THE INFLUENCE OF CHANNEL AND SOURCE DEGRADATIONS ON INTELLIGIBILITY AND PHYSIOLOGICAL MEASUREMENTS OF EFFORT

Maximillian Paulus

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

Doctor of Philosophy

Department of Speech, Hearing and Phonetic Sciences

University College London (UCL)

25. 11. 2020

Declaration

I, Maximillian Paulus, 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.

Maximillian Paulus

Abstract

Despite the fact that everyday listening is compromised by acoustic degradations, individuals show a remarkable ability to understand degraded speech. However, recent trends in speech perception research emphasise the cognitive load imposed by degraded speech on both normal-hearing and hearing-impaired listeners. The perception of degraded speech is often studied through channel degradations such as background noise. However, source degradations determined by talkers' acoustic-phonetic characteristics have been studied to a lesser extent, especially in the context of listening effort models. Similarly, little attention has been given to speaking effort, i.e., effort experienced by talkers when producing speech under channel degradations. This thesis aims to provide a holistic understanding of communication effort, i.e., taking into account both listener and talker factors.

Three pupillometry studies are presented. In the first study, speech was recorded for 16 Southern British English speakers and presented to normal-hearing listeners in quiet and in combination with three degradations: noise-vocoding, masking and time-compression. Results showed that acoustic-phonetic talker characteristics predicted intelligibility of degraded speech, but not listening effort, as likely indexed by pupil dilation. In the second study, older hearing-impaired listeners were presented fast time-compressed speech under simulated room acoustics. Intelligibility was kept at high levels. Results showed that both fast speech and reverberant speech were associated with higher listening effort, as suggested by pupillometry. Discrepancies between pupillometry and perceived effort ratings suggest that both methods should be employed in speech perception research to pinpoint processing effort. While findings from the first two studies support models of degraded speech perception, emphasising the relevance of source degradations, they also have methodological implications for pupillometry paradigms. In the third study, pupillometry was combined with a speech production task, aiming to establish an equivalent to listening effort for talkers: speaking effort. Normal-hearing participants were asked to read and produce speech in quiet or in the presence of different types of masking:

stationary and modulated speech-shaped noise, and competing-talker masking. Results indicated that while talkers acoustically enhance their speech more under stationary masking, larger pupil dilation associated with competing-speaker masking reflected higher speaking effort.

Results from all three studies are discussed in conjunction with models of degraded speech perception and production. Listening effort models are revisited to incorporate pupillometry results from speech production paradigms. Given the new approach of investigating source factors using pupillometry, methodological issues are discussed as well. The main insight provided by this thesis, i.e., the feasibility of applying pupillometry to situations involving listener and talker factors, is suggested to guide future research employing naturalistic conversations.

Impact statement

Speech is an integral part of everyday communication. Humans have developed sophisticated auditory and cognitive mechanisms to deal with acoustic variability introduced by listening environments, talkers and accents. However, such abilities and capacities have a natural limit which is often exceeded because of high noise levels, and hearing and cognitive impairments. The shift from physical to virtual communication has introduced further barriers, owed to signal processing constraints. One step in tackling those challenges is to understand and quantify the abilities and capacities of both normal-hearing and hearing-impaired listeners in adverse listening conditions.

While much research has quantified acoustic challenges by measuring recognition, i.e., the amount of information retrieved from a degraded signal, more recent approaches acknowledge the fact that acoustic challenges extend beyond recognition. Effort and fatigue are downsides of impaired communication that are not captured by traditional measures. Therefore, the current thesis investigates pupillometry as an objective measure of listening effort, for both normal-hearing and hearing-impaired listeners. Specifically, the presented research focused on acoustic challenges beyond environmental degradations (e.g. noise) by taking talker-related factors into account (e.g. speaking rate).

The research presented in this thesis was conducted as part of a larger research training network (ENRICH) funded by the European Horizon 2020 programme. Industry placements were part of the research collaboration. As such, I spent three months at a hearing-aid manufacturer (Sonova) in Switzerland, conducting research with hearing-impaired listeners. The work explored several novel aspects of applied hearing research that have the potential to benefit users of hearing aids in the future. First of all, pupillometry is a relatively novel method that measures listening effort objectively. As part of the project, I investigated its application in more realistic laboratory settings involving simulations of existing room acoustics using a sophisticated loudspeaker setup and simulation technique (higher-order ambisonics). Such setups allow the evaluation of novel

hearing aid algorithms in acoustic environments that correspond to everyday listening.

Training within ENRICH focused not only on technical and research skills, but also on public engagement. On two occasions, I presented my work to public audiences at the Royal Institution in London (July 2019 and March 2020). The presentations demonstrated the use of specific tools (eye tracking and transcranial magnetic stimulation) for research purposes, but also real-world applications in clinical audiology. Audiences were encouraged to test the equipment themselves, for instance by listening to degraded speech while observing their pupil dilation being tracked. The enthusiasm received from a mixed audience highlighted the importance of providing the public insights into scientific projects.

List of Tables

1 Descriptive statistics for acoustic-phonetic features. ME13 = mean energy (dB), F0M = f0 Median (Hz), F0SD = f0 standard deviation in semitones (st) and Hz, SR = speaking rate (syllables/s), VSD = vowel space dispersion in mels and Hz. 55

2 Linear regression results for intelligibility. ME13 = mean energy, F0M = f0 Median, F0SD = f0 standard deviation, SR = speaking rate, VSD = vowel space dispersion. All continuous predictors are mean-centered and scaled to have SD = 1. *** p < 0.001; ** p < 0.01; * p < 0.05. 69

3 Linear regression results for peak pupil dilation. ME13 = mean energy, SR = speaking rate, VSD = vowel space dispersion. All continuous predictors are mean-centered and scaled to have SD = 1. *** p < 0.001; ** p < 0.01; * p < 0.05. 70

4 Linear regression results for peak latency. ME13 = mean energy, VSD = vowel space dispersion, SR = speaking rate. All continuous predictors are mean-centered and scaled to have SD = 1. *** p < 0.001; ** p < 0.01; * p < 0.05. 71

5 Linear regression results for recognition, perceived effort, and peak dilation. Difference scores were calculated between slow speech in dry and fast speech in reverb (test session). Higher scores indicate less detrimental effects of reverb and time-compression. Age, PTA = pure tone average, PS = processing speed. All continuous predictors are mean-centered and scaled to have SD = 1. *** p < 0.001; ** p < 0.01; * p < 0.05. 107

List of Figures

1 The Framework for Understanding Effortful Listening (FUEL). Adapted and modified from Kahneman (1973) and Pichora-Fuller et al. (2016). The original components of the Capacity Model for Attention (Kahneman, 1973) are shown with names modified according to FUEL. In yellow: evaluation components. In orange: model inputs. In blue: model outputs. 27

2 Yerkes-Dodson law. Adapted and modified from Kahneman (1973). 29

3 Interpretation of the Framework for Understanding Effortful Listening with emphasis on the relationship between arousal and capacity. Adapted and modified from Kahneman (1973) and Pichora-Fuller et al. (2016). The evaluation components of the capacity model have been summarised. Two listening scenarios are depicted with inspirations taken from Lemke and Besser (2016). 31

4 Proposed relationship between arousal/effort and task demands. Three discrete levels of performance are indicated as sections of the curve. Based on results from Wendt et al. (2018). 34

5 Vowel category centers (points) and vowel space centers (crosses) for all talkers. Vowels are represented by their first and second formants, F1 and F2, respectively. 53

6 Trial events with duration. Rectangles represent displays with central fixation cross. 57

7 Distributions of the average and rate of keywords recognized correctly in all conditions. Adaptation rate was the slope of the linear fit to an individual's performance, based on the first half of trials within each condition block (trials 1-24). Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range. 63

8 Percent of keywords recognized correctly as a function of trial number, averaged across participants. Lines indicate linear fits to the averaged data in the first (1-24) and the second half (25-48) of trials. Individual adaptation slopes were obtained from the first half of trials. 64
9 Distributions of peak pupil dilation (top) and latency (bottom) in all conditions. Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range. 65
10 Change in baseline pupil size across trials. Lines indicate linear fits to the averaged data in the first (1-24) and the second half (25-48) of trials. For visualisation purposes, baseline pupil size is displayed as percent change from first trial. Statistical analyses were conducted on raw pupil size data. 66
11 Intelligibility, i.e., average percent of keywords recognized correctly aggregated by talker and experimental condition. Each symbol therefore represents one talker average. Noise-vocoding is plotted against masking and time-compression. 68
12 Audiogram averaged across all 18 listeners. Points indicate means across participants and bars indicate one standard error around the mean. 90
13 Perceived effort rating display (visual analogue scale 0-100%) and game pad. The movable slider was initially located at 50%. Effort minimum (-) and maximum (+) were always indicated in the display. 94
14 Visualisation of the two contrasts used for analysis. Contrast 1 investigated the main effects and interaction of speech and room while contrast 2 investigated the main effects and interaction of speed and program. 96
15 Word recognition performance in the pilot experiment. Points indicate means across participants and bars indicate one standard error around the mean. 97

16 Perceived effort in the pilot experiment. Points indicate means across participants and bars indicate one standard error around the mean. 98
17 Average pupil dilation in the pilot experiment. For visualisation purposes, curves have been smoothed with a moving-average filter. 98
18 Word recognition performance for Contrast 1 (Room). Points indicate means across participants and bars indicate one standard error around the mean. 100
19 Word recognition performance for Contrast 2 (Program). Points indicate means across participants and bars indicate one standard error around the mean. 101
20 Distributions of effort ratings for Contrast 1 (Room), split by room (dry vs. reverb). Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range. Jittered points represent individual effort ratings. 102
21 Effort ratings for Contrast 1 (Room). Points indicate means across participants and bars indicate one standard error around the mean. 103
22 Effort ratings for Contrast 2 (Program). Points indicate means across participants and bars indicate one standard error around the mean. 103
23 Average pupil traces for Contrast 1 (Room). Bands indicate one standard error around the mean. Smoothed with a 5-point moving average filter. 104
24 Average pupil traces for Contrast 2 (Program). Bands indicate one standard error around the mean. Smoothed with a 5-point moving average filter. 105

25 Distributions of peak pupil dilation for Contrast 1 (Room), with combined data across test and retest sessions. Boxes represent values from the first to the third quartile with the median indicated by a solid line. Whiskers extend up to 1.5 times the interquartile range. Jittered points represent individual peaks. 106
26 Trial events with duration. Rectangles represent displays with central fixation cross. 125
27 Mean energy in the 1-3 kHz range in each condition. Points indicate means across participants and bars indicate one standard error around the mean. 128
28 Speaking rate and speech onset in each condition. Points indicate means across participants and bars indicate one standard error around the mean. 129
29 Vowel space dispersion in each condition. Points indicate means across participants and bars indicate one standard error around the mean. 129
30 Fundamental frequency in each condition. Points indicate means across participants and bars indicate one standard error around the mean. 130
31 Average pupil traces with trial events. Comparison of two baseline correction methods, resting state baseline (upper panel) and noise baseline (lower panel). The solid line indicates average speech offset (3.05 s). Coloured bands indicate the standard error. 132
32 Mean and peak dilation during the preparation and speaking phase, respectively (resting baseline). Points indicate means across participants and bars indicate one standard error around the mean. 133

33 Mean and peak dilation during the preparation and speaking phase, respectively (noise baseline). Points indicate means across participants and bars indicate one standard error around the mean. 134
34 Mean energy for all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2. 188
35 Fundamental frequency of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2. 189
36 Fundamental frequency standard deviation (SD) of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2. 189
37 Speaking rate of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2. 190
38 Vowel space dispersion of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all vowel productions and vowel categories. 190
39 Distributions of mean energy in each condition (Chapter 4), separate for female and male participants. 191
40 Distributions of fundamental frequency in each condition (Chapter 4), separate for female and male participants. 191
41 Distributions of speaking rate in each condition (Chapter 4), separate for female and male participants. 192
42 Distributions of speech onset in each condition (Chapter 4), separate for female and male participants. 192

43 Distributions of vowel space dispersion in each condition (Chapter 4), separate for female and male participants. 193

Contents

List of Tables 6

List of Figures 7

Chapter 1: General introduction 11

1.1 Thesis context . . . 11
1.2 Source and channel degradations . . . 12
1.2.1 Source degradations . . . 14
1.2.2 Channel degradations . . . 16
1.3 Listener factors in degraded speech perception . . . 24
1.4 Listening effort . . . 26
1.5 Pupillometry and listening effort . . . 32
1.5.1 Physiology of the pupil dilation response . . . 32
1.5.2 Pupil dilation and degraded speech . . . 33
1.6 Summary and thesis outline . . . 38

Chapter 2: Source and channel degradations and their effect on intelligibility and effort 42

2.1 Introduction . . . 42
2.1.1 Intelligibility under source and channel degradations . . . 42
2.1.2 Listening effort and adaptation under source degradations 46
2.1.3 Aims of the current study . . . 49
2.2 Methods . . . 51
2.2.1 Speech materials . . . 51
2.2.2 Acoustic analyses . . . 52
2.2.3 Participants . . . 54
2.2.4 Listening conditions . . . 55
2.2.5 Design & Procedure . . . 57
2.2.6 Dependent variables . . . 58
2.2.7 Statistical analysis . . . 61
2.3 Results: channel degradations . . . 63
2.3.1 Intelligibility and adaptation . . . 63

2.3.2 Pupillometry . . . 64
2.3.3 Interim summary . . . 67
2.4 Results: source degradations . . . 67
2.4.1 Intelligibility across degradations . . . 67
2.4.2 Intelligibility and adaptation . . . 68
2.4.3 Pupillometry . . . 70
2.5 Discussion . . . 72
2.5.1 Channel degradations . . . 72
2.5.2 Source degradations: intelligibility and adaptation . . . 74
2.5.3 Source degradations: listening effort . . . 77
2.5.4 Limitations . . . 79
2.6 Conclusion . . . 79

Chapter 3: Listening effort experienced by hearing-impaired listeners processing fast speech with simulated room acoustics 82

3.1 Introduction . . . 82
3.1.1 Listening effort at high intelligibility . . . 83
3.1.2 Processing time-compressed and reverberant speech: effects of hearing impairment and aging . . . 85
3.1.3 Aims of the current study . . . 88
3.2 Methods . . . 90
3.2.1 Participants . . . 90
3.2.2 Materials . . . 91
3.2.3 Equipment . . . 92
3.2.4 Design & Procedure . . . 93
3.2.5 Preprocessing and statistical analysis . . . 95
3.3 Results: pilot experiment . . . 97
3.4 Results: main experiment . . . 99
3.4.1 Word recognition . . . 99
3.4.2 Effort ratings . . . 101
3.4.3 Pupillometry . . . 104
3.4.4 Individual differences . . . 106
3.5 Discussion . . . 108

3.5.1 Performance . . . 108
3.5.2 Listening effort . . . 110
3.5.3 Limitations . . . 113
3.5.4 Conclusion . . . 113

Chapter 4: Measuring speaking effort during speech production in background noise **116**

4.1 Introduction . . . 116
4.1.1 Lombard effect . . . 117
4.1.2 Pupillometry and speech production . . . 120
4.1.3 Aims of the current study . . . 123
4.2 Methods . . . 123
4.2.1 Participants . . . 123
4.2.2 Materials & Design . . . 124
4.2.3 Procedure . . . 125
4.2.4 Dependent variables and preprocessing . . . 126
4.2.5 Statistical analysis . . . 127
4.3 Results: acoustics . . . 128
4.4 Results: pupillometry . . . 132
4.4.1 Resting baseline . . . 133
4.4.2 Noise baseline . . . 134
4.5 Discussion . . . 135
4.5.1 Lombard effect . . . 135
4.5.2 Pupillometry . . . 137
4.5.3 Limitations . . . 139
4.6 Conclusion . . . 140

Chapter 5: General discussion **142**

5.1 Source and channel degradations and their effect on intelligibility and effort (RQ1 and RQ2) . . . 142
5.1.1 Limitations and future directions . . . 147
5.2 Pupillometry and older hearing-impaired listeners: fast speech and room acoustics (RQ3 and RQ4) . . . 149

5.2.1 Limitations and future directions 151
5.3 Pupillometry during speech production as a measure of speaking effort 153
5.3.1 Limitations and future directions 155
5.4 Conclusion 156

References **159**

Appendix **188**

Acknowledgements

In the last three years, I was fortunate to meet many interesting people, some of whom I call friends now. First of all, I would like to thank my supervisors, Patti and Valerie. Both were very supportive over the years, while leaving me a lot of space to explore and find my own way - that's how I imagined the PhD journey. I am thankful to Patti for always keeping me focused and not letting myself being distracted by all the interesting topics the field has to offer. To stick to a few realisable ideas and exploring them in depth, that's what science is, and that's what I had to learn. There are plenty of brilliant individuals that kept me company during this time; I will do my best to mention them all. I would like to thank the entire *Speech on the Brain* lab: Gwijde, Shego, Tony, Han, Dan and Andrew. Gwijde, I am glad that we managed to spend some leisure time together, as well - jamming was fun! Thanks to Dan who supported me in the beginning and Andrew who was always available for technical support. I haven't seen my PhD room fellows during the last stage of thesis writing (thanks Covid), but I was lucky to have such nice "room mates" during the three years: Giulia, Julie, Shiran, Katherina, Rachel, Anqi and everyone else. Giulia, mi ricordo bene il primo anno, Brighton, Liguria, lo scivolo alto (con Anna), parlare del lavoro (in italiano?!), sere con Antonio e Massimo… Grazie non solo per l'aiuto professionale ma anche per rendere l'inizio più comodo per me. Anche se non ci vediamo spesso, rimani sempre un'amica vicina.

I am grateful for all the experience I gained as part of the ENRICH network: all the workshops, collaborations, presentations and public engagement events were stressful at times, but shaped me as a person. Thanks to Valerie, Martin and Monica for putting so much effort into organising all those interesting events. Thanks to my fellow ENRICHies: Elif, Katherine, Chen, Amy, Gerard, Dip, Shifas, Julie, Avashna, Sneha, Olina and Carol. The network allowed me to collaborate with Sonova in Switzerland where I conducted part of my research. There I met another bunch of great, supportive people: Matthias, Juliane, Schelle, Markus, Basti, Daniel, Stefan, Diego, Hannes, Nick, Marlene and Julia. Thanks for introducing me to the best Cordon Bleu I've ever eaten (and

the rooftop sauna). With Anna, I shared both ENRICH and Chandler House life. I always appreciated her healthy view on both the academic and the real world. Danke Anna! And then, Emma. She is the reason why I haven't given up and why I always stayed optimistic. Our hikes, conversations, and laughs made every difficult situation bearable. I'm looking forward to the next phase with you - Goodbye London, Welkom in Amsterdam! And last, but not least, my small, chaotic but warm-hearted family: Villmols Merci Mama, Papa, Ute, Hubi, Steve, Gemma and Freddy. E grazie Nonna.

Chapter 1: General introduction

1.1 Thesis context

Communication can be divided into four components (Kiessling et al., 2003; Lemke & Besser, 2016): *Hearing*, i.e., receiving sounds physically and passively through the mechanics of the ear; *Listening*, i.e., perceiving and intentionally allocating attention to sounds; *Comprehending*, i.e., decoding a speaker's message and embedding it into its broader context; *Communicating*, i.e., the entire process of speech perception and production with one or more interlocutors. The definition of the latter component, *Communicating*, is relatively broad and could potentially be divided further into *Speaking*, i.e., emphasising the role of the talker, and other communicative behaviour such as turn-taking. The literature review presented in this chapter will touch upon all these components of communication to some extent. However, emphasis will be on the intersection of hearing and listening which are the components at the core of listening effort models. In Chapter 4 and 5, these models will then be refined to account for speech production, as well.

Research into speech perception has provided an in-depth understanding of the human ability to perceive speech despite the lack of invariance, i.e., the variability in speech production across and within talkers (Casserly & Pisoni, 2010). Models of speech perception are often based on research conducted in optimal listening environments. However, the remarkable ability of humans to hear and perceive speech is often challenged in everyday listening situations due to degradations of the communication channel and/or the characteristics of the talker (Mattys et al., 2012). Such adverse conditions not only affect the intelligibility of speech, but also induce effort and fatigue, especially in hearing-impaired listeners (McGarrigle et al., 2014; Pichora-Fuller et al., 2016).

Effort research in speech science has predominantly focused on the listener, with emphasis on acoustic degradations that affect the communication channel, such as background noise. Source degradations that comprise properties of

the speaker at accent, anatomical and physiological levels have received little attention (Van Engen & Peelle, 2014). Source degradations have been known to affect intelligibility (Bradlow et al., 1996; Hazan & Markham, 2004), which suggests that they contribute to effort, as well - in particular for hearing-impaired listeners. Conversely, while it is known that talkers modify their speech when the communication channel is degraded, little research has focused on the effort experienced by talkers. This thesis therefore aims to take a holistic approach to communication effort, taking into account both listener and talker. On the one hand, the speech perception studies presented in this thesis aimed to evaluate whether talker characteristics affect listening effort for normal-hearing (Chapter 2) and hearing-impaired listeners (Chapter 3). On the other hand, a speech production study aimed to establish an equivalent to listening effort for talkers, which will be termed "speaking effort" (Chapter 4).

In the following, acoustic degradations and their effect on speech perception are presented. This is followed by an introduction of the notion of listening effort and its objective assessment via pupillometry. Underlying theoretical frameworks are discussed. The chapter is concluded with an outline of the research aims addressed in this thesis.

1.2 Source and channel degradations

Adverse conditions have been defined as "any factor leading to a decrease in speech intelligibility on a given task relative to the level of intelligibility when the same task is performed in optimal listening situations, i.e., healthy native listeners hearing carefully recorded speech in a quiet environment and under focused attention" (Mattys et al., 2012, p. 953). From a cognitive perspective, adversity has been described as "the mismatch between external demands and internal resources to meet these demands" (Lemke & Besser, 2016, p. 79S). The definition by Lemke & Besser (2016) has been formulated in the context of listening effort models, which I will turn to in Section 1.4 of this literature review. According to Mattys et al. (2012), degradations can be divided into two

main categories: environmental or transmission channel degradations (hereinafter *channel degradations*) and *source degradations*. The term "channel degradation" is a generalisation of the terminology used by Mattys et al. (2012) and includes environmental degradations (masking, reverberation), and spectro-temporal modifications such as noise-vocoding or time-compression. Source degradations refer to characteristics of the talker that result in lower intelligibility, such as accents or speech disorders. The definition of a source degradation used in this thesis goes further in including any feature of a talker's acoustic-phonetic profile that results in a perceptual disadvantage for that talker (e.g., faster speaking rate).

Channel degradations are typically employed in speech perception and production research to simulate acoustic phenomena encountered in real life. For instance, masking and reverberation have been extensively applied to simulate challenging acoustic environments such as cocktail parties and auditoria (Cherry, 1953; Knudsen, 1929). Spectral degradations have been applied to simulate reduced frequency selectivity under sensorineural hearing loss or with cochlear implants (Nejime & Moore, 1998; Shannon et al., 1995). To study temporal aspects of speech perception, temporal degradations such as time-compression, interrupted speech and temporal envelope flattening have been applied (Dupoux & Green, 1997; Ghitza, 2012; Miller & Licklider, 1950). Time-compression has been a popular method to simulate fast speaking rates and to study effects of both acoustic degradation and increased information rate. In that respect, time-compression can be treated as artificial source degradation; however, typical compression rates applied in time-compression studies result in speaking rates outside the range of conversational speech (Koch & Janse, 2016).

The following sections provide an overview of source as well as channel degradations. The effects of acoustic degradations are discussed mainly for speech perception - a detailed discussion for speech production will be provided in Chapter 4 (Section 4.1).

1.2.1 Source degradations

Due to a multitude of factors, including differences in vocal tract shape and size, accents, or idiosyncratic features, some talkers are more intelligible than others (Bradlow et al., 1996; Hazan & Baker, 2011; Hazan & Markham, 2004; Munro & Derwing, 1995). *Idiosyncratic features* will be used as a term in this thesis to refer to characteristics of an individual speaking style that has not been elicited by specific task instructions. With regard to its behavioural consequences, such as reduced intelligibility, accented speech (and possibly other source degradations) has been compared to other forms of degraded speech (i.e., channel degradations) (Van Engen & Peelle, 2014). In the absence of anatomical-physiological and accent constraints, speech production is highly dynamic and dependent on the production context. Speech that has been acoustically modified to result in intelligibility benefits has been termed *clear speech* as opposed to unmodified conversational speech (see Smiljanic & Bradlow, 2009 for a review). On the one hand, clear speech occurs when talkers accommodate for listener constraints such as hearing loss (Cooke et al., 2014). On the other hand, talkers modify their speech based on the acoustic environment, a phenomenon referred to as *Lombard speech* (Lombard, 1911). Lombard speech has similar acoustic characteristics to listener-directed clear speech (Smiljanic & Bradlow, 2009), even though talkers have been shown to adapt to the specific needs of different listener groups (Cooke et al., 2014). Lombard speech modifications will be discussed in detail in Chapter 4 (Section 4.1). Even when talkers are recorded in the absence of an interlocutor, and without explicit instructions to speak clearly, acoustic-phonetic talker differences can be observed that are associated with higher or lower intelligibility (Bradlow et al., 1996; Hazan & Markham, 2004). To distinguish this type of speech from deliberate clear speech with communicative intent, it has been described as 'intrinsically' clear speech (see Chapter 2, Section 2.1, for a detailed discussion).

Acoustically, both types of clear speech have been characterised by global and segmental (or fine-grained) measurements (Bradlow et al., 1996; Smiljanic & Bradlow, 2009). Global measurements include spectral features such as energy

in specific frequency regions and fundamental frequency mean and range; and temporal features such as speaking rate, pause duration and amplitude modulations (Bradlow et al., 1996; Hazan & Markham, 2004; Krause & Braida, 2004; Picheny et al., 1986). Segmental measurements include consonant-vowel ratios, segment duration and vowel space expansion (Smiljanic & Bradlow, 2009). Expanded vowel spaces in particular have been observed in many studies (Bradlow et al., 1996; Cooke et al., 2014; Moon & Lindblom, 1989; Picheny et al., 1986), even though it has been debated whether expansion applies to the overall vowel space or only to specific vowel contrasts (Cooke et al., 2014). In conversational speech, individual vowel spaces are often characterised by underarticulation (or undershoot), i.e., the failure of a vowel production to reach its acoustic target (Lindblom, 1963, 1990). Acoustic enhancements (e.g., overarticulation) as part of clear-speech modifications are often attributed to the *hyper-hypo* model of speech production (H&H) (Lindblom, 1990). Acoustic modifications are considered to be the result of the talker trading off (listener) demands against system constraints, i.e., the aim to speak with minimal effort. For instance, clear speech elicited by talkers when communicating with an interlocutor exhibits less extreme acoustic enhancements than clear speech elicited solely through task instructions (Hazan & Baker, 2011). H&H explains this difference as follows: when a talker communicates with an interlocutor, acoustic modifications tend to fluctuate with listener demands, since there is no need for the talker to enhance her/his speech (minimal effort) when the listener signals comprehension (low demand). On the other hand, when asked to speak clearly, talkers enhance their speech more consistently since demands are determined by task instructions only.

Source degradations have been shown to interact with channel degradations; some studies demonstrated that clear speech can consistently improve speech intelligibility under different types of channel degradations (e.g., Green et al., 2007; Bent et al., 2009). The effect of talker acoustics on intelligibility under different channel degradations will be discussed in detail in Chapter 2 (Section 2.1).

1.2.2 Channel degradations

1.2.2.1 Environmental degradations: masking and reverberation

Even though many everyday conversations take place in quiet (Smeds et al., 2015), external sounds can interfere to some extent with the transmission of speech. This additive masking process has been studied extensively and two dimensions of masking are generally distinguished: energetic and informational masking (Mattys et al., 2012). This division corresponds to the respective effect of masking on specific sections of the auditory pathway (Wegel & Lane, 1924). The energetic interference of a masker with the speech target occurs at the auditory periphery, for instance when sounds of both target and masker excite a similar area on the basilar membrane (Wegel & Lane, 1924). On the other hand, informational masking affects central auditory processing and is elicited by "intelligible and meaningful content" (Mattys et al., 2012, p. 956), for instance in the presence of a competing talker (Cooke et al., 2008). Energy of a competing-talker masker tends to fluctuate so that its masking potential is usually lower than that found for stationary maskers (Festen & Plomp, 1990; Koelewijn et al., 2012). However, its perceptual interference with semantic and phonological processing (Heinrich et al., 2008) results in disruptions of a different kind usually reflected in increased cognitive load (Koelewijn et al., 2012; Wendt et al., 2018). Differences in masking potential and perceptual interference also affect speech production (Cooke & Lu, 2010). These effects will be discussed in more detail as part of a literature review on Lombard speech in Chapter 4 (Section 4.1).

Informational masking is often considered a 'catch-all' term, signifying any intelligibility-related effect of a masker that has not been accounted for by energetic masking (Cooke et al., 2008; Kidd, Mason, et al., 2008). However, this definition has been deemed too simplified (Kidd, Mason, et al., 2008). Instead, it has been suggested that informational masking is determined by the extent of perceptual overlap between target and masker, affecting attention and memory processes and creating perceptual uncertainty (Kidd, Mason, et al., 2008; Ren-

nies et al., 2019). For instance, Kidd, Best, et al. (2008) measured sentence recognition performance in an interleaved-word experiment. Target and masker sentences were presented in an interleaved fashion, i.e., odd-numbered words belonged to the target while even-numbered words belonged to the masker. Kidd, Best, et al. (2008) observed increased performance when target words were linked together, either by using recordings of the same talker, or by presenting all target words to the same ear. Since performance in a control condition with an unintelligible noise masker was similar to a control condition in quiet, it was suggested that the effect was attributable to a reduction in informational masking and not energetic masking.

Despite its excessive use as experimental degradation, masking does not account for all environmental degradations encountered in everyday life. Instead, as a large part of communication happens indoors, room acoustics assume a major role in degraded speech perception (Zahorik & Brandewie, 2016). Reverberation, which is defined as the rate of growth and decay of sound, is a main characteristic of room acoustics (Knudsen, 1929). It is determined by room size, wall surfaces and angles (Picou et al., 2016). Reverberant speech is characterised by direct energy arriving from the source, and late reflections, that cause masking and temporal smearing (Bolt & MacDonald, 1949; Picou et al., 2016; n.d.). Early reflections tend to "fuse" with the direct sound and might therefore even increase its intensity; however, late reflections can interfere with the direct sound, causing masking or even self-masking if a reflection of a sound overlaps with its direct component (Nabelek & Robinette, 1978). Reverberation has been described as temporal degradation (Gordon-Salant & Fitzgibbons, 1993); however, the acoustic effects of reverberation are more complex than those of purely temporal degradations such as uniform time-compression (see Section 1.2.2.3). The amount of reverberation is usually defined by the duration required for a signal level to decay by a fixed amount (Picou et al., 2016). For instance, a commonly used measure is *RT60*, which describes the time that it takes for a signal to decay by 60 dB. Typical classrooms exhibit reverberation times between 0.21 to 0.62 seconds (RT60), with the recommended limit being 0.6 seconds (Lucus et al., 2011).

Room acoustics can be simulated by means of head-related transfer functions that model the spatial relationship between talker and listener (Zahorik & Brandewie, 2016). However, speech convolved with such transfer functions has to be presented over headphones which bears methodological disadvantages. For instance, in order to evaluate the efficacy of hearing aids, researchers are constrained to pre-process the speech signal with individual testing algorithms (Oreinos, 2015). On the other hand, mathematical models have been proposed that are capable of reconstructing room acoustics in ecological multi-loudspeaker setups. For instance, Higher Order Ambisonics (HOA, Pulkki, 2001) encodes sound field recordings of any type of room that can be decoded for any type of loudspeaker setup; this feature allows the same room acoustics to be simulated in different laboratory settings.

Masking and reverberation are environmental degradations, i.e., they are dominant characteristics of natural communication environments. On the other hand, spectral degradations can occur due to technological imperfections such as a reduced bandwidth for telephone calls. Furthermore, spectral degradations are experienced by hearing-impaired listeners; signal-processing techniques such as noise-vocoding have been applied to simulate such degradations for normal-hearing listeners.

1.2.2.2 Spectral degradations: noise-vocoding

Speech can be degraded spectrally in a multitude of ways using filtering techniques. Spectral degradations are typically employed to simulate reduced frequency selectivity as a result of hearing impairment or to mimic hearing aids and prosthetics (Dubno & Schaefer, 1992; Nejime & Moore, 1998; Shannon et al., 1995). One method, noise-vocoding, has been applied in many studies to date; it is the process of spectrally distorting the spectrum of a speech signal while preserving its amplitude envelope (Shannon et al., 1995). Noise-vocoding is considered an approximation of the operating principles within a cochlear implant: speech is transduced, filtered into discrete frequency channels and sent directly to the auditory nerve of the patient, bypassing damaged areas (Schnupp

et al., 2011). In cochlear implants, the number of channels depends on the number of electrodes placed along the cochlea which is determined by insertion depth and by the minimal distance possible between electrodes (Schnupp et al., 2011). The noise-vocoding approach is only a crude approximation of speech perceived by cochlear implant users as the carrier signal is an electrical impulse and electrode arrays are usually inserted partly, causing spectral shifts that further distort the speech signal (Rosen et al., 1999).

The poor spectral resolution of noise-vocoded speech can lead to reduced intelligibility. However, listeners are generally able to achieve high performance, even with a small number of frequency channels. For instance, Shannon et al. (1995) showed that while performance in recognising consonants, vowels and sentences increased with the number of frequency channels, high performance was obtained with as few as four channels. In the case of vowel recognition, even the relative amount of energy in adjacent channels can be a cue to vowel formants (Dorman et al., 1997). However, it has been suggested that with both 4- and 8-channel noise-vocoding, which more closely mimics performance of cochlear implant users, listeners rely more on durational cues (Winn et al., 2012). In their study, normal-hearing listeners were presented synthetic vowels that varied along a continuum of spectral properties (e.g., formant change) and duration. Both spectral and durational cues can be used to distinguish between lax and tense vowels - in the case of Winn et al. (2012), the contrast /ɪ/-/i/. The authors showed that while listeners used both spectral and durational cues to distinguish between the two vowels, they relied heavier on durational cues when speech was noise-vocoded.

Due to its consistent and predictable modifications to the speech signal, listeners are generally able to adapt to noise-vocoded speech, i.e., improve word recognition performance with only little amount of training (Davis et al., 2005; Hervais-Adelman et al., 2008; Peelle & Wingfield, 2005). This *perceptual learning* (Samuel & Kraljic, 2009) is particularly strong when listening to noise-vocoded speech is accompanied by written or auditory feedback, allowing to re-map sounds to meaning (Davis et al., 2005; Hervais-Adelman et al., 2008).

Huyck & Johnsrude (2012) investigated the effect of attention on adaptation to noise-vocoded speech. During training, all participants were presented simultaneously with noise-vocoded sentences, auditory bursts and visual stimuli. Participants were divided into three groups; each group had to attend to only one of these inputs. A fourth control group did not receive any training. Only the group that attended to vocoded speech during training showed significant improvements over the control group. These results suggest that attentional processes play a significant role in perceptual learning of noise-vocoded speech.

While noise-vocoding has been applied to study the effects of spectral degradation on speech perception, time-compression has been used to investigate temporal degradations on speech perception.

1.2.2.3 Temporal degradations: time-compression

Time-compression comprises a range of methods that modify speech by removing segments, therefore increasing speaking (or information) rate. Even though time-compression can be applied to simulate fast speech, modern signal-processing techniques usually do not modify the spectral properties of speech, while natural fast speech exhibits temporal as well as spectral distortions (Koch & Janse, 2016). It is therefore not surprising that time-compressed speech is generally more intelligible than natural fast speech at similar rates since its spectral properties remain unchanged (Adank & Janse, 2009; Gordon-Salant et al., 2014). A popular time-compression method that maintains the spectral characteristics of speech is pitch-synchronous overlap-and-add (PSOLA, Moulines & Charpentier, 1990): shorter signal duration is achieved by removing pitch periods or fixed intervals of unvoiced speech. The idea for overlap-and-add techniques in general is derived from studies on interrupted speech that demonstrated preserved intelligibility despite the removal of entire speech segments (Janse, 2003; Miller & Licklider, 1950). It is evident, however, that intelligibility drops when too many segments are removed. Versfeld & Dreschler (2002) measured sentence recognition at different time-compression rates for two talkers. Psychometric functions were fitted to determine the proportion

of sentences that were recognised correctly as a function of speaking rate. Intelligibility dropped significantly from about 85-90% to 10% for speaking rates between about 10 to 16 syllables per second. Miller & Licklider (1950) proposed that the loss of intelligibility is purely acoustic: by removing more and more non-redundant speech cues, words become unrecognisable. More specifically, the removal of durational cues impairs the perception of several phonetic contrasts for vowels (e.g., short vs. long) and consonants (e.g., voiced vs. voiceless fricatives) (Klatt, 1976).

However, it has been suggested that a purely acoustic explanation, characterised by the removal of non-redundant cues, is not sufficient to explain certain findings related to time-compressed speech. For instance, it has been demonstrated that the loss of intelligibility through severe time-compression can in parts be restored when inserting intervals of silence into the speech signal (Ghitza & Greenberg, 2009). The authors found that inserting silence at regular intervals following 40 ms segments of time-compressed speech (factor 3, i.e., 33% of its original duration) considerably improved intelligibility. The improvement was dependent on the duration of the silence intervals, with maximal improvement for a duration of 80 ms. The authors linked their findings to oscillation accounts of speech perception (Ghitza, 2011; Peelle & Davis, 2012). The fundamental hypothesis of these models is that listeners 'entrain' to quasi-rhythmic fluctuations in amplitude that correspond to syllabic units, i.e., they employ brain oscillations of similar rate to track and resolve those units (Ghitza, 2014). In particular, oscillation models take into account not only the acoustic information per se (the 'what'), but also the distribution of acoustic information along the time axis (the 'when'). The TEMPO model (Ghitza, 2011) suggests two processes, parsing and decoding, that follow peripheral auditory processing. Speech is decoded through template matching, calculating coincidence across frequencies and time. The parsing mechanism at the heart of the model controls the speech decoding process. Similar to hands of a clock, parsing is done by an array of oscillators that work in a cascaded fashion. The master oscillator is *theta* (4-10 Hz) responsible for parsing syllables (2-8 Hz). The *beta* oscillator is a multiple of *theta* and parses segments within a syllable called dyads. These

dyads are located at the boundary between two phones and reflect the movement of articulators. Further down the cascade is the *gamma* oscillator which is a multiple of *beta* and associated with rapid spectro-temporal transitions, e.g., formant transitions. TEMPO suggests that intelligibility of time-compressed speech remains high as long as the insertion of silent parts results in syllabic segments that correspond to the oscillatory *theta* range (4-10 Hz). In Versfeld & Dreschler (2002), speaking rates resulting in the largest drop in intelligibility were between 10 and 16 syllables per second. This rate is precisely outside the range of *theta*, thereby demonstrating that intelligibility decrements might have been driven by disrupted syllable parsing. Conversely, Ghitza & Greenberg (2009) found that inserting silence intervals of 80 ms after 40 ms speech segments resulted in a syllable (or 'packet') rate of 8.3 per second, a rate that lies within the range of *theta*. In summary, TEMPO provides a theoretical framework that supports the involvement of a temporal component (*information speed*) in predicting intelligibility decrements through time-compression, in addition to acoustic degradations.

Similar to noise-vocoding, listeners have been found to adapt to time-compressed speech. Dupoux & Green (1997) observed that listeners quickly adapted to compression rates of 38% and 45% of the original duration, i.e., word recognition improved by about 10%. For 38% time-compression, performance stabilised after 15 sentences while for 45% time-compression, performance stabilised after only 10 sentences. Learning was found to be generalisable to other talkers: switching talkers after 10 sentences resulted in a brief drop in performance from which participants recovered after only two sentences. In contrast to noise-vocoded speech, adaptation to time-compressed speech is supposed to occur at the phonological level (Dupoux & Green, 1997; Golomb et al., 2007; Kennedy-Higgins et al., 2020; Pallier et al., 1998). Pallier et al. (1998) conducted a cross-linguistic study using time-compressed speech. In one experiment, monolingual Spanish speakers were trained on either 10 Spanish or 10 Catalan sentences and then tested on 5 Spanish sentences. All sentences were time-compressed to 38% of their original duration. A control group did not receive any training. Surprisingly, training improved

speech recognition even when conducted in Catalan, which was an unfamiliar language to all participants. Interestingly, the effect depended on whether the training and test languages were phonologically related to each other: while training on Catalan sentences improved speech recognition for Spanish native speakers, as outlined above, training on French sentences did not improve speech recognition for English native speakers. In contrast, training on Dutch sentences improved speech recognition for English native speakers, as the two languages are phonologically related.

Peelle & Wingfield (2005) investigated adaptation to time-compressed and noise-vocoded speech in comparison to speech masked by noise. The authors hypothesised that adaptation to time-compressed and noise-vocoded speech could be linked to increased task familiarity, and not necessarily to perceptual learning. If this was the case, adaptation effects should also be observed for masked speech. However, despite improved performance for time-compressed and noise-vocoded speech, masked speech was not associated with adaptation over time. This result indicated that adaptation to time-compressed and noise-vocoded speech was not solely due to task familiarity. It has been suggested that noise obscures the speech without systematically modifying it, which would be a requirement for perceptual learning (Bent et al., 2009; Mattys et al., 2012; Peelle & Wingfield, 2005).

Source and channel degradations presented above have been classified as adverse listening conditions (Mattys et al., 2012). However, adversity is ultimately dependent on the listener, as it refers to the "mismatch between external demands and internal resources to meet these demands" (Lemke & Besser, 2016, p. 79S). For instance, listener (or receiver) limitations can be considered internal factors contributing to the adversity of listening conditions. The definition of adversity in terms of individual listener constraints is important for understanding *listening effort,* that is assumed to arise when listening performance is maintained under adversity (Lemke & Besser, 2016).

1.3 Listener factors in degraded speech perception

Amplifiers of the detrimental effects of both source and channel degradations are limitations on the side of the listener (or receiver, see Mattys et al., 2012) such as hearing impairments. The hearing impairments discussed in this thesis will be limited to the common age-related peripheral sensorineural hearing loss, denoted by hair cell and neural cell loss (Gordon-Salant et al., 2011). Peripheral hearing loss is characterised by poorer sensitivity and frequency selectivity (Nejime & Moore, 1998). While many studies have investigated the influence of hearing impairment under specific types of degradations, the general problem for listeners is the loss of redundant speech cues (Mattys et al., 2012) which can occur under any type of degradation. For instance, peripheral hearing loss is associated with an age-independent decline in speech perception performance with energetic maskers (Goossens et al., 2017). Similarly, hearing loss significantly impairs speech perception under reverberation (Gelfand & Hochberg, 1976). In fact, challenging room acoustics are a frequent complaint, even when the effects of hearing loss are partly compensated for by hearing aids (Johnson et al., 2010; Zahorik & Brandewie, 2016).

In contrast to energetic maskers, peripheral hearing loss contributes less to difficulties experienced by older listeners with informational maskers. In fact, older listeners perform worse than young listeners even when both groups have normal hearing (e.g., Schoof & Rosen, 2014; Goossens et al., 2017). It has been suggested that the difficulties of older normal-hearing listeners with informational maskers and other types of degraded speech might be linked to temporal processing deficits (e.g., Gordon-Salant & Fitzgibbons, 1993; Goossens et al., 2017). Temporal processing can be diminished due to cochlear synaptopathy, i.e., the loss of cochlear nerve fibres, which can occur before observable changes in peripheral hearing (Sergeyenko et al., 2013). The role of temporal information for speech perception has been described in detail by Rosen (1992). Three temporal speech components are generally distinguished: temporal envelope (e.g., manner cues), temporal fine structure (e.g., formant patterns) and periodicity (cues to periodic vs. aperiodic segments). Temporal processing is usually

measured as an individual's ability to process temporal envelope and temporal fine structure cues (Goossens et al., 2017; Moore, 2008; Pichora-Fuller et al., 2007; Schoof & Rosen, 2014). Temporal processing deficits have also been linked to the difficulty in processing time-compressed speech (Gordon-Salant & Fitzgibbons, 1993). A detailed review of this literature will be provided in Chapter 3 (Section 3.1).

