

# Evaluation of pupillometry as a diagnostic tool

**Neagu, Mihaela-Beatrice**

*Publication date:*
2022

*Document Version*
Publisher's PDF, also known as Version of record

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 54

*Mihaela-Beatrice Neagu*

# Evaluation of pupillometry as a diagnostic tool

**DTU Health Tech**
Department of Health Technology

# Evaluation of pupillometry as a diagnostic tool

PhD thesis by
Mihaela-Beatrice Neagu

Preliminary version: June 17, 2022


Technical University of Denmark

2022

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Health Technology, Technical University of Denmark.

The project was financed by William Demant Foundation.

## Supervisors

**Prof. Torsten Dau**
Hearing Systems Section
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark
**Principal Scientist Dorothea Wendt**
Eriksholm Research Centre
Snekkersten, Denmark
**Assist. Prof. Abigail Anne Kressner**
Hearing Systems Section
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark

# Abstract

Pupillometry is a widely used tool for measuring listening effort in hearing science. Changes in pupil dilation during speech-in-noise tasks have been linked to differences in speech intelligibility, masker type, hearing status and noise reduction schemes. Recent technological progress has allowed for pupillometry's use on a larger scale, thus opening its potential for clinical application where it can be relevant to measure hearing-impaired (HI) listeners' effort expenditure. However, the link between effort and the pupil response has not yet been validated for the individual listener, nor has the method's test-retest reliability been thoroughly evaluated. Moreover, an understanding of the impact of listener factors on the variation of the pupil response observed during speech recognition is still missing and little is known about the relationship between the evoked pupil response and the subjective effort investment perceived by the listener. This thesis assessed the individual pupillary response as an outcome measure of listening effort by investigating its reliability and sensitivity within speech-in-noise tasks.

The first study of this thesis assessed the reliability of a broad range of pupil features in normal-hearing (NH) and hearing-impaired (HI) individuals while performing a speech-in-noise test. It was found that some features of the pupil response (the rise and fall around the peak and the mean pupil dilation) showed high reliability independent of the listener group, while other pupil features' reliability varied depending on the listener group. Furthermore, a cluster analysis performed on the temporal characteristics of the pupil response showed that the signal-to-noise ratio (SNR) was, contrary to expectation, not a good predictor to cluster these pupil features.

The second study expanded the reliability analysis to include more SNRs, multiple visits and different normalization procedures. The results showed that data normalization procedures have a strong impact on the reliability of the pupil features. In particular, subtractive baseline correction in combination with a range normalization applied to the individual pupil response across all visits resulted in the highest reliability. Furthermore, the results suggested that the SNR and the number of visits only have a minor impact on the reliability of the pupil response. The most reliable pupil features were the traditional mean pupil dilation (MPD) and peak pupil dilation (PPD). The outcome of the first and second studies helped to identify test conditions and parameters as well as the pre-processing data analysis under which highly reliable pupil features can

be obtained.

The third study explored the impact of individual listener factors, such as age, hearing status, cognitive abilities, motivation, and fatigue, on PPD and MPD and their variation across multiple visits. Furthermore, this study examined the effect of these listener factors on the dynamic range of the pupil response measured in several tasks (a speech-in-noise task, a cognitive task and at rest). The results identified motivation as the main listener factor affecting PPD and MPD. In addition, PPD was modulated by daily-life fatigue and age. At the same time, MPD was highly affected by the interaction of cognitive abilities with visits, resulting in changes in MPD across visits based on the listeners' cognitive abilities.

The final study investigated the sensitivity of the pupil response to changes in SNR and its relation to the perceived listening effort in a novel paired-sentence paradigm. The concept of a 'just noticeable difference in effort' (JND in effort) was introduced, reflecting the minimum increase in SNR necessary for a person to perceive a difference in perceived effort. The results were related to corresponding pupil responses at the JND in effort and two additional behavioural JNDs, the 'JND in clarity' and the 'JND in meaning' that have been reported in earlier studies. The results showed that, on average, the JND in effort was between the JND in clarity and JND in meaning but varied substantially across individuals. The pupil responses showed a difference between the pairs of sentences at the SNRs corresponding to the JND in effort and the JND in meaning for particular time-windows (i.e., retention period and listening time, respectively), whereas no difference between the pairs was found at the SNRs corresponding to the JND in clarity.

Together, the findings of this thesis suggest that pupillometry has potential for future applicability as a clinical measure of individual listening effort. More specifically, depending on the test conditions (e.g., the SNR) and the normalization procedures, highly reliable pupil features can be obtained, which is a prerequisite for a clinically feasible measure. However, listener factors have been shown to contribute to the variability in the pupil response, meaning that such factors need to be considered when interpreting the pupil response. Finally, the assessment of the behavioral JND in effort appears very relevant for the interpretation of the individual's pupil response as a marker of effort investment. Overall, this work may provide a valuable basis for developing a clinical tool to assess listening effort, which will facilitate more comprehensive evaluations of speech communication that extend beyond audibility and speech intelligibility.

# Resumé

Pupillometri er en udbredt metode til måling af lytteanstrengelse indenfor høreforskning. Ændringer i pupiludvidelse under tale-i-støj-opgaver er blevet forbundet med forskelle i taleforståelighed, maskeringstype, hørestatus og støjreduktionsalgoritmer. Nyere tids teknologiske fremskridt har gjort det muligt at anvende pupillometri i større skala, hvilket har udvidet dets potentiale for klinisk brug , hvor det kan være relevant at måle hørehæmmede (eng.: *hearing-impaired*, HI) personers lytteindsats. Sammenhængen mellem indsats og pupiludvidelse er imidlertid endnu ikke valideret for individer, ligesom metodens test-retest-reliabilitet ikke er blevet grundigt evalueret. Ydermere mangler der fortsat en forståelse for påvirkningen af lytterfaktorer på variationen af pupilreaktioner observeret under talegenkendelse, og der vides kun lidt om forholdet mellem fremkaldt pupilresponsog den subjektive lytteinvestering oplevet af den lyttende. Denne afhandling vurderede individuelle pupilreaktioner som resultat af lytteanstrengelse ved at undersæge målingens reliabilitet og sensitivitet indenfor tale-i-støj-opgaver.

Afhandlingens første studie vurderede reliabiliteten af en bred vifte af pupilfunktioner hos normalthørende (NH) og hørehæmmede (HI) individer, mens de udførte en tale-i-støj-test. Det blev fundet, at visse funktioner af pupilreaktioner (stigning og fald omkring toppen samt den gennemsnitlige pupiludvidelse) viste høj reliabilitet uafhængigt af deltagergruppe, mens andre pupilfunktioners reliabilitet varierede afhængigt af deltagergruppe. Yderligere viste en clusteranalyse foretaget på de temporale karakteristika af pupilresponset, at signal-støj-forholdet (eng.: *signal-to-noise ratio*, SNR) – i modsætning til det forventede – ikke var en god prædiktor til at samle disse pupilfunktioner.

Det andet studie udvidede reliabilitetsanalysen til at inkludere flere SNR'er, gentagne besøg og forskellige normaliseringsprocedurer. Resultaterne viste, at datanormaliseringsprocedurer har stor indflydelse på reliabiliteten af pupilfunktionerne. I særdeleshed gav subtraktiv baseline -korrektion i kombination med range normalization fokuseret på den individuelle pupil over alle besøg den højeste reliabilitet. Yderligere indikerede resultaterne, at SNR samt antal besøg kun har mindre indflydelsepå pupilreaktionens respons. De mest pålidelige (eng.: *reliable*) pupilfunktioner var den traditionelle gennemsnitlige pupiludvidelse (eng.: *mean pupil dilation*, MPD) and maksimal pupiludvidelse (eng.: *peak pupil dilation*, PPD). Udfaldet af første og andet studie bidrog til at identificere testforhold og -parametre såvel som den preprocesseringsdataanalyse hvorunder stærkt pålidelige pupilfunktioner kan måles.

Det tredje studie udforskede betydningen af individuelle personfaktorer såsom aldr, hørestatus, kognitive evner, motivation og træthed på PPD og MPD samt deres variation over flere besøg. Deruduver undersøgte studiet effekten af disse personfaktorer på den dynamiske rækkevidde af pupilreaktionerne målt under flere opgaver (en tale-i-støj-opgave, en kognitiv opgave og i hvile). Resultaterne identificerede motivation som den væsentligste personfaktor til at påvirke PPD og MPD. Desuden var PPD moduleret af hverdags-træthed og alder. Samtidigt viste MPD sig at være stærkt påvirket af interaktionen mellem kognitive evner og besøg, hvilket resulterede i ændringer i MPD henover besøgene baseret på deltagerens kognitive evner.

Det sidste studie undersøgte følsomheden af pupilreaktionerne overfor ændringer i SNR og dennes relation til den oplevede lytteanstrengelse i et nyudviklet parret-sætning-paradigme. Konceptet 'netop mærkbar forskel i anstrengelse' (eng.: *'just noticeable difference in effort'*, JND i anstrengelse) blev introduceret som et udtryk for den mindste ændring i SNR der var nødvendig for, at en person mærkede en ændring i sin oplevede lytteanstrengelse. Resultaterne var relateret til korresponderende pupilreaktioner ved JND i anstrengelse og to yderligere adfærdsmæssige JND'er: 'JND i klarhed' og 'JND i betydning' som er blevet rapporteret i tidligere studier. Resultaterne vist, at JND i anstrengelse i gennemsnai var mellem JND i klarhed og JND i betydning, men varierede betydeligt mellem individer. Pupilreaktionerne viste en forskel mellem sætningsparved SNR'erne svarende til JND i anstregelse og JND i betydning for specifikke tidsvinduer (hhv. tilbageholdelsestid og lyttetid), hvorimod ingen forskel blev fundet mellem parrene ved SNR svarende til JND i klarhed.

Alt i alt indikerer resultaterne i denne afhandling, at pupillometri har potentiale for fremtidig brug som klinisk måling af individuel lytteanstrengelse. Mere specifikt kan meget troværdige pupilfunktioner måles - afhængigt af testsituationen (f. eks. SNR) og normaliseringsprocedurerne – hvilket er en forudsætning for en klinisk brugbar måling. Personlige deltagerfaktorer har imidlertid vist sig at bidrage til variabiliteten i pupilreaktionerne, hvilket betyder at sådanne faktorer skal medregnes, når man fortolker pupilreaktionene. Sluttelig forekommer vurderingen af adfærdsmæssig JND i anstrengelse yderst relevant for fortolkningen af et individs pupilreaktion som en markør af lytteindsats.Overordnet kan dette værk give en værdifuld basis for at udvikle kliniske redskaber til at vurdere lytteanstrengelse, hvilket vil facilitere mere omfattende evalueringer af talekommunikation der rækker ud over hørbarhed og taleforståelighed.

# Acknowledgments

My PhD studies have been an intense journey with challenging periods at times. I had the opportunity to apply my knowledge in a field with huge benefit to people with hearing impairment, and I am grateful for getting the chance to bring a contribution to this field by conducting experimental work using new technologies and by developing new scientific, technical and interpersonal skills. I would like to express my gratitude to people I had the chance to work with, coming from different backgrounds to colleagues, family and friends.

First, I would like to sincerely thank my supervisors, Torsten Dau, Dorothea Wendt and Abigail Anne Kressner, for giving me this opportunity and guiding me through the entire process. Although there were, at times, difficult periods with misalignments in opinions, openness to discussions and communication brought us a step further in my PhD journey. I am grateful for learning from your vast knowledge and for improving my scientific skills, which gave me the chance to progress.

I would also like to thank the other members of the PUPILS team, more specifically Per Bækgaard and Helia Relaño Iborra, for their support within the technical aspects of the field and statistical modelling. Your help was very important in shaping different chapters of this story and in wrapping up this entire process.

Special thanks go to Rikke Skovhøj Sørensen for her help with the participants in the experiments, measuring audiograms, scoring hours of HINT sentences, also supporting with the translation of the abstract to Danish.

Next, I would like to thank the William Demant Foundation for funding this project. Of course, all these studies would not have been possible without the people participating in the experiments. Thank you for taking the time and effort to participate in these studies. A big thank you also to Andreas Madsen for help, at times, with setting-up experiments and piloting.

I would also like to thank all my colleagues at the Hearing Systems and the laboratory manager, for their support in setting-up experiments and for high-

level academic discussions. I will bring here special thanks to my colleagues from office 128. We not only shared an office, but we shared our doubts, our frustrations, our uncertainty, our victories, our ideals and our dreams.

Lastly, I am highly grateful to my friends and family for their encouragement and their great support along the way. Thank you so much for always being there for me.

# Related publications

## Journal papers

- Neagu, M. B., Kressner, A. A., Relaño Iborra, H., Bækgaard, P., Dau, T., and Wendt, D. (**2022a**). "Exploring the reliability of pupillometry under different task demands, normalization procedures and at multiple visits" Trends in Hearing (*under revision*)

- Neagu, M. B., Kressner, A. A., Relaño Iborra, H., Bækgaard, P., Dau, T., and Wendt, D. (**2022b**). "Towards a better understanding of the impact of listener factors on pupil responses in a speech in noise paradigm" (*in preparation*)

- Neagu, M. B., Kressner, A. A., Relaño Iborra, H., Bækgaard, P., Dau, T., and Wendt, D. (**2022c**). "Exploring the relationship between perceptual effort investment and the evoked pupil response during speech perception" (*in preparation*)

- Rico Jensen, K. M., Neagu, M. B., Kressner, A. A., Relaño Iborra, H., Bækgaard, P., Dau, T., and Wendt, D. (**2022**). "Investigating the consistency of the evoked pupil response in hearing-impaired listeners during a speech-in-noise task" (*in preparation*)

- Relaño Iborra, H., Wendt, D., Neagu, M. B., Kressner, A. A., Dau, T. and Bækgaard, P. (**2022**). "Baseline pupil size modulates the temporal dynamics of the task-evoked pupillary response in a speech-in-noise task" Trends in Hearing (*under revision*)

## Conference papers

- Neagu, M. B., Dau, T., Hyvärinen, P., Bækgaard, P., Lunner, T. and Wendt, D. (**2019**). "Investigating pupillometry as a reliable measure of individual's

listening effort” Proc. of International Symposium on Auditory and Audiological Research. 7, 365-372.

# Contents

# 1

# General introduction

Speech communication is an essential part of human interaction. In everyday life, successful communication is reliant on an ability to understand speech in listening situations that can be challenging. Adverse listening situations can be taxing for all, but particularly for people with hearing impairment (HI). This is reportedly the case even when people with HI use a hearing aid (HA) that provides full audibility, as these listeners nonetheless experience additional challenges (Ng et al., 2013).

## 1.1 Listening effort - beyond audibility and speech intelligibility

Whereas eligibility for HAs, and the HA fitting itself, are mainly based on measures of sensitivity and the restoration of it, the evaluation of the benefit the individual listener receives from a HA traditionally revolves around measures of speech understanding. Speech understanding is measured clinically with speech-in-noise tests, which are tests designed to measure the proportion of correctly repeated speech items, usually single words or single sentences, in simulated versions of controlled, adverse listening situations (Hagerman, 1984; Nielsen and Dau, 2011; Plomp and Mimpen, 1979). The outcome of such tests is classically reported as a speech reception threshold (SRT), which reflects the signal-to-noise ratio (SNR) at which 50% of the items have been correctly recognised. However, for most clinically available tests, SRT estimates are obtained at relatively low SNRs, where communication would be challenged to an unrealistic level (Smeds et al., 2015). Moreover, clinical speech-in-noise tests tend to overestimate speech understanding performance when compared to evaluations conducted in more realistic listening scenarios (Best et al., 2015; Mansour et al., 2021; Miles et al., 2022).

Even when clinically measured speech intelligibility outcomes indicate that hearing aids (HA) restore speech understanding for an individual with HI, these

listeners often still report having difficulties understanding speech in their everyday lives, especially with regard to how effortful it can be. In fact, several studies have demonstrated that hearing loss can lead to reduced speech communication in everyday listening scenarios, increased cognitive demands and slowed down speech processing (Duquesnoy, 1983; Mattys et al., 2012; Plomp, 1986; Wendt et al., 2014). These outcomes can eventually have psychosocial consequences, such as increased levels of mental distress, withdrawal from social situations and isolation due to an increased effort required for listening (Htu et al., 1994; Kramer et al., 2006; Weinstein and Ventry, 1982). This growing body of literature highlights the importance of listening effort in speech communication. Therefore, to resolve the difficulties listeners with HI face in everyday speech communication, measures of performance must extend beyond audibility and speech intelligibility to include additional aspects of speech communication, such as listening effort.

There are various ways to measure listening effort. One of them is based on pupillometry, which has recently gained popularity within the hearing community. The pupil response is considered to be an objective indicator of effort and, thus, has the potential to provide estimates of effort with a smaller bias than its subjective counterparts. To date, changes in pupil dilation have been linked to speech intelligibility, masker type, hearing status and hearing aid signal processing (Koelewijn et al., 2012b; Kramer et al., 1997; Kuchinsky et al., 2013; Ohlenforst et al., 2017a,b; Wendt et al., 2017; Zekveld et al., 2010, 2011). However, these studies have solely been based on an analysis at a listener group level (i.e., the pupil responses in a particular condition have been averaged across listeners, and the mean responses have been compared across conditions). While such investigations are valuable for understanding the fundamental role of listening effort in speech communication and for illustrating the relationships between specific aspects of speech communication and the pupil response, they do not facilitate the characterisation of a specific individual's pupil response to a specific listening scenario. Indeed, an individual approach to the characterisation of the pupil response is crucial when evaluating the potential of pupillometry as a clinical tool. The main objective of this thesis was to better understand the characteristics of the pupil responses in individuals.

## 1.2 Physiology of pupil dilation

It has been demonstrated that physiological functions of the eye, and in particular, pupil dilation, are regulated by the autonomic nervous system (ANS) (Bremner, 2009; May et al., 2019; Wang et al., 2016). ANS plays an important role in maintaining stability and balance in the body. Its activity consists of both sympathetic and parasympathetic responses. The part of the ANS involved in the wakefulness state (i.e., preparing the body for high-energy activity and a "fight" response) is the sympathetic activity. In contrast, parasympathetic activity is involved in homeostasis (i.e., keeping the body in a stable condition and a "flight" response). Physiologically, the pupil size is controlled by the iris sphincter (constrictor) muscle and the iris dilator muscle. The iris dilator muscle is controlled by the sympathetic nervous system, connecting a status of arousal with an enlarged pupil size. Conversely, the iris sphincter muscle is innervated by the parasympathetic nervous system, connecting a state of rest with a relatively small pupil size (Borgdorff, 1975; Glasser, 2011). Thus, the balance between the sympathetic and parasympathetic nervous systems determines the pupil size (Loewenfeld, 1993; Wang et al., 2018b), as the two systems are complementary.

The relative contributions to the nervous system, however, can vary as a function of luminance, cognitive activity and fatigue (Steinhauer et al., 2004; Wang et al., 2018a,b). Reimer et al., 2016 suggested that non-luminance changes in pupil size might be determined by the locus coeruleus (LC), which has been shown to be a noradrenergic source for the cortex (Aston-Jones and Cohen, 2005; Carter et al., 2010; Jones, 2004; Lee and Dan, 2012). Several studies showed that LC activity determines parasympathetic inhibition which thereby causes inhibition of the constrictor muscle of the pupil and ultimately leads to a dilation of the pupil (Eckstein et al., 2017; Wang et al., 2016). Broadly speaking, pupil dilation has been shown to be correlated with several different cognitive processes: attention (Koelewijn et al., 2012a), task demands (Beatty, 1982; Janisse, 1977), memory (Van Der Meer et al., 2010; Zekveld et al., 2011) and mental load (Just et al., 2010; Kramer et al., 2013).

## 1.3 Pupillometry in hearing science

The pupil response has been shown to be sensitive to changes in the allocation of cognitive resources in response to auditory stimuli (Beatty and Lucero-Wagoner, 2000; Hepach and Westermann, 2016; Laeng et al., 2012; Schmidtke, 2018). Importantly, Koelewijn et al., 2012a showed that in conditions with similar levels of speech intelligibility, the listening effort can vary depending on the linguistic properties of the speech. Moreover, Wendt et al., 2017 demonstrated that listening effort is highly affected by the masker type and its semantic content. Wendt et al., 2017 also observed an inverted U-shaped relationship between peak dilation and speech intelligibility both in listeners with NH and HI, across a broad range of SNRs. Furthermore, Ohlenforst et al., 2017b showed that listeners with HI invested more effort than their NH counterparts in situations where the level of speech intelligibility was relatively high. As a consequence, HI listeners might have fewer cognitive resources available in more adverse listening situations. This can lead to a "give up effect" where less effort is invested in an increasingly difficult listening situation. Ohlenforst et al., 2017a and Wendt et al., 2017 further demonstrated that HA signal processing can reduce listening effort by applying noise reduction schemes.

However, a major challenge is that a substantial variability typically characterises the pupil response within and across individual listeners. Various stimulus-related and listener-related factors contribute to such variability. While various studies investigated how stimulus-related factors, including SNR or noise characteristics, affect the task demand and, thereby, the pupil response, only very little is known about the variability within individuals when stimulus-related factors are controlled for. It has not yet been evaluated systematically how reliable specific pupil features are when retesting individuals under the same conditions. Only a few studies examined the test-retest reliability of markers of listening effort. Alhanbali et al., 2019 studied the reliability of different markers of effort and found that the peak pupil dilation (PPD) and the mean pupil dilation (MPD) of the pupil response showed higher reliability than other physiological measures such as skin conductance or alpha power. Similarly, Giuliani et al., 2020 reported the highest reliability for pupillometry among different measures tested, such as perceived effort, reaction time or skin conductance, even though the corresponding level of reliability was only moderate. However, not much is known about the test parameters and conditions under

which high reliability can be observed, for example, how the task demand or the normalization procedure affect the reliability of the response.

In an overview study, Zekveld et al., 2018 provided reports of various listener factors affecting the individual pupil response to an auditory task. It was concluded that pupil size could be sensitive to factors such as hearing status, age, and cognitive abilities. At the same time, existing theories on listening effort suggest that the level of motivation, but also the fatigue status of the listener, can have an impact on effort allocation and, hence, on the pupil response in a speech-in-noise task (Brehm and Self, 1989; Pichora-Fuller et al., 2016). Pichora-Fuller et al., 2016 highlighted how task demands and motivation could interactively modulate effort. Furthermore, it has been suggested that fatigue depends on motivation since the expenditure of effort can lead to fatigue if the task is not meaningful to the listener or if it was not self-initiated (Hockey, 2013; Hornsby et al., 2016). This literature emphasises the importance of assessing listener factors when studying listening effort in individuals. Despite this increasing evidence of the impact of listener factors on the pupil response, the results remain nonetheless conflicting, as an increase in a given listening factor leads, in some cases, to an increase in the pupil response and in other cases to a decrease.

Finally, even though listening effort has been studied with different measures (such as self-reported effort, behavioural performance or physiological measures) in many studies (Hornsby, 2013; Larsby et al., 2005; Picou et al., 2011; Zekveld et al., 2010), the discussion about their relationships has been controversial. Measures of listening effort have rarely shown reliable correlations, and some studies even suggest that pupil dilations and subjective ratings of effort are uncorrelated (Alhanbali et al., 2019; Lau et al., 2019; Wendt et al., 2016; Zekveld and Kramer, 2014; Zekveld et al., 2011), possibly because objective measures of effort and subjective self-reports tap into different dimensions of effort (Alhanbali et al., 2019; Francis et al., 2016; Hornsby et al., 2016; Wendt et al., 2017). The sensitivity of an individual listener to a change in behavioural listening effort might vary largely across individual listeners, and the corresponding change in the pupil response may be more strongly correlated with the individual's behavioural response than what is reflected at a given fixed SNR at the 'average' response across listeners.

Thus, overall, various factors contribute to the variability underlying the pupil response, and it seems important to delineate, as much as possible, the

main sources of variability underlying the individual's pupil response. This would allow optimizing stimulation paradigms, task selection and analysis methods and could be the basis for applications of pupillometry in a clinical context.

## 1.4 Overview of the thesis

*Chapter 2* of this thesis assesses the reliability of the pupil responses in normal-hearing and hearing-impaired individuals while performing a speech-in-noise test. The reliability of a broad range of pupil features (i.e., the traditional pupil features PPD and MPD as well as temporal features of the pupil response function extracted using growth curve analysis) is analysed to identify features with high reliability. Furthermore, the impact of HI on the reliability of those features is examined.

*Chapter 3* further investigates the effect of specific parameters and test conditions on the reliability of the pupil response in a speech-in-noise test. The reliability of these features is studied across a broad range of signal-to-noise ratios, across several visits and using different normalization procedures (i.e., baseline correction, range normalization, z-score, and baseline correction combined with range normalization).

*Chapter 4* investigates the impact of listener factors and their relative contributions to the variation in the pupil response. The listener factors, including age, cognitive abilities, fatigue and motivation, are explored in relation to different features of the pupil responses and evaluated across visits. Moreover, the relationships between dynamic ranges of the pupil response measured in three different tasks (i.e., speech-in-noise task, cognitive task, pupil measured at rest in darkness and light) are examined.

*Chapter 5* introduces a novel experimental paradigm to study the relationship between subjective and objective measures of listening effort. A paired-sentence paradigm is employed to estimate the change of the pupil response that corresponds to the behavioural just noticeable difference (JND) of the perceived listening effort. The results are related to corresponding pupil responses at two additional behavioural JNDs, the 'JND in SNR' and the 'just meaningful difference', which have been reported in earlier investigations.

Finally, *Chapter 6* summarizes the main findings of the individual chapters, presents a broader discussion of these contributions and their implications,

and provides an outlook for future work.

# 2

# Investigating pupillometry as a reliable measure of individual's listening effort [a]


## Abstract

Pupillometry as a tool indicating listening effort has been extensively analyzed on a group level, but less is known about how reliable pupil dilation is as an indicator of an individual's listening effort. The aim of this study was to investigate the reliability of the pupil dilation measured during a speech-in-noise task as an indicator of an individual's listening effort. The pupil dilation of 27 normal-hearing (NH) and 24 hearing-impaired (HI) participants was recorded while they performed a speech-in-noise test on two different days. Measures of intraclass correlation coefficient (ICC) absolute agreement were considered in the analysis. The ICC was applied to the peak and mean pupil dilation as well as to the different terms resulting from fitting a third-order orthogonal polynomial within Growth Curve Analysis (intercept, 1st order, 2nd order and 3rd order terms), which are assumed to provide further information about temporal changes of the pupil dilation. High values of test-retest reliability were found on some measures of the pupil response. Furthermore, a Bland-Altman analysis was applied as a graphical representation of the reliability of the pupillometry. The results showed different levels of reliability depending on the different features of the pupil response (slope, rise-fall and mean pupil dilation for the HI listeners; rise-fall, delay and mean pupil dilation for NH).


[a] This chapter is based on Neagu et al., (2019)

## 2.1 Introduction

Pupillometry has been considered as a tool for reflecting listening effort, particularly in HI people who typically have higher listening effort than NH listeners in a given condition (Kramer et al., 2006, Wendt et al., 2016). Changes in listening effort as indicated by changes in the pupil size have been demonstrated on a group level (Zekveld et al., 2010, Wendt et al., 2016). The mentioned studies used speech-in-noise tests in combination with pupillometry to examine the impact of intelligibility, signal-to-noise ratio (SNR) and type of noise on listening effort as indicated by changes in the pupil dilation. However, the reliability of pupillometry as an indicator of individual listening effort has not been systematically studied yet.

The current study investigated the reliability of pupillometry as an objective listening effort measure in individuals, while they perform a speech-in-noise test. The most common methods for assessing test-retest reliability are the Intraclass Correlation Coefficient (ICC), proposed by Hays et al., 1993, and the Bland and Altman, 1986 approach. Alhanbali et al., 2019 showed a good reliability (ICC $>$ 0.85) of the mean and the peak pupil dilation (PPD). However, the reliability of other pupil dilation characteristics, such as time-dependent features of the pupil response was not considered in their study. Therefore, the present study focused on the reliability of the pupil dilation as a measure of listening effort by considering features such as the average height of the pupil response function, the slope, the rise and fall around the inflection point and the inflexions at the extremities of the function. These features were extracted when applying the growth curve analysis (GCA) model developed by Mirman et al., 2008. Furthermore, this study explored the visual representation of the reliability by using the Bland and Altman, 1986 approach describing the individual differences of the two visits against their average. Another element of this study was to perform a cluster analysis on the individual responses of the pupil. The purpose of the cluster analysis was to identify the main features of the pupillary response function that could best characterize listening effort.

## 2.2 Methods

### 2.2.1 Data set

Two different data sets were analysed as reported in Wendt et al., 2018 and Ohlenforst et al., 2018. The first data set was collected by Wendt et al., 2018 for a group of 27 NH listeners while the second data set was recorded by Ohlenforst et al., 2018 for a group of 24 HI listeners. The pupil dilation was recorded while people performed a speech-in-noise test (HINT, Nielsen and Dau, 2011) at 8 different SNRs. Only two subsets were considered for assessing reliability (two out of eight SNRs for each group, NH and HI: 0 dB and 4 dB, each tested at a different date) and three subsets for the cluster analysis (8 dB, 0 dB, -8 dB for NH and HI). Four to six weeks were considered in between the two different dates, to avoid learning effects with respect to the sentence material since the sentences were repeatedly used. A list of 25 sentences per condition was presented to the participants in a block-based design. The pupil data were processed using MATLAB, 2018 and R Core Team, 2019. To remove any initial effects, the first five sentences (out of 25) of the pupil traces from a list were excluded from the analysis. Data cleaning was performed as reported in Wendt et al., 2018. Trials with less than 80% reliable data were removed from the analysis and the other traces were baseline corrected. In total, 40 recordings of each individual were compared between the two dates (2x20x27NH, 2x20x24HI). The mean pupil dilation was calculated as the average pupil dilation over the trials. The PPD was calculated between the 3rd and 8th second of the stimulus presentation as in Zekveld et al., 2010.

### 2.2.2 Growth curve analysis (GCA)

To examine temporal changes of the pupil response function for the two different dates, GCA was applied twice for the 2 different dates. According to Mirman et al., 2008, GCA fits orthogonal polynomial terms to time series data with the purpose of showing different variations in the function among individuals. To describe the shape of the function, three orthogonal polynomials ($p_1$, $p_2$ and $p_3$) were used. Pupil size was considered as a dependent variable in the model, predicted by a series of fixed and random effects (2.1). The temporal features of the pupil response for the two dates extracted through GCA were considered when calculating test-retest reliability. According to Kalénine et al., 2012, the

intercept term represents the averaged height of the pupil response, the linear term reflects the slope, the quadratic term reflects the rise and fall around the central inflection point of the response function, and the cubic term reflects the inflexions at the extremities of the curve referred to as delay in the current study. In other words, an estimate of the 3 coefficients and the intercept were obtained, representing the GCA terms of different orders.

$$\text{pupil} \sim (p_1 + p_2 + p_3) * \text{participant} + (1 + p_1 + p_2 + p_3 | \text{sentence}) \tag{2.1}$$

### 2.2.3 ICC

Intraclass correlation coefficient (ICC) is one of the most used reliability indices in test-retest studies. The ICC can reflect either the degree of consistency or the agreement between measurements. The agreement assumes that the values measured on two different dates are expected to be equal for each respondent. Consistency considers that the values measured on two different occasions are correlated in an additive manner. Thus this measure is less relevant in the current analysis, but is nevertheless still reported. ICC agreement was calculated according to Hays et al., 1993, as reflected in (2.2), where $MS_B$ is the mean square between subjects, $MS_T$ is the mean square between trials, $MS_E$ is the mean square for error and $n$ is the number of subjects.

$$ICC_{\text{agreement}} = \frac{MS_B - MS_E}{MS_B + \frac{MS_T - MS_E}{n}} \tag{2.2}$$

### 2.2.4 Bland-Altman (BA) approach

To apply the BA approach, the first step was to calculate the limits of agreement (LoA) as the mean ± 1.96 standard deviation of the two similarly conditioned tests. The plot is designed to show the difference between the two visits against their mean, according to Bland and Altman, 1986. The bias is an important aspect in the interpretation of the BA approach, and it was calculated as a mean applied to the difference between the value determined in the first visit and the value determined in the second visit.

