–75.

#### 4.2.2.3 Cognitive Failiures Questionnaire

The Cognitive Failures Questionnaire (CFQ, Broadbent et al., 1982) was administered as an 'Android App' on a Blackberry Playbook. Appendix D shows the questions of the CFQ and describes the scoring procedure.

#### 4.2.2.4 P.835

The ITU P.835 evaluation ("Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm" ITU-T Rec. P.835, 2003, p. 1) is a telecommunications-industry standardised test, specifically aimed at evaluating DNR systems (Hu & Loizou, 2007). Participants are required to make ratings on three aspects of a speech signal: speech distortion, intrusion of background noise, and overall speech quality.

Stimuli for the P.835 task were created using the same the 50 CVC singular nouns (see table 2.2 (p. 44) and six speakers used in the auditory flanker tasks. Words and speakers were selected at random for each trial and padded with 1000 ms of silence, mixed with a random segment of the pre-processed (i.e., amplitude smoothed) babble noise to 0 or +8 dB SNR (LO and HG SNR conditions respectively). They were then processed with the three DNR algorithms (MM, SB, SP conditions) or left without any DNR processing (NN), creating a total of eight audio (SNR × DNR) conditions. Stimuli were presented in mono to both ears (as required by the P.835 standard).

For each stimulus, participants were required to rate the speech distortion, intrusion of background noise, and overall speech quality (SC, BK and OV rating conditions, respectively) on a discrete 5-point Likert scale with the anchors shown in figure 4.3 (p. 125). Participants had a maximum of 10 s to make a response, and could not listen to the stimulus more than

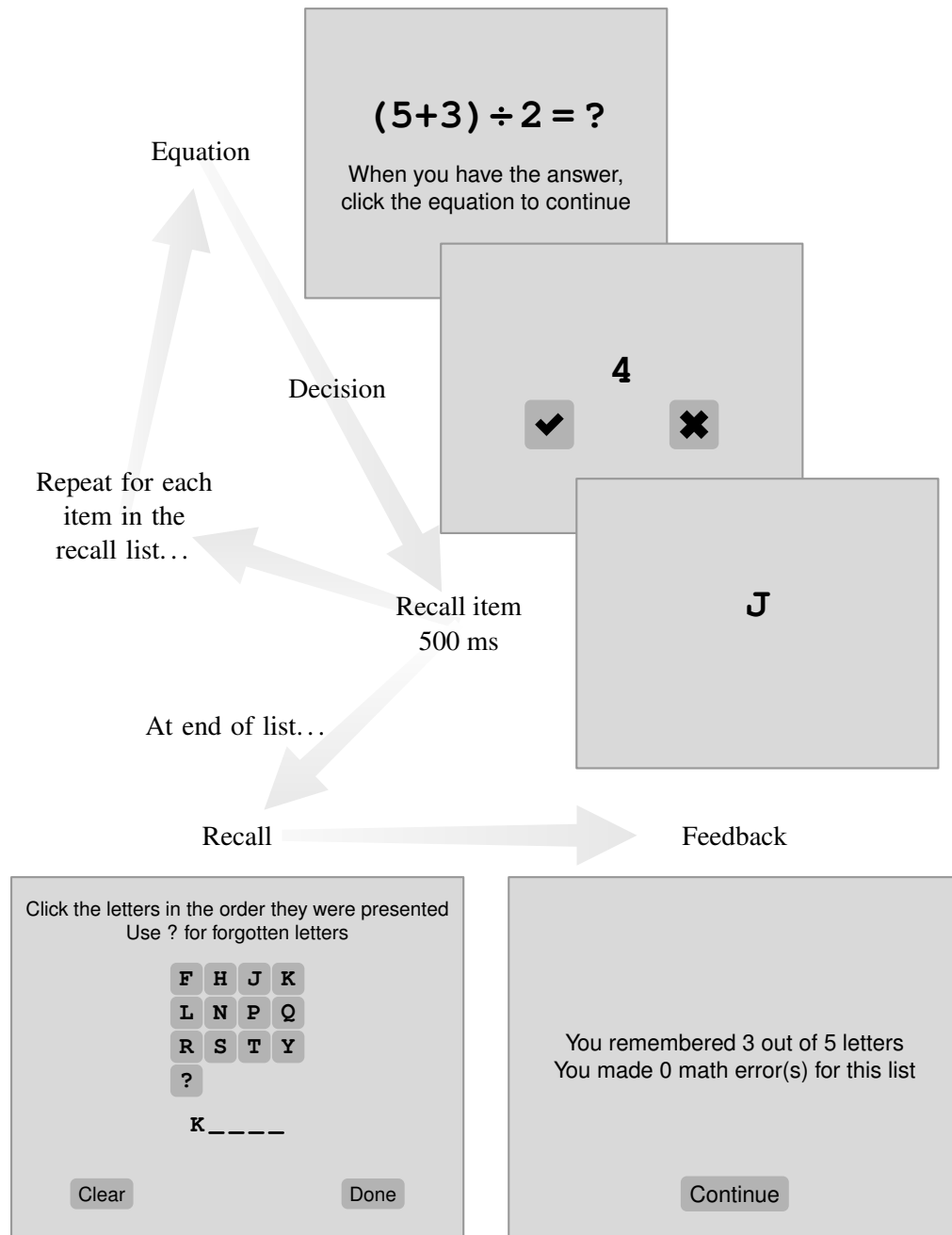


Figure 4.2: Illustration of a single AOSpan trial.

once. The next trial began 500 ms after a response was given (or after 10 s if no response was given). Stimuli were presented in triplets, with each triplet taken from the same audio condition and each stimulus in the triplet requiring either a SC, BK, and OV rating. The P.835 standard recommends collecting only two samples of each rating type per audio condition, however two ratings on five point scale only provides a 'measurement space' of nine points per participant

when averaged. So, in the version of the ratings test used in this experiment participants were required to provide ten ratings for each rating type in each audio condition to improve the precision in the ratings. In total, participants completed 240 ratings, in four blocks of 60 (i.e., 20 stimulus triplets) and the order of SC, BK and OV ratings was randomised between triplets.

#### 4.2.2.5 Intelligibility and listening effort

Intelligibility and listening effort scores were collected simultaneously. Participants were played a sequence of five words, and were required to repeat back each word as they heard it. At the end of each list, participants rated how much effort they believed they made to hear the words in the list. For each list, five words and five speakers were selected at random from the 50 CVC singular nouns (see table 2.2 (p. 44) and six speakers used in the auditory flanker tasks, subject to the constraint that no two words in the list were identical. Each word was padded with 250 ms of silence at each end, and mixed with a random segment of amplitude-smoothed babble noise which was scaled to create 0 or +8 db SNR (LO and HG SNR conditions respectively). The noisy words were then processed with the three DNR algorithms (MM, SB, SP DNR conditions) or left without any DNR processing (NN DNR condition) to create the same eight audio conditions used in the flanker/probe and P.835 tasks. Each word was prefixed with a 500 ms 500 Hz auditory fixation tone, and all the words in a single list were concatenated so that the fixation tones occurred at 2000 ms intervals. Finally, the word list was mixed to the left or right channel of a stereo signal to create the final stimulus.

Eighty word lists were created for each participant, ten lists for each of the eight audio conditions; half the word lists were created for the left target ear, and the other half for the right target ear. A further sixteen word lists were created with no background noise or DNR processing to be used as training. Participants were instructed to respond as soon as they were able to identify the word. Each word list was presented to either the left or right ear (i.e., each sequence of five words was presented to the same ear).

The intelligibility scores and effort ratings were collected using the identical equipment and same listening environment as the other tasks. The presentation of stimuli and recording of responses were controlled by a version of PsytoolKit (Stoet, 2010), customised to permit audio recording. The format for each trial is illustrated in figure 4.4 (p. 127). Prior to each list participants were shown an on screen instruction indicating the target ear (e.g., “Pay attention

Attend only to the SPEECH signal  
The SPEECH SIGNAL in this sample was

- 1 Not distorted
- 2 Slightly distorted
- 3 Somewhat distorted
- 4 Fairly distorted
- 5 Very distorted

Attend only to the BACKGROUND  
The BACKGROUND in this sample was

- 1 Not noticeable
- 2 Slightly noticeable
- 3 Noticeable but not intrusive
- 4 Somewhat intrusive
- 5 Very intrusive

Attend only to the OVERALL SPEECH SAMPLE  
The OVERALL SPEECH SAMPLE in this sample was

- 1 Bad
- 2 Poor
- 3 Fair
- 4 Good
- 5 Excellent

Figure 4.3: Representation of the user interface for the ITU-T P.835 subjective ratings, showing the verbal anchors for 5-point Likert scales used for rating the speech signal (top), background noise (middle) and overall speech sample (bottom).

to your LEFT ear”), followed by a dot in the centre of the screen while the audio stimuli were played and the participants responses were recorded. After the responses had been recorded for each list, a screen showing the listening effort rating screen was displayed and participants were required to rate the effort they had made to hear the words in the list. Participants had ten seconds to indicate their effort rating and responded by clicking on a continuous scale which had numerical and verbal anchors on either side. The verbal anchors were based on the Borg CR-10 scale for the perception of exertion (Borg, 1990, p. 57, figure 4). Participants completed 80 lists in blocks of 20 lists, and between each block had the opportunity to take breaks. Prior to completing the main task, participants completed 16 practice trials (with no background noise) in order to familiarise themselves with the task.

### 4.2.3 Results

#### 4.2.3.1 Auditory flanker task

An analysis of the flanker task was carried out to confirm that drift-interference had been eliminated as it had been in experiment IV. One participant’s results were omitted as their accuracy in the flanker task was consistently more than 3 MAD below the mean accuracy for the other participants (Leys et al., 2013). Trials in which participants failed to make a response in the flanker trial were discarded, as were trials where participants failed to respond correctly in the probe task (in case participants had ‘traded-off’ performance in the flanker trial for performance in the memory-probe trial). This led to a loss of 348 trials (13.35% of the data). The DDM was fitted to the remaining data with the  $v$ ,  $a$  and  $t$  parameters dependent on the full interaction of cognitive load  $\times$  congruency  $\times$  DNR  $\times$  and SNR conditions, with participant’s parameters estimated under a group level variance (Wiecki et al., 2013). The MCMC algorithm was run for 58332 steps and 200 samples were discarded as burn-in, with no thinning. Convergence was assessed visually and with the Geweke statistic which revealed no concerns. Model fit was assessed visually (see figure 4.5 for representative plots) and the fit was considered acceptable (MSE = 0.07).

Table 4.1 shows the group posterior means and standard deviations for the drift-rate parameter and figure 4.6 shows the group posterior drift-rate modes, means and 95% HDIs. There is little to separate the drift-rates, although some SNR trends are indicated with drift-rate appearing to be higher at VH SNR compared to MD SNR. But the overlapping HDIs suggest that few

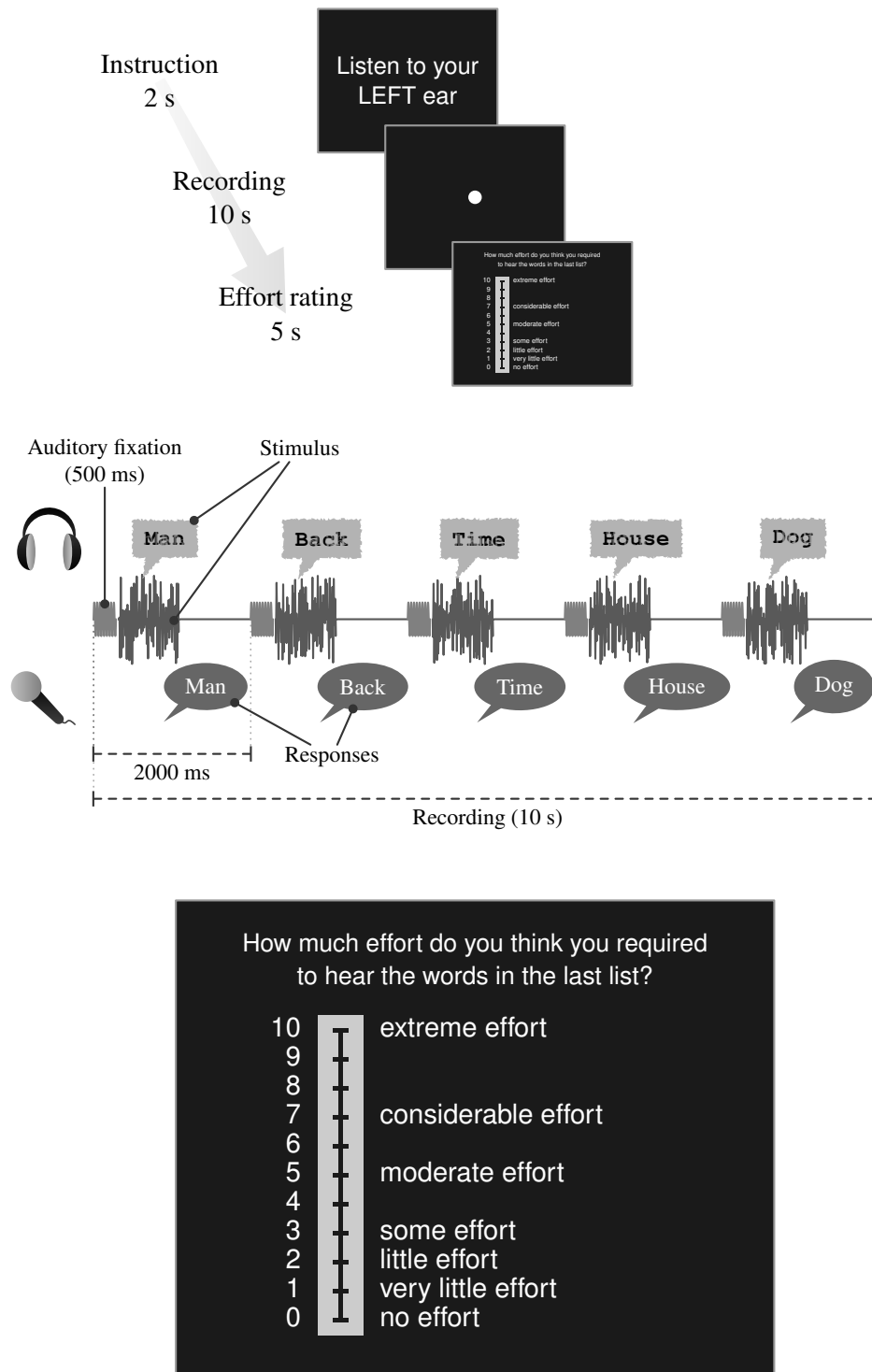


Figure 4.4: The format for a single trial in the intelligibility and listening effort task (top), the presentation of the list of five words in a trial (middle) and an illustration of the rating scale used by participants to rate the effort required to identify the words in the trial (bottom).

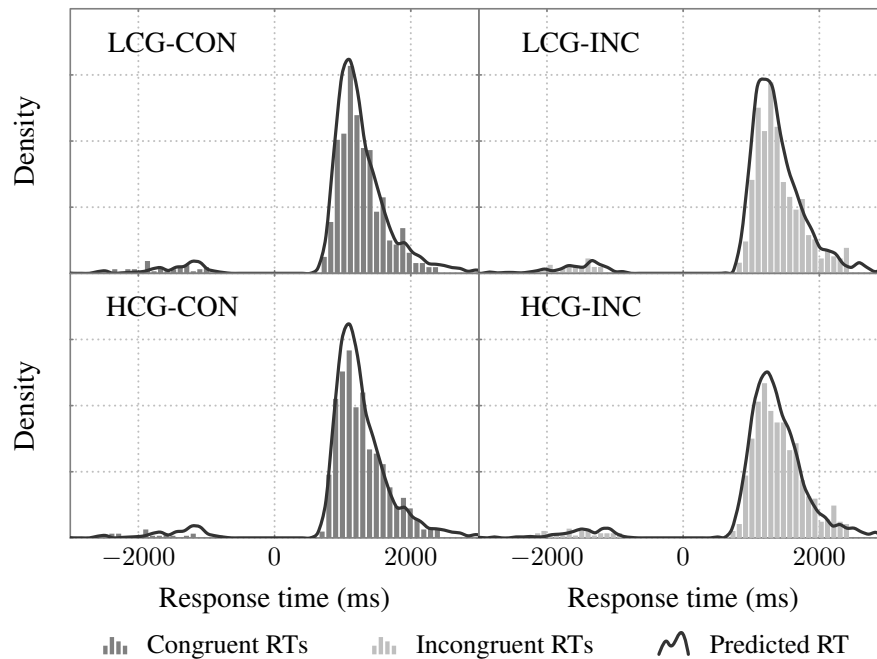


Figure 4.5: Predicted response time (RT) densities overlaid onto the observed RT histograms for the flanker task in congruent (CON) and incongruent (INC) trials under high (HCG) and low (LCG) load. Negative RTs indicate incorrect responses.

(if any) of these differences would be credible. In particular, no obvious differences in drift-rate are indicated between CON or INC trials suggesting that the elimination of drift-interference found in experiment III has been replicated in the current experiment.

Planned comparisons for drift-interference are summarised in table 4.2 (p. 130), showing that drift-interference has been eliminated in all combinations of load, SNR and DNR, with the exception of the HCG-VH-MM condition (comparison 14) where the credible difference only constitutes a credibly small effect size. With the elimination of drift-interference, no further comparisons are warranted. However, it is worth noting that the credible difference in the HCG-VH-MM condition, which was not found in the LCG-VH-MM (i.e., the difference was not found under low cognitive load), cannot be unambiguously attributed to the difference in cognitive load, as a comparison between drift-interference in the HCG-VH-MM and LCG-VH-MM figure 4.7 (p. 131) demonstrated that while the difference approached the margin of credibility, it did not constitute a credible effect size ( $\hat{d} < 0.10$ ).

With an absence of credible drift-interference in all but one of the experimental conditions



		Cognitive load							
		LCG				HCG			
		Congruency				Congruency			
DNR	SNR	CON		INC		CON		INC	
		mean	sd.	mean	sd.	mean	sd.	mean	sd.
MD	NN	1.76	0.82	1.33	0.41	1.80	0.31	1.84	0.65
MD	MM	1.54	0.52	1.95	0.26	1.92	0.38	1.44	0.26
MD	SB	1.37	0.31	1.74	0.39	1.43	0.42	1.80	0.28
MD	SP	1.62	0.31	1.26	0.30	1.63	0.55	1.26	0.60
VH	NN	2.11	0.58	1.84	0.30	2.92	0.58	2.27	0.37
VH	MM	2.21	1.07	2.45	0.81	3.18	0.49	1.65	0.32
VH	SB	2.16	0.39	2.28	0.29	2.52	1.33	2.71	0.50
VH	SP	2.86	0.55	2.30	0.43	2.89	0.85	2.38	0.49

Table 4.1: Posterior means and standard deviations for the flanker drift-rate parameter in all experimental conditions.

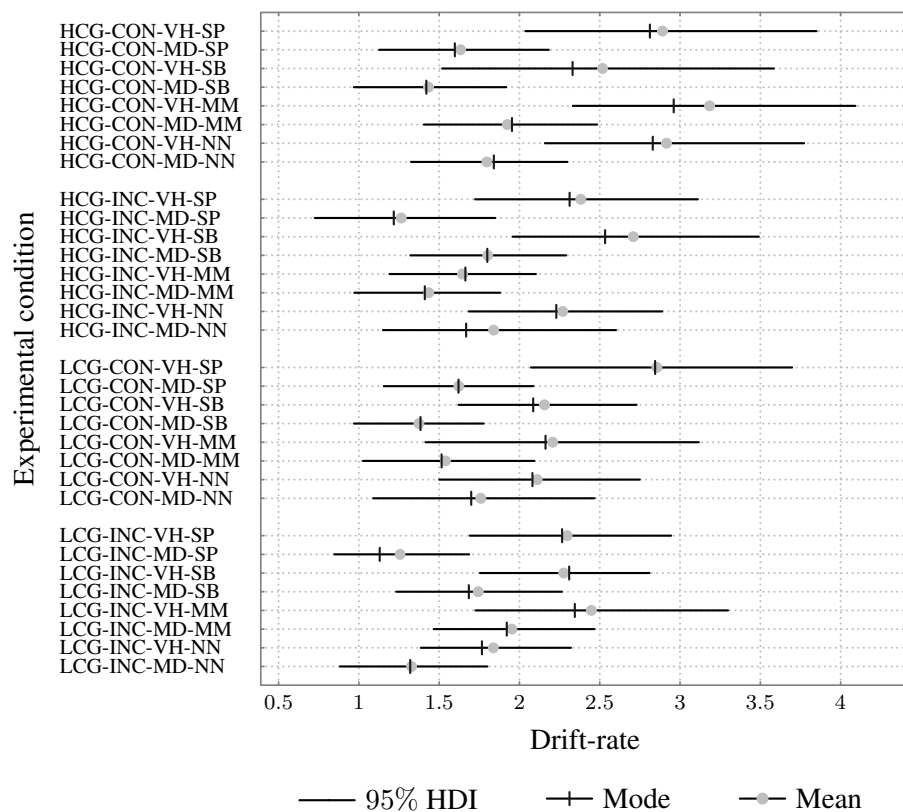


Figure 4.6: Posterior means, modes and 95% HDIs for the drift-rate parameter in the auditory flanker task for all experimental conditions.

Load	Comparison		Mode	Distribution around null	Null outside		Effect size $\hat{d}$
	Congruency	SNR			DNR	HDI	
1. LCG	CON-INC	MD	NN	0.45	14.42% $\leq 0 < 85.58\%$	No	32.26%
2. LCG	CON-INC	MD	MM	-0.40	87.19% $\leq 0 < 12.81\%$	No	23.78%
3. LCG	CON-INC	MD	SB	-0.34	86.87% $\leq 0 < 13.13\%$	No	35.14%
4. LCG	CON-INC	MD	SP	0.37	12.11% $\leq 0 < 87.89\%$	No	24.92%
5. HCG	CON-INC	MD	NN	-0.01	51.93% $\leq 0 < 48.07\%$	No	47.81%
6. HCG	CON-INC	MD	MM	0.46	8.03% $\leq 0 < 91.97\%$	No	19.94%
7. HCG	CON-INC	MD	SB	-0.45	86.61% $\leq 0 < 13.39\%$	No	28.05%
8. HCG	CON-INC	MD	SP	0.34	16.16% $\leq 0 < 83.84\%$	No	31.30%
9. LCG	CON-INC	VH	NN	0.29	24.29% $\leq 0 < 75.71\%$	No	35.08%
10. LCG	CON-INC	VH	MM	-0.14	66.29% $\leq 0 < 33.71\%$	No	41.78%
11. LCG	CON-INC	VH	SB	-0.15	62.90% $\leq 0 < 37.10\%$	No	39.65%
12. LCG	CON-INC	VH	SP	0.52	13.81% $\leq 0 < 86.19\%$	No	20.62%
13. HCG	CON-INC	VH	NN	0.66	9.91% $\leq 0 < 90.09\%$	No	18.42%
14. HCG	CON-INC	VH	MM	1.44	0.08% $\leq 0 < 99.92\%$	Yes	0.00%
15. HCG	CON-INC	VH	SB	-0.33	62.60% $\leq 0 < 37.40\%$	No	42.84%
16. HCG	CON-INC	VN	SP	0.51	19.24% $\leq 0 < 80.76\%$	No	26.32%

Table 4.2: Summaries of the planned comparisons for the change in drift-rate in the auditory flanker task between congruent (CON) and incongruent (INC) trials (i.e., drift-interference) at every level of Cognitive load, SNR and DNR.

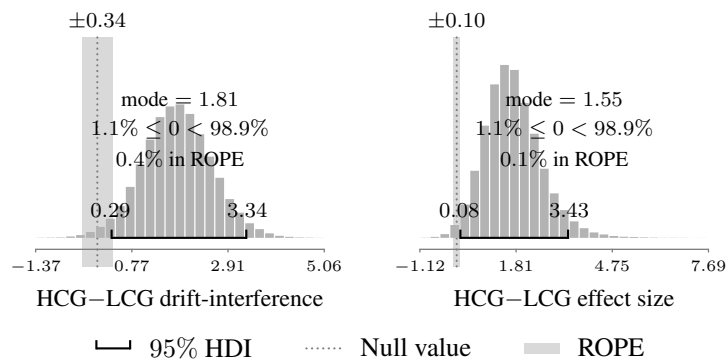


Figure 4.7: Comparison of drift-interference between HCG and LCG load at VH SNR with the MM DNR.

where the effect was only very small, it was considered that there was sufficient evidence to conclude that participants had maintained performance in the flanker task and any impact on performance due to the SNR and DNR conditions would (as in experiment III) be found in the memory probe task.

#### 4.2.3.2 Memory probe task

Trials where participants failed to respond in the memory probe task or responded incorrectly in the listening task were discarded (269 trials, 8.76% of the data). A DDM was fitted to the data with the  $v$ ,  $t$  and  $a$  parameters dependent on the four-way cognitive load  $\times$  congruency  $\times$  SNR  $\times$  DNR interaction, with participant's parameters estimated under a group level variance (Wiecki et al., 2013). The MCMC sampling was run for 58332 steps, with 173 samples discarded as burn-in, and no thinning. Convergence was assessed with a visual inspection of the posterior traces and the Geweke statistic which revealed no concerns. Model was assessed visually (see figure 4.8 for representative plots) and the fit was considered reasonable (MSE = 0.03).