While peripheral hearing loss and temporal processing deficits represent diminished auditory processing abilities, cognitive processing abilities are another important factor for degraded speech processing (Lemke & Besser, 2016). For instance, processing difficulties with time-compressed and natural fast speech, as well as informational masking have been attributed to cognitive decline (Goossens et al., 2017; Janse, 2009; Salthouse, 1996), which refers to the decline of executive functions with age (Craik & Bialystok, 2006). Even though executive functions are not of primary interest in this thesis, they are crucial for models of effortful listening. To obtain a working definition of effort, I will therefore provide an overview of the three main executive functions as discussed in Diamond (2013).

Inhibitory control refers to the ability to control attention, behaviour, thoughts and emotions, in favour of existing predispositions. For instance, selectively attending to one voice while ignoring others in the famous cocktail party situation (Cherry, 1953) falls under inhibitory control. *Cognitive flexibility* is the process of adopting a different perspective (spatial or interpersonal) or strategy based on changes in demand. Cognitive flexibility requires inhibitory control in order to inhibit previous perspectives and strategies. *Working memory* (WM) is the process of temporarily storing information and also manipulating it, which distinguishes it from short-term memory (Baddeley & Hitch, 1994). WM is assumed to have limited capacity (Rönnberg et al., 2013) and has been assigned a major role in speech processing (Lemke & Besser, 2016; Rönnberg et al., 2013). The working memory model for Ease of Language Understanding (ELU, Rönnberg et al., 2008, 2013, 2019) classifies two types of processing, implicit and explicit processing. Incoming speech information is stored in an episodic buf-

fer and phonological information is constantly matched against lexical items stored in long-term memory. If the match is successful, then speech is implicitly processed, without demanding additional cognitive resources. However, in case of a mismatch, explicit processing sets in, invoking mainly executive functions that rely on WM capacity (Rönnberg et al., 2013). It is therefore not surprising that a larger WM capacity has been associated with improved explicit processing capabilities (Rönnberg et al., 2013). However, this relationship has been questioned recently for normal-hearing listeners. A meta-analysis by Füllgrabe & Rosen (2016) found that for normal-hearing listeners, there is little evidence for a role of WM capacity in the ability to identify speech in noise, which led to an adjustment of the ELU model (Rönnberg et al., 2019).

By name, the ELU model emphasises the 'ease' with which humans can implicitly process speech. However, speech processed under acoustic degradations generally requires explicit processing. It is therefore not surprising that research has started to focus on the 'unease' of processing speech and the *listening effort* associated with it. The following section will discuss the concept of listening effort and how physiological measures can be harnessed to predict effort.

1.4 Listening effort

The *Capacity Model of Attention* (Kahneman, 1973) can be considered the foundation of the concept of listening effort. A fundamental assumption of the capacity model is that cognitive resources are limited, but not fixed, so that a listener's motivation can ultimately influence how many resources are used (Wingfield, 2016). The capacity model has been adapted more recently in the *Framework for Understanding Effortful Listening* (FUEL, Pichora-Fuller et al., 2016). While there is substantial overlap between the two models, FUEL includes functions that are specific to listening. FUEL is shown in Figure 1. Kahneman coined the term "allocation policy", an executive function responsible for allocating cognitive resources to a task. The allocation policy is influenced predominantly by two factors: voluntary (or intentional) attention, and invol-

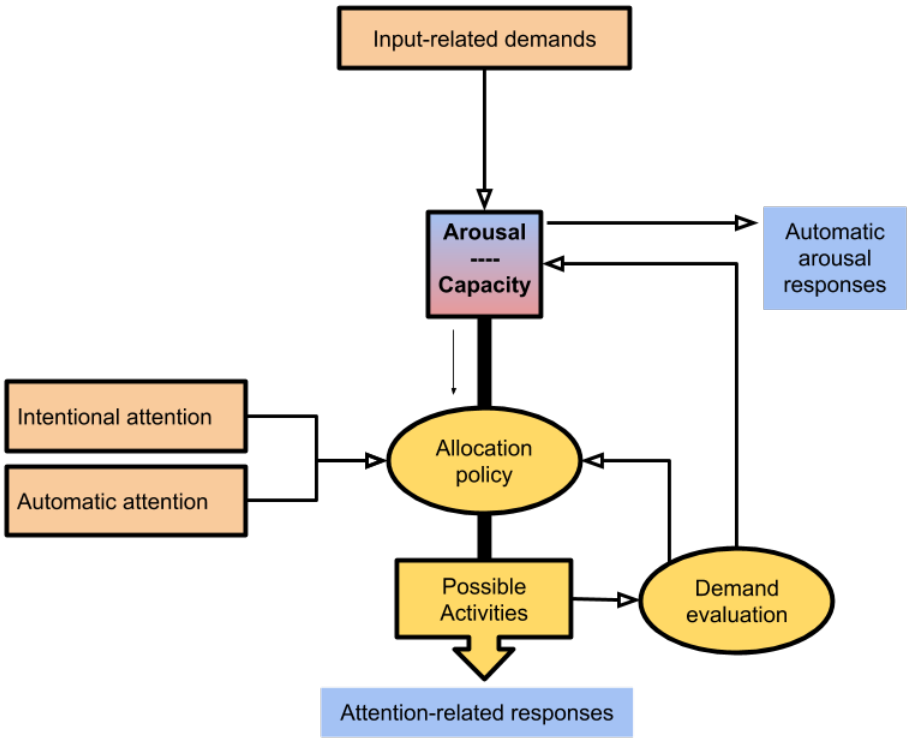

Figure 1: *The Framework for Understanding Effortful Listening (FUEL). Adapted and modified from Kahneman (1973) and Pichora-Fuller et al. (2016). The original components of the Capacity Model for Attention (Kahneman, 1973) are shown with names modified according to FUEL. In yellow: evaluation components. In orange: model inputs. In blue: model outputs.*

untary (or automatic) attention controlled by dispositions. Such dispositions can manifest as preferences to allocate more resources to novel stimuli (Kahneman, 1973). Task demands are constantly evaluated and can influence both the amount of available resources (i.e., capacity) and the allocation policy, i.e., how those resources are allocated. What is effort according to this model? A working definition of listening effort has been provided by McGarrigle et al. (2014): "the mental exertion required to attend to, and understand, an auditory message" (p. 434). With respect to FUEL, effort can be quantified as the amount of resources released from the capacity container and allocated to a task (see Figure 1). Lemke & Besser (2016) use a stricter definition of effort

that accounts only for the "extra processing load" that is required to maintain listening-task performance under increased external input demands.

Measures of task performance, such as the number of words recognised correctly, implicitly index the extent to which resources are allocated to a task. Especially for normal-hearing listeners, it is reasonable to assume that some amount of effort has been exerted when only half of all words in a sentence are recognised correctly. However, this assumption is contingent on the listener's state of engagement or motivation. If engagement is low, then performance might suffer, while effort remains low, as well. Similarly, if engagement is high, good performance might not indicate absence of effort: a certain amount of effort might be required to achieve good performance. It is therefore a plausible assumption that task performance does not adequately reflect effort.

Since resource allocation is influenced by subjective factors (e.g., motivation and capacity), asking listeners to quantify the amount of effort spent on a listening task might provide a suitable estimate of effort. Such perceived effort ratings are easy to collect and therefore widely used in hearing research (McGarrigle et al., 2014). However, perceived effort is subjective and thresholds for what is considered effortful vary across individuals (McGarrigle et al., 2014). Furthermore, it has been shown that perceived effort can be biased by subjective performance: in an online experiment employing the text reception threshold test, participants judged similarly effortful stimulus sets as more effortful when those sets contained more so-called skip trials, i.e., trials that were impossible to complete and therefore not expected to elicit effort at all (Moore & Picou, 2018). In addition, self-reports of perceived effort are collected offline after speech processing is complete (Wendt et al., 2017). It is therefore likely that such ratings do not only reflect listening demands, but also post-hoc decision making. While other behavioural measures of listening effort have been frequently used, such as speed of processing or performance on a secondary task (Pichora-Fuller et al., 2016), physiological measures tap into bodily responses to effort or stress (Mackersie & Cones, 2011). Such measures might therefore be more objective as they bypass the listener's interpretation of effort, and instead

index its immediate (yet indirect) effect on physiological arousal.

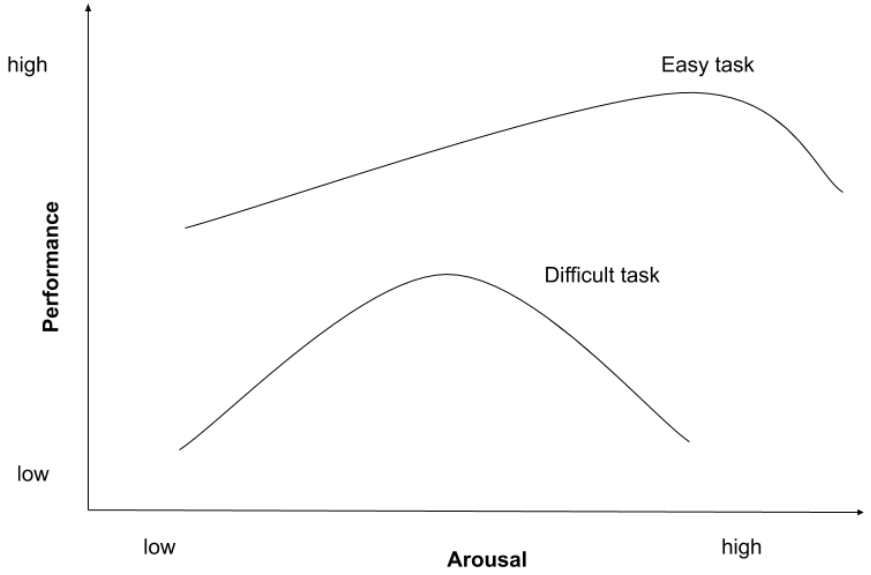

Figure 2: *Yerkes-Dodson law. Adapted and modified from Kahneman (1973).*

Arousal is supposed to "occur when an input change produces a measurable incrementing of a physiological […] or behavioural […] indicator over a baseline" (Pribram & McGuinness, 1975, p. 116). Effort has been characterised as a special case of physiological arousal that applies only to situations in which the agent is exerting her- or himself (Kahneman, 1973). For instance, while noise exposure can lead to a general increase in arousal, it does not necessarily mean that effort is exerted, as well. The practical implication of this definition is the following: when the level of engagement is known, physiological arousal can index effort. Given the simplicity of the experimental paradigms presented in this thesis, i.e., listening to semantically plausible but contextless sentences with known length, it can be assumed that intrinsic task engagement is similar across listeners. In other words, I will make the assumption that listeners are generally motivated to do the task, even though effective engagement might differ, depending on task performance.

The relationship between arousal and performance is non-linear: while some level of arousal is necessary to release cognitive resources, too much arousal will hinder access to resources (Gilzenrat et al., 2010; Kahneman, 1973). The Yerkes-Dodson law of arousal (Yerkes & Dodson, 1908) states that the relationship between performance and arousal can be expressed as an inverted U-shape (Figure 2). For easy tasks with higher baseline performance, more arousal is necessary to achieve optimal performance. On the other hand, for difficult tasks, similarly high arousal levels would result in a performance drop. Instead, moderate arousal is required for optimal performance. This principle has been demonstrated by Broadbent (1954): inducing physiological arousal through background noise exposure improved performance of a visual task when the task was easy, but impaired performance when the task was difficult. In single-task listening experiments conducted under controlled laboratory settings, arousal is usually not manipulated by secondary inputs. The Yerkes-Dodson law is therefore only in parts applicable, as the arousal level is primarily determined by task demands, which in turn also affect performance.

Figure 3 shows a modified version of FUEL that has been adapted for the purpose of this thesis. First of all, the two types of degradations considered in this thesis are highlighted: source and channel degradations. Furthermore, the model emphasises the relationship between arousal (for simplification equated with effort) and capacity. Two different listening scenarios are illustrated, normal listening and effortful listening (Lemke & Besser, 2016). For normal listening, input demands cause low to moderate levels of arousal, and the respective pressure on capacity releases relevant resources. For effortful listening, input demands cause moderate to high levels of arousal; the increased pressure on capacity leads to extra resources being released. The two extreme cases are not displayed: while very low arousal levels might not raise enough pressure to release resources, very high arousal levels might exceed capacity and the resulting pressure would prevent the release of resources entirely.

In summary, listening effort can be estimated by measuring physiological arousal in response to changing input demands. Assumptions are that arousal

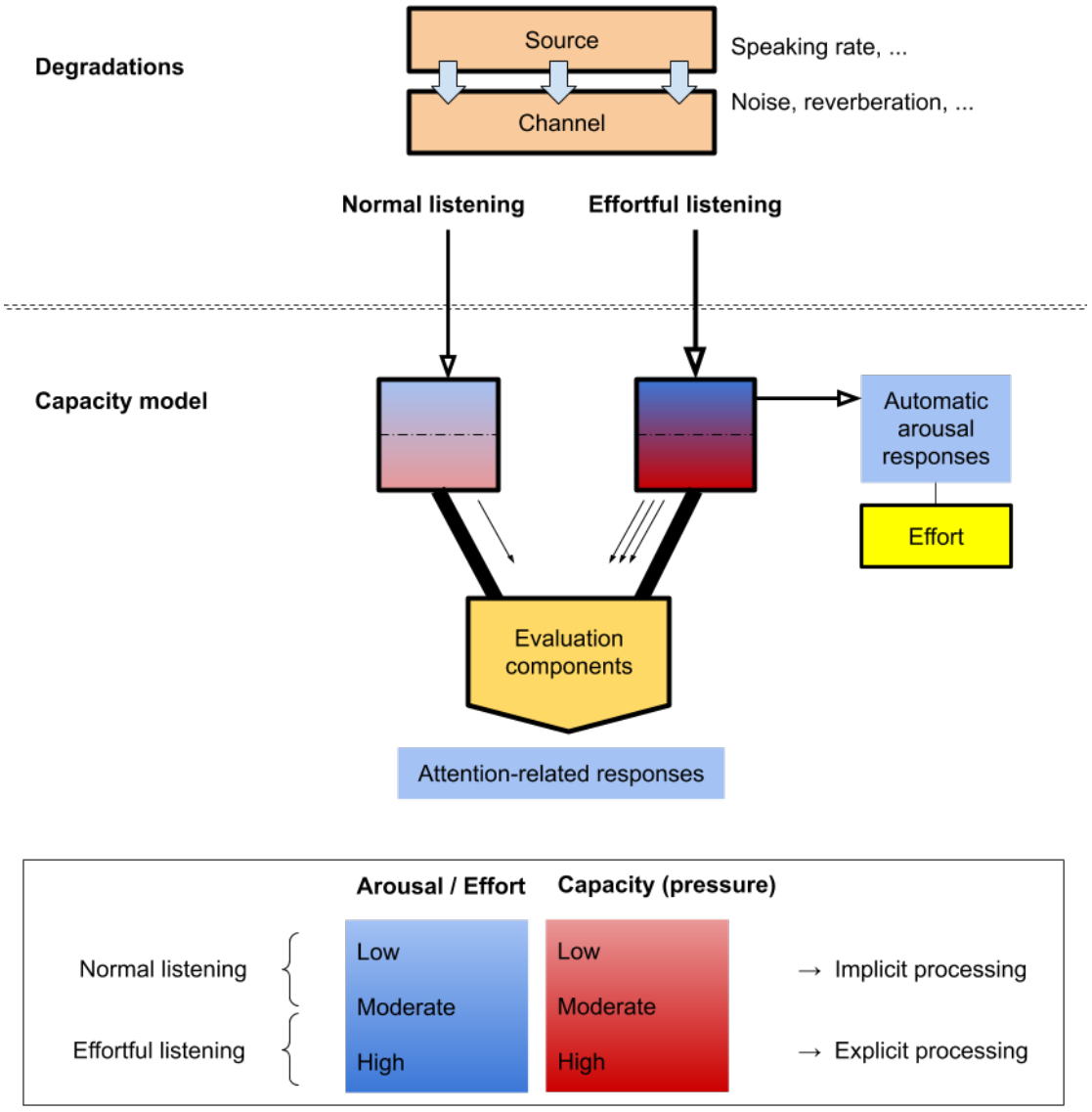

Figure 3: *Interpretation of the Framework for Understanding Effortful Listening with emphasis on the relationship between arousal and capacity. Adapted and modified from Kahneman (1973) and Pichora-Fuller et al. (2016). The evaluation components of the capacity model have been summarised. Two listening scenarios are depicted with inspirations taken from Lemke and Besser (2016).*

is neither too low nor too high and that listeners are sufficiently engaged in the task. In the next section, I will discuss how arousal can be indexed by sympathetic nervous system activity. Specifically, I will explain how measures of pupil dilation can be used to quantify effort in degraded listening conditions.

1.5 Pupillometry and listening effort

1.5.1 Physiology of the pupil dilation response

As outlined in the previous section, effort can be linked to physiological arousal. The body's arousal response is mainly initiated by the locus coeruleus (LC) in the brain stem that is responsible for release of the neuromodulator norephinephrine (Eckstein et al., 2017). Locus coeruleus activity is closely coupled with pupil dilation and constriction (Rajkowski et al., 1993) given its inhibitory projections to the Edinger-Westphal nucleus and excitatory projections to sympathetic divisions of the spinal cord (Eckstein et al., 2017). On the one hand, inhibitory activity within the parasympathetic Edinger-Westphal complex indirectly causes the pupil to dilate by countering its constriction by the sphincter pupillae muscle. On the other hand, excitatory activity in the sympathetic nervous system causes the pupil to dilate. Specifically, sympathetic neurons in the spinal cord (C8-T2, the ciliospinal center) enter the sympathetic trunk through the ramus communicans albus, forming synapses with postganglionic neurons in the superior cervical ganglion. Those neurons project to the eye where they innervate the dilator pupillae muscle, causing the pupil to dilate (Faller & Schünke, 1995).

The most intuitive observation is that pupil size changes in response to light - specifically, the pupil constricts. The average onset latency of this reflex ranges from 248 ms (high light intensity) to 322 ms (low light intensity), with intersubject variability (Bergamin & Kardon, 2003). On the other hand, the pupil dilation reflex is observed in response to a range of cognitive tasks involving short-term memory (Kahneman & Beatty, 1966; Piquado et al., 2010), pitch dis-

crimination (Kahneman & Beatty, 1967) and mental arithmetic (Klingner et al., 2011). Pupil dilation in response to cognitive processing is relatively fast and can reach a maximum within one second of stimulus presentation (Kahneman, 1973). Piquado et al. (2010) investigated the pupil dilation response during listening to spoken-digit lists of varying length (4, 6 or 8 items). Pupil size increased over the course of the list presentation, indicating local peaks following digit arrival and a global peak in the retention interval (3s) prior to list recall. The global peak was interpreted as the cumulative memory load, which was dependent on list length, with larger peaks observed for longer lists.

Pupil dilation has also been measured previously in response to motor planning and execution. In general, movement-related pupil responses (MRPRs) are large and can take up a considerable proportion (up to 70%) of the task-evoked pupil dilation (Hupé et al., 2009; McCloy et al., 2016; Richer & Beatty, 1985). Richer & Beatty (1985) investigated MRPRs during finger flexion. They showed that peak amplitude at around 500 ms post-movement increased as a function of the number of subsequent finger flexions (1, 2 or 3). Interestingly, the pupil response started before movement execution, indicating sensitivity to both motor planning and execution. Despite its high sensitivity to movement, pupil dilation has also been shown to reflect cognitive processing during speech production. For instance, recent studies investigated syntactic and semantic processing, as well as turn-taking behaviour (Barthel & Sauppe, 2019; Papesh & Goldinger, 2012; Sauppe, 2017) (see Chapter 4, Section 4.1). However, none of these studies have investigated speech production in noise, which will be the focus of Chapter 4.

1.5.2 Pupil dilation and degraded speech

For audiological researchers, the main application of pupillometry is the study of sentence processing under varying degrees of degradation. Ideally, such measures benefit the decision making concerning interventions (e.g., hearing aids) or serve as a general diagnostic tool (McGarrigle et al., 2014). For sen-

tence processing, the pupil dilation reflex typically sets in 0.5-1.3 s following sentence onset and peaks around 0.7-1.0 s after sentence offset (Winn et al., 2018). This morphology applies generally to normal-hearing listeners, but differences in number and latency of peaks have been observed for hearing-impaired populations such as cochlear implant users (Wagner et al., 2019). A standardised pupillometry protocol for sentence processing has emerged in recent years. The sentence is presented while pupil size is measured simultaneously using an eye-tracker. Sentence offset is followed by a retention period to allow the pupil dilation to reach its peak. Afterwards, participants are usually asked to repeat back as many words as they could identify. Outcome measures of such a paradigm are therefore word recognition performance (usually in %) and various components of the pupil dilation function, such as peak and mean dilation, and peak latency (Winn et al., 2018).

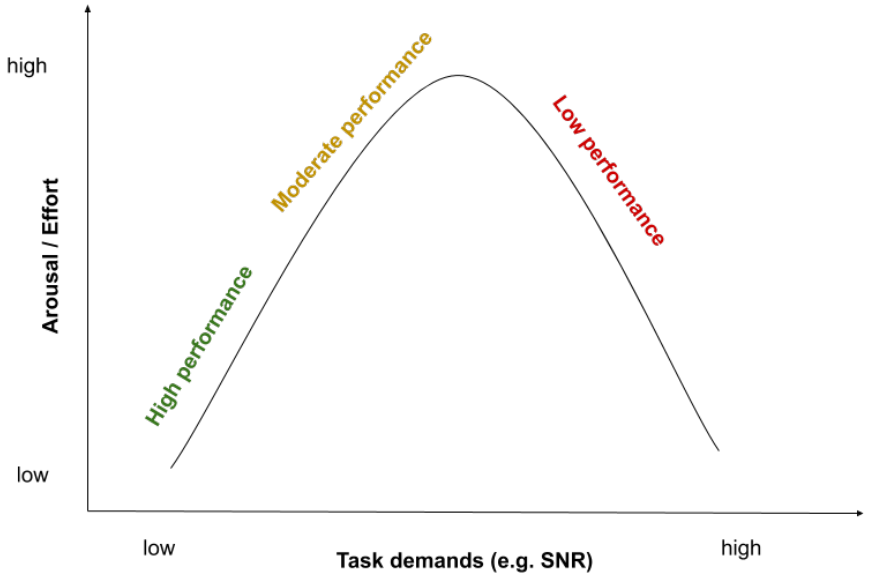

Figure 4: *Proposed relationship between arousal/effort and task demands. Three discrete levels of performance are indicated as sections of the curve. Based on results from Wendt et al. (2018).*

Wendt et al. (2018) measured pupil dilation during sentence recognition, at

various signal-to-noise ratios (SNRs). Stationary noise was added at +8 to -20 dB SNR in 4 dB steps. Moreover, the noise was filtered to simulate the spectral masking properties of speech. Performance was co-registered by asking participants to repeat back as many words as possible. While word recognition was stable around 100% between +8 to 0 dB SNR, peak pupil dilation increased from +4 dB SNR onwards. As performance slowly decreased below 0 dB SNR, peak pupil dilation increased drastically to reach its ceiling value at -4 dB SNR. During the steepest performance drop (-4 to -8 dB SNR), peak pupil dilation did not increase further, but started to decrease from -12 dB SNR onwards. The peak pupil dilation as a function of word recognition performance at varying SNR can therefore be described as an inverted U-shape (Figure 4). There are three crucial properties of this function: (1) despite high performance, arousal rapidly increases; (2) arousal reaches its ceiling when performance is still moderate; (3) arousal remains briefly at its peak when performance declines significantly and then drops at very low performance, possibly due to disengagement. A similar relationship between word recognition and peak pupil dilation has been found for noise-vocoded speech. Winn et al. (2015) investigated the effect of varying numbers of noise-vocoder channels on the peak pupil dilation. Similar to Shannon et al. (1995), the amplitude envelope of speech was used to modulate noise bands. The resulting noise-vocoded signal was chosen to have either 32, 16, 8 or 4 channels. While word recognition only slowly declined, with its largest drop in performance between 8 and 4 channels, peak pupil dilation increased even when sentences were still fully intelligibility. As performance did not fall below 40%, Winn et al. (2015) did not report any decline in pupil dilation due to disengagement. Taken together, studies that vary acoustic degradations along the dimension of a single parameter (e.g., SNR) tend to observe a U-shaped relationship between intelligibility and pupil dilation (e.g., Winn et al., 2015; Ohlenforst et al., 2017; Wendt et al., 2018). Other studies have specifically compared the effects of different stimulus 'dimensions' on the pupil dilation (e.g., energetic vs. informational masking), when controlling for intelligibility levels (Koelewijn et al., 2012; Wendt et al., 2018). For instance, Koelewijn et al. (2012) presented listeners with speech at fixed intelligibility levels (50%

and 84%) by adaptively changing signal-to-noise ratios to determine suitable speech-reception thresholds (SRTs) for each listener. In addition, SRTs were determined independently for different masker types (stationary noise, fluctuating noise and competing-talker masker). Speech presented in the presence of a competing-talker masker elicited a larger pupil dilation response, compared to stationary and fluctuating noise, indicating higher listening effort. At the same time, SRTs were lowest for the competing-talker and the fluctuating noise masker, as would be expected given the release from masking associated with energetic dips (Festen & Plomp, 1990). The effort exerted to obtain the respective SRTs, as indexed by pupil dilation, appears to be determined by masker characteristics inherent to the competing talker. As outlined in Section 1.2.2.1, it is likely that the intelligible and meaningful content of the masker interfered with speech perception on phonological or semantic levels, leading to increased effort (Kidd, Mason, et al., 2008; Kidd, Best, et al., 2008; Koelewijn et al., 2012; Wendt et al., 2018).

A number of other studies have employed adaptive procedures to equate for intelligibility levels across different targets, maskers, and listener groups (Borghini & Hazan, 2018, 2020; Paulus et al., 2019; Zekveld et al., 2010, 2014). For instance, Borghini & Hazan (2018) established that non-native listeners exert more effort than native listeners despite similar word recognition performance. While adaptive procedures have been applied in many pupillometry studies, Winn et al. (2018) advised against the use of pupillometry during the adaptive procedure itself due to the noisiness of single trial measurements. Further issues with adaptive tracking arise because the uncertainty about the stimulus can influence pupil dilation. For instance, Francis et al. (2018) could show that pupil dilation was larger when stimuli with varying difficulty were presented in a less predictable mixed (event-related) design, as opposed to a block design, which is commonly employed in pupillometry studies.

Only recently has pupillometry been applied to the study of time-compression: Müller et al. (2019) observed larger pupil dilation in response to fast than slow speech. In their study, speech was time-compressed to 75% of its original dura-

tion, corresponding to a speaking rate of 5.1 syllables per second. Additionally, speech was time-elongated to 125%, resulting in a speaking rate of 3 syllables per second. Intelligibility was fixed at 80% by applying an adaptive procedure with a single-talker masker. Peaks for fast speech were not only found to be larger, but also occurred earlier, possibly linked to shorter sentence duration.

Apart from the task-evoked pupil dilation reported above, several other measures have been derived from pupil size recordings. However, these measures have often been discarded in past research (Winn et al., 2018). For instance, McGarrigle et al. (2017) investigated how pupil dilation changes over the course of a sustained listening task using short text passages (13-18 s) instead of sentences. Speech was presented in babble noise at either +15 dB SNR (easy) or -8 dB SNR (hard), which resulted in high intelligibility in both conditions. Results showed that after an initial dilation peak (i.e., the task-evoked pupil response), pupil size steadily declined over the course of a text passage. In the second half of a listening block (after ~10 min), the authors observed that pupil dilation declined faster in the hard compared to the easy listening condition, which was interpreted as an earlier onset of fatigue.

Another related pupil size measure that has only recently been employed in hearing science is the baseline pupil size (Ayasse & Wingfield, 2020; Wagner et al., 2019). While both task-evoked pupil dilation and baseline pupil size have been linked to arousal (McGarrigle et al., 2017; Wagner et al., 2019), baseline pupil size refers to the more general state of arousal: while small baselines have been associated with states of inattentiveness, large baselines indicate states of distraction (Unsworth & Robison, 2016). At intermediate baseline levels, attention is assumed to be focused; this assumption links to the Yerkes-Dodson law of arousal that predicts higher performance at intermediate arousal levels. Baseline pupil size changes across trials have been analysed to differentiate between sustained attention demands for normal-hearing and hearing-impaired listeners (Wagner et al., 2019). Ayasse & Wingfield (2020) measured baseline pupil size changes across trials for listeners with varying hearing thresholds. As expected, task performance, in this case sentence comprehension, was poorer

for listeners with poorer hearing. In addition, baseline pupil size declined faster for listeners with poorer hearing. Initially, it was suggested that this effect indicated higher fatigue in listeners with poorer hearing. However, results also showed that baseline pupil size at the beginning of the experiment was larger for listeners with poorer hearing, suggesting higher arousal under elevated task demands which then progressively declined during the experiment. This hypothesis was corroborated by the finding that task performance improved across trials.

1.6 Summary and thesis outline

The presented literature review discussed the physical properties of different types of degraded speech and their perceptual consequences. Furthermore, receiver limitations were discussed and it was outlined how listeners can cope with degraded speech by drawing on cognitive resources. I presented models of listening effort and discussed how physiological arousal can serve as an index of effort in listening tasks involving degraded speech. Despite the popularity of pupillometry, several gaps can be identified in the literature. First of all, research has focused almost exclusively on the involvement of channel degradations in effortful listening. Various types of background noise have been evaluated with respect to their impact on pupil dilation during sentence recognition. While focusing on a limited set of (mostly environmental) channel degradations research has missed out on other potential contributors to listening effort. Acoustic-phonetic differences between talkers affect the intelligibility of speech, especially in combination with channel degradations (Hazan & Markham, 2004) and it is conceivable that such source factors also influence listening effort. Furthermore, while objective measures of effort are sought of to understand the difficulties of listeners in real-world acoustic environments, little research has focused on situations that are in fact realistic. Such situations can involve both talker and listener constraints (source degradations and receiver limitations) and often occur in quiet, with room acoustics altering the acoustic environment. Finally, while many pupillometry studies have focused

on speech perception in noise, no study to date has applied pupillometry to speech production in noise. Since communication is rarely one-directional, but involves both listening and speaking, it is surprising that little research has aimed at developing a definition of *speaking effort*, analogous to listening effort (McGarrigle et al., 2014). Given those gaps in research, the aim of this thesis is therefore to take a holistic approach to the study of effort in spoken communication. Specifically, the overarching research question can be stated as follows:

- How do acoustic degradations affect individuals at cognitive and acoustic levels, in both communicative roles, i.e., during listening and speaking?

To address the overarching research question, several experiments were conducted. Each experiment targeted one or more individual research questions (RQs). In order to answer each RQ, a number of hypotheses were tested, as shown below. The specific patterns of data expected to confirm these hypotheses are provided in each chapter's methods section (see Section 2.2.7, 3.2.5 and 4.2.5).

Chapter 2

- RQ1: Do source degradations that affect intelligibility interact with channel degradations?
  - H1: The same source characteristics that predicted intelligibility in previous studies will also predict intelligibility in the current study.
  - H2: Talker intelligibility will persist across channel degradations, as reflected by correlations amongst condition pairs.
  - H3: The same source characteristics that predicted intelligibility will also predict adaptation, i.e., the change in intelligibility over time.
- RQ2: Do source degradations interact with channel degradations that affect listening effort, as measured by pupil dilation?
  - H4: Pupillometry results with respect to degraded speech will be replicated, indicating higher effort when listening to degraded speech than speech in quiet.

  - H5: Pupillometry applied across trials will reflect differences between degradations with respect to adaptation.
  - H6: Source characteristics that predict intelligibility will also predict listening effort, as measured by pupil dilation.

To test hypotheses H1-H6, a multi-talker corpus was recorded and analysed; the sentence material was then used in a listening experiment combined with pupillometry, conducted with normal-hearing listeners.

Chapter 3

- RQ3: Does speaking rate affect effort experienced by hearing-impaired listeners even when intelligibility is high?
  - H7: Fast speech will be associated with higher listening effort, as measured by pupil dilation and perceived effort ratings.
  - H8: The effect of speaking rate on listening effort will persist in a second testing session (retest).
- RQ4: Do room acoustics affect listening effort experienced by hearing-impaired listeners even when intelligibility is high?
  - H9: Reverberation will be associated with higher listening effort, as measured by pupil dilation and perceived effort ratings.
  - H10: The effect of reverberation on listening effort will persist in a second testing session (retest).
  - H11: Reverberation will amplify the detrimental effect of a fast speaking rate on listening effort.
  - H12: The effect of reverberation on listening effort will be reduced by a dereverberation algorithm.

To test hypotheses H7-H12, a study with hearing-impaired listeners in realistic acoustic environments was conducted, involving talker-, listener- and environmental constraints. Realistic listening situations were simulated using higher-order ambisonics in conjunction with time-compression at high intelligibility levels. Time-compression was used as a proxy for faster speaking rates, which have been shown to be particularly problematic for hearing-impaired listeners.

Chapter 4

- RQ5: Can pupillometry be combined with a speech production paradigm to measure speaking effort, as equivalent to listening effort?
  - H13: Talkers' acoustic adaptations observed under channel degradations in previous studies will also be observed in the current study.
  - H14: Findings from speech perception regarding pupil dilation under channel degradations can be replicated for speech production, indicating higher effort under informational masking.

To test hypotheses H13-H14, pupillometry was combined with a speech production paradigm in which participants were asked to produce speech in quiet and in background noise.

Chapter 5

In Chapter 5, the empirical work presented in this thesis is discussed in the context of models of speech perception and production, and listening effort presented in this chapter. The models are refined to accommodate the findings presented in the following chapters. Furthermore, practical implications, but also limitations of the presented work are highlighted.

Chapter 2: Source and channel degradations and their effect on intelligibility and effort

2.1 Introduction

In Chapter 1 of this thesis, channel degradations were introduced and their effect on intelligibility and listening effort were discussed. Source degradations have been studied to a lesser extent than channel degradations with respect to the effort they impose on the listener. Furthermore, channel and source degradations are known to interact (Hazan & Markham, 2004), but the specifics of this interaction, specifically for varying channel degradations (Bent et al., 2009) have yet to be uncovered. The current chapter will first examine how talkers can be classified in terms of their acoustic-phonetic features and how these source characteristics affect intelligibility under different channel degradations. I will then discuss the few studies investigating the effect of source characteristics on listening effort, employing pupillometry paradigms. An experiment is presented that investigated the combined effect of channel and source degradations on intelligibility, effort and adaptation.

2.1.1 Intelligibility under source and channel degradations

Acoustic variations between talkers arise due to accent differences (Bradlow & Bent, 2008), but also due to idiosyncratic and anatomical-physiological differences (Hazan & Markham, 2004) (see Chapter 1, Section 1.2.1). Acoustic-phonetic features such as vowel space, energy in speech-critical bands, and speaking rate predict intelligibility in quiet or in noise (Bradlow et al., 1996; Hazan & Markham, 2004). However, the extent to which such acoustic-phonetic features affect the intelligibility of talkers varies in the results of those studies. It is possible that differences between studies might stem from the range of possible speaking styles that talkers can employ to achieve high intelligibility. Bradlow et al. (1996) linked acoustic-phonetic characteristics

of 20 talkers to intelligibility in quiet. Vowel space dispersion was chosen as a measure of articulatory precision. Vowel space dispersion measures the articulatory distances of vowel productions from the vowel space center with larger distances associated with clear as opposed to conversational speech (Bradlow et al., 1996; Picheny et al., 1986). The authors observed that vowel space dispersion was significantly correlated with intelligibility for the 10 most intelligible talkers [$\rho = +0.698, p = 0.036$], even though this correlation did not reach significance when all 20 talkers were considered [$\rho = +0.431, p = 0.060$]. In addition, female talkers were overall more intelligible, so that fundamental frequency was also (weakly) correlated with intelligibility. Hazan & Markham (2004) analysed a corpus of 45 talkers and linked acoustic-phonetic features to intelligibility. In contrast to Bradlow et al. (1996), who presented speech in quiet, Hazan & Markham (2004) presented speech in noise (babble at $+6$ dB SNR), to avoid ceiling effects. Interestingly, the authors did not find correlations between vowel space size and intelligibility. However, word duration (reflecting speaking rate) was correlated significantly with intelligibility [$r = 0.382, p = 0.01$], contrary to Bradlow et al. (1996). The correlation was particularly strong for adult male talkers [$r = 0.672, p = 0.006$]. It has to be noted that stimuli in Hazan & Markham (2004) were words and not sentences as in Bradlow et al. (1996). Studies that manipulated speaking rate artificially were usually not successful in achieving higher speech intelligibility (Picheny et al., 1989; Uchanski et al., 1996). It has therefore been suggested that a slower speaking style only indirectly contributes to higher intelligibility because it is linked to more precise articulation (Hazan et al., 2018; Hazan & Markham, 2004). On the other hand, results in Bradlow et al. (1996) also showed that articulatory precision, as reflected by greater vowel space dispersion, does not necessarily coincide with a slower speaking rate. The importance of vowel spaces has been emphasised by models of speech production as discussed in Chapter 1 (Liljencrants & Lindblom, 1972; Lindblom, 1990). Perceptually, listeners prefer overarticulated vowels (Johnson et al., 1993; Johnson, 2000). In Johnson (2000), listeners chose the best example amongst 330 versions of a word, each with a different synthesised vowel production.

Vowels were modified from recordings of one male talker, sampling the vowel space across the $F_1 - F_2$ plane. In comparison to the vowel space derived from the original vowel productions, listeners preferred an overall larger vowel space, the so-called hyperspace.

To summarise the research outlined above, talkers appear to differ in their 'intrinsic' intelligibility and some acoustic-phonetic features have been associated more strongly with intelligibility than others. It has to be noted that intrinsic intelligibility in this case refers to speech produced without communicative intent, i.e., recordings with isolated sentences and without the presence of an interlocutor. Talkers are known to modify their speech in the presence of noise and even more so when communicating (e.g., Cooke & Lu, 2010, see Chapter 4, Section 4.1).

While the above-mentioned studies have investigated talker intelligibility in quiet and in noise, there has also been recent interest in other channel degradations. For instance, a study by Johnson et al. (2020) investigated effects of talker acoustics and gender on intelligibility under time-compression. The speech of two female and two male talkers was presented in babble noise (0 dB SNR), at their original speaking rate or time-compressed to 66.7%. While original speaking rates were similar for all four talkers (4.83-4.97 syllables/s), vowel space perimeter was largest for the two male talkers ($female: 13.88, 14.23$ Bark; $male: 15.31, 16.03$ Bark). For the female talker with the smallest vowel space perimeter, both time-compressed and normal-rate speech was less intelligible. Even though the second female talker was the most intelligible at the original speaking rate, the intelligibility decrement through time-compression was largest for both female talkers. Reduced intelligibility under time-compression was thus linked to vowel space differences. However, given the small number of talkers in the study, these findings cannot be considered conclusive.

Some studies investigated the effect of talker differences on both masked and vocoded speech (Bent et al., 2009; Green et al., 2007). Bent et al. (2009) presented speech by 20 talkers either in babble noise (0 dB SNR) or under sine-wave vocoding (eight channels). Intelligibility of each talker in each condition

was determined based on recognition data from ten unique listeners. Intelligibility in babble noise was significantly correlated with intelligibility under vocoding [$r = 0.73, p < 0.001$]. Similarly, Green et al. (2007) investigated intelligibility of six talkers across three degradations: babble noise (+6 dB SNR) and noise-vocoding with four and eight channels. In addition, speech was presented to cochlear implant users. Talkers were divided into high and low intelligibility groups, based on results from Hazan & Markham (2004). Consistently across degradations and listeners, performance was better for speech by high- than low-intelligibility talkers. Noise and noise-vocoding degrade speech in different ways, and it is therefore conceivable that combinations of acoustic-phonetic features were responsible for preserved talker differences. Green et al. (2007) suggested that temporal properties such as longer word duration benefited intelligibility since spectral detail is removed by the speech processor in cochlear implants. In their study, word duration and mean energy in the 1-3 kHz range were both found to be positively correlated with intelligibility. As both acoustic-phonetic measures were also inter-correlated that might explain why some talkers were more intelligible in both conditions: increased energy in the 1-3 kHz range benefited word recognition in noise while longer word duration benefited word recognition with (simulated) cochlear implant speech processor. However, the set of talkers was a small subset (N = 6) taken from Hazan & Markham (2004) (N = 45) that did not observe this correlation of energy and word duration.

Taken together, talker differences in intelligibility appear to be preserved across different channel degradations. However, it is not yet known which acoustic-phonetic features enable higher intelligibility across degradations. It is also possible that features promoting higher intelligibility differ between talkers, as suggested by Hazan & Markham (2004). For instance, high intelligibility for talker A could be signified by a very slow speaking rate, while high intelligibility for talker B could be a combination of a moderately slow speaking rate and a moderately large vowel space.

2.1.2 Listening effort and adaptation under source degradations

As outlined in Chapter 1 of this thesis, input demands in the FUEL framework (Pichora-Fuller et al., 2016) have also been divided into channel and source factors. Pupillometry studies have predominantly focused on channel degradations, showing that pupil dilation increases as intelligibility decreases, for instance due to decreasing signal-to-noise ratio (Wendt et al., 2018). However, the pupillometry literature investigating source degradations is sparse. Some studies have aimed at quantifying listening effort during accented speech processing. A recent study using Chinese-accented speech showed that pupil dilation increased with the intelligibility of the accent, following a similar pattern as observed for channel degradations (Porretta & Tucker, 2019). Similarly, McLaughlin & Van Engen (2020) investigated the effect of Chinese-accented speech on pupil dilation; however, the authors only used high-intelligibility sentences, with scores obtained in a separate experiment. English sentences were presented, spoken by either native English or native Mandarin-Chinese speakers. Results indicated larger pupil dilation and higher subjective effort when listening to accented compared to native speech. Another recent study also measured subjective effort for Mandarin-Chinese accented speech in comparison to native speech and found that it was more effortful to process for native listeners under various background noise and reverberation conditions (Peng & Wang, 2019).

Other studies have applied pupillometry to measure listening effort elicited by speech from native speakers. Koch & Janse (2016) investigated the effect of speaking rate on listening effort in quiet. Talkers ($N = 49$) were taken from a corpus of conversational speech, i.e., speaking rate differences were based on idiosyncratic talker characteristics. Listeners were sampled from different age groups. Koch & Janse (2016) did not observe any systematic relationship between speaking rate and pupil dilation, regardless of listener age group. Since sentences were spoken by native speakers, it is possible that speaking rate effects would only emerge when speech is perceived in background noise; as channel and source degradations interact, the overall difficulty of the task in-

creases. For instance, Simantiraki et al. (2018) observed that Lombard speech (as opposed to conversational speech) was associated with increased intelligibility and decreased pupil dilation when presented in stationary noise (-1, -3 and -5 dB SNR). Similarly, Borghini & Hazan (2020) investigated the effects of clear speech (elicited by task instructions) on pupil dilation at fixed intelligibility levels. An adaptive procedure was applied to equate intelligibility across participants at 50%. Interestingly, while the resulting signal-to-noise ratios were lower for clear speech, the elicited pupil dilation was reduced in comparison to conversational (plain) speech. The results indicated that despite less favourable noise levels, listening effort was overall reduced for clear speech.

To counter the detrimental effects of channel degradations, listeners implement cognitive strategies such as perceptual learning, as discussed in Chapter 1 of this thesis. Even short-term exposure to degraded speech can thus improve speech recognition. Listeners have also been shown to adapt to source degradations, e.g., accents (Banks et al., 2015), but also to idiosyncratic talker differences. For instance, Adank & Janse (2009) investigated adaptation to both time-compressed and natural fast speech. The authors recorded sentences spoken by one male Dutch speaker, asked to produce speech at normal and fast rates, resulting in 4.7 and 10.2 syllables per second, respectively. Time-compressed versions of the normal-rate sentences were created by matching duration per sentence to that of the fast-rate sentences. Comprehension was tested by asking participants to judge whether a sentence was true or false (e.g., *Beavers build dams in the river*, translated from Dutch). While time-compressed speech resulted in overall high accuracy, natural fast speech was significantly less intelligible, resulting in less accurate responses. Participants adapted to natural fast speech within the experimental block containing 60 sentences. Since baseline accuracy under time-compression was too high overall, no adaptation was observed. However, adaptation to time-compressed speech has been shown repeatedly in other studies (see Chapter 1, Section 1.2.2.3). Bent et al. (2009) observed differences in adaptation to speech by different talkers in combination with noise and sine-wave vocoding. However, it was not investigated further which acoustic-phonetic characteristics drove those differences.