### 2.2.5 Cluster analysis

The aim of applying a clustering algorithm was to identify whether the data points will group according to the different levels of SNRs, or with respect to the different characteristics of the pupil traces from the individuals. The *k-means* (k=number of clusters) clustering algorithm applied in this study divides the data into different clusters, based on the distance between points (Euclidean distance). Given the distance between all data points and the centroids (the center of the cluster), the measurement will be assigned to the cluster with the nearest centroid.

## 2.3 Results

### 2.3.1 Pupillometry data

Figure 2.1 shows the pupil response of the most representative 10 (out of 27) individual NH listeners for the two test-retest pupil data sets. Significant effects were obtained on the GCA terms (intercept, linear, quadratic and cubic) with small p-values of polynomials estimates for both visits (between $1.18 \cdot 10^{-08}$- 0.009). Similarly, Figure 2.2 shows the pupil response of the 10 most representative (out of 24) individual HI listeners for the two test-retest pupil data sets. Significant effects were obtained on the GCA terms as indicated by small p-values for both visits (between $5.32 \cdot 10^{-15}$- 0.012).

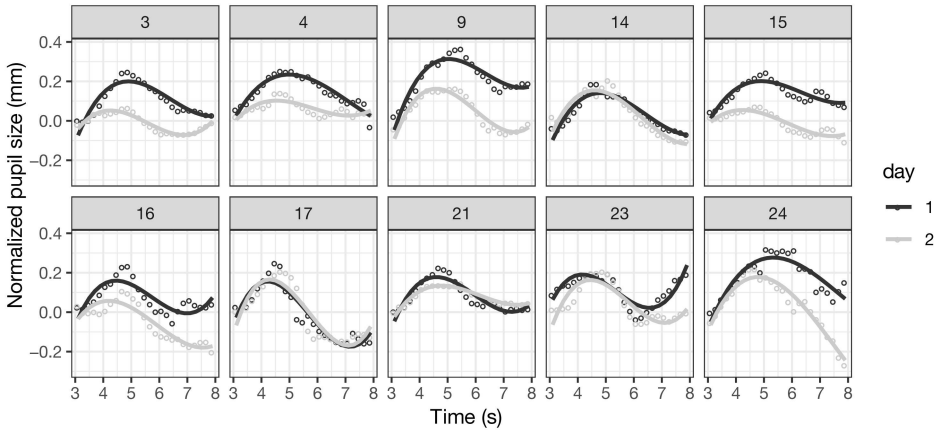



Figure 2.1: Growth Curve Analysis for individual NH listeners. Examples of the most relevant pupil responses as a function of time, on the two different visits (black and grey). The open circles represent the actual data, while the lined functions show the fitted GCA model. The numbers in the figure represent the test subjects.

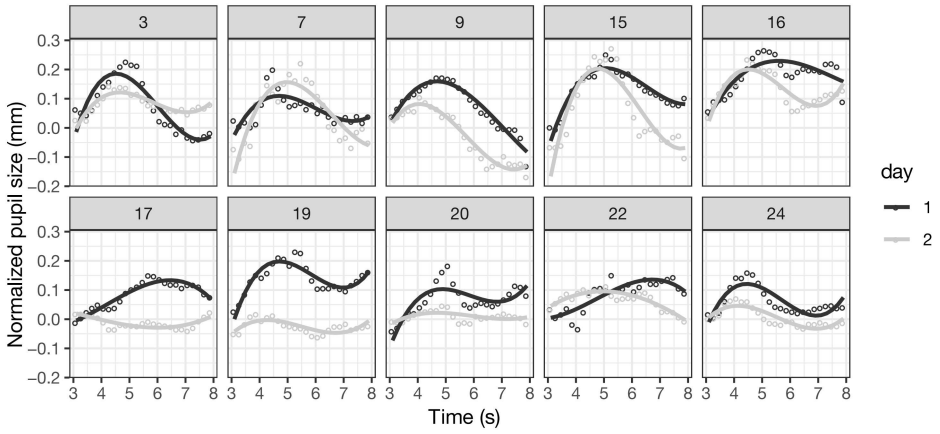



Figure 2.2: Growth Curve Analysis for individual HI listeners. Examples of the most relevant pupil responses as a function of time, on the two different visits (black and grey). The open circles represent the actual data, while the lined functions show the fitted GCA model. The numbers in the figure represent the test subjects

Both figures show that there were individual listeners with comparable pupil responses obtained at the two visits (e.g. NH 14, 17, 21, HI 15). However, there were also individuals showing clearly different responses (e.g. NH 9, HI 17, 19) at the two visits. The dissimilarity could be explained by the difference in the condition tested (0-4 SNR) at the two visits or by other individual factors that need to be identified.

### 2.3.2 ICC

The classical interpretation of the ICC states that an excellent reliability is reached when ICC values are over 0.75, a good one when ICC is between 0.60 and 0.74 and a fair one for values between 0.4 and 0.59 (Cicchetti, 1994). In the current study, the correlation coefficient was calculated for the mean, peak pupil dilation and the time-dependent terms obtained when applying the GCA model. Table 2.1 shows the ICC values obtained by assessing the reliability of the different features of the pupil response indicating the individual listening effort.

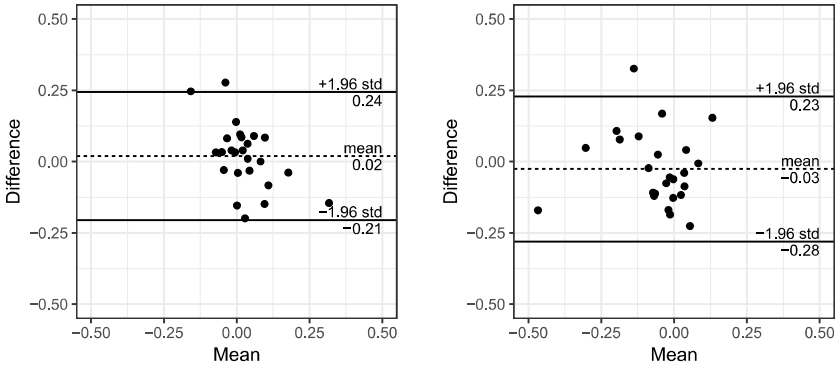
| ICC | NH | | HI | |
|---|---|---|---|---|
| | Agreement | Consistency | Agreement | Consistency |
| GCA Average peak | 0.6 | **0.62** | 0.41 | 0.54 |
| GCA Slope | 0.56 | 0.58 | **0.74** | **0.73** |
| GCA Rise-fall | **0.60** | **0.69** | **0.64** | **0.66** |
| GCA Delay | **0.74** | **0.86** | 0.27 | 0.47 |
| Peak pupil dilation | 0.48 | **0.60** | 0.48 | **0.64** |
| Mean pupil dilation | **0.63** | 0.59 | **0.60** | **0.64** |

Table 2.1: ICC agreement and consistency for mean, peak pupil dilation and for different terms of GCA. The ICC values reflect test-retest reliability and bold values are the ones showing good reliability

Different features of the pupil response are reliable for the two listener groups (rise-fall, delay and mean pupil for the NH listener group; slope, rise-fall and mean pupil dilation for the HI listener group).

### 2.3.3 Bland-Altman visual approach

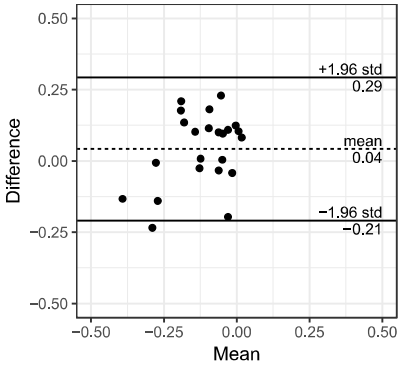
Figure 2.3 shows some examples of the agreement between tests taken on two separate visits as suggested by Bland-Altman. The difference between the two visits is shown against the mean of the two. Sometimes the value obtained on one visit was higher than the other, while sometimes the opposite was found. This contributes to a bias close to zero. If it is not close to zero, the values of the two visits systematically produce different results, and this represents a low agreement of the method.



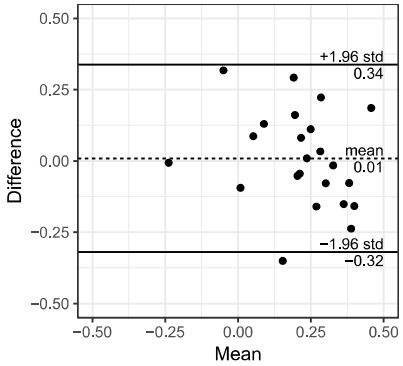

(a) Bland Altman Delay NH




(b) Bland Altman Rise-fall NH




(c) Bland Altman Rise-fall HI

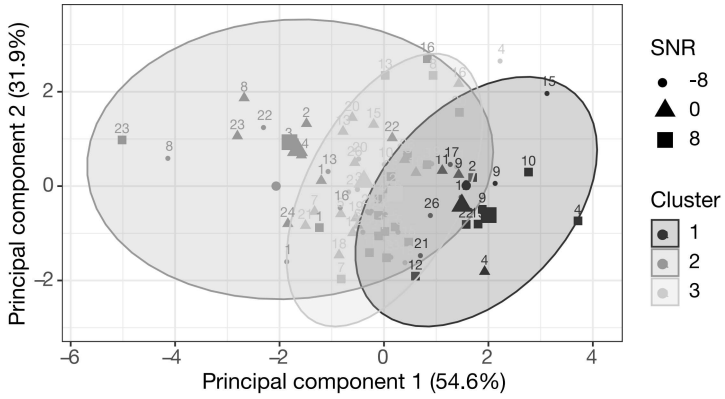



(d) Bland Altman Slope HI

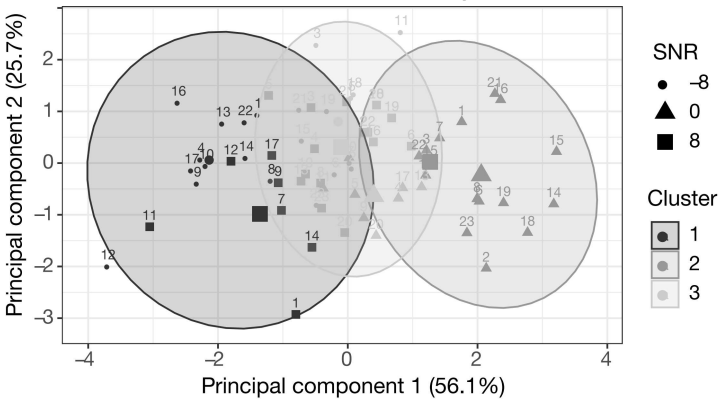
Figure 2.3: Example of Bland-Altman plots for NH (a,b) and HI (c,d) groups. The difference between two tests was plotted against their mean. Figures 2.3a and b show the BA agreement for delay and rise-fall features (NH group) while the figures 2.3c and d show the BA agreement for the rise-fall and slope features (HI group).

Panels a and b of Figure 2.3 show the results for the NH listeners. Most of the data points representing the delay were positioned within the LoA, as in the Figure 2.3a. The bias was close to zero showing that there were no significant differences between the two visits. Panels c and d of Figure 2.3 show corresponding results for the HI listeners. According to Figure 2.3d, the agreement of slope was good, with large LoA values, but the bias was still close to zero. This reflects good agreement, given that the spread of the data points was broader. These results were consistent with the ICC results. Thus, the test-retest reliability was considered as good.

### 2.3.4 Cluster analysis




(a) Clusters GCA terms for NH group, 3 SNRs




(b) Clusters GCA terms for HI group, 3 SNRs

Figure 2.4: Clustering of GCA terms for 3 different SNRs (k=3). One point represents one value of the measurement per participant per SNR.

Figure 2.4 shows the results of clustering the GCA terms for the NH (a) and HI (b) groups at 3 SNR conditions (-8 dB, 0 dB, 8 dB). The choice of the SNR levels to be analysed was made as in Wendt et al., 2018. Three different SNRs (out of the eight SNRs contained by the dataset) with a large range between their PPD were chosen for the cluster analysis. The cluster analysis was applied to both groups, NH and HI, and the results were similar. Listeners with the same SNR were expected to be assigned to the same cluster. According to Figure 2.4, the points belonging to the same cluster were data points at different SNRs, suggesting that these clusters could be formed on the base of other factors than those that were considered here.

## 2.4 Discussions and conclusion

This study showed a good reliability for some of the pupil responses features (slope, rise-fall and mean pupil dilation for the HI listeners; rise-fall, delay and mean pupil dilation for the NH listeners). The results obtained with the BA approach were consistent with the ICC results. As Alhanbali et al., 2019 also reported, the mean pupil size seems to be a reliable measure for both listeners groups. However, PPD was found to be less reliable than other measures in the current study. Moreover, the time-dependent features of the pupil response seem to be useful for evaluating the reliability of the method. Also, the slope seems to be more reliable for the HI group than for the NH group and it might be an important feature to explore in future studies.

The GCA model reported significant pupil features according to the small p-values of the polynomial estimates. The differences between individual functions obtained with the GCA for the two visits suggest that there could be other factors explaining the variance in the pupil curves (such as listener-dependent factors), apart from the difference in the level conditions (SNR). Zekveld et al., 2018 addressed some of these factors and emphasized that further investigations of the individual factors and the effects on the pupil response are required.

The cluster analysis suggested that SNR is not sufficient to classify listening effort, but that there might be some other factors needed for a classification such as listener-dependent factors like age, cognitive abilities and fatigue. Thus, future investigations of the data could consider such individual factors as input features. Furthermore, classification of the listening effort could be modeled with a supervised machine learning algorithm or even a time series analysis.

One of the limitations of the study was the use of different SNR conditions to test the pupil response reliability. It would be valuable to evaluate the reliability of pupillometry in the same acoustic conditions. Eventually, identifying and controlling the factors that can provide insights in cognitive understanding of listening situations will improve the accuracy of pupillometry as an objective measure of listening effort.

Overall, this study showed that rise-fall and mean pupil dilations seem to be important features of the pupil response, demonstrating that the signal is reliable enough in both listener groups. Other time-dependant features seemed to be reliable for one of the groups (Slope for HI and Delay for NH). The reliability results of the method are an important prerequisite for future experimental

analysis and for developing pupillometry and the test protocol towards a standardized test for clinical use.

# 3

# Exploring the reliability of pupillometry under different task demands, normalization procedures and at multiple visits [b]


## Abstract

Recordings of the pupillary response have been used in numerous studies to assess listening effort during a speech-in-noise task. Most studies focused on averaged response across listeners, whereas less is known about pupil dilation as an indicator of the individuals' listening effort. The present study investigated the reliability of several pupil features as potential indicators of individual listening effort and the impact of different normalization procedures on the reliability. The pupil dilations of 31 normal-hearing listeners were recorded during multiple visits while performing a speech-in-noise task. The signal-to-noise ratios (SNRs) of the stimuli ranged from -12 dB to +4 dB. All listeners were measured twice at separate visits, and 11 were re-tested at a third visit. To examine the reliability of the pupil responses across visits, the intraclass correlation coefficient was applied to the peak and mean pupil dilation and to the temporal features of the pupil response, extracted using growth curve analysis. The reliability of the pupillary response was assessed in relation to SNR and different normalization procedures over multiple visits. The most reliable pupil features were the traditional mean and peak pupil dilation. The highest reliability results were obtained when the data were baseline-corrected and normalized to the individual pupil response range across all visits. The present study results showed


[b] This chapter is based on Neagu et al., (2022a), under revision

an impact of the normalization procedure on the reliability of the pupil response across multiple visits. The reliability varied across SNR. Overall, the results are an important basis for developing a standardized test for pupillometry in the clinic.

## 3.1 Introduction

Listening effort, defined as 'the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task' (Pichora-Fuller et al., 2016), has been a growing topic in the auditory field over the last couple of decades. Among different measures of listening effort, pupillometry, i.e., tracking of the pupil size, has been recognized to be the 'most useful autonomic indication' of effort (Kahneman, 1973). Pupillometry has been demonstrated to provide a measure of listening effort during speech-in-noise tests both in normal-hearing (NH) and hearing-impaired (HI) listeners (Koelewijn et al., 2012b; Kramer et al., 1997; Zekveld et al., 2010, 2011). For example, Ohlenforst et al., 2017b indicated that HI listeners showed an increased pupil dilation indicating increased allocation of resources to reach similar speech intelligibility performance compared to NH listeners. Several studies examined the impact of the level of speech intelligibility, signal-to-noise-ratio (SNR), linguistic complexity and hearing-aid signal processing on listening effort (Kuchinsky et al., 2014, 2013; McGarrigle et al., 2014; Wendt et al., 2018; Winn, 2016; Zekveld et al., 2011). For instance, pupillometry has been shown to be sensitive to changes in the acoustic signal caused by hearing-aid signal processing. Specifically, a reduction in listening effort has been reported with noise-reduction schemes for HI listeners at SNRs reflecting ecologically valid listening situations at a high level of speech intelligibility (Ohlenforst et al., 2017b; Wendt et al., 2017). These studies support the hypothesis that a more complete characterization of the difficulties in speech understanding arising as a consequence of hearing impairment, and the potential benefit of hearing aid interventions, can be gained when measuring listening effort in addition to speech intelligibility.

So far, pupillometry as a measure of listening effort during a speech-in-noise task has only been evaluated on a listener group level (as averaged responses across listeners) and little is known about the sensitivity and reliability of this method for individual listeners. However, such sensitivity and reliability of the method on an individual listener's level would be crucial for pupillometry to

be used as a basis for individualized rehabilitation strategies. The transition from pupillometry assessed on a group level to an individual listener level is challenging because the pupil response has numerous sources of variation (Koelewijn et al., 2012b; Partala and Surakka, 2003; Wang et al., 2018a; Zekveld et al., 2018, 2011). For example, the pupil response are affected by environmental factors, such as luminance, masking noise or communication technologies (e.g., hearing aids). Furthermore, listener-specific factors, such as cognitive abilities, hearing impairment or the level of fatigue, can affect the pupil response (Kuchinsky et al., 2016; Pichora-Fuller et al., 2016; Wang et al., 2018a; Zekveld et al., 2018).

A few studies investigated the reliability of the pupil response assessed during speech recognition. Alhanbali et al., 2019 explored the reliability of several physiological measures during a digit-in-noise recognition task performed under individualized listening conditions, whereby the level of speech intelligibility performance was fixed at 71%. The authors reported that among the assessed physiological measures, pupillometry (specifically, the mean pupil dilation, MPD, and the peak pupil dilation, PPD, of the response) showed the highest reliability with an intraclass correlation coefficient (ICC $>$ 0.85) as compared to EEG and skin conductance. Similarly, Giuliani et al., 2020 investigated the sensitivity and reliability of different measures of listening effort (including skin conductance, pupillometry and self-reported listening effort using a dual-task paradigm). The authors assessed listening effort during sentence recognition at SNR levels of 0, -3 dB and -5 dB. Consistent with Alhanbali et al., 2019, Giuliani et al., 2020 reported the highest reliability for pupillometry among all tested measures, even though the corresponding level of ICC was only fair (ICC $<$ 0.5). ICC is a reliability index that reflects the degree of agreement between similar measurements. Both studies showed that investigated pupil features were equally reliable to the subjective measures of listening effort (NASA Task Load Index - NASA-TLX, Hart and Staveland, 1988 and another self-reported effort question).

These studies focused on the analysis of the MPD and PPD only, following the traditional characterization of the pupil response (Koelewijn et al., 2012b; Zekveld et al., 2010, 2011). However, more recently, Kuchinsky et al., 2013 showed that growth curve analysis (GCA) can be used to detect changes in the shape of the pupil response over time, allowing for an independent evaluation of different temporal characteristics of the pupil response (Mirman et al., 2008;

Winn et al., 2015). GCA fits orthogonal polynomial terms to time series data to show different variations in the function among individuals (Mirman et al., 2008). Not much is known, though, about the reliability of the traditional nor GCA pupil features across multiple visits.

Only a few studies evaluated the reliability of various measures other than pupil features over more than two visits (e.g., psycho-physiological measures: intrinsic attentive selection of one of two lateralized visual cues, Aday and Carlson, 2019; daytime sleepiness, Zwyghuizen-Doorenbos et al., 1988). Aday and Carlson, 2019 showed that attention biases were not reliable until participants had fairly extensive experience with the task. They suggested that more visits could reduce the noise in the data related to task familiarity and increase the reliability. These studies showed, in fact, an increase of the reliability of the tests with increasing number of visits. However, the reliability of pupillometry assessed within a speech-in-noise task paradigm over multiple visits has not yet been studied. Furthermore, Alamia et al., 2019 and Widmann et al., 2018 showed that the pupil dilates following increased surprise or, more generally, following global arousal, and that emotional arousal to novel sounds enhances the sympathetic contribution to the pupil dilation response. Thus, it follows that an arousal effect observed in the pupil response when performing a novel task (i.e., at the first visit, Visit 1) could result in lower reliability of pupillometry between Visit 1 and 2 than a comparison between the responses in subsequent visits. A common approach to avoid arousal effects has been to remove the first trials (within a condition) from the analysis, and thus, to reduce the impact of any initial effects (Winn et al., 2018). However, a more general arousal effect (i.e., novel task, novel environment, unknown experimenter) is difficult to control. Thus, the present study investigated the reliability of the pupil response over multiple visits.

Furthermore, regarding the changes in the reliability of the pupil response, with changing SNR, results differed remarkably across studies. Giuliani et al., 2020 found a fair reliability irrespective of their considered SNR changes from 0 to -3 dB and from -3 to -5 dB, respectively. In contrast, other studies suggested that task demands impact the reliability such that increasing task demands lead to a higher index of pupillary activity (Duchowski et al., 2018), higher inter-trial change in pupil dilation (Krejtzid et al., 2018) and prospective memory (Einstein et al., 1997).

Finally, different methods of pupil dilation normalization have been pro-

posed in the literature (e.g., Winn et al., 2018). A common approach when assessing listening effort in a speech-in-noise task paradigm is baseline correction. Baseline-corrected responses represent a change in the pupil size relative to a particular temporal window before the stimulus, known as baseline, (Winn et al., 2018). However, while some studies argued that the normalization of task-evoked changes in pupil size should be done independently of the baseline pupil size (Beatty, 1982; Bradshaw, 1969), others stated that different ways of baseline scaling could produce disparities in the reported pupil size results (Mathôt et al., 2018; Reilly et al., 2019). Moreover, relatively large interindividual differences in the dynamic range of the pupil dilation have been observed and several other approaches have been proposed to target these differences. For example, Piquado et al., 2010 obtained a dynamic range of the pupil response based on changes in the luminance (dark versus light), which was then used for range normalization. Furthermore, McCloy et al., 2016 applied z-score transformation and Winn, 2016 considered a proportional change within the individual between a reference condition and the task condition. However, the impact of normalization procedure on the reliability of different pupil features has not been studied.

The present study aimed to obtain a better understanding of the reliability of pupillometry as an objective indicator of an individual's listening effort. Different features of the pupil response, assessed in a speech-in-noise paradigm, were extracted and the impact of task demands (i.e., changing SNR) and data normalization procedures on the reliability of those features were systematically investigated. The test-retest reliability of pupillometry was investigated by assessing the pupil response over three visits. It was hypothesized that the reliability of different pupil features would increase with decreasing SNR (i.e., higher task demands). Furthermore, it was hypothesized that the reliability of different pupil features would be affected by applying distinctive normalization procedures.

## 3.2 Methods

### 3.2.1 Participants

Thirty-five participants (aged from 18 to 65 years, mean 38) took part in this study. All participants were native Danish speakers. They had pure-tone hearing

thresholds of 20 dB hearing level (HL) or better at low frequencies (below 6 kHz) in both ears and thresholds of 30 dB HL or better at frequencies above 6 kHz. The participants had no history of eye diseases or eye operations. Exclusion criteria also included caffeine intake less than 3 hours prior to the test time. The data of four participants out of the thirty-five were excluded from the analysis because of their withdrawal from the study after the first visit. The research procedures were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391), and all participants provided written informed consent for the study procedures and received monetary compensation for their participation.

### 3.2.2 Procedure and stimuli

Participants were asked to perform a speech-in-noise test with sentences from the Danish Hearing in Noise Test (HINT, (Nielsen and Dau, 2011)). HINT sentences were presented in a 4-talker babble masker which was created by overlapping two male and two female talkers (all reading different excerpts from a newspaper) with the same long-term average frequency spectrum as the HINT sentences. For each measurement trial, the masker onset started 3 seconds prior to sentence onset and stopped 3 seconds after sentence offset, as the vertical lines in Figure 3.1 indicate. The length of each trial varied depending on the length of the presented HINT sentence, which have a mean duration of about 1.5 s. After masker offset, the participants were asked to repeat back the HINT sentence. Two seconds of silence were established before noise onset to allow for the pupil to return to pre-task levels (i.e., recovery). Sentences were presented at 5 different SNRs: 4 dB, 0 dB, -4 dB, -8 dB and -12 dB. Different conditions were presented in a block design with 25 trials containing 25 sentences for each SNR. Trials were randomized within each block, and the presentation order of each condition was randomized across participants. The stimuli were presented through Sennheiser HD650 headphones using an SPL Audio Phonitor Mini amplifier. The noise level was fixed to a sound pressure level (SPL) of 70 dB for both ears while the level of the target speech varied depending on the SNR.

The participants were instructed to fix their gaze at a grey cross in the middle of a black screen during the speech-in-noise task and to repeat the HINT sentence after the noise offset. The responses were scored on a word-level basis (all recognized words from the sentence were marked as correct).

The participants were tested at two different visits (Visit 1 and Visit 2) using

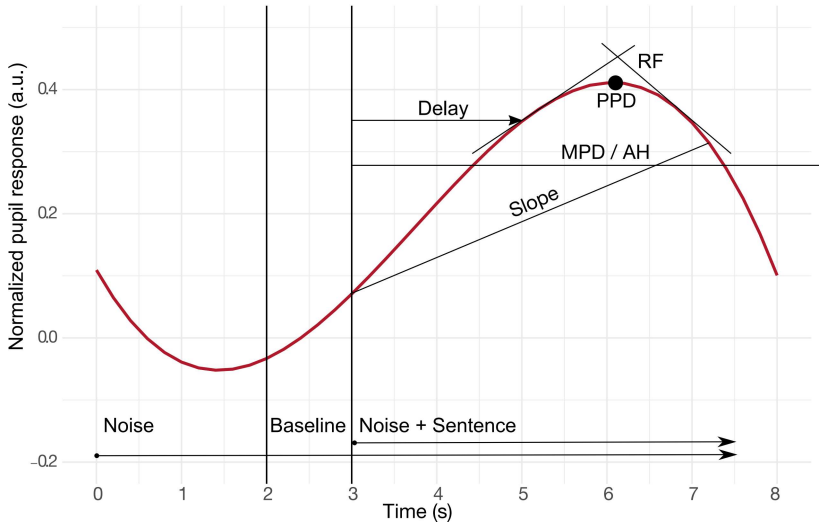



Figure 3.1: Schematic illustration of the pupil response within the speech-in-noise test with sentence onset at second 3. All analyzed pupil features (traditional and GCA features) are schematically represented.

a repeated measures design. Eleven out of the thirty-one participants were re-tested additionally at a third visit (Visit 3). The visits were spaced three to six weeks apart to avoid any learning effects of the sentence material (Bramsløw et al., 2016). The subsequent visits were scheduled at the same time of the day and at the same period of the week (i.e., beginning, middle, or end) as for Visit 1 to minimize the potential effect of fatigue at different times during a day or at different days of the week and to control for circadian rhythm effects (Daguet et al., 2019). The procedure was the same at the second and third visits with the same presentation order of the conditions and the same sentences but in different order, per condition for each of the listeners.

### 3.2.3 Apparatus and pupillometry data processing

Eye-tracking data were continuously recorded during the speech-in-noise test using a desktop mounted eye-tracker (EyeLink 1000; SR-Research Ltd., Mississauga, Ontario, Canada). Pupil sizes were recorded from the left eye with a sampling frequency of 500 Hz. The measurements were performed in the same booth with same luminosity levels across visits (screen and ambient light). The screen's luminance and ambient light were controlled to prevent any changes in pupil response that could be attributed to changes in ambient or screen light intensity. The ambient light was measured at 75 lx for the tasks performed in

light. The screen had an approximate brightness of 9 cd/m2 during the speech-in-noise task, where the screen displayed a black background with a grey cross in the middle. The distance from the middle of the participant's eyes to the centre of the screen varied between 50-70 cm.

The pupil data were processed using (MATLAB, 2018) and R (R Core Team, 2019). In order to remove any initial arousal effects, the pupil traces of the first three trials within a block were excluded from the analysis. Since a decreasing trend of the pupil within each block was observed, the entire block recording was linearly detrended. For the eye-blink removal, the mean pupil dilation with standard deviation was calculated across the whole trial. Pupil dilation values more than three standard deviations smaller than the mean were coded as eye-blinks. Eye-blinks were removed by a linear interpolation that started about 80 ms before and ended 150 ms after the blinks. Data were then smoothed using a moving average filter with a symmetric rectangular window of 117 ms. Trials with more than 20% missing data, eye blinks or artefacts were removed from the analysis. All remaining traces were scaled using each of the four normalization procedures presented in Section 3.2.5 below.

### 3.2.4 NASA-TLX and perceived effort

After each block, participants were asked to answer the NASA-TLX (Hart and Staveland, 1988) questionnaire to assess a measure of the perceived listening effort. The NASA-TLX uses a 0-20 scale (low/high). NASA-TLX has six subitems: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration. The score was rescaled to a 0-10 scale and was calculated as a mean score of each of the subitems. Additionally, another measure of self-reported listening effort was provided by each participant after each SNR block. On a 0-to-10 scale (with 0 indicating low effort and 10 indicating high effort), participants were asked to answer the following question: 'Hvor meget anstrengte du dig for at høre sætningerne?' which translates to English as, 'How much effort did you put into hearing the sentences?'

### 3.2.5 Data normalization

Four different normalization procedures were applied. First, baseline correction (Eq. 3.1) was applied by subtracting the mean pupil size measured in the 1 s

period preceding the sentence onset within each trial.

$$x_{\text{baseline corrected}} = x - \sigma_{\text{baseline}} \tag{3.1}$$

where $x$ is the pupil dilation at a given sample, and $\sigma$ is referring to the mean pupil size within the baseline time window (i.e., between the 2nd and 3rd second). The baseline was established 1 s prior to the sentence onset, as recommended by Winn et al., 2018.

Alternatively, a range normalization procedure was applied for each individual for each trial. The pupil range was calculated by extracting the maximum and the minimum pupil dilation across all trials of all conditions and visits for each individual. All trials were then range normalized (Eq. 3.2).

$$x_{\text{range}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{3.2}$$

where $x$ is the pupil dilation at a given sample and $x_{\text{max}}$ and $x_{\text{min}}$ refers to the overall maximum and minimum pupil dilation over all trials and visits.

As another option, a Z-score normalization was applied, which subtracts the mean pupil dilation for an individual from each pupil sample and divides the result by the standard deviation of the mean pupil dilation (Eq. 3.3).

$$x_{\text{Z-score}} = \frac{x - \sigma}{\mu} \tag{3.3}$$

where $x$ is the pupil dilation at a given sample, $\sigma$ refers to the mean of the pupil dilation per individual and $\mu$ to the standard deviation of the pupil dilation per individual.

Finally, a range normalization procedure (Eq. 3.4) was applied on the baseline corrected data using formulas (Eq. 3.1) and (Eq. 3.2), referred to here as 'baseline range' normalization.

$$x_{\text{baseline range}} = \frac{x_{\text{baseline corrected}} - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{3.4}$$

### 3.2.6 Feature extraction

The MPD was calculated as the average pupil dilation in the interval between sentence onset and masker offset (see Figure 3.1 arrow Noise + Sentence). The PPD was calculated as the maximum dilation in the same interval.

