UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Science
School of Engineering

Listening Hard: New Tools for Monitoring Effort in Listening to Speech in Noise

by

Xiaoxue Wang
ORCiD: 0009-0003-8894-2376

A thesis for the degree of
Doctor of Philosophy

December 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Science
School of Engineering

Doctor of Philosophy

Listening Hard: New Tools for Monitoring Effort in Listening to Speech in Noise

by Xiaoxue Wang

Abstract

Background Listening effort refers to the cognitive exertion required to understand auditory information, particularly in challenging environments. Excessive effort can negatively affect communication, well-being, and cognitive function, especially in individuals with hearing loss or older adults. However, defining and measuring listening effort remains complex, as subjective reports, behavioural performance, and physiological indices often diverge, and substantial individual variability exists in how listeners respond to auditory challenges. Existing theoretical frameworks like the Framework for Understanding Effortful Listening (FUEL) and the Ease of Language Understanding (ELU) model highlight the interplay of cognitive resources, task demands, and listener factors, but a deeper understanding of the dynamic physiological mechanisms and individual response patterns is still needed.

Aims This research aims to provide a better understanding individual differences in physiological responses to effortful listening in normal and hearing impaired listeners and during a range of task difficulty, and relating these to subjective perception and word recognition scores.. The novelty and focus of this work lie in analysing the **temporal dynamics** of physiological responses (i.e., response shape), rather than reducing time-series data to isolated summary values to identify patterns of response. This approach enables a richer and more dynamic understanding of listening effort and listening experiments..

Methods A two-study approach was employed. **Study 1** involved secondary analysis of data from 30 older adults (aged 51–80) with varying hearing loss, who performed a digit-in-noise recall task under adaptive signal-to-noise ratios (SNRs). Measures included multi-channel EEG (alpha power), GSR, pupillometry, subjective workload (NASA-Task Load Index), and accuracy. Analyses focused on characterising individual response time-courses, assessing response shape consistency across sessions, clustering based on waveform similarity, and examining links with hearing level (PTA) and outcomes.

Study 2 involved a new experiment with 30 normal-hearing adults (aged 18–40), who completed a complex sentence-in-noise task using a word-matrix procedure at four fixed SNR levels (–16, –11, –6, and +12 dB). Data were collected across multiple measurements, including single-channel EEG (Pz), galvanic skin response (GSR), pupillometry, electrocardiography (ECG), respiration, subjective ratings, and word identification accuracy. Analyses examined the effects of SNR on subjective effort and difficulty, as well as response dynamics across physiological measures. Analysis especially focusing on within-subject consistency, shape-based clustering across individuals within each SNR condition, and inter-measurement correlations and clustering agreement.

Both studies emphasised time-course shape analysis within defined windows, complementing traditional point-based metrics. Study 2 extended Study 1 by using a range of SNRs, sentence stimuli, and additional physiological measures (ECG, respiration) to broaden physiological insight.

Results Both studies confirmed substantial individual variability in physiological responses, alongside within-subject consistency—particularly for respiration (across all SNRs in Study 2), and for GSR, pupil, and EEG under specific conditions. These patterns suggest stable individual response styles. Physiological activity was modulated by task events and, in Study 2, systematically by SNR, with greater difficulty typically producing stronger or altered responses.

Time-course-based clustering consistently revealed subgroups, but cluster membership generally did not predict behavioural accuracy or subjective ratings, highlighting a gap between physiology and perceived effort. In Study 2, shifting cluster membership across SNR levels suggested task-difficulty-dependent, rather than fixed, response patterns. Low agreement between clusters across physiological modalities further indicated that each system—EEG, GSR, pupil, heart rate, and respiration—captures distinct yet complementary aspects of listening effort. Notably, GSR change during listening was consistently associated with higher accuracy and lower perceived difficulty.

Conclusions Listening effort emerges as a dynamic, multi-system physiological process marked by significant individual variability. While individuals show consistent response patterns, these are modulated by task difficulty and context, and do not directly align with behavioural performance or perceived effort. Different physiological measures offer complementary perspectives, highlighting the value of multi-modal, time-course analysis focused on response shape. The dissociation between physiology and outcomes suggests a key role for internal effort regulation and cognitive strategies—factors not fully captured by standard behavioural or subjective metrics. These findings support dynamic models of listening effort and point to the need for more individualised assessment.

Contents

List of Figures

List of Tables

Glossary

ABR Auditory Brainstem Response 29

ANS Autonomic Nervous System 41

ASA Auditory Scene Analysis 11–13

ATR Average Trial Response ix, 74, 79, 83, 87, 89, 92, 101, 106, 137

ECG Electrocardiography x, 6, 57, 59, 60, 105, 106, 110, 116, 117, 119, 127, 203, 214

EEG Electroencephalogram x, xii, 6, 42, 45–47, 56, 57, 59, 60, 65–67, 69, 70, 73, 74, 82, 83, 101, 102, 105–107, 110, 116, 117, 120, 127, 131, 137, 202–204, 207–209, 212, 214, 217

ELU Ease of Language Understanding 22, 26, 27, 31, 38, 53, 56, 218

fMRI functional Magnetic Resonance Imaging 45–47

fNIRS Functional Near-Infrared Spectroscopy 47

FUEL Framework for Understanding Effortful Listening xii, 22, 24–27, 32, 53, 56, 210–212, 218

GSR Galvanic Skin Response ix, x, xii, xiii, 6, 32, 42, 57, 59, 60, 65–67, 70, 73, 86, 87, 89, 101, 102, 105–107, 110, 115–117, 119, 125, 127, 129, 130, 137, 153, 154, 202, 203, 207–210, 212, 214, 217

HHL Hidden Hearing Loss 28, 29

HR Heart Rate 40, 106, 107, 203

HRV Heart Rate Variability 40

ISVR Institute of Sound and Vibration Research 122

OLSA Oldenburger Satztest, or Oldenburg Sentence Test xiii, 110, 111

PNS Parasympathetic Nervous System 41

PTA Pure Tone Average xvii, 58, 59, 69, 77, 78, 86, 101, 112, 113

RR Respiratory Rate 40, 41, 106, 107, 166, 168, 170, 171, 176, 179, 203

RT Reaction Time 36, 38, 39

SNR Signal to Noise Ratio x, xi, xiii, 34, 40, 41, 56–60, 77, 101, 102, 105–107, 110, 112, 113, 134, 137, 143, 146–149, 154, 159–162, 168, 172, 173, 180, 183, 185, 192, 199, 202, 203, 207–212, 215

SNS Sympathetic Nervous System 41, 42

TCR Time Course Response ix, 74, 80, 87, 88, 101, 137

TV Tidal Volume 41

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   X. Wang, S. Alhanbali, R. Millman, K. Munro, D. Simpson, Individual differences in physiological responses to effortful listening, International Journal of Psychophysiology, Volume 188, Supplement, 2023, DOI:10.1016/j.ijpsycho.2023.05.211.

Signed:..... Date:.....

Acknowledgements

I would like to express my sincere gratitude to all those who have supported me throughout the course of my research and the writing of this thesis.

First and foremost, I would like to thank my supervisors, Professor David Simpson, Professor Stefan Bleeck, and Dr Katie Plant, for their invaluable guidance, encouragement, and support. Their expertise and thoughtful feedback have been essential to my development as a researcher.

I am also deeply grateful to my collaborators, Dr Sara Alhanbali and Professor Kevin Munro from University of Manchester, for providing the dataset used in the first study and for their valuable contributions to this work.

I would like to extend my sincere thanks to Professor Paul White and Professor Filippo Fazi for their support, trust, insightful advice, and readiness to help whenever needed throughout this journey.

I am also grateful to the wider academic staff of the Signal Processing Audio and Hearing Group (SPAH), Institute of Sound and Vibration Research (ISVR), for providing a collaborative and supportive environment during my time here.

I would like to thank my colleagues and friends in the SPAH / ISVR, particularly those with whom I shared an office, for their companionship, stimulating discussions, and the insights that have significantly enriched my research.

I am thankful to my family for their encouragement and kindness throughout this journey.

I am especially grateful to my husband, *Yi*, for his unwavering support, patience, and belief in me throughout this journey. His encouragement has been a constant source of strength.

Finally, I would like to thank my friends for their support and understanding, which have meant a great deal during the more challenging moments of this process.

Without the guidance, assistance, and encouragement of all those mentioned above, this thesis would not have been possible.

Part I

Background

Chapter 1

Introduction

Introduction

Broad Context and Definition Navigating the complexities of the auditory world, from engaging in conversations in bustling social settings to discerning critical warning sounds, human's ability to process auditory information is remarkable. However, this process is not always effortless. In everyday situations-communicating across a noisy restaurant, understanding a speaker with an unfamiliar accent, or deciphering speech over a poor phone connection-we engage in "listening effort".

This term involves the deliberate cognitive exertion required to attend to, process, and comprehend auditory signals, particularly when they are degraded, masked, or otherwise challenging (McGarrigle et al., 2014). Far from being a passive reception of sound, listening under such conditions demands active allocation of finite cognitive resources, including attention, working memory, and executive control (Pichora-Fuller et al., 2016c).

Significance and Impact The requirement for sustained or excessive listening effort affects far beyond communication struggles. Chronic exertion can leads to substantial cognitive fatigue, a pervasive sense of exhaustion distinct from physical tiredness that can impair daily functioning (Edwards et al., 2016; Hornsby, 2013). This listening-related fatigue, alongside the inherent stress of difficult communication, can negatively impact mental well-being and often leads individuals to avoid socially demanding situations, fostering social withdrawal and isolation (Heffernan et al., 2016). This impact is felt most strongly by vulnerable groups.

Older adults often experience age-related declines in auditory processing and cognitive function that increase effort (Peelle, 2018), while the estimated 466 million individuals globally with hearing impairment face a constant struggle against degraded auditory input (World Health Organization, 2021). For both groups, heightened listening effort is

not merely an inconvenience but a factor linked to reduced quality of life (Heffernan et al., 2016) and potentially contributing to accelerated cognitive decline and increased dementia risk (Lin et al., 2013). Beyond the individual, the societal costs are also considerable, revealing as barriers to learning in acoustically challenging classrooms (Zekveld et al., 2018) and impacting workplace productivity and safety in numerous professions (Hornsby & Kipp, 2016; Mattys et al., 2018).

Research Problem and Gap Despite widespread recognition of its importance, listening effort remains difficult to capture and comprehend. A central challenge lies in its measurement. Traditional approaches relying on subjective self-reports, behavioural task performance (like recognition accuracy), and objective physiological indices frequently produce disconnected results - a high score on one measure does not reliably predict the outcome on another (Alhanbali et al., 2018; Ohlenforst et al., 2017). Standard clinical audiological assessments, such as pure-tone audiometry conducted in quiet environments, often fail to reflect the real-world difficulties experienced by listeners in noise (Ohlenforst et al., 2017; Tremblay & Backer, 2015), further complicating diagnosis and intervention. Furthermore, listeners exhibit substantial individual variability in their responses to auditory challenges, driven by factors like hearing status, cognitive capacity, motivation, and potentially even personality traits (Koelewijn et al., 2012; Peelle, 2018; Zekveld et al., 2011).

These individual differences are often masked by research focusing on group averages. Existing theoretical frameworks, notably the Framework for Understanding Effortful Listening (FUEL) (Pichora-Fuller et al., 2016c), which extends Kahneman's capacity model (Kahneman, 1973), and the Ease of Language Understanding (ELU) model (Rönnberg et al., 2013), which emphasises the role of working memory in compensating for perceptual mismatch, provide crucial insights. However, a deeper investigation is required to understand the dynamic, multi-system physiological mechanisms underlying effortful listening and to characterise the consistency and differences of individual physiological response patterns.

Research Aim The overarching aim of this research program is to achieve a more comprehensive, multi-dimensional understanding of listening effort by focusing explicitly on its physiological responses (see Chapter 5, Page 55 for detail). In addition to examine static or peak measures, this research especially focus on investigating the dynamic time-course of physiological signals - how responses in systems like the brain (EEG), and autonomic nervous system (GSR, pupillometry, heart rate, and respiration) interact during effortful listening tasks (Koelewijn et al., 2018; Winn et al., 2016). The specific goals are as follows : (1) to characterise individual differences in these dynamic physiological response patterns, (2) to assess the consistency of these patterns within individuals over repeated exposures, and (3) to examine how these physiological

responses are modulated by systematically varied task difficulty (signal-to-noise ratio) and listener characteristics (hearing status), (4) to examine the relationship of different measurements of listening effort..

Methodology Overview To address these multifaceted aims, a two-study research strategy was employed. Study 1 focuses on an existing dataset, performing a secondary analysis of multi-channel EEG, GSR, and pupillometry data from older adults (aged 51-80) with varying degrees of hearing loss undertaking a digit-recall task with individually adapted SNRs. This allowed an initial characterisation of individual response consistency and physiological subgroups in a clinically relevant population. Study 2 involved conducting experiment with younger (aged 18-40), normal-hearing adults performing a more complex and more ecological sentence-in-noise recognition task (the Oldenburg Sentence Test, OLSA, *Oldenburger Satztest* in German; (Hey et al., 2014; Neumann et al., 2012)) at four fixed SNR levels spanning a wide difficulty range.

Study 2 also expanded the physiological assessment to include ECG (for heart rate), and respiration(for respiration rate), alongside EEG, GSR and pupillometry, enabling a systematic investigation of task difficulty effects and cross-modal relationships. In both studies, the analysis focused on the dynamic shape of individual responses over time, rather than averaged data points, were used in permutation testing for consistency and correlation-based clustering for identifying response patterns.

By integrating findings from both studies, this research contributes to a deeper understanding of listening effort through multi-dimensional measurements—combining subjective reports, task performance, and physiological responses. It places particular emphasis on individual differences (beyond group averages), dynamic time-course patterns (beyond static data points), and the relationships between these different measures. Together, these insights provide an empirical foundation for developing more sensitive assessment tools and potentially more personalised interventions.

To understand the motivation behind this research programme, it's important to first explore the key foundational concepts. The following chapter will provide a detailed background review, exploring the mechanisms of human hearing, the cognitive processes involved in listening, the evolution of the listening effort concept, existing theoretical models, factors influencing effort, and established measurement techniques.

Thesis Structure

This thesis is organised as follows:

- **Part I: Background** - This section provides an overview of the theoretical foundations, previous research, and key concepts relevant to this study.
- **Part II: Study One** - The first study investigates the relationship between listening effort and various physiological responses, including EEG, GSR, and pupillometry.
- **Part III: Study Two** - The second study broadens the exploration of effortful listening by incorporating a wider range of difficulty levels in listening tasks, using sentences instead of simple digits as stimuli, and analyzing additional physiological indices such as ECG and respiration.
- **Part IV: Discussion and Conclusion** - This final section discusses the key findings of the research, their broader implications, limitations of the study, and potential directions for future research.

This structure ensures a coherent progression from theoretical foundations to empirical findings and practical applications.

Chapter 2

Sense of Hearing

In this chapter, we introduce the fundamental principles of human hearing and explore how sound is perceived and processed by the auditory system. We begin with the physical properties of sound and how the ear and brain process basic auditory stimuli. We then describe the cognitive mechanisms that support listening, from attention to executive function, and how these underlie higher-level processes such as auditory scene analysis and speech perception. Finally, we discuss how these mechanisms are affected by hearing impairment, linking to broader cognitive consequences.

2.1 The Auditory System

Peripheral Auditory Processing

Physical Characteristics of Sound Sound is a form of energy that propagates through a medium, such as air, water, or solids (Moore, 2012; Rossing et al., 2002). It is characterised by several physical properties, including frequency, amplitude, and timbre. Those physical features influence how sound is perceived by the auditory system (Gazzaniga et al., 2014).

Frequency refers to the number of cycles per second of a sound wave, measured in hertz (Hz), which is closely related to pitch, a perceptual experience of frequency (Howard & Angus, 2017). Human auditory system can typically detect frequencies ranging from 20 to 20,000 Hz, which tends to diminish with age, especially at high frequencies (Moore, 2012).

Another feature of sound, amplitude, represents the magnitude of air pressure variation in the sound wave, which correlates with the perception of loudness. Amplitude of sound is measured in decibels (dB), while loudness, the subjective experience is measured by Phon. Perceived loudness is affected not by amplitude, but by pitch as well (see figure 2.1

). Normal conversation typically falls the around 60dB, whilst prolonged exposure to sounds exceeding 85 dB may cause hearing damage (World Health Organization, 2018).

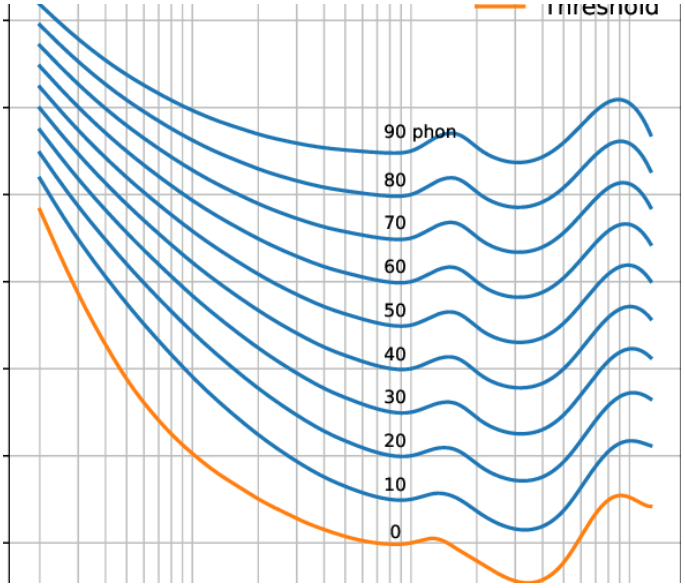


Figure. 2.1. Equal Loudness Contours (Veronesi & Mauney, 2022).

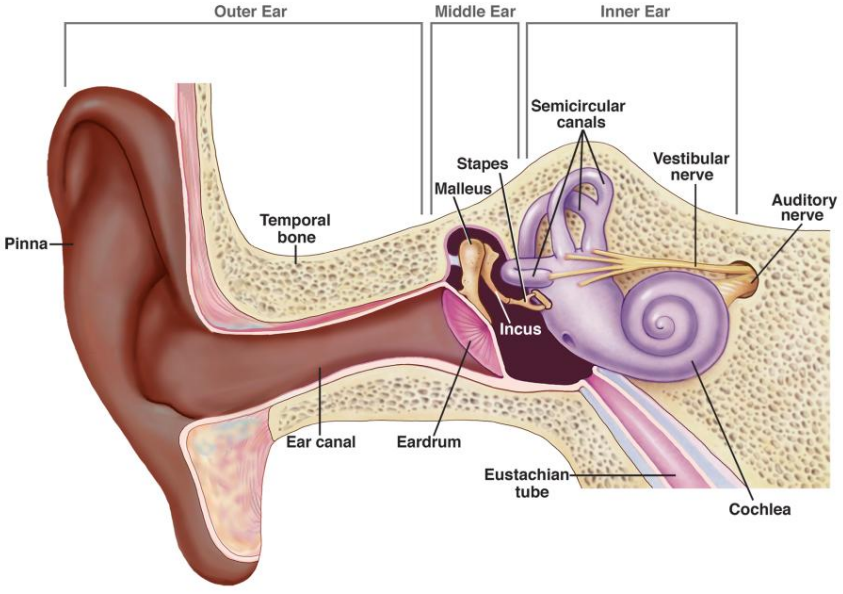


Figure. 2.2. Structure of the Ear (National Institute on Deafness and Other Communication Disorders (NIDCD), 2024).

2.2 Auditory Perception

2.2.1 Perception of Simple Sounds

Plack provides an in-depth explanation of sound and human hearing in the book The Sense of Hearing (Plack, 2018). Human hearing is a complex process that involves detecting sound across a wide range of frequencies and intensities. The mechanism of hearing can be divided into three main stages: sound transmission in the ear (including the outer, middle, and inner ear; see Figure 2.2), signal transduction in the cochlea, and neural processing in the brain (Gazzaniga et al., 2014; Moore, 2012).

The hearing process begins with the outer ear. Sound waves travel past the pinna and through the ear canal. Its resonant properties amplify frequencies in the 200 to 5000 Hz range by approximately 10 to 15 dB (Pickles, 2012). This amplification is particularly useful in human conversation, as typical speech sounds fall within this range (Howard & Angus, 2017).

The tympanic membrane, also known as the eardrum, vibrates as sound waves reach the end of the ear canal. It transmits these vibrations to the middle ear, which contains three tiny bones: the malleus (hammer), incus (anvil), and stapes (stirrup). These bones further amplify the vibrations and transfer them to the oval window of the cochlea (Pickles, 2012).

The cochlea, a spiral-shaped, fluid-filled structure in the inner ear, serves to convert mechanical vibrations into neural signals. Inside the cochlea, the basilar membrane vibrates in response to different frequencies: higher frequencies are processed at the base, while lower frequencies stimulate the apex (Plack, 2018).

Within the cochlea, hair cells located inside the organ of Corti on the basilar membrane contain tiny structures called stereocilia. These bend in response to fluid movement and vibrations, releasing neurotransmitters that transmit signals through the auditory nerve to the brain (Moore, 2012; Plack, 2018).

2.2.2 Auditory Pathway

2.2.3 Central Auditory Processing

Once the brain receives information from the cochlea, it is processed through a system called the auditory pathway. The auditory pathway consists of both ascending (bottom-up) and descending (top-down) pathways that interact with each other (Pickles, 2012). These pathways create a feedback loop that allows for more effective and selective sound processing, enabling humans to adapt their responses to sound more efficiently (Suga & Ma, 2008).

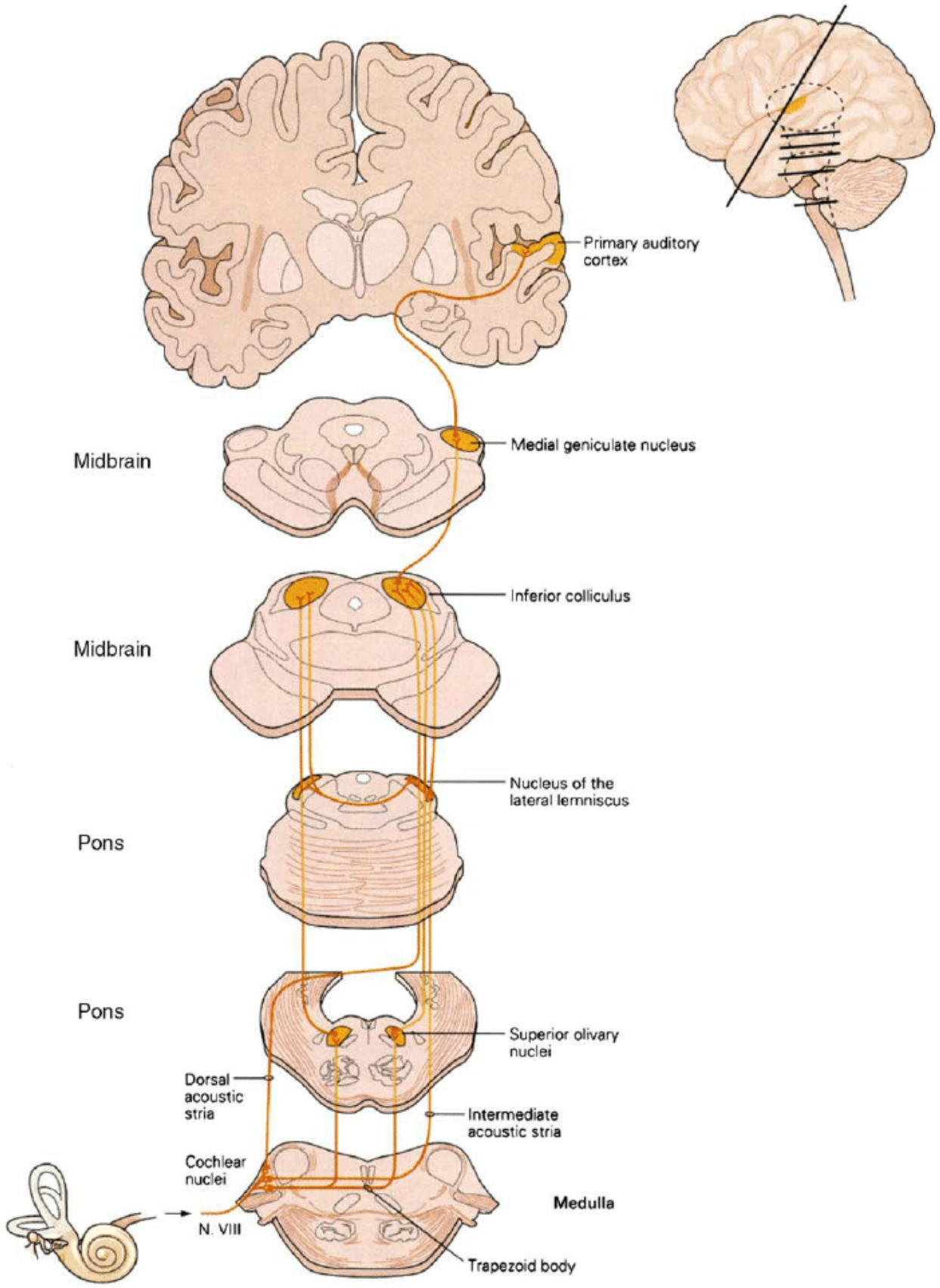


Figure. 2.3. Central Auditory Pathways(Graven & Browne, 2008).

Ascending Pathway The ascending pathway refers to the process by which auditory information travels from the ear to the brain. It begins in the cochlea, passes through the auditory nerve to the brainstem and midbrain, and finally reaches the auditory cortex in the temporal lobe (Moore, 2012), where conscious sound perception occurs (Kandel et al., 2013).

Descending Pathway The brain does not simply receive information from the ears; it also sends instructions through the descending pathways (Suga & Ma, 2008). This provides a feedback loop based on attention, context, and learning, enhancing responses to important sounds while suppressing irrelevant ones (Bajo et al., 2010). For example, this interaction enables us to focus on a teachers voice in a noisy classroom, increasing adaptability and enhancing learning outcomes (Guinan, 2006).

From Neural Processing to Cognitive Functions The knowledge of sound and neural pathways described above provides the foundation for the human listening process. Effective listening involves much more than transmission of auditory signals. Higher order cognitive mechanisms actively shape how human perceive, interpret, and respond to auditory information (Moore, 2012). The following section explains the cognitive mechanisms which makes complex listening experience possible.

2.2.4 Auditory Scene Analysis

Our ears receive multiple sounds from different sources everyday. Humans can focus on a specific sound source, such as a friend's voice in a noisy restaurant, or distinguish the violin playing in an orchestra. This ability does not only rely on our peripheral auditory system, but on a more complex cognitive process known as Auditory Scene Analysis (ASA) (Bregman & McAdams, 1994).

Principles of Auditory Scene Analysis Auditory Scene Analysis (ASA) is the perceptual process by which the human brain processes seemingly chaotic sound environments into meaningful inputs. Bregman's work plays a critical role in explaining and understanding ASA (Bregman & McAdams, 1994). It was proposed that, similar to visual analysis, ASA also unfolds in two stages: segregation, and grouping or organisation.

Segregation and Grouping Segregation refers to the brain's ability to break down complicated sound into simple acoustic elements, such as frequency (pitch), amplitude (loudness), temporal patterns (timing), and location. The brain uses these features to further determine which sound elements need to be separated, and which to be grouped

into one meaningful input for better understanding. For example, sounds that occur closer in timing or frequency tend to be grouped into a single source, whilst sounds from different locations would be interpreted as different sounds.

Different Types of Grouping: Primitive and Schema-Based Processes There are two types of grouping mechanism in ASA: Primitive and schema-based processes (Darwin, 2007). Primitive processes, widely understood as bottom-up processing, are mostly automatic and stimulus-driven. Sound enters the ear and is processed by the cochlea, passed by the auditory nerves to the brain, to analyse basic acoustic features of the sound, such as timing and loudness. It focuses on building and understanding auditory foundation from raw materials.

Schema-based processes, on the other hand, are understood largely as top-down analysis of the brain. They rely more on prior knowledge and cognitive capacity of the brain. For example, even facing the same situation like waiting for the bus on the street, one who has learned what the local bus would sound like would be more likely to notice the bus they are waiting for, compared to someone who is new to the area

In real life, the two cognitive processes cannot be completely separated, which makes researching the process of listening more challenging. The brain takes information provided from the primitive processes, and the schema-based process can influence the primitive one through feedback loops. Taking the previous bus waiting example: when one is expecting the sound from the bus, the brain would enhance the vigilance level for the specific input.

Subconsciousness Feature of Auditory Scene Analysis

A significant portion of ASA is processed automatically, without extra effort or attention of the brain. The basic acoustic feature extraction and grouping operate largely without a conscious level. This arrangement of the brain function is important because humans are constantly facing complex sound mixtures. Unlike visual analysis, we cannot simply "shut our ears" to stop listening. This feature also allows humans to quickly respond to potential threats, such as a sudden loud noise or an approaching car, without constant conscious effort.

2.3 Speech Perception

ASA provides foundation for process complex sound environments, such as speech perception. Speech perception is the process by which humans decode and understand spoken language. It allows humans to transform acoustic features of speech into

meaningful information. While ASA is largely processed subconsciously and automatically, speech perception requires more active attention and cognitive effort to engage in extracting meanings. This leads to the importance of understanding cognitive effort in the process of speech perception and comprehension.

2.3.1 Features of Speech

Speech is highly structured with predictable patterns which make it different from random environmental noise (Liberman et al., 1957). Sounds in speech include both periodic sounds, such as vowels, and non-periodic sounds, such as consonants (Tallal et al., 1993). Speech contains segments with distinct spectral and temporal characteristics (Peterson & Barney, 1952).

Speech is highly hierarchical. At the bottom level, speech consists of the same acoustic features as other sounds, such as frequencies and amplitude. These combine to form phonemes, the smallest sound units which could differentiate meanings in language, such as "b" and "p" in "bat" and "pat" (Liberman et al., 1957).

Phonemes then combine to form syllables, which typically consist of a vowel sound and one or two consonant sounds before and/or after the vowel, such as "cat" in English, or "la" in Italian. Syllables continue to form words, the basic units to transfer meaning (Saffran et al., 1996), and further to form phrases and sentences.

The hierarchical nature of speech illustrated above represents the universal characteristic of human language. For different languages, however, how these elements are structured and combined varies. This variation improves further understanding of human language, and brought challenges when brain tries to engage in a language different from their native tongue.

Speech Across Different Languages The acoustic features of speech vary significantly across different languages. Some sounds exist in certain languages but not others, making speech perception particularly challenging in such situations. The clustered consonants structure in English, such as "string," which consists of "s," "t," "r" at the same time, is generally not permitted in Japanese language, where the syllables consist of only one consonant and one vowel, such as "ka" (Ishida, 2007). The trilled or rolling "r" sound in Italian and Spanish, for instance, does not exist in English (Harris, 1969; Miller, 2006). Another example is tonal languages like Mandarin and Thai, where changing the pitch would change the meaning of the word completely, even with the same consonant and vowel, making learning and understanding such languages especially difficult for English speakers (Yip, 2002).

2.3.2 Speech and Speech-in-noise perception

The hierarchical structure of speech and the differences of speech structure from different languages, mean that speech perception would rely largely on the previous knowledge of the certain speech structure. The hierarchical feature creates redundancies in speech perception. Based on previous knowledge, the human brain can recognise and predict speech based on the context, even when certain sounds are masked by noise. On the other hand, how effectively one can sustain and extract such knowledge varies largely between different individuals. The complexity of speech perception becomes more apparent when comparing to other forms of auditory processing, such as music perception.

In realistic listening environments, background noise often interferes with speech perception. Such conditions, known as speech-in-noise (SiN), challenge the auditory systems capacity to isolate relevant signals from irrelevant acoustic input (Kalikow et al., 1977). SiN tasks provide a controlled way to examine how listeners manage degraded or masked speech.

Listeners must engage attentional and memory resources to compensate for missing or distorted information. Two key types of masking are typically involved: energetic masking, where the target signal is acoustically buried beneath the noise, and informational masking, where competing speech or noise draws on similar linguistic processing pathways (Brungart, 2001; Mattys et al., 2012). These effects are more pronounced at lower signal-to-noise ratios (SNRs), which increase the perceptual and cognitive load (Zekveld et al., 2010).

2.4 Cognitive Mechanisms in Listening

2.4.1 Attention and Selective Attention

Attention is a cognitive function that enables individuals to focus on specific stimuli while filtering out irrelevant information. It is crucial for learning, perception, memory, and decision making (Posner & Petersen, 1990). Attention can be understood as different types, including selective attention, sustained attention, and divided attention.

Selective Attention Selective attention refers to the ability to choose and focus on one source of information while filtering out irrelevant noise (Suga & Ma, 2008). The "cocktail party effect", where one can concentrate on a single conversation with competing background noise, demonstrates how selective attention works effectively (Cherry, 1953).

Sustained Attention Sustained attention, or vigilance, is the ability to maintain a certain level of focus on a task over time. It is particularly important with tasks that require long-term vigilance, such as driving. Studies suggest that sustained attention relies on neural circuits involving the prefrontal cortex and parietal lobes (Posner & Petersen, 1990), and varies largely between different individuals (Esterman et al., 2014; Petersen & Posner, 2012).

Divided Attention Divided attention, on the other hand, is commonly known as multitasking. It involves processing multiple information sources simultaneously. While the human brain is able to manage several tasks at once, this distribution of resources often leads to a decline in performance and increased error rates (Pashler, 1994).

From Attention to Executive Processes While attention directs cognitive resources toward specific stimuli, successful listening in complex environments - such as on a crowded train - requires much more than attentional mechanisms. Executive functions work alongside attention to support efficient listening and comprehension (Diamond, 2013).

2.4.2 Executive Function

Executive function is a set of cognitive processes that enable individuals to manage tasks effectively. These functions are essential for goal-directed behaviour and self-regulation (Miyake et al., 2000). Executive function primarily relies on prefrontal cortical networks, though it engages distributed systems throughout the brain. Core components of executive function include inhibitory control, working memory, and cognitive flexibility (Diamond, 2013).

Inhibitory Control Inhibitory control is a fundamental component of executive function that allows the human brain to suppress irrelevant or habitual responses in order to engage in goal-directed behaviour (Diamond, 2013). It is primarily governed by the prefrontal cortex, particularly the right inferior frontal gyrus (rIFG), which is responsible for suppressing inappropriate responses (Aron et al., 2011).

It involves several distinct mechanisms:

- **Perceptual Inhibition:** Suppressing attention from distracting sources, allowing humans to focus on task-relevant information (Miyake et al., 2000).
- **Cognitive Inhibition:** Suppressing irrelevant thoughts, outdated information, or intrusive memories, preventing them from interfering with current cognitive tasks (Hasher & Zacks, 1999).

- **Response Inhibition:** The ability to suppress automatic behavioural responses. This is crucial for impulse control and delayed gratification (Diamond, 2013).

In auditory perception, inhibitory control enables humans to filter out background noise, resist processing meaning for irrelevant speech, and refrain from interrupting speakers despite having immediate thoughts or reactions (Diamond, 2013). Although inhibitory control shares conceptual and neural overlap with *attention*, the key difference is that inhibitory control is primarily suppressive while attention is primarily attentive. In human development, mature inhibitory control develops much later than certain aspects of attention (Diamond, 2013).

Working Memory in Listening The effectiveness of inhibitory control is closed linked to another executive function: working memory. Inhibitory control depends on working memory to maintain task-related goals while suppressing distractions (McNab & Klingberg, 2008). This relationship is important for effective auditory processing, where in real life, listeners must hold multiple speech content in working memory while simultaneously filtering out unwanted information (McNab & Klingberg, 2008).

Working memory is a cognitive system that temporarily holds and manipulates information for further complex tasks, such as reasoning, learning, and comprehension (Baddeley, 1992). It has a limited capacity, where most healthy adults can hold up to 4 to 7 components simultaneously.

Proposed by Braddeley and colleagues, the multicomponent model of working memory provided a framework for understanding how brain process information. This model identifies several components, each plays different roles in auditory processing, as illustrated in Figure 2.4):

- **Central Executive:** The supervisory system that controls attention and coordinates information.
- **Phonological Loop:** Processes verbal and auditory information. It holds, rehearses and refreshes the information.
- **Visuospatial Sketchpad:** Stores and manipulates visual and spatial information.
- **Episodic Buffer:** Integrates information and connects working memory with long-term memory.

In listening tasks, the *phonological loop* temporarily stores information such as numbers or names, while the brain processes for meaning. The *central executive* focuses attention on critical information, the *visuospatial sketchpad* may create mental images about the conversation, and the *episodic buffer* may connect this new information with existing knowledge to further process meaning.

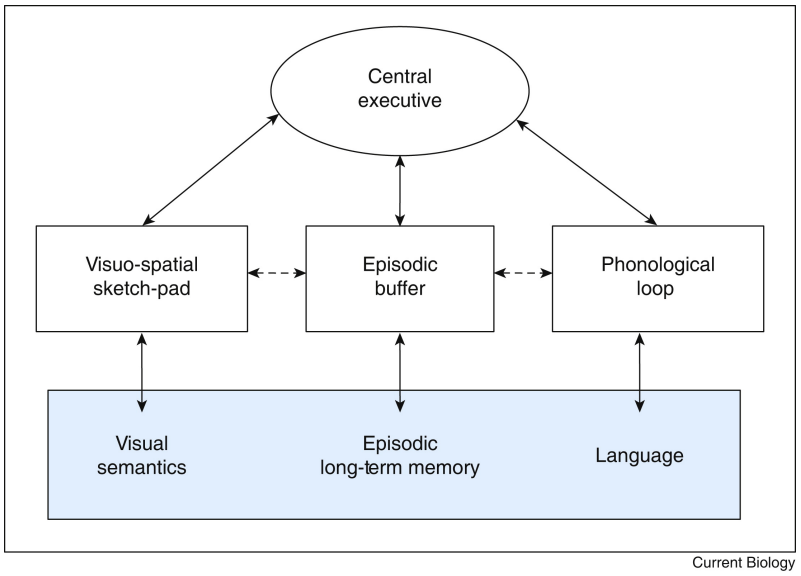


Figure. 2.4. Multicomponent Model of Working Memory (Graven & Browne, 2008).

Cognitive Flexibility Working memory and cognitive flexibility are closely linked, as working memory maintains and updates relevant information to enable the brain to switch between different tasks (Miyake et al., 2000). Cognitive flexibility refers to the mental ability to adapt to new situations, switch between tasks, and update thinking in response to changing environments (Monsell, 2003). It is the ability to adjust our perspective, approach, and thinking patterns when circumstances change (Diamond, 2013).

In listening tasks, cognitive flexibility plays important roles, particularly in dynamic or complex auditory environments. In daily conversation, it allows the brain to follow multiple speakers, switch attention between different voices and topics. When listening conditions change, like walking from a quiet path into a noisy restaurant, cognitive flexibility allows rapid adaptation of listening strategies and management of background noise.

The cognitive mechanisms - attention, inhibitory control, working memory, and cognitive flexibility - function as an integrated system that supports the brain's ability to navigate complex auditory environments. These mechanisms do not operate alone but work as an integrated function to support effective listening.

2.5 Hearing Impairment and Broader Consequences

2.5.1 Age-related and Sensorineural Hearing Loss

Hearing loss is a prevalent sensory impairment, particularly among older adults. One of the most common forms is age-related hearing loss, also known as presbycusis, which typically affects high-frequency sounds and gradually progresses over time (Gates & Mills, 2005; Plack et al., 2014). This form of hearing loss is often sensorineural in nature, resulting from the degeneration of hair cells in the cochlea, the auditory nerve, or both (Moore, 2012).

Sensorineural hearing loss not only affects audibility but also impairs temporal and spectral resolution. Even when sounds are made loud enough through amplification, individuals with this type of hearing loss may still struggle with clarity, especially in noisy environments (Peelle et al., 2011; Tremblay et al., 2003). This is partly due to disrupted phase locking and decreased frequency selectivity in the auditory system, leading to poor representation of fine-grained speech features (Hopkins & Moore, 2008).

Furthermore, hearing loss is associated with changes in central auditory processing. Studies using neuroimaging have shown that individuals with hearing impairment may exhibit reduced activation in auditory cortices and increased recruitment of frontal areas, suggesting a shift towards compensatory cognitive strategies (Campbell & Sharma, 2016; Peelle et al., 2011). These changes may also reflect increased listening effort and cognitive load in everyday communication situations.

Although hearing aids and cochlear implants can provide access to auditory input, they may not fully restore natural processing. Residual deficits in spatial hearing, sound localisation, and speech-in-noise perception often remain (Choi et al., 2011). Therefore, sensorineural hearing loss presents challenges not only to auditory perception but also to broader cognitive functioning and quality of life.

2.5.2 Hearing Loss and Dementia Risk

Recent years have seen growing interest in the relationship between hearing loss and cognitive decline. Epidemiological studies have consistently reported that individuals with hearing loss are at significantly increased risk of developing dementia, including Alzheimer's disease (Lin et al., 2011; Livingston et al., 2020). Hearing loss has been identified as a potentially modifiable risk factor, with estimates suggesting that addressing hearing impairment could delay or prevent up to 9% of dementia cases worldwide (Livingston et al., 2020).

Several mechanisms have been proposed to explain this association. One hypothesis is the *cognitive load* theory, which suggests that hearing loss increases the mental effort

required for auditory processing, thereby diverting resources from other cognitive operations such as memory encoding or executive control (Peelle, 2014; Tun et al., 2009). Over time, this increased effort may lead to accelerated cognitive fatigue and decline.

Another explanation is *sensory deprivation*, where reduced auditory input leads to structural and functional changes in the brain. Longitudinal studies have shown that individuals with hearing loss exhibit greater atrophy in auditory and frontal brain regions compared to those with normal hearing (Lin et al., 2013; Peelle et al., 2011). The reduced stimulation may compromise neuroplasticity and limit engagement in cognitively enriching activities.

Social isolation may also play a contributing role. Hearing loss can reduce participation in conversations and social activities, which are known protective factors against cognitive decline (Mick et al., 2014). The combination of cognitive effort, sensory decline, and reduced social interaction creates a multifactorial pathway linking hearing impairment to dementia.

Understanding these interactions is crucial for developing interventions that extend beyond hearing aid provision alone. Comprehensive strategies addressing both auditory and cognitive health are increasingly viewed as essential in promoting healthy ageing and mitigating dementia risk.

The previous sections have outlined the fundamental components of the human auditory system, from the physical properties of sound to the neural and cognitive mechanisms supporting complex listening behaviours. In the next chapter, we will examine how listening effort is defined and measured, the conditions that increase it, and why it matters for individuals with and without hearing loss.

Chapter 3

Listening Effort

3.1 Definition of Listening Effort

3.1.1 Historical Definitions Focus

The early works on listening effort mainly focused on understanding it as cognitive resource allocation for perceiving and understanding auditory information. In Downs' work, listening effort was defined as "the allocation of additional attentional resources to auditory tasks" (Downs, 1982). Similarly, Feuerstein defined listening effort as "the attention and cognitive resources required to understand speech" (Feuerstein, 1992).

3.1.2 Current Definitions of Listening Effort

Compared to earlier research which understood listening effort as cognitive exertion due to unsatisfactory auditory input, contemporary research has developed more comprehensive frameworks for listening effort, viewing it as a complex interaction of cognitive, physiological, and environmental factors.

The view of working memory shifted from seeing listening effort as competing for resources (Downs, 1982; Rabbitt, 1968), into recognising working memory as a supporting role in challenging listening situations (Peelle, 2018; Winn & Moore, 2018). Modern definitions also integrate motivation into understanding listening effort (Hicks & Tharpe, 2013; Pichora-Fuller et al., 2016a), recognising the importance of motivation in performing a listening task. Further, recent definitions of listening effort attempt to embed within established frameworks, such as capacity theory (Kahneman, 1973) and FUEL (Pichora-Fuller et al., 2016c), rather than being mainly descriptive. Table 3.1 illustrates the progression of definitions of listening effort.

Aspect	Early Definitions (1960s-1990s)	Modern Definitions (2000s-Present)
Primary Focus	Defined as *increased cognitive load* due to degraded speech (Anderson-Hsieh & Koehler, 1988; Rabbitt, 1968). Characterized as "the allocation of additional attentional resources" (Downs, 1982) and "the attention and cognitive resources required to understand speech" (Feuerstein, 1992).	A *multidimensional process* involving *cognitive, motivational, and physiological* factors (McGarrigle et al., 2014; Pichora-Fuller et al., 2016c). Defined as "a multidimensional construct encompassing the cognitive, motivational, and emotional resources deployed during auditory tasks" (Ohlenforst et al., 2017).
Working Memory	*Listening effort competes with working memory*, reducing resources for information retention (Downs, 1982; Rabbitt, 1968).	*Working memory actively compensates* for auditory challenges, improving speech comprehension (Peelle, 2018; Winn & Moore, 2018). Integrated into frameworks like Ease of Language Understanding (ELU) (Rönnberg et al., 2013), where "explicit working memory resources are brought into play" when bottom-up and top-down information do not match.
Motivation	*Not considered a key factor* (Anderson-Hsieh & Koehler, 1988).	*Motivation influences effort allocation*; listeners exert more effort when engaged (Hicks & Tharpe, 2013; Pichora-Fuller et al., 2016c). FUEL framework emphasises motivation and explains listening effort as: "the deliberate allocation of mental resources to overcome obstacles in goal pursuit" (Pichora-Fuller et al., 2016b).
Adaptability	Effort was viewed as *static* and *task-dependent*, varying only with listening difficulty (Downs, 1982).	*Effort is flexible*; listeners *adjust cognitive resources* based on goals and task demands (Mattys & Wiget, 2009; McGarrigle et al., 2014). "The application of cognitive resources to overcome obstacles... reflecting both task demands and the motivation of the listener" (Lemke & Besser, 2016).
Theoretical Foundation	Often *lacked explicit theoretical grounding* in cognitive science; more descriptive than explanatory.	*Embedded in broader cognitive frameworks* like capacity theory (Kahneman, 1973) and explicitly developed frameworks like FUEL (Pichora-Fuller et al., 2016b).

Table. 3.1. Comparison of Early and Modern Definitions of Listening Effort

3.2 Theoretical Models of Listening Effort

The development of listening effort definitions has led to more comprehensive frameworks in understanding this concept. The following section examines several key theoretical models in understanding listening effort.

3.2.1 Kahneman's Capacity Model of Attention and Listening

Developed in the early 1970s, Kahneman's Capacity Model of attention has been highly influential in understanding listening effort (Kahneman, 1973). This model is based on the key principle of the limited capacity of attention and a single resource pool of mental capability. According to Kahneman, cognitive resources constitute one general resource pool, and can be distributed based on factors including task demands and arousal level (see Figure 3.1).

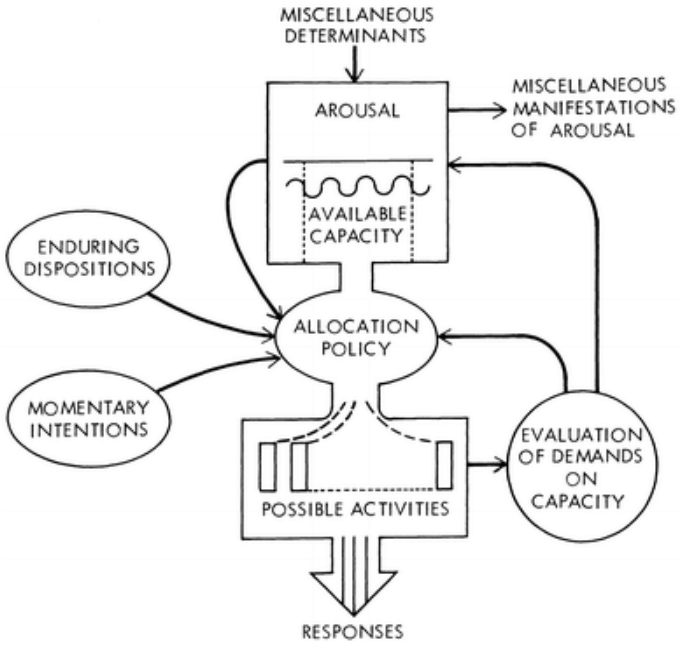


Figure. 3.1. Kahneman's Capacity Model Theory (Kahneman, 1973).

In understanding listening effort, Kahneman's model provides insight into why listening effort varies across different situations:

- **Resource allocation:** When listening conditions become challenging, more attentional resources must be allocated for basic perceptual processing, thus fewer resources are left for higher processes such as extracting meaning and comprehension.

- **Arousal and effort:** Including arousal in this model establishes listening effort as a dynamic rather than static process. It not only explains why listening effort may vary based on motivation (increased arousal) and fatigue (lack of arousal), but has laid the theoretical foundation for measuring listening effort through **physiological responses**.
- **Individual differences:** This model can account for differences in listening effort among individuals with similar levels of hearing ability, as the differences in cognitive capacity and cognitive resource allocation strategy vary between individuals.

While Kahneman's model provides a valuable framework, researchers have identified limitations when investigating listening effort. Especially, research suggests that rather than a single pool of cognitive resources, there exist multiple resources for different tasks (Pichora-Fuller et al., 2016c). Kahneman's framework has been extended through the Framework for Understanding Effortful Listening (FUEL) (Pichora-Fuller et al., 2016b), and the Ease of Language Understanding (ELU) model (Rönnberg et al., 2013), both of which elaborate on how perceptual and cognitive factors interact during speech processing.

3.2.2 Framework for Understanding Effortful Listening (FUEL)

The FUEL framework (Figure 3.2) represents an integration of Kahneman's model (page 23) with modern understanding specifically for listening and hearing. While Kahneman's Capacity Model was designed for understanding general cognitive processes, FUEL was tailored specifically for auditory processing. It explains how various factors, such as cognitive resources, motivation, and task demand, influence the effort required for listening tasks (Pichora-Fuller et al., 2016b).

FUEL further emphasises the role of motivation and arousal in modulating available cognitive resources. It places motivation at the centre of its model, emphasises how factors like reward or importance of success directly influence effort allocation. Individual differences in performing a listening task are addressed in FUEL. It accounts for how individual factors, such as age, hearing status, and cognitive abilities influence both available capacity and allocation strategies.

FUEL offers a further explanation of challenges faced by people with hearing impairment, with possible rehabilitation approaches. it accounts not only the level of hearing loss, but incorporated social-cognitive dimensions. It explains that in social situations, motivation for connection may override discomfort one experiencing comparing to other similar acoustic challenging environments, which leads to significantly increased effort and frustration (Pichora-Fuller, 2017). The clinical

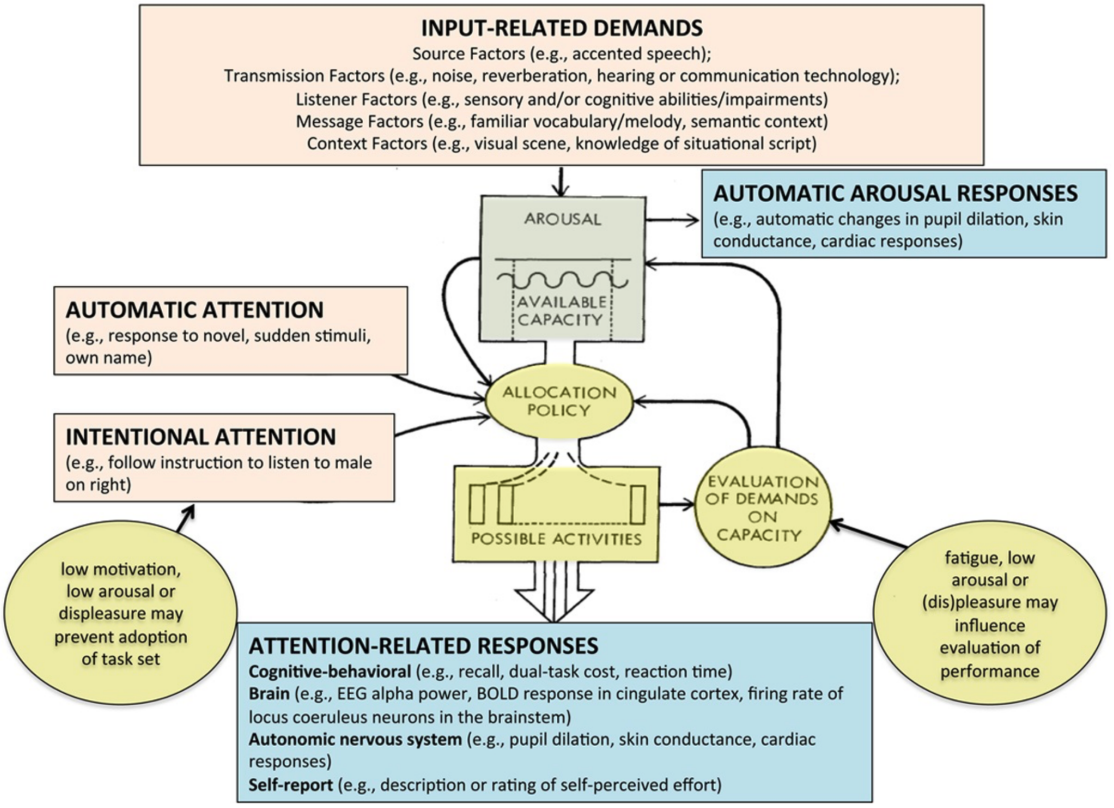


Figure. 3.2. FUEL: Expanding Kahneman's Capacity Model in Relation to Listening Effort and Fatigue (Pichora-Fuller et al., 2016b).

FUEL further expanded Kahneman's model in understanding cognitive effort which tailored for listening effort specifically. It emphasises the role of motivation and arousal in auditory processing, building a feedback loop based on importance and fulfillment of the task.

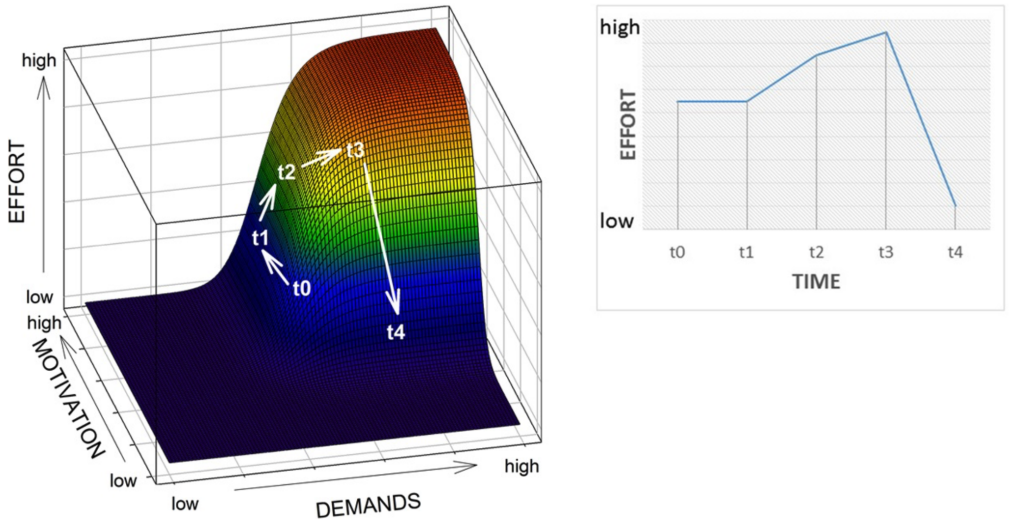


Figure. 3.3. Relationship between Demands, Motivation, and Effort (Pichora-Fuller et al., 2016b).

This 3D plot shows how effort changes based on task demand and motivation. A path on the figure shows how effort changes overtime when affected by these factors. no specific units were provided for this plot.

implications suggests that besides improving audibility, a broader approach which considers cognitive costs of listening, such as training protocols which addresses attentional control, would greatly supporting the effectiveness of listening.

3.2.3 Ease of Language Understanding (ELU) Model

The *Ease of Language Understanding* (ELU) model, proposed by Rönnberg and colleagues (Rönnberg et al., 2013), took a different perspective. It puts an emphasis on speech perception specifically rather than attention in FUEL and Kahneman's model. Rooted in theory of working memory, ELU explains how individuals with stronger working memory capacity can better compensate for degraded auditory input.

One key element of the ELU model is its integration of bottom-up and top-down processes in listening (see page 9), and the importance of working memory. The bottom-up processing encodes the acoustic features of speech, matches these features against stored phonological library in long-term memory. When it matches, understanding occurs. However, when they don't match, the top-down process is engaged to compensate.

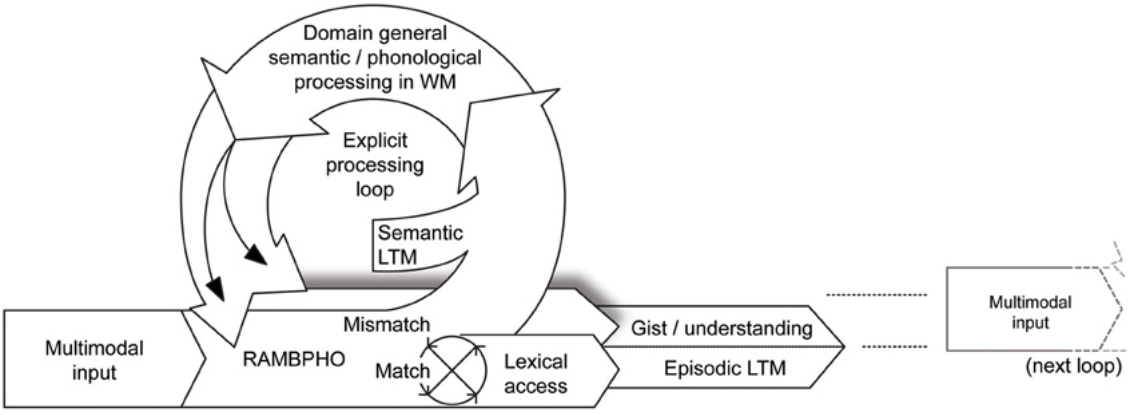


Figure. 3.4. The Ease of Language Understanding Model (Rönnberg et al., 2013).

ELU model explains the process of speech perception. In ideal listening situations, auditory input aligns well with stored phonological patterns, allows quick and automatic recognition. However, in challenging listening circumstances, when there is a mismatch, working memory (WM) would step in to assist and fill in the gaps for missing information. Listening effort increases when needs constant assistance from working memory.

Relationship Between Theoretical Models. These models, while developed with different emphases, provide important perspectives that help us understand the nature of listening effort. Kahneman's Capacity Model proposes a limited pool for cognitive resources, yet lacks the emphasis on motivation and didn't address listening specifically. FUEL expands this model by explicitly incorporating motivation and value for reward in

understanding the listening process specifically. The ELU model, on the other hand, provides additional understanding of the role of working memory in speech perception.

Together, these frameworks create a more complete picture of listening effort. It's not merely a response to demanding auditory signals, but an active cognitive process influenced by multiple elements. More importantly, as suggested by FUEL, rather than thinking of listening effort as an issue to overcome, optimal effort is necessary for task performance. The difference, perhaps, lies in the optimal level of effort required to complete the current listening task. What complicates matters is that this optimal level, would be different for each individual.

3.3 Elements Affecting Listening Effort

The journey from sound waves to meaningful auditory perception forms the foundation for understanding listening effort. Based on previous explanations, this chapter examines the key elements which affect listening effort, including environmental, signal, and individual factors. It explores why understanding this concept is important and sets the stage for measurement approaches that will be presented in the next chapter.

3.3.1 Hearing Impairment and Listening Effort

Types of Hearing Loss

Sensorineural Hearing Loss There are different types of hearing loss, each brings different challenge for hearing impairment individuals. Sensorineural hearing loss, which caused by damage to the cochlea or auditory nerve, often involves reduced frequency selectivity and temporal resolution. It typically have permanent effects on both audibility and distorted signal processing (Hornsby, 2013).

Conductive Hearing Loss Conductive hearing loss, which results from malfunctions in the outer or middle ear which affects sound reaching the inner ear, often involves physiological blockage or damage to the structures that conduct sound. Comparing to Sensorineural hearing loss, it usually involves less listening effort once sounds are made audible enough (Downs, 1982).

Mixed Hearing Loss The more complicated cases are mixed hearing loss, which combines both sensorineural and conductive hearing loss, making addressing the issue more challenging due to both attenuation and distortion of sound (Dillon, 2012). Standard treatment such as hearing aids, can compensate for the aspect of conductive loss, but cannot fully address the sensoroneural distortion effects, especially in noisy environments (Moore, 2012).

Hidden Hearing Loss (HHL) Hidden Hearing Loss (HHL) adds another fascinating layer of hearing loss which is difficult to address in clinical settings. It represents a relatively recent advancement in understanding auditory processing deficits which conventional audiometry cannot detect. Individuals appear to be "normal" in standard pure-tone audiometry, nonetheless experience significant difficulties understanding speech in noisy environment (Plack et al., 2014).

HHL is typically related to synaptopathy, or damage of synapses between auditory nerve fibres and inner hair cells of the cochlea. It first emerged from animal studies which found that noise exposure could cause permanent loss of synaptic connections between inner hair cells and auditory nerve fibres without affecting audiometric thresholds (Kujawa & Liberman, 2009). It revealed that up to 50% of auditory nerve fibres could be damaged while still appear to have normal hearing for pure tone threshold tests (Liberman & Kujawa, 2016).

Despite the difficulty of detecting HHL in standard audiometry, HHL can be measured through electrophysiological measures. The amplitude of Wave I (first peak) in the Auditory Brainstem Response (ABR), which reflects activity from auditory nerve fibre, is found reduced in individuals who were exposed to long-term noise (Stamper & Johnson, 2015). Current research aims to develop and validate clinical tests which could detect HHL, which would improve early diagnosis and interventions (Guest et al., 2018; Plack et al., 2014).

Other Types of Hearing Loss Other types of hearing loss include Central Auditory Processing Disorder (CAPD), where individuals have deficient sound localisation mechanisms (Iliadou et al., 2017), and Auditory Neuropathy Spectrum Disorder (ANSD), where outer hair cells in the cochlea function normally in responding to frequencies but fail to align timing when passing information to the brain. This disruption in ANSD distorts the temporal resolution of sound and leads to increased listening effort (Rance, 2008).

Relationship between Hearing Loss and Listening Effort Hearing loss introduces diverse challenges that typically increase the listening effort required for speech comprehension. While some forms, like conductive hearing loss, primarily reduce the sound level reaching the inner ear and may impose less effort once audibility is restored (Downs, 1982), other types involve more complex processing deficits. Sensorineural hearing loss, for instance, often impairs frequency and temporal resolution, leading to signal distortion that necessitates greater cognitive resources for interpretation, even if the sound is loud enough (Hornsby, 2013). Mixed hearing loss combines these issues of attenuation and distortion, making compensation particularly demanding (Dillon, 2012).

Furthermore, conditions like Hidden Hearing Loss (HHL), potentially caused by cochlear synaptopathy (Kujawa & Liberman, 2009), can lead to significant difficulties and effort in noisy environments despite normal audiograms (Plack et al., 2014), highlighting limitations in standard hearing tests. Other deficits, such as Central Auditory Processing Disorder (CAPD) affecting sound localisation (Iliadou et al., 2017) or Auditory Neuropathy Spectrum Disorder (ANSD) impacting temporal processing (Rance, 2008), also distort auditory information in ways that demand increased listening effort.

Therefore, the type and nature of the hearing impairment significantly shape the cognitive burden associated with listening.

3.3.2 Other Listener-Related Factors

Research has traditionally focused on understanding how the average population reacts to difficult listening situations, aiming to identify external factors that account for increased listening effort. Even within studies designed to capture average responses, however, results demonstrate large individual variability (Koelewijn et al., 2012; Zekveld et al., 2011). These individual variation has prompted a growing interest in examining listener-specific factors in listening effort (Peelle, 2018). Better understanding of these individual differences would support developing personalised approaches to address their unique needs (McGarrigle et al., 2014; Pichora-Fuller et al., 2016b).

Age Aging is typically linked to increase risk for hearing loss. However, it affects listening effort independently from hearing status. Functional neuroimaging studies reveal that older adults show greater activation in prefrontal cortical regions comparing to younger listeners, even when controlling hearing sensitivity (Eckert et al., 2008; Peelle et al., 2010). This increased prefrontal cortical engagement correlates with successful speech comprehension, but at a cost for more cognitive resources (Wong et al., 2009).

Beyond changes in neural activation patterns, cognitive changes associated with aging also contribute to increased listening effort. The executive functions typically decline as one age, which adds to the increased effort in listening. Older adults generally experience declining temporal processing abilities which affects speech perception, especially for fast speed speeches (Anderson et al., 2012). Reduced inhibitory control,,which typically occurs as a symbol of cognitive aging, diminishes the ability to suppress irrelevant information, making speech-in-noise particularly challenging and effortful (Sommers & Huff, 2015).

Cognitive Factors

As introduced in Section 2.4, page 14 exploring the role of cognitive mechanisms in listening, individual differences in cognitive functions significantly influence the listening process and the result of listening task performance.

Attention and Inhibitory Control The ability to focus on certain stimulus while suppressing irrelevant information, varies among individuals which affects listening effort. Listeners with stronger inhibitory control, shows different patterns of neural resources allocation during challenging listening tasks, resulting in improved speech

processing (Sommers & Huff, 2015; Zekveld et al., 2012). With noisy backgrounds, specifically, studies show that people with better inhibitory control, tested through classic Stroop tasks, tend to understand speech better in noisy settings (Dryden et al., 2017).

Processing Speed Individual differences in processing speed result from interactions of neurological, developmental, genetic factors, and experiences. Processing speed refers to how quickly listeners can process and understand auditory information. Research found constant correlation between processing speed and speech comprehension under challenging listening conditions, even after controlling other cognitive factors (Gordon-Salant et al., 2014). This relationship becomes particularly obvious when processing compressed or rapid speech (Schneider et al., 2005).

Neuroimaging studies provide insights of the neural basis of processing speed. individuals with faster processing speed demonstrate more efficient neural resource allocation, showed more activation in core language regions including superior temporal gyrus ad left inferior frontal gyrus. As listening conditions become more challenging, slower processors showed greater increases in frontal lobe, than individuals with faster processing speed, suggesting an increased effort performing the task (Peelle et al., 2010).

Working Memory Capacity Individual differences in working memory significantly affect how one performing a listening task and the result of speech perception. Listeners with greater capacity often demonstrate improved results in speech comprehension, especially in difficult listening conditions (Lunner, 2003; Rudner et al., 2012). The ELU model specifically highlights the importance of working memory when processing speech (see page 26).

The relationship between working memory and listening effort is more complex than a simple linear relationship. When measuring listening effort with pupillometry, studies have shown that individuals with higher working memory capacity may actually exert more effort, potentially reflecting their greater ability to sustain attention rather than disengaging when conditions become difficult (Wendt et al., 2018). It suggests that the benefits brought by larger working memory capacity maybe modulated by resource allocation strategy (Pichora-Fuller et al., 2016c).

Task Switching Task switching represents another critical executive function that influences listening effort, particularly in dynamic listening environments. Task switching refers to the ability to shift attention between different activities or mental tasks. This executive function plays an important role in auditory processing, especially when involving multiple speakers (Monsell, 2003). Evidence from research demonstrates that, the ability to switch task efficiently can predict performance in complex listening environments (Getzmann et al., 2015).

Psychological Factors

Motivation Motivation refers to the internal process that initiate, guide, and sustain goal directed behaviour (Atkinson, 1964). It varies among individuals based on the perceived importance and achievement motivation (Picou & Ricketts, 2016). As described in FUEL (see page 24) which emphasises the role of motivation, individual differences in motivation when responding to listening challenges significantly affect how much effort to be allocated to listening tasks (Pichora-Fuller et al., 2016c).

Studies support the importance of motivation, demonstrate that the level of willingness to exert effort in communication, predicts real-world hearing aid use and satisfaction beyond audiometric factors (Picou & Ricketts, 2016). Research manipulated task importance through rewarding system show that higher motivation can temporarily overcome the negative effects of acoustic challenges or fatigue, when listeners show sustained effort and engagement under compromising conditions (McGarrigle et al., 2014).

Personality Traits Personality traits create systematic patterns in how individuals engage with listening challenges, influencing both subjective effort perception and objective physiological responses. How individuals engage in listening tasks varies in both subjective effort perception, and objective physiological responses which is affected through the arousal level of individual. Traits like introversion-extroversion and anxiety sensitivity, were particularly examined in relationship with listening effort.

The introversion-extraversion dimension affects listening effort through higher level or arousal resulted from heightened sensitivity. Studies show that introverts typically show greater sensitivity to sensory inputs (Geen, 1984), renders higher baseline arousal level, and may lead to greater distress and effort in noisy environments (Hockey, 2013). They show a high level of GSR during challenging listening tasks comparing to extraverts (Mackersie et al., 2015).