It has been suggested that the consistent speech patterns found for time-compressed, noise-vocoded or accented speech promote perceptual learning (Mattys et al., 2012; Peelle & Wingfield, 2005). This hypothesis is intuitive as the deviation from the known is what requires adaptation in the first place. In that sense, it is conceivable that adaptation might be modulated by variability in talkers' acoustic-phonetic profiles. However, Dupoux & Green (1997) could show that changing talkers while listeners adapted to time-compressed speech only briefly disrupted the learning process, without reducing intelligibility to baseline levels. Nevertheless, it is possible that listeners adapt differently to time-compressed and noise-vocoded speech depending on specific talker characteristics. Specifically, adaptation might be modulated by acoustic-phonetic features determining baseline intelligibility as discussed in detail in the beginning of this introduction. For instance, Bradlow & Bent (2008) observed faster adaptation to accented speech by talkers with higher baseline intelligibility.

One recent study investigated changes in pupil dilation while listeners adapted to accented speech (Brown et al., 2020). Sentence materials were the same highly intelligible Chinese-accented speech as in McLaughlin & Van Engen (2020). While pupil dilation was largest when listeners processed accented speech, in line with McLaughlin & Van Engen (2020), pupil dilation also declined across trials (50 in total). This decline in task-evoked pupil dilation has been associated with fatigue (McGarrigle et al., 2017). However, Brown et al. (2020) also observed that the decrease in pupil dilation over time was largest for accented speech, which was interpreted as reflecting adaptation to accented speech. Pupil dilation is generally associated with listening effort in sentence recognition tasks; it is however questionable whether a reduction in the overall pupil dilation also corresponds to adaptation. Since intelligibility was at maximum in Brown et al. (2020), adaptation in the traditional sense (i.e., perceptual learning) was not indicated by behavioural measures. It is conceivable that the larger decrease in pupil dilation for accented speech (compared to native speech) merely reflected a more extreme fatigue-related decline, given the larger initial pupil dilation observed for accented speech. In a recent

study investigating changes in pupil dilation across trials for normal-hearing and hearing-impaired listeners, Wagner et al. (2019) observed that baseline pupil size indicated more sustained attention for hearing-impaired listeners. Baseline pupil size corresponds to the task-independent state of attention (Unsworth & Robison, 2016) and might reflect adaptation more adequately given that successful adaptation has been shown to require sustained attention (Huyck & Johnsrude, 2012).

To sum up, idiosyncratic talker differences have been investigated previously with respect to intelligibility in noise. Fewer studies have attempted to expand those findings to spectral and temporal degradations such as noise-vocoding and time-compression. Since talker differences have been shown to be preserved across different channel degradations, it is conceivable that specific features in talkers' acoustic-phonetic profiles, such as greater vowel spaces, contribute to their overall intelligibility benefit. While previous studies have linked accented speech to listening effort and adaptation (or both), there exists no comprehensive study that aimed to link idiosyncratic talker differences to listening effort and adaptation. Specifically, it is unclear how listening effort and adaptation are affected by an interaction of talker-specific acoustic features, and spectral and temporal degradations.

2.1.3 Aims of the current study

The current study was conducted to investigate two principal research questions (see Chapter 1, Section 1.6). First, I asked whether source degradations that affect intelligibility interact with channel degradations (RQ1). Second, I asked whether source degradations interact with channel degradations that affect listening effort, as measured by pupil dilation (RQ2). I conducted a listening experiment combined with pupillometry, measuring keyword recognition performance, as a measure of intelligibility, adaptation and task-evoked pupil response for noise-vocoded, time-compressed, and masked speech, as well as speech in quiet. Prior to the listening experiment, sentences were recorded

from sixteen Southern British English speakers. A range of acoustic-phonetic features were analysed and linked to intelligibility and effort of, and adaptation to, degraded speech.

I hypothesised that acoustic-phonetic features, such as vowel space dispersion that have been linked to intelligibility in previous studies (e.g., Bradlow et al., 1996) would also affect intelligibility in the current study (H1). Additionally, I hypothesised that talker intelligibility would persist across channel degradations (H2). To expand on previous findings (Bent et al., 2009), I hypothesized that talkers who were more intelligible under degradations that affect the spectral detail of speech (noise-vocoded and masked speech) would also be more intelligible under temporal degradations (time-compressed speech). This hypothesis was based on the assumption that spectral features such as precise articulation can be linked to temporal features, for instance slower speaking rates (Hazan & Markham, 2004). Similarly, I hypothesised that acoustic-phonetic features driving intelligibility benefits would be linked to adaptation (H3), as it has been shown previously that adaptation is linked to intelligibility (Bradlow & Bent, 2008).

With respect to listening effort, I aimed to replicate previous pupillometry studies showing larger pupil dilation for degraded speech (e.g., Wendt et al., 2018) (H4). An additional hypothesis made for this study was that differences in baseline pupil size changes would link to conditions with higher adaptation, i.e. noise-vocoding and time-compression (H5). For noise-vocoding, it has been shown that adaptation requires sustained attention (Huyck & Johnsrude, 2012) while sustained attention is reflected in a slower decline in baseline pupil size (Wagner et al., 2019). Finally, since pupil dilation has been shown to vary as a function of intelligibility, I hypothesised that acoustic-phonetic features driving intelligibility benefits would also be associated with larger pupil dilation (H6).

2.2 Methods

2.2.1 Speech materials

Sixteen speakers were recorded: eight older adults [four females; M_{age} = 71 (5.1) years; $range_{age}$: 61-77 years] and eight younger adults [four females; M_{age} = 26.8 (3.2) years; $range_{age}$: 22-33 years]. Speakers were sampled across different age groups and both sexes to include a wide range of speaker-related anatomical-physiological variation. All participants were native speakers of Southern British English. Each speaker read aloud 720 Harvard sentences (Institute of Electrical and Electronics Engineers, 1969), which are commonly used in speech perception experiments given their low semantic predictability and normed phonetic structure and length (e.g., Banks et al., 2015). During each recording session, breaks were permitted if needed. Recordings were made in an anechoic chamber using a Bruel & Kjaer 2231 Sound Level Meter fitted with a type 4165 condenser microphone. The signal was digitized with a Focusrite 2i2 USB audio interface at a sampling rate of 44100 Hz and a bit-depth of 16 bits. Sentences were displayed on a screen facing the participant and the experimenter controlled the timing of sentence presentation. ProRec (Huckvale, 2014) was used for sentence recording and segmentation, including removal of silent parts. Recordings were manually checked and any remaining silent parts at the beginning and end of each sentence were cut at zero crossings using Praat (Boersma & Weenink, 2018). Of all 720 sentences, those with unexpected noise or mis-pronunciations for any of the speakers were removed from the final sentence set of all speakers; 237 sentences were removed in total. Of the remaining sentences, 192 were randomly selected for the experiment. The same subset of sentences was selected from each speaker and only this subset was analysed acoustically. Sentences were converted to mono, down-sampled to 22050 Hz and high-pass filtered at 50 Hz, removing gross fluctuations. Sentences were root-mean-square (rms) normalized. Sentences were automatically annotated and aligned using the Montreal forced aligner (McAuliffe et al., 2017). The aligner is trained on raw speech and word-level transcription of each sentence

and outputs aligned text grids with word- and phone-level annotation. These text grids were checked to ensure that no processing errors occurred.

2.2.2 Acoustic analyses

Acoustic analyses were conducted using custom-made scripts written in Python that access Praat through the Parselmouth interface (Jadoul et al., 2018). All acoustic analyses were based on rms-normalised signals. Single values of each measure for each talker were obtained by calculating the mean across all 192 sentences.

Mean energy in 1-3 kHz range (ME13): Mean energy in mid-range frequencies (1-3 kHz) has reliably shown a relationship with intelligibility (Green et al., 2007; Hazan & Markham, 2004). To calculate ME13, a band-pass filter (Hann, 1-3 kHz) was first applied to each sentence. The intensity contour was then extracted and the mean calculated across the entire sentence.

Fundamental frequency (f_0): Bradlow et al. (1996) found mean fundamental frequency (f_0) to be correlated with intelligibility, which was driven by increased mean f_0 and higher intelligibility for female talkers. They also found a tendency for a correlation between wider f_0 range and higher intelligibility. I included f_0 median (in Hz) and f_0 standard deviation (in semitones) as acoustic features. Semitones were used in order to compare across a range of talkers with different fundamental frequencies (Hazan & Markham, 2004). Periodicity detection was performed by applying the auto-correlation method implemented in Praat (Boersma, 1993), using a 10 ms frame duration. Upper and lower boundaries were set to $q65 * 1.92$ and $q15 * 0.83$ with q representing the respective quantiles. The formulas were optimized to reduce artefacts such as octave jumps (De Looze & Hirst, 2008).

Speaking rate (SR): Even though speaking rate is not consistently linked to intelligibility and effort, I hypothesized that slow speech would be more beneficial for intelligibility than fast speech when speech is time-compressed. I estimated

speaking rate by dividing the canonical number of syllables in a sentence by the duration of the sentence. The number of syllables was obtained for each sentence transcription using the package *quanteda* in R (Benoit, 2018). Syllables per second were then defined as a measure of speaking rate. Speaking rate was strongly negatively correlated with vowel duration [$r = -0.92, p < 0.001$].

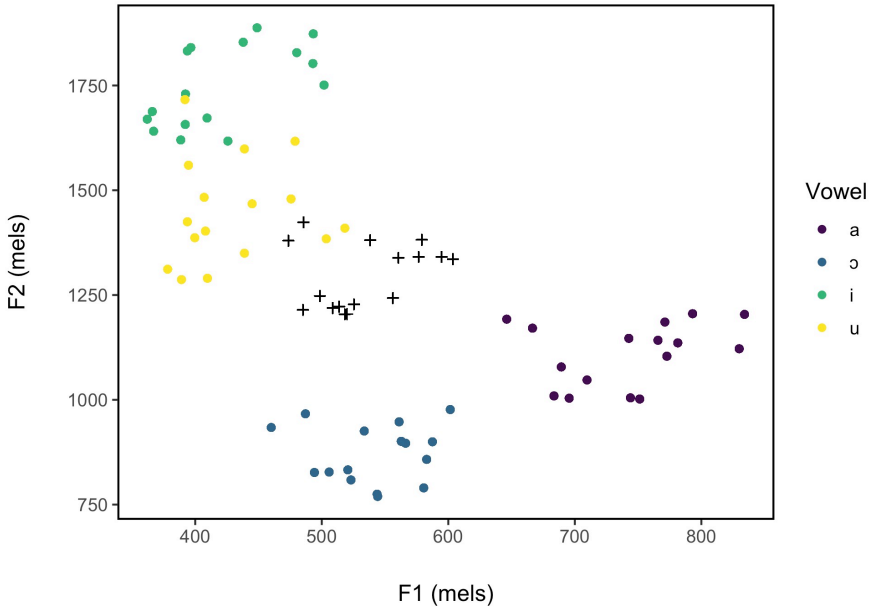

Figure 5: *Vowel category centers (points) and vowel space centers (crosses) for all talkers. Vowels are represented by their first and second formants, F1 and F2, respectively.*

Vowel space dispersion (VSD): More disperse peripheral vowels in the $F_1 - F_2$ space relate to higher intelligibility (Bradlow et al., 1996). Estimates of a talker's vowel space can be obtained by measuring the dispersion of vowel productions from the vowel space center; dispersion is also linked to the overall vowel space perimeter (Bradlow et al., 1996). Vowel space measurements are usually based on the three point vowels (/aiu/) (Cooke et al., 2014). Due to allophonic variation in American English, Bradlow et al. (1996) replaced measurements of /u/ by measurements of /o/. In the current study, category averages of /u/ indicated fronting which is a common phonetic phenomenon, especially for Southern British English (Strycharczuk & Scobbie, 2017). Category averages of /u/ did therefore not represent corners of the respective vowel spaces (see Figure 5). I therefore decided to use measurements of /ɔ/ instead, to better approximate a triangular shape. Vowel space dispersion was then calculated

as the Euclidean distance of /ɑiɔ/ from the centroid of the vowel space. Only vowels from content words were analysed (ɑ = 41, i = 89, ɔ = 62). Formants were measured at the vowel center, applying short-term spectral analysis with a 25 ms window size. The formant maximum was adjusted for speaker gender (male = 5000 Hz; female = 5500 Hz). For each speaker and vowel category, outliers were removed (2 standard deviations above or below the mean) given the possibility of measurement errors with the automised procedures (Hazan et al., 2018; Hazan & Baker, 2010). The average number of vowel productions per participant and condition was not substantially different after outlier exclusion (average: a = 38, i = 83.2, ɔ = 57.5). Formants were converted to the mel scale which is a widely accepted perceptually-motivated transformation (Bradlow et al., 1996; Cooke & Lu, 2010; Fant, 1973; Wieland et al., 2015). The formula is shown in equation (1) with M and F representing frequencies in mels and Hz, respectively.

$$M = (1000/log_{10}2) * log_{10}((F/1000) + 1) \tag{1}$$

Corners of the vowel space were defined as the vowel category means of $F1$ and $F2$, and the centroid was obtained as the arithmetic mean of the corners (Wieland et al., 2015). The Euclidean distance of each vowel realization (F_1, F_2) from the vowel space center $(\overline{F_1}, \overline{F_2})$ was calculated using the Formula in equation (2).

$$d(F_1, F_2) = \sqrt{(F_1 - \overline{F_1})^2 + (F_2 - \overline{F_2})^2} \tag{2}$$

The average distance was first obtained for each vowel category and then across vowels, resulting in a single measure of vowel space dispersion.

Descriptive statistics for all acoustic-phonetic measures are shown in Table 1. Differences between talker groups (female vs. male, older vs. younger) were not of primary interest in the current study, but plots are provided in the Appendix.

2.2.3 Participants

Sixty-four normal-hearing native speakers of British English were recruited for the experiment [40 females; M_{age} = 22.3 (4.3) years; $range_{age}$: 18-37 years].

Table 1: *Descriptive statistics for acoustic-phonetic features. ME13 = mean energy (dB), F0M = f0 Median (Hz), F0SD = f0 standard deviation in semitones (st) and Hz, SR = speaking rate (syllables/s), VSD = vowel space dispersion in mels and Hz.*

Feature	Mean	SD	Min	Max
ME13	59.24	2.77	54.34	63.96
F0M	156.32	38.59	101.79	210.18
F0SD(st)	2.81	0.55	1.76	3.69
F0SD(Hz)	25.91	6.63	17.39	36.61
SR	3.80	0.23	3.41	4.18
VSD(mels)	377.25	28.62	320.73	428.30
VSD(Hz)	635.91	73.64	504.56	773.34

They were either reimbursed for their participation following the guidelines of the Division of Psychology and Language Sciences at the University College London or given course credit. Hearing ability was established by a standardised audiometric test at the beginning of the testing session. Participants had hearing thresholds equal or better than 25 dB HL at all tested octave frequencies between 0.25 and 4 kHz. This threshold is in line with similar studies (e.g., Wendt et al., 2018; Wagner et al., 2019). Two participants were excluded because their hearing exceeded those thresholds (30 dB HL at 0.5 and 2 kHz respectively). Dependent measures from one listener in the noise-vocoding condition were removed entirely since almost no keywords were recognized correctly [$M = 2.5\%$]. The remaining conditions of this participant were then excluded, as well, since the reason for the poor performance under noise-vocoding could not be explained.

2.2.4 Listening conditions

From the 192 sentences, four lists of 48 items each were created. Even though pupil dilation effects during listening can be detected with as few as 20-25 items (Winn et al., 2018), more trials are necessary to sufficiently estimate adaptation to noise-vocoded speech (e.g., Erb et al., 2012). The lists were optimised so

that the mean duration was roughly matched across lists [$M_{duration}$ = 2.246 s (0.016)]. It was ensured that the same keyword did not appear more than twice within the same list.

Sentences were presented in quiet and in three degradations: time-compression, noise-vocoding and masking (noise). Noise-vocoding and masking have been used in a previous study that found talker differences to be preserved across these conditions (Bent et al., 2009). A similar effect was expected for time-compression. In addition, in accordance with Peelle & Wingfield (2005), time-compression and noise-vocoding were expected to show robust adaptation effects, in contrast to masking. While masking noise is random and only obscures speech (Peelle & Wingfield, 2005), modifying speech by time-compression and noise-vocoding introduces systematic changes that can be adapted to.

Parameters were chosen based on experimental test runs with lab members indicating that intelligibility was not too low overall, avoiding disengagement effects on the pupil measures. At the same time, it was ensured to leave enough room for possible adaptation effects. A secondary aim was to achieve equal intelligibility across conditions; however, the large number of talkers and limited amount of test runs ultimately resulted in intelligibility differences, as demonstrated in the results section. Sentences were time-compressed to 37% of their original duration by applying the pitch-synchronous overlap-add implementation in Praat. Pilot data showed that this rate was sufficient to elicit adaptation effects without a significant drop in intelligibility that can be observed when increasing compression rate further (e.g., Versfeld & Dreschler, 2002). For noise-vocoding, the original signal was divided into six frequency bands spaced according to the cochlear frequency-position function (Greenwood, 1990). Amplitude envelopes were extracted from each band by applying a 4th-order Butterworth low-pass filter with a cutoff frequency at 256 Hz and half-wave rectification. The envelopes were then used to modulate white noise. For masking, speech-shaped noise was created by obtaining the long-term average spectrum (LTAS) of a separate set of sentences from a non-experimental female talker. Noise was then generated with the same LTAS and added to the sentence at a

signal-to-noise ratio (SNR) of -1 dB.

2.2.5 Design & Procedure

Each listener was presented all conditions in four blocks of 48 sentences. Blocks were counterbalanced across listeners using a Latin square design. All 48 sentences in one block were spoken by the same talker, as it has been shown that changing talkers interferes with adaptation (e.g., Dupoux & Green, 1997). Talkers were counterbalanced across listeners and blocks so that each talker was heard by 16 listeners in total and by four listeners per condition. Lists and sentences within each list were randomized. The large number of sentences required and the talker change constraint imposed by the adaptation measure limited the number of talkers that could be presented within one testing session. In addition, the acquisition of pupillometry data requires monitoring by the experimenter, making larger-scale studies such as conducted by Bent et al. (2009) unfeasible.

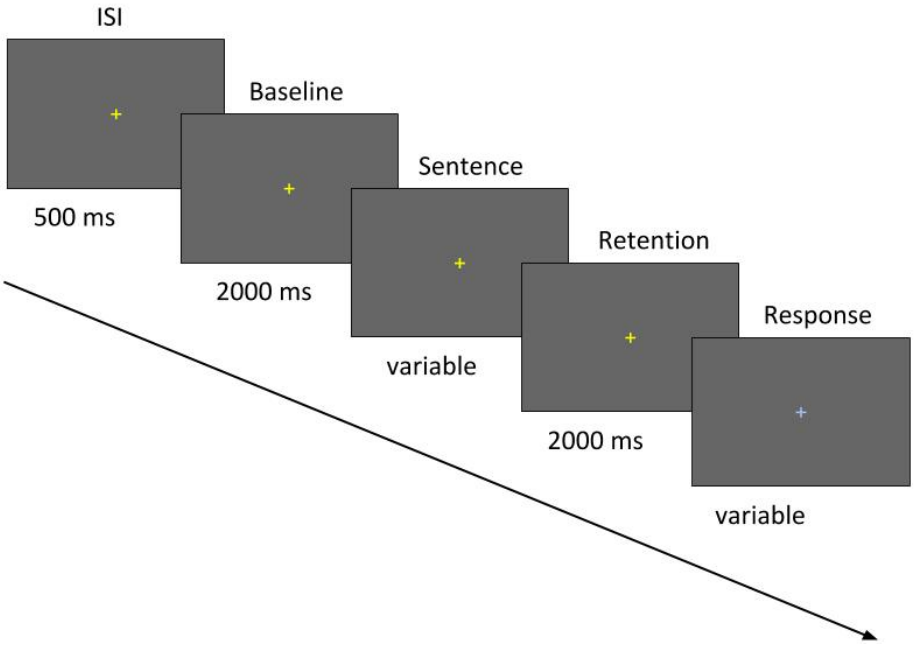

Figure 6: *Trial events with duration. Rectangles represent displays with central fixation cross.*

Participants wore headphones (Sennheiser HD 25 SP II) throughout the experiment with output levels at 70 dB SPL. They were asked to put their head comfortably on a table-mounted chin rest, to minimize head movements. Glasses had to be removed for the duration of the experiment. Pupil recordings were obtained using an EyeLink 1000 table-mounted eye tracker (SR Research; Oakville, Canada) at a distance of 55 cm from the participant's head. A sampling rate of 500 Hz was used. The light level was kept constant at 130 lux, but for participants with very large or very small resting state pupil sizes, the light level was adjusted as required (cf. Wendt et al., 2018). The experiment started with eight practice trials in which sentences in quiet were presented to the participants. These sentences were not included in the main experiment. They were taken from a non-experimental talker whose recordings were not included in the corpus. Each trial followed the same procedure (see Figure 6): after an inter-stimulus interval of 500 ms, the baseline pupil size was recorded for 2000 ms, which was followed by the sentence onset. After the offset of the sentence, the pupil size was tracked for another 2000 ms in quiet since the dilation usually peaks around 0.7 to 1.2 s after stimulus offset (Winn et al., 2018). The fixation cross changed colour to signal the end of the retention period and the start of the response. Participants repeated back words to the experimenter who logged correctly identified keywords (i.e., content words) on a separate control screen. Each sentence contained five keywords. A keyword was considered correctly identified despite incorrect plural (-s) or tense (-ed) endings (cf. Banks et al., 2015). For instance, given the sentence *The pigs were fed chopped corn and rubbish*, the response *pig* was counted as correct while the response *peg* was counted as wrong. Participants were asked to blink as little as possible during the trial up to the point that a response had to be given. The experiment was implemented in MATLAB (R2016a).

2.2.6 Dependent variables

Recognition: To measure word recognition performance, the proportion of keywords correctly identified out of five was calculated and averaged across

trials. This means that for each listener, recognition averages were based on 48 sentences (i.e., 240 keywords). To obtain single measures of adaptation, Erb et al. (2012) obtained linear slopes from word recognition data across all trials (N = 100). However, specifically for time-compression, adaptation is usually fast and levels off after as little as 15 sentences (Dupoux & Green, 1997, for a similar time-compression rate at 38%). In the current study therefore, trials were divided into four blocks of 12 trials each (cf. Kennedy-Higgins et al., 2020). Adaptation was defined as a significant increase in recognition performance from one block of trials to the next. Slopes were then obtained from linear model fits for each individual's performance data within the range of trials (blocks) showing adaptation.

Pupillometry: to preprocess pupil data, the guidelines by Winn et al. (2018) and functions provided by Geller et al. in the GazeR package (https://github.com/dmirman/gazer) were applied. Only data collected between the onset of the inter-stimulus interval and the verbal response prompt were included due to the possibility that articulatory movements interfered with the measure (Richer & Beatty, 1985). Trials that contained more than 20% missing data within the specified interval were excluded. In order to obtain representative pupil trace averages, blocks with fewer than 24 remaining trials (50%) were removed. Two blocks of one participant were therefore removed. The remaining data set included an average of 47.16 trials for each listener and condition. Blinks were marked as missing values by the Eyelink and were interpolated linearly. Before interpolation, gaps of missing values were extended to 100 ms before and 100 ms after the gap. Data was then smoothed using a 5-point moving average filter. Additionally, rapid pupil size disturbances were detected and removed using the median absolute deviation. Subtractive baseline correction was applied using the median of the baseline recorded 1000 ms before the onset of the sentence. Several baseline correction methods have been applied in previous research, the two common ones being subtractive and divisive baseline correction (Winn et al., 2018). However, it has rightfully been argued that it is paradoxical that both methods coexist, as divisive baseline correction assumes a non-linear relationship between

baseline and task-evoked pupil dilation, while subtractive baseline correction assumes a linear relationship (Reilly et al., 2019). Measuring pupil dilation in a pure-tone detection task, Reilly et al. (2019) observed that different baseline pupil sizes, induced by changing light settings, did not influence the magnitude of the task-evoked pupil dilation. Since the authors employed a pure-tone detection task, results might not necessarily generalise to sentence recognition. However, non-linear pupil size scaling might be problematic as it introduces biases when baseline pupil size is not controlled for between participants. Furthermore, most sentence recognition studies to date applied subtractive baseline correction. The same correction technique was therefore also applied in this thesis.

Pupil traces were then time-aligned with the end of the sentence (offset) and down-sampled to 20 Hz. Aligning pupil traces to either sentence onset or offset does usually not change the general shape of the data (Winn et al., 2018). Pupil traces were averaged for each listener and condition. All average pupil traces were inspected for anomalies in overall shape and magnitude. One participant was excluded since average pupil traces in each condition showed decreasing pupil size (see Winn et al., 2018). It is possible that the pupil size for this participant was only affected by the motor response while slowly returning to baseline during the trial. Peak dilation and latency were obtained from average pupil traces. For standardised speech perception tasks combined with pupillometry, as described in Chapter 1 of this thesis, these traditional measures usually suffice; reporting peak dilation only is common for pupillometry studies (e.g., Wendt et al., 2017, 2018; Müller et al., 2019). Other studies have reported mean dilation, as well, which models information about both pupil dilation and constriction following the peak (Verney et al., 2001; Zekveld et al., 2010). However, similar to other analysis techniques aiming to model the shape of the pupil dilation curve, no consensus has been achieved about interpreting components other than peaks (Wendt et al., 2018). Larger peak dilation is usually interpreted as increased processing (or memory) load, reflecting higher listening effort (e.g., Zekveld et al., 2010; Borghini & Hazan, 2018, 2020; Koelewijn et al., 2012; Wendt et al., 2017, 2018).

The peak dilation is the maximum value of each average trace within a specified time window. Since sentence duration varied largely between time-compressed speech and all other conditions, latency of dilation maximums for each participant and condition were first inspected. The search space ranged from -1418 to +1500 ms (quiet, masking and noise-vocoding) and -525 to +1500 ms (time-compression) with respect to sentence offset. The lower boundary was the respective duration of the shortest sentence and the upper boundary was 500 ms before the onset of the response. A "buffer" of 500 ms was subtracted from the total retention time (2000 ms) to account for possible pre-motor effects on the pupil dilation (Richer & Beatty, 1985). The pupil dilation in similar paradigms typically occurs within 500-1000 ms after sentence offset (Winn et al., 2018), but given the use of shorter time-compressed sentences in the current study, the analysis window was extended. Indeed, the majority of peaks in the current study were located within the retention period (see Results section). This finding is in line with observations in other pupillometry studies (Winn et al., 2018). Peak dilation and latency were extracted with respect to the specified time window.

Following Wagner et al. (2019), changes in baseline pupil size across trials were analysed, hypothesising that sustained attention as indexed by a slower decline in baseline pupil size would relate to conditions requiring adaptation. Note however that Wagner et al. (2019) compared baseline pupil size changes across the entire experiment, with a pre-experiment baseline as reference. In the current study, statistical analyses were conducted on raw pupil size data.

2.2.7 Statistical analysis

To analyse differences between listening conditions, linear mixed models (LMMs) were fitted using *lme4* in R (Bates et al., 2015). In all models, random intercepts were allowed for listeners, given the repeated measures design. F-tests were performed on all models with Satterthwaite degrees of freedom approximations, implemented in *lmerTest* (Kuznetsova et al., 2017). To obtain

p-values from LMMs, two valid approaches have been suggested, model comparison and degrees of freedom approximations (Luke, 2017). Satterthwaite approximations have been shown to be fairly conservative with acceptable Type 1 error rates, independent of sample size, when applied to restricted maximum likelihood models (REML). Pairwise comparisons were done using the function *estimate_contrasts* from the library *modelbased* in R. For pairwise comparisons, Bonferroni adjustments were made. Results of this analysis were expected to show a main effect of condition, with lower intelligibility for degraded speech, but larger and delayed peak pupil dilation (H4, see Chapter 1, Section 1.6). Furthermore, a linear mixed model with trial as additional fixed effect, was expected to show a slower decline of baseline pupil size across trials for conditions with adaptation (H5).

Talker intelligibility across conditions was investigated by means of Pearson's product moment correlations for each condition pair. Recognition performance was therefore averaged across all four listeners for each talker and condition. Each talker average was based on maximally four listeners and 192 sentences in total. Results were expected to show that talker intelligibility under one channel degradation would correlate with talker intelligibility under all other channel degradations (H2). The influence of acoustic-phonetic features on intelligibility, adaptation and pupil size measures was investigated by means of multiple linear regression. Results were expected to show that the same source characteristics that predicted intelligibility in previous studies (e.g., vowel space dispersion) would also predict intelligibility in the current study (H1). The same source characteristics that would predict intelligibility in the current study were expected to predict adaptation rates, and peak pupil dilation and latency (H3 and H6).

2.3 Results: channel degradations

2.3.1 Intelligibility and adaptation

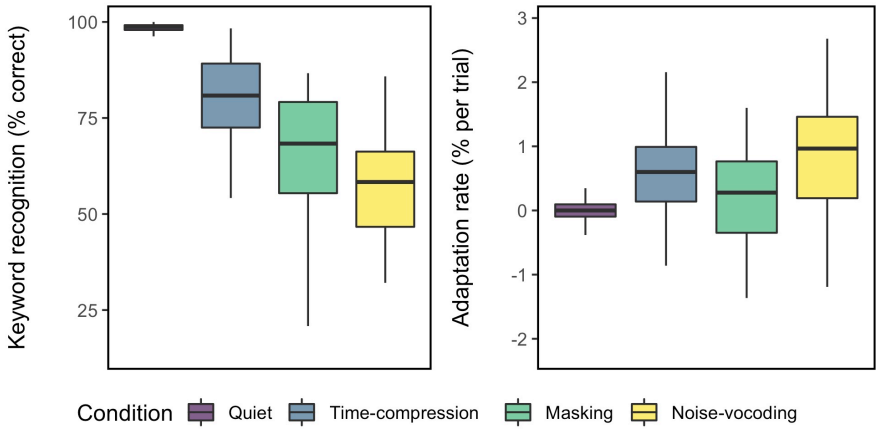

Figure 7: *Distributions of the average and rate of keywords recognized correctly in all conditions. Adaptation rate was the slope of the linear fit to an individual's performance, based on the first half of trials within each condition block (trials 1-24). Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range.*

First, the effect of degradation type on keyword recognition was investigated. There was a main effect of condition [$F(3, 180) = 154, p < 0.001$]. Pairwise comparisons showed that recognition was poorer for all types of degradations compared to quiet [$p < 0.001$]. Recognition for noise-vocoding was poorer than for masking and time-compression [$p < 0.001$] while recognition for masking was poorer than for time-compression [$p < 0.001$] (Figure 7). These differences should not be understood as an effect of degradation type, but rather as reflecting the degree of degradation chosen a priori for each condition. In the second part of this results section, I will show how talker differences can explain the variances observed in each condition.

Linear mixed effects model analysis revealed a significant interaction between block and condition [$F(9, 900) = 3.20, p < 0.001$]. Pairwise comparisons showed that for time-compressed ($p = 0.01$) and noise-vocoded speech ($p < 0.001$), recognition improved from block 1 to block 2. This result indicated that listeners adapted to noise-vocoded and time-compressed speech,

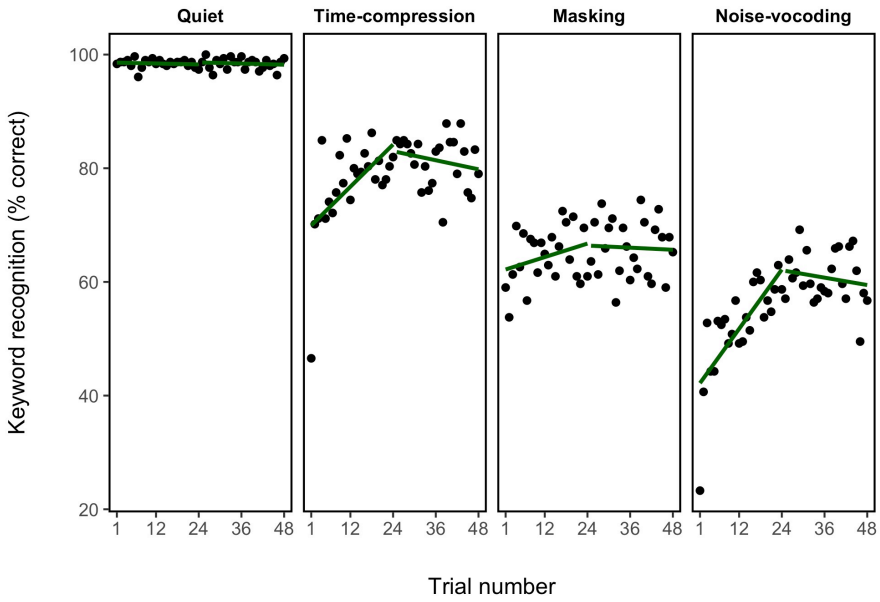

Figure 8: *Percent of keywords recognized correctly as a function of trial number, averaged across participants. Lines indicate linear fits to the averaged data in the first (1-24) and the second half (25-48) of trials. Individual adaptation slopes were obtained from the first half of trials.*

but not to masked speech. Linear models were created for each listener and condition, only including trials within the first two blocks (N = 24; see Figure 8). The slopes were then used as estimates of each individual's adaptation rate.

2.3.2 Pupillometry

It was investigated whether pupil dilation measures followed the same trend as recognition scores, reflecting increased effort for degraded speech (Figure 9). There was a main effect of condition for peak dilation [$F(3, 174.62) = 35.58, p < 0.001$]. Pairwise comparisons showed that peak dilation was larger for all three degradations compared to quiet ($p < 0.001$), but there was no difference between degradations ($p > 0.05$). Figure 9 also shows distributions of peak latency for all individuals in all conditions. It is apparent that while the majority of peaks occurred within the retention period (i.e., peak latency $>$ 0 ms), individual differences can be observed, showing that for some listeners, pupil dilation peaked before the end of the sentence (specific-

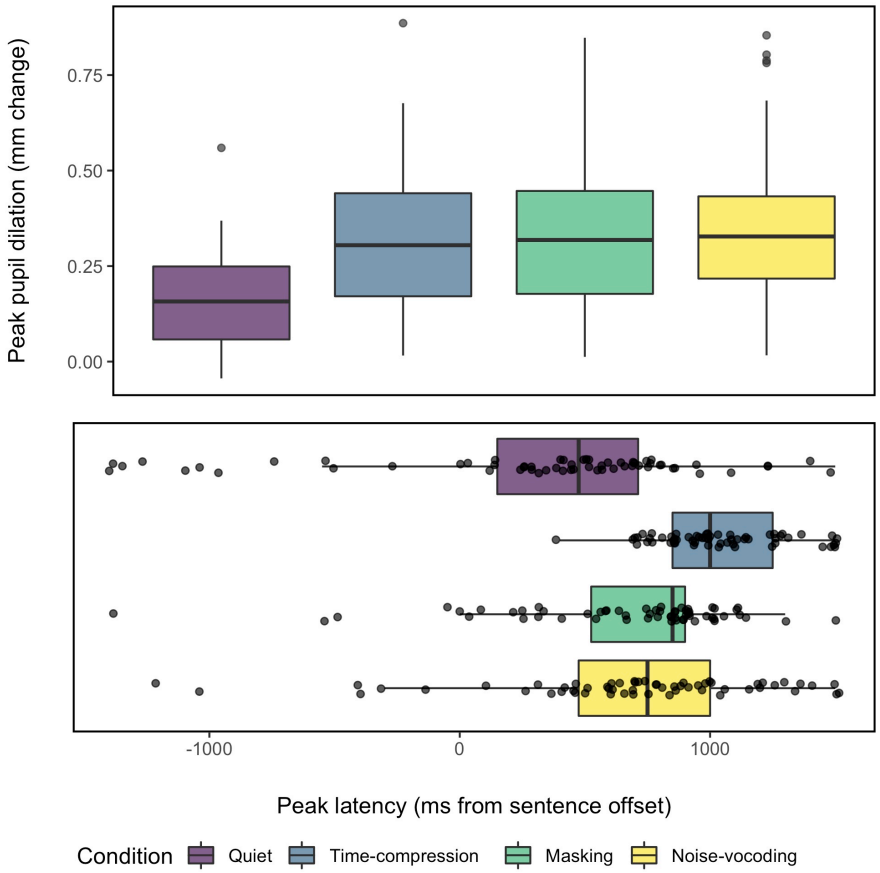

Figure 9: *Distributions of peak pupil dilation (top) and latency (bottom) in all conditions. Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range.*

ally in quiet). Peak latency also showed a significant main effect of condition [$F(3, 175.54) = 27.26, p < .001$]. Pairwise comparisons indicated larger peak latency for masking, noise-vocoding and time-compression compared to quiet ($p < 0.001$). These results indicate that the task-evoked pupil response peaked later when speech was degraded, likely reflecting increased effort as indicated by previous studies (e.g., Zekveld et al., 2010). Pairwise comparisons also indicated larger peak latency for time-compression compared to masking and noise-vocoding ($p < 0.001$). It has to be noted that pupil traces were aligned to sentence offset so that these results indicate a delayed peak response for time-compressed speech measured from the end of the sentence. When measuring peak responses from sentence onset, shorter time-compressed sentences elicit faster peaks, given the shorter sentence duration (cf. Müller et al., 2019).

Overall, the results presented above indicate that larger pupil dilation and delayed peaks were associated with conditions yielding lower recognition performance. This finding likely indicates higher listening effort, as expected based on previous studies (Wendt et al., 2018; Zekveld et al., 2010) and models of arousal (Kahneman, 1973). At the same time, pupil dilation measures were less sensitive to differences between conditions, possibly driven by larger variability between listeners (Winn et al., 2018).

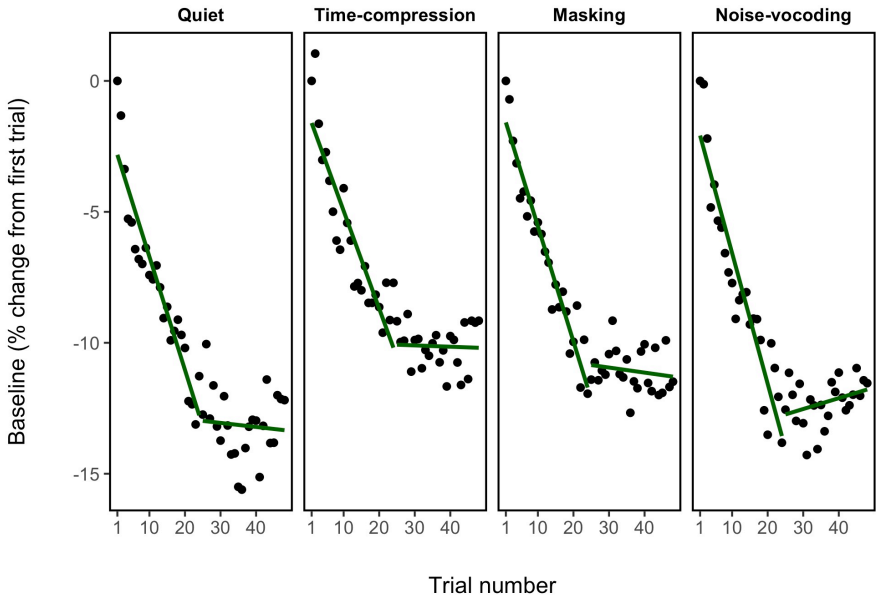

Figure 10: *Change in baseline pupil size across trials. Lines indicate linear fits to the averaged data in the first (1-24) and the second half (25-48) of trials. For visualisation purposes, baseline pupil size is displayed as percent change from first trial. Statistical analyses were conducted on raw pupil size data.*

In accordance with adaptation data, baseline pupil size was analysed in two time windows, (1) trial 1-24, and (2) trial 25-48 (see Figure 10). The reasoning was that overall arousal, as indexed by the baseline pupil size, would be more sustained in conditions with adaptation, i.e., noise-vocoding and time-compression. In time window 1, a main effect of trial [$F(1, 5566) = 803.32, p < 0.001$] and condition emerged [$F(3, 5566.1) = 5.51, p < 0.001$]. While baseline pupil size generally declined over the course of the first 24 trials, pairwise comparisons indicated that baseline pupil size was larger under masking and time-compression, compared to noise-vocoding and quiet ($p < 0.001$). In time

window 2, there was no effect of trial ($p = 0.77$). However, a main effect of condition [$F(3, 5544) = 3.32, p = 0.019$] indicated larger baseline pupil size under masking and time-compression, compared to noise-vocoding and quiet ($p < 0.001$). In addition, pairwise comparisons indicated larger baseline pupil size under noise-vocoding compared to quiet ($p < 0.001$).

It has to be noted that there was no main effect of condition on the baseline pupil size in the first trial [$p = 0.46$]. It has been suggested that a faster decline in baseline pupil size might be driven by larger initial arousal levels (Ayasse & Wingfield, 2020); this assumption did not apply in this study.

2.3.3 Interim summary

Intelligibility differed between listening conditions and was optimal for speech in quiet. It was observed that listeners adapted to noise-vocoded and time-compressed speech only. Pupil dilation results showed the inverse pattern to intelligibility data, suggesting more effortful processing for less intelligible speech. Degraded speech generally elicited a peak pupil dilation that occurred later compared to speech in quiet. Overall baseline pupil size was smaller for noise-vocoded speech and speech in quiet, reflecting lower arousal. In the next section, the dependent measures presented in this section were linked to acoustic-phonetic talker differences.

2.4 Results: source degradations

2.4.1 Intelligibility across degradations

To investigate talker intelligibility across listening conditions, correlation analyses were conducted for each pair of conditions. Talker intelligibility was determined as the mean recognition score across listeners for each talker and condition. There was a significant correlation between talker intelligibility under masking and noise-vocoding [$r_{12} = 0.52, p = 0.037$] and noise-vocoding

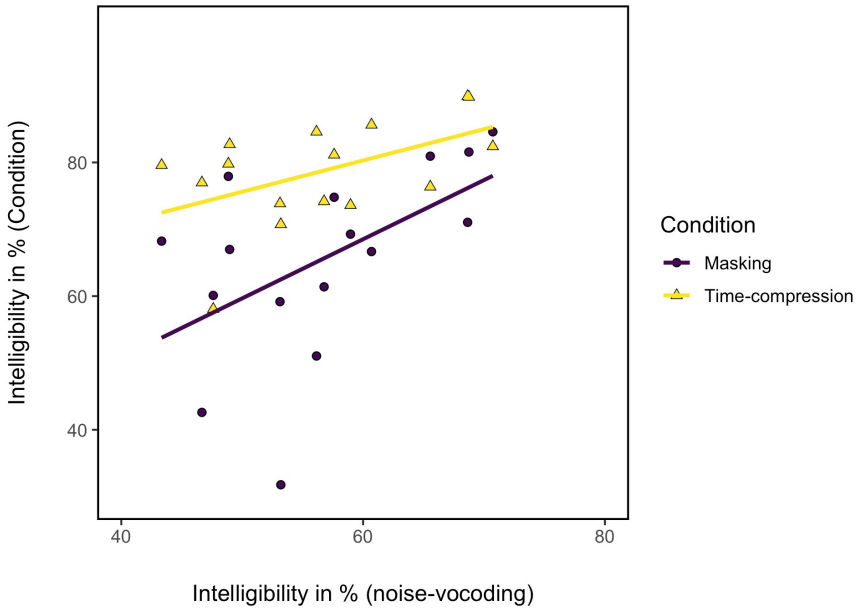

Figure 11: *Intelligibility, i.e., average percent of keywords recognized correctly aggregated by talker and experimental condition. Each symbol therefore represents one talker average. Noise-vocoding is plotted against masking and time-compression.*

and time-compression [$r_{23} = 0.51, p = 0.046$] (see Figure 11). These results indicate that talkers who were more intelligible under noise-vocoding were also more intelligible under masking and time-compression. For the masking/time-compression pair, a correlation was found at $r_{13} = 0.41$ that did not reach significance [$p = 0.12$]. However, William's test indicated that correlation r_{12} was not significantly stronger than correlation r_{13}, given correlation r_{23} [$t = 0.47, p < 0.64$]. In the following section, it was investigated which acoustic-phonetic profiles were linked to higher intelligibility across and within degradations.

2.4.2 Intelligibility and adaptation

For each condition and dependent measure, linear regression analyses were conducted to investigate the relevance of each acoustic-phonetic feature. Multiple linear regression was applied including all acoustic-phonetic features as predictors and dependent measures of each individual listener. R^2 is reported unadjusted and adjusted for the number of predictors. The variance inflation

factor was below 2 for each feature, ruling out collinearity (Menard, 2002). Given the small number of listeners assigned to each talker, regression analyses were conducted on the entire data set, i.e., without averaging listener data for each talker, as this approach eliminates potentially meaningful variability observed across listeners. The explained variances are therefore generally smaller than those observed for comparable studies that conducted regression analyses on averaged data (e.g., Green et al., 2007).