In order to account for effects reflected in the time-course of the pupillary response, growth curve analysis (GCA) was applied (Mirman et al., 2008). GCA is a multi-level regression technique that fits orthogonal polynomials to time course data. A third-order (cubic) orthogonal polynomial was applied to the overall time course of the pupil dilation within a time window starting at 2 s (i.e., at the baseline onset) until 8 s of stimulus presentation (see Figure 3.1). A third-order polynomial function including the intercept through cubic terms was considered to provide a good fit to the shape of the pupil response across time (Kuchinsky et al., 2014, 2016). The feature extraction is described in (Eq. 3.5). Pupil size was considered as a dependent variable in the model, predicted by a series of fixed and random effects (individual and trial number, respectively).

$$\begin{aligned} &\text{pupil feature} \sim (1 + p_1 + p_2 + p_3) * \text{participant} \\ &+ \big(1 + p_1 + p_2 + p_3 | \text{trial}\big) \end{aligned} \tag{3.5}$$

A schematic representation of the GCA features can be seen in Figure 3.1. The intercept term represents the average height (AH) of the pupil response, the linear term ($p_1$) reflects the slope, the quadratic term ($p_2$) reflects the rise and fall (RF) around the central inflexion point of the response function, and the cubic term ($p_3$) reflects the inflexions at the extremities of the curve, referred to as “delay” in the current study.

### 3.2.7 Reliability analysis

The reliability of the pupil features was assessed using Spearman’s correlation coefficient, which reveals how consistent the results are across the different visits, as well as the intraclass correlation coefficient (ICC), which evaluates the test-retest reliability (Cicchetti, 1994; Koo and Li, 2016). Sperman’s correlation sorts the observations by rank and evaluates how similar the ranks are. Their values lies between -1 and 1 with 1 indicating strong relationship. Sperman’s correlation coefficient is calculated as in Eq. 3.6.

$$\text{Spearman}_{\text{coef}} = \frac{\text{Cov}(\text{rank}_{V_1}, \text{rank}_{V_2})}{\mu_{V_1} \mu_{V_2}} \tag{3.6}$$

where $\text{Cov}(\text{rank}_{V_1}, \text{rank}_{V_2})$ are the covariances between the ranks of the pupil measures at Visit 1, respectively Visit 2, while $\mu$ refers to the standard deviation of the same ranks.

The ICC assesses the group reliability by comparing the variability within different visits of the same participant's pupil dilation to the total variation across all visits and all participants. Here, the ICC was calculated to evaluate the reliability of different features of the pupil response (see Section 3.2.6) between Visit 1 and 2 for 31 participants, and between Visit 2 and 3 for the subgroup of 11 participants who came for a third visit. The latter was compared to the ICC values measured for the same 11 participants between Visit 1 and 2.

The ICC was calculated as a two-way mixed-effects model with two measurements, as reflected in (Eq. 3.7), where $MS_B$ is the mean square between subjects, $MS_T$ is the mean square between trials, $MS_E$ is the mean square error, $n$ is the number of subjects and $k$ is the number of measurements.

$$ICC_{\text{agreement}} = \frac{MS_B - MS_E}{MS_B + (k-1)MS_E + \frac{k}{n}(MS_T - MS_E)} \tag{3.7}$$

To assess the test-retest reliability between two visits, ICC was calculated for each combination of normalization technique (i.e., baseline correction, range normalization, Z-score, baseline range normalization) and feature (i.e., PPD, MPD, and GCA features), and between Visit 1 and Visit 2, and Visit 2 and 3, for all combinations of normalization type and pupil feature.

## 3.3 Results

### 3.3.1 Group average data

Although this study focused on the reliability of individual's pupil dilation, a group-level analysis was conducted first to provide an anchor to previous literature and to gauge group-level reproducibility. The pupil traces for the different normalization procedures averaged across all participants are shown in Figure 3.2. Overall, it can be seen that, regardless of the normalization procedure, the general trend of increasing pupil response with decreasing SNR remains. By visually inspecting the traces, it appears that larger differences between the two visits occur in the Z-score (low right panel) and range normalization (low left panel) procedure compared to the other two, especially for -8, -4 and 4 dB SNR. A quantitative analysis of these differences will be provided below on an individual level (Results subsections 3.3.3 and 3.3.4).

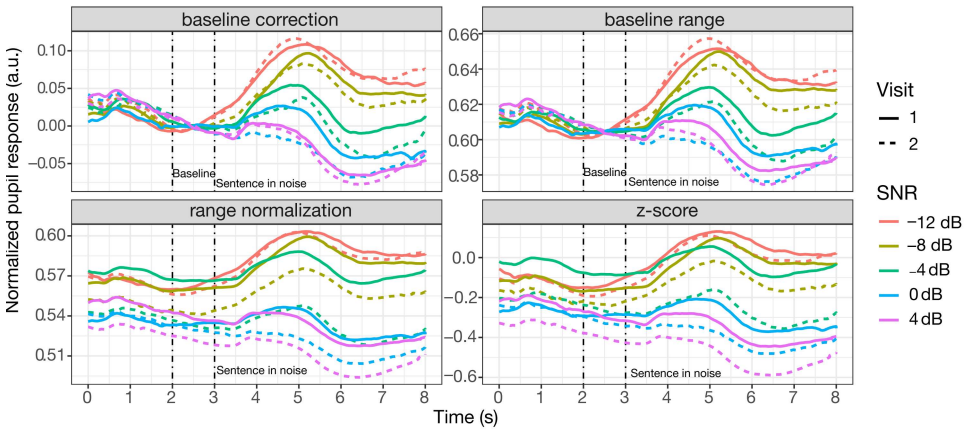



Figure 3.2: Pupil traces averaged across participants, normalized with different procedures. The SNRs tested are presented in different colors and the Visits are presented in different line types.

### 3.3.2 Group level pupil features across visits and SNRs

The six different pupil features extracted from the group-averaged, baseline-corrected pupil response are displayed in Figure 3.3 for all 3 visits and all 5 SNRs (-12, -8, -4, 0 and 4 dB). The visits are presented in different colors, such that the figure depicts how the distribution over each feature varies as a function of SNR and visit. All features except delay showed a slightly decreasing trend with increasing SNR. An increasing trend of delay with increasing SNR indicates that the peak dilation is reached later with increasing SNR.

A two-way ANOVA was performed to investigate the impact of SNR and visit on each pupil feature for each normalization procedure. The results are displayed in Table 3.1. Significant effects are highlighted in bold. There was no impact of the visit number on the group-level analysis for any of the pupil features except delay, suggesting that average features were reliable across multiple visits. There was an effect of SNR for some of the features when certain normalization procedures were applied (i.e., slope, RF and delay for all normalization procedures and MPD only for baseline correction and range normalization) (p-value $< 0.05$). Interestingly, significant effect of SNR on PPD occurs only for some for some normalization procedures for high SNRs (i.e., baseline range -4 to 4 dB SNR and range normalization 0 to 4 dB SNR).

| Normalization | Features | PPD | MPD | AH | Slope | RiseFall | Delay |
|---|---|---|---|---|---|---|---|
| **Baseline correction** | Intercept | **0.622** *** | **0.080** *** | **0.057** *** | **0.145** *** | **0.128** *** | **-0.211** *** |
| | Visit2 | -0.034 | -0.011 | -0.005 | -0.033 | -0.021 | 0.047 * |
| | Visit3 | 0.038 | 0.002 | 0.001 | -0.003 | -0.018 | -0.002 |
| | -8 dB | -0.006 | **-0.028** * | **-0.016** * | -0.061 | -0.058 | 0.041 |
| | -4 dB | 0.013 | **-0.072** *** | **-0.04** *** | **-0.242** *** | **-0.107** ** | **0.114** *** |
| | 0 dB | -0.054 | **-0.095** *** | **-0.052** *** | **-0.312** *** | **-0.128** *** | **0.156** *** |
| | 4 dB | **-0.117** * | **-0.112** *** | **-0.061** *** | **-0.311** *** | **-0.185** *** | **0.210** *** |
| **Range normalization** | Intercept | **0.653** *** | **0.621** *** | **0.650** *** | **0.099** *** | **0.06** *** | **-0.116** *** |
| | Visit2 | -0.022 | -0.023 | -0.03 | -0.027 | -0.007 | **0.026** ** |
| | Visit3 | 0.007 | 0.004 | -0.004 | -0.015 | -0.009 | 0.009 |
| | -8 dB | -0.012 | -0.012 | -0.015 | -0.033 | -0.027 | 0.019 |
| | -4 dB | -0.034 | -0.034 | -0.028 | **-0.125** *** | **-0.059** *** | **0.063** *** |
| | 0 dB | **-0.065** ** | **-0.065** ** | -0.014 | **-0.165** *** | **-0.067** *** | **0.084** *** |
| | 4 dB | **-0.077** ** | **-0.077** ** | -0.027 | **-0.17** *** | **-0.089** *** | **0.110** *** |
| **Z-score** | Intercept | **2.382** *** | **0.230** *** | 0.077 | **0.641** *** | **0.415** *** | **-0.754** *** |
| | Visit2 | 0.019 | **-0.051** * | **-0.143** * | -0.153 | -0.054 | **0.144** ** |
| | Visit3 | -0.061 | -0.012 | -0.08 | -0.0512 | -0.058 | 0.049 |
| | -8 dB | 0.023 | -0.066 | 0.001 | -0.251 | **-0.21** * | **0.163** * |
| | -4 dB | 0.015 | **-0.239** *** | -0.016 | **-0.796** *** | **-0.382** *** | **0.408** *** |
| | 0 dB | -0.051 | **-0.294** *** | -0.004 | **-1.063** *** | **-0.465** *** | **0.542** *** |
| | 4 dB | -0.059 | **-0.326** *** | -0.001 | **-1.086** *** | **-0.625** *** | **0.742** *** |
| **Baseline range** | Intercept | **0.626** *** | **0.588** *** | **0.581** *** | **0.105** *** | **0.066** *** | **-0.127** *** |
| | Visit2 | -0.006 | -0.008 | -0.004 | -0.029 | -0.01 | **0.027** ** |
| | Visit3 | 0.003 | 0.004 | -0.009 | -0.016 | -0.013 | 0.0129 |
| | -8 dB | -0.011 | -0.011 | 0.004 | -0.031 | -0.028 | 0.02 |
| | -4 dB | **-0.039** *** | **-0.039** *** | -0.026 | **-0.129** *** | **-0.062** *** | **0.068** *** |
| | 0 dB | **-0.051** *** | **-0.052** *** | 0.007 | **-0.175** *** | **-0.073** *** | **0.092** *** |
| | 4 dB | **-0.06** *** | **-0.059** *** | 0.006 | **-0.178** *** | **-0.096** *** | **0.122** *** |

Table 3.1: Estimates obtained when applying a two-way ANOVA to investigate the effect of SNR and visit on different pupil features for different normalization procedures. The intercept is represented by Visit 1, -12 dB SNR. Significant effects are highlighted in bold ($p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***)

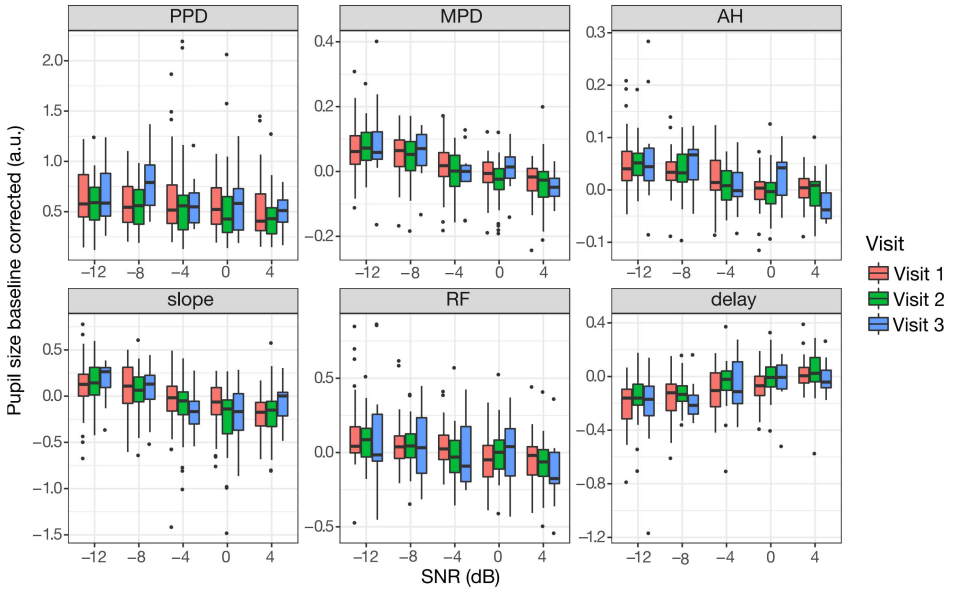



Figure 3.3: Boxplots of the pupil features PPD, MPD, AH, slope, RF, and delay indicated in the different panels are shown as a function of SNR for three different visits (Visit 1, Visit 2, and Visit 3) indicated by different colors. The mid-line of the boxes represents the median values while the vertical line is the standard deviation.

### 3.3.3 Consistency across visits and normalization procedures

To investigate the impact of the normalization procedure on the consistency of each pupil feature across visits, a Spearman's correlation analysis was performed with each of the pupil features. Spearman's correlation coefficients for Visit 1 versus 2 (31 participants), and Visit 1 versus 2 and 2 versus 3 (11 participants) are shown in Table 3.2 and the individual correlations are shown in Figures 3.4 and. For 31 participants, the highest correlation coefficients between visits 1-2 were observed for two pupil features, MPD and PPD, for three out of four normalization procedures (i.e., for baseline correction, baseline range and range normalization but not for the Z-score). From the GCA features, the delay and slope were the most consistent features across the normalization procedures, with correlations above 0.5.

Almost all correlations were significant (with p-values $< 0.0001$ (***) and p-values $< 0.001$ (**) as indicated in Table 3.2). The lowest correlation, and even some negative correlations, were observed for the Z-score normalization procedures (ranging between -.78 and 0.5). Among all the normalization procedures applied in this study, the baseline-corrected data combined with a range normalization procedure showed the highest correlations across visits (between

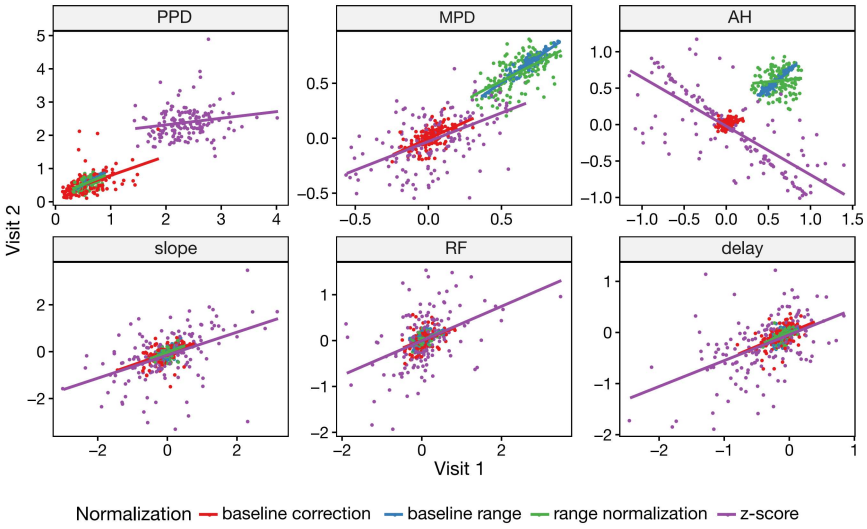



Figure 3.4: Scatter plot depicting the correlation between Visit 1 and 2 per individual across all SNRs and for each pupil feature (PPD, MPD, AH, slope, RF, delay) indicated in the different panels and for each normalization procedure (baseline correction, range normalization, Z-score, and baseline range) as indicated by different colors.

R=0.43 and R=0.94).

Due to differences in the sample size (i.e., 31 participants for Visit 1 versus Visit 2 and 11 participants for Visit 2 versus Visit 3), a comparison was also made with the pupil features of the same 11 participants at Visit 1 versus 2. Overall, for MPD and PPD, higher consistency was obtained between Visit 1 and 2 than between Visit 2 and 3 for all normalization procedures except for the baseline range normalization. The GCA features showed no clear trend in consistency across visits. Among GCA features a high correlation was only observed in the delay values for the subsample of 11 participants between both, Visit 1-Visit 2 and Visit 2-Visit 3.

### 3.3.4 ICC

To examine the reliability of the pupil features on an individual level, ICC values were calculated with 95% confidence intervals and are summarized in Table 3.3 for Visit 1 and 2 and in Table 2 of the supplemental material for the subsample of 11 participants for the three session.

The results were categorized according to Cicchetti, 1994, who defined excellent reliability for ICCs above 0.75 and good reliability for ICCs above 0.6. Good reliability is indicated in bold, while excellent results are highlighted in

| Spearman correlation | PPD | MPD | AH | Slope | RF | Delay | PPD | | MPD | | AH | | Slope | | RF | | Delay | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Visit 1-2 (31) | Visit 1-2 (31) | Visit 1-2 (31) | Visit 1-2 (31) | Visit 1-2 (31) | Visit 1-2 (31) | Visit 1-2 (11) | Visit 2-3 (11) | Visit 1-2 (11) | Visit 2-3 (11) | Visit 1-2 (11) | Visit 2-3 (11) | Visit 1-2 (11) | Visit 2-3 (11) | Visit 1-2 (11) | Visit 2-3 (11) | Visit 1-2 (11) | Visit 2-3 (11) |
| Baseline correction | **0.6** *** | 0.59 *** | 0.45 *** | 0.5 *** | 0.41 *** | 0.53 *** | 0.52 *** | 0.5 *** | **0.64** *** | 0.53 *** | 0.57 *** | 0.5 *** | 0.48 *** | 0.39 ** | 0.53 *** | 0.45 ** | **0.72** *** | **0.72** *** |
| Range normalization | **0.63** *** | **0.63** *** | 0.12 | 0.51 *** | 0.44 *** | 0.54 *** | **0.66** *** | 0.46 *** | **0.66** *** | 0.46 *** | 0.32 * | 0.28 * | 0.49 *** | 0.39 ** | 0.58 *** | 0.47 *** | **0.64** *** | **0.67** *** |
| Z-score | 0.21 ** | 0.54 *** | -0.78 *** | 0.5 *** | 0.45 *** | 0.5 *** | -0.03 | 0.036 | **0.6** *** | 0.52 *** | -0.47 *** | -0.4 ** | 0.45 *** | 0.37 ** | 0.59 *** | 0.45 ** | **0.65** *** | **0.67** *** |
| Baseline Range | **0.87** *** | **0.87** *** | **0.94** *** | 0.54 *** | 0.43 *** | 0.58 *** | **0.73** *** | **0.81** *** | **0.73** *** | **0.81** *** | **0.91** *** | **0.93** *** | 0.49 *** | 0.38 ** | 0.59 *** | 0.53 *** | **0.65** *** | **0.68** *** |

Table 3.2: Spearman correlations between two consecutive visits for all pupil features calculated through different normalization procedures. The values above 0.6 are highlighted in bold, representing good correlation.

| ICC | Feature | PPD | MPD | AH | Slope | RF | Delay |
|---|---|---|---|---|---|---|---|
| Baseline correction | All SNRs | **0.65** | **0.73** | 0.51 | **0.70** | 0.52 | **0.66** |
| | -12 dB | **0.67** | **0.72** | 0.56 | **0.66** | 0.5 | 0.56 |
| | -8 dB | **0.72** | 0.56 | 0.5 | 0.46 | 0.48 | 0.5 |
| | -4 dB | **0.71** | **0.7** | ***0.77*** | ***0.78*** | 0.17 | ***0.77*** |
| | 0 dB | 0.58 | 0.16 | 0.44 | 0.31 | 0.29 | 0.44 |
| | 4 dB | **0.71** | ***0.81*** | 0.49 | 0.53 | ***0.79*** | 0.49 |
| Range normalization | All SNRs | 0.59 | 0.58 | ***0.97*** | **0.64** | **0.74** | **0.67** |
| | -12 dB | ***0.77*** | 0 | 0 | ***0.77*** | 0.47 | **0.7** |
| | -8 dB | **0.72** | 0.19 | 0.19 | 0.33 | 0.58 | 0.46 |
| | -4 dB | 0.43 | 0 | 0 | **0.62** | 0.44 | **0.68** |
| | 0 dB | ***0.8*** | 0.58 | 0.58 | 0.35 | 0.43 | 0 |
| | 4 dB | ***0.8*** | 0.3 | 0.3 | 0.56 | ***0.82*** | 0.45 |
| Z-score | All SNRs | 0.39 | 0.33 | 0 | 0.55 | **0.71** | **0.66** |
| | -12 dB | 0.36 | 0.49 | 0 | ***0.76*** | 0.46 | **0.72** |
| | -8 dB | 0.55 | 0.42 | 0 | 0.23 | **0.61** | 0.32 |
| | -4 dB | 0 | 0.34 | 0 | **0.62** | 0.49 | **0.67** |
| | 0 dB | 0.26 | 0 | 0 | 0.37 | 0.28 | 0 |
| | 4 dB | 0.24 | 0.51 | 0 | 0.45 | ***0.81*** | 0.52 |
| Baseline Range | All SNRs | ***0.88*** | ***0.90*** | ***0.98*** | **0.71** | **0.64** | **0.69** |
| | -12 dB | ***0.98*** | ***0.98*** | ***0.96*** | ***0.79*** | 0.51 | **0.68** |
| | -8 dB | ***0.98*** | ***0.98*** | ***0.99*** | 0.49 | **0.6** | **0.61** |
| | -4 dB | ***0.98*** | ***0.98*** | ***0.97*** | **0.6** | 0.54 | **0.74** |
| | 0 dB | ***0.98*** | ***0.98*** | ***0.98*** | 0.5 | 0.5 | 0 |
| | 4 dB | ***0.99*** | ***0.99*** | ***0.99*** | **0.61** | ***0.76*** | 0.54 |

Table 3.3: ICC values for all normalization procedures and SNRs between Visit 1 and 2. Values between 0.6 and 0.75, representing good reliability, are highlighted in bold and values above 0.75, representing excellent reliability, are highlighted in italic bold.

bold italic in the table. Negative ICC values were truncated to zero.

For all features using baseline correction, the ICC analysis showed good to excellent reliability, with ICC values equal to or greater than 0.6. The ICCs for all SNRs have comparable values to Spearman correlations. However, the ICC values varied across SNR without following a general trend. For both the PPD and MPD, high ICC values were observed for most of the SNRs (see Table 3.3) when comparing Visit 1 and 2. When applied on the GCA features of the pupil traces, good to excellent reliability (ICC above 0.6) was only found for 2 out of 5 of the SNRs for the slope and 1 out of 5 of the SNRs for the other features (AH, RF, delay). Thus, across all SNRs, the PPD and the MPD showed overall higher ICC values compared to the GCA features.

The range normalization provided good to excellent reliability for the traditional PPD with 4 out of the 5 SNRs when comparing Visit 1 and 2. Interestingly, none of the ICC values were above 0.6 for the MPD. The GCA features showed,

overall, poor-to-fair reliability between Visit 1 and 2 with a few exceptions (delay at -12 dB and -4 dB, slope at -12 dB and RF at 4 dB). When Z-score was applied as a normalization procedure, poor-to-fair reliability was obtained for PPD and MPD for all SNRs between Visit 1 and 2. Good-to-excellent reliability was obtained for only some of the GCA features (i.e., for RF, slope and delay), at only 2 out of the 5 SNRs.

When the data were baseline corrected and then range normalized within individuals, very high ICC values were observed for PPD, MPD and AH, indicating that these were the most reliable features across all SNRs between Visit 1 and 2.

The NASA-TLX was analysed to assess the perceived effort for each condition (Hart and Staveland, 1988). Participants were also asked to evaluate their effort on a scale from 0 to 10 after each condition. Reliability values (ICC) for the subjective listening effort assessments are summarized in Table 3.4 for Visit 1 and 2 and in Table 1 of the supplemental material for the subsample of 11 subjects for the three visits. For both NASA-TXL and the subjective self-report, good to excellent ICC values (above 0.6) were observed for -12, -8 and 4 dB SNR between Visit 1 and 2 but not for -4 and 0 dBs.

| ICC | Feature | Nasa Tlx | Subjective effort |
|---|---|---|---|
| | All SNRs | ***0.77*** | ***0.84*** |
| | -12 dB | ***0.87*** | **0.67** |
| | -8 dB | ***0.76*** | **0.68** |
| | -4 dB | 0.55 | 0.14 |
| | 0 dB | 0.42 | 0.57 |
| | 4 dB | ***0.84*** | ***0.75*** |

Table 3.4: ICC values for the subjective measures of effort, comparisons between Visit 1 and 2. Values between 0.6 and 0.75, representing good reliability, are highlighted in black bold and values above 0.75, representing excellent reliability, are highlighted in italic bold.

## 3.4 Discussion

The present study examined the reliability of the evoked pupil response in a speech-in-noise test paradigm to identify test conditions and analysis techniques that provide the highest test re-test reliability. Specifically, it was analyzed how task demands (manipulated through SNR changes) and data normalization impact the reliability of the evoked pupil response. Overall, the results showed

that data normalization procedures have the strongest impact and that certain procedures lead to high reliability in the pupil response.

It was hypothesized that reliability would be affected by the normalization procedure of the extracted pupil response. Thus, various normalization procedures that were recommended in previous literature were considered (McCloy et al., 2016; Piquado et al., 2010; Winn et al., 2018). These procedures included baseline correction, two different range normalization procedures and a Z-score normalization. The results indicate that the baseline correction procedure combined with range normalization provides the highest reliability results. High agreement (ICC results) was observed for the stationary features (i.e., PPD and MPD), but also for the AH feature extracted from the GCA. Similar values of AH and MPD were obtained using this normalization procedure, as expected. However, the Z-scores produced totally different results that might be explained by the different time period considered for the GCA features extraction than for MPD. A normalization procedure that takes into account the dynamic range of the pupil response has been suggested when comparing groups of different ages, or even when testing on different days (Piquado et al., 2010; Winn et al., 2018). The combination of a baseline correction and range normalization addresses the reactivity of the pupil response (i.e., high versus small dynamic range) and removes variance in the individual pupil response, which, provides high within-subject reliability across different visits as shown by the results presented here.

The lowest agreement across all conditions was obtained with the Z-score. Z-score calculations use the two statistical values (i.e., mean and standard deviation) to address inter-individual differences in variability in dilation. However, the Z-score assumes a normal distribution, and the pupil traces do not actually follow a normal distribution for all participants. In addition, not having a baseline on a trial level established when calculating the Z-score prior to the normalization process, produces higher disparities across SNRs and visits.

It was hypothesized that changes in task demands (manipulated through the SNR) would affect the reliability of the pupil features, such that higher reliability would be obtained for higher task demands. This was based on previous literature indicating increased reliability with increasing task difficulty (Aday and Carlson, 2019; Zwyghuizen-Doorenbos et al., 1988). Overall, the ICC values varied widely across SNRs, ranging from poor agreement to excellent agreement, and there was no clear trend between the SNR and the agreement. This is in line

with other previous literature suggesting that reliability is independent of SNR. For example, Giuliani et al., 2020 reported fair reliability for all test conditions, independent of SNR, and they also did not find a clear trend across SNR. While Giuliani et al., 2020 studied only relatively high SNR conditions ranging between 0 and -5 dB, the present study addressed a broader range of SNRs, including more challenging SNRs up to -12 dB (corresponding to an average of 25- 40% intelligibility). The results obtained in the present study were thus unexpected, rejecting the hypothesis of an increasing reliability with SNR.

Note that the task demands were manipulated by varying the SNR. However, participants differed in their performance for a given SNR, meaning that the task demands could differ across individuals at similar SNRs. Thus, examining the reliability at similar performance or intelligibility levels (instead of SNRs) might reveal a clearer relationship between reliability and task demand.

In contrast to Wendt et al., 2018, there was no evidence of disengagement in the group level analysis, which would have been illustrated by a reduced pupil response at the lowest SNRs (e.g., -12 dB) where speech recognition performance tends to be low. Despite this, some individuals did show some level of disengagement, as larger pupil reponses were observed at higher (e.g., -8 or -4 dB SNR as compared to lower (e.g., -12 dB SNR) indicating a reduction in effort investment when processing and studying individual's pupil response. The fact that disengagement was observed in only some individuals and that task demands seemed to differ across individuals for a given SNR could, taken together, explain why reliability was not increasing with SNR and, as was originally hypothesized.

A higher reliability for each of the pupil features was expected to be obtained between Visit 2 and 3 compared to Visit 1 and 2. This expectation was attributed to a potential global arousal or to the learning effect due to the novelty of the task that could occur in the first visit compared to the subsequent visits (Alamia et al., 2019; Widmann et al., 2018). However, there was no clear trend between reliability and the visit number in this study, and no overall arousal effect was observed across the visits either. Furthermore, these results do not support the assumption of potential learning effects either, even though several studies showed that the learning effect due to repeating the task over multiple sessions could be reflected in a decrease of PPD across repeated measurements (Foroughi et al., 2017; Sibley et al., 2011). In general, the results of this study show no significant impact of the visit on most of the pupil features, supporting the

assumption of reliability. Moreover, these results suggest that with a minimum of 3 weeks between the visits, no systematic change in the pupil response is seen with respect to its reliability. Instead, it further shows that high reliability can already be obtained with two visits when the data is normalized with baseline correction in combination with range normalization.

Note that only 11 participants out of the 31 were tested in the third visit, and, consequently, a comparison between the reliability at different visits was performed for only a subsample of 11 participants. Since ICC analysis requires a minimum of 30 participants in order to provide sufficient power Koo and Li, 2016, a Spearman correlation on this subsample of participants was performed to verify the conclusions. The ICC and Spearman's correlation results for Visit 1-2 were similar, such that no trend of correlation coefficients was found with increasing number of visits. Further testing with a larger sample of subjects participating in three visits would be needed to better clarify how the reliability changes with more than two visits.

Overall, it seems that the traditional pupil features (i.e., PPD and MPD) are more reliable than the temporal features. This finding is in line with other studies that only considered PPD and MPD as relevant features (Kramer et al., 1997; Wendt et al., 2018; Zekveld et al., 2010). Nonetheless, all the pupil features in the current study were, in one way or another, aggregated values of a time series of the pupil response. The aggregation of the pupil response over all trials and within the final trial can limit the understanding of the entire time series's and its associated reliability. This aspect was partly addressed by including the GCA temporal features. However, assessment of the reliability of the pupil response using non-aggregating methods could lead to a different conclusion.

The reliability of the subjective ratings of the listening effort (i.e., NASA Tlx and the subjective effort) was assessed, and the perceived listening effort showed in most of the cases reliability that was on par with the pupil features, in line with previous literature (Alhanbali et al., 2019; Giuliani et al., 2020). This study reported slightly higher, or similar reliability between measures of perceived effort and the PPD or MPD, irrespective of the normalization procedure applied. Similarly with the pupil features results, no clear patterns in the reliability of subjective effort across all of the pupil features, SNRs and normalization procedures.

Overall, several pupil features as potential indicators of listening effort, revealed high reliability only in some particular cases (i.e., baseline range normal-

ization procedure). Therefore, careful consideration of the data normalization procedure used when processing and studying individual's pupil response is recommended.

## 3.5 Summary and conclusion

The current study examined the reliability of pupillometry with several normalization procedures and feature extraction methods, while also assessing the impact of SNR and the number of visits on the resulting reliability. Overall, the results suggest that SNR and the number of visits only have a minor impact on the reliability of the pupil response, at least within a speech-in-noise test paradigm. Moreover, to obtain the highest reliability across SNRs, baseline correction combined with range normalization is recommended when analyzing the pupil response of individual listeners. Moreover, the stationary features (i.e., PPD and MPD) are the most reliable features. Overall, these reliability results provide valuable insights for determining the future of pupillometry as a potential diagnostic tool in the clinic.