As in experiment III, drift-rates were averaged across CON and INC trials for further analysis. Table 4.3 (p. 133) shows the probe drift-rate posterior means and standard deviations and figure 4.9 (p. 133) illustrates the posterior means, modes and 95% HDIs. No clear trends are apparent across any of the conditions of cognitive load, SNR, or DNR. Table 4.4 (p. 134) shows planned comparisons for the probe drift-rate, revealing no credible differences in drift-

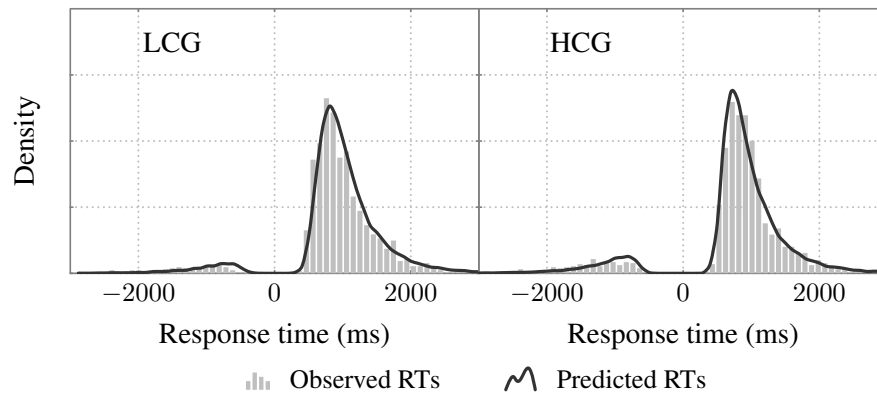


Figure 4.8: Predicted response time (RT) densities overlaid onto the observed RT histograms for the probe task under high (HCG) and low (LCG) load. Negative RTs indicate incorrect responses.

rate between VH and MD SNR in any DNR at LCG load (comparisons 1–4) or HCG load (comparisons 5–8) suggesting that there were no credible effects on the probe task from attending to noisy or de-noised flanker targets, at any SNR with or without DNR processing, under any kind of cognitive load. Furthermore, there were no credible differences in probe drift-rate in trials where the flanker target was DNR-processed speech (MM, SB, SP conditions) compared to trials where the flanker target was left unprocessed (NN), at any SNR or under any cognitive load (comparisons 10–20), although at VH SNR in the LCG load condition the comparisons just failed to reach the margins of credibility (comparisons 12–14).

Taken together, both the flanker task and the probe task failed to yield any credible differences in drift-rate due to cognitive load, DNR, SNR or (in the flanker task) congruency. In addition, the effect found in the probe task in experiment III where increasing the SNR led to an increase in probe drift-rate was not replicated. The inability of the probe/flanker task to measure changes in distraction or to distinguish between different levels of SNR, changes in DNR or cognitive load, demonstrate that in its current form, the flanker task is an inappropriate task for measuring listening effort or evaluating speech technologies. This will be discussed further below.

		Cognitive load			
		LCG		HCG	
DNR	SNR	mean	sd.	mean	sd.
MD	NN	1.62	0.35	1.67	0.65
MD	MM	1.56	0.27	1.62	0.45
MD	SB	1.63	0.24	1.57	0.59
MD	SP	1.46	0.34	1.69	0.84
VH	NN	2.02	0.45	1.55	0.68
VH	MM	1.56	0.36	1.46	0.48
VH	SB	1.39	0.36	1.29	0.34
VH	SP	1.55	0.40	1.53	0.49

Table 4.3: Posterior means and standard deviations for the probe drift-rate parameter averaged across congruent and incongruent trials.

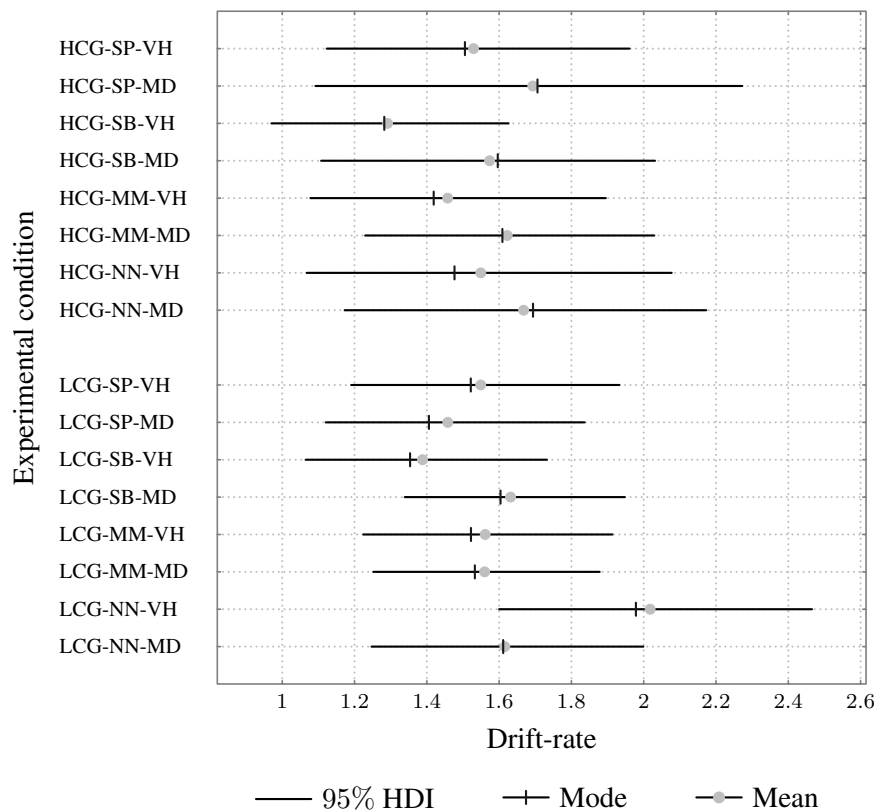


Figure 4.9: Posterior means, modes and 95% HDIs for the drift-rate parameter in the memory probe task averaged across congruent and incongruent trials.

Load	Comparison		Mode	Distribution around null	Null outside HDI	% HDI in ROPE	Effect size $\hat{d}$
	SNR	DNR					
1. LCG	VH-MD	NN	0.43	$7.80\% \leq 0 < 92.20\%$	No	15.80%	
2. LCG	VH-MD	MM	-0.05	$50.21\% \leq 0 < 49.79\%$	No	40.04%	
3. LCG	VH-MD	SB	-0.23	$85.95\% \leq 0 < 14.05\%$	No	24.52%	
4. LCG	VH-MD	SP	0.05	$35.98\% \leq 0 < 64.02\%$	No	32.03%	
5. HCG	VH-MD	NN	-0.09	$63.13\% \leq 0 < 36.87\%$	No	39.72%	
6. HCG	VH-MD	MM	-0.22	$72.31\% \leq 0 < 27.69\%$	No	34.88%	
7. HCG	VH-MD	SB	-0.23	$84.05\% \leq 0 < 15.95\%$	No	26.70%	
8. HCG	VH-MD	SP	-0.07	$67.49\% \leq 0 < 32.51\%$	No	33.77%	
9. LCG	MD	NN-MM	0.08	$41.19\% \leq 0 < 58.81\%$	No	45.26%	
10. LCG	MD	NN-SB	-0.07	$53.79\% \leq 0 < 46.21\%$	No	46.53%	
11. LCG	MD	NN-SP	0.20	$26.73\% \leq 0 < 73.27\%$	No	37.43%	
12. LCG	VH	NN-MM	0.43	$4.76\% \leq 0 < 95.24\%$	No	9.05%	
13. LCG	VH	NN-SB	0.65	$1.08\% \leq 0 < 98.92\%$	Yes	0.92%	
14. LCG	VH	NN-SP	0.43	$5.20\% \leq 0 < 94.80\%$	No	8.79%	
15. HCG	MD	NN-MM	0.07	$44.65\% \leq 0 < 55.35\%$	No	34.75%	
16. HCG	MD	NN-SB	0.13	$39.17\% \leq 0 < 60.83\%$	No	35.00%	
17. HCG	MD	NN-SP	0.10	$52.39\% \leq 0 < 47.61\%$	No	35.76%	
18. HCG	VH	NN-MM	0.08	$38.95\% \leq 0 < 61.05\%$	No	41.57%	
19. HCG	VH	NN-SB	0.27	$19.77\% \leq 0 < 80.23\%$	No	33.15%	
20. HCG	VH	NN-SP	0.07	$47.50\% \leq 0 < 52.50\%$	No	42.43%	

Table 4.4: Summaries of the selected comparisons for the change in drift-rate in the memory-probe task.

DNR	SNR	Cognitive load			
		LCG		HCG	
		mean	sd.	mean	sd.
MD	NN	0.92	0.12	0.67	0.32
MD	MM	0.93	0.07	0.66	0.33
MD	SB	0.94	0.06	0.64	0.35
MD	SP	0.93	0.11	0.56	0.29
VH	NN	0.95	0.07	0.62	0.32
VH	MM	0.93	0.11	0.65	0.32
VH	SB	0.95	0.08	0.60	0.29
VH	SP	0.94	0.08	0.64	0.29

Table 4.5: Prior means and standard deviations for the recall accuracy averaged across congruent and incongruent trials.

#### 4.2.3.3 Recall

Trials where an incorrect response or no response were given in the flanker trial or the probe trial were discarded (491 trials 15.98% of the data). Accuracy was averaged over CON and INC trials and summarised in table 4.5. A Bayesian mixed effects logistic regression was fitted to the data (see section B.6 for the graph and parameter settings) with a three-way cognitive load  $\times$  SNR  $\times$  DNR interaction, with separate intercepts and slopes for each participant. The MCMC sampling was run for 100000 steps with 50000 samples discarded as burn-in. Visual inspection of the MCMC traces suggested some auto-correlation, so the MCMC sampling was re-run and thinned by discarding every second sample. Convergence was assessed visually and using the Geweke statistic revealing no concerns. Model fit was reasonable (MSE = 0.03). Figure 4.10 shows the posterior means, modes and 95% HDIs. The expected distinction between LCG and HCG load conditions is apparent, with the log-odds of recall being higher under LCG load compared to HCG load, but no other clear trends are apparent.

Planned comparisons for recall are shown in table 4.6 (p. 137). As expected the effect of cognitive load was credible (comparison 1) with considerable certainty, showing that the log-odds of correct serial recall in LCG load increases by 2.78 compared to HCG load, corresponding to an increase from 63% accuracy to 92% accuracy. This increase constituted a large effect size ( $\hat{d} = 1.52$ ). In addition, the effect of cognitive load was credible across all SNRs and DNRs (comparisons 2–9), in each case with a high degree of certainty and a large effect size. However, all recall comparisons for trials between SNR levels were non-credible

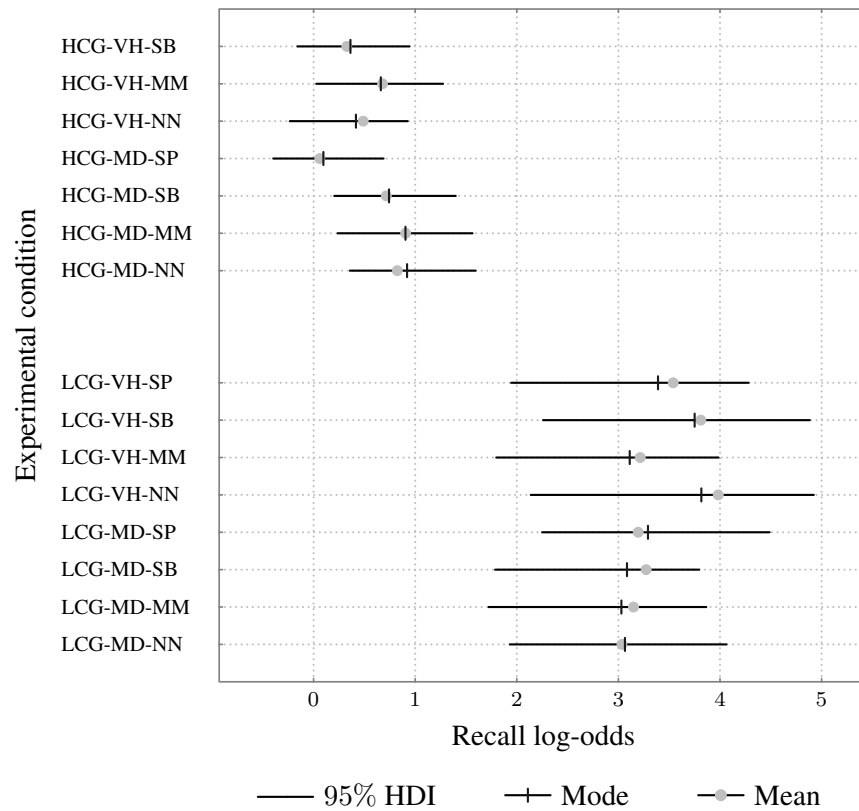


Figure 4.10: Posterior means, modes and 95% HDIs for the recall task, averaged across congruent and incongruent conditions.

(comparisons 10–17). Recall comparisons (not shown in table 4.6) between trials with noisy flanker-targets (NN) and DNR processed flanker-targets (MM, SB, SP) were also non-credible although the the comparison between NN and SP at MD SNR (comparison 18) approached the margins of credibility. However, if this comparison had been accepted as a credible difference it would have barely constituted a small effect.



Load	Comparison		DNR	Mode log-odds	Distribution around null	Null outside HDI	% HDI in ROPE	Effect size $\hat{d}$
	SNR	MD						
1.	HCG-LCG			2.78	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.52
2.	HCG-LCG	MD	NN	2.23	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	0.85
3.	HCG-LCG	MD	MM	2.25	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	0.88
4.	HCG-LCG	MD	SB	2.49	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.07
5.	HCG-LCG	MD	SP	3.05	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.38
6.	HCG-LCG	VH	NN	3.49	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.56
7.	HCG-LCG	VH	MM	2.57	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.02
8.	HCG-LCG	VH	SB	3.38	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.56
9.	HCG-LCG	VH	SP	2.82	0.00% $\leq$ 0 < 100.00%	Yes	0.00%	1.29
10.	LCG	VH-MD	NN	0.80	5.53% $\leq$ 0 < 94.47%	No	9.06%	
11.	LCG	VH-MD	MM	0.05	45.23% $\leq$ 0 < 54.77%	No	33.29%	
12.	LCG	VH-MD	SB	0.62	18.82% $\leq$ 0 < 81.18%	No	21.60%	
13.	LCG	VH-MD	SP	0.48	28.11% $\leq$ 0 < 71.89%	No	26.34%	
14.	HCG	VH-MD	NN	-0.45	87.39% $\leq$ 0 < 12.61%	No	32.39%	
15.	HCG	VH-MD	MM	-0.14	76.91% $\leq$ 0 < 23.09%	No	42.77%	
16.	HCG	VH-MD	SB	-0.46	90.47% $\leq$ 0 < 9.53%	No	26.43%	
17.	HCG	VH-MD	SP	0.60	3.63% $\leq$ 0 < 96.37%	No	15.03%	
18.	HCG	MD	NN-SP	0.82	0.42% $\leq$ 0 < 99.58%	Yes	1.14%	0.11

Table 4.6: Summaries of the selected comparisons for full recall in the probe flanker task. With the exclusion of comparison 18, effect sizes are shown only for credible comparisons.

Participant	Span	Math accuracy %	CFQ
1	68	97	43
2	63	100	70
3	54	99	68
4	22	96	52
5	3	92	38
6	48	97	71
7	62	97	45
8	55	99	60
9	75	97	57
10	54	97	52
11	61	97	44
12	48	97	62

Table 4.7: AOSpan and CFQ scores for each participant in experiment IV.

#### 4.2.3.4 AOSpan and CFQ

Table 4.7 shows the AOSpan scores along with the CFQ scores. A considerable range of AOSpan span scores are shown but importantly the math accuracy for all participants is greater than 85% which is the criterion for adequate math performance (Unsworth et al., 2005) demonstrating that participants did not trade off math performance for recall performance. Figure 4.11 shows the posterior correlation between the standardised AOSpan and CFQ scores (see section B.3 for the graph and hyperprior settings). To form the posterior correlation, the MCMC sampling process was run for 200000 iterations with 50000 samples discarded as burn-in. Visual inspection of the posterior samples in pilot runs of the MCMC sampling revealed a degree of auto-correlation in the samples so thinning was used with every second sample discarded. Visual checks for convergence and the Geweke statistic revealed no concerns.

The most credible correlation between the AOSpan and CFQ scores was weakly negative  $\rho = -0.23$ . However, it must be noted that the null value 0 (representing no correlation) was one of the 95% most credible values of the  $\rho$  parameter so no strong conclusions could be drawn. In addition, previous research has reported no correlation between CFQ scores and AOSpan scores (McVay & Kane, 2009). This may be because the AOSpan reflects the control of attention in a more fine grained manner, whereas the CFQ reflects lapses of attention which are so severe they produce consequences that are noticeable by the individual (cf. Cheyne, Carriere & Smilek, 2006).

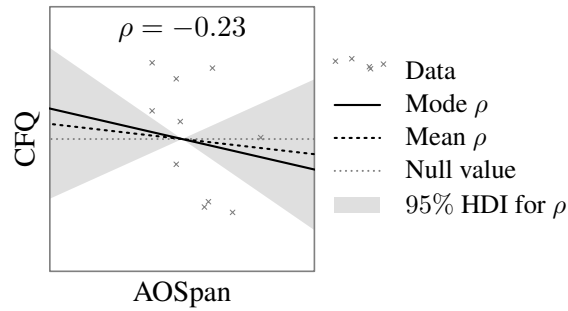


Figure 4.11: Posterior correlation for scores on the automated operation span (AOSpan) and cognitive failures questionnaire (CFQ) scores. The data were standardised (by converting to z-scores) for the graph.

SNR	DNR	P.835					
		SD		BK		OV	
		mean	sd.	mean	sd.	mean	sd.
MD	NN	2.58	0.59	5.27	0.55	3.57	0.49
	MM	4.31	0.56	4.93	0.33	3.10	0.28
	SB	4.09	0.73	5.07	0.34	3.05	0.32
	SP	4.91	0.41	5.27	0.36	2.77	0.41
VH	NN	2.18	0.20	4.24	0.55	4.58	0.45
	MM	2.79	0.23	3.85	0.37	4.29	0.31
	SB	2.85	0.35	3.91	0.48	4.26	0.51
	SP	2.93	0.30	4.27	0.40	4.08	0.40

Table 4.8: Mean and standard deviation ratings for speech distortion (SC), background intrusiveness (BK) and overall speech quality (OV) from the P.835 speech technology evaluation task.

#### 4.2.3.5 P.835

Trials where participants failed to make a response were discarded (1 trial, < 1% of the responses). Table 4.8 shows the group mean and standard deviation ratings for each of the P.835 rating scales and figure 4.12 illustrates the distribution of ratings in each SNR and DNR condition.

#### Speech distortion

Table 4.8 and figure 4.12 suggest that DNR processed speech (MM, SB and SP DNR conditions) was considered more distorted compared to unprocessed speech (NN) and this distortion was considered worse at lower (MD) SNR compared to higher (VH) SNR. However,

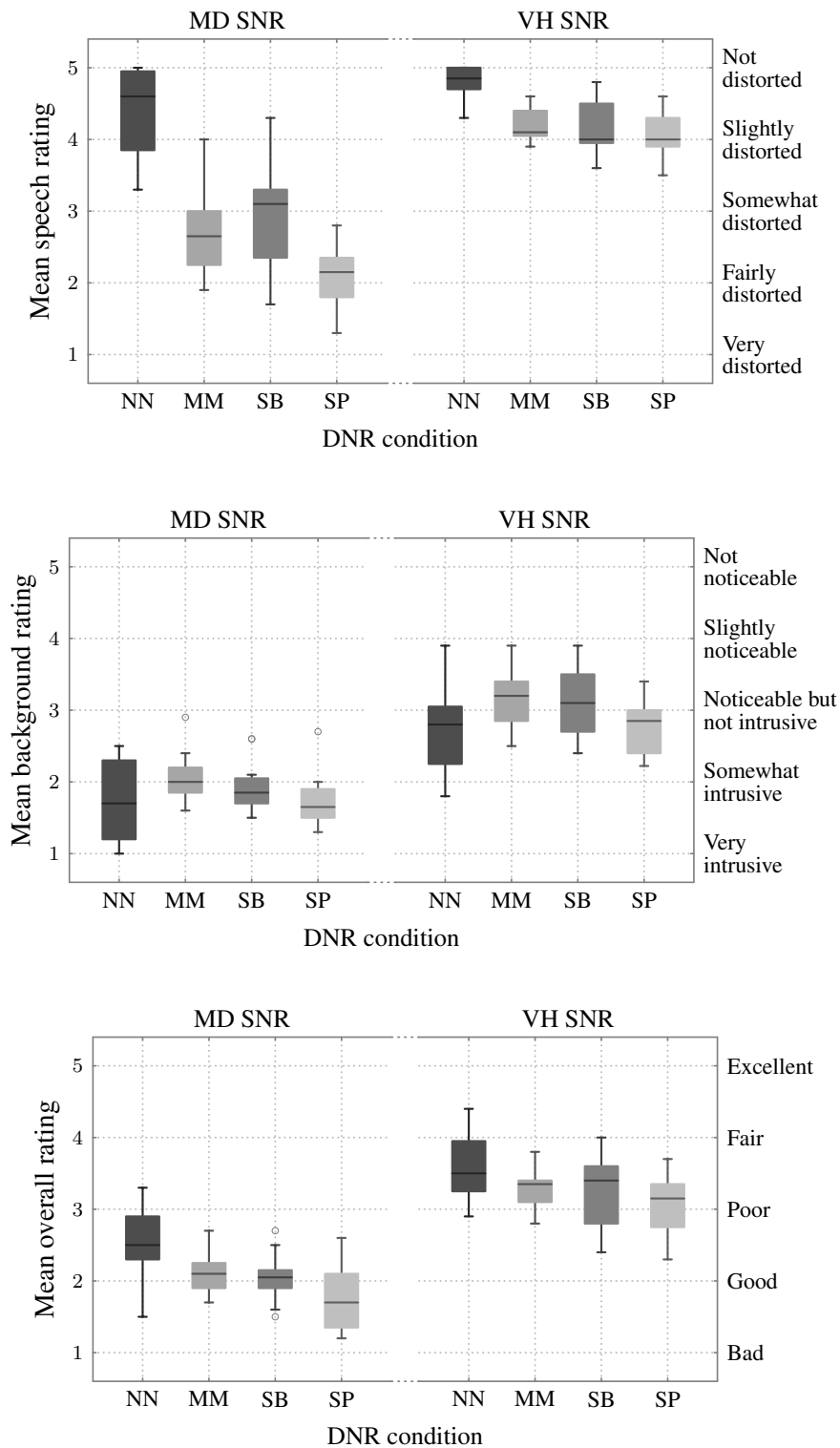


Figure 4.12: Box-plots of the ITU-T P.835 speech quality ratings for speech at 0 dB (MD) and +8 dB (VH) with DNR processing (MM, SB, SP) or without DNR processing (NN). The ratings shown are speech distortion (top), background intrusiveness (middle) and overall speech quality (bottom). Verbal anchors for each rating scale are shown on the right.

there is little to suggest that speech distortion ratings differed between DNR algorithms at each level of SNR. To investigate these observations, a Bayesian ANOVA was fitted to the speech distortion ratings (see figure B.2 for the graph specification). The MCMC algorithm was run for 100000 steps, with a thinning factor of 2 (i.e., every second sample discarded to reduce auto-correlation) and 50000 samples discarded as burn-in. Convergence was assessed with a visual inspection of the chains and the Geweke statistic which revealed no concerns. Model fit was more than adequate ( $MSE \ll 0.01$ ).

Table 4.9 (p. 143) summarises the relevant post-hoc comparisons. The average ratings were credibly different with considerable certainty at VH SNR compared to MD (comparison 1 in table 4.9), with VH speech rating less distorted (1.46 units on the speech distortion scale, a moderately large effect  $\hat{d} = 0.39$ ). This difference was considered not to be due to changes in SNR *per se* (i.e., participants were not confusing energetic masking with distortion) as there were no credible differences in ratings in the NN condition between the SNRs (comparison 2).

There were credible differences between ratings for DNR processed speech compared to unprocessed speech in the MD condition (comparison 3), with distortion rated worse for the DNR processed conditions (MM, SB, SP) compared to the unprocessed condition (NN), a difference of 1.85 on the speech distortion rating scale which constituted a moderate effect size ( $\hat{d} = 0.30$ ) with considerable certainty. Although ratings for DNR processed speech were worse compared to unprocessed speech in the VH condition, this was a very small change (0.69 units on the speech distortion scale) which was not credible (comparison 4); the difference between DNR-processed and processed speech was credibly different between MD and VH SNRs (comparison 5), albeit with a small effect size ( $\hat{d} = 0.12$ ).

There were credible differences in speech quality ratings for all of the DNRs compared individually to the NN condition at MD SNR (comparisons 6, 7 and 8) with all the algorithms judged to have increased the speech distortion relative to the NN condition, up to over 2 points on the distortion rating scale. The differences were apparent with considerable certainty and all differences constituted credible effect sizes.

There were no credible differences for the individual DNRs compared to the NN condition at VH SNR (comparisons 9, 10 and 11). Furthermore, no differences were found in the speech distortion ratings between any of the DNR algorithms at MD SNR (comparisons 12, 13 and 14)

or VH SNR (comparisons 15, 16, 17) and, in general, there was considerable certainty in this lack-of-difference (i.e., the percentage of the HDI in the ROPE was high), with the exception of the SB–SP comparison at MD SNR (comparison 14) where the difference of 0.83 (i.e., SP speech was considered less distorted than SB speech) only marginally failed to reach credibility.

### Background intrusion

Table 4.8 (p. 139) and figure 4.12 (p. 140) suggest that (as might be expected) background noise in the MD SNR condition (0 dB SNR) was considered slightly more intrusive than in the VH SNR condition (+8 dB SNR). However, no clear patterns are suggested between the DNR conditions. To investigate these observations, a Bayesian ANOVA was fitted to the speech distortion ratings (see figure B.2 for the graph specification). The MCMC algorithm was run for 100000 steps, with a thinning factor of 2 (i.e., every second sample discarded to reduce auto-correlation) and 50000 samples discarded as burn-in. Convergence was assessed with a visual inspection of the chains and the Geweke statistic which revealed no concerns. Model fit was adequate (MSE = 0.03). Table 4.10 (p. 144) summarises the relevant post-hoc comparisons.