Table 2: *Linear regression results for intelligibility. ME13 = mean energy, F0M = f0 Median, F0SD = f0 standard deviation, SR = speaking rate, VSD = vowel space dispersion. All continuous predictors are mean-centered and scaled to have SD = 1. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.*

	Masking		Time-compression		Noise-vocoding	
(Intercept)	65.2 ***	(1.6)	79.2 ***	(1.6)	56.4 ***	(1.6)
ME13	11.3 ***	(1.8)	1.1	(1.8)	2.8	(1.9)
F0M	0.1	(1.8)	-0.7	(1.9)	-0.3	(1.9)
F0SD	2.2	(2.0)	-1.5	(2.0)	2.1	(2.1)
SR	0.9	(2.0)	-4.9 *	(2.0)	-0.7	(2.1)
VSD	6.7 **	(2.0)	1.4	(2.0)	6.6 **	(2.1)
R sq	0.44		0.14		0.23	
R sq (adj)	0.39		0.06		0.16	

Regression results for intelligibility are shown for each condition in Table 2. The model for masking showed a significant contribution of mean energy and vowel space dispersion [$R^2_{adj} = 0.39$]. The model for noise-vocoding contained vowel space dispersion as significant predictor [$R^2_{adj} = 0.16$]. On the other hand, the model for time-compression showed a significant contribution of speaking rate [$R^2_{adj} = 0.06$], but none for the other predictors. It has to be noted that for time-compressed speech, little of the observed variance was explained by the chosen parameters (see Discussion). In summary, the regression results indicate that preserved talker differences in masking and noise-vocoding might

be driven by talkers with larger vowel space dispersion. However, there was no common acoustic-phonetic feature for noise-vocoding and time-compression. A correlation between vowel space dispersion and speaking rate did not reach significance [$r = -0.32, p = 0.22$].

The effect of talker acoustics on adaptation was investigated by means of regression analyses, predicting adaptation rates with acoustic-phonetic measures across sentences for each talker. As outlined in the introduction to this chapter, it was hypothesised that listeners adapt more to talkers with acoustic-phonetic profiles that also relate to higher baseline intelligibility (Bradlow & Bent, 2008). However, none of the predictors reached significance ($p > .05$).

2.4.3 Pupillometry

Table 3: *Linear regression results for peak pupil dilation. ME13 = mean energy, SR = speaking rate, VSD = vowel space dispersion. All continuous predictors are mean-centered and scaled to have SD = 1. *** p < 0.001; ** p < 0.01; * p < 0.05.*

	Quiet		Masking		Time-c.		Noise-v.	
(Intercept)	0.17 ***	(0.02)	0.33 ***	(0.02)	0.32 ***	(0.02)	0.35 ***	(0.03)
ME13	0.00	(0.02)	0.00	(0.03)	-0.02	(0.03)	0.03	(0.03)
SR	-0.02	(0.02)	0.01	(0.03)	0.04	(0.02)	-0.01	(0.03)
VSD	-0.04 *	(0.02)	-0.00	(0.03)	-0.03	(0.03)	0.02	(0.03)
R sq	0.12		0.01		0.09		0.03	
R sq (adj)	0.08		-0.05		0.04		-0.03	

As outlined in the introduction to this chapter, it was hypothesised that acoustic-phonetic features promoting intelligibility would also be linked to pupil dilation. Multiple linear regression analyses were conducted to establish the relationship between acoustic-phonetic features, and peak pupil dilation and latency. Specifically, features that showed to be relevant for intelligibility, i.e., mean energy, vowel space dispersion and speaking rate were investigated.

Table 4: *Linear regression results for peak latency. ME13 = mean energy, VSD = vowel space dispersion, SR = speaking rate. All continuous predictors are mean-centered and scaled to have SD = 1. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.*

	Quiet		Masking		Time-c.		Noise-v.	
(Intercept)	308 ***	(87)	664 ***	(64)	1038 ***	(33)	686 ***	(72)
ME13	-39	(101)	25	(72)	7	(37)	-22	(84)
SR	-59	(94)	22	(69)	52	(35)	129	(79)
VSD	-257 *	(106)	-30	(76)	0	(39)	65	(90)
R sq	0.11		0.01		0.04		0.05	
R sq (adj)	0.06		-0.04		-0.01		0.00	

Regression results are shown for each condition in Table 3 and 4.

In quiet, peak pupil dilation [$R^2_{adj} = 0.08$] and latency [$R^2_{adj} = 0.06$] were predicted by vowel space dispersion, with larger peaks and latencies associated with smaller vowel space dispersion. However, it has to be stressed that for both measures, explained variances were low, possibly reflecting individual differences in pupil dilation and latency (see Discussion).

2.5 Discussion

The aim of the current study was to establish the combined effect of channel and source degradations on intelligibility and listening effort. In particular, the focus was on the interaction of talker-specific acoustic-phonetic features with masking, spectral and temporal degradations. As listeners can adapt to both channel and source degradations, it was also hypothesized that adaptation to spectral and temporal degradations would be modulated by talker-specific acoustic features. A listening experiment was conducted combined with pupillometry that required participants to listen to and repeat back sentences in degraded and quiet conditions. Sentences were taken from a corpus including sixteen speakers recorded especially for this study.

2.5.1 Channel degradations

Average intelligibility differed across listening conditions, reflecting the parameters chosen for each degradation type. Intelligibility was on average low for noise-vocoded speech (56%) and masked speech (66%), and high for time-compressed speech (79%) and speech in quiet (98%). In accordance with my predictions, listeners adapted to noise-vocoded and time-compressed speech, indicated by intelligibility improvements from the first to the second block of 12 trials. There was no improvement for masked speech and it is likely that adaptation was degradation-specific and not due to task familiarity (Peelle & Wingfield, 2005).

Peak pupil dilation was larger for degraded speech than for speech in quiet, which was attributable to lower intelligibility and likely higher effort (Wendt et al., 2018). There was no significant difference in pupil dilation between degradations despite large differences in intelligibility. As suggested previously, there is no linear relationship between pupil dilation and recognition scores; generally, pupil dilation reaches a plateau at moderate intelligibility levels (e.g., ~80% at -4 dB SNR for stationary maskers, based on Figure 5 in Wendt et al., 2018). Differences in effort between degradations might simply not have

been detectable as the measure is not sensitive enough in this intelligibility range (56% for noise-vocoded speech; 79% for time-compressed speech). At the same time, different talkers were presented in each condition, which possibly introduced further variability. On the other hand, studies employing adaptive procedures have observed differences between listening conditions even when intelligibility is fixed at levels that are expected to yield maximum pupil dilation (e.g., 50% in Koelewijn et al., 2012; Borghini & Hazan, 2020). It appears that a direct comparison between the two methods (i.e., fixed acoustic parameters and fixed intelligibility) is not possible. While adaptive procedures measure the effort exerted by listeners to reach a certain level of performance (e.g., word recognition), the approach chosen in the current study measures the effort exerted by listeners to cope with a certain amount of degradation.

Peaks occurred predominantly within the retention period (0-1500 ms after sentence offset), which is in line with previous studies using sentence materials (Winn et al., 2018). For some listeners, particularly in quiet, peaks occurred however before the end of the sentence. Individual differences have recently been considered for pupillometry data (Lõo et al., 2016). Since pre-sentence peaks were observed mostly in quiet, the current findings possibly indicate that the task was less engaging for normal-hearing listeners. In fact, when speech is not degraded, lexical access can occur before words have been processed entirely (Mattys et al., 2012; Radeau et al., 2000). Latency of the peak pupil dilation was larger for degraded speech compared to speech in quiet, indicating higher effort for degraded speech (Zekveld et al., 2010). This finding is in accordance with the Ease of Language Understanding model (ELU), which predicts delayed lexical access when speech input does not match phonological representations (Rönnberg et al., 2008, 2013). Latency was also larger for time-compressed speech compared to all other conditions. Since intelligibility was overall high for time-compressed speech, this effect might not solely be due to increased demands. Even though dilation peaks usually appear with a delay of 0.7-1.2 s after sentence offset (Winn et al., 2018), it seems that a shorter sentence duration prolongs the peak. Note however that this delay only applies when pupil traces are aligned to sentence offset. In fact, peaks for time-

compressed speech occurred earlier when measured from sentence onset (see also Müller et al., 2019). On the other hand, it possible that complete sentence processing occurred at a later stage, despite lower demands for time-compressed than for noise-vocoded speech - the pupil dilation peak has been suggested to reflect cumulative memory load (Piquado et al., 2010). In fact, the time-course of word recognition is not sequential when speech is degraded (Mattys et al., 2012). For instance, Radeau et al. (2000) found, for faster speaking rates, a disruption of pre-lexical recognition (the "uniqueness point effect").

Baseline pupil size declined within the first half of a block (24 trials), but was overall smaller for noise-vocoding and in quiet. This finding might be linked to overall intelligibility differences observed between conditions. As noise-vocoded speech and speech in quiet were respectively the most and least challenging conditions, the smaller baseline pupil size likely indicates two separate processes. On the one hand, the so-called phasic state of arousal - referring to task-related neural activity in the locus coeruleus (Aston-Jones & Cohen, 2005) - is characterised by a small baseline pupil size and a larger task-evoked dilation (Gilzenrat et al., 2010). Indeed, these observations were made for noise-vocoded speech. However, the phasic state is usually associated with elevated task performance whereas performance under noise-vocoding was lowest. Alternatively, it has been suggested that demanding tasks lead to a decrease in arousal (Ayasse & Wingfield, 2020) which is the more likely scenario under noise-vocoding in this study. On the other hand, a lower arousal level for speech in quiet might reflect less sustained attention, which was reported by Wagner et al. (2019) for listeners who were less challenged by the task. In summary, it appears that the interpretation of baseline pupil size measures requires similar levels of task performance as performance influenced overall arousal levels.

2.5.2 Source degradations: intelligibility and adaptation

Talkers who were more intelligible under noise-vocoding were also more intelligible under masking and time-compression. This finding confirmed my hy-

pothesis, replicating and extending results from Bent et al. (2009), who found that talkers intelligible under sine-wave vocoding were also more intelligible under babble noise. The authors argued that both types of degradations affected the spectral characteristics of talkers so that similar acoustic-phonetic properties were responsible for the effect. However, this explanation does not extend to time-compression because the signal processing technique used in this experiment (pitch synchronous overlap and add, Moulines & Charpentier, 1990) does ideally not change the spectral properties of speech. For masking, mean energy in the 1-3 kHz region predicted intelligibility, which is in line with previous studies (Green et al., 2007; Hazan et al., 2018; Hazan & Markham, 2004; Krause & Braida, 2004); this result is also expected given the predictions of the speech intelligibility index (SII). For masking, also vowel space dispersion predicted intelligibility. Larger vowel space dispersion is generally associated with overarticulated clear speech and has been found to predict intelligibility in previous studies (Bradlow et al., 1996; Lindblom, 1990; Picheny et al., 1986).

In the current study, vowel space dispersion also predicted intelligibility under noise-vocoding. As noise-vocoding leads to spectral broadening, especially with a small number of channels (Bent et al., 2009), improved discriminability of vowels and possibly other phonemes possibly improved word identification. Shannon et al. (1995) observed high vowel and sentence recognition even with only four noise-vocoder channels. Despite the use of two additional channels in the current study, the unpredictable sentence material combined with talker-specific factors rendered word recognition under noise-vocoding relatively challenging. However, it has to be noted that reduced frequency selectivity, as simulated by noise-vocoding, typically results in heavier reliance on temporal cues (Rosen, 1992). For instance, Winn et al. (2012) observed that specifically for lax-tense vowel contrasts, (simulated) cochlear implant users relied predominantly on durational cues (e.g., vowel duration change) and less on spectral cues (e.g., formant change). It is therefore possible that greater vowel space dispersion was also accompanied by temporal modifications enhancing phonetic contrasts. On the other hand, there is evidence that spectral formant cues can be obtained to some extent under noise-vocoding through

cross-channel comparisons (Roberts et al., 2011).

For time-compressed speech, only speaking rate contributed to intelligibility. Given the relatively high compression (37% of the original duration), it appears intuitive that overall longer speech segments provided more redundant and robust acoustic cues. As uniform time-compression affects all parts of speech equally (Dupoux & Green, 1997), any durational acoustic correlate of a slower speaking style might have been more robust against time-compression. For instance, longer vowel durations, which have also been associated with overarticulated speech (Lindblom, 1990), could have been less detrimentally affected by time-compression. Indeed, this suspicion was confirmed as speaking rate was strongly correlated with vowel duration. A recent study found that a larger vowel space was linked to higher intelligibility under time-compression (Johnson et al., 2020); however, only four talkers were used so that those findings were not conclusive. Speaking rate was correlated with vowel space dispersion in this study, but the correlation failed to reach significance.

For time-compressed and noise-vocoded speech, acoustic-phonetic features explained little of the variance observed in recognition performance across listeners. Performance in degraded speech perception has been linked to individual differences on cognitive measures such as working memory and vocabulary knowledge (Kennedy-Higgins et al., 2020; McLaughlin et al., 2018), so that listener-related factors might have been responsible for some of the variability, as well. Furthermore, it has been suggested that different talkers might employ different strategies to achieve high intelligibility so that there might not necessarily exist a single 'catch-all' feature that determines intelligibility (Hazan & Markham, 2004). Listeners adapted to noise-vocoded and time-compressed speech, but their improvement could not be linked back to specific acoustic-phonetic measures. Talkers in the current study were from the same accent group and talkers and listeners shared the same native language background. Therefore, it seems likely that listeners were familiar with the accent- or language-specific acoustic-phonetic characteristics so that talker-specific adaptation was not required. It can be assumed that adaptation

mainly functioned to overcome the channel degradation and not the source degradation. Systematic changes in the signal, introduced by time-compression and noise-vocoding, allowed for perceptual learning to occur. The finding that acoustic-phonetic features did not interfere with adaptation is line with previous studies on time-compressed speech. For instance, listeners have been shown to quickly adjust to talker changes under time-compression (Dupoux & Green, 1997), which might be considered even more disruptive than the talker differences considered here. The robustness of adaptation has also been shown in other studies that found changes in time-compression rate to not interfere with adaptation (Adank & Janse, 2009; Golomb et al., 2007). For instance, Golomb et al. (2007) found that participants adapted equally well to time-compressed speech (30%) with or without disruptive insertions of sentences spoken at a normal rate.

2.5.3 Source degradations: listening effort

For speech in quiet, talkers with greater vowel space dispersion were associated with a more attenuated pupil response, indicating reduced listening effort. Vowel space dispersion also emerged as a relevant feature for noise-vocoded and masked speech. Even though intelligibility for speech in quiet was high, it appears that the smaller pupil dilation indicated ease of processing for speech by talkers with a clear speaking style. Listeners have been shown to prefer over-articulated vowels when tasked to choose the best example amongst a range of synthesised vowels (Johnson et al., 1993). This subjective preference might be reflected in pupil dilation responses, as well, as indicated in the current study. These results also fit recent pupillometry findings regarding accented speech that was also highly intelligible (McLaughlin & Van Engen, 2020). Accented speech elicited larger pupil dilation than native speech. Applied to the current study, intrinsically less clear speech might have had a similar effect on pupil dilation than accented speech in McLaughlin & Van Engen (2020). However, two factors have to be taken into account when considering the presented findings. First of all, the overall variance of pupil dilation in quiet explained by

vowel space dispersion was low, indicating that talker differences did not contribute largely to individual pupil dilation and latency. Secondly, the fact that intelligibility was at ceiling in quiet poses the crucial question as to whether pupil dilation can be interpreted as listening effort. As normal-hearing listeners are unlikely to experience high effort for speech presented in quiet, it is possible that pupil dilation measures merely reflect attentional processes and are not a proxy for listening effort (Govender et al., 2019). Interestingly, speech presented in quiet in the current study does not meet the definition of an adverse condition as defined by Mattys et al. (2012). According to the authors, conversational speech can be considered adverse only when "it reduces intelligibility relative to citation form" (Mattys et al., 2012, p. 954). On the other hand, McLaughlin & Van Engen (2020) found elevated perceived effort ratings for accented speech, as well, which indicates that pupil dilation might indeed index effort, even at high intelligibility.

Since intelligibility differed between degradations, the current results indicated an effect of degradation level on the sensitivity of the pupil dilation measure. For stationary maskers, Wendt et al. (2018) found that peak dilation remained large when increasing the signal-to-noise ratio (SNR) from -8 dB to -4 dB, while improvement in sentence recognition was steepest (30-80%). Increasing the SNR from 0 dB to 4 dB resulted in a significant decrease in peak dilation, but virtually no difference in intelligibility, due to a ceiling effect.

The non-linearity of the pupil dilation response can explain why talker differences were not apparent for degraded speech in the current study: even though talker differences contributed to intelligibility in these conditions, the pupil dilation was not modulated further. For speech in quiet, even small differences in intelligibility associated with the acoustic-phonetic characteristics of the talkers were reflected in the pupil dilation response. Previously, also subjective ratings of listening effort have been shown to be more sensitive at higher intelligibility levels (Morimoto et al., 2004; Rennies et al., 2019).

2.5.4 Limitations

Firstly, despite the aim to target similar intelligibility levels across conditions, the large number of talkers and limited amount of test runs resulted in measurable differences. Therefore, direct comparisons of pupil dilation between conditions should also consider the impact of differing intelligibility levels on pupil dilation. Another limitation was the small number of listeners assigned to each talker. This decision was due to constraints imposed on the experimental design by adaptation and pupillometry measures, as outlined in the methods section. In particular, analyses conducted on talker averages (intelligibility across degradations) might have been affected by listener variability.

2.6 Conclusion

Intelligibility under different channel degradations was predicted by acoustic-phonetic features, confirming hypothesis H1: vowel space dispersion under noise-vocoding and masking, mean energy under masking and speaking rate under time-compression. Results also showed that talkers intelligible under noise-vocoding were also intelligible under masking and time-compression, confirming hypothesis H2. This finding extends previous research that found this effect for sine-wave vocoded and masked speech (Bent et al., 2009). Even though adaptation to noise-vocoded and time-compressed speech was observed, talker differences did not modulate the effect, in contrast to hypothesis H3. With respect to RQ1 (Chapter 1, Section 1.6), results presented in the current chapter showed that source degradations indeed interacted with channel degradations. However, results indicated that acoustic-phonetic source characteristics determining intelligibility vary across degradation types.

Pupillometry results showed that peak pupil dilation was larger and more delayed for degraded speech compared to speech presented in quiet, in accordance with hypothesis H4. On the other hand, a faster decline of baseline pupil size possibly indicated fatigue for both the easiest and hardest

listening condition (quiet and noise-vocoding, respectively). These results were contrary to hypothesis H5 that predicted more sustained baseline pupil size for conditions with adaptation (time-compression and noise-vocoding). Furthermore, a relationship between talker acoustics and listening effort (as indexed by pupil dilation) was not revealed by results of the current study, contrary to hypothesis H6. With respect to RQ2, rejection of hypothesis H6 suggests that source degradations do not interact with channel degradations that affect pupil dilation.

Pupillometry findings resulted in *two important conclusions*. First of all, peak dilation proved to be more reliable at high intelligibility levels, indicating differences in effort between time-compressed speech and speech in quiet, but not between degradations, i.e. noise-vocoding, masking and time-compression. The results of the current study therefore confirm that for speech perception, pupillometry as physiological measure of effort might be particularly useful when intelligibility is overall high. However, the current study could also show possible downsides of applying pupillometry at high intelligibility. Results indicated that speech by talkers with smaller vowel space dispersion was associated with larger peak dilation in quiet. While it is possible that this finding indicates higher listening effort for talkers with reduced vowel spaces, the fact that (1) intelligibility was at ceiling and (2) listeners had no hearing impairment raises doubts as to whether effort was the measured quantity or merely increased arousal - which would be the case if the task did not require mental exertion (Kahneman, 1973). The *second conclusion* from the current pupillometry findings concerns the experimental design. Individual differences between listeners were another possible reason for the lack of talker differences on the pupil dilation measures under degraded speech. Without methods to reduce variability between listeners, it is therefore advisable to employ within-subjects designs instead.

The study presented in the next chapter was built on the two principal conclusions for pupillometry shown above. While the aim of the study was to employ pupillometry for conditions at high intelligibility, it was expected that

such conditions would not require additional effort for normal-hearing listeners. Therefore, the study was conducted with hearing-impaired listeners who are expected to exert effort, even at high intelligibility levels (Pichora-Fuller et al., 2016). Testing conditions with high intelligibility allowed for the employment of time-compression at conversational speaking rates, which were more reflective of realistic listening situations. To circumvent the possibility that pupil dilation would index arousal independent of effort, subjective ratings were collected, as well.

Chapter 3: Listening effort experienced by hearing-impaired listeners processing fast speech with simulated room acoustics

3.1 Introduction

The study presented in the current chapter was conducted as part of a secondment at hearing aid manufacturer Sonova in Switzerland. The goal of the industry placement was to apply the pupillometry research tools developed in the academic setting to help build solutions for hearing aid users. Beneficiaries of objective measures of listening effort are ultimately hearing-impaired individuals, with many studies aiming to simulate some form of hearing loss or testing those groups directly (Ohlenforst et al., 2017; Wagner et al., 2019; Wendt et al., 2017; Winn et al., 2015, 2018). Pupillometry could potentially be used to steer hearing-aid interventions (McGarrigle et al., 2014; Wendt et al., 2020).

Findings presented in Chapter 2 confirmed previous findings in the pupillometry literature (e.g., Wendt et al., 2018), showing that pupil dilation in experiments with fixed acoustic parameters (e.g., time-compression rate) is more sensitive to differences between listening conditions at high intelligibility. While peak dilation was significantly larger under time-compression (~80% average intelligibility) compared to quiet, there was no difference between degradations, despite large differences in intelligibility (e.g., 56% average intelligibility under noise-vocoding). While a number of pupillometry studies have employed adaptive procedures to account for differences in intelligibility, adaptively adjusting noise levels might not necessarily be warranted when testing populations other than normal-hearing listeners (Winn et al., 2015). In fact, for hearing-impaired listeners, many everyday listening scenarios involve positive signal-to-noise ratios (Smeds et al., 2015). Moreover, hearing-impaired listeners are known to experience increased effort even under conditions with optimal audibility and intelligibility (Pichora-Fuller et al., 2016). As traditional

performance-based measures are not sensitive to differences between optimal and near optimal listening conditions, audiological researchers are interested in measures such as pupillometry to quantify effort objectively (Pichora-Fuller et al., 2016).

The experiment presented in this chapter was conducted to evaluate pupillometry in realistic listening environments with hearing-impaired listeners. Realistic listening environments were defined as a combination of source and channel degradation resulting in optimal intelligibility, but increased effort for hearing-impaired individuals. Specifically, speaking rate was manipulated and different room acoustics were simulated using a spheric arrangement of loudspeakers. I investigated whether hearing-impaired listeners would exert higher effort when presented with speech at fast speaking rates, even at high intelligibility levels (RQ3; see Chapter 1, Section 1.6). Additionally, I investigated whether listening effort would be modulated depending on the room acoustics, i.e., with and without reverberation (RQ4).

The following literature review complements the themes of Chapter 1 and 2, i.e., listening effort under channel and source degradations. Specifically, I will focus on existing research that investigated listening effort in more realistic environments, i.e., at high levels of intelligibility, with and without reverberation. I will also summarise research on the relationship between age- and hearing-related factors, and the ability to recognise reverberant and time-compressed speech. While perception of time-compressed speech has been discussed in much detail in the previous chapters, its relationship to aging and hearing loss has not been discussed in detail so far.

3.1.1 Listening effort at high intelligibility

The detrimental effects of degraded speech on intelligibility and effort are generally amplified by hearing impairment and aging (Gordon-Salant & Fitzgibbons, 1993; Kramer et al., 1997; Ohlenforst et al., 2017). It has been suggested that the allocation of additional cognitive resources can partly compensate for

such deficits (Heinrich et al., 2016; Rönnberg et al., 2013; Wingfield et al., 2005). However, the associated increase in effort can disrupt *downstream operations* such as working memory encoding. For instance, it has been shown that noise exacerbates word recall, even if it does not disrupt recognition itself (e.g., Rabbitt, 1968; Heinrich et al., 2008). Hearing-aid algorithms such as noise reduction have been shown to benefit intelligibility and lower effort, as indicated by pupillometry and dual-task paradigms (Ohlenforst et al., 2018; Sarampalis et al., 2009). Moreover, noise-reduction has been shown to lower effort even when intelligibility is close to optimal (Wendt et al., 2017). In their study, Wendt et al. (2017) evaluated listening effort during speech perception in noise for 24 hearing-impaired listeners. Listeners were fitted with hearing aids, and a noise reduction program was either turned on or off. Speech perception tests were conducted at two intelligibility levels (50% and 95%), determined beforehand for each listener using an adaptive procedure. While noise reduction resulted in significantly better performance at low intelligibility, there was no performance improvement at high intelligibility. It should however be noted that significance was marginal ($p = .07$) with noise reduction yielding better performance. Peak pupil dilation was significantly larger without noise-reduction, at both intelligibility levels. This result indicated lower effort for the noise reduction program, even when intelligibility was overall high (95%). The authors did not report subjective ratings of effort so that no comparisons between objective and subjective measures were made.

As discussed in Chapter 1 (Section 1.2.2.1), room acoustics play an important role in everyday communication. However, in pupillometry studies, reverberation has been largely ignored, despite posing a fundamental problem for hearing-impaired listeners (Picou et al., 2016; Zahorik & Brandewie, 2016). In studies with normal-hearing listeners, perceived effort is increased for reverberant speech, even when intelligibility is high (Morimoto et al., 2004; Rennies et al., 2019). Only few pupillometry studies have incorporated reverberation in their experimental design. McCloy et al. (2017) investigated the effect of attention switching and reverberation on the pupil dilation. Participants were asked to attend to one of two simultaneously presented auditory streams of spoken al-

phabet letters, while maintaining or switching streams halfway through a trial. The task was to respond to the occurrence of the letter "O" in the attended stream by pressing a button. Whether participants had to maintain or switch attention was indicated by a cue in the beginning of each trial. The initially attended auditory stream always contained speech by the same male talker while the second stream contained speech by the same talker or another talker. Speech was presented either with or without simulated reverberation. While pupil dilation emerged as significantly larger for switch than maintain trials following the initial cue, reverberation or talker type did not affect pupil dilation. McCloy et al. (2017) concluded that the type of stimulus, i.e., letter streams instead of sentences, resulted in a null effect for reverberation. Specifically, it was hypothesised that the lack of context information normally provided by full sentence material did not provide listeners with the opportunity to reconstruct the degraded speech streams. Another possibility is that reverberation did not elicit enough effort given that listeners were normal-hearing individuals. Even though behaviourally, participants performed better without reverberation - as indicated by faster reaction times and higher hit rate - similar amounts of effort might have been exerted for conditions with and without reverberation.

3.1.2 Processing time-compressed and reverberant speech: effects of hearing impairment and aging

Older hearing-impaired listeners are more prone to the detrimental effect of reverberation and other temporal degradations such as time-compression (e.g., Nabelek & Robinette, 1978; Gordon-Salant & Fitzgibbons, 1993, 2004). In a study using reverberant as well as time-compressed speech, Gordon-Salant & Fitzgibbons (1993) investigated the relationship between recognition performance and age, hearing loss as well as auditory processing measures. While hearing loss was associated with lower performance when processing reverberant and time-compressed speech, higher age was associated with lower performance, as well, independent of hearing loss. It was concluded that the general age-related cognitive decline contributed to this effect. The lifespan trajectory

of cognitive function follows an inverse U-shape and refers to a decline in fluid cognitive operations (or cognitive control) rather than crystallised knowledge representations, which are in fact maintained at older age (Craik & Bialystok, 2006). Age-related cognitive decline is expressed as a decline in executive functions such as inhibitory control and working memory (West, 1996) which have been discussed in more depth in Chapter 1 of this thesis.

In the context of fast (time-compressed) speech, researchers have investigated cognitive slowing, i.e., the speed at which cognitive functions operate (Janse, 2009; Salthouse, 1996; Schneider et al., 2005). For instance, Janse (2009) found that older adults' difficulties with time-compressed speech (50% and 67% of the original duration) were associated with hearing loss, but also cognitive processing speed, as measured by a reading speed test. Similar to Gordon-Salant & Fitzgibbons (1993), findings therefore supported the role of acoustic as well as cognitive factors for the perception of time-compressed speech. However, such results might be contingent on the severity of the chosen time-compression rate. For instance, Wingfield et al. (2003) presented time-compressed speech (80% and 65% of the original duration) to older and younger listeners. For such mild compression rates, with potentially little acoustic degradation, the authors showed that older listeners achieved high comprehension performance with subject-relative sentences that was comparable to the performance of young listeners. At the same time, response times were generally slower for older listeners and increased with more severe compression rates, indicating slower processing time. This finding was potentially linked to age-related cognitive slowing.

In contrast, other studies have promoted a more dominant role of acoustic factors. Gordon-Salant & Fitzgibbons (2001) investigated different time-compression techniques to establish whether the acoustic degradation or faster information rate contributes to difficulty with time-compressed speech for older listeners. Sentences were compressed to 50% of their original duration using either uniform time-compression, or selective time-compression of pauses, vowels, or consonants. All time-compression techniques resulted in

different sentence durations since the overall proportion and length of the segments differed. Therefore, uniform time-compression resulted in the largest intelligibility drop, specifically for older hearing-impaired listeners. Moreover, it was observed that the difficulty of older hearing-impaired listeners with selective time-compression was mostly driven by the shortening of consonants. Since perception of consonants is particularly diminished for listeners with peripheral hearing loss (Helfer & Wilber, 1990), this finding was highly expected. While the shortening of vowels led to some intelligibility decrement, no effect was observed for shortening pause segments. However, Gordon-Salant & Fitzgibbons (2001) noted that pause segments were rare and short so that time-compression did not substantially reduce their overall duration.

Another argument in favour of acoustic factors was provided by Schneider et al. (2005) who investigated different time-compression methods, as well: deletion of every third amplitude sample, every other 10-ms segment or steady-state segments (67% time-compression rate). Older and younger listeners were differentially affected by different time-compression methods, suggesting the contribution of acoustic factors. While deleting every third amplitude sample was detrimental for older listeners, it had the smallest effect on younger listeners. This method severely alters the spectral characteristics of speech by shifting frequencies upward, with shorter formant transitions and stop consonant gaps (Schneider et al., 2005). In contrast, when only removing steady-state segments, i.e., leaving the spectral characteristics widely intact, there was no performance difference between older and younger listeners - specifically for low-context sentences and partly for high-context sentences. The authors therefore argued that the acoustic effects of time-compression, e.g., through consonant shortening (Gordon-Salant & Fitzgibbons, 2001), led to increased difficulty for older listeners.

Oscillation accounts of speech perception (e.g., Ghitza, 2011) suggest that acoustic degradations introduced by time-compression can partly be compensated by altering information rate through insertion of silence intervals (Ghitza & Greenberg, 2009). Similarly, Wingfield et al. (1999) observed that

listeners' performance with time-compressed speech (68% and 55%) increased when silence intervals were inserted after clause and sentence endings. Performance improved for both older and younger listeners, suggesting the involvement of cognitive rather than acoustic factors since inserting silence intervals did not alter the original speech acoustics. However, while for young adult listeners performance returned to baseline with the faster rate (55%), this was not the case for older adult listeners, which led the authors to conclude that both cognitive and acoustic factors contributed to older listeners' difficulties with time-compressed speech.

In summary, it appears that acoustic degradations introduced through time-compression (specifically consonant shortening) have a detrimental effect on intelligibility. However, cognitive decline, in particular slower processing speed associated with higher age, might contribute to the difficulty experienced when processing time-compressed speech (Salthouse, 1996). It therefore seems reasonable to suggest that the effort exerted by older hearing-impaired listeners when processing time-compressed speech might be elevated even without observing severe intelligibility decrements. Conditions with high intelligibility are of particular interest since those are more likely to reflect realistic listening environments. Such environments are often corrupted by reverberant room acoustics which are particularly problematic for hearing-impaired listeners (Zahorik & Brandewie, 2016). However, little pupillometry research to date has investigated reverberant speech, and none has aimed to evaluate pupillometry in an ecological multi-loudspeaker setup that allows to simulate room acoustics in the laboratory.

3.1.3 Aims of the current study

The current study was conducted to investigate two research questions (see Chapter 1, Section 1.6). First, I asked whether listening effort experienced by hearing-impaired listeners would be modulated by speaking rate even at overall high intelligibility levels (RQ3). Second, I asked whether room acoustics would

influence listening effort at high intelligibility levels (RQ4).

Time-compression was applied to mimic fast speaking rates. Note that time-compression was introduced as channel degradation in Chapter 1 and 2, but is treated as source degradation in the current study. The reason to consider time-compression as channel degradation was that many studies target speaking rates that are outside the range of natural speaking rates. For instance, the study presented in Chapter 2 of this thesis used time-compression resulting in speaking rates of 10.3 syllables per second, given an average speaking rate of 3.8 syllables per second and a compression rate of 37%. Speaking rates between 4-6 syllables per second are considered normal for conversational speech in West Germanic languages such as English (Koch & Janse, 2016). In the current study, I therefore applied time-compression to mimic faster speaking rates, similar to the observed idiosyncratic speaking rate differences in Chapter 2.

Drawing on studies suggesting that age-related cognitive slowing contributes to difficulties with time-compressed speech (Janse, 2009; Wingfield et al., 2003), I hypothesised increased effort for older hearing-impaired listeners when processing time-compressed speech, even with optimal intelligibility (H7). I expected this effect to persist in a retest session (H8) given that the individual cognitive factors responsible for increased effort would not change in such a short period of time (cf. Kuchinsky et al., 2014).

Furthermore, I hypothesised that listening effort would be modulated by room acoustics, in both test and retest session (H9 and H10). Reverberation was expected to be more detrimental for fast speech than slow speech (H11) because of the combined effect of acoustic degradations (Gordon-Salant & Fitzgibbons, 1995; Picou et al., 2016). As the main goal was to stay as close to realistic communication environments as possible, room acoustics with and without reverberation were simulated using Higher Order Ambisonics (HOA, Pulkki, 2001). Participants were older hearing-impaired listeners who conducted a listening experiment combined with pupillometry, similar to the one reported in Chapter 2. In addition, participants were fitted hearing aids equipped with a dereverberation feature that was either turned on or off. As noise reduction has been

previously linked to reduced listening effort in noise even at high intelligibility (Wendt et al., 2017), it was hypothesised that dereverberation would have a similar effect under reverberant room acoustics (H12).

3.2 Methods

3.2.1 Participants

Nineteen hearing-impaired listeners were recruited for the experiment from which five took part in the pilot study [one female; $Age_M = 73.8(4.2)$ years] and thirteen in the main study [five females; $Age_M = 74.6(6.5)$ years]. All participants were native speakers of German. Potential participants were invited to take part in a screening session, which consisted of an anamnesis, audiometric and cognitive tests.

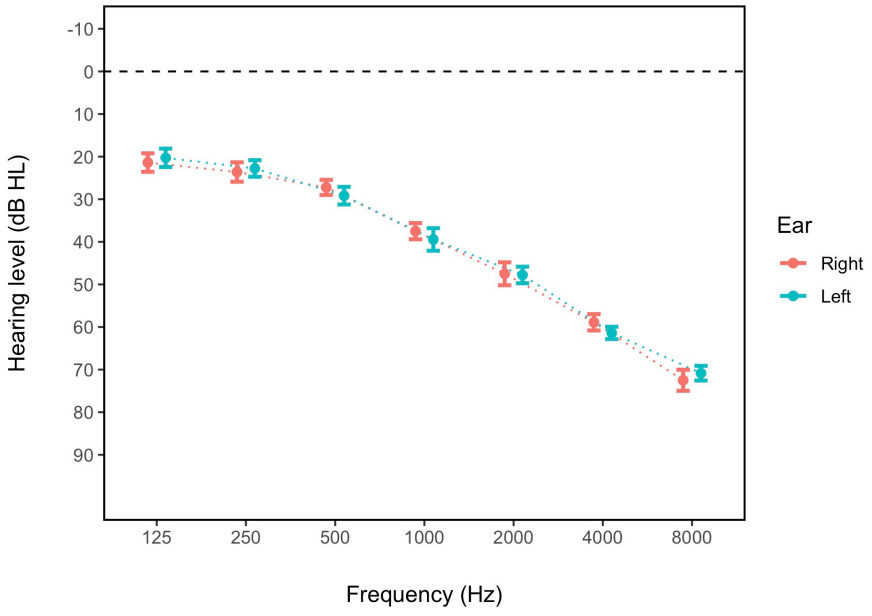

Figure 12: *Audiogram averaged across all 18 listeners. Points indicate means across participants and bars indicate one standard error around the mean.*

The audiological inclusion criterion was a symmetrical sensorineural hearing loss in the range N2-N4 (mild to moderate-severe). One participant was excluded because of sloping hearing loss. Audiogram averages for both ears are shown in Figure 12. Nine participants were hearing aid users (Pilot: three) and

experience ranged up to 17.5 years (Pilot: 30 years). A Trail-Making Test (TMT; Reitan, 1958) was conducted to measure executive function. Participants were also screened for medication with possible effects on pupil dilation (see Winn et al., 2018). Ethical approval was obtained by the Ethics Commission of the Canton of Zurich, Switzerland. Participants in the main experiment took part in a test, as well as a retest session, which were approximately one week apart.

3.2.2 Materials

Sentences were taken from the Oldenburger Satztest (OLSA, Wagener et al., 1999). These matrix sentences were constructed from random permutations of words from ten basis sentences with an average German phoneme distribution and low semantic predictability. Each sentence consists of five words and has the same syntactic structure: name, verb, number, adjective, object (e.g., German: *Peter bekommt drei grosse Blumen*; English: *Peter receives three large flowers*). Lists consist of 20 sentences so that every word from the basis set occurs exactly two times. Sentences have equal rms levels. The first six lists were randomly assigned to conditions for each participant. The first 18 sentences (pilot: 16) of each list were then randomly presented to participants. Eight sentences used for practice trials were randomly selected from the remaining sentences of each list.

Sentences were time-compressed using Waveform Synchronous Overlap and Add (WSOLA; Verhelst & Roelands, 1993). WSOLA optimises the quality of the output by removing segments at places of maximum correlation with the previous segment. In comparison to PSOLA (Moulines & Charpentier, 1990), WSOLA does not require pitch marking which makes it more suitable for real-time applications (Roebel, 2010). For the main experiment, sentences were time-compressed to 70% of their original duration (fast speech). This compression rate was chosen based on pilot data, indicating high intelligibility for hearing-impaired listeners with and without reverberation. The complete results of the pilot experiment are reported in Section 3.3.

The speaking rate of the original (slow) speech was 3.88 syllables per second (Wagener et al., 1999) so that the speaking rate of fast speech after time-compression was 5.54 syllables per second. This rate is within the range of average conversational speaking rates in West Germanic languages (Koch & Janse, 2016).

3.2.3 Equipment

Two different room acoustics were simulated. To that end, the original and time-compressed sentences were convoluted with two impulse responses that were recorded in a large foyer (reverb) and in a park (dry). Mean reverberation times (RT60) were calculated for each impulse response across the full spectrum and were 0.98 s (reverb) and 0.05 s (dry). The RT60 measure indicates the time needed for the sound pressure level to decrease by 60 dB. Impulse responses were equalised for rms. Before convoluting sentence and impulse response, sentences were padded with 1s of silence to account for late reflections. In the experiment, the original sound field was recreated using Higher Order Ambisonics (HOA, Pulkki, 2001) with a spheric arrangement of 32 loudspeakers. Output levels were calibrated to 53 dBA SPL in the sweet spot of the loudspeaker arrangement, using a Norsonic Type 140 sound level meter.

The hearing aids used in the experiment were Phonak Audéo M 90-312. None of the participants was a habitual user of this specific type of hearing aid. They were bilaterally fitted with receiver-in-canal (RIC) and disposable power dome, i.e., without venting, to maximise the effectiveness of the hearing aid program (Magnusson et al., 2013). Two different programs were used in the main experiment, *Calm situation* and *Comfort in Echo*. Both are readily implemented in the Phonak Audéo M hearing aid. The dereverberation feature of *Comfort in Echo* was expected to lower effort in the reverberant room, in comparison to *Calm situation*. The feature is based on a method described by Lebart et al. (2001). The programs are henceforth referred to as NoDR (no dereverberation) and DR

(dereverberation).

3.2.4 Design & Procedure

In total, testing conditions consisted of two speaking rates (slow and fast), two rooms (dry and reverb) and two hearing aid programs (NoDR and DR). Since DR was designed to show improvements in reverberant listening conditions, it was only tested in reverb, at two speaking rates. This resulted in six conditions in total. All four conditions nested under NoDR (two speaking rates, two rooms) were counterbalanced according to a Latin square design. The two conditions nested under DR and the order of hearing aid programs were also counterbalanced. This resulted in 16 different combinations from which 13 were randomly selected. In the retest session, the presentation order of conditions was reversed for each participant.

Each session started with the hearing aid fitting procedure. The procedure was automated by the product software (*Phonak Target*) with gains based on individual audiograms. It was ensured that no feedback was present. Participants were seated in the center of the loudspeaker arrangement, facing the experiment screen at a distance of 85 cm. The screen was located next to the loudspeaker representing the target location. Participants were asked to use a chin rest mounted on a table, located at the sweet spot of the sound field. The sweet spot is the location at which the sound field reconstruction is considered to be the least affected by aliasing. The eyetracker (Eyelink Portable Duo; SR Research, Oakville, Canada) was placed on the same table at a distance of 45 cm from the participant's eye. The light level was kept constant at 100-110 lux (light source from ceiling; left side). Due to the lack of a visible pupil response, the light level for one participant was increased by switching on an additional light (ceiling; right side), resulting in 130-140 lux. Participants were allowed to wear their glasses during the experiment. Cases of morphology changes in the pupil dilation have been previously reported when vision correction was removed (Wagner et al., 2019). The screen background was grey [$hsv = (0, 0, .2)$]

at 20% luminance. The fixation cross was yellow [$hsv = (.1558, 1, .5)$] during the trial and changed to blue [$hsv = (.561, 1, .5)$] to indicate response onset. Both fixation crosses had the same luminance (50%).

Figure 13: *Perceived effort rating display (visual analogue scale 0-100%) and game pad. The movable slider was initially located at 50%. Effort minimum (-) and maximum (+) were always indicated in the display.*

Before the start of each trial, the experimenter ensured that pupil size and the number of blinks had returned to a stable baseline. Each trial started with a pure tone (500 Hz) played for 500 ms from the target loudspeaker position. Simultaneously, the fixation cross appeared on the screen. After an inter-stimulus-interval of 500 ms, the baseline pupil size was recorded for 2000 ms. Then, the sentence was played, which was followed by a retention interval of 2000 ms. Each sentence was padded with one extra second to account for late reflections so that the effective retention interval was 3000 ms. The fixation cross changed colour to indicate that a response was requested from the participant. Participants repeated back as many words as possible which were then logged by

the experimenter. As sentences followed a normed structure (e.g., only plural nouns), scoring was strict in comparison to the experiment presented in Chapter 2, in which missing plural endings were counted as correct. The fixation cross disappeared after the response was logged. Participants were asked to blink between trials if necessary to lower the probability of within-trial blinks. Subjective ratings of listening effort were requested from participants after the end of each block/condition, using a visual analogue scale. Participants used a game pad to move a slider onto a position between - (minimum effort) and + (maximum effort). The default location of the slider was at the center of the scale (see Figure 13).

3.2.5 Preprocessing and statistical analysis

Word recognition was calculated for each listener and condition by averaging the percentage of words recognised correctly across trials. As described in detail below however, the statistical analysis was conducted on the total number of words recognised in each condition. Perceived effort ratings were only requested once in each condition so that percentage scores were obtained directly from the visual analogue scale. Pupillometry data were preprocessed in the same way as described in Chapter 2. No block was excluded after preprocessing due to an insufficient number of trials ($<50\%$). On average, 17.76 trials were included for each listener and condition. Peak dilation and peak latency were extracted.