# 4

# Towards a better understanding of the impact of listener factors on pupil responses in a speech-in-noise paradigm[c]


## Abstract

In the past decades, numerous studies examined pupillometry as a measure of listening effort. Various listener factors affecting the pupil response in an auditory task have been identified, including age, cognitive abilities and hearing loss. However, there has been conflicting evidence on the direction of the effects and their interaction. The present study examined a broad range of listener factors and their relative contributions to the variation of the pupil response. Thirty-one normal-hearing listeners participated in the study. The pupil response was measured during a speech-in-noise task at two different visits. Several individual factors (age, cognitive abilities, fatigue and motivation) were explored to evaluate their contribution to the variability of the peak pupil dilation (PPD) and the mean pupil dilation (MPD) across participants and visits. The results showed that motivation, age, and daily-life fatigue had the most substantial impact on the pupil response variability, whereby their relative contributions depended on the specific pupil feature (i.e., PPD vs. MPD). Furthermore, the listener factors' impact on the pupil response's dynamic range was examined during three different conditions: a speech-in-noise perception task, a mental arithmetic task and a no-task condition at rest (i.e., pupil response measured at rest in dark and light). The results showed age as the main contributor to the dynamic range of the pupil size extracted in the speech-in-noise condition as well as in the condition at rest,


[c] This chapter is based on Neagu et al., (2022b), in prep.

while the listeners' cognitive abilities, as well as their motivation, mainly affected the dynamic range in the mental arithmetic task. Overall, the findings may contribute to a better understanding of the role of listener factors on the observed pupil response variability.

## 4.1 Introduction

Pupillometry has been used as a physiological measure to assess listening effort in a listening task. Several studies suggested that the individual pupil response can be affected by interindividual factors related to the listener (Koelewijn et al., 2012b; Kramer et al., 2016; Kuchinsky et al., 2014; Peelle, 2018; Steinhauer et al., 2022; Tryon, 1975; Wang et al., 2018b; Zekveld et al., 2018, 2011). Zekveld et al., 2018 provided an overview of literature indicating that various listener factors affect the individual pupil response to auditory stimuli. The authors concluded that the pupil size could be sensitive to various factors, including hearing status, age, cognitive abilities, fatigue, or motivation, and emphasized the importance of assessing those factors when studying listening effort in individuals. Despite increasing evidence of the impact of listener factors on the pupil response, conflicting results have been reported regarding the direction of the effect, i.e., whether an increase of a given listener factor leads to an increase or a decrease in the pupil response. Furthermore, the relative contribution of each factor and the interaction between factors on the pupil response have not yet been evaluated. Disentangling the contribution of listener factors to the variability of the individuals' pupil responses is an important prerequisite for a valid interpretation of the pupil size towards a clinical application. The present study aimed to identify the most relevant listener factors and their relative contribution to the pupil size and to the dynamic range measured within a speech-in-noise task.

Age has been identified as a contributor that modulates pupil response. Numerous studies have examined the role of age on the pupil size and its dynamic range (Bitsios et al., 1996; Kim et al., 2000; Koch and Janse, 2016; Morris et al., 1997; Steel et al., 2015; Winn et al., 1994; Zekveld et al., 2011). It has been shown that increasing age relates to smaller pupil size in adults as compared to infants (Karatekin, 2004; Wetzel et al., 2016). Furthermore, a reduced dynamic range of the pupil has been observed in older listeners when compared with younger listeners (Piquado et al., 2010). Consequently, Piquado et al., 2010

suggested normalization methods that account for differences in the dynamic range depending on age. However, several other studies could not confirm such a relationship between age and pupil response (Ayasse and Wingfield, 2020; Chaney et al., 1989; Koelewijn et al., 2012b; Kuchinsky et al., 2016; Morris et al., 1997). Hence, the contribution and importance of age as a listener factor impacting the pupil remains unclear.

A few studies examined the hearing status as another factor impacting the individual pupil response. While Kuchinsky et al., 2014 observed that more severe hearing loss was associated with a flatter pupil dilation response, other studies did not find any effects of the hearing status on the pupil response (Koelewijn et al., 2017; Kuchinsky et al., 2016). Overall, the literature seems to have shown mixed results.

The association between cognitive abilities and the pupil response to an auditory task has been examined in several studies (see Zekveld et al., 2018 for an overview). A direct link between different aspects of cognitive abilities and the pupil response measured during speech processing has been found in some studies (Koch and Janse, 2016; Koelewijn et al., 2012b; Kuchinsky et al., 2016; Wendt et al., 2017), while other studies did not find an effect (Koelewijn et al., 2014; Zekveld and Kramer, 2014). For example, Koelewijn et al., 2012b showed that linguistic abilities, as indicated by higher performance in a text reception task suggesting better inhibition of irrelevant speech, were positively correlated with larger pupil dilation. Furthermore, Wendt et al., 2016 reported that a higher working memory capacity (WMC) was associated with increased pupil responses for people with normal hearing. However, for people with hearing impairment, a higher WMC was negatively correlated with the pupil response during sentence recognition (Wendt et al., 2017). The contradictory findings regarding the effect of cognitive abilities on pupil size depending on the hearing status may reflect an interaction of different listener factors, but their relative contributions need to be disentangled to explain the variability in the pupil size.

The Framework for Understanding Effortful Listening (FUEL, Pichora-Fuller et al., 2016) assumes that mental fatigue impacts effort allocation and, furthermore, the pupil size during a listening task (Pichora-Fuller et al., 2016). Fatigue has been defined as a mood state or subjective experience (Hornsby et al., 2016). Bafna and Hansen, 2021 described mental fatigue as a 'subjective feeling associated with a reduction in mental resources and a reduced motivation that

develops with sustained cognitive effort over time and that can impact task performance'. Wang et al., 2018a investigated the effects of daily-life fatigue (assessed by the Need for Recovery questionnaire; Van Veldhoven and Broersen, 2003 on the pupil response and reported that higher fatigue levels were associated with smaller pupil size. Furthermore, Hopstaken et al., 2015 reported that increased mental fatigue coincided with diminished stimulus-evoked pupil dilation. While the aforementioned studies provide evidence that a higher fatigue level led to reduced pupil size, some studies indicated that motivation might reduce or even eliminate the impact of fatigue on the pupil response. The role of motivation has been increasingly studied within the past decade, and several studies demonstrated that the pupil response was sensitive to the listener's motivation within a listening task (Koelewijn et al., 2018; Peelle, 2018; Pichora-Fuller et al., 2016; Pielage et al., 2021). It was argued that, although a person might have sufficient cognitive capacity to perform a task, low motivation can result in task disengagement (Peelle, 2018; Pichora-Fuller et al., 2016). Koelewijn et al., 2018 showed that higher motivation results in increased pupil dilation during listening and, consequently, in increased listening effort. Similarly, Hopstaken et al., 2015 indicated that sufficient rewards (to manipulate motivation) helped restore fatigue's diminishing effect on pupil dilation and argued that the listener's motivation is relevant when predicting engagement versus disengagement during fatigue.

In summary, numerous listener factors have been suggested to impact the individual pupil size. Some factors are more static (such as age or hearing status), while others are more dynamic (such as motivation and mental fatigue) and may change over time during the task or across the different visits. Although some studies reported an interaction of various listener factors with respect to the pupil response, most of the aforementioned studies focused on the impact of only a single factor. For example, the impact of cognitive abilities has typically been studied independently from other potential contributors, such as fatigue or motivation. Furthermore, Zekveld et al., 2018 emphasized a lack of knowledge with respect to the relative contribution of those different listener factors on pupil size. Hence, gaining a complete picture of the interplay of listener factors and their relative contribution to pupil size would be valuable in order to validly interpret individual listening effort.

Furthermore, this study explores the impact of the listener factors mentioned above (both static and dynamic) on the *dynamic range* of the pupil

response since some literature suggested to account for individual differences by applying a range normalization procedure (Einhäuser et al., 2008; Winn et al., 2018). Some studies extracted the dynamic range at rest during dark and light conditions (Piquado et al., 2010); others argued in favor of measuring the dynamic range using a cognitive task (Ayasse and Wingfield, 2020; Winn et al., 2018). However, little is known about how the listener factors affect the dynamic ranges of the pupil response. Ayasse and Wingfield, 2020 investigated the correlations of hearing status and age on the dynamic range of the pupil response, showing a smaller dynamic range in older as compared to younger adults. No correlation between the dynamic range of the pupil response and the hearing status was measured. The difference in the dynamic ranges between individuals can, nevertheless, also be driven by other (more dynamic) listener factors that might vary over the period of an experiment, such as motivation or fatigue.

Overall, this study aimed to explore the relative contribution of listener factors such as fatigue, motivation, cognitive abilities, age and hearing status to: 1) the variability of different pupil features, including peak and mean pupil dilation across individuals; 2) the change of different features across multiple visits; and 3) the dynamic ranges of the pupil response.

## 4.2 Methods

### 4.2.1 Participants

Thirty-five normal-hearing (NH) participants (between18-65 years, mean 31 years) participated in this study. The participants were native Danish speakers and had pure-tone hearing thresholds in both ears of 20 dB hearing level (HL) or better at frequencies below 6 kHz and 30 dB HL or better at frequencies above 6 kHz. The participants had no history of eye disease or eye operations and were asked to avoid any caffeine intake for the 3 hours leading up to the experimental sessions to avoid caffeine-induced arousal effects. Four participants out of the thirty-five were excluded from the analysis as they withdrew from the study during the course of the investigation. The research procedure was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). All participants read and signed informed consent to participate in this study and were offered monetary compensation for their participation.

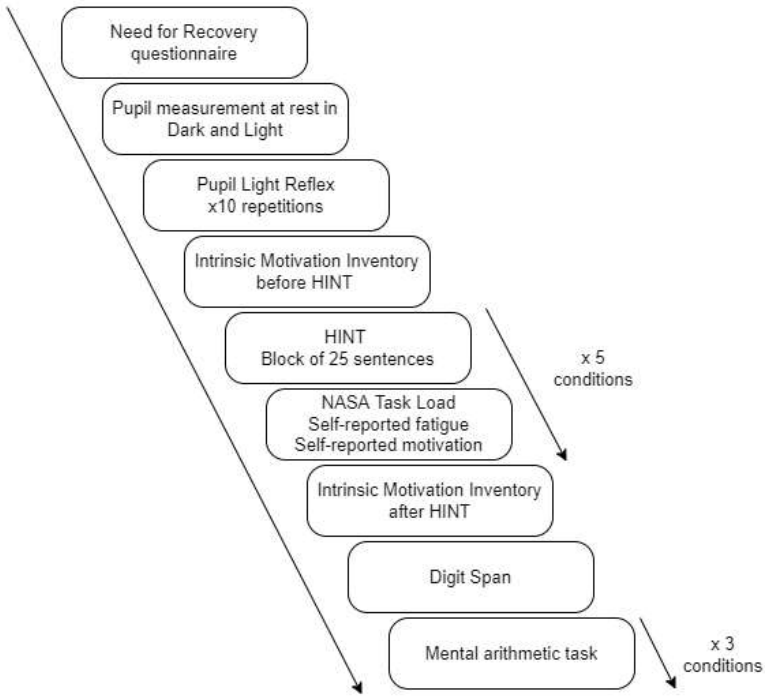



Figure 4.1: Experimental procedure within a visit containing different measurements: Need for Recovery questionnaire, pupil measurements at rest in dark and light conditions, Pupil Light Reflex, Intrinsic Motivation Inventory before and after HINT, HINT test at five SNRs ranging between -12 to 4 dB, self-reported measures (NASA Task Load, self-reported fatigue, self-reported motivation), Digit Span, Mental arithmetic task.

### 4.2.2 Procedure and stimuli

The investigation was conducted using a repeated measurements study design, where all participants attended at least two visits, and eleven of them attended a third visit. The visits were separated by an interval of at least three weeks, but at most six weeks, to avoid any learning effects of the speech material (Bramsløw et al., 2016). All visits were scheduled at the same time of the day and at the same period of the week (i.e., beginning, middle, or end) to minimize potential changes in fatigue throughout the day or on different days of the week (Daguet et al., 2019). The procedure consisted of a sequence of tests, as illustrated in Figure 4.1. The procedure was kept fixed across visits. The individual tests are described below. The pupil response was recorded throughout the procedure.

***Need for Recovery***

The participants completed the Need for Recovery (NfR; Van Veldhoven and Broersen, 2003) questionnaire, assessing the individuals' subjective chronic fatigue at the beginning of the session. The questionnaire contained 11 questions. The participants were asked to provide 'yes/no' answers to each question, and the percentage of 'yes' answers was calculated. The scores were normalized to a scale from zero to one.

***Pupil response at rest***

At the beginning of each visit, the pupil response at rest was measured for 30 seconds in bright light (ambient light 102 lx) and, after 20 seconds of adaptation, in complete darkness (0.2 lx) for another 30 seconds. The participants were instructed to fix their gaze on a cross in the middle of the screen.

***Pupil light reflex (PLR)***

Next, the pupil light reflex (PLR) was measured. The PLR task was performed in darkness after 20 seconds of adaptation to the dark, and the stimuli consisted of 10 consecutive light flashes (light green screen) presented every 15 seconds. To quantify the parasympathetic activity of the autonomic system provided through PLR, the maximum constriction velocity (MCV) was extracted (Wang et al., 2018b).

***Speech-in-noise test (HINT)***

The Danish Hearing in Noise Test (HINT; Nielsen and Dau, 2011) was performed at five different signal-to-noise ratios (SNRs) of 4 dB, 0 dB, -4 dB, -8 dB and -12 dB in the presence of a four-talker babble masker with the same long-term average spectrum as the HINT sentences. Specifically, the masker consisted of voices from two male and two female talkers reading various overlapping excerpts from a newspaper. A list of 25 sentences was presented for each SNR in a block-based design. The sound pressure level (SPL) of the noise was fixed at 70 dB at both ears, and the noise started 3 seconds prior to the sentence and ended 3 seconds after the sentence. A baseline was established one second before the sentence onset, as recommended by Winn et al., 2018. The participants were instructed to fix their gaze on a cross in the middle of the screen and repeat the

target sentence aloud after the noise offset. The responses were scored on a word-level basis.

#### *NASA Task Load*

After each SNR-block of sentences for HINT, the participants answered the NASA Task Load (NASA TLX; Hart and Staveland, 1988) questionnaire. The questionnaire contained six subitems: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration (translated in Danish to “psykisk krav”, “fysisk efterspørgsel”, “midlertidig efterspørgsel”, “præstation”, “indsats” and “frustration”), which participants rated on a scale from 1 (low) to 20 (high).

#### *Self-rated tiredness and tendency to give up*

Directly following each NASA Task Load questionnaire presentation, the participants also rated their level of tiredness and their tendency to give up on the task after each block of sentences. To do so, the participants were asked, on a scale from 1 to 10, to answer the following questions: ‘How tired were you while listening to the sentences?’ and ‘How often did you have to give up on understanding the sentence?’.

#### *Intrinsic Motivation Inventory*

The Intrinsic Motivation Inventory (IMI; Ryan, 1982) questionnaire was performed at the beginning, middle, and the end of the entire HINT task (i.e., before the first block, after the third block and after the last block) to assess the participants’ engagement in the task. The questionnaire contained 22 questions with four subscales: enjoyment, perceived choice, perceived competence, and pressure. The enjoyment subscale is considered the self-report measure of intrinsic motivation. Perceived choice and perceived competence are considered to be positive predictors of self-reported and behavioral measures of intrinsic motivation (Ryan, 1982). Pressure is considered to represent a negative predictor of intrinsic motivation (Ryan, 1982). The participants were asked to answer these questions on a scale from 1 to 7. Following the recommended scoring procedure, a subset of the questions was scored using reverse scoring (i.e., the response was subtracted from 8). The final scores were normalized between zero and one.

***Digit span test***

After the five blocks of HINT and the corresponding questionnaires, working memory capacity (WMC) was measured using a digit span task, both forwards and backwards (Wechsler, 1981). In the forward version, the participants were asked to repeat the sequence of digits presented. The number of digits presented varied from two to eight. At every third presentation, the number of digits increased by one. In the backward version (reversed digit span), the participants were asked to repeat the sequence of digits in reverse order. According to the traditional scoring, one point was awarded for each correctly repeated sequence (Tewes, 1991) and presented as a percentage of correct scores (reflecting how many out of 14 possible sequences were repeated correctly). The scores were then normalized to a scale from zero to one.

***Mental arithmetic test***

A mental arithmetic test was conducted at three levels of difficulty: easy, intermediate, and difficult (Klingner, 2010; Marquart and De Winter, 2015). The test included ten calculations (multiplications between two numbers) at each level of difficulty such that a stable pupil response was obtained by averaging the pupil responses within each level of difficulty. The difficulty levels of the multiplications were established as recommended in (Marquart and De Winter, 2015).

### 4.2.3 Physical setup

The participants were seated in a sound-isolated booth on a chair fixed in place in front of a desk, which had a computer screen and mounted desktop-eye tracker placed on top of it. Different graphical interfaces, implemented in Matlab (MATLAB, 2018) corresponding to the different measurements, were running on a computer outside the booth, which then synchronized with the screen inside the booth using Psychtoolbox 3 (Brainard and Vision, 1997; Kleiner et al., 2007; Pelli and Vision, 1997). The answers to the questionnaires were provided on paper. For the speech recognition, digit span and mental arithmetic measurements, verbal responses provided by the participants were sent through a Shure WH20 microphone to the experimenter sitting outside of the booth actively scoring the responses. The stimuli were presented through HD650 headphones using an SPL Audio Phonitor Mini amplifier. The experimenter

could communicate with the participants through a talk-back t.bone GM5212 microphone during breaks in testing. A Fireface 802 sound card was used to connect to the amplifier, the headphones and the microphones.

### 4.2.4 Apparatus

Eye-tracking data were recorded using an EyeLink 1000 long-range eye tracker. The pupil size was recorded from the left eye at a rate of 500 Hz. The ambient light was measured at the head of the participant and corresponded to 75 lx for the tasks performed in light and to 0.2lx for the tasks performed in darkness. The screen displayed a black background with a grey cross in the middle, resulting in brightness of approx. 9 cd/m2. The distance from the participant's eyes to the middle of the screen was 50-70 cm.

### 4.2.5 Pupillometry data processing

The pupil data were processed using MATLAB (MATLAB, 2018) and R (R Core Team, 2019). During the HINT task, the pupil traces from the first three sentences of each block (out of 25) were excluded from the analysis (Wendt et al., 2018; Winn et al., 2018). The entire sentence block was linearly detrended to remove any decreasing trend observed towards the end of a block. Trials with less than 80% reliable data were removed from the analysis, and the remaining traces were baseline corrected. The data cleaning was otherwise performed, as reported in Wendt et al., 2018.

Similarly, in the PLR and the mental arithmetic task, the eye blinks were removed by a linear interpolation that started 80 ms before and ended 150 ms after the blinks. The data were smoothed using a moving average filter with a symmetric rectangular window of 117 ms. Trials with less than 80% reliable data were removed from the analysis. All remaining traces in the PLR task were scaled using a baseline correction normalization.

### 4.2.6 Data analysis and feature extraction

The peak pupil dilation (PPD) and mean pupil dilation (MPD) were extracted from the pupil trace of each individual participant in the speech-in-noise task, averaged across conditions and visits. PPD and MPD were extracted in the time window between stimulus presentation and noise offset, i.e., between 3 seconds and 8 seconds. Furthermore, three different dynamic ranges were

| Listener factors | Abbreviation | Definition |
|---|---|---|
| Age | Age | Years of age at the time of testing |
| Pure Tone Average | PTA | Average of the pure-tone hearing thresholds at frequencies tested between 500 Hz and 4000 kHz, obtained for the better ear. |
| Motivation before HINT task | Motiv_start | Response to the Intrinsic Motivation Inventory (Ryan, 1982) questionnaire at the beginning of HINT task |
| Motivation after HINT task | Motiv_end | Response to the Intrinsic Motivation Inventory (Ryan, 1982) questionnaire at the end of the speech-in-noise task |
| Need for Recovery | NfR | Response to the Need for Recovery questionnaire (Van Veldhoven and Broersen, 2003) as a measure of subjective chronic fatigue |
| Maximum Constriction Velocity | MCV | Maximum slope of the constriction of the pupil in the Pupil Light Reflex task (PLR) |
| Reverse Digit Span | RDS | Response to the Reverse Digit Span (Tewes, 1991) as a measure of working memory capacity calculated according to traditional scoring |

Table 4.1: List of listener factors asses in the experiment, their abbreviations and definitions.

extracted from the pupil response measured in three different tasks. First, the dynamic range during the speech-in-noise task (DR_SiN) was computed for each individual by subtracting the minimum pupil dilation from the maximum pupil dilation across all conditions (SNRs) and visits. Second, the dynamic range at rest (DR_DL) was estimated by assessing the pupil dilation in darkness (to extract the maximum pupil dilation) and in a light condition (to extract the minimum pupil dilation). Third, the dynamic range was extracted during the cognitive task, i.e., while participants performed the mental arithmetic task (DR_Cog), by subtracting the minimum pupil dilation (extracted in the easy condition) from the maximum pupil dilation (extracted in the difficult condition). In order to minimize the impact of outliers, the minimum and the maximum dilation were calculated by estimating the 95% and 5% percentile of the calculated MPDs. The dynamic ranges were not considered to change across visits.

**The impact of listener factors on pupil dilation**

All listener factors (i.e., factors that are related to the individual) were standardized to have a mean of zero and a standard deviation of one. The complete list of listener factors is described in Table 4.1.

In order to investigate the impact of listener factors on PPD and MPD, a Pearson's correlation analysis was performed. PPD and MPD were extracted for 31 participants across all conditions for visits 1 and 2. Subsequently, the magnitude of the listener factors' contribution to the pupil features was estimated using a mixed-effects model. The model included the factors 'participant' and 'SNR' as random effects and the 'listener factors' as fixed effects. The SNR was considered as a random effect in order to address the pupil response as an overall measure and not at specific SNRs. The model is described in Eq. 4.1. In order to account for factors that affect the variability of PPD and MPD across visits, the base model in Eq. 4.1 includes 'Visit' (and all possible interactions) as fixed effects.

$$\begin{aligned}\text{pupil feature} \sim \text{Visit} * (&\text{Age} + \text{PTA} + \text{Motiv\_start} + \text{Motiv\_end} + \\ &\text{NfR} + \text{MCV} + \text{RDS}) + (1|\text{participant}) + (1|\text{SNR})\end{aligned} \tag{4.1}$$

While the correlation results purely address the direction and strength of the linear association between two variables (listener factor and pupil feature), the mixed-effect model was chosen to provide more information regarding the hierarchy of the contributions of these factors to the pupil features. Moreover, the model allows exploring the pupil features changes across visits. The advantage of the mixed-effect model is that the predictions could provide valuable information on how pupil features are affected by different categories in the explanatory variables (such as high/low motivation and high/low cognitive abilities). Only the data from the first two visits were analyzed in the statistical model, as the sample size for the third visit was significantly smaller.

The effect size of each factor and its interactions with Visit was calculated using Eq. 4.2 (Brysbaert and Stevens, 2018; Westfall et al., 2014). The numerator in Eq. 4.2 represents the coefficient extracted from the model results for each fixed effect. The denominator represents the square root of the sum of the variance for each random effect considered in the model and the residual.

$$d = \frac{\text{estimate}}{\sqrt{\text{varintercept}_{\text{participant}} + \text{varintercept}_{\text{SNR}} + \text{var}_{\text{residual}}}} \tag{4.2}$$

**Dynamic range**

To study the impact of listener factors on the dynamic range of the pupil response, first, a correlation analysis between every two dynamic ranges was assessed using Pearson correlations (i.e., DR_SiN vs. DR_Cog, DR_SiN vs. DR_DL, DR_Cog vs. DR_DL). Subsequently, the magnitude of the listener factors' contribution to these dynamic ranges was estimated with a mixed-effect model. The correlation analysis provides information in terms of how strongly these dynamic ranges are related to each other. Based on this analysis, it was decided whether an explorative analysis of the listener factors on the dynamic range should be applied individually for each dynamic range. The model included the participant as a random effect and the listener factors as fixed effects. The model is described in Eq. 4.3.

$$\begin{aligned}\text{Dynamic range} \sim (&\text{Age} + \text{PTA} + \text{Motiv\_start} + \text{Motiv\_end} + \text{NfR} + \\ &\text{MCV} + \text{RDS}) + (1|\text{participant})\end{aligned} \tag{4.3}$$

The effect sizes of each of these factors were thereby calculated using Eq. 4.4 (Brysbaert and Stevens, 2018; Westfall et al., 2014).

$$d = \frac{\text{estimate}}{\sqrt{\text{varintercept}_{\text{participant}} + \text{var}_{\text{residual}}}} \tag{4.4}$$

## 4.3 Results

### 4.3.1 Listener factors and their contribution to PPD & MPD

**Correlation analysis**

Figure 4.2 shows the results of the correlation analysis between all listener factors. A moderate correlation of PTA with age (Corr = 0.494, p<0.001) and a strong correlation between Motiv_start and Motiv_end (Corr = 0.794, p<0.001) were found.

Table 4.2 shows the Pearson correlations of each listener factor with PPD and MPD. The results represent significant correlations of both PPD and MPD

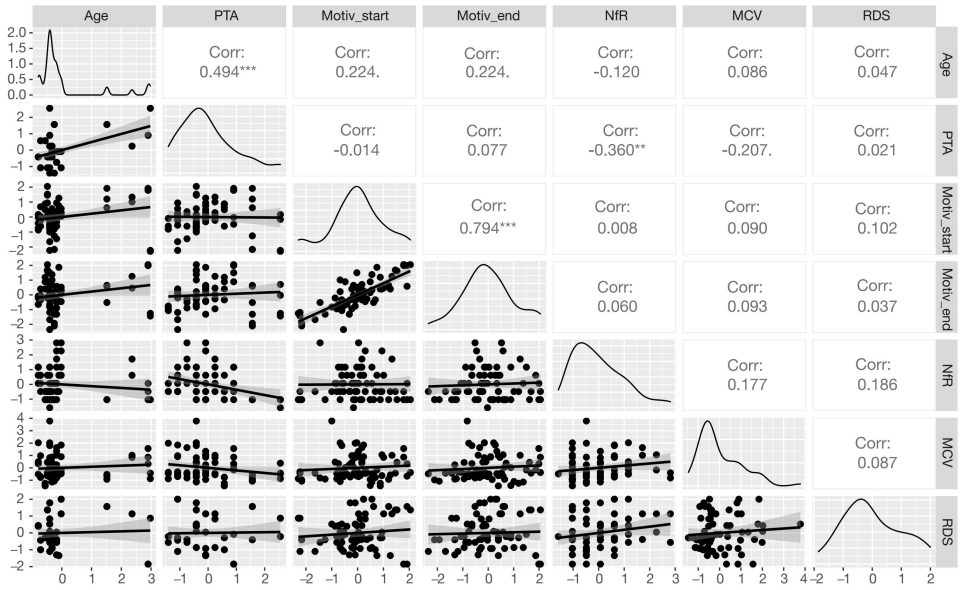



Figure 4.2: Correlations analysis of all listener factors. The correlations' values (Corr) and significance levels are displayed in the upper part of the plot. The density of each variable is displayed on the diagonal. The lower part displays scatter plots representing Pearson correlations between different variables. Significance levels were defined as follows: * p< 0.05, ** p< 0.01, ** p< 0.001.

| | **Age** | **PTA** | **Motiv_start** | **Motiv_end** | **NfR** | **MCV** | **RDS** |
|---|---|---|---|---|---|---|---|
| ***PPD*** | -0.175 | -0.099 | 0.306** | 0.250* | -0.214 | 0.206 | 0.070 |
| ***MPD*** | 0.158 | 0.196 | 0.418*** | 0.469*** | -0.061 | 0.064 | 0.168 |

Table 4.2: Correlation coefficients between pupil features (PPD and MPD) and listener factors. Significance levels were defined as follows: * p< 0.05, ** p< 0.01, ** p< 0.001.

with Motiv_start (PPD: Corr = 0.306, p<0.01; MPD: Corr = 0.418, p<0.001) as well as with Motiv_end (PPD: Corr = 0.250, p<0.05; MPD: Corr = 0.469, p<0.001), indicating increased pupil features (PPD, MPD) with increasing motivation both at the beginning and the end of the speech-in-noise task. No other significant correlations were found between the pupil features (PPD, MPD) and the other listener factors.

### Mixed-effects model

All listener factors were investigated as predictors of the two pupil features (i.e., PPD, MPD) using mixed-effects models. A model reduction was performed for each model by stepwise elimination of statistically nonsignificant terms until

only the significant effects were kept in the model. Table 4.3 shows the results of the final models. The model showed that motivation at the end of the task (Motiv_end) has a significant positive effect on both PPD ($p<0.05$) and MPD ($p<0.001$). Thus, on average, the participants who reported higher motivation levels at the end of the task tended to have high PPD and MPD during the task. Moreover, the results showed a significant ($p<0.05$) negative impact of fatigue (NfR) on PPD but not on MPD, indicating lower PPDs for individuals that reported higher fatigue. The model also revealed a significant ($p<0.05$) effect of age on PPD, such that younger people tended to have a higher PPD than older people.

Based on the model's results (regarding the impact of age, motivation and fatigue on PPDs), a test dataset was constructed to predict the PPD. This test dataset contained the same variables as the initial dataset (training dataset) with all combinations of extreme values for motivation (high vs. low) and fatigue (NfR high vs. low), while age was divided into two categories ('young' representing people below 30 and 'old' representing people over 50). The predictions of PPDs for this test dataset are illustrated in Figure 4.3. Overall, PPD was higher for young vs. old participants, as well as for highly vs. lower motivated participants. Furthermore, the highest PPD was found for young people with high motivation at the end of the task and a low level of fatigue (right panel, blue line). The lowest PPD values were obtained for elderly participants that reported low motivation and higher fatigue (left panel, yellow line in Figure 4.3).

Regarding the random effects, the variance explained by SNR (4.6%), as a random effect in the model predicting PPD, was much lower than that explained by the individual (38.12%).

The model showed no impact of Visit on PPD and MPD. However, significant positive interactions of Visit with RDS ($p<0.01$, effect size = 0.269) and significant negative interactions with age ($p<0.05$, effect size = -0.194) on MPD were found, indicating that both listener factors had an impact on the variability in MPD from Visit 1 to 2, with cognitive abilities showing a stronger interaction, as indicated by the effect size.