There was a credible difference between average ratings in MD and VH SNR conditions (comparison 1, table 4.10) with VH speech being rated an average of 1.11 units (on the intrusiveness scale) less intrusive than MD speech. However, this only constituted a small effect ( $\hat{d} = 0.16$ ). No other comparisons were credible, so there were no credible differences in ratings between DNR-processed speech compared to unprocessed speech at MD SNR (comparison 2, table 4.10) or VH SNR (comparison 3). Furthermore, there were no credible differences in ratings for any of the DNRs compared individually to the NN condition, at MD SNR (comparisons 4, 5 and 6) or VH SNR (comparisons 7, 8 and 9), and no credible differences in between the different DNRs at MD SNR (comparisons 10, 11 and 12) or VH SNR (comparisons 13, 14 and 15). With the exception of the MM–SB comparison at MD SNR, the lack of credibility in these comparisons was very certain.

	SNR	Comparison DNR	Mode	Distribution around null	Null outside HDI	% HDI in ROPE	Effect size $\hat{d}$
1.	MD-VH	$\frac{1}{4}(\text{NN} + \text{MM} + \text{SB} + \text{SP})$	1.46	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	0.39
2.	MD-VH	NN	0.40	$0.00\% \leq 0 < 100.00\%$	Yes	94.99%	
3.	MD	$\frac{1}{3}(\text{MM} + \text{SB} + \text{SP}) - \text{NN}$	1.85	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	0.30
4.	VH	$\frac{1}{3}(\text{MM} + \text{SB} + \text{SP}) - \text{NN}$	0.69	$0.00\% \leq 0 < 100.00\%$	Yes	15.97%	
5.	MD-VH	$\frac{1}{5}(\text{MM} + \text{SB} + \text{SP}) - \text{NN}$	1.19	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	0.12
6.	MD	NN - MM	-1.75	$100.00\% \leq 0 < 0.00\%$	Yes	0.00%	0.22
7.	MD	NN - SB	-1.51	$100.00\% \leq 0 < 0.00\%$	Yes	0.00%	0.19
8.	MD	NN - SP	-2.34	$100.00\% \leq 0 < 0.00\%$	Yes	0.00%	0.31
9.	VH	NN - MM	-0.64	$100.00\% \leq 0 < 0.00\%$	Yes	89.75%	
10.	VH	NN - SB	-0.67	$100.00\% \leq 0 < 0.00\%$	Yes	80.02%	
11.	VH	NN - SP	-0.73	$100.00\% \leq 0 < 0.00\%$	Yes	44.74%	
12.	MD	MM - SP	0.58	$0.00\% \leq 0 < 100.00\%$	Yes	89.22%	
13.	MD	MM - SB	-0.21	$98.77\% \leq 0 < 1.23\%$	Yes	94.99%	
14.	MD	SB - SP	0.83	$0.00\% \leq 0 < 100.00\%$	Yes	13.24%	
15.	VH	MM - SP	0.58	$0.00\% \leq 0 < 100.00\%$	Yes	89.22%	
16.	VH	MM - SB	0.07	$21.73\% \leq 0 < 78.27\%$	No	94.99%	
17.	VH	SB - SP	0.10	$18.50\% \leq 0 < 81.50\%$	No	95.00%	

Table 4.9: Summaries of the post-hoc posterior comparisons for the speech distortion ratings in the P.835 task.

SNR	Comparison		Mode	Distribution around null	Null outside HDI	% HDI in ROPE	Effect size $\hat{d}$
	SNR	DNR					
1. MD-VH	$\frac{1}{4}$ (NN + MM + SB + SP)		1.11	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	0.16
2. MD	$\frac{1}{3}$ (MM + SB + SP) - NN		0.00	$69.87\% \leq 0 < 30.13\%$	No	90.08%	
3. VH	$\frac{1}{3}$ (MM + SB + SP) - NN		-0.39	$84.97\% \leq 0 < 15.03\%$	No	70.46%	
4. MD	NN - MM		0.45	$12.03\% \leq 0 < 87.97\%$	No	75.16%	
5. MD	NN - SB		-0.15	$71.30\% \leq 0 < 28.70\%$	No	95.00%	
6. MD	NN - SP		0.03	$41.98\% \leq 0 < 58.02\%$	No	94.31%	
7. VH	NN - MM		0.02	$20.03\% \leq 0 < 79.97\%$	No	81.69%	
8. VH	NN - SB		0.86	$6.57\% \leq 0 < 93.43\%$	No	54.16%	
9. VH	NN - SP		0.29	$38.37\% \leq 0 < 61.63\%$	No	93.02%	
10. MD	MM - SP		0.42	$16.03\% \leq 0 < 83.97\%$	No	79.16%	
11. MD	MM - SB		0.92	$7.05\% \leq 0 < 92.95\%$	No	58.72%	
12. MD	SB - SP		-0.51	$75.00\% \leq 0 < 25.00\%$	No	91.20%	
13. VH	MM - SP		0.42	$16.03\% \leq 0 < 83.97\%$	No	79.16%	
14. VH	MM - SB		-0.83	$78.65\% \leq 0 < 21.35\%$	No	83.59%	
15. VH	SB - SP		0.57	$4.75\% \leq 0 < 95.25\%$	No	75.28%	

Table 4.10: Summaries of the post-hoc posterior comparisons for the background intrusiveness ratings in the P.835 task.



SNR	Comparison DNR	Mode	Distribution around null	Null outside HDI	% HDI in ROPE	Effect size $\hat{d}$
1. MD-VH	$\frac{1}{4}(\text{NN} + \text{MM} + \text{SB} + \text{SP})$	-1.45	$100.00\% \leq 0 < 0.00\%$	Yes	0.00%	0.83
2. MD	$\frac{1}{3}(\text{MM} + \text{SB} + \text{SP}) - \text{NN}$	-0.72	$93.16\% \leq 0 < 6.84\%$	No	13.65%	
3. VH	$\frac{1}{3}(\text{MM} + \text{SB} + \text{SP}) - \text{NN}$	0.52	$34.77\% \leq 0 < 65.23\%$	No	30.84%	
4. MD	NN - MM	1.01	$3.19\% \leq 0 < 96.81\%$	Yes	11.22%	
5. MD	NN - SB	-0.17	$29.99\% \leq 0 < 70.01\%$	No	29.60%	
6. MD	NN - SP	1.00	$7.07\% \leq 0 < 92.93\%$	No	12.93%	
7. VH	NN - MM	0.89	$36.48\% \leq 0 < 63.52\%$	No	35.63%	
8. VH	NN - SB	0.13	$39.59\% \leq 0 < 60.41\%$	No	43.58%	
9. VH	NN - SP	-1.07	$90.56\% \leq 0 < 9.44\%$	No	14.64%	
10. MD	MM - SP	0.14	$42.51\% \leq 0 < 57.49\%$	No	37.92%	
11. MD	MM - SB	0.97	$18.42\% \leq 0 < 81.58\%$	No	19.17%	
12. MD	SB - SP	-0.50	$71.43\% \leq 0 < 28.57\%$	No	31.66%	
13. VH	MM - SP	0.14	$42.51\% \leq 0 < 57.49\%$	No	37.92%	
14. VH	MM - SB	0.02	$49.95\% \leq 0 < 50.05\%$	No	45.93%	
15. VH	SB - SP	1.02	$0.56\% \leq 0 < 99.44\%$	Yes	5.84%	

Table 4.11: Summaries of the post-hoc posterior comparisons for the overall speech quality ratings in the P.835 task.

### Overall speech quality

Table 4.8 (p. 139) and figure 4.12 (p. 140) suggest that ratings are higher in the VH condition compared to the MD condition. There is also the suggestion that DNR-processed speech (MM, SB and SP conditions) is considered lower speech quality than unprocessed speech (the NN condition), although these differences appear to be minor in the VH SNR condition. To investigate these observations, a Bayesian ANOVA was fitted to the speech distortion ratings (see figure B.2 for the graph specification). The MCMC algorithm was run for 100000 steps, with a thinning factor of 2 (i.e., every second sample discarded to reduce auto-correlation) and 50000 samples discarded as burn-in. Convergence was assessed with a visual inspection of the chains and the Geweke statistic which revealed no concerns. Model fit was adequate ( $MSE < 0.01$ ). Table 4.11 (p. 145) summarises the relevant post-hoc comparisons.

There was a highly credible difference between the average speech quality ratings in the VH and MD SNR conditions (comparison 1, table 4.11), with ratings being higher in the VH condition (1.45 units on the speech quality scale). This effect had high certainty and constituted a large effect size ( $\hat{d} = 0.83$ ) and suggests that overall speech quality ratings may have been reflected in the level of background noise. None of the other comparisons were credible however. So, there were no credible differences in overall speech quality between DNR-processed speech compared to unprocessed speech at MD SNR (comparison 2, table 4.11) or VH SNR (comparison 3). Furthermore, there were no credible differences in overall speech quality for any of the DNRs compared individually to the NN condition at MD SNR (comparisons 4, 5 and 6) or VH SNR (comparisons 7, 8 and 9) and no credible overall speech quality differences between the individual DNR algorithms at MD SNR (comparisons 10, 11 and 12) or VH SNR (comparisons 13, 14 and 15).

#### 4.2.3.6 Intelligibility

Table 4.12 shows the group mean and standard deviation intelligibility scores, and figure 4.13 summarises the observed intelligibility scores. Intelligibility is generally high although there is clear distinction between MD and VH SNRs. In addition, the medians in the box-plots suggest that there may be differences in intelligibility in the different DNR conditions.

SNR	DNR	Intelligibility		Effort	
		mean	sd.	mean	sd.
MD	NN	3.60	0.46	4.89	1.55
	MM	3.13	0.55	5.15	1.65
	SB	2.91	0.34	5.61	1.14
	SP	2.69	0.60	6.22	1.80
VH	NN	4.59	0.19	1.64	0.89
	MM	4.47	0.23	2.23	1.55
	SB	4.38	0.31	2.01	0.95
	SP	4.36	0.31	2.60	1.11

Table 4.12: Mean and standard deviations for the intelligibility scores and listening effort ratings.

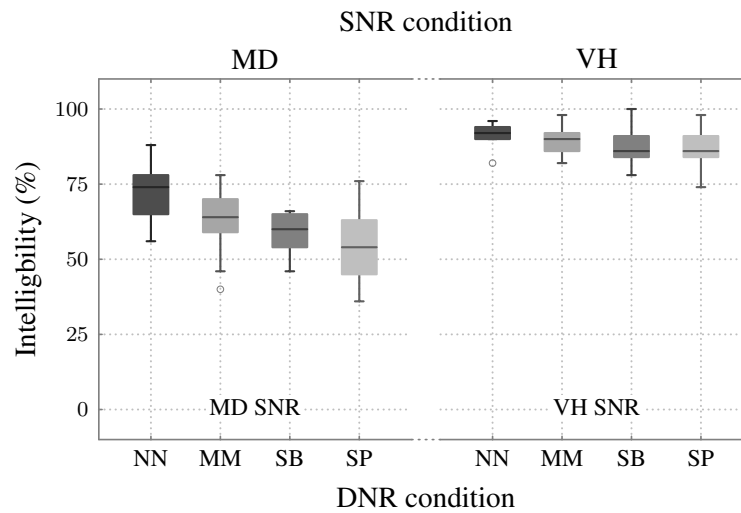


Figure 4.13: Box-plots of the intelligibility scores for speech at 0 dB (MD) and +8 dB (VH) with DNR processing (MM, SB, SP) or without (NN).

SNR	Comparison DNR	Mode log odds	Distribution around null	Null outside HDI	% HDI in ROPE	Effect size $\hat{d}$	
1.	VH-MD	$\frac{1}{4}(\text{NN} + \text{MM} + \text{SB} + \text{SP})$	1.68	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	6.68
2.	MD	$\text{NN} - \frac{1}{3}(\text{MM} + \text{SB} + \text{SP})$	0.62	$0.03\% \leq 0 < 99.97\%$	Yes	0.00%	0.71
3.	VH	$\text{NN} - \frac{1}{3}(\text{MM} + \text{SB} + \text{SP})$	0.40	$1.27\% \leq 0 < 98.73\%$	Yes	0.00%	0.13
4.	VH-MD	$\text{NN} - \frac{1}{3}(\text{MM} + \text{SB} + \text{SP})$	-0.13	$80.72\% \leq 0 < 19.28\%$	No	20.32%	
5.	MD	NN - MM	0.48	$0.06\% \leq 0 < 99.94\%$	Yes	0.00%	0.40
6.	MD	NN - SB	0.68	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	0.96
7.	MD	NN - SP	0.79	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	1.37
8.	MD	MM - SB	-0.25	$93.90\% \leq 0 < 6.10\%$	No	11.34%	
9.	MD	MM - SP	-0.37	$99.87\% \leq 0 < 0.13\%$	Yes	0.00%	0.35
10.	MD	SB - SP	-0.22	$91.46\% \leq 0 < 8.54\%$	No	14.71%	
11.	VH	NN - MM	0.30	$12.16\% \leq 0 < 87.84\%$	No	13.06%	
12.	VH	NN - SB	0.45	$0.43\% \leq 0 < 99.57\%$	Yes	0.00%	0.19
13.	VH	NN - SP	0.44	$1.02\% \leq 0 < 98.98\%$	Yes	0.00%	0.25
14.	VH	SB - SP	-0.03	$56.41\% \leq 0 < 43.59\%$	No	23.75%	
15.	VH-MD	NN	1.42	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	2.89

Table 4.13: Summary of the posterior comparisons made for the intelligibility scores. Effect sizes are only shown for credible differences with considerable certainty.

A Bayesian mixed-effects logistic regression model was fitted to the data (see section B.6 for the graphical model and initial parameter settings) regressing the intelligibility scores on to the two-way SNR  $\times$  DNR interaction with separate intercepts and random slopes for each participant. The MCMC sampling was run for 100000 steps, with 25000 samples discarded as burn-in and a thinning factor of 2 (i.e., every second sample was discarded to avoid auto-correlation in the posterior samples). Convergence was assessed with visual inspection of the traces and the Geweke statistic, revealing no concerns. Model fit was adequate MSE = 0.15.

Table 4.13 (p. 148) shows the parameters of various post-hoc posterior comparisons. In particular, table 4.13 shows the expected effect of SNR (comparison 1) with the mode log-odds of a correct response increasing by 1.68 between the MD and VH SNR, which corresponded to an increase in accuracy from 62.22% in the MD condition to 89.11% in the VH condition and constituted a large credible effect size ( $\hat{d} = 6.68$ ). In addition, intelligibility was credibly worse for DNR processed speech in general (i.e., averaged across the MM, SB, and SP DNR conditions) compared to speech without DNR processing (NN) in both the MD SNR condition (comparison 2) where the difference constituted a moderately large effect size ( $\hat{d} = 0.71$ ) and in the VH SNR condition (comparison 3) where the difference only constituted a small effect size ( $\hat{d} = 0.13$ ).

DNR comparisons at MD SNR (comparisons 5, 6 and 7), revealed that intelligibility was worse for speech processed with the individual DNR algorithms compared to speech with no DNR processing. In particular, the SP algorithm decreased the log-odds of correct identification by 0.79, which corresponded to a decrease from 72.48% to 52.90% (where participants were essentially performing at chance). There was considerable certainty in these differences, and effect sizes ranged from medium to large, providing strong evidence that the use of DNR algorithms was making intelligibility worse at MD SNR. However, differences between the individual DNRs (comparisons 8, 9 and 10) at MD SNR showed only that intelligibility was credibly different between the MM and SP DNR algorithms ( $\hat{d} = 0.35$ ). Similar comparisons at VH SNR (comparisons 11, 12 and 13) revealed smaller differences in intelligibility between individual MM, SB and SP DNR algorithms and the NN condition. These differences were only credible for the SB and SP algorithms, constituting moderate to large effect sizes in each case ( $\hat{d} = 0.46$  and  $\hat{d} = 0.93$ , respectively). However, there was no credible difference in intelligibility between the SB and SP algorithms (comparison 14).

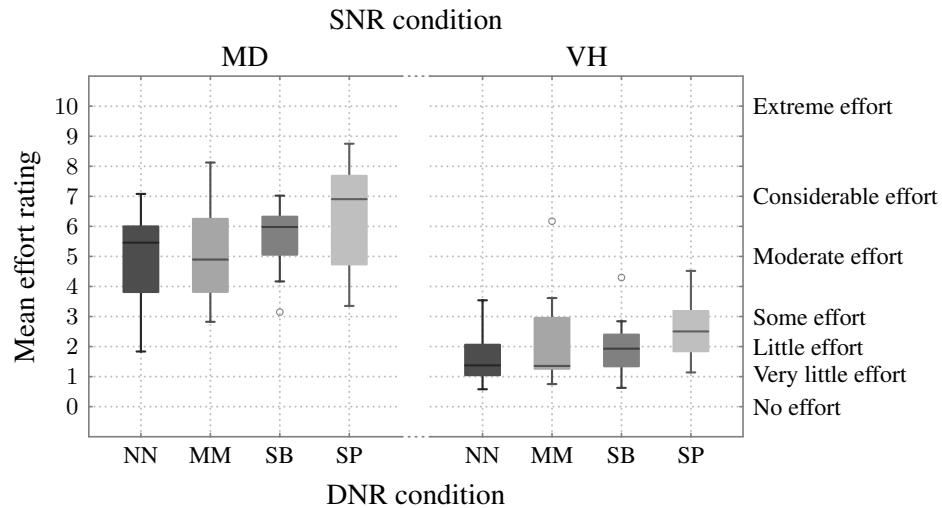


Figure 4.14: Box-plots of the perceived effort ratings for speech at 0 dB SNR (MD) and +8 dB SNR (VH) with DNR processing (MM, SB, SP) or without (NN).

Taken together, the intelligibility results have demonstrated an unsurprising drop in speech intelligibility as the SNR decreases but a surprising drop in intelligibility for DNR processed speech compared to unprocessed speech. In particular, at low SNRs the application of DNR made credible reductions in intelligibility and in one case (the SP condition) reduced intelligibility to almost chance levels. In addition, on the basis of the current data the intelligibility test is able to distinguish between DNR processed and unprocessed speech, with greater distinctions found for lower SNRs.

#### 4.2.3.7 Listening effort

Table 4.12 (p. 147) shows the group mean and standard deviation effort ratings, and figure 4.14 summarises the observed ratings. Effort ratings appear to be higher for the MD SNR condition compared to the VH SNR condition, but there appear to be little differences between the DNR conditions at each level of SNR.

A two-way SNR  $\times$  DNR ‘repeated-measures’ Bayesian ANOVA (Kruschke, 2010a) was carried out on the effort ratings (see section B.4 for the graph and initial parameter settings). The MCMC chain was run for 100000 steps with 50000 samples discarded as burn in. A visual inspection of the traces suggested some auto-correlation so the trace was ‘thinned’ by discarding every second sample. Convergence was assessed by visual assessment of the chains and the

Geweke statistic which revealed no concerns and model fit was acceptable ( $MSE \ll 0.01$ ).

Table 4.14 summarises the key post-hoc posterior comparisons for perceived listening effort. Perceived effort ratings averaged over the DNR conditions were credibly higher in MD SNR trials compared to VH SNR trials with a large difference between ratings (3.37 units on the effort rating scale, with an effect size  $\hat{d} = .83$ , see table 4.14, comparison 1), demonstrating with considerable certainty that perceived effort ratings were considerably higher at 0 dB SNR compared to +8 dB SNR. However, as comparison 2 and comparison 3 show, the differences between the ratings averaged over the DNR algorithms (the MM, SB and SP conditions) and no DNR processing (NN) were less certain, as although perceived effort ratings for the DNR algorithms were generally lower, in both MD and VH SNR conditions, the 95% HDI was overlapping with the ROPE in both MD and VH SNR conditions. The uncertainty regarding the differences was greater in the MD condition than in the VH condition (i.e., the percentage of the HDI overlapping the ROPE was nearer 50% in the MD condition).

Individual DNR comparisons are also summarised in table 4.14 for VH SNR (comparisons 4, 5 and 6) and MD SNR (comparisons 7, 8 and 9). There is considerable certainty in the lack of a credible difference in effort ratings for the NN–MM comparison in both MD and VH SNR and for the NN–SB comparison at LO SNR. Less certainty can be ascribed to the NN–SP comparison at MD SNR and the NN–SB comparison at VH SNR despite the fact that the effort ratings were consistently lower for the DNR processed conditions compared to the NN DNR condition. However, with only a 1.5% overlap between the HDI and the ROPE in the NN–SP condition at VH SNR there is more certainty in the credible difference for 1.33 units on the rating scale, although this only constitutes a very small effect ( $\hat{d} = 0.10$ ).

SNR	Comparison		Mode	Distribution around null	Null outside		Effect size $\hat{d}$
	DNR	DNR			HDI	ROPE	
1. MD-VH	$\frac{1}{4}(NN + MM + SB + SP)$	$\frac{1}{4}(NN + MM + SB + SP)$	3.37	$0.00\% \leq 0 < 100.00\%$	Yes	0.00%	0.83
2. VH	$NN - \frac{1}{3}(MM + SB + SP)$	$NN - \frac{1}{3}(MM + SB + SP)$	0.77	$0.02\% \leq 0 < 99.98\%$	Yes	27.76%	
3. MD	$NN - \frac{1}{3}(MM + SB + SP)$	$NN - \frac{1}{3}(MM + SB + SP)$	0.61	$0.43\% \leq 0 < 99.57\%$	Yes	56.39%	
4. VH	NN - MM	NN - MM	0.24	$14.39\% \leq 0 < 85.61\%$	No	95.00%	
5. VH	NN - SB	NN - SB	0.69	$0.65\% \leq 0 < 99.35\%$	Yes	66.18%	
6. VH	NN - SP	NN - SP	1.35	$0.00\% \leq 0 < 100.00\%$	Yes	1.50%	0.10
7. MD	NN - MM	NN - MM	0.60	$2.04\% \leq 0 < 97.96\%$	Yes	83.43%	
8. MD	NN - SB	NN - SB	0.34	$6.42\% \leq 0 < 93.58\%$	No	94.19%	
9. MD	NN - SP	NN - SP	0.80	$0.15\% \leq 0 < 99.85\%$	Yes	22.56%	

Table 4.14: Summary of the posterior comparisons made for perceived effort ratings. Effect sizes are only shown for credible differences with considerable certainty.



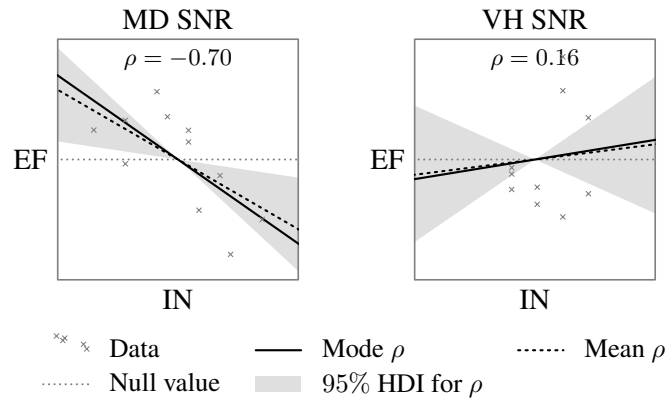


Figure 4.15: Correlations between (standardised) intelligibility (IN) scores and effort ratings (EF) at low SNR (left) and high SNR (right).

With the only credible difference found in perceived effort ratings with maximal certainty between the SNR conditions, this suggests that perceived effort may be based simply on the perceived level of background noise. Alternatively, given that the effort ratings were made after completing an intelligibility trial, the effort ratings may have been based on perceived intelligibility accuracy. Correlations between intelligibility and effort ratings at MD and VH SNR without DNR (i.e., the NN condition) are shown in 4.15. The parameters of the correlation were derived using the Bayesian graph and parameter settings shown in section B.3. MCMC sampling ran for 200000 steps with 50000 samples discarded as burn-in, and every second sample discarded as thinning to reduce auto-correlation in the traces. Convergence was assessed visually and with the Geweke statistic which revealed no concerns. Figure 4.15 shows a large negative correlation between intelligibility and effort ratings (mode  $\rho = -0.70$ ) at MD (i.e., 0 dB) SNR and with the null value outside the 95% HDI this could be interpreted as a credible correlation (although considerable caution should be exercised when interpreting a correlation with so few participants, particularly without a ROPE).<sup>2</sup> No credible correlation is suggested at VH (+8 dB) SNR as the null value is one of the 95% most credible correlations). This suggests that participants may have based their effort ratings on perceived intelligibility at lower SNRs.

Although the effort ratings for the no-DNR condition (NN) were consistently lower than the effort ratings for the DNR algorithm conditions (MM, SB, SP), suggesting that perceived

<sup>2</sup>A ROPE was omitted for the correlations as it wasn't clear what values of the correlation coefficient would be equivalent to the null.

effort was increased when listening to DNR processed speech, only one of these differences (NN–SP at VH SNR) was found to be credible with a high level of certainty (which only constituted a small effect), showing either that the perceived listening effort scale was not sensitive enough to distinguish between unprocessed and processed speech, or that any distortions introduced into the speech by the DNR algorithms counteracted the benefits of any reduction in background noise that they provided.

#### 4.2.4 Discussion

The effect of noise on the probe drift-rate found experiment III was not replicated in the current experiment. No consistent effects of noise were found on attention in the flanker task or the probe task. The only effect found in probe/flanker task was a credible effect of cognitive load on the full recall task, showing an unsurprising decrease in accuracy under high cognitive load. Although a marginally credible effect of DNR was found on flanker drift-rate in trials where the flanker target was mixed with 0 dB (MD) SNR and processed with the Sub-space (SB) DNR algorithm, this would have only constituted a very small effect size.