Anxiety sensitivity impact listening experience through heightened vigilance level during difficult listening situations. Measured through pupil diameter, anxious individuals tend to demonstrate higher level of vigilance during uncertain listening conditions (Zekveld et al., 2011), and sustain effort allocation and vigilance, even when speech becomes nearly intelligible (Koelewijn et al., 2018).

Experiential Factors Lifetime experiences-particularly those involving specialised listening activities-can significantly shape how individuals experience and manage listening effort. Certain professionals, such as musicians and simultaneous interpreters, developed higher skills in listening which may affect their experience in listening effort. Research demonstrate that musicians developed higher level of speech-in-noise

perception, and potentially more efficient when processing complex auditory signals (Parbery-Clark et al., 2011). This advantage appears to be affected by the duration and intensity of musical training, suggests neural plasticity and possible rehabilitation direction for hearing impairment.

3.3.3 Contextual Factors

Environmental Acoustics Room acoustics would largely affect the intelligibility of speech. Reverberation, in particular, would significantly increase listening effort by degrading listening signals. It affects the quality of speech by creating temporal smearing of speech sounds and blurring boundaries between phonemes and syllables, which increases effort applied in listening tasks (Rennies et al., 2014).

Spatial configuration, where physical arrangements of sound sources, would affect listening effort in a way that brain is better to distinguish sounds when decide that sounds are coming from difference resources (Best et al., 2018). This benefit of reducing listening effort by segregating space cues, however, appears to decrease with age and hearing loss (Gallun et al., 2013).

Visual Cues Visual cues, especially facial movements and lip patterns during speech production, provide another factor affecting listening effort. In noisy environment, visual cues become particularly valuable. The McGurk effect, for example, demonstrates how visual information (lip movements) influences what we hear. When the brain hear the sound /ba/ but see a speaker mouthing /ga/, your brain combines the two and perceives /da/ instead (McGurk & MacDonald, 1976). This effect persists, even when participants are aware of the illusion (Tiippana, 2014).

Social Context of Listening. The social context of communication would significantly influence listening effort. When one values the connection and acceptance by others in the conversation, listening effort would typically increase (Munro & Derwing, 2006). When the situation was perceived as high-stake, for example, an interview, or presentation, listeners demonstrate greater willingness to make and sustain effort despite increased effort and fatigue (Picou et al., 2011).

3.3.4 Signal-Related Factors

While contextual factors play an important role in the listening experience, the acoustic signal itself presents challenges that directly affect speech perception.

Signal-noise-ratio Level SNR is an important concept in understanding speech perception and listening effort. It represents the relationship between the level of desired signal and the level of background noise. A positive SNR means that speech is louder than background noise, and a negative SNR indicated that the noise is stronger than speech, making comprehension difficult. When SNR gets lower, listening effort would increase dramatically, requiring more attention and cognitive resources.

SNR level is frequently used in research to manipulate different listening conditions. The relationship between SNR level and listening effort, however, is not linear. When equalising listening effort to reaction time, it was found that at very poor SNR levels, listeners may actually reduce their effort as they feel increasingly furtile in performing the task. It creates an inverted U-shaped function where effort /reaction time peaked at moderate SNR levels rather than the most difficult ones(Wu et al., 2016).

Signal Degradation Signal degradation includes various types of acoustic distortion other than increasing background noise which can influence listening effort. Spectral degradation, for example, refers to the reduction of frequency resolution in speech signals, which occurs in both cochlear hearing loss, and paradoxically occurs in cochlear implants and hearing aids-devices intended to enhance auditory perception. As both processes divide the speech spectrum into a limited number of frequency bands (Friesen et al., 2001; Strelcyk & Dau, 2009). Research demonstrated listeners exhibit increased pupil dilation, indicating greater cognitive effort, when processing speech with compromised resolution (Winn et al., 2015).

Beyond spectral degradation, signal compression introduces additional challenges by altering the natural amplitude envelope of speech (Jenstad & Souza, 2003). Research found that even when intelligibility remained relatively high, aggressive compression ratios increased listening effort measured through dual-task paradigm (Arehart et al., 2013). This effect is particularly high in complex acoustic environments and for older listeners (Souza & Arehart, 2015).

3.4 Measurements of Listening Effort

In the previous chapters, we explored how the listening process occurs and the various factors that influence humans' ability to process sound and speech. We also introduced the concept of listening effort, including different frameworks that attempt to explain it and the importance of measuring listening effort.

In this chapter, we examine the different methodologies that research has adopted to quantify listening effort. These measurements range from subjective and behavioural assessments to more recent physiological approaches. Each method offers additional insights into listening effort.

3.4.1 Self-Report Measures

Subjective reports remain the most direct approach to assessing listening effort. This method is widely used in both experimental and clinical settings.

Different Questionnaires Several questionnaires are frequently used to assess listening effort. The NASA Task Load Index (NASA-TLX) is one of the most commonly employed tools for measuring subjective cognitive load (Hart & Staveland, 1988). However, it was not specifically designed for measuring listening effort. Research in this area has attempted to focus on specific elements or adapt the original questionnaire to better capture listening effort.

The Speech, Spatial and Qualities of Hearing Scale (SSQ) was developed for understanding listening experience (but not limited to listening effort) with items asks questions in listening effort (Gatehouse & Noble, 2004). This questionnaire is more sensitive and comprehensive which picks up on different aspects of listening (Jensen et al., 2016). The Abbreviated Profile of Hearing Aid Benefit (APHAB), on the other hand, addresses listening effort from a clinical perspective, focusing on communication difficulties (Cox & Alexander, 1995).

For studies specifically focusing on listening effort, the Effort Assessment Scale (EAS) was developed to target the cognitive dimensions of listening. It asks participants to rate mental exertion on a 5-point scale (Luts et al., 2010). The Vanderbilt Fatigue Scale for Hearing Aid Users (VFS-AH), by contrast, focuses on the cumulative effects of effort rather than a single moment (Hornsby & Kipp, 2016).

Simplified Visual Analogue Scales (VAS) (Huskisson, 1974) and the Effort Assessment Scale (EAS) (Krueger et al., 2017) contain only a single item for rating listening effort. These have been proven to be valuable and efficient in listening effort experiments (Rudner et al., 2012).

A comparison of different questionnaires used to measure listening effort is shown in Table 3.2.

Strengths and Limitations of Self Report Self-report measures have the advantage of being convenient to use compared to other methods that may require specialised equipment (McCormack et al., 2004). However, subjective measures have inherent limitations precisely because they are subjective. Individual differences in self-reflection and response bias can influence a person's answers (McGarrigle et al., 2014).

When examining the relationship between subjective reports and objective measures, the correlation is often only moderate (Ohlenforst et al., 2017). Combining subjective measures with objective ones, such as behavioural and physiological assessments, would provide a more comprehensive and balanced understanding of listening effort.

3.4.2 Behavioural Measures

Behavioural measures provide an objective perspective on listening effort by assessing task performance in real-time rather than relying on retrospective recall which inevitably invites bias. They primarily focus on methods such as dual-task paradigms to reflect cognitive resource allocation. The main behavioural indices include reaction time and accuracy across various paradigms.

Primary Indicators

Reaction Time Reaction Time (RT) is theoretically grounded in capacity theories of attention, which suggest that cognitive resources are limited when completing a task (Kahneman, 1973). As task difficulty increases, the brain requires greater resources, leading to slower processing speeds. Reaction time is particularly valuable for assessing listening effort under conditions of mild to moderate difficulty (McGarrigle et al., 2014).

Accuracy Accuracy reflects how successfully participants perform a listening task. It is primarily measured by the percentage of spoken words, phonemes, or sentences correctly identified or repeated (McGarrigle et al., 2014; Pichora-Fuller et al., 2016b). Similar to reaction time, it can be used to assess either a primary (Zekveld & Kramer, 2014) or secondary task during listening (Sarampalis et al., 2009).

Key Paradigms Several established paradigms are used in behavioural measurements when assessing listening effort.

Table. 3.2. Comparison of Subjective Measures of Listening Effort

Questionnaire	What It Measures	Number of Items	Completion Time	Strengths	Limitations
Ecological Momentary Assessment (EMA) of Listening Effort (Wu et al., 2015)	Real-time effort ratings in natural environments	Varies (typically 1-5 items per assessment)	<1 minute per assessment, multiple times per day	• High ecological validity • Captures real-world variation • Minimizes recall bias • Contextual information available	• Requires smartphone/technology • Participant compliance challenges • Analysis complexity • Potential sampling biases
Visual Analog Scale (VAS) for Listening Effort (Rudner et al., 2012)	Direct rating of perceived effort	1 item (typically)	<1 minute	• Extremely quick administration • Can track moment-to-moment changes • Minimal participant burden • Easy to implement • Focused specifically on effort	• Single-item measure with limited reliability • Subject to anchoring and scaling biases • Limited context for interpretation • Limited validation compared to other measures
Effort Assessment Scale (EAS) (Luts et al., 2010)	Perceived mental exertion during listening tasks	5-7 items (varies by version)	2-3 minutes	• Brief administration • Sensitive to different listening conditions • Extensively validated across domains	• Fewer normative data • May lack sensitivity to small changes • Not specific to listening tasks
NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988)	Six dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration	6 core items + 15 pairwise comparisons (in full version)	2-5 minutes	• Multidimensional assessment • Good sensitivity to task difficulty • Available in many languages	• Requires explanation for proper use • Raw vs. weighted scoring debate
Vanderbilt Fatigue Scale for Hearing Aid Users (VFS-AH) (Hornsby & Kipp, 2016)	Listening-related fatigue and cumulative effects of effort	10-16 items (depending on version)	5-8 minutes	• Links effort to fatigue outcomes • Measures prolonged effects • Relevant to real-world impacts	• Focuses on consequences rather than immediate effort • Specific to hearing aid users • Retrospective rather than immediate assessment
Abbreviated Profile of Hearing Aid Benefit (APHAB) (Cox & Alexander, 1995)	Hearing difficulties in different environments, including ease of communication	24 items	10 minutes	• Well-established clinical tool • Good for pre/post intervention • Normative data available	• Measures difficulty rather than effort directly • Limited to communication scenarios • Less sensitive to subtle differences
Speech, Spatial and Qualities of Hearing Scale (SSQ) (Gatehouse & Noble, 2004)	Three domains: speech hearing, spatial hearing, and qualities of hearing (including effort)	49 items (14 items in SSQ-12 short form)	15-20 minutes (full version) 5-7 minutes (short form)	• Hearing-specific • Comprehensive assessment • Sensitive to technology differences • Good ecological validity	• Length can be prohibitive • Some items may be difficult for participants to conceptualize • Not exclusively focused on effort

Single-Task Paradigm Single-task paradigms measure listening effort through a single auditory task, assuming that increased demands will lead to longer RT and/or decreased performance (McGarrigle et al., 2014). The key advantage of this approach is its simplicity compared to other designs, such as dual-task paradigms (Houben et al., 2013).

Dual-Task Paradigm Dual-task paradigms require participants to perform a listening task while simultaneously completing a secondary task that competes for cognitive resources (Gagné et al., 2017). A decline in performance or an increase in RT serves as an indicator of increased listening effort for the primary task. Common secondary tasks include visual tracking of another stimulus or memory tasks (e.g., digit recall).

Other Working Memory Paradigms These paradigms build on working memory and the ELU model (Section 3.2.3, Page 26), which posits that more challenging listening conditions place greater demands on working memory (Rönnberg et al., 2013; Rudner et al., 2012). One example is the reading span or listening span task, where participants are presented with a series of sentences and tested for comprehension. At the end of the task, they are asked to recall a specific word from a designated position within the material they have read or heard. The maximum number of items they can recall represents their "span," which is expected to decrease as task demands increase (Akeroyd, 2008; Daneman & Carpenter, 1980; Desjardins & Doherty, 2013; Souza & Arehart, 2015).

Table. 3.3. Summary of Typical Behavioural Measures in Listening Effort Research

Measure	**Description**	**Typical Paradigm**	**Interpretation** (Higher Listening Effort typically leads to...)	**Example Metric(s)**
Secondary Task Reaction Time (RT)	Speed of response to a simple, non-listening task performed concurrently.	Dual-Task Paradigm	Slower / Increased RT	Milliseconds (ms)
Accuracy	Accuracy of identifying/understanding spoken words, sentences, etc.	Primary Task Performance	Lower accuracy (as task difficulty increases, effort usually increases up to task failure)	Percentag Correct, Score
Secondary Task Accuracy	Correctness of response on the concurrent secondary task.	Dual-Task Paradigm	Lower Accuracy / Increased Errors	% Correct, Error Rate
Recall / Comprehension Performance	Ability to remember content or answer questions after a listening period.	Primary Task (Post-hoc)	Poorer Recall / Lower Comprehension Scores	Items recalled, Percentage Correct
Task Endurance / Time-on-Task Effects	How long performance is sustained or changes over a prolonged session.	Task Engagement/ Sustained	Shorter Endurance / Faster Performance Decline	Time (min), Performance change rate
Choice Behaviour / Task Selection	Participant's preference when choosing listening conditions or tasks.	Decision Making Task	Avoidance of more demanding conditions / Choice of less effortful options	Choice Frequency, Selected SNR

Strengths and Limitations of Behavioural Measures

Objectivity of Behavioural measures Compared to self-reports, behavioural measures such as RT and accuracy offer greater objectivity in assessing listening effort. They are based on clearly measurable external criteria, reducing the risk of reporting biases found in subjective methods, such as the tendency to meet expectations. Nonetheless, the choice of method should be carefully considered in research design.

Individual Differences Individuals vary in their baseline RT and accuracy, which can complicate the interpretation of listening effort (Wendt et al., 2016). Factors such as hearing status, cognitive capacity, and resource allocation strategies can significantly influence RT and accuracy results (Rudner et al., 2012). Failure to account for these differences may obscure meaningful patterns in group-level analyses of listening effort research (Gosselin & Gagné, 2017).

Speed-Accuracy Tradeoffs One of the most notable limitations of using either RT or accuracy in isolation is the phenomenon of speed-accuracy trade-offs. Some participants may prioritise response speed, while others may focus on maintaining accuracy (Houben et al., 2013). This effect can be attributed to inherent individual differences, experimental instructions (e.g., directing participants to prioritise accuracy), or task difficulty (as more challenging tasks tend to encourage accuracy over speed) (McMahon et al., 2016; Zekveld & Kramer, 2019).

A more balanced approach would be to combine RT and accuracy when measuring listening effort. Integrated methods such as Inverse Efficiency Scores (IES), which calculate a final score by dividing RT by the proportion correct (accuracy), provide a way to account for speed-accuracy trade-offs (Bruyer & Brysbaert, 2011).

As behavioural measures such as RT and accuracy aim to assess cognitive processes related to listening, including cognitive capacity and working memory, subjective reports focus more on individuals' perceived listening effort. To gain a more comprehensive understanding, a multimodal approach incorporating behavioural measures alongside subjective reports and physiological indices could be beneficial (Gagné et al., 2017; Pichora-Fuller et al., 2016b). This will be discussed in the next section.

3.4.3 Physiological Measures

Listening effort is essentially a cognitive process, but it is closely linked to physiological and brain activities, which makes physiological measurements possible. This connection arises from the body's reaction when performing cognitive tasks, the resource mobilisation system ensures brain receives necessary energy and oxygen to sustain cognitive work. As listening demands increase, the body reacts in pushing physiological adjustments to meet those needs (Kahneman, 1973; Kramer et al., 2016).

Automatic Nervous System Measures

Electrocardiography (ECG) The body's cardiovascular system provides valuable insights into cognitive effort during listening tasks. Heart Rate (HR) and Heart Rate Variability (HRV) are frequently used to assess listening effort.

Heart Rate is typically calculated by measuring the number of heartbeats per minute. Research by Mackersie and Cones demonstrated that HR consistently increases as SNR decreases, indicating more challenging listening conditions (Mackersie & Cones, 2011). Hicks and Tharpe found that children with hearing loss exhibit significantly greater HR elevations compared to their peers with normal hearing (Hicks & Tharpe, 2002), a finding later echoed by Mackersie and colleagues in adult participants (Mackersie et al., 2015).

Heart Rate Variability, on the other hand, refers to fluctuations in the time interval between individual heartbeats. It measures the regularity of heart rate over time. When individuals are relaxed, their heartbeat naturally fluctuates, allowing for greater variation and flexibility. However, under challenging conditions, when the sympathetic nervous system (associated with stress and effort) becomes more dominant, HRV tends to decrease. This reduction enables the body to maintain focus in a *fight or flight* mode, requiring less physiological regulation (Berntson et al., 1997; Hjortskov et al., 2004; Mackersie et al., 2015).

In listening effort research, studies have shown that HRV decreases as SNR increases in speech-in-noise tasks (Mackersie et al., 2015). Similarly, Wendt and colleagues (2018) used time-frequency analysis of HRV, revealing moment-by-moment fluctuations that correspond to specific linguistic and acoustic challenges (Wendt et al., 2018). When examining populations with hearing impairment, research suggests that listeners with hearing loss not only demonstrate greater HRV suppression during difficult listening tasks but also exhibit a delayed recovery once the task is completed (Hornsby & Kipp, 2016).

Respiration Respiration provides valuable insights into autonomic regulation during difficult listening conditions, though it has been studied less extensively than cardiovascular or pupillometry measures. Several key indices of respiration are used in listening effort research:

Respiration Rate is the most commonly used index, quantified as breaths per minute. Bernardi and colleagues demonstrated that during challenging speech-in-noise tasks, Respiratory Rate (RR) increases compared to quiet listening conditions (Bernardi et al., 2014). Similarly, Richter (2016) observed significant RR changes correlated with SNR (Richter et al., 2016a). Auer and colleagues further revealed that a sustained increase in

RR precedes performance decline in challenging listening tasks, suggesting RR as a potential early marker of listening effort (Auer et al., 2021).

Respiratory Variability examines the consistency of breathing patterns during effortful listening. Studies have found that in difficult listening conditions, there is a significant change in breathing interval variability (Vlemincx et al., 2013). However, Grassmann and colleagues (2016) revealed that as total variability in respiratory rate is not systematically affected by cognitive load, though the correlated fraction - the proportion of respiratory variability that shows regular, structured, or predictable patterns over time. decreases (Grassmann et al., 2016).

Tidal Volume (TV), another index used in measuring listening effort, refers to the volume of air moved into or out of the lungs with each breath. It is typically measured in millilitres and serves as an indicator of breathing depth during listening tasks. Compared to RR and respiratory variability, which require post-task analysis, TV provides a faster measure of breathing changes, improving temporal accuracy.

Studies have shown that during more demanding listening tasks, TV decreases by 15-20% compared to baseline measures (Grassmann et al., 2016). Research monitoring respiratory parameters while manipulating SNR during speech perception tasks has demonstrated progressive decreases in TV as noise levels increase (Bernardi et al., 2014). When examining individuals with hearing impairments, Alhanbali and colleagues (2019) found that during challenging listening conditions, individuals with hearing loss show **more pronounced reductions in TV** compared to those with normal hearing (Alhanbali et al., 2019).

Pupillometry Pupil diameter reflects the interplay of the Autonomic Nervous System (ANS) (Loewenfeld, 1999; McDougal & Gamlin, 2015). When the Sympathetic Nervous System (SNS)-which is associated with the *fight or flight* response and effortful processing-is more active, arousal levels increase, leading to pupil dilation (Loewenfeld, 1999; Szabadi, 2011). Conversely, when the Parasympathetic Nervous System (PNS)-often referred to as the *rest and digest* system-is more dominant, pupil size tends to constrict (Szabadi, 2011).

Pupillometry provides a sensitive and robust method for assessing listening effort. Research has consistently demonstrated a correlation between increased pupil diameter and listening difficulty. Kramer and colleagues (1997) provided early evidence of a link between pupil size and noise levels during listening tasks (Kramer et al., 1997). More recent research has continued to confirm the strong relationship between pupil dilation and task difficulty (Kuchinsky et al., 2013; Zekveld et al., 2010).

Galvanic Skin Response GSR, or skin conductance, measures electrodermal activity, which fluctuates due to microscopic sweat secretions triggered by the SNS (Mackersie et al., 2015).

GSR has proven particularly useful for investigating stress responses in challenging listening situations. Studies have demonstrated increased GSR when individuals process degraded speech signals (Mackersie & Cones, 2011). Furthermore, individual differences in GSR appear to predict susceptibility to listening fatigue, suggesting that some listeners may experience greater physiological strain than others during demanding auditory tasks (Francis & Love, 2016).

Electroencephalography (EEG) Electroencephalography is a non-invasive technique that records electrical activity generated by the brain. It captures voltage fluctuations resulting from ionic currents within neurons (Nunez & Srinivasan, 2006). EEG offers high temporal resolution, allowing researchers to track rapid brain changes associated with cognitive processes, including listening effort (Kraus & Slater, 2015)

Multi-channel EEG systems use multiple electrodes positioned across the scalp according to the International 10-20 system (Oostenveld & Praamstra, 2001). This setup gathers spatial information about brain activity, enabling researchers to localise cognitive processes to specific brain regions and study connections between different areas. In listening effort research, increased activity in frontal cortical regions has been observed during difficult listening tasks (Dimitrijevic et al., 2019).

Single-channel EEG uses a single electrode to record brain activity from a specific location. While it provides limited spatial resolution, single-channel EEG offers advantages such as lower cost and easier setup. It is particularly suitable for real-world applications and ecological testing environments.

EEG Signal Components Several indices are commonly used in EEG data when investigating cognitive effort.

Alpha wave (8-13 Hz) are dominant when a person is awake but relaxed (Niedermeyer & Lopes da Silva, 2005). They are typically suppressed during cognitively demanding tasks (Klimesch et al., 2007; Obleser et al., 2012; Pfurtscheller & Lopes da Silva, 1999).

Beta wave (13 - 30 Hz) are associated with active mental states, including alertness, concentration, and problem-solving. In listening effort research, beta power has been observed to increase during speech-in-noise tasks (Weisz et al., 2011).

Theta wave (4 -8 Hz.) are more prominent during drowsiness and shallow sleep and in certain meditative states (Niedermeyer & Lopes da Silva, 2005). In cognitive effort

studies, theta wave activity has been found to increase during difficult listening tasks (Bernarding et al., 2017) and challenging speech comprehension (Wiśniewski et al., 2017).

Other commonly used EEG indices include delta waves, P300, and N400. A detailed summary table is presented in Table 3.4.

Table. 3.4. EEG Indices in Cognitive and Listening Effort Research

EEG Index (Frequency/Timing)	General Functional Associations	Key Findings in Effort Research
Alpha (8-13 Hz)	• Relaxed wakefulness • Cortical inhibition • Attentional gating • Internal focus (Jensen & Mazaheri, 2010)	• Power decreases (suppression) in temporal/parietal regions during difficult listening conditions (Obleser et al., 2012). • Suppression magnitude correlates with subjective listening effort ratings (McMahon et al., 2016). • Increased power in task-irrelevant regions may reflect active inhibition during focused listening (Klimesch et al., 2007).
Beta (14-30 Hz)	• Active mental states • Motor function • Cognitive stability • Top-down control (Engel & Fries, 2010)	• Increased frontal power observed during challenging speech-in-noise tasks (Weisz et al., 2011). • Activity linked to executive control mechanisms engaged by effortful listening (Engel & Fries, 2010). • May reflect maintenance of the current cognitive or attentional state during sustained listening (McMahon et al., 2016).

Continued on next page

Table 3.4 – continued from previous page

EEG Index (Frequency/Timing)	General Functional Associations	Key Findings in Effort Research
Theta (4-7 Hz)	• Cognitive control • Working memory • Error monitoring • Executive function (esp. Frontal Midline Theta) (Cavanagh & Frank, 2014)	• Enhanced frontal midline theta power occurs during difficult speech comprehension (Wiśniewski et al., 2017). • Power often positively correlates with increasing cognitive demand in listening tasks (Bernarding et al., 2017). • Activity levels can relate to performance limits or decrements in sustained listening (Cavanagh & Frank, 2014).
Delta (0.5-4 Hz)	• Deep sleep • Signal detection • Motivational processes • Decision-making (Harmony, 2013)	• Increased power observed during effortful listening, particularly in noisy conditions (Dimitrijevic et al., 2019). • Thought to be associated with cognitive resource allocation during demanding speech processing (Wöstmann et al., 2015). • Potentially linked to semantic integration efforts under challenge (Kösem & van Wassenhove, 2018).

Continued on next page

Table 3.4 – continued from previous page

EEG Index (Frequency/Timing)	General Functional Associations	Key Findings in Effort Research
P300 (ERP) (Positive peak 300ms post-stimulus)	• Context updating • Stimulus evaluation • Attentional allocation • Working memory updating (Polich, 2007)	• Amplitude is often reduced under conditions demanding high listening effort (Ohlenforst et al., 2017). • Latency tends to increase in more challenging listening situations (Polich, 2007). • Reflects demands on attention allocation and working memory during speech processing (Kramer et al., 2016).
N400 (ERP) (Negative peak **400ms** post-stimulus)	• Semantic processing • Language comprehension • Expectancy violations • Meaning integration (Kutas & Federmeier, 2011)	• Amplitude typically increases (becomes more negative) with greater speech comprehension difficulty or semantic incongruity (Kutas & Federmeier, 2011). • Reflects increased semantic processing load or effort in challenging conditions (Strauss et al., 2013). • Sensitive to linguistic prediction errors encountered during effortful listening (Bidelman & Dexter, 2017).

Advantages and Limitations of EEG in Listening Effort Research EEG offers significant advantages due to its high temporal resolution and non-invasive nature. It enables researchers to monitor rapid changes in neural activity during cognitive tasks without the need for invasive procedures.

However, EEG also has limitations, as its spatial resolution is lower than that of functional Magnetic Resonance Imaging (fMRI). It is particularly susceptible to artefacts from muscle movements and eye blinks. Additionally, it imposes task-related constraints, requiring participants to remain relatively still to minimise movement-induced noise, which may reduce the ecological validity of listening experiments (Debener et al., 2012).

The practical use of EEG on experiment settings presents several challenges. Current EEG systems mainly use either wet or dry electrodes for measuring. Wet electrodes, which rely on conductive gels or pastes, offer better quality but require considerable time to prepare and clean up afterwards (Ferree et al., 2001). Bringing much longer experiment periods when testing multiple participants (Chi et al., 2010).

Dry electrodes, on the other hand, have emerged as a more convenient alternative and reducing preparation time significantly. however, to compensate for reduced conductivity, these electrodes often feature claw-like or pin-based designs which rely on more pressure on the scalp for better conductivity, which causes discomfort for longer experiments.

Future research could benefit from the integration of both types of conductive electrodes. Furthermore, advancements in dry electrode EEG technology may enhance participant comfort while simultaneously improving conductivity.

Functional Magnetic Resonance Imaging (fMRI) Functional magnetic resonance imaging (fMRI) is a neuroimaging technique used to measure brain activity. Unlike EEG, which relies on electrical currents generated by neuronal firing, fMRI measures changes in blood flow and oxygen levels associated with neural activity (Huettel et al., 2014). fMRI provides excellent spatial resolution, enabling researchers to pinpoint the specific brain regions activated during a listening task. However, its temporal resolution is lower than that of EEG (Logothetis et al., 2001).

Key findings of listening effort in fMRI are shown in Table 3.5

Table. 3.5. Summary of fMRI Findings on Brain Regions Activated During Increased Listening Effort

Brain Region(s)	General Function(s) Relevant to Effort	Typical Finding in Listening Effort Research	Key References
Dorsolateral Prefrontal Cortex (dlPFC)	Executive functions, working memory, top-down attention, strategic control	Increased activation with higher task difficulty or noise levels	(Peelle et al., 2010; Wild et al., 2012)
Anterior Cingulate Cortex (ACC)	Conflict/performance monitoring, error detection, effort scaling/perception	Increased activation reflecting monitoring demands and perceived difficulty	(Eckert et al., 2009; Wild et al., 2012)

Continued on next page

Table 3.5 – continued from previous page

Brain Region(s)	General Function(s) Relevant to Effort	Typical Finding in Listening Effort Research	Key References
Inferior Frontal Gyrus (IFG)	Cognitive control, language processing (syntax, semantics), inhibition	Increased activation, potentially for compensatory language processing/control	(Peelle et al., 2010)
Parietal Cortex (e.g., IPS)	Attention allocation, working memory maintenance, sensory integration	Increased activation linked to attentional demands and processing load	(Peelle et al., 2010; Wild et al., 2012)
Anterior Insula (AI)	Salience detection, interoception, task-level control, awareness of effort	Increased activation, often co-activated with ACC, linked to effort awareness	(Eckert et al., 2009; Peelle et al., 2010)
Auditory Cortex (STG/STS)	Sensory processing of sound, speech feature extraction	Modulated activity; complex patterns, can increase with intelligibility/processing demands	(Davis et al., 2011; Peelle et al., 2010)
Motor / Premotor Cortex	Motor planning and execution, simulation (e.g., covert articulation)	Increased activation sometimes reported, possibly linked to articulatory simulation	(Hickok et al., 2009; Wild et al., 2012)

Functional Near-Infrared Spectroscopy (fNIRS) Functional Near-Infrared Spectroscopy (fNIRS) is a relatively new method for measuring brain activity. It is a non-invasive neuroimaging technique that monitors changes in oxygenated and deoxygenated haemoglobin (a type of protein) concentrations in cerebral blood vessels (Ferrari & Quaresima, 2012). Compared to fMRI, it is lower in cost and offers better spatial resolution, but its temporal resolution is lower than that of EEG.

One advantage of Functional Near-Infrared Spectroscopy (fNIRS) is its ability to tolerate noise better than EEG. Additionally, it can be used simultaneously with hearing aids or cochlear implants without the risk of electromagnetic interference, unlike EEG (Lawrence et al., 2018). It is also more portable and easier to use, allowing for more natural experimental settings. Furthermore, it has greater tolerance to noise introduced by movement or other electrical interference (Peelle, 2017).

In summary, listening effort is a complex, multidimensional construct involving cognitive, motivational, and physiological components. Understanding how effort is experienced and measured is especially important for those with hearing loss, where the impact of effort may extend beyond the listening task itself. The next chapter considers why listening effort matters - not only at the individual level, but also in clinical, social, and educational contexts.

Chapter 4

The Significance of Listening Effort

Listening effort, defined as the cognitive resources deployed during auditory tasks (McGarrigle et al., 2014; Pichora-Fuller et al., 2016c), is a complex aspect of human perception and communication. As established in the preceding chapters, understanding speech and other relevant sounds, particularly in adverse conditions, is not merely a passive reception of stimuli but an active cognitive process influenced by contextual, signal-related, and listener-specific factors. This chapter presents the importance of understanding listening effort, examine the impact of excessive listening effort on individual well-being, particularly for those with hearing impairment, its role in shaping clinical audiological practices and hearing technology development, its broader societal relevance in educational and occupational settings, and its importance for specific populations including neurodivergent individuals and older adults. Understanding why listening effort matters underscores the critical need for the robust measurement techniques that will be discussed subsequently

4.1 Impact on Individuals with Hearing Impairment

4.1.1 Quality of Life, Fatigue, and Mental Well-being

For individuals with hearing impairment, the consequences of heightened listening effort extend far beyond simple communication difficulties. The constant exertion required to process degraded auditory signals can lead to significant cognitive fatigue, a sense of exhaustion distinct from physical tiredness (Edwards et al., 2016; Hornsby, 2013). This listening-related fatigue has been strongly linked to reduced overall well-being (Heffernan et al., 2016).

The sustained cognitive load can contribute to increased stress levels and negatively impact mental health (Edwards et al., 2016). Faced with the persistent strain of

communication, individuals may begin to avoid socially demanding situations, leading to social withdrawal, isolation, and a compromised quality of life.

Understanding listening effort provides a crucial framework for recognising and quantifying these often-hidden costs associated with hearing loss, moving beyond simple audibility towards a more comprehensive view of the listener's experience (Pichora-Fuller et al., 2016c). Studies have also shown that people who feel more affected by their hearing loss often report higher levels of fatigue, showing the real-world impact (Alhanbali et al., 2018).

4.1.2 Challenges in Clinical Assessment

The significance of listening effort also highlights limitations in traditional audiological assessments. Standard tests like Pure Tone Audiometry (PTA), typically conducted in quiet environments, often fail to capture the real-world difficulties experienced by individuals, particularly in noisy settings (Ohlenforst et al., 2017). This leads to the common clinical scenario where patients report significant communication problems ("I can hear, but I can't understand") despite having seemingly adequate hearing thresholds on the audiogram (Tremblay & Backer, 2015).

Listening effort research underscores the need to incorporate assessments that evaluate auditory processing under more realistic, challenging conditions and consider the cognitive resources involved. This motivates the exploration of supplementary clinical tools, potentially including more specific self-report questionnaires, behavioural paradigms that tax cognitive resources (like dual-tasks), and further, physiological measures, to gain a more accurate diagnosis and tailor interventions effectively (Ohlenforst et al., 2017)

4.1.3 Driving Hearing Technology and Rehabilitation

Recognising the burden of listening effort has shifted goals in hearing aid and cochlear implant development (Lunner et al., 2016). The focus extends beyond simply restoring audibility towards designing technologies that actively reduce the cognitive load associated with listening. Features such noise reduction algorithms, directional microphones, and frequency compression techniques are increasingly evaluated not just for their impact on speech intelligibility scores, but for their ability to lessen perceived effort and fatigue (Ng et al., 2015).

Understanding the physiological correlates of effort, can inform the design and evaluation of these features. Furthermore, acknowledging the role of cognitive factors opens avenues for rehabilitation strategies beyond technology, such as cognitive training or auditory training programs aimed at improving specific skills (like working memory

or attention) that support listening in challenging conditions. The ultimate aim, potentially informed by identifying individual physiological response patterns, is to move towards more personalised hearing solutions tailored to individual needs and processing styles

4.2 Societal and Functional Consequences

4.2.1 Educational Environments

In educational settings, increased listening effort puts a significant barrier to learning, affecting not only students with diagnosed hearing loss but also typically developing children in acoustically challenging classrooms (Zekveld et al., 2018). Poor room acoustics (e.g., high reverberation) and background noise force students to allocate excessive cognitive resources simply to perceive the teacher's instruction, leaving fewer resources available for comprehension, learning, and memory consolidation.

This increased cognitive load can result in reduced attention spans, increased fatigue throughout the school day, and ultimately, poorer academic outcomes. Recognising listening effort as a factor highlights the importance of optimising classroom acoustics and implementing supportive teaching strategies, such as using visual aids or remote microphone systems, to create more accessible learning environments for all students (Mackersie et al., 2015).

4.2.2 Workplace Productivity and Safety

The consequences of high listening effort extend significantly into the workplace. In professions requiring critical communication in noisy environments (e.g., emergency services, aviation, construction, call centres), the constant effort needed to understand speech can lead to substantial mental fatigue, reduced productivity, and an increased risk of communication errors (Hornsby & Kipp, 2016).

For safety-critical occupations, where clear communication and situational awareness are paramount, excessive listening effort can directly compromise performance and potentially lead to accidents (Mattys et al., 2018). Understanding and mitigating listening effort through environmental modifications, improved communication technologies, or appropriate work-rest schedules is therefore crucial for maintaining both worker well-being and operational safety.

4.3 Significance for Specific Populations

4.3.1 Neurodivergent Populations

Listening effort presents unique challenges for neurodivergent individuals, such as those with Autism Spectrum Disorder (ASD) or Attention-Deficit/Hyperactivity Disorder (ADHD). These individuals often experience significant difficulties processing auditory information, particularly in complex social or noisy settings, even when standard hearing tests (audiograms) are normal (Rance et al., 2014; Robertson & Baron-Cohen, 2020).

Research suggests they may expend greater cognitive resources (i.e., higher listening effort) than their neurotypical peers to achieve similar levels of performance in auditory tasks (Gosselin & Gagné, 2017). Recognising the role of listening effort in these populations is vital for accurate diagnosis, avoiding misattribution of difficulties solely to attention or behaviour, and developing tailored educational and therapeutic support strategies (Rance et al., 2014).

4.3.2 The Aging Population and Cognitive Health

Understanding listening effort is increasingly critical given global demographic shifts towards aging populations. Older adults often report experiencing greater listening effort, even with relatively normal hearing thresholds, potentially due to age-related declines in central auditory processing and cognitive functions like working memory and inhibitory control (Peelle, 2018).

This carries significant implications for cognitive health, as a potential detrimental cycle often exist: age-related cognitive decline can increase listening effort, while the chronic exertion of high listening effort may, in turn, deplete cognitive resources needed for other functions, potentially accelerating cognitive decline or increasing the risk of dementia (Lin et al., 2013). Addressing listening effort in older adults, through both hearing interventions and cognitive support, may therefore be crucial for maintaining cognitive function and overall quality of life.

4.4 Research Importance

4.4.1 Relevance to listening Effort Theory

Beyond its practical consequences, the study of listening effort holds considerable theoretical importance. It serves as a valuable paradigm for investigating fundamental cognitive processes such as attention, working memory, and executive control under

demanding conditions. The challenges in measuring effort consistently across different modalities and contexts push researchers to refine theoretical models like FUEL (Pichora-Fuller et al., 2016c) or ELU (Rönnberg et al., 2013), leading to a more nuanced understanding of how cognitive resources are allocated and managed during complex perceptual tasks. Furthermore, the documented significance across so many domains directly motivates the ongoing research into developing more sensitive, reliable, and ecologically valid measurement techniques - the focus of the subsequent chapter.

4.4.2 Listening Effort as a Bridge Between Fields

Listening effort bridges multiple fields: auditory neuroscience, cognitive psychology, audiology, and human factors research. It links bottom-up acoustic processing with top-down cognitive control, including attention, memory, and executive function. From a theoretical standpoint, listening effort has helped clarify why intelligibility does not always predict listening ease. Someone may accurately repeat what was said, but only with significant exertion. This dissociation challenges simplistic assumptions that better performance always means better experience. By providing a framework for studying this disconnect, listening effort helps refine models of speech perception and cognitive load (Pichora-Fuller et al., 2016c; Rönnberg et al., 2013).

In summary, listening effort represents a significant cognitive load with implications across numerous aspects of life. From impacting the daily well-being, social participation, and mental health of individuals with hearing loss, to influencing clinical practice and the design of assistive technologies. Furthermore, listening effort plays a critical role in educational attainment, workplace safety and productivity, and presents unique challenges for neurodivergent individuals and the rapidly growing aging population, potentially interacting with cognitive health trajectories. The significance highlights the importance of understanding its underlying mechanisms and developing effective methods for its measurement and management.

Chapter 5

Overall Research Aim

In the previous chapters, we have presented the importance of listening effort research, and measurements of listening effort from three fields: subjective reports, behavioural measures, and physiological measures. Each method has its own merits and limitations. The multifaceted nature of listening effort, which involves the biological foundations of the auditory system and brain, the cognitive functions of the brain, and the complexity of speech and language, makes measuring listening effort particularly challenging.

5.1 Addressing the Gaps

The process of listening, particularly in challenging acoustic environments, demands significant cognitive resources, commonly referred to as listening effort. As established in the preceding chapters, sustained or excessive listening effort carries substantial consequences, impacting not only communication quality but also contributing to cognitive fatigue, reduced quality of life, and social withdrawal, especially for individuals with hearing impairment and older adults (Edwards et al., 2016; Heffernan et al., 2016; Hornsby, 2013; Lin et al., 2013).

In addition, the cognitive load associated with listening effort has implications in crucial societal contexts, including educational attainment and workplace productivity and safety (Hornsby & Kipp, 2016; Mattys et al., 2018; Zekveld et al., 2018). Understanding the nature and mechanisms of listening effort is therefore of considerable theoretical and practical importance. Despite its significance, accurately defining and measuring listening effort remains a considerable challenge (Chapter 3.1, 3.4). While subjective self-reports offer direct insight into perceived exertion and behavioural measures like dual-task performance can index resource allocation (Gagné et al., 2017; Hart & Staveland, 1988; Luts et al., 2010), these methods have limitations.

Subjective ratings can be influenced by bias and introspection, while behavioural outcomes may not fully capture the internal cost incurred (McGarrigle et al., 2014; Ohlenforst et al., 2017). Physiological measures (e.g., pupillometry, electrodermal activity, cardiovascular responses, EEG) render more objective indices of the body's response to cognitive demands (Kramer et al., 2016; Mackersie et al., 2015; McMahon et al., 2016; Zekveld et al., 2010). However, a key unresolved issue, highlighted throughout the literature and foreshadowed in the initial analysis of Study 1, is the frequent dissociation observed between these different measurement domains - physiological responses do not always align predictably with behavioural performance or subjective reports (Alhanbali et al., 2018; Ohlenforst et al., 2017).

Furthermore, much research has focussed on group-level averages, potentially obscuring the substantial individual variability known to exist in how listeners experience and respond to auditory challenges (Chapter 3.3.2; (Koelewijn et al., 2012; Peelle, 2018; Zekveld et al., 2011)). Existing theoretical frameworks like FUEL and ELU provide valuable models (Chapter 3.2), but a deeper understanding requires empirical investigation into the dynamic, multi-system physiological patterns that characterise individual responses, their consistency over time, and how they adapt to varying task demands. Specifically, there is a need to move beyond static or peak measures to analyse the full time-course of physiological signals, which may reveal more nuanced aspects of effort regulation (Koelewijn et al., 2018; Winn et al., 2016).

Therefore, the current research programme is motivated by the need to address these specific gaps. It employs a multi-dimensional approach, integrating behavioural, subjective, and a diverse range of physiological measures (autonomic and central) to gain a more comprehensive understanding of listening effort. Crucially, this research focuses on analysing the dynamic temporal patterns of physiological responses and directly investigates individual differences and within-subject consistency. By examining these aspects across different listener groups (hearing-impaired and normal-hearing) and systematically varying task difficulty (adaptive and fixed SNRs), this work aims to provide a more nuanced characterisation of the physiological manifestations of listening effort and bridge the gap between internal processing costs and observable outcomes.

5.2 Overall Research Aim and Strategy

Given the limitations and gaps identified in the current understanding of listening effort, particularly the dissociation between measurements and the role of individual variability, the overarching aim of this research programme is to achieve a deeper, multi-dimensional understanding of listening effort. This work specifically seeks to investigate how listening effort manifests through various physiological systems, how these response

patterns differ between individuals and remain consistent within them, and how they are modulated by task demands and listener characteristics.

To address the multifaceted nature of listening effort, a two-study research strategy was adopted. **Study 1** involves a secondary analysis of physiological data (EEG, GSR, Pupillometry) collected from older adults with hearing impairment performing a digit-in-noise task. This allows for an initial exploration of individual physiological consistency and response patterns within a population known to experience significant listening challenges in daily life.

Study 2, building on Study 1, involves normal-hearing participants conducting performing a more complex, sentence-based speech-in-noise test under systematically varied, fixed signal-to-noise ratios (SNRs). This controlled design enables a clearer investigation of how task difficulty modulates physiological responses (including additional measures of ECG and Respiration) and performance, and facilitates a more direct examination of the relationships between different measurements. Together, these studies allow for both the characterisation of effort responses in a clinically significant group (Study 1) with a simple digit-in-noise task, and examination of underlying mechanisms and task-difficulty effects (Study 2) with speech-in-noise test, offering a richer, more integrated understanding of listening effort.

5.3 Research Questions and Hypothesis

Overarching Goal of this research is to investigate the physiological responses of listening effort, explore individual differences in these responses, and examine how these responses are modulated by task difficulty (SNR) and listener characteristics (hearing status).

Physiological Correlates and Task Difficulty

Research Question 1: How do distinct physiological systems (autonomic: pupillometry, GSR, heart rate, respiration; central: EEG alpha power) respond dynamically during effortful listening tasks, and how are these responses modulated by varying levels of task difficulty (SNR)?

- Hypothesis 1.1 (Difficulty Effect): Increased task difficulty (lower SNR) will lead to greater physiological activation across multiple systems, reflected by:
  - Increased pupil dilation (Pupillometry).
  - Increased skin conductance levels/responses (GSR).
  - Increased heart rate and/or decreased heart rate variability (ECG).

  - Changes in respiration rate or variability (Respiration).
  - Greater suppression of EEG alpha power.

- Hypotheses 1.2 (Dynamic Response): Physiological responses will show distinct temporal patterns related to task phases (e.g., listening vs. retention), with greater modulation (e.g., larger peaks/troughs, slower recovery) observed under higher task difficulty.

Individual Differences and Consistency

Research Question 2: Do individuals exhibit consistent, characteristic physiological response patterns (signatures) to listening effort across repeated experiments, and can individuals be reliably grouped based on these patterns?

- Hypotheses 2.1 (Within-Subject Consistency): Individuals will demonstrate significantly higher similarity in their physiological response time-courses across repeated experimental sessions compared to similarity between different individuals, indicating stable individual response styles.
- Hypotheses 2.2 (Between-Subject Differences and Clustering): Cluster analysis applied to physiological time-courses will reveal distinct subgroups of participants exhibiting qualitatively different response patterns (e.g., different magnitudes, timings, or shapes of response) within specific conditions.

Research Question 3: How do these individual physiological response patterns (clusters) relate to listener characteristics (hearing status - Study 1) and behavioural/subjective outcomes (accuracy, perceived effort, perceived difficulty - Study 1 and 2)?

- Hypotheses 3.1 (Hearing Status - Study 1): Individuals with greater hearing loss (higher PTA) will exhibit physiological patterns indicative of higher effort (e.g., belonging to clusters with greater activation) and report higher subjective effort, even when performance is matched via adaptive SNR.
- Hypotheses 3.2 (Physiology-Behaviour Link - Study 1 and 2): Membership in distinct physiological clusters will be associated with differences in behavioural accuracy and/or subjective ratings of effort/difficulty, although the relationship may not be clear (e.g., some high-activation clusters might correspond to better performance due to effective compensation, while others might link to poorer performance or higher reported effort).
- Hypotheses 3.3 (Physiology-Subjective Link - Study 1 and 2): Direct correlations will exist between the magnitude of physiological change during demanding task

periods (e.g., listening window) and subjective ratings, such as increased physiological activation correlating with higher reported effort or difficulty. Initial findings suggest GSR change correlates significantly with accuracy and difficulty.

Cross-Modal Relationships and Task Complexity

Research Question 4: To what extent do different physiological measures provide convergent or divergent information about listening effort? Is there significant agreement in how individuals are classified based on different physiological signals?

- Hypotheses 4.1 (Cross-Modal Divergence): Clustering agreement across different physiological modalities (e.g., comparing GSR-based clusters to pupil-based clusters) will be low to moderate, indicating that different systems capture distinct aspects of the overall effort response.

Research Question 5 (Study 2 Focus): How does a more complex listening task (sentence recognition vs. digit recall) and the inclusion of additional physiological measures (ECG, Respiration) refine the understanding of listening effort compared to previous findings (Study 1)?

- Hypotheses 5.1 (Task Complexity Effect): The more complex sentence task will elicit more pronounced and potentially more differentiated physiological responses compared to simpler digit tasks, particularly in measures sensitive to higher cognitive load (e.g., EEG, pupillometry).
- Hypotheses 5.2 (Added Value of ECG/Respiration): Heart rate and respiration measures will provide complementary information about autonomic regulation during listening effort, potentially revealing different temporal dynamics or sensitivities compared to GSR and pupillometry.

5.4 Comparison of the Studies

To achieve the research aims outlined above, this thesis presents findings from two distinct but complementary empirical studies.

Study 1, detailed in Part III, involved a secondary analysis of existing data collected from a group of 30 older adults (aged 51-80 years) with varying degrees of hearing loss (mean PTA 41.7 dB HL). Participants performed a listening and memory task based on the Sternberg paradigm, requiring them to recall digits presented in unmodulated noise. Crucially, the signal-to-noise ratio (SNR) was adaptively adjusted for each individual to target a consistent performance level (71% accuracy).

The available data included multi-channel electroencephalography (EEG - focusing on alpha power, analysing data from Pz electrode, referencing the electrode on the right ear lobe), galvanic skin response (GSR), and pupillometry, alongside subjective workload ratings (NASA-TLX) and behavioural accuracy. The primary focus of this analysis was to characterise individual differences in physiological response patterns, assess their consistency across repeated sessions (test-retest reliability), and explore potential relationships between these physiological signatures and listener characteristics (hearing level) or outcomes (subjective effort, performance).

Study 2, presented in Part IV, describes a new experiment conducted with approximately 30 younger adults (aged 18-40 years) with normal hearing. This study employed a more complex and arguably more ecologically valid listening task: the Oldenburg Sentence Test (OLSA), adapted for British English, requiring participants to identify five-word sentences presented in multi-talker babble noise. In contrast to Study 1, task difficulty was systematically manipulated using four fixed SNR levels (-16, -11, -6, and 12 dB), chosen to span a wide range of intelligibility.

The physiological measures were expanded to include electrocardiography (ECG, for heart rate analysis) and respiration, alongside single-channel EEG, GSR, and pupillometry (using wearable eye-tracking glasses). In Study 2, EEG was recorded at Pz using a single active electrode referenced to a forehead electrode positioned at Fpz, with a separate ground electrode placed behind the right earlobe(the mastoid area); the single-channel configuration provides only the differential signal between Pz and the Fpz reference. Subjective ratings of perceived effort and difficulty were collected using simple visual scales after blocks of trials at each SNR.

The key objectives of Study 2 were to examine how different levels of task difficulty modulate behavioural, subjective, and multi-system physiological responses; to investigate the specific contributions of heart rate and respiration dynamics; and to further explore individual consistency and the relationships between different measures under controlled conditions.

A brief summary of two studies in this research are shown in Table 5.1.

Outline of Subsequent Chapters The next sections will proceed as follows: Part III (from page 65) will present Study 1, and Part IV (from page 105) will introduce Study 2. Finally, Part V (from page 207) will provide a general discussion and outline potential directions for future research.

Table. 5.1. Summary of Study One and Study Two

	Study 1 (Secondary Data Analysis)	**Study 2** (Full Experiment)
Participants	Patients with hearing impairment	People with normal hearing
Sample size	Around 30 (valid data varies between different measures)	Around 30 (valid data varies between different measures)
Stimulus	Digit Test	Speech-in-noise Test
Signal-to-Noise Ratio	One adaptive level: Adjusted for each individual to achieve 71% accuracy	Four set levels: -16, -11, -6, 12 dB
Measurements	Subjective Reports Accuracy Electroencephalogram Galvanic Skin Response Pupillometry	Subjective Reports Accuracy Electroencephalogram Galvanic Skin Response Pupillometry Electrocardiogram Respiration

Part II

Study 1: Secondary Data Analysis

Chapter 6

Introduction and Research Aims

6.1 Introduction

This first study involves a secondary analysis of a rich dataset acquired previously at the University of Manchester (data collection led by Dr Sara Alhanbali; Faculty Ethics Committee reference ERGO ID: 72142). While the original analyses primarily focussed on group-level trends and relationships between average responses across different measures (e.g., Alhanbali et al. (2018, 2021)). The current study adopts a different perspective. Rather than group-level trends, it explicitly focuses on the the consistency and differences of individual listeners.

The original research investigated listening effort in 30 older adults with varying degrees of hearing loss, employing subjective reports (NASA-TLX), behavioural measures (accuracy in a digit-in-noise task), and multiple physiological indices, including (EEG), galvanic skin response (GSR), and pupillometry. This existing dataset presents a valuable opportunity to explore the physiological dimensions of listening effort within a clinically relevant population known to experience significant communication challenges.

6.2 Aim of Study

The primary aim of Study 1 is therefore to characterise the nature and consistency of individual physiological response patterns (EEG alpha, GSR, pupillometry) during an effortful listening task within older adults with hearing impairment. By examining individual time-course data and exploring clustering results based on physiological signatures, this study seeks to understand whether stable, distinct response profiles exist and how they might relate to listener characteristics or outcomes, moving beyond group-average summaries.

6.3 Research Questions and Hypotheses

This secondary data analysis is guided by the following specific research questions and corresponding hypotheses:

Research Question 1.1 To what extent do individuals show distinct physiological response patterns (in terms of time-course dynamics) in EEG alpha activity, GSR, and pupillometry during the listening effort task?

- Hypotheses 1.1 (Individual Differences): Significant between-subject variability will be observed in the time-course patterns of EEG alpha, GSR, and pupillometry responses.

Research Question 1.2 Are these individual physiological response patterns (EEG alpha, GSR, pupillometry) consistent and reliable within the same person across two experimental sessions conducted one week apart?

- Hypotheses 1.2 (Consistency): Individuals will exhibit statistically significant within-subject consistency (test-retest reliability), measured by correlation, in their physiological response patterns (EEG, GSR, Pupillometry) across the two experimental sessions, exceeding the consistency observed between randomly paired individuals.

Research Question 1.3 Can participants be clustered meaningfully based on the similarity of their physiological response time-courses for each measure (EEG, GSR, Pupillometry)?

- Hypotheses 1.3 (Clustering): Cluster analysis applied to the time-course of physiological response patterns (EEG alpha, GSR, pupillometry) will successfully identify a small number (e.g., 2-3) of distinct subgroups of participants for each measure, based on standard clustering validation metrics (e.g., elbow method).

Research Question 1.4 If distinct physiological clusters are identified, are they significantly associated with participants' hearing level (PTA), overall subjective effort ratings (NASA-TLX), or behavioural performance (accuracy)?

- Hypotheses 1.4 (Cluster Correlates): Membership in the identified physiological clusters will be significantly associated with variations in participants' hearing level (PTA), subjective effort ratings (NASA-TLX), and/or behavioural performance

(accuracy). (Note: Based on the literature's physiology-behaviour gap, this hypothesis is tentative).

Research Question 1.5 How much agreement exists between the participant groupings derived independently from EEG, GSR, and pupillometry? Do these different physiological signals classify individuals in similar ways?

- Hypotheses 1.5 (Cross-Modal Agreement): Cluster assignments derived independently from EEG, GSR, and pupillometry will show statistically significant, albeit potentially moderate, agreement (e.g., measured by ARI), suggesting these measures capture partially overlapping aspects of the effort response.

Addressing these questions and testing these hypotheses through the analysis of this existing dataset will provide foundational insights into the nature of individual physiological responses to listening effort in an older, hearing-impaired group, setting the stage for the experimental investigations in Study 2. The subsequent chapters in this Part will detail the specific analysis methods applied to the dataset and present the corresponding results. The following chapter details the experiment design and data acquisition procedures.

Chapter 7

Experiment Design

Original data was collected by Alhanbali from the University of Manchester. The description of experiment design is mainly referenced from study 3 in Alhanbali's thesis (Alhanbali, 2017).

7.1 Participants

Participants were native speakers recruited from database of three UK National Health Service audiology departments, through flyers at the University of Manchester, or through social groups. Thirty participants were recruited for the study. Due to differences in data quality across different measures, the number of participants included in each analysis varies slightly. Each measure was analysed using the participants with valid data for that modality, and overlapping datasets were used when comparisons between measures were required (e.g., valid EEG datasets: n = 29; datasets with both EEG and behavioural data: n = 27).

Of the participants, 50% being male, participated in this study with repeated design. They ranged in age from 51 to 80 years (M = 69.9, SD = 6.37), and hearing level (PTA) ranged from 7.5 to 78.75 dB HL (M = 41.7, SD = 17.74). Participants with hearing threshold $\leq$ 30 dB HL were classified as having normal hearing (n = 8, age: 60–78 years).

7.2 Listening Task

The listening task was based on a modified version of the Sternberg paradigm (Sternberg, 1966). Participants with hearing impairment performed the task whilst wearing their hearing aids with their everyday setting. In contrast to the 3-digit sequences used for

SNR determination, the main study employed sequences of 6 digits to increase the cognitive demands of the task.

The speech material consisted of digit from 1 to 9, recorded from a male speaker (McShefferty et al., 2013). The bisyllabic number 7 was excluded to maintain consistent syllable count across stimuli. Digits were presented with background noise that started five seconds before the first digit, and continued until one second after the final digit ended (Alhanbali et al., 2018).

The Signal-noise-ratio was adapted for each participant to achieve 71% correct identification of the digits. This was done using sequences of three digits in an adaptive 2-down, 1−up procedure with a 2−dB step size. A response was only considered correct if the participant identified all three digits and in the right order. The average SNR determined was -4 dB (SD 5 dB). Noise applied here are unmodulated continuous noise, presented at 65 dBA. (Alhanbali et al., 2018).

7.3 Measurements

Self-report The NASA Task Load Index was employed to measure self-reported listening effort (Hart & Staveland, 1988). The questionnaire comprises six items: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration. Participants completed the questionnaire using a 20-step scale (see Appendix A, Page 220) at the end of the experiment. These ratings were subsequently averaged and converted to percentages for analysis. Task performance (accuracy) was recorded throughout the experiment.

EEG EEG was collected through Nexus - physiological recording system, sampled at 256Hz with BioTrace software. No online filtering was applied. Four electrodes were positioned according to the international 10-20 system: Cz, Pz, P3, and P4 (channels 1, 2, 3 and 4, respectively; see Figure 7.1). Pz was referenced to a negative electrode placed on the right ear lobe. The ground electrode was placed at the forehead. EEG data were cleaned and filtered to the Alpha band (8 to 13 Hz) for subsequent analysis.

GSR Skin conductance (Galvanic skin response) was recorded from two electrodes on the non dominant hand (index and middle finger). It was using the same system as EEG - Nexus-10 system, sampling at 32 Hz. Participants were asked to keep their hand palm-up to minimise movement artefacts.

Pupillometry Pupillometry was recorded through an Eyelink 1000 eye-tracker. Room lighting and screen brightness were individually adjusted for each participant following

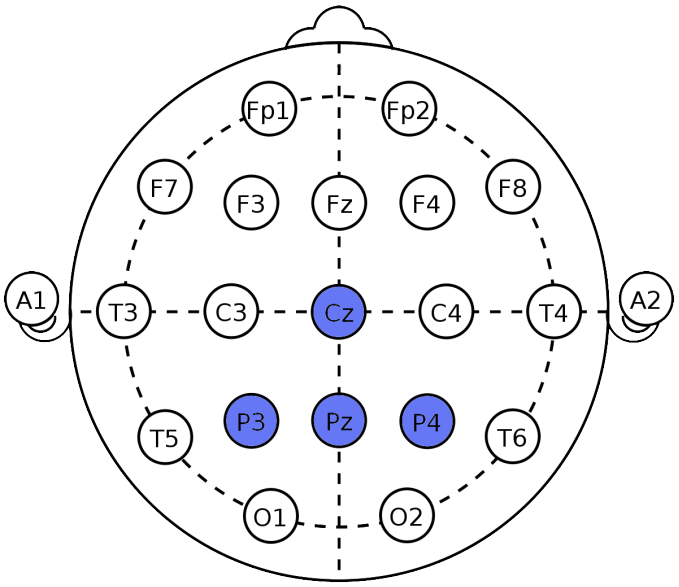


Figure. 7.1. EEG Electrodes Applied in Experiment. The coloured positions indicate electrode placements used in the experiment.

the procedures described by Zekveld et al. (2010). A standard 9-point eye point calibration was performed before the experiment. Room lighting and screen brightness were adjusted for each participant to place their baseline pupil size in a mid-rage, avoiding floor - ceiling effect.

7.4 Experiment Procedure

Participant Setup and Preparation

Environment and Positioning Participants were seated comfortably in a sound-treated booth facing a computer monitor. A chin rest was used to help maintain a stable head position during the experiment. Participants who regularly used hearing aids wore their own devices, using the settings they typically use in everyday life.

Physiological Monitoring Setup EEG electrodes were attached to the scalp according to the 10-20 system, skin conductance sensors were placed on the non-dominant hand, and an Eyelink 1000 eye-tracker was positioned and calibrated for each participant to record pupillometry data. Room lighting and screen brightness were adjusted individually before pupillometry recording commenced.

Baseline Acclimatisation. Prior to the main experimental tasks, participants watched a documentary for 10 minutes. This served as an acclimatisation period and allowed for the collection of baseline physiological data, particularly for skin conductance.

Trial Structure

- Listening Phase Each trial began when the participant indicated they were ready (e.g., by pressing ENTER). The word "Listen" appeared on the screen. Unmodulated background noise commenced via loudspeakers. After 5 seconds of noise, a sequence of six spoken digits (1-9, excluding 7) was presented within the noise at the participant's predetermined SNR. Each digit won't repeat more than twice. The noise stopped 1 second after the final digit.
- Retention Phase A fixation cross appeared on the screen for 3 seconds. During this silent period, participants were instructed to mentally rehearse and memorise the six digits they had just heard.
- Response Phase A single digit appeared on the screen next to a question mark. Simultaneously, an audible alert tone (beep) sounded as a cue to respond. Participants had to indicate as quickly as possible whether the probe digit was one of the six digits presented in that trial by pressing a "Yes" or "No" button on a response box.
- Recovery Phase A silent 4-second recovery period followed the participant's response before the next trial could be initiated.

Experiment Structure Participants performed 10 practice trials using the main task format (six-digit sequences) at their determined SNR to ensure they understood the procedure before the recorded session began. The main task involved 50 trials, with each trial following the sequence below. The total task duration was approximately 15 minutes. This design allowed for the collection of time-locked behavioural and physiological data across a consistent task structure, providing a foundation for assessing individual response patterns and test-retest reliability.

Chapter 8

Data Analysis Method

Alhanbali and colleagues has done substantial works which can be found in Alhanbali et al., 2018, 2019, 2021, where the group average response of physiological responses were explored. This study is focusing on the individual differences, particularly focusing on time-course response - the overall shape of the whole trial, in addition to a extracted data point.

Two types of response measures were derived: the Average Trial Response (ATR), which captures the average shape of a physiological signal across all trials, and the Time Course Response (TCR), which reflects how these signals evolve over the full session, trial by trial.

8.1 Data Structure

A total of 31 participants were included in the initial dataset, with each participant providing two data files from repeated measures. The data were analysed using MATLAB (R2021b). Raw data were first cleaned to remove artefacts, and some files were excluded due to measurement faults or missing data. Ultimately, 29 participants had valid EEG data, 26 participants had valid GSR data, and 26 participants had valid pupillometry data. Participants with valid data varied across the three physiological measures; for example, participant 7 was included in EEG but not in GSR. To maximise the available data for analysis, valid participants' data files were selected for each corresponding analysis rather than restricting the dataset to only those participants who overlapped across all measurements (18 participants overlapped across all measures).

8.2 Data Preparation and Derived Measures

8.2.1 Average Trial Response (ATR) and Time Course Response (TCR)

Raw data were initially examined and cleaned before calculating and analysing Average Trial Response (ATR) and Time Course Response (TCR) (examples are shown in Figures 9.3 and 9.4). Average Trial Response (ATR) represents the average response across all trials, reflecting task-related responses (50 trials in total; see EEG analysis, page 79). Time Course Response (TCR) was computed by averaging all samples within each trial, resulting in 50 data points per data file. These points were then connected to illustrate response changes over time.

8.3 Analysis of Individual Consistency and Differences

8.3.1 Permutation Test of Correlation

To assess whether participants were more similar to themselves than to others, permutation tests were used to determine correlation significance (Good, 2013). Originally introduced by R.A. Fisher in the 1930s, permutation tests evaluate whether observed differences are greater than those obtained by chance. The null hypothesis assumes that all samples originate from the same distribution.

First, the observed difference between two groups (d1) is computed. The dataset is then randomly shuffled to create new groups, and the difference between these randomised groups (dR) is measured. This process is repeated 1,000 times to generate a distribution of random differences. A one-sided or two-sided p-value (depending on test requirements) is then calculated to determine whether d1 is significantly different from dR (Good, 2013).

In this case, correlation rather than difference was tested. Since each participant contributed data from two experimental sessions, the null hypothesis posited no difference between the correlation within the same participant and the correlation between randomly paired participants. The correlation between the same participant's data were first computed and averaged. Next, datasets were randomly reassigned to different participants, and the correlation was recalculated and averaged. This process was repeated 1,000 times to generate a distribution of random correlations. The significance of the observed correlation was assessed by its position within this distribution (see Figure 9.5 for an example).

8.3.2 Cluster Analysis

Cluster analysis was conducted to determine whether participants, despite their individual differences, could be grouped based on similar response patterns. Cluster analysis is a statistical technique that groups observations with similar characteristics and encompasses various classification methods.

Clustering methods can be hierarchical or non-hierarchical. Hierarchical clustering, which utilises a dendrogram approach, is more appropriate for smaller datasets, whereas non-hierarchical methods, such as K-means clustering, are preferable for larger datasets as they require the researcher to specify the number of clusters (Ziegel, 2000).

Clustering in Study 1 was carried out in Python using the scikit-learn implementation of K-means. K-means clustering was selected for this study due to its advantages: (1) faster computational time, (2) suitability for large datasets, and (3) flexibility in defining clustering criteria (Wu, 2012). The method begins by selecting random centroids, then assigns each observation to the closest cluster based on the chosen algorithm. The centroids are iteratively adjusted until convergence is reached, meaning that cluster assignments remain stable (Boccard & Rudaz, 2013).

For this analysis, correlation was used as the clustering criterion rather than Euclidean distance. This ensured that participants clustered together exhibited similar response patterns rather than mere spatial proximity. Given that K-means clustering is sensitive to initial centroid selection, the procedure was repeated 1,000 times to ensure stability and minimise error.

The elbow method was used to evaluate the optimal number of clusters by examining the reduction in clustering error as the number of clusters (k) increased. As k increases, the average distortion or clustering error decreases. When clustering is based on correlation, error is calculated as the sum of (1 - correlation coefficient). However, as k continues to increase, the improvement in clustering becomes marginal. The optimal number of clusters is identified at the 'elbow' point of the plot, where additional clusters no longer yield substantial improvements in error reduction (Dangeti, 2017) (see Figure 9.7 for an example). Further details are provided in the subsequent analysis.

8.3.3 Relationship between Physiological Responses and Other Measures

After clustering participants based on their physiological responses, we examined whether these groupings were associated with subjective effort, hearing thresholds (PTA), or age. More importantly, we also assessed the consistency of clustering across different physiological measures - that is, whether individuals grouped together based on one measure tended to be grouped together based on another.

Chapter 9

Results and Discussion

9.1 Subjective and Behavioural Response

This section presents descriptive data of subjective effort, hearing level (PTA), and performance. Participants (n = 32) include mainly senior adults, age range from 51 to 80 years(Mean = 68.74 years, Median = 70 years, SD = 6.39). Participants with various degrees of hearing loss and using their own hearing aids when performing the task.

Subjective listening effort was assessed using the NASA Task Load Index (NASA-TLX), a general measure of cognitive workload. (Hart & Staveland, 1988). Self-reported effort scores showed considerable variation, ranging from 5 to 100 (Mean = 29.44, Median = 24.17, SD = 20.57 , see Figure 9.1). This wide range likely reflects the significant individual differences within the participant group.

Behavioural performance, measured as accuracy in identifying the probe digit, also exhibited substantial variability, ranging from 68% to 98% (Mean = 88.19%, Median = 90%, SD = 7.1). Interestingly, this variation occurs despite the adaptive method to set SNR to achieve a consistent 71% accuracy level. This discrepancy suggests that factors other than the SNR level strongly influenced task success.

Correlation between PTA, Subjective Effort, and Performance To explore the relationships between hearing ability, perceived effort, and task performance, correlations were examined. Following normality tests which indicated that only PTA data were normally distributed, non-parametric Spearman correlations were computed between PTA, subjective effort (NASA-TLX), and performance (accuracy).

As shown in Table 9.1, none of these correlations reached statistical significance at the conventional $\alpha = 0.05$ level. However, the direction of the observed relationships aligns with theoretical expectations: higher PTA values (poorer hearing) tended to be associated

with higher subjective effort ($\rho = 0.29$) and lower performance ($\rho = -0.33$), and higher subjective effort showed a weak association with lower performance ($\rho = -0.12$).

The lack of statistical significance could be attributed to several factors, including the modest sample size ($N = 30 - 32$ for these correlations), the high degree of individual variability within this group of older adults with hearing loss, potential non-linear relationships between these variables, the limitations inherent in using PTA and a general workload measure like NASA-TLX, and possibly the adaptive SNR procedure itself which might have compressed the range of task difficulty experienced by participants.

Table. 9.1. Correlation between Subjective Effort, PTA, and Performance

Relationship	Spearman's ρ	p-value	Significant	n
PTA vs Subjective Effort	0.29	0.12	No	30
Subjective Effort vs Performance	-0.13	0.48	No	32
PTA vs Performance	-0.34	0.06	No	30

The relationship between subjective effort and performance accuracy for each participant is shown in Figure 9.1. The plot further underscores the considerable individual variability in response to the listening task, showing a wide scatter of participants across the effort and performance dimensions.

Overall, the subjective and behavioural data reveal significant individual differences in perceived effort and performance, even under individually adapted noise conditions. While the correlations align directionally with expectations, they were not statistically significant in this sample.