Based on the model's results (regarding the impact of cognitive abilities, motivation, age, and Visit on MPD), a test dataset was constructed to predict the MPD for groups of high vs. low motivation as well as high vs. low levels of cognitive abilities for both visits. Figure 4.4 shows the MPD predictions for low motivation (left panel) vs. high motivation (right panel) and its changes for

| | *PPD* | | | *MPD* | | |
|---|---|---|---|---|---|---|
| | ***Estimate*** | ***Effect sizes*** | ***p-values*** | ***Estimates*** | ***Effect sizes*** | ***p-values*** |
| Visit | | | | -0.122 | -0.135 | 0.1292 |
| Age | -0.282* | -0.315 | 0.0155 | 0.109 | 0.121 | 0.2456 |
| PTA | | | | | | |
| Motiv_start | | | | | | |
| Motiv_end | 0.195* | 0.218 | 0.0350 | 0.275*** | 0.305 | 0.0006 |
| NfR | -0.292*** | -0.326 | 0.0001 | | | |
| MCV | 0.142* | 0.159 | 0.0320 | | | |
| RDS | | | | -0.068 | -0.075 | 0.4756 |
| Interaction Visit : Age | | | | -0.175* | -0.194 | 0.0337 |
| Interaction Visit : PTA | | | | | | |
| Interaction Visit : Motiv_start | | | | | | |
| Interaction Visit : Motiv_end | | | | | | |
| Interaction Visit : NfR | | | | | | |
| Interaction Visit : MCV | | | | | | |
| Interaction Visit : RDS | | | | 0.242** | 0.269 | 0.0039 |
| Constant | -0.014 | -0.016 | 0.783 | 0.048 | 0.053 | 0.8362 |

Table 4.3: Coefficients of the models, effect sizes and level of significance of the listener factors were used to predict PPD and MPD. The interaction of the listener factors with Visit provides information about the impact of listener factors on the variability of pupil features. Significance levels were defined as follows: * p<0.05, ** p<0.01, *** p<0.001.

different visits depending on different levels of cognitive abilities, represented by the blue and yellow symbols reflecting low vs. high levels of cognitive ability, respectively. The results show a drop in MPD across visits for people with low cognitive abilities, while the opposite occurs for people with high cognitive

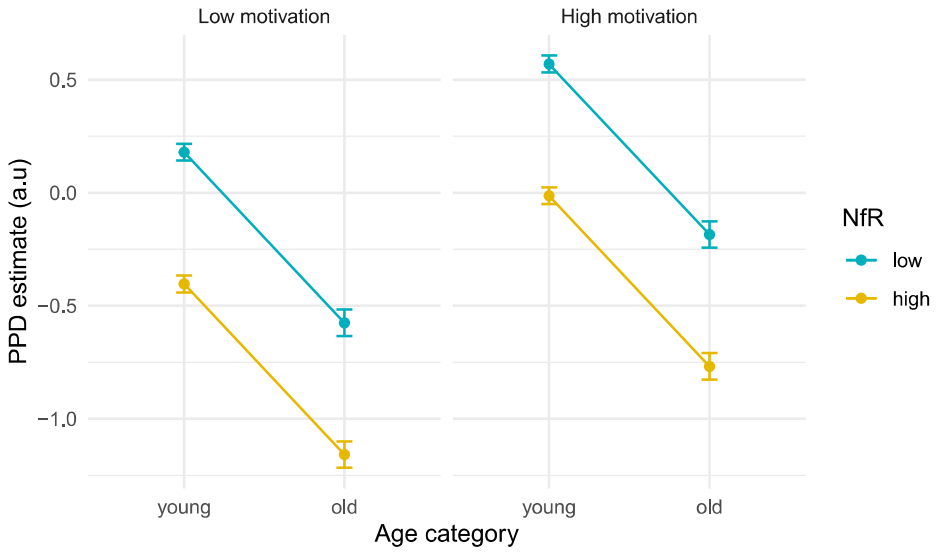



Figure 4.3: PPD estimates of the model for two age categories (young/old) in the cases of low-high motivation and low-high NfR score. The different panels represent low or high motivation at the end of the task, while the different colors represent low or high NfR scores. The error bars indicate the standard deviations of PPD for the participants with low vs. high NfR.

abilities. Interestingly, motivation (rated at the end of the task) had a stronger impact ($p<0.001$, effect size = 0.305) than the interaction effects since the MPD seemed to be overall higher for people with high motivation (Motiv_end) than for people with relatively low motivation, regardless of their cognitive abilities (RDS) and regardless of 'Visit'.

Furthermore, the variance explained by SNR (25.27%), as a random effect in the model predicting MPD, was higher than that explained by the individual (18.95%).

### 4.3.2 The impact of listener factors on dynamic ranges of pupil response

Figure 4.5 shows dynamic ranges (DR_SiN, DR_Cog, DR_DL) extracted for the individuals. The dynamic ranges were normalized to the minimum of the dark-light condition as it was assumed that the minimal dilation should be measured at rest for the light condition, i.e., the minimum of the DR_DL was set to 0.

A visual inspection of Figure 4.5 shows that the largest dynamic range was obtained for most of the participants at rest when changing the light condition (DR_DL). The speech-in-noise task (DR_SiN) tested at a broad range of SNRs

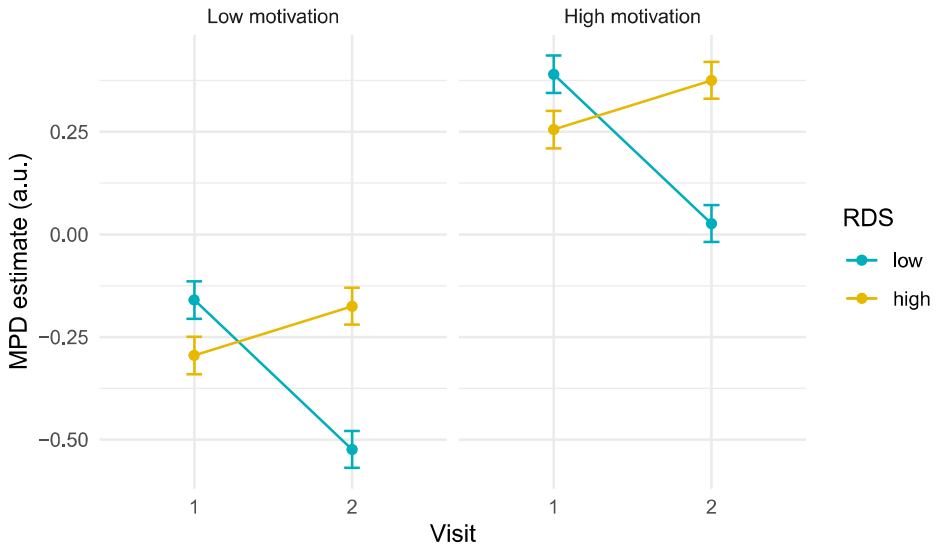



Figure 4.4: MPD estimates of the model for the two repeated visits in the cases of low-high motivation and low-high RDS (cognitive abilities). The different panels represent low or high motivation at the end of the task, while the different colors represent low or high RDS (cognitive abilities). The error bars indicate the standard deviations of MPD for the participants with low vs. high RDS.

produced a larger dynamic range as compared to the cognitive task (DR_Cog) for most of the participants, indicating a larger range of effort allocation than the mental arithmetic test.

It was expected that DR_SiN and DR_Cog ranges would be within the dark-light dynamic range. However, there were a few participants for which the minimum pupil response in the speech-in-noise task was even lower than in the light condition (e.g.,TP26, TP27, TP10, TP6). This can be explained by the fact that the minimum pupil dilation measured at rest was not an absolute individual minimum (i.e., generated by a potential brightest condition) since the luminance was lower than a more realistic scenario (daylight). Thus, a higher constriction in the speech-in-noise task is not surprising and can be caused by other cognitive processes.

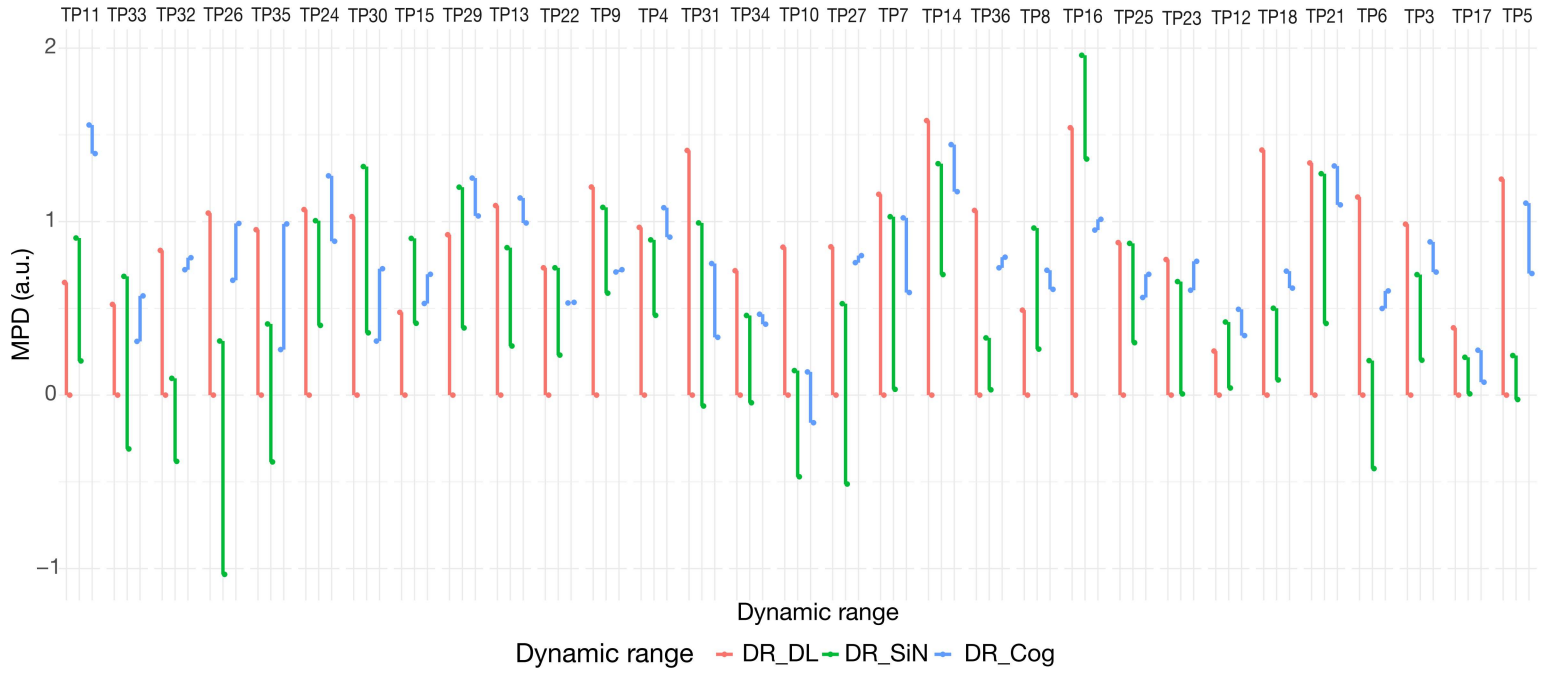



Figure 4.5: Individual dynamic ranges obtained in three different tasks as indicated by different colors, i.e., DR_DL in red, DL_SiN in green and DL_Cog in blue. The dynamic range is represented by the vertical lines drawn between the minimum and maximum values in each of the three tasks. The dynamic ranges were normalized according to the minimum of the DR_DL (which was set to 0). The participants are ordered based on the dynamic range obtained in the speech-in-noise task (green), with the participants with the largest and smallest dynamic range displayed as the first and the last, respectively.

| | *DR_SiN* | | | *DR_Cog* | | | *DR_DL* | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Estimates* | *Effect size* | *p-values* | *Estimates* | *Effect size* | *p-values* | *Estimates* | *Effect size* | *p-values* |
| Age | -0.347* | -0.365 | 0.019 | | | | -0.331* | -0.350 | 0.029 |
| PTA | | | | | | | | | |
| Motiv_start | | | | -0.405* | -0.441 | 0.020 | | | |
| Motiv_end | | | | 0.590*** | 0.643 | 0.001 | 0.294* | 0.311 | 0.034 |
| NfR | | | | | | | | | |
| MCV | | | | | | | | | |
| RDS | | | | 0.289* | 0.315 | 0.048 | | | |
| Constant | 0.009 | 0.009 | 0.947 | -0.504 | -0.549 | 0.887 | -0.015 | -0.016 | 0.916 |

Table 4.4: Coefficients of the models, effect sizes and level of significance of the listener factors used to predict the dynamic ranges of the pupil size measured in different tasks. Significance levels were defined as follows: * p<0.05, ** p<0.01, ** p<0.001.

A correlation analysis between the different dynamic ranges was explored. No significant correlation was found for any of the combinations (DR_SIN vs. DR_Cog: Corr = 0.032, p = 0.86, DR_SIN vs. DR_DL: Corr = 0.13, p = 0.5, DR_Cog vs. DR_DL Corr = 0.061, p = 0.74). Hence, all listener factors were investigated as predictors of the three different dynamic ranges (i.e., DR_Sin, DR_Cog, DR_DL) using mixed-effects models. Table 4.4 shows the corresponding results. Age was found to have a strong impact on the dynamic range for both the DR_SiN (p<0.05, effect size = -0.365) and DR_DL (p<0.05, effect size = -0.350). Older participants tended to have a smaller dynamic range in the speech-in-noise task and at rest. In contrast, motivation affected the dynamic range during the cognitive task (DR_Cog) and at rest (DR_DL), such that people with higher-rated motivation at the end of the speech-in-noise task tended to have a larger pupil range in the mental arithmetic task (p<0.001, effect size = 0.643) as well as at rest during dark vs. light (p<0.05, effect size = 0.311). The motivation rated at the end vs. the beginning of the task seemed to have opposite effects on the dynamic range in the cognitive task, with an overall larger effect size regarding the motivation rated towards the end of the task (Motiv_start: p<0.05, effect size = -0.441; Motiv_end: p<0.001, effect size = 0.643). Furthermore, cognitive abilities impact DR_Cog (p<0.05, effect size = 0.315), suggesting that higher

cognitive abilities result in a large cognitive dynamic range.

## 4.4 Discussion

The current study investigated the impact of various listener factors on pupil features measured in a speech-in-noise task across repeated visits. A broad variation of listener factors was considered, including factors that either have been reported to affect the pupil response during an auditory task (such as age, hearing status or cognitive abilities; see Zekveld et al., 2018 for a review) or were expected to impact effort allocation as predicted by existing theories on resource allocation (such as fatigue or motivation; see Pichora-Fuller et al., 2016). Specifically, the impact of listener factors on two commonly used pupil features (PPD and MPD) as well as on the variations of the pupil features across visits were examined. Finally, the association of listener factors and the dynamic range of the pupil response was explored. Overall, the results indicated that most of the considered listener factors interfered with the pupil features and their variation across repeated visits. Also, the dynamic range of the pupil response seemed to be affected by various listener factors. Out of all factors tested in the current study, motivation and fatigue showed the strongest impact on the pupil response. However, their relative impact depended on the pupil feature (PPD vs. MPD).

### 4.4.1 Role of motivation and fatigue

The data obtained in the present study indicated a moderate effect of motivation on pupil features. While a correlation analysis (Table 4.2) suggested a positive correlation between the motivation (both rated at the beginning as well as the end of the task) and pupil dilation, the model revealed that motivation rated at the end of the task mainly interfered with the PPD and MPD. Existing concepts, including FUEL, highlight the role of motivation in explaining effort allocation by applying a general theory of motivation: Motivation Intensity Theory (MIT,Brehm and Self, 1989; Pichora-Fuller et al., 2016). MIT assumes that the importance of success on a task moderates people's decision to invest effort. Previous studies reported that motivation could be further moderated by task difficulty, preference for the task and perceptual performance while resolving the task (Pichora-Fuller et al., 2016; Pittman et al., 1982; Reeve, 1989); see also

Carolan et al., 2022 for a recent review on the impact of motivational factors on listening effort). For example, Pittman et al., 1982 showed that a preference for tasks that are difficult and challenging (but not impossible) would involve high motivation. Moreover, Reeve, 1989 reported that perceptual performance (i.e., the self-perception of individual performance) influences enjoyment, which contributes to intrinsic motivation by maintaining the willingness to continue the task. More recent studies examined how motivation manipulated by monetary or social reward affects the pupil response in a speech recognition paradigm and reported a higher dilation of the pupil with increased motivation (Koelewijn et al., 2018, 2021; Pielage et al., 2021). While the aforementioned studies identified that extrinsic motivation (external manipulation) modulates the pupil response, the results from the present study clearly indicated that more intrinsic motivation rated by the listener affects the pupil response as well. Among all listener factors, motivation rated toward the end of the test seems to have the strongest impact on the MPD, as indicated by the high effect size. It is speculated that after becoming knowledgeable regarding the task, listeners' perception of their success importance in the experiment and their expectations might change; therefore, their effort allocation might decrease over time. It could, furthermore, be speculated that the motivation ratings towards the end of the task could interact with the fatigue level of the listener.

Mental fatigue is known to increase with time performing a task which then impacts effort allocation (Pattyn et al., 2018). Recent research demonstrated an interactive effect of fatigue and motivation, such that in a situation of increased mental fatigue, effort investment depends on the individual's motivation to perform the task (Müller and Apps, 2019). In fact, the model results obtained in the current study suggest that daily-life fatigue is a major contributor to the pupil response, as indicated by the largest effect size. A high need for recovery was negatively associated with the PPD, which is in line with recent literature (Wang et al., 2018b). Generally, previous investigations suggested that effort mobilization for listening can be influenced by fatigue (see Bafna and Hansen, 2021 for a review). Is it speculated that the correlation of daily-life fatigue with PPD is due to increased parasympathetic nervous system activity, which might have led to a reduced pupil response (Steinhauer et al., 2004; Wang et al., 2018b). Interestingly, the results from the present study showed a strong effect of daily-life fatigue only on PPD, whereas no impact of daily-life fatigue on the MPD was found. Potential differences in the effect of listener factors on the PPD and

MPD will be discussed at a later point in this discussion.

### 4.4.2 MPD variation across days is driven by cognitive abilities

Results from the mixed-effects model indicated that variations in the pupil features across visits were mainly affected by the listeners' cognitive abilities. More specifically, for people with comparable low WMC, MPDs decreased from one visit to another, indicating lower effort deployment. At the same time, a reversed pattern was observed for people with a comparable high WMC, i.e., an increase in MPD with an increasing number of visits (see Figure 4.4). According to FUEL, the cognitive capacity available for an individual plays a central role in effort mobilization in a way that lower available resourses might lead to decreased effort mobilization. These results might be rooted in previous literature, which assumes that listeners conduct a cost-benefit analysis, evaluating the benefit of the effort expended and the cost of the cognitive capacity allocation (Eckert et al., 2016). Hence, it is speculated that the cost-benefit analysis might change after the first visits and participants with low cognitive abilities could have changed their expectations and success evaluation across visits.

Interestingly, the current study did not find an overall effect of WMC neither on PPD nor on MPD. This is in contrast to previous studies which examined the role of WMC and its relation to the pupil response in a speech recognition task (Dingemanse and Goedegebure, 2021; Wendt et al., 2016, 2017; Zekveld et al., 2011). Some studies (Wendt et al., 2016; Zekveld et al., 2011) reported a positive correlation between WMC and pupil response in normal-hearing listeners. In contrast, other studies found that for people with HI the pupil response in a speech recognition task was negatively correlated with WMC (Dingemanse and Goedegebure, 2021; Wendt et al., 2017).

### 4.4.3 Impact of listener factors on the dynamic range

The results from the present study showed that the dynamic range of the pupil response was mainly affected by three factors: age, motivation and fatigue. Age impacted the dynamic range of the pupil, both within a speech-in-noise task as well as at rest, such that older participants tended to exhibit a smaller dynamic range. This is consistent with previous investigations showing a smaller dynamic range for older individuals (Ayasse and Wingfield, 2020; Bitsios et al., 1996). This impact of age on the dynamic range should be considered when extracting

different pupil features by applying normalization procedures (Piquado et al., 2010; Winn et al., 2018). However, it should be noted that this study had an unbalanced sample size for the different age groups (with only four listeners in the older age group), such that the findings related to age need to be considered with caution. Nevertheless, the results were consistent with previous findings addressing age as an important contributor to pupil sizes assessed within a speech-in-noise paradigm.

The strongest impact on the dynamic range was found for the motivation rated towards the end of the task. This effect was most pronounced for the cognitive task (DR_Cog) but also significant for the dynamic range measured at rest (DR_DL). Since the cognitive task (arithmetic mental task) was reported to be rather difficult to accomplish by the participants, it is speculated that the high task demands could have impacted the participants' motivation and their cost-benefit evaluation of the task, which, in turn, might have affected their effort investment in the task (according to MIT, Brehm and Self, 1989. In other words, if the task was too difficult to accomplish, the success importance further affected their effort investment.

Finally, WMC was found to affect the dynamic range. However, this was only the case in the cognitive task and not in the speech-in-noise task. This might be explained by the fact that a cognitive task, such as the mental arithmetic task, might require similar cognitive processes as those involved in the Digit span test. However, an investigation of other cognitive aspects, such as unconscious memory (i.e., priming) and its impact on the dynamic range, would be valuable to explore.

### 4.4.4 MPD vs. PPD contributors

The current study focused on two different pupil features, namely the PPD and MPD, as both have been common measures to study listening effort within a speech-in-noise paradigm (e.g., Winn et al. 2018). The findings from the present study showed that both pupil features were affected by different listener factors. While the variations in PPD values were mainly driven by daily-life fatigue and age, MPD was found to be most affected by motivation and cognitive abilities (as an interaction effect with visits). The findings regarding the PPD are in line with previous literature indicating smaller PPDs for elderly people (Bitsios et al., 1996) and for people that reported higher levels of daily-life fatigue (Wang et al., 2018b). No significant interactions of listener factors across visits were

found, suggesting that the PPD might be a more stable feature when repeated measurements are applied. This is in accordance with the results from *Chapter 3* of this thesis, exploring the reliability of these features.

It has been speculated that both features might reflect effort allocation related to different aspects of the task (Wagner et al., 2019). For most of the participants, the PPD occurred approximately 1-2 seconds after the sentence offset and was thought to encapsulate effort allocation for listening to the target speech (Winn et al., 2015). In contrast, the MPD is assessed within a longer time window, including the retention interval, which is the time between the sentence offset and the participant's response. Hence, it has been argued that the MPD does not only reflect aspects of listening but further incorporates cognitive aspects of processing the speech, such as linguistic processing, or even preparation for the response (Wagner et al., 2019; Winn et al., 2015). Thus, the impact of WMC observed only for the MPDs might be explained by the fact that the MPD reflects those additional cognitive aspects of the listening task. It was out of the scope for this study to further explore how the different pupil features would reflect different cognitive processes involved in such a task. However, since these two features (PPD and MPD) have typically been used as listening effort indicators, it is important to distinguish between them since they are driven by distinctive individual listener factors.

### 4.4.5 Future directions

Based on the findings of the current study, further research might further explore the complex relationship between motivation, fatigue and cognitive capacity and its impact on these different pupil features. Including self-reports of success importance or perceptual preference for the task in the experimental procedure could be valuable to analyze the individual decision-making process of effort expenditure. Since the data from the current study were mainly obtained from young listeners (only four old listeners were included in the analysis), it would be valuable to further explore the impact of age by including a more balanced age group. A better understanding of the magnitude of the contribution of hearing loss to the pupil response would bring research closer to developing rehabilitation techniques for people with hearing impairment.

### 4.4.6 Conclusion

Three main observations were made in the present study. First, motivation and daily-life fatigue were found to be the main contributors to modulating the pupil response in a speech-in-noise task. Second, the contribution of different listener factors (such as fatigue and motivation) was highly dependent on the pupil features. While fatigue seemed to be the dominant factor explaining variability for the PPD, MPD was mainly driven by the motivation rated towards the end of the task. Third, the dynamic range assessed within a speech-in-noise task was only affected by the age of the listeners. However, the impact of the listener factors on the dynamic range depended highly on the task. The findings showed that motivation, as well as cognitive abilities, can further impact the dynamic range of the pupil within a cognitive task or at rest.

Overall, these findings suggest that both motivation and fatigue provide major contributions to the individual variability of the pupil response, whereby their relative contributions change depending on the pupil feature (MPD vs. PPD). Furthermore, differences in the dynamic range based on the age level may be accounted for by applying an appropriate normalization procedure. These results may contribute to a better understanding of the impact of different listener factors on the variation of the individual pupil response, which is essential when developing pupillometry towards a more clinically feasible and relevant measure.

# 5

# Exploring the relationship between perceptual effort investment and the evoked pupil response during speech perception [d]


## Abstract

Pupillometry is commonly used as an objective measure of listening effort. However, the sensitivity of the task-evoked pupil response and its relation to perceived listening effort in speech-in-noise perception conditions are not yet fully understood. In the present study, the just noticeable difference (JND) in clarity, the JND in perceived listening effort and the just meaningful difference (JMD; McShefferty et al., 2016), here referred to as the JND in meaning, were explored when varying the speech signal-to-noise ratio (SNR). To estimate the JND in clarity, participants were asked to identify which sentence in a given pair of sentences was perceived as clearer. To estimate the JND in effort, participants were asked to indicate if they noticed a difference in the effort allocation when listening to the two sentences. To obtain the JND in meaning, the participants were asked if they would change to new headphones that would provide an improvement in the clarity they perceived between the two sentences. The participants listened to blocks of paired sentences presented at two different SNRs: a reference SNR at 0 dB and a target SNR ranging from 0.5 to 8 dB. A psychometric function was fitted to each listener's perceptual ratings, and the respective JNDs were extracted. At a listener group level, the results showed an average JND in effort at a value between those obtained for the JND in clarity


[d] This chapter is based on Neagu et al., (2022c), in prep.

and JND in meaning. However, substantial variability was observed across the individual listeners. The relationship between the different JNDs and corresponding changes in the pupil responses were investigated. Regarding the evoked pupil responses, no difference was found between the reference and the target SNR conditions in terms of the JND in clarity, whereas significant changes were observed for the JNDs in effort and meaning, with smaller responses in the target conditions than in the reference conditions. Overall, the results suggest that differences in the pupil response become first noticeable when people perceive a difference in effort, on average corresponding to a 4 dB SNR change, whereas the differences in the pupil response do not capture the JND in clarity. The results may contribute to a better understanding of the relationship between perceptual listening effort and the evoked-pupil response and might be relevant for potential applications of pupillometry as a clinical tool.

## 5.1 Introduction

Speech communication is an essential ground for human interaction. Understanding speech in the presence of background noise can, however, be challenging, especially for people with hearing impairment (HI) who sometimes report tendencies to withdraw from social interactions (Edwards, 2007; Hornsby, 2013; Kramer et al., 2006; Ogawa et al., 2019) due to increased levels of mental distress and fatigue (Kramer et al., 1997, 2006; Stephens and Hétu, 1991). Studying listening effort has in recent years received increasing attention in the audiological research field. Different measures of listening effort have been considered, including physiological paradigms assessing changes in the autonomic nervous system's activity as well as behavioural measures or self-reports and subjective ratings to assess listeners' self-perceived effort. However, Alhanbali et al., 2019 indicated that different measures of listening effort, such as the mean and peak pupil size, reaction time, skin conductance, alpha power (during speech processing and retention), as well as perceived effort might represent largely independent aspects of listening effort. Alhanbali et al., 2019 suggested careful consideration of these different measures when choosing a given paradigm to investigate listening effort. Pupillometry, representing one of the physiological

measures, has been used extensively as an objective measure of listening effort during speech-in-noise paradigms (Wendt et al., 2018; Winn, 2016; Zekveld and Kramer, 2014). For example, several studies explored the impact of the signal-to-noise ratio (SNR) and the level of speech intelligibility on the pupil response as a measure of listening effort (Krueger et al., 2017; Wendt et al., 2018; Zekveld et al., 2010). Other studies (Alhanbali et al., 2019; Koelewijn et al., 2012a; Zénon et al., 2014) investigated the correlation between the pupil response and subjective ratings of listening effort, using NASA (Hart and Staveland, 1988) or other self-reported measures. Even though some of the research investigated both physiological as well as perceived measures of effort, the relationship between such measures, particularly with respect to pupillometry, has remained unclear. Understanding the link between pupil response and perceived effort seems crucial for evaluating the potential of pupillometry as a clinical tool. This present study attempted to contribute to such understanding.

A common intervention for people with hearing loss is to provide hearing aids (HA). Some devices can improve the SNRs by applying noise reduction (NR) schemes which, in turn, may improve speech intelligibility and decrease listening effort. There has also been increasing evidence that HA signal processing affects listening effort and might provide some benefits for the user. Ohlenforst et al., 2017a provided a systematic review exploring the effects of hearing impairment and HA amplification on listening effort. Some studies reported a reduced listening effort due to active NR schemes (Fiedler et al., 2021; Ohlenforst et al., 2017a; Wendt et al., 2017). However, changes in listening effort due to changes in the SNR have mainly been studied on a listener group level, while much less has been reported about how changes in listening effort are represented in the individual listeners and how robust such measures are. Furthermore, an improvement in SNR provided through a hearing device might not always lead to a change in a listener's perception of listening effort (McClymont et al., 1991; Saunders and Forsline, 2006). It is, therefore, important to explore how a change in SNR that leads to a change in perceived effort is related to a given person's just noticeable difference in SNR and how this varies across people.

The current study considered the concept of a just noticeable difference in effort (JND in effort), reflecting the minimum increase in SNR necessary for a person to perceive a difference in effort. In addition, the minimum difference in SNR that caused a just noticeable difference in the clarity of a sentence presented

in noise was measured and considered to represent the 'limit of resolution' with respect to a listener's sensitivity to SNR changes. Finally, the just meaningful difference (JMD) in SNR was measured as proposed in McShefferty et al., 2016, where the participants were asked to rate the minimum increase in SNR at which they would seek an intervention (e.g. a replacement of their hearing device). McShefferty et al., 2016 showed that the average JMD in SNR was at 6-8 dB and, thus, clearly above the JND in SNR, corresponding to values at about 3 dB. Therefore, the present study investigated three different 'thresholds' corresponding to three different JNDs obtained at different values along the same dimension, the SNR: (i) the JND in SNR, referred here to as the 'JND in clarity' since listeners were asked to judge the just noticeable difference of the perceived clarity of the speech associated with a small change in SNR; (ii) the JND in effort where listeners report a difference in perceptual effort and (iii) the JMD, referred here to as the 'JND in meaning' where listeners report to seek intervention. It was anticipated that the JND in effort would lie between the JND in clarity and the JND in meaning.

An important aspect of the study was to examine how the perceived JNDs would correspond to changes in the pupil response assessed during listening. Pupil responses were recorded while the participants listened to the presented sentences. Since previous literature did not find significant changes in the peak pupil dilation for SNR changes of 2-3 dB (Giuliani et al., 2020; Wendt et al., 2018), it was hypothesized that no changes in the evoked pupil response might occur at the JND in clarity which is expected to lie around 3 dB SNR according to McShefferty et al., 2016. Regarding the JND in effort, significant changes in the pupil response would be expected to occur based on Wendt et al., 2018, who found differences in the pupil peak dilation (PPD) for changes in SNR of 4 dB and above. Hence, when people report differences in their perceived listening effort, i.e. at the JND in effort, this might be reflected by a significant change in the corresponding evoked pupil response. Furthermore, based on the results obtained in McShefferty et al., 2016 and assuming that individuals consider a form of intervention only when the effort provokes a discomfort, an increased difference in the evoked pupil response might be expected at the JND in meaning.

Another important aspect of this study was the analysis of results obtained in individual listeners, whereas previous approaches reported JNDs and 'meaningful differences' as averaged results across individuals (Killion and others,

2004; McShefferty et al., 2015). The analysis of the variability of results across individuals should help evaluate the robustness of pupillometry as a potential correlate of listening effort.

## 5.2 Methods

### 5.2.1 Participants

Twenty-nine native Danish speaking participants (between 18 and 40 years of age, with a mean of 25 years) took part in this study. They had pure-tone hearing thresholds of 20 dB hearing level (HL) or better at frequencies below 6 kHz and 30 dB HL or better at frequencies above 6 kHz. Exclusion criteria included no history of eye diseases or eye operations. Additionally, participants were asked to avoid any caffeine intake at least 3 hours before the test time. All participants provided informed consent for the tests included in the study and received monetary compensation for their participation. The research procedures were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

### 5.2.2 Measurement setup and stimuli

The participants sat in a sound-isolated booth on a chair fixed in place in front of a desk, which had a computer screen and keyboard placed on top of it, to perform the tasks. A graphical interface was implemented in Matlab (MATLAB, 2018) running on a computer outside of the booth, which then synchronized with the screen inside the booth using Psychtoolbox 3 (Brainard and Vision, 1997; Kleiner et al., 2007; Pelli and Vision, 1997). This setup facilitated the coordination between the presentation of the stimuli and the collection of the responses to the questions that participants submitted via the keyboard. The verbal responses from the participants for the speech recognition task were sent through a Shure WH20 microphone to an experimenter sitting outside of the booth actively scoring the responses. Stimuli were presented through HD650 headphones using an SPL Audio Phonitor Mini amplifier. The experimenter could communicate with the participants through a talk-back t.bone GM5212 microphone during breaks in testing. A Fireface 802 sound card was used to connect to the amplifier, the headphones and the microphones.