In the P.835 task, speech was rated more distorted with DNR processing, but only at 0 dB SNR suggesting that DNR processing runs the risk of distorting speech at lower SNRs. Background intrusiveness and overall speech quality ratings were only credibly different between 0 dB and +8 dB (VH) SNR. There were no other credible differences found in any of the P.835 ratings. The intelligibility results demonstrated the largest number of credible differences in ratings. In particular, a large credible difference in intelligibility was demonstrated between the SNRs used in the experiment without DNRs. The listening effort ratings also produced a large credible difference between SNRs with lower SNRs being rated as requiring more listening effort. However, the listening effort and intelligibility test were performed simultaneously, with listening effort ratings being provided after an intelligibility trial, so the ratings may have been influenced by each participant's perception of their performance in the intelligibility trial.

### 4.3 General discussion

Experiment IV compared the auditory flanker task with three conventional speech technology evaluation tasks: speech intelligibility, perceived effort ratings and the ITU-T P.835 standard for evaluating speech technologies. The aims were to see if the credible (but unexpected) effect of

noise found in the probe task in experiment III could be replicated, and if the other evaluations were any more sensitive than the auditory flanker task to experimental manipulations of noise and digital noise reduction.

The auditory flanker task performed poorly with all effects in the probe task being eliminated and with only one condition in the flanker task providing a very small effect. Thus, the flanker task in its current form, with or without the probe task is insensitive to relatively large changes in SNR, and has provided relatively little insight into how processing noisy targets affects attention.

If the replicated elimination of the flanker effect was due to increased cognitive load, perhaps a systematic effect resulting from the interaction of attention and the requirement of task coordination to complete the flanker and probe tasks, it is not clear why the flanker effect was eliminated, given that previous research has suggested that increased cognitive load results in increased interference from distractors (Lavie et al., 2004; Dalton et al., 2009; Francis, 2010). The measures of attentional control (AOSPan) and susceptibility to distraction (CFQ) revealed that the participants covered a range of possible scores for both the AOSpan and the CFQ (rather than all being clumped at one end) suggesting that the elimination of the flanker effect in experiment IV was not due to the chance recruitment of a group of participants who were unusually good at resisting distraction. One possibility for the elimination of the flanker effect in this experiment is that participants' rehearsal of the recall set is likely to have involved cycling through the each number in the recall set in order, and responses in the flanker task may have been delayed until the end of a particular rehearsal cycle. However, if that was the case, it is reasonable to assume that the same rehearsal strategy was used in experiment III, so responses in low load trials (one digit) would be uniformly faster than responses in high load trials (six digits), which from the largely overlapping HDIs shown in figure 3.18 (p. 106) does not appear to be the case.

Another possibility is that the number of responses in each condition was quite small, resulting in biased fitting of the DDM parameters. With 256 trials fully crossed between two loads (LCG, HCG), two 'congruencies' (CON, INC), two SNRs (MD, VH) and four DNR conditions (NN, MM, SB, SP), this resulted in only eight responses in each condition. Ratcliff (1979) argues that with some RT models (e.g., the Gamma distribution) at least 100 responses are required, and low numbers of reaction times (e.g., less than 10) need to be pooled in order to

avoid bias in any subsequent estimates. One indirect way of examining this issue is to take the results from experiment Ia and randomly sample small numbers of responses from participants, fit the DDM model to each subset and establish if there is still a credible difference between congruent and incongruent trials. If the large credible effect of congruency on drift-rate disappears with smaller samples sizes, that could suggest that the small number of responses used in experiment IV (and possibly experiment III, where there only 12 responses in each condition) may have resulted in biased estimation of the DDM parameters, which led to the apparent elimination of the flanker effect. If however, the large credible effect of congruency on drift-rate remains with smaller samples sizes, then that could suggest that the elimination of the flanker effect in experiment IV was not a result of the small number of responses. It should be stressed that is a *very* informal test. The responses in experiment I were collected under very different conditions, and no assessment of participant's attentional control or susceptibility to distraction was made. The very large flanker effect observed in experiment I could, therefore, be the result of a set of participants that were unusually susceptible to distraction. However, the test would give some indication of how drift-interference might be affected with a reduced number of responses.

Figure 4.16 shows the results from 100 simulations using random subsets of the data from experiment Ia, with 8, 12 or 16 responses from each participant in each condition (i.e., congruent and incongruent). A DDM model was fitted to each subset (with 11000 MCMC steps, 1000 burn-in samples discarded and no thinning) and the credibility of the difference between congruent and incongruent trials was calculated (i.e., whether the ROPE was outside the 95% HDIs with 100% certainty *and* the drift-rate was higher in congruent trials compared to incongruent trials). The resulting proportion of credible differences is shown on the left in figure 4.16, and the effect sizes for these credible differences is shown on the right (with some jittering along the x-axis to make the distribution of effect sizes clearer). It can be seen that with only 8 samples from each participant in each condition, over 75% of the models result in a credible difference, which rises to almost 100% for 16 samples per participant in each condition. In addition, the effect sizes are all large effect sizes, and only increase slightly as the number of responses per participant per condition increases.

But with approximately 25% of the models failing to produce a credible flanker effect using the same number of samples per participant per condition from experiment Ia as were used

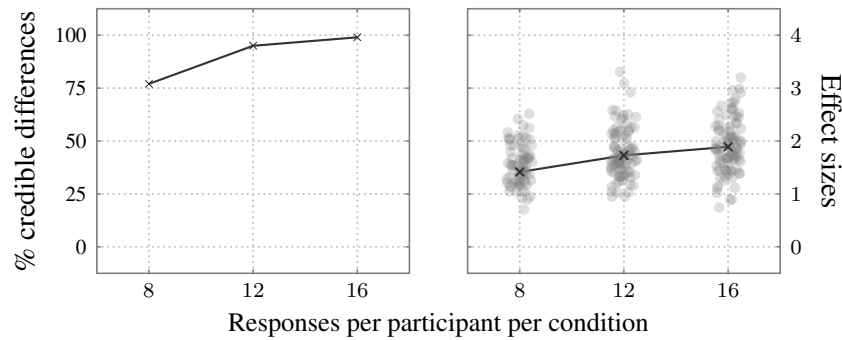


Figure 4.16: Plots of the proportion of credible differences in drift-rate from reduced samples sizes in experiment Ia (left) and the effect sizes for the credible differences (right).

in experiment IV, one very cautious conclusion (given that the flanker data from experiments Ia and experiments IV were collected in under substantially different conditions) is that the elimination of the flanker effect in experiment IV *may* have been due to the fact that the number of responses in each condition was too small. It should be stressed, that this is only a very informal conclusion, and the only way for this conclusion to be verified would be to run experiment IV again with more trials.

However, even if more trials were run, it should be noted that existing speech technology evaluations are recommended to last (at most) for forty minutes to an hour (Thorpe, 1998) to avoid participant fatigue and boredom. In addition many evaluations are carried out by companies that specialise in particular evaluations which can involve considerable expense (e.g., Hu & Loizou, 2007). So, in practice, the increased time required for evaluating each participant may result not only in participant fatigue and boredom, but also increased financial costs. Taken together, these considerations suggest that the probe/flanker task in its current form is not a suitable task for evaluating speech technologies.

Out of the other speech technology evaluations, only the intelligibility test was able to provide consistent distinctions, and these were mostly unsurprising, as intelligibility was credibly reduced when the SNR was reduced. Listening effort, speech distortion, background intrusiveness and overall speech quality only produced credible distinctions which could mostly be accounted for by differences in SNR. However, intelligibility was made *worse* by all the DNR algorithms at relatively low SNRs, and made worse by two of the DNR algorithms at a

relatively high SNR as well. On the basis of the current experiment, simply in terms of the number of credible effects that were produced, the intelligibility test would appear to be the most valid test for distinguishing between noisy and DNR processed speech, at least with the SNRs and DNR algorithms used in experiment IV.

## Chapter 5

# Conclusions

In the first chapter it was proposed that if listening to noisier speech increases listening effort then the success of a digital noise reduction (DNR) system should be measured in terms of the reduction in listening effort it provides. Furthermore, it was proposed that a simple behavioural measure of listening effort would be an efficient and cost-effective way of measuring listening effort and could be used to evaluate speech technologies (cf. Houben et al., 2013).

With listening effort linked to the control and maintenance of attention (Bernarding, Strauss, Hannemann, Seidler & Corona-Strauss, 2013) and the flankers task having been shown to be a valid behavioural task for investigating attentional control (Lavie et al., 2004; Francis, 2010), the flankers task was selected as the behavioural task that would be used to investigate changes in attention when listening to noisy speech. The reversal of any changes found would then form the criteria by which DNR systems would be evaluated.

### 5.1 Experimental summary

Chapter 2 developed the auditory flanker task, which simulated telephone use with monaural targets and binaural distractors and could use an arbitrary number of monosyllabic words and speakers (although in practice only fifty words and six speakers were used). As shown in previous research in auditory flanker tasks (e.g., Chan et al., 2005; Francis, 2010), response times were faster in trials where the target and distractor words were congruent (associated with the same response) and lower when the targets and distractors were incongruent (associated with difference responses). In addition, accuracy was higher in congruent trials and lower in incongruent trials. Flanker performance was also analysed in experiment Ia using the Drift-Diffusion Model (DDM, e.g., Voss et al., 2013) which models performance in 2AFC tasks

(like the flanker task) as a noisy decision process that accumulates information for one of two decisions until a threshold has been crossed and a response is initiated. When information from the flanker target is processed, the decision process ‘drifts’ towards the threshold associated with the correct response, and when information from the flanker distractor is processed, the decision process drifts towards the threshold associated with the incorrect response.

In congruent trials, slips or leaks of attention to information from the distractor results in the decision process drifting towards the correct-response threshold more quickly, generating faster, correct responses. In incongruent trials, slips or leaks of attention to information from the distractor results in the decision process drifting towards the threshold associated with the correct response more slowly, generating slower, correct responses. With more severe slips or leaks of attention, the decision process drifts towards the incorrect-response threshold resulting in slower, incorrect responses. So, the drift-rate represents the average rate of information accumulation in the decision process, and models response time and accuracy simultaneously.

Experiment Ia showed that the drift-rate in congruent trials was higher than in incongruent trials, and that the variation in the stimuli with the increased numbers of words and stimuli did not credibly alter this difference. In addition, experiment Ib showed that the differences in drift-rate between congruent and incongruent trials were not due to energetic masking from the distractor overlapping with the target, as the effects were still shown when energetic masking was eliminated by using ‘dichotic’ distractors which were mixed to the opposite stereo channel to the target. Assuming participants were, on average, equally distracted in congruent and incongruent trials, and therefore processed the same ‘amount’ of distractor information in both types of trial, the difference in drift-rate between congruent and incongruent trials provided a measure of the degree to which the distractor had ‘interfered’ with the participants’ decision process. So, the difference in drift-rate between congruent and incongruent trials, or ‘drift-interference’ was proposed as the measure of distraction, by analogy with the measures of interference derived from RTs in previous research using the flanker task (Lavie et al., 2004; Francis, 2010; Wyatt & Machado, 2013).

Experiment II took an initial look at how targets corrupted with background noise would effect attentional control and hence induce changes in (drift) interference. According to a prominent model of attention, ‘Load theory’ (Lavie, 2005), background noise could have one of two effects, depending on which of two limited capacity sub-systems the background noise affected.



If the background noise affected the perceptual sub-system, which passively processes all incoming stimuli (subject to capacity limitations), then the background noise would constitute a ‘perceptual load’, the available perceptual capacity would be reduced, and less information from flanker distractors would be processed. This would result in a decrease in drift-interference. If, however, the background noise affected the cognitive sub-system, which actively maintains the focus of attention according to task demands, then the background noise would constitute a ‘cognitive load’, the available cognitive capacity would be reduced leading to a loss of attentional control, and more information from flanker distractors would be processed. This would result in an increase in drift-interference.

Experiment II found that there were credible effects of noise on drift-rate in congruent and incongruent trials separately, but differences were only for quite large changes in SNR, and were not credibly different between the congruent and incongruent trials, so drift-interference was not credibly altered with the addition of noise. However, the results, while not credible, suggested that drift-interference increased as SNR decreased, but that the basic flanker task was not sensitive enough to measure the difference. In addition, the addition of DNR to noisy flanker targets introduced no measurable change in drift-interference.

With an indication (albeit non-credible) that decreasing the SNR increased the drift-interference, this implied that the increased SNR constituted a cognitive load (according to load theory). So, it was argued that by explicitly introducing an additional cognitive load, listeners could be ‘pushed’ into using more cognitive capacity. This would result in participants losing attentional control to a greater degree, to the extent that any losses in attentional control due to changes in SNR would be more noticeable, resulting in larger, measurable changes drift-interference. Experiment III embedded in the flanker trial inside a memory probe task which had already been demonstrated to disrupt attentional control under conditions of high cognitive load compared to low cognitive load (Lavie et al., 2004; Francis, 2010). It was hoped that this would make the flanker task more sensitive to SNRs. However, somewhat unexpectedly, drift-interference was eliminated entirely, at all SNRs, with or without digital noise reduction, under both high and low cognitive load.

It was considered possible that participants may have shifted their priorities in the combined probe/flanker task, and instead of maintaining performance in the probe task, so that the increased cognitive load affected the flanker performance, participants had instead maintained

performance in the flanker task. As a consequence, it was possible that any effects of cognitive load and SNR would be evident in the probe task. In line with research suggesting that attending to noisy speech can have an impact on memory (Rabbitt, 1968; Howard et al., 2010; Sarampalis et al., 2009) it was possible that attending to noisy targets in the flanker trial would divert attentional resources from maintaining the memory trace for the probe recall set. Noisier targets would require more attentional resources, resulting in greater difficulty in maintaining the memory trace, and as a consequence, the memory trace would become more ‘degraded’ and contain less information about the recall set. The probe-task could also be modelled using the drift-diffusion model (like the flanker task it was a 2AFC task) and the probe drift-rate (averaged across congruent and incongruent trials) acted as an indicator of the degradation in the memory trace (Ratcliff et al., 2004) and the accumulation of information from the memory trace was expected to be slower for more degraded memory traces.

Credible effects of SNR were found, with probe drift-rate lower in trials where the corresponding flanker trial had lower SNR targets compared to trials where the flanker trial had higher SNR targets. This suggested that the memory trace was degraded to a greater degree in trials where flanker targets had a greater level of background noise. These effects were found under both high and low cognitive load but mainly for large differences in SNR (8 dB)

With a credible reduction in probe drift-rate with decreased SNR, it was possible to establish a criterion for evaluating the digital noise reduction systems: if decreasing the SNR in the flanker targets increased the degradation in the memory trace for the recall set in the probe task (as reflected by decreased probe drift-rates), then the application of DNR to flanker targets should result in less degradation in memory trace for the probe recall set (which would be reflected in higher drift-rates). However, no credible differences in probe drift-rate could be found between conditions with DNR-processed or unprocessed flanker targets, which was taken to suggest that the speech distortion introduced by the DNR processing counteracted any benefits provided by the reduction in background noise (Hu & Loizou, 2007). Nevertheless, experiment III showed that the probe/flanker task could measure *something* related to SNR, but without much sensitivity. Furthermore, it was not possible to establish if the increased distortion in the probe recall set (as measured by probe drift-rate) was simply due attending to the noisy targets or the combination of attending to the noisy targets *and* resisting distraction from the distractor.

Experiment IV aimed to replicate the effects of experiment III, and also compare the flanker/probe task with some established speech technology evaluations: speech intelligibility, subjective listening effort, and a version of the ITU-T P.835 standardised speech technology evaluation protocol used in the telecommunications industry to evaluate the ‘quality’ of DNR systems. This time the probe/flanker task failed to provide any credible distinctions between any of the experimental conditions, under high or low cognitive load, at different levels of SNR or with or without DNR processing, in either the flanker task or the probe task. The only effect was found in the full recall task which had been added to the end of each trial where participants had to recall the entire probe recall set, but the effect was only related to contrast between high and low cognitive load with the expected lowering of accuracy under high cognitive compared to low cognitive load. The lack of an effect could not be attributed to recruiting participants who were unusually resistant to distraction, as an objective measure of attentional control the (automated) Operation span (Unsworth et al., 2005) and a subjective measure of distraction in daily life, the Cognitive Failures Questionnaire (Broadbent et al., 1982) both suggested that participants were susceptible to distraction to a different degree. So, an explanation for the lack of any effects due to a uniformly high propensity for attentional control across all the participants was not plausible.

If no plausible theoretical reason for the elimination of the flanker effect could be established then an alternative suggestion was there were too few responses in each condition resulting in biased estimates of the DDM parameters resulting in the elimination of the flanker effect. However, increasing the trials, while possible, may not have been practical as existing speech technology evaluations are recommended to last (at most) for forty minutes to an hour (Thorpe, 1998) to avoid participant fatigue and boredom. As the probe/flanker task already lasted an hour, and several participants found it considerably taxing, adding more trials to the experiment may have increased fatigue and boredom to the extent that performance on the task may have suffered (Cheyne et al., 2006) Therefore, the auditory flanker task in its current form could not be considered a practical measure of listening effort and was not a suitable task to complement existing speech technology evaluations.

However, the other tasks did not fair much better, with most of the small number of credible differences in the perceived effort ratings, speech distortion, background noise intrusiveness and overall speech quality relating to SNR. Only the intelligibility test provided the largest number

of credible differences between the audio conditions, including differences between DNR processed and unprocessed speech at a relatively high SNR (+8 dB). On the basis of the current work, it would appear that it is only the intelligibility test that provides measurable distinctions between levels of noise and distinctions between processing with or without technologies.

## 5.2 Research themes

The current work aimed to investigate the use of listening effort as criterion for speech technology evaluation. Although it was acknowledged that listening effort may be difficult to measure in adults with no hearing, reading or cognitive impairments, if the production of a novel measure of listening effort with a wide range of application was the criterion for success then the it has to be acknowledged that this criterion was not met. Nevertheless, during the design, execution and analysis of the experimental work presented above, interesting areas of research have been touched upon.

### 5.2.1 The auditory flanker task

The auditory flanker task developed in this thesis (chapter 2), represents a considerable extension of existing auditory flanker tasks using spoken words (Francis, 2010; Chan et al., 2005) or letters (Murphy et al., 2013). The flanker effect was replicated using a substantial increase in the number of stimulus words from six (e.g., Chan et al., 2005) to fifty, and the flanker effect proved to be considerably robust to the resulting stimulus variability.

In addition, the flanker task developed above utilised a method of stimulus alignment which does not appear to have been used in attention research before in order to make the targets and distractors appear to occur simultaneously. The perceptual centre, or ‘p-centre’ is the point in a word when it is perceived to occur (Fowler, 1979) and was operationalised as the point at which a critical frequency band reached 50% of the maximum energy for the word (Scott, 1994). This proved effective for the monosyllabic singular CVC nouns used in the current work and this approach could be used to extend the stimuli used (in terms of both words and speakers) in speech-based flanker tasks to investigate the interaction of attention and speech perception along other dimensions lexical frequency, lexical neighbourhood or semantic similarity.

Given that selective attention is being given an increasing prominent role in accounting for

how listeners organise acoustically complex environments into perceptually distinct sources (Shinn-Cunningham, 2008; Shamma, 2008; Ihlefeld & Shinn-Cunningham, 2008; Cusack, Deeks, Aikman & Carlyon, 2004; Carlyon & Cusack, 2005; Ruggles, Bharadwaj & Shinn-Cunningham, 2011) the flankers task, either in the free-field (Chan et al., 2005) or simulated over headphones (Francis, 2010 and experiment Ia and experiment Ib above), may provide a useful paradigm for establishing more precisely the role that attention plays in perceptual organisation, although perhaps with more constrained stimuli.

#### 5.2.1.1 Drift-diffusion model

Performance in the flanker task (and the probe task) was modelled as a drift-diffusion model (DDM, Ratcliff, 1978; Voss et al., 2013; Ratcliff et al., 2004) where RTs and accuracies in 2AFC tasks are the result of a noisy decision process which continually accumulates information for two possible responses until sufficient information is accumulated for one of them, and a response is initiated. The use of the DDM in the experiments reported above adds not only to the growing body of research using the DDM, but also to the much smaller body of research that has used the DDM to analyse performance in flanker tasks. In addition, it appears that this is first known use of the DDM to analyse auditory flanker tasks. The lack of effects in the latter experiments are not felt to reflect badly on the DDM, as experiment Ia and experiment Ib both provided interesting perspectives on the modelling flanker performance using the DDM and interpreting attentional processes using the parameters of the model.

In addition, the DDM provides an interesting alternative to analysing reaction times and accuracies separately in other speech based research. Conceivably, a straightforward 2AFC listening test (i.e., without flanking distractors) could be analysed with the DDM and the other parameters which were not credibly relevant in the flanker task (i.e., the non-decision time and the threshold separation parameters) may provide additional insight into how listeners process noisy speech and enhance existing theories regarding speech perception in noise.

However, although the descriptions of the DDM given above have made reference to a noisy decision process that accumulates ‘information’ over time for two responses in a 2AFC task (see section 2.2), nothing specific has been said about the nature of that information. In part, that is because most applications of the DDM to date have also been vague about the specifics of what constitutes ‘information’ in the decision process of the DDM, generally referring to

“sensory evidence” (e.g., Zhang & Rowe, 2014, figure 1). But, even when recognising isolated words listeners do not necessarily rely solely on the sensory input and may supplement sensory information (particularly when it is degraded) with lexical or sub-lexical information depending on the degradation or masking in the stimulus or the level of cognitive load (Mattys et al., 2009; Mattys & Wiget, 2011).

If interference in the flanker task is due “in large part” (Eriksen, 1995, p. 101) to response competition, then the distractor must have been processed to the point that it could activate the response associated with the distractor (Santee & Egeth, 1980). If this is the case then, it is reasonable to assume that other sources of information that are used by listeners when recognising isolated spoken words (i.e., lexical or sub-lexical information) may be used in the decision process, and conceivably contribute to the overall drift-rate. Therefore, one important aspect of the DDM which needs to be clarified if it is to be useful in future in speech-based experiments (perhaps by more careful control of the experimental stimuli and conditions that occurred in the experiments reported above), is exactly what ‘information’ is being accumulated by the decision process of the DDM, or whether the exact nature of the information accumulated can be experiment-specific depending on the experimental manipulation.

### 5.2.2 Empirical Bayesian analysis

The experimental data collected here were analysed using empirical Bayesian methods, updating prior beliefs (models) with observed evidence (experimental data) to form posterior beliefs (distributions of model parameter values) in order to assess the credibility of the prior beliefs (i.e., test hypotheses). Although Bayesian principles predate the more usual NHST analyses found in psychology, they are rarely used in current research, perhaps due to the computational complexity of forming the posterior distributions of parameters (Kruschke, 2010a). The analyses presented in the current work demonstrate a number of ‘traditional’ analyses (correlation, ANOVA, mixed effects linear regression and mixed effects logistic regression) from a Bayesian perspective. While methods for hypothesis testing in Bayesian analysis are a focus of active research (Wiecki et al., 2013) so do not necessarily have such widely agreed system of interpretation that is apparent in the more established NHST approaches, the analyses shown in the current work constitutes a small but practical introduction to applied Bayesian analysis in psychological research.

### 5.2.3 Listening effort

Despite the growing body of evidence pointing to effects of listening to noisy speech that go beyond simple intelligibility effects, it is still the case that ‘listening effort’ is an ill-defined concept (McGarrigle et al., 2014) with multiple definitions (Picou et al., 2011; Gosselin & Gagné, 2010; Bourland-Hicks & Tharpe, 2002; McGarrigle et al., 2014). These definitions are so broad that almost anything that isn’t related to intelligibility could be viewed as listening effort. Perhaps because of these broad definitions, research into listening effort uses a wide variety of stimuli, including sounds, syllables, words or sentences (Bertoli & Bodmer, 2014; Bernarding et al., 2013; Sarampalis et al., 2009; McCreery & Stelmachowicz, 2013; Mackersie & Cones, 2011), a wide variety of tasks, such as recognition, memory, vigilance and motor control (Howard et al., 2010; Hornsby, 2013), and a wide variety of measures, including response times to a primary stimulus or secondary stimulus, recall accuracy, pupil dilation, ERP, skin conductance and brain activity (Houben et al., 2013; Hornsby, 2013; Howard et al., 2010; Zekveld et al., 2010; Bertoli & Bodmer, 2014; Mackersie & Cones, 2011; Wild et al., 2012). With such a wide variety of stimuli, tasks and measures, it might be reasonable to ask if everyone can really be measuring the same thing (McGarrigle et al., 2014). Underlying listening effort research is the idea that attention is under the active control of a limited capacity system, and that in challenging listening situations, the resources of this system are redirected to deal with the listening challenge, resulting in some loss in attentional control. However, there is no complete agreement about the structure or function of the system that actively controls attention (see the various chapters in Miyake and Shah, 1999 for different perspectives) or how to assess its capacity (Oberauer, Süß, Schulze, Wilhelm & Wittmann, 2000).

Like the majority of listening effort research, the research reported above has referenced, but not committed to any specific model of attention, attentional control or memory. Instead, some general form of limited capacity processing has been assumed where some of the processing is under the control of the participants and some the processing is not. Although considerable reference has been made to load theory (e.g., Lavie, 2005), the current work has not aimed to validate load theory in the auditory domain (cf. Gomes et al., 2008; Francis, 2010; Murphy et al., 2013) or address particular specifics of load theory, but merely used certain aspects of the theory to frame predictions regarding changes in the drift-diffusion model parameters under specific experimental manipulations. No consideration has been given to the fact

that the concept of ‘perceptual load’ is controversial (Benoni & Tsal, 2010; Benoni & Tsal, 2013), and may not have the same effects in the auditory domain that have been demonstrated in the visual domain (Gomes et al., 2008; Murphy et al., 2013). It may be that this kind of non-committal ‘hand-waving’ approach to specific psychological models (which is typical of listening effort research) will lead to an increasing number of studies which show ‘promising’ or ‘interesting’, but nonetheless inconclusive results (see McGarrigle et al., 2014 for similar arguments).