9.2 Electroencephalography (EEG)

9.2.1 Data Pre-processing

Electroencephalography (EEG) data were analysed to investigate the neural correlates of listening effort during the task. Four EEG channels were used, positioned according to the International 10-20 system: Pz, P3, P4, and Cz (channels 1, 2, 3, and 4, respectively), as illustrated in Figure 9.2.

The analysis is focusing on Pz(channel 1), as other channels generate similar results, with an emphasis on the alpha frequency band (8–13 Hz) as alpha wave, which is often associated with increased cognitive load and listening effort. The raw EEG data were filtered within this band and analysed using MATLAB.

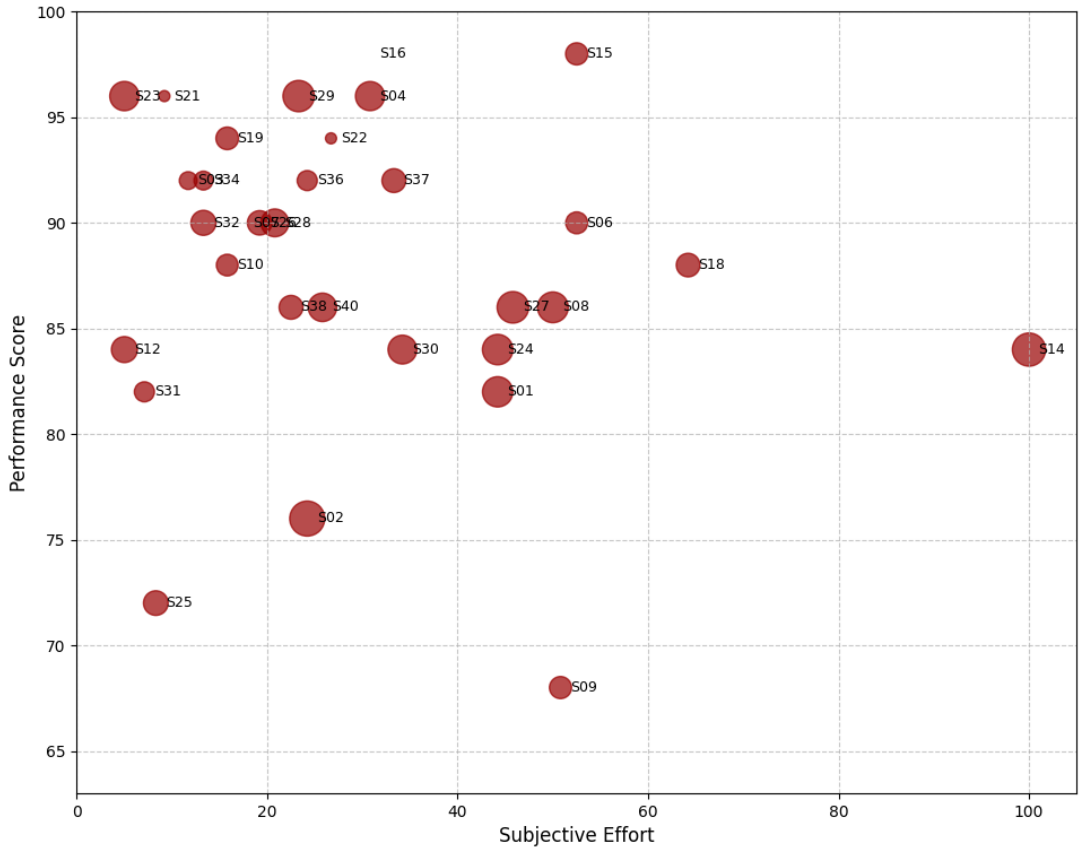


Figure. 9.1. Relationship between effort, performance, and hearing level (PTA)

Relationship between subjective effort, task performance, and hearing level (PTA). Each bubble represents a participant, with position based on subjective effort rating (y-axis) and task accuracy (x-axis), and bubble size corresponding to PTA (poorer hearing = larger bubble). The scatter illustrates substantial individual variability across all three measures. Despite the adaptive design targeting consistent accuracy, performance and effort levels varied widely, and no significant correlations were found between these variables.

Average Trial Response (ATR) Calculation To characterise the EEG response across trials, the Average Trial Response (ATR) was calculated for each participant by averaging the EEG alpha amplitude across all 50 trials. Before the listening task begins, there is a baseline period lasting approximately 10 minutes during which participants relax and watch a documentary. Each trial lasted approximately 18 seconds, resulting in a total experiment duration of 25 to 30 minutes. The same pre-analysis procedure was also performed in GSR and pupillometry data.

Figure 9.3 shows ATR from all the participants with valid EEG data (N=29) from one experiment (participant repeat the same experiment after one week interval). The vertical lines indicate key task events: the cue to memorise digits and the cue to respond.

The figure also shows the event evoked response, as the dip of alpha wave before and after each event. Details of event-related results were published in Alhanbali et al., 2021.

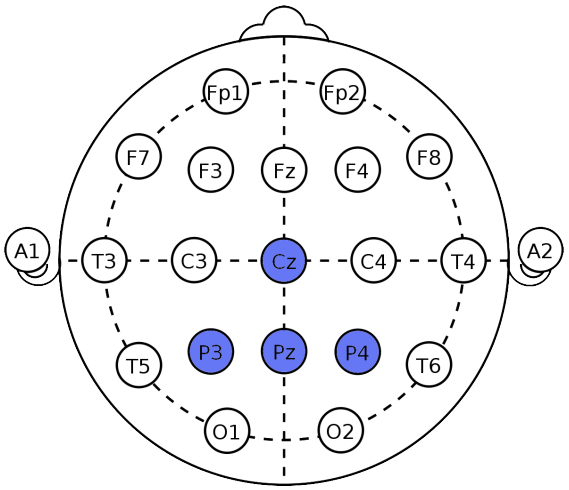


Figure. 9.2. Position of EEG electrodes

Electrode positions used for EEG recording, based on the international 10-20 system. Coloured points indicate the four channels recorded in this study (Pz, P3, P4, Cz). The Pz channel was selected for analysis as it showed representative response patterns, and earlier exploratory tests indicated similar outcomes across all channels.

This study is mainly focusing on exploring individual differences.

Time Course Response (TCR) To examine how the alpha response evolved over the course of the experiment, the Time Course Response (TCR) was computed. This involved averaging all EEG samples within each individual trial, yielding 50 data points per experiment for each participant. The TCR can be viewed as a representation of the change in average alpha activity from trial to trial (see Figure 9.4).

9.2.2 Individual Consistency and Difference

Individual Consistency: Permutation Test of Correlation To statistically evaluate the observation of high within-subject consistency in EEG responses, permutation tests (see Page 74 for detail about this method) were conducted on the correlation coefficients. This non-parametric approach tests whether the similarity of responses within the same individual (across the two experimental sessions) is significantly greater than the similarity between responses from randomly paired individuals.

As shown in Figure 9.5 for ATR data (channel 1 example), the observed average within-subject correlation (red circle) was significantly higher than expected by chance ($p < 0.001$).

A similar significant result was found for the TCR data (Figure 9.6, $p = 0.004$). These findings were consistent across other channels, confirming that individuals' EEG alpha

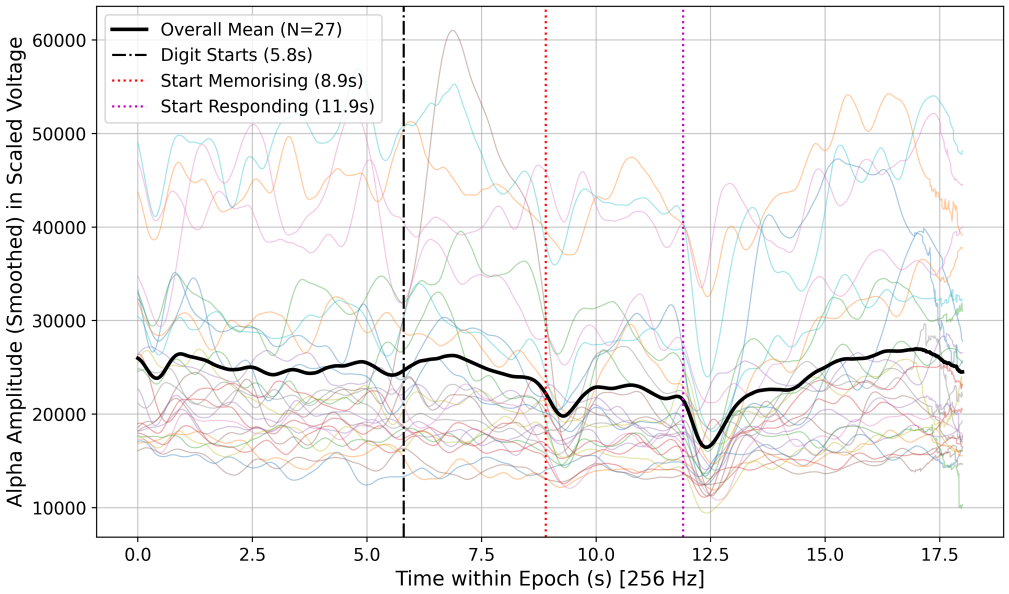


Figure. 9.3. Example of average EEG average trial response (ATR) alpha band for all participants (channel Pz)

The plot shows average alpha amplitude at channel Pz across 50 trials, with vertical lines marking trial onsets (digit starts, memorising/retention start, and responding start)). Each line represents one participant's data: averaging 50 trials across 2 experiments. The black line thick line reflects the mean across all valid trials. Despite variability between trials, a consistent dip in alpha activity can be seen around the onset of different events, indicating a repeatable task-evoked neural response.

responses to the task were significantly more correlated within themselves than with other participants' responses.

Finding Patterns: Cluster Analysis Given the confirmed individual differences, cluster analysis (see page 75 for detail of the method) was employed to investigate whether participants could be grouped based on the *shape* of their average neural response (ATR). The ATR was selected for this shape-based clustering because, it provides a clearer representation of the underlying task-evoked response pattern compared to the TCR; the latter reflects considerable trial-to-trial variability (as seen in Figure 9.6), making it more challenging to identify consistent waveform shapes suitable for grouping.

The analysis was performed on the average ATR across the two experiments for each participant. Each ATR spanned 18 seconds of post-stimulus activity sampled at 256 Hz, every participant was represented by a 4,608-point time-series vector (30 vectors in total, each of dimension 1 x 4608). Prior to clustering, the ATRs were aligned in length, imputed where occasional missing values occurred, baseline-corrected at stimulus onset, and z-normalised.

To determine the optimal number of clusters (k), the elbow method was used, plotting the clustering error (inertia) against k. As seen in Figure 9.7, the plot suggested an optimal k of 2 or 3. Initial clustering with k=3 (Figure 9.8) revealed two clusters with very

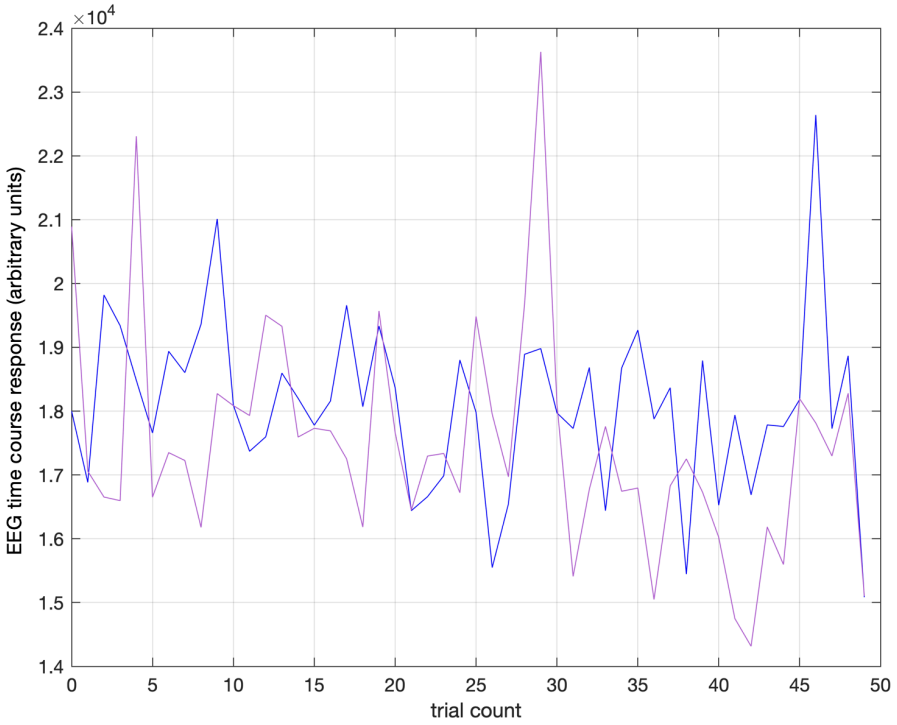


Figure. 9.4. EEG time course response (TCR) for one participant (channel Pz)

The TCR was calculated by averaging alpha amplitude within each trial, resulting in 50 time points per participant. Each line represents one participants average trial-wise alpha response across the experiment. Two different lines represent two experiments(repeated measures). This format captures how neural activity evolves over time, highlighting inter-individual variability and potential trends in task adaptation or fatigue across repeated exposures..

similar average waveforms. Therefore, a 2-cluster solution was adopted, the results of which are shown in Figure 9.9.

This cluster results (Figure 9.8 and 9.9) displays the centroid (average ATR waveform) for each of the two clusters, representing the general trend for that group. The shaded regions indicate the standard error of the mean (SEM) (Cacioppo et al., 2007; Luck, 2014) at each time point, illustrating the variability of individual waveforms within each cluster around the average; narrower bands indicate higher waveform similarity among participants within that cluster.

The two clusters appear to exhibit distinct patterns of alpha modulation during the trial, particularly around the listening phase. In summary, EEG ATR shapes revealed consistent patterns across sessions for most participants, with clusters differing in the timing and prominence of post-retention peaks. These differences suggest individual variability in task-related cortical engagement and possible distinctions in listening or memory effort.

Relationship between EEG Clustering Result and subjective Effort, Hearing Level (PTA), and Performance To determine if the identified EEG response patterns related to

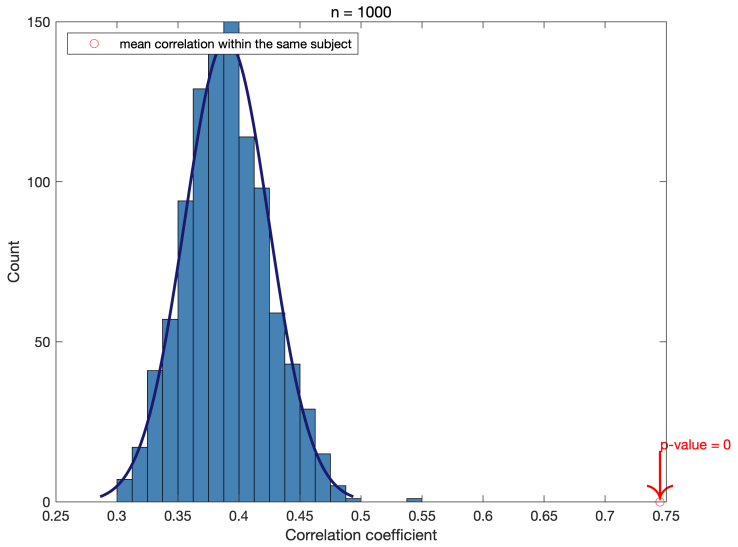


Figure. 9.5. Permutation test of within-subject correlations in EEG average trial response

The red circle marks the average correlation between each participants EEG ATR data across the two sessions. This is compared to a null distribution created by randomly re-pairing data across participants. The observed within-subject correlation was significantly higher than would be expected by chance ($p < 0.001$), confirming consistent individual EEG response patterns across sessions

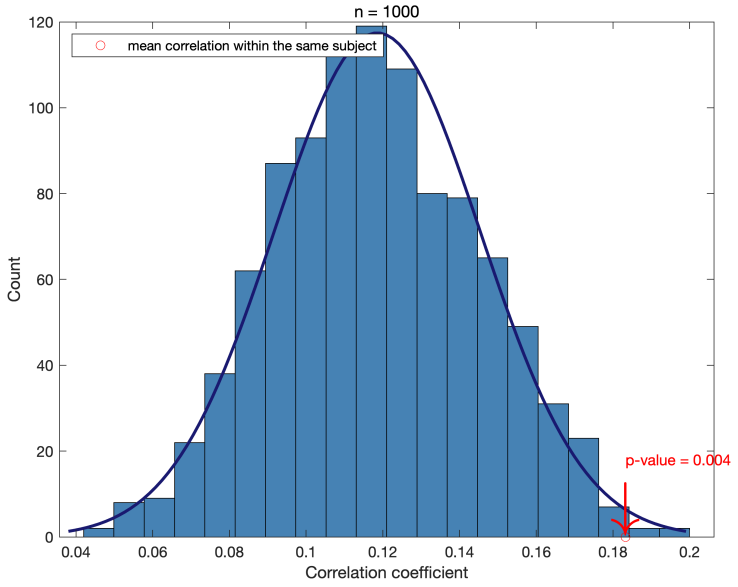


Figure. 9.6. Permutation test of within-subject correlations in EEG time course response

The red circle marks the average correlation between each participants EEG TCR data across the two sessions. This is compared to a null distribution generated by randomly re-pairing participant data. The observed within-subject correlation was significantly higher than expected by chance ($p = 0.004$), indicating that trial-to-trial response trends were consistent within individuals over time.

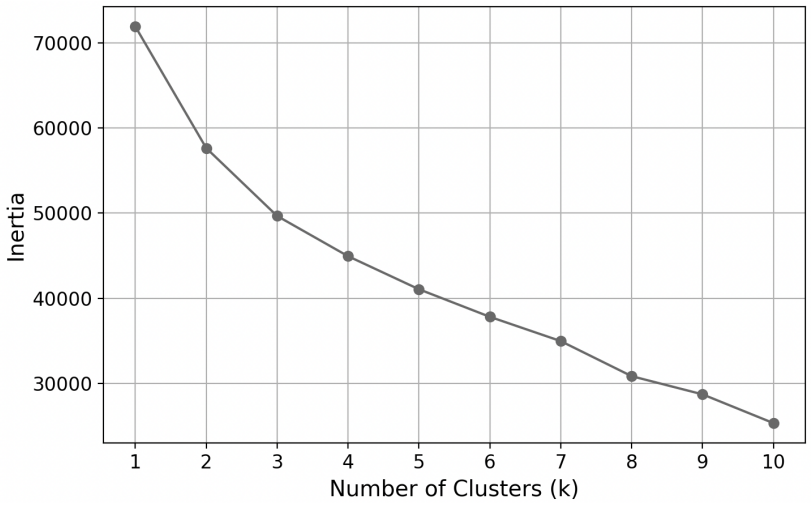


Figure. 9.7. K-means elbow plot for EEG average trial response clustering

The elbow method was used to identify the optimal number of clusters (K) for grouping participants based on their EEG ATR waveforms. The plot displays clustering error (inertia) across different values of K. The turning point (elbow) suggests that 2 or 3 clusters provide the most meaningful separation, with diminishing gains from further increasing K..

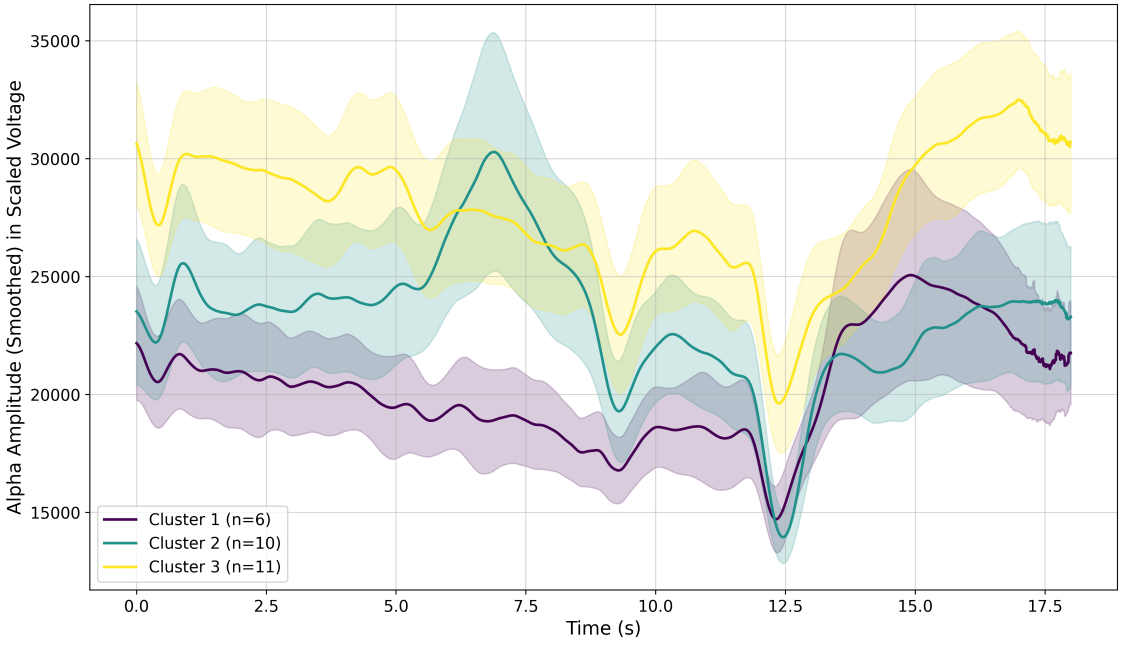


Figure. 9.8. EEG clustering results of alpha-band average trial response (3 groups)

Participants were clustered based on their EEG alpha-band ATR waveforms at channel Pz. Each line represents the average waveform (cluster centroid) for one of the three identified groups, with shaded areas indicating within-cluster variability. Although three distinct clusters were extracted, two showed highly similar shapes, motivating the use of a two-cluster solution for interpretation in subsequent analyses.

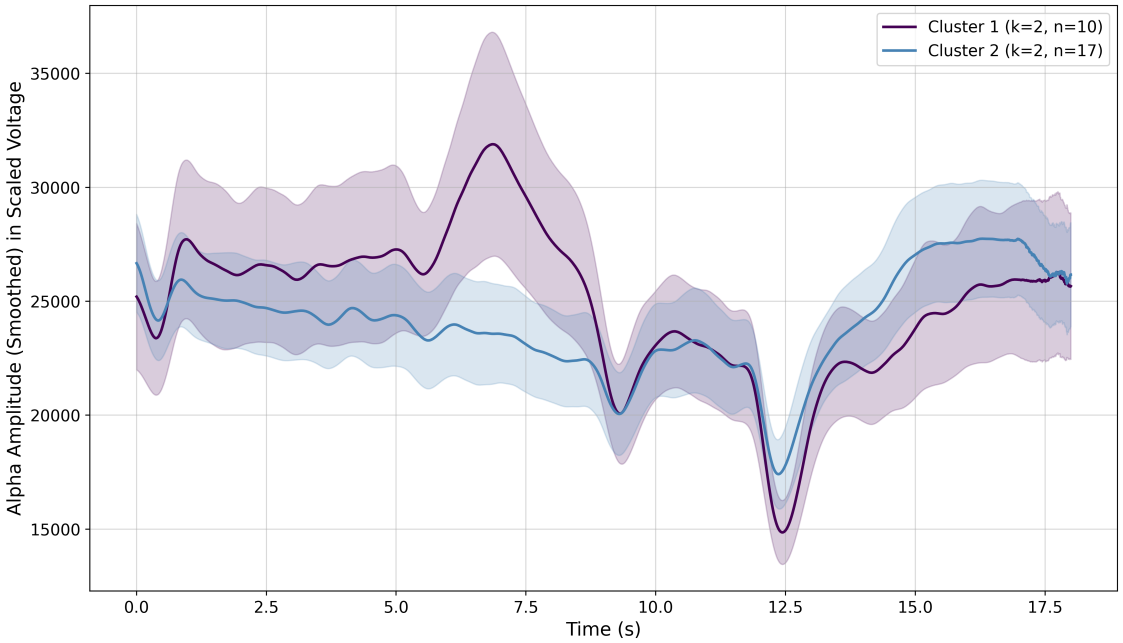


Figure. 9.9. EEG clustering results of alpha-band average trial response (2 groups)

Participants were grouped into two clusters based on the shape of their EEG alpha-band ATR waveforms at channel Pz, averaged across both sessions. The centre lines represent the cluster centroids, showing the typical response shape for each group. Shaded areas indicate the standard error of the mean (SEM) across participants within each cluster. This result were used in subsequent analyses examining relationships with subjective effort, hearing thresholds, and performance.

other measures, the two clusters were compared based on their PTA, subjective effort (Self-report from NASA-TLX), and behavioural performance (Accuracy). Due to the relatively small sample size within clusters (N=18 in Cluster 1, N=11 in Cluster 2 for Self-report) and non-normal distribution of some variables, non-parametric Mann-Whitney U tests were applied.

Figure 9.10 presents the distribution of PTA, Self-report, and Performance for each cluster using boxplots. While some visual differences between the groups might be suggested by the plots, the Mann-Whitney U tests revealed no statistically significant differences between the two EEG clusters for PTA, subjective effort, or performance (all $p > 0.1$, see Table 9.2.

It is noted that sample sizes is limited and varied slightly for comparisons due to missing data for some participants on specific measures. The lack of a significant relationship suggests that these distinct EEG alpha response patterns, while consistent within individuals, may not directly map onto these specific behavioural or subjective outcomes

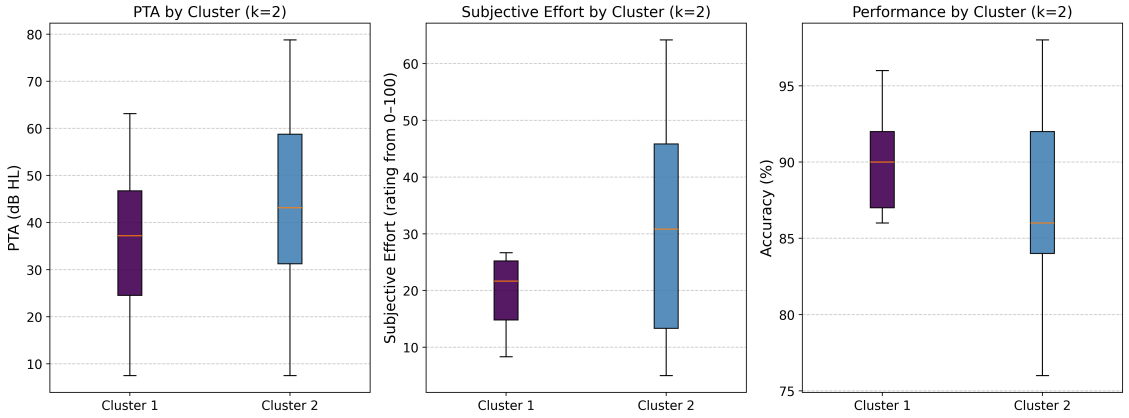


Figure. 9.10. Subjective effort, hearing level (PTA), and performance across EEG alpha ATR cluster groups

Subjective effort, hearing level (PTA), and performance across EEG clustering groups. Boxplots show the distributions of self-reported effort (NASA-TLX), pure-tone average (PTA), and task accuracy for the two EEG alpha ATR clusters ($k = 2$). Colours match the cluster labels in Figure 9.9. Although visual differences appear between groups, Mann-Whitney U tests revealed no statistically significant differences for any measure (all $p > 0.1$; see Table 9.2). These findings suggest that differences in EEG alpha response patterns may not directly correspond to behavioural or subjective outcomes.

Table. 9.2. Mann-Whitney U Test Results Comparing Cluster Groups (k=2) on PTA, Subjective Effort and Performance

Variable	N (Cluster 1 / Cluster 2)	Median (C1)	Median (C2)	U Statistic	p-value
PTA	10 / 17	37.19	43.13	59.5	0.209
Self-report	10 / 17	21.67	30.83	66.0	0.352
Performance	10 / 17	90.0	86.0	105.0	0.324

Note: C1 = Cluster 1, C2 = Cluster 2, based on k=2 clustering. P-values rounded to three decimal places. Medians and U statistic rounded to two decimal places. Sample size inside the groups were not consistent due to missing data.

9.3 Galvanic Skin Response (GSR)

9.3.1 Data Pre-processing

Galvanic Skin Response (GSR), also known as skin conductance or electrodermal activity, was measured as another physiological indicator of listening effort, reflecting sympathetic nervous system arousal. Data were recorded at the same time with EEG using electrodes placed on the participant's non-dominant hand, sampled at 32 Hz. The analysis procedure for GSR data followed the same process outlined for the EEG data. Raw data were first inspected and cleaned to remove artefacts.

Average Trial Response (ATR) and Time Course Response (TCR) Caculation The Average Trial Response (ATR) for GSR was calculated by averaging the baseline–corrected skin conductance level across all 50 trials for each participant. An example of a participant's GSR ATR is shown in Figure 9.11. This represents the average phasic response related to the cognitive demands of the listening and memory task within a trial.

The Time Course Response (TCR) was also computed by averaging all GSR samples within each trial, yielding 50 data points per experiment. This aimed to illustrate how average GSR levels changed over the duration of the experiment.

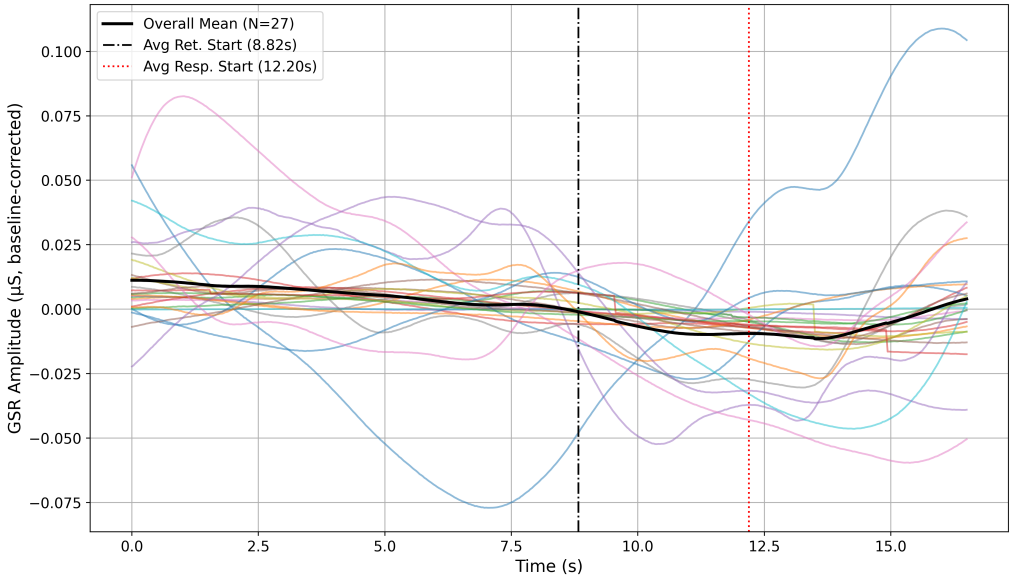


Figure. 9.11. Example of GSR average trial response(ATR) from all participants from one experiment

Average trial response(ATR) was calculated for participants skin conductance response (across 50 trials). Vertical lines indicate trial onsets. The black line shows the average of all response The plot reflects the typical phasic GSR pattern evoked by the listening and memory task, showing event-related modulations in arousal level across trials.

9.3.2 Individual Consistency and Difference

Permutation Test of Correlation (Individual Difference) As with the EEG data, permutation tests were performed to statistically assess within-subject consistency for GSR responses. The tests confirmed that the correlation between a participant's own two ATR recordings was significantly higher than the correlation between randomly paired participants (Figure 9.12, $p < 0.001$). A similar significant result was found for the TCR data (Figure 9.13, $p < 0.001$). This indicates that individuals' GSR response patterns during the task were highly consistent within themselves across the two sessions.

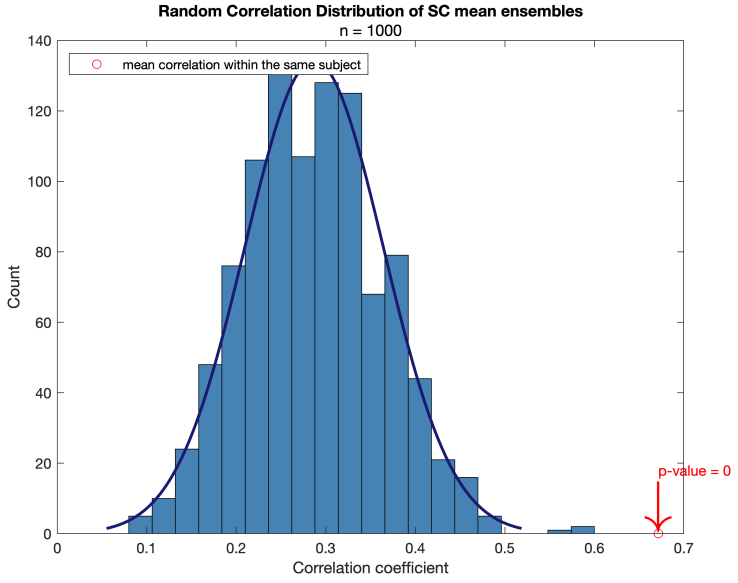


Figure. 9.12. Permutation test of within-subject correlations in GSR average trial response

The red circle indicates the average correlation between each participants GSR ATR data across the two sessions. This is compared against a null distribution created by randomly re-pairing participants. The observed within-subject correlation is significantly higher than would be expected by chance ($p < 0.001$), demonstrating that individuals exhibited consistent skin conductance response patterns across sessions.

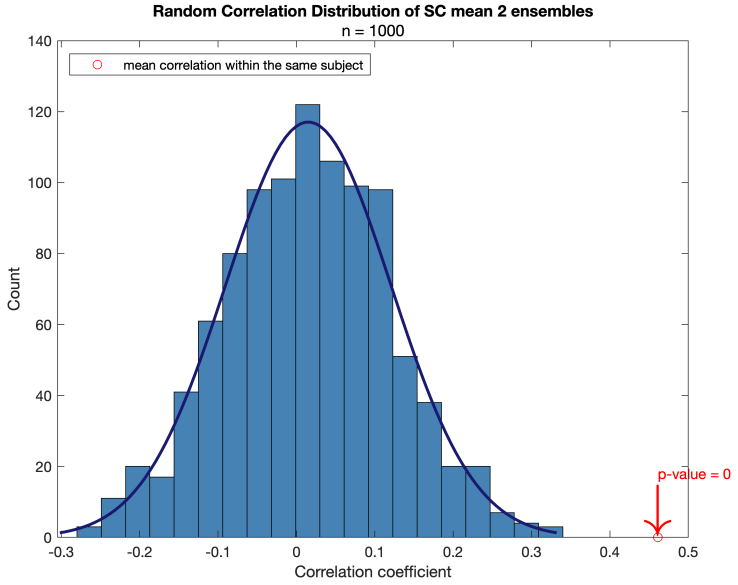


Figure. 9.13. Permutation test of within-subject correlations in GSR time course response

Permutation Test of Pupillometry TCR in Correlation. The red circle is the average correlation between the same participant's data (experiment 1 and 2). The p−value is significant, which means that the correlation within the same participant's data is significantly higher than the correlation between random participants' data.

Finding Patterns: Clustering Analysis To explore potential subgroupings based on GSR response patterns, K-means cluster analysis was applied to the GSR ATR data, again using correlation as the distance metric. The ATR was chosen over the TCR for shape-based clustering due to its clearer representation of the average task-evoked response pattern, free from the higher trial-to-trial variability inherent in TCR which complicates pattern identification.

The elbow method (Figure 9.14) suggested an optimal number of 2 or 3 clusters. After examination, a 2-cluster solution was adopted. Figure 9.15 presents the results, showing the centroid (average GSR ATR waveform) and SEM for each cluster. The two clusters appear to represent different temporal dynamics in the average skin conductance response during the trials.

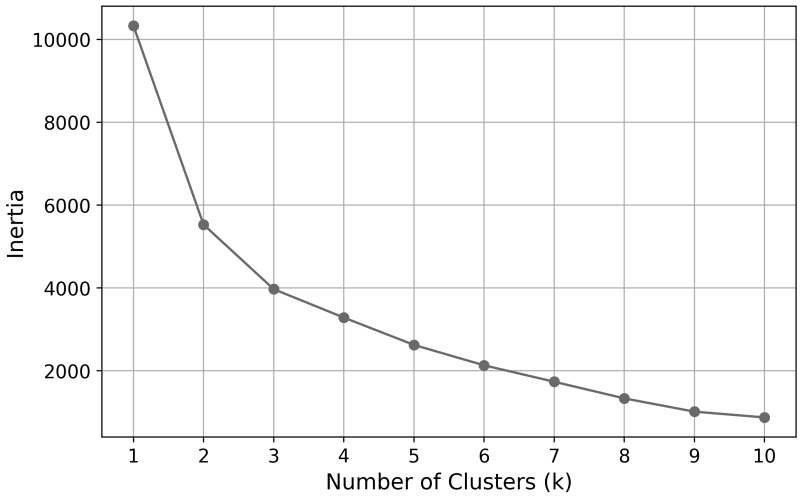


Figure. 9.14. K-Means Elbow Result of GSR

The elbow method was used to determine the optimal number of participant clusters based on GSR ATR waveforms. The plot shows the within-cluster sum of squares (WCSS) across different values of K. A sharp decrease followed by a plateau suggests that 2 or 3 clusters provide the most meaningful separation of participants based on their skin conductance response patterns.

GSR ATR waveforms showed typical sympathetic arousal patterns, with group differences in peak amplitude and recovery slope. These patterns may reflect varying levels of physiological effort or arousal across participants.

Relationship between Clustering and Subjective Effort The relationship between these GSR-based clusters and other participant measures (PTA, subjective effort, performance) was investigated using Mann-Whitney U tests, appropriate for the sample sizes and potential non-normality. The distributions of these variables for each cluster are visualised in Figure 9.16

Similar to the findings for the EEG clusters, despite any visual suggestions in the plots, no statistically significant differences were found between the two GSR clusters for PTA,

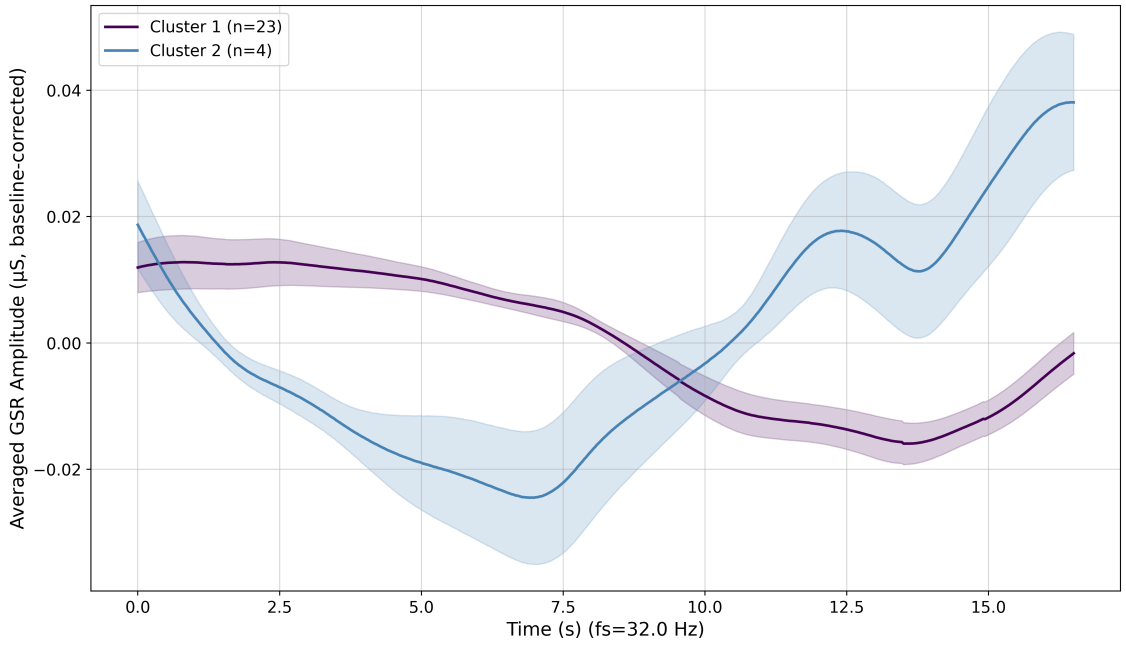


Figure. 9.15. GSR clustering results of average trial response (2 groups)

Participants were grouped into two clusters based on the shape of their average skin conductance response (GSR ATR). Each line represents the centroid of a cluster, reflecting the typical GSR waveform for that group. Shaded regions represent within-cluster variability (1 SEM). The two clusters display distinct temporal profiles, suggesting differences in autonomic engagement during the task.

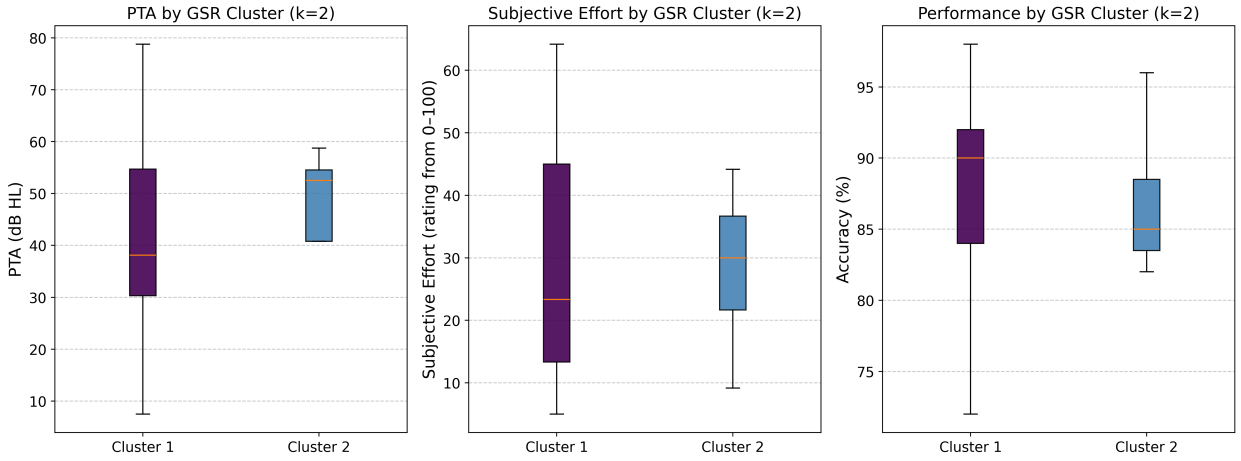


Figure. 9.16. Subjective effort, hearing level (PTA), and performance across GSR alpha ATR cluster groups

Subjective effort, hearing level (PTA), and performance across GSR ATR clusters. Boxplots show the distributions of NASA-TLX scores (subjective effort), pure-tone average (PTA), and task accuracy across the two GSR clustering groups ($k = 2$). Colours match cluster labels from Figure 9.15. Despite apparent visual trends, Mann-Whitney U tests found no statistically significant differences between clusters (all $p > 0.6$; see Table 9.3), suggesting that physiological response patterns did not align with behavioural or subjective outcomes.

subjective effort (Self-report), or performance (all $p > 0.6$, see Table 9.3). Sample sizes for these comparisons varied (e.g., N=23/4 for PTA, N=23/5 for Self-report and Performance) due to missing data across different measures. These results suggest that the distinct patterns identified in average GSR responses do not straightforwardly align with individual differences in hearing thresholds, overall subjective workload ratings, or task accuracy in this study.

Table. 9.3. Mann-Whitney U Test Results Comparing GSR Cluster Groups (k=2) on PTA, Subjective Effort, and Performance.

Variable	N (Cluster 1 / Cluster 2)	Median (C1)	Median (C2)	U Statistic	p-value
PTA	23 / 4	38.13	52.50	41.50	0.785
Subjective Effort	23 / 4	23.33	30.00	41.50	0.785
Performance	23 / 4	90.00	85.00	54.50	0.583

Note: C1 = Cluster 1, C2 = Cluster 2, based on k=2 clustering of SC data. Ns show sample sizes for each cluster in the comparison for that variable (after dropping NaNs). P-values rounded to three decimal places. Medians and U statistic rounded to two decimal places. Sample size inside the groups may differ between variables due to missing data.

9.4 Pupillometry

9.4.1 Data Pre-processing

Pupillometry, the measurement of pupil diameter, was utilised as a further physiological index of listening effort, reflecting cognitive load and autonomic nervous system activity. Pupil size data were recorded using an Eyelink 1000 eye-tracker at a sampling frequency of 1000 Hz. To ensure stable recordings, participants used a chin rest, and room lighting and screen brightness were individually adjusted prior to data collection.

Preprocessing steps typically include artefact removal (e.g., correcting for blinks), although the units in Figure 9.17 are noted as arbitrary and the caption questions whether baseline correction was applied; we assume standard preprocessing suitable for task-related analysis was performed.

Individual Consistency: Permutation Test of Correlation The consistency of pupillary responses within individuals across the two experimental sessions was assessed using permutation tests on correlation coefficients. The results demonstrated that the correlation between a participant's own two ATR recordings was significantly higher than that between randomly paired participants (Figure 9.18, $p < 0.001$). A similar

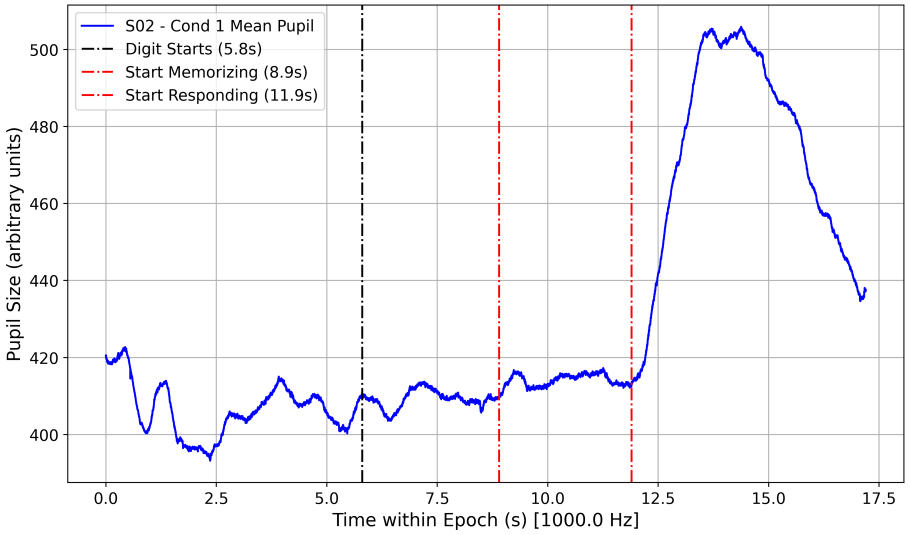


Figure. 9.17. Example of Pupillometry average trial response

Pupil diameter traces from one participant, averaged across 50 trials. Vertical lines indicate the onset of each trial. The plot illustrates task-evoked pupillary responses during the listening and memory task, reflecting event-related changes in cognitive load over time.

significant finding was obtained for the TCR data (Figure 9.19, $p < 0.001$). These outcomes confirm that, like the EEG and GSR responses, individual pupillometry patterns during the task were highly consistent within participants.

9.4.2 Finding Patterns: Cluster analysis

K-means cluster analysis was applied to the pupillometry ATR data to identify potential subgroups based on the shape of the average task-evoked pupillary response, using correlation as the distance metric. The ATR was selected for this shape-based clustering because, by averaging across trials, it provides a clearer representation of the underlying task-evoked response pattern compared to the TCR; the latter reflects considerable trial-to-trial variability, making it more challenging to identify consistent waveform shapes suitable for grouping.

The elbow method (Figure 9.20) suggested an optimal number of 2 or 3 clusters. Consequently, a 2-cluster solution was examined, as presented in Figure 9.21. The figure shows the centroid (average pupil ATR waveform) and SEM for each cluster. Visual inspection suggests the two clusters primarily differ in the magnitude and potentially the timing of the peak task-related pupil dilation.

Pupillometry ATR traces revealed groups with more sustained dilation during retention versus those with earlier constriction, suggesting differences in cognitive load or sustained effort during the memory phase.

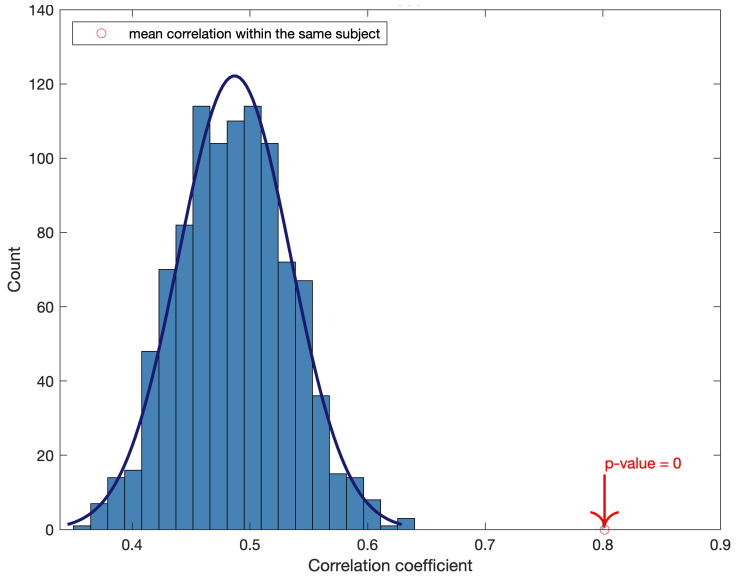


Figure. 9.18. Permutation test of within-subject correlations in pupil diamter average trial response

The red circle indicates the average correlation between each participants ATR data across the two experimental sessions. This is compared against a null distribution generated by randomly re-pairing participants. The observed within-subject correlation is significantly higher than expected by chance ($p < 0.001$), confirming consistent individual pupillary response patterns across sessions.

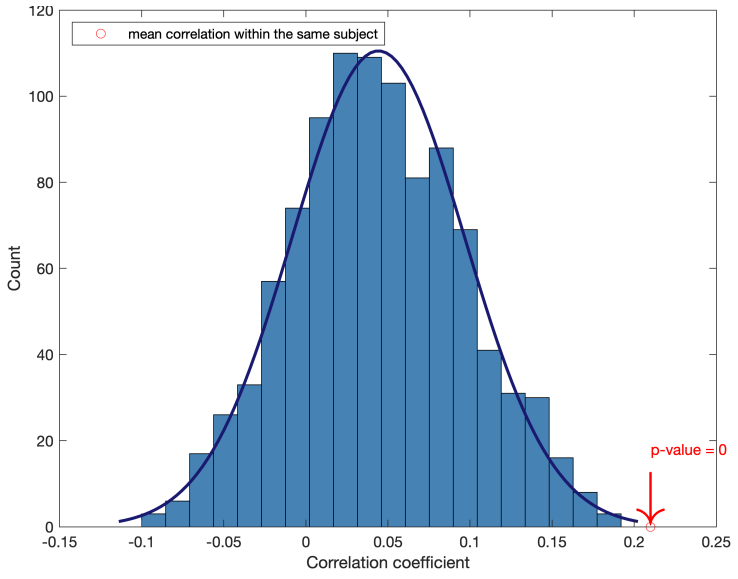


Figure. 9.19. Permutation test of within-subject correlations in pupil diameter time course response

The red circle indicates the average correlation between each participants TCR data across the two sessions. This is compared against a null distribution generated by randomly re-pairing participants. The significantly higher within-subject correlation ($p < 0.001$) supports the presence of stable trial-wise pupil response patterns across experimental sessions.

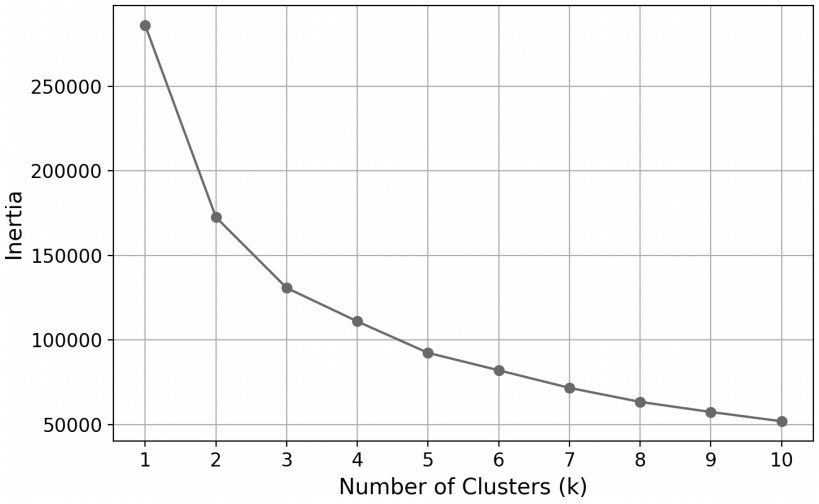


Figure. 9.20. K-means elbow plot for pupil diameter average trial response clustering

The elbow method was used to identify the optimal number of clusters for pupil ATR data. The plot shows the within-cluster sum of squares (WCSS) across increasing values of K. The turning point around $K = 2$ or 3 suggests that these values provide the most meaningful grouping of participants based on their average pupillary response shape

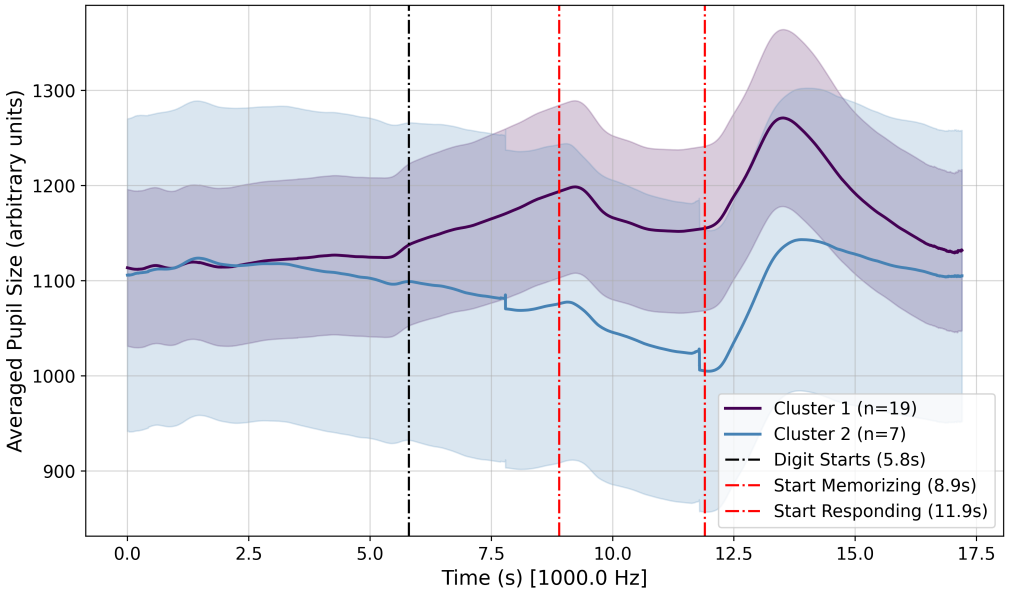


Figure. 9.21. Pupil diameter clustering results of average trial response (2 groups)

Participants were grouped into two clusters based on the shape of their average pupillometry response. Each line represents the cluster centroid, showing the mean pupil waveform for that group, and the shaded regions indicate the standard error of the mean (SEM). The y-axis reflects pupil size in arbitrary units, corresponding to the eye-tracker's native pupil-size measure rather than physical diameter; values therefore represent relative differences in pupil dilation over time. Vertical dashed lines mark the onset of the digit sequence (5.8 s), the memorisation period (8.9 s), and the response phase (11.9 s). The two clusters differ primarily in the magnitude and timing of task-evoked pupil dilation, suggesting distinct physiological response patterns during the listening and memory task.

Relationship between Pupillometry and Hearing Level (PTA), Subjective Effort, and Performance Finally, the relationship between the two pupillometry-derived clusters and the participants' PTA, subjective effort (Self-report), and performance was examined using Mann-Whitney U tests. The distributions are illustrated in Figure 9.22. Consistent with the findings for EEG and GSR clusters, no statistically significant differences were found between the two pupillometry clusters for PTA, subjective effort, or performance (all p > 0.3, see Table 9.4).

It is important to note the differing sample sizes within the clusters for these comparisons (N=19 in Cluster 1, N=7 in Cluster 2) and that performance data were noted as unavailable for this specific cluster comparison in the original analysis documentation. Thus, the distinct average pupillary response patterns identified did not significantly correspond to variations in hearing thresholds or subjective effort ratings among these participants.

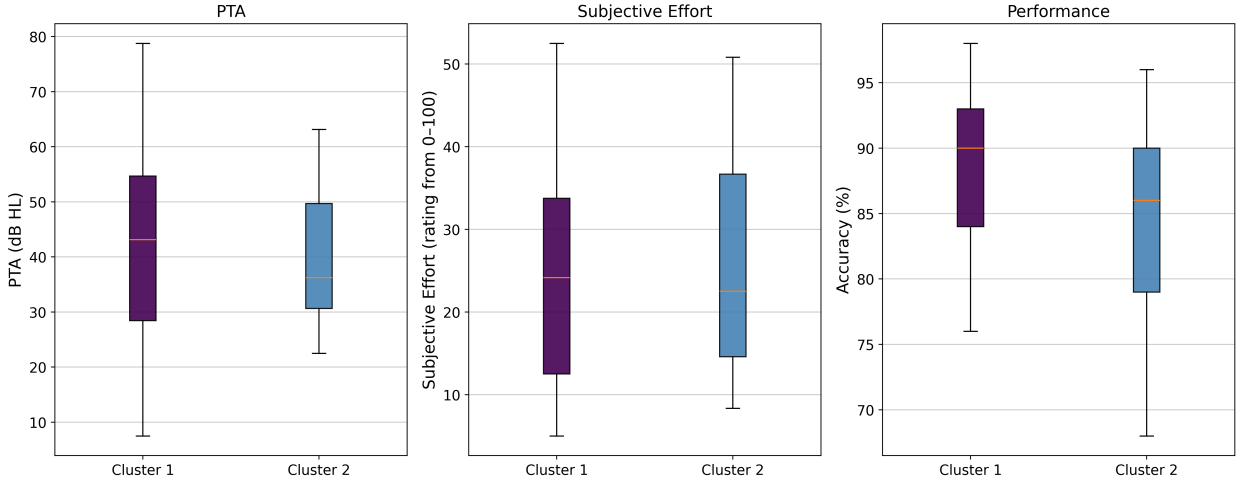


Figure. 9.22. Subjective effort, hearing level (PTA), and performance across pupil diameter ATR cluster groups

Subjective effort, hearing level (PTA), and performance across pupillometry ATR clusters. Boxplots show the distribution of NASA-TLX scores (self-reported effort), pure-tone average (PTA), and task accuracy across the two participant groups identified by pupillometry ATR clustering ($K = 2$). Cluster colours match those in Figure 9.21. Despite visual trends, Mann-Whitney U tests revealed no statistically significant group differences (all $p > 0.3$; see Table 9.4). This suggests that the physiological pupil response patterns do not directly map onto behavioural or subjective outcomes.

9.5 Clustering Agreement between Physiological Measures

Having derived 2-cluster solutions independently for each of the three physiological measures (EEG alpha ATR, GSR ATR, and Pupillometry ATR), this section examines the extent to which these different measures classified participants into similar groups.

Table. 9.4. Mann-Whitney U Test Results Comparing Pupillometry Cluster Groups (k=2) on PTA and Subjective Effort.

Variable	N (Cluster 1 / Cluster 2)	Median (C1)	Median (C2)	U Statistic	p-value
PTA	7 / 19	36.25	43.13	60.00	0.729
Subjective Effort	7 / 19	22.50	25.00	66.50	0.868
Performance	7 / 19	86.00	90.00	51.00	0.303

Note: C1 = Cluster 1, C2 = Cluster 2, based on k=2 clustering of Pupillometry data. Ns show sample sizes for each cluster in the comparison for that variable (after dropping NaNs). Results for 'Performance' were not available. P-values rounded to three decimal places. Medians and U statistic rounded to two decimal places. Sample size inside the groups may differ between variables due to missing data.

Understanding the agreement between these cluster assignments provides insight into whether these distinct physiological signals capture related underlying response patterns to the listening task.

Figure 9.23 presents a detailed visualisation of the clustering agreement across the three measures for each participant included in the overlapping analysis. Participant identifiers are shown along the x-axis, and the corresponding three-digit code below each indicates the cluster assignment (1 or 2) for Pupillometry, GSR, and EEG, respectively. This allows for inspection of individual agreement patterns (e.g., participant S29 assigned to '211', meaning Cluster 2 for Pupillometry, Cluster 1 for GSR, and Cluster 1 for EEG).

To summarise the overall consistency, Figure 9.24 illustrates the percentage of participants who were assigned to the same cluster (i.e., both assigned to Cluster 1 or both assigned to Cluster 2) across pairs of physiological measures, and across all three measures. The highest level of agreement was observed between the EEG and GSR clustering results, with 64% of participants falling into the same cluster category for both measures.

Agreement involving pupillometry was considerably lower (Pupil-GSR: 40%; Pupil-EEG: 28%). Notably, only a small fraction of participants (16%, n=4) were classified into the same cluster group across all three physiological measures simultaneously. This pattern suggests a greater degree of similarity between the response types captured by EEG alpha and GSR in this task, compared to pupillometry, and highlights substantial divergence in participant classification depending on the physiological measure used.

Finally, an exploratory analysis examined whether participants who showed agreement in their cluster assignments across different measure combinations also exhibited distinct characteristics in terms of their hearing level (PTA), self-reported effort, or task performance. Figure 9.25 displays the average PTA, Self-Reported Effort, and Performance scores for groups of participants defined by their agreement status (e.g., those agreeing on Pupil-GSR clusters, those agreeing on GSR-EEG clusters, etc.). While

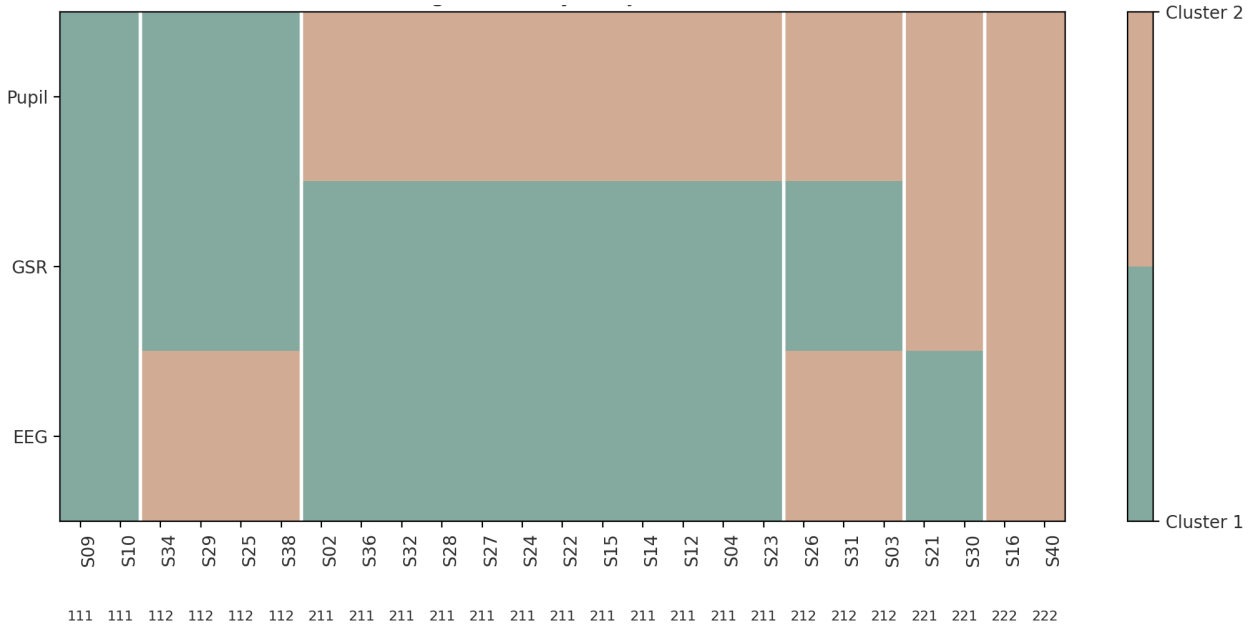


Figure. 9.23. Clustering Group Agreement Across Different Physiological Measures

The x-axis shows participant numbers, and the y-axis represents different physiological measures.

Clustering group agreement across three physiological measures: GSR, EEG, and pupillometry. A two-cluster solution was used. The digits at the bottom indicate the group assignment for each measure-'1' for cluster 1, and '2' for cluster 2. The three-digit strings represent the grouping pattern for each participant across the three measures, in the order: pupillometry, GSR, and EEG.

It can be observed that the highest percentage of agreement in group membership occurs between EEG and GSR.

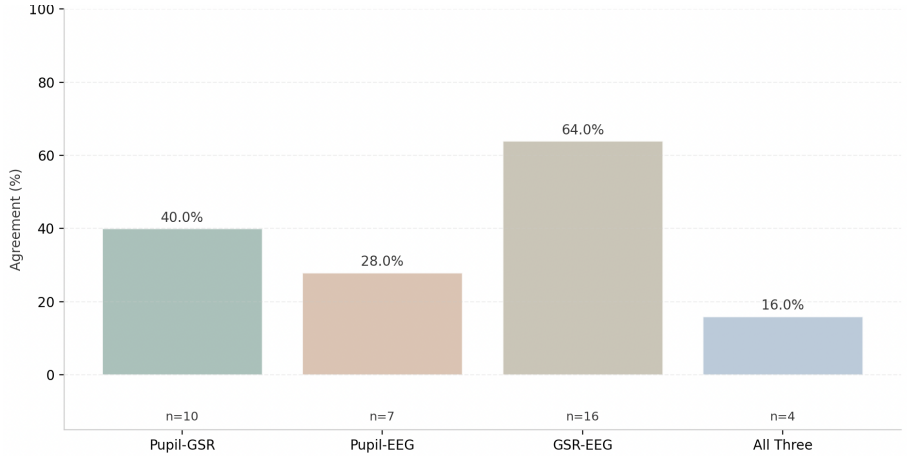


Figure. 9.24. Clustering Group Agreement Across Different Physiological Measures (overlapping meausres)

Clustering group agreement across three physiological measures, focusing on the overlapping measures, instead of detailed groups as Figure 9.23. It shows thatGSR and EEG clustering groups has the most overlapping members in their grouping result, resulting 64% of the participants.

This result shows higher compatibility of participants' response of EEG and GSR, comparing to Pupillometry and GSR, or Pupillometry and EEG. It's notable that only 4 participants were grouped in the same clustering group for all three measurements, indicating disagreement across different measures of listening effort.

some visual differences between these agreement groups are apparent in the plots, subsequent statistical comparisons indicated no significant differences between these groups for PTA, Self-Reported Effort, or Performance.

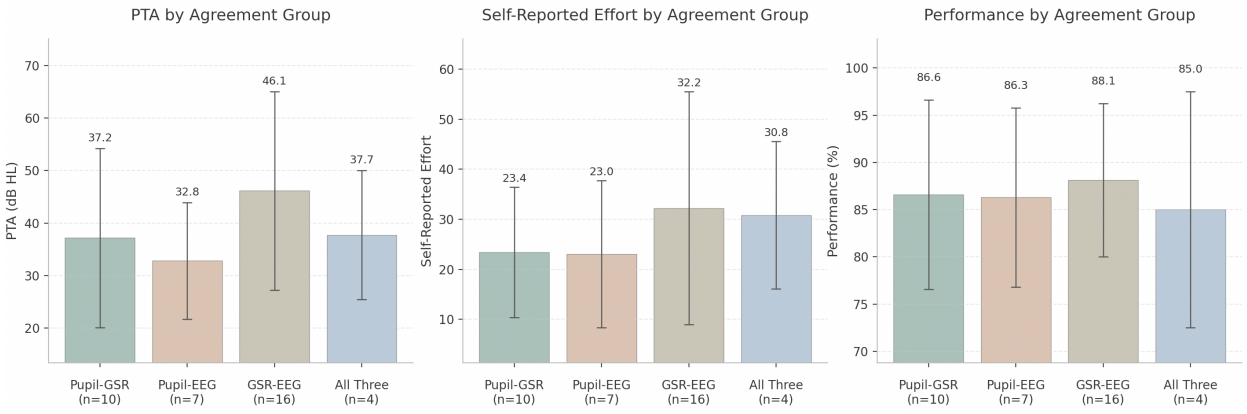


Figure. 9.25. Subjective Effort, PTA, and Performance over Clustering Group Agreement

Result of hearing level(PTA), Self-reported effort, and Performance, over those groups which agrees on K-Means clustering. There's a visible difference between each group, though no statistic significance was found between those groups.