Eye-tracking data were recorded during the two listening paradigms using

a desktop mounted eye-tracker (EyeLink 1000; SR-Research Ltd., Mississauga, Ontario, Canada). The eye-tracking camera was placed on the desk in front of the screen below the participant's line of view. The distance from the eyes of the participants to the camera was between 50 cm and 70 cm, depending on the individual and the individual's exact position in the chair. During the listening tasks, the participants were instructed to fix their gaze at a grey cross in the middle of a black screen. Pupil sizes were then recorded from the left eye with a sampling frequency of 500 Hz. The luminance of the screen and the ambient light were controlled to prevent any changes in pupil response that could be influenced by changes in luminosity. The ambient light was measured at 75 lx for the tasks. The approximate brightness of the black screen with the grey cross in the middle was 9 cd/m2.

Sentences from the Danish Dagmar, Asta, or Tine (DAT) corpus (Nielsen et al., 2014) were used in this study. Each sentence consists of five words with a fixed structure: "<Name> thought about <keyword> and <keyword> yesterday". These sentences were presented in a 4-talker babble masker, which was created by superimposing two male and two female talkers in an interval of 1 minute. The masker was chosen as a random sequence of about 7.5 seconds for each sentence. The sound pressure level (SPL) of the sentences was 63 dB ±1 dB, including a roving element similar to (McShefferty et al., 2016).

A paired-sentence paradigm was applied in which two sentences were presented with the masker onset starting 3 seconds prior to each sentence onset and ending 3 seconds after each sentence offset. Each pair of sentences consisted of a reference sentence presented at 0 dB SNR and a target sentence presented at a higher SNR ranging between 0.5 and 8 dB. The target and the reference SNR were always presented in a pseudo-randomized order such that half of the trials in each block contained the reference first and vice versa.

### 5.2.3 Perceptual measures

Figure 5.1 shows a schematic representation of the overall paired-sentence paradigm. The participants were instructed to listen to each sentence and repeat the keywords after the noise offset. The responses were scored on a word-level basis. Measurements were conducted in two experimental series. In the first series, the JND in clarity was measured, and in the second series, the JND in effort, as well as the JND in meaning, were obtained.

The clarity-JND measurement consisted of five blocks of 12 paired sentences.

Five target SNRs were considered: 0.5 dB, 1 dB, 2 dB, 3 dB and 4 dB SNR, while the reference SNR was at 0 dB. Twelve trials for each target SNR were presented in a pseudo-randomly manner across the blocks such that the number of paired sentences containing each SNR level was balanced within a block. After each trial, the participants had to answer a question visually presented on a screen: "Which sentence was clearer?". The participants could then indicate whether they thought the first or the second sentence was clearer by pressing the left or right key on a keyboard. Each participant was given a score for each target SNR level that quantified the percentage of trials wherein they correctly identified the target sentence.

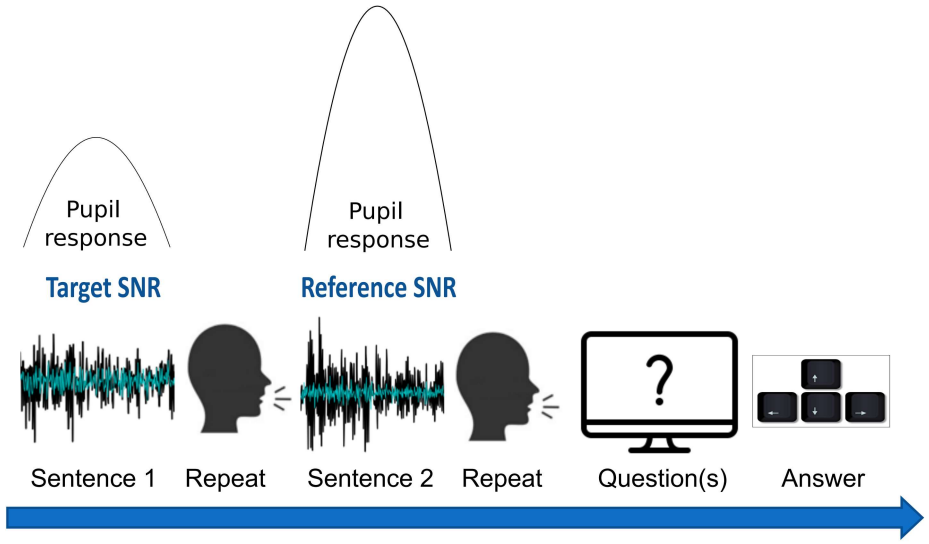



Figure 5.1: Schematic representation of the paired-sentence paradigm illustrated for a single trial. The participants listened to a sentence (sentence 1) and, after a short retention interval of 2 seconds, were asked to repeat back as many keywords as possible. This procedure was then repeated for another sentence (sentence 2). Both sentences were presented at different SNRs, i.e., at the reference SNR (0 dB) and a target SNR (above the reference SNR). Target and reference SNRs were presented in random order. After repeating the keywords, a question was displayed on a screen where participants were asked to rate which sentence (1 vs. 2) was clearer to estimate the JND in clarity. In the effort and meaning JND tasks, participants were asked if they perceived changes in effort (yes/no) and if they would opt to change the device (yes/no). The pupil response was measured throughout the whole trial to assess the evoked pupil response during listening to the sentences.

The effort-JND and meaning-JND measurements also consisted of five blocks of 12 trials each. Five target SNRs were considered, which were designed to span a larger range than in the clarity-JND task: 0.5 dB, 2 dB, 4 dB, 6 dB and 8 dB SNR, while the reference SNR was constant at 0 dB. The target SNR was

pseudo-randomly presented across the blocks, such that the number of paired sentences containing each SNR level was balanced within a block. At the end of each trial, i. e., after each pair of sentences and the associated speech recognition task responses, the participants had to answer two questions visually presented on the screen. The first question was: "Did you perceive any difference in the listening effort you allocated for these sentences?“. The participants could then answer *yes* or *no* by pressing the left or right key on a keyboard, respectively. Thereafter, the participants were asked the second question: "If you would experience this improvement in the clarity you perceived between the two sentences, would you change your current headphones?". The participants could then answer *yes* or *no* by pressing the left or right key on the keyboard. The duration of each trial, including the pair of sentences, the question and the response, varied depending on the duration of the presented sentence and on the time the participants needed to perform the speech recognition task and respond to the questions. Each participant was given a score for each SNR level and each question, and the score quantified the percentage of trials wherein they answered *yes* to the question.

To estimate the individual listener's JND in clarity, JND in effort and JND in meaning, psychometric functions were fitted to each of the participants' scores as a function of SNR using logistic regression. The SNR at which the fitted psychometric functions reached the 50% threshold was defined as the corresponding JND in the respective tasks.

### 5.2.4 Pupil data analysis

The pupil data were processed using MATLAB (MATLAB, 2018) and R (R Core Team, 2019). All pupil traces within a block were included in the analysis. The post-processing of the raw pupil data included several steps. First, blink removal was performed, where pupil dilation values more than three standard deviations smaller than the mean were considered as eye-blinks. Afterwards, a linear interpolation, starting about 80 ms before and ending 150 ms after each blink, was applied. Thereafter, the data were smoothed using a moving average filter with a rectangular window of 117 ms. Trials containing more than 20% missing data, eye blinks or artefacts were removed from the analysis. All remaining traces were cut into individual sentences and then baseline corrected by subtracting the mean pupil size measured 1 s prior to the sentence onset. Additionally, pupil responses were explored both at the reference SNR and the target SNRs.

Furthermore, the pupil responses were extracted at the SNRs associated with the three JNDs in clarity, effort and meaning for the individual participants. The peak pupil dilation (PPD) and mean pupil dilation (MPD) were extracted from the averaged pupil trace of each individual for each SNR (target and reference SNR) within the paired-sentence paradigm. PPD and MPD were extracted in the time window between stimulus onset and noise offset.

## 5.3 Results

### 5.3.1 Perceptual JNDs in clarity, effort and meaning

Figure 5.2 shows the perceptual results obtained in the present study. The upper left panel represents the mean data, averaged across all listeners. The other panels show the data for the individual listeners. The blue data points indicate the data obtained in the clarity-JND task. The percentages of correct responses (left ordinate) are represented as a function of Delta SNR. The blue solid curves represent the corresponding psychometric functions fitted to the data. The green data points and functions show the results obtained in the effort-JND task. Here, the percentages of *yes* responses (right ordinate) are indicated as a function of Delta SNR. The red data points and solid functions represent the corresponding percentages of *yes* responses obtained in the meaning-JND task.

Regarding the mean data (upper left panel), a clarity-JND of 2 dB, an effort-JND of 4 dB and a meaning-JND of 8 dB were found. The values for the clarity- and meaning-JNDs are consistent with those reported in McShefferty et al., 2016. As anticipated, at a listener group level, the effort-JND lies in between the clarity-JND and the meaning-JND, such that the magnitude of Delta SNR that is required to produce a noticeable change in the perceived effort is larger than that required for a clarity-JND and smaller than that evoking the perception of a meaningful difference.

Regarding the individual results, some of the listeners showed very close results to the average patterns (e.g., TP7, TP28). However, there were also listeners who largely deviated from the average behaviour. For example, for listener TP6, the effort-JND matched the clarity-JND, and for listeners TP4 and TP8, the effort-JND matched the meaning-JND. Furthermore, for some individuals, it was not possible to obtain a clarity-JNDs (e.g., TP2, TP15) or a meaning-JND (e.g., TP2, TP7, TP11, TP20) within the analysed SNR range, i.e., the listeners

either did not reach 50% correct responses or reached the thresholds outside the range of the tested SNRs. Overall, clarity-JNDs could be estimated in 24 listeners, and effort-JNDs could also be estimated in 24 listeners, even though not in the same ones (see Table 5.1). Meaning-JND could only be estimated in 19 listeners. In total, for 18 of the listeners, all three JNDs could be derived. TP12 was excluded from further analysis because of inconsistent behavioural results, indicating that this listener may have misunderstood the task.

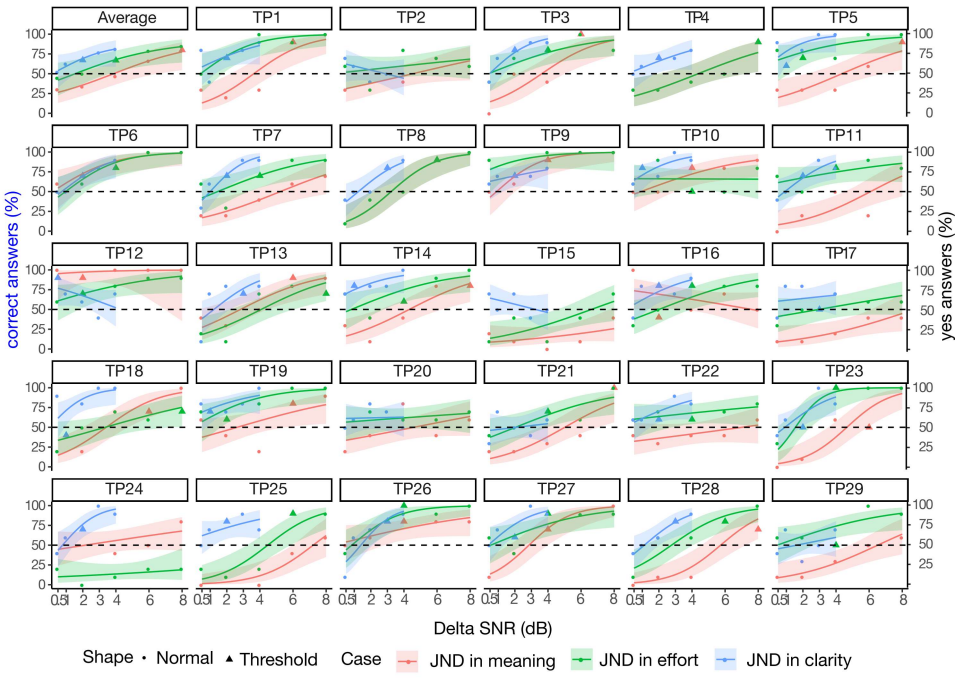



Figure 5.2: The proportion of correct responses (left ordinate) as a function of Delta SNR in the clarity JND task (blue symbols). Percentage of YES responses (right ordinate) as a function of Delta SNR in the effort (green symbols) and meaning (red symbols) JND tasks. Results averaged across listeners are shown in the upper left panel. Results obtained for the individual listeners (TP1-TP29) are shown in the other panels. Psychometric functions fitted to the data (dots) are shown by the corresponding colored solid functions. The respective JNDs were estimated as the first tested SNR where the confidence interval of the psychometric function exceeded 50% (dotted line).

| *Participant* | JND in SNR | JND in Effort | JMD in SNR |
|---|---|---|---|
| **TP1** | **2** | **2** | **6** |
| TP2 | NA | NA | NA |
| **TP3** | **2** | **4** | **6** |
| **TP4** | **2** | **8** | **8** |
| **TP5** | **1** | **2** | **8** |
| **TP6** | **2** | **4** | **2** |
| TP7 | 2 | 4 | NA |
| **TP8** | **3** | **6** | **6** |
| **TP9** | **2** | **2** | **4** |
| **TP10** | **1** | **4** | **4** |
| TP11 | 2 | 4 | NA |
| **TP12** | **0.5** | **2** | **2** |
| **TP13** | **3** | **8** | **6** |
| **TP14** | **1** | **4** | **8** |
| TP15 | NA | NA | NA |
| TP16 | 2 | 4 | 2 |
| TP17 | 3 | NA | NA |
| **TP18** | **1** | **8** | **6** |
| **TP19** | **1** | **2** | **6** |
| TP20 | NA | NA | NA |
| TP21 | NA | 4 | 8 |
| TP22 | 2 | 4 | NA |
| **TP23** | **2** | **4** | **6** |
| TP24 | 2 | NA | NA |
| TP25 | 2 | 6 | NA |
| **TP26** | **3** | **4** | **4** |
| **TP27** | **2** | **4** | **4** |
| **TP28** | **3** | **6** | **8** |
| TP29 | NA | 4 | NA |

Table 5.1: Estimated JND in clarity, effort and meaning obtained in all 29 individual listeners. If the threshold was not reached in the 0.5-8 dB interval, the value was marked as NA. The observations highlighted in bold represent the listeners where all three JND thresholds could be obtained.

### 5.3.2 Pupil responses

***Effects of presentation order and SNR on peak pupil dilation and mean pupil dilation.***

To assess a potential order effect (i.e., whether the reference SNR was presented in the first or the second sentence), pupil responses were obtained and averaged for all conditions where the first sentence (sentence 1) was presented at the reference SNR as compared to when the second sentence (sentence 2) was presented at the reference SNR. Figure 5.3 shows the results for the reference sentences, where the SNR was the same (0 dB) across all trials. Trials from the clarity-JND task are indicated in the left panel, and trials from the effort-

and meaning-JND tasks are shown in the right panel. Note that the order of the reference and target SNR presentations was randomized in a balanced way across trials, while the order of the tasks was not since the clarity JND experiment was always conducted before the other experiments.

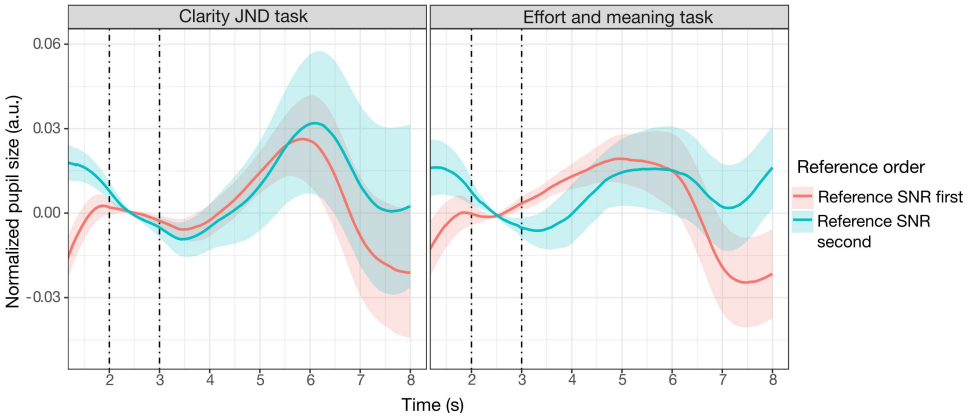



Figure 5.3: Average pupil traces were recorded while the reference SNR was presented in the clarity JND measurement (left panel) and the effort/meaning JND measurement (right panel). The red curve shows the mean pupil trace, averaged across all participants when the reference SNR condition was presented first. The red shaded areas indicate the standard errors of the mean. The blue curves showed the corresponding results when the reference SNR condition was presented second (i.e., after the target SNR condition). The noise started at time 0 and sentence onset at 3 seconds. Dashed vertical lines indicate the baseline interval, i.e., 1-second preceding sentence onset.

A visual inspection of the pupil traces shows similar pupil curves independent of when the reference SNR was presented (first vs. the second sentence) for the clarity-SNR measurement (left panel in Figure 5.3) and the effort and meaning-JND measurement (right panel in Figure 5.3). In order to investigate if the order of the reference SNR had an impact on the main pupil features, a two-way ANOVA analysis was applied to the peak pupil dilation (PPDs) and mean pupil dilations (MPDs). No effect was found on the PPDs neither for the clarity-JND measurement ($F = 1.192$, p-value $= 0.28$) nor in the effort- and meaning-JND measurements ($F = 0.119$, p-value $= 0.731$). Similarly, a two-way ANOVA analysis showed no effect of the reference SNR order on the MPDs (clarity JND: $F = 0.062$, p-value $= 0.805$; effort- and meaning-JND: $F = 0.051$, p-value $= 0.822$). Nevertheless, comparisons across pupil responses of different measurements, i.e., to assess the clarity JND vs. effort- and meaning JND, should be done with caution.

Figure 5.4 shows the pupil responses obtained at the target SNR ranging

from 0.5-8 dB. The left panel shows the results obtained during the clarity-JND measurements, and the right panel shows the results obtained during the effort-/meaning-JND measurements. The different colors indicate the different target SNR conditions. Regarding the results in the clarity-JND conditions (left), a two-way ANOVA revealed that there was no effect of the SNR on the PPD extracted from the pupil traces ($F = 1.4$, p-value $= 0.237$). Similarly, in the effort- and meaning-JND conditions (right), a two-way ANOVA analysis showed no effect of SNR on the PPDs ($F = 0.497$, p-value $= 0.783$).

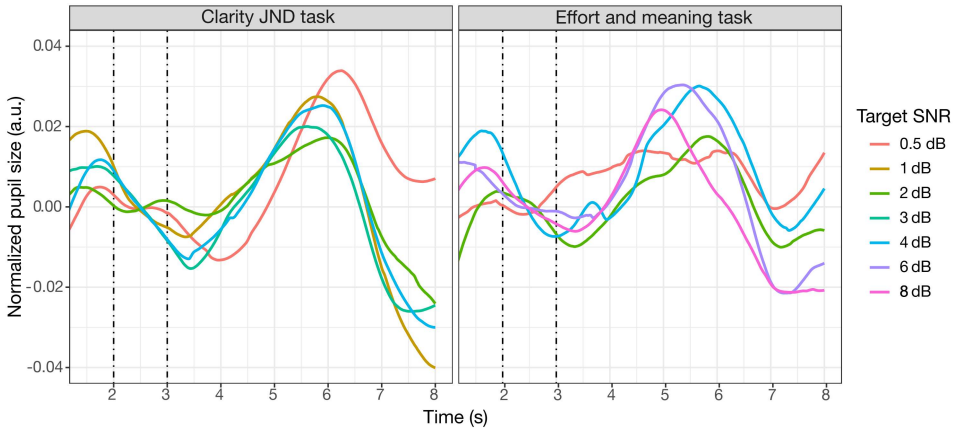



Figure 5.4: Mean pupil traces recorded at different target SNRs, averaged across all participants and trials. Different colors indicate the different target SNRs considered in the behavioral tasks. Left: Results obtained during the clarity-JND task. Right: Results obtained during the effort- and meaning-JND tasks. The noise started at time 0 and sentence onset at 3 seconds. Dashed vertical lines indicate the baseline interval, i.e., 1-second preceding sentence onset.

***Pupil responses at perceptual JNDs in clarity, effort and meaning.***

Although no systematic effect of the SNR on the PPD was found, clear pupil responses were generated in the different JND conditions. The paired-sentence paradigm might enable the extraction of evoked pupil responses characterizing different behavioral sensitivities, as reflected by the considered JNDs in the present study.

The pupil traces were extracted for 17 participants at the SNRs corresponding to the JNDs in clarity, effort and meaning (as listed in Table 5.1). Figure 5.5 depicts the evoked pupil response for five different conditions: at the smallest applied Delta SNR between the reference SNR and the target SNR (0.5 dB; panel a); at the individual's JND in clarity (panel b); at the individual's JND in effort

(panel c); at the individual's JND in meaning (panel d) and at the largest applied Delta JND (8 dB). The red functions indicate the results for the different target SNRs, and the blue functions represent the responses at the reference SNR (0 dB). Unexpectedly, the pupil response at the reference SNR (0 dB) was found to be smaller in the effort- and meaning-JND conditions than in the clarity-JND condition. However, a much larger *difference* between the pupil responses at the target and the reference SNRs was observed at the JND in effort (panel c) and the JND in meaning (panel d) as compared to the difference between the pupil responses at the JND in clarity (panel b) and at the target SNR of 0.5 dB (panel a).

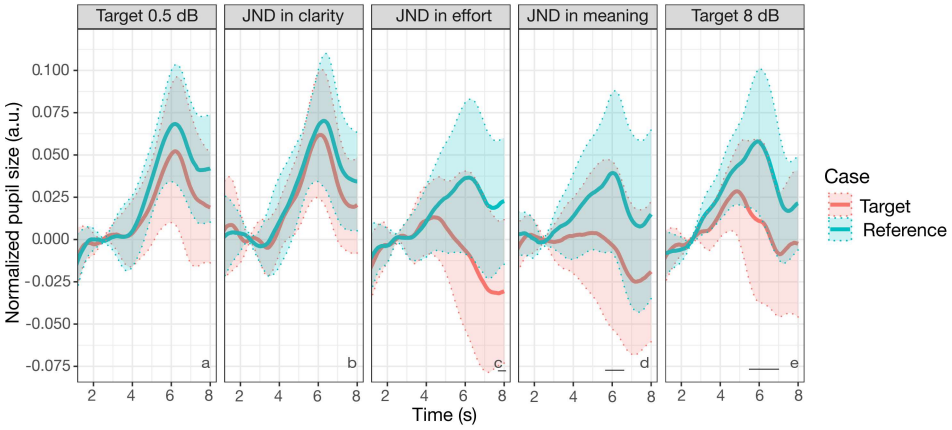



Figure 5.5: Panel a: Pupil responses obtained at the smallest applied change in SNR between the target and reference SNR (reference at 0 dB SNR; target at 0.5 dB SNR). Corresponding pupil response extracted at the individual's JND in clarity (panel b), effort (panel c) and meaning (panel d). Panel e shows the results at the largest applied change in SNR (target at 8 dB SNR). The red curves represent the responses during the target SNR presentations, and the blue functions show the responses for the reference SNR presentations, both averaged across those 17 participants for whom all three perceptual JNDs could be estimated (see Table 5.1). Shaded areas represent the 95% confidence intervals. The black horizontal lines represent the intervals of time calculated from the 3rd second onwards, where the pupil response at target and reference SNRs was significant different.

Multiple-comparisons paired t-tests were performed to evaluate whether there were any significant differences between the pupil traces for the target and the reference SNRs conditions in the interval between 3 and 8 seconds of the paired-sentence paradigm. This analysis accounted for individual pupil responses. The time series was divided into time bins of 100 ms each, and multiple paired t-tests were performed between the data samples contained within each time bin (50 observations each). In total, 17 comparisons were performed for each time bin. Bonferroni correction was applied to compensate

for the Type I error generated by the multiple comparisons. In the condition with a target SNR of 0.5 dB, no significant differences were found between target and reference (p-value > 0.5). Similarly, no significant differences were found between target and reference at the JND in clarity. For the JND in effort, significant differences were found between the target and the reference during the time interval of 7.7 to 8 seconds (p-value< 0.5). At the JND in meaning, the pupil responses differed significantly between target and reference in the interval of 5.7 to 6.6 seconds (p-value < 0.05). Finally, in the condition with a target SNR of 8 dB, significant changes between the target and the reference conditions were found in the interval between 5.4 to 7 seconds (p-value < 0.05).

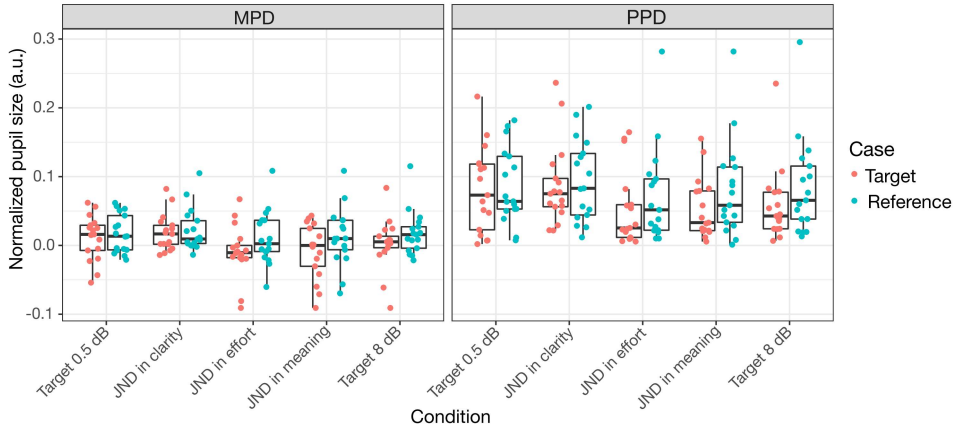



Figure 5.6: MPDs (left panel) and PPDs (right panel) obtained in the references SNR (blue dots) conditions and in the target SNR (red dots) conditions at five different 'operation points': the smallest applied SNR (0.5 dB), the listeners' individual JND in clarity, the listeners' individual JND in effort, the listeners' individual JND in meaning, and the largest applied Delta SNR (8 dB). The values are shown in different colors representing the MPD/PPD extracted at the target SNR (red dots) or the reference SNR (blue dots).

Furthermore, the pupil features, mean pupil dilation (MPD) and peak pupil dilation (PPD), were extracted from the average pupil traces. Figure 5.6 shows boxplots of the MPD (left panel) and the PPD (right panel) of all participants for the reference SNR conditions (blue dots) and the same target SNRs conditions (red dots). The differences between the reference and the target SNRs were evaluated using multiple paired t-tests comparisons. After correcting for family-wise Type 1 error by conducting Bonferroni correction, no significant differences between the target and reference SNRs were found, neither for the MPD (at the JND in clarity: t = 0.52, p-value = 0.609; at the JND in effort: t = 0.829, p-value = 0.413; at the JND in meaning: t = 1.28, p-value = 0.212) nor for the PPD (at the

JND in clarity: t = 0.66, p-value = 0.514; at the JND in effort: t = 1.41, p-value = 0.169, at the JND in meaning: t = 1.19, p-value = 0.241).

## 5.4 Discussion

This is the first study that used a paired-sentence paradigm to study the relationship between perceptual behaviour (JNDs) and an objective measure of listening effort (pupil response). Three main observations were made in the present study, as further discussed below: (i) JNDs in effort could be obtained with the paired-sentence paradigm, which amounted to be about 4 dB SNR at a listener group level; (ii) evoked pupil responses could be extracted at various JNDs which allowed exploring the relationship between the perceived speech-in-noise and the invested listening effort; and (iii) behavioral JNDs in clarity, effort and meaning could be obtained at an individual-listener level for most participants, in contrast to previous studies where JNDs were obtained at a listener-group level.

Regarding the JNDs in clarity and meaning, similar values on a listener-group level were found in the present study as in previous investigations (McShefferty et al., 2015, 2016). A mean JND in clarity of 2 dB and a mean JND in meaning of 8 dB were found, as compared to 3 and 6-8 dB, respectively, in McShefferty et al., 2016. The JND in effort (4 dB SNR) obtained in the present study was between the JND in clarity and the JND in meaning. McShefferty et al., 2016 argued that a 'JMD in SNR' (termed JND in meaning in the present study) has more clinical importance than the 'JND in SNR' (termed JND in clarity in the present study) because it indicates a change in SNR that is required to motivate people to seek intervention. Similarly, it is argued here that the JND in effort, introduced in the present study, is more indicative of people's everyday listening experiences than just detectable changes in the SNR as reflected by the JND in clarity. In particular, people with hearing impairment often report that their listening is effortful, tiring or stressful, even when speech can be recognized accurately. Sustained effort investment in speech communication can have rather severe consequences, such as social withdrawal or increasing days of sick leave from work due to mental distress (Kramer, 2009; Kramer et al., 2006). Hence, exploring the JND in effort should be relevant when studying everyday challenges in speech communication, particularly for people with hearing impairment.

Regarding the evoked pupil responses, no differences in the responses obtained at the target vs. the reference SNR conditions were found when evaluated at the JND in clarity (Figure 5.5b). This is consistent with earlier studies reporting that changes in SNR of about 3 dB did not necessarily lead to significant changes in the pupil response during a speech recognition task (Giuliani et al., 2020; Wendt et al., 2018). In contrast, when evaluated at the JNDs in effort (Figure 5.5c) and meaning (Figure 5.5d), differences between the pupil responses were found during the listening or retention interval. Interestingly, the changes in the pupil response at the JND in effort did not differ from those observed at the JND in meaning. A larger change of the pupil dilation (for the target vs. the reference SNR condition) was expected at the JND in meaning than at the JND in effort. However, at both JNDs, the pupil responses obtained in the target SNR conditions (red functions in Figure 5.5c and 5.5d) exhibited relatively low amplitudes. This suggests a rather small effort investment at those SNRs (4 dB and 6 dB), which could be represented by a ceiling effect.

While the (mean) pupil response functions indicated differences between the target and reference SNR conditions when evaluated at the effort and meaning JNDs (but not the clarity JND), the responses features MPD and PPD did not indicate any significant differences in any SNR condition, including those representing the effort and meaning JNDs (Figure 5.6). This is interesting since the MPD and PPD have been considered in various studies to characterize the pupil response (Winn et al., 2018; Zekveld et al., 2018). Although significant differences were only found for short time intervals, our results suggest that these 'static' features might not be appropriate to represent a correlate of behavioral listening effort since they do not capture dynamic characteristics in the pupil response function. Such dynamic characteristics appear to contribute to the representation of listening effort and are better reflected in an analysis that takes the whole pupil response function into consideration.

An important aspect of the present study was the analysis of the data on an individual-listener level. This is the first study that evaluated pupil response data obtained at individual listeners' JNDs. Such analysis should be of particular relevance for a better understanding of how a given perceptual 'resolution limit' (such as the three different JNDs representing different 'operating points' along the SNR axis) is reflected in the pupil response obtained during the behavioral task. Thus, such an analysis allows a correspondence between an individual listener's change in sensation (as, for example, reflected by an effort JND) and

the corresponding change in the pupil response instead of analyzing the pupil response for the same SNR or intelligibility level across listeners. The results from the present study suggest that, indeed, the pupil response may be sensitive to a change in behavioral listening effort when evaluated at the corresponding SNRs, which can differ markedly across the individual listeners (Figure 5.2). Consistent with earlier studies, the present study also showed a larger variability of the individual listener's pupil response than the same listener's behavioral data. The way toward a potential clinical application of pupillometry might thus still be far. However, the results obtained in this study might provide a valuable basis for investigating the effects of hearing loss and also rehabilitation strategies on JNDs in clarity, effort and meaning and their correlates in the respective pupil responses.