Data from the main experiment were analysed separately for two different contrasts (see Figure 14). These planned comparisons were justified since hypotheses differed (see Introduction of this Chapter, Section 3.1.3). Contrast 1 (henceforth *Room*) investigated main effects and interactions of speed, room and session, while program was set to NoDR. In accordance with H7 and H9, I expected larger and more delayed peak pupil dilation as well as higher perceived effort ratings for fast speech compared slow speech, and for reverberant speech compared to dry speech (main effects of speed and room). Given the combined effect of fast and reverberant speech, I further expected an interac-

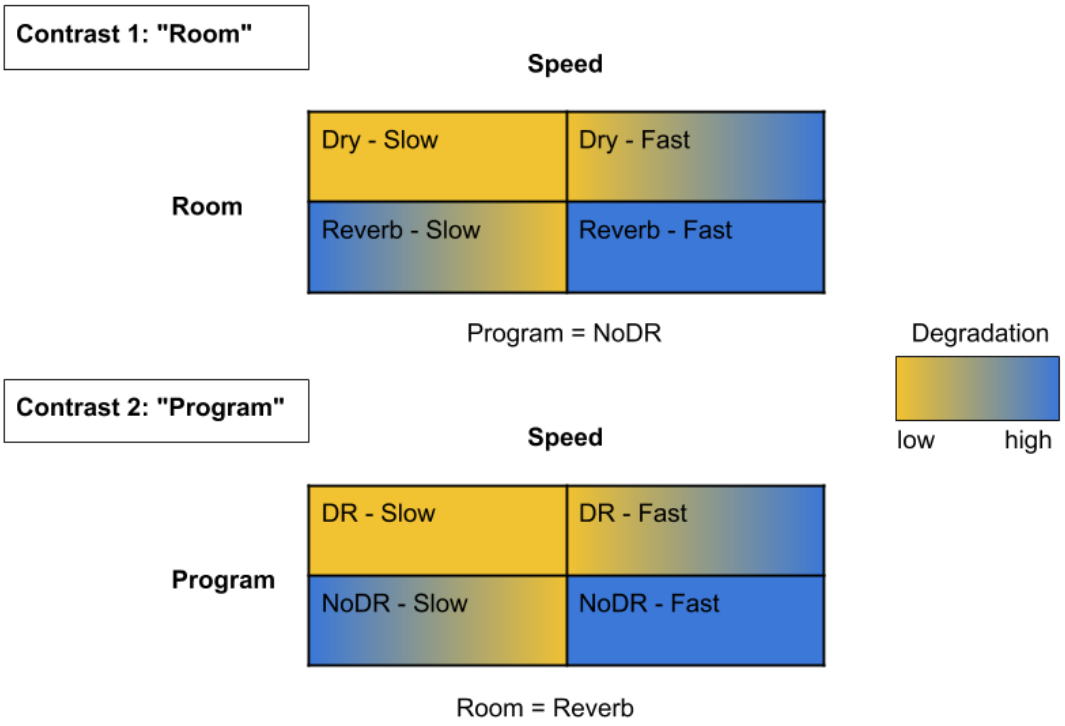

Figure 14: *Visualisation of the two contrasts used for analysis. Contrast 1 investigated the main effects and interaction of speech and room while contrast 2 investigated the main effects and interaction of speed and program.*

tion of speed and room (H11). In addition, I hypothesised that there would be no interactions with session, showing that the effects and interaction of speed and room would persist in a retest session (H8 and H10).

Contrast 2 (henceforth *Program*) investigated main effects and interactions of speed, program and session, while room was reverberant-only. In accordance with H12, I expected that dereverberation would reverse the effect of reverberation (see H9 and H11), leading to decreased perceived effort ratings and peak pupil dilation as well as latency (main effect of program and interaction of program and speed).

To analyse perceived effort and pupil dilation, linear mixed models (LMMs) were fitted using *lme4* in R (Bates et al., 2015). In all models, random intercepts for listeners were allowed. LMMs were analysed using F-tests provided by *lmerTest* (Kuznetsova et al., 2017), similar to the analysis presented in Chapter 2. To account for near-ceiling performance, word recognition data was ana-

lysed using mixed effects logistic regression, with number of words (out of 90, i.e., 5 keywords times 18 trials) identified or not identified as binomial outcome variable. Since mixed effects logistic regression models cannot be analysed using F-tests with Satterthwaite approximation (Kuznetsova et al., 2017), model comparison was then applied to determine significance by sequentially adding fixed effects and interactions.

3.3 Results: pilot experiment

I conducted a pilot experiment to establish the difficulty of each listening condition. To estimate performance in the main experiment, participants in the pilot were from the same population as participants in the main experiment. The original rate and two time-compression rates (60% and 70%) were tested in both rooms, dry and reverb. Similar hearing aids were fitted, and program NoDR was switched on throughout the experiment.

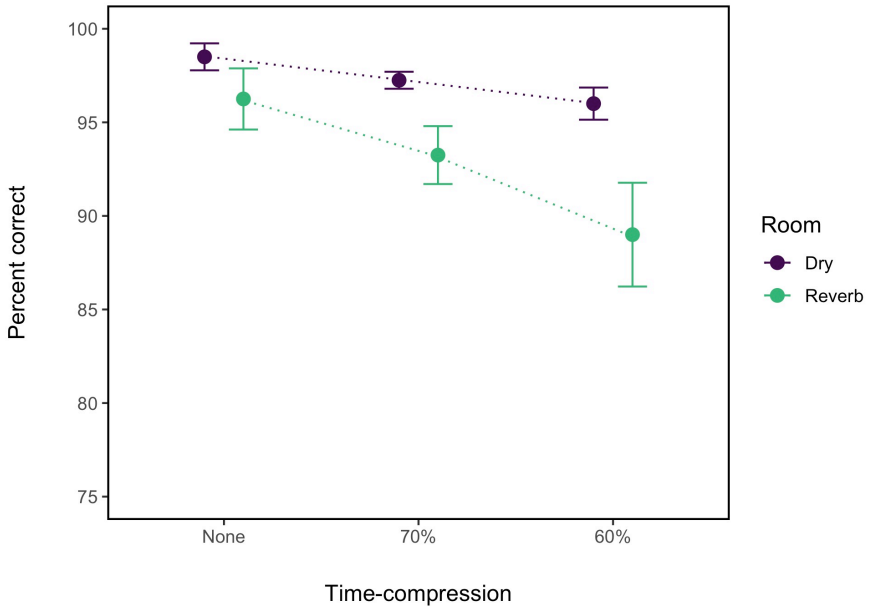

Figure 15: *Word recognition performance in the pilot experiment. Points indicate means across participants and bars indicate one standard error around the mean.*

Word recognition performance (see Figure 15) was overall high in dry [$M_{100\%} = 98.5\%(2.1), M_{70\%} = 97.2\%(2.2), M_{60\%} = 96.0\%(3.1)$] and reverb [$M_{100\%} = 96.2\%(3.8), M_{70\%} = 93.2\%(4.2), M_{60\%} = 89.0\%(7.8)$]. Perceived effort

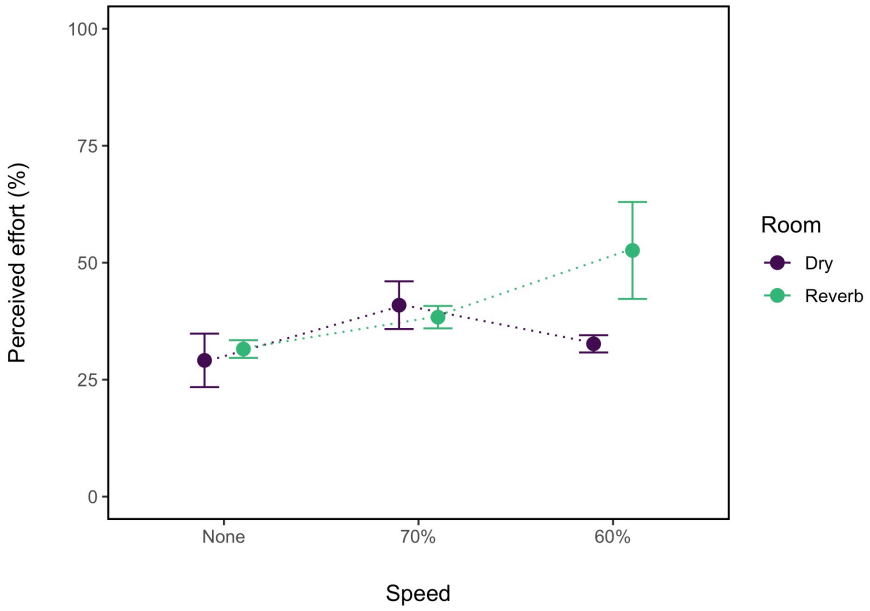

Figure 16: *Perceived effort in the pilot experiment. Points indicate means across participants and bars indicate one standard error around the mean.*

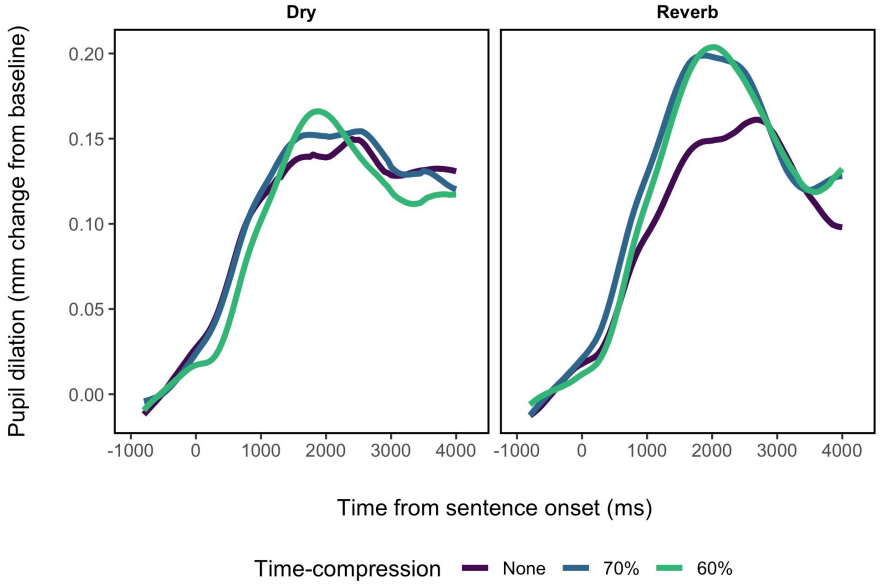

Figure 17: *Average pupil dilation in the pilot experiment. For visualisation purposes, curves have been smoothed with a moving-average filter.*

in reverb increased monotonically with time-compression rate (see Figure 16). However, in dry, perceived effort increased at 70% time-compression, but decreased again at 60%. This observation does not correspond to word recognition performance, which monotonically decreased with time-compression rate. Pupil traces averaged over all 5 listeners showed larger dilation for fast reverberant speech (see Figure 17). Visually, there was only a small difference between fast and slow speech in dry. In summary, pilot data indicated that while the performance decrease in dry was negligible when speech was time-compressed to 70% of its original duration (1.3% on average), perceived effort increased. However, this finding was not corroborated by pupillometry data. In reverb, both perceived effort and pupillometry data indicated higher effort for speech time-compressed to 70%. The decline in intelligibility was again relatively small (3% on average). I therefore chose 70% as time-compression parameter in the main experiment.

3.4 Results: main experiment

3.4.1 Word recognition

I first analysed word recognition performance. Mixed effects logistic regression models were fitted, testing for main effects and interactions of speed, room and session (Contrast 1), as well as speed, program and session (Contrast 2). For Contrast 1 (Room) (see Figure 18), the best model included a fixed effect of speed [$\chi^2(1) = 38.95, p < 0.001$], but also an interaction between speed and room [$\chi^2(1) = 7.53, p = 0.006$]. Submodels for room indicated smaller log odds for fast speech in reverb [$\beta = -1.00, z = -6.49, p < .001$], but no difference between slow and fast speech in dry [$p = .34$]. The best model also included a fixed effect of room [$\chi^2(1) = 77.10, p < 0.001$], with smaller log odds for speech in reverb [$\beta = -0.63, z = -2.91, p = .004$], and a fixed effect of session [$\chi^2(1) = 24.32, p < 0.001$], with larger log odds in the retest session [$\beta = 0.61, z = 4.88, p < .001$]. In summary, results showed that intelligibility was lower in reverb, but improved overall in the retest session. Furthermore,

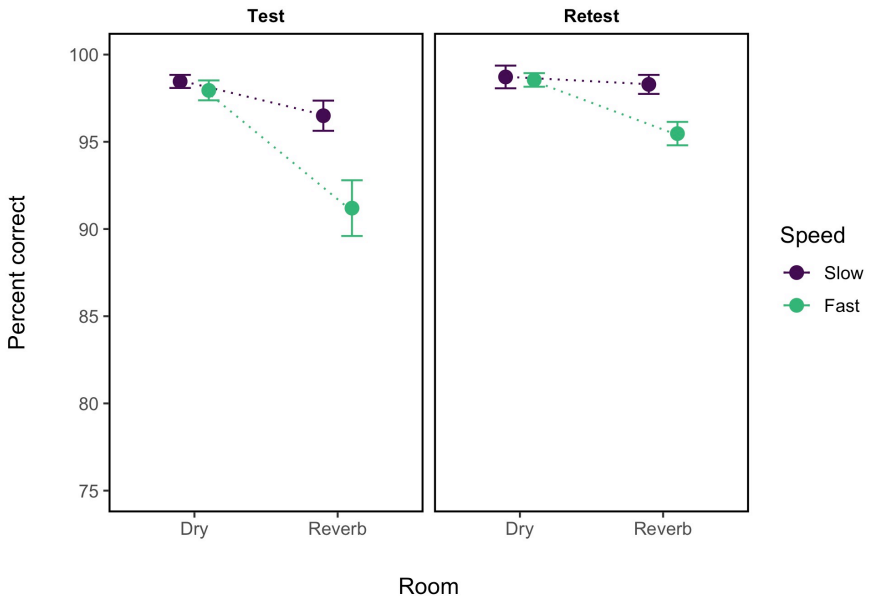

Figure 18: *Word recognition performance for Contrast 1 (Room). Points indicate means across participants and bars indicate one standard error around the mean.*

while fast speech was less intelligible than slow speech in reverb, it was not less intelligible than slow speech in dry.

Word recognition performance was also analysed for Contrast 2 (Program) (see Figure 19). The best model included a fixed effect of program [$\chi^2(1) = 6.27, p = 0.01$], but also an interaction between program and session [$\chi^2(1) = 4.23, p = 0.04$]. Submodels indicated smaller log odds for program DR than program NoDR in the retest session [$\beta = -0.48, z = -3.16, p = .002$], but no difference between programs in the test session. The best model also included a fixed effect of speed [$\chi^2(1) = 64.36, p < 0.001$], with smaller log odds for fast speech [$\beta = -0.78, z = -7.84, p < 0.001$]. Furthermore, the model included a fixed effect of session [$\chi^2(1) = 27.72, p < 0.001$], with larger log odds in the retest session [$\beta = 0.73, z = 4.93, p < 0.001$]. In summary, results indicated that fast speech was less intelligible than slow speech (in reverb) and that intelligibility overall improved in the retest session. Furthermore, dereverberation (DR) yielded lower intelligibility than no dereverberation (NoDR) in the retest session.

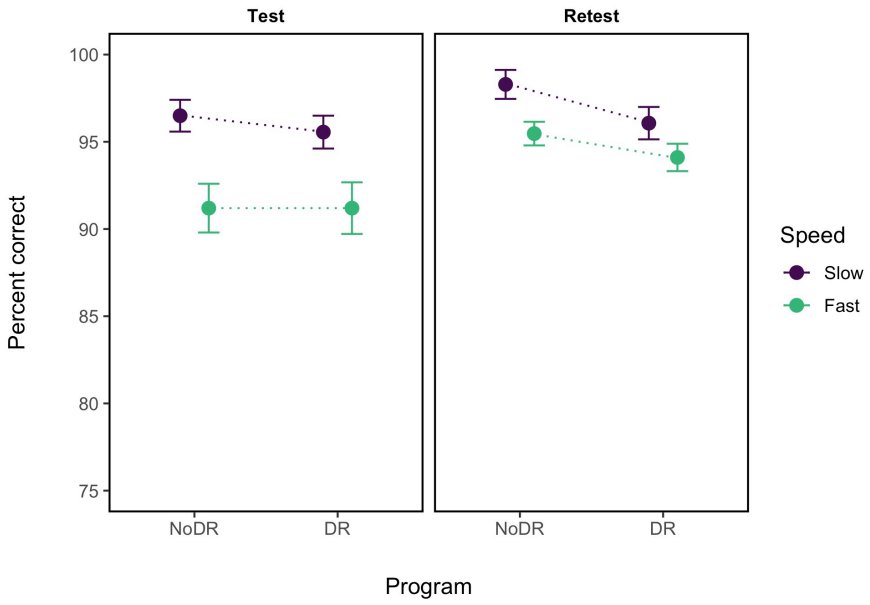

Figure 19: *Word recognition performance for Contrast 2 (Program). Points indicate means across participants and bars indicate one standard error around the mean.*

3.4.2 Effort ratings

I then analysed subjective ratings. Linear mixed models were fitted with effort ratings as dependent variable, testing for main effects and interactions of speed, room and session (Contrast 1), as well as speed, program and session (Contrast 2). For Contrast 1 (Room), there was a main effect of room [$F(1, 84) = 13.77, p < .001$], with higher effort reported for reverberant speech (see Figure 21). While on average, there was a tendency for listeners to rate reverberant speech at 51%, i.e., around the slider's initial position, individual responses were spread across the scale (see Figure 20).

There were no other main effects or interactions. I used Bayes factors to determine whether there was sufficient evidence to support the null hypothesis, i.e., the model with room as only fixed effect, in favour of models including the remaining fixed effects and interactions. According to Jeffreys (1961), a Bayes factor below 0.3 would indicate sufficient evidence, a factor between 0.3-3 insufficient evidence for the null hypothesis. Bayes factor analysis indicated that there was sufficient evidence to support the null hypothesis in almost all complex models ($B < 0.3$). However, there was insufficient evidence to

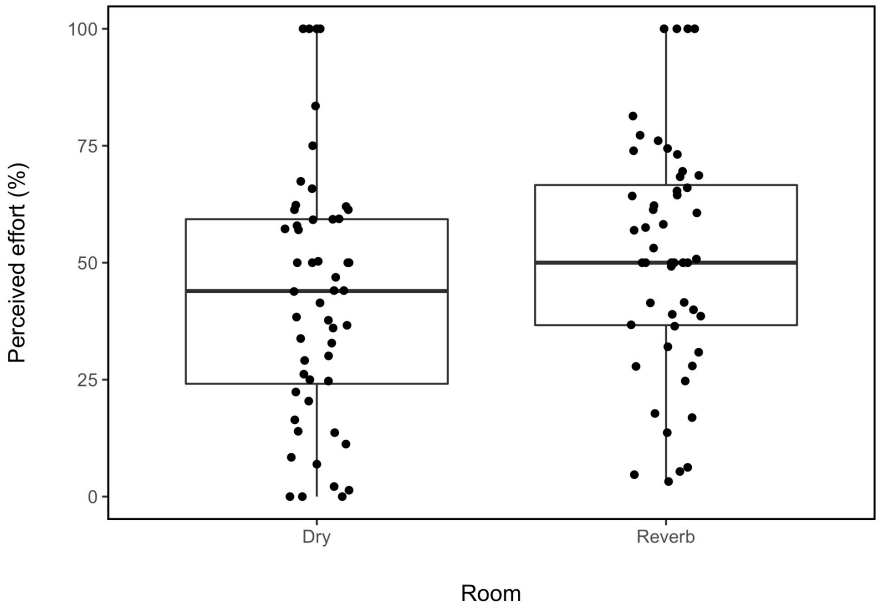

Figure 20: *Distributions of effort ratings for Contrast 1 (Room), split by room (dry vs. reverb). Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range. Jittered points represent individual effort ratings.*

support the null hypothesis for the inclusion of speed as fixed effect ($B = 0.58$). For Contrast 2 (Program), subjective ratings showed a main effect of speed [$F(1, 84) = 4.55, p = 0.036$], with higher effort reported for fast speech (see Figure 22). There were no other main effects or interactions and Bayes factors indicated that there was sufficient evidence to support the null hypothesis, i.e., a model with speed as only fixed effect ($B < 0.3$).

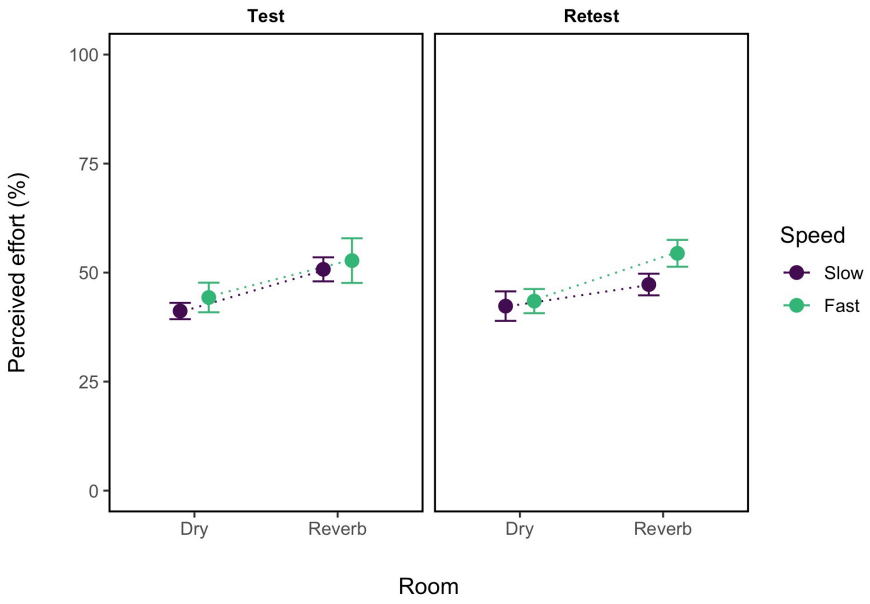

Figure 21: *Effort ratings for Contrast 1 (Room). Points indicate means across participants and bars indicate one standard error around the mean.*

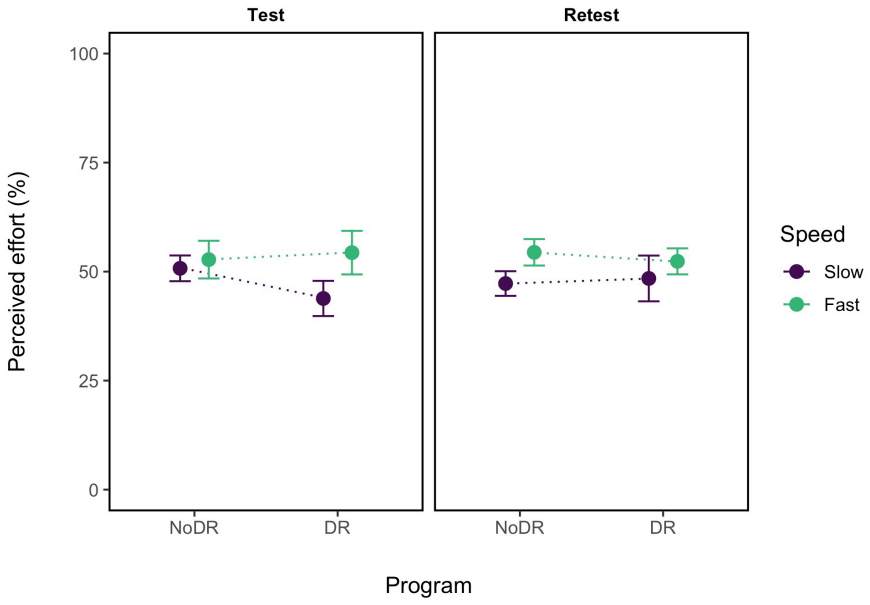

Figure 22: *Effort ratings for Contrast 2 (Program). Points indicate means across participants and bars indicate one standard error around the mean.*

3.4.3 Pupillometry

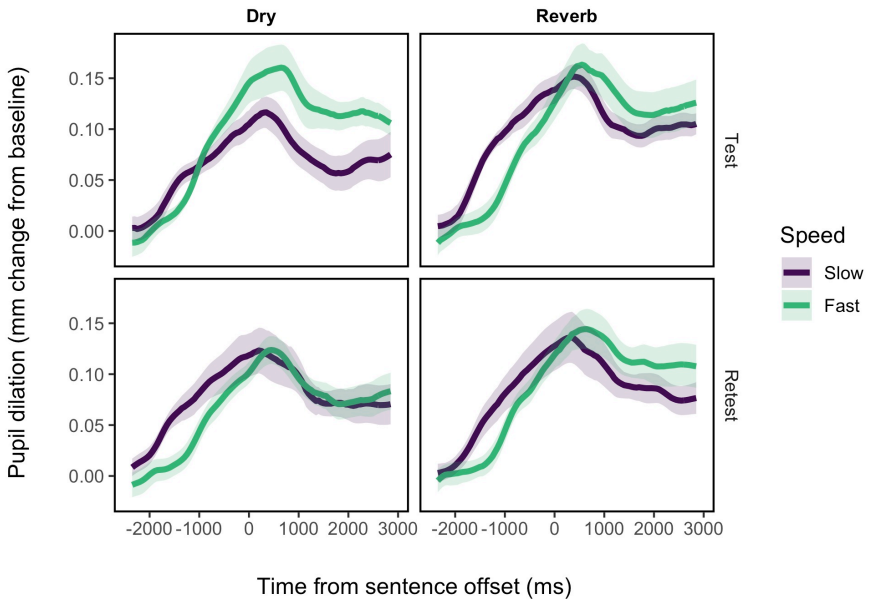

Figure 23: *Average pupil traces for Contrast 1 (Room). Bands indicate one standard error around the mean. Smoothed with a 5-point moving average filter.*

Average pupil traces are shown in Figure 23 and 24, for Contrast 1 (Room) and Contrast 2 (Program), respectively. Linear mixed models were fitted with peak pupil dilation and latency as dependent variables, testing for main effects and interactions of speed, room and session (Contrast 1), as well as speed, program and session (Contrast 2). For Contrast 1 (Room), peak dilation showed a main effect of speed [$F(1, 84) = 4.79, p = 0.03$], with larger peaks for fast speech. There was also a main effect of room [$F(1, 84) = 5.47, p = 0.02$], with larger peaks for speech in reverb.

There were no interactions between speed, room and session. I used Bayes factors to determine whether there was sufficient evidence to support the null hypothesis, i.e., the model with speed and room as fixed effects, in favour of models including interactions. While there was sufficient evidence supporting the null hypothesis against inclusion of an interaction between speed and room, there was insufficient evidence against inclusion of a three-way interaction between speed, room and session ($B = 0.87$). It is therefore possible that a decrease in peak dilation for fast speech in dry at retest would have been

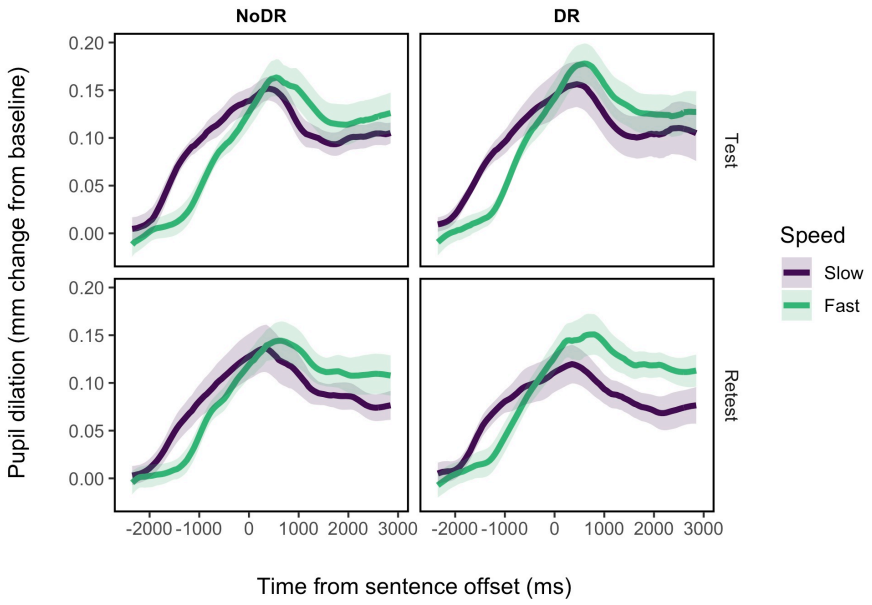

Figure 24: *Average pupil traces for Contrast 2 (Program). Bands indicate one standard error around the mean. Smoothed with a 5-point moving average filter.*

evident with a larger sample size.

For peak latency, there was a main effect of speed [$F(1, 84) = 4.80, p = 0.03$], with larger latencies observed for fast speech. There were no further main effects or interactions between speed, room and session, which was supported by Bayes factor analysis ($B < 0.3$).

Figure 25 shows distributions of peak dilation for conditions in Contrast 1 (Room). The boxplots show that distributions were characterised by few extreme peak values, indicating higher pupil reactivity. Given the small sample size and possible influences of a few but extreme outliers, I re-fitted all linear mixed models using robust estimation (Koller, 2016). The method assigns robustness weights to individual data points, thereby discounting outliers by assigning lower weights, without needing to discard them. While the main effect of speed survived robust estimation ($p = 0.03$), the main effect of reverb was only marginally significant after robust estimation ($p = 0.07$).

For Contrast 2 (Program), peak dilation showed a main effect of speed [$F(1, 84) = 5.44, p = 0.02$], with larger peaks for fast speech, which did however not survive robust estimation ($p = 0.61$). There was no effect of

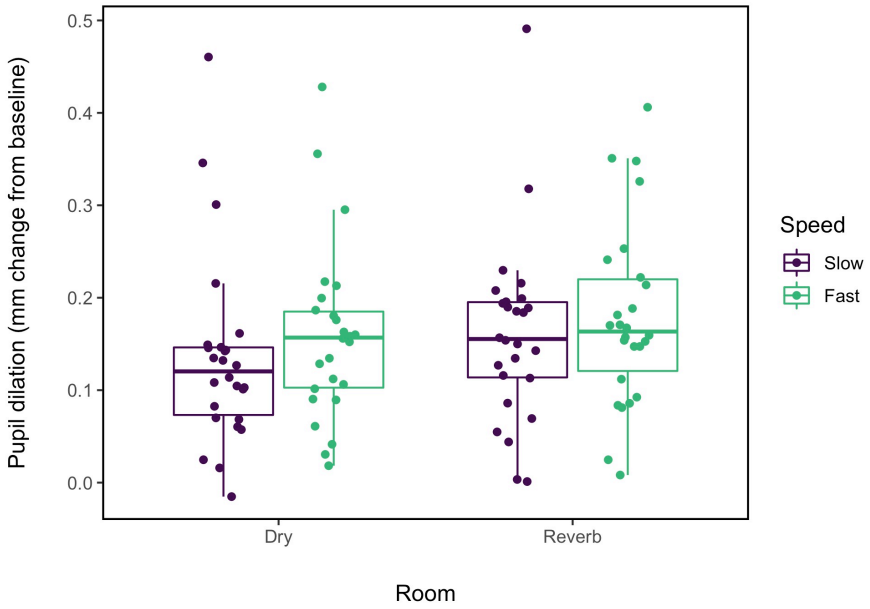

Figure 25: *Distributions of peak pupil dilation for Contrast 1 (Room), with combined data across test and retest sessions. Boxes represent values from the first to the third quartile with the median indicated by a solid line. Whiskers extend up to 1.5 times the interquartile range. Jittered points represent individual peaks.*

program nor interactions between speed, program and session. Bayes factors indicated that there was sufficient evidence to support the null hypothesis, i.e., the model with speed as fixed effect, in favour of more complex models ($B < 0.3$). Only when including an interaction of speed and session, Bayes factors indicated that there was not sufficient evidence to support the null hypothesis ($B = 0.35$). For peak latency, there were no effects or interactions of speed, program and session, with Bayes factors indicating sufficient evidence to support the null hypothesis, i.e., the intercept-only model ($B < 0.3$). For the fixed effect of speed, Bayes factors indicated insufficient evidence to support the null hypothesis ($B = 0.38$).

3.4.4 Individual differences

I tested whether individual differences between listeners on several outcome measures were linked to background measures. Outcome measures were recognition performance, perceived effort and peak pupil dilation; they were cal-

Table 5: *Linear regression results for recognition, perceived effort, and peak dilation. Difference scores were calculated between slow speech in dry and fast speech in reverb (test session). Higher scores indicate less detrimental effects of reverb and time-compression. Age, PTA = pure tone average, PS = processing speed. All continuous predictors are mean-centered and scaled to have SD = 1. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.*

	Recognition		Effort		Peak dilation	
(Intercept)	-7.3 **	(1.5)	-11.6	(5.2)	-68.6	(35.1)
Age	-4.3	(1.9)	3.7	(6.5)	26.5	(43.5)
PS	1.8	(1.9)	-0.0	(6.5)	-50.4	(43.6)
PTA	1.4	(1.6)	3.0	(5.4)	-34.8	(36.7)
R sq	0.40		0.08		0.21	
R sq (adj)	0.20		-0.23		-0.05	

culated as difference scores between the hardest and easiest condition, i.e., slow speech in dry and fast speech in reverb (in the test session). A measure of processing speed was obtained from the results of the trail making test (part A), i.e., the time required to connect ordered numbers (Fellows et al., 2017). Pure-tone averages (PTA) were obtained from the better ear across 0.5, 1, 2 and 4 kHz. Participant age was also entered into the analysis. Multiple linear regression results only indicated a marginal contribution of age to word recognition performance ($p = 0.05$), with a more detrimental effect of reverberation and time-compression with increasing age ($R^2_{Adj} = 0.20$, see Table 5).

3.5 Discussion

The present study investigated the listening effort experienced by older hearing-impaired listeners when processing fast speech with and without reverberant room acoustics. Specifically, conditions with high intelligibility were targeted that reflected realistic communication environments. A pupillometry experiment was conducted in two testing sessions, simulating fast speech by means of time-compression, and room acoustics by means of higher-order ambisonics. It was hypothesised that older hearing-impaired listeners would experience elevated effort when processing fast or reverberant speech, even at high intelligibility levels.

3.5.1 Performance

Word recognition performance was overall high ($>$90%), but differences between listening conditions were nevertheless observed. Speech in reverb was less intelligible than speech in dry in the test session. This finding is expected as reverberation degrades speech, with late reflections masking the direct sound arriving from the source (Bolt & MacDonald, 1949; Nabelek & Robinette, 1978; Picou et al., 2016). In particular, reverberation can be challenging for older hearing-impaired listeners (Gordon-Salant & Fitzgibbons, 1993, 1995; Nabelek & Robinette, 1978; Zahorik & Brandewie, 2016). It has been shown that the detrimental effect of reverberation can be amplified by time-compression (Gordon-Salant & Fitzgibbons, 1995). In the current study, fast speech in reverb was indeed less intelligible than slow speech in reverb. On the other hand, there was no significant difference in intelligibility between slow and fast speech in dry, due to a ceiling effect. Given the 'mild' time-compression rate (70% of the original sentence duration), it is possible that the acoustic degradation through time-compression (e.g., consonant shortening) was too weak to have detrimental effects on intelligibility. This finding is consistent with a previous study using similar compression rates (80% and 65%) (Wingfield et al., 2003).

The multiple linear regression analysis hinted towards an involvement of age in recognition performance: the detrimental effect of reverberation and time-compression was exacerbated by age, with older listeners achieving lower recognition performance. It has been suggested before that the decline of cognitive functions with age might explain difficulties in processing temporally degraded speech, specifically reverberant and time-compressed speech (Gordon-Salant & Fitzgibbons, 1993). One such cognitive function is processing speed (Salthouse, 1996). Janse (2009) observed a link between processing speed and accuracy in recognising time-compressed speech. In the current study, a measure of processing speed (Trail Making Test - Part A) did not predict intelligibility. It has to be noted that Janse (2009) measured processing speed during reading which provided a more specific assessment of processing speed in the language domain. Another possible explanation for age-related difficulties in processing fast reverberant speech is that temporal processing degrades with age (Goossens et al., 2017; Moore, 2008; Pichora-Fuller et al., 2007; Schoof & Rosen, 2014; Sergeyenko et al., 2013). However, no measure of temporal processing was obtained in the current study.

Overall, word recognition performance improved in the retest session. Several explanations are possible for this finding. First of all, prior exposure to room acoustics has been shown to improve intelligibility (Watkins, 2005; Zahorik & Brandewie, 2016). Specifically, Zahorik & Brandewie (2016) showed that listeners tolerated higher noise levels when reverberant target phrases were presented within a reverberant carrier phrase, allowing perceptual adjustment to the reverberant speech, similar to adaptation observed for time-compressed and noise-vocoded speech. Interestingly, these adaptation effects were constrained to room acoustics with reverberation times between 0.4 and 1 s, which also applies to the current study (RT60 = 0.98 s). However it is questionable whether these context effects extend to multiple testing sessions. It is also conceivable that listeners adapted to time-compressed speech, as demonstrated in Chapter 2 of this thesis and in numerous studies before (e.g., Dupoux & Green, 1997; Golomb et al., 2007; Kennedy-Higgins et al., 2020; Peelle & Wingfield, 2005). Moreover, Schlueter et al. (2015) observed within- and between-

session adaptation for time-compressed speech using the same sentence material (OLSA, Wagener et al., 1999). However, in the current study, intelligibility improvements were also observed for slow speech in reverb which had not been time-compressed. It is therefore more likely that learning was either specific to reverberation or both reverberation and time-compression. Furthermore, improved recognition might reflect familiarisation with the task or the hearing aids, possibly in combination with degradation-specific learning. Since adaptation or task familiarisation was not the primary focus of this study, no further conclusions can be drawn.

Dereverberation did not benefit word recognition. However, it was observed that in the retest session, word recognition performance was in fact lower when dereverberation was active. This difference was not apparent in the test session, in which performance was overall lower. The nature of this performance difference is unclear. However, it has to be noted that intelligibility was near ceiling so that no strong conclusions should be drawn from this performance difference. Dereverberation is mainly targeted at continuous speech so that it is possible that the functionality of the algorithm was compromised by the short sentence duration and breaks between sentence presentations. Indeed, preliminary follow-up tests indicate that with increasing pause duration between sentence presentations, dereverberation becomes less active. Further investigations of a possible interaction between dereverberation and sentence as well as pause duration are necessary to disentangle the effect found in the current study. Another confound that has to be considered is that the experimental design was not fully crossed (4 blocks with program NoDR, 2 blocks with program DR), i.e., listeners were exposed more to the default setting (NoDR, i.e., no dereverberation).

3.5.2 Listening effort

Perceived effort ratings underpinned performance results in part, showing that speech in reverb was perceived as more effortful. The rating scale in the current

study was continuous, ranging from 0% (minimum effort) to 100% (maximum effort). With an average rating of 51% in reverb and 43% in dry, it appears that despite those differences in perceived effort, the overall effort imposed by reverberation was still relatively low. Similarly, Sato et al. (2012) observed that for older listeners, intelligibility within a range from 80%-100% was associated with listening difficulty ratings of 0-60%. Stimuli were similar to the ones used in the current experiment, i.e., speech convolved with room impulse responses; however, with the addition of background noise at varying levels (35 to 60 dB at 5 dB intervals). In a study with hearing-impaired listeners, Kramer et al. (2016) measured an average self-rated effort of 3.1 ($low = 0, high = 10$) in quiet, and 7.6 in noise at 50% speech-reception threshold (SRT). In the current study, the average lower boundary was 41% (slow speech in dry) and the average upper boundary 53% (fast speech in reverb), both measured in the test session. Even the hardest condition was therefore perceived as substantially less effortful than the 50% SRT condition in Kramer et al. (2016). The level of perceived effort for the hardest condition in the current study was comparable to levels measured by Koelewijn et al. (2012) for young normal-hearing adults at 84% intelligibility in the presence of stationary noise (5.3, with $low = 0, high = 10$). The rating scale was similar to the one used in the current study, ranging continuously from 0 (no effort) to 10 (very effortful).

Peak pupil dilation was also larger in reverb, supporting the notion that reverberant speech was more effortful to process, as reflected by perceived effort ratings. It has to be noted however that the effect of reverberation was only marginally significant when robust linear mixed model estimation was applied. Peak pupil dilation was also sensitive to differences in speaking rate. Similarly, Müller et al. (2019) observed larger peak dilation for fast speech compared to slow speech at a fixed speech-reception threshold of 80%. The study was conducted with normal-hearing listeners, but given the lower intelligibility, the results might be comparable to the ones presented here for hearing-impaired listeners (cf. Koelewijn et al. (2012) for perceived effort). The magnitude of the pupil dilation in the current study was around 0.15 mm which would be indicative of an easy task compared to other speech perception studies (see Winn

et al., 2015 for a magnitude comparison). However, it has to be taken into account that older adults generally exhibit smaller pupil responses (Winn et al., 2015, 2018). Compared to a previous study conducted with hearing-impaired listeners ($M_{Age} = 59$) (Wendt et al., 2017), pupil dilation magnitude observed in the current study for harder conditions (fast and/or reverberant speech) corresponds to conditions between 50% and 95% intelligibility (without noise reduction).

Surprisingly, the difference between slow and fast speech was largest for dry in the test session. In this condition, there was no performance difference between slow and fast speech. Several explanations for this finding are possible. On the one hand, it has been suggested that the difficulty of older hearing-impaired adults to process time-compressed speech stems from acoustic degradations introduced by uniformly compressing speech segments, specifically consonants (Gordon-Salant & Fitzgibbons, 2001; Schneider et al., 2005). However, since intelligibility was very high in the current study, it is less likely that acoustic degradation was the driving factor leading to larger pupil dilation.
On the other hand, the increased information rate might have posed higher cognitive demands on older listeners who are more susceptible to cognitive decline. However, speaking rates for both slow and fast speech were well within the range of conversational speech (Koch & Janse, 2016). Oscillation-based models of speech perception would therefore predict optimal performance as speaking rate falls within the so-called theta range, that allows for successful syllable parsing (Ghitza, 2011; Peelle & Davis, 2012). Indeed, word recognition was at ceiling, indicating successful perception. It is therefore plausible that larger pupil dilation indicated the extra processing effort that was required by listeners to process faster speech in spite of receiver limitations (Lemke & Besser, 2016; Mattys et al., 2012) such as reduced processing speed (Janse, 2009; Salthouse, 1996).

Since perceived effort ratings did not indicate higher effort for fast speech, other explanations have to be taken into account, as well. It is possible that this type of fast speech was particularly novel for listeners. The capacity model of at-

tention (Kahneman, 1973, see also Chapter 1) suggests that resource allocation prioritises novel stimuli which leads in turn to increased arousal. It is therefore possible that artificial changes in speaking rate were perceived as unnatural and novel, triggering an arousal response of the autonomic nervous system. This explanation would fit results of a previous study that showed no effect of natural variation in speaking rate on pupil dilation (Koch & Janse, 2016). Interestingly, some participants in the current study reported that both fast and slow sentences sounded as if they were spoken in real time, which would argue against the unnaturalness of time-compressed speech. These reports were however anecdotal and could therefore not be analysed statistically. It is important to emphasise that the lack of subjective effort for fast speech does not preclude that pupillometry findings can be attributed to effort, rather than novelty. Moore & Picou (2018) suggested that subjective ratings are driven by task difficulty and not necessarily effort alone. Therefore, in the current study, discrepancies between subjective and physiological effort might have arisen because there were no performance decrements for fast speech in dry.

3.5.3 Limitations

Given the small sample size, some of the visually observed differences might have reached significance with a higher statistical power. For instance, some interactions appeared to be informative such as the observation that the large difference in pupil dilation between fast and slow speech in dry disappeared in the retest session. Further experimentation with larger sample sizes will be necessary to investigate possible retest effects further.

3.5.4 Conclusion

Results suggest that listening effort exerted by older hearing-impaired listeners when processing fast speech was elevated at high intelligibility, as indicated by larger pupil dilation (H7, see Chapter 1, Section 1.6). It has to be noted however that H7 was only partly confirmed, since perceived effort ratings did

not show such an effect. It will therefore be necessary to investigate whether pupil dilation indeed indexed listening effort or whether increased arousal was due to other factors such as the novelty of the stimulus. Statistically, pupillometry results showed no effect of session, indicating that the effect of speaking rate on listening effort persisted in the retest session (H8). However, given the small sample size, it is possible that retest effects were left undetected; visually, the large difference in pupil dilation between fast and slow speech in dry was less apparent in the retest session. With regard to the research questions outlined in Chapter 1 (Section 1.6), this study could show that the effort exerted by older hearing-impaired listeners when processing fast speech was elevated even at high intelligibility (RQ3). However, more studies employing pupillometry across multiple sessions will be necessary to decide whether this effect was confined to the test session and to clarify why results diverged for perceived effort ratings. While time-compression in Chapter 2 resulted in overall high intelligibility for normal-hearing listeners, pupil dilation was larger compared to speech presented in quiet. Compared to results in the current study, it appears that while hearing-impaired listeners achieved overall high intelligibility, effort was elevated for fast speech, as well. However, normal-hearing listeners were able to tolerate more severe compression (37% of the original duration). Direct comparisons of the magnitude of pupil dilation in both studies are however not possible since older listeners typically exhibit lower pupil reactivity (Winn et al., 2015, 2018).