In summary, the analysis of clustering agreement revealed limited concordance across the three physiological measures, particularly when involving pupillometry. The highest agreement found between EEG and GSR suggests these two measures might reflect more closely related aspects of the physiological response in this context. The overall low agreement implies that EEG alpha, GSR, and pupillometry likely capture different, potentially complementary, facets of an individual's complex response to listening effort.

9.6 Summary of Key Findings

The results presented in the previous chapter provide several key insights into the subjective, behavioural, and physiological responses of older adults with hearing impairment to a demanding listening task.

Significant Individual Variability in Subjective and Behavioural Responses The study highlights the significant individual variability inherent in this population. Both self-reported listening effort, measured using the NASA-TLX, and behavioural performance accuracy demonstrated considerable variation across participants. This occurred even when the signal-to-noise ratio was individually adapted to target a consistent performance level (71% accuracy), indicating that factors beyond basic audibility profoundly influence the listening experience and task outcome.

Performance Variation Beyond Audibility The substantial range in performance accuracy (68–98%) underscores the importance of cognitive factors, such as the working memory load imposed by the 6-digit task, and potentially the effectiveness of individual hearing aid processing in challenging conditions, extending beyond simple audibility adjustments achieved via SNR adaptation.

Weak Association Between Standard Measures The investigation into the relationships between standard clinical or global measures revealed weaker-than-expected associations in this cohort. While the correlations between hearing loss (PTA), subjective effort, and performance accuracy trended in the expected directions (e.g., poorer hearing associated with greater effort and lower performance), none reached statistical significance.

This lack of significant findings might be attributed to the study's sample size, the high degree of inherent variability within the participant group, the limitations of the specific measures used (PTA, NASA-TLX), or potentially the adaptive SNR procedure itself masking some underlying relationships.

Consistent Individual Physiological Signatures The physiological data revealed that individuals possess highly consistent and reliable physiological response patterns. Within-subject consistency across repeated experimental sessions was statistically confirmed for EEG alpha activity, galvanic skin response (GSR), and pupillometry using permutation tests. This suggests that individuals have distinct, stable physiological 'signatures' when engaging in this listening task.

Distinct Physiological Patterns: Clustering Results Despite this high within-subject consistency, clear differences exist *between* individuals in these physiological patterns. Cluster analysis based on the average trial response (ATR) shape identified distinct subgroups (typically two clusters) for each physiological measure.

Physiological Patterns Do Not Directly Map to Global Measures However, a key finding was that these physiologically-defined clusters did not significantly align with variations in participants' hearing thresholds (PTA), overall subjective effort ratings (NASA-TLX), or behavioural performance (Accuracy). This implies that the specific temporal dynamics captured by these physiological measures reflect aspects of neural or autonomic processing that are not directly mapped by these broader outcome variables.

Divergence Between Physiological Measures The study found limited agreement in how participants were classified across the different physiological measures. Comparing the cluster assignments derived independently from EEG, GSR, and pupillometry

revealed low agreement overall, with only 16% of participants falling into the same cluster category across all three modalities.

Agreement was highest between EEG and GSR (64%), suggesting these two measures might capture more closely related physiological processes in this context than pupillometry. This divergence strongly indicates that EEG alpha, GSR, and pupillometry likely reflect distinct facets of the complex, multidimensional response to listening effort.

9.7 Discussion

This study explored listening effort in a group of older adults with varying degrees of hearing loss, using a digit-in-noise task where the signal-to-noise ratio (SNR) was adaptively adjusted to target approximately 71% accuracy. However, despite this adaptive design, participants' actual performance varied notably, as did their reported effort using the NASA-TLX (Hart & Staveland, 1988). This outcome underscores the presence of substantial individual differences, even when performance is normalised by design (Koelewijn et al., 2012; Zekveld et al., 2011).

No statistically significant correlations were found between hearing thresholds (PTA), subjective effort, and task accuracy. While trends followed expected directions-such as poorer hearing being loosely linked to greater perceived effort and reduced performance-the absence of significant effects may reflect limitations in sample size or variability inherent to the older, hearing-impaired group.

Physiological measures (EEG alpha, GSR, and pupillometry) provided further insight into effort-related responses. A particularly clear outcome was the strong within-subject consistency across the two sessions spaced one week apart. Permutation tests confirmed that both average trial responses (ATR) and trial-by-trial time courses (TCR) were more similar within individuals than between them, suggesting reliable individual physiological profiles under these testing conditions.

However, clustering analyses revealed a disconnect between these physiological profiles and other data. For each modality- EEG alpha, GSR, and pupil size-two distinct clusters were identified, reflecting different characteristic response shapes across participants. Yet these clusters were not significantly associated with hearing level, effort ratings, or task performance. This echoes prior findings suggesting that physiological and subjective measures of effort may not always align (Ohlenforst et al., 2017).

Additionally, there was minimal consistency in how participants were grouped across the different physiological modalities. Pupillometry, in particular, showed low agreement with EEG and GSR, although moderate alignment was observed between the latter two (64%). Only a small proportion of participants (16%) were consistently assigned to the same cluster across all three measures. This suggests that these modalities may capture distinct physiological aspects of listening effort (Gagné et al., 2017; Pichora-Fuller et al., 2016b), further highlighting the challenge of identifying a singular physiological marker.

Several limitations should be acknowledged. The modest sample size (around 30) may have reduced the ability to detect subtle effects or between-group differences. Additionally, heterogeneity within the participant group likely contributed to variability in subjective and behavioural data.

There were also methodological constraints. The NASA-TLX, while widely used, may not be specific enough to isolate the perceptual and cognitive demands of listening. Similarly, the adaptive SNR procedure-although designed to equalise difficulty-did not fully achieve its goal, adding interpretive complexity. Finally, missing data on certain physiological measures further reduced the number of participants included in some analyses.

9.8 Conclusion

Study 1 examined listening effort in older adults with hearing loss using a digit-in-noise task, adaptive SNR, subjective workload ratings (NASA-TLX), performance accuracy, and physiological responses (EEG alpha, GSR, and pupillometry). While subjective effort and behavioural performance varied widely across individuals, physiological responses showed strong within-subject consistency across sessions.

Clustering revealed distinct response patterns within each modality, but these were not linked to hearing thresholds, subjective ratings, or task outcomes. Cross-modality agreement was limited, implying that EEG, GSR, and pupillometry capture different facets of the physiological processes underlying listening effort.

Taken together, these findings highlight the individual nature of physiological effort responses and the complexity of linking them to more traditional behavioural and subjective outcomes. They also point to the value of multimodal approaches when investigating listening effort, especially in heterogeneous populations.

Part III

Study 2: Listening Effort Experiment

Chapter 10

Introduction and Research aims

10.1 Introduction and Connection to Study 1

Following the secondary analysis presented in Study 1, which characterised individual physiological response patterns (EEG, GSR, Pupillometry) and their consistency in older adults with hearing impairment, Study 2 embarks on a primary experimental investigation designed to extend these findings and address remaining questions for this project. Study 1 confirmed significant individual variability and notable within-subject consistency in physiological responses to listening effort (Faculty Ethics Commitee reference: ERGO/FEPS/ 87716).

However, it also highlighted the limited association between these physiological patterns and behavioural or subjective outcomes under adaptively controlled signal-to-noise ratios (SNRs), reinforcing the known physiology-behaviour gap. Furthermore, the adaptive SNR design precluded a systematic examination of how varying levels of task difficulty influence these responses, and the analysis was limited to the available EEG, GSR, and pupillometry measures.

Study 2 aims to build directly upon these findings and address these limitations. By conducting a new experiment with normal-hearing participants, a key change is the use of four fixed SNR levels (-16, -11, -6, 12 dB), allowing for a systematic investigation of how varying degrees of task difficulty modulate physiological and behavioural responses across a wide performance range.

Furthermore, Study 2 employs a more complex and arguably more ecologically valid sentence-in-noise task (OLSA) compared to the digit recall task in Study 1. The physiological assessment is also broadened to include ECG and respiration measurements alongside single-channel EEG (Pz), GSR, and pupillometry, providing a more comprehensive multi-system perspective on autonomic regulation during effortful listening.

10.2 Aim of Study

The overall aim of Study 2 is therefore to systematically investigate how task difficulty (manipulated via fixed SNRs) influences dynamic physiological responses across multiple systems, behavioural performance, and subjective experience in normal-hearing listeners. It further aims to explore the nature of individual differences, response consistency, and the relationships between these various measures under controlled conditions, using insights from the temporal dynamics of the trial (ATR in study 1) of physiological signals.

10.3 Research Questions and Hypotheses

Study 2 is guided by the following specific research questions and corresponding hypotheses:

Research Question 2.1 How does using a more complex listening task (speech-in-noise sentences) influence the individual variability and consistency relationships observed in Study 1 within a normal-hearing population?

- Hypothesis 2.1 (Consistency): Individuals will exhibit statistically significant within-subject consistency in physiological responses (Pupil diameter, GSR, HR, RR) across repeated sessions.

Research Question 2.2 How do additional physiological measures (Respiration rate, Heart Rate/ ECG) respond during effortful listening, particularly in relation to varying task difficulty?

- Hypothesis 2.2 (New Measures Modulation): HR (derived from ECG) and RR will show significant task-evoked modulation (i.e., distinct changes related to task events like listening vs. retention). Furthermore, their dynamics (e.g., HR recovery during retention, change in RR during retention) will be significantly sensitive to SNR level, reflecting differential autonomic regulation based on task difficulty.

Research Question 2.3 What is the effect of different, fixed SNR levels on behavioural (accuracy, subjective effort, subjective difficulty) and physiological (Pupillometry, GSR, Heart Rate, Respiration Rate, EEG) measurements?

- Hypothesis 2.3 (SNR Effect - Behavioural): Decreasing SNR (representing increased difficulty) will lead to significantly lower behavioural accuracy, higher subjective effort ratings, and higher subjective difficulty ratings.
- Hypothesis 2.4 (SNR Effect - Physiological): Decreasing SNR will lead to significant changes across physiological measures indicative of increased effort. This includes, but is not limited to, greater task-evoked pupil dilation, increased GSR amplitude or change during listening/retention, altered HR dynamics (e.g., reduced recovery), altered RR dynamics, and greater EEG alpha suppression during relevant task phases.

Research Question 2.4 What are the relationships between the behavioural measures (accuracy, subjective effort, subjective difficulty) and the various physiological measures, particularly considering the different SNR levels?

- Hypothesis 2.5 (Clustering within SNR): Cluster analysis applied to the time-course of physiological responses (Pupil, GSR, HR, RR) will identify distinct subgroups of participants within each specific SNR level, reflecting different physiological response styles to a given difficulty.
- Hypothesis 2.6 (Cluster Correlates): Consistent with Study 1 findings and the recognised physiology-behaviour gap, membership in these physiological clusters within a given SNR level is not expected to be significantly associated with concurrent behavioural accuracy or subjective ratings.
- Hypothesis 2.7 (Cross-Modal Agreement): Agreement between cluster assignments derived from different physiological measures (e.g., HR vs GSR vs Pupil vs RR) will be low to moderate within each SNR level, indicating that these measures capture distinct or complementary aspects of the physiological response.
- Hypothesis 2.8 (Physiology-Behaviour Correlation): The magnitude of change in specific physiological measures during the listening period (particularly GSR and potentially pupillometry, based on literature and Study 1 trends) will significantly correlate with behavioural accuracy and/or subjective ratings across participants, potentially revealing links between physiological engagement and outcomes.

By addressing these questions and testing these hypotheses, Study 2 aims to provide a detailed, multi-dimensional account of how listening effort is physiologically expressed and modulated by task difficulty, and the relationship between physiological measures and subjective effort, difficulty, and performance.

Chapter 11

Experiment Design

This chapter explains the experimental design employed to investigate listening effort. It describes the listening task, a speech-in-noise test, outlines the participant selection process, specific inclusion criteria, and illustrates the experimental procedure involving task blocks and randomised condition sequences. It further explains what measures have been taken to minimise the influence of confounding factors.

11.1 Participants

Sample Size The experiment aimed to recruit thirty participants, a sample size chosen on the basis of an a priori G*Power analysis for a repeated-measures within-participants design and consistent with previous work using similar paradigms (Faul et al., 2007; McGarrigle et al., 2014; Zekveld et al., 2011). The power analysis assumed a medium effect size ($f = 0.25$), $\alpha = .05$, and 90% power for a design with two listener groups and four SNR levels, yielding a required total sample size of $N = 30$.

Inclusion Criteria Participants were recruited via posters displayed on campus and direct email invitations.

To be eligible for the experiment, participants had to:

a. be native English speakers;

b. be aged between 18 and 40 years;

c. have normal hearing;

d. have no known neurological diseases, physical or cognitive impairments that might affect responses;

e. have no severe skin allergies or irritation in areas where physiological measurement methods would be applied.

Screening questions were completed to decide if participant can take part in the experiment. After they met the criteria after the screening questions, participants were given a Participant Information Sheet to read through the details of the experiment design before deciding whether to participate. When signing up, participants were asked to book two experimental sessions one week apart. They were paid £40 for completing the full experiment.

We recruited only native speakers because the task involved understanding English language. Previous research suggests that native and non-native speakers differ in both performance and subjective effort when understanding language (Peng & Wang, 2019).

The age range of 18 to 40 was chosen to ensure participants were mature enough to understand the requirements of the experiment while not being at an age associated with increased risk of hearing loss. Additionally, age is known to contribute to listening effort when performing speech-in-noise tasks, regardless of hearing loss level. This effect has been attributed to the decline in working memory that occurs with ageing (Zekveld et al., 2011).

Thirty-three participants took part in the experiment. One participant dropped out after the first session. For each physiological measurement (EEG, ECG, GSR, pupillometry, and respiration), different numbers of valid data sets were used in the analysis. For pupillometry, for example, data from 26 participants were included in the final analysis, others were excluded because the noise is too high for reliable analysis..

11.2 Listening Test: Speech-in-noise Test (OLSA)

The primary tasks were speech-in-noise tests. These tests were originally developed in German, named Oldenburger Satztest, or Oldenburg Sentence Test (OLSA) (Wagener et al., 1999). This test was adapted from German into British English.All sentences were read by a female voice speaking British English.

The speech stimuli consisted of five-word sentences presented against a background of multi-talker babble noise (see Figure 11.1). Babble noise was selected to increase ecological validity by simulating real-world listening environments.

Different conditions were created by varying the SNR levels at -16 dB, -11 dB, -6 dB, and 12 dB. These levels were established through pilot experiments, aiming to achieve approximate accuracy rates of 20%, 50%, 80%, and 100%, respectively. Babble noise was used throughout the test.

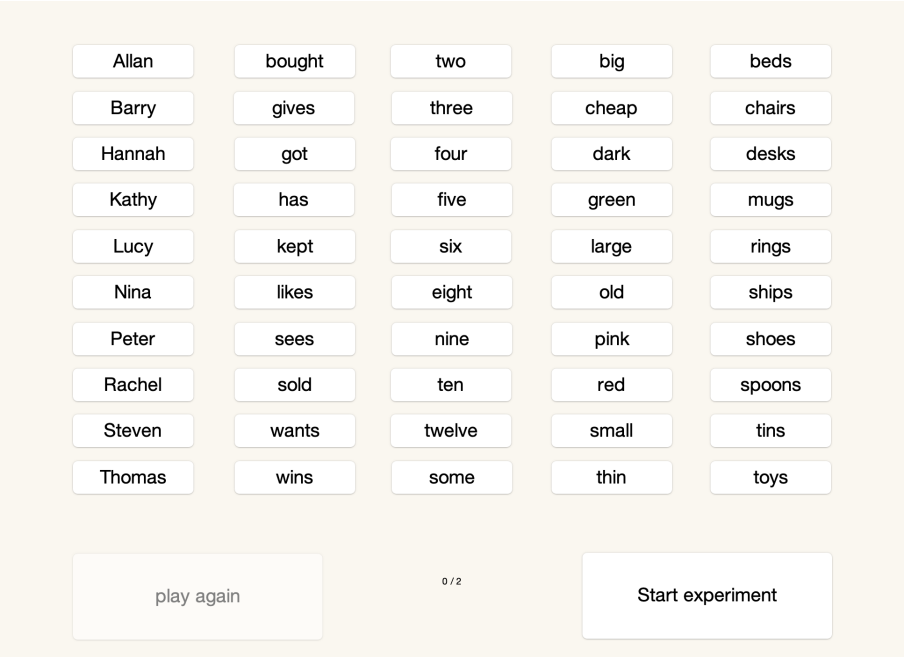


Figure. 11.1. Speech-in-noise Matrix Test (Oldenburger Satztest, or Oldenburg Sentence Test (OLSA))

This figure illustrates the speech-in-noise test adapted from German version (Wagener et al., 1999). The sentence consisted of five words selected randomly from each matrix column (e.g., "Alan bought two big beds").

After hearing the sentence, participants touched the screen to select the words they perceived. Words in the matrix were arranged alphabetically and numerically. The 'play again' button was disabled to standardise exposure.

For each listening task, noise began 500 ms before the sentence. Sentences were read by a female voice in British English, with each sentence comprising five words. Each word was randomly selected from the corresponding column of the matrix (see Figure 11.1). The nouns in the response matrix were arranged in alphabetical order, and the numbers were arranged in numerical order.

To complete each task, participants touched the screen to select the words they had heard. The screen was adjusted to eye level to ensure comfort and high-quality pupillometry data. Participants were instructed to minimise their movements during the test. As shown in Figure 11.1, the "play again" button was disabled. participants cannot replay the stimulus during the listening task.

Although the OLSA Matrix test is designed to minimise semantic predictability, some degree of procedural learning can still occur over repeated trials. Participants can become more familiar with the response matrix, the talker's voice, and the fixed five-word sentence structure, which may lead to modest improvements in performance over time. These effects are most likely to appear at intermediate SNR levels where the task is neither too easy nor too difficult.

To mitigate these potential effects, the order of the SNR conditions was fully randomised across the experiment. Each SNR level was divided into two blocks, resulting in eight blocks in total. The session was further split into two parts, each containing four blocks with different randomised SNR levels. This design ensured that no participant encountered the SNR levels in a predictable or increasing-difficulty order, thereby reducing systematic learning effects across the test.

11.3 Repeated Experiment Design

Each participant completed two experimental sessions. The second session took place one week after the first and was scheduled at the same time of day to control for diurnal variation in physiological responses. The one-week interval was selected for several methodological reasons. Firstly, it aligns with the protocol used by Alhanbali et al. (2018) in study 1, allowing for direct comparison of findings. Secondly, it represents a compromise between methodological rigour and participant retention: longer gaps risk higher attrition, whereas shorter intervals may introduce carryover effects (Hausknecht et al., 2007).

The one-week interval is well-established in psychophysiological research paradigms, as it corresponds to natural circadian and behavioural cycles, thereby minimising extraneous variables whilst maintaining experimental control (Fallon et al., 2013). Additionally, scheduling sessions at the same time of day for each participant was implemented to control for diurnal variations in physiological responses, which have been shown to affect measures such as pupil dilation, cortisol levels, and EEG patterns (Schmidt et al., 2007).

This repeated-measures design enhances data reliability by enabling the assessment of both within-session fatigue effects and between-session variability in listening effort—factors particularly relevant to studies involving hearing-impaired population (McGarrigle et al., 2014).

11.4 Experiment Procedure

An overview of the experimental structure is shown in Figure 11.2. The full experiment comprised eight blocks and one short training block. The training block consisted of five listening trials presented at a high SNR level (+40 dB), allowing participants to familiarise themselves with the task without difficulty. The main experiment included 160 trials in total, with each block containing 20 trials at a fixed SNR level.

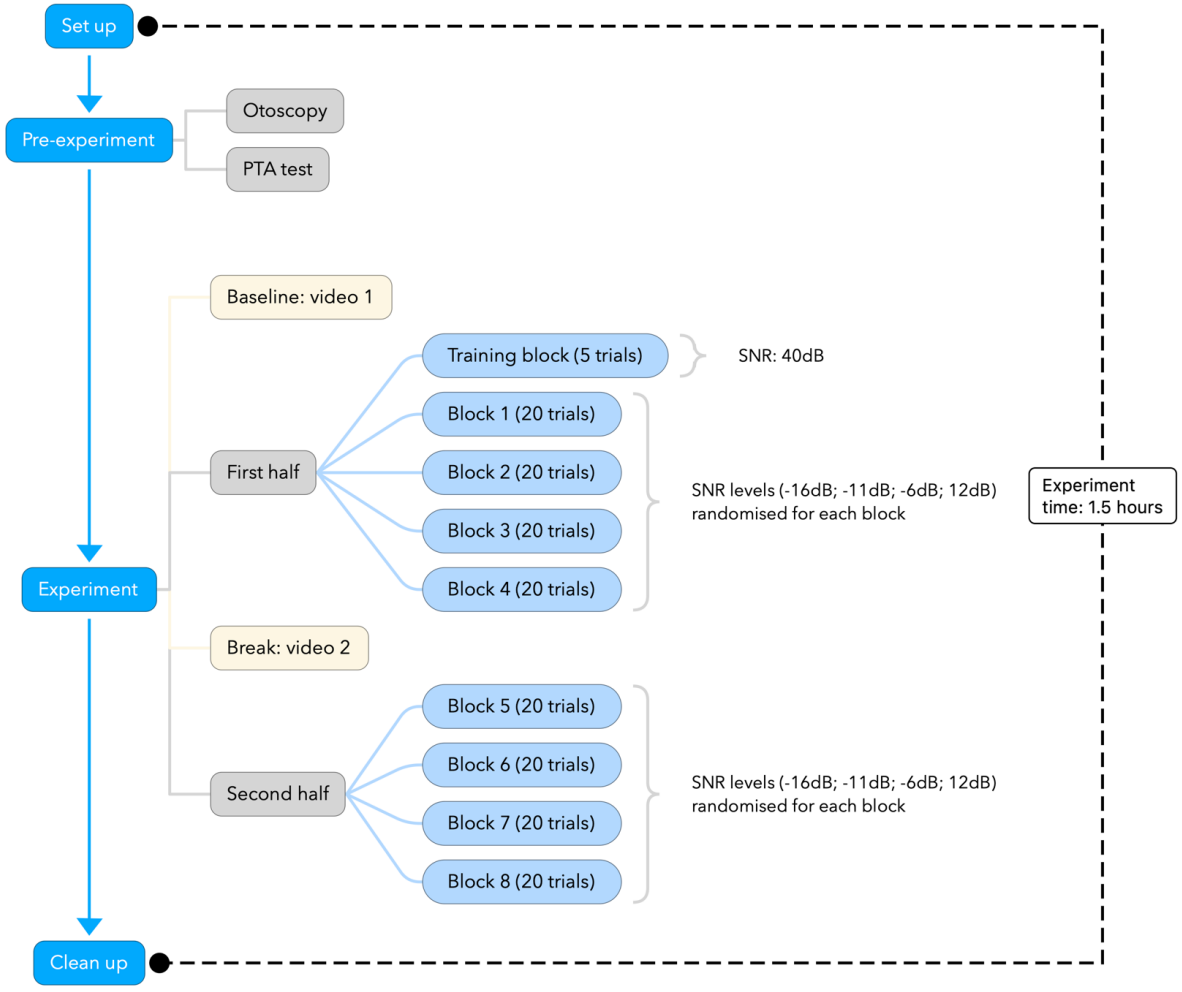


Figure. 11.2. Experiment Design

After setting up, the experiment began with otoscopy and a hearing test to confirm participants had healthy ear canals and normal hearing. Afterwards, participants viewed a two-minute documentary designed to help them settle, with part of this data serving as baseline measurements.
The listening tests began with a training block (consisting of five trials with low-level noise, Signal to Noise Ratio (SNR): 40dB). Following this, the first half of the experiment commenced, comprising four blocks of listening tasks. Each block included 20 trials with the same SNR level. The sequence of SNR levels was randomised.

This structure provided sufficient repeated trials per SNR condition (40 trials each) to enable reliable averaging and reliable data analysis, while keeping the total experiment duration under two hours to minimise participant fatigue.

Experiment Preparation Participants were tested in the Small Anechoic Chamber at the University of Southampton. Following initial setup, they underwent an otoscopic examination to assess the condition of their ear canals. The experimenter, trained by a

professional audiologist, ensured the procedure was conducted safely and accurately. Once participants were confirmed to have healthy ear canals, they were fitted with insert headphones to complete a PTA screening.

Stimulus presentation Sound was presented through insert headphones (E-A-Rtone 3A). It was calibrated through soundcard to ensure a stable sound level - 65 dB SPL, and also to control the channel so participant won't hear the click sound which was used for the researcher to mark the start of events (see Figure 11.3).

Hearing Test: Pure Tone Average (PTA) To verify normal hearing, a simplified PTA test was conducted. Pure tones were played through MATLAB. Each pure tone lasted approximately 2 seconds, four frequencies (500, 1000, 2000, and 4000 Hz) were each presented at three sound levels (10, 20, and 35 dB), producing a total of 12 unique stimuli. Sounds were played in a random order. The sound level was calibrated to 65 dB SPL based on the stimulus of the experiment (speech-in-noise test). Participants indicated whether they could hear all of the pure tones played.

Ensuring Participant Comfort Water and cups were provided within reach. Participants were requested to minimise movement as much as possible. They were informed that they could take breaks between tasks (approximately 12 seconds for each task) and could raise their hands if they had any questions. The researcher remained in the same room with the participant throughout the experiment.

Task Introduction Participants were informed about the procedure of the experiment verbally first and later through text on the screen. As the experiment proceeded, instruction dialogues appeared on the screen when needed.

Baseline Stabilisation The experiment began with a 2-minute nature documentary (Video 1), designed to help participants settle and to establish baseline measurements. After watching the baseline video, participants completed a training session. Instructions were presented on the screen, and participants completed 5 listening tasks (MatrixMat listening test, see Figure 11.1).

Training Session After completing 5 training listening tasks, participants rated their experience by answering two questions: "How difficult was it to understand what was said in the previous tasks?" and "For the last questions, how much effort did you put into understanding what was said?" (see figure 11.5) Participants used a slider to rate each question from 0 (not difficult/no effort) to 100 (very difficult/extreme effort).

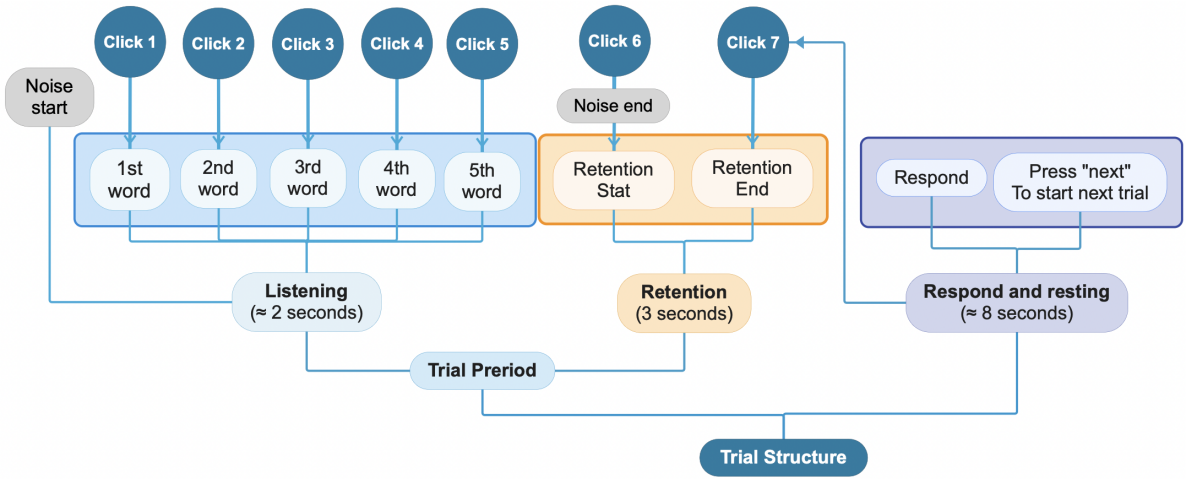


Figure. 11.3. Structure of Each Listening Trial

Each trial began with 500 ms of background noise(babble), followed by a five-word sentence embedded in noise. Each click / impulse was set to mark the start to each word, and the start and end of retention period, but has been specially treated through sound card so participant could not hear. After listening, a three-second retention period followed, where participant cannot respond. Participants then responded by selecting the words they heard on the screen.

Full Experiment After the training session, participants received on-screen instructions to begin the experiment. There were 8 blocks of tasks, with each block consisting of 20 questions. To minimise learning effects and sequence effects (participants performing worse as the experiment progressed), difficulty levels were randomised between blocks 1 to 8.

11.5 Measurements

11.5.1 Rationale for Each Measurement

Self-report This method directly assesses the individual's subjective experience of effort or task demand. Rating scales are commonly used due to their ease of administration and good face validity for capturing perceived exertion (McGarrigle et al., 2014). While subjective, they provide important insight into the listener's on experience of task difficulty and exerted effort (Lemke & Besser, 2016).

Performance (accuracy) Accuracy (word recall) to provide an objective measure of listening comprehension success under varying conditions. . It serves as a crucial information about how successful people are in completing the task, regardless of how effortful they might feel (subjective report) and how they react physiologically. Research shows that the relationship between performance and subjective effort isn't always linear,

and significant effort may be expended even when accuracy is high (Francis & MacPherson, 2021; Ohlenforst et al., 2018; Pichora-Fuller et al., 2016b).

Pupillometry Changes in pupil dilation are a well-established physiological correlate of cognitive load and mental effort (McGarrigle et al., 2017; Zekveld et al., 2018). Pupil size typically increases with greater task demand, reflecting high arousal during effortful listening. Task-evoked pupillary responses provide a temporally reliable measure of moment-to-moment fluctuations (Zekveld et al., 2018) and correlate reliably with performance and subjective effort (Zekveld & Kramer, 2014).

GSR GSR reflects changes in skin conductance due to sympathetic nervous system activation, providing an index of physiological arousal, stress, and cognitive effort. Increases in skin conductance level can occur with increased auditory task demand (Figner & Murphy, 2011), making it a candidate measure for listening effort (Mackersie & Calderon-Moultrie, 2016).

ECG Cardiovascular measures, particularly heart rate (HR) and heart rate variability (HRV), reflect autonomic nervous system activity and are sensitive to stress and cognitive load (Forte et al., 2020; Kim et al., 2018). Increased heart rate was linked to greater cognitive demand or stress associated with challenging listening conditions (Mackersie et al., 2015; Richter et al., 2016a).

Respiration Comparing to GSR, ECG, and Pupillometry, respiration was examined less in previous listening effort research. However, studies show that respiration patterns (rate, depth, variability) are linked to cognitive load and emotional state (Grassmann et al., 2016). Monitoring respiration provides complementary physiological information and helps control for potential confounds in other autonomic measures like ECG (Laborde et al., 2017).

EEG EEG provides direct measures of brain activity related to cognitive processing and attention. Specifically, alpha power (8-13 Hz) modulations in EEG recordings provide insights into attentional allocation and inhibitory processes during challenging listening conditions (Strauß et al., 2014). Baseline alpha power has been associated with greater pre-task engagement and predictive of better subsequent task performance (Alhanbali et al., 2021; Hanslmayr et al., 2005).

A summary of how the experiment was set up is presented in Figure 11.4. Two computers were used concurrently: computer 1 collected data on self-reported effort,

performance(accuracy), and pupillometry; computer 2 recorded EEG, ECG, GSR, and respiration via the Biopac system.

11.5.2 Self-report

As shown in Figure 11.5, participants responded to two questions after each block of listening tasks (comprising 20 speech-in-noise trials at the same signal-to-noise ratio, SNR). The two-question format was designed to track subjective ratings of effort and difficulty across SNR levels, while keeping the questionnaire brief enough to avoid participant fatigue. The questions were intentionally simple and intuitive, requiring participants to respond by adjusting a slider.

The two questions addressed: (1) subjective effort - how much effort participants felt they exerted, and (2) subjective difficulty - how difficult they perceived the task to be. Both were rated on a 0 to 100 scale. Importantly, the instructions clarified that effort and difficulty were not necessarily the same. For example, a participant might perceive the task as very difficult but report low effort if they had mentally disengaged or given up.

11.5.3 Accuracy

Accuracy were immediate recorded after each trial (160 in total) through Matlab where the listening task was presented. It was calculated as the percentage of correct word identified in the 5 word sentence (see Figure 11.1 Page 111). For example, if two words were chosen correctly, the percentage would be 40 %. Collected accuracy was later averaged based on the same SNR level.

11.5.4 Pupillometry

Pupillometry data were recorded using Pupil Core eye-tracking glasses in conjunction with Pupil Labs software, with a nominal sampling rate of 100 Hz. Compared to the EyeLink system used in Study 1, this setup was more user-friendly, as it did not require participants to rest their chins on a fixed support. However, this convenience came at the cost of increased noise and reduced data reliability. For instance, although the system was set to sample at 100 Hz, the actual sampling rate - based on recorded timestamp - fluctuated between approximately 80 and 120 Hz.

Prior to each experiment, a calibration procedure was performed to ensure accurate pupil size measurements. Height of the screen was adjusted accordingly. Participants who wore prescription glasses were able to complete the experiment while wearing their own glasses, as the device is designed to accommodate them comfortably.

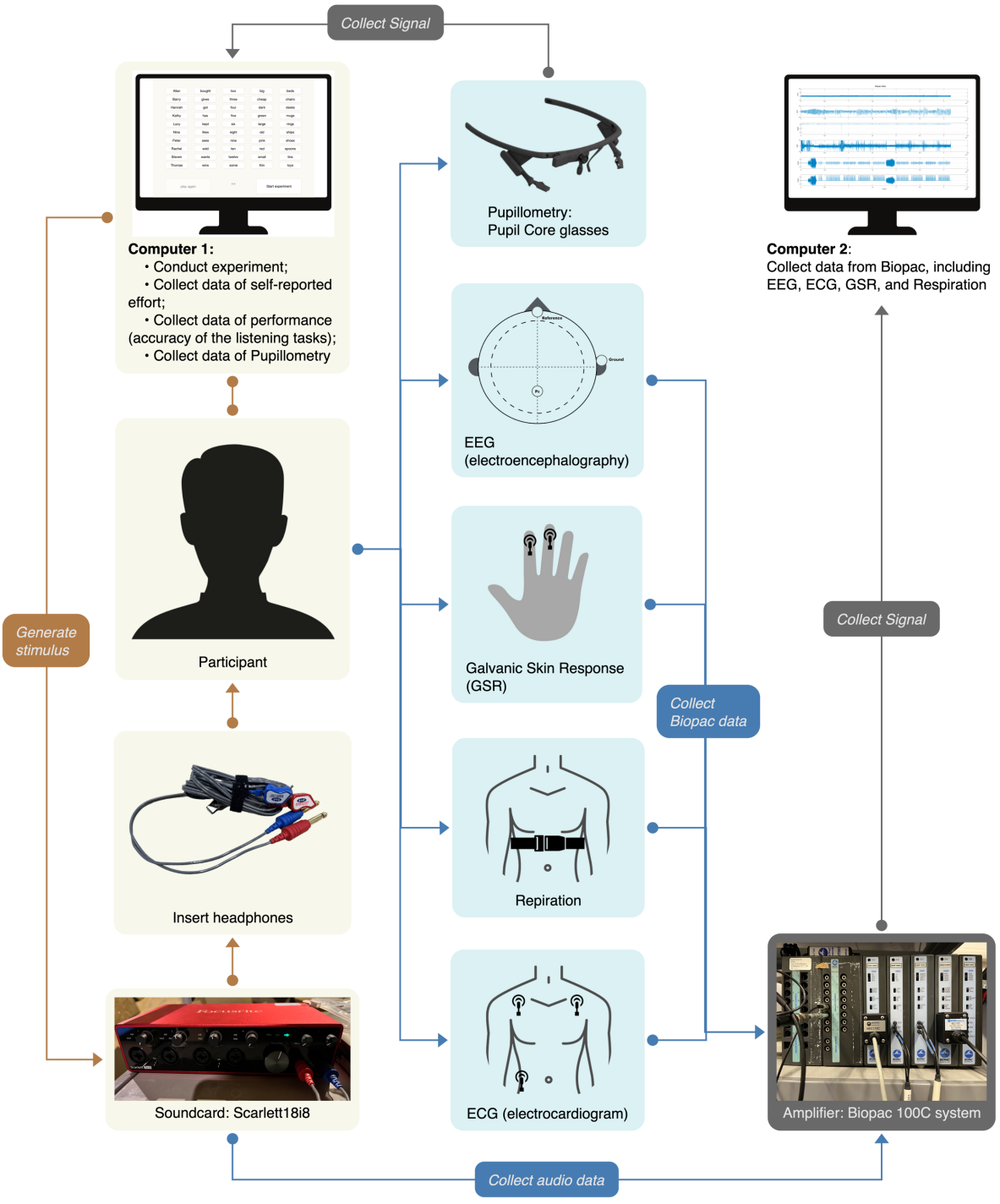


Figure. 11.4. An Overview of Experiment Measurements and Data Collection

To conduct the experiment, computer 1 was set up to generate and play stimulus. The stimulus was routed through a soundcard to reach a calibrated sound level of 65 dB SPL and then delivered through insert headphones. Computer 1 also collected self-reported effort, performance in the listening task, and pupillometry data. Physiological measures including EEG, ECG, GSR, and respiration were collected through the Biopac system and recorded on computer 2.

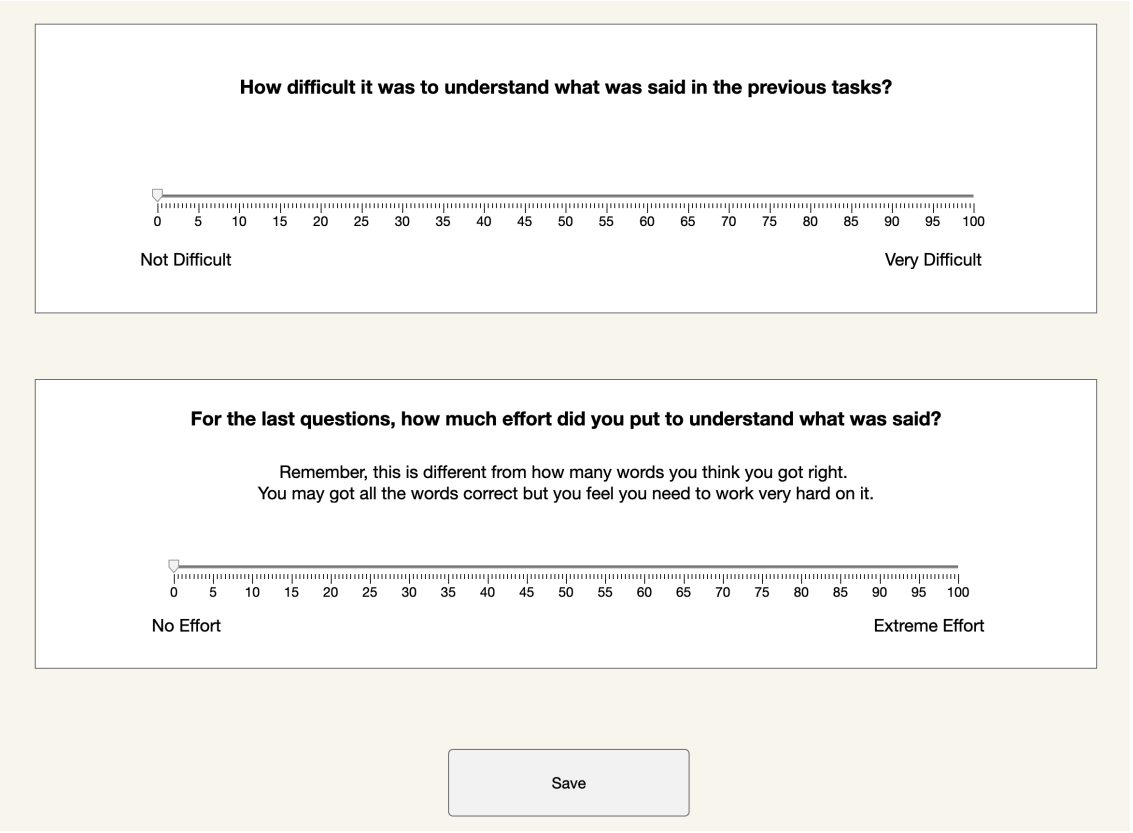


Figure. 11.5. Rating of Self-reported Effort

After each block of listening tasks (20 speech-in-noise tests at the same signal-noise-ratio), participants rated their effort level using a slider from 0 to 100 to answer each question. Data were recorded in Matlab.

11.5.5 GSR

In this study, GSR was recorded using the Biopac 100C system. Electrodes were placed on the palmar surface of the non-dominant hand (the one not used in selecting words on the screen). This placement minimised movement artifacts whilst ensuring sensitivity to autonomic arousal. Participants were asked to move as less as possible when performing the task.

11.5.6 ECG

ECG was recorded using the Biopac 100C system, same as GSR. Electrodes were placed on the upper left and right chest, just below the clavicles, and on the lower right torso, below the ribcage (see Figure11.4). This standard setup captures heart activity by measuring the electrical potential differences between these points, allowing for reliable detection of heart rate and rhythm during the task.

11.5.7 Respiration

Respiratory patterns were recorded using a Biopac 100C respiration amplifier with a strain gauge transducer placed below the participant's chest. This setup allowed for continuous, non-invasive monitoring of breathing throughout the experiment (see Figure11.4).

11.5.8 EEG

For this study, single-channel EEG (Pz) data were collected through electrodes connected to the Biopac system. Three electrodes were applied: one electrode at the Pz position for EEG data collection, one electrode behind the right earlobe(the mastoid area) serving as ground, and one electrode on the high forehead as reference.

Since no existing multi-channel EEG caps were compatible with the Biopac system while also allowing simultaneous ECG, GSR, and respiration recordings, a custom EEG cap was developed. The researcher independently learned 3D design and fabrication to build a cap specifically for holding a single electrode at the Pz location. Prior to each session, conductive gel was applied inside the electrode holder to reduce impedance and ensure signal quality.

Prior to each recording session, the electrode surface was prepared with conductive gel, applied into the electrode hold, to ensure good signal quality and low impedance. Electrode impedance was checked during the initial system setup to ensure that all channels met the manufacturer's recommended standards. Impedances were confirmed to be within acceptable limits (typically below 10–20 kΩ for the amplifier used), and the system provided stable recordings throughout data collection. However, impedances were not re-measured for each individual participant.

While the live signal appeared reasonable during data collection-for instance, alpha waves were visibly enhanced when participants closed their eyes-the recorded data remained relatively noisy and challenging to analyse.

Matlab Application Design The entire experiment was developed and implemented in MATLAB, with a custom-designed app to automate transitions and streamline the participant experience. Afterwards, the researcher further adapted the core script * and designed a Matlab app to incorporate the single test into and interactive application. The procedure, lasting approximately one hour, was programmed to progress naturally -

*Core script was designed by Professor Stefan Bleeck, Institute of Sound and Vibration Research (ISVR), University of Southampton, and updated by Professor David Simpson ISVR, University of Southampton, for this experiment.

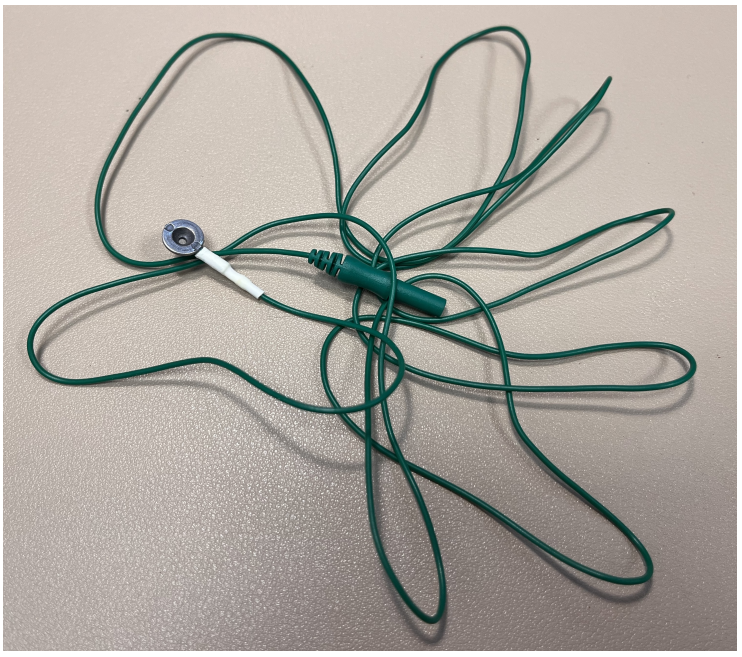


Figure. 11.6. Reusable single EEG electrode with lead wire

This image shows the single reusable EEG electrode used with the custom EEG cap. The electrode features a flat metal surface intended for skin contact, but maintaining stable contact on the scalp is challenging due to hair. To address this, a dedicated holder was 3D-printed and integrated into the cap to secure the electrode at the Pz site (see Figure 11.7).

moving through instructions, training sessions, experimental trials, breaks, and subjective rating prompts-with minimal researcher intervention.

Details of adaption as follow:

- **Matrix Interface Update**
  - Background colour was matched to the desktop environment (light beige) to minimise distraction and effect of window change on pupil size.
  - Button layout and font alignment were adjusted for improved readability and usability.
- **Integrated App Design:**

  Master script development: A central control script was developed to orchestrate the full experimental procedure.

  During the experiment, this script:

  1. Selects a predefined script sequence from a file to structure the trial flow (contains randomised SNR sequence for speech-in-noise test).
  2. Automatically launches and coordinated multiple scripts in sequence.

Figure. 11.7. Design and assembly of the self-constructed EEG cap

This figure illustrates the custom-built EEG cap used for recording scalp potentials in the study. The design includes modular components such as 3D-designed and 3D-printed electrode holders, joints, and fastening bolts, which allow flexible placement and secure attachment of electrodes. Elastic fabric straps were threaded through loops in a white plastic frame, creating an adaptable cap structure suitable for different head sizes. Each electrode module can be screwed into place using a specially designed installation tool, ensuring consistent electrode positioning and contact pressure. The cap was tested on a prior to use in participants, enabling a reliable, reconfigurable, and cost-effective EEG setup.

3. Manages the following components:
   - *Instruction screens* to guide the participant through the experiment stages.
   - *Full-screen video playback* (e.g., nature scenes) before and during the session to encourage relaxation.
   - *Countdown screens* to prepare participants for upcoming segments or to manage break durations.
   - *Speech-in-noise Matrix screen* for participants to complete the task.
   - *Subjective rating interfaces* for reporting perceived effort and task difficulty after each block.

- **Participant-Paced Interaction:** All scripts were designed to complete their tasks and wait for participant input before continuing. Participants could regulate the pace by pausing on any screen until they were ready to proceed.

11.6 Minimising Confounding Factors

Minimising Light Effects on Pupillometry To minimise the effect of screen luminance changes on pupillary responses, a consistent neutral beige background was implemented throughout the experiment. The laptop display background was first set to a subtle beige tone, and then the experiment interface background was matched precisely to this colour using a digital colour sampling tool. This consistent background luminance across all experimental screens was maintained to control for pupil dilation responses that might otherwise be triggered by changes in screen brightness rather than by cognitive load associated with the listening tasks.

Stimulus Calibration The air-conducted sounds were calibrated to 65 dB SPL using the Bruel and Kjaer occluded ear simulator (type 4157) connected to a sound level meter (type 2250). It was chosen to ensure the sound level is comfortable from the insert headphones for the appoximately one hour long experiment.

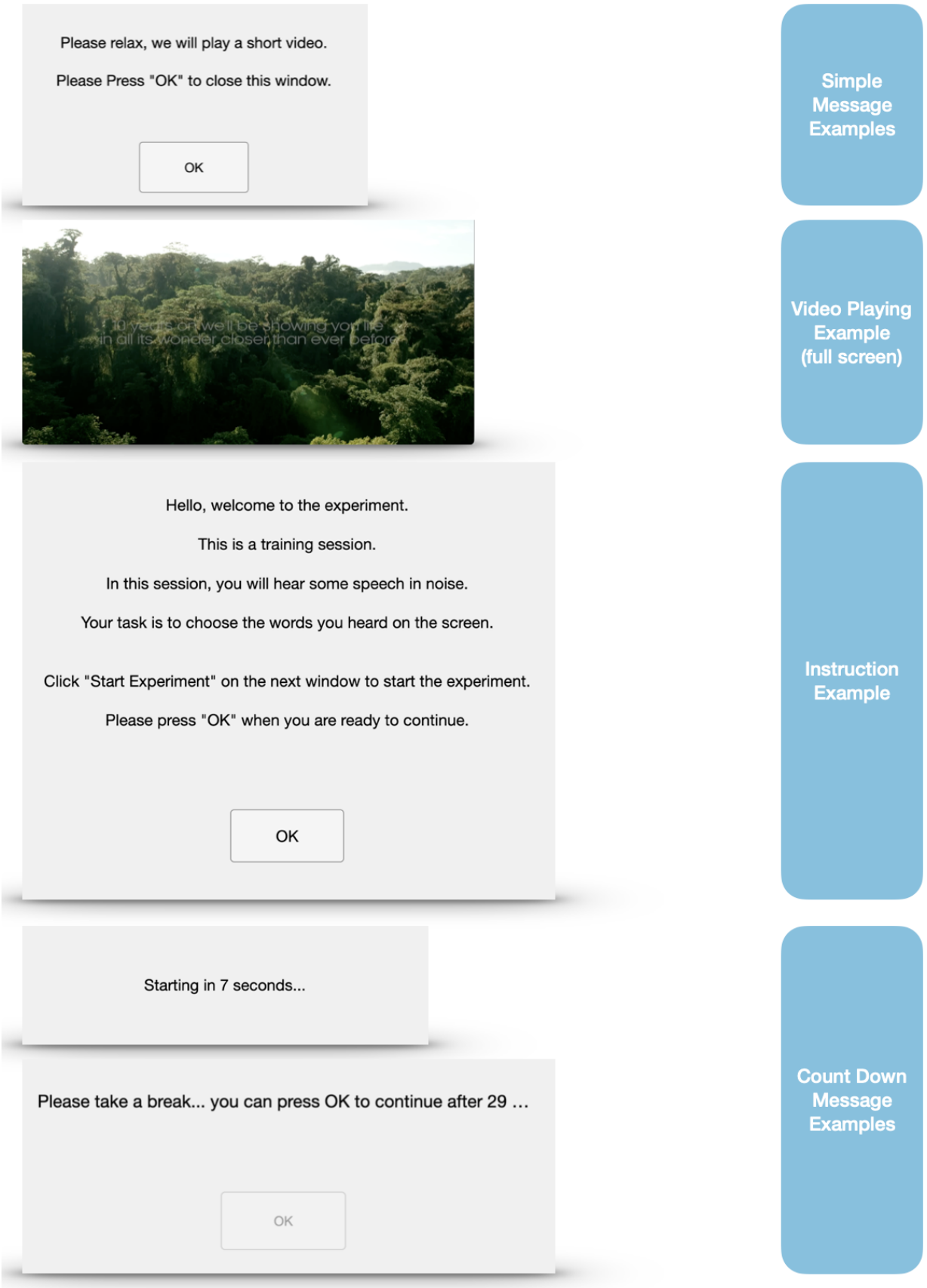


Figure. 11.8. Automated experimental interface developed in MATLAB

Automated interface were developed for the full experiment through Matlab App function. These are examples developed using MATLAB app design function streamline participant experience. Top: simple message prompts used to provide instructions or confirmations. Middle: full-screen nature video used to help participants relax prior to or during the session. Below: instructional screens that introduced each experimental block, and countdown displays to prepare participants for the next segment or manage breaks. All screens were designed for clarity, ease of use, and to reduce cognitive load and experimenter intervention.

Chapter 12

Data Pre-Analysis

12.1 Time Alignment

The first challenge of data analysis is aligning timing of the events. During the full experiment, clicks (sudden spikes) were added to mark the start of each word, and start and end of retention period. Hence, each trial consists 7 clicks, and the full experiment would contain 1155 clicks (see Figure 12.1). Figure 12.2 shows clicks during the full experiment (example data from GSR).

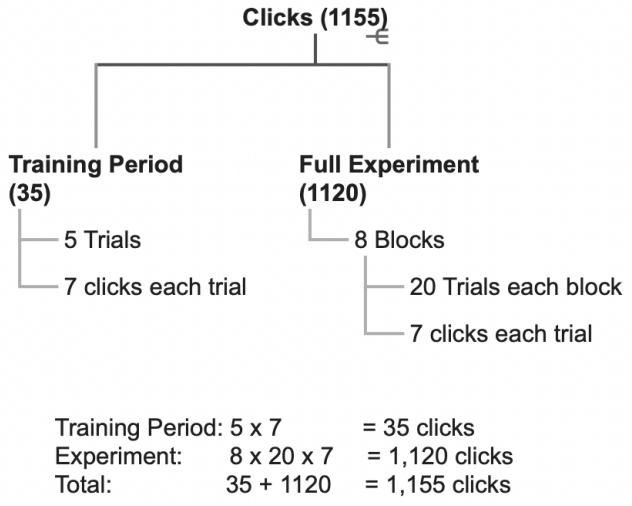


Figure. 12.1. Clicks Number Content

Figure shows the structure of the click numbers. Each experiment has 1155 clicks as a mark of the start of different event.

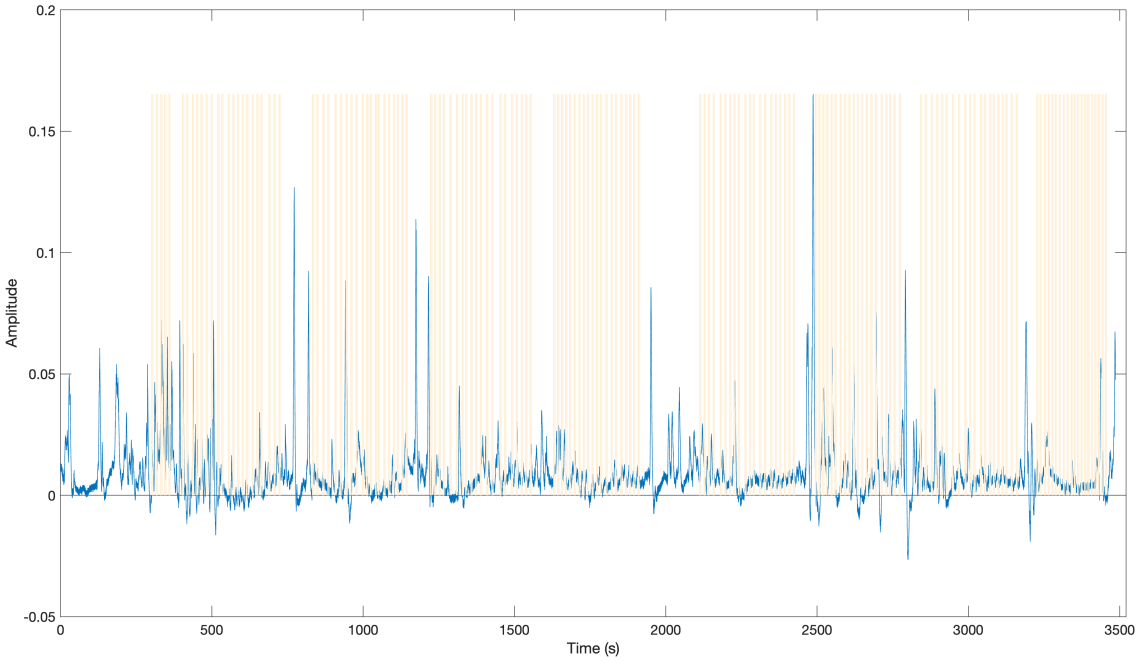


Figure. 12.2. Example of clicks across Full Experiment (GSR)

Figure shows the structure of the click numbers (yellow spikes). Each experiment has 1155 clicks as a mark of the start of different event. There are 7 clicks for each trial. The first 5 signals where each word in the sentence starts, the 6th click represents when retention period starts, right after the 5th word finish. The 7th click means where the retention end and responding start.

12.2 Subjective and Behavioural Measures

Accuracy was calculated through Matlab automatically, where it presents the listening task. It stores the accuracy for each trial (140 trails for full experiment). Accuracy was calculated based on how many words participant correctly recall. For example, if participants chose one word correctly out of five, the accuracy would be 20%.

To minimise the effect that at the start of each block(after 1 minute force break), participant were still adapting to the listening task at a new SNR level, the first trial was excluded when calculating the average accuracy through the block (20 trials). The accuracy was further averaged for the same SNR level (each has 2 blocks, therefore 40 trials).

Different from accuracy which was calculated for each trial, subjective ratings, including subjective effort and subjective difficulty were presented after each block (20 trials). Hence eight subjective ratings were collected for the full experiment, which were later averaged in to four ratings for each SNR level.

12.3 Pupillometry Data Cleaning

Synchronisation and Temporal Standardisation Pupillometry data processing started by averaging data from both eyes and aligning it using system time recordings. The raw data's inconsistent sampling rate (around 90 to 120 Hz) was corrected to 100Hz according to the timestamps recorded by the software to make sure the time alignment. This ensured easier alignment with other physiological data from Biopac (EEG, ECG, GSR and Respiration), which were sampled at 1000 Hz.

In pupillometry, time of each sample was recorded as timestamps in exported data. It's a relative time which can be translated into computer system time. Therefore, recording of Biopac and Pupillometry can be aligned using the computer system time. Based on pilot data, the difference of Matrixmat and Pupillometry time recording range from 20 to 200 ms, maximum.

Initial Filtering and Artefact Detection The standardised data underwent initial filtering where physiologically implausible diameter values (below 0.1 mm) were marked as NaN. Artefacts like eye blinks were then identified. This involved using dual median filters (long and short windows) to capture signal trends, followed by flagging points with high local variance (>5x median variance) or existing NaN values.

Artefact Detection and Interpolation Detected artefacts were handled based on their duration. Brief artefacts, those lasting less than 0.5 seconds, were corrected using linear interpolation between adjacent valid data points. This approach aimed to preserve signal continuity for momentary disruptions without introducing artificial dynamics.

Final Signal Conditioning Longer artefactual segments (>0.5s) were explicitly retained as NaN, acknowledging the unreliability of interpolating over extended gaps. Additionally, physiologically brief (< 0.1s) patches of seemingly valid data between artefacts were also marked as NaN to ensure signal plausibility. The final cleaned pupil diameter was then estimated via a moving average of the remaining valid data, balancing noise reduction and signal preservation.

12.4 Respiration Rate Extraction

To extract respiration rate from the raw respiratory signal, a multi-step process was applied to ensure both robustness and temporal precision. First, the signal was filtered and smoothed that preserves peak shapes while reducing high-frequency noise (see

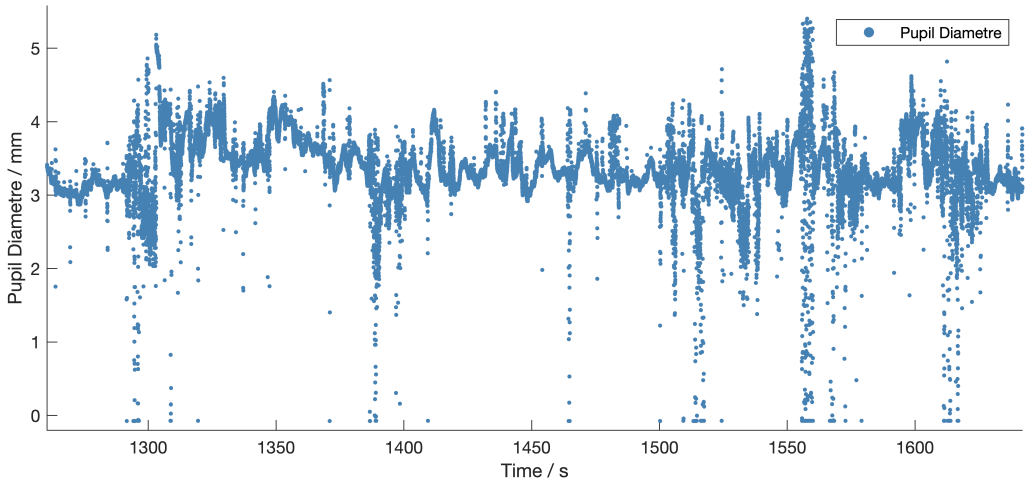


Figure. 12.3. One Section of Pupillometry Data

The figure above shows pupil diametre data(in millimetres) in about 300 seconds. Human Pupil diametre generally range from 2 to 8mm Fawcett et al., 2022. Therefore, data below 1mm are considered as blinks(eyes closed leading to a big drop of pupil diamtre) and removed.

Figure 12.4). This step enhances signal clarity without significantly distorting the respiratory waveform, making it well suited for detecting subtle breathing cycles.

After peaks were detected with an adaptive peak detection method, the time intervals between consecutive peaks were calculated and converted into breathing rates (in breaths per minute) the interpolated into the same sampling frequency ($fs = 1000$) as the original data, so to ensure data is easily comparable with other physiological measures at the same sampling frequency.

12.5 Heart Rate Extraction

The heart rate extraction process started by removing outliers from the raw ECG signal. Data points exceeding ± 4 standard deviations from the mean were classified as outliers, replaced with NaN, and then linearly interpolated to maintain signal continuity. The cleaned ECG signal was then bandpass filtered between 5 Hz and 30 Hz using a first-order Butterworth filter. This removed baseline drift and high-frequency noise while preserving the QRS complex. Zero-phase filtering (`filtfilt`) was used to prevent phase distortion.

Figure 12.5 shows an example of part of ECG original data and extracted heart rate.

R-peaks were detected from the integrated signal using an adaptive threshold based on the signal's mean. Instantaneous heart rate was calculated from the time intervals between consecutive R-peaks (RR intervals). Each RR interval (in seconds) was converted

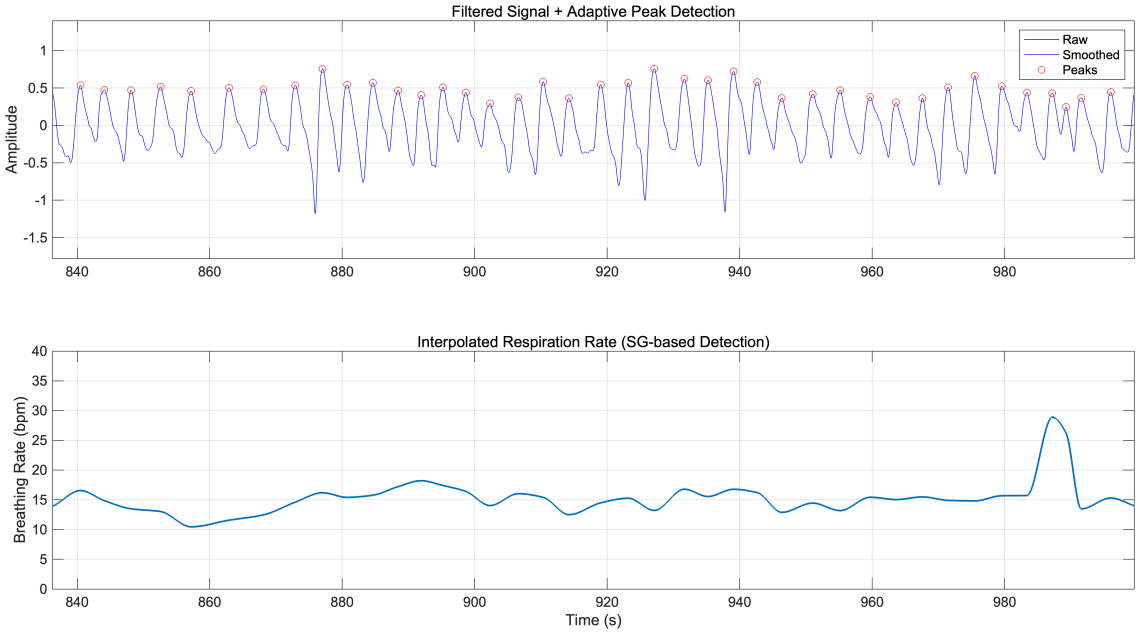


Figure. 12.4. Example respiration signal and derived breathing rate for one participant (Subject 8, Experiment 1)

The top panel shows the raw and smoothed respiration signal, with detected peaks marked in red. The bottom panel displays the corresponding breathing rate (in breaths per minute, bpm), estimated using peak-to-peak intervals and linearly interpolated across time. This method provides a continuous and robust estimate of respiration dynamics across the listening task. Filtering improves peak clarity while preserving underlying waveform structure, facilitating accurate rate calculation even in moderately noisy recordings.

to beats per minute using the formula: Heart Rate = 60 / RR interval. To get a continuous heart rate signal aligned with the original ECG timeline, the discrete, instantaneous heart rate values (associated with the midpoints of RR intervals) were linearly interpolated with the same sapling frequency (1000) of the ECG signal.

12.6 GSR Data Cleaning

Negative Signal Artefact Removal Large negative spikes in the GSR data were addressed first. Following the removal of negative spikes, general outliers were detected using a Z-score approach. Any data point with an absolute Z-score exceeding a threshold of 5 was classified as an outlier. These identified outliers were also replaced with NaN and subsequently managed using linear interpolation to restore signal continuity.

To reduce high-frequency noise and enhance signal quality, the cleaned GSR data was smoothed using a fourth-order Butterworth low-pass filter with a 4th order and cut-off frequency of 1 Hz. Zero-phase filtering(Matlab function *filtfilt*) was employed to prevent phase distortion, ensuring essential signal characteristics were preserved (see Figure 12.6).

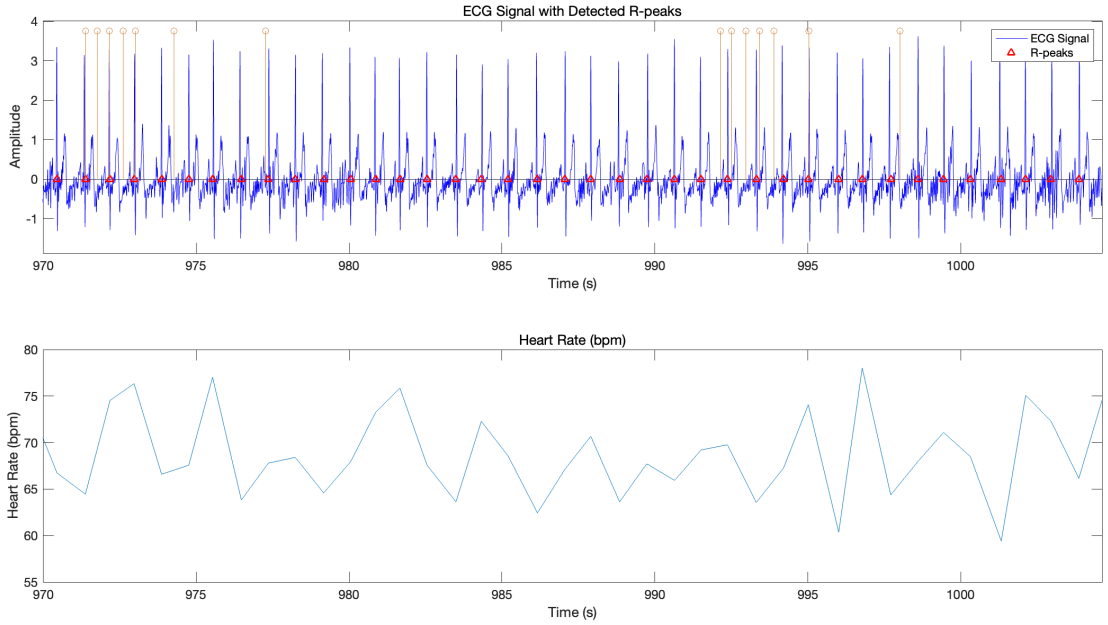


Figure. 12.5. Example of Heart Rate Extraction

The top panel shows the raw ECG signal (blue) with detected R-peaks marked by red triangles. Orange stems indicate click events used to structure the experimental trial, with seven clicks per trial: five clicks marking word onset times and two clicks marking the retention start and retention end, respectively. The bottom panel displays the extracted heart rate (beats per minute, bpm) over time, calculated from the R-R intervals..