The present study had several limitations that might be addressed in future investigations. For example, the pupil response showed different behavior in the clarity-JND task compared to the effort-JND and meaning-JND tasks that were always performed after the clarity-JND task. A smaller amplitude and a less steep function were observed in the effort- and meaning-JND tasks than in the clarity-JND task. The pupil responses may have been affected by the task order, such that fatigue or motivation effects might have occurred towards the end of the session. Further investigations might explore the cause of this different pupil behavior between the different paradigms.

Another limitation was the rather high target SNRs (ranging from 0.5 to 8 dB), which might have caused ceiling effects in some conditions. The listening effort at an SNR of 4 dB might already be at a low level, such that no further reduction in effort might occur at further increased SNRs (e.g., at 6-8 dB). Wendt et al., 2018 reported that speech intelligibility at those SNRs is at the ceiling for people with normal hearing, and no major changes in the pupil response might be obtained at SNRs of 4 dB and above in such a task. The positive target SNRs have been chosen in the present study to represent ecological valid scenarios that are more relevant for everyday communication. However, it would be valuable to also apply negative SNRs to avoid potential ceiling effects in effort allocation.

Overall, the results from the present study provided some insights into the relationship between individual listening effort and its representation in the pupil response function. This work might provide a valuable basis for further investigations that explore listening effort allocation in people with hearing loss and its benefit for rehabilitation strategies for the individual.

## 5.5 Summary and conclusion

This study introduced a new concept to assess the JND in effort together with the pupil size on the individual-listener level. The JND in effort indicated the minimal change in SNR needed for the listener to perceive a difference in effort investment during sentence recognition. Mean JND in effort was around 4 dB SNR at a listener group level, while JND in effort ranges between 2 dB and 8 dB SNR at an individual-listener level. The JND in clarity (when listeners detect a change in the SNR) and JND in meaning (when listener would change their device) were measured at the same time. Listener group data suggest a JND in clarity of around 2 dB and a JND in meaning of around 8 dB, which are in line with previous literature (McShefferty et al., 2016). Pupil responses measured at the individual JNDs in clarity, effort and meaning suggested no changes in effort investment at the JND in clarity but significant changes in pupil response (between target and reference SNR) at the JND in effort/meaning, supporting the perceptual behaviour of the listener. Overall, this study provides a better understanding of the relationship between individuals' perception of effort investment and a more objective indicator of listening effort, their pupil response.

# 6

# Overall discussion

Many studies have shown the potential of pupillometry (i.e., the measure of the pupil dilation) as an objective measure of listening effort during speech perception in background noise. The pupil dilation in such listening conditions has been shown to be sensitive to the intelligibility of the speech, to changes in masker type, hearing status and hearing aid (HA) signal processing (Koelewijn et al., 2012b; Kramer et al., 1997; Kuchinsky et al., 2013; Ohlenforst et al., 2017a,b; Wendt et al., 2017; Zekveld et al., 2010, 2011). However, while most of these studies analyzed results averaged across individuals, a systematic analysis of individual pupil responses has not yet been undertaken.

People with hearing impairment often report increased listening effort during communication in their everyday life. This thesis aimed to evaluate the feasibility of pupillometry for clinical use and, hence, to understand its potential for adequately addressing the problems people with hearing impairment face in everyday communication. To move toward this goal, this thesis was set out to assess the potential of the individual pupil response as an outcome measure of listening effort by investigating the reliability and sensitivity of pupillometry within a speech-in-noise test. This includes the examination of the test-retest reliability of the pupillary response for two groups: normal-hearing (NH) vs. hearing-impaired (HI) listeners, across a broad range of pupil features such as peak pupil dilation (PPD), mean pupil dilation (MPD) and growth curve analysis (GCA) parameters (*Chapter 2*), as well as across various signal-to-noise ratios (SNRs), multiple visits and different normalization procedures (*Chapter 3*). Furthermore, this thesis investigated the listener factors impacting different pupil features and the variance of the pupil response within the speech-in-noise test across multiple visits (*Chapter 4*). Finally, a new concept was introduced, the just noticeable difference (JND) in effort, to investigate the sensitivity of the pupil response corresponding to the minimal change in the acoustic stimuli (the SNR) that causes a change in the perceived listening effort (*Chapter 5*).

## 6.1 Summary of main results

In the first study of this thesis (*Chapter 2*), the reliability of the pupil response in terms of the PPD, MPD and temporal features extracted using GCA was compared for data from Wendt et al., 2018 and Ohlenforst et al., 2017b obtained in two groups of listeners (NH vs. HI). The results indicated that two pupil features, the rise and fall of the pupil trace and the MPD provided high reliability in both listener groups. The reliability of the other considered pupil features (e.g., PPD, delay of the curve) varied across listener groups. Furthermore, a cluster analysis performed on these GCA pupil features showed that listeners tested at the same SNR were not necessarily assigned to the same cluster. The cluster analysis results suggested that SNR is not sufficient to classify pupil traces, but that there might be other factors that need to be considered for such a classification, such as data normalization procedures, as well as listener-dependent factors.

In *Chapter 3*, the investigation of the test-retest reliability of pupil features was extended to assess the impact of task demand (manipulated through variations of SNR) and different data normalization procedures on the reliability of the pupil response across multiple visits. The intraclass correlation coefficient (ICC; Cicchetti, 1994) and Spearman correlations showed that data normalization procedures have the most substantial impact on the reliability of the pupil response. Specifically, baseline-corrected data normalized to the range of the individual's pupil response across multiple visits and conditions was shown to lead to the highest reliability. In contrast to what was initially hypothesized, there was no evidence for higher reliability at lower SNRs (i.e., higher task demands). The most reliable pupil features were the traditional MPD and PPD, while the GCA features provided low reliability for all SNRs.

*Chapter 4* explored how various listener factors (namely, age, hearing status, motivation, daily-life fatigue, maximum constriction velocity and cognitive abilities) affected the standard pupil features (i.e., PPD and MPD) as well as their variation across visits. Furthermore, the individual dynamic ranges of the pupil response were measured in multiple tasks: a speech-in-noise task, a cognitive task and at rest, and the contribution of the listener factors to these dynamic ranges was analyzed. The results showed a high impact of motivation on both PPD and MPD. However, the two pupil features seemed to be driven by different factors. On the one hand, the daily-life fatigue and age had the

highest impact on PPD, with the youngest, least fatigued and most motivated people showing the largest PPD. On the other hand, MPD increased or decreased across visits based on listeners' cognitive abilities, such that the highest MPD was obtained for highly motivated people with low cognitive abilities in Visit 1, while a decrease in their MPD in Visit 2 was observed. An effect of the visit was only found for the MPD, but not for PPD. In terms of the dynamic ranges, the results showed that age affects the dynamic ranges in the speech-in-noise task and in the condition at rest, while the cognitive range was driven by cognitive abilities and motivation.

Finally, *Chapter 5* investigated the relationship between changes in the pupil response and the perceptual effort identified by the listener. A new paired-sentence paradigm was introduced to assess the JND in effort, which reflects the minimal change in SNR that is required for a listener to report a change in their effort investment. Based on the participants' ratings along the same SNR dimension, three different JND thresholds were extracted: the JND in clarity, the JND in effort and the JND in meaning. The results showed that, at the group level, the JND in effort lies between the JND in clarity and the JND in meaning, while the individual thresholds varied substantially across individuals. No significant differences in the changes in the pupil response at the JND in clarity were found. However, pupil responses differed significantly in the retention period at the JND in effort and during the listening time-window at the JND in meaning. Thus, a link between the three perceptual SNR regions and the pupil measurements was established to better understand the relationship between perceptional and physiological changes related to effort within an individual listener.

### 6.1.1 Reliable pupillary responses across visits: Optimal experimental conditions, pre-processing and analysis

Characterizing the reliability of pupillometry as a measure of listening effort is an essential outcome in hearing research. The results presented in the thesis showed that the test-retest reliability of the pupil response depends on various aspects, including test conditions, data processing and, not least, individual listener factors.

It was hypothesized that the temporal features of the pupil responses would provide more valuable information than the stationary features of the pupil, commonly used in the literature (i.e., PPD and MPD) because they capture

time-dependent aspects of the pupil traces. However, the ICC and Spearman correlation results showed that GCA features were less reliable. This thesis supports that PPD and MPD are better choices for providing higher reliability within speech-in-noise tasks (*Chapter 3*). These results are consistent with Alhanbali et al., 2019, who found a high reliability on the same features (i.e., PPD and MPD). Further investigations might explore other methods, such as generalized additive mixed modelling (Hastie and Tibshirani, 1990; Lin and Zhang, 1999; Rij et al., 2019; Wood, 2011; Wood et al., 2017), to characterize the pupil traces. The generalized additive mixed modelling method was shown to handle the variability of the pupil response through nonlinear random effects and to allow nonlinear interactions with two or more numeric predictors.

While choosing the appropriate feature is important, the data pre-processing seems to be crucial when establishing the reliability of pupil measurements. Careful consideration is required at each processing step (e.g., interpolation window, removal of noisy trials, smoothing filters) as suggested by Winn et al., 2018, but most importantly, this thesis showed that the procedure chosen to normalize the data affects the reliability of the results. Even though subtractive baseline correction is widely used in the literature (Mathôt et al., 2018), it was shown here that an additional range normalization after baseline correction increases the reliability of different pupil features. However, one should be cautious when choosing the data normalization strategy since the scope of each study should be what points towards the appropriate normalization procedure. For example, some normalization methods address interindividual differences in the variability of the dilation while others correct for average differences instead (Winn et al., 2018).

Furthermore, the fact that no significant impact of the visit on most of the pupil features was found in the regression models supports the assumption of reliability. It also showed that high reliability in the data can already be found at the second visit, when the data are normalized with baseline correction in combination with range normalization. Nevertheless, as the sample size for the third visit was significantly smaller than for the other two visits, a fair comparison of the reliability between Visits 1-2 and Visits 2-3 was not possible (*Chapter 3*). However, one could argue that when measuring a mental process such as effort, the test-retest analysis may be incorporating a correlate that reflects the objective physical manipulation of SNRs and not listening effort. Therefore, further approaches to test-retest analysis for additional validation of

the pupillometry should be performed not only at equal SNRs but also at equal performance levels and equal self-reported effort, to identify whether the pupil data are consistent in terms of which aspects are reflected beyond the acoustic properties of the stimuli.

### 6.1.2 The contribution of listener factors to changes in pupil features

Pupillometry has been previously shown to be affected by individual factors such as age, hearing status, gender, cognitive abilities, general level of daily-life fatigue or motivation (Hopstaken et al., 2015; Koch and Janse, 2016; Koelewijn et al., 2012b; Kuchinsky et al., 2016; Peelle, 2018; Wendt et al., 2017); for a review, see Zekveld et al., 2018. These factors were further investigated in this thesis.

This thesis showed that motivation plays an important role and affects the pupil response within individuals. Motivation reported at the end of the task (but not at the beginning of the task) showed the highest impact on both PPD and MPD which might have been due to the fact that the listener had no prior knowledge about the task at the start of the experiment. Furthermore, motivation measured at the end of the task might also interact with the fatigue level of the listener since fatigue is known to increase with time while the task is performed. Additionally, no interaction of motivation with visit was found in any of the pupil features indicating that motivation has no impact on the reliability of the tested pupil features. This thesis showed that daily-life fatigue is a major contributor to pupil response, being negatively associated with PPD.

In contrast to previous studies that suggested task disengagement at acoustic challenging conditions (Wendt et al., 2018), the results in this thesis did not show any task disengagement at a listener group level. However, individual pupil responses seemed to indicate disengagement (i.e., smaller pupil responses) at low SNRs for some listeners, while others seemed to be still engaged under the same conditions, i.e., at the same SNR. Hence, it is suggested to furthermore study the role of disengagement on an individual basis and how it might affect the reliability of the individual pupil response.

Both pupil features were shown to be driven by distinctive factors. PPD was mainly driven by daily-life fatigue and age, and MPD was mainly affected by motivation and the interaction of visit with cognitive abilities. This difference between the two features may be because they reflect different aspects of the task (Wagner et al., 2019). While PPD is a momentary measurement that occurs at one moment in time, shortly after the stimulus presentation, MPD is assessed

within a time window, often including the retention period and, hence, might reflect aspects related to response preparation and short-term memory. Thus, the fact that the impact of cognitive abilities was observed only for the MPD might be explained by the MPD capturing additional cognitive aspects beyond the listening task.

Moreover, listener factors were shown to impact changes in MPD across multiple visits (i.e., interaction with visit was shown for cognitive abilities and age factors) while no variation across visits was found in PPD. These results suggest that PPD might be a more reliable and stable feature, even though it still incorporates aspects that are difficult to control for, similarly to MPD. However, if task complexity manipulated through SNRs and its reflection on the pupil trace is what the experimenter is interested in, then a closer look at MPD feature is recommended, given its higher sensitivity to SNR.

The linear mixed-effects models presented in this thesis showed differences in the pupil response of individuals with distinctive cognitive abilities, represented in the interaction of cognitive abilities with visit for the MPD. More specifically, for people with low cognitive capacity, MPDs decreased from one visit to another, indicating lower effort allocation. At the same time, a reversed pattern was observed for people with a comparable high cognitive capacity. Clearly, correlation measurements for test-retest reliability (such as ICC and Spearman correlations) were not able to capture the impact of such cognitive processes on the reproducibility of pupillometry. Thus, a more complex analysis such as linear mixed-effects models can provide a better understanding of pupil responses' reliability.

Overall, these results emphasized the necessity to factor in individual cognitive capacity, motivation or fatigue when assessing listening effort in a speech-in-noise paradigm.

### 6.1.3 Pupil response dynamics at meaningful individual hearing thresholds

Even though self-reports have been intensively used to assess perceptual effort and several studies have tried to relate reported or subjectively rated effort to physiological measures of effort such as the pupil response (Alhanbali et al., 2019, 2020), the connection between both markers of effort is not yet fully understood (Winn et al., 2018). A better understanding of the relationship

between the pupil response and the perceived effort would be crucial to better link the physiological measure with the listeners' own judgment of their mental process (*Chapter 5*).

At a listener group level, an average JND in effort at 4 dB SNR was found which lies between those obtained for the JND in clarity (2 dB SNR) and JND in meaning (8 dB SNR). The group results are in line with the literature (McShefferty et al., 2016). However, high variability was observed across the listeners indicating that the listener's perception regarding effort investment is highly dependent on the individual. McShefferty and colleagues argued that JND in meaning has more clinical importance as compared to the 'JND in clarity' (see McShefferty et al., 2016 referring to the 'Just Meaningful Difference' and 'JND in SNR') since it indicates the acoustic conditions (change in SNR) under which people are motivated to seek intervention.

Since people with hearing impairment often report listening as being effortful, tiring, or stressful, which can have severe consequences, it is argued here that the JND in effort can be indicative of challenges people experience in everyday listening scenarios. Increased effort investment has been related to higher levels of fatigue, social withdrawal or increasing days of sick leave from work due to mental distress (Kramer, 2009; Kramer et al., 2006). By assessing the JND in effort it is possible to detect changes in SNR needed in order to report changes in effort investment on an individual basis, which may be relevant for the choice of hearing-aid compensation strategies and future rehabilitation techniques.

## 6.2 The future of pupillometry as a marker of individual listening effort

Pupillometry has the potential to be a valuable tool for evaluating listening effort invested during speech recognition. There are several advantages of pupillometry over other markers of effort such as self-reports. One main argument is that the pupil response is assessed 'online', i.e., during listening and performing the task. Thus, pupillometry allows disentangling the effort exertion needed for the different cognitive processes related to the task, while subjective ratings, that are performed retrospectively, might reflect accumulative effort investment. In addition, the pupil response is an objective measure and, therefore, not dependent on the individual's interpretation of effort, which has been speculated

to refer to individuals' performance accuracy (or task difficulty) rather than mental effort.

The findings of this thesis showed that under certain test conditions, high test-retest reliability can be obtained using pupillometry. In addition, several listening factors have been identified impacting the variance of the pupil response both across individuals as well as across repeated visits. These findings are a prerequisite for the development of a measure that has the potential to assess listening effort within individuals and, hence, to become part of the diagnostic protocol in clinics. Such a measure or tool would be highly relevant since it has been shown that effort can hinder hearing rehabilitation (Hornsby, 2013).

Previous studies have shown a tendency of people with HI to withdraw from social interactions by demonstrating that they have reduced performance and increased cognitive demands in speech recognition, due to increased effort (Duquesnoy, 1983; Mattys et al., 2012; Plomp, 1986). More recent research further showed different patterns of PPDs across SNRs between NH and HI listeners (Ohlenforst et al., 2017b; Wagner et al., 2019), which indicates that effort allocation during listening may differ for NH and HI. Thus, further research is needed for specifically exploring the reliability of pupil features for HI listeners.

Furthermore, eye-tracking technology has developed rapidly within the last years allowing data collection to expand from laboratories to daily-life environments. From desktop eye-trackers and flex mount designs to eye-tracking glasses, advances in webcams quality have made it possible to capture eye-tracking data even with tablets or smartphones (Bott et al., 2017; Valliappan et al., 2020). Because of this, scientific enquiry might be potentially expanded to experimental procedures outside laboratories to simulate real-life scenarios. At the same time, the technical barriers and financial burden that might come with testing pupillometry in the clinic have been reduced. However, some care needs to be taken in the application of these new technologies, since pupillometry is a measurement sensitive to light. Also, an evaluation of the applicability of these new technologies is needed to better understand their feasibility in a clinical setting.

Another expansion of the current research could involve experimental setups which include pupil dilation recordings in more complex speech tasks that go beyond sentence repetition because in real life, most listening situations involve conversations with continuous discourse (Ala et al., 2020; MacPherson and Akeroyd, 2013; Speaks et al., 1972). Assessing listening effort during con-

versational scenarios could provide more realistic information on the effort invested by a listener by simulating more ecological situations.

## 6.3 Conclusions

This thesis maps out the potential of pupillometry as a marker of individual listening effort. Test conditions (e.g., SNR) and normalization procedures have been studied to understand the reliability of a selection of pupil features. Overall, high reliability can be observed for some of the pupil features assessed in a speech-in-noise task at multiple visits. Several listener factors contribute to the variability of the pupil response and, hence, are suggested to be taken into consideration when interpreting the individual's pupil response as a marker of listening effort. Finally, to better relate the pupil response to an individual's perception of listening effort, which can be relevant for the interpretation of the individual's pupil response, the JND in effort was introduced. As a whole, the thesis brings pupillometry one step further towards the development of a tool that captures individuals' listening effort.

# Bibliography


Aday, J. S. and J. M. Carlson (Feb. 2019). “Extended testing with the dot-probe task increases test–retest reliability and validity”. In: *Cognitive Processing* 20.1, pp. 65–72.

Ala, T. S., C. Graversen, D. Wendt, E. Alickovic, W. M. Whitme, and T. Lunner (July 2020). “An exploratory Study of EEG Alpha Oscillation and Pupil Dilation in Hearing-Aid Users During Effortful listening to Continuous Speech”. In: *PLOS ONE* 15.7, e0235782.

Alamia, A., R. VanRullen, E. Pasqualotto, A. Mouraux, and A. Zenon (July 2019). “Pupil-Linked Arousal Responds to Unconscious Surprisal”. In: *The Journal of Neuroscience* 39.27, pp. 5369–5376.

Alhanbali, S., P. Dawes, R. E. Millman, and K. J. Munro (2019). “Measures of Listening Effort Are Multidimensional”. In: *Ear and Hearing* 40.5, pp. 1084–1097.

Alhanbali, S., K. J. Munro, P. Dawes, P. J. Carolan, and R. E. Millman (2020). “Dimensions of self-reported listening effort and fatigue on a digits-in-noise task, and association with baseline pupil size and performance accuracy”. In: *International Journal of Audiology* 0.0, pp. 1–11.

Aston-Jones, G. and J. D. Cohen (2005). “An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance”. In: *Annual review of neuroscience* 28, pp. 403–450.

Ayasse, N. D. and A. Wingfield (Jan. 2020). “Anticipatory Baseline Pupil Diameter Is Sensitive to Differences in Hearing Thresholds”. In: *Frontiers in Psychology* 10, p. 2947.

Bafna, T. and J. P. Hansen (June 2021). “Mental fatigue measurement using eye metrics: A systematic literature review”. In: *Psychophysiology* 58.6, e13828.

Beatty, J. and B. Lucero-Wagoner (2000). *The pupillary system*. Pp. 142–162.

Beatty, J. (1982). *Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources*. Tech. rep. 1, pp. 276–292.

Best, V., G. Keidser, J. M. Buchholz, and K. Freeston (Oct. 2015). "An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment". In: *International Journal of Audiology* 54.10, pp. 682–690.

Bitsios, P., R. Prettyman, and E. Szabadi (Nov. 1996). "Changes in Autonomic Function with Age: A Study of Pupillary Kinetics in Healthy Young and Old People". In: *Age and Ageing* 25.6, pp. 432–438.

Bland, J. M. and D. G. Altman (Feb. 1986). "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The Lancet* 327.8476, pp. 307–310.

Borgdorff, P. (1975). *Respiratory fluctuations in pupil size*. Vol. 228. 4. American Physiological Society, pp. 1094–1102.

Bott, N. T., A. Lange, D. Rentz, E. Buffalo, P. Clopton, and S. Zola (June 2017). "Web camera based eye tracking to assess visual memory on a visual paired comparison task". In: *Frontiers in Neuroscience* 11.JUN, p. 370.

Bradshaw, J. L. (Oct. 1969). "Background light intensity and the pupillary response in a reaction time task". In: *Psychonomic Science* 14.6, pp. 271–272.

Brainard, D. H. and S. Vision (1997). "The Psychophysics Toolbox". In: *Spatial vision* 10.4, pp. 433–436.

Bramsløw, L, L. B. Simonsen, M El Hichou, R Hashem, and R. K. Hietkamp (2016). "Learning effects as result of multiple exposures to Danish HINT". In: *Poster presented at the International Hearing Aid Conference, Lake Tahoe, CA, USA*.

Brehm, J. W. and E. A. Self (1989). "The intensity of motivation". In: *Ann. Rev. Psychol* 40, pp. 109–140.

Bremner, F. (Apr. 2009). "Pupil evaluation as a test for autonomic disorders". In: *Clinical Autonomic Research* 19.2, pp. 88–101.

Brysbaert, M. and M. Stevens (Jan. 2018). "Power analysis and effect size in mixed effects models: A tutorial". In: *Journal of Cognition* 1.1.

Carolan, P. J., A. Heinrich, K. J. Munro, and R. E. Millman (Jan. 2022). "Quantifying the Effects of Motivation on Listening Effort: A Systematic Review and Meta-Analysis:" in: *Trends in Hearing* 26.

Carter, M. E. et al. (Oct. 2010). "Tuning arousal with optogenetic modulation of locus coeruleus neurons". In: *Nature Neuroscience 2010 13:12* 13.12, pp. 1526–1533.

Chaney, R. H., C. A. Givens, M. F. Aoki, and M. L. Gombiner (Oct. 1989). "Pupillary responses in recognizing awareness in persons with profound mental retardation". In: *Perceptual and Motor Skills* 69.2, pp. 523–528.

Cicchetti, D. V. (Dec. 1994). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." In: *Psychological Assessment* 6.4, pp. 284–290.

Daguet, I., D. Bouhassira, and C. Gronfier (2019). "Baseline Pupil Diameter Is Not a Reliable Biomarker of Subjective Sleepiness". In: *Frontiers in neurology* 10.FEB.

Dingemanse, G. and A. Goedegebure (Dec. 2021). "Listening Effort in Cochlear Implant Users: The Effect of Speech Intelligibility, Noise Reduction Processing, and Working Memory Capacity on the Pupil Dilation Response". In: *Journal of Speech, Language, and Hearing Research* 65.1, pp. 392–404.

Duchowski, A. T. et al. (2018). "The Index of Pupillary Activity Measuring Cognitive Load vis-à-vis Task Difficulty with Pupil Oscillation". In.

Duquesnoy, A. J. (1983). "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons". In: *The Journal of the Acoustical Society of America* 74, p. 739.

Eckert, M. A., S Teubner-Rhodes, and K Vaden (2016). "Neuroimaging of adaptive control during speech and language processing". In: *Ear Hear* 37, 101S–110S.

Eckstein, M. K., B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge (June 2017). "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?" In: *Developmental Cognitive Neuroscience* 25, pp. 69–91.

Edwards, B. (Aug. 2007). "The Future of Hearing Aid Technology". In: *Trends in Amplification* 11.1, pp. 31–45.

Einhäuser, W., J. Stout, C. Koch, and O. Carter (Feb. 2008). "Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.5, pp. 1704–1709.

Einstein, G. O., R. E. Smith, M. A. McDaniel, and P. Shaw (Sept. 1997). "Aging and prospective memory: The influence of increased task demands at encoding and retrieval". In: *Psychology and Aging* 12.3, pp. 479–488.

Fiedler, L., T. S. Ala, C. Graversen, E. Alickovic, T. Lunner, and D. Wendt (2021). "Hearing aid noise reduction lowers the sustained listening effort during

continuous speech in noise-a combined pupillometry and EEG study". In: *Ear and Hearing*, pp. 1590–1601.

Foroughi, C. K., C. Sibley, and J. T. Coyne (Oct. 2017). "Pupil size as a measure of within-task learning". In: *Psychophysiology* 54.10, pp. 1436–1443.

Francis, A. L., M. K. MacPherson, B. Chandrasekaran, and A. M. Alvar (Mar. 2016). "Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort". In: *Frontiers in Psychology* 7.MAR, p. 263.

Giuliani, N. P., C. J. Brown, and Y.-H. Wu (Sept. 2020). "Comparisons of the Sensitivity and Reliability of Multiple Measures of Listening Effort". In: *Ear & Hearing* Publish Ah, pp. 1–10.

Glasser, A (2011). *Adler's physiology of the eye*. Ed. by L. Levin, S. Nilsson, J. Hoeve, and S. Wu. Elsevier, pp. 502–503.

Hagerman, B. (1984). "Clinical Measurements of Speech Reception Threshold in Noise". In: *Scandinavian Audiology* 13.1, pp. 57–63.

Hart, S. G. and L. E. Staveland (Jan. 1988). "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in Psychology* 52.C, pp. 139–183.

Hastie, T. J. and R. J. Tibshirani (Jan. 1990). "Generalized additive models". In: *Generalized Additive Models* 4.2, pp. 1–335.

Hays, R. D., R. Anderson, and D. Revicki (1993). "Psychometric considerations in evaluating health-related quality of life measures". In: *Quality of Life Research* 2.6, pp. 441–449.

Hepach, R. and G. Westermann (May 2016). "Pupillometry in Infancy Research". In: *Journal of Cognition and Development* 17.3, pp. 359–377.

Hockey, R. (2013). *The psychology of fatigue : work, effort and control*. Cambridge University Press.

Hopstaken, J. F., D. van der Linden, A. B. Bakker, and M. A. Kompier (Sept. 2015). "The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics". In: *Biological Psychology* 110, pp. 100–106.

Hornsby, B. W. (Sept. 2013). "The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands". In: *Ear and Hearing* 34.5, pp. 523–534.

Hornsby, B. W., G. Naylor, and F. H. Bess (2016). "A Taxonomy of Fatigue Concepts and Their Relation to Hearing Loss". In: *Ear and hearing* 37.Suppl 1, 136S.

Htu, R., L. Getty, and S. Waridel (1994). “Attitudes towards co-workers affected by occupational hearing loss II: Focus groups interviews”. In: *British Journal of Audiology* 28.6, pp. 313–325.

Janisse, M. P. (1977). *Pupillometry: The psychology of the pupillary response*. Hemisphere Pub.

Jones, B. E. (2004). “Activity, modulation and role of basal forebrain cholinergic neurons innervating the cerebral cortex”. In: *Progress in brain research* 145, pp. 157–169.

Just, M. A., P. A. Carpenter, and A. Miyake (2010). “Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work”. In: *Theoretical Issues in Ergonomics Science* 4.1-2, pp. 59–88.

Kahneman, D. (1973). *Attention and effort*. Citesser.

Kalénine, S., D. Mirman, E. L. Middleton, and L. J. Buxbaum (Sept. 2012). “Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38.5, pp. 1274–1295.

Karatekin, C. (Mar. 2004). “Development of attentional allocation in the dual task paradigm”. In: *International Journal of Psychophysiology* 52.1, pp. 7–21.

Killion, M. C. and others (2004). “Myths about hearing in noise and directional microphones”. In: *Hearing Review* 11.2, pp. 14–21.

Kim, M., D. Q. Beversdorf, and K. M. Heilman (2000). “Arousal response with aging: Pupillographic study”. In: *Journal of the International Neuropsychological Society* 6.3, pp. 348–350.

Kleiner, M., D. Brainard, and D. Pelli (2007). “What’s new in Psychtoolbox-3?” In: *Perception* 36.

Klingner, J. (2010). *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking*. Stanford University.

Koch, X. and E. Janse (Apr. 2016). “Speech rate effects on the processing of conversational speech across the adult life spana)”. In: *The Journal of the Acoustical Society of America* 139.4, p. 1618.

Koelewijn, T., B. G. Shinn-Cunningham, A. A. Zekveld, and S. E. Kramer (June 2014). “The pupil response is sensitive to divided attention during speech processing”. In: *Hearing Research* 312, pp. 114–120.

Koelewijn, T., N. J. Versfeld, and S. E. Kramer (Oct. 2017). “Effects of attention on the speech reception threshold and pupil response of people with impaired and normal hearing”. In: *Hearing Research* 354, pp. 56–63.

Koelewijn, T., A. A. Zekveld, J. M. Festen, and S. E. Kramer (Mar. 2012a). "Pupil dilation uncovers extra listening effort in the presence of a single-talker masker". In: *Ear and Hearing* 33.2, pp. 291–300.

Koelewijn, T., A. A. Zekveld, J. M. Festen, J. Rönnberg, and S. E. Kramer (2012b). "Processing Load Induced by Informational Masking Is Related to Linguistic Abilities". In: *International Journal of Otolaryngology* 2012, pp. 1–11.

Koelewijn, T., A. A. Zekveld, T. Lunner, and S. E. Kramer (Sept. 2018). "The effect of reward on listening effort as reflected by the pupil dilation response". In: *Hearing Research* 367, pp. 106–112.

Koelewijn, T., A. A. Zekveld, T. Lunner, and S. E. Kramer (July 2021). "The effect of monetary reward on listening effort and sentence recognition". In: *Hearing Research* 406, p. 108255.

Koo, T. K. and M. Y. Li (June 2016). "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". In: *Journal of Chiropractic Medicine* 15.2, pp. 155–163.

Kramer, S. E. (Nov. 2009). "Hearing impairment, work, and vocational enablement". In: *International Journal of Audiology* 47.SUPPL. 2, pp. 124–130.

Kramer, S. E., T. S. Kapteyn, J. M. Festen, and D. J. Kuik (Jan. 1997). "Assessing Aspects of Auditory Handicap by Means of Pupil Dilatation". In: *International Journal of Audiology* 36.3, pp. 155–164.

Kramer, S. E., T. S. Kapteyn, and T. Houtgast (Sept. 2006). "Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work". In: *International Journal of Audiology* 45.9, pp. 503–512.

Kramer, S. E., A. Lorens, F. Coninx, A. A. Zekveld, A. Piotrowska, and H. Skarzynski (2013). "Processing load during listening: The influence of task characteristics on the pupil response". In: *Language and Cognitive Processes* 28.4, pp. 426–442.

Kramer, S. E., C. E. Teunissen, and A. A. Zekveld (2016). "Cortisol, chromogranin A, and pupillary responses evoked by speech recognition tasks in normally hearing and hard-of-hearing listeners: A pilot study". In: *Ear and Hearing* 37.July, 126S–135S.

Krejtzid, K., A. T. Duchowski, A. Niedzielska, B. Cezary, and I. Krejtz (2018). "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze". In.