Results presented in this chapter showed both larger pupil dilation and higher perceived effort ratings when listeners processed reverberant speech, suggesting higher listening effort (H9). It has to be noted that pupillometry effects were only marginally significant when applying more robust model estimation. Since intelligibility was also overall lower under reverb, pupil dilation appeared to follow a U-shaped curve, with increasing pupil dilation associated with decreasing intelligibility (see Chapter 1, Figure 4). Neither pupillometry nor perceived effort ratings showed interactions with session, indicating that the effect of reverberation on listening effort persisted in the retest session (H10). In contrast to hypothesis H11, there was no evidence that the effect of fast speech

was amplified by reverberation. Results also did not indicate any benefit of dereverberation for listening effort under reverb, in contrast to hypothesis H12. With regard to the fourth research question (RQ4), this study showed that room acoustics affected listening effort when intelligibility was high, as indicated by pupillometry and perceived effort ratings.

Two main conclusions were drawn from the current study. First of all, this study demonstrated the feasibility of collecting pupillometry data in a complex acoustic setting, simulating both source degradations using time-compression and realistic room acoustics by means of higher-order ambisonics. One issue raised by results presented in Chapter 2 and the current Chapter 3 is that pupillometry research has focused largely on speech perception, following a standardised experimental protocol. Specifically, one experimental design has emerged in recent years, in which sentences are presented in a blocked fashion, requiring participants to listen and repeat. Only few studies have varied this standard protocol to investigate different aspects of communication such as speech production or turn-taking (Barthel & Sauppe, 2019; Papesh & Goldinger, 2012; Sauppe, 2017). It is likely that this one-sided approach is owed to the fact that there is considerable corporate and clinical interest to operationalise objective measures of listening effort (Pichora-Fuller et al., 2016; Wendt et al., 2020). However, while behavioural speech research has shifted to paradigms involving realistic communication settings, i.e., spontaneous speech with an interlocutor (Beechey et al., 2019; Hazan et al., 2018; Van Engen et al., 2010), no such attempt has been made for pupillometry research. One step in achieving this goal would be to complement the standardised pupillometry paradigm for speech perception with a paradigm for speech production. The next chapter therefore presents a study that applied pupillometry as a tool to investigate speaking effort, as a complement to listening effort.

Chapter 4: Measuring speaking effort during speech production in background noise

4.1 Introduction

In Chapter 2, I demonstrated that acoustic-phonetic differences between talkers affect intelligibility. Intelligibility-promoting acoustic-phonetic features bear similarity to acoustic modifications achieved by talkers when asked to produce clear speech (Krause & Braida, 2004; Smiljanic & Bradlow, 2009). Similarly, when producing speech in noise (i.e., Lombard speech), talkers enhance certain acoustic-phonetic parameters to increase audibility such as vocal intensity (e.g., Cooke & Lu, 2010). In Chapter 3, I showed that even in realistic acoustic environments, i.e., conditions at high intelligibility involving source and channel degradations, older hearing-impaired listeners exerted higher effort when processing fast speech. This finding implies that pupillometry as a measure of listening effort might be useful even when applied to such realistic communication settings. The current chapter focuses on another side of communication that is likely to be effortful, as well, when the acoustic environment is degraded: the process of speaking.

Studies aiming to recreate naturalistic communication in laboratory settings have employed a range of tasks that require participants to interact by collaboratively solving puzzles (Beechey et al., 2019; Cooke & Lu, 2010; Van Engen et al., 2010). Such tasks require participants to both listen and talk while typically being exposed to different acoustic environments, some more challenging than others. Subjective ratings have been used to quantify the amount of effort experienced by participants when communicating in such naturalistic environments (e.g., Beechey et al., 2019; Hazan et al., 2019). However, as discussed in Chapter 1, subjective ratings have methodological disadvantages. While pupillometry can be employed to objectively quantify effort in speech perception tasks, as shown in Chapter 2 and 3, only few studies have applied pupillometry to speech production. To the best of my knowledge, no study exists so far that

has applied pupillometry specifically to speech production under channel degradations. The current chapter addresses this gap and investigates whether in analogy to listening effort, pupillometry can be applied to quantify speaking effort, i.e., the effort exerted by talkers when producing speech in the presence of different masker types (RQ5, see Chapter 1, Section 1.6).

The current chapter first describes the speech modifications that typically occur under different types of masking, and the underlying mechanisms involved. Then, the relevant literature on movement-related pupil responses is discussed and the few speech production experiments conducted so far are examined. An experiment is presented that applied pupillometry to a speech production in noise task, aiming to replicate results from the speech perception literature.

4.1.1 Lombard effect

Speech production (of sentences) has been described as a multi-stage process (for a review see Ferreira, 2010). Message encoding involves two parallel complex processes, one that forms content representations and another one that forms structure representations. The two processes interact to form coherent semantic-syntactic representations that are subsequently articulated. In a similar way as the perceptual system is able to detect inconsistencies in an interlocutor's speech, it also monitors one's own production (Levelt, 1983). This "perceptual loop" allows to readjust one's production system based on auditory feedback.

When speaking in the presence of masking noise, talkers modify their speech to promote intelligibility (Lombard effect, Lombard, 1911; Brumm & Zollinger, 2011; Cooke et al., 2014). Talkers generally increase their vocal intensity to surpass the spectral energy of the masker (Cooke et al., 2014). This increase in vocal effort is also associated with other acoustic modifications such as a decreased spectral slope (Sundberg & Nordenberg, 2006) and increased fundamental frequency (Titze, 1989). Garnier et al. (2010) investigated the Lombard effect for two types of maskers (stationary and babble noise) at a range of in-

tensity levels (40 [quiet], 62, 70, 78, and 86 dB SPL). Participants were involved in a communicative task, i.e., they exchanged words with an interlocutor who was seated in front of them. As masker levels increased, talkers increased their overall vocal intensity, fundamental frequency and vowel duration. These effects were observed for both masker types. In comparison, when the task was conducted alone, without the presence of an interlocutor, acoustic adaptations were overall reduced. It was concluded that the Lombard effect is determined by both automatic regulation of vocal intensity and communicative intent. Similarly, the so-called diapix task (Van Engen et al., 2010) has been employed in a number of studies to elicit spontaneous speech with communicative intent (Hazan et al., 2016, 2018; Hazan & Baker, 2011). The task requires participants to spot differences between picture pairs while communicating with or without the presence of background noise. Hazan et al. (2018) compared acoustic adaptations by younger and older adults completing the diapix task. Differences between the two groups arose specifically in quiet, with slower speaking rates and lower vocal intensity observed for older adults, which was possibly linked to age-related changes in physiology. Interestingly, in noise, older adults with hearing loss showed increased vocal effort, as indicated by a simultaneous increase in fundamental frequency and vocal intensity. It was suggested that increased vocal effort might eventually result in higher fatigue; however, no fatigue measure was reported.

Cooke & Lu (2010) compared acoustic adaptations when talkers produced speech under different degrees of energetic and informational masking. Acoustic analyses were based on productions of digit words, recorded while participants solved sudoku puzzles alone (non-communicative) or together with an interlocutor (communicative). Speech production occurred either in quiet or in the presence of one of three maskers: stationary or fluctuating speech-shaped noise, or intelligible speech from a competing talker. All maskers were presented at 82 dB SPL. Compared to quiet, all maskers elicited Lombard effects that were characterised by an increase in vocal intensity and fundamental frequency, as well as a flatter spectral tilt, similar to results obtained by Garnier et al. (2010). Furthermore, the Lombard effect was generally

more pronounced in the communicative compared to the non-communicative task. For stationary noise, acoustic adaptations were larger compared to both fluctuating noise and competing-talker masker. It is known that at matched levels, stationary maskers have a higher energetic masking potential than fluctuating maskers (Festen & Plomp, 1990) and it has been suggested that the Lombard effect is proportional to the amount of energetic masking (Lu & Cooke, 2008).

In summary, when speech is produced in the presence of noise, talkers generally exert higher vocal effort, which is reflected in acoustic-phonetic measures such as higher vocal intensity, flatter spectral tilt, and higher fundamental frequency. Furthermore, when the masking potential is decreased due to energy fluctuations, or lower overall levels, acoustic adaptations are less pronounced. It has to be noted that despite acoustic similarities between Lombard speech and clear speech directed towards specific listener groups (Smiljanic & Bradlow, 2009), clear-speech modifications have been shown to vary in accordance with changing listener demands. For instance, Hazan & Baker (2011) showed that intensity and fundamental frequency, which are typically increased when speaking in noise, are in fact reduced when talkers direct their speech to an individual exposed to a cochlear implant simulation. Since such simulations compromise the spectral resolution of speech, increased loudness and pitch would not benefit intelligibility so that talkers adapt their production strategy to this constraint.

While most Lombard studies focused on behavioural outcome measures, fewer studies have been conducted to investigate cognitive aspects of producing speech in noise. Some neuroimaging studies have investigated the neural mechanism underlying the Lombard effect. It has been argued that greater energetic masking results in reduced auditory feedback, with subsequent acoustic modifications to compensate for this vocal-auditory mismatch. For instance, Christoffels et al. (2007) showed that activity in the superior temporal sulcus was increased in conditions with energetic masking. The superior temporal sulcus is involved in a number of processes underlying speech perception

(Hickok & Poeppel, 2007). Christoffels et al. (2007) argued that the increased activity reflects a mismatch between an efference copy of the speech motor commands and the actual speech input. Meekings et al. (2016) identified a confound in the original study: since increased vocal intensity when speaking in noise would counter the feedback mismatch, Christoffels et al. (2007) asked participants to suppress raising their voice. Meekings et al. (2016) suggested that this task instruction induced higher cognitive effort which was reflected in an elevated activity in the superior temporal sulcus. To address the issue, the authors compared neural responses when producing speech in noise at different levels of energetic and informational masking. The amount of informational content in the masker was parametrised by presenting white noise (low information level), speech-modulated noise, rotated speech and unmodified speech (high information level). While vocal intensity was highest in the (stationary) white noise condition (cf. Cooke & Lu, 2010), activation in the superior temporal sulcus was actually decreased compared to unmodified speech. Meekings et al. (2016) suggested that unattended intelligible speech interfered with speech production, with the lexical competition leading to an increase in cognitive effort. However, as vocal intensity was also higher for white noise, the results also support the hypothesis made by Christoffels et al. (2007): increasing vocal intensity reduces the vocal-auditory feedback mismatch, possibly lowering activation in the superior temporal sulcus.

For speech perception, the role of informational masking has been studied previously using pupillometry, as discussed in Chapter 1 of this thesis (Section 1.2.2.1). A competing talker generally elicits a larger pupil dilation than stationary or fluctuating noise (e.g., Koelewijn et al., 2012; Wendt et al., 2018). However, whether these findings can be replicated for speech production is unclear.

4.1.2 Pupillometry and speech production

While the pupil has been shown to dilate in response to motor planning and execution [Richer & Beatty (1985); see also Chapter 1, Section 1.5.1], only few studies to date have investigated the relationship between speech production and task-evoked pupil response. Papesh & Goldinger (2012) tasked participants to name high or low frequency words. Single words were visually presented for 500 ms which was followed by a variable delay period (250, 500, 1000 or 2000 ms). After the delay period, a tone was played that signalled one of two tasks: saying the previously displayed word out loud (high-pitch tone) or saying *blah* out loud (low-pitch tone). Pupil dilation peaks were extracted from several trial events: word presentation, delay, tone, preparation, response and post-response. Peaks were normalised by the average pupil size during a fixation period before the onset of the word presentation. Firstly, Papesh & Goldinger (2012) observed that peak dilation during the delay period increased with the duration of the delay. For the longest delay, there was an effect of word frequency, with larger peak dilation for low-frequency words. This effect was also present during response preparation, response and post-response. Word frequency also affected peak dilation in conditions where *blah* had to be produced instead of the displayed word. The authors suggested that the increased cognitive demands for retrieving low-frequency words were carried over to speech production.

Sauppe (2017) measured pupil dilation during a picture-naming speech production study. Specifically, sentences with active and passive voice were compared for two languages, German and Tagalog. While German is considered to have an asymmetrical voice system, i.e., active voice is unmarked while passive voice is marked, Tagalog is considered to have a symmetrical system, with morphologically marked voice forms. Participants were tasked to describe pictures using single sentences. Beforehand, pictures were rated for their tendency to elicit an active or passive description. This procedure was chosen to ensure that a similar number of sentences with both voice types was expected to be produced during the experiment. In accordance with the predictions, passive sentences

elicited a larger pupil dilation only in German, as they exhibit a more complex syntactic structure. In contrast, no difference between active and passive sentences was observed in Tagalog, for which syntactic complexity was similar across both voice forms.

In a more recent study, Barthel & Sauppe (2019) investigated pupil dilation during speech production in a turn-taking scenario. Participants had to describe objects shown in a picture in turn with a confederate interlocutor. Speech planning occurred either in silence, or in overlap with the interlocutor's speech. The two production contexts were created by presenting the interlocutor's speech either ending on a verb or an object; in the former condition, speech planning occurred in overlap, as the verb indicated a turn end, while in the latter condition, speech planning occurred in silence as all objects had to be named before the sentence could be produced by the participant. Barthel & Sauppe (2019) showed that pupil dilation was larger when speech planning occurred in overlap with the interlocutor's speech. It was suggested that turn-taking requires speakers to tolerate higher cognitive load while planning speech in overlap.

Taken together, the research on pupillometry during speech production conducted so far suggests a sensitivity to a range of linguistic manipulations (Barthel & Sauppe, 2019; Papesh & Goldinger, 2012; Sauppe, 2017), despite the strong influence of movement-related pupil responses (Hupé et al., 2009; Richer & Beatty, 1985). It is therefore reasonable to hypothesise that the well-known effect of informational masking on the pupil dilation (Koelewijn et al., 2012; Wendt et al., 2018) should be replicable in a speech production experiment. While acoustic modifications under different masker types have been studied extensively (Cooke et al., 2014), it has been suggested that measures of effort and fatigue could provide complementary information (Hazan et al., 2018), potentially reflecting speaking effort, analogous to listening effort. Thus, pupillometry has the potential to be used as an objective measure of speaking effort, complementing acoustic measures and subjective ratings when evaluating naturalistic communication environments (Beechey et al., 2019; Hazan et al., 2019).

4.1.3 Aims of the current study

In accordance with the research question described in Chapter 1 (RQ5, see Section 1.6), the current study was designed to investigate the feasibility of employing pupillometry in a speech production in noise paradigm to measure speaking effort. In a Lombard-style experiment, participants were asked to read and then produce sentences in quiet and under different types of masking: stationary and modulated speech-shaped noise (energetic masking) and a competing talker (informational masking). Pupil size was simultaneously recorded for the entire duration of the experiment.

I hypothesised that talkers' acoustic adaptations observed under channel degradations in previous studies would also be observed in the current study (H13). For instance, Lombard studies typically show that talkers exhibit higher fundamental frequency and vocal energy under stationary maskers compared to modulated maskers (e.g., Cooke & Lu, 2010). Furthermore, I hypothesised that previous findings in the speech perception literature with respect to channel degradations would be replicated with a speech production paradigm (H14). Specifically, higher effort as indicated by a larger pupil dilation was predicted for producing speech in a competing-talker background compared to a stationary-noise background (Koelewijn et al., 2012; Wendt et al., 2018).

4.2 Methods

4.2.1 Participants

Twenty-three normal-hearing native speakers of British English were recruited for the experiment [11 females; M_{age} = 23.8 (4.8) years; $range_{age}$: 19-36 years]. Recruitment and reimbursement was comparable to Chapter 2, following the guidelines of the Division of Psychology and Language Sciences at the University College London. Hearing ability was established by a standardised audiometric test at the beginning of the testing session. Participants had hear-

ing thresholds equal or better than 25 dB HL at all tested octave frequencies between 0.25 and 4 kHz.

4.2.2 Materials & Design

Speech production material consisted of a subset of 128 sentences (32 per condition) from the set of experimental sentences used in Chapter 2. Similar to Cooke & Lu (2010), participants were asked to produce speech in three different masking conditions and in quiet. To construct maskers, speech by a female talker was taken from the corpus recorded for Chapter 2. Stationary noise was created by generating speech-shaped noise with the same long-term average spectrum as the female talker. For modulated noise and competing-talker masker, sentences were concatenated to result in 48 masker streams with a duration of 16 s, sufficient to cover the maximum trial duration. It was ensured that none of the keywords in the sentences used for speech production appeared in the sentences used for masking. All maskers were rms-normalised. Four blocks of 32 trials (one per condition) were presented to each participant, with presentation order of conditions counterbalanced using a Latin square design. Each masker type was therefore presented in a separate block, similar to perception studies (e.g., Koelewijn et al., 2012; Wendt et al., 2018). Lists, sentences within each list, and masker streams (out of 48) were assigned randomly to each participant and condition.

4.2.3 Procedure

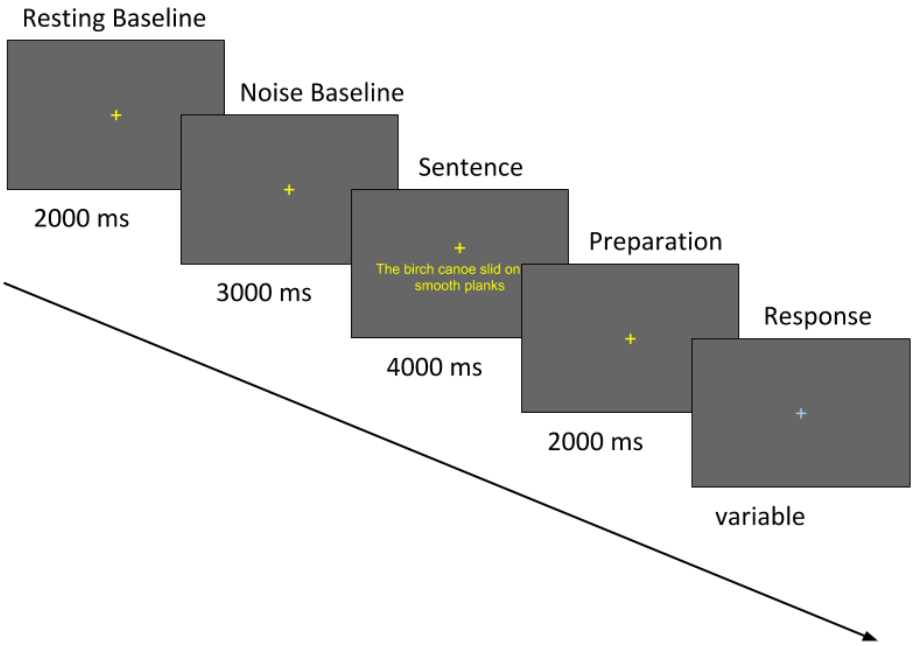

Figure 26: *Trial events with duration. Rectangles represent displays with central fixation cross.*

The same laboratory was used as in Chapter 2. The technical equipment and the procedure through which data was collected were therefore largely the same: participants wore headphones (Sennheiser HD 25 SP II) through which noise was played at 80 dB SPL. Only the forehead cushion of the previously used chin rest was used in the current study, to facilitate speaking, while minimising head movements at the same time. The light level was kept at 70 lux, but was adjusted individually if a too large or too small pupil size was observed. Before the start of the experiment, four practice trials were presented to familiarise participants with the task, one for each condition. See Figure 26 for an overview of all trial events. Each trial started with the appearance of a fixation cross at the centre of the screen. The resting baseline was recorded for 2 s, after which noise presentation started. After 3 s, the sentence appeared at the centre of screen and stayed there for 4 s. Font size was 30 pt with font type *Helvetica*. Participants were instructed to read the sentence and to return

to the fixation cross as soon as the sentence disappeared. After 2 s, the colour of the fixation cross changed to blue and participants were instructed to initiate their response at that point (response cue). The luminance of the two fixation crosses was kept equal as pilot data indicated a pupil light reflex with unmatched luminance. The trial was terminated by the experimenter after participants' verbal responses were recorded. Pupil size was tracked for another 2 s given the possibility that peaks followed task offset, which is commonly found in speech perception experiments (Winn et al., 2018). For instance, as shown in Chapter 2, pupil dilation tended to peak after sentence offset, and was delayed further for degraded speech, possibly reflecting slowed lexical access (see Figure 9).

4.2.4 Dependent variables and preprocessing

Acoustic analyses: Following the procedure presented in Chapter 2, recordings were first annotated and aligned using the Montreal forced aligner (McAuliffe et al., 2017). Annotations were manually checked and corrected if necessary, and sentence productions with errors were discarded. Mispronunciations, word omissions, and erroneous word order were counted as errors. Function word modifications were accepted if they resulted in a semantically plausible sentence (*the/a,his/her,these/those*; e.g. *Add the sum to the product of these three* vs. *Add the sum to the product of those three*). Similar to Chapter 2, mean energy, fundamental frequency, speaking rate and vowel space dispersion were obtained. The analysis window ranged from speech onset to speech offset. Additionally, speech-onset time, measured from the response cue, was analysed. For one participant, acoustic measurements, specifically vowel formant measurements, contained a large number of tracking errors due to technical issues with the microphone. This participant was therefore excluded from acoustic analyses. It has to be noted that since vowels were extracted from maximally 32 sentences (per participant and condition), the number of vowels was substantially smaller compared to Chapter 2 (average: a = 6.4, i = 15.0, ɔ = 9.8). After outlier exclusion (2 standard deviations above and below the mean), the

average number of vowel productions per participant and condition did not drastically change (average: a = 6.2, i = 13.6, ɔ = 9.2).

Pupillometry: Preprocessing was done in accordance with the procedure described in detail in Chapter 2. However, since trial events were not comparable to the other two experiments presented in this thesis (i.e., perception-only), normalisation and analysis differed. I followed the procedure described in Papesh & Goldinger (2012), who corrected pupil dilation to a baseline recorded before the visual presentation of the target word (in this study a full sentence). However, Papesh & Goldinger (2012) did not add masking noise. Since noise increases arousal and therefore baseline pupil size (Antikainen & Niemi, 1983), two ways of baseline correction were explored: (1) correcting pupil size to the resting state baseline recorded 1 s prior to noise onset, and (2) correcting pupil size to the baseline recorded with the presence of noise 1 s prior to visual sentence display (see Figure 26 and 31). Two time windows were considered for analysis, the preparation phase 2 s before the speech onset trigger, and the speaking phase with variable window size. The duration of the speaking phase was calculated as the time from the speech onset trigger to the end of the response (mean across trials), with an additional 2 s for possibly delayed peaks. Peak dilation and latency were then extracted from average pupil traces, in accordance with Chapter 2 and 3.

4.2.5 Statistical analysis

Similar to within-subjects analyses in Chapter 2 and 3, linear mixed models (LMMs) were fitted using *lme4* in R (Bates et al., 2015). In all models, random intercepts for talkers were allowed. LMMs were analysed using F-tests from the *lmerTest* package (Kuznetsova et al., 2017). With respect to acoustic measures, I expected enhancements across all analysed features under masking compared to quiet (H13): higher mean energy in mid-range frequencies and fundamental frequency, slower speaking rate and greater vowel space dispersion. Enhancements were also expected to be more pronounced under sta-

tionary noise given its higher masking potential. With respect to pupillometry (H14), I expected larger and more delayed peak dilation during speech production under competing-talker masking compared to other masker types and speech produced in quiet. Furthermore, I expected the effect of competing-talker masking to emerge already during speech planning, as reflected by a larger mean dilation in the preparation phase.

4.3 Results: acoustics

Incorrectly produced sentences were discarded (see Methods section) so that on average 28.2 out of 32 sentences were included per condition ($SD = 3.79$). Mixed effects logistic regression indicated that the number of included trials (vs. excluded trials) differed significantly between conditions [$\chi^2(3) = 54.05, p < 0.001$], with fewer trials included for the competing-talker condition compared to all other conditions ($p < 0.001$). The average number of included trials in the competing-talker condition was 25.9 ($SD = 5.0$).

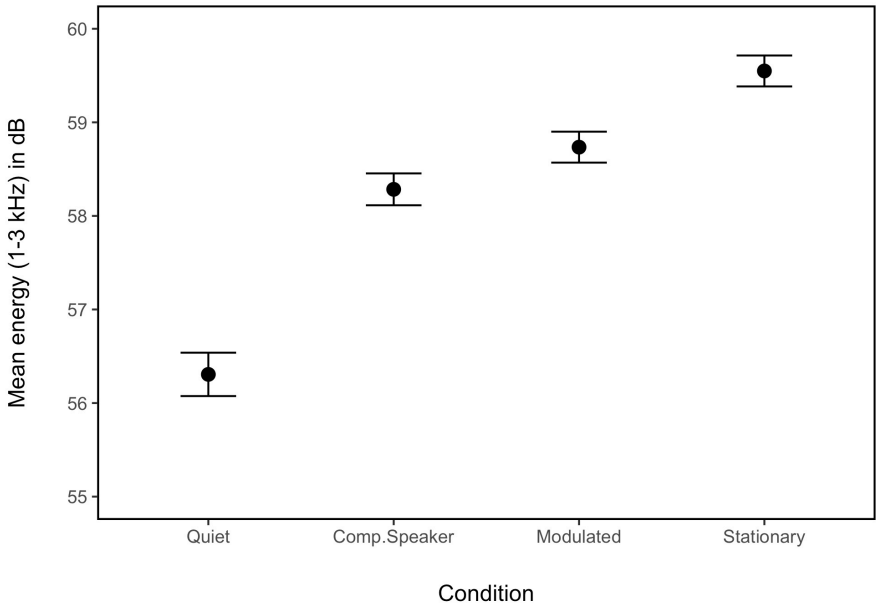

Figure 27: *Mean energy in the 1-3 kHz range in each condition. Points indicate means across participants and bars indicate one standard error around the mean.*

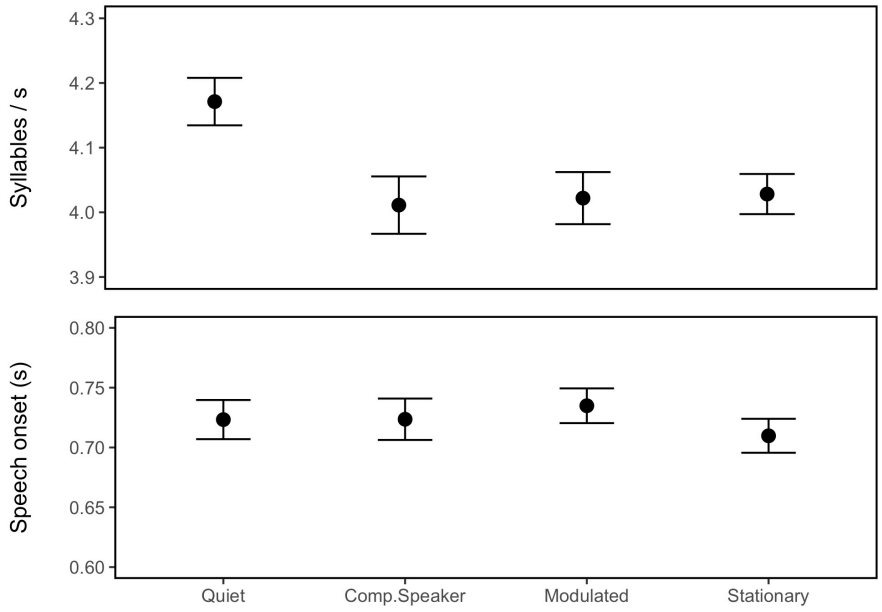

Figure 28: *Speaking rate and speech onset in each condition. Points indicate means across participants and bars indicate one standard error around the mean.*

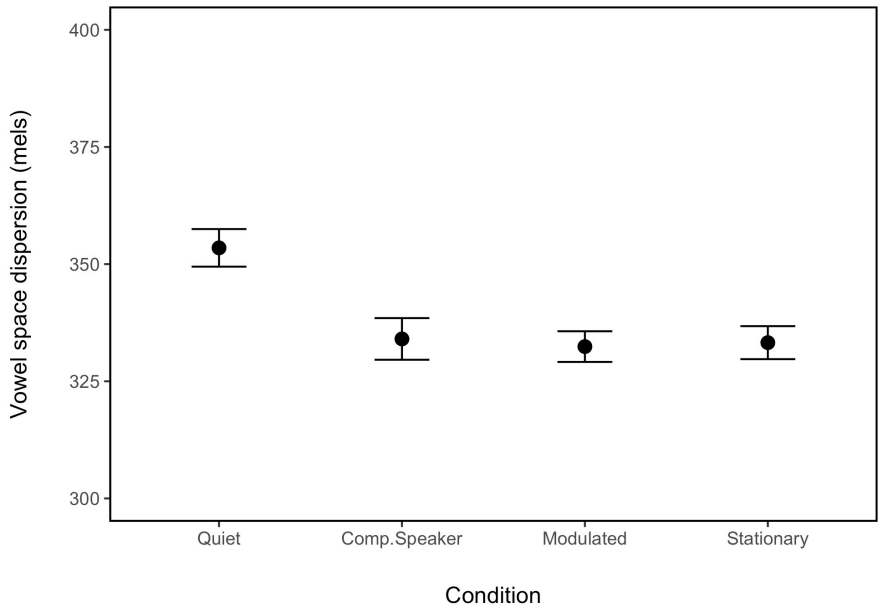

Figure 29: *Vowel space dispersion in each condition. Points indicate means across participants and bars indicate one standard error around the mean.*

The effect of masker type on acoustic measures was investigated with linear mixed models, including masker type as fixed effect. For mean energy (Figure 27), there was a main effect of condition [$F(3, 63) = 55.28, p < 0.001$]. Pairwise comparisons showed that mean energy was lower in quiet compared to all other

conditions ($p < 0.001$). Mean energy was also lower under competing-talker masking ($p < 0.001$) and modulated noise ($p = 0.017$) compared to stationary noise. For speaking rate (Figure 28), there was a main effect of masker type [$F(3, 63) = 3.89, p = 0.013$], with faster rates in quiet compared to competing-talker masker ($p = 0.03$) and modulated noise ($p = 0.047$). There was no effect of masker type on speech onset (Figure 28) [$F(3, 63) = 0.43, p = 0.73$]. For vowel space dispersion (Figure 29), there was a main effect of condition [$F(3, 63) = 8.35, p < 0.001$], with larger dispersion in quiet compared to all masker types ($p = 0.001, p = 0.001, p < 0.001$ for competing talker, modulated and stationary noise, respectively).

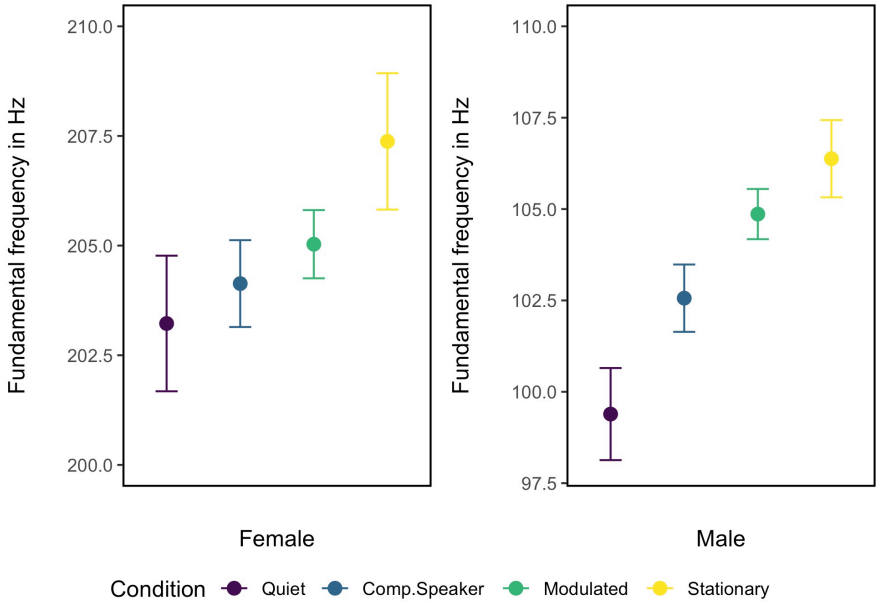

Figure 30: *Fundamental frequency in each condition. Points indicate means across participants and bars indicate one standard error around the mean.*

Due to known differences between female and male speakers, fundamental frequency (f_0, Figure 30) was analysed with a linear mixed model including fixed effects of masker type and gender. As expected, there was a main effect of gender [$F(1, 20) = 218.87, p < 0.001$], with higher f_0 observed for female speakers. Furthermore, there was a main effect of condition [$F(3, 60) = 8.60, p < 0.001$]. Pairwise comparisons showed that f_0 was lower in quiet compared to modulated ($p = 0.014$) and stationary noise ($p < 0.001$). Additionally,

f_0 was lower with the competing-talker masker compared to stationary noise ($p = 0.018$).

In summary, increased vocal effort, as reflected by higher mean energy and fundamental frequency, was observed for speech produced under masking, compared to speech produced in quiet. Furthermore, the effect was larger for stationary noise. Faster speaking rates and greater vowel space dispersion was observed for speech produced in quiet, while speech onset was not significantly different between conditions. It has to be noted that all maskers were based on recordings from a female talker. Boxplots showing distributions of acoustic measures by gender are provided in the appendix.

4.4 Results: pupillometry

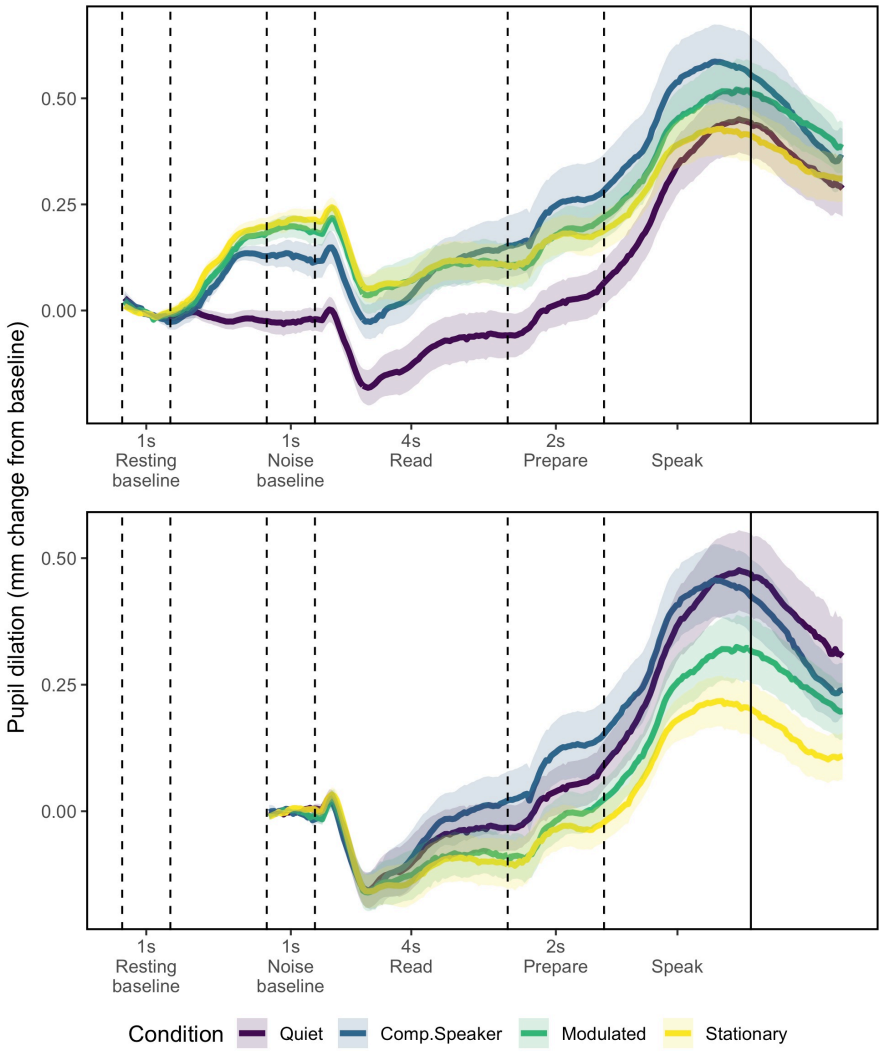

Figure 31: *Average pupil traces with trial events. Comparison of two baseline correction methods, resting state baseline (upper panel) and noise baseline (lower panel). The solid line indicates average speech offset (3.05 s). Coloured bands indicate the standard error.*

Figure 31 shows pupil traces for the entire trial, averaged across all participants. It can be seen that noise induced an increase in the pupil baseline, with respect to the resting baseline. This was reflected in a main effect of condition [$F(3, 66) = 26.18, p < 0.001$], with larger baseline in noise than in quiet ($p < 0.001$). Pairwise comparisons also indicated a larger baseline under stationary noise compared to competing-talker masking ($p = 0.04$). This noise-induced change in pupil size was corrected for by using the average pupil size

1s prior to sentence display. The following analyses will investigate effects of masker type separately for both correction methods.

4.4.1 Resting baseline

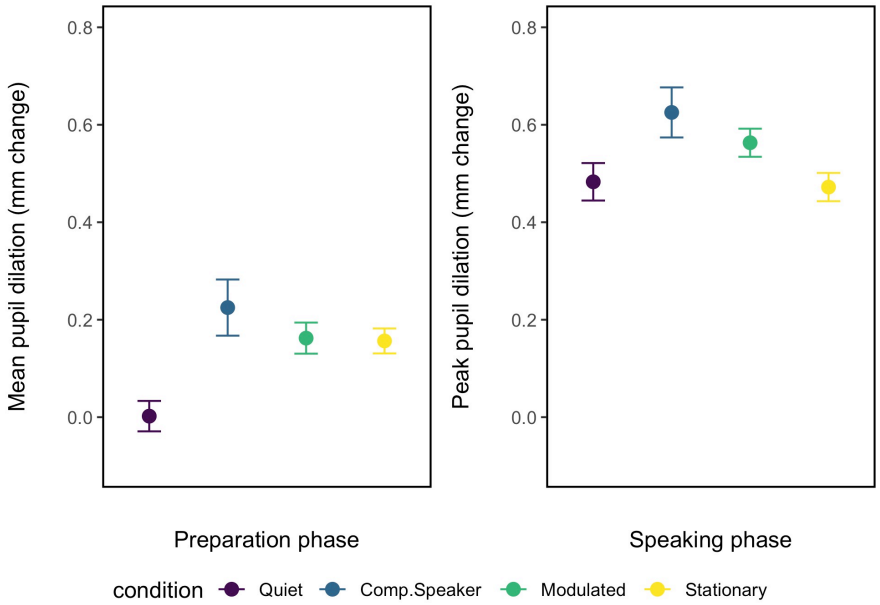

Figure 32: *Mean and peak dilation during the preparation and speaking phase, respectively (resting baseline). Points indicate means across participants and bars indicate one standard error around the mean.*

First, I analysed pupil dilation during the preparation phase (see Figure 32). Since peaks occurred during the speaking phase (see Figure 31), mean dilation was used instead. Linear mixed models were analysed with masker type as fixed effect. When correcting to resting baseline, there was a main effect of condition [$F(3, 66) = 6.02, p = 0.001$]. Pairwise comparisons indicated that mean dilation under all masker types was significantly larger than mean dilation in quiet (competing talker: $p < 0.001$, stationary noise: $p = 0.04$, modulated noise: $p = 0.03$). I then analysed peak pupil dilation during the speaking phase (see Figure 32). There was a main effect of condition [$F(3, 66) = 3.59, p = 0.02$]. Pairwise comparisons indicated larger peak dilation under competing-talker masking than stationary noise ($p = 0.04$). There was no significant difference between conditions for peak latency [$F(3, 66) = 1.61, p = 0.20$].

4.4.2 Noise baseline

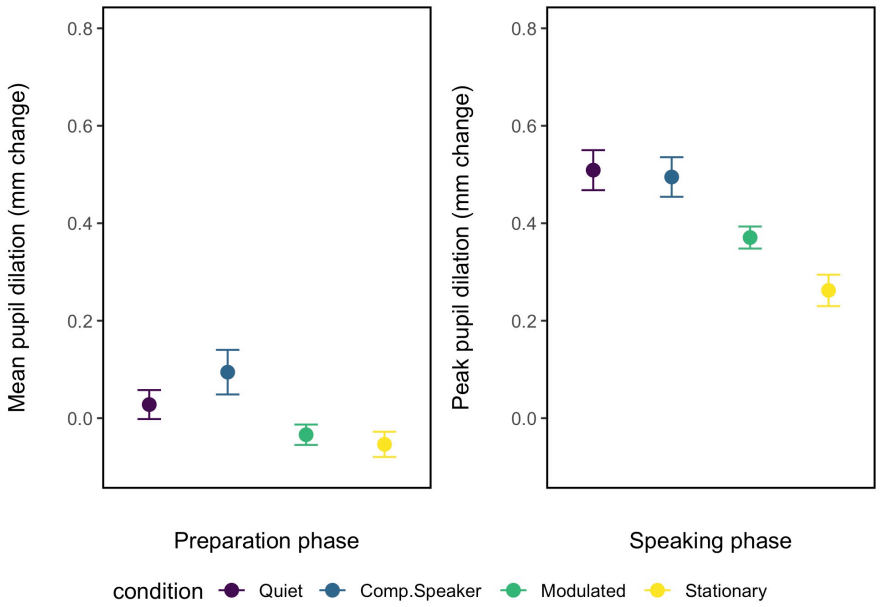

Figure 33: *Mean and peak dilation during the preparation and speaking phase, respectively (noise baseline). Points indicate means across participants and bars indicate one standard error around the mean.*

When correcting to noise baseline (see Figure 33), there was also a main effect of condition in the preparation phase [$F(3, 66) = 4.39, p = 0.007$]: mean dilation under competing-talker masking was larger than mean dilation under stationary noise ($p = 0.01$) and modulated noise ($p = 0.04$). During the speaking phase, there was also a main effect of condition [$F(3, 66) = 11.02, p < 0.001$]. Pairwise comparisons showed that peak dilation was larger under competing-talker masking than stationary noise ($p < 0.001$). Peak dilation was also larger in quiet compared to both stationary noise ($p < 0.001$) and modulated noise ($p = 0.03$). There was no significant difference between conditions for peak latency [$F(3, 66) = 1.61, p = 0.20$].

4.5 Discussion

In a speech production experiment combined with pupillometry, participants read and produced sentences in quiet and under different types of masking (stationary and modulated noise, and with a competing talker). While acoustic adaptations (Lombard effect) were expected in conditions with higher energetic masking, higher cognitive (or speaking) effort, as indicated by the pupil dilation response, was expected in the presence of a competing talker.

4.5.1 Lombard effect

Mean energy in the 1-3 kHz range was increased for all masking conditions, but was also significantly higher under stationary masking. This finding is consistent with Cooke & Lu (2010) and expected as stationary maskers have a higher energetic masking potential than fluctuating maskers (Festen & Plomp, 1990). Similarly, fundamental frequency was increased under masking, and even more so under stationary masking. Both results are consistent with the hypothesis that vocal effort is linked to both a decrease in spectrum slope (Sundberg & Nordenberg, 2006) and an increase in fundamental frequency (Titze, 1989). Speaking rate was slower under masking compared to quiet. This result is consistent with other Lombard studies (Aubanel et al., 2011; Webster & Klumpp, 1962) and clear speech in general, even though speaking slowly does usually not contribute substantially to the clear-speech intelligibility benefit (Cooke et al., 2014; Picheny et al., 1986). On the other hand, the experimental procedure might have played an important role in this finding. Since the reading task was temporally separated from the speaking task, due to constraints of the pupillometry procedure, sentences had to be stored briefly in short-term memory. Lexical retrieval might then have been slowed under masking, resulting in slower speaking rates. However, with this reasoning, an even slower speaking rate would have been expected for competing-talker masking, for which perceptual interference is more severe. For instance, Aubanel et al. (2011) observed that speaking rates decreased with increasing number of competing talkers, provid-

ing greater perceptual interference. Perceptual interference was possibly observed in the current study, reflected in the number of sentences with pronunciation errors, which was higher for competing-talker masking. Even though it was ensured that target keywords were not contained in the masker, even phonologically similar distractor words have been linked to a larger number of speech production errors (Saito & Baddeley, 2004).