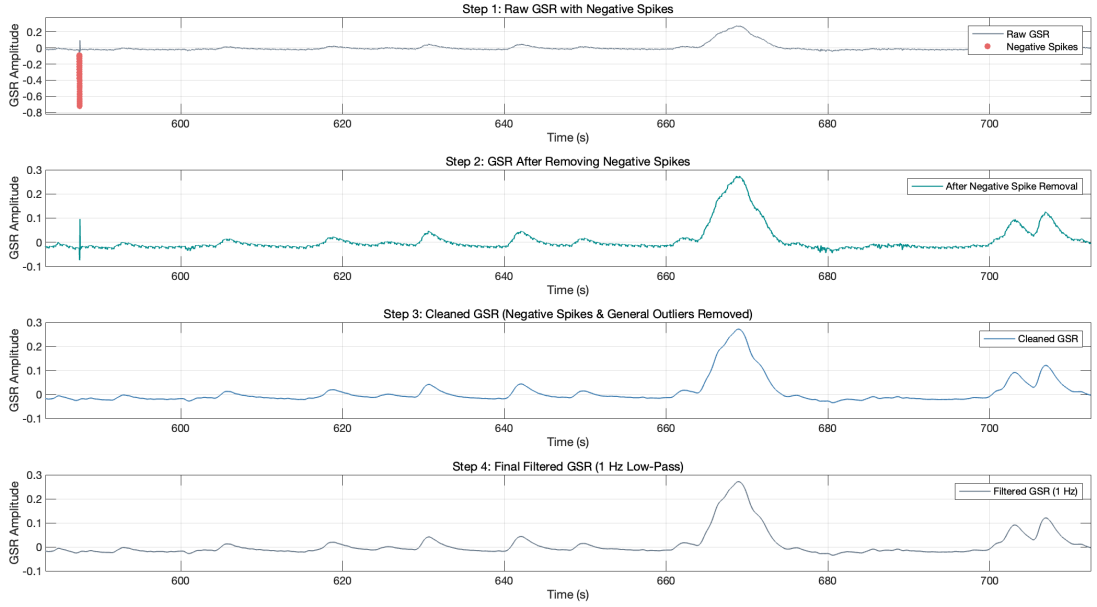


Figure. 12.6. GSR Data Cleaning Example

This figure present an example of the process of GSR data cleaning. Unnecessary spikes and outliers were removed. Then data were further smoothed through filtering.

12.7 Single-Channel EEG Pre-processing

Initial processing involved adjusting the EEG signal to remove any constant DC offset. Standard filtering techniques were applied to remove slow drifts and high-frequency noise (like muscle activity and mains interference). Following this, the cleaned signal was further filtered to isolate activity within specific brainwave frequency bands of interest, such as Theta (4-8 Hz), Alpha (8-13 Hz), and Beta (13-30 Hz).

However, due to the device limitation, EEG data remains noisy and results difficult to interpret. Full results will be presented int the next chapter.

Chapter 13

Results and Discussion

Experiment Design Overview Thirty-three participants with normal hearing were initially recruited. Due to data quality requirements across different physiological measures (Pupillometry, GSR, Respiration, ECG, EEG), the final analyses included approximately 26 complete datasets for each measurements.

The study employed a within-subjects, repeated-measures design. Participants attended two identical experimental sessions separated by a one-week interval. Both sessions were scheduled at the same time of day for each participant. This repetition and timing protocol was implemented to ensure consistency and allow for direct comparison with findings from Study 1.

Participants completed a speech-in-noise recognition task involving five-word sentences spoken by a female speaker with a standard British English accent. These sentences were embedded in multi-talker babble noise and presented at four signal-to-noise ratio (SNR) levels: -16, -11, -6, and 12 dB. These SNR values were determined through pilot testing to elicit target performance levels of approximately 20%, 50%, 80%, and near 100%, respectively.

Participants reported the sentence heard by selecting words from the screen, where it presentsa 10x5 response matrix (see Section 11.2, Page 110 for details). Participants were required to select one word within each column to reconstruct the sentence (5 words), Each column contained the target word and nine semantically similar distractors.

Subjective ratings were collected after each block containing 20 trials at the same SNR level. Participants rated two questions: 1) Perceived Difficulty ('How difficult it was to understand what was said in the previous tasks?') and 2) Invested Effort ('For the last questions, how much effort did you put to understand what was said?'). Both scales ranged from 0 ('Not Difficult' / 'No Effort') to 100 ('Very Difficult' / 'Extreme Effort').

Along with behavioural measures (subjective effort, subjective difficulty, and accuracy), physiological measures including Pupillometry, Galvanic Skin Response (GSR),

Respiration, Electrocardiography (ECG), and Electroencephalography (EEG) were also collected during the experiment.

13.1 Behavioural results: Subjective Effort, Subjective Difficulty, and Accuracy

The influence of the Signal-to-Noise Ratio (SNR) on task performance and subjective experience was analysed. Figure 13.1 provides a visual summary of accuracy, subjective effort, and subjective difficulty across the four SNR levels (-16 dB, -11 dB, -6 dB, and 12 dB), displaying group averages (bold lines), individual participant data (faint lines), and response distributions (boxplots). Statistical analyses confirmed significant effects for all three measures.

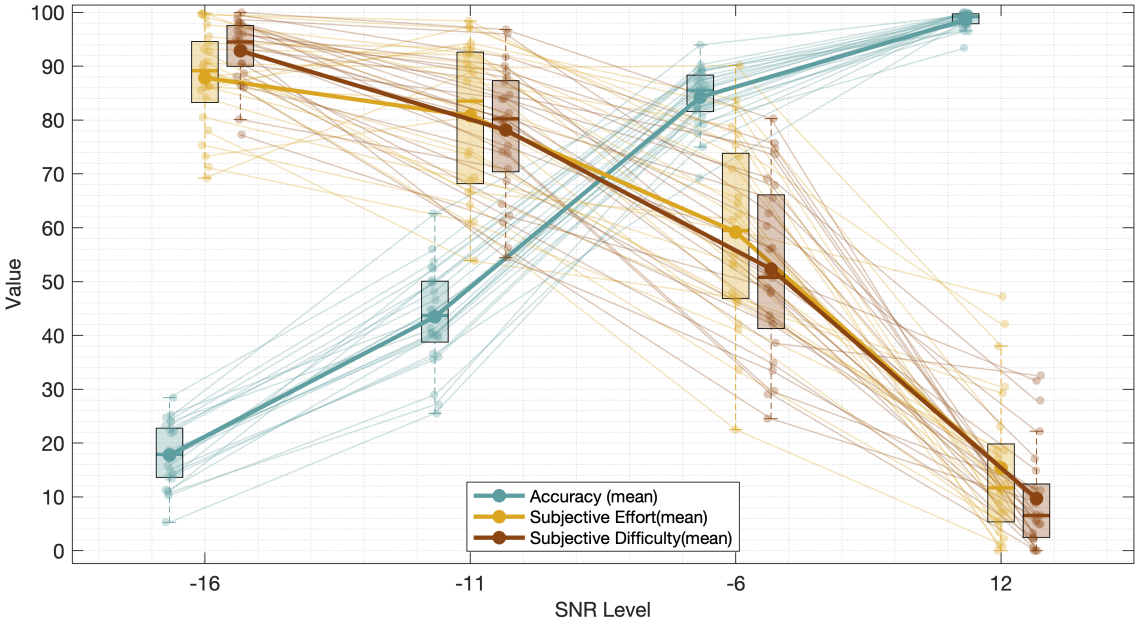


Figure. 13.1. Accuracy, subjective effort, and perceived difficulty across different SNR levels

Individual participant are shown as faint dots and connected by lines, while bold lines represent group averages. Boxplots display the distribution of responses at each SNR level. Accuracy improves with SNR, while effort and difficulty ratings decline.

Performance across different SNR Levels Performance, measured as the percentage of correctly identified words, was significantly affected by the SNR level. Friedman test indicated a highly significant overall difference across conditions ($p < 001$). To pinpoint where these differences occurred, post-hoc tests were conducted. These revealed significant differences between all pairs of SNR levels ($p < 0.05$). As clearly depicted by the upward trend of the bold green line in Figure 16.1, accuracy improved significantly

with each successive increase in SNR, demonstrating the effectiveness of the SNR manipulation on task performance

Subjective Effort across different SNR Levels Participants' ratings of invested effort also varied significantly across SNR levels, as shown by a repeated ANOVA test($p < .001$). Post-hoc analysis using Tukey HSD tests provided further detail. While effort generally decreased significantly as SNR improved (most pairwise $p < 0.001$), there was notably no significant difference in reported effort between the two most challenging conditions, -16 dB and -11 dB ($p < 0.26$).

Significant reductions in effort were found between all other adjacent and non-adjacent levels (e.g., -11 dB vs -6 dB, -6 dB vs 12 dB, -11 dB vs 12 dB, etc.). This pattern is visible in Figure 13.1, where the mean effort ratings (bold yellow line) for -16 dB and -11 dB are relatively close, before dropping more substantially at -6 dB and further still at 12 dB.

Subjective difficulty across different SNR Levels Perceived task difficulty was likewise significantly affected by SNR. Friedman test confirmed a significant overall effect ($p < .001$). Post-hoc Dunn's tests indicated that, similar to accuracy, perceived difficulty differed significantly between all pairs of SNR levels (all $p < 0.05$). Figure 16.1 illustrates this clearly, with the mean difficulty rating (bold dark red line) decreasing significantly at each step increase in SNR

Complementing the behavioural findings presented above, the subsequent sections report on the analysis of the physiological data collected throughout the experiment. Results from Pupillometry, Galvanic Skin Response (GSR), Respiration, Electrocardiography (ECG), and Electroencephalography (EEG) will now be presented.

Correlation Between Subjective Effort, Difficulty, and Accuracy Figure 13.2 presents the relationships between subjective effort ratings, subjective difficulty ratings, and recognition accuracy across all SNR conditions in data averaged from experiment 1 and 2. Subjective effort and difficulty were both measured using participant ratings from 0 to 100, while recognition accuracy reflects the percentage of words correctly recognised (also scaled 0 to 100).

Figure 13.2 illustrates the relationships between subjective effort ratings, subjective difficulty ratings, and recognition accuracy in Experiment 3. For each participant, ratings were averaged across SNR conditions to obtain a single between-subject estimate for each measure. Subjective effort and difficulty were obtained using participant ratings on a 0–100 scale, while recognition accuracy reflects the percentage of correctly recognised words (0–100%).

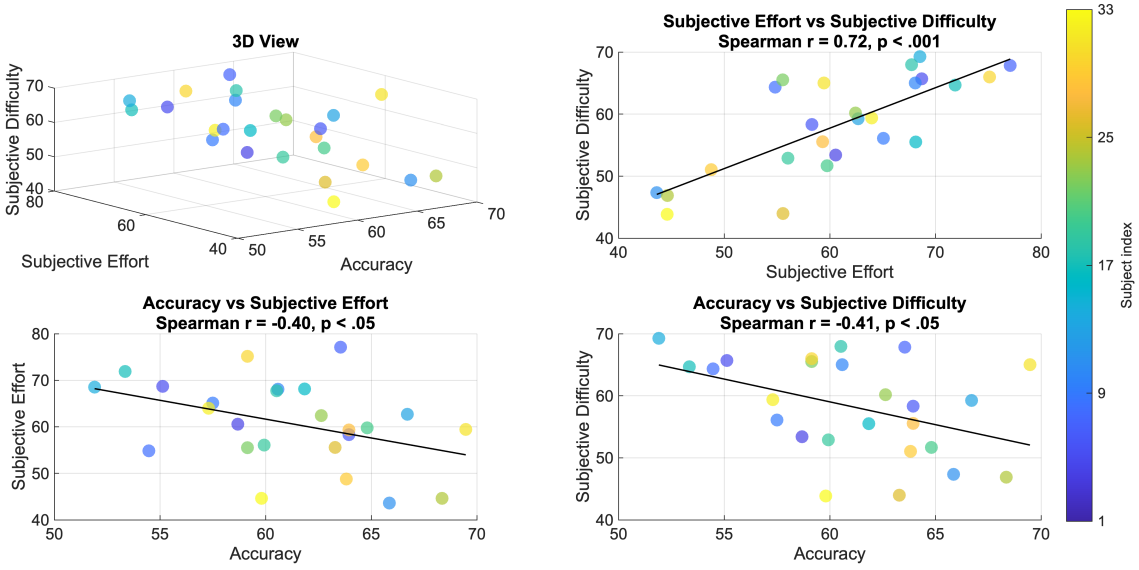


Figure. 13.2. Correlations among subjective effort, subjective difficulty, and recognition accuracy

This figure illustrates the between-subject relationships among subjective effort ratings, subjective difficulty ratings, and recognition accuracy in Experiment 3. Each point represents an individual participant, averaged across SNR conditions, and is colour-coded by subject index. Effort and difficulty ratings were collected on a 0–100 subjective scale, while recognition accuracy reflects the percentage of correctly identified words (0–100%).

Difficulty ratings were not normally distributed, associations were quantified using Spearman rank correlations. A strong positive relationship was observed between subjective effort and subjective difficulty ($\rho = 0.72$, $p < .001$). Recognition accuracy showed moderate negative associations with both effort ($\rho = -0.40$, $p < .05$) and difficulty ($\rho = -0.41$, $p < .05$), indicating that participants who reported greater perceived effort or difficulty tended to achieve lower accuracy.

Linear trend lines are overlaid on each 2D projection to visualise the direction of these associations. Overall, the results highlight a consistent coupling between subjective experience and objective performance at the between-subject level.

All associations were quantified using Spearman rank correlations because subjective difficulty ratings exhibited non-normality, The analysis revealed systematic relationships between the three measures. Subjective effort and subjective difficulty were strongly positively correlated ($\rho = 0.72$, $p < .001$), indicating that participants who reported greater effort also tended to perceive the task as more difficult. Recognition accuracy showed moderate negative associations with both effort ($\rho = -0.40$, $p < .05$) and difficulty ($\rho = -0.41$, $p < .05$), suggesting that participants who experienced the task as more effortful or more difficult generally achieved lower objective performance.

These between-subject patterns highlight a reliable coupling between subjective experience and objective task performance, even when SNR-specific fluctuations are averaged out.

13.2 Physiological Data Analysis Methods Overview

Extracting Average Trial Response Compared to Study 1, where both ATR and TCR were extracted from physiological responses (EEG, GSR, and pupillometry), this study focused solely on average trial response (ATR in study 1) analysis. In Study 1, only a single SNR level was used, allowing analysis of response changes across trials within a consistent condition.

In contrast, Study 2 involved multiple SNR conditions, each presented in two blocks of 20 trials, with block order randomised across participants to reduce sequence effects. As a result, trial-by-trial (i.e., TCR) analysis was not pursued, since systematic trends over time were not expected. Although intra-block dynamics could be explored in future work, the current analysis focuses on condition-level comparisons across SNR levels (see Figure 13.3).

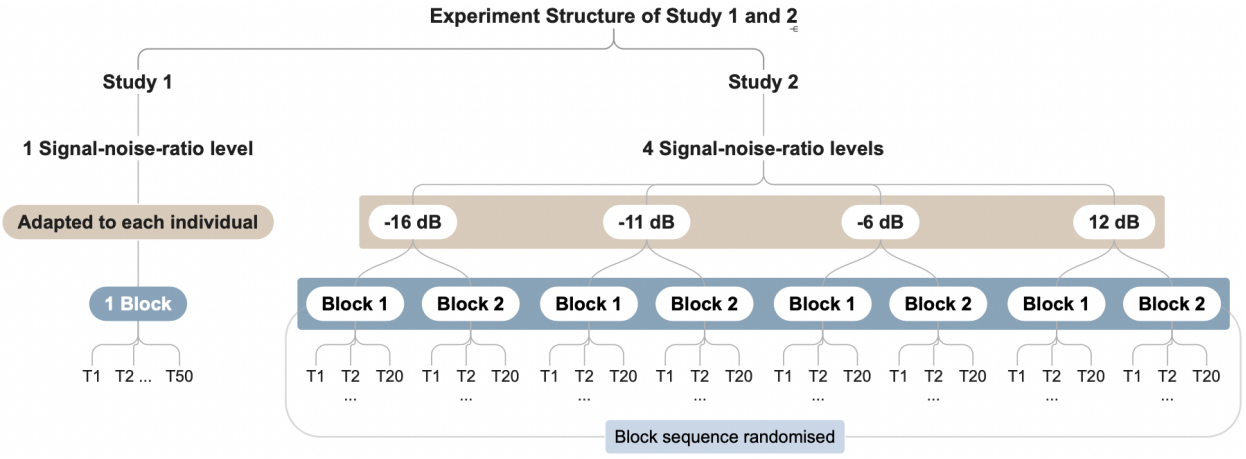


Figure. 13.3. Comparison of experimental structure in Study 1 and Study 2

Study 1 involved a single signal-to-noise ratio (SNR) level adapted to each individual, with 50 trials presented in one block. Study 2 included four fixed SNR levels, each presented in two blocks of 20 trials. The order of blocks was randomised to minimise sequence effects. In Study 1, trial-wise responses could be tracked over time (TCR), whereas in Study 2, analysis was conducted at the block (SNR) level. Note: T means Trial in the figure.

Data Extraction Strategy Physiological data were first extracted over a defined trial window, starting 1.5 seconds before the onset of the first word (baseline period) and extending to the end of the retention period. For some measures known to exhibit slower response dynamics (e.g., GSR, respiration), data extraction was extended into close to the beginning of the subsequent trial to ensure that late responses were fully captured.

Within each trial, values were further extracted at specific key events: baseline, word start (first word onset), retention start (following the final word), and retention end

(conclusion of the retention phase). These points were selected to reflect distinct stages of cognitive and sensory processing during the listening and memory task.

In addition to extracting values at these discrete time points, changes between points were also computed to quantify the level of physiological adjustment during task phases. For example, the difference between the value at retention start and word start was calculated to assess the magnitude of response modulation during the listening period. Such change scores were used to compare dynamic physiological responses across different SNR conditions.

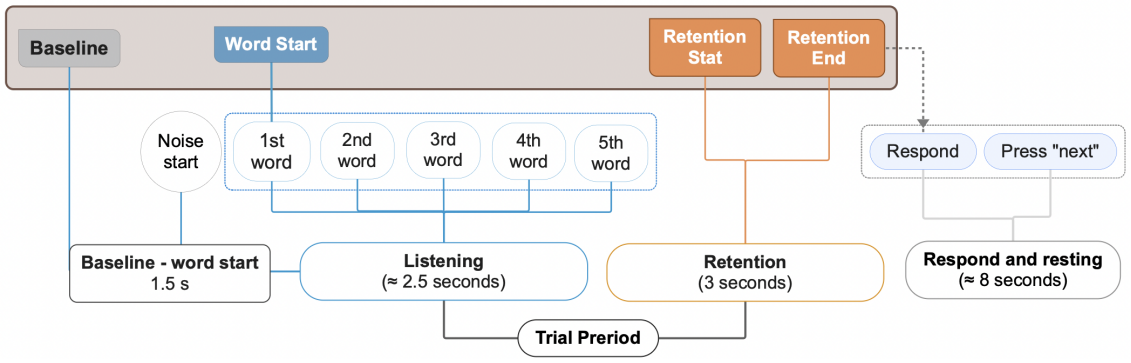


Figure. 13.4. Trial structure: Listening, Retention, and Response periods

Trial design for pupil data analysis. Each trial starts with a noise onset, followed by the presentation of five words during the Listening period (approximately 2 seconds). Participants then enter a 3-second Retention period where no new auditory information is given but memory is engaged. Finally, participants respond and rest during an 8-second interval before proceeding to the next trial.

Data were first extracted and averaged at this time window for each SNR level, from baseline to retention end (or to start of next trial for some measures). Later, data were further extracted at specific time points at those events (e.g., Word start, Retention start, Retention end), to compare the effects of onset or events (e.g. word start comparing to baseline), or the response change during listening and retention phase across different SNRs (e.g. change between word start and retention start, across different SNRs).

Within Individual Consistency - Permutation Test To evaluate whether participants' physiological responses were more consistent with themselves than with others, permutation testing was used. Specifically, we tested whether the correlation between a participant's own response time-courses across sessions was significantly higher than the correlations between randomly paired participants.

The procedure involved first computing the average correlation between each participant's two experimental sessions. Next, participant labels were randomly reassigned, and correlations between mismatched pairs were calculated and averaged. This random reassignment was repeated 1,000 times to generate a distribution of average correlations expected by chance.

The observed within-participant correlation was then compared against this distribution to assess its statistical significance. A high within-participant correlation relative to the

random distribution would indicate that participants exhibited stable temporal dynamics in their physiological responses across experiments. Permutation testing was chosen as it does not assume normality and provides a robust non-parametric framework for significance testing (Good, 2000).

Between Individual Differences - Clustering Analysis K-means clustering is a widely used unsupervised machine learning algorithm that partitions a dataset into K non-overlapping clusters by minimising the within-cluster variance (Jain et al., 1999; MacQueen, 1967). Each cluster is represented by the centroid (mean) of the points assigned to it. K-means is computationally efficient, scalable to large datasets, and produces partitions that are easily interpretable.

In this study, rather than clustering based on absolute amplitudes and Euclidean distance, we computed pairwise similarity between participant response curves using Pearson's correlation coefficient. This approach emphasises the *shape* or *temporal dynamics* of physiological responses, allowing participants to be grouped based on the similarity in the structure of their responses over time, rather than on overall signal amplitude.

To determine the optimal number of clusters K, we employed the elbow method. This technique involves plotting the total within-cluster sum of squares (WCSS) against increasing values of K (See figure 13.11, Page 146 in for an example). The WCSS measures the compactness of clusters(how close together the points are inside each cluster), with lower values indicating tighter, more coherent groupings. Initially, adding more clusters substantially reduces WCSS, but after a certain point, the rate of decrease diminishes. The "elbow" of the curve — where the reduction in WCSS begins to level off — suggests an appropriate trade-off between model simplicity and data fit (Thorndike, 1953). The value of K corresponding to this elbow point was selected for clustering.

Following the identification of the optimal number of clusters, the K-means algorithm was applied using the correlation-based similarity measure to partition participants into groups exhibiting similar temporal response profiles.

We subsequently compared subjective effort ratings, perceived difficulty, and performance accuracy across the identified clusters to evaluate whether group membership was associated with differences in behavioural outcomes.

Clustering Result Agreement across different SNR levels Following the clustering analysis, we evaluated the consistency of cluster memberships across different SNR levels using the Adjusted Rand Index (ARI). The ARI quantifies the similarity between two clustering results by examining all possible pairs of participants and checking whether each pair was grouped together or separately in both clustering solutions (Hubert & Arabie, 1985).

It then adjusts this similarity score by accounting for the amount of agreement that would be expected purely by chance. An ARI of 1 indicates perfect agreement, meaning the clustering structures are identical, while an ARI close to 0 suggests that any agreement is likely random. Negative ARI values indicate less agreement than would be expected by chance. This analysis allowed us to assess whether participants' physiological response patterns remained stable across different listening conditions. An illustrative example of ARI-based comparison is shown in Figure 13.18, Page 152.

13.3 Pupillometry

13.3.1 Data Overview

Pupillometry was recording thought Pupil Core glasses and Pupil Lab software. Pupilometry data analysed here are pupil diameter, which was cleaned and averaged two eyes.

Figure 13.6 illustrates the average pupil response in data averaged from experiment 1 and 2 across four signal-to-noise ratio (SNR) conditions: -16, -11, -6, and 12 dB. To highlight task-evoked changes more clearly, the first time point of each individual trace was subtracted, aligning all responses to a common zero baseline.

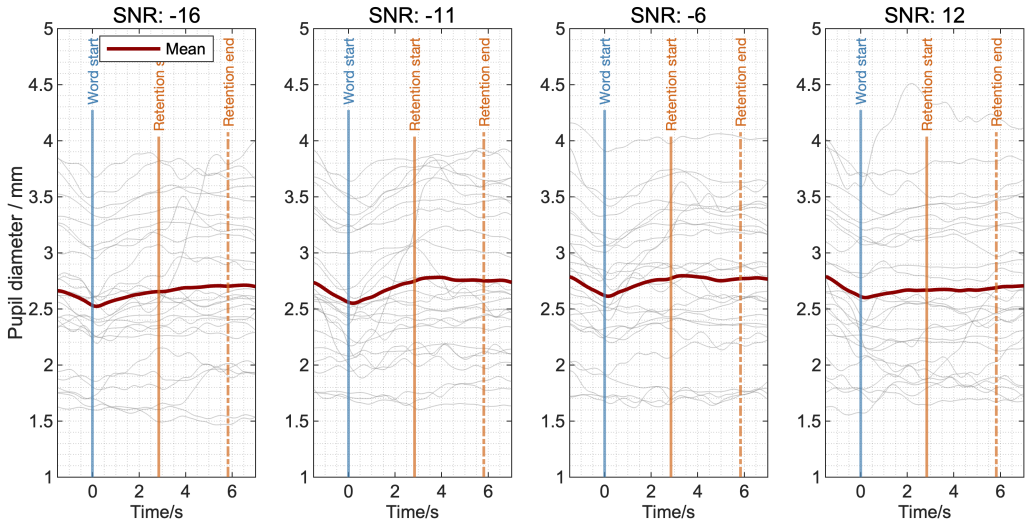


Figure. 13.5. Pupil diameter grand average (original scale) - data averaged from experiment 1 and 2, all SNR levels

Grand average pupil responses across all participants and trials, combining data from Experiment 1 and Experiment 2. Participants repeated the same paradigm in the second session after a one-week gap. Unlike baseline-subtracted plots, this figure shows the pupil diameter in its original scale, providing a view of absolute changes in pupil size across time and SNR levels.

Despite natural variability, a consistent pattern emerges: an initial constriction following word onset, followed by dilation through the sentence and retention period, becoming more pronounced as intelligibility increases.

Across all SNR conditions, the pupil exhibits an initial dip shortly after sentence onset, followed by a recovery and a prominent peak that generally coincides with the retention phase. These fluctuations are consistent with the interpretation that pupil size reflects cognitive processing demands: the early dip may be associated with auditory onset or attentional orienting, while the subsequent peak reflects sustained effort during memory retention.

Interestingly, both the magnitude and timing of these features vary with SNR. In the more degraded conditions (e.g. -16 dB), the pupil dilates more rapidly and to a greater extent, whereas in the clearest condition (12 dB), the response is more gradual and comparatively subdued.

13.3.2 Task-Evoked Changes in Pupil Diameter

To better understand the pupil dynamics seen in the average traces (Figure 13.5), figure 13.6 shows the the same data as Figure 13.5, with the first value of the data subtracted, to shift all data start from zero. There is a clear change in pupil size reacting to the event, for instance, pupil size drop before the first word start, and then increase considerably when retention start.

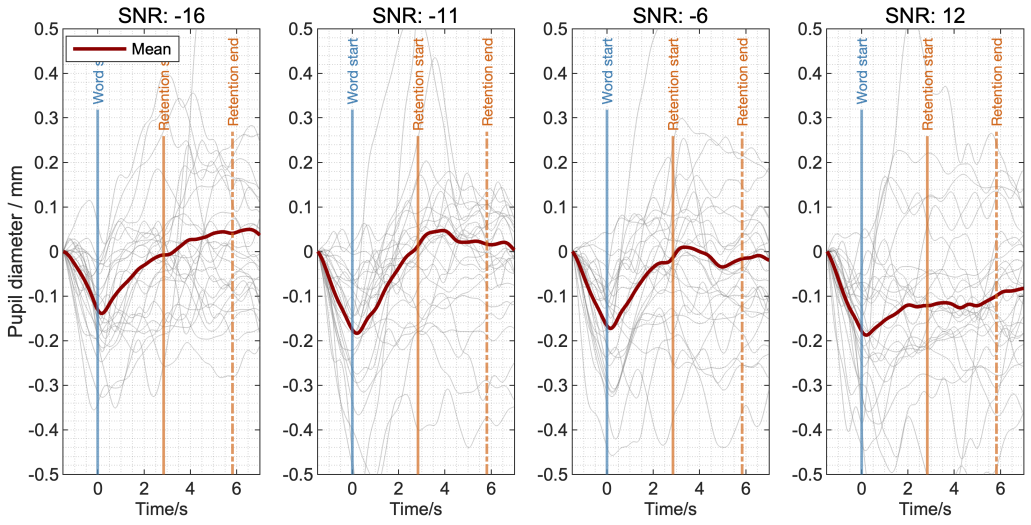


Figure. 13.6. Pupil diameter grand average (First-Point Subtracted) - data averaged from experiment 1 and 2, all SNR levels

Grand average pupil diameter is the average of Experiments 1 and 2 across SNR levels. different from Figure 13.5 , data plotted here subtracting the first time point from each data trial. This subtraction centres the data to a zero baseline, highlighting relative change from the beginning of sentence processing. Clear differences in the timing and magnitude of the initial dip and subsequent peak emerge across SNR conditions.

We extracted pupil diameter at four events: before the sentence began (Baseline), at the onset of the first word (Word Start), and at the sentence offset (Retention Start), and Retention End (time points as shown in Figure 13.5 and 13.6). These four time points

mark the transitions from rest to listening, and from listening to remembering. We then compared the pupil size between these time points across all SNR levels (see Figure 13.7).

Paired t-tests was used for comparing the difference when data is normally distributed, otherwise non-parametric Wilcoxon signed-rank tests were applied.

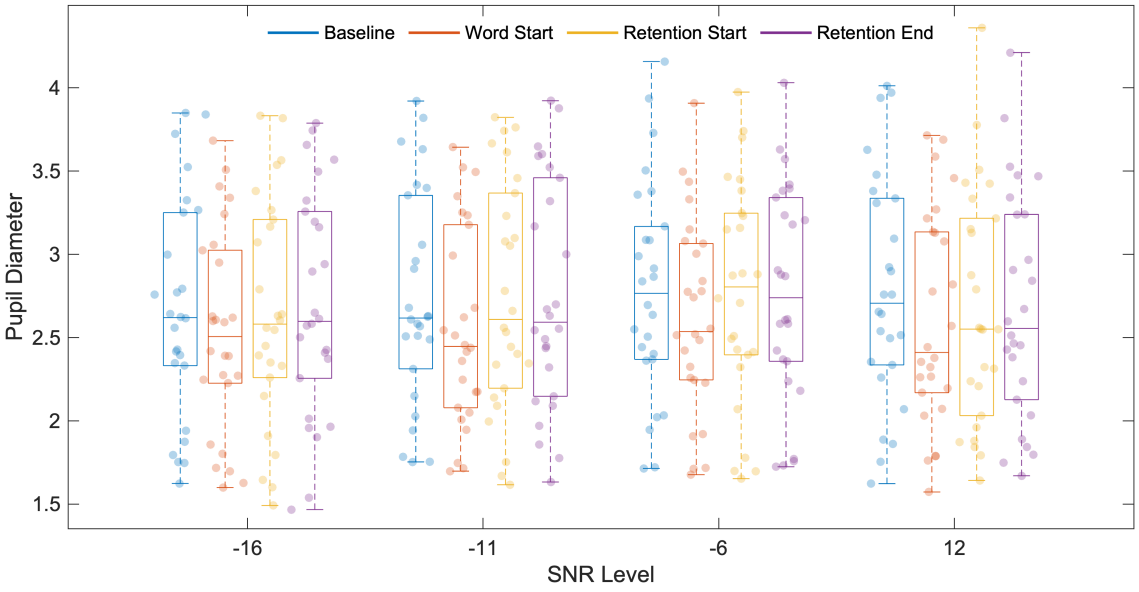


Figure. 13.7. Pupil Diameter at four key events across SNR Levels - average experiment 1 and 2

Grouped boxplots showing pupil diameter for four cognitive events - Baseline, Word Start, Retention Start, and Retention End - at each signal-to-noise ratio (SNR) level in Experiment 3. Each coloured box represents one variable per SNR group.

This layout visualises how pupil responses evolve through the task timeline under different listening conditions, highlighting individual variability, event-specific (e.g. onset of first word), and SNR sensitive response.

As shown in Tables 13.1 and 13.2, pupil diameter significantly decreased from Baseline to Word Start across all SNRs ($p < 0.001$), and significantly increased from Word Start to Retention Start in all conditions except the highest SNR (12 dB), where the effect was not significant ($p = 0.0912$). These consistent changes suggest that the pupil reflects different phases of the listening task - an initial orienting response followed by increased mental effort during memory retention.

Table. 13.1. Task Evoked Response Comparison Result: Baseline vs Word Start

SNR	N	Test Used	p-value
-16	26	Paired t-test	< 0.001
-11	26	Paired t-test	< 0.001
-6	26	Paired t-test	< 0.001
12	26	Paired t-test	< 0.001

Table. 13.2. Statistical Comparison: Word Start vs Retention Start (Data Averaged from Two Experiments)

SNR	N	Test Used	p-value
-16	26	Paired t-test	< 0.001
-11	26	Wilcoxon	< 0.001
-6	26	Wilcoxon	< 0.001
12	26	Wilcoxon	0.0912

13.3.3 Pupil Diameter Changes at Different SNR Levels

To investigate whether pupil responses were affected by task difficulty, we compared average pupil diameter across the four SNR conditions. As illustrated in Figures 13.5 and 13.7, there appeared to be observable differences across SNR levels.

We examined pupil diameter at several key points: Baseline, Word Start, Retention Start, Retention End, the average across the Listening period, and the average across the Retention period. Group-level comparisons using the Friedman test revealed no significant overall differences in mean pupil size between SNR levels for any of these variables ($p > 0.05$).

However, subsequent pairwise comparisons revealed specific differences between certain SNR levels. A summary of these contrasts is presented in Figure 13.8.

The significance matrix (Figure 13.8) highlights specific contrasts where pupil diameter differed significantly between SNR conditions. Notably, pupil size at Retention Start showed significant increases between the lowest SNR level (-16 dB) and both mid-range conditions (-11 dB and -6 dB), with the latter reaching a higher level of significance ($p < .01$).

Similarly, the average pupil diameter during the Listening period was significantly greater at -6 dB compared to -16 dB. The average Retention period also showed significant differences between low and mid SNRs (-16 dB vs -11 dB and -6 dB), though not between the higher SNRs.

Interestingly, the change in pupil size from Word Start to Retention End-a dynamic marker of task-evoked dilation-showed significant differences between the mid (-11 dB, -6 dB) and high (12 dB) SNR conditions, indicating stronger pupil diameter difference during listening period phases as intelligibility improves.

In contrast, other comparisons such as Baseline, Word Start, and Retention End did not reveal significant differences across SNRs, suggesting that pupil modulation was more

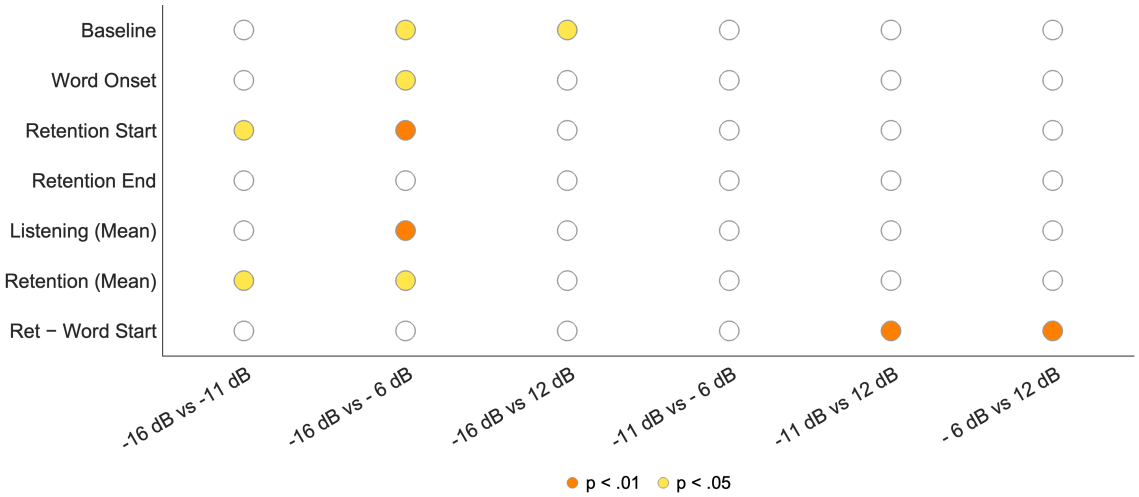


Figure. 13.8. Significance of Pairwise Comparisons of Pupil Diameter - Averaged data from Experiment 1 and 2

Pairwise statistical significance matrix showing pupil diameter differences across SNR conditions in Pupil diameter averaged from Experiment 1 and 2. Each row represents a time point or period of interest (e.g., Baseline, Word Onset, Retention phases), and each column a comparison between two SNR levels.

Circle colour indicates significance level. The results highlight consistent modulation of pupil responses across SNR conditions during Retention Start, Retention Mean, and Listening periods, with significant contrasts especially between low and high intelligibility conditions.

sensitive to SNR during the early and sustained phases of cognitive load rather than at static time points.

13.3.4 Within Individual Consistency - Permutation Test Result of Pupil Diameter

To assess the consistency of pupil responses across experiments, a permutation test was conducted for each SNR condition (Figure 13.10). The results revealed significant within-subject correlation at SNR –6 and SNR –16, whereas SNR –11 showed no statistical evidence of cross-experiment similarity.

13.3.5 Clustering Result of Pupil Diameter (presenting per SNR level

To explore individual differences in pupil responses under low intelligibility, we first applied the elbow method to determine the optimal number of clusters for pupil diameter time courses at SNR –16 dB. As shown in Figure 13.11, a clear drop in within-cluster variance indicated that two clusters ($k = 2$) best captured the structure in the data.

The clustering analysis was based on data from the trial period, defined as 1.5 seconds before word onset to 2 seconds after retention offset. This time window was selected to

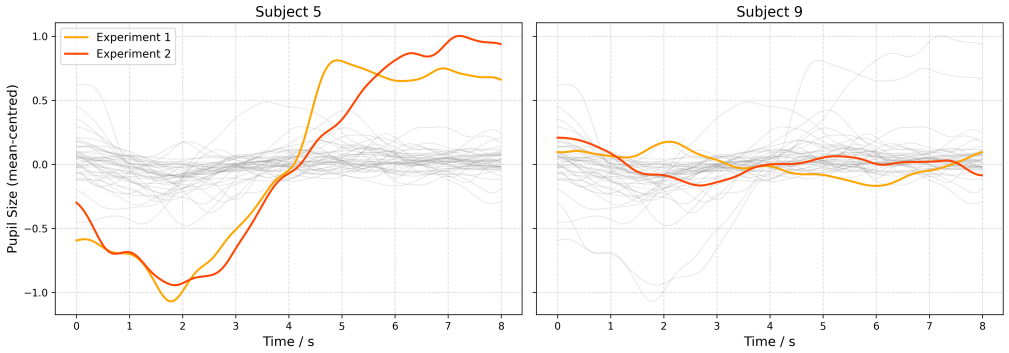


Figure. 13.9. Example of Individual Consistency across Two Experiments (Pupil Diameter, SNR - 6 dB)

Example of two participants whose pupil traces showed high within-subject similarity at SNR -6 - a condition with statistically significant correlation between Experiment 1 and 2 ($p = 0.000$).

The highlighted orange and red lines represent the same subject across experiments, while grey traces represent other participants. Signals are mean-centred to emphasise shape rather than absolute size. This figure illustrates the strong temporal consistency that underpins the group-level result.

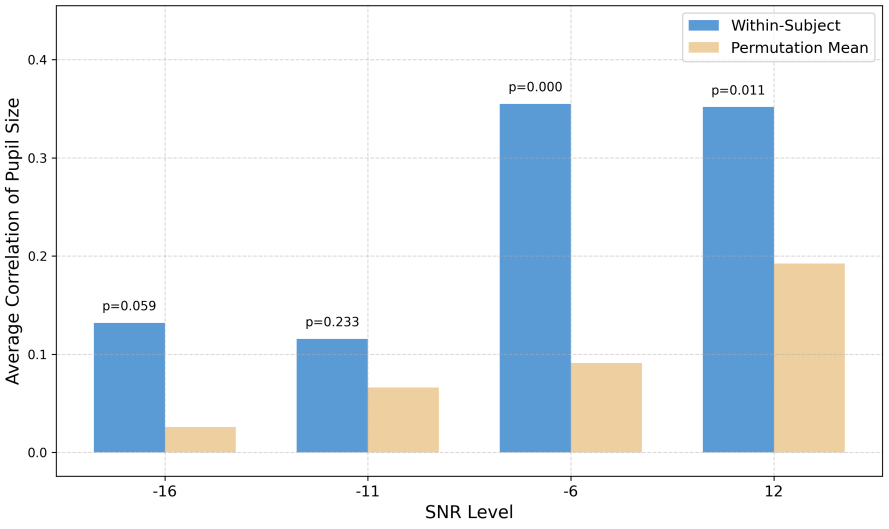


Figure. 13.10. Permutation test of within-subject correlations in pupil diameter average trial response

Average within-subject pupil correlation across SNR conditions, compared against a permutation-based null distribution. Blue bars represent average Pearson correlations between Experiment 1 and 2 for each subject at each SNR.

Beige bars show the average similarity under random subject pairing. P-values from the permutation test are annotated above each bar group, indicating whether within-subject consistency was significantly greater than expected by chance.

capture the full event-related pupil response while avoiding overlap with subsequent trials. Unlike other physiological signals such as GSR, respiration, or heart rate—which often require longer segments due to their slower response dynamics—pupil diameter changes occur more rapidly and with minimal latency.

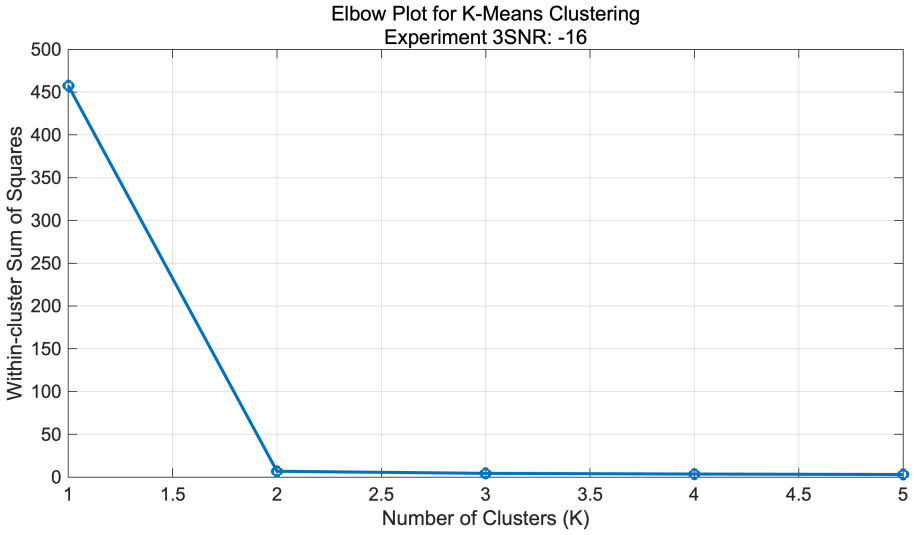


Figure. 13.11. K-Means elbow plot of cluster sum of squares (WCSS) in within subject correlations for Pupil Diameter Clustering (-16 dB)

Elbow method applied to K-means clustering for pupil data at SNR -16 in data averaged from experiment 1 and 2. The plot shows the within-cluster sum of squares (WCSS) as a function of the number of clusters (K). A clear "elbow" is observed at $K = 2$, suggesting that two distinct clusters best capture the underlying variation in pupil response patterns at this noise level.

Clustering result at SNR : -16 dB (Pupil Diameter) Moreover, because pupil measurements fluctuate quickly and are sensitive to noise, using a longer time window could introduce additional variability and obscure meaningful patterns. The chosen time frame therefore balances temporal resolution with signal clarity for effective clustering.

Based on the elbow result, we performed *k*-means clustering and identified two distinct pupil response profiles. Figure 13.12 shows the averaged time courses for each cluster, with shaded standard error. Cluster 1 showed a stronger dip and later recovery, while Cluster 2 displayed flatter, more sustained constriction.

To investigate whether these physiological clusters reflected meaningful behavioural or subjective differences, we compared the groups on three measures: task accuracy, perceived difficulty, and self-reported effort. These comparisons are shown in Figure 13.13.

Although Group 1 tended to show higher accuracy and report lower effort and difficulty than Group 2, no statistically significant differences were found. This suggests that the physiological differences in pupil dynamics do not clearly map onto behavioural outcomes in a categorical way.

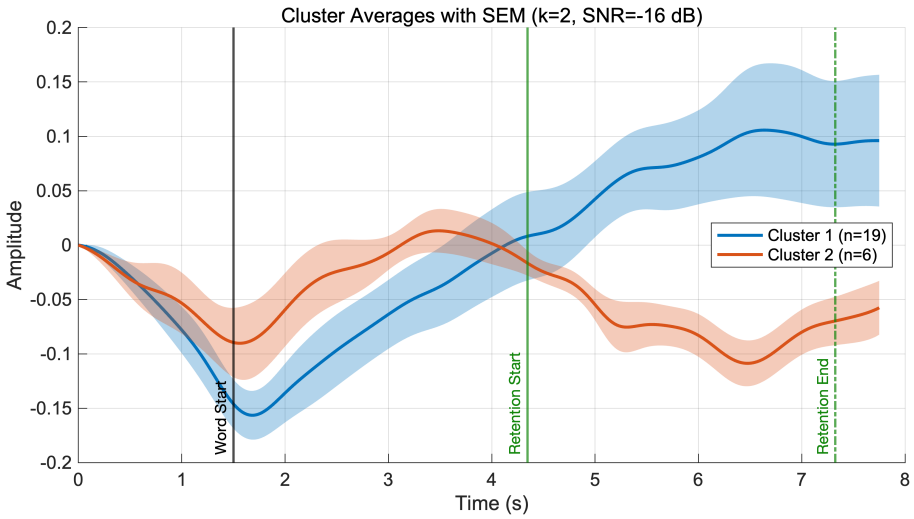


Figure. 13.12. Clustering result - average trial response of pupil diameter with standard error shading (baseline corrected, k = 2, SNR = -16 dB)

Mean pupil diameter time courses for each cluster (k = 2) at SNR - 16 dB (starting point corrected to zero). To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Shaded regions indicate the standard error of the mean (SEM) across participants. Cluster 1 (blue, $n = 19$) showed stronger pupil constriction and subsequent dilation, while Cluster 2 (orange, $n = 6$) displayed more sustained constriction with flatter recovery.

Still, the visible trends may reflect subtle variation in how participants regulate effort or adapt to degraded speech. The consistent clustering pattern suggests that pupil dynamics capture stable internal response styles, even if these are not always mirrored in task performance.

Together, the elbow plot, time course clusters, and behavioural comparisons highlight how pupil diameter can reveal individual differences in cognitive processing during listening, offering insight beyond what behavioural metrics alone can provide.

Clustering result at SNR : -11 dB (Pupil Diameter) To explore individual differences in pupil responses under SNR level at -11 dB, we applied the elbow method to determine the optimal number of clusters for pupil diameter time courses at SNR –11 dB. Using the Elbow method, the best group numbers were suggested at 2.

Based on this, we applied *k*-means clustering and identified two distinct pupil response profiles. Figure 13.14 shows the average time course for each cluster, with shaded standard error. Cluster 1 showed a stronger and more prolonged constriction with a delayed recovery, while Cluster 2 exhibited a flatter initial dip followed by a faster and more sustained recovery over time.

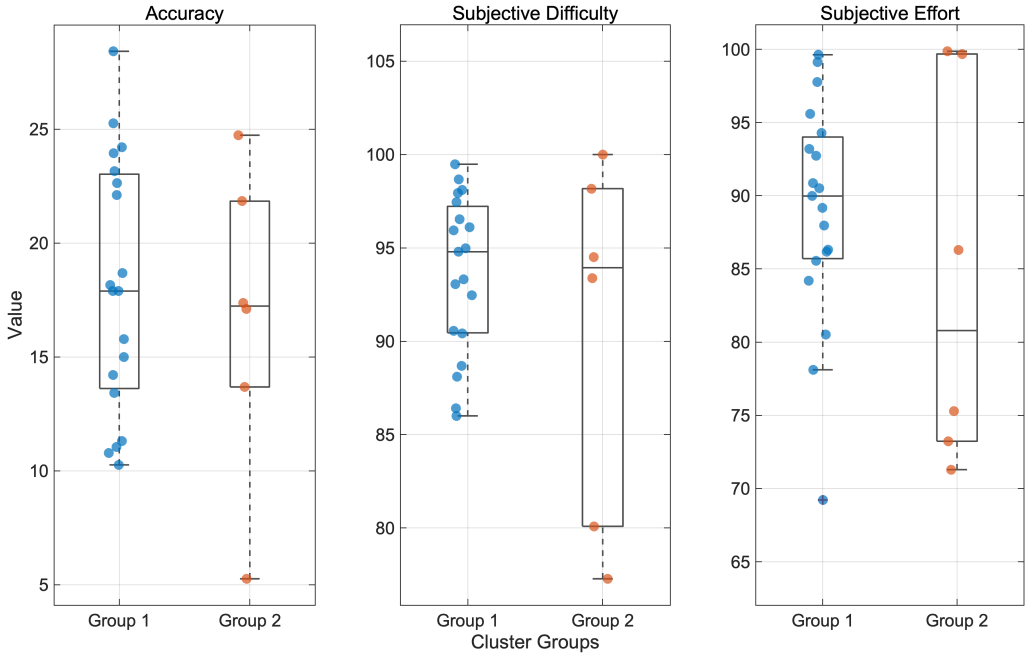


Figure. 13.13. Differences in Performance and Subjective Ratings Between Pupil Diameter Clusters (SNR -16 dB)

Boxplots showing group differences in task performance and self-reported experience between the two clusters identified from GSR time courses at SNR -16 dB. Group 1 and Group 2 correspond to participants classified via clustering analysis.

Plotted measures include task accuracy (left), subjective difficulty ratings (middle), and perceived effort (right). Each point represents an individual participant. Differences in scores between groups highlight how physiological response patterns may relate to both behavioural outcomes and subjective experience.

Although this clustering revealed stable differences in physiological response profiles, there were no statistically significant differences in behavioural or subjective measures between the two groups.

Together, these results demonstrate how pupil dynamics at SNR -11 dB continue to reveal distinct internal response styles, consistent with findings at lower intelligibility levels.

Clustering result at SNR : -6 dB (Pupil Diameter) To examine individual variability in pupil dynamics under relatively intelligible conditions, we applied the elbow method to determine the optimal number of clusters for pupil diameter time series at SNR - 6 dB. The Elbow test shows most substantial reduction in within-cluster variance occurred at $k = 2$, suggesting After applying Elbow method, a two-cluster solution was most appropriate.

Using this criterion, we performed k-means clustering and identified two distinct patterns of pupil responses. As shown in Figure 13.15, Cluster 1 was characterised by a more pronounced and prolonged constriction with relatively little recovery, whereas

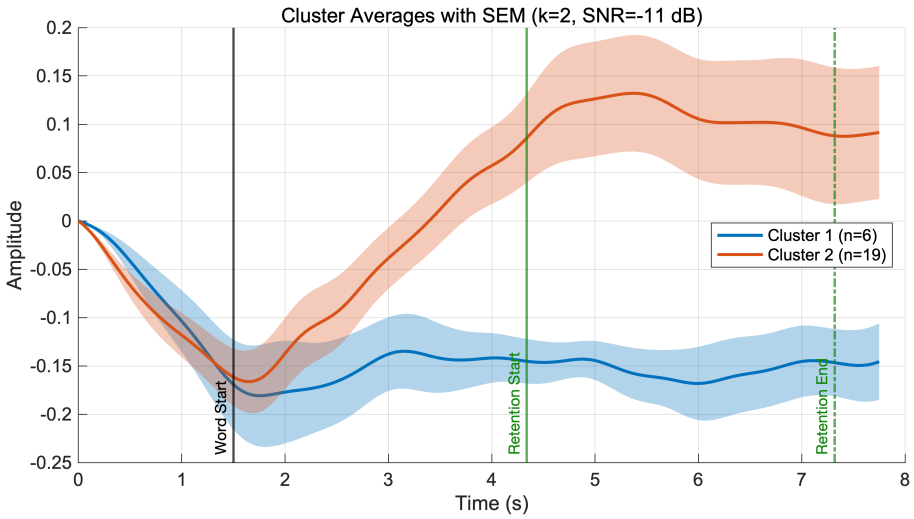


Figure. 13.14. Clustering result - average trial response of pupil diameter with standard error shading (baseline corrected, k = 2, SNR = -11 dB)

Clustered pupil response time courses at SNR –11 dB. Two clusters were identified using k-means clustering ($k = 2$), based on the elbow criterion. The plot displays mean pupil diameter across time for each group, with shaded areas indicating the standard error of the mean (SEM). To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Cluster 1 (blue, $n = 7$) exhibits a larger constriction followed by a more delayed and attenuated recovery. Cluster 2 (red, $n = 19$) shows a shallower constriction and a faster, sustained recovery during the retention interval.

Cluster 2 demonstrated a milder initial dip followed by a strong and sustained dilation throughout the retention phase.

Despite these physiological distinctions, behavioural performance and subjective reports did not significantly differ between clusters.

Clustering result at SNR : 12 dB (Pupil Diameter) To explore individual variability in pupil dynamics at the clearest speech level, we applied the elbow method to determine the optimal number of clusters at SNR 12 dB. After applying Elbow method, the presence of two primary response patterns were decided.

The resulting average time courses for each cluster are plotted in Figure 13.16. Although both clusters initially constricted following stimulus onset, Cluster 1 displayed a more sustained and deeper constriction during the retention period, whereas Cluster 2 recovered more rapidly and stabilised around baseline levels. This pattern suggests some divergence in cognitive or autonomic engagement even at favourable listening conditions.

However, similar to previous SNR levels, no significant behavioural or subjective differences were observed between the two pupil-based clusters.

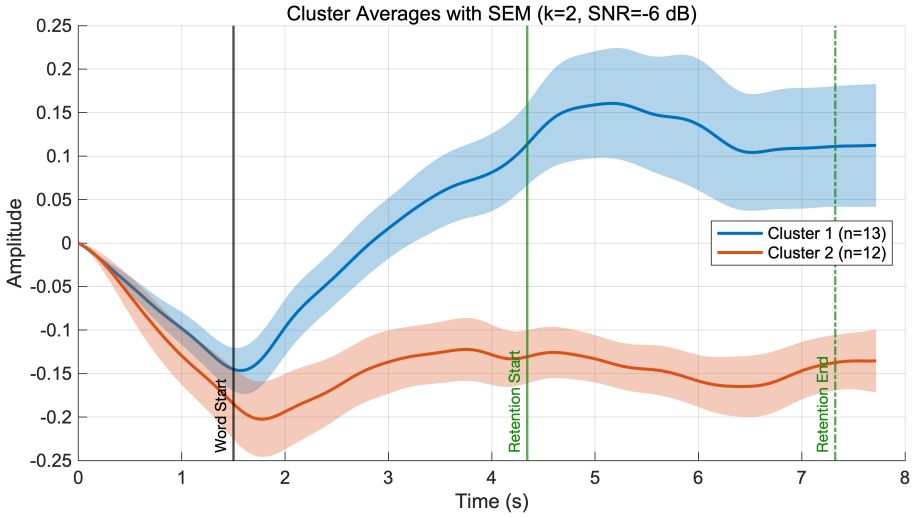


Figure. 13.15. Clustering result - average trial response of pupil diameter with standard error shading (baseline corrected, k = 2, SNR = -6 dB))

Pupil diameter responses for the two clusters identified at SNR –6 dB. Cluster 1 (blue, $n = 13$) showed a larger early constriction followed by a flatter retention period. Cluster 2 (orange, $n = 13$) exhibited a shallower initial constriction but a strong sustained dilation across the retention phase. To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero.

Despite balanced cluster sizes, statistical comparisons between key retention-related timepoints revealed no significant group differences, suggesting inter-subject variation without robust cluster separation.

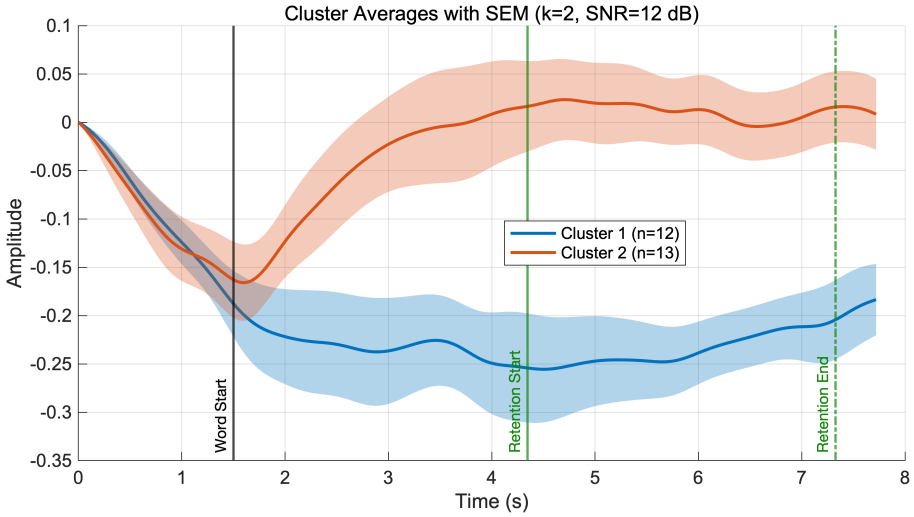


Figure. 13.16. Clustering result - average trial response of pupil diameter with standard error shading (baseline corrected, k = 2, SNR = 12 dB)

Cluster-averaged pupil diameter time courses with standard error for SNR 12 dB. To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Two distinct response profiles were identified using *k*-means clustering. Cluster 1 exhibited deeper and more sustained constriction, while Cluster 2 showed a faster recovery, despite no differences in behavioural or subjective performance between groups.

Having established that pupil responses can be meaningfully clustered at each SNR level, we next examined whether these groupings remained consistent across listening conditions. Specifically, we assessed the degree of agreement in cluster membership across SNR levels to determine whether the same individuals tended to exhibit similar pupil response patterns in different SNR levels.

13.3.6 Clustering Result Agreement Across Different SNR Levels (Pupil Diameter)

To assess whether individual participants exhibited stable clustering behaviour across SNR levels, we compared pupil-based cluster assignments at each noise condition. Figure 13.17 visualises cluster membership by subject across all four SNRs. Each row corresponds to one SNR level, while each column represents an individual subject.

This visualisation suggests a high degree of variability. Only a handful of participants (e.g., Subject 18, 20, and 31) consistently fell into the same cluster across all conditions. In contrast, many switched between clusters depending on the listening difficulty, highlighting the state-dependent nature of the pupil-based response profiles.

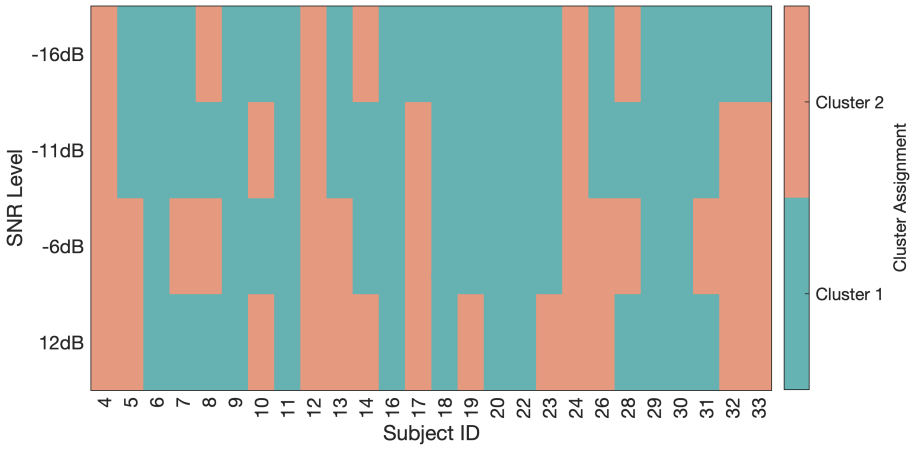


Figure. 13.17. Cluster membership across SNR levels - pupil diameter

Cluster assignments for each participant across all SNR levels. Each row represents an SNR condition, and each column indicates a subjects cluster membership. Colour indicates whether the participant was assigned to Cluster 1 or Cluster 2 at that SNR.

To quantify the agreement between cluster labels, we computed the Adjusted Rand Index (ARI) for each SNR pair, shown in Figure 13.18. The ARI measures the similarity of clustering solutions while adjusting for random chance, with values close to 1 indicating high similarity and values near 0 suggesting random agreement.

The results confirm the visual pattern: agreement between clustering solutions was generally low. The highest ARI value (0.537) was observed between SNR –6 dB and 12 dB, suggesting moderate consistency in these conditions. All other comparisons yielded ARI values below 0.36, with the lowest being 0.075 between SNR –16 dB and 12 dB.

These findings imply that pupil response pattern - and the groupings they imply - are highly sensitive to contextual demands. Rather than reflecting stable traits, cluster membership appears to shift with task difficulty, suggesting dynamic regulation of cognitive or listening strategies rather than fixed individual profiles.

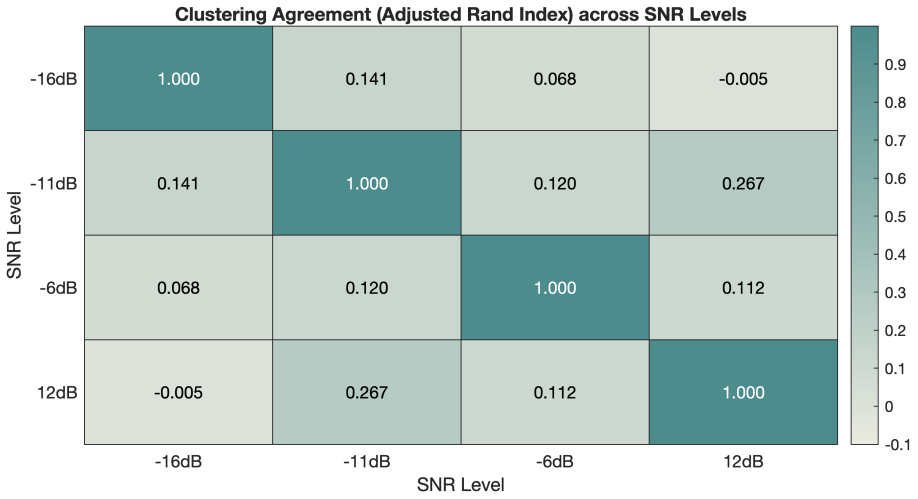


Figure. 13.18. Clustering membership agreement - adjusted rand Index (ARI) between SNR Pairings (pupil diameter)

Pairwise Adjusted Rand Index (ARI) values are presented for clustering outcomes across SNR levels. ARI assesses the similarity in cluster assignments between two conditions, correcting for chance (ARI = 1 indicates perfect agreement, 0 indicates random alignment, and negative values suggest disagreement).
Overall, ARI values remain low across comparisons, with the highest agreement observed between –11 dB and 12 dB (ARI = 0.267). These results reinforce that pupil-based cluster memberships are highly sensitive to contextual listening conditions and are not consistently preserved across noise levels.

Across all SNR levels, pupil diameter time courses revealed reliable clustering structure, with the elbow method consistently identifying two distinct clusters per condition. These clusters captured differences in response dynamics-typically contrasting stronger, more prolonged constriction with flatter or more transient patterns.

While the physiological differences between groups were consistent and visually distinct across SNR –16 dB, –11 dB, –6 dB, and 12 dB (Figures 13.12-13.16), no statistically significant differences were found in behavioural performance or subjective ratings between clustered groups.

This suggests that pupil-based clustering reveals internal cognitive or physiological response styles that do not neatly map onto accuracy, effort, or difficulty ratings. Furthermore, agreement analysis (Figures 13.17 and 13.18) showed limited consistency in cluster membership across SNR levels, with most participants switching clusters depending on the noise condition.

These results indicate that pupil clustering reflects dynamic, rather than stable individual traits. As such, pupil dynamics offer a valuable, nuanced window into

moment-to-moment listening strategies that are not easily inferred from behavioural data alone.

13.4 Galvanic Skin Response (GSR)

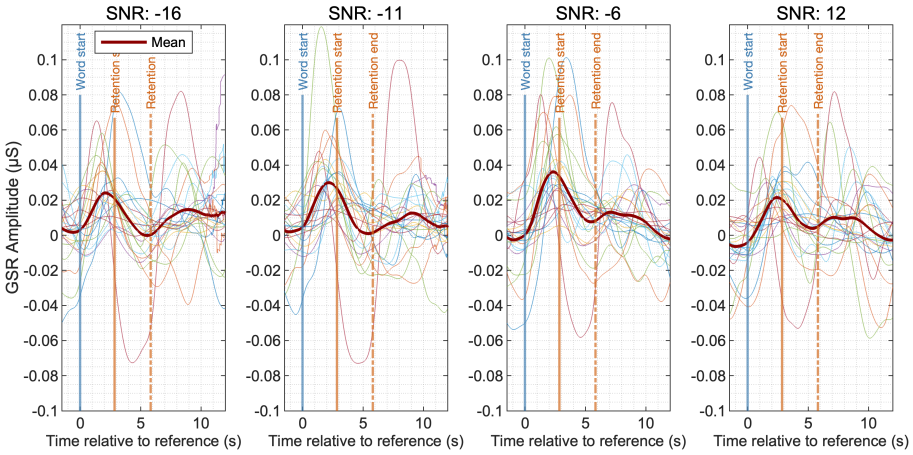


Figure. 13.19. GSR grand average - data averaged from experiment 1 and 2, all SNR levels

Average Galvanic Skin Response (GSR) across different Signal -to -Noise Ratio (SNR) levels, averaged over Experiments 1 and 2. Each panel shows a different SNR condition (-16, -11, -6, 12). The X -axis represents time in seconds relative to a reference point, and the Y -axis represents the GSR signal.
Faint lines show individual responses, and the thick red line indicates the mean response. Vertical lines mark the timing of experimental events: Word Start (light blue), Word End (green), Retention Start (orange solid), and Retention End (orange dot -dashed).

13.4.1 Data Overview

The original GSR data were first cleaned (see Section 12.6) and segmented into individual trials. Figure 13.19 displays the average signal across all trials and participants, combining data from Experiment 1 and Experiment 2.

GSR is known to exhibit a physiological latency in its response to stimuli (Benedek & Kaernbach, 2010) To accommodate this delay, we included the full trial duration in our analysis, thereby extending the analysis window and allowing sufficient time for the GSR signal to evolve in response to the stimulus.

We chose not to shift the data along the time axis for two main reasons. First, there is no consensus in the existing literature regarding the precise latency of the GSR response, making it difficult to determine an appropriate and justifiable time shift ((Laine et al., 2009; Sjouwerman & Lonsdorf, 2018)). Second, we believe it is more informative to preserve the natural time course of the GSR response, allowing it to reflect the

participants actual physiological dynamics without the imposition of an artificial alignment.

13.4.2 Task-Evoked Changes in GSR

Difference of Peak GSR across SNR Levels We analysed the difference between the word start and retention start (see Figure 13.19 for where the time events are during a trial). To analyse if there's a difference. We analysed that for each SNR level, Results shows that there's a clear difference between word start and retention start, as after the sentence start, GSR rise significantly (Wilcoxon signed -rank p = 0.0000).

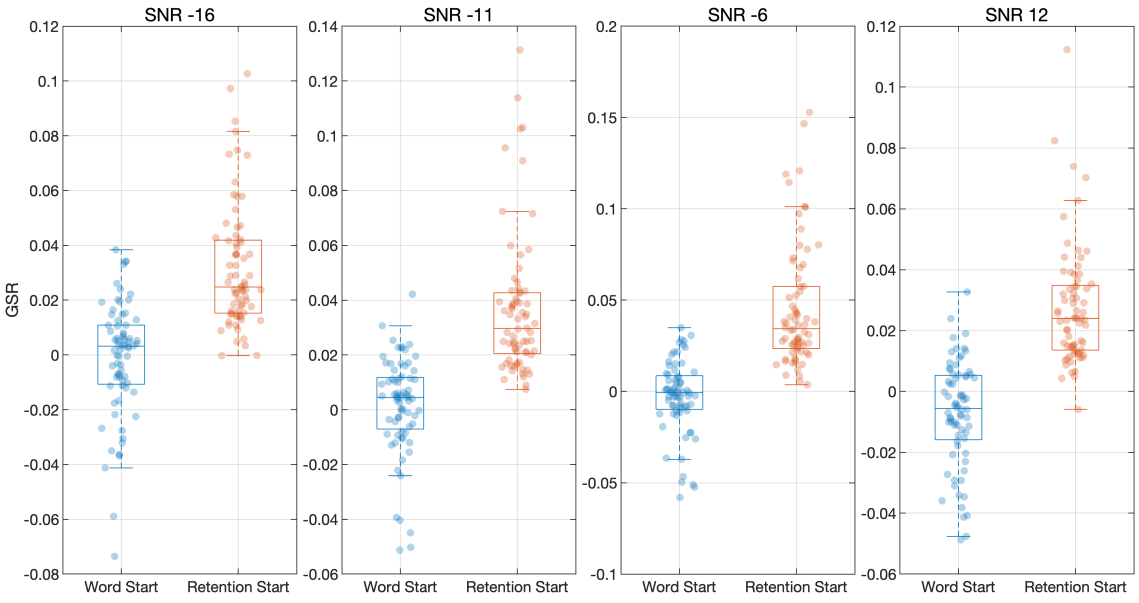


Figure. 13.20. Effect of Start of The Stimulus and Retention (GSR)

Each subplot shows the distribution of individual GSR values (scatter) and group -level spread (boxplots) at two key event timings: word onset (lowest GSR value within 500 ms) and retention start (maximum GSR within 1000 ms). These are plotted separately for each SNR level: -16 dB, -11 dB, -6 dB, and 12 dB.
Across all SNR levels, GSR tends to be higher at retention start compared to word onset, suggesting increased physiological arousal during early memory retention. This effect is most pronounced at moderate SNRs (e.g., -11 dB), where task difficulty may have triggered greater listener engagement. Scatter points represent within -subject variance, and boxplots show the distribution of responses across all participants and sessions. Colours correspond to event type: blue for word onset, and red for retention start. Jitter is applied to enhance visibility of overlapping data points.