Mattys, Davis, Bradlow and Scott (2012) provide a list of different conditions which present various kinds of challenge (‘adverse conditions’) for listeners which have an impact on the listener in different ways. For example, energetic masking (i.e., from background noise) was estimated to have severe effects on intelligibility and available attentional capacity, but less severe effects on available memory capacity. In contrast, cognitive load was estimated to have a severe effect on available attentional capacity and memory capacity, but less severe effects on intelligibility. In addition, signal degradation that does not result from energetic masking was estimated to have a less severe effect on intelligibility, available attentional capacity and memory capacity. While Mattys et al. are careful to point out that these are only the authors’ subjective judgements about the relative importance of adverse conditions on listeners, taken together, they do point to a number of (possibly interacting) effects of listening adverse in conditions which existing listening effort research (including the research presented above) often fails to address.

In the experiments reported above, targets were presented with multiple ‘adverse conditions’ such as energetic masking (background noise) non-energetic signal degradation (digital noise reduction), and cognitive load. With a distractor stimulus thrown in, a complex, possibly interacting series of effects might be expected. Failing to carefully distinguish between the possible effects (assuming they are dissociable to some degree) and their interactions may have contributed to the disappointing results reported in experiment II, experiment III and experiment IV. More credible effects may have been found if the experiments were designed with the recommendations of Mattys et al. (2012) in mind, limiting the locus of the ‘adverse conditions’ to the extent that the experimental manipulations would yield specific, measurable effects.

In summary, while ‘listening effort’ is an interesting concept, its broad and rather vague definitions leave it open to a wide range of interpretations. Future listening effort research should



be more specific about the type of challenge that is being imposed on listeners in listening effort tasks, as well as showing a greater commitment to specific psychological theories, to order be clearer about the causes and effects of listening effort that are being addressed.

Finally, the majority of listening effort research considers special or clinical populations sometimes specifically selecting the participant population (such as the elderly) because it may produce more measurable results (Bertoli & Bodmer, 2014). The research reported here used adult listeners with no known hearing, reading or cognitive difficulties. As acknowledged from the start, these listeners may have been able to complete the listening effort task developed above without measurable difficulty, or with sufficient ease that variability in trial by trial performance was greater than the variation caused by experimental manipulations. Furthermore, measurable changes in performance for unimpaired listeners may only result in the presence of extreme noise or distortion which renders the speech unintelligible. In this case, a simple intelligibility test would be sufficient. Assuming the theoretical assumptions behind the motivation of the research reported above were correct, and that no serious methodological issues exist in the experimental work (although it is possible that insufficient trials were run), a credible suggestion is that listening effort may not be an issue for adult listeners with no hearing or cognitive impairments. Alternatively, if listening effort is an issue for these listeners, it does not produce effects that can be measured using the behavioural tests developed above. Nevertheless, it is possible that with other populations, such as the elderly, hearing impaired or (perhaps with some modifications to the vocabulary) children, the speech-based auditory flanker task developed in the current work may show credible effects of noise on interference from distractors.

### **5.3 Conclusions**

This thesis set out to investigate listening effort in unimpaired adults listeners, in order to establish a new method of evaluating speech technologies based on the reduction of listening effort that the technology provides. The failure to establish a new method does not mean that listening effort is an invalid concept, or though it might lend credence to an argument suggesting that listening effort is only relevant for special populations such as the very young, very old, or populations with sensory or cognitive impairments.

Nevertheless, it is still the case that the use of speech technologies is an increasing fact of life for many individuals, that evaluation is an important part of the development of speech

technologies, and that there is widespread belief that conventional evaluations may be failing to capture important factors that affect users of speech technologies in real world environments. So the argument put forward at the beginning of the current work is still valid, and it is believed that this argument should form the basis of the development of any future speech technologies: the evaluation of any technology designed to ease communication should include a demonstration that the technology does actually make communication easier, and that should be the principle criterion used to determine the technology's success.

# Appendices

## Appendix A

# Bayesian inference

### A.1 Bayesian statistics

There are numerous introductions to Bayesian statistics which discuss in detail the practical and philosophical differences between Bayesian and frequentist approaches to statistical inference and in particular the relationship between Bayesian inference and null hypothesis significance testing (NHST) (e.g., Glickman and van Dyk, 2007; Kruschke, 2010a, 2010b, 2011, 2013; Fisher and Wolfe, 2012; Feinberg and Gonzalez, 2012; Gelman and Shalizi, 2012; Pullenayegum, Guo and Hopkins, 2012; Dienes, 2011; Wagenmakers, 2007, see also Bishop, 2006). The purpose of this appendix is to provide a general outline to the Bayesian analyses used throughout this thesis. The use of Bayesian statistics in this thesis is motivated in part by the fact that performance in the flanker task is modelled as a drift-diffusion process (see section 2.2) and the tool used to estimate the parameters of the model empirical Bayesian estimation (Wiecki et al., 2013) making the use of Bayesian inference appropriate. Although some researchers have used Bayesian estimation to establish model parameters for various experimental manipulations and then performed a traditional NHST ANOVA on these parameters (e.g., Zhang & Rowe, 2014), the suspicion is that the use of NHST was motivated by requirements that arose during the peer-review process, as inferences are possible on the distributions of the estimated parameters without resorting to NHST (Kruschke, 2010a).

The fundamental principle behind Bayesian statistics is that the experimenter's prior beliefs about some observable event can be modified once the event has been observed according to the credibility that the observations give to the prior beliefs (Kruschke, 2010a). The 'prior beliefs' are represented as model of the observable event. The parameters of the model are

probability distributions (Glickman & van Dyk, 2007) which form the initial (i.e., prior) beliefs about plausible or ‘credible’ parameter values for the model (Fisher & Wolfe, 2012). Updating the prior beliefs involves evaluating their credibility given some observed data to form the *posterior* beliefs on which inferences can be made. (Kruschke, 2013).

## A.2 Bayesian model specification

The model representing the experimenter’s prior beliefs consists of a number of parameters, the relationships between the parameters and a specification of possible values each parameter may take. Each parameter is specified as a random variable with a particular distribution (the ‘priors’) with constraints on the values that the distribution may take (‘hyperpriors’). The priors and hyperpriors represent the experimenter’s prior beliefs about the relative ‘credibility’ of the parameter values in the model before the data are observed (Kruschke, 2013). Priors and hyperprior distributions can be ‘informative’ or ‘uninformative’. Informative (hyper-) priors may be initialised with values based on the observed data or prior knowledge, which may help reduce the number of steps required for the posterior distributions to stabilise (see below). Uninformative (or ‘vague’) (hyper-) priors are initialised to random or extreme values and are usually favoured in order to counter the argument that the algorithm generating the posterior may have been restricted to ranges of parameters that guaranteed the final values of parameters. However in practice, sensibly chosen informative priors can be used effectively (Wiecki et al., 2013).

As an example of this kind of model specification figure A.1 shows a simple example of a Bayesian network for a linear regression model. Each prior distribution (coloured dark grey) in the network represents a parameter in the model. The hyperprior distributions (coloured light greys) represent constraints on the parameters of the prior distributions. At the bottom of the network  $y_i$  represents the  $i$ th observed data point and is assumed to be a normally distributed random variable with mean  $\mu_y$  and precision  $\tau_y$ . The mean  $\mu_y$  depends on the sum of multiple (random variables)  $\beta_0$  and  $\beta_1$  representing the intercept and ‘slope’. Each ‘ $\beta$ ’ is modelled as a normally distributed random variable, and the  $\mu_i$  parameter of the  $\beta_i$  random variable (i.e., prior) is assumed to derive from a normally distributed random variable (hyperprior) with mean  $M_i$  and precision  $T_i$  and each  $\tau_i$  parameter is assumed to derive from a (transformed) uniform hyperprior with limits  $L_i$  and  $H_i$  representing the standard deviation.

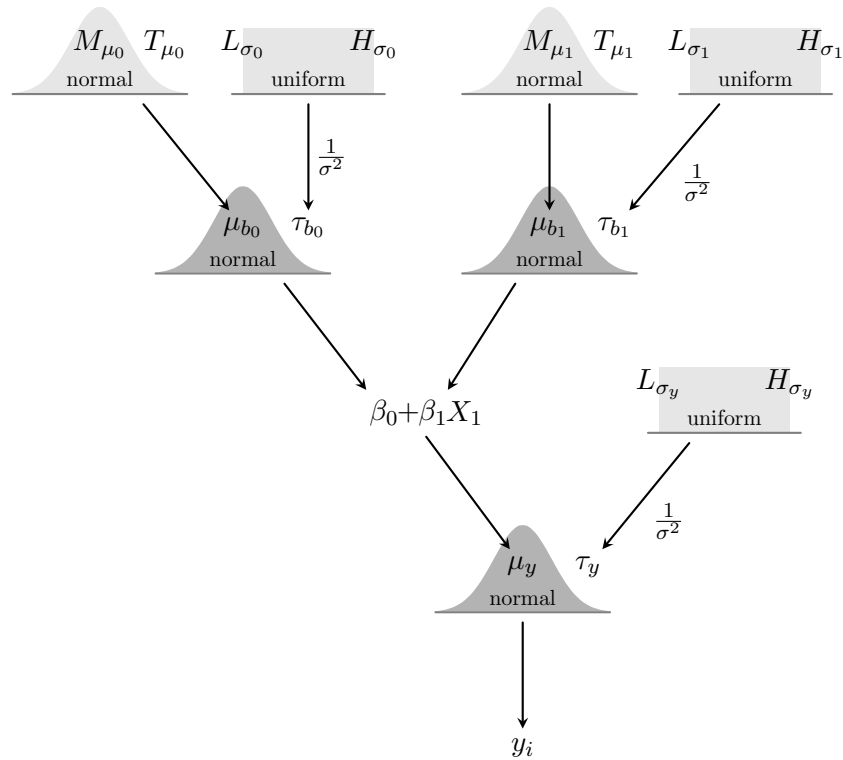


Figure A.1: Model dependencies for a Bayesian network regression for linear regression with prior distributions shown in dark grey and hyperprior distributions (i.e., the prior distributions for the prior distribution parameters) shown in lighter grey.

### A.3 Parameter estimation: Markov-Chain Monte Carlo

Having specified the prior model and collected data from the observed event which the model represents, Bayesian parameter estimation then reassigns the credibility of model resulting in posterior distributions of parameter values (or simply ‘posteriors’) which taken together represent the credibility of the model after the data has been observed (Fisher & Wolfe, 2012). Forming the posteriors of a model invariably uses Markov-Chain Monte Carlo (MCMC) methods due to the complexity of the joint distribution of the priors. Each configuration of the model parameter values can be thought as representing a point in a multi-dimensional space and MCMC sampling involves randomly moving around the parameter space until a representative sample of the points have been visited. Taken together the points that are visited in the parameter space represent the posterior distribution of the parameters space (and, therefore, the model). Detailed mathematical descriptions of MCMC principles are given in Andrieu et al. (2003), Brooks (1998) and Bishop (2006, chapter 11). In what follows a fairly high level description of

basic MCMC sampling is given based on Geyer (2011), Kruschke (2010a) and Davidson-Pilon (2015) using the Metropolis-Hastings algorithm (e.g., Bishop, 2006, chapter 11).

At each step in the MCMC sampling process, a point in the parameter space is generated by sampling from the prior distributions for the parameters and a decision is made whether to move to this point (i.e., whether to accept or reject the proposed sample of the parameters). The probability of accepting the proposed model is calculated by according to the probability of the proposed model (i.e., the configuration of the proposed parameters) and the probability of the current model (the configuration of the current parameters) given the data. Then a uniform random number is generated and if this number is greater the probability of accepting the proposed model then the MCMC records the values of the current parameters, moves to the point representing the proposed parameters and continues the process again. After a sufficient number steps the MCMC sampling ‘converges’ (the accepted values for each parameter form a stationary distribution) and the record of the accepted parameter values forms the ‘trace’ for that parameter, and when the MCMC terminates it is the trace that represents the posterior distribution of each parameter.

### A.3.1 MCMC convergence

Although more sampling iterations increases the probability that the posterior distributions have converged (provided appropriate prior distributions have been specified), if the process is stopped too early the posterior distribution created by the samples in the chain may not have converged making inferences unreliable. Depending on the initial values of the prior distributions the initial accepted parameter samples are likely to be unrepresentative of the final distribution (see figure A.2). These samples are referred to as ‘burn-in’ samples (Raftery & Lewis, 1992) and are discarded before further inferences are made with the posterior distribution. In addition, to avoid correlation in the MCMC trace samples where sequences of samples tend to ‘clump’ together (Kruschke, 2011) and samples can be discarded at regular intervals (e.g., every 3rd or 4th sample) a process known as ‘thinning’.

A key issue when using MCMC sampling is being able to assess (or ‘diagnose’ Cowles and Carlin, 1996) whether the posterior distributions have converged. Informal diagnostics involve plotting the traces for each parameter and visually inspecting the result (Wiecki et al., 2013, see figure A.2). In addition, a number formal methods for diagnosing the convergence

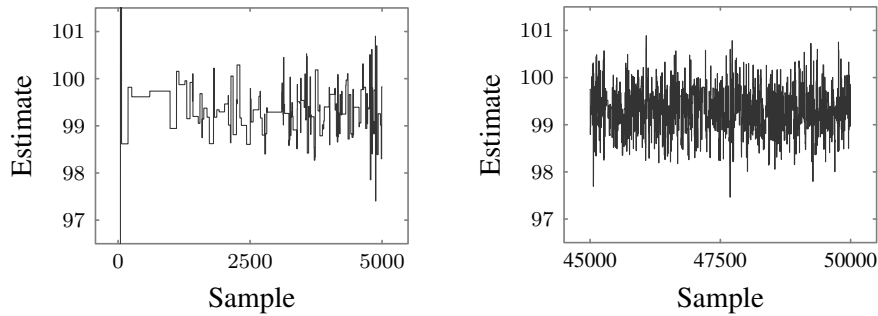


Figure A.2: Examples of an unconverged trace (left) from the beginning of an MCMC sampling process for a normally distributed prior and a converged chain (right) from the end of the same MCMC sampling process.

of an MCMC process (e.g., Gelman and Rubin, 1992; Geweke, 1992) or for proposing the minimum number of steps required for convergence (Raftery and Lewis, 1992, see Cowles and Carlin, 1996 for a discussion of these and other methods). One approach which is used for some of the models in this thesis is to use the ‘Raftery-Lewis diagnostic’ (Raftery & Lewis, 1992) to obtain initial recommendations for the minimum number of steps, burn-in and thinning for each parameter. After a short ‘pilot’ run of the model, diagnostics are generated for each parameter to ensure 0.5% accuracy at the 2.5% and 97.5% quartiles with a probability of 0.95 in the posterior distribution. The values subsequently used in the MCMC sampling were the maximum value for the number of steps, burn-in and thinning, after considering all parameters to ensure that every parameter had converged. However, while the recommended number of samples were discarded as ‘burn-in’ the recommended thinning was ignored for some models (unless visual inspection of the samples suggested otherwise) as more precision in the estimates can be obtained without thinning (Link & Eaton, 2012) and the prevalence of thinning may have arisen due to historical constraints on computer memory (Wiecki et al., 2013) which generally was not an issue with the models used in the experiments reported above.

While diagnostics can be useful indicators of the number of MCMC sampling steps required for convergence, in practice it is possible to diagnose convergence prior to the recommended number of steps. This can be particularly useful for very complex models with large amounts of data which may take a long time to complete if diagnostics are followed to the letter. So, for some models reported above, a procedure based on Geweke (1992) was used to evaluate if the distributions of the parameters had converged before the recommended number



of steps. After a fixed sampling interval (e.g., every 50000 steps) the distribution of the samples for that interval were compared to the distribution of the samples from previous intervals to see if they they were credibly different. If there was no credible difference between all pairwise comparisons of distributions over a number of intervals the sampling was halted and the posterior distribution constructed from the samples in the last few intervals. In addition, a visual check of the plot of the traces and auto-correlation plots were used to check convergence.

### A.3.2 Model fit

The assessment of model fit can be carried out by ‘posterior predictive sampling’ where samples are generated from the fitted model and these predictive samples are then compared to the observed data. Visual checks of the model fit can be obtained by overlaying the (averaged) posterior predictive sample RT densities on top of normalised histograms the observed participants (e.g., Zhang & Rowe, 2014).

Comparing the fit of models can be used to infer whether a particular parameter has an important role in a model. Several statistics have been proposed for comparing the model fit and relative merit of hierarchical Bayesian models including the mean square error (MSE, Gelman, Hwang and Vehtari, 2014) and the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin and van der Linde, 2002). The DIC is a generalisation of the Akaike Information Criterion (AIC, Akaike, 1973; Burnham and Anderson, 2004; Wagenmakers and Farrell, 2004) for hierarchical models for which the AIC and Bayesian Information Criterion (BIC, Nielsen, Christensen and Jensen, 2013; Wagenmakers, 2007) are inappropriate measures as although posterior distributions are formed using the likelihood of the model given the observed data, the posteriors are not formed by maximising the overall likelihood of the model. The DIC consists of a measure of model fit and a penalisation term for the number of parameters in the model, and while the DIC cannot be interpreted in isolation, it can be used to compare models (Zhang & Rowe, 2014): smaller DIC values represent a ‘better’ model.

Although the MSE may be less appropriate for parameters that are not normally distributed (Gelman et al., 2014), the DIC (like the AIC) is biased towards more complex models (Wiecki et al., 2013) and its use is controversial (Spiegelhalter, Best, Carlin & van der Linde, 2014). So, in the experiments above, model fit is reported from visual checks of predicted and observed values, and also using the MSE which is calculated by generating expected values  $E$  from

the posterior model comparing them to the observed data  $O$  as follows:  $\frac{1}{n} \sum_{i=1}^n (O_i - E_i)^2$ . To account for sampling errors introduced during the MCMC sampling process, the MSE is calculated for multiple data sets of expected values (typically 100 to 500 data sets depending on the complexity of the model) and averaged over all the data sets.

In addition, following the recommendations of Kruschke (2011) model comparison is eschewed in favour of parameter estimation, as in most cases the focus of interest is not simply the presence of an effect, but the direction and magnitude of the effect. For example, knowing simply that there is an effect of background noise on resistance to distraction, while interesting in itself, would be less informative than knowing the direction of the effect (is the resistance to distraction increased or decreased?) or the magnitude of the effect (is the change in resistance a large easily measurable effect or not?). In these cases, model comparison is less informative than parameter estimation (Kruschke, 2013) and so parameter estimation and parameter comparison will be used in preference to model comparison.

## A.4 Hypothesis testing and comparisons

The advantage to estimating the posterior distribution for each parameter is that these posteriors represent replications of the experiments that produced the observed data (Kruschke, 2010b) and hypotheses can be tested simply by performing element-wise arithmetic on the MCMC traces examining the resulting distributions (Kruschke, 2010a). For example, to test if two means  $\mu_1$  and  $\mu_2$  are credibly different it is simply necessary to form the ‘posterior difference distribution’ by subtracting the trace of for the distribution of  $\mu_1$  from the trace of the distribution of  $\mu_2$  (i.e., the value of the  $i$ th sample of trace  $\mu_1$  is subtracted from the value of the  $i$ th sample of  $\mu_2$ ) and examining the resulting distribution. If the null value (representing no difference) is outside the credible values for the posterior distribution for the comparison then it can be concluded that the data provides credible evidence for the difference between the two means. More complex comparisons are possible, and multiple comparisons do not require corrections (Kruschke, 2010b; Gelman, Hill & Yajima, 2012) and there are no  $p$ -values (Kruschke, 2010b, 2011, for further arguments against  $p$ -values see Wagenmakers, 2007 and Carver, 1993).

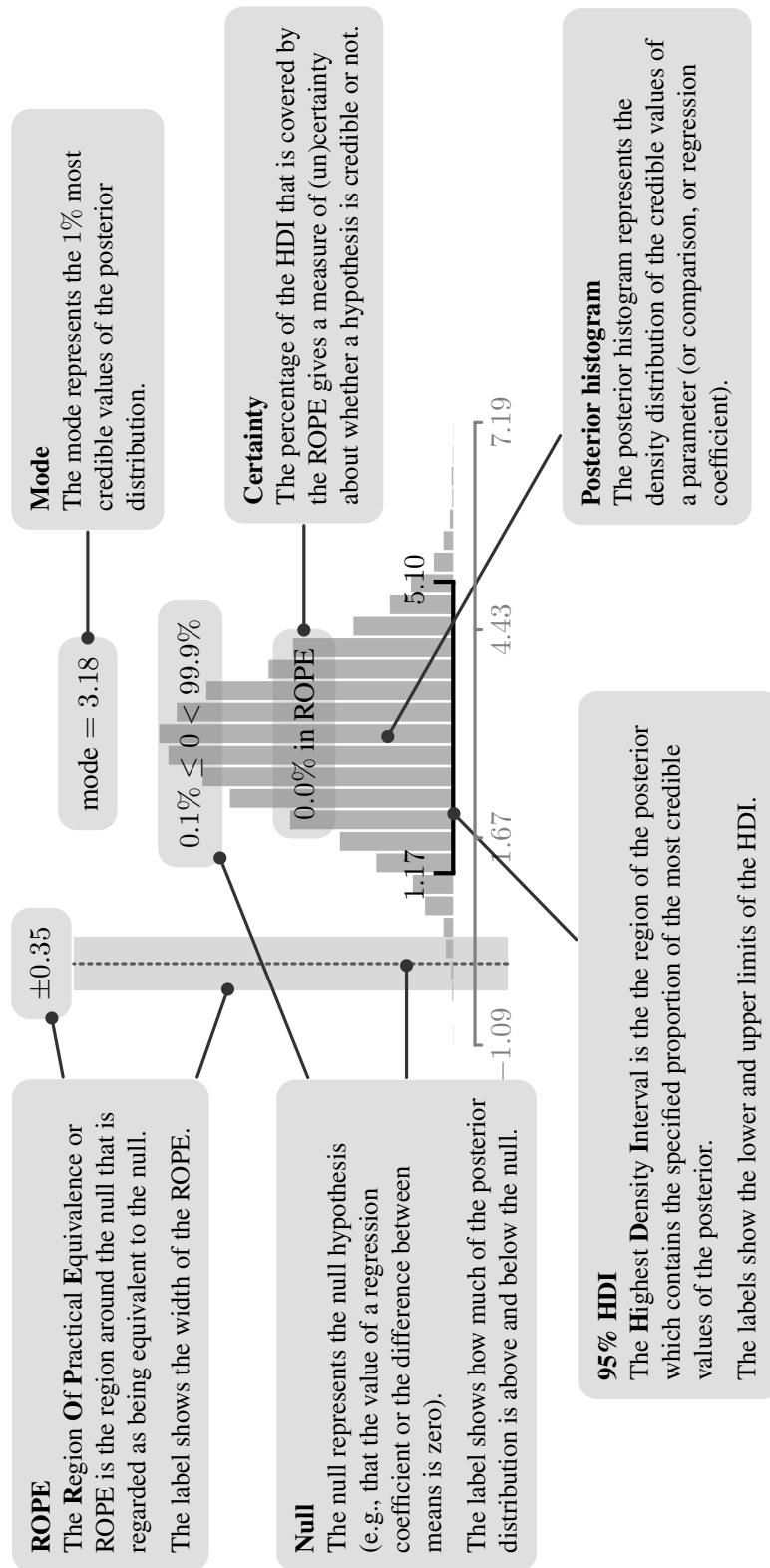


Figure A.3: The anatomy of a posterior plot based on examples from Kruschke (2010a, 2010b, 2011). See section A.4 for details.

However, it is also useful to indicate the size of an effect, not just whether there is an effect (as discussed in section A.3.2). As MCMC produces traces for the standard deviations as well as the means of the parameters, it is simple to derive the posterior distributions of effect sizes by manipulating the traces as follows:  $(\mu_1 - \mu_2) / \sqrt{\frac{1}{2}(\sigma_1^2 + \sigma_2^2)}$ , where  $\sigma_1$  and  $\sigma_2$  represent the traces of the standard deviation posteriors. Assuming the null value (representing no effect) is outside the credible values for the effect size posterior, the effect-size can be approximated as distance of the null value (or the region equivalent to the null value) from the nearest credible value.

A simple test to determine if the null value is one of the credible values of the posterior is to use a posterior plot as shown in figure A.3 (p. 179), and plots of this type are used extensively to illustrate hypothesis tests in the analyses reported above. The posterior histogram represents the posterior density resulting from a comparison of posterior means or regression coefficients. The graphical hypothesis test involves determining whether the null value (which is application dependent, but 0 for all the experiments reported above) is one of the credible values of the posterior (Kruschke, 2010a).

To establish what constitutes a credible value of the posterior distribution involves the construction of a credible interval. One such interval is the Highest Density Interval (HDI) which represents the region which accounts for most of the credible values of the posterior distribution (Curran, 2005; Kruschke, 2011) (given the observed data) and the 95% HDI is used throughout this thesis to define the credible values for the posterior distribution. Note, that the 95% HDI should not be confused with the frequentist 95% confidence interval which indicates that if an infinite number of finite samples of the total population was taken the ‘true’ mean would be expected to occur within the confidence interval on 95% of occasions (Curran, 2005).

When making comparisons, whether between two or means (similar to an ANOVA) or between a mean and a null value (e.g., comparing a regression coefficient) there may be some measurable change in values, but it is necessary to consider whether that change is remarkable, and there may well be cases where a measurable change may not be worth remarking upon. The range of parameter values which are considered unremarkable are those maximum deviations from the null value that are for practical purposes equivalent to the null value (Freedman, Lowe & Macaskill, 1984). This this range is known as a ‘Region Of Practical Equivalence’ (ROPE, e.g., Kruschke, 2010a) also some times called a ‘range of equivalence’ (Spiegelhalter, Freedman

& Parmar, 1994) or “indifference interval” (Brutti & De Santis, 2008, p. 1577). A wider ROPE decreases the probability of mistakenly inferring a credible parameter value as non-credible but also decreasing the probability of mistakenly inferring a non-credible parameter value as credible (Kruschke, 2010a).<sup>1</sup>

The ROPE is usually established using prior knowledge of the domain in which the model is being applied from existing research or from expert opinion (Spiegelhalter et al., 1994). However, in many cases there is no choice but to set an arbitrary ROPE (Kruschke, 2010a). For example, Kruschke (2010b) points out that when testing effect sizes, a difference of 0.1 is considered a small effect (although only for some designs Cohen, 1992) so a ROPE of  $\pm 0.1$  around the null value is an appropriate choice and represents the belief that any effect size less than a small effect is, for practical purposes, equivalent and unremarkable.