Surprisingly, vowel space dispersion was greater in quiet compared to all masking conditions, suggesting hyperarticulation. Even though it appears counterintuitive that vowel space dispersion was smaller in noise, this finding has been observed previously: Perkell et al. (2007) asked participants to produce words containing the vowels /i/, /u/, /ɛ/ and /æ/, while speech-shaped noise was presented at different levels (highest level: 95 dB SPL). Vowel contrasts were measured as the distances between vowel pairs in the $F_1 - F_2$ space. Perkell et al. (2007) observed that vowel contrasts initially increased with increasing noise level, as would be expected in light of the Lombard effect. However, when noise levels increased further, vowel contrasts actually decreased compared to conditions without noise. It was argued that the perception of vowel contrasts was diminished at higher noise levels, which implicitly interfered with production given the lack of auditory feedback. In the current study, this effect was possibly even further amplified by the use of headphones without auditory feedback mechanism. It also has to be noted that acoustic modifications are usually amplified when headphones are used instead of loudspeakers (Garnier et al., 2010). Using closed headphones has been suggested to induce a Lombard effect even in quiet, as auditory feedback is attenuated (Garnier et al., 2010). On the other hand, the average vowel space dispersion in quiet in the current study (353 mels) was similar to the average vowel dispersion for talkers in Chapter 2 (377 mels), who were recorded in quiet and without the use of headphones. This observation suggests that the lack of auditory feedback in the current study contributed to smaller vowel space dispersion under masking, in accordance with Perkell et al. (2007). When auditory feedback is provided, an increase in vocal intensity is usually accompanied by an expanded vowel space (Koenig & Fuchs, 2019). Studies that manipulated auditory feedback

explicitly observed that talkers counter acoustic perturbations by up- or down-regulating respective features; for instance, Patel et al. (2011) observed that talkers increase their fundamental frequency and intensity to compensate for downward-shifted fundamental frequency (pitch perturbation). It is possible that diminished vowel perception under masking in the current study led to compensation by increasing fundamental frequency and energy in mid-range frequencies.

4.5.2 Pupillometry

The general curvature of the pupil dilation function was in accordance with previous speech production studies (Papesh & Goldinger, 2012; Sauppe, 2017). Pupil dilation increased during speech preparation and peaked during speech production. To account for trial-to-trial fluctuations in pupil size (Mathôt et al., 2018), I attempted two types of baseline correction, with a resting and noise-exposure phase as reference, respectively. In speech perception studies, baselines are usually obtained from pre-trial measurements (see Chapter 2 and 3 in this thesis, and Winn et al., 2018). Since noise increases arousal and therefore baseline pupil size (Antikainen & Niemi, 1983), the baseline used for correction usually incorporates the effect of background noise which starts prior to stimulus presentation. However, Zekveld et al. (2010) showed that for a relatively small range in levels (55-63 dB SPL) noise levels did not affect baseline pupil size.

In the current study, the presentation of any masker type significantly increased baseline pupil size in comparison to quiet, as expected given the large level differences (80 dB SPL). It is therefore not surprising that after accounting for the noise-induced increase in pupil size, pupil dilation during speech production was found to be larger in quiet compared to both stationary and fluctuating noise. Interestingly, there was no significant difference between pupil dilation during speech production in quiet and under competing-talker masking, suggesting that the level differences alone were not the sole contributors to pupil

dilation differences. However, the fact that baseline pupil size was by default larger under noise makes it difficult to interpret any differences in pupil dilation between quiet and masker conditions.

During speech preparation, pupil dilation was increased for all masker types compared to quiet when resting baseline correction was applied. This effect was likely driven by heightened noise-induced arousal, as it disappeared when noise baseline correction was applied. When factoring in the effect of noise on the baseline pupil size, mean dilation during speech preparation was larger for competing-talker masking than for both stationary and modulated noise. Since the noise-induced baseline was also larger in these two conditions, baseline correction might have possibly discounted a task-induced increase in pupil size. On the other hand, during speech production, pupil dilation under competing-talker masking was significantly increased over stationary noise, both with and without correcting for noise-induced arousal levels. This finding is consistent with speech perception studies, reflecting higher cognitive effort with a perceptually interfering signal (Kidd, Mason, et al., 2008; Kidd, Best, et al., 2008; Koelewijn et al., 2012; Wendt et al., 2018).

The magnitude of the peak pupil dilation (in mm) in the competing-talker condition was large when measured as change from the resting baseline ($M = 0.62$ mm); in comparison, maximum pupil dilation during speech production was around 0.5 mm in Papesh & Goldinger (2012). However, the magnitude of the pupil dilation in the current study was probably overall amplified by an initial increase in arousal due to masker levels. In fact, when factoring in this initial noise baseline, average peak dilation decreased ($M = 0.50$ mm), which was in line with Papesh & Goldinger (2012). Pupillometry studies using masking noise typically factor in such initial noise-induced baseline levels; comparing the magnitude of the current results to speech perception and other cognitive tasks, shows that the average magnitude of dilation corresponds to a 'hard task' such as processing 4-channel noise-vocoded speech (cf. Winn et al., 2015).

Interestingly, even without the presence of a masker, speech production elicited a very large pupil dilation response ($M = 0.48$ mm). This response likely stems

from the generally large contribution of motor planning and execution (Hupé et al., 2009; McCloy et al., 2016; Richer & Beatty, 1985). It is therefore questionable whether the pupil dilation magnitude in speech production tasks can be directly compared to the magnitude in speech perception or other cognitive tasks. One possibility to allow comparison between speech perception and production tasks would be to measure individuals' pupil dilation response to a simplified articulation task, i.e., without involving lexical or semantic processes. For instance, participants could be asked to produce individual phonemes or nonsense syllables (see Papesh & Goldinger, 2012). This (speech) motor-evoked pupil response could then be used to normalise the task-evoked pupil response.

The initial increase in arousal can be considered a confound of the current study. Several alternatives are therefore conceivable for the current experimental design. First of all, noise could be presented continuously to allow pupil size to return to resting-state levels. This option was not chosen in the current study as it was expected that continuous noise would lead to an early onset of fatigue given the duration of the experiment (~70 minutes). It might be sufficient to provide a longer pre-trial noise exposure ($>$ 3 s used in this study). Another option could be to present maskers at overall lower levels that are also able to induce Lombard effects (see Garnier et al., 2010). It is possible that the smaller the influence of other factors such as noise, the more sensitive pupil dilation would be to the task manipulation.

4.5.3 Limitations

Maskers were presented continuously throughout a trial, i.e., they started before the visual sentence presentation and ended after speech offset. Perceptual interference, as suggested by larger pupil dilation for competing-talker masking, could therefore have occurred at multiple stages of the speech production process. It is therefore possible that a larger pupil dilation reflected increased demands on reading, rehearsal and articulation. The choice for continuous masker presentation was ecological as this corresponds to everyday speaking

environments.

4.6 Conclusion

Results showed that in accordance with speech perception studies, pupil dilation was larger when speech was produced in the presence of a competing talker, compared to a purely energetic stationary noise masker. This observation was made despite the fact that stationary noise elicited larger acoustic modifications such as increased energy in mid-range frequencies, in accordance with other Lombard studies (e.g., Cooke & Lu, 2010). Therefore, it can be concluded that larger pupil dilation during speech production reflected higher cognitive rather than physical demands, despite the relatively large contribution of motor planning and execution on the pupil dilation (e.g., Richer & Beatty, 1985). The results therefore confirmed my hypotheses as outlined in Chapter 1 (H13 and H14, see Section 1.6). With respect to research question RQ5, results presented in the current chapter suggest that pupillometry can indeed be applied to measure speaking effort. Similar to the definition of listening effort (McGarrigle et al., 2014, p. 434), speaking effort indicates the mental exertion required to convey an auditory message.

Results of the current study showed that when producing speech in noise, talkers modify acoustic-phonetic parameters that have been shown to promote intelligibility in Chapter 2, such as energy in mid-range frequencies and speaking rate. However, results also emphasise that talker acoustics are dynamic, with modifications depending on the acoustic environment.

Findings of both Chapter 3 and the current Chapter 4 indicate that pupillometry can be applied to more realistic communication settings. In particular, Chapter 3 showed that pupillometry can reveal elevated listening effort for older hearing-impaired listeners when processing fast speech, even at high intelligibility. Furthermore, the current Chapter 4 showed that pupillometry can measure speaking effort, i.e., higher cognitive demands when speaking in the presence of a competing talker. Taken together, these results have implications

not only for current models of communication, but also for the design of future studies. Specifically, the studies presented in this thesis are predominantly first attempts at extending the use of pupillometry to more realistic communication, aiming to incorporate the role of the talker in both speech perception and production. Implications and recommendations for future studies will therefore be outlined in the next chapter, alongside discussions of the current results with respect to models of communication.

Chapter 5: General discussion

The overarching research question of this thesis can be stated as follows:

- How do acoustic degradations affect individuals at cognitive and acoustic levels, in all communicative roles, i.e., during listening and speaking?

The aim of this thesis was therefore to expand on existing research and to incorporate factors related to the talker in models of effort. On the one hand, this aim demanded consideration of talker-related factors in speech perception. On the other hand, it required an account of speaking effort, i.e., the mental exertion required by talkers to convey an auditory message. To address the overarching research question, I conducted three studies, presented in Chapters 2-4 respectively, each aiming to answer one or more sub-questions. In Chapter 2, I investigated the interaction between source and channel degradations and their effect on both intelligibility (RQ1) and effort (RQ2). In Chapter 3, I focused specifically on speaking rate as simulated using time-compression, and investigated whether older hearing-impaired listeners exert more effort when processing fast speech, even at high intelligibility (RQ3). Moreover, I investigated the effect of room acoustics on listening effort (RQ4). Finally, in Chapter 4, I applied pupillometry in a speech production study, to investigate whether pupil dilation would index speaking effort, similar to the concept of listening effort used in speech perception research (RQ5).

5.1 Source and channel degradations and their effect on intelligibility and effort (RQ1 and RQ2)

Results presented in Chapter 2 showed that acoustic-phonetic differences between talkers predicted intelligibility under different types of channel degradations: masking, noise-vocoding and time-compression. At the same time, correlation analyses indicated that talkers who were more intelligible under noise-vocoding were also more intelligible under masking and time-

compression. Taken together, these results confirmed hypotheses H1 and H2. With respect to RQ1, results presented in Chapter 2 therefore showed that source degradations affecting intelligibility interacted with channel degradations.

Intelligibility under masking was driven by both higher energy in mid-range frequencies, as predicted by the speech intelligibility index, and by greater vowel space dispersion. Similarly, larger vowel space dispersion benefited intelligibility of noise-vocoded speech. Greater vowel spaces have been associated with higher (intrinsic) intelligibility in a previous study using similar sentence material (Bradlow et al., 1996). Vowel space expansion is linked to hyperarticulated speech within the hyper-hypo model of speech production (H&H, Lindblom, 1990). However, while Chapter 2 investigated sentences recorded by talkers in isolation, H&H specifically addresses the interaction between talker and listener, i.e., it predicts that talkers adjust their production depending on listener constraints such as hearing impairment (Cooke et al., 2014). For instance, Granlund et al. (2018) showed that children communicating with hearing-impaired peers adjust their speech by expanding their vowel spaces and increasing energy in mid-range frequencies (amongst other modifications). When asked to produce clear speech in the absence of an interlocutor, talkers typically show similar acoustic modifications (Smiljanic & Bradlow, 2009). However, it has been demonstrated that these changes are larger compared to communicative situations (Hazan & Baker, 2011). In light of H&H, it has been argued that in the presence of an interlocutor, acoustic modifications fluctuate with listener demands, i.e., when listener feedback signals successful comprehension, talkers will adjust their speech to a smaller extent. Subsequently, average acoustic modifications will be smaller than when talkers are specifically asked to speak clearly.

Acoustic-phonetic differences observed in Chapter 2 were neither induced by the experimenter nor by listener demands, as talker and experimenter were seated in separate rooms. However, while distinctions are often made between ‘intrinsically’ and ‘deliberately’ clear speech (e.g., Hazan & Markham, 2004), it

is plausible that most acoustic differences, other than anatomic-physiological, are to some extent deliberate. Therefore, when asked to read speech out loud in quiet, it is possible that some talkers deliberately decided to speak more clearly than others.

Vowel space dispersion was not predictive of intelligibility under time-compression. However, it has to be noted that the time-compression rate (37% of the original duration) resulted in overall high intelligibility (~80%), which was significantly higher than intelligibility under both noise-vocoding and masking. It is conceivable that the relevance of specific acoustic-phonetic features changes with decreasing intelligibility. For instance, a recent study hinted towards a possible effect of vowel space on intelligibility of time-compressed speech (Johnson et al., 2020); however, in their study, time-compression (66.7% of the original duration) resulted in intelligibility levels below 60%. While the time-compression rate was less severe than here, Johnson et al. (2020) simultaneously presented babble noise at 0 dB SNR. The additional noise could have been the driving factor explaining the relevance of greater vowel spaces, similar to findings for masked speech in the current thesis (Chapter 2). I showed that speaking rate was associated with intelligibility of time-compressed speech. Speaking rate was also strongly correlated with vowel duration. It is therefore plausible that time-compression was even more detrimental for initially shorter segments. For instance, durational cues are relevant for the distinction between short and long vowels (Klatt, 1976). The detrimental effect of time-compression on intelligibility was likely due to the removal of acoustic information given the high compression rate (Gordon-Salant & Fitzgibbons, 2001; Schneider et al., 2005). Since uniform time-compression as used in Chapter 2 shortens vowels and consonants to the same extent, it is not clear which acoustic cues were more affected by shortening. Previously, it has been shown that selective time-compression of consonants is more detrimental to intelligibility (Gordon-Salant & Fitzgibbons, 2001). Besides acoustic degradation, the increased information rate under time-compression might have contributed to the intelligibility reduction. Oscillation models of speech perception (e.g., TEMPO, Ghitza, 2011) predict

disruptions in syllabic parsing at syllable rates outside the theta range (4-10 Hz). Indeed, as syllable rates were on average 3.8 syllables per second, time-compression by 37% resulted in rates of 10.3 syllables per second, which is just outside the range for theta. The drop in intelligibility can therefore not only be explained by acoustic degradation, but also by disrupted temporal parsing, i.e., the diminished ability to track and resolve syllabic units (Ghitza, 2011). Syllable rates would have to be within the theta range to enable tracking by neural oscillators and subsequent decoding by memory processes.

Results presented in Chapter 2 showed that listeners adapted to noise-vocoded and time-compressed speech. However, contrary to hypothesis H3, talker acoustics did not predict adaptation. Therefore, while overall intelligibility was determined to some extent by acoustic differences between talkers, adaptation was not. It is possible that variances observed in individual adaptation slopes were rather attributable to listeners' cognitive abilities, which has been shown in a range of studies employing degraded (accented) speech (Adank & Janse, 2010; Banks et al., 2015; Janse & Adank, 2012; McLaughlin et al., 2018). Furthermore, it is conceivable that despite measurable acoustic differences between talkers, listeners were familiar with such talker characteristics as they shared the same native language background. In fact, adaptation has been shown to remain intact even in the face of talker change (Dupoux & Green, 1997) or when adjusting parameters such as time-compression rate (Adank & Janse, 2009; Golomb et al., 2007). With respect to physiology, results presented in Chapter 2 did not show that baseline pupil size changes followed the observed adaptation trends, contrary to hypothesis H5. While adaptation was seen for noise-vocoded and time-compressed speech, baseline pupil size was overall smaller for noise-vocoded speech and speech in quiet. In fact, a decrease in arousal indicated by decreasing baseline pupil size has been associated with elevated task demands (Ayasse & Wingfield, 2020). At the same time, this explanation would not extend to speech in quiet, which resulted in high intelligibility. Here, overall lower arousal levels might have indicated less sustained attention (see Wagner et al., 2019). The difficulty in interpreting findings from Chapter 2 related to baseline pupil size can be

attributed to the fact that intelligibility differed between conditions. This limitation is discussed further below (see Section 5.1.1).

Pupillometry results presented in Chapter 2 showed that while peak pupil dilation was larger and more delayed for degraded speech compared to speech in quiet (H4), acoustic-phonetic features did not modulate this effect, in contrast to hypothesis H6. With regard to RQ2, results presented in Chapter 2 therefore suggest that channel degradations that affect listening effort, as measured by pupil dilation, did not interact with source degradations. However, limitations have to be taken into account, such as large individual differences in the pupil dilation measure and overall intelligibility levels (see also Section 5.1.1). On the other hand, results showed that while intelligibility was relatively high under time-compression (~80%), pupil dilation was significantly larger compared to quiet. Since time-compression resulted in very fast speaking rates, it is remarkable that listeners were at all able to yield such high intelligibility. The larger pupil dilation was likely to reflect the "extra effort" required to maintain relatively high intelligibility at fast speaking rates (Lemke & Besser, 2016), as suggested by models of listening effort (Pichora-Fuller et al., 2016).

Furthermore, time-compressed speech was also associated with delayed peaks, when measured from sentence offset. The Ease of Language Understanding model (ELU) predicts delayed lexical access for degraded speech input (Rönnberg et al., 2008, 2013). Indeed, delayed peak dilation was observed for all degradation types in Chapter 2. Interestingly, peaks were further delayed for time-compressed speech, despite overall higher intelligibility compared to both noise-vocoded and masked speech. It is possible that a combination of two processes led to latency differences between time-compressed speech and other types of degraded speech: (1) delayed lexical access due to acoustic degradations and (2) increased information rate.

Another possibility is that peak dilation was simply delayed because it took a certain amount of time for the pupil dilation to reach its peak (Winn et al., 2018); in fact, when pupil traces were aligned to sentence onset, peak dilation occurred earliest for time-compressed speech. To further investigate this ques-

tion, an experiment would have to be designed in which overall sentence duration could be maintained while speech segments are time-compressed. However, since insertion of pauses has been shown to restore intelligibility partly (Ghitza & Greenberg, 2009) such paradigms would be confounded.

5.1.1 Limitations and future directions

5.1.1.1 Intelligibility

Fixed acoustic parameters (e.g., time-compression ratio) allowed intelligibility to vary between conditions. Despite the observation that acoustic-phonetic differences between talkers affected intelligibility in each listening condition, it is possible that these results depended on the baseline intelligibility, as determined by the chosen acoustic parameters. For instance, the relevance of talkers' vowel spaces on intelligibility might be higher at overall lower intelligibility levels (see Johnson et al., 2020). However, Johnson et al. (2020) presented time-compressed speech in noise which could have influenced their results. For instance, listeners rely more on temporal fine structure in the presence of background noise (Moore, 2008) while time-compression at mild rates affects mostly the temporal envelope. This potential differential effect of time-compression and masking on talker intelligibility should be investigated further in future studies. Future studies should also consider the influence of different baseline intelligibility levels; for instance, speech by different talkers could be presented at two time-compression ratios (low and high).

5.1.1.2 Pupillometry

Contrary to intelligibility results, pupillometry was not a sensitive enough measure to capture acoustic-phonetic talker differences. As physiological measure, pupil dilation is susceptible to listeners' individual differences (Winn et al., 2018). As described in Chapter 2, different talkers were presented to different listeners (between-subjects design) so that listeners' physiological differences might have contributed to the null result. Furthermore, since intelligibility was

determined by chosen acoustic parameters, individual speech-perception skills also likely contributed to greater variability. Many pupillometry studies specifically control for intelligibility across listeners and conditions by employing adaptive procedures. For instance, Borghini & Hazan (2020) investigated differences in listening effort induced by conversational and instructed clear speech. Intelligibility levels were individually adjusted by varying signal-to-noise ratios per condition and speech style to result in 50% intelligibility. Results showed that listeners not only tolerated more noise when processing clear speech, but also likely exhibited less effort when doing so, as indexed by a smaller pupil dilation. While adaptive procedures add testing time and therefore only allow the consideration of few experimental manipulations, they might reduce variability across listeners and conditions, allowing for higher statistical power.

On the other hand, by adapting signal-to-noise ratios separately for each testing condition, initial noise levels differ between conditions. Since baseline correction is usually applied with the noise-induced baseline pupil size as reference (e.g., Koelewijn et al., 2012; Borghini & Hazan, 2020), conditions with lower signal-to-noise ratios (i.e., higher initial noise levels) might potentially be biased by baseline correction. This issue is less problematic when the task-evoked pupil dilation is expected to show similar patterns as behavioural measures. For instance, Koelewijn et al. (2012) observed higher noise levels (lower signal-to-noise ratios) and larger pupil dilation for a competing-talker masker compared to stationary noise. Subtracting a larger noise baseline discounts the overall larger task-evoked pupil dilation so that results are eventually more conservative (cf. Chapter 4). However, problematic cases are conditions that result in lower signal-to-noise ratios (i.e., high noise levels), but smaller pupil dilation. Baseline correction might then cause the pupil dilation to appear smaller, even though this is not necessarily the case. Future studies with sentences presented at different noise levels should therefore consider a range of baseline correction methods (cf. Chapter 4) to inform about possible biases introduced by the method itself.

5.2 Pupillometry and older hearing-impaired listeners: fast speech and room acoustics (RQ3 and RQ4)

Results presented in Chapter 3 showed that listening effort, as likely indexed by pupil dilation, was higher when listeners processed fast speech than slow speech, even in the absence of reverberation when intelligibility was near ceiling. Listeners were older hearing-impaired individuals who were fitted hearing aids to compensate for frequency-specific gain loss. Results therefore confirmed hypotheses H7 and H8 and answered research question RQ3 by showing that speaking rate did indeed affect listening effort even when intelligibility was high. However, certain limitations of the study have to be taken into account, as discussed below (Section 5.2.1).

The high intelligibility levels achieved by listeners and the relatively mild time-compression rate (70%, i.e., 5.54 syllables per second) indicated that there was little acoustic degradation. In comparison, time-compression in Chapter 2 (37%) resulted in speaking rates nearly twice as fast (10.3 syllables per second). The speaking rate in Chapter 3 was therefore well within the range of conversational speaking rates in West Germanic languages (Koch & Janse, 2016). Furthermore, the speaking rate was within the theta range of syllable parsing, according to the TEMPO model of speech perception (Ghitza, 2011). According to TEMPO, syllabic parsing should be intact for such rates. Indeed, results showed optimal intelligibility for speech presented in quiet without reverberation. However, pupillometry findings showing larger dilation in response to fast speech cannot be explained by TEMPO. Previous studies have shown that older listeners can achieve a performance similar to that of young listeners with time-compressed speech at low compression rates (Wingfield et al., 2003). At the same time, Wingfield et al. (2003) showed that older listeners were generally slower to respond to time-compressed sentences. In fact, the difficulty of older listeners with time-compressed speech has been previously associated with cognitive decline, specifically slower processing speed (Janse, 2009; Salthouse, 1996). Larger pupil dilation might therefore reflect the extra processing effort exerted when processing fast speech under such cognitive

constraints. Since results presented in Chapter 3 do not indicate compromised intelligibility, they therefore contribute to the body of research emphasising the involvement of cognitive factors in older listeners' experienced difficulty in processing fast speech (e.g., Salthouse, 1996; Janse, 2009; Wingfield et al., 2003). The results presented in Chapter 3 also suggest that intact syllabic parsing (see TEMPO, Ghitza, 2011) does not preclude absence of listening effort experienced by listeners when processing fast speech.

Results presented in Chapter 3 also showed that added reverberation by simulating different room acoustics led to an increase in listening effort, confirming my hypotheses (H9 and H10). While reverberation amplified the effect of speaking rate on intelligibility, similar to previous studies (e.g., Gordon-Salant & Fitzgibbons, 1995), this was not the case for listening effort, dis-confirming hypothesis H11. Similarly, listening effort under reverberation was not diminished by dereverberation, dis-confirming hypothesis H12. Possible implications for future studies are discussed below (Section 5.2.1). With respect to RQ4, results therefore showed that room acoustics did indeed affect listening effort even when intelligibility was high. It has to be noted that the increase in effort was also accompanied by a decrease in intelligibility, reflecting acoustic degradation through mechanisms such as masking and self-masking (Nabelek & Robinette, 1978). While higher listening effort under reverberation was indicated by perceived effort ratings, pupillometry effects were only marginally significant under robust LMM estimation. The results were possibly masked because fast speech also elicited a larger pupil dilation in dry. These results have implications regarding the non-linearity of the pupil dilation with respect to intelligibility (e.g., Winn et al., 2015; Wendt et al., 2018). Specifically, pupil dilation appears to be sensitive to very small differences in sentence intelligibility at high intelligibility.

However, in Chapter 3, despite subtle differences in intelligibility between speech in dry and speech in reverb, fast speech appeared to elicit pupil responses that were similarly large in dry and in reverb. This finding indicates that intelligibility might not be necessarily the driving factor involved. It is

possible that the generally assumed U-shape curve only applies to situations in which stimulus difficulty is manipulated by adapting one continuous parameter such as signal-to-noise ratio (e.g., Ohlenforst et al., 2017; Wendt et al., 2018). However, when adapting different stimulus characteristics such as speaking rate and reverberation (i.e., source and channel) as in Chapter 3, the U-shape might not apply any more. The principle of changed stimulus complexity (i.e., varying in different dimensions) has been described before by Koelewijn et al. (2012), who argued that the fact that informational masking leads to larger pupil dilation than energetic masking, despite fixed intelligibility, is due to a change in masker complexity, rather than signal intensity. Furthermore, other limitations and possible alternative explanations have to be taken into account, as well, as discussed below.

5.2.1 Limitations and future directions

Several limitations have to be taken into account regarding experimental design and interpretation of findings from Chapter 3. As discussed in Chapter 1, resource allocation in the capacity model of attention prioritises novel stimuli. Since pupil dilation is linked to physiological arousal, an increase for fast speech might have indicated states of higher attention towards unusually fast speech. Similarly, a study with normal-hearing listeners has shown that pupil dilation increased for temporally-modified speech even when intelligibility was fixed at 50% (Paulus et al., 2019). In this study, speech was modified by local time-compression and time-elongation, possibly rendering speech to sound unnatural. Time-compression is usually more intelligible than natural fast speech because it preserves the spectral characteristics of the original speech (Janse, 2009). Despite its rate being similar to that of conversational speech, it is possible that time-compressed speech in Chapter 3 was perceived as less natural because it was not accompanied by spectral modifications. It is therefore conceivable that the larger peak dilation observed for fast speech might disappear when listeners have received training with this novel stimulus. Indeed, results hinted at a reduction of pupil dilation in the retest session (on average).

Little pupillometry research has investigated such training effects. One study by Kuchinsky et al. (2014) applied pupillometry to a speech-perception training protocol. Across multiple sessions, participants in the experimental group were trained to recognise speech presented in noise. After training, it was observed that the average task-evoked pupil dilation increased for the training group, but not for the control group, which was interpreted as increased attentiveness. Furthermore, the pupil dilation was found to increase faster after training, which was interpreted as faster speech in noise discrimination. Even though those results appear to contradict the interpretation of trends shown in Chapter 3, differences in sentence difficulty and degradation have to be taken into account. Nevertheless, both studies indicate changes in pupil dilation between test and retest sessions, suggesting that some results in the pupillometry literature might be confined to single sessions. There is potential for future research to take such retest effects into account and to evaluate whether pupillometry is able to detect a reduction in listening effort with auditory training. These measures could be useful in determining the success of specific training protocols.

While the current study investigated session effects, it did not explore differences in intelligibility levels, in contrast to previous studies. It therefore remains unclear to which extent overall high intelligibility influenced results with respect to dereverberation, i.e., lower intelligibility in the retest session, but no effect on listening effort. As hearing aid programs are designed to improve challenging listening conditions, it is not unusual to observe unexpected effects at such boundary conditions. Future studies would therefore ideally consider at least two levels of intelligibility (low and high). In addition, given the current results with respect to dereverberation, it is conceivable that the standard design of pupillometry studies is less appropriate when evaluating hearing aid programs targeted at continuous speech. A short sentence duration as well as gaps of silence between trials possibly interfered with the activation cycle of the algorithm. Pupillometry designs allowing for continuous speech might therefore be more appropriate when evaluating such programs (e.g., McGarrigle et al., 2017).

Another downside of the study presented in Chapter 3 is that a direct comparison between younger and older listeners was not made. While it is possible to compare results presented in Chapter 2 with Chapter 3 to some extent - with respect to time-compression effects - I did not specifically compare between age groups. However, between-group comparisons are difficult because of age-related changes in pupil reactivity (e.g., Piquado et al., 2010; Winn, 2016). Therefore, comparisons across listener groups usually require a form of normalisation. Piquado et al. (2010) proposed to adjust pupil size based on individual dynamic ranges, measured by the pupil light reflex. Winn (2016) implicitly normalised pupil size by measuring effort reduction, i.e., the difference between harder and easier conditions. Recently, it has been suggested to not only correct for pupil dynamic ranges, but also 'cognitive dynamic ranges' (Winn et al., 2018), which entails measuring an individual's pupil reactivity to a range of cognitive task manipulations (e.g. memory load). Future pupillometry studies comparing both younger and older adults, with and without hearing impairments, could conduct between-subject analyses by factoring in dynamic ranges (both anatomic and cognitive reactivity), based on standardised tests such as the light-reflex test at different luminance levels (anatomic reactivity) and digit memory tests with varying numbers of items (cognitive reactivity).

5.3 Pupillometry during speech production as a measure of speaking effort

Results presented in Chapter 4 replicated previous Lombard studies showing stronger acoustic enhancements when producing speech in the presence of different masker types, confirming hypothesis H13. Specifically, acoustic enhancements were reflected across a range of features such as higher fundamental frequency and mean energy in mid-range frequencies, as well as slower speaking rate. Furthermore, results showed larger pupil dilation under competing-talker masking compared to stationary masking, likely indicating higher speaking effort. These results confirmed the second hypothesis ad-

dressed in Chapter 4 (H14). The chapter answered research question RQ5 by showing that pupillometry can be applied during speech production to quantify speaking effort, analogous to listening effort. Similar to the definition of listening effort provided by McGarrigle et al. (2014), speaking effort can be considered the mental exertion required to convey an auditory message.

The results presented in Chapter 4 are in accordance with speech perception findings, showing larger pupil dilation when processing speech with a competing talker background (cf. Koelewijn et al., 2012; Wendt et al., 2018 for speech perception). Since informational masking leads to perceptual interference, it is not surprising that higher effort has to be exerted both when perceiving and producing speech. Interference by informational masking can occur at phonological and semantic levels (Heinrich et al., 2008; Saito & Baddeley, 2004; Schneider et al., 2007). For instance, Saito & Baddeley (2004) asked participants to read and then produce single target words repeatedly (10-12 times) while each production was preceded by tones, or phonologically similar or dissimilar distractor words. Using this so-called speech-error induction technique, the authors observed speech production errors when the production was preceded by a distractor word, with a higher number of errors observed when the distractor word was phonologically similar. Similarly, in Chapter 4, it was shown that perceptual interference was also reflected by a higher number of speech production errors observed for speech produced in the presence of a competing talker. In addition, distractor words in the current study were entire semantically plausible sentences, i.e., interference was possibly caused on the semantic level, as well. Higher effort was then required by talkers to focus on the articulation of the target sentence while inhibiting processing of the competing speech (cf. Schneider et al., 2007 for speech perception). In accordance with the capacity model of attention (Kahneman, 1973) and the framework for understanding effortful listening (Pichora-Fuller et al., 2016), the additional task demands under informational masking (e.g., inhibitory processes) lead to elevated arousal, as indexed by a larger pupil dilation. The allocation of additional processing resources then allows talkers to maintain task performance, i.e., to accurately produce sentences. However, similar to results from speech

perception presented in Chapter 2, the release of additional resources did not help to maintain optimal task performance, as indexed by behavioural measures (speech perception and production errors). On the other hand, it is conceivable that higher effort, as indexed by larger pupil dilation, would be observed even without decrements in behavioural measures, similar to results presented in Chapter 3. The sentence material used, i.e., unpredictable IEEE sentences, was potentially difficult enough to elicit production errors that would not have occurred otherwise. In analogy to models of effortful speech perception (Rönnberg et al., 2008), it is possible that talkers with better working memory would require less effort when producing speech in the presence of a competing talker. Such background measures were not obtained in the current study, but should be considered in future speech production studies.

5.3.1 Limitations and future directions

In Lombard studies, maskers are usually presented continuously, reflecting ecologically valid real-life communication (e.g., Beechey et al., 2019). Hence, in Chapter 4, perceptual interference with the competing talker could have occurred at multiple stages of the speech production process. It is therefore possible that the larger pupil dilation under competing-talker masking was an accumulation of increased demands on reading, rehearsal and articulation. To test whether interference can occur at any of these stages, further studies should be conducted that vary masker onset time. However, as shown in Chapter 4, increasing baseline pupil dilation following masker onset might make it difficult to disentangle noise-induced and cognitive effects on the pupil dilation. In addition, as the aim of the current thesis was to evaluate the feasibility of pupillometry for naturalistic communication, varying masker on- and offsets might discount the ecological validity of the paradigm.

A next step towards the application of pupillometry in naturalistic communication research would be to combine both listening and speaking tasks (cf. Chapters 2, 3 and 4), before moving towards spontaneous speech tasks

(e.g., Van Engen et al., 2010; Beechey et al., 2019). Spontaneous speech might be methodologically challenging as pupillometry relies on the timing of stimulus events. On the other hand, alternative measures have been proposed to measure more gradual physiological changes, independent of the task-evoked pupil response which is time-locked to a stimulus. For instance, a method proposed by Wagner et al. (2019) measures changes in baseline pupil size across trials as an index of sustained attention under increased task demands (see also Ayasse & Wingfield, 2020). This method could be applied to a speech production task, quantifying how different talker groups employ attention throughout a communication task such as diapix (Van Engen et al., 2010). In communicative settings, deliberate speech modifications are often dependent on listener factors (Cooke et al., 2014). For example, speech directed towards hearing-impaired listeners often has clear-speech characteristics such as slower speaking rates and fewer vowel reductions (Hazan et al., 2018; Picheny et al., 1986). While Chapter 4 presented a study with younger normal-hearing listeners, a direction for future research could be to compare normal-hearing and hearing-impaired, as well as younger and older talker groups. Specifically, it has been suggested that cognitive load might play an important role in how different talker groups modify their speech (Hazan et al., 2018). Identifying the extra processing load in the presence of noise through pupillometry could pinpoint when and why talkers acoustically modify their speech.

5.4 Conclusion

This thesis showed that acoustic degradations affected individuals not only at acoustic, but also at cognitive levels. Most importantly, I showed that the concept of cognitive effort can be applied not only to listening, but also to speaking, which are both key communicative roles.

The three studies presented in this thesis provided new insights into several aspects of speech perception and production under acoustic degradations. The results of the first study showed that acoustic-phonetic talker differences such as

vowel space dispersion predicted intelligibility under different channel degradations. Furthermore, despite relatively high intelligibility, listeners exhibited larger pupil dilation when processing fast time-compressed speech compared to speech presented at normal rates, suggesting higher listening effort. The results of the second study showed that older hearing-impaired listeners also exhibited larger pupil dilation when processing fast time-compressed speech, even without loss of intelligibility. These results add to the growing realisation that older hearing-impaired listeners experience higher effort even in highly intelligible listening situations. The results of the third study showed that pupillometry can be used as a tool to measure speaking effort. Specifically, larger pupil dilation was observed when talkers produced speech in the presence of a competing talker, suggesting perceptual interference.

Findings and limitations presented in this thesis should guide further research. While many studies have shown that pupil dilation varies as a function of intelligibility, this thesis suggests that listening situations with high intelligibility are most appealing for several reasons. First, they reflect realistic everyday acoustic environments. Second, pupillometry appears to be most sensitive at high intelligibility levels, indicating effort where traditional measures are not necessarily able to capture differences between listening conditions. However, two limitations warrant further research. Discrepancies found between pupillometry and perceived effort ratings suggest a possible influence of stimulus novelty on the pupil dilation response. The indication of a possible reduction of pupil dilation in the retest session supports this hypothesis. Both limitations prompt further investigation of both behavioural and physiological measures to determine thresholds at which effort becomes a critical problem for listeners. To pinpoint processing effort more precisely, both methods should be employed in conjunction.

This thesis provided first insights into applying pupillometry to a speech production in noise paradigm, paving the way for more sophisticated experimental designs. One direction of future research would be to apply pupillometry to naturalistic conversations, involving both speech perception and production.

Pupillometry would then complement traditional acoustic-phonetic measures and reveal the amount of effort exerted by different talker or listener groups. Another direction of future research would be to use tonic (baseline) pupil size measurements, collected across the duration of an entire experiment, to pinpoint the fatigue experienced when producing speech in noise.

References

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *J. Acoust. Soc. Am.*, *126*(5), 2649–2659. https://doi.org/10.1121/1.3216914

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging*, *25*(3), 736–740. https://doi.org/10.1037/a0020054

Antikainen, J., & Niemi, P. (1983). Neuroticism and the pupillary response to a brief exposure to noise. *Biological Psychology*, *17*(2-3), 131–135. https://doi.org/10.1016/0301-0511(83)90013-3

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Aubanel, V., Cooke, M., Villegas, J., & Lecumberri, M. L. G. (2011). Conversing in the presence of a competing conversation: Effects on speech production. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, August*, 2833–2836.

Ayasse, N. D., & Wingfield, A. (2020). Anticipatory Baseline Pupil Diameter Is Sensitive to Differences in Hearing Thresholds. *Frontiers in Psychology*, *10*(January), 1–7. https://doi.org/10.3389/fpsyg.2019.02947

Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology*, *8*(4), 485–493. https://doi.org/https://doi.org/10.1037/0894-4105.8.4.485

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *137*(4), 2015–2024. https://doi.org/10.1121/1.4916265

Barthel, M., & Sauppe, S. (2019). Speech Planning at Turn Transitions in Dialog Is Associated With Increased Processing Load. *Cognitive Science*, *43*(7), 1–16. https://doi.org/10.1111/cogs.12768

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beechey, T., Buchholz, J. M., & Keidser, G. (2019). Eliciting Naturalistic Conversations: A Method for Assessing Communication Ability, Subjective Experience, and the Impacts of Noise and Hearing Impairment. *Journal of Speech Language and Hearing Research*, *62*(February), 470–484. https://doi.org/10.1044/2018_JSLHR-H-18-0107

Benoit, K. (2018). *quanteda: Quantitative Analysis of Textual Data*. https://doi.org/10.5281/zenodo.1004683

Bent, T., Buchwald, A., & Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *J. Acoust. Soc. Am.*, *126*(5), 2660–2669. https://doi.org/10.1121/1.3212930

Bergamin, O., & Kardon, R. H. (2003). Latency of the pupil light reflex: Sample rate, stimulus intensity, and variation in normal subjects. *Investigative Ophthalmology and Visual Science*, *44*(4), 1546–1554. https://doi.org/10.1167/iovs.02-0468

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to- noise ratio of a sampled sound. *IFA Proceedings*, *17*, 97–110.

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer (Version 6.0.40) [Computer software]*. http://www.praat.org/

Bolt, R. H., & MacDonald, A. D. (1949). Theory of Speech Masking by Reverberation. *Journal of the Acoustical Society of America*, *21*(6), 577–580. https://doi.org/10.1121/1.1906551

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, *12*(152), 1–13. https://doi.org/10.3389/fnins.2018.00152

Borghini, G., & Hazan, V. (2020). Effects of acoustic and semantic cues on listening effort during native and non-native speech perception. *The Journal of the Acoustical Society of America*, *147*(6), 3783–3794. https://doi.org/10.1121/10.0001126

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

Broadbent, D. E. (1954). Some Effects of Noise on Visual Performance. *Quarterly Journal of Experimental Psychology*, *6*(1), 1–5. https://doi.org/10.1080/17470215408416643

Brown, V. A., Mclaughlin, D. J., Strand, J. F., & Engen, K. J. V. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*, *73*(9), 1431–1443. https://doi.org/10.1177/1747021820916726

Brumm, H., & Zollinger, A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*(11-13), 1173–1198. https://doi.org/10.1163/000579511X605759

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 629–647. https://doi.org/10.1002/wcs.63

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Christoffels, I. K., Formisano, E., & Schiller, N. O. (2007). Neural correlates of verbal feedback processing: An fMRI study employing overt speech. *Human Brain Mapping*, *28*(9), 868–879. https://doi.org/10.1002/hbm.20315

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427. https://doi.org/10.1121/1.2804952

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech and Language*, *28*(2), 543–571. https://doi.org/10.1016/j.csl.2013.08.003

Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, *128*(4), 2059–2069. https://doi.org/10.1121/1.3478775

Craik, F. I., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences*, *10*(3), 131–138. https://doi.org/10.1016/j.tics.2006.01.007

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

De Looze, C., & Hirst, D. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Speech Prosody*, 135–138.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*(4), 2403–2411. https://doi.org/10.1121/1.419603

Dubno, J. R., & Schaefer, A. B. (1992). Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners. *The Journal of the Acoustical Society of America*, *91*(4), 2110–2121. https://doi.org/10.1121/1.403697

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *Journal of Experimental Psychology. Human Perception and Performance*, *23*(3), 914–927. https://doi.org/10.1037/0096-1523.23.3.914

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia*, *50*(9), 2154–2164. https://doi.org/10.1016/j.neuropsychologia.2012.05.013

Faller, A., & Schünke, M. (1995). *Der Körper des Menschen* (12th ed., pp. 425–435). Thieme.

Fant, G. (1973). *Speech sounds and features*. MIT press.