13.4.3 GSR Changes at Different SNR Levels

Difference of GSR when Retention Start The **highest GSR value around Retention Start** showed a significant difference across SNR levels (p = .0078), indicating that listening difficulty influenced physiological responses even after the sentence had ended. This peak marks the point at which the stimulus concludes and the memory retention

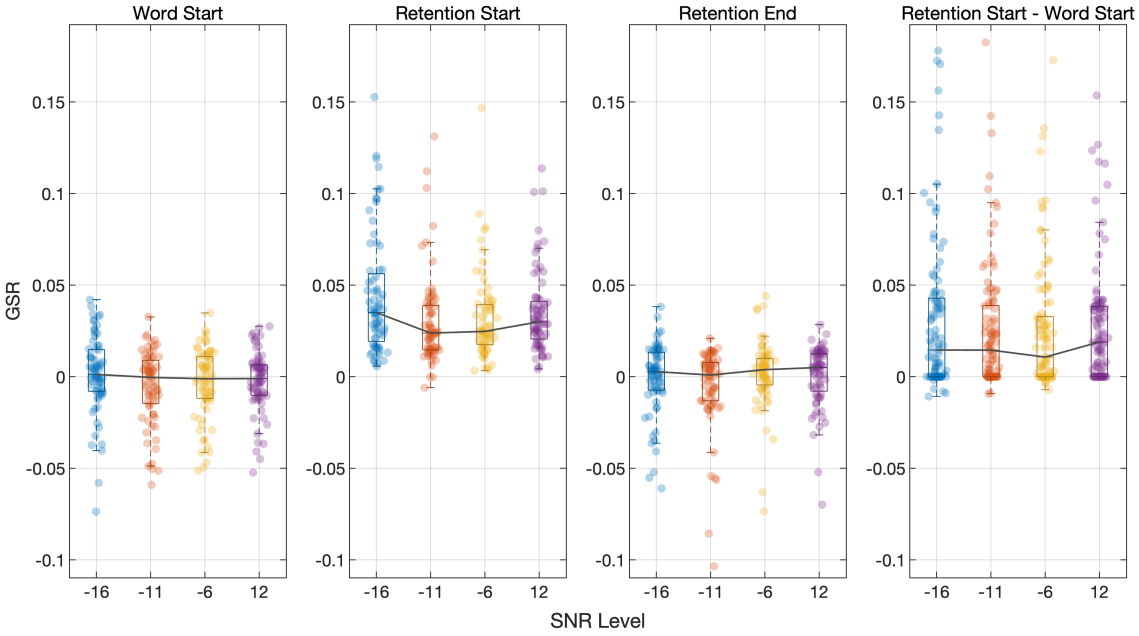


Figure. 13.21. GSR Response at Word Start, Retention Start, and Retention End

GSR responses are shown around key task events, derived from trial -level data during a speech -in -noise memory task. Word Start marks the lowest GSR value within 500 ms of the first word; Retention Start refers to the maximum GSR within 1000 ms of sentence offset; and Retention End captures the lowest GSR within 500 ms of the end of the retention phase.
A significant effect of SNR was observed only at Retention Start (p = .0078), suggesting increased physiological arousal during early memory retention. No significant differences were found at Word Start (p = .2783), Retention End (p = .2675), or for the difference between Retention Start and Word Start (p = .7229).

phase begins. As such, it likely reflects the level of cognitive effort or arousal maintained during the listening phase and carried forward into retention. A higher peak at more difficult SNRs may suggest greater listening effort being sustained through to the end of the sentence.

In contrast, there were **no significant differences** across SNR levels for GSR values around **Word Start** (p = .2783) or **Retention End** (p = .2675). These points represent the beginning of the auditory stimulus and the conclusion of the retention interval, respectively. The lack of significant variation suggests that GSR responses at these timepoints were less sensitive to changes in SNR, and may not capture the same degree of sustained listening effort as the Retention Start peak.

Together, these results highlight the Retention Start as a sensitive physiological marker of listening effort that accumulates during sentence processing, whereas earlier or later moments in the task may not exhibit strong modulation by signal quality.

13.4.4 Within Individual Consistency - Permutation Test Result of GSR

To assess whether participants exhibited consistent physiological responses across experimental sessions, a permutation test was performed using galvanic skin response (GSR) data. For the *within -subject correlation* (true pairing), each participants GSR time series from Experiment 1 and Experiment 2 was compared at the same signal -to -noise ratio (SNR) level. The data were aligned from the onset of the target word to the end of the retention phase, and Pearson correlation was computed for each participants paired traces.

Figure 13.22 shows individual examples of galvanic skin response (GSR) traces recorded during this period, for two participants (Subject 8 and Subject 10), at the same signal -to -noise ratio (SNR) level, across two experimental sessions. Each subplot depicts GSR activity aligned in time from the onset of the first word to the end of the retention phase, allowing comparison of physiological responses to a structured auditory task.

The coloured traces represent data from Experiment 1 and Experiment 2 for the same individual, overlaid to visualise the consistency of physiological patterns across sessions. These traces provide qualitative evidence of within -subject similarity. In contrast, the background traces in light grey represent all other participants' GSR responses at the same SNR level, included to provide a baseline of population -level variability. Overall, these plots suggest that certain individuals exhibit reliable and recognisable patterns of arousal or engagement in response to the same task, even when repeated a week later.

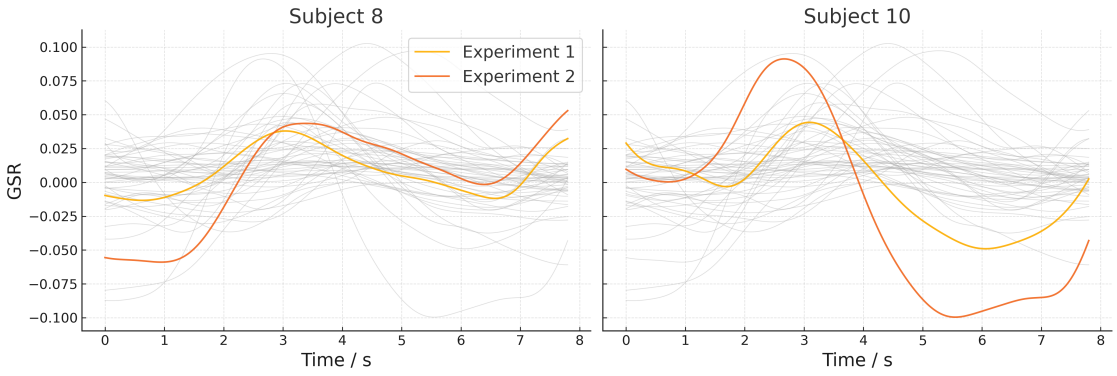


Figure. 13.22. Example of Participants GSR data During Trail Period

Examples of GSR traces from two participants at the same SNR level, across two experimental sessions. Each subplot shows one participants GSR response aligned from word onset to the end of the retention phase. Data from the two sessions (Experiment 1 and Experiment 2) are overlaid in colour for direct comparison.
Light grey traces in the background represent GSR data from all other participants at the same SNR level, providing visual context. Traces are trimmed to a maximum duration of 7.8 seconds to remove noisy or unstable tails. Time is shown in seconds, based on a sampling rate of 1000 Hz.

To formally test whether this apparent within -subject similarity exceeds what could be expected by chance, a permutation test was performed. The results are displayed in

Figure 13.23. In this analysis, Pearson correlation coefficients were calculated between each participants GSR time series in Experiment 1 and their own corresponding series in Experiment 2, at each SNR level.

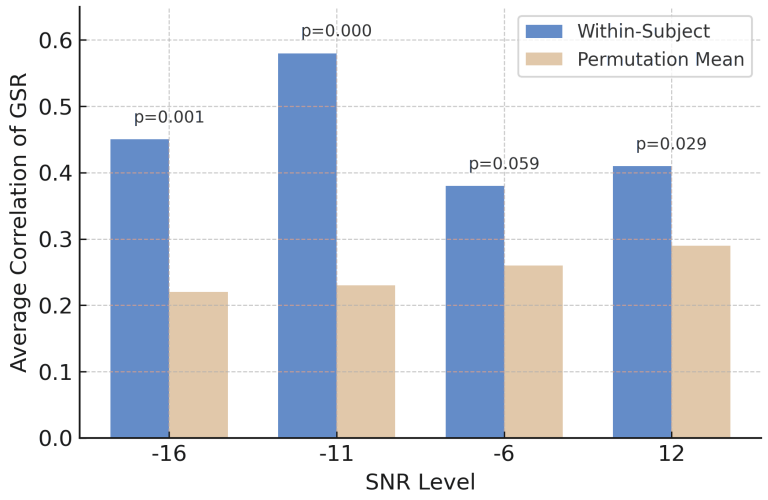


Figure. 13.23. Permutation test of within-subject correlations in GSR average trial response

Average within -subject and between -subject (permutation -based) correlation of GSR responses across two experimental sessions, shown separately by SNR level. Each bar represents the mean Pearson correlation coefficient between time -aligned GSR traces for the same or different participants. The data used for comparison was taken from the period starting at word onset and ending at the close of the retention phase.

Within -subject correlations (dusty rose) reflect the similarity of each participants GSR response across a one -week interval. Permutation -based values (beige) represent the average correlation when participant identities are randomly mismatched across sessions. P -values above bars indicate whether within -subject similarity significantly exceeds chance -level similarity. Error bars are omitted for visual clarity.

This provided a measure of within -subject similarity based on the full trace, from word onset to the end of the retention phase. These true within -subject values were then averaged across all participants, and the resulting means are shown as the green bars in the figure.

To determine whether this level of similarity was statistically meaningful, a permutation -based comparison was conducted. In each permutation, participant identities in the second session were randomly shuffled and paired with traces from different individuals in the first session. The correlation between these mismatched pairs was calculated in the same manner and repeated 1000 times, generating a null distribution that reflects the level of similarity expected under random pairing.

The average value of these shuffled correlations is shown as the beige bar for each SNR level. P -values were then calculated by counting how many permutations yielded a correlation equal to or greater than the true within -subject average. These p -values are shown above each pair of bars.

The results show that at SNR -16 dB and -11 dB, within -subject GSR similarity was significantly greater than the permutation baseline ($p = .001$ and $p < .001$, respectively), indicating that physiological responses at these moderate -to -difficult levels of listening demand were reliably reproduced across sessions.

Interestingly, at - 6 dB, the difference approached statistical significance ($p = .059$), suggesting a possible trend towards consistency that may have been affected by greater response variability. At the easiest condition (12 dB), a statistically significant result was still observed ($p = .029$), although the effect size appeared smaller. This could imply that while some level of consistent physiological response remains even when the task is less demanding, engagement and effort -related arousal may vary more subtly or be less uniformly sustained across individuals.

Together, Figure 13.22 and Figure 13.23 provide evidence that participants exhibit individually stable GSR patterns in response to the same auditory task when repeated over time. This suggests that GSR traces are not only sensitive to task difficulty but may also capture trait -like physiological response profiles that are robust enough to be detected across sessions, even after a delay of one week.

13.4.5 Clustering Result of GSR (presenting per SNR level

To examine individual patterns in physiological responses, a clustering analysis was performed on the GSR data. The goal was to identify whether participants exhibited distinct, consistent response types during the task, particularly in response to varying signal -to -noise ratios (SNRs).

Data was extracted to include aligning each trial according to defined event markers (e.g., word onset and retention start), then computing the average GSR signal for each participant across all relevant trials. These participant -wise average traces were then cleaned to remove extreme artefacts and truncated at a consistent endpoint.

Clustering analysis was performed to examine whether individuals could be grouped according to the similarity of their physiological responses, specifically GSR (galvanic skin response) signals. The algorithm employed was k -means clustering, a widely used unsupervised learning method. Unlike supervised approaches, which rely on pre -labelled data, unsupervised clustering attempts to uncover inherent structure in the dataset without prior knowledge of group membership.

In this analysis, clustering was applied to GSR time series recorded during a defined trial period. Each participants data was represented as a single time series, and clustering was based not on absolute differences in amplitude, but on the similarity in shape and temporal pattern. To achieve this, the clustering used a correlation -based distance metric

- specifically, one minus the Pearson correlation coefficient - rather than the default Euclidean distance.

This choice of correlation captures whether two time series follow a similar temporal pattern, regardless of their magnitude. For instance, two GSR traces with similar rises and falls over time but different amplitudes would be grouped together under a correlation -based metric, whereas they might not under Euclidean distance.

To identify the optimal number of clusters (K), the within -cluster sum of squares (WCSS) was calculated for K ranging from 2 to 5 (see Figure 13.24). WCSS quantifies how similar the members of each cluster are to their cluster centroid; lower values indicate tighter clustering. The "elbow method" was used to determine the best number of clusters, with the point at which additional clusters no longer substantially reduce WCSS indicating the most appropriate K.

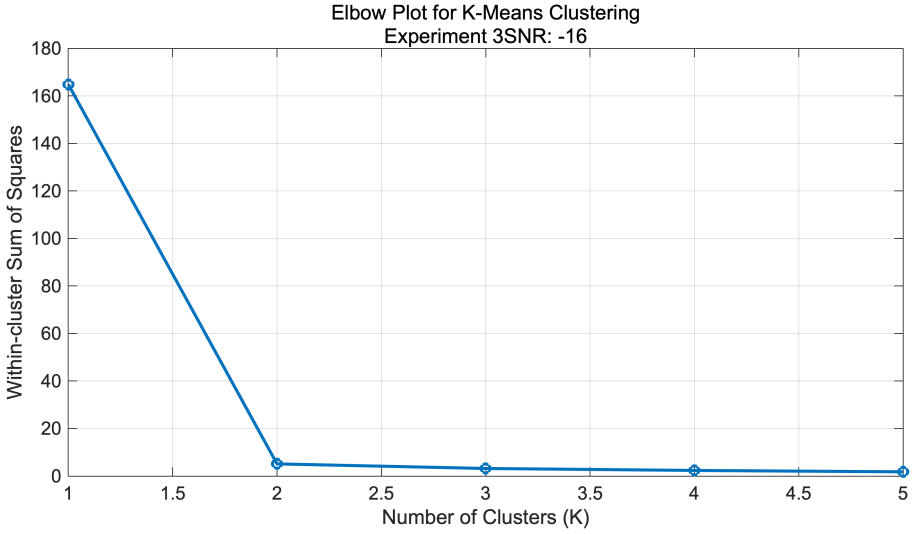


Figure. 13.24. K-Means elbow plot of cluster sum of squares (WCSS) in within subject correlations for GSR clustering (-16 dB)

The within -cluster sum of squares (WCSS) were used to indicated the best cluster group number. The elbow method identifies the optimal K as the point where adding more clusters yields diminishing returns in reducing WCSS. Here, the sharp drop from $K = 1$ to $K = 2$ followed by a plateau suggests that $K = 2$ offers the most meaningful separation of physiological response patterns.

Clustering result at SNR : -16 dB (GSR) To examine whether the GSR -derived clusters reflected meaningful differences in behavioural or subjective outcomes, statistical comparisons were conducted on task accuracy, subjective difficulty, and subjective effort between the two groups. As the data did not follow a normal distribution, the non -parametric Wilcoxon rank -sum test was used for group comparisons.

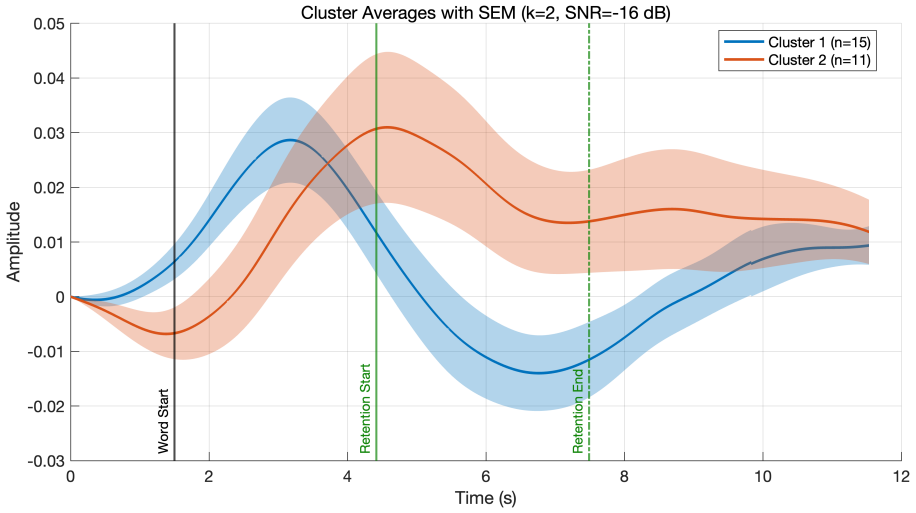


Figure. 13.25. Clustering result - average trial response of GSR with standard error shading (baseline corrected, k = 2, SNR = -16 dB)

Two clusters of participants' GSR responses are shown, averaged across individuals within each cluster. To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The solid lines represent the mean GSR trace for each group, and shaded regions indicate ± 1 standard error of the mean (SEM). Clustering was based on the temporal similarity of GSR patterns using correlation distance.
Vertical lines mark key task events: word onset (black), retention start (solid green), and retention end (dashed green). Cluster 1 (blue, $n = 15$) shows an early peak around retention onset followed by a decline, whereas Cluster 2 (orange, $n = 11$) shows a slower rise and a broader, sustained response across the retention phase, indicating differing patterns of physiological engagement.

The results revealed no statistically significant differences across any of the three measures ($p > 0.05$), suggesting that the physiological clustering did not correspond directly to differences in behavioural performance or self -reported experience.

The clustering analysis revealed distinct patterns in participants' physiological responses, indicating that individuals reacted differently during the listening task. Despite this divergence in GSR activity, statistical comparisons of task accuracy, perceived difficulty, and reported effort between the clusters showed no significant differences.

This apparent disconnect suggests that while participants' physiological arousal varied, it may not have translated into observable behavioural or subjective differences. Possible explanations include inter -individual variability in autonomic regulation, differences in physiological sensitivity, or cognitive strategies that are not fully captured by self -report or performance metrics.

Clustering result at SNR : -11 dB (GSR) To determine the optimal number of GSR response patterns, the elbow method was applied at SNR -11 dB. The WCSS dropped steeply from $K = 1$ to $K = 2$, indicating that two clusters offered a reasonable trade-off between complexity and explanatory power.

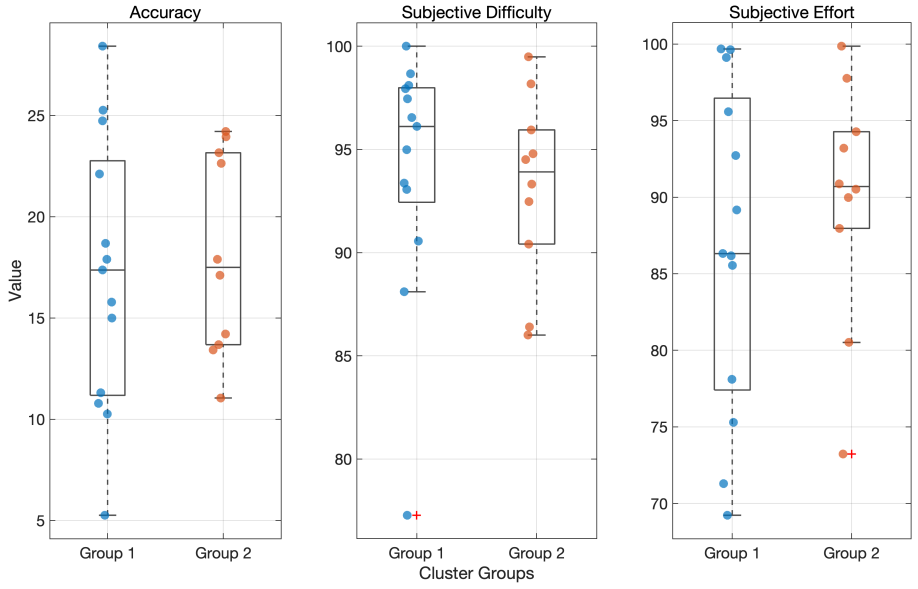


Figure. 13.26. Differences in Performance and Subjective Ratings Between GSR Clusters (SNR -16dB)

Boxplots show differences in accuracy, subjective difficulty, and subjective effort between the two groups identified through correlation -based clustering of GSR responses (as shown in Figure 13.25). Each dot represents one participant, and groupings are derived from the GSR trace similarity during the trial period at -16 dB SNR.
Group 1 and Group 2 show partially overlapping distributions across metrics. Notably, Group 1 exhibited slightly lower accuracy but reported marginally higher subjective effort. These results suggest that differences in GSR patterns may be associated with subtle variations in task performance and perceived effort.

Despite the observed physiological differences, the two identified groups did not differ significantly in task performance. Accuracy, subjective effort, and subjective difficulty scores were comparable between clusters, as confirmed by non-parametric Wilcoxon tests.

Nevertheless, the average GSR traces (Figure 13.27) revealed temporally distinct patterns between clusters. One group showed a rapid GSR rise during retention, while the other displayed a slower and more sustained increase, indicating diverging physiological engagement profiles.

These findings suggest that, at an SNR of -11 dB, participants exhibit meaningful physiological differences in their GSR responses, as captured by unsupervised clustering.

However, these physiological patterns do not appear to directly translate into measurable differences in behavioural performance. This discrepancy may indicate that individuals adopt different regulatory or coping strategies during the task, which manifest physiologically but yield similar performance outcomes.

Clustering result at SNR : -6 dB (GSR) At the -6 dB SNR level, the elbow plot indicated that $K = 2$ was the optimal number of clusters, supporting the presence of two

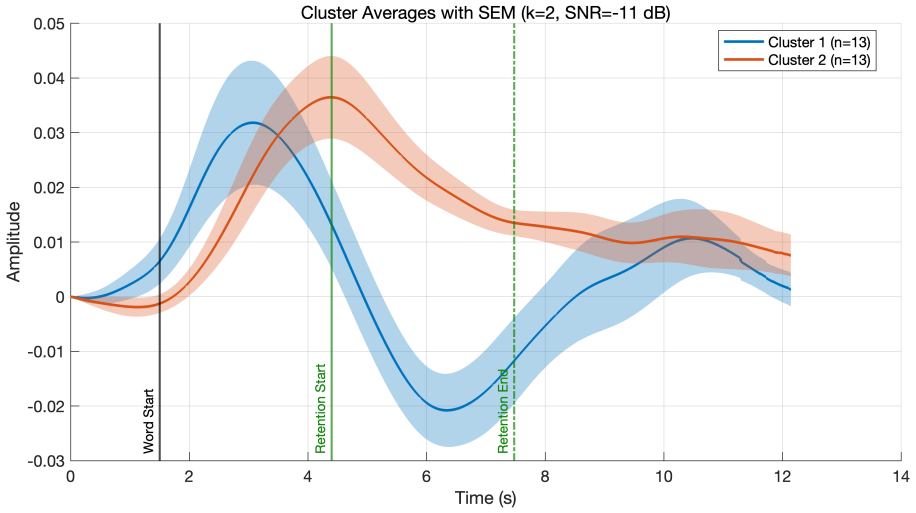


Figure. 13.27. Clustering result - average trial response of GSR with standard error shading (baseline corrected, k = 2, SNR = -11 dB)

Two clusters of participants' GSR responses are shown, averaged across individuals within each cluster. The solid lines represent the mean GSR trace for each group, and shaded regions indicate ± 1 standard error of the mean (SEM). To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Clustering was based on the temporal similarity of GSR patterns using correlation distance.
Vertical lines mark key task events: word onset (black), retention start (solid green), and retention end (dashed green). Cluster 1 (blue, $n = 12$) shows an early peak around retention onset followed by a decline, whereas Cluster 2 (orange, $n = 14$) shows a slower rise and a broader, sustained response across the retention phase, indicating differing patterns of physiological engagement.

distinct physiological response patterns.

These clusters were further visualised in the SEM plot, which revealed clear differences in GSR dynamics across time-one group showed a sharp increase followed by decline, while the other exhibited a slower, sustained rise.

However, when comparing behavioural accuracy, subjective difficulty, and perceived effort between these groups, no significant differences were found. This pattern of physiological divergence without behavioural separation echoes findings at -11 and -16 dB, suggesting a stable dissociation between autonomic response profiles and overt task performance across conditions.

Clustering result at SNR : 12 dB (GSR) Clustering results at 12 dB SNR showed that using two clusters ($K = 2$) provided a good fit to the data. This was based on the elbow method, where the drop in within-cluster variance became much smaller after $K = 2$, suggesting no strong benefit from adding more clusters.

The corresponding cluster-averaged GSR signals are shown in Figure 13.29. Cluster 1 ($n = 8$) exhibited an earlier, sharper peak shortly after word onset, followed by a more rapid decline. In contrast, Cluster 2 ($n = 18$) displayed a slower rise and more prolonged

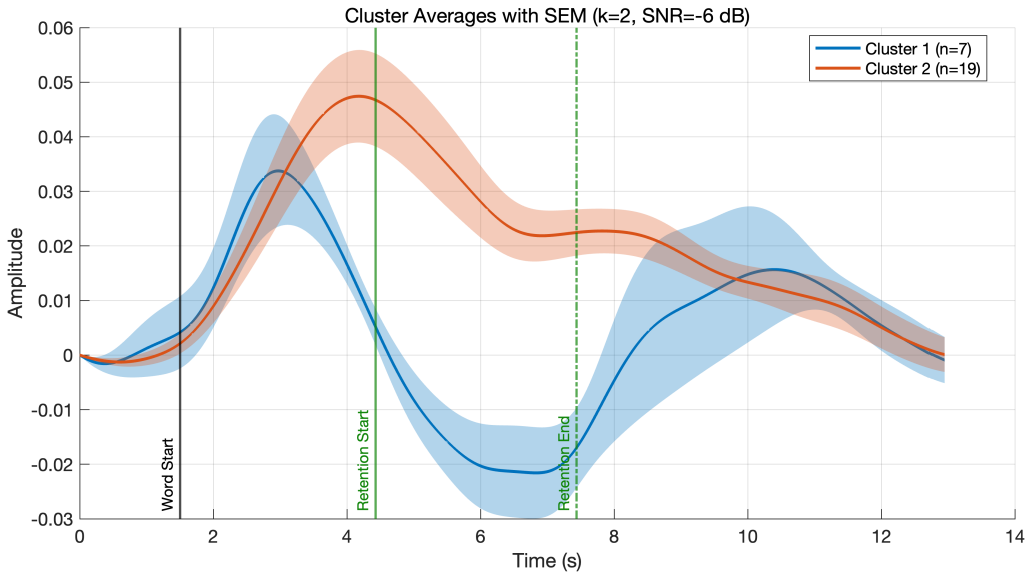


Figure. 13.28. Clustering result - average trial response of pupil diameter with standard error shading (baseline corrected, k = 2, SNR = -6 dB)

Two clusters of participants' GSR responses are shown, averaged across individuals within each cluster. The solid lines represent the mean GSR trace for each group, and shaded regions indicate ± 1 standard error of the mean (SEM). To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Clustering was based on the temporal similarity of GSR patterns using correlation distance.
Vertical lines mark key task events: word onset (black), retention start (solid green), and retention end (dashed green). Cluster 1 (blue, $n = 12$) shows an early peak around retention onset followed by a decline, whereas Cluster 2 (orange, $n = 14$) shows a slower rise and a broader, sustained response across the retention phase, indicating differing patterns of physiological engagement.

elevation across the retention period, indicating qualitatively distinct physiological profiles.

Despite this divergence in GSR dynamics, no significant differences were observed in behavioural accuracy, subjective difficulty, or self-reported effort between the two clusters. These metrics remained statistically comparable across groups, based on Wilcoxon rank-sum tests ($p > .05$).

This dissociation between physiological clustering and task outcomes aligns with previous findings at lower SNR levels (-16, -11, and -6 dB). Together, these results suggest that autonomic response patterns may reflect differences in internal processing or engagement, without a clear behavioural correlate.

Such findings imply that GSR-based clustering may capture subtle cognitive or emotional dynamics not easily accessed through performance scores or subjective ratings. This highlights the potential of psychophysiological measures to reveal latent individual differences in task-related responses beyond overt behaviour.

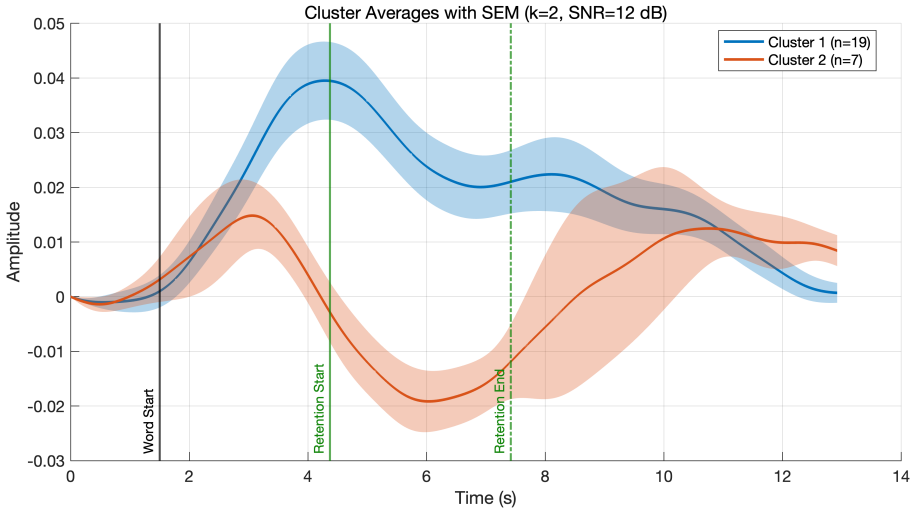


Figure. 13.29. Clustering result - average trial response of GSR with standard error shading (baseline corrected, k = 2, SNR = 12 dB)

Two clusters of participants' GSR responses are shown, averaged across individuals within each cluster. To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The solid lines represent the mean GSR trace for each group, and shaded regions indicate ± 1 standard error of the mean (SEM). Clustering was based on the temporal similarity of GSR patterns using correlation distance.
Vertical lines mark key task events: word onset (black), retention start (solid green), and retention end (dashed green). Cluster 1 (blue, n = 12) shows an early peak around retention onset followed by a decline, whereas Cluster 2 (orange, n = 14) shows a slower rise and a broader, sustained response across the retention phase, indicating differing patterns of physiological engagement.

13.4.6 Clustering Result Agreement Across Different SNR Levels (GSR)

The clustering assignments across SNR levels are illustrated in Figure 13.30, which maps each participant's group membership at four different auditory conditions. Each row represents a distinct SNR level, ordered from the most difficult -16 dB to the easiest (12 dB), and each column corresponds to an individual subject.

The coloured bands indicate whether a participant was assigned to Cluster 1 or Cluster 2 under each condition. While several participants remained in the same cluster across all conditions-suggesting stable physiological response profiles-many others shifted clusters depending on the listening difficulty.

This variability implies that some individuals exhibit flexible physiological patterns that adapt with task demands, whereas others respond more consistently regardless of difficulty level. Such observations provide support for subject -specific response profiles that are either stable or modulated by task challenge.

To quantify the similarity of clustering solutions across noise conditions, the Adjusted Rand Index (ARI) was computed between every pair of SNR levels (see Figure 13.31).

Results show that clustering solutions tend to be more consistent between neighbouring SNR levels, while comparisons across more extreme noise contrasts exhibit weaker

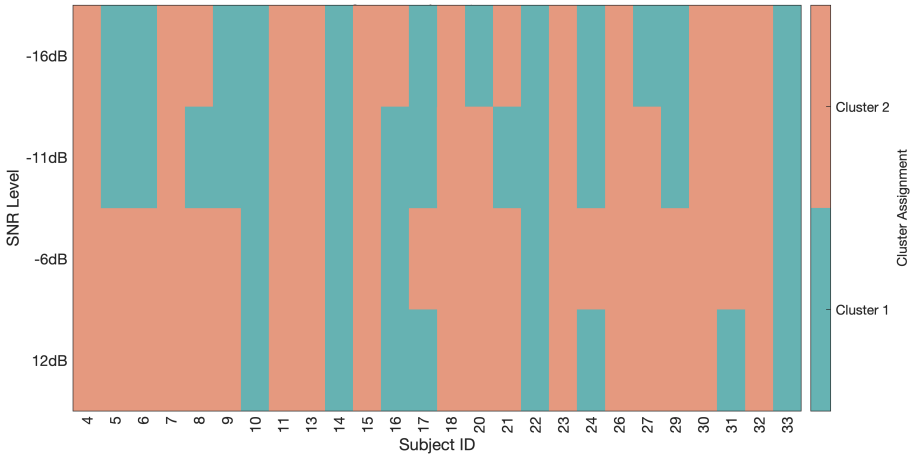


Figure. 13.30. Cluster membership across SNR levels - GSR

Cluster assignments are shown by participant across different SNR levels. Each column represents an individual, and each row corresponds to one SNR condition (from -16 dB to 12 dB). Colours indicate cluster identity (Cluster 1 or Cluster 2).
Some participants maintain consistent group membership across all conditions, while others switch between clusters depending on the noise level. This pattern may reflect a distinction between stable and more contextually adaptive physiological response styles.

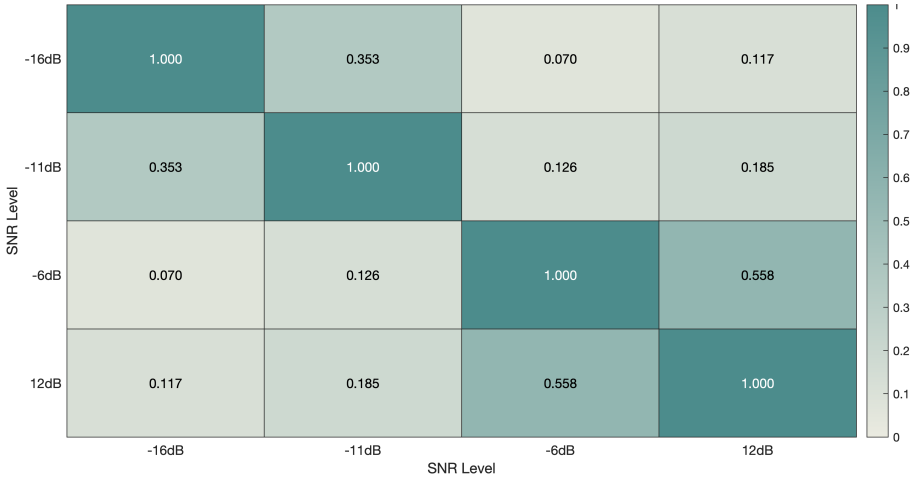


Figure. 13.31. Clustering membership agreement - adjusted rand Index (ARI) between SNR Pairings (GSR)

Clustering agreement across SNR conditions was evaluated using the Adjusted Rand Index (ARI), which measures similarity between cluster assignments while correcting for chance. Higher values indicate greater consistency between clustering solutions at different SNRs.
Moderate agreement was observed between adjacent levels (e.g., -16 dB and -11 dB, ARI = 0.353), whereas low agreement (e.g., ARI = 0.070 between -16dB and -6dB) suggests that physiological response patterns shift substantially with changes in task difficulty.

agreement. This pattern suggests that participants' physiological response profiles adapt dynamically to task difficulty, leading to different groupings across conditions.

13.5 Respiration Rate

13.5.1 Data Overview

Respiration data was first cleaned and extracted respiration rate (RR) (see Page 127 for cleaning details). RR was intepolated at frequency of 1000, to match other physiological data. To examine whether respiration patterns were modulated by task demands or auditory clarity, respiration traces were averaged across participants and repeated experiments.

Time-locked data were extracted from the onset of the first spoken word and analysed across four levels of signal-to-noise ratio (SNR). A consistent dip was observed following retention onset, becoming more pronounced at higher intelligibility (e.g., SNR = 12), suggesting possible respiratory adaptation linked to cognitive processing during retention.

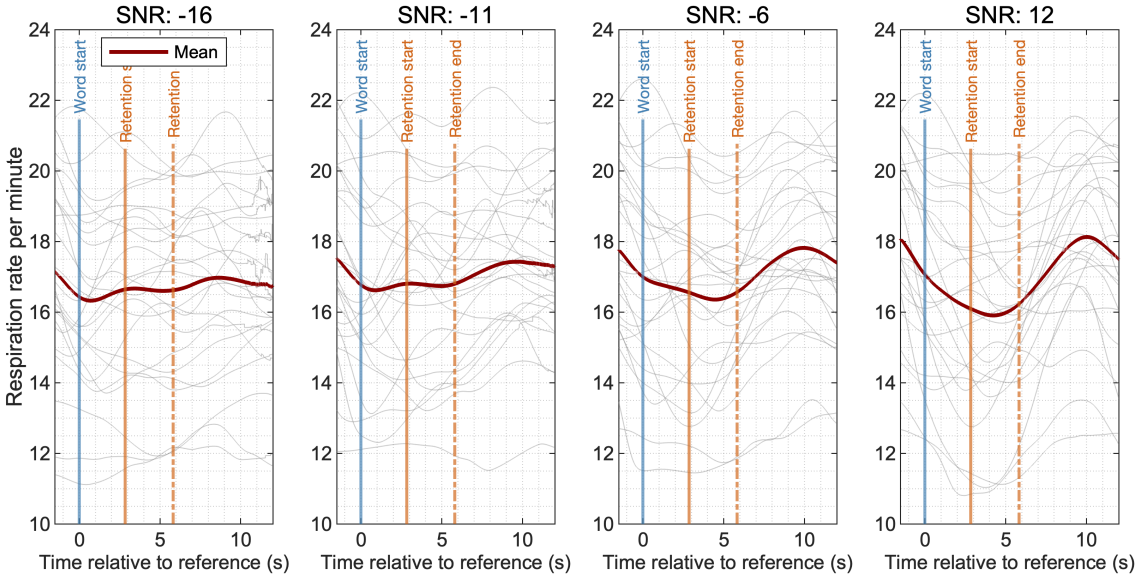


Figure. 13.32. Respiration rate grand average (original) - data averaged from experiment 1 and 2, all SNR levels

Each panel shows respiration signals averaged across two experiments for four SNR conditions (-16, -11, -6, and 12 dB). Grey lines represent individual participant averages, and the red trace indicates the grand average.
Time is aligned to the onset of the first word (blue line). The solid and dashed orange lines mark the start and end of the memory retention phase, respectively.

To assess task-related changes in respiration, signals were mean-subtracted and time-aligned to the onset of the first spoken word. Averaged across two experiments, the

respiration traces show a consistent dip following the onset of the retention period, particularly at higher SNRs.

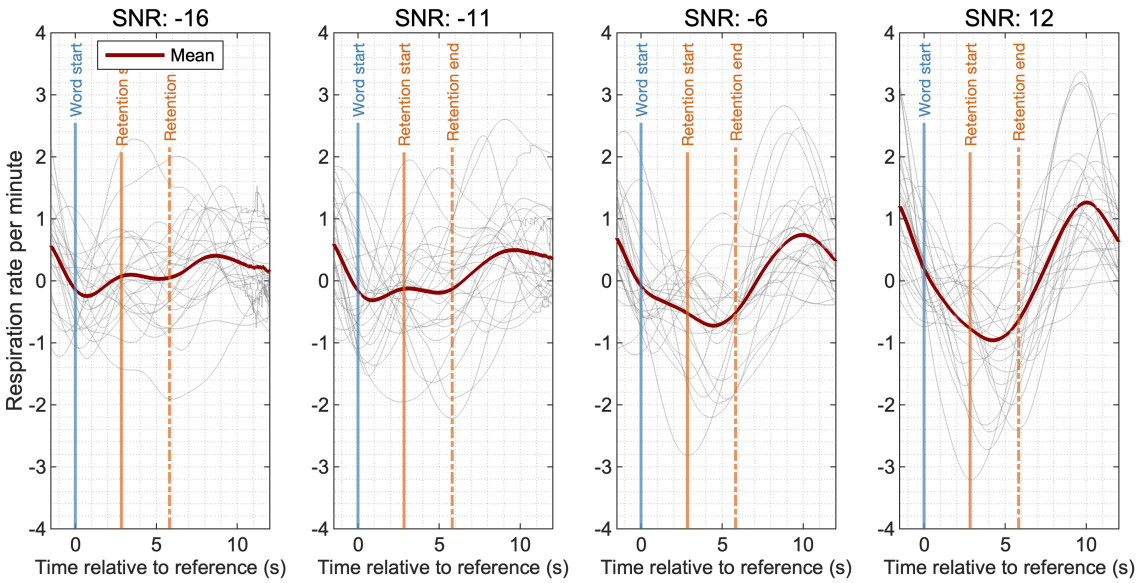


Figure. 13.33. Respiration rate grand average (mean-subtracted) - data averaged from experiment 1 and 2, all SNR levels

Grand average respiration signals (red line) and individual participant traces (grey lines) are plotted across four SNR levels (-16, -11, -6, 12 dB), based on data averaged from two repeated experiments.
Signals are time-aligned to word onset (blue line), with the start and end of the memory retention phase marked by solid and dashed orange lines, respectively. The y-axis reflects respiration amplitude following mean subtraction within each trace.

At SNR = 12, an obvious recovery and peak is observed after the retention period, whereas this pattern is less evident at lower SNRs. These dynamics suggest that respiration may be modulated by both task structure and stimulus intelligibility, with clearer speech potentially evoking more distinct respiratory adjustments during memory retention.

13.5.2 Task-Evoked Changes in Respiration Rate

Significant differences in respiration rate were observed between Word Start and Retention Start across all SNR levels (-16, -11, -6, and 12 dB). Wilcoxon signed-rank tests were applied when data is not normal (SNRs -16 and -11), while paired t-tests were used where data were normally distributed (SNRs -6 and 12).

All comparisons yielded highly significant differences ($p = .000$), indicating that respiration rate systematically increased over the retention period. These findings suggest that task-evoked respiration dynamics are sensitive to both the acoustic clarity and the cognitive phase of the task.

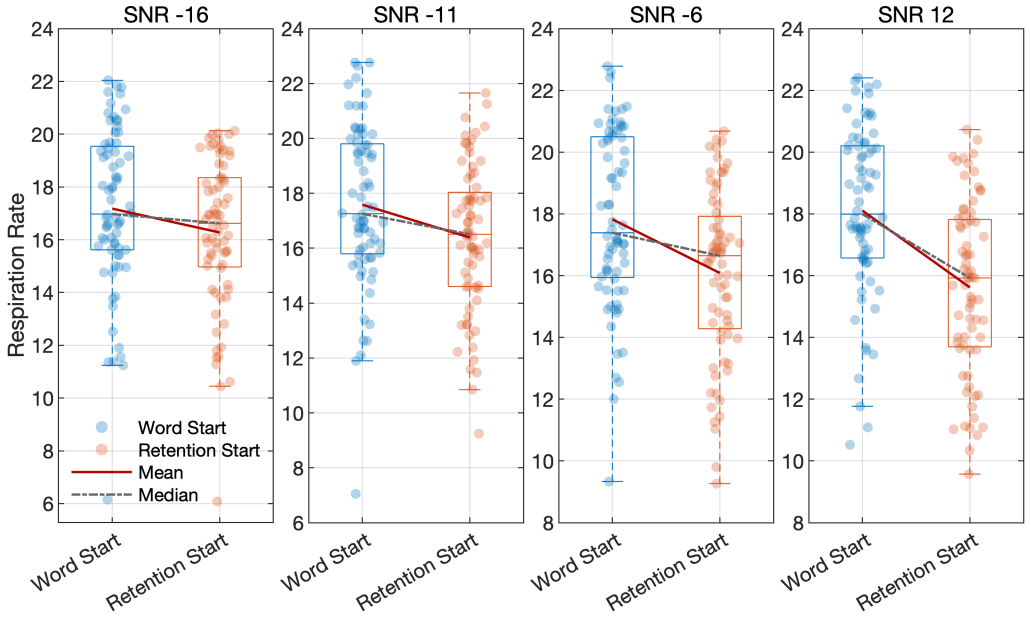


Figure. 13.34. Respiration rate comparison between word and retention onset across SNR levels

These paired boxplots compare respiration rates measured at word onset and retention onset across four levels of signal-to-noise ratio (SNR). Each colour-coded pair reflects average participant responses at a given SNR, combining data from repeated experiments. A consistent downward shift is visible from word to retention onset within individuals.

Overlaid lines indicate the mean (solid red) and median (dashed grey) values per condition. These patterns suggest a subtle decrease in respiration rate as participants transition into the memory retention phase, potentially reflecting a shift in cognitive or attentional demand.

13.5.3 Respiration Rate Changes at Different SNR Levels

To explore how respiratory dynamics are modulated by listening difficulty and task engagement, we analysed respiration rate across different time windows and across signal-to-noise ratio (SNR) conditions.

Statistical testing showed no significant overall differences in RR at word onset, retention start, or retention end (Friedman test all $p > .16$). However, post hoc comparisons revealed that the most degraded condition (SNR -16) differed significantly from both -6 and 12 dB at several time points, including word onset and retention offset (Figure 13.36).

To further examine how RR changed over time, we computed within-subject differences between key task periods (Figure 13.37). The difference between retention onset and word onset showed a small but consistent negative shift across SNR levels, though not statistically robust overall.

In contrast, the difference between retention offset and retention onset was more strongly modulated by SNR, with a significant main effect detected (Friedman test $p < 0.05$). Greater respiration rate increases were observed at higher SNRs, suggesting clearer speech may induce stronger physiological changes during the memory retention phase.

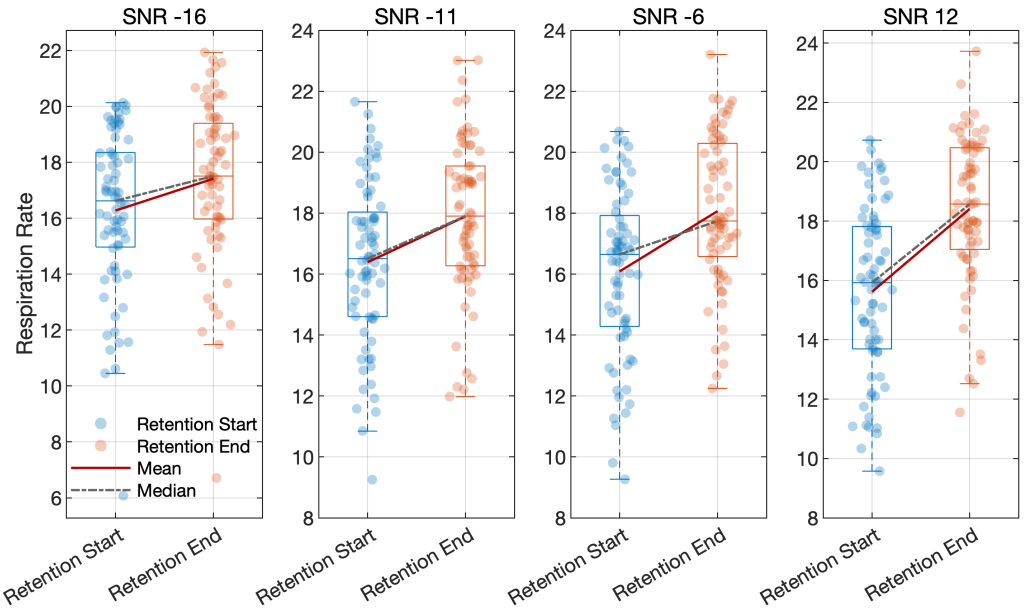


Figure. 13.35. Respiration rate comparison between retention onset and offset across SNR levels

Paired boxplots illustrate the change in respiration rate from the beginning to the end of the retention phase at each signal-to-noise ratio (SNR). Each point represents a participant's averaged rate per phase across experiments, allowing comparison within-subject.

Across all four SNR levels, respiration rate tended to increase from retention onset to offset, suggesting a progressive shift in physiological engagement during the memory retention interval. Trends were most pronounced at higher SNRs, where task difficulty was reduced.

These findings indicate that while respiration rate at isolated time points is highly variable, relative changes across task segments reveal systematic modulation by listening condition. This supports the idea that clearer auditory input elicits more distinct physiological dynamics over the course of cognitive processing.

13.5.4 Within Individual Consistency - Permutation Test Result of Respiration Rate

To examine whether participants' respiration patterns were consistent across experiments, we conducted a within-subject similarity analysis. Specifically, we tested whether each participant's respiration signal in one experiment was more similar to their own signal in the other experiment than to those of other participants. Similarity was quantified using Pearson's correlation coefficient.

Figure 13.38 illustrates two representative cases at SNR –11 dB. In both examples, the participants respiration traces from Experiment 1 and Experiment 2 closely resembled each other, with notable divergence from other participants' responses (shown in grey). These cases visually highlight within-subject consistency amid group-level variability.

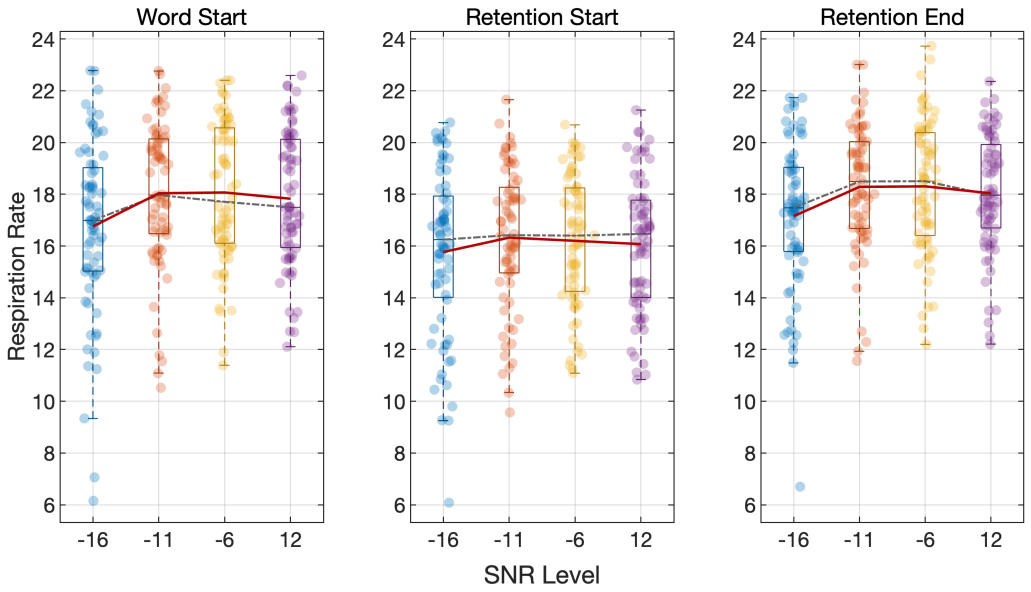


Figure. 13.36. Respiration rate across conditions at different signal-to-noise ratios (SNRs)

Boxplots illustrate respiration rate (in breaths per minute) extracted at three task-relevant moments: word onset, retention onset, and retention offset. Each point represents an individual participant's average RR across two experiments. SNR levels range from the most degraded (-16 dB) to the most intelligible (12 dB).

Overlaid lines indicate the mean (solid red) and median (dashed grey) values per SNR level. Respiration at retention offset shows a visible increase at clearer SNRs, although trends are modest across conditions.

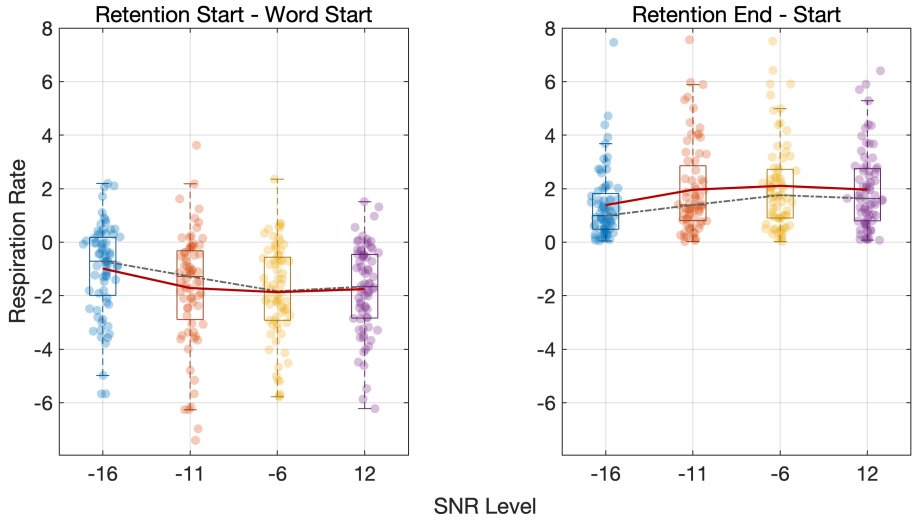


Figure. 13.37. Respiration rate differences across SNR levels.

Each panel shows the respiration rate change between two task stages: retention onset minus word onset (left), and retention offset minus retention onset (right). These metrics reflect dynamic changes in breathing patterns during memory retention.

Values are averaged across two experiments per participant. Points represent individuals, while boxplots and overlaid lines display the distribution, mean (solid red), and median (dashed grey) per SNR level. The largest relative increase from retention start to end appears at intermediate to higher SNRs.

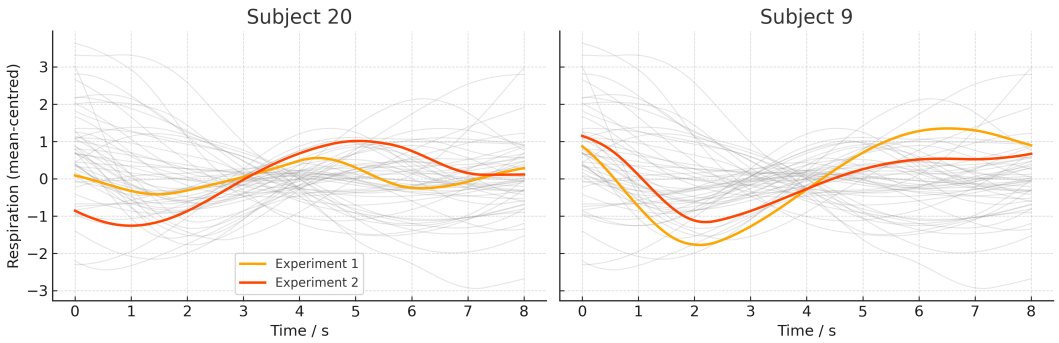


Figure. 13.38. Example of within-subject respiration similarity across experiments at SNR -11 dB

Two participants (Subjects 2 and 9) show consistent RR patterns across Experiment 1 (orange) and Experiment 2 (red), overlaid on background traces (grey) representing all other participants at the same SNR level.

Data have been mean-centred within each condition to improve vertical alignment. These examples illustrate the higher within-subject similarity compared to the variability observed across the group.

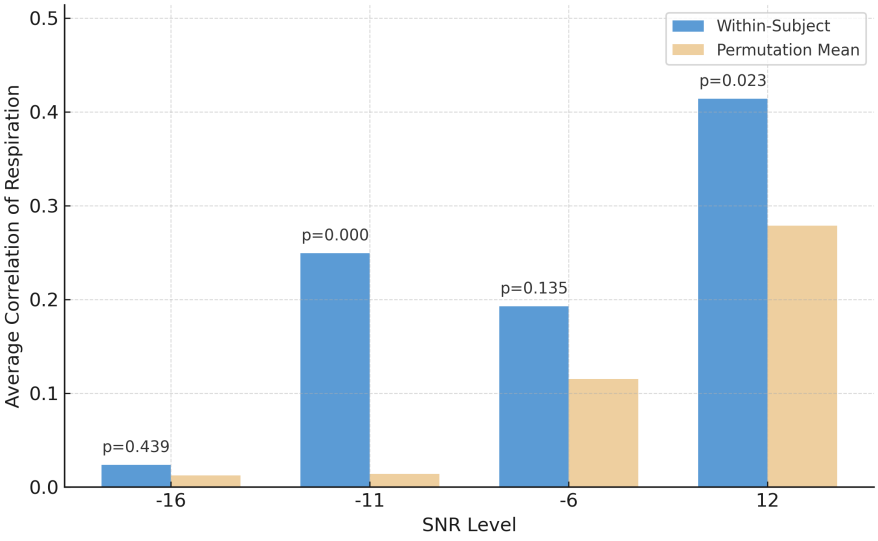


Figure. 13.39. Permutation test of within-subject correlations in respiration rate average trial response

Bars represent average correlation coefficients between respiration signals from the same individuals across two experiments (within-subject; blue) compared to the mean correlation from random pairings (permutation mean; beige).

Each SNR level is tested separately. Annotated p-values indicate whether within-subject similarity was significantly greater than expected by chance, based on 1000 permutations.

To statistically assess this effect across the sample, a permutation test was conducted for each SNR level. The average within-subject correlation was compared to a null distribution generated by randomly pairing subjects across experiments. As shown in Figure 13.39, within-subject similarity was significantly higher than chance at SNR –11 dB and 12 dB, but not at SNR –16 dB or –6 dB. These findings suggest that individual consistency in respiration patterns emerges more reliably at moderate-to-clear SNRs.

A comparable analysis was previously applied to GSR data (Figure 13.23, Page 157). GSR responses showed stronger within-subject consistency across all SNRs, particularly at SNR –11 dB and –16 dB. In contrast, respiration exhibited weaker alignment under degraded SNRs, potentially reflecting its slower timescale or greater sensitivity to general arousal and attentional shifts.

13.5.5 Clustering Result of Respiration Rate

To explore whether participants exhibited distinct physiological patterns in response to challenging listening conditions, a k-means clustering analysis was applied to the respiration data from averaging Experiment 1 and 2 at Different SNR levels.

Clustering result at SNR : -16 dB (RR) At SNR -16 dB level, the elbow method indicated that a two-cluster solution was optimal.These clusters reflected two broad groups of individuals with differing respiration trajectories during the listening and retention periods.

Despite the physiological separation, no significant group differences were found in behavioural performance or subjective ratings (same as GSR results). indicates that individual physiological reactivity may not correspond directly with explicit behavioural or self-reported measures of listening effort.

Clustering result at SNR : -11 dB (RR) Participants were clustered into two groups based on their respiration waveforms at SNR = -11 dB, using k-means clustering. The number of clusters ($k = 2$) was determined via an elbow analysis of the within-cluster sum of squares. This approach was applied to data averaged across Experiments 1 and 2, providing greater statistical stability across repeated testing sessions.

The resulting clusters reflected distinct respiration response profiles during the listening and retention period. However, when comparing behavioural metrics between the two groups-accuracy, perceived difficulty, and perceived effort-no significant differences were observed. This suggests that while participants differed in physiological expression, this did not correspond to observable differences in behavioural outcomes.

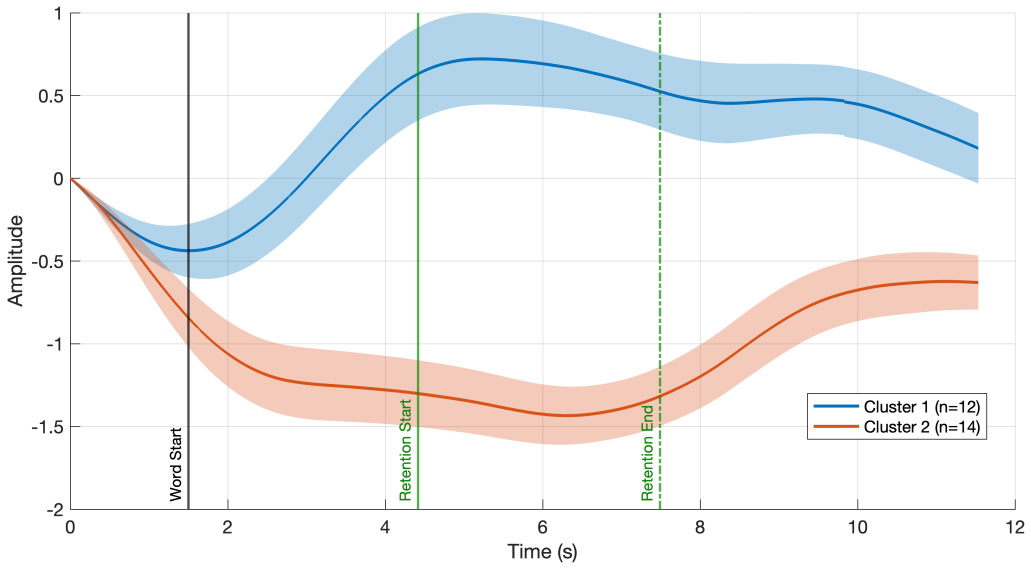


Figure. 13.40. Clustering result - average trial response of respiration rate with standard error shading (baseline corrected, k = 2, SNR = -16 dB)

Average respiration traces are shown for two participant clusters identified at SNR −16 dB in RR data averaged from experiment 1 and 2, with shading indicating the standard error of the mean (SEM). To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Key task events-word onset, retention start, and retention end-are marked with vertical lines.

Participants in Cluster 1 (blue) exhibited increased respiration amplitude during the retention period, while Cluster 2 (orange) showed a sustained decrease. These diverging patterns may reflect distinct engagement or regulation strategies under highly degraded acoustic conditions.

These findings mirror the clustering analysis conducted on GSR data, where similarly, no reliable behavioural distinctions emerged between physiological groups. This convergence reinforces the idea that physiological variation does not always map directly onto performance or subjective experience.

Clustering result at SNR : -6 dB (RR) At SNR −6dB, clustering analysis again suggested the presence of two distinct participant subgroups based on their respiration patterns Participants in Cluster 1 exhibited stronger respiratory suppression during the early retention phase, while those in Cluster 2 showed relatively shallower modulation (Figure 13.42).

Importantly, no significant differences were observed between these two groups across behavioural measures such as accuracy, perceived difficulty, or effort. This is consistent with the clustering findings at other SNR levels.

Clustering result at SNR : 12 dB (Respiration Rate) K-means clustering applied to respiration traces at SNR 12 dB also identified a clear two-cluster structure. This was

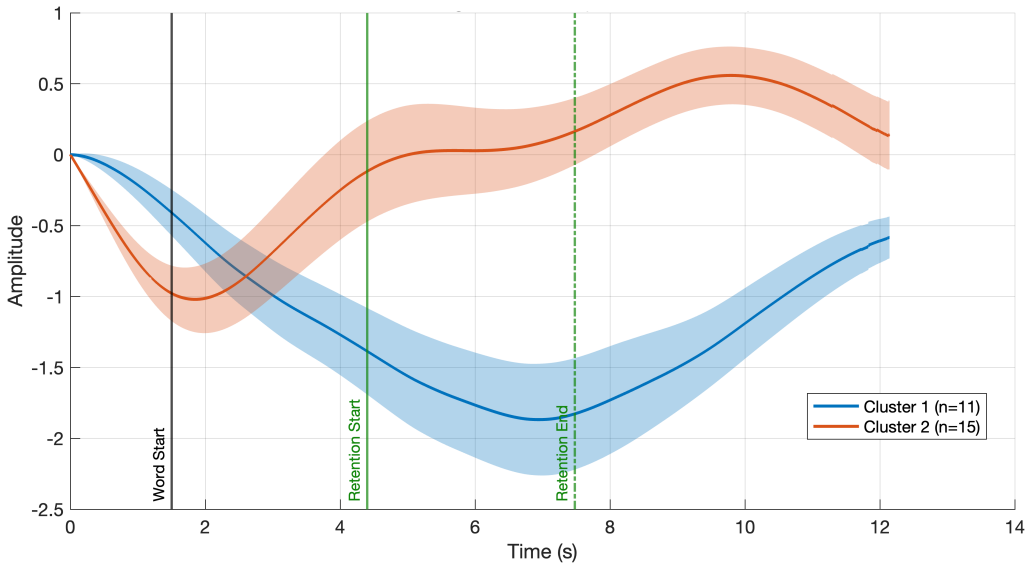


Figure. 13.41. Clustering result - average trial response of respiration rate with standard error shading (baseline corrected, k = 2, SNR = -11 dB)

Cluster-averaged respiration traces with SEM shading are shown for participants grouped via k-means clustering ($k = 2$) at SNR -11 dB. Data were averaged across Experiments 1 and 2. To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. Timepoints corresponding to word onset, retention start, and retention end are marked with vertical lines.

The two clusters exhibit distinct respiratory patterns, particularly during the retention phase. One group (Cluster 2) shows a more rapid initial drop followed by a larger increase in amplitude compared to the other. These differences may reflect individual variation in task-related respiratory dynamics.

supported by the elbow plot, where the within-cluster sum of squares sharply declined at $k = 2$ and plateaued thereafter

The corresponding cluster-averaged traces (Figure 13.43) showed substantial divergence in respiratory patterns across the task period. However, as with the other SNR levels, no statistically significant group differences were found in task accuracy, subjective effort, or perceived difficulty. This suggests that although individuals differ in physiological reactivity, these differences do not appear to predict behavioural outcomes.

Clustering results revealed the presence of distinct patterns in participants' respiration profiles, despite identical acoustic conditions and task demands. At each signal-to-noise ratio (SNR) level, a two-cluster solution was consistently supported by the elbow method, indicating the existence of two dominant physiological response types.

These results suggest that listeners differ in their physiological engagement during the task. Such variation could reflect differences in cognitive strategies, attention, or emotional regulation. Importantly, these response types emerged organically from the data, without any prior behavioural grouping.

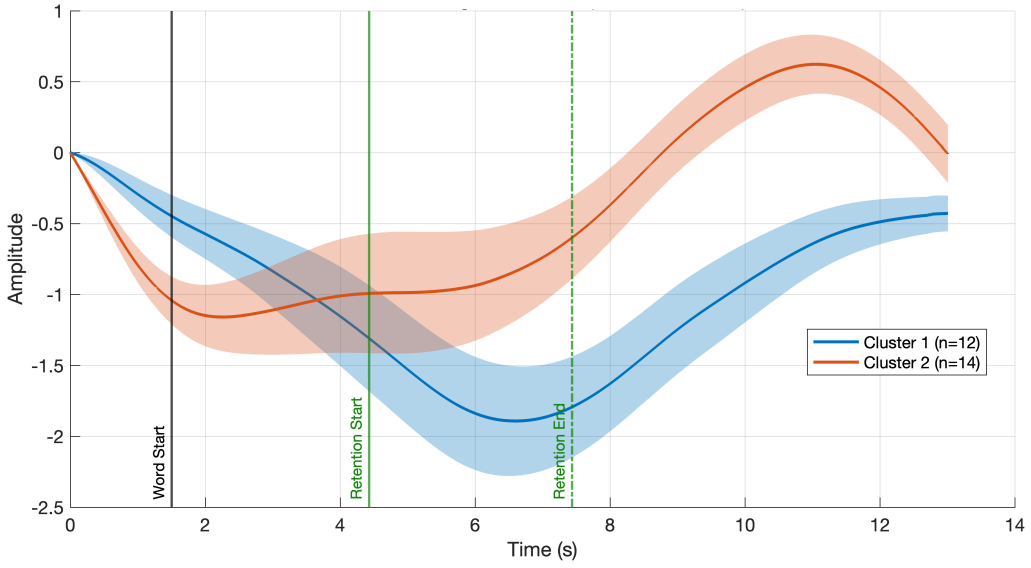


Figure. 13.42. Clustering result - average trial response of respiration rate with standard error shading (baseline corrected, k = 2, SNR = -6 dB)

At SNR -6dB, the clustering of respiration traces revealed two groups with distinct temporal patterns. To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The lines show mean respiration per cluster, and the shaded areas represent the standard error of the mean (SEM).

Cluster 1 (blue, $n = 12$) displayed a deeper suppression during the retention phase, followed by gradual recovery, while Cluster 2 (orange, $n = 14$) followed a shallower and earlier-shifting pattern. These profiles suggest variability in physiological modulation under moderately degraded listening conditions.