As the work in this thesis appears to represent the first example of using a flanker task to measure changes in distraction in a challenging listening task, there is no prior research or expert knowledge to guide the setting of the ROPE size. So, in the current work, the ROPE will be established following the recommendation of Kruschke (2011) and use a  $\pm 0.01$  ROPE that represents a small effect size on the scale of the comparison. This means that differences that constitute an effect size which is smaller than a small effect size may be considered non-credible. When evaluating evidence against a null hypothesis, if the entire ROPE is outside the 95% HDI for the posterior for the difference of the means, the null value will be considered not to be one of the 95% most credible values of the posterior and provide evidence for regarding the rejection of the null hypothesis as plausible. When evaluating evidence for the null hypothesis if the ROPE is inside the 95% HDI for the posterior of for the difference of means the null value will be considered one of the 95% most credible values of the the posterior and thus provides evidence that the null hypothesis is plausible.

In this latter case, however, the null hypothesis can only be rejected with considerable certainty if the entire 95% HDI is inside the ROPE so that 95% of the most credible values of the posterior are equivalent to the null. Typically, some of the ROPE may overlap with the HDI or the entire ROPE is inside the HDI but only overlaps with a small part of it. In these cases it is possible to specify a degree of uncertainty regarding the decision to accept or reject a hypothesis

---

<sup>1</sup>This is analogous to Type-I and Type-II errors in NHST.

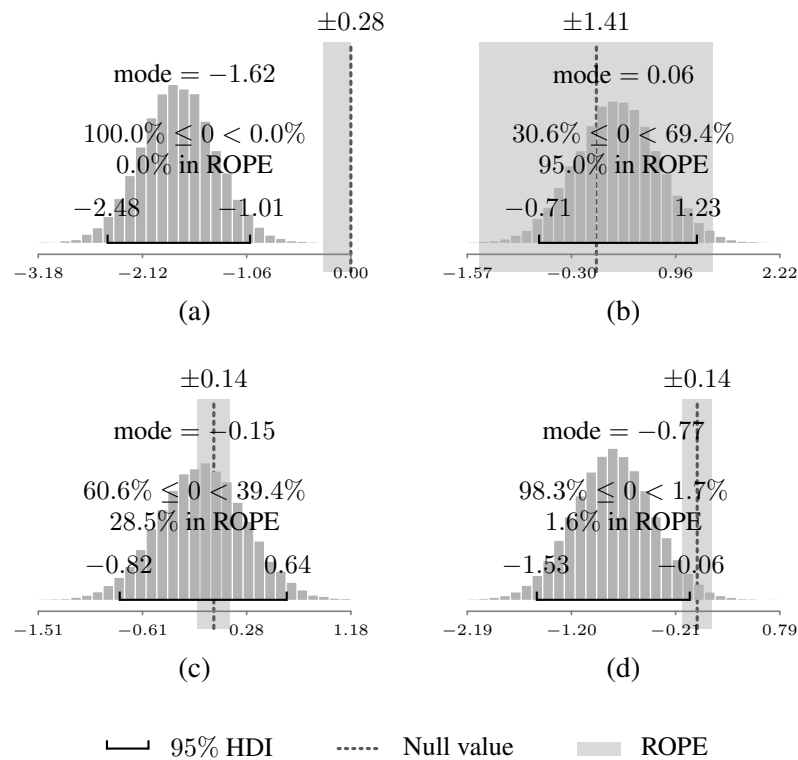


Figure A.4: Four examples of posterior plots for the comparison of two parameters. Comparisons are made by subtracting the posterior trace for one parameter from the posterior trace for the other.

in terms of the percentage of the HDI that overlaps with the ROPE. When testing hypotheses, if 0% of the HDI is inside the ROPE (i.e., 100% is outside the ROPE) then a null hypothesis could be rejected with considerable certainty (or absence of uncertainty) and an alternative hypothesis could be accepted with considerable certainty. Conversely, if 100% of the HDI is inside the ROPE then a null hypothesis could be accepted with considerable certainty and an alternative hypothesis could be rejected with considerable certainty. In more complex cases (which is most realistic applications) less extreme overlaps may occur in which interpreting inferences on the posterior must be done with care. For example, if 50% of the HDI overlaps with the ROPE then no conclusions can be drawn with any certainty.

Figure A.4 shows four examples of posterior plots resulting from the comparison of different pairs of parameters (the data are simulated but representative of some of the situations encountered in the analyses presented above). In (a) the plot demonstrates a comparison with a

clear interpretation: 100% of the posterior is below the null, the null value and its ROPE are not one of the 95% most credible values of the posterior so a plausible interpretation is that there is a credible difference between the two parameters. In addition, with 0% of the HDI overlapping the ROPE there is considerable certainty in this conclusion. In (b) another comparison with a clear interpretation is demonstrated: the posterior is much less unequally distributed around the null, the null value and its ROPE are one of the 95% most credible values of the posterior so no difference is a credible interpretation. In addition, with 100% of the HDI overlapping the ROPE there is considerable certainty in this conclusion as if samples were drawn from the posterior then 100% of them would be equivalent to the null.

Less clear-cut (and unfortunately more realistic) situations are depicted in (c) and (d). In (c) the null value and the ROPE are inside the HDI so a credible interpretation is that there is no difference between the parameters. However, with only approximately 30% of the HDI overlapping the ROPE, if samples were taken from the posterior then approximately 70% of them would not be equivalent to the null. In this case, asserting certainty in the lack of a difference is less easy, but as the ROPE is so close to the mode (the most credible 1% of the posterior) it would be uncontroversial to assign a reasonable degree of certainty to the credible lack of an effect.

The situation (d) represents the case where almost 98% of the posterior is below the null, just under 2% of the HDI overlaps with the ROPE and the null is just outside the HDI, suggesting a credible difference with a high degree of certainty. However, the ROPE is very slightly overlapping the HDI. Recalling that one purpose of the ROPE is to protect against mistakenly assigning credibility to a non credible difference, considerable caution has to be taken to make an interpretation either way. Under a strict decision rule, the difference would be assigned as non-credible, but a reasonable interpretation is that the difference approaches the margins of credibility. An examination of the posterior effect size for the comparison may help decide if the comparison should be considered marginally credible or not.

## **A.5 Reporting Bayesian analysis**

Bayesian analysis is an active area of research (Wiecki et al., 2013) and so unlike the more established NHST there is less agreement regarding what should be reported from a Bayesian analysis. Nevertheless, there many recommendations for reporting Bayesian statistics (e.g.,

Spiegelhalter, Myles, Jones & Abrams, 2000; Sung et al., 2005; Pullenayegum et al., 2012; Kruschke, 2010a). The analyses presented in the work above follows the common elements of these recommendations, and in addition to the inferences regarding the credibility of model parameters and comparisons between these parameters, reports (i) model structures (or references for model structures) (ii) prior distributions (including initial parameter settings) (iii) sampling procedures (i.e., algorithms, iterations, burning, thinning, convergence diagnostics) (iv) specification (graphical or otherwise) of credible intervals, and regions of practical equivalence.



## Appendix B

# Graphs for Bayesian estimation

### B.1 Model specifications

This appendix shows the mode graphs and the specification of the initial values for the priors and hyperpriors used for empirical Bayesian estimation in the analysis of the experimental data above. Graphs are shown for correlations (section B.3), repeated measures Bayesian ANOVA (section B.4), mixed-effects linear regression (section B.5) and mixed-effects logistic regression (section B.6). Note that Gaussian/Normal random variables are specified using mean  $\mu$  and precision  $\tau$ , where  $\tau = \frac{1}{\sigma^2}$ .

### B.2 HDDM

Graphs for the HDDM and the initial values for the hyperpriors are given in (Wiecki et al., 2013). Note that HDDM uses informative priors (rather than the uninformative priors used in the models below). All HDDM models used in the analyses presented above included a parameter  $\eta$  to model inter-trial variability to take into account trial-by-trial variations in stimuli (Voss et al., 2013).

### B.3 Correlation

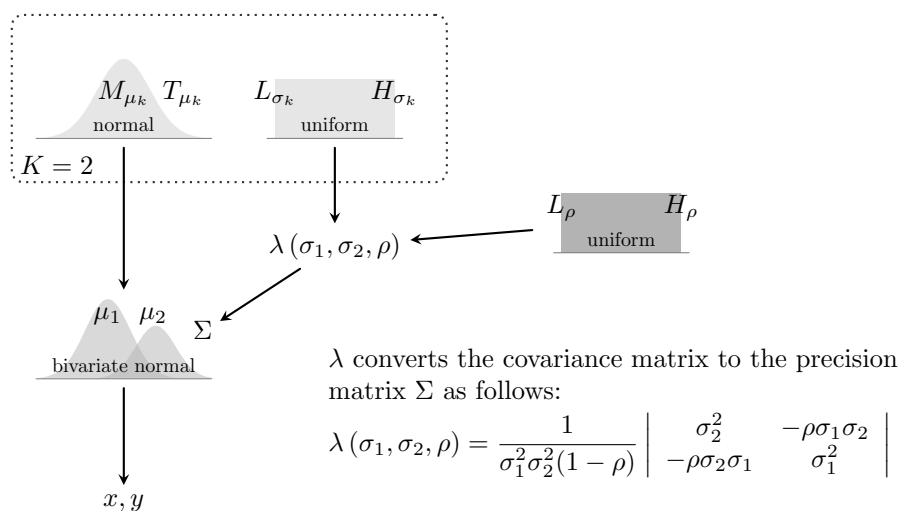


Figure B.1: Graph for empirical Bayesian estimation of the correlation coefficient  $\rho$ .

(Hyper)-prior parameters	Initial value
$L_{\sigma_\rho}$	-1.0
$H_{\sigma_\rho}$	1.0
$L_{\sigma_k}$	0.0
$H_{\sigma_k}$	100.0
$M_{\mu^{\mu_k}}$	0.0
$T_{\mu^{\mu_k}}$	0.0001

Table B.1: Initial values for the (hyper)-prior parameters in figure B.1 All other prior parameters were initialised by random sampling from the hyperprior distributions.

### B.4 Repeated measures BANOVA

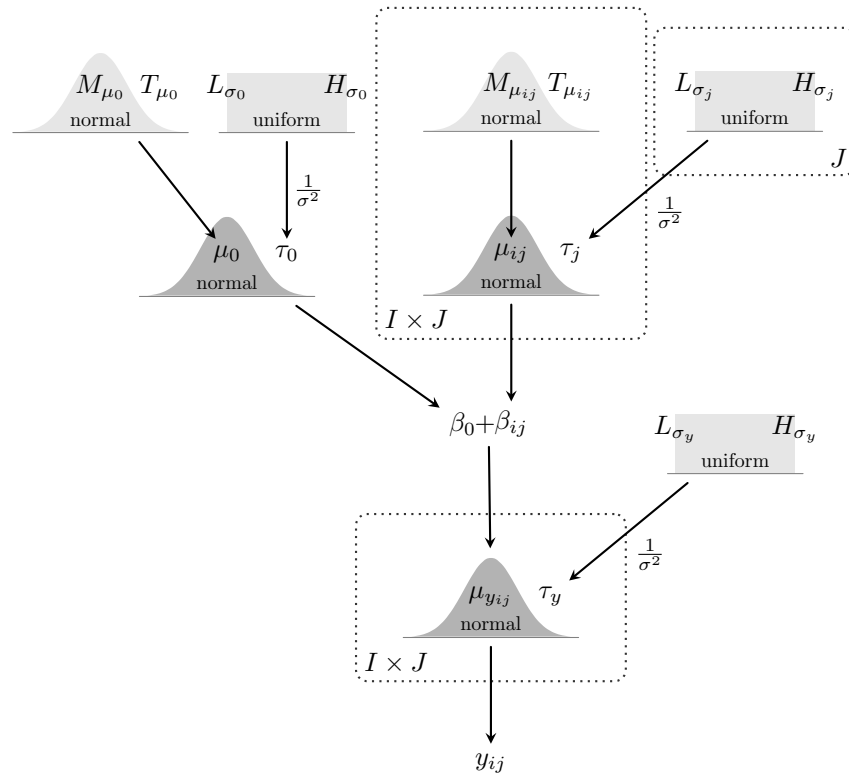


Figure B.2: Graph of prior and hyperprior random variables for a repeated measures Bayesian ANOVA (BANOVA) with  $I$  levels/groups and  $J$  participants. Note, the variance  $\sigma_j^2$  is different for each participant but the same across groups. The plates (the rectangles surrounding around the random variables) indicate duplication of the random variable according to the index in the corner so, for example, there are  $I \times J$  random variables, one for each group where  $\mu_{ij}$  is the mean of the prior for the  $i$ th group and the  $j$ th participant.

Hyperprior parameters			Initial value
$L_{\sigma_0}$	$L_{\sigma_j}$	$L_{\sigma_y}$	0.0
$H_{\sigma_0}$	$H_{\sigma_j}$	$H_{\sigma_y}$	100.0
$M_{\mu_0}$	$M_{\mu_{ij}}$		0.0
$T_{\mu_0}$	$T_{\mu_{ij}}$		0.0001

Table B.2: Initial values for the hyperprior parameters in figure B.2 All other prior parameters were initialised by random sampling from the hyperprior distributions.

### B.5 Mixed effects linear regression

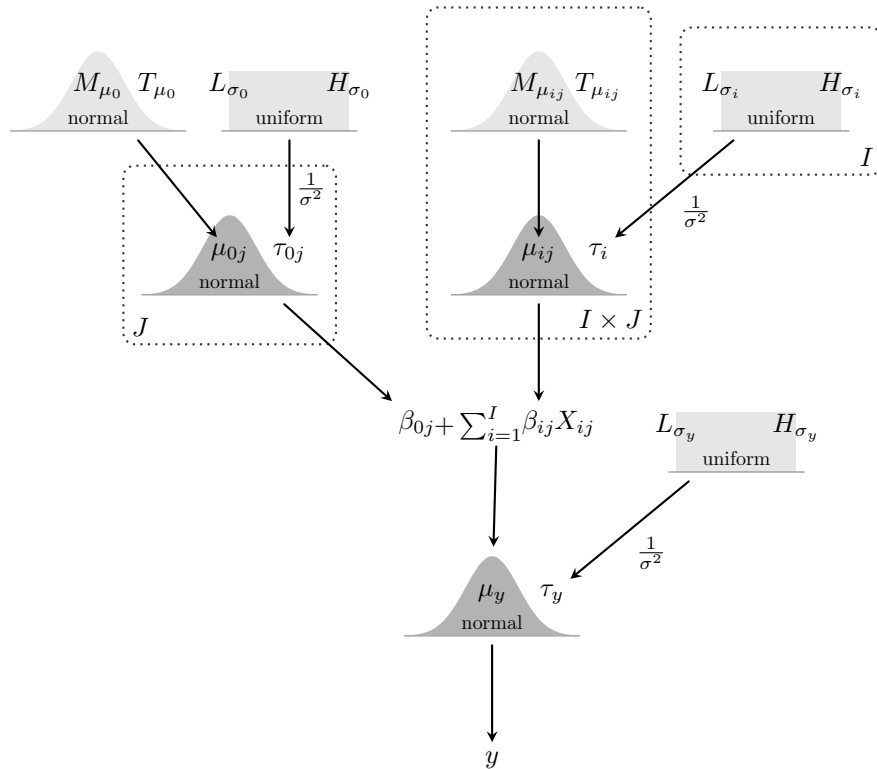


Figure B.3: Graph of prior and hyperprior random variables for a ‘mixed’ effects regression with  $I$  levels/groups and  $J$  participants. Note that here there variance  $\sigma_i^2$  is shared by participants but different across groups, although this does not have to be the case. The plates (the rectangles surrounding around the random variables) indicate duplication of the random variable according to the index in the corner so, for example, there are  $J$   $\mu_0$  random variables for the intercepts, one for each participant.  $X_{ij}$  is the matrix that maps the  $\beta_{ij}$  on to the observed data  $y$ .

Hyperprior parameters			Initial value
$L_{\sigma_j}$	$L_{\sigma_{ij}}$	$L_{\sigma_y}$	0.0
$H_{\sigma_j}$	$H_{\sigma_{ij}}$	$L_{\sigma_y}$	100.0
$M_{\mu_{0j}}$	$M_{\mu_{ij}}$		0.0
$T_{\mu_{0j}}$	$T_{\mu_{ij}}$		0.0001

Table B.3: Initial values for the hyperprior parameters in figure B.3 All other prior parameters were initialised by random sampling from the hyperprior distributions.

## B.6 Mixed effects logistic regression

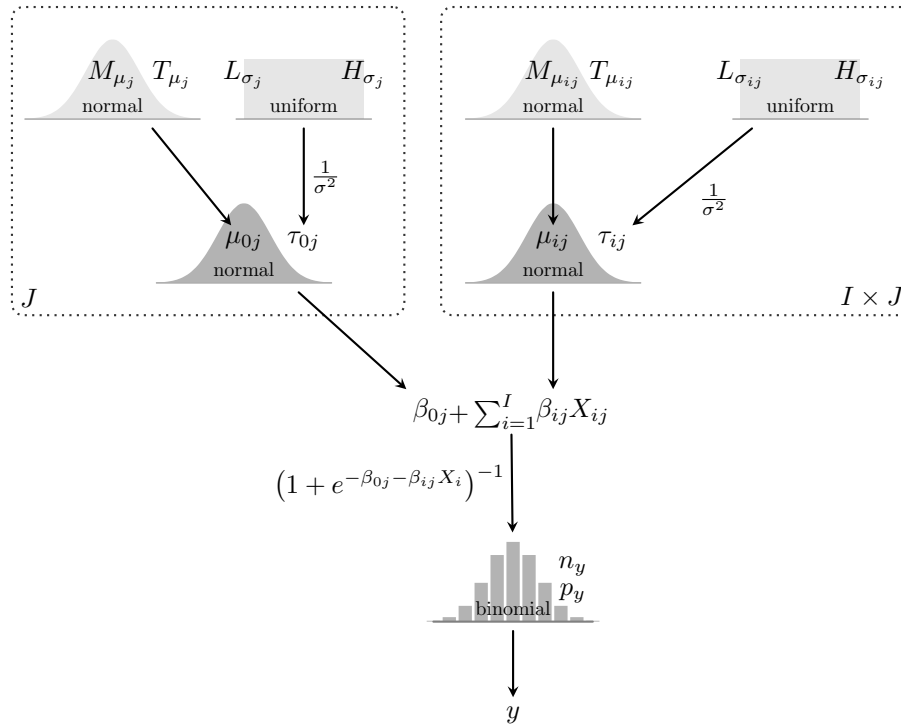


Figure B.4: Graph for a mixed effects logistic regression with  $I$  levels/groups and  $J$  participants. Note, the variance  $\sigma_{ij}^2$  is different for each participant in each level/group but this need not be the case. The plates (the rectangles surrounding around the random variables) indicate duplication of the random variable according to the index in the corner so, for example, there are  $J$   $\mu_0$  random variables for the intercepts, one for each participant.  $X_{ij}$  is the matrix that maps the  $\beta_{ij}$  on to the observed data  $y$  via the ‘inverse logit’ link function.

(Hyper)prior parameters		Initial value
$L_{\sigma_j}$	$L_{\sigma_{ij}}$	0.0
$H_{\sigma_j}$	$H_{\sigma_{ij}}$	100.0
$M_{\mu_{0j}}$	$M_{\mu_{ij}}$	0.0
$T_{\mu_{0j}}$	$T_{\mu_{ij}}$	0.0001
$\beta_{ij}$		0.0

Table B.4: Initial values for the hyperprior parameters in figure B.4 All other prior parameters were initialised by random sampling from the hyperprior distributions.

## Appendix C

# Simulating telephone use in the auditory flanker task

The auditory flanker task used in all the experiments in the current work was designed to simulate an everyday task: listening to speech over a telephone. Figure 2.4 (p. 44) illustrates the listening situation simulated in these experiments: the listener receives a telephone call from a friend. As her friend speaks (the target) a person in front of the listener also speaks (the distractor) and the listener must attend to the target while ignoring the distractor. This situation is simulated throughout the experiments reported above with the target and distractor presented using headphones.

Although some auditory research emulates listening situations in free field, for example, using speakers positioned around the listener in a sound proof listening booth (e.g., Chan et al., 2005). The choice of simulating the listening environment over headphones is based on a criterion of simplicity: should a task prove to be useful, a test administered over headphones (e.g., Brungart & Simpson, 2007) would be simpler and less expensive than a test requiring multiple precisely position loudspeakers in an anechoic chamber (e.g., Seeber, Kerber & Hafter, 2010).

### C.1 Simulating the distractor speaker using head-related transfer functions

To simulate distracting speech coming from in front of the listener, one could simply use a stereo signal (cf. Francis, 2010). However, the approach taken in this work is to use a *head-related*

*transfer function* (HRTF) which characterises the spectral and temporal information that the auditory system uses to determine the location of a sound in the free-field as it is filtered by the torso, head, and external parts of the ear (pinna) before it enters the ear (Cheng & Wakefield, 2001). In particular it can model the difference in intensity and timing as a free-field signal reaches the left and the right ears. The HRTF is derived by taking the Fourier transform of the head-related impulse response (HRIR) obtained by recording a free field impulse signal from a loudspeaker using a microphone situated at the entrance to the ear canal. To then simulate a sound originating from the free-field, the Fourier transform of the sound is multiplied by the HRTF or the sound itself is convolved with the HRIR (Kayser et al., 2009). Although the HRIR can be measured for individual listeners, HRIR can also be measured using a dummy head and torso with rubber pinna, with the loudspeaker placed at various angles originating from the centre of the head in the horizontal plane (i.e., the angles of *azimuth*) and the vertical plane (i.e., angles of *elevation*), and sometimes also at different distances from the centre of the head (Gardner & Martin, 1995; Kayser et al., 2009; Drullman & Bronkhorst, 2000).

Throughout this thesis, the HRIRs used are taken from the MIT KEMAR database (Gardner & Martin, 1995). The term *projected* will be used to describe a signal that has been convolved with the appropriate HRIR for a particular azimuth and elevation for both the left and right ear, and the results mixed to the left and right channels of a stereo signal, respectively. The main advantage of using HRTFs is that multiple free-field sound sources can be simulated at a range of azimuths and elevations over headphones, although this was not used in the experiments reported above. However, as HRTFs are usually derived from measurements taken using a head-and-torso dummy, they do not take into account the fact that the dimensions of heads, shoulders, torsos, and pinna, differ considerably between listeners. Differences in physiological dimensions for any one of these features will result in quantitatively different HRTFs for each listener, and each listener's auditory system will have adapted to the subtle differences in temporal and spectral information that result from the individual HRTF (Drullman & Bronkhorst, 2000). This may account (at least in part) for misjudgements of distance (although mostly the near field, Kan, Jin and van Schaik, 2009), and also 'front-back' confusions when using HRTFs (Wenzel, Wightman & Kistler, 1991). However, in the experiments reported below the projected location of the distracting speech is kept constant and location judgements are not part of the listening tasks, so these problems may be considered irrelevant.

Furthermore, Drullman and Bronkhorst (2000) found no significant difference in performance (i.e., for intelligibility, speaker recognition, and localisation) between using non-individualised and individualised HRTFs.

## **C.2 Simulating telephone use with a monaural speaker**

Telephone use involves monaural listening when using a handset (Noonan & Axelrod, 1981; Mantokoudis et al., 2012) or an ear-piece (Picou & Ricketts, 2011; Ferlazzo, Fagioli, Nocera & Sdoia, 2008), and binaural listening when using ear-phones (Picou & Ricketts, 2011) or ‘hands-free’ devices (Strayer & Johnston, 2001; Briem & Hedman, 1995). However, listeners’ attentional requirements are similar whether they are using a telephone monaurally or binaurally (Strayer & Johnston, 2001). In the experiments reported below telephone use is simulated using a monaural speaker.

Given the advantages of HRTFs described above, it might be expected that HRTFs would be used to simulate the location of a telephone held to one ear. However, HRTF databases do not usually contain data for near-field sources (i.e., less than 1 m) as changes in the HRTF in the near-field are much greater than in the far-field, requiring an extra number of measurements that would increase the size of the database fourfold (Brungart, 2002). Although techniques for deriving near-field HRTFs from far-field HRTFs exist (e.g., Kan et al., 2009), the smallest distance considered appears to 10–12 cm (Kan et al., 2009; Brungart & Rabinowitz, 1999), which — by informal observation — is much greater than the distance that users typically hold a handset to their head. Furthermore, studies investigating physiological effects of electromagnetic radiation from mobile phones have considered handset to head distances of 5.4–20 mm (e.g., De Salles, Bulla & Fernández Rodríguez, 2006; Bernardi, Cavagnaro, Pisa & Piuzzi, 2000; Martínez-Búrdalo, Martín, Anguiano & Villar, 2004; Jensen & Rahmat-Samii, 1995), and some informal calculations based on the recommendations for handset position for standardised evaluation of telephones (ITU-T Rec. P.64, 2007, annex E) suggest a distance of 13 mm, so HRTFs collected from a distance of 10 cm would clearly be inadequate.

So, in the absence of relevant HRTF data for distances less than 10 cm, and the considerable inconvenience of collecting novel HRTFs (Brungart, 2002), the approach taken in the experiments reported below is to approximate telephone listening using the left or right channel of a stereo signal presented over headphones. This is clearly some way from a realistic hand-held



telephone environment (even considering that the entire acoustic scene was being simulated) as it ignores any acoustic effects caused the closure between the handset and the pinna on both the target and distractor speech, and also the possibility that an imperfect closure between the handset and the pinna may lead to the target being audible to the non-target ear (although extrapolating the measurements shown in Brungart and Rabinowitz, 1999, figure 9, p. 1474 suggests that the target signal would be inaudible in the target ear). Ultimately, however, this approach was considered superior to other approaches claiming to simulating telephone use with monaural speech, for example, Mantokoudis et al. (2012), who used a free field source 1 m in front of the listener and an ear-plug in the non-target ear.