Fellows, R. P., Dahmen, J., Cook, D., & Schmitter-Edgecombe, M. (2017). Multicomponent analysis of a digital Trail Making Test. *Clinical Neuropsychologist*, *31*(1), 154–167. https://doi.org/10.1080/

13854046.2016.1238510

Ferreira, V. S. (2010). Language production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 834–844. https://doi.org/10.1002/wcs.70

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736. https://doi.org/10.1121/1.400247

Francis, A. L., Tigchelaar, L. J., Zhang, R., & Zekveld, A. A. (2018). Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1815–1830. https://doi.org/10.1044/2018_JSLHR-H-17-0254

Füllgrabe, C., & Rosen, S. (2016). On The (Un) importance of Working Memory in Speech-in-Noise Processing for Listeners with Normal Hearing Thresholds. *Frontiers in Psychology*, *7*(1268), 1–8. https://doi.org/10.3389/fpsyg.2016.01268

Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language, and Hearing Research*, *53*(3), 588–609. https://doi.org/https://doi.org/10.1044/1092-4388(2009/08-0138)

Gelfand, S., & Hochberg, I. (1976). Binaural and Monaural Speech Discrimination Under Reverberation. *Audiology*, *15*(1), 72–84. https://doi.org/10.3109/00206097609071765

Ghitza, O. (2014). Behavioral evidence for the role of cortical Θ oscillations in determining auditory channel capacity for speech. *Frontiers in Psychology*, *5*(JUL), 1–12. https://doi.org/10.3389/fpsyg.2014.00652

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the

input rhythm. *Frontiers in Psychology*, *2*(JUN), 1–13. https://doi.org/10.3389/fpsyg.2011.00130

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology*, *3*(JUL), 1–12. https://doi.org/10.3389/fpsyg.2012.00238

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–126. https://doi.org/10.1159/000208934

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective and Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Golomb, J. D., Peelle, J. E., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *The Journal of the Acoustical Society of America*, *121*(3), 1701–1708. https://doi.org/10.1121/1.2436635

Goossens, T., Vercammen, C., Wouters, J., & Wieringen, A. V. (2017). Masked speech perception across the adult lifespan: Impact of age and hearing impairment. *Hearing Research*, *344*, 109–124. https://doi.org/10.1016/j.heares.2016.11.004

Gordon-Salant, S., & Fitzgibbons, P. J. (1993). Temporal factors and speech recognition performance in young and elderly listeners. *Journal of Speech and Hearing Research*, *36*(6), 1276–1285. https://doi.org/10.1044/jshr.3606.1276

Gordon-Salant, S., & Fitzgibbons, P. J. (1995). Recognition of multiply degraded speech by young and elderly listeners. *Journal of Speech and Hearing Research*, *38*(5), 1150–1156. https://doi.org/https://doi.org/

10.1044/jshr.3805.1150

Gordon-Salant, S., & Fitzgibbons, P. J. (2001). Sources of age-related recognition difficulty for time-compressed speech. *Journal of Speech, Language, and Hearing Research*, *44*(4), 709–719. https://doi.org/10.1044/1092-4388(2001/056)

Gordon-Salant, S., & Fitzgibbons, P. J. (2004). Effects of stimulus and noise rate variability on speech perception by younger and older adults. *The Journal of the Acoustical Society of America*, *115*(4), 1808–1817. https://doi.org/10.1121/1.1645249

Gordon-Salant, S., Fitzgibbons, P., & Yeni-Komshian, G. (2011). Auditory temporal processing and aging: implications for speech understanding of older people. *Audiological Research*, *1*(1), 9–15. https://doi.org/10.4081/audiores.2011.e4

Gordon-Salant, S., Zion, D. J., & Espy-Wilson, C. (2014). Recognition of time-compressed speech does not predict recognition of natural fast-rate speech by older listeners. *The Journal of the Acoustical Society of America*, *136*(4), EL268–EL274. https://doi.org/10.1121/1.4895014

Govender, A., Wagner, A. E., & King, S. (2019). Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, September*, 1551–1555. https://doi.org/10.21437/Interspeech.2019-1783

Granlund, S., Hazan, V., & Mahon, M. (2018). Children's acoustic and linguistic adaptations to peers with hearing impairment. *Journal of Speech, Language, and Hearing Research*, *61*(5), 1055–1069. https://doi.org/10.1044/2017_JSLHR-S-16-0456

Green, T., Katiri, S., Faulkner, A., & Rosen, S. (2007). Talker intelligibility differences in cochlear implant listeners. *J. Acoust. Soc. Am.*, *121*(6), EL223–9. https://doi.org/10.1121/1.2720938

Greenwood, D. D. (1990). A cochlear frequency position function for several species - 29 years later. *J. Acoust. Soc. Am.*, *87*(6), 2592–2605. https://doi.org/10.1121/1.399052

Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, *130*(4), 2139–2152. https://doi.org/10.1121/1.3623753

Hazan, V., & Baker, R. (2010). Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? *Proceedings of Diss-Lpss Joint Workshop*, 7–10.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *J. Acoust. Soc. Am.*, *116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., & Brungart, D. (2018). Clear speech adaptations in spontaneous speech produced by young and older adults. *J. Acoust. Soc. Am.*, *144*(3), 1331–1346. https://doi.org/10.1121/1.5053218

Hazan, V., Tuomainen, O., & Pettinato, M. (2016). Suprasegmental Characteristics of Spontaneous Speech Produced in Good and Challenging Communicative Conditions by Talkers Aged 9–14 Years. *Journal of Speech, Language, and Hearing Research*, *59*(6), S1596–S1607. https://doi.org/10.1044/2016_JSLHR-S-15-0046

Hazan, V., Tuomainen, O., & Taschenberger, L. (2019). Subjective evaluation of communicative effort for younger and older adults in interactive tasks with energetic and informational masking. *Proceedings of 20th Annual Conference of the International Speech Communication Association - Interspeech*, 3098–3102. https://doi.org/10.21437/Interspeech.2019

Heinrich, A., Gagné, J.-P., Viljanen, A., Levy, D. A., Ben-David, B. M., & Schneider, B. A. (2016). Effective Communication as a Fundamental

Aspect of Active Aging and Well-Being: Paying Attention to the Challenges Older Adults Face in Noisy Environments. *Social Inquiry into Well-Being*, *2*(1), 51–69. https://doi.org/10.13165/SIIW-16-2-1-05

Heinrich, A., Schneider, B. A., & Craik, F. I. (2008). Investigating the influence of continuous babble on auditory short-term memory performance. *Quarterly Journal of Experimental Psychology*, *61*(5), 735–751. https://doi.org/10.1080/17470210701402372

Helfer, K. S., & Wilber, L. A. (1990). Hearing loss, aging, and speech perception in reverberation and noise. *Journal of Speech and Hearing Research*, *33*(1), 149–155. https://doi.org/10.1044/jshr.3301.149

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual Learning of Noise Vocoded Words: Effects of Feedback and Lexicality. *Journal of Experimental Psychology*, *34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech understanding. *Nature*, *8*(May), 393–402. https://doi.org/doi:10.1038/nrn2113

Huckvale, M. (2014). *ProRec: a program for field workers (Version 1.45) [Computer software]*. https://www.phon.ucl.ac.uk/

Hupé, J. M., Lamirel, C., & Lorenceau, J. (2009). Pupil dynamics during bistable motion perception. *Journal of Vision*, *9*(7), 1–19. https://doi.org/10.1167/9.7.10

Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech requires attention. *J. Acoust. Soc. Am.*, *131*(3), EL236–EL242. https://doi.org/10.1121/1.3685511

Institute of Electrical and Electronics Engineers. (1969). IEEE recommended practices for speech quality measurements. *IEEE Trans. Aud. Electroacoust*, *17*, 227–246.

Jadoul, Y., Thompson, B., & Boer, B. de. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*(2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Janse, E. (2003). *Production and perception of fast speech* [PhD thesis]. Universiteit Utrecht.

Janse, E. (2009). Processing of fast speech by elderly listeners. *The Journal of the Acoustical Society of America*, *125*(4), 2361–2373. https://doi.org/10.1121/1.3082117

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *Quarterly Journal of Experimental Psychology*, *65*(8), 1563–1585. https://doi.org/10.1080/17470218.2012.658822

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford University Press.

Johnson, E. M., Morgan, S. D., & Ferguson, S. H. (2020). Does Time Compression Decrease Intelligibility for Female Talkers More Than for Male Talkers? *Journal of Speech, Language, and Hearing Research*, *63*(4), 1083–1092. https://doi.org/10.1044/2020_JSLHR-19-00301

Johnson, J. A., Cox, R. M., & Alexander, G. C. (2010). Development of APHAB Norms for WDRC Hearing Aids and Comparisons with Original Norms. *Ear & Hearing*, *31*(1), 47–55. https://doi.org/10.1097/AUD.0b013e3181b8397c

Johnson, K. (2000). Adaptive dispersion in vowel perception. *Phonetica*, *57*(2-4), 181–188. https://doi.org/10.1159/000028471

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: phonetic targets are hyperarticulated. *Linguistic Society of America*, *69*(3), 505–528. https://doi.org/10.2307/416697

Kahneman, D. (1973). *Attention and Effort*. https://doi.org/10.2307/1421603

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Sci-*

ence, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kahneman, D., & Beatty, J. (1967). Pupillary responses in a pitch-discrimination task. *Perception & Psychophysics*, *2*(3), 101–105. https://doi.org/10.3758/BF03210302

Kennedy-Higgins, D., Devlin, J. T., & Adank, P. (2020). Cognitive mechanisms underpinning successful perception of different speech distortions. *The Journal of the Acoustical Society of America*, *147*(4), 2728–2740. https://doi.org/10.1121/10.0001160

Kidd, G., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, *124*(6), 3793–3802. https://doi.org/10.1121/1.2998980

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational Masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–189). Springer. https://doi.org/10.1007/978-0-387-71305-2

Kiessling, J., Pichora-Fuller, M. K., Gatehouse, S., Stephens, D., Arlinger, S., Chisolm, T., Davis, A. C., Erber, N. P., Hickson, L., Holmes, A., Rosenhall, U., & Von Wedel, H. (2003). Candidature for and delivery of audiological services: Special needs of older people. *International Journal of Audiology*, *42*(SUPPL. 2). https://doi.org/10.3109/14992020309074650

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208–1221. https://doi.org/10.1121/1.380986

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323–332. https://doi.org/10.1111/j.1469-8986.2010.01069.x

Knudsen, V. O. (1929). The hearing of speech in auditoriums. *Journal of the Acoustical Society of America*, *1*(1), 56–82. https://doi.org/10.1121/1.1901470

Koch, X., & Janse, E. (2016). Speech rate effects on the processing of conversational speech across the adult life span. *J. Acoust. Soc. Am.*, *139*(4), 1618–1636. https://doi.org/10.1121/1.4944032

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*(2), 291–300. https://doi.org/10.1097/AUD.0b013e3182310019

Koenig, L. L., & Fuchs, S. (2019). Vowel Formants in Normal and Loud Speech. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1278–1295. https://doi.org/https://doi.org/10.1044/2018_JSLHR-S-18-0043

Koller, M. (2016). Robustlmm: An R package for Robust estimation of linear Mixed-Effects models. *Journal of Statistical Software*, *75*(1). https://doi.org/10.18637/jss.v075.i06

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing Aspects of Auditory Handicap by Means of Pupil Dilatation. *Audiology*, *36*(3), 155–164. https://doi.org/10.3109/00206099709071969

Kramer, S. E., Teunissen, C. E., & Zekveld, A. A. (2016). Cortisol, Chromogranin A, and Pupillary Responses Evoked by Speech Recognition Tasks in Normally Hearing and Hard-of-Hearing Listeners. *Ear and Hearing*, *37*(2015), 126S–135S. https://doi.org/10.1097/AUD.0000000000000311

Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.*, *115*(1), 362–378. https://doi.org/10.1121/1.1635842

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eck-

ert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology*, *51*(10), 1046–1057. https://doi.org/10.1111/psyp.12242

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lebart, K., Boucher, J. M., & Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acustica United with Acustica*, *87*(3), 359–366.

Lemke, U., & Besser, J. (2016). Cognitive Load and Listening Effort : Concepts and Age-Related Considerations. *Ear & Hearing*, *37*, 77S–84S. https://doi.org/10.1097/AUD.0000000000000304

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. https://doi.org/https://doi.org/10.1016/0010-0277(83)90026-4

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, *48*(4), 839–862. https://doi.org/10.2307/411991

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *Journal of the Acoustical Society of America*, *35*(11), 1773–1781. https://doi.org/10.1121/1.1918816

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Mal. De L'Oreille et Du Larynx*, 101–119.

Lõo, K., Rij, J. van, Jarvikivi, J., & Baayen, H. (2016). Individual Differences in Pupil Dilation during Naming Task. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 550–555.

Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, *124*(5), 3261–3275. https://doi.org/10.1121/1.2990705

Lucus, M. N., Goshorn, E. L., & Kemker, B. E. (2011). Ambient noise levels and reverberation times in Mississippi school rooms. *Proceedings of Meetings on Acoustics*, *9*, 1–5. https://doi.org/10.1121/1.3556447

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Mackersie, C. L., & Cones, H. (2011). Subjective and Psychophysiological Indexes of Listening Effort in a Competing-Talker Task. *Journal of the American Academy of Audiology*, *22*(2), 113–122. https://doi.org/10.3766/jaaa.22.2.6

Magnusson, L., Claesson, A., Persson, M., & Tengstrand, T. (2013). Speech recognition in noise using bilateral open-fit hearing aids: The limited benefit of directional microphones and noise reduction. *International Journal of Audiology*, *52*(1), 29–36. https://doi.org/10.3109/14992027.2012.707335

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

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. https://doi.org/10.1080/01690965.2012.705006

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proc. Interspeech*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McCloy, D. R., Larson, E. D., Lau, B., & Lee, A. K. C. (2016). Temporal alignment of pupillary response with stimulus events via deconvolution. *The Journal of the Acoustical Society of America*, *139*(3), EL57–EL62. https://doi.org/10.1121/1.4943787

McCloy, D. R., Lau, B. K., Larson, E., Pratt, K. A. I., & Lee, A. K. C. (2017). Pupillometry shows the effort of auditory attention switching. *The Journal of the Acoustical Society of America*, *141*(4), 2440–2451. https://doi.org/10.1121/1.4979340

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, *54*(2), 193–203. https://doi.org/10.1111/psyp.12772

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*(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception, and Psychophysics*, *80*(6), 1559–1570. https://doi.org/10.3758/s13414-018-1537-4

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *J. Acoust. Soc. Am.*, *147*(2), EL151–EL156. https://doi.org/10.1121/10.0000718

Meekings, S., Evans, S., Lavan, N., Boebinger, D., Krieger-Redwood, K., Cooke, M., & Scott, S. K. (2016). Distinct neural systems recruited when speech production is modulated by different masking sounds. *The Journal of the Acoustical Society of America*, *140*(1), 8–19. https://doi.org/10.1121/1.4948587

Menard, S. (2002). *Quantitative Applications in the Social Sciences: Applied logistic regression analysis*. SAGE Publications. https://doi.org/10.4135/9781412983433

Miller, G. A., & Licklider, J. C. (1950). The Intelligibility of Interrupted Speech. *Journal of the Acoustical Society of America*, *22*(2), 167–173. https://doi.org/10.1121/1.1906584

Moon, S., & Lindblom, B. (1989). Formant undershoot in clear and citation-form speech: a second progress report. *Speech Transmission Laboratory - Quarterly Progress Status Report*, *30*(1), 121–123.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, *9*(4), 399–406. https://doi.org/10.1007/s10162-008-0143-x

Moore, T. M., & Picou, E. M. (2018). A potential bias in subjective ratings of mental effort. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2405–2421. https://doi.org/10.1044/2018_JSLHR-H-17-0451

Morimoto, M., Sato, H., & Kobayashi, M. (2004). Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces. *J. Acoust. Soc. Am.*, *116*(3), 1607–1613. https://doi.org/10.1121/1.1775276

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453–467. https://doi.org/https://doi.org/10.1016/0167-6393(90)90021-Z

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Müller, J. A., Wendt, D., Kollmeier, B., Debener, S., & Brand, T. (2019). Effect of speech rate on neural tracking of speech. *Frontiers in Psychology*, *10*(MAR), 1–15. https://doi.org/10.3389/fpsyg.2019.00449

Nabelek, A. K., & Robinette, L. (1978). Influence of the precedence effect on word identification by normally hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, *63*(1), 187. https://doi.org/10.1121/1.381711

Nejime, Y., & Moore, B. C. J. (1998). Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *The Journal of the Acoustical Society of America*, *103*(1), 572–576. https://doi.org/10.1121/1.421123

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., & Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hearing Research*, *365*, 90–99. https://doi.org/10.1016/j.heares.2018.05.003

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., & Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. https://doi.org/10.1016/j.heares.2017.05.012

Oreinos, C. (2015). *Virtual Acoustic Environments for the Evaluation of Hearing Devices* [PhD thesis]. Macquarie University.

Pallier, C., Sebastian-Gallés, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory and Cognition*, *26*(4), 844–851. https://doi.org/10.3758/BF03211403

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, and*

Psychophysics, *74*(4), 754–765. https://doi.org/10.3758/s13414-011-0263-y

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech, Language, and Hearing Research*, *54*(4), 1051–1059. https://doi.org/10.1044/1092-4388(2010/10-0162)

Paulus, M., Hazan, V., & Adank, P. (2019). Talker intelligibility and listening effort with temporally modified speech. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, 3128–3132. https://doi.org/10.21437/Interspeech.2019-1402

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*(SEP), 1–17. https://doi.org/10.3389/fpsyg.2012.00320

Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1315–1330. https://doi.org/10.1037/0096-1523.31.6.1315

Peng, Z. E., & Wang, L. M. (2019). Listening effort by native and nonnative listeners due to noise, reverberation, and talker foreign accent during english speech perception. *Journal of Speech, Language, and Hearing Research*, *62*(4), 1068–1081. https://doi.org/10.1044/2018_JSLHR-H-17-0423

Perkell, J. S., Denny, M., Lane, H., Guenther, F., Matthies, M. L., Tiede, M., Vick, J., Zandipour, M., & Burton, E. (2007). Effects of masking noise on vowel and sibilant contrasts in normal-hearing speakers and postlingually deafened cochlear implant users. *The Journal of the Acoustical Society of America*, *121*(1), 505–518. https://doi.org/10.1121/1.2384848

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational

speech. *J. Acoust. Soc. Am.*, *29*(4), 434–446. https://doi.org/10.1044/jshr.2904.434

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1989). Speaking Clearly for the Hard of Hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, *32*(3), 600–603.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37*, 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Pichora-Fuller, M. K., Schneider, B. A., MacDonald, E., Pass, H. E., & Brown, S. (2007). Temporal jitter disrupts speech intelligibility: A simulation of auditory aging. *Hearing Research*, *223*(1-2), 114–121. https://doi.org/10.1016/j.heares.2006.10.009

Picou, E. M., Gordon, J., & Ricketts, T. A. (2016). The effects of noise and reverberation on listening effort for adults with normal hearing. *Ear & Hearing*, *37*(1), 1–13. https://doi.org/10.1097/AUD.0000000000000222

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Porretta, V., & Tucker, B. V. (2019). Eyes Wide Open : Pupillary Response to a Foreign Accent Varying in Intelligibility. *Frontiers in Communication*, *4*(8), 1–12. https://doi.org/10.3389/fcomm.2019.00008

Pribram, K. H., & McGuinness, D. (1975). Arousal, activation, and effort in the control of attention. *Psychological Review*, *82*(2), 116–149. https://doi.org/10.1037/h0076780

Pulkki, V. (2001). *Spatial sound generation and perception by amplitude panning techniques* [PhD thesis]. Helsinki University of Technology.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, *20*(3), 241–248. https://doi.org/10.1080/14640746808400158

Radeau, M., Morais, J., Mousty, P., & Bertelson, P. (2000). The Effect of Speaking Rate on the Role of the Uniqueness Point in Spoken Word Recognition. *Journal of Memory and Language*, *42*(3), 406–422. https://doi.org/10.1006/jmla.1999.2682

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1993). Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. *Society for Neuroscience Abstracts*, *19*, 974.

Reilly, J., Kelly, A., Kim, S. H., Jett, S., & Zuckerman, B. (2019). The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry. *Behavior Research Methods*, *51*(2), 865–878. https://doi.org/10.3758/s13428-018-1134-4

Reitan, R. M. (1958). Validity of the Trail Making Test as an Indicator of Organic Brain Damage. *Perceptual and Motor Skills*, *8*(3), 271–276. https://doi.org/https://doi.org/10.2466/pms.1958.8.3.271

Rennies, J., Best, V., Roverud, E., & Kidd, G. (2019). Energetic and Informational Components of Speech-on-Speech Masking in Binaural Speech Intelligibility and Perceived Listening Effort. *Trends in Hearing*, *23*, 1–21. https://doi.org/10.1177/2331216519854597

Richer, F., & Beatty, J. (1985). Pupillary Dilations in Movement Preparation and Execution. *Psychophysiology*, *22*(2), 204–207. https://doi.org/https://doi.org/10.1111/j.1469-8986.1985.tb01587.x

Roberts, B., Summers, R. J., & Bailey, P. J. (2011). The intelligibility of

noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes. *Proceedings of the Royal Society of London, Series B: Biological 2 Sciences*, *278*(1711), 1595–1600. https://doi.org/10.1098/rspb.2010.1554

Roebel, A. (2010). A Shape-Invariant Phase Vocoder for Speech Transformation. *Int. Conference on Digital Audio Effects (Dafx-10)*, 1–8.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Phil. Trans. R. Soc. Lond. B*, *336*(1278), 367–373. https://doi.org/10.1098/rstb.1992.0070

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America*, *106*(6), 3629–3636. https://doi.org/10.1121/1.428215

Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *International Journal of Audiology*, *58*(7), 1–15. https://doi.org/10.1080/14992027.2018.1551631

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, *7*(July), 31. https://doi.org/10.3389/fnsys.2013.00031

Rönnberg, 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*(March), S99–S105. https://doi.org/10.1080/14992020802301167

Saito, S., & Baddeley, A. D. (2004). Irrelevant sound disrupts speech production: Exploring the relationship between short-term memory and experimentally induced slips of the tongue. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *57*(7), 1309–1340.

https://doi.org/10.1080/02724980343000783

Salthouse, T. A. (1996). The Processing-Speed Theory of Adult Age Differences in Cognition. *Psychological Review*, *103*(3), 403–428. https://doi.org/10.1037/0033-295X.103.3.403

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics*, *71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

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*(5), 1230–1240. https://doi.org/10.1044/1092-4388(2009/08-0111)

Sato, H., Morimoto, M., & Wada, M. (2012). Relationship between listening difficulty rating and objective measures in reverberant and noisy sound fields for young adults and elderly persons. *The Journal of the Acoustical Society of America*, *131*(6), 4596–4605. https://doi.org/10.1121/1.4714790

Sauppe, S. (2017). Symmetrical and asymmetrical voice systems and processing load: Pupillometric evidence from sentence production in Tagalog and German. *Language*, *93*(2), 288–313. https://doi.org/10.1353/lan.2017.0015

Schlueter, A., Brand, T., Lemke, U., Nitzschner, S., Kollmeier, B., & Holube, I. (2015). Speech perception at positive signal-to-noise ratios using adaptive adjustment of time compression. *The Journal of the Acoustical Society of America*, *138*(5), 3320–3331. https://doi.org/10.1121/1.4934629

Schneider, B. A., Daneman, M., & Murphy, D. R. (2005). Speech comprehension difficulties in older adults: Cognitive slowing or age-related changes in hearing? *Psychology and Aging*, *20*(2), 261–271. https://doi.org/10.1037/0882-7974.20.2.261

Schneider, B. A., Li, L., & Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology*, *18*(7), 559–572. https://doi.org/10.3766/jaaa.18.7.4

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*. MIT press.

Schoof, T., & Rosen, S. (2014). The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners. *Frontiers in Aging Neuroscience*, *6*, 1–14. https://doi.org/10.3389/fnagi.2014.00307

Sergeyenko, Y., Lall, K., Liberman, M. C., & Kujawa, S. G. (2013). Age-Related Cochlear Synaptopathy: An Early-Onset Contributor to Auditory Functional Decline. *Journal of Neuroscience*, *33*(34), 13686–13694. https://doi.org/10.1523/JNEUROSCI.1783-13.2013

Shannon, R. V., Zeng, F.-g., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Simantiraki, O., Cooke, M., & King, S. (2018). Impact of different speech types on listening effort. *Proc. Interspeech*, 2267–2271. https://doi.org/10.21437/Interspeech.2018-1358

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology*, *26*(2), 183–196. https://doi.org/10.3766/jaaa.26.2.7

Smiljanic, R., & Bradlow, A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Lang Linguist Compass*, *3*(1), 236–264. https://doi.org/10.1111/j.1749-818X.2008.00112.x

Strycharczuk, P., & Scobbie, J. M. (2017). Fronting of Southern British English high-back vowels in articulation and acoustics. *Journal*

of the Acoustical Society of America, *142*(1), 322–331. https://doi.org/10.1121/1.4991010

Sundberg, J., & Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The Journal of the Acoustical Society of America*, *120*(1), 453–457. https://doi.org/10.1121/1.2208451

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *Journal of the Acoustical Society of America*, *85*(2), 901–906. https://doi.org/10.1121/1.397562

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech, Language, and Hearing Research*, *39*(3), 494–509. https://doi.org/https://doi.org/10.1044/jshr.3903.494

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective and Behavioral Neuroscience*, *16*(4), 601–615. https://doi.org/10.3758/s13415-016-0417-4

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, *53*(4), 510–540. https://doi.org/10.1177/0023830910372495

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*, 1–4. https://doi.org/10.3389/fnhum.2014.00577

Verhelst, W., & Roelands, M. (1993). An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. *Proc. ICASSP*, 554–557. https://doi.org/10.1109/ICASSP.1993.319366

Verney, S. P., Granholm, E., & Dionisio, D. P. (2001). Pupillary responses and processing resources on the visual backward masking task. *Psychophysiology*, *38*(1), 76–83. https://doi.org/10.1017/S0048577201990195

Versfeld, N. J., & Dreschler, W. A. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *J. Acoust. Soc. Am.*, *111*(1), 401–408. https://doi.org/10.1121/1.1426376

Wagener, K., Kühnel, V., & Kollmeier, B. (1999). Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test. *Zeitschrift Für Audiologie*, *38*(1), 1–32.

Wagner, A. E., Nagels, L., Toffanin, P., Opie, J. M., & Başkent, D. (2019). Individual Variations in Effort: Assessing Pupillometry for the Hearing Impaired. *Trends in Hearing*, *23*, 1–18. https://doi.org/10.1177/2331216519845596

Watkins, A. J. (2005). Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, *118*(1), 249–262. https://doi.org/10.1121/1.1923369

Webster, J. C., & Klumpp, R. G. (1962). Effects of Ambient Noise and Nearby Talkers on a Face-to-Face Communication Task. *The Journal of the Acoustical Society of America*, *34*(7), 936–941. https://doi.org/10.1121/1.1918224

Wegel, R. L., & Lane, C. E. (1924). The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physical Review*, *23*(2), 266. https://doi.org/https://doi.org/10.1103/PhysRev.23.266

Wendt, D., Hietkamp, R. K., & Lunner, T. (2017). Impact of noise and noise reduction on processing effort: A pupillometry study. *Ear and Hearing*, *38*(6), 690–700. https://doi.org/10.1097/AUD.0000000000000454

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, *369*, 67–78. https://doi.org/10.1016/j.heares.2018.05.006

Wendt, D., Lunner, T., Książek, P., & Alickovic, E. (2020). *Method for adjusting hearing aid configuration based on pupillary information* (pp. 1–26).

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin*, *120*(2), 272–292. https://doi.org/10.1037/0033-2909.120.2.272

Wieland, E. A., Burnham, E. B., Kondaurova, M., Bergeson, T. R., & Dilleya, L. C. (2015). Vowel Space Characteristics of Speech Directed to Children With and Without Hearing Loss. *Journal of Speech, Language, and Hearing Research*, *58*(2), 254–267. https://doi.org/10.1044/2015_JSLHR-S-13-0250

Wingfield, A. (2016). Evolution of Models of Working Memory and Cognitive Resources. *Ear & Hearing*, *37*, 35S–43S. https://doi.org/10.1097/AUD.0000000000000310

Wingfield, A., Peelle, J. E., & Grossman, M. (2003). Speech Rate and Syntactic Complexity as Multiplicative Factors in Speech Comprehension by Young and Older Adults. *Aging, Neuropsychology, and Cognition*, *10*(4), 310–322. https://doi.org/10.1076/anec.10.4.310.28974

Wingfield, A., Tun, P. A., Koh, C. K., & Rosen, M. J. (1999). Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech. *Psychology and Aging*, *14*(3), 380–389. https://doi.org/10.1037/0882-7974.14.3.380

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood: What it is and how it interacts with cognitive performance. *Current Directions in Psychological Science*, *14*(3), 144–148. https://doi.org/

10.1111/j.0963-7214.2005.00356.x

Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, *20*, 233121651666972. https://doi.org/10.1177/2331216516669723

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America*, *131*(2), 1465–1479. https://doi.org/10.1121/1.3672705

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The Impact of Auditory Spectral Resolution on Listening Effort Revealed by Pupil Dilation. *Ear and Hearing*, *36*(4), e153–e165. https://doi.org/10.1097/AUD.0000000000000145

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing*, *22*, 1–32. https://doi.org/10.1177/2331216518800869

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*, 459–482. https://doi.org/10.1002/cne.920180503

Zahorik, P., & Brandewie, E. J. (2016). Speech intelligibility in rooms: Effect of prior listening exposure interacts with room acoustics. *The Journal of the Acoustical Society of America*, *140*(1), 74–86. https://doi.org/10.1121/1.4954723

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil Response as an Indication of Effortful Listening: The Influence of Sentence Intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

Appendix

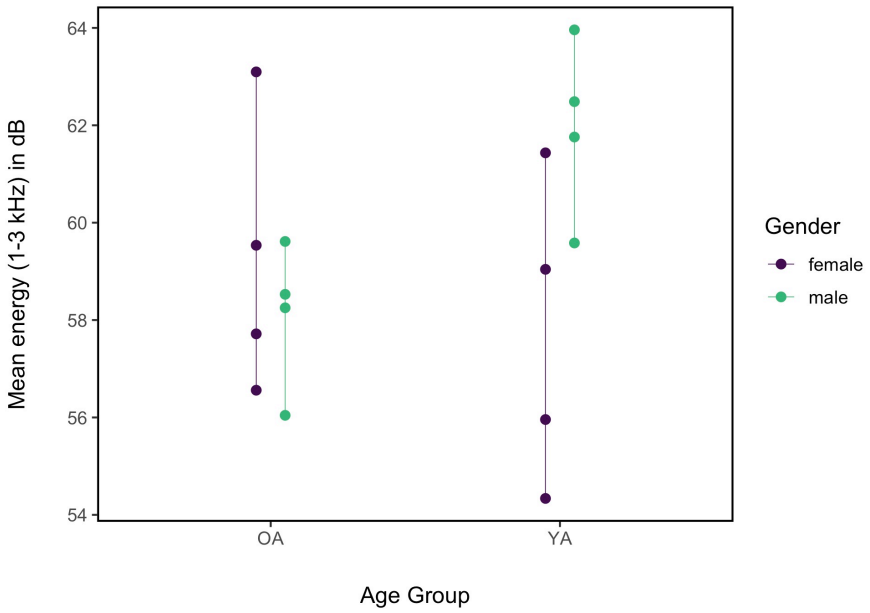

Figure 34: *Mean energy for all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2.*

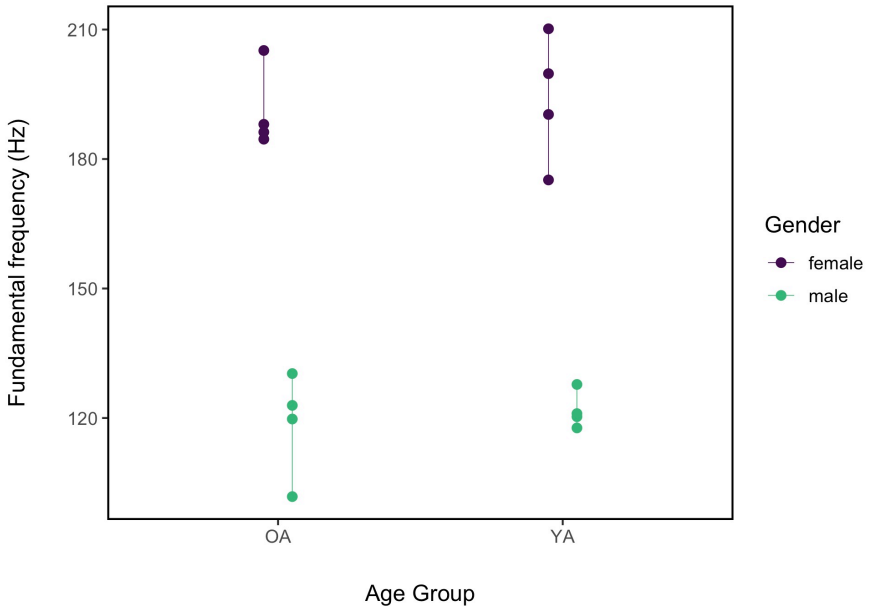

Figure 35: *Fundamental frequency of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2.*

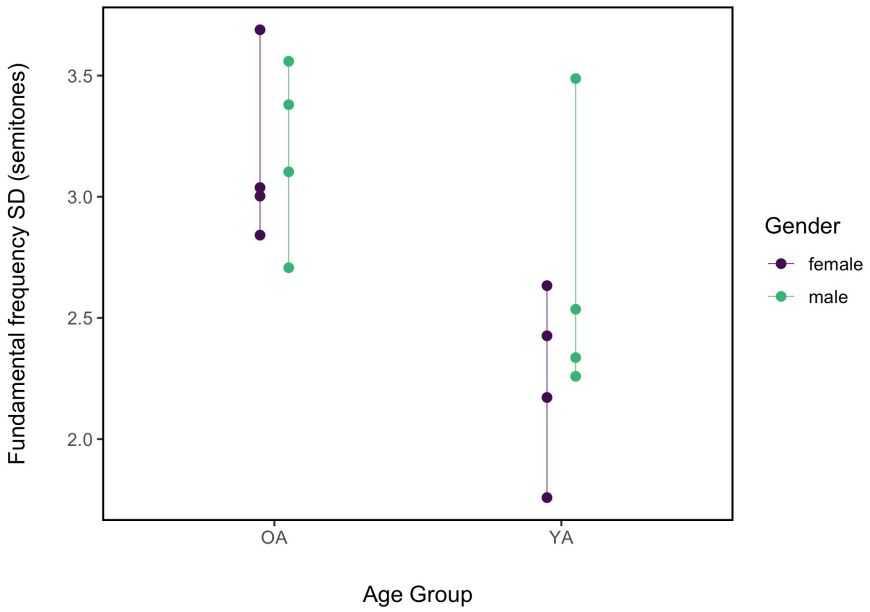

Figure 36: *Fundamental frequency standard deviation (SD) of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2.*

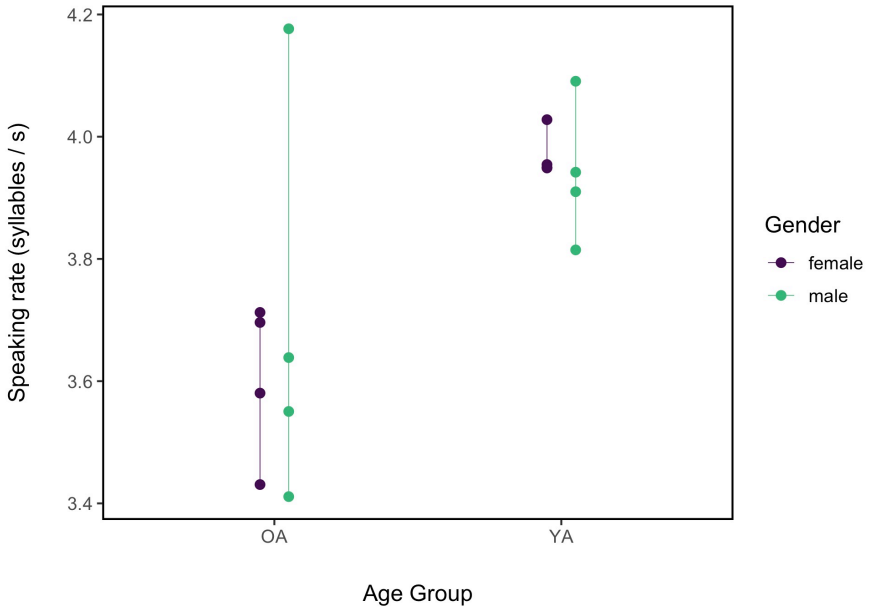

Figure 37: *Speaking rate of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all 192 sentences used in Chapter 2.*

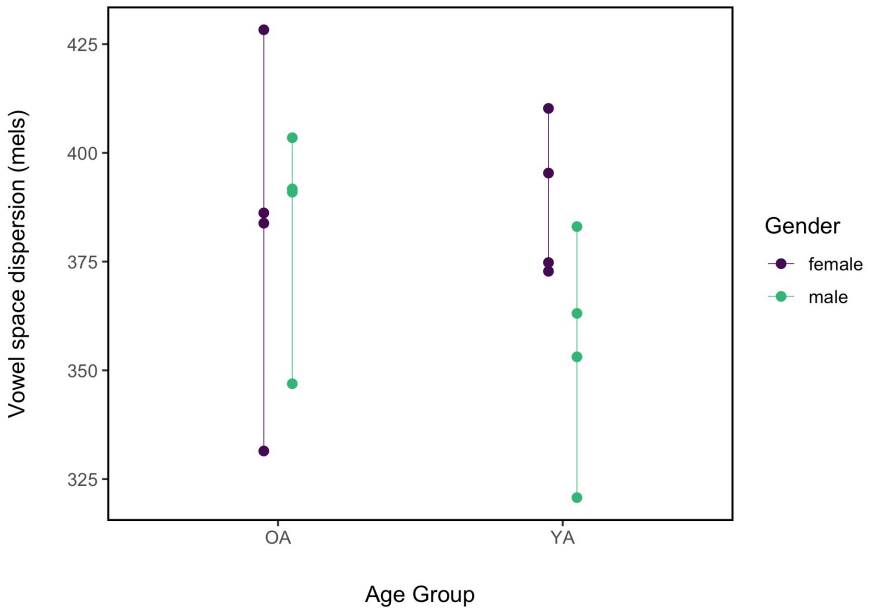

Figure 38: *Vowel space dispersion of all talkers, displayed by age group (OA = older adults, YA = younger adults) and gender. Points indicate individual talker means across all vowel productions and vowel categories.*

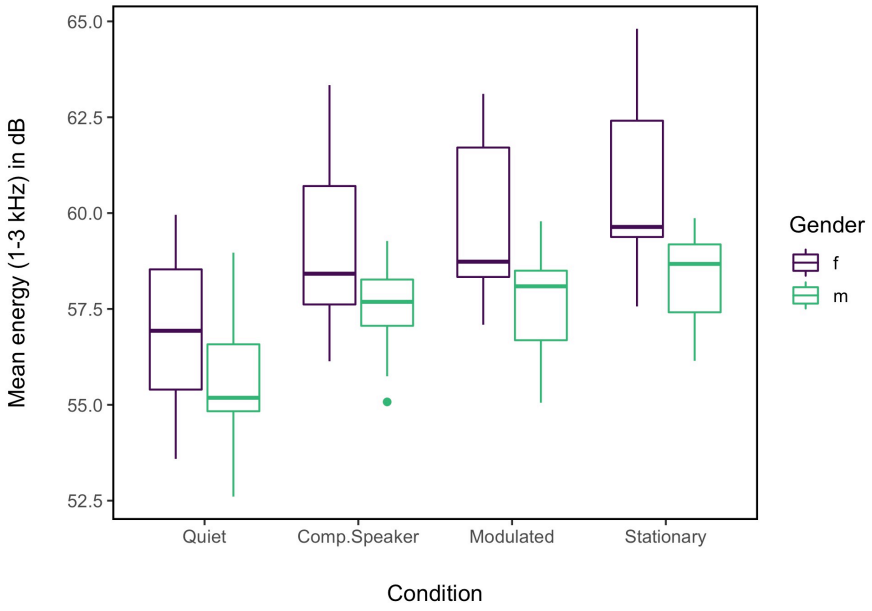

Figure 39: *Distributions of mean energy in each condition (Chapter 4), separate for female and male participants.*

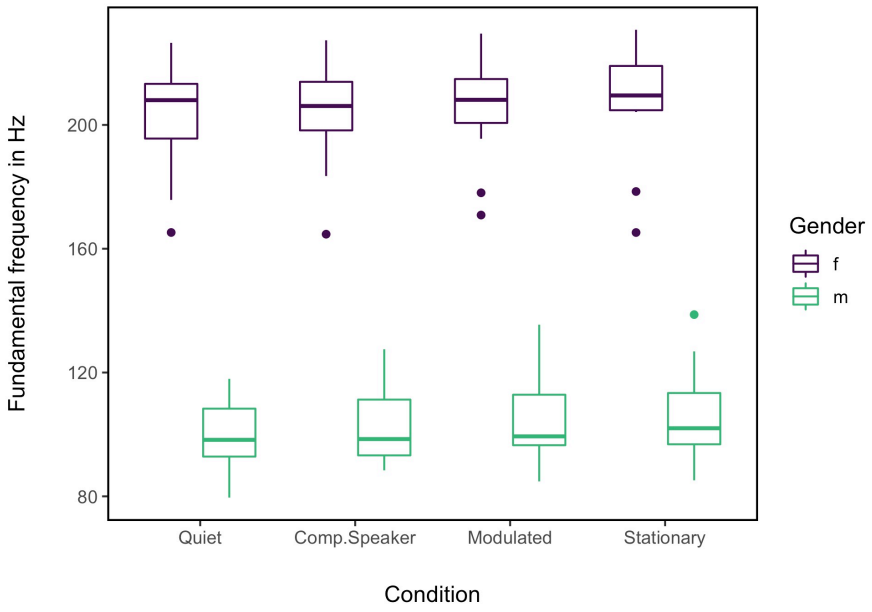

Figure 40: *Distributions of fundamental frequency in each condition (Chapter 4), separate for female and male participants.*

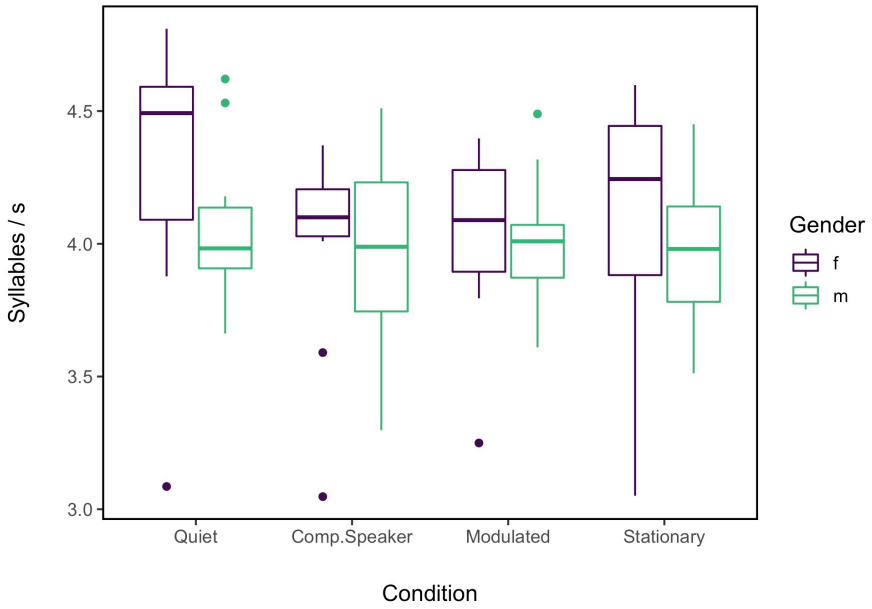

Figure 41: *Distributions of speaking rate in each condition (Chapter 4), separate for female and male participants.*

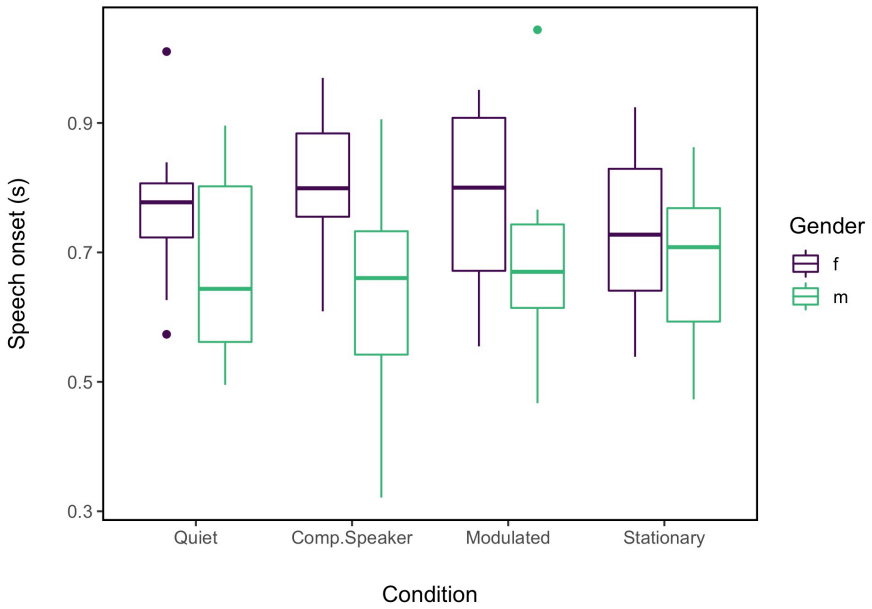

Figure 42: *Distributions of speech onset in each condition (Chapter 4), separate for female and male participants.*

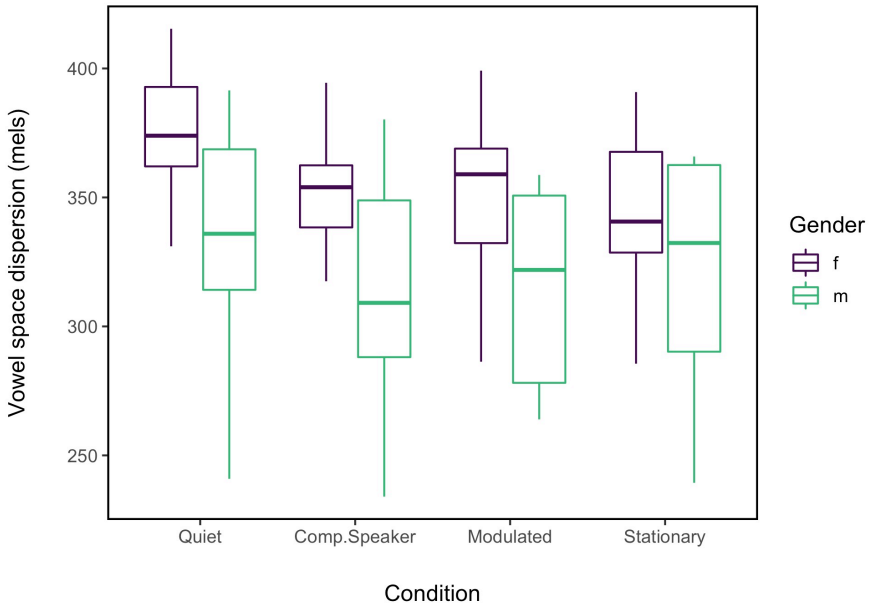

Figure 43: *Distributions of vowel space dispersion in each condition (Chapter 4), separate for female and male participants.*