However, comparisons of behavioural measures - including accuracy, subjective difficulty, and reported effort - revealed no significant differences between the clusters. This pattern mirrors the findings observed in galvanic skin response (GSR) data, where similar groupings emerged in the absence of behavioural divergence.

Together, these findings imply that physiological responses can capture individual differences in task engagement or processing style that remain invisible to traditional behavioural measures. Clustering thus offers a valuable complementary perspective for interpreting listener variability.

However, it is also possible that the absence of observable behavioural differences may reflect limitations in statistical power, as the group sizes in each cluster were relatively small. Thus, while clustering reveals individual variability in physiological dynamics, further data may be needed to confirm whether such patterns are functionally meaningful.

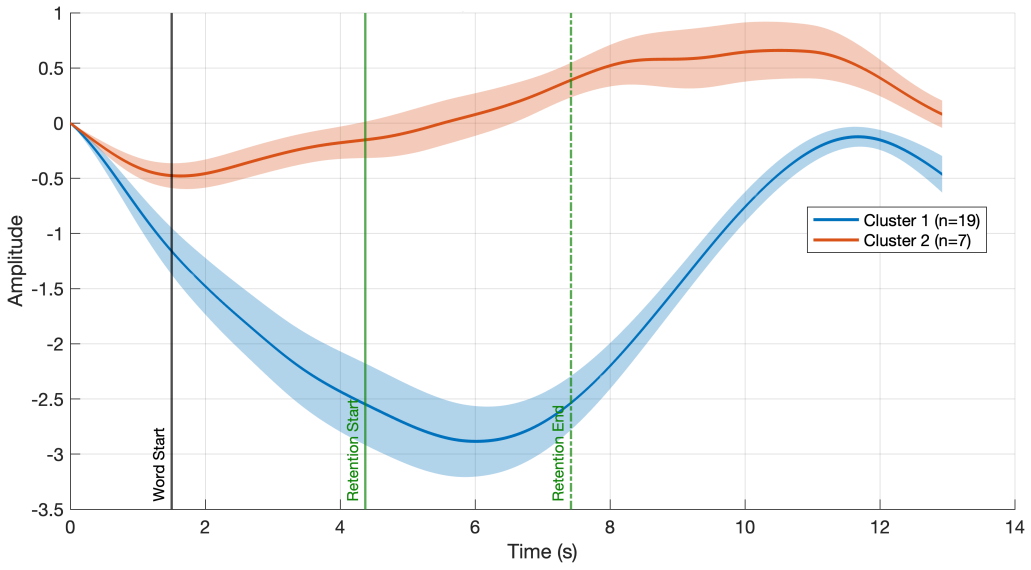


Figure. 13.43. Clustering result - average trial response of respiration rate with standard error shading (baseline corrected, k = 2, SNR = 12 dB)

Clustered respiration traces at SNR 12dB reveal two distinct group-level patterns in temporal dynamics. Solid lines represent the mean across individuals within each cluster, while shaded regions indicate the standard error of the mean (SEM). To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero.

Group 1 (in blue, $n = 19$) showed a marked dip during the retention period followed by a rebound, whereas Group 2 (in orange, $n = 7$) exhibited a shallower trajectory throughout. These differences suggest variation in physiological engagement strategies under high intelligibility.

13.5.6 Clustering Result Agreement Across Different SNR Levels (Respiration Rate)

To assess the consistency of clustering assignments across signal-to-noise ratio (SNR) levels, individual participant groupings were visualised in Figure 13.44. Although the two-cluster solution was optimal at each SNR level, the figure shows that most participants changed cluster assignments across conditions, indicating variability in group membership depending on acoustic clarity.

This observation is quantified in Figure 13.45, which presents the Adjusted Rand Index (ARI) as a measure of similarity in clustering solutions between SNR levels. The ARI accounts for agreement expected by chance, with values near 1 indicating strong consistency and values near 0 or negative suggesting little to no agreement (ARI ranges from -1 to 1). Most ARI values were close to zero or negative, implying that the composition of clusters was highly variable across acoustic conditions.

The one exception was the comparison between -11 dB and -6 dB SNRs, which showed moderate agreement (ARI = 0.575), suggesting some stability in participant groupings between these two moderately degraded conditions. However, the overall lack of consistency highlights that participants do not fall into fixed RR responses. Instead, their

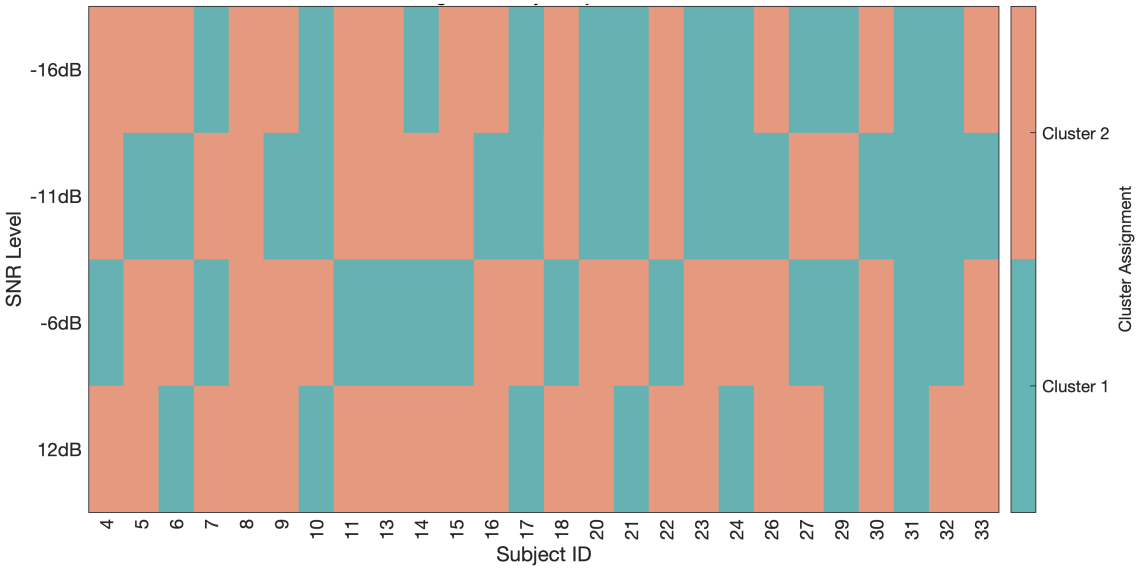


Figure. 13.44. Cluster membership across SNR levels - respiration rate

Heatmap displays each subject's assigned cluster (Cluster 1 or Cluster 2) at each SNR level. Rows represent SNR conditions (-16, -11, -6, and 12 dB), while columns denote individual participants. Colour indicates cluster identity, highlighting how participant classification varies across acoustic clarity.

While some individuals consistently fall into the same cluster, many shift across conditions, suggesting context-dependent physiological profiles rather than stable group membership. This variability may reflect fluctuations in internal state or task engagement rather than fixed individual traits.

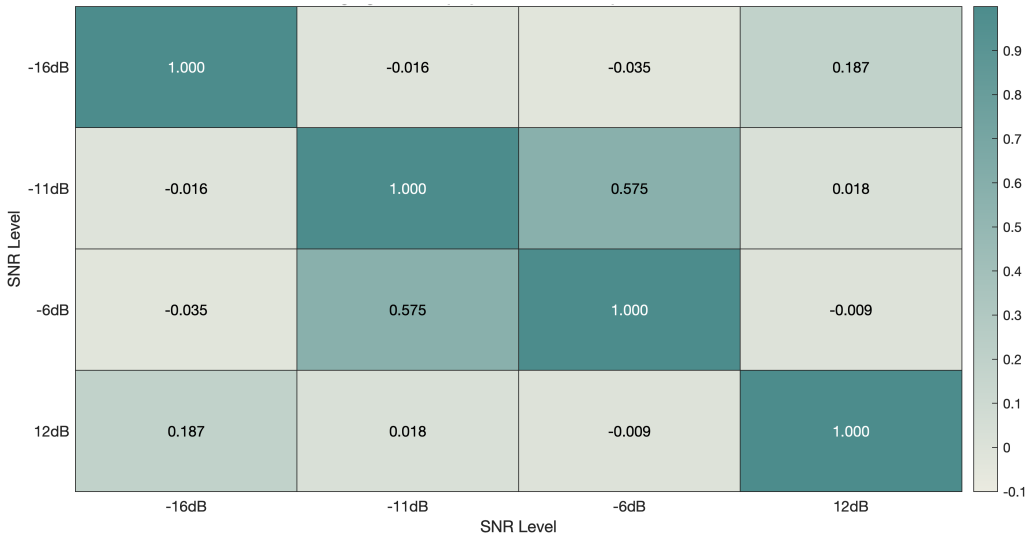


Figure. 13.45. Clustering membership agreement - adjusted rand Index (ARI) between SNR Pairings (respiration rate)

This matrix shows pairwise agreement between cluster solutions at each signal-to-noise ratio (SNR) level, using the Adjusted Rand Index (ARI). Values close to 1 indicate high consistency, while those near 0 suggest random or inconsistent assignments.

Relatively high agreement was observed between -11 and -6 dB (ARI = 0.575), but agreement was otherwise low or negative. This suggests that clustering solutions were not consistent across SNR conditions, reinforcing the idea that group structure may vary as a function of task acoustics.

respiration patterns-and by extension, their engagement or regulation strategies-may adapt dynamically in response to the task difficulty level.

13.6 Heart Rate

13.6.1 Data Overview

Heart rate was extracted from the original ECG recordings and interpolated to a sampling frequency of 1000 Hz (see Section 12.5, Page 128, for details). Data were then segmented within a time window spanning from 1.5 seconds before word onset to 5 seconds after the end of the retention period.

As shown in Figure 13.46, the resulting heart rate signals were mean-subtracted and averaged across participants for each signal-to-noise ratio (SNR) condition. This allowed visualisation of the general temporal profile of heart rate modulation in response to task events across varying listening conditions.

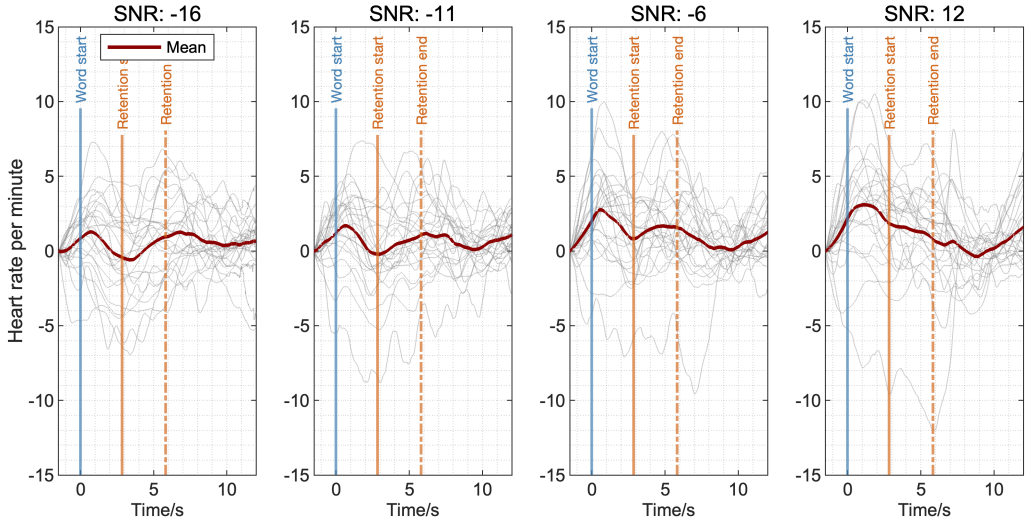


Figure. 13.46. Heart rate grand average (mean-subtracted) - data averaged from experiment 1 and 2, all SNR levels

This figure shows the average heart rate response (mean-subtracted) across four signal-to-noise ratio (SNR) conditions: -16, -11, -6, and 12 dB. The thick red line represents the grand average across participants, while the grey lines show individual responses. Key task events are marked: word onset (blue), retention start (solid orange), and retention end (dashed orange).

A clear increase in heart rate is visible around word onset across all SNRs, followed by dips during retention. The general shape is preserved across SNRs, though the amplitude and timing of features vary.

Despite inter-individual variability, a consistent phasic pattern was observed: a sharp increase in heart rate following word onset, followed by a dip during the retention interval. This pattern appeared across all SNR levels, although the amplitude and timing

varied. These results suggest that heart rate dynamically tracks key cognitive stages in the task.

13.6.2 Task-Evoked Changes in Heart Rate

To examine the effect of SNR on heart rate modulation, we extracted heart rate values at four key time points within each trial: baseline (1.5 seconds before word onset), word start, retention start, and retention end (see Figure 13.46). A Friedman test was conducted for each time point across SNR levels, but no statistically significant differences were found.

We calculated four difference scores to capture how heart rate changed from one phase to the next: RR Baseline - 1.5 seconds before the word start, Word start, Retention Start, Retention End.

Within each SNR level, heart rate changed clearly in response to the task: it rose after word onset, dropped during the retention period, and rose again toward the end of the trial. These shifts were statistically significant across nearly all comparisons (paired-wise t-test, $p < 0.01$), showing a consistent pattern across different difficulty levels.

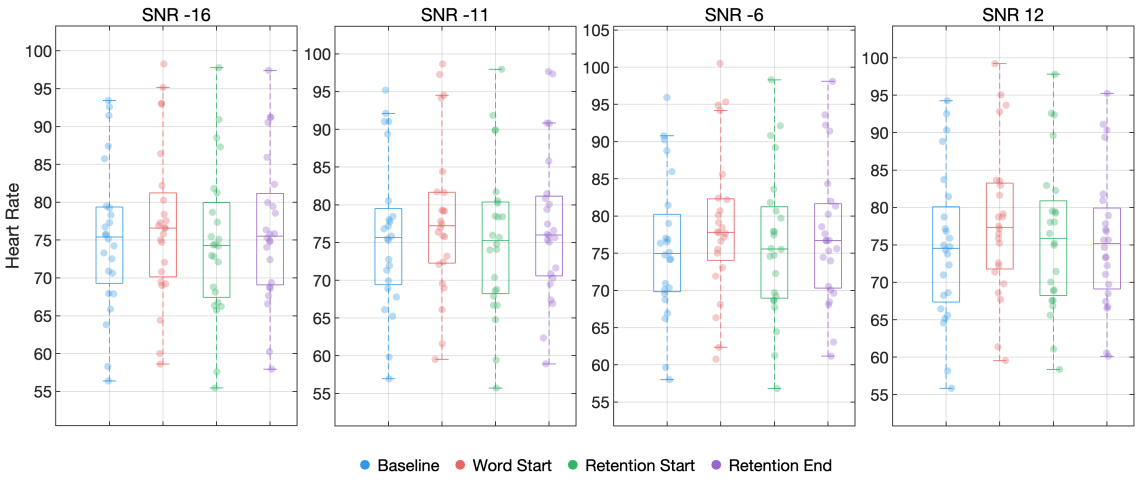


Figure. 13.47. Heart rate at four key task timepoints across SNR levels

This figure presents heart rate values at four key timepoints within each trial: baseline (1.5 seconds before word onset), word start, retention start, and retention end. Each subplot corresponds to a signal-to-noise ratio (SNR) condition: -16, -11, -6, and 12 dB. Colours reflect timepoints: baseline (blue), word start (red), retention start (green), and retention end (purple).
Heart rate increases sharply from baseline to word start across all SNRs, followed by a dip during the retention period. The degree of heart rate rise after retention varies by condition, with a more pronounced reactivation observed at lower SNRs. Pairwise statistical comparisons within each SNR level all shows significant difference (t-test, $p < 0.01$)).

13.6.3 Heart Rate Changes at Different SNR Levels

To examine the effect of SNR on heart rate modulation, we extracted heart rate values at four key time points within each trial: baseline (1.5 seconds before word onset), word start, retention start, and retention end (see Figure 13.46). A Friedman test was conducted for each time point across SNR levels, but no statistically significant differences were found.

We next investigated the dynamic change in heart rate over time by computing differences between peak and dip points. Specifically, we quantified heart rate recovery during the retention period as the difference between the peak at retention end and the dip at retention start. This revealed a statistically significant difference across SNR levels ($p < 0.01$), indicating that heart rate recovery varies systematically with signal clarity (see Figure 13.48.

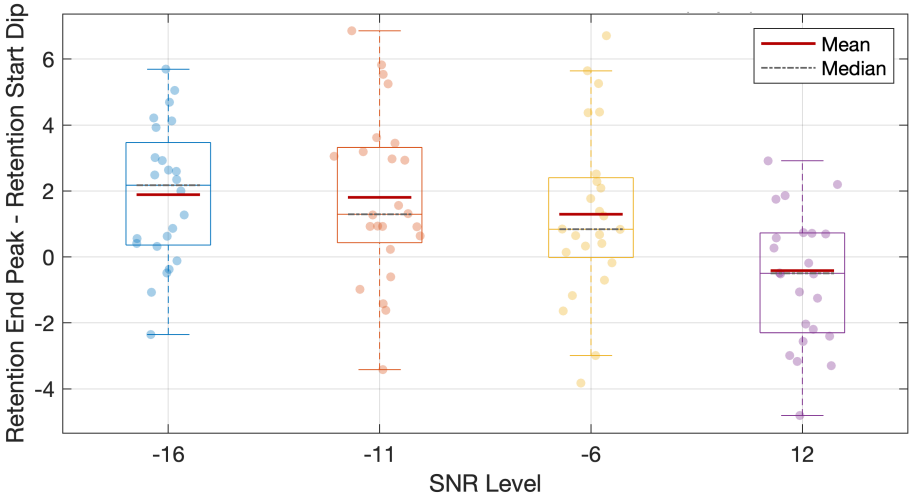


Figure. 13.48. Heart rate difference between retention end and retention start across SNR levels

This boxplot shows the difference in heart rate between the peak at the end of the retention period and the dip at the start of the retention period across four signal-to-noise ratio (SNR) conditions: -16, -11, -6, and 12 dB. Each dot represents a participant. The red line indicates the mean, while the grey dashed line shows the median.
A larger positive difference reflects greater heart rate recovery during the retention period. Differences tend to be smaller at higher SNR levels (e.g., 12 dB), suggesting reduced physiological recovery as clarity improves, although variability is high.
Results of Friedman test suggests that there is a significant difference of heart rate recovery across different SNR levels ($p < 0.01$.)

The magnitude of change between retention start and retention end reflects differences in task-related engagement. In more challenging conditions (e.g., lower SNRs), heart rate tends to rise again before the response, possibly reflecting increased cognitive effort or arousal. In contrast, under easier conditions (e.g., 12 dB SNR), heart rate remains low, suggesting that participants stay relaxed even while preparing to respond.

13.6.4 Within Individual Consistency - Permutation Test Result of Heart Rate

To evaluate the consistency of heart rate responses within individuals across repeated experimental sessions, we conducted a permutation test comparing within-subject correlations to a null distribution generated by random pairing. As shown in Figure 13.49, within-subject correlations were significantly higher than would be expected by chance across all SNR conditions ($p < 0.001$), indicating robust individual response patterns.

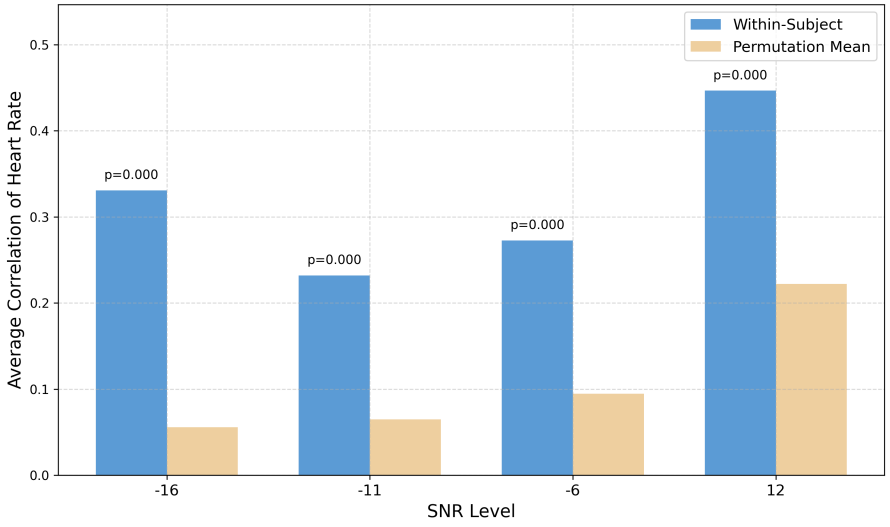


Figure. 13.49. Permutation test of within-subject correlations in heart rate average trial response

This figure presents the average within-subject heart rate correlation (blue bars) across four SNR conditions: -16, -11, -6, and 12 dB. Each blue bar is contrasted with the corresponding permutation-based null distribution mean (tan bars), representing the expected correlation under random pairing.

All within-subject correlations were significantly higher than chance (permutation mean), with p-values from the permutation tests shown above each bar. These results indicate robust individual consistency in heart rate responses within each SNR condition.

Figure 13.50 provides an illustrative example using two participants at the -11 dB SNR level. Despite notable inter-individual variability (grey traces), both participants displayed highly similar heart rate trajectories across experiments, underscoring the presence of consistent physiological dynamics at the individual level.

Given the strong within-subject consistency observed across experiments, we next explored whether participants could be grouped based on similarities in their heart rate response patterns. We applied cluster analysis to assess whether distinct physiological response profiles emerge across individuals, and whether these clusters relate to behavioural measures.

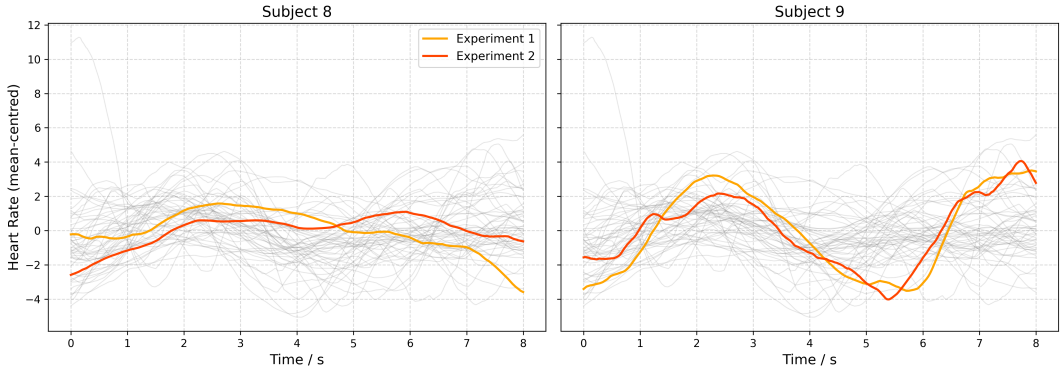


Figure. 13.50. Example of within-subject heart rate consistency at -11 dB SNR

This figure shows heart rate responses (mean-centred) for two example participants (Subject 8 and Subject 9) at -11 dB SNR, illustrating within-subject consistency across two experimental sessions. Thick lines represent the average response in Experiment 1 (orange) and Experiment 2 (red). Grey traces in the background show responses from other participants for reference.

13.6.5 Clustering Result of Heart Rate (presenting per SNR level

To further examine the structure underlying individual variability in heart rate responses, we applied cluster analysis to group participants with similar physiological profiles. Given the significant within-subject consistency observed across all SNR levels (Figure 13.49), we reasoned that between-subject differences may reflect meaningful subgroup patterns rather than noise.

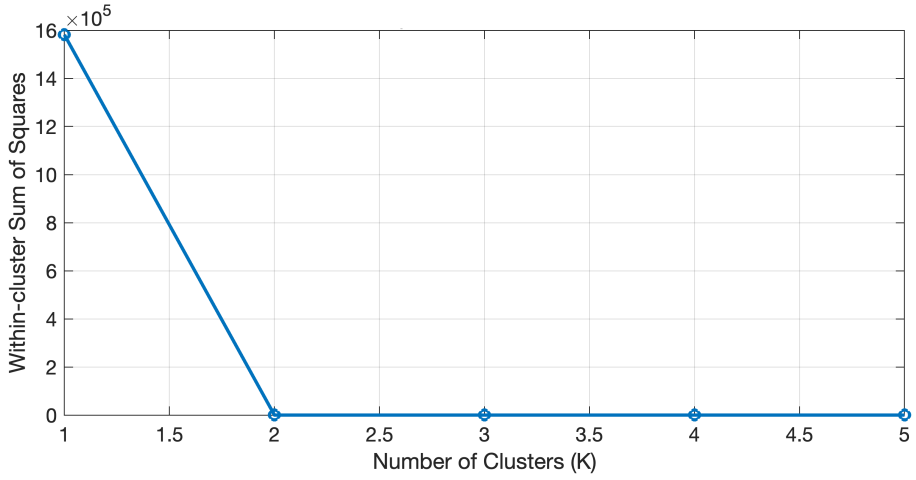


Figure. 13.51. K-Means elbow plot of cluster sum of squares (WCSS) in within subject correlations for heart rate clustering (-16 dB)

This elbow plot shows the within-cluster sum of squares as a function of the number of clusters (K) for k-means clustering applied to heart rate response features at -16 dB SNR. A clear inflection is observed at $K = 2$, suggesting that two clusters provide an optimal balance between explanatory power and parsimony.

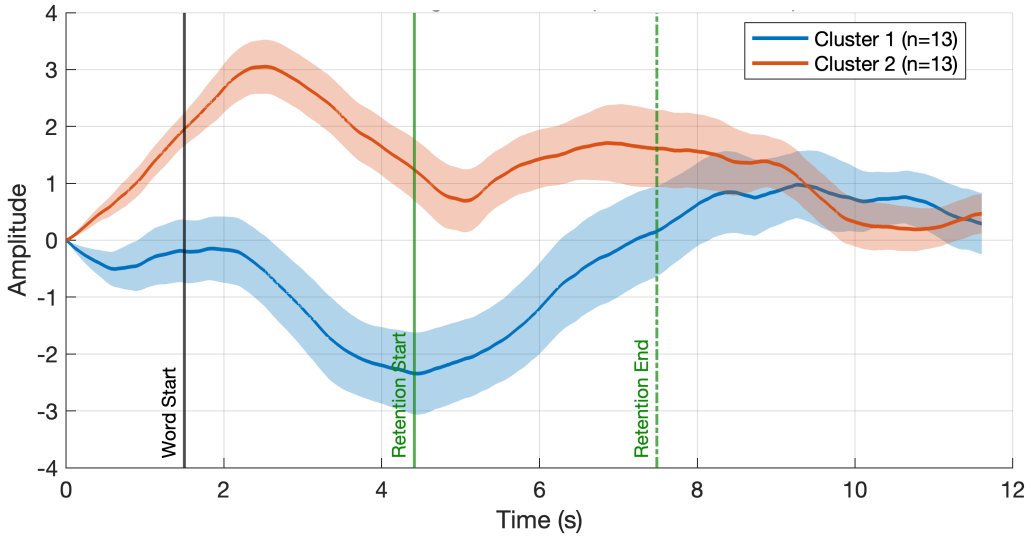


Figure. 13.52. Clustering result - average trial response of heart rate with standard error shading (baseline corrected, k = 2, SNR = -16 dB)

This figure shows the average heart rate responses for the two clusters identified via k-means clustering ($k = 2$) at -16 dB SNR. Each line represents the mean response across participants in a cluster, with shaded regions indicating the standard error of the mean (SEM). Key task events are marked: word onset (black), retention start (green), and retention end (dashed green).

To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The two clusters display distinct response profiles, particularly around the retention period, suggesting meaningful differences in physiological engagement or regulation strategies among participants.

Clustering result at SNR : -16 dB (Heart Rate) An elbow plot of within-cluster sum of squares (Figure 13.51) indicated that two clusters provided the optimal solution for the -16 dB SNR condition. The resulting cluster averages (Figure 13.52) revealed two distinct patterns of heart rate modulation, especially during the retention period, suggesting potential differences in cognitive or physiological engagement across individuals.

To assess whether the observed physiological clusters corresponded to meaningful behavioural differences, we compared task accuracy, subjective difficulty, and subjective effort between the two groups (Figure 13.53). Statistical comparisons revealed no significant differences across any of the three measures.

Clustering result at SNR : -11 dB (Heart Rate) As with the -16 dB SNR condition, we also examined whether the physiological clusters identified at -6 dB SNR corresponded to differences in behavioural or subjective measures. Participants were grouped into two distinct clusters based on heart rate response patterns (Figure 13.55), and their accuracy, perceived task difficulty, and effort were compared.

However, consistent with the findings at -16 dB, no significant differences were observed between the two groups across any of these measures, suggesting that physiological

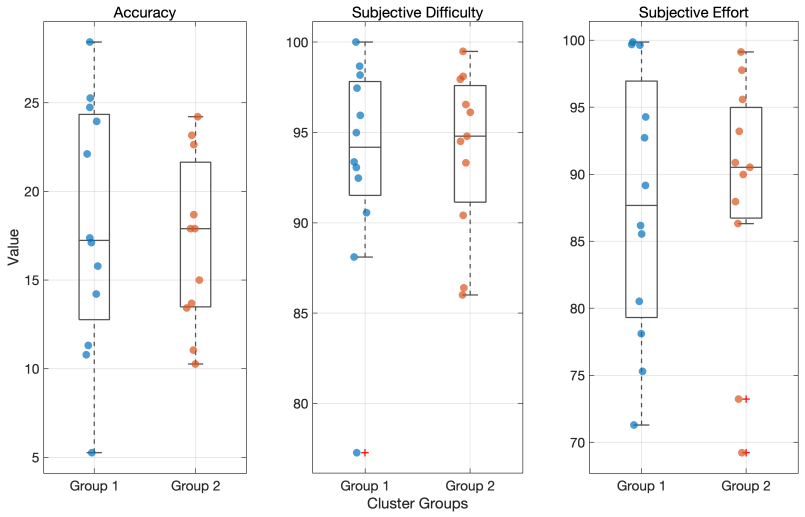


Figure. 13.53. Comparison of behavioural and subjective measures between heart rate response clusters at -16 dB SNR

This figure compares accuracy, subjective difficulty, and subjective effort between the two participant groups identified through heart rate clustering at -16 dB SNR. Each dot represents an individual, with boxplots overlaid to indicate group distributions. Despite similar levels of perceived difficulty and effort, Group 1 showed more variability in accuracy, suggesting differing behavioural outcomes associated with the physiological response profiles.

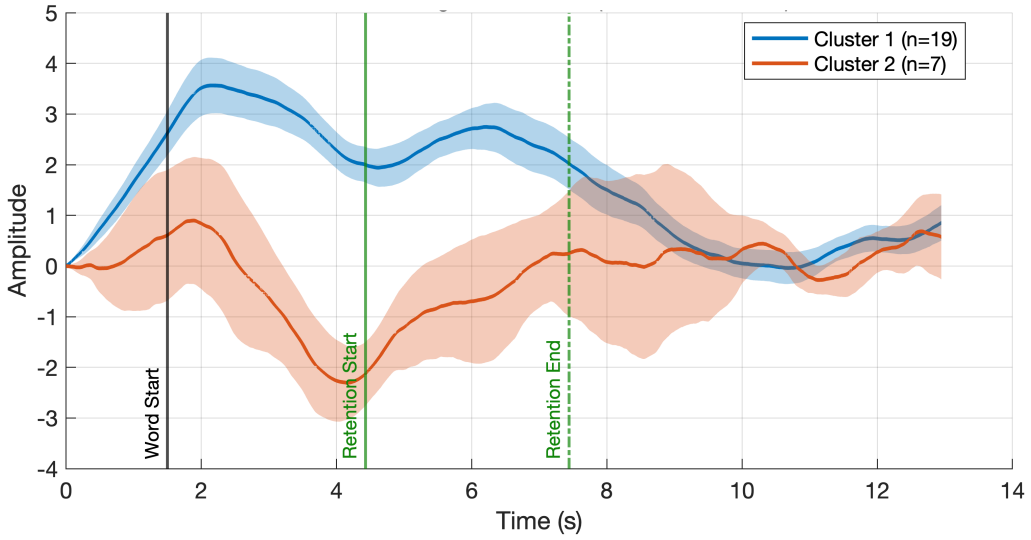


Figure. 13.54. Clustering result - average trial response of heart rate with standard error shading (baseline corrected, k = 2, SNR = -11 dB

This figure shows the average heart rate responses for the two clusters identified via k-means clustering ($k = 2$) at -6 dB SNR. Each line represents the mean response across participants in a cluster, with shaded regions indicating the standard error of the mean (SEM). Task events are marked at word onset (black), retention start (green), and retention end (dashed green).

To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The clusters display distinct amplitude and timing characteristics, particularly surrounding the retention period. These results suggest that differences in physiological modulation may reflect underlying cognitive or emotional processing differences across participants.

response profiles did not systematically map onto task performance or self-reported experience at this SNR level.

Clustering result at SNR : -6 dB (Heart Rate) At −6 dB SNR, *k*-means clustering again revealed two distinct heart rate response profiles (Figure 13.55). The clusters differed in both amplitude and temporal dynamics, particularly during the retention period, suggesting differential physiological engagement across participants under moderate noise conditions.

To assess whether these physiological groupings reflected differences in task performance or subjective experience, we compared behavioural accuracy, perceived task difficulty, and self-reported effort between the two clusters. However, consistent with findings from the −16 dB and 12 dB SNR conditions, no significant differences were found in any of these measures, indicating that physiological divergence did not correspond to measurable behavioural or experiential outcomes at this SNR level.

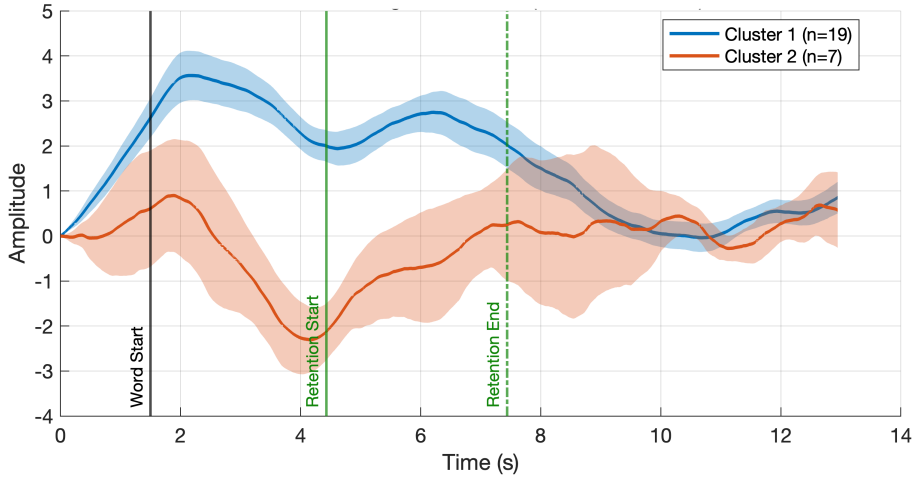


Figure. 13.55. Clustering result - average trial response of heart rate with standard error shading (baseline corrected, k = 2, SNR = -6 dB)

This figure shows the average heart rate responses for the two clusters identified via *k*-means clustering ($k = 2$) at -6 dB SNR. The solid lines represent the mean response of each cluster, and the shaded areas denote the standard error of the mean (SEM). Task markers indicate word onset (black), retention start (green), and retention end (dashed green).

To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The clusters exhibit clearly divergent temporal dynamics, with notable differences during the retention phase. These patterns suggest that individual participants engaged with the task in physiologically distinct ways despite being presented with the same acoustic conditions.

Clustering result at SNR: 12 dB (Heart Rate) At 12 dB SNR, the *k*-means clustering again produced two distinct heart rate response profiles (Figure 13.56), with Cluster 1 showing a stronger initial increase followed by a gradual decline, and Cluster 2 displaying a sustained dip throughout the retention phase.

Despite the clear physiological divergence, we found no significant differences in accuracy, subjective difficulty, or effort ratings between the two groups.This pattern is consistent with the results observed at -16 dB, -11 dB, and -6 dB SNR, further supporting the conclusion that while physiological responses can vary markedly across individuals, these differences are not necessarily reflected in task performance or self-reported experience, even under high SNR conditions.

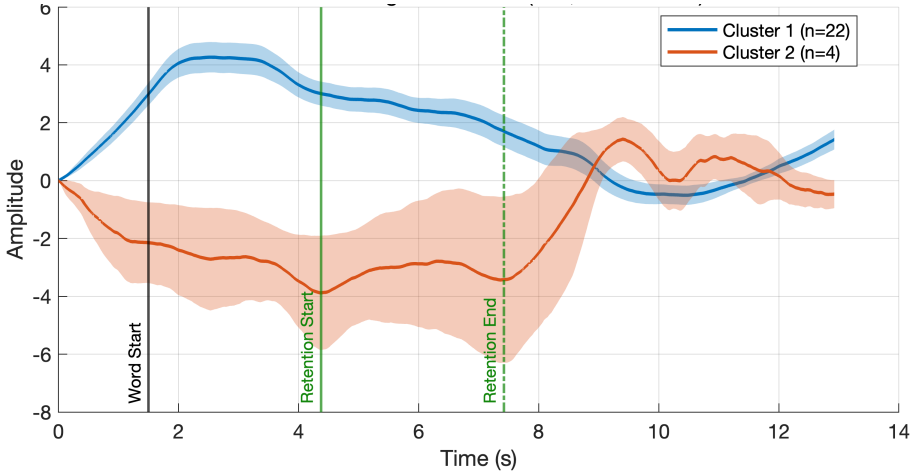


Figure. 13.56. Clustering result - average trial response of heart rate with standard error shading (baseline corrected, k = 2, SNR = 12 dB)

This figure presents the mean heart rate responses of two participant clusters identified using k-means clustering ($k = 2$) at 12 dB SNR. Solid lines indicate cluster averages, and shaded areas represent the standard error of the mean (SEM). Key task events are marked: word onset (black), retention start (green), and retention end (dashed green).
To allow for easier comparison, the first value was subtracted, aligning all traces to a common starting point of zero. The two clusters exhibit distinct response patterns throughout the trial, particularly during the retention phase. These differences may reflect individual variations in physiological regulation under low cognitive demand conditions.

Together, the clustering analyses across all SNR levels revealed consistent evidence of individual variation in heart rate responses, yet no clear behavioural or subjective correlates. While the physiological profiles were distinguishable within each SNR condition, it remained unclear whether these response patterns were stable within individuals across different acoustic environments.

To address this, we next evaluated the consistency of cluster membership across SNR levels, both visually and quantitatively.

13.6.6 Clustering Result Agreement Across Different SNR Levels (Heart Rate)

To explore whether participants tended to exhibit stable physiological profiles across listening conditions, we examined the agreement of cluster assignments across SNR levels.

Figure 13.57 illustrates how cluster assignments varied across SNR conditions for each participant. While some individuals remained in the same cluster across multiple conditions, many others shifted cluster membership depending on the SNR level, suggesting that physiological response patterns may not be entirely trait-like.

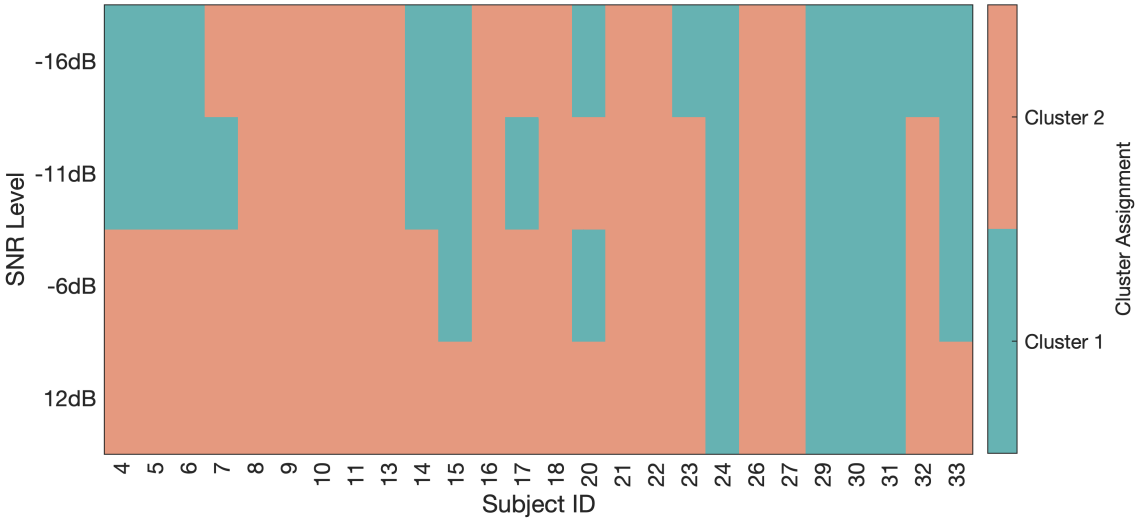


Figure. 13.57. Cluster membership across SNR levels - heart rate

This heatmap shows the clustering assignments of each participant (columns) across the four SNR conditions (rows). Each cell indicates the cluster label assigned at a specific SNR level, enabling visual comparison of assignment consistency across conditions.

To quantify the consistency of clustering structures, we computed the Adjusted Rand Index (ARI) between every pair of SNR-specific clustering solutions (Figure 13.58). The ARI values were generally low to moderate, with the highest agreement observed between -6 dB and 12 dB SNR (ARI = 0.537).

These results indicate only limited consistency in clustering outcomes across SNR levels, implying that the observed groupings may be partially condition-specific rather than reflecting stable individual differences.

Overall, the clustering analyses revealed meaningful variation in physiological response patterns within each SNR condition, consistently forming two distinct groups. However, these groupings did not correspond to significant differences in task accuracy or subjective ratings, indicating a disconnect between physiological profiles and behavioural or experiential measures.

Moreover, when comparing cluster assignments across SNR levels, moderate agreement was observed, with most participants shifting group membership depending on the acoustic condition. This suggests that while individual differences in heart rate dynamics exist, they still vary by the specific demands of the listening environment rather than reflecting stable, trait-like patterns. These findings underscore the context-sensitive nature of physiological responses in challenging listening tasks.

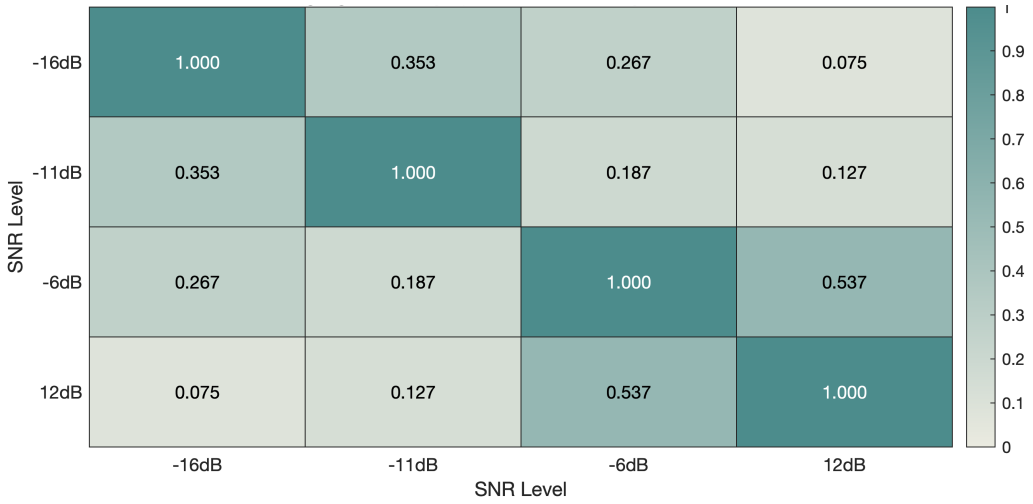


Figure. 13.58. Clustering membership agreement - adjusted rand Index (ARI) between SNR Pairings (heart rate)

This matrix displays the Adjusted Rand Index (ARI) values quantifying the similarity between clustering results across all SNR levels. ARI values range from 0 (no agreement) to 1 (perfect agreement), with diagonal entries representing self-comparisons.

13.7 Electroencephalography (EEG)

13.7.1 Data Overview

This section presents EEG alpha power responses during three critical task phases: listening, retention, and responding. Figures 13.59, 13.60, and 13.61 display alpha envelopes baseline-corrected to the one-second window preceding each respective event.

Data Processing and Alignment EEG was recorded at the Pz channel and bandpass-filtered for the alpha band (8–13 Hz). The envelope was extracted using the Hilbert transform. Data were epoched relative to task events using timing markers derived from auditory clicks: Word Start (Click 1), Retention Start (Click 6), and Respond Start (Click 7). For each phase, the signal was baseline-corrected using a window from -1 to 0 seconds relative to the alignment point.

Listening Phase In the listening phase (Figure 13.59) , alpha amplitude rises from noise onset and peaks just after Word Start (0 s), followed by a progressive decline towards Retention Start. Notably, lower SNRs (e.g., -16 dB) show stronger alpha suppression post-Word Start, potentially reflecting greater listening effort. The vertical line at -0.5 s indicates Noise Start.

Retention Phase During the retention period (Figure 13.60), alpha amplitude remains relatively low and stable for high-demand conditions (e.g., -16 dB), indicating sustained

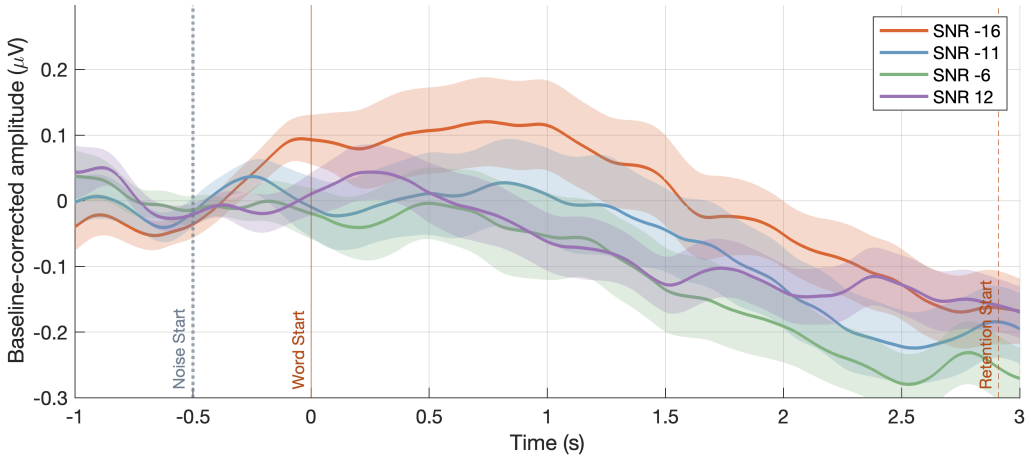


Figure. 13.59. EEG Alpha envelope around word start across SNR levels with SEM shading

This figure displays the baseline-corrected EEG alpha band envelope during the listening period, aligned to the onset of the target word (0 s), across four SNR conditions. Lines indicate the mean response, and shaded regions represent standard error of the mean (SEM) across participants. The vertical dashed line at -0.5 s indicates the onset of background noise. Word onset and retention onset are marked with solid and dashed vertical lines, respectively. A clear alpha suppression follows word onset, with stronger suppression under more favourable SNRs (e.g., +12 dB), reflecting higher engagement with clearer auditory input. In contrast, higher noise levels (e.g., -16 dB) exhibit weaker suppression and elevated alpha power throughout, suggesting reduced auditory encoding or increased disengagement. These findings align with cognitive load theories of alpha dynamics, where lower alpha is associated with increased sensory processing demands during speech perception.

cognitive load. In contrast, higher SNRs (e.g., $+12$ dB) show earlier recovery. A notable rise begins just before the Respond Start marker, particularly in easier SNRs.

Responding Phase Alpha power increases sharply after the Respond Start (0 s) across all SNRs, particularly at higher SNRs (see Figure 13.61). This rebound appears prolonged and more sustained than in prior literature. The shaded regions represent standard error of the mean across subjects.

Comparison to Study 1 Study 1 (Alhanbali, 2017), observed a brief alpha rebound peaking around 1 s after response, using a binary Yes/No speech-in-noise task. In contrast, the current task involves a more demanding 5-word identification process, combining auditory, verbal, visual, and motor demands. This likely explains the more sustained and pronounced alpha rebound observed here.

A comparison between the current study and the listening memory paradigm from Study 1 (Alhanbali, 2017) is summarised in Table 13.3. The task used in Study 1 involved binary verbal responses to speech in noise, with a relatively short response window and moderate cognitive demand. Their alpha rebound, observed in posterior channels, was brief and peaked roughly 1 second after response onset.

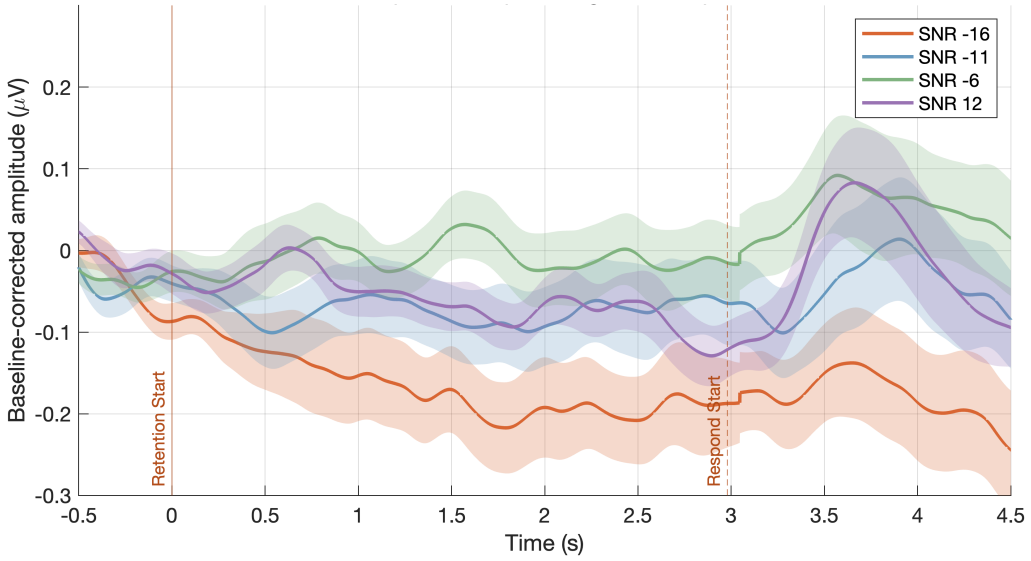


Figure. 13.60. EEG Alpha envelope during the retention period across SNR levels with SEM shading

This figure illustrates the EEG alpha band envelope during the retention period, baseline-corrected to the 1 s interval preceding retention onset (0 s). The mean alpha response is shown for each SNR condition, with shaded areas representing the standard error of the mean (SEM) across participants. All SNR conditions exhibit a suppression of alpha power following retention onset. However, suppression is most pronounced under difficult listening conditions (e.g., −16 dB), consistent with sustained cognitive load during memory maintenance. In contrast, clearer speech (e.g., +12 dB) shows less suppression and a sharper increase prior to response. The vertical dashed line indicates the onset of the response window. The rising alpha trend approaching this marker, particularly under higher SNRs, may reflect anticipatory disengagement or motor preparation. These trends contrast with Alhanbali et al. (2018), where a short rebound followed binary responses, likely due to task simplicity.

In contrast, the current study (Study 2) employed a more cognitively demanding paradigm. Participants listened to a noisy sentence and subsequently selected five words from a 10-by-5 visual matrix. This involved auditory decoding, working memory retention, lexical identification, visual scanning, and motor execution, sustained over several seconds.

These differences are reflected in the alpha trajectories. During the retention phase (Fig. 13.60), we observe a gradual decline in alpha power, consistent with sustained cognitive engagement. Notably, after Respond Start, there is a clear and sustained alpha rebound that lasts beyond 2 seconds (Fig. 13.61). This rebound is stronger and more prolonged at higher SNRs, potentially reflecting reduced effort or earlier response resolution under easier listening conditions.

The data suggest that alpha activity during retention reflects memory maintenance load, while the post-response rebound reflects release from effort. The magnitude and duration of the rebound scale with task complexity, consistent with increased cognitive unburdening after response.

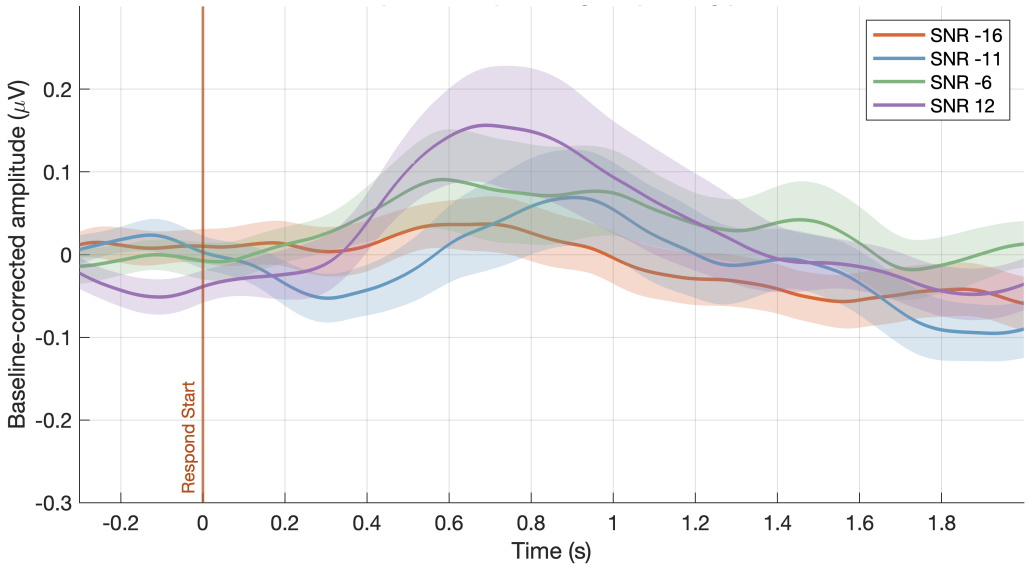


Figure. 13.61. EEG Alpha envelope during the responding period across SNR levels with SEM shading

This figure focuses on the EEG alpha response aligned to the start of the response window (0 s), zooming in on the period surrounding participant interaction. Each curve shows the mean alpha envelope by SNR condition, and shaded regions reflect the standard error of the mean (SEM).

A robust increase in alpha power is observed following the response onset, most prominently under high-SNR conditions (e.g., +12 dB). This rebound is likely linked to cognitive disengagement or reduced attentional demands after the task is completed.

Compared with Alhanbali et al. (2018), who reported a brief post-response alpha burst, the current task elicits a broader and more sustained rebound, possibly due to the longer, more effortful response procedure in the current paradigm.

Table. 13.3. Comparison of Task Design and Alpha Response between Study 1 and 2

Feature	**Study 1**	**Study 2**
Response type	Binary (Yes/No)	Multi-step (5-word selection)
Response duration	Short (∼1 s)	Extended (several seconds)
Effort structure	Listen → Hold → Decide	Listen → Identify → Search matrix → Click 5 words
Cognitive load	Moderate	High (verbal + auditory + visual + motor)
Alpha rebound	Brief, peaking ∼1 s post-response	Sustained, stronger, prolonged

13.7.2 Task-Evoked Changes in EEG

As reported above, clear task-evoked alpha modulations were observed across listening, retention, and responding phases, with notable SNR-dependent differences. Alpha suppression was strongest during high-demand listening and retention phases, while alpha rebound was prominent post-response, especially under high-SNR conditions.

13.7.3 EEG Changes at Different SNR Levels

SNR effects were evident in the magnitude and dynamics of alpha responses. Poor SNRs (e.g., -16 dB) elicited sustained suppression during listening and retention, reflecting heightened cognitive demands. In contrast, clearer SNRs led to earlier alpha recovery and stronger post-response rebound, suggesting reduced processing load.

13.7.4 Within-Individual Consistency - Permutation Test Results

To assess consistency of response shape within individuals, permutation tests were applied using the listening window (Word Start to Retention Start). However, these tests failed to show robust intra-subject correlation across trials or SNR levels. One possible explanation is the high variability in individual alpha responses under noisy conditions, which may have masked any stable shape patterns.

Another key factor is the data preprocessing approach: we employed a trial-wise adaptive method that extracted epochs based on task-aligned events (click markers), and applied trial-specific baseline correction and smoothing. While this enhances task-locked responses, it may attenuate or distort broader signal dynamics that clustering and permutation-based shape analyses typically rely on.

13.7.5 Clustering Considerations

Clustering was not performed in this analysis due to the high degree of variability observed in trial-level EEG responses. The envelopes showed substantial fluctuations across subjects and trials, with no visually or statistically consistent shapes emerging. Given the complex nature of the task and the trial-specific preprocessing (e.g., adaptive segmentation and baseline correction), clustering would likely yield unstable or uninterpretable groupings.

Further, exploratory clustering would require stronger within-subject consistency and more uniform signal timing than is currently observed. At this stage, we prioritised clarity of average task-evoked responses over unsupervised pattern discovery.

Summary We observed clear task-related alpha power modulations, shaped by both SNR level and task phase. Alpha suppression reflected effort during listening and memory maintenance, while the post-response rebound suggested disengagement. However, further intra-individual or pattern-based analyses were not pursued due to high variability and lack of within-subject consistency. The adaptive, trial-wise preprocessing method used-while effective for capturing event-locked trends-may have diminished the comparability of trial shape features needed for permutation testing and clustering analyses.

13.8 Relationship between Different measurements

Two analyses were conducted to examine the relationship between performance, subjective effort, subjective difficulty, and physiological measurements. The first focused on clustering results to assess the consistency of grouping based on physiological responses. The second involved correlation analysis.

Specifically, we extracted changes in physiological indices during the listening period - defined as the time window between sentence onset and the start of the retention phase (i.e., sentence end). These changes were then correlated separately with performance, perceived effort, and perceived difficulty. The results are presented in the following section

13.8.1 Clustering Agreement Across Different Physiological Measures

To investigate how consistently participants are grouped based on their physiological responses, we conducted a clustering agreement analysis across four physiological measures: heart rate, galvanic skin response (GSR), pupil size, and respiration rate. Each measure was analysed independently at four signal-to-noise ratio (SNR) levels: -16, -11, -6, and 12 dB, with clustering performed separately for each condition.

Participants were grouped into two clusters per measure based on their physiological signal patterns. To assess the similarity of clustering structures across modalities, we compiled the subject-wise cluster labels into a matrix and calculated the pairwise Adjusted Rand Index (ARI) between measures. This approach allowed us to evaluate both global cross-modal consistency and subject-specific group membership under varying SNR conditions.

The ARI values quantify the agreement between two clustering outcomes, where 1 indicates perfect alignment, 0 reflects random similarity, and negative values suggest systematic disagreement. The resulting agreement patterns are visualised in the ARI

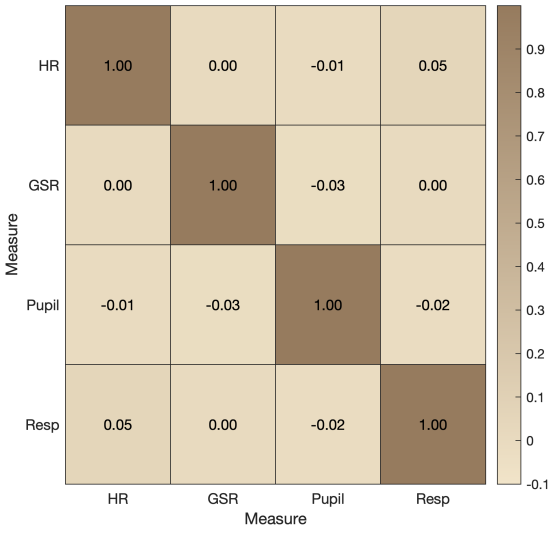


Figure. 13.62. Clustering membership agreement - adjusted rand index (ARI) across measures (SNR = - 16 dB)

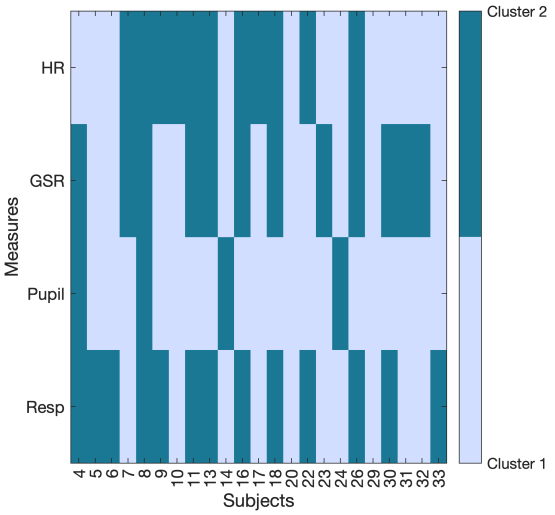


Figure. 13.63. Clustering membership assignment across measures (SNR = - 16 dB)

Clustering agreement (left) and subject-wise cluster assignments (right) across physiological measures (Heart Rate, GSR, Pupil, Respiration Rate) for SNR = −16 dB). The Adjusted Rand Index heatmap shows low agreement between measures, while the block plot visualises individual subject assignments to clusters 1 or 2.

heatmaps shown in Figure 13.62, Figure 13.64, Figure 13.66, and Figure 13.68 for SNR levels of −16, −11, −6, and 12 dB, respectively.

To complement these heatmaps, the block plots in Figure 13.63, Figure 13.65, Figure 13.67, and Figure 13.69 display the individual subject cluster assignments for each physiological measure. These plots reveal whether participants are consistently grouped across modalities or vary significantly in their cluster membership.

Across all SNR conditions, the clustering agreement between measures is generally low. Most ARI values remain close to zero or slightly negative, with the highest observed value (approximately 0.35) occurring between heart rate and respiration rate at 12 dB. This suggests that different physiological signals tend to yield divergent clustering structures, even under favourable acoustic conditions.

This pattern is confirmed by the block plots, where many subjects change cluster assignment depending on the modality. For example, a participant in cluster 1 for heart rate may appear in cluster 2 for GSR or pupil size. Such variability indicates limited overlap in the way different physiological systems reflect participant responses.

Overall, these findings suggest that the physiological measures are not capturing a single unified state but instead reflect complementary, modality-specific processes. The low cross-modal clustering agreement highlights the multidimensionality of physiological reactivity to speech-in-noise, with each measure likely sensitive to distinct autonomic, perceptual, or cognitive components of the task.

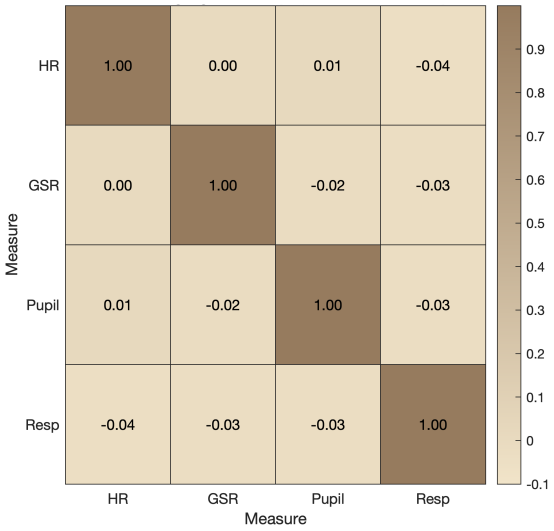


Figure. 13.64. Clustering membership agreement - adjusted rand index (ARI) across measures (SNR = - 11 dB)

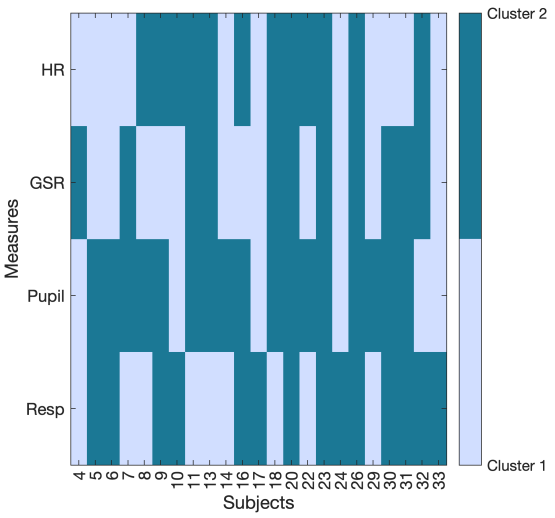


Figure. 13.65. Clustering membership assignment across measures (SNR = - 11 dB)

Clustering agreement (left) and subject-wise cluster assignments (right) across physiological measures (Heart Rate, GSR, Pupil, Respiration Rate) for SNR = -11 dB. The Adjusted Rand Index heatmap shows near-zero or negative agreement, indicating low consistency across modalities. The right plot visualises which subjects were grouped together for each measure.

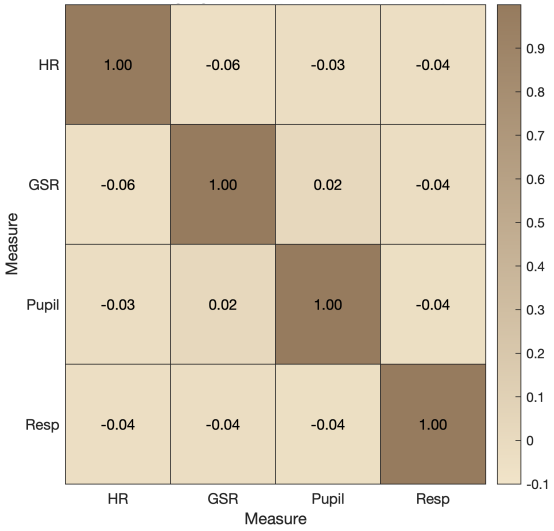


Figure. 13.66. Clustering membership agreement - adjusted rand index (ARI) across measures (SNR = - 6 dB)

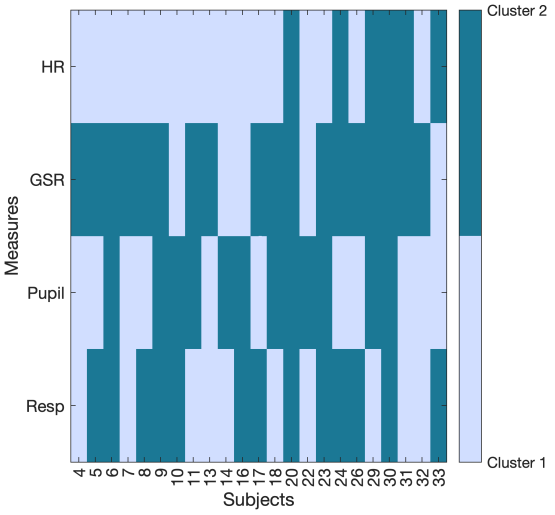


Figure. 13.67. Clustering membership assignment across measures (SNR = - 6 dB)

Clustering agreement (left) and subject-wise cluster assignments (right) across physiological measures (Heart Rate, GSR, Pupil, Respiration Rate) for SNR = -6 dB. While Heart Rate and Respiration Rate show some consistency, most pairwise agreements remain close to zero, as shown in the ARI heatmap. Subject-level groupings vary notably across modalities.

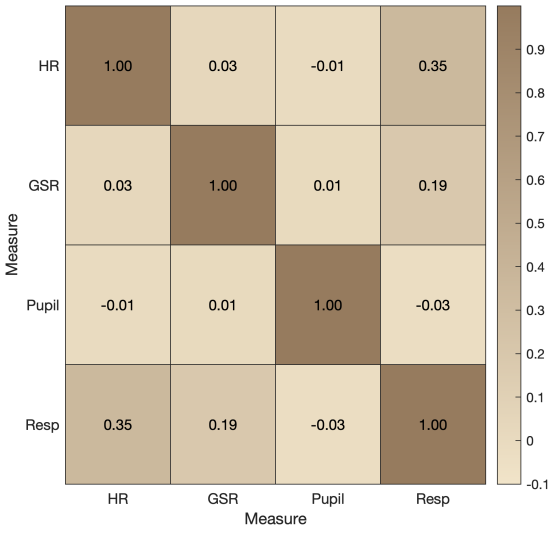


Figure. 13.68. Clustering membership agreement - adjusted rand index (ARI) across measures (SNR = 12 dB)

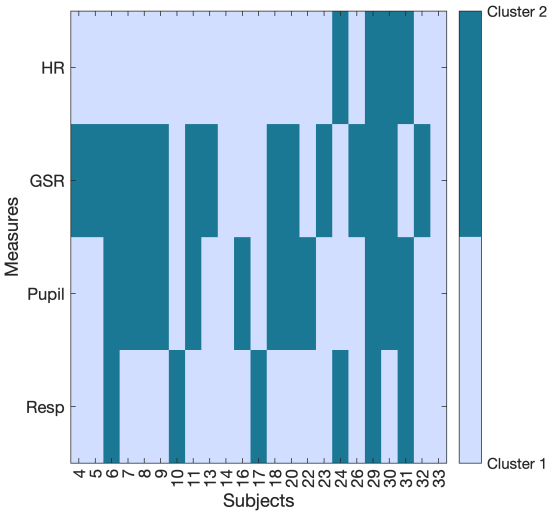


Figure. 13.69. Clustering membership assignment across measures (SNR = 12 dB)

Clustering agreement (left) and subject-wise cluster assignments (right) across physiological measures (Heart Rate, GSR, Pupil, Respiration Rate) for SNR = 12 dB. The Adjusted Rand Index heatmap shows moderate agreement between Heart Rate and Respiration Rate, and lower agreement between other pairs. The block plot displays individual subject cluster memberships.