## Appendix D

# Cognitive failures questionnaire

The 'Cognitive Failures Questionnaire' (CFQ — Broadbent et al., 1982) was administered using a Blackberry playbook tablet. This appendix summarises the instructions, questions and scoring for the CFQ.

### Instructions

Participants were presented with following instruction taken directly from the original paper version of the CFQ:

The following questions are about minor mistakes which everyone makes from time to time, but some of which happen more often than others. We want to know how often these things have happened to you in the last six months. Please circle the appropriate number. (Broadbent et al., 1982, p. 15, appendix 1).

### Scoring

The CFQ scoring scale is an integer scale from 0 (i.e., not susceptible to distraction) to 100 (i.e., very susceptible to distraction). The scores and anchors for individual questions are 4 - Very often, 3 - Quite often 2 - Occasionally 1 - Rarely and 0 - Very rarely, and the overall CFQ score is calculated by summing the score for each question.

### Questions

1. Do you read something and find you haven't been thinking about it and must read it again?
2. Do you find you forget why you went from one part of the house to the other?
3. Do you fail to notice signposts on the road?

4. Do you find you confuse right and left when giving directions?
5. Do you bump into people?
6. Do you find you forget whether you've turned off a light or a fire or locked the door?
7. Do you fail to listen to people's names when you are meeting them?
8. Do you say something and realise afterwards that it might be taken as insulting?
9. Do you fail to hear people speaking to you when you are doing something else?
10. Do you lose your temper and regret it?
11. Do you leave important letters unanswered for days?
12. Do you find you forget which way to turn on a road you know well but rarely use?
13. Do you fail to see what you want in a supermarket (although it's there)?
14. Do you find yourself suddenly wondering whether you've used a word correctly?
15. Do you have trouble making up your mind?
16. Do you find you forget appointments?
17. Do you forget where you put something like a newspaper or a book?
18. Do you find you accidentally throw away the thing you want and keep what you meant to throw away — as in the example of throwing away the matchbox and putting the used match in your pocket?
19. Do you daydream when you ought to be listening to something?
20. Do you find you forget people's names?
21. Do you start doing one thing at home and get distracted into doing something else (unintentionally)?
22. Do you find you can't quite remember something although it's 'on the tip of your tongue'?
23. Do you find you forget what you came to the shops to buy?
24. Do you drop things?
25. Do you find you can't think of anything to say?

## Bibliography

- Action on Hearing Loss. (2011). Facts and figures on hearing loss and tinnitus. Retrieved August 12, 2013, from <http://www.actiononhearingloss.org.uk/supporting-you/factsheets-and-leaflets/deaf-awareness.aspx>
- Ahmed, L. & de Fockert, J. W. (2012). Focusing on attention: the effects of working memory capacity and load on selective attention. *Public Library of Science (PLoS ONE)*, 7(8), e43101. doi:10.1371/journal.pone.0043101
- Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2), 255–265. doi:10.2307/2334537
- Alain, C. & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience*, 5, 202–212.
- Alain, C. & Izenberg, A. (2003). Effects of attentional load on auditory scene analysis. *Journal of Cognitive Neuroscience*, 15(7), 1063–1073. doi:10.1162/089892903770007443
- Allen, J. B. (2005). *Articulation and intelligibility*. Morgan & Claypool.
- Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50, 5–43.
- Baldwin, C. L. (2012). *Auditory cognition and human performance: research and applications*. CRC Press.
- Barker, J. & Shao, X. (2009, March). Energetic and informational masking effects in an audiovisual speech recognition system. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3), 446–458. doi:10.1109/TASL.2008.2011534
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Benesty, J., Makino, S. & Chen, J. (Eds.). (2005). *Speech enhancement*. Springer.
- Benesty, J., Sondhi, M. M. & Huang, Y. (Eds.). (2008). *Springer handbook of speech processing*. Springer. doi:10.1007/978-3-540-49127-9
- Benoni, H. & Tsal, Y. (2010). Where have we gone wrong? perceptual load does not affect selective attention. *Vision Research*, 50, 1292–1298. doi:10.1016/j.visres.2010.04.018
- Benoni, H. & Tsal, Y. (2013). Conceptual and methodological concerns in the theory of perceptual load. *Frontiers in Psychology*, 4(522). doi:10.3389/fpsyg.2013.00522
- Bentler, R. & Chiou, L.-K. (2006). Digital noise reduction: an overview. *Trends in Amplification*, 10, 67–82. doi:10.1177/1084713806289514

- Bernardi, P., Cavagnaro, M., Pisa, S. & Piuze, E. (2000). Specific absorption rate and temperature increases in the head of a cellular-phone user. *IEEE Transactions on microwave theory and techniques*, 48(7), 1118–1126. doi:10.1109/22.848494
- Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H. & Corona-Strauss, F. I. (2013). Neural correlates of listening effort related factors: influence of age and hearing impairment. *Brain Research Bulletin*, 91, 21–30. doi:10.1016/j.brainresbull.2012.11.005
- Bertoli, S. & Bodmer, D. (2014). Novel sounds as a psychophysiological measure of listening effort in older listeners with and without hearing loss. *Clinical Neurophysiology*, 125, 1030–1041. doi:10.1016/j.clinph.2013.09.045
- Best, V., Gallun, F. J., Carlile, S. & Shinn-Cunningham, B. G. (2007). Binaural interference and auditory grouping. *Journal of the Acoustical Society of America*, 121(2), 1070–1076.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Boike, K. T. & Souza, P. E. (2000). Effect of compression ratio on speech recognition and speech-quality ratings with wide dynamic range compression amplification. *Journal of Speech Language and Hearing Research*, 43, 456–468.
- Boothroyd, A. & Nittrouer, S. (1988). Context effects in phoneme and word recognition by young children and older adults. *Journal of the Acoustical Society of America*, 84(1), 101–114.
- Borg, G. (1990). Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian Journal of Work and Environmental Health*, 16(supplement 1), 55–58.
- Borowsky, R., Esopenko, C., Gould, L., Kuhlmann, N., Sarty, G. & Cummine, J. (2013). Localisation of function for noun and verb reading: converging evidence for shared processing from fMRI activation and reaction time. *Language and Cognitive Processes*, 28(6), 789–809. doi:10.1080/01690965.2012.665466
- Boulenger, V., Hoen, M., Ferragne, E., Pellegrino, F. & Meunier, F. (2010). Real-time lexical competitions during speech-in-speech comprehension. *Speech Communication*, 52, 246–253. doi:10.1016/j.specom.2009.11.002
- Bourland-Hicks, C. & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language and Hearing Research*, 45(3), 573–84.
- Bregman, A. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge, Mass.: MIT Press.
- Briem, V. & Hedman, L. R. (1995). Behavioural effects of mobile telephone use during simulated driving. *Ergonomics*, 38(12), 2536–2562.
- Broadbent, D. E. (1958). *Perception and communication*. Pergamon.
- Broadbent, D. E., Cooper, P. F., FitzGerald, P. & Parkes, K. R. (1982). The cognitive failures questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, 21, 1–16.
- Brons, I., Houben, R. & Dreschler, W. A. (2012). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear & Hearing*, 34(1), 29–41.

- Brookes, M. (2003). VOICEBOX: speech processing toolbox for MATLAB. Retrieved June 30, 2011, from <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>
- Brooks, S. P. (1998). Markov Chain Monte Carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1), 69–100.
- Brungart, D. S. (2002). Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environments*, 11(1), 93–106. doi:10.1162/105474602317343686
- Brungart, D. S. & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3), 1465–1479. doi:10.1121/1.427180
- Brungart, D. S. & Simpson, B. D. (2007). Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task. *Journal of the Acoustical Society of America*, 122(3), 1724–1734.
- Brungart, D. S., Simpson, B. D., Ericson, M. A. & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, 110(5), 2528–2548.
- Brutti, P. & De Santis, F. (2008). Robust Bayesian sample size determination for avoiding the range of equivalence in clinical trials. *Journal of Statistical Planning and Inference*, 138, 1577–1591.
- Bruyer, R. & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, 51(1), 5–13.
- BSA. (2011). Recommended procedure: pure-tone air-conduction and bone-conduction threshold audiometry with and without masking. Amended 6<sup>th</sup> February 2012 British Society of Audiology.
- Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33, 261–304. doi:10.1177/0049124104268644
- Carlyon, R. & Cusack, R. (2005). Effects of attention on auditory perceptual organization. In L. Itti, G. Rees & J. Tsotsos (Eds.), *The neurobiology of attention*. Academic Press.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61(A), 287–292.
- Chan, J. S., Merrifield, K. & Spence, C. (2005). Auditory spatial attention assessed in a flanker interference task. *Acta Acustica United with Acustica*, 91, 554–563.
- Chateau, N., Gros, L., Durin, V. & Macé, A. (2006). Redrawing the link between customer satisfaction and speech quality. In *2nd ISCA/DEGA tutorial and research workshop on perceptual quality of systems* (pp. 89–95).
- Cheng, C. I. & Wakefield, G. H. (2001). Representations of HRTFs in time, frequency, and space. *Journal of the Audio Engineering Society*, 49(4), 231–249.
- Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustic Society of America*, 25, 975–979.
- Cheyne, J. A., Carriere, J. S. A. & Smilek, D. (2006). Absent-mindedness: lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition*, 15, 578–592.

- Chung, K. (2004). Challenges and recent developments in hearing aids : part I. speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends in Amplification*, 8, 83–124. doi:[10.1177/108471380400800302](https://doi.org/10.1177/108471380400800302)
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Colflesh, G. J. H. & Conway, A. R. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic Bulletin & Review*, 14, 699–703.
- Conway, A. R., Cowan, N. & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, 12(5), 331–335.
- Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O. & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12(5), 769–786.
- Cooke, M., Lecumberri, M. L. G. & Barker, J. (2008). The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *Journal of the Acoustical Society of America*, 123(1), 414–427.
- Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A. & Conway, A. R. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51(1), 42–100. doi:[10.1016/j.cogpsych.2004.12.001](https://doi.org/10.1016/j.cogpsych.2004.12.001)
- Cowles, M. K. & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91, 883–904.
- Curran, J. M. (2005). An introduction to bayesian credible intervals for sampling error in dna profiles. *Law, Probability and Risk*, 4, 115–126. doi:[10.1093/lpr/mgi009](https://doi.org/10.1093/lpr/mgi009)
- Cusack, R., Deeks, J., Aikman, G. & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 643–656. doi:[10.1037/0096-1523.30.4.643](https://doi.org/10.1037/0096-1523.30.4.643)
- Dalton, P., Santangelo, V. & Spence, C. (2009). The role of working memory in auditory selective attention. *The Quarterly Journal of Experimental Psychology*, 62(11), 2126–2132. doi:[10.1080/17470210903023646](https://doi.org/10.1080/17470210903023646)
- Darwin, C. J. & Hukin, R. W. (1999). Auditory objects of attention: the role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance*, 25(3), 617–629. doi:[10.1037/0096-1523.25.3.617](https://doi.org/10.1037/0096-1523.25.3.617)
- Davidson, D. J. & Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta Psychologica*, 144, 83–96. doi:[10.1016/j.actpsy.2013.04.016](https://doi.org/10.1016/j.actpsy.2013.04.016)
- Davidson-Pilon, C. (2015). Bayesian methods for hackers: using Python and PyMC. Retrieved 2010, from <http://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/>
- De Salles, A. A., Bulla, G. & Fernández Rodríguez, C. E. (2006). Electromagnetic absorption in the head of adults and children due to mobile phone operation close to the head. *Electromagnetic Biology and Medicine*, 25, 349–360. doi:[10.1080/15368370601054894](https://doi.org/10.1080/15368370601054894)

- Dhamani, I., Leung, J., Carlile, S. & Sharma, M. (2013). Switch attention to listen. *Scientific reports*, 3. doi:[10.1038/srep01297](https://doi.org/10.1038/srep01297)
- Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science*, 6(3), 274–290. doi:[10.1177/1745691611406920](https://doi.org/10.1177/1745691611406920)
- Dishon-Berkovits, M. & Algom, D. (2000). The Stroop effect: It is not the robust phenomenon that you have thought it to be. *Memory & Cognition*, 28(8), 1437–1449.
- Dittrich, K. & Stahl, C. (2012). Selective impairment of auditory selective attention under concurrent cognitive load. *Journal of Experimental Psychology: Human Perception and Performance*, 38(3), 618–627.
- Downs, D. W. (1982). Effects on hearing aid use on speech discrimination and listening effort. *Journal of Speech and Hearing Disorders*, 47, 189–193.
- Downs, D. W. & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research*, 21(4), 702–714.
- Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92, 53–78.
- Driver, J. & Tipper, S. P. (1989). On the nonselectivity of “selective” seeing: contrasts between interference and priming in selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 15(2), 304–304.
- Drullman, R. & Bronkhorst, A. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America*, 107(4), 2224–2235.
- Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S. & Kidd Jr., G. (2003). Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *Journal of the Acoustical Society of America*, 114(1).
- Dyson, B. J. & Quinlan, P. T. (2003). Feature and conjunction processing in the auditory modality. *Perception & Psychophysics*, 65(2), 254–272.
- Edwards, B. (2007). The future of hearing aid technology. *Trends in Amplification*, 11(1), 31–46.
- Egan, J. (1948). Articulation testing methods. *Laryngoscope*, 58, 955–981.
- Ephraim, Y. & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal processing*, ASSP-33(2), 443–445.
- Eriksen, B. A. & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1), 143–149.
- Eriksen, C. W. (1995). The flankers tasks and response competition: a useful tool for investigating a variety of cognitive problems. *Visual Cognition*, 2(3), 101–118.
- Fairbanks, G. (1958). Test of phonemic differentiation: the rhyme test. *Journal of the Acoustical Society of America*, 30(7).
- Feinberg, F. M. & Gonzalez, R. (2012). Bayesian modeling for psychologists: an applied approach. In H. Cooper (Ed.), *APA handbook of research methods in psychology vol.2 : research designs*. Magination Press, (American Psychological Association). doi:[10.1037/13620-024](https://doi.org/10.1037/13620-024)



- Ferlazzo, F., Fagioli, S., Nocera, F. D. & Sdoia, S. (2008). Shifting attention across near and far spaces: implications for the use of hands-free cell phones while driving. *Accident Analysis and Prevention*, 40, 1859–1864. doi:10.1016/j.aap.2008.07.003
- Feuerstein, J. (1992). Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort. *Ear and Hearing*, 13(2), 80–86.
- Finkelman, J. M. & Glass, D. C. (1970). Reappraisal of the relationship between noise and human performance by means of a subsidiary task measure. *Journal of Applied Psychology*, 54(3), 211–213.
- Fisher, C. R. & Wolfe, C. R. (2012). Teaching Bayesian parameter estimation, bayesian model comparison and null hypothesis significance testing using spreadsheets. *Spreadsheets in Education (eJSiE)*, 5(3), Article 3. Retrieved from <http://epublications.bond.edu.au/ejsie/vol5/iss3/3>
- Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Oliver and Boyd.
- Fletcher, H. & Steinberg, J. C. (1929). Articulation testing methods. *Bell System Technical Journal*, 8(4), 806–854.
- Forster, S. & Lavie, N. (2007). High perceptual load makes everybody equal eliminating individual differences in distractibility with load. *Psychological Science*, 18(5), 377–381.
- Fowler, C. A. (1979). “Perceptual centers” in speech production and perception. *Perception & Psychophysics*, 25(375), 375–388.
- Francis, A. L. (2010). Improved segregation of simultaneous talkers differentially affects perceptual and cognitive capacity demands for recognizing speech in competing speech. *Attention, Perception, & Psychophysics*, 72(2), 501–516. doi:10.3758/APP.72.2.501
- Francis, A. L. & Nusbaum, H. C. (1999). The effect of lexical complexity on intelligibility. *International Journal of Speech Technology*, 3, 15–25.
- Freedman, L. S., Lowe, D. & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, 40(3), 575–586.
- Frigge, M., Hoaglin, D. C. & Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician*, 43(1), 50–54. doi:10.2307/2685173
- Gaizauskas, R. (1998). Evaluation in language and speech technology. *Computer Speech and Language*, 12, 249–262.
- Gardner, W. G. & Martin, K. D. (1995). HRTF measurements of a KEMAR. *Journal of the Acoustical Society of America*, 97(6), 3907–3908.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. 27, 2865–2873. doi:10.1002/sim.3107
- Gelman, A., Hill, J. & Yajima, M. (2012). Why we (usually) don’t have to worry about multiple comparisons. 5, 189–210. doi:10.1080/19345747.2011.618213
- Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016. doi:10.1007/s11222-013-9416-2
- Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

- Gelman, A. & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1), 8–38. doi:10.1111/j.2044-8317.2011.02037.x
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith (Eds.), *Bayesian statistics 4*. Clarendon Press, Oxford, UK.
- Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.
- Glickman, M. E. & van Dyk, D. A. (2007). Basic Bayesian methods. In W. T. Ambrosius (Ed.), *Methods in molecular biology* (Vol. 404: Topics in Biostatistics). Humana Press Inc., Totowa, NJ.
- Gomes, H., Barrett, S., Duff, M., Barnhardt, J. & Ritter, W. (2008). The effects of interstimulus interval on event-related indices of attention: an auditory selective attention test of perceptual load theory. *Clinical Neurophysiology*, 119(3), 542–55. doi:10.1016/j.clinph.2007.11.014
- Goodman, D. & Nash, R. (1982). Subjective quality of the same speech transmission conditions in seven different countries. *Communications, IEEE Transactions on*, 30(4), 642–654. doi:10.1109/TCOM.1982.1095507
- Gosselin, P. A. & Gagné, J.-P. (2010). Use of a dual-task paradigm to measure listening effort. *Canadian Journal of Speech-Language Pathology and Audiology*, 34(1), 43–51.
- Goulding, M. & Bird, J. (1990). Speech enhancement for mobile telephony. *Vehicular Technology, IEEE Transactions on*, 39(4), 316–326. doi:10.1109/25.61353
- Gracharov, V. & Kleijn, W. B. (2008). Speech quality assessment. In J. Benesty, M. M. Sondhi & Y. Huang (Eds.), *Handbook of speech processing* (Chap. 5, pp. 83–99). Springer.
- Greenspan, S. L., Bennet, R. W. & Syrdal, A. K. (1998). An evaluation of the diagnostic rhyme test. *International Journal of Speech Technology*, 2, 201–214.
- Griffiths, T. D. & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5, 887–892. doi:10.1038/nrn1538
- Guerreiro, M. J. S., Murphy, D. R. & Gerven, P. W. M. V. (2010). The role of sensory modality in age-related distraction: a critical review and a renewed view. *Psychological Bulletin*, 136(6), 975–1022. doi:10.1037/a0020731
- Gutschalk, A., Micheyl, C. & Oxenham, A. J. (2008). Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biology*, 6(6), 1160–1165. doi:10.1371/journal.pbio.0060138
- Harsin, C. (1997). Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception & Psychophysics*, 59(2), 243–251.
- Hazeltine, E., Poldrack, R. & Gabrieli, J. D. E. (2000). Neural activation during response competition. *Journal of Cognitive Neuroscience*, 12(Supplement 2), 118–129.
- Hecker, M. H. L., Stevens, K. N. & Willaims, C. E. (1966). Measurements of reaction time in intelligibility tests. *Journal of the Acoustical Society of America*, 39(6), 1188–1189.

- Hervey, A. S., Epstein, J. N., Curry, J. F., Tonev, S., Arnold, L. E., Conners, C. K., . . . Hechtman, L. (2006). Reaction time distribution analysis of neuropsychological performance in an ADHD sample. *Child Neuropsychology*, *12*(2), 125–140.
- Hintze, J. L. & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184. doi:[10.2307/2685478](https://doi.org/10.2307/2685478)
- Hirschman, L. (1998). The evolution of evaluation: lessons from the message understanding conferences. *Computer Speech and Language*, *12*, 281–305.
- Hirsh, I. J., Reynolds, E. G. & Joseph, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America*, *26*(4), 530–538.
- Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, *34*(5), 523–34. doi:[10.1097/AUD.0b013e31828003d8](https://doi.org/10.1097/AUD.0b013e31828003d8)
- Houben, R., van Doorn-Bierman, M. & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology*, *52*, 753–761. doi:[10.3109/14992027.2013.832415](https://doi.org/10.3109/14992027.2013.832415)
- House, A. S., Williams, C. E., Hecker, M. H. L. & Kryter, K. D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America*, *37*(1), 158–166.
- Howard, C. S., Munro, K. J. & Plack, C. J. (2010). Listening effort at signal-to-noise ratios that are typical of the school classroom. *International Journal of Audiology*, 1–5.
- Howell, P. (1988). Prediction of p-center location from the distribution of energy in the amplitude envelope: I. *Perception & Psychophysics*, 90–93.
- Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29*(2), 296–305.
- Hu, J. & Lee, M. (2009). Speech enhancement for mobile phones based on the imparity of two-microphone signals. In *International Conference on Information and Automation, ICIA '09* (pp. 606–611). doi:[10.1109/ICINFA.2009.5204994](https://doi.org/10.1109/ICINFA.2009.5204994)
- Hu, Y. & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, *49*, 588–601. doi:[10.1016/j.specom.2006.12.006](https://doi.org/10.1016/j.specom.2006.12.006)
- Hughes, A. & Trudgill, P. (1997). *English accents and dialects*. Arnold.
- Hughes, R. W., Hurlstone, M. J., Marsh, J. E., Vachon, F. & Jones, D. M. (2012). Cognitive control of auditory distraction: impact of task difficulty, foreknowledge, and working memory capacity supports duplex-mechanism account. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 539–53. doi:[10.1037/a0029064](https://doi.org/10.1037/a0029064)
- Hughes, R. W. & Jones, D. M. (2003). Indispensable benefits and unavoidable costs of unattended sound for cognitive functioning. *Noise & Health*, *6*(21), 63–76.
- Ihlefeld, A. & Shinn-Cunningham, B. G. (2008). Spatial release from energetic and informational masking in a selective speech identification task. *Journal of the Acoustical Society of America*, *123*(6), 4369–4379.
- ITU-T Rec. P.64. (2007, November). Determination of sensitivity/frequency characteristics of local telephone systems. International Telecommunication Union.