Krueger, M. et al. (Oct. 2017). "Relation Between Listening Effort and Speech Intelligibility in Noise". In: *American Journal of Audiology* 26.3S, pp. 378–392.

Kuchinsky, S. E., J. B. Ahlstrom, S. L. Cute, L. E. Humes, J. R. Dubno, and M. A. Eckert (Oct. 2014). "Speech-perception training for older adults with hearing loss impacts word recognition and effort". In: *Psychophysiology* 51.10, pp. 1046–1057.

Kuchinsky, S. E. et al. (Jan. 2013). "Pupil size varies with word listening and response selection difficulty in older adults with hearing loss". In: *Psychophysiology* 50.1, pp. 23–34.

Kuchinsky, S. E. et al. (2016). "Task-Related Vigilance during Word Recognition in Noise for Older Adults with Hearing Loss". In: *Experimental Aging Research* 42.1, pp. 64–85.

Laeng, B., S. Sirois, and G. Gredebäck (Jan. 2012). "Pupillometry: A Window to the Preconscious?" In: *Perspectives on psychological science : a journal of the Association for Psychological Science* 7.1, pp. 18–27.

Larsby, B., M. Hällgren, B. Lyxell, and S. Arlinger (Mar. 2005). "Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects". In: *International journal of audiology* 44.3, pp. 131–143.

Lau, M. K., C. Hicks, T. Kroll, and S. Zupancic (May 2019). "Effect of Auditory Task Type on Physiological and Subjective Measures of Listening Effort in Individuals With Normal Hearing". In: *Journal of Speech, Language, and Hearing Research* 62.5, pp. 1549–1560.

Lee, S. H. and Y. Dan (Oct. 2012). "Neuromodulation of Brain States". In: *Neuron* 76.1, pp. 209–222.

Lin, X. and D. Zhang (Apr. 1999). "Inference in generalized additive mixed modelsby using smoothing splines". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.2, pp. 381–400.

Loewenfeld, I. E. (1993). *The pupil: Anatomy, physiology, and clinical applications*. Vol. 2. Iowa State University Press.

MATLAB (2018). *9.7.0.1190202 (R2019b)*. Natick, Massachusetts: The MathWorks Inc.

MacPherson, A. and M. A. Akeroyd (May 2013). "The Glasgow Monitoring of Uninterrupted Speech Task (GMUST): A naturalistic measure of speech intelligibility in noise". In: *Proceedings of Meetings on Acoustics* 19.1, p. 050068.

Mansour, N., M. Marschall, T. May, A. Westermann, and T. Dau (Apr. 2021). "Speech intelligibility in a realistic virtual sound environment". In: *The Journal of the Acoustical Society of America* 149.4, p. 2791.

Marquart, G. and J. De Winter (Aug. 2015). "Workload assessment for mental arithmetic tasks using the task-evoked pupillary response". In: *PeerJ Computer Science* 2015.8, e16.

Mathôt, S., J. Fabius, E. Van Heusden, and S. Van der Stigchel (Feb. 2018). "Safe and sensible preprocessing and baseline correction of pupil-size data". In: *Behavior Research Methods* 50.1, pp. 94–106.

Mattys, S. L., M. H. Davis, A. R. Bradlow, and S. K. Scott (2012). "Language and Cognitive Processes Speech recognition in adverse conditions: A review". In.

May, P. J., A. Reiner, and P. D. Gamlin (May 2019). "Autonomic Regulation of the Eye". In: *Oxford Research Encyclopedia of Neuroscience*.

McCloy, D. R., E. D. Larson, B. Lau, and A. K. C. Lee (Mar. 2016). "Temporal alignment of pupillary response with stimulus events via deconvolution". In: *The Journal of the Acoustical Society of America* 139.3, EL57–EL62.

McClymont, L. G., G. G. Browning, and S. Gatehouse (1991). "Reliability of patient choice between hearing aid systems". In: *British Journal of Audiology* 25.1, pp. 35–39.

McGarrigle, R. et al. (2014). "Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'". In: *International Journal of Audiology* 53.7, pp. 433–445.

McShefferty, D., W. M. Whitmer, and M. A. Akeroyd (Jan. 2015). "The just noticeable difference in speech-to-noise ratio". In: *Trends in Hearing* 19.

McShefferty, D., W. M. Whitmer, and M. A. Akeroyd (Jan. 2016). "The Just Meaningful Difference in Speech-to-Noise Ratio". In: *Trends in Hearing* 20.

Miles, K., T. Beechey, V. Best, and J. Buchholz (Mar. 2022). "Measuring Speech Intelligibility and Hearing-Aid Benefit Using Everyday Conversational Sentences in Real-World Environments". In: *Frontiers in Neuroscience* 16.

Mirman, D., J. A. Dixon, and J. S. Magnuson (Nov. 2008). "Statistical and computational models of the visual world paradigm: Growth curves and individual differences". In: *Journal of Memory and Language* 59.4, pp. 475–494.

Morris, S. K., E. Granholm, A. J. Sarkin, and D. V. Jeste (Oct. 1997). “Effects of schizophrenia and aging on pupillographic measures of working memory”. In: *Schizophrenia Research* 27.2-3, pp. 119–128.

Müller, T. and M. A. Apps (Feb. 2019). “Motivational fatigue: A neurocognitive framework for the impact of effortful exertion on subsequent motivation”. In: *Neuropsychologia* 123, pp. 141–151.

Ng, E. H. N., M. Rudner, T. Lunner, and J. Rönnberg (Dec. 2013). “Relationships between self-report and cognitive measures of hearing aid outcome”. In: *Speech, Language and Hearing* 16.4, pp. 197–207.

Nielsen, J. B. and T. Dau (Mar. 2011). “The Danish hearing in noise test”. In: *International Journal of Audiology* 50.3, pp. 202–208.

Nielsen, J. B., T. Dau, and T. Neher (Jan. 2014). “A Danish open-set speech corpus for competing-speech studies”. In: *The Journal of the Acoustical Society of America* 135.1, p. 407.

Ogawa, T. et al. (July 2019). “Hearing-impaired elderly people have smaller social networks: A population-based aging study”. In: *Archives of Gerontology and Geriatrics* 83, pp. 75–80.

Ohlenforst, B., D. Wendt, S. E. Kramer, G. Naylor, A. A. Zekveld, and T. Lunner (Aug. 2018). “Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response”. In: *Hearing research* 365, pp. 90–99.

Ohlenforst, B. et al. (2017a). “Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review”. In: *Ear and Hearing* 38.3, pp. 267–281.

Ohlenforst, B. et al. (Aug. 2017b). “Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation”. In: *Hearing Research* 351, pp. 68–79.

Partala, T. and V. Surakka (July 2003). “Pupil size variation as an indication of affective processing”. In: *International Journal of Human Computer Studies* 59.1-2, pp. 185–198.

Pattyn, N., J. Van Cutsem, E. Dessy, and O. Mairesse (Sept. 2018). “Bridging exercise science, cognitive psychology, and medical practice: Is "cognitive fatigue" a remake of "the emperor’s new clothes"?” In: *Frontiers in Psychology* 9.SEP, p. 1246.

Peelle, J. E. (2018). “Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior”. In: *Ear and Hearing* 39.2, p. 204.

Pelli, D. G. and S. Vision (1997). “The VideoToolbox software for visual psychophysics: Transforming numbers into movies”. In: *Spatial vision* 10, pp. 437–442.

Pichora-Fuller, M. K. et al. (2016). “Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL)”. In: *Ear and Hearing* 37, 5S–27S.

Picou, E. M., T. A. Ricketts, and B. W. Hornsby (Oct. 2011). “Visual cues and listening effort: individual variability”. In: *Journal of speech, language, and hearing research : JSLHR* 54.5, pp. 1416–1430.

Pielage, H., A. A. Zekveld, G. H. Saunders, N. J. Versfeld, T. Lunner, and S. E. Kramer (2021). “The presence of another individual influences listening effort, but not performance”. In: *Ear and Hearing*, pp. 1577–1589.

Piquado, T., D. Isaacowitz, and A. Wingfield (May 2010). “Pupillometry as a measure of cognitive effort in younger and older adults”. In: *Psychophysiology* 47.3, pp. 560–569.

Pittman, T. S., J. Emery, and A. K. Boggiano (1982). “Intrinsic and extrinsic motivational orientations: Reward-induced changes in preference for complexity”. In: *Journal of Personality and Social Psychology* 42.5, pp. 789–797.

Plomp, R. (1986). “A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired”. In: *Journal of Speech and Hearing Research* 29.2, pp. 146–154.

Plomp, R. and A. M. Mimpen (1979). “Speech-reception threshold for sentences as a function of age and noise level”. In: *The Journal of the Acoustical Society of America* 66.5, pp. 1333–1342.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Reeve, J. (June 1989). “The interest-enjoyment distinction in intrinsic motivation”. In: *Motivation and Emotion 1989 13:2* 13.2, pp. 83–103.

Reilly, J., A. Kelly, S. H. Kim, S. Jett, and B. Zuckerman (Apr. 2019). “The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry”. In: *Behavior Research Methods* 51.2, pp. 865–878.

Reimer, J. et al. (Nov. 2016). “Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex”. In: *Nature Communications 2016 7:1* 7.1, pp. 1–7.

Rij, J. van, P. Hendriks, H. van Rijn, R. H. Baayen, and S. N. Wood (May 2019). “Analyzing the Time Course of Pupillometric Data”. In: *Trends in Hearing* 23.

Ryan, R. M. (1982). “Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory”. In: *Journal of Personality and Social Psychology* 43.3, pp. 450–461.

Saunders, G. H. and A. Forsline (June 2006). “The Performance-Perceptual Test (PPT) and its relationship to aided reported handicap and hearing aid satisfaction”. In: *Ear and Hearing* 27.3, pp. 229–242.

Schmidtke, J. (Sept. 2018). “Pupillometry in Linguistic Research: An Introduction and Review for Second Language Researchers”. In: *Studies in Second Language Acquisition* 40.3, pp. 529–549.

Sibley, C., J. Coyne, and C. Baldwin (Sept. 2011). “Pupil Dilation as an Index of Learning:” in: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pp. 237–241.

Smeds, K., F. Wolters, and M. Rung (Feb. 2015). “Estimation of signal-to-noise ratios in realistic sound scenarios”. In: *Journal of the American Academy of Audiology* 26.2, pp. 183–196.

Speaks, C., B. Parker, C. Harris, and P. Kuhl (1972). “Intelligibility of Connected Discourse”. In: *Journal of speech and hearing research* 15.3, pp. 590–602.

Steel, M. M., B. C. Papsin, and K. A. Gordon (Feb. 2015). “Binaural Fusion and Listening Effort in Children Who Use Bilateral Cochlear Implants: A Psychoacoustic and Pupillometric Study”. In: *PLOS ONE* 10.2, e0117611.

Steinhauer, S. R., M. M. Bradley, G. J. Siegle, K. A. Roecklein, and A. Dix (Apr. 2022). “Publication guidelines and recommendations for pupillary measurement in psychophysiological studies”. In: *Psychophysiology* 59.4, e14035.

Steinhauer, S. R., G. J. Siegle, R. Condray, and M. Pless (2004). “Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing”. In: *International Journal of Psychophysiology* 52, pp. 77–86.

Stephens, D. and R. Hétu (1991). “Impairment, disability and handicap in audiology: Towards a consensus”. In: *International Journal of Audiology* 30.4, pp. 185–200.

Tewes, U. (1991). *Hamburg-Wechsler Intelligenztest für Erwachsene: HAWIE-R.* Huber.

Tryon, W. W. (Jan. 1975). “Pupillometry: A Survey of Sources of Variation”. In: *Psychophysiology* 12.1, pp. 90–93.

Valliappan, N. et al. (Sept. 2020). “Accelerating eye movement research via accurate and affordable smartphone eye tracking”. In: *Nature Communications* 11.1, pp. 1–12.

Van Der Meer, E. et al. (Jan. 2010). “Resource allocation and fluid intelligence: Insights from pupillometry”. In: *Psychophysiology* 47.1, pp. 158–169.

Van Veldhoven, M. and S. Broersen (June 2003). “Measurement quality and validity of the "need for recovery scale"”. In: *Occupational and environmental medicine* 60 Suppl 1.Suppl 1.

Wagner, A. E., L. Nagels, P. Toffanin, J. M. Opie, and D. Başkent (May 2019). “Individual Variations in Effort: Assessing Pupillometry for the Hearing Impaired”. In: *Trends in Hearing* 23.

Wang, Y., S. E. Kramer, D. Wendt, G. Naylor, T. Lunner, and A. A. Zekveld (2018a). “The Pupil Dilation Response During Speech Perception in Dark and Light: The Involvement of the Parasympathetic Nervous System in Listening Effort”. In: *Trends in Hearing* 22, pp. 1–11.

Wang, Y. et al. (2016). “Parasympathetic nervous system dysfunction, as identified by pupil light reflex, and its possible connection to hearing impairment”. In: *PLoS ONE* 11.4, pp. 1–26.

Wang, Y. et al. (2018b). “Relations Between Self-Reported Daily-Life Fatigue, Hearing Status, and Pupil Dilation During a Speech Perception in Noise Task”. In: *Ear and hearing* 39.3, pp. 573–582.

Wechsler, D. (1981). *Wechsler adult intelligence scale-revised (WAIS-R).* Psychological Corporation.

Weinstein, B. E. and I. M. Ventry (1982). “Hearing impairment and social isolation in the elderly”. In: *Journal of Speech and Hearing Research* 25.4, pp. 593–599.

Wendt, D., T. Brand, and B. Kollmeier (2014). “An eye-tracking paradigm for analyzing the processing time of sentences with different linguistic complexities”. In: *PLoS ONE* 9.6.

Wendt, D., T. Dau, and J. Hjortkjær (2016). “Impact of background noise and sentence complexity on processing demands during sentence comprehension”. In: *Frontiers in Psychology* 7.MAR, pp. 1–12.

Wendt, D., R. K. Hietkamp, and T. Lunner (Nov. 2017). "Impact of Noise and Noise Reduction on Processing Effort: A Pupillometry Study". In: *Ear & Hearing* 38.6, pp. 690–700.

Wendt, D., T. Koelewijn, P. Książek, S. E. Kramer, and T. Lunner (Nov. 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". In: *Hearing Research* 369, pp. 67–78.

Westfall, J., D. A. Kenny, and C. M. Judd (2014). "Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli". In: *Journal of Experimental Psychology: General* 143.5, pp. 2020–2045.

Wetzel, N., D. Buttelmann, A. Schieler, and A. Widmann (Apr. 2016). "Infant and adult pupil dilation in response to unexpected sounds". In: *Developmental Psychobiology* 58.3, pp. 382–392.

Widmann, A., E. Schröger, and N. Wetzel (Mar. 2018). "Emotion lies in the eye of the listener: Emotional arousal to novel sounds is reflected in the sympathetic contribution to the pupil dilation response and the P3". In: *Biological Psychology* 133, pp. 10–17.

Winn, B, D Whitaker, D. B. Elliott, and N. J. Phillips (Mar. 1994). "Factors affecting light-adapted pupil size in normal human subjects". In: *Investigative Ophthalmology & Visual Science* 35.3, pp. 1132–1137.

Winn, M. B. (Jan. 2016). "Rapid Release From Listening Effort Resulting From Semantic Context, and Effects of Spectral Degradation and Cochlear Implants". In: *Trends in Hearing* 20, p. 233121651666972.

Winn, M. B., J. R. Edwards, and R. Y. Litovsky (2015). "The impact of auditory spectral resolution on listening effort revealed by pupil dilation". In: *Ear and Hearing* 36.4, e153–e165.

Winn, M. B., D. Wendt, T. Koelewijn, and S. E. Kuchinsky (Jan. 2018). "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started". In: *Trends in Hearing* 22, p. 233121651880086.

Wood, S. N. (Jan. 2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1, pp. 3–36.

Wood, S. N., Z. Li, G. Shaddick, and N. H. Augustin (July 2017). “Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data”. In: *Journal of the American Statistical Association* 112.519, pp. 1199–1210.

Zekveld, A. A., T. Koelewijn, and S. E. Kramer (Jan. 2018). “The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge”. In: *Trends in Hearing* 22, p. 233121651877717.

Zekveld, A. A. and S. E. Kramer (2014). “Cognitive processing load across a wide range of listening conditions: Insights from pupillometry”. In: *Psychophysiology* 51.3, pp. 277–284.

Zekveld, A. A., S. E. Kramer, and J. M. Festen (Aug. 2010). “Pupil Response as an Indication of Effortful Listening: The Influence of Sentence Intelligibility”. In: *Ear & Hearing* 31.4, pp. 480–490.

Zekveld, A. A., S. E. Kramer, and J. M. Festen (July 2011). “Cognitive Load During Speech Perception in Noise: The Influence of Age, Hearing Loss, and Cognition on the Pupil Response”. In: *Ear & Hearing* 32.4, pp. 498–510.

Zénon, A., M. Sidibé, and E. Olivier (Aug. 2014). “Pupil size variations correlate with physical effort perception”. In: *Frontiers in Behavioral Neuroscience* 8.AUG, p. 286.

Zwyghuizen-Doorenbos, A., T. Roehrs, M. Schaefer, and T. Roth (Sept. 1988). “Test-Retest Reliability of the MSLT”. In: *Sleep* 11.6, pp. 562–565.

# Appendix

| ICC | Feature | Nasa Tlx | | Subjective effort | |
|---|---|---|---|---|---|
| | | Visit 1-2 (11) | Visit 2-3 (11) | Visit 1-2 (11) | Visit 2-3 (11) |
| | All SNRs | ***0.84*** | ***0.88*** | ***0.86*** | ***0.86*** |
| | -12 dB | **0.64** | ***0.93*** | 0.52 | **0.67** |
| | -8 dB | ***0.83*** | ***0.89*** | ***0.9*** | 0.45 |
| | -4 dB | ***0.85*** | ***0.91*** | 0.4 | ***0.76*** |
| | 0 dB | ***0.93*** | ***0.91*** | **0.72** | ***0.75*** |
| | 4 dB | ***0.88*** | ***0.86*** | ***0.94*** | ***0.78*** |

Table 1: ICC values for the subjective measures of effort, comparisons between Visit 1, 2 and 3 for a subsample of 11 participants. Values between 0.6 and 0.75, representing good reliability, are highlighted in black bold and values above 0.75, representing excellent reliability, are highlighted in italic bold.

| ICC | Feature | PPD Visit 1-2 (11) | PPD Visit 2-3 (11) | MPD Visit 1-2 (11) | MPD Visit 2-3 (11) | AH Visit 1-2 (11) | AH Visit 2-3 (11) | Slope Visit 1-2 (11) | Slope Visit 2-3 (11) | RF Visit 1-2 (11) | RF Visit 2-3 (11) | Delay Visit 1-2 (11) | Delay Visit 2-3 (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline correction | All SNRs | **0.67** | ***0.80*** | ***0.75*** | **0.74** | 0.59 | ***0.75*** | 0.57 | 0.56 | **0.64** | **0.64** | **0.68** | ***0.84*** |
| | -12 dB | 0.59 | **0.64** | **0.6** | ***0.83*** | 0.28 | ***0.83*** | **0.68** | **0.74** | **0.68** | ***0.83*** | **0.62** | ***0.89*** |
| | -8 dB | ***0.79*** | **0.65** | 0.52 | 0.34 | ***0.8*** | ***0.82*** | 0.27 | 0 | **0.73** | **0.67** | **0.6** | ***0.8*** |
| | -4 dB | 0.11 | 0.45 | **0.64** | **0.65** | 0 | 0.06 | **0.67** | 0.56 | 0 | 0 | ***0.78*** | ***0.91*** |
| | 0 dB | **0.62** | 0.56 | 0.41 | **0.6** | 0.47 | 0.57 | 0.3 | 0.26 | 0.2 | **0.71** | 0.42 | ***0.78*** |
| | 4 dB | ***0.87*** | 0.3 | **0.72** | 0.27 | **0.74** | 0.14 | 0.23 | 0.59 | ***0.92*** | **0.68** | 0.51 | 0.36 |
| Range normalization | All SNRs | 0.59 | 0.59 | 0.58 | 0.59 | ***0.97*** | ***0.98*** | **0.64** | 0.57 | **0.74** | **0.69** | **0.67** | ***0.77*** |
| | -12 dB | **0.67** | ***0.82*** | **0.67** | ***0.79*** | 0 | 0.27 | **0.62** | **0.71** | ***0.79*** | ***0.79*** | ***0.79*** | **0.66** |
| | -8 dB | **0.74** | **0.64** | ***0.81*** | **0.71** | **0.66** | 0.35 | **0.65** | 0 | 0 | ***0.77*** | 0.42 | **0.66** |
| | -4 dB | ***0.77*** | ***0.85*** | ***0.8*** | **0.74** | 0.33 | 0.41 | 0.39 | 0.41 | 0.5 | 0 | **0.7** | ***0.85*** |
| | 0 dB | ***0.83*** | **0.81** | ***0.8*** | **0.73** | **0.64** | 0.42 | 0.49 | 0.45 | 0.42 | **0.62** | 0 | 0.49 |
| | 4 dB | ***0.78*** | ***0.83*** | ***0.81*** | ***0.85*** | 0.53 | **0.68** | ***0.91*** | **0.64** | 0.31 | **0.63** | 0.48 | 0.41 |
| Z-score | All SNRs | 0.39 | 0.36 | 0.33 | 0.34 | 0 | 0 | 0.55 | 0.46 | **0.71** | 0.53 | **0.66** | **0.72** |
| | -12 dB | **0.64** | 0.34 | 0.17 | **0.74** | 0 | 0 | ***0.79*** | **0.66** | **0.62** | ***0.8*** | ***0.77*** | **0.66** |
| | -8 dB | ***0.75*** | 0.47 | 0.23 | 0.12 | 0 | 0 | 0 | 0 | **0.68** | ***0.78*** | 0.35 | **0.64** |
| | -4 dB | 0 | 0 | 0 | 0.11 | 0 | 0 | 0.5 | 0.39 | 0.48 | 0 | **0.72** | ***0.85*** |
| | 0 dB | 0.25 | 0 | 0.29 | 0.51 | 0 | 0 | 0.19 | 0.26 | 0.17 | 0.5 | 0 | 0.41 |
| | 4 dB | 0.59 | 0 | 0.47 | 0.34 | 0 | 0 | 0 | 0.55 | ***0.92*** | 0.51 | 0.5 | 0.5 |
| Baseline Range | All SNRs | ***0.82*** | ***0.87*** | ***0.86*** | ***0.87*** | ***0.97*** | ***0.98*** | **0.64** | 0.57 | **0.74** | **0.69** | **0.67** | ***0.77*** |
| | -12 dB | **0.72** | 0.59 | ***0.9*** | ***0.95*** | ***0.86*** | ***0.95*** | ***0.78*** | **0.72** | **0.64** | ***0.81*** | **0.72** | ***0.79*** |
| | -8 dB | ***0.86*** | **0.68** | ***0.9*** | ***0.88*** | ***0.99*** | ***0.99*** | 0.1 | 0 | 0.58 | ***0.79*** | 0.52 | **0.69** |
| | -4 dB | 0 | 0 | ***0.92*** | ***0.94*** | ***0.95*** | ***0.95*** | 0.17 | 0.28 | 0.5 | 0 | **0.71** | ***0.85*** |
| | 0 dB | 0.23 | 0.18 | ***0.92*** | ***0.95*** | ***0.99*** | ***0.98*** | **0.63** | 0.54 | **0.62** | ***0.76*** | 0 | 0.47 |
| | 4 dB | ***0.83*** | 0.44 | ***0.95*** | ***0.93*** | ***0.98*** | ***0.98*** | 0.49 | **0.62** | ***0.91*** | **0.64** | 0.55 | 0.47 |

Table 2: ICC values for all normalization procedures, SNRs and comparisons between Visit 1, 2 and 3 for a subsample of 11 participants. Values between 0.6 and 0.75, representing good reliability, are highlighted in black bold and values above 0.75, representing excellent reliability, are highlighted in italic bold.

# Contributions to Hearing Research

**Vol. 1:** *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.
External examiners: Mark Lutman, Stefan Stenfeld

**Vol. 2:** *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.
External examiners: Brian Moore, Kathrin Krumbholz

**Vol. 3:** *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.
External examiners: Michael Akeroyd, Armin Kohlrausch

**Vol. 4:** *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.
External examiners: Jesko Verhey, Steven van de Par

**Vol. 5:** *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.
External examiners: Björn Hagerman, Ejnar Laukli

**Vol. 6:** *Helen Connor*, Hearing aid amplification at soft input levels, 2010.
External examiners: Inga Holube, Birgitta Larsby

**Vol. 7:** *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.
External examiners: Birger Kollmeier, Ray Meddis

**Vol. 8:** *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.
External examiners: David Kemp, Stephen Neely

**Vol. 9:** *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.
External examiners: Bernhard Seeber, Michael Vorländer

**Vol. 10:** *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.
External examiners: Christopher Plack, Christian Lorenzi

**Vol. 11:** *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.
External examiners: Joost Festen, Jürgen Tchorz

**Vol. 12:** *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.
External examiners: Bob Burkard, Stephen Neely

**Vol. 13:** *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.
External examiners: Stuart Rosen, Christian Lorenzi

**Vol. 14:** *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.
External examiners: Michael Stone, Oded Ghitza

**Vol. 15:** *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.
External examiners: John Culling, Martin Cooke

**Vol. 16:** *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.
External examiners: Lawrence Rosenblum, Matthias Gondan

**Vol. 17:** *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.
External examiners: Shihab Shamma, Guy Brown

**Vol. 18:** *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.
External examiners: Sascha Spors, Ville Pulkki

**Vol. 19:** *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.
External examiners: Bernhard Seeber, Steven van de Par

**Vol. 20:** *Michał Feręczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.
External examiners: Christopher Plack, Enrique Lopez-Poveda

**Vol. 21:** *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.
External examiners: Steven van de Par, John Culling

**Vol. 22:** *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.
External examiners: Ingrid Johnsrude, Christian Lorenzi

**Vol. 23:** *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.
External examiners: Judy Dubno, Martin Cooke

**Vol. 24:** *Johannes Käsbach*, Characterizing apparent source width perception, 2016.
External examiners: William Whitmer, Jürgen Tchorz

**Vol. 25:** *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.
External examiners: Thomas Brand, Armin Kohlrausch

**Vol. 26:** *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.
External examiners: Laurel Carney, Bob Carlyon

**Vol. 27:** *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.
External examiners: Volker Hohmann, Piotr Majdak

**Vol. 28:** *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.
External examiners: Maria Chait, Christian Lorenzi

**Vol. 29:** *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.
External examiners: Ville Pulkki, Pavel Zahorik

**Vol. 30:** *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.
External examiners: Roland Schaette, Ian Bruce

**Vol. 31:** *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.
External examiners: Enrique Lopez-Poveda, Tim Jürgens

**Vol. 32:** *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.
External examiners: Armin Kohlrausch, James Kates

**Vol. 33:** *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.
External examiners: Stefan Bleeck, Jürgen Tchorz

**Vol. 34:** *François Guérit*, Temporal charge interactions in cochlear implant listeners, 2018.
External examiners: Julie Arenberg, Olivier Macherey

**Vol. 35:** *Andreu Paredes Gallardo*, Behavioral and objective measures of stream segregation in cochlear implant users, 2018.
External examiners: Christophe Micheyl, Monita Chatterjee

**Vol. 36:** *Søren Fuglsang*, Characterizing neural mechanisms of attention-driven speech processing, 2019.
External examiners: Shihab Shamma, Maarten de Vos

**Vol. 37:** *Borys Kowalewski*, Assessing the effects of hearing-aid dynamic-range compression on auditory signal processing and perception, 2019.
External examiners: Brian Moore, Graham Naylor

**Vol. 38:** *Helia Relaño Iborra*, Predicting speech perception of normal-hearing and hearing-impaired listeners, 2019.
External examiners: Ian Bruce, Armin Kohlrausch

**Vol. 39:** *Axel Ahrens*, Characterizing auditory and audio-visual perception in virtual environments, 2019.
External examiners: Pavel Zahorik, Piotr Majdak

**Vol. 40:** *Niclas A. Janssen*, Binaural streaming in cochlear implant patients, 2019.
External examiners: Tim Jürgens, Hamish Innes-Brown

**Vol. 41:** *Wiebke Lamping*, Improving cochlear implant performance through psychophysical measures, 2019.
External examiners: Can Gnasia, David Landsberger

**Vol. 42:** *Antoine Favre-Félix*, Controlling a hearing aid with electrically assessed eye gaze, 2020.
External examiners: Jürgen Tchorz, Graham Naylor

**Vol. 43:** *Raul Sanchez Lopez*, Clinical auditory profiling and profile-based hearing-aid fitting, 2020.
External examiners: Judy R. Dubno, Pamela E. Souza

**Vol. 44:** *Juan Camilo Gil Carvajal*, Modeling audiovisual speech perception, 2020.
External examiners: Salvador Soto-Faraco, Kaisa Maria Tippana

**Vol. 45:** *Charlotte Amalie Emdal Navntoft*, Improving cochlear implant performance with new pulse shapes: a multidisciplinary approach, 2020.
External examiners: Andrew Kral, Johannes Frijns

**Vol. 46:** *Naim Mansour*, Assessing hearing device benefit using virtual sound environments, 2021.
External examiners: Virginia Best, Pavel Zahorik

**Vol. 47:** *Anna Josefine Munch Sørensen*, The effects of noise and hearing loss on conversational dynamics, 2021.
External examiners: William McAllister Whitmer, Martin Cooke

**Vol. 48:** *Thirsa Huisman*, The influence of vision on spatial localization in normal-hearing and hearing-impaired listeners, 2021.
External examiners: Steven van de Par, Christopher Stecker

**Vol. 49:** *Florine Lena Bachmann*, Subcortical electrophysiological measures of running speech, 2021.
External examiners: Samira Anderson, Tobias Reichenbach

**Vol. 50:** *Nicolai Pedersen*, Audiovisual speech analysis with deep learning, 2021.
External examiners: Zheng-Hua Tan, Hani Camille Yehia

**Vol. 51:** *Aleksandra Koprowska*, Auditory Training Strategies to Improve Speech Intelligibility in Hearing-Impaired Listeners, 2022.
External examiners: Ulrich Hoppe, David Jackson Morris

**Vol. 52:** *Mie Lærkegård Jørgensen*, Exploring innovative Hearing Aid Techniques for Tinnitus Treatment, 2022.
External examiners: Pim van Dijk, Tobias Kleinjung

**Vol. 53:** *Chiara Casolani*, Electrophysiological characterization of tinnitus in listeners with normal audiogram and supra-threshold hearing deficits, 2022.
External examiners: Pim van Dijk, Holger Schulze

*The end.*

*To be continued…*

Participating in a conversation is an essential part of human social interaction. However, speech communication involving background noise can become challenging for everyone, but in particular for people with hearing impairment. Listening effort is a common complaint among people with hearing impairment that could, eventually, have psycho-social consequences leading to hearing-impaired people's withdrawing from social interactions and becoming socially isolated. Therefore, one of the most important outcomes of hearing rehabilitation is to alleviate hearing-impaired people's ability to participate in social interaction, by addressing listening effort. Pupillometry was widely used as an objective measure of listening effort in the past decades providing results as averages across individuals. However, to develop this tool toward clinical use, a closer look at individual pupil response indicating listening effort was needed together with an investigation of pupillometry's reliability and sensitivity. This thesis identified the conditions at which different pupil features provide reliable results. Moreover, this thesis demonstrated that listener factors such as motivation and fatigue have a strong impact on pupil features, while cognitive abilities and age seem to affect their reliability. Finally, this thesis established a link between changes in pupil response and perceptual effort, when investigating its sensitivity to signal-to-noise ratios. The results are promising for the prospect of using pupillometry as an objective measure to evaluate individual listening effort in clinics.

**DTU Health Tech**
Department of Health Technology

Ørsteds Plads
Building 352
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 71 34 85 99
www.dtu.dk