- ITU-T Rec. P.835. (2003, November). Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. International Telecommunication Union.
- James, C., Cheesman, M., Cornelisse, L. & Miller, L. (1994). Response times to sentence verification tasks (SVTs) as a measure of effort in speech perception. In *Fifth Australian international conference on speech science & technology II* (pp. 600–605).
- Jekosch, U. (2002). Speech quality in modern telecommunications. *Revista de Acústica*, 33(3 & 4), 26–30.
- Jekosch, U. (2010). *Voice and speech quality perception: assessment and evaluation*. Springer.
- Jensen, M. & Rahmat-Samii, Y. (1995). EM interaction of handset antennas and a human in personal communications. *Proceedings of the IEEE*, 83(1), 7–17. doi:10.1109/5.362755
- Kahneman, D. (1973). *Attention and effort*. Prentice Hall.
- Kalikow, D. N., Stevens, K. N. & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61(5), 1337–1351.
- Kan, A. H., Jin, C. T. & van Schaik, A. (2009). A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. *The Journal of the Acoustical Society of America*, 125(4), 2233–42. doi:10.1121/1.3081395
- Kayser, H., Ewert, S. D., Anemuller, J., Rohdenburg, T., Hohmann, V. & Kollmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *European Association for Signal Processing Journal on Advances in Signal Processing*, 2009, 10. doi:10.1155/2009/298605
- Kidd Jr., G., Richards, V. M., Mason, C. R., Gallun, F. J. & Huang, R. (2008). Informational masking increases the costs of monitoring multiple channels. *JASA express letters*, 124(4), EL223–EL229.
- Klauer, K. C., Voss, A., Schmitz, F. & Teige-Mocigemba, S. (2007). Process components of the implicit association test: a diffusion-model analysis. *Journal of Personality and Social Psychology*, 93(3), 353–368. doi:10.1037/0022-3514.93.3.353
- Kondo, K. (2012). *Subjective quality measurement of speech: its evaluation, estimation and applications*. Springer.
- Konig, C. J., Buhner, M. & Murling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. *Human Performance*, 18(3), 243–266. doi:10.1207/s15327043hup1803\_3
- Krajbich, I., Armel, C. & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10), 1292–1293.
- Krajbich, I., Lu, D., Camerer, C. & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in psychology*, 3, 1–14. doi:10.3389/fpsyg.2012.00193
- Kruschke, J. K. (2010a). *Doing Bayesian data analysis: a tutorial with R and BUGS* (1st ed.). Academic Press Inc.
- Kruschke, J. K. (2010b). What to believe: bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14, 293–300. doi:10.1016/j.tics.2010.05.001

- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3), 299–312. doi:10.1177/1745691611406925
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, 142(2), 573–603. doi:10.1037/a0029146
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS and Stan* (2nd ed.). Academic Press Inc.
- Kubovy, M. & Valkenburg, D. V. (2001). Auditory and visual objects. *Cognition*, 80, 97–126.
- Lachaud, C. M. & Renaud, O. (2011). A tutorial for analyzing human reaction times: how to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, 32, 389–416.
- Lachter, J., Forster, K. I. & Ruthruff, E. (2004). Forty-five years after Broadbent (1958): still no identification without attention. *Psychological Review*, 111(4), 880–913.
- Lachter, J., Ruthruff, E., Lien, M.-C. & McCann, R. S. (2008). Is attention needed for word identification? evidence from the stroop paradigm. *Psychonomic Bulletin & Review*, 15(5), 950–955. doi:0.3758/PBR.15.5.950
- Lamey, D., Leber, A. B. & Egeth, H. E. (2012). Selective attention. In A. Healy & R. Proctor (Eds.), *Experimental psychology, volume 4: experimental psychology*. New York: Wiley.
- Larsby, B., Hällgren, M., Lyxell, B. & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, 44, 131–143.
- Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3), 451–468.
- Lavie, N. (2000). Selective attention and cognitive control: dissociating attentional functions through different types of load. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: attention and performance xviii* (pp. 175–194). MIT Press, Cambridge.
- Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, 9(2), 75–82.
- Lavie, N. (2010). Attention, Distraction, and Cognitive Control Under Load. *Current Directions in Psychological Science*, 19(3), 143–148.
- Lavie, N. & De Fockert, J. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin and Review*, 12(4), 669–674.
- Lavie, N. & De Fockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics*, 65(2), 202–212.
- Lavie, N., de Fockert, A. H. J. W. & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3), 339–354. doi:10.1037/0096-3445.133.3.339
- Lavie, N. & Fox, E. (2000). The role of perceptual load in negative priming. *Journal of Experimental Psychology: Human Perception and Performance*, 26(3), 1038–1052.
- Leech, G., Rayson, P. & Wilson, A. (2011). Companion website for ‘Word frequencies in written and spoken English: based on the British National Corpus’. Retrieved August 6, 2013, from <http://ucrel.lancs.ac.uk/bncfreq/>

- Legge, G. E. & Bigelow, C. A. (2011). Does print size matter for reading? A review of findings from vision science and typography. *Journal of Vision*, *11*(8), 1–22. doi:[10.1167/11.5.8](https://doi.org/10.1167/11.5.8)
- Legge, G. E., Mansfield, J. S. & Chung, S. T. (2001). Psychophysics of reading XX. linking letter recognition to reading speed in central and peripheral vision. *Vision Research*, *725*–*743*. doi:[10.1016/S0042-6989\(00\)00295-9](https://doi.org/10.1016/S0042-6989(00)00295-9)
- Levitt, H. (2001). Noise reduction in hearing aids: a review. *Journal of Rehabilitation Research and Development*, *38*(1), 111–121.
- Levitt, H. & Rabiner, L. R. (1967). Use of a sequential strategy in intelligibility testing. *Journal of the Acoustical Society of America*, *42*(3), 609–612.
- Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*, 764–766.
- Link, W. A. & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *(3)*, 112–115. doi:[10.1111/j.2041-210X.2011.00131.x](https://doi.org/10.1111/j.2041-210X.2011.00131.x)
- Loizou, P. C. (2007). *Speech enhancement: theory and practice*. CRC Press.
- Luce, P. A., Feustel, T. C. & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *25*(17). doi:[10.1177/001872088302500102](https://doi.org/10.1177/001872088302500102)
- Lutfi, R. A. (1990). How much masking is informational masking? *Journal of the Acoustical Society of America*, *88*(6), 2607–2610. doi:[10.1121/1.399980](https://doi.org/10.1121/1.399980)
- Macken, W. J., Phelps, F. G. & Jones, D. M. (2009). What causes auditory distraction? *Psychonomic Bulletin & Review*, *16*(1), 139–144. doi:[10.3758/PBR.16.1.139](https://doi.org/10.3758/PBR.16.1.139)
- Mackersie, C. L., Boothroyd, A. & Prida, T. (2000). Use of a simultaneous sentence perception test to enhance sensitivity to ease of listening. *Journal of Speech, Language, and Hearing Research*, *43*, 475–682.
- Mackersie, C. L. & Cones, H. (2011). Subjective and psychophysiological indices of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*, 113–122.
- Mackersie, C. L., Neuman, A. C. & Levitt, H. (1999). A comparison of response time and word recognition measures using a word-monitoring and closed-set identification task. *Ear & Hearing*, *20*(2), 140–148.
- Mackie, K., Dermody, P. & Katsch, R. (1987). Assessment of evaluation measures for processed speech. *Speech Communication*, *6*, 309–316.
- Mantokoudis, G., Dubach, P., Pfiffner, F., Kompis, M., Caversaccio, M. & Senn, P. (2012). Speech perception benefits of internet versus conventional telephony for hearing-impaired individuals. *Journal of Medical Internet Research*, *14*(4), 21. doi:[10.2196/jmir.1818](https://doi.org/10.2196/jmir.1818)
- Marcus, S. M. (1981). Acoustic determinants of perceptual center (p-center) location. *Perception & Psychophysics*, *30*(3), 247–256.
- Martínez-Búrdalo, M., Martín, A., Anguiano, M. & Villar, R. (2004). Comparison of FDTD-calculated specific absorption rate in adults and children when using a mobile phone at

- 900 and 1800 MHz. *Physics in Medicine and Biology*, 49, 345–354. doi:[10.1088/0031-9155/49/2/011](https://doi.org/10.1088/0031-9155/49/2/011)
- Mathewson, K. J., Arnell, K. M. & Mansfield, C. A. (2008). Capturing and holding attention: the impact of emotional words in rapid serial visual presentation. *Memory & Cognition*, 36(1), 182–200.
- Mattys, S. L., Brooks, J. & Cooke, M. (2009). Recognizing speech under a processing load: dissociating energetic from informational factors. *Cognitive Psychology*, 59, 203–243. doi:[10.1016/j.cogpsych.2009.04.001](https://doi.org/10.1016/j.cogpsych.2009.04.001)
- Mattys, S. L., Davis, M. H., Bradlow, A. R. & Scott, S. K. (2012). Speech recognition in adverse conditions: a review. *Language and Cognitive Processes*, 27(7–8), 953–978.
- Mattys, S. L. & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, 65, 145–160.
- McCreery, R. & Stelmachowicz, P. (2013). The effects of limited bandwidth and noise on verbal processing time and word recall in normal-hearing children. *Ear and Hearing*, 34(5), 585–91. doi:[10.1097/AUD.0b013e31828576e2](https://doi.org/10.1097/AUD.0b013e31828576e2)
- McDermott, J. M., Pérez-Edgar, K. & Fox, N. A. (2007). Variations of the flanker paradigm: assessing selective attention in young children. *Behavior Research Methods*, 39(1).
- McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G. & Amitay, S. (2014). Listening effort and fatigue: what exactly are we measuring? a British Society of Audiology cognition in hearing special interest group ‘white paper’. *International Journal of Audiology*, 53(433–445). doi:[10.3109/14992027.2014.890296](https://doi.org/10.3109/14992027.2014.890296)
- McKoon, G. & Ratcliff, R. (2013). Aging and predicting inferences: a diffusion model analysis. *Journal of Memory and Language*, 68, 240–254. doi:[10.1016/j.jml.2012.11.002](https://doi.org/10.1016/j.jml.2012.11.002)
- McVay, J. C. & Kane, M. J. (2009). Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35(1), 196–204. doi:[10.1037/a0014104](https://doi.org/10.1037/a0014104)
- McVay, J. C. & Kane, M. J. (2012). Drifting from slow to “d’oh!”: working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 525–549. doi:[10.1037/a0025896](https://doi.org/10.1037/a0025896)
- Merkt, J., Singmann, H., Goossens-Merkt, S. B. H., Kappes, A., Wendt, M. & Gawrilow, C. (2013). Flanker performance in female college students with ADHD: a diffusion model analysis. *ADHD Attention Deficit Hyperactive Disorder*, 5, 321–341. doi:[10.1007/s12402-013-0110-1](https://doi.org/10.1007/s12402-013-0110-1)
- Miller, G. A., Heise, G. A. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41(5), 329–335.
- Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27(2), 338–342.
- Milosavljevic, M., Malmaud, J., Huth, A., Koch, C. & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, 5(6), 437–449.
- Miyake, A. & Shah, P. (Eds.). (1999). *Models of working memory*. Cambridge.

- Moller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I. & Weiss, B. (2009). A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *International workshop on quality of multimedia experience (QoMEX 2009)* (pp. 7–12). doi:10.1109/QOMEX.2009.5246986
- Möller, S. (2000). *Assessment and prediction of speech quality in telecommunications*. Kluwer Academic Publishers.
- Moore, B. C. J. & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3), 750–753. doi:10.1121/1.389861
- Moray, N. (1959). Attention in dichotic listening: affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 56–61.
- Morton, J., Marcus, S. & Prankish, C. (1976). Perceptual centers (p-centers). *Psychological Review*, 83(5), 405–408.
- Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W. & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience*, 32(7), 2335–2343. doi:10.1523/jneurosci.4156-11.2012
- Mullane, J. C., Corkum, P. V., Klein, R. M. & McLaughlin, E. (2009). Interference control in children with and without ADHD: a systematic review of flanker and Simon task performance. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, 15(4), 321–342.
- Murphy, S., Fraenkel, N. & Dalton, P. (2013). Perceptual load does not modulate auditory distractor processing. *Cognition*, 129, 345–355. doi:10.1016/j.cognition.2013.07.014
- Nagle, K. F. & Eadie, T. L. (2012). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders*, 45, 235–245.
- Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231, 289–337.
- Nielsen, J. K., Christensen, M. G. & Jensen, S. H. (2013, May). Bayesian model comparison and the BIC for regression models. In *IEEE international conference on acoustics, speech and signal processing (icassp 2013)* (pp. 6362–6366). doi:10.1109/ICASSP.2013.6638890
- Nilsson, M., Soil, S. D. & Sullivan, J. A. (1993). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95(2), 1085–1099.
- Noisex. (1990). NOISE-ROM-0. NATO: AC243/(Panel 3)/RSG-10, ESPRIT: Project 2589-SAM. Samples downloaded from <http://web.archive.org/web/20120718175524/http://spib.rice.edu/spib/data/signals/noise/babble.html>.
- Noonan, M. & Axelrod, S. (1981). Earedness (ear choice in monaural tasks): its measurement and relationship to other lateral preferences. *The Journal of Auditory Research*, 21, 263–277.
- Norman, D. A. & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44–64.



- Nudds, M. (2007). What are auditory objects? In N. Bullock & P. Egré (Eds.), *Objects and sound perception. European review of philosophy* (Vol. 7).
- Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O. & Wittmann, W. W. (2000). Working memory capacity — facets of a cognitive ability construct. *Personality and Individual Differences*, 29, 1017–1045.
- Ofcom. (2013). Communications market report. Retrieved August 21, 2013, from [http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr13/2013\\_UK\\_CMV.pdf](http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr13/2013_UK_CMV.pdf)
- Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9, 97–113.
- Paquet, L. (2001). Eliminating flanker effects and negative priming in the flankers task: evidence for early selection. *Psychonomic Bulletin & Review*, 8(2), 301–306.
- Park, S., Kim, M. & Chin, M. M. (2007). Concurrent working memory load can facilitate selective attention: evidence for specialized load. *Journal of Experimental Psychology: Human Perception and Performance*, 33(5), 1062–1075. doi:10.1037/0096-1523.33.5.1062
- Patel, A. D., Löfqvist, A. & Naito, W. (1999, August). The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem. In *Proceedings of the 14th international congress of phonetic sciences*.
- Patil, A., Huard, D. & Fonnesbeck, C. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4), 1–81.
- Pearson, B., Raškevičius, J., Bays, P. M., Pertzov, Y. & Husain, M. (2014). Working memory retrieval as a decision process. *Journal of Vision*, 14(2:2), 1–15. doi:10.1167/14.2.2
- Pérez, F. & Granger, B. E. (2007, May). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3), 21–29. doi:10.1109/MCSE.2007.53
- Picou, E. M. (2011, August). *The effect of individual variability on listening effort in unaided and aided conditions* (Doctoral dissertation, Hearing and Speech Sciences, Vanderbilt University).
- Picou, E. M. & Ricketts, T. A. (2011). Comparison of wireless and acoustic hearing aid-based telephone listening strategies. *Ear & Hearing*, 32(2), 209–220. doi:10.1097/AUD.0b013e3181f53737
- Picou, E. M., Ricketts, T. A. & Hornsby, B. W. Y. (2011). Visual cues and listening effort: individual variability. *Journal of Speech, Language, and Hearing Research*, 54, 1416–1430.
- Pisoni, D. B., Manous, L. M. & Dedina, M. J. (1987). Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, 2(3–4), 303–320.
- Polkosky, M. D. & Lewis, J. R. (2003). Expanding the MOS: development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*, 6, 161–182.
- Pollack, I., Rubenstein, H. & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America*, 31(3), 273–297. doi:10.1121/1.1907712

- Pourmand, N., Parsa, V. & Weaver, A. (2013). Computational auditory models in predicting noise reduction performance for wideband telephony applications. *International Journal of Speech Technology*, 1–17. doi:10.1007/s10772-013-9189-1
- Praamstra, P., Stegeman, D. F., Cools, A. R. & Horstink, M. W. I. M. (1998). Reliance on external cues for movement initiation in Parkinson's disease evidence from movement-related potentials. *Brain*, 121, 167–177.
- Pratt, R. L. (1981). On the use of reaction time as a measure of intelligibility. *British Journal of Audiology*, 15, 253–255.
- Pullenayegum, E. M., Guo, Q. & Hopkins, R. B. (2012). Developing critical thinking about reporting of Bayesian analyses. *Journal of Statistics Education*, 20(1), 1–14.
- Punch, J. L. & Beck, L. B. (1986). Relative effects of low-frequency amplification on syllable recognition and speech quality. *Ear & Hearing*, 7(2), 57–62.
- R Core Team. (2013). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Rabbitt, P. M. A. (1964). Ignoring irrelevant information. *British Journal of Psychology*, 55(4), 403–414.
- Rabbitt, P. M. A. (1966). Memory for words correctly heard in noise. *Psychonomic Science*, 6(8), 383–384.
- Rabbitt, P. M. A. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, 20(3), 241–248.
- Raftery, A. E. & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. Berger, A. Dawid & A. Smith (Eds.), *Bayesian statistics 4* (pp. 763–773). Oxford University Press.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 2, 59–108.
- Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, 86(3), 446–461.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3), 510–532.
- Ratcliff, R. & Rouder, J. N. (1998). Modelling response times for two-choice decisions. *Psychological Science*, 9(5), 347–356.
- Ratcliff, R., Thapar, A. & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, 50, 408–424. doi:10.1016/j.jml.2003.11.002
- Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3), 438–481. doi:10.3758/BF03196302
- Rix, A. (2004). Perceptual speech quality assessment — a review. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP '04)* (Vol. 3, pp. 1056–1059). doi:10.1109/ICASSP.2004.1326730
- Robinson, T. (1996). The British English Example Pronunciation (BEEP) dictionary. Retrieved September 27, 2010, from <ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries>

- Rönnerberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model. *International Journal of Audiology*, *42*, S68–S76.
- Rönnerberg, J., Andersson, J., Andersson, U., Johansson, K., Lyxell, B. & Samuelsson, S. (1998). Cognition as a bridge between signal and dialogue: communication in the hearing impaired and deaf. *Scandinavian Audiology*, *49*, 101–8.
- Rönnerberg, J., Rudner, M., Foo, C. & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *International Journal of Audiology*, *47*(Suppl. 2), S99–S105.
- Rönnerberg, J., Rudner, M., Lunner, T. & Zekveld, A. A. (2010). When cognition kicks in: working memory and speech understanding in noise. *IJ*(49), 263–269.
- Rönnerberg, J., Rudner, M. & Zekfeld, A. A. (2009). Cognitive hearing science: the role of a working memory system for speech understanding in old age. In L. Hickson (Ed.), *Hearing care for adults*. Phonak.
- Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P. & Heathcote, A. (2014). The lognormal race: a cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 1–23. doi:10.1007/s11336-013-9396-3
- Rubenstein, H., Decker, L. & Pollack, I. (1959). Word length and intelligibility. *Language and Speech*, *2*, 175–178. doi:10.1177/002383095900200401
- Ruggles, D., Bharadwaj, H. & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences*, *108*(37), 15516–15521. doi:10.1073/pnas.1108912108
- Santee, J. L. & Egeth, H. E. (1980). Interference in letter identification: a test of feature-specific inhibition. *Perception & Psychophysics*, *27*(4), 321–330.
- Sarampalis, A., Kalluri, S., Edwards, B. & Hafter, E. (2009). Objective measures of listening effort: effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, *52*, 1230–1240.
- Schmidt-Nielsen, A. (1993). *Intelligibility and acceptability testing for speech technology* (tech. rep. No. AD-A252 015). Naval Research Laboratory. Washington DC.
- Scott, S. K. (1994). *Perceptual centers in speech — acoustic determinants* (Doctoral dissertation, University College London).
- Scott, S. K. (1998). The point of p-centres. *Psychological Research*, *61*, 4–1.
- Seeber, B. U., Kerber, S. & Hafter, E. R. (2010). A system to simulate and reproduce audio-visual environments for spatial hearing research. *Hearing Research*, *260*(1–2), 1–10. doi:10.1016/j.heares.2009.11.004
- Shahar, N., Teodorescu, A. R., Pereg, M. U. M. & Meiran, N. (2014). Selective influence of working memory load on exceptionally slow reaction times. *I43*(5), 1837–1860. doi:10.1037/a0037190
- Shamma, S. (2008). On the emergence and awareness of auditory objects. *PLoS Biology*, *6*(6), 1141–1143. doi:10.1371/journal.pbio.0060155
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

- Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. doi:[10.1016/j.tics.2008.02.003](https://doi.org/10.1016/j.tics.2008.02.003)
- Shinn-Cunningham, B. G. & Wang, D. (2008). Influences of auditory object formation on phonemic restoration. *Journal of the Acoustical Society of America*, *123*(1), 295–301.
- Simon, J. R. (1967). Ear preference in a simple reaction-time task. *Journal of Experimental Psychology*, *75*(1), 49–55.
- Simon, J. R. (1969). Reactions toward the sources of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176.
- Slaney, M. (1993). *An efficient implementation of the Patterson-Holdsworth auditory filter bank*. Apple Computer Technical Report #35. Apple Computer, Inc.
- Smith, N. A. & Trainor, L. J. (2011). Auditory stream segregation improves infants' selective attention to target tones amid distractors. *Infancy*, *16*(6), 655–668. doi:[10.1111/j.1532-7078.2011.00067.x](https://doi.org/10.1111/j.1532-7078.2011.00067.x)
- Sörqvist, P. (2010). *The role of working memory capacity in auditory distraction* (Doctoral dissertation, Luleå University of Technology, Universitetstryckeriet, Luleå).
- Spiegelhalter, D. J., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*(4), 583–639.
- Spiegelhalter, D. J., Best, N., Carlin, B. & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society B*, *76*(3), 485–493.
- Spiegelhalter, D. J., Freedman, L. S. & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A*, *157*(3), 357–416.
- Spiegelhalter, D. J., Myles, J. P., Jones, D. R. & Abrams, K. R. (2000). Bayesian methods in health technology assessment: a review. *Health Technology Assessment*, *4*(38).
- Stelmachowicz, P., Lewis, D., Hoover, B., Nishi, K., McCreery, R. & Woods, W. (2010). Effects of digital noise reduction on speech perception for children with hearing loss. *Ear and Hearing*, *31*(3), 345–355.
- Stoet, G. (2010). PsyToolkit: a software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104.
- Strayer, D. L. & Johnston, W. A. (2001). Driven to distraction: dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, *12*(6), 462–466.
- Studebaker, G. A. & Sherbecoe, R. L. (1988). Magnitude estimations of the intelligibility and quality of speech in noise. *Ear and Hearing*, *9*(5), 259–267.
- Sung, L., Hayden, J., Greenberg, M. L., Koren, G., Feldman, B. M. & Tomlinson, G. A. (2005). Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology*, *58*, 261–268.
- Swaffield, J. & Richards, D. (1959). Rating of speech links and performance of telephone networks. *Proceedings of the Institute of Electrical Engineers Part B: Radio and Electronic Engineering*, *106*(26), 65–76. doi:[10.1049/pi-b-1.1959.0017](https://doi.org/10.1049/pi-b-1.1959.0017)
- Swensson, R. G. (1972). The elusive tradeoff: speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics*, *12*(1), 16–32. doi:[10.3758/BF03212837](https://doi.org/10.3758/BF03212837)

- Thorpe, L. A. (1998). Subjective evaluation of speech compression codecs and other non-linear voice-path devices for telephony applications. *International Journal of Speech Technology*, 2, 273–288.
- Tun, P. A., Benichov, J. & Wingfield, A. (2010). Response latencies in auditory sentence comprehension: effects of linguistic versus perceptual challenge. *Psychology and Aging*, 25(3), 730–735.
- Turner, M. L. & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, 28, 127–154.
- Ulrich, R. & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123(1), 34–80.
- Unsworth, N., Heitz, R. P., Schrock, J. C. & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505.
- van der Walt, S., Colbert, S. C. & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13, 22–30. doi:10.1109/MCSE.2011.37
- Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, 7(3), 424–465.
- Vandekerckhove, J., Tuerlinckx, F. & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16(1), 44–62. doi:10.1037/a0021765
- Villing, R., Repp, B., Ward, T. & Timoney, J. (2011). Measuring perceptual centers using the phase correction response. *Attention Perception Psychophysics*, 73, 1614–1629. doi:10.3758/s13414-011-0110-1
- Villing, R., Ward, T. & Timoney, J. (2003, July). P-centre extraction from speech: the need for a more reliable measure. In *Irish signals and systems conference (issc) 2003*. University College Cork.
- Voiers, W. D. (1967). *Performance evaluation of speech processing devices iii. diagnostic evaluation of speech intelligibility* (tech. rep. No. AF19(628)-4987). Air Force Cambridge Research Laboratories. United States Air Force, Bedford, Massachusetts.
- Voss, A., Nagler, M. & Lerche, V. (2013). Diffusion models in experimental psychology: a practical introduction. *Experimental Psychology*, 60(6), 385–402. doi:10.1027/1618-3169/a000218
- Voss, A., Rothermund, K. & Voss, J. (2004). Interpreting the parameters of the diffusion model: an empirical validation. *Memory & Cognition*, 32(7), 1206–1220.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779–804.
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 21(5), 641–671. doi:10.1080/09541440802205067
- Wagenmakers, E.-J. & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192–196.

- Wältermann, M. (2013). *Dimension-based quality modeling of transmitted speech*. T-Labs series in telecommunication services. Springer Publishing Company, Incorporated. doi:[10.1007/978-3-642-35019-1](https://doi.org/10.1007/978-3-642-35019-1)
- Wang, Y., Stables, R. & Reiss, J. (2010). Audio latency measurement for desktop operating systems with onboard soundcards. In *128th convention of the Audio Engineering Society*. London, UK.
- Wenzel, E. M., Wightman, F. L. & Kistler, D. J. (1991). Localization with non-individualized virtual acoustic display cues. In *Proceedings of the Special Interest Group on Computer-Human Interaction (SIGCHI) conference on human factors in computing systems* (pp. 351–359). New Orleans, Louisiana, USA: ACM. doi:[10.1145/108844.108941](https://doi.org/10.1145/108844.108941)
- Whalen, D. H., Cooper, A. M. & Fowlert, C. A. (1991). *P-center judgments are generally insensitive to the instructions given*. Haskins Laboratories Status Report on Speech Research SR-107/108. Haskins Laboratories.
- White, C. N., Ratcliff, R. & Starns, J. J. (2011). Diffusion models of the flanker task: discrete versus gradual attentional selection. *Cognitive Psychology*, *63*, 210–238. doi:[10.1016/j.cogpsych.2011.08.001](https://doi.org/10.1016/j.cogpsych.2011.08.001)
- Whitney, P., Arnett, P. A., Driver, A. & Budd, D. (2001). Measuring central executive functioning: what's in a reading span? *Brain and Cognition*, *45*, 1–14. doi:[10.1006/brcg.2000.1243](https://doi.org/10.1006/brcg.2000.1243)
- Whittaker, S. (2002). Theories and methods in mediated communication. In A. Graesser, M. Gernsbacher & S. Goldman (Eds.), *The handbook of discourse processes* (pp. 243–286). Erlbaum, NJ.
- Wiecki, T. V., Sofer, I. & Frank, M. J. (2013). HDDM: hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *17*(14), 1–10. doi:[10.3389/fninf.2013.00014](https://doi.org/10.3389/fninf.2013.00014)
- Wightman, F. L. & Kistler, D. J. (2005). Informational masking of speech in children: effects of ipsilateral and contralateral distracters. *Journal of the Acoustical Society of America*, *118*(5), 3164–3188. doi:[10.1121/1.2082567](https://doi.org/10.1121/1.2082567)
- Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H. & Johnsrude, I. S. (2012). Effortful listening: the processing of degraded speech depends critically on attention. *The Journal of Neuroscience*, *32*(40), 14010–14021.
- Wood, N. & Cowan, N. (1995). The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 255–260.
- Wright, M., Cassidy, R. J. & Zbyszyński, M. F. (2004). Audio and gesture latency measurements on Linux and OSX. In *International computer music conference 2004* (pp. 423–429). Miami, FL: International Computer Music Association.
- Wyatt, N. & Machado, L. (2013). Evidence inhibition responds reactively to the salience of distracting information during focused attention. *PLoS ONE*, *8*(4), e62809. doi:[10.1371/journal.pone.0062809](https://doi.org/10.1371/journal.pone.0062809)
- Yeshurun, Y. & Marciano, H. (2013). Degraded stimulus visibility and the effects of perceptual load on distractor interference. *Frontiers in Psychology*, *4*(289), 1–14.

- Zeguers, M. H., Snellings, P., Tijms, J., Weeda, W. D., Tamboer, P., Bexkens, A. & Huizenga, H. M. (2011). Specifying theories of developmental dyslexia: a diffusion model analysis of word recognition. *Developmental Science*, *14*(6), 1340–1354. doi:[10.1111/j.1467-7687.2011.01091.x](https://doi.org/10.1111/j.1467-7687.2011.01091.x)
- Zekveld, A. A., Kramer, S. E. & Festen, J. M. (2010). Pupil response as an indication of effortful listening: the influence of sentence intelligibility. *Ear & Hearing*, *31*, 480–490.
- Zhang, J. & Rowe, J. B. (2014). Dissociable mechanisms of speed-accuracy tradeoff during visual perceptual learning are revealed by a hierarchical drift-diffusion model. *Frontiers in Neuroscience*, *8*(69), 1–13. doi:[10.3389/fnins.2014.00069](https://doi.org/10.3389/fnins.2014.00069)