13.8.2 Correlation between Behaviour Measures and Physiological Measurements

Following the clustering analysis, we conducted correlation analyses to further explore the relationships between performance (accuracy), subjective effort, subjective difficulty, and physiological responses. These analyses were based on data extracted from the listening period, defined as the time window between sentence onset and the start of the retention phase (i.e., sentence end).

For each trial, we computed the change in physiological signals-specifically GSR, heart rate, pupil diameter, and respiration rate-over this interval. These change scores were then correlated separately with performance and self-report measures across participants.

We selected Pearson or Spearman correlation based on both normality and linearity. Pearson's was used when variables were normally distributed and linearly related; otherwise, Spearman's rank correlation was applied to account for non-normal or monotonic relationships. Although we examined all combinations of physiological and behavioural measures, only statistically significant results are reported and discussed below.

Pupil diameter showed marginal significance, but only under specific task conditions. As presented in Figure 13.70, there was a trend-level positive association between Performance / accuracy and task-evoked pupil dilation at SNR −11 (Spearman

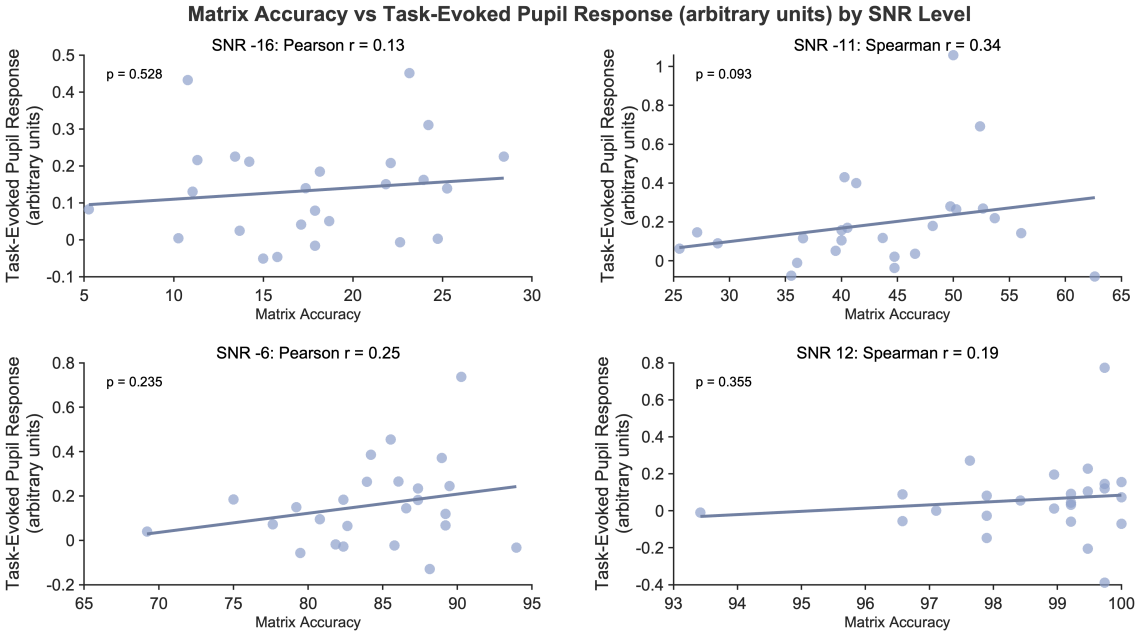


Figure. 13.70. Correlation between accuracy vs task-evoked pupil response change across SNR levels

This figure presents the correlation between accuracy and Task-Evoked Pupil Response, separately for each SNR level. Scatter plots include linear trend lines for visual reference.

Although none of the correlations reached statistical significance, the SNR -11 condition showed a marginal positive association ($\rho = .34$, $p = .093$), suggesting a potential link between task accuracy and pupil dilation at intermediate noise levels. Other SNR conditions (e.g., -6, -16, and $+12$ dB) showed weaker or negligible associations.

These findings may reflect subtle modulations of cognitive effort or arousal linked to intelligibility, with pupil responses being more sensitive to moderate listening demands.

correlation, $\rho = .34$, $p = .093$), possibly reflecting increased cognitive effort in moderately challenging listening environments. However, this relationship did not hold across other SNR levels.

In comparison, heart rate, GSR and respiration rate did not show any significant correlations with behavioural or subjective measures. One possible explanation for this difference is that these signals may have relatively slower response dynamics and smaller amplitude changes over short task epochs compared to pupil dilation. Their temporal resolution may limit their sensitivity to the transient changes in cognitive or emotional load that occur during brief listening periods.

13.9 Summary of Key Findings

This study employed the speech-in-noise test (OLSA; (Hey et al., 2014; Neumann et al., 2012)) and systematically varied the SNR levels to manipulate task difficulty. Data were collected across multiple methods, including subjective reports, task performance, and

physiological responses. Below is a summary of the key findings for each type of measurement:

Behavioural Results Participants' accuracy in identifying words significantly improved as the signal-to-noise ratio (SNR) increased, meaning they performed better when the speech was clearer. In comparison, their ratings of how much effort they invested and how difficult they found the task significantly decreased as the SNR improved. Interestingly, at the hardest level (-16 dB SNR), participants rated the task as more difficult than effortful, perhaps suggesting they actually reduced their effort when knowing the task was very challenging.

Pupillometry (Pupil Diameter) Pupil diameter changed reliably during the task, typically constricting when words started and dilating during the listening phase, with pupil diameter being larger in noisier conditions. While overall pupil size didn't significantly differ across SNRs at key moments, some specific comparisons showed differences, especially during active listening and retention periods between low and mid/high SNRs. Pupil responses showed some consistency within individuals between sessions, mainly at -6 dB and 12 dB SNR.

To check if individuals had consistent pupil response patterns across the two experimental sessions, a permutation test was used. This revealed significant within-subject consistency at the -6 dB and 12 dB SNR levels, but not at -11 dB.

Clustering analysis grouped participants into two distinct pupil response types at each SNR level. However, these physiological groupings didn't align with differences in task performance or subjective ratings. Furthermore, most participants didn't consistently stay in the same cluster across different noise levels, indicating these pupil response patterns are likely influenced by the immediate task difficulty rather than being stable individual traits.

Galvanic Skin Response (GSR) GSR, reflecting sympathetic nervous system activity, typically rose significantly from the start of the words to the start of the memory retention period. The peak GSR measured around the start of retention was significantly affected by the noise level (SNR), suggesting this is a sensitive indicator of the listening effort built up during the sentence. Participants showed significantly consistent GSR patterns across the two experimental sessions, particularly in the more difficult noise conditions (-16 dB and -11 dB).

Similar to pupillometry, clustering analysis identified two groups with distinct GSR patterns at each SNR level. These physiological groupings, however, same as Pupillometry, did not correspond to significant differences in accuracy or subjective

ratings. Cluster membership agreement across different SNRs was generally low to moderate.

Respiration Rate Respiration patterns changed during the task, with rate generally increasing from the start of the words through the retention period. While the rate at specific moments didn't show an overall significant difference across noise levels, the change in respiration rate during the retention phase was significantly modulated by SNR, with clearer speech prompting larger increases. Individual consistency in respiration patterns across sessions was only significant at moderate to clear noise levels (-11 dB and 12 dB SNR).

Again, clustering revealed two distinct respiratory patterns at each SNR level. Consistent with other measures, these physiological clusters did not align with significant differences in behaviour or subjective reports. Agreement on cluster membership across SNRs was very low, suggesting respiration patterns adapted dynamically to task difficulty.

Heart Rate Heart rate followed a distinct pattern during the task: increasing sharply after word onset, dipping during the retention period, and then rising again towards the end. While the absolute heart rate at key moments didn't differ significantly across SNRs, the amount of heart rate recovery during the retention phase (the rise from the dip to the end peak) was significantly affected by the noise level. Participants showed very strong consistency in their heart rate patterns across the two experimental sessions for all noise levels.

Clustering analysis identified two distinct heart rate profiles at each SNR. Yet again, these physiological groupings did not correspond to significant differences in accuracy or subjective ratings. Cluster membership showed only limited to moderate consistency across the different noise levels, suggesting heart rate patterns were influenced by the listening condition rather than being purely stable traits.

EEG (Alpha Power) Brain activity measured by EEG, specifically alpha wave power, showed clear changes related to the task phase and noise level. Alpha power was generally suppressed (lower) during the demanding listening and retention phases, especially in difficult noise conditions, reflecting cognitive effort. After participants responded, alpha power showed a rebound (increase), which was stronger and more sustained in easier conditions, likely indicating cognitive disengagement or release from effort. Compared to a previous, simpler study (Study 1), the alpha rebound in this more complex task was more prolonged. However, due to high variability in individual responses, and the noise induced through specific data collection methods, tests for within-subject consistency were not significant, and clustering analysis was not performed.

Relationship Between Measures When comparing the clustering results from different physiological measures (e.g., heart rate clusters vs. GSR clusters), there was generally very low agreement at all noise levels. This indicates that the different physiological systems (heart, skin conductance, pupils, breathing) likely reflect different aspects of the listening effort, rather than a single, unified state.

Correlation analyses examined the direct link between physiological changes during listening and the behavioural/subjective outcomes. GSR emerged as the most consistent indicator, showing a significant positive correlation with task accuracy and a significant negative correlation with subjective difficulty across all noise levels. Pupil diameter change showed a marginal link to accuracy at one noise level. Heart rate and respiration rate changes did not significantly correlate with performance or subjective ratings in this analysis.

13.10 Discussion

The results of Study 2 provide valuable insights into the physiological and subjective dimensions of listening effort under varying noise conditions. The following discussion interprets the findings through central research questions by integrating perspectives from the broader literature on listening effort and cognitive processing.

When examining consistency in individual responses across the two testing sessions spaced a week apart, permutation tests revealed a varied outcome. While heart rate patterns were highly consistent within individuals across all conditions (Mackersie et al., 2015), the consistency of GSR, pupillometry, and respiration rate was often dependent on the specific noise level, achieving statistical significance only under certain SNRs.

EEG alpha patterns, however, did not show significant consistency across sessions; the use of a custom-built single-channel EEG setup, noted in the experiment design as yielding more noisy data. Nonetheless, the overall variable consistency across measures suggests a complex interplay between stable individual physiological tendencies and adaptive responses to the task's demands.

In exploring individual differences, Study 2 revealed substantial variability in how participants responded physiologically, as shown by the spread of data and the distinct groupings identified through clustering analyses. This finding is consistent with existing literature that highlights the importance of listener-specific factors (Koelewijn et al., 2012; Zekveld et al., 2011), reinforcing the argument that such individual variability should be taken into account in the study of listening effort (McGarrigle et al., 2014; Peelle, 2018).

Further exploring these differences, clustering analyses consistently grouped participants into two distinct physiological response patterns for pupillometry, GSR, respiration, and heart rate within each noise condition. This indicates that identifiable subgroup response styles exist (Peelle, 2018).

However, these cluster memberships across different noise conditions were not stable; participants frequently shifted groups as the SNR changed, confirmed by low agreement scores (ARI). This strongly suggests these particular physiological patterns reflect dynamically regulated, task-difficulty-dependent responses influenced by immediate task difficulty rather than representing fixed, trait-like individual characteristics.

In contrast to Study 1, Study 2 offered further insights into listening effort by examining additional physiological signals, specifically focusing on respiration and cardiac activity. Respiration rate increased during the demanding retention phase, and importantly, the magnitude of change in respiration rate during retention was sensitive to the noise level. This aligns with literature suggesting respiration reflects cognitive load and effort in auditory tasks (Auer et al., 2021; Bernardi et al., 2014; Richter et al., 2016a).

Heart Rate (HR), derived from ECG, exhibited clear phasic changes linked to task events (onset peak, retention dip). Study 2's finding that HR recovery during the retention phase was significantly modulated by SNR adds nuance to previous work showing overall HR increases with difficulty (Hicks & Tharpe, 2002; Mackersie & Cones, 2011; Mackersie et al., 2015), suggesting that the dynamics of recovery are also a sensitive aspect of the cardiovascular response to effort.

Study 2 reinforced well-established findings regarding the influence of different SNR levels: as SNR increased (i.e., noise decreased), participants demonstrated better accuracy and reported lower levels of perceived difficulty and effort, in line with prior research (Alhanbali et al., 2018; McGarrigle et al., 2017, 2021; Reinten et al., 2021; Wu et al., 2016).

The physiological measures also demonstrated sensitivity to different SNR levels, although often in specific metrics or task phases. SNR levels showed effects on pupil dilation dynamics (Kuchinsky et al., 2013; Zekveld et al., 2010), peak amplitude of the GSR response (Mackersie & Cones, 2011), change in respiration rate during retention (Richter et al., 2016a), and further, heart rate recovery (Mackersie et al., 2015), and the suppression and rebound characteristics of EEG alpha activity (McMahon et al., 2016; Obleser et al., 2012), indicating their different levels and aspects to distinctive listening demand.

Looking at how behavioural outcomes related to physiological signals, Study 2 revealed a complex and layered pattern of results. While distinct physiological response patterns (clusters) were identified, membership in these clusters generally did not predict significant differences in task accuracy or subjective ratings. This apparent disconnection suggests that different physiological states might not always show as overt performance differences, aligning with literature noting only moderate correlations between different types of effort measures (Alhanbali et al., 2018; Ohlenforst et al., 2017).

The lack of strong correlations for changes in HR, GSR and RR with behaviour in this specific analysis, despite their documented sensitivity (Bernardi et al., 2014; Hicks & Tharpe, 2002; Mackersie & Cones, 2011; Mackersie et al., 2015; Richter et al., 2016a), might stem from the analysis window, or specific task demands. Similarly, pupillometry also showed a marginal association (Wendt et al., 2018). It is important to consider potential limitations of Study 2 when interpreting these findings.

The final sample size for analysis was reduced due to data quality requirements across multiple measures (approx. 26 datasets per measure), which could limit the power to detect smaller effects or subgroup differences. Furthermore, the findings are based on normal-hearing participants performing a specific speech-in-noise task involving recognising sentence heard from a visual matrix; generalisability to individuals with hearing impairment or different listening tasks requires more cautious understanding. Finally, inherent limitations exist for each physiological measure, such as EEG's susceptibility to artefacts, which might influence results.

13.11 Conclusion

In conclusion, Study 2 investigated listening effort using a multi-modal approach, combining behavioural, subjective, and diverse physiological measures (pupillometry, GSR, ECG, respiration, EEG) during a speech-in-noise task across different difficulty levels. The results confirmed that behavioural accuracy and subjective ratings of effort and difficulty reliably tracked signal-to-noise ratio. Physiological measures also showed sensitivity to task demands and noise levels, revealing complex, dynamic response patterns tied to specific task phases.

While significant individual variability was evident, attempts to group individuals based on physiological response patterns (clustering) showed these patterns were often task-difficulty-dependent rather than stable traits across conditions, and generally did not predict behavioural outcomes. Overall, Study 2 highlights the intricate nature of physiological responses to listening effort and underscores the value of different measures in capturing its various dimensions.

Part IV

General Discussion and Conclusion

Chapter 14

General Discussion

14.1 Introduction

Listening effort remains a central challenge in auditory and cognitive research, particularly in the context of increasingly noisy environments and an ageing population. Although the concept has been widely explored, the mechanisms underlying listening effort - and the ways in which it is physiologically expressed across individuals - are still not fully understood (McGarrigle et al., 2014; Pichora-Fuller et al., 2016c; Rönnberg et al., 2013).

This research addresses that gap through two studies, combining behavioural, subjective, and physiological data to explore effort across different populations and task demands.

A key strength of this work is its two-level analysis: traditional summary measures (e.g., peak or mean response) are used alongside full time-course analyses. This approach captures how effort-related responses unfold dynamically - their shape, timing, and evolution across task phases - which are often missed by single-point measures (Koelewijn et al., 2018; Winn et al., 2016).

Study 1 focused on older adults with varying hearing levels. It asked whether individuals show stable physiological response patterns across repeated sessions (individual consistency) and whether participants could be clustered based on those patterns (individual differences). It also explored whether these clusters relate to task performance, perceived effort, and how different physiological signals (EEG, GSR, pupil size) relate to one another within individuals.

Study 2 extended this by using a more complex and ecologically valid sentence-in-noise task with fixed SNRs. It introduced additional physiological signals (heart rate and respiration) and tested whether individual response patterns remained consistent in this more demanding context. It also examined how task difficulty (via SNR manipulation)

modulates physiological responses, and whether these changes align with self-reported effort and behavioural accuracy.

Together, this research aim to deepen our understanding of listening effort as a dynamic, context-sensitive, and individually mediated process. The following sections examine how the findings from each study contribute to this broader goal.

14.2 Summary of findings between Two studies

A primary outcome emerging from both Study 1 and Study 2 is the obvious individual variability characterising responses to listening effort tasks. Despite different participant demographics (older, hearing-impaired vs. younger, normal-hearing) and listening tasks and experiment design (adaptive vs. fixed SNR, digit vs. sentence stimuli), both studies observed considerable variation in subjective reports, behavioural accuracy, and physiological response patterns. This finding across diverse conditions reinforces the importance of considering listener-specific factors, as highlighted in the literature (Koelewijn et al., 2012; Peelle, 2018; Zekveld et al., 2011), and suggests that group averages may conceal crucial aspects of the listening effort experience.

Individual Consistency Across Studies Participants frequently demonstrated stable, individual-specific physiological response patterns. In Study 1, EEG alpha power, GSR, and pupil diameter exhibited high test-retest reliability, confirmed by permutation tests. Study 2 showed similar within-subject consistency in GSR, heart rate, and respiration at certain SNR levels, despite greater task complexity. However, some measures, particularly EEG and pupillometry, showed condition-dependent reliability, potentially influenced by data quality and task design.

These findings suggest that physiological response patterns to listening tasks can remain consistent within individuals over time, yet are also modulated by factors such as task complexity, SNR structure, and measurement itself. Importantly, this consistency does not predict subjective effort or behavioural performance, indicating that physiological engagement may reflect deeper or unconscious cognitive mechanisms.

Individual Differences and the Physiology–Behaviour Gap Both studies identified distinct physiological response clusters through time-course analysis, typically revealing two subgroups per modality. However, these clusters showed limited alignment with behavioural accuracy or self-reported effort. This repeated dissociation between physiology and outcome measures reinforces a well-documented gap in the literature (Alhanbali et al., 2019; Mackersie et al., 2015; Ohlenforst et al., 2017), and supports that

listening effort cannot be directly inferred from one single measurement (Wendt et al., 2018).

Notably, Study 2 revealed that physiological clustering was sensitive to SNR, with individuals shifting cluster membership across conditions. This highlights that physiological strategies may not be fixed traits, but context-dependent states that adapt to task demands. Moreover, even at the most challenging SNRs (e.g., –16 dB), distinct response patterns persisted, suggesting both trait-like stability and state-driven modulation.

The comparison of study designs underscores the role of task structure in shaping effort responses. The sentence-in-noise task in Study 2 elicited a more prolonged EEG alpha rebound than the digit task in Study 1, possibly reflecting disengagement following high cognitive load.

Relationships Between Physiological Measures A recurring theme across both studies is the lack of strong agreement between physiological modalities. Clusters derived from EEG, GSR, pupil, heart rate, and respiration often differed, and cross-modal alignment was generally low. This implies that each system may be attuned to different components of listening effort - such as cognitive load, emotional arousal, attentional engagement, or autonomic regulation - and that no single measure can act as a universal proxy for effort (Gagné et al., 2017; Pichora-Fuller et al., 2016c).

Using full time-course data was particularly valuable in uncovering these patterns. Static measures often failed to capture subtle but important differences in response dynamics - such as earlier versus later peaks, sustained versus transient activity - which were revealed through time-course clustering. These different patterns may reflect diverse cognitive strategies, such as anticipatory preparation versus reactive coping, which would remain invisible using peak-only analyses (Beatty & Lucero-Wagoner, 2000; Richter et al., 2016b; Zekveld & Kramer, 2014).

Conclusion Taken together, these findings demonstrate that listening effort is not a unidimensional construct, but a complex, dynamic phenomenon expressed through multiple physiological systems. Individuals differ not only in the intensity of their responses, but also in the temporal dynamics and system-specific patterns through which those responses manifest. Integrating conventional metrics with full time-course analysis was essential in revealing these insights and suggests a move toward more nuanced, person-specific models of listening effort in future research.

14.3 Theoretical Integration and Alignment with Previous Research

This thesis contributes to the evolving theoretical landscape of listening effort by offering empirical findings that both support and challenge current models. Two frameworks are particularly relevant: the *Framework for Understanding Effortful Listening* (FUEL) (Pichora-Fuller et al., 2016c) and the *Ease of Language Understanding* (ELU) model (Rönnberg et al., 2013). Together with empirical studies on physiological and behavioural effort responses, these frameworks provide a foundation for interpreting the complex dynamics uncovered in both studies.

Support for Dynamic Resource Models The FUEL model conceptualises listening effort as a dynamic allocation of cognitive resources in response to task demands. It highlights the role of working memory, attention, and motivation, and posits that effort is modulated both by environmental challenges (e.g., SNR) and listener-specific factors (e.g., hearing loss, cognitive capacity) (Kahneman, 1973; Pichora-Fuller et al., 2016c).

This dynamic perspective is strongly supported by Study 2, which demonstrated that physiological response profiles shifted across different SNR levels. Participants changed cluster membership depending on noise level, and within-subject consistency in measures like GSR and respiration was condition-dependent (Alhanbali et al., 2019). These findings reinforce the FUEL models claim that effort is not a fixed trait, but a context-sensitive state shaped by ongoing demands.

FUEL Model and Individual Strategies The FUEL model emphasises mismatch-driven listening effort. Increased cognitive load arises when speech input does not match stored phonological representations, prompting resource-intensive processing (Rönnberg et al., 2013). The observed individual variability across both studies supports this idea.

Despite adaptive (Study 1) or fixed (Study 2) SNR designs, participants showed distinct physiological strategies. Cluster analyses revealed consistent subgroups with different response dynamics, such as early peaks or prolonged activation (McGarrigle et al., 2014). These may reflect alternative compensatory strategies that align with the FUEL framework, particularly under challenging or ambiguous conditions.

The Physiology–Behaviour Gap Both studies consistently found that physiological patterns did not align with behavioural accuracy or subjective ratings. Participants grouped into physiological clusters showed no significant differences in performance or self-reported effort (Alhanbali et al., 2019; Mackersie et al., 2015; Ohlenforst et al., 2017).

This recurring dissociation mirrors earlier studies showing weak or inconsistent links between subjective, behavioural, and physiological indices of effort. It suggests that physiological activation does not map simplistically onto task outcomes, possibly due to different cognitive strategies.

For instance, individuals with higher cognitive capacity may engage more resources, reflected in heightened physiological activity, yet perform well due to effective compensation (Wendt et al., 2018). Others may disengage under high difficulty, showing low physiological activity and poorer performance (Wu et al., 2016). Still others may prioritise speed over accuracy (Houben et al., 2013), masking underlying effort in their behavioural data.

Varying reliance on top-down mechanisms, as described in the FUEL model, could also lead to similar behavioural outcomes via different internal processing routes and costs. Moreover, Study 2 showed that physiological patterns were task-difficulty-dependent, shifting across SNRs (Kahneman, 1973; Pichora-Fuller et al., 2016c). This indicates effort expression is not fixed, but contextually adaptive.

These findings collectively underscore that physiological effort signals and behavioural or subjective outcomes do not follow a one-to-one relationship. However, rather than viewing this as a limitation, it points toward a more nuanced conceptualisation of listening effort — one in which subjective experience and autonomic arousal reflect distinct but complementary systems (Alhanbali et al., 2019; McGarrigle et al., 2014).

This has practical and theoretical implications. For example, an individual may show elevated physiological reactivity without reporting much difficulty, suggesting unconscious or automatic engagement (Critchley & Garfinkel, 2017). Conversely, someone might feel overwhelmed or mentally fatigued but display a blunted physiological profile — possibly due to disengagement, emotional coping, or individual differences in interoceptive awareness (Richter et al., 2016a).

Recognising this dissociation helps avoid overly simplistic interpretations of either measure alone. Instead, it encourages a multi-dimensional approach to effort that acknowledges how cognitive, affective, and physiological systems may respond differently to the same task (Pichora-Fuller et al., 2016c).

In applied contexts, such as clinical audiology or assistive technology design, this reinforces the value of using physiological indicators not to replace self-report or performance, but to capture otherwise invisible forms of cognitive strain — particularly in populations where verbal reporting may be unreliable or underdeveloped (e.g., children, non-native speakers, or cognitively impaired individuals).

Taken together, the results support the view that effort is not a singular experience, but an adaptive, context-sensitive process shaped by internal capacity, task demands, and conscious awareness (Kahneman, 1973; Pichora-Fuller et al., 2016c). Physiological

profiling may therefore prove valuable not in predicting outcomes directly, but in tailoring support strategies and adaptive systems to better align with an individual's internal state.

Limitations of Subjective and Behavioural Measures Another contributor to the physiology–behaviour gap may be the limitations of global subjective scales. NASA-TLX (Hart & Staveland, 1988) and single-item effort ratings may not be sensitive enough to reflect subtle or dynamic effort processes (McGarrigle et al., 2014). Similarly, behavioural accuracy provides only an endpoint, not the cognitive cost incurred during task performance (Sarampalis et al., 2009). This calls into question the validity of relying on any single outcome to infer effort.

Context-Dependency and Dynamic Adaptation Study 2 offered strong evidence for the context-sensitive nature of effort responses. Cluster membership changed with SNR, supporting flexible resource allocation models such as those proposed by Kahneman (Kahneman, 1973) and FUEL (Pichora-Fuller et al., 2016c). Our findings suggest that physiological responses are shaped by real-time demands rather than being fixed traits. This supports the view that strategies are adapted based on perceived challenge, motivation, or cognitive reserve (Mattys & Wiget, 2009; McGarrigle et al., 2014).

This adds complexity to interpretation, as the same physiological pattern may reflect different effort states depending on task context. It also contrasts with trait-focused models (e.g., personality-related baseline arousal (Geen, 1984; Mackersie et al., 2015)), reinforcing the need for adaptive frameworks.

Divergence Between Physiological Modalities Study 1 found limited agreement between EEG-, GSR-, and pupil-based clusters. Study 2 showed similar modality-specific clustering for pupil, GSR, respiration, and heart rate, reinforcing their independence. This aligns with prior work suggesting that these signals reflect different physiological processes: pupil dilation with central arousal (Zekveld et al., 2010), GSR with sympathetic activation (Mackersie et al., 2015), and cardiorespiratory patterns with broader autonomic regulation (Grassmann et al., 2016). While interrelated via shared control systems, these modalities are not redundant. Each provides complementary information, reinforcing the value of multi-modal measurement strategies (Alhanbali et al., 2019; Kramer et al., 2016).

Conclusion In summary, this research reveals the complexity of listening effort. It provides empirical evidence that effort is dynamic, system-specific, and individually mediated. The integration of full time-course analysis with traditional metrics helped uncover nuanced physiological signatures that static measures may overlook. These

findings argue for moving beyond one-size-fits-all models toward personalised, context-aware approaches to listening effort.

14.4 Implications and Future Directions

The findings presented in this thesis offer several practical, methodological, and conceptual implications for research on listening effort.

Practical and Clinical Implications The identification of consistent, individual-specific physiological response patterns suggests that personalised strategies for managing listening effort may be warranted. While the clustering analyses revealed consistent individual physiological response patterns, these did not reliably predict subjective effort, perceived difficulty, or task accuracy. This apparent disconnect reflects the broader physiology–behaviour gap noted in prior literature, suggesting that autonomic and cognitive markers of effort may operate somewhat independently of subjective awareness.

Rather than predicting outcomes directly, these physiological profiles may instead reflect individualised styles of task engagement — for example, varying degrees of autonomic regulation, emotional reactivity, or sustained attention. One participant may show high physiological arousal without reporting high effort, while another may report high effort despite relatively flat physiological responses.

This variability suggests that individual profiles could still hold value in clinical and applied contexts, not for universal prediction, but for tailoring interventions. For instance, recognising a patient's typical physiological "signature" could inform more personalised auditory training regimes or adaptive hearing technologies that respond to the individual's real-time engagement style rather than assuming a one-size-fits-all model.

Methodological Implications A major methodological contribution of this work is the demonstration that full time-course analysis provides critical insights that peak or mean metrics often miss. Researchers relying solely on summary values may overlook important aspects of how effort unfolds across a task.

The integration of dynamic, system-specific time-series data into listening effort research provides a richer and more accurate representation of physiological engagement. This supports a move toward more sophisticated modelling and analysis techniques, including machine learning classifiers that can account for dynamic features and inter-individual variability.

Theoretical Development and Modelling These findings also inform theoretical modelling of listening effort. The evidence for context-sensitivity, system-specific responses, and the persistent physiology-behaviour dissociation suggests that effort is not a unitary construct, but an emergent outcome of multiple interacting systems. Future models should explicitly incorporate temporal dynamics, cross-system divergence, and person-specific adaptation. This aligns with recent proposals advocating for systems-level, multidimensional models of listening effort (Gagné et al., 2017; Herrmann & Johnsrude, 2020).

Directions for Future Research Future research should explore how stable these physiological profiles remain over longer timescales and across more varied tasks or real-world environments. Larger samples with greater demographic diversity, as well as the inclusion of additional cognitive and psychological measures, would help clarify the sources of individual variability observed here. Finally, validating dynamic, multimodal physiological profiles as predictive tools for listening-related fatigue, performance decline, or benefit from intervention represents a promising translational direction.

In addition, future work could integrate all physiological modalities into a single multivariate framework to capture how different systems jointly respond during challenging listening. While the present clustering analyses examined each physiological system separately, combining measures such as EEG, GSR, pupillometry, and respiration would enable the identification of multimodal physiological profiles that may better reflect the coordinated dynamics of listening effort.

Unsupervised techniques such as multivariate clustering or dimensionality reduction could reveal latent subgroups of listeners who share characteristic cross-system response patterns – for example, individuals showing synchronised increases across measures versus those who engage predominantly through a single modality. Such an approach would allow researchers to move beyond system-specific interpretations and consider listening effort as a distributed, multi-system process. Importantly, establishing whether these multimodal profiles predict real-world outcomes, such as susceptibility to listening-related fatigue or benefit from hearing support technologies, represents a valuable direction for future translational research.

Overall, this thesis highlights the importance of adopting a dynamic, individualised, and multi-system view of listening effort, both in research and in practice.

14.5 Limitations

While the studies presented in this thesis offer novel insights into physiological correlates of listening effort, several limitations should be considered when interpreting the

findings.

Multi-modal Complexity and Signal Noise One potential limitation of the current study arises from the simultaneous recording of multiple physiological signals (pupillometry, GSR, ECG, respiration, and EEG) during each trial. While this multi-system approach allows for a rich and comprehensive perspective on listening effort, it also introduces a substantial technical challenge: the risk of increased noise or artefacts due to overlapping hardware, signal interference, and participant movement constraints.

For example, single-channel EEG recordings were found to be particularly susceptible to noise, potentially due to compromises made in cap design to accommodate concurrent ECG and respiration monitoring. In addition, electrode impedance was checked only during the initial system setup for the entire experiment, rather than before each participant session, which may have contributed to variability in the single-channel recordings.

Similarly, while the Pupil Core glasses offer portability and participant comfort, their data quality may have been affected by subtle shifts in head position or interaction with ambient light and screen interface, especially during longer sessions.

This complexity highlights an important trade-off in multi-modal psychophysiological research: the breadth of insight gained may come at the cost of reduced data quality or interpretability in individual channels. Future studies might consider optimising for fewer modalities per session or testing the stability of each measure in isolation first, particularly if signal sensitivity is critical for hypothesis testing.

Measurement Constraints and Signal Quality One limitation of the EEG recordings in this study relates to signal quality monitoring. Although electrode impedance was checked before the full experiment, and appeared within an acceptable range at the beginning of each experiment recording, continuous impedance monitoring during recording was not available with the custom-built EEG system used.

As a result, variations in electrode contact quality across participants or gradual changes over time could not be systematically tracked. It is therefore possible that some of the noise observed in the EEG data may be partly attributed to electrode conductance during recording, particularly in cases where the signal appeared noisier despite initial impedance checks. Future studies would benefit from using EEG systems that allow for real-time impedance monitoring to ensure more consistent data quality across participants.

Sample Size and Generalisability Both studies included relatively small and demographically homogeneous samples, particularly in terms of age, hearing status, and cultural background. This limits the generalisability of the findings to broader populations. Future studies with more diverse participant group are needed to confirm whether the observed individual response profiles and cross-modal dissociations hold across different demographic and clinical groups.

Task Design Differences and Metric Compatibility The two studies used different tasks (digits vs. sentences), SNR structures (adaptive vs. fixed), and subjective scales (NASA-TLX vs. single-item ratings), which may complicate direct comparisons. While these differences reflect ecologically valid variations in task complexity and listening contexts, they also make it harder to isolate the effect of individual task characteristics on effort responses.

Subjective Ratings and Interpretability Subjective effort measures, particularly the NASA-TLX and single-item scales, may not fully capture the nuanced or dynamic experience of effort during listening. Although Study 2 improved temporal resolution by collecting ratings after each block, these measures still depend on post-task reflection and may be influenced by memory, expectation, or individual interpretation. More continuous or other behaviour measures such as reaction time, could further enhance sensitivity and offer a more direct complement to the physiological data.

Analytical Scope and Clustering Assumptions While clustering was effective for uncovering individual response patterns, it is inherently sensitive to parameter choices (e.g., number of clusters, distance metrics). Additionally, clustering was performed within each modality, limiting insight into potential cross-modal synergy. Future work could apply multi-view clustering or dimensionality reduction techniques to jointly model multimodal data streams.

Despite these limitations, the combined use of time-course analysis, multimodal measures, and within-subject designs represents a substantial methodological advance. The limitations noted here do not undermine the central claims of this thesis, but rather provide a roadmap for refinement and future exploration.

Chapter 15

Conclusion

This thesis set out to investigate how listening effort is expressed physiologically, with a focus on individual variability, dynamic response patterns, and multimodal measurement. Across two studies involving distinct populations and task designs, the research demonstrated that physiological responses to effortful listening are both consistent within individuals and highly variable across them. These responses were shaped not only by task difficulty but also by the unique physiological dynamics of each listener.

Study 1 demonstrated stable and differentiable patterns of dynamic physiological responses across EEG, GSR, and pupillometry in a hearing-impaired population. These patterns were diversed within each trial, capturing how responses evolved across time and task phases. The ability to cluster these time-course profiles into distinct physiological reactions provided valuable insights into individualised effort strategies, while Study 2 extended these findings to normal-hearing adults and included additional measures such as heart rate and respiration.

Both studies showed that while these dynamic physiological clusters reflected consistent and interpretable patterns of reactivity, they did not align neatly with behavioural performance or subjective ratings. This recurring physiology-behaviour dissociation, along with the divergence between physiological modalities, underscores the complexity of the listening effort construct, which is widely noted in prior literature, suggesting that autonomic and cognitive markers of effort may operate somewhat independently of subjective awareness.

Importantly, this research showed that listening effort is not a fixed quantity but a context-sensitive process that evolves over time. By using full time-course analyses rather than relying solely on static metrics, the studies revealed nuanced patterns in the shape, timing, and regulation of physiological responses. These dynamic features provided a richer view of how effort is mobilised and managed by listeners in real time.

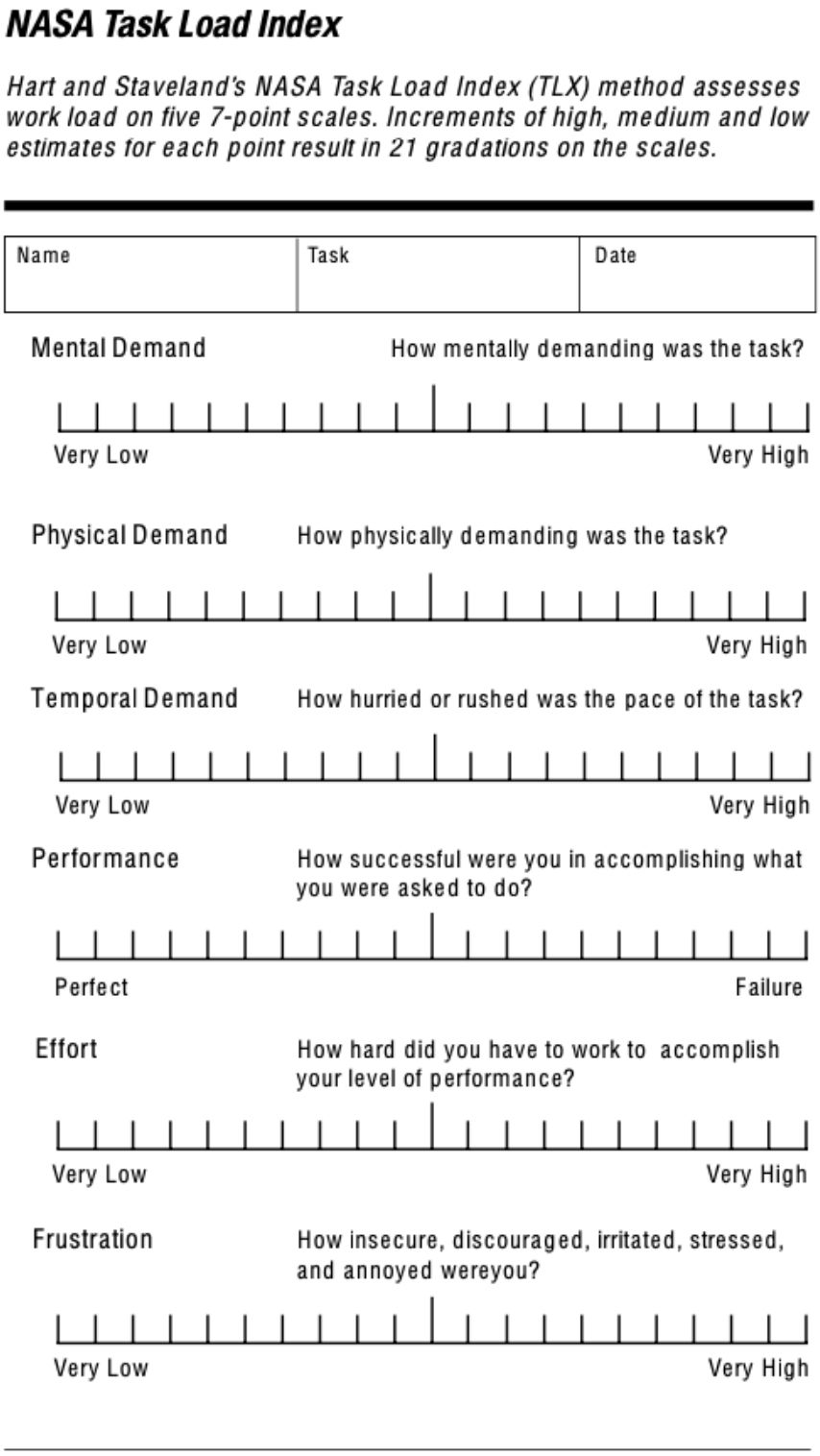
The findings support and expand theoretical models such as the FUEL and ELU frameworks, reinforcing the idea that listening effort involves flexible resource allocation in response to task demands and perceptual mismatch. At the same time, the results challenge simplistic assumptions about direct correlations between effort, performance, and subjective experience, and call for more sophisticated, individualised models.

Beyond theoretical contributions, this thesis offers methodological and applied value. It demonstrates the importance of integrating multiple physiological systems and dynamic analysis techniques in effort research. It also suggests potential for person-specific profiles to inform clinical practice, auditory training, and the design of adaptive technologies.

In sum, this research adds knowledge in understanding of listening effort as a multi-dimensional, dynamic, and individualised phenomenon. It highlights the value of looking beyond group means and singular indicators, and toward a richer, more person-centred account of how effort is experienced and expressed. Seeing listening effort as a dynamic and individual process creates opportunities for future research and practical use - such as designing better listening support, tools, or environments that work well for different types of listeners.

Appendix A

Questionnaires



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?
Very Low — Very High

Physical Demand — How physically demanding was the task?
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?
Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?
Very Low — Very High

Figure. A.1. NASA Task Load Index (TLX) rating form used for assessing subjective workload

The NASA TLX assesses perceived workload across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants rate each item on a 21-point scale, enabling fine-grained estimation of task difficulty. This form was used in (Alhanbali et al., 2018)to measure perceived listening effort after a speech-in-noise task.

Study Title: Individual Differences in Listening Effort

Researcher: Yuki Wang

	Participant Number:		
	Date:	/	/
1	Do you feel you hear better with a specific ear?	No	Yes
2	Do you have a known hearing impairment?	No	Yes
3	Have you had any pain, tenderness, infections, discharge, surgery or bleeding from either of your ears?	No	Yes
4	Do you have tinnitus (persistent ringing or buzzing in either of your ears)?	No	Yes
5	Do you regularly use any prescribed medication that affects your hearing?	No	Yes
6	Are you aware of any skin sensitivity/conditions (eczema, dermatitis)?	No	Yes
7	Are you aware of any allergies to electrode paste?	No	Yes
8	Do you have problem using alcohol pad to clean the skin (for applying electrodes)?	No	Yes
9	Do you have any neurological, muscular or joint conditions that may affect your ability to understand speech or respond on a computer screen?	No	Yes
10	Do you have any neurological, cardiovascular or respiratory condition that may affect your physiological responses (heart beat, breathing, etc)?	No	Yes
11	Do you struggle to use computer monitors (even with glasses if needed)?	No	Yes

Thank you very much for your time!

Figure. A.2. Screening Questions before Participation (study 2)

The screening question were used in Study 2. Before participant join the experiment, they need to meet all the criteria and has no conditions stated in the questions above

Bibliography

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? a survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(sup2), S53–S71. https://doi.org/10.1080/14992020802301142

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2018). Hearing Handicap and Speech Recognition Correlate With Self-Reported Listening Effort and Fatigue. *Ear & Hearing*, *39*(3), 470–474. https://doi.org/10.1097/AUD.0000000000000515

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing*, *40*(5), 1084–1097. https://doi.org/10.1097/AUD.0000000000000697

Alhanbali, S., Munro, K. J., Dawes, P., Perugia, E., & Millman, R. E. (2021). Associations between pre-stimulus alpha power, hearing level and performance in a digits-in-noise task [Alhanbali, Sara Munro, Kevin J Dawes, Piers Perugia, Emanuele Millman, Rebecca E eng England Int J Audiol. 2021 Apr 1:1-8. doi: 10.1080/14992027.2021.1899314.]. *International Journal of Audiology*, 1–8. https://doi.org/10.1080/14992027.2021.1899314

Alhanbali, S. W. (2017). *Measuring listening effort and fatigue in adults with hearing impairment* [Doctoral dissertation, University of Manchester] [Thesis submitted for the degree of Doctor of Philosophy in the Faculty of Biology Medicine and Health, School of Health Sciences, Division of Human Communication, Development and Hearing].

Anderson, S., Parbery-Clark, A., White-Schwoch, T., & Kraus, N. (2012). Aging affects neural precision of speech encoding. *Journal of Neuroscience*, *32*(41), 14156–14164. https://doi.org/10.1523/JNEUROSCI.2176-12.2012

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language learning*, *38*(4), 561–613.

Arehart, K. H., Souza, P., Baca, R., & Kates, J. M. (2013). Working memory, age, and hearing loss: Susceptibility to hearing aid distortion. *Ear and Hearing*, *34*(3), 251–260. https://doi.org/10.1097/AUD.0b013e318271aa5e

Aron, A. R., Poldrack, R. A., & Robbins, T. W. (2011). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *15*, 85–94.

Atkinson, J. W. (1964). *An introduction to motivation*. Van Nostrand.

Auer, E., Bernstein, L. E., & Tak, S. (2021). Listening effort: Respiratory indicators. *Trends in Hearing*, *25*, 23312165211010255. https://doi.org/10.1177/23312165211010255

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Bajo, V. M., Nodal, F. R., Moore, D. R., & King, A. J. (2010). The descending corticocollicular pathway mediates learning-induced auditory plasticity. *Nature Neuroscience*, *13*(2), 253–260. https://doi.org/10.1038/nn.2466

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 142–162). Cambridge University Press.

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

Bernardi, L., Porta, C., Gabutti, A., Spicuzza, L., & Sleight, P. (2014). Modulation of respiratory rate by hypnotic suggestion. In *Hypnosis and conscious states: The cognitive neuroscience perspective* (pp. 143–158). Oxford University Press.

Bernarding, C., Strauss, D. J., Hannemann, R., & Corona-Strauss, F. I. (2017). Neural correlates of listening effort: Neurophysiological markers of cognitive resource allocation in listening tasks. *International Journal of Psychophysiology*, *118*, 120–132.

Berntson, G. G., Thomas Bigger Jr, J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., et al. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Best, V., Ahlstrom, J. B., Mason, C. R., Roverud, E., Streeter, T. M., Gallun, F. J., & Dubno, J. R. (2018). Spatial release from masking in older adults: The role of audibility. *Journal of the Acoustical Society of America*, *143*(4), 2055–2065. https://doi.org/10.1121/1.5030518

Bidelman, G. M., & Dexter, L. (2017). Blindness impairs predictions during speech processing as revealed by auditory event-related potentials. *Neuropsychologia*, *99*, 269–276. https://doi.org/10.1016/j.neuropsychologia.2017.03.031

Boccard, J., & Rudaz, S. (2013). Mass Spectrometry Metabolomic Data Handling for Biomarker Discovery. In *Proteomic and Metabolomic Approaches to Biomarker Discovery* (pp. 425–445). Elsevier. https://doi.org/10.1016/B978-0-12-394446-7.00027-3

Bregman, A. S., & McAdams, S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. *The Journal of the Acoustical Society of America*, *95*(2), 1177–1178. https://doi.org/10.1121/1.408434

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (ies) a better dependent variable than

the mean reaction time (rt) and the percentage of errors (pe)? *Psychologica Belgica*, *51*(1), 5–13. https://doi.org/10.5334/pb-51-1-5

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (Eds.). (2007). *Handbook of psychophysiology* (3rd). Cambridge University Press. https://www.cambridge.org/core/books/handbook-of-psychophysiology/EACAC4007D68C77D20B912D18C78A370

Campbell, J., & Sharma, A. (2016). Compensatory changes in cortical resource allocation in adults with hearing loss. *Frontiers in Systems Neuroscience*, *10*, 71. https://doi.org/10.3389/fnsys.2016.00071

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Chi, Y. M., Jung, T.-P., & Cauwenberghs, G. (2010). Dry-contact and noncontact biopotential electrodes: Methodological review. *IEEE reviews in biomedical engineering*, *3*, 106–119.

Choi, J. H., Lotto, A. J., Lewis, D., Hoover, B., & Stelmachowicz, P. G. (2011). Children's recognition of speech in noise using amplitude-modulated maskers: Effects of hearing loss and presentation level. *Ear and Hearing*, *32*(5), 593–599. https://doi.org/10.1097/AUD.0b013e31820f475c

Cox, R. M., & Alexander, G. C. (1995). The abbreviated profile of hearing aid benefit. *Ear and Hearing*, *16*(2), 176–186.

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current Opinion in Psychology*, *17*, 7–14. https://doi.org/10.1016/j.copsyc.2017.04.020

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Darwin, C. J. (2007). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1011–1021.

Davis, M. H., Ford, M. A., Kherif, F., & Johnsrude, I. S. (2011). Does semantic context benefit speech understanding through "top–down" processes? evidence from time-resolved sparse fmri. *Journal of Cognitive Neuroscience*, *23*(12), 3914–3932. https://doi.org/10.1162/jocn_a_00084

Debener, S., Minow, F., Gandras, K., De Vos, M., & Jaeger, M. (2012). How about taking an eeg appliance into the wild? *Psychophysiology*, *49*(11), 1617–1621. https://doi.org/10.1111/j.1469-8986.2012.01458.x

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing*, *34*(3), 261–272. https://doi.org/10.1097/AUD.0b013e31826d0ba4

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

Dillon, H. (2012). *Hearing aids* (2nd). Thieme.

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in children and adolescents with bilateral cochlear implants. *Hearing Research*, *371*, 85–95. https://doi.org/10.1016/j.heares.2018.11.007

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *Journal of Speech and Hearing Disorders*, *47*(2), 189–193. https://doi.org/10.1044/jshd.4702.189

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis. *Trends in Hearing*, *21*, 233121651774467. https://doi.org/10.1177/2331216517744675

Eckert, M. A., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2008). Age-related effects on word recognition: Reliance on cognitive control systems with structural declines in speech-responsive cortex. *Journal of the Association for Research in Otolaryngology*, *9*(2), 252–259. https://doi.org/10.1007/s10162-008-0113-3

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the cocktail party: Brain systems underlying speech intelligibility. *Journal of the Association for Research in Otolaryngology*, *10*(1), 153–163. https://doi.org/10.1007/s10162-008-0149-4

Edwards, E. J., Edwards, M. S., & Lyvers, M. (2016). Cognitive trait anxiety, situational stress, and mental effort predict shifting efficiency: Implications for attentional control theory. *Emotion*, *16*(8), 1130–1142. https://doi.org/10.1037/emo0000188

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology*, *20*(2), 156–165. https://doi.org/10.1016/j.conb.2010.02.015

Esterman, M., Rosenberg, M. D., & Noonan, S. K. (2014). Intrinsic fluctuations in sustained attention and distractor processing. *The Journal of Neuroscience*, *34*, 1724–1730.

Fallon, M., Trehub, S. E., & Schneider, B. A. (2013). Reliability and stability of the cortical response to acoustically modified speech in healthy adults. *European Journal of Neuroscience*, *38*(10), 3146–3154. https://doi.org/10.1111/ejn.12301

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fawcett, C., Nordenswan, E., Yrttiaho, S., Häikiö, T., Korja, R., Karlsson, L., Karlsson, H., & Kataja, E.-L. (2022). Individual differences in pupil dilation to others' emotional and neutral eyes with varying pupil sizes. *Applied Cognitive Psychology*, *36*(3), 657–666. https://doi.org/10.1002/acp.3951

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application. *Neuroimage*, *63*(2), 921–935.

Ferree, T. C., Luu, P., Russell, G. S., & Tucker, D. M. (2001). Scalp electrode impedance, infection risk, and eeg data quality. *Clinical Neurophysiology*, *112*(3), 536–544.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort. *Ear and Hearing*, *13*(2), 80–86. https://doi.org/10.1097/00003446-199204000-00003

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kühberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research: A critical review and user's guide* (pp. 163–184). Psychology Press.

Forte, G., Favieri, F., & Casagrande, M. (2020). Decision making and heart rate variability: A systematic review. *Applied Cognitive Psychology*, *34*(3), 486–501.

Francis, A. L., & MacPherson, M. K. (2021). Perspective taking in listening effort: Subjective and physiological responses to relational and personal listener-oriented messages. *Journal of Speech, Language, and Hearing Research*, *64*(6S), 2259–2279. https://doi.org/10.1044/2021_JSLHR-20-00538

Francis, A. L., & Love, J. (2016). Autonomic correlates of speech perception in noise. *Attention, Perception, & Psychophysics*, *78*(2), 611–621.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America*, *110*(2), 1150–1163. https://doi.org/10.1121/1.1381538

Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, *21*, 233121651668728. https://doi.org/10.1177/2331216516687287

Gallun, F. J., Diedesch, A. C., Kampel, S. D., & Jakien, K. M. (2013). Independent impacts of age and hearing loss on spatial release in a complex auditory environment. *Frontiers in Neuroscience*, *7*, 252. https://doi.org/10.3389/fnins.2013.00252

Gatehouse, S., & Noble, W. (2004). The speech, spatial and qualities of hearing scale (ssq). *International Journal of Audiology*, *43*(2), 85–99. https://doi.org/10.1080/14992020400050014

Gates, G. A., & Mills, J. H. (2005). Presbycusis. *The Lancet*, *366*(9491), 1111–1120. https://doi.org/10.1016/S0140-6736(05)67423-5

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2014). *Cognitive neuroscience: The biology of the mind* (4th ed.). New York, NY: WW Norton.

Geen, R. G. (1984). Preferred stimulation levels in introverts and extraverts: Effects on arousal and performance. *Journal of Personality and Social Psychology*, *46*(6), 1303–1312. https://doi.org/10.1037/0022-3514.46.6.1303

Getzmann, S., Golob, E. J., & Wascher, E. (2015). Focused and divided attention in a simulated cocktail-party situation: Erp evidence from younger and older adults. *Neurobiology of Aging*, *36*(5), 1900–1911. https://doi.org/10.1016/j.neurobiolaging.2014.11.019

Good, P. I. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd). Springer. https://doi.org/10.1007/978-1-4757-3235-1

Good, P. I. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd ed.). Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-3235-1

Gordon-Salant, S., Zion, D., & Espy-Wilson, C. (2014). Recognition of time-compressed speech does not predict recognition of natural-fast speech sentences by older listeners. *The Journal of the Acoustical Society of America*, *136*(3), EL268–EL274. https://doi.org/10.1121/1.4895014

Gosselin, P. A., & Gagné, J.-P. (2017). Use of a dual-task paradigm to measure listening effort. *Canadian Journal of Speech-Language Pathology and Audiology*, *35*(1), 43–51.

Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J. M., & Van den Bergh, O. (2016). Respiratory changes in response to cognitive load: A systematic review. *Neural Plasticity*, *2016*, 1–16. https://doi.org/10.1155/2016/8146809

Graven, S. N., & Browne, J. V. (2008). Auditory development in the fetus and infant. *Newborn and Infant Nursing Reviews*, *8*(4), 187–193. https://doi.org/10.1053/j.nainr.2008.10.010

Guest, H., Munro, K. J., Prendergast, G., Howe, S., & Plack, C. J. (2018). Tinnitus with a normal audiogram: Relation to noise exposure but no evidence for cochlear synaptopathy. *Hearing Research*, *344*, 265–274. https://doi.org/10.1016/j.heares.2016.12.002

Guinan, J., John J. (2006). Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear and Hearing*, *27*(6), 589–607. https://doi.org/10.1097/01.aud.0000240507.83072.e7

Hanslmayr, S., Sauseng, P., Doppelmayr, M., Schabus, M., & Klimesch, W. (2005). Increasing individual upper alpha power by neurofeedback improves cognitive performance in human subjects. *Applied Psychophysiology and Biofeedback*, *30*(1), 1–10. https://doi.org/10.1007/s10484-005-2169-8

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neuroscience*, *7*, 83. https://doi.org/10.3389/fnint.2013.00083

Harris, J. (1969). The spanish trill. *Journal of the International Phonetic Association*, *3*(1), 13–16.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183, Vol. 52). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Hasher, L., & Zacks, R. T. (1999). Inhibitory processes in attention, memory, and language. *Inhibitory processes in attention, memory, and language*, 1–42.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Heffernan, E., Coulson, N. S., & Ferguson, M. A. (2016). Development of the social participation restrictions questionnaire (sparq) through consultation with adults with hearing loss and health professionals. *International Journal of Audiology*, *55*(9), 540–549. https://doi.org/10.1080/14992027.2016.1187760

Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing Research*, *397*, 108016. https://doi.org/10.1016/j.heares.2020.108016

Hey, M., Hocke, T., Hedderich, J., & Müller-Deile, J. (2014). Investigation of a matrix sentence test in noise: Reproducibility and discrimination function in cochlear implant patients. *International Journal of Audiology*, *53*(12), 895–902.

Hickok, G., Houde, J., & Rong, F. (2009). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/j.neuron.2009.01.001

Hicks, C. B., & Tharpe, A. M. (2002). Listening Effort and Fatigue in School-Age Children With and Without Hearing Loss. *Journal of Speech, Language, and Hearing Research*, *45*(3), 573–584. https://doi.org/10.1044/1092-4388(2002/046)

Hicks, C. B., & Tharpe, A. M. (2013). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *56*(3), 1029–1039. https://doi.org/10.1044/1092-4388(2012/12-0235)

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, *92*(1), 84–89.

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

Hopkins, K., & Moore, B. C. J. (2008). The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise. *Journal of the Acoustical Society of America*, *123*(3), 1140–1155. https://doi.org/10.1121/1.2839016

Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, *34*(5), 523–534. https://doi.org/10.1097/AUD.0b013e31828003d8

Hornsby, B. W. Y., & Kipp, A. M. (2016). Subjective ratings of fatigue and vigor in adults with hearing loss are driven by perceived hearing difficulties not degree of

hearing loss. *Ear and Hearing*, *37*(1), e1–e10. https://doi.org/10.1097/AUD.0000000000000203

Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology*, *52*(11), 753–761. https://doi.org/10.3109/14992027.2013.832415

Howard, D. M., & Angus, J. A. S. (2017). *Acoustics and psychoacoustics* (5th). Routledge. https://doi.org/10.4324/9781315686424

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/bf01908075

Huettel, S. A., Song, A. W., & McCarthy, G. (2014). *Functional magnetic resonance imaging* (3rd). Sinauer Associates.

Huskisson, E. C. (1974). Measurement of pain. *The Lancet*, *304*(7889), 1127–1131. https://doi.org/10.1016/S0140-6736(74)90884-8

Iliadou, V., Ptok, M., Grech, H., Pedersen, E. R., Brechmann, A., Deggouj, N., Kiese-Himmel, C., Śliwińska-Kowalska, M., Nickisch, A., Demanez, L., Veuillet, E., Thai-Van, H., Sirimanna, T., Callimachou, M., Santarelli, R., Kuske, S., Barajas, J., Hedjever, M., Konukseven, O., … Bamiou, D. E. (2017). A european perspective on auditory processing disorder-current knowledge and future research focus. *Frontiers in Neurology*, *8*, 622. https://doi.org/10.3389/fneur.2017.00622

Ishida, T. (2007). Voiceless lateral in japanese: A phonetic investigation of its perception and production. *Journal of Phonetics*, *35*(4), 299–315.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jensen, N. S., Alickovic, E., Pontoppidan, N. H., MacDonald, E. N., & Andersen, T. D. (2016). Comparison of the speech, spatial and qualities of hearing scale (ssq) and the hearing handicap inventory for the elderly (hhie) in assessing the benefit of hearing aid amplification. *International Journal of Audiology*, *55*(10), 582–590. https://doi.org/10.1080/14992027.2016.1192858

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Frontiers in Human Neuroscience*, *4*, 186. https://doi.org/10.3389/fnhum.2010.00186

Jenstad, L. M., & Souza, P. E. (2003). Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility. *Journal of Speech, Language, and Hearing Research*, *46*(5), 1233–1243. https://doi.org/10.1044/1092-4388(2003/096)

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351. https://doi.org/10.1121/1.381436

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2013). *Principles of neural science* (5th ed.). McGraw-Hill Education.

Kim, H.-J., Cheon, E.-J., Bai, D.-H., Lee, Y.-H., & Koo, B.-H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, *15*(3), 235–245.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). Eeg alpha oscillations: The inhibition–timing hypothesis. *Brain research reviews*, *53*(1), 63–88.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil Dilation Uncovers Extra Listening Effort in the Presence of a Single-Talker Masker. *Ear and Hearing*, *33*(2), 291–300. https://doi.org/10.1097/AUD.0b013e3182310019

Koelewijn, T., Zekveld, A. A., Lunner, T., & Kramer, S. E. (2018). The effect of reward on listening effort as reflected by the pupil dilation response. *Hearing Research*, *367*, 106–112. https://doi.org/10.1016/j.heares.2018.07.011

Kösem, A., & van Wassenhove, V. (2018). Distinct contributions of low- and high-frequency neural oscillations to speech segmentation. *Journal of Neuroscience*, *38*(28), 6306–6317. https://doi.org/10.1523/JNEUROSCI.3125-17.2018

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

Kramer, S. E., Teunissen, C. E., & Zekveld, A. A. (2016). Physiological indicators of listening effort in noise. *Ear and Hearing*, *37*(Supplement 1), 118S–124S.

Kraus, N., & Slater, J. (2015). Music and language. In *Handbook of clinical neurology* (pp. 207–222, Vol. 129). Elsevier. https://doi.org/10.1016/B978-0-444-62630-1.00012-3

Krueger, M., Schulte, M., Zokoll, M. A., Wagener, K. C., Meis, M., Brand, T., & Holube, I. (2017). Relation Between Listening Effort and Speech Intelligibility in Noise. *American Journal of Audiology*, *26*(3S), 378–392. https://doi.org/10.1044/2017_AJA-16-0136

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss: Pupil size in older adults. *Psychophysiology*, *50*(1), 23–34. https://doi.org/10.1111/j.1469-8986.2012.01477.x

Kujawa, S. G., & Liberman, M. C. (2009). Adding insult to injury: Cochlear nerve degeneration after "temporary" noise-induced hearing loss. *Journal of Neuroscience*, *29*(45), 14077–14085. https://doi.org/10.1523/JNEUROSCI.2845-09.2009

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology*, *8*, 213.

Laine, C. M., Spitler, K. M., Mosher, C. P., & Gothard, K. M. (2009). Behavioral triggers of skin conductance responses and their neural correlates in the primate amygdala.

Journal of Neurophysiology, *101*(4), 1749–1754. https://doi.org/10.1152/jn.91110.2008

Lawrence, R. J., Wiggins, I. M., & Hartley, D. E. (2018). Functional near-infrared spectroscopy as a measure of listening effort in cochlear implant users: A proof of concept study. *Journal of Hearing Science*, *8*(3), 9–21. https://doi.org/10.17430/JHS.2018.8.3.1

Lemke, U., & Besser, J. (2016). Cognitive Load and Listening Effort: Concepts and Age-Related Considerations. *Ear and Hearing*, *37*(Suppl 1), 77S–84S. https://doi.org/10.1097/AUD.0000000000000304

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358.

Liberman, M. C., & Kujawa, S. G. (2016). Cochlear synaptopathy in acquired sensorineural hearing loss: Manifestations and mechanisms. *Hearing Research*, *349*, 138–147. https://doi.org/10.1016/j.heares.2016.03.013

Lin, F. R., Metter, E. J., O'Brien, R. J., Resnick, S. M., Zonderman, A. B., & Ferrucci, L. (2011). Hearing loss and incident dementia. *Archives of Neurology*, *68*(2), 214–220. https://doi.org/10.1001/archneurol.2010.362

Lin, F. R., Yaffe, K., Xia, J., Xue, Q.-L., Harris, T. B., Purchase-Helzner, E., Satterfield, S., Ayonayon, H. N., Ferrucci, L., & Simonsick, E. M. (2013). Hearing loss and cognitive decline in older adults. *JAMA Internal Medicine*, *173*(4), 293–299. https://doi.org/10.1001/jamainternmed.2013.1868

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., … Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The Lancet*, *396*(10248), 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

Loewenfeld, I. E. (1999). *The pupil: Anatomy, physiology, and clinical applications*. Butterworth-Heinemann.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature*, *412*(6843), 150–157.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd). MIT Press. https://mitpress.mit.edu/9780262525855/an-introduction-to-the-event-related-potential-technique/

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology*, *42*(S1), S49–S58. https://doi.org/10.3109/14992020309074624

Lunner, T., Rudner, M., & Rönnberg, J. (2016). Cognition and hearing aids. *Scandinavian Journal of Psychology*, *57*(2), 153–161. https://doi.org/10.1111/sjop.12264

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Meister, H., Brand, T., & Kollmeier, B. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. *The Journal of the Acoustical Society of America*, *127*(3), 1491–1505.

Mackersie, C. L., & Calderon-Moultrie, N. (2016). Autonomic Nervous System Reactivity During Speech Repetition Tasks: Heart Rate Variability and Skin Conductance. *Ear & Hearing*, *37*(1), 118S–125S. https://doi.org/10.1097/AUD.0000000000000305

Mackersie, C. L., & Cones, H. (2011). Subjective and Psychophysiological Indexes of Listening Effort in a Competing-Talker Task. *Journal of the American Academy of Audiology*, *22*(02), 113–122. https://doi.org/10.3766/jaaa.22.2.6

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of Hearing Loss on Heart Rate Variability and Skin Conductance Measured During Sentence Recognition in Noise. *Ear & Hearing*, *36*(1), 145–154. https://doi.org/10.1097/AUD.0000000000000091

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2018). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mattys, S. L., & Wiget, L. (2009). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *61*(2), 145–160. https://doi.org/10.1016/j.jml.2009.04.002

McCormack, G., Giles-Corti, B., Lange, A., Smith, T., Martin, K., & Pikora, T. (2004). An update of recent evidence of the relationship between objective and self-report measures of the physical environment and physical activity behaviours. *Journal of Science and Medicine in Sport*, *7*(1), 81–92. https://doi.org/10.1016/S1440-2440(04)80282-2

McDougal, D. H., & Gamlin, P. D. (2015). Autonomic control of the eye. *Comprehensive Physiology*, *5*(1), 439–473. https://doi.org/10.1002/cphy.c140014

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology*, *53*(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

McGarrigle, R., Munro, K. J., & Stewart, A. J. (2017). Pupillometry as a measure of listening effort: Steady state visually evoked potentials as a trigger for pupil dilation. *International Journal of Audiology*, *56*(3), 156–161.

McGarrigle, R., Rakusen, L., & Mattys, S. (2021). Effortful listening under the microscope: Examining relations between pupillometric and subjective markers of effort and

tiredness from listening. *Psychophysiology*, *58*(1), e13703. https://doi.org/10.1111/psyp.13703

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McMahon, C. M., Boisvert, I., Ibrahim, R. K., Galloway, J., & Stewart, H. (2016). Electrophysiological measures of listening effort in hearing loss and normal aging. *Frontiers in Aging Neuroscience*, *8*, 128.

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*, 103–107.

McShefferty, D., Whitmer, W. M., Swan, I. R. C., & Akeroyd, M. A. (2013). The effect of experience on the sensitivity and specificity of the whispered voice test: a diagnostic accuracy study. *BMJ open*, *3*(4), e002394.

Mick, P., Kawachi, I., & Lin, F. R. (2014). The association between hearing loss and social isolation in older adults. *Otology & Neurotology*, *35*(3), 543–548. https://doi.org/10.1097/MAO.0000000000000025

Miller, L. (2006). The italian r and the role of the tongue. *Journal of the International Phonetic Association*, *36*(1), 105–111.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*, 134–140.

Moore, B. C. J. (2012). *An introduction to the psychology of hearing* (6th). Brill.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in esl pronunciation instruction: An exploratory study. *System*, *34*(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

National Institute on Deafness and Other Communication Disorders (NIDCD). (2024). How do we hear? [Accessed: 28 March 2025]. https://www.nidcd.nih.gov/health/how-do-we-hear

Neumann, K., Baumeister, N., Baumann, U., Sick, U., Euler, H. A., & Weißgerber, T. (2012). Speech audiometry in quiet with the oldenburg sentence test for children. *International journal of audiology*, *51*(3), 157–163.

Ng, E. H. N., Rudner, M., Lunner, T., Pedersen, M. S., & Rönnberg, J. (2015). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *International Journal of Audiology*, *54*(1), 20–28. https://doi.org/10.3109/14992027.2014.920112

Niedermeyer, E., & Lopes da Silva, F. H. (Eds.). (2005). *Electroencephalography: Basic principles, clinical applications, and related fields* (5th). Lippincott Williams & Wilkins.

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of eeg* (2nd). Oxford university press.

Obleser, J., Herrmann, B., & Henry, M. J. (2012). Neural oscillations in speech: Don't be enslaved by the envelope. *Frontiers in Human Neuroscience*, *6*, 250.

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

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

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution eeg and erp measurements. *Clinical Neurophysiology*, *112*(4), 713–719.

Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., & Kraus, N. (2011). Musical experience and the aging auditory system: Implications for cognitive abilities and hearing speech in noise. *PLoS ONE*, *6*(5), e18082. https://doi.org/10.1371/journal.pone.0018082

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220–244.

Peelle, J. E., Troiani, V., Wingfield, A., & Grossman, M. (2010). Neural processing during older adults' comprehension of spoken sentences: Age differences in resource allocation and connectivity. *Cerebral Cortex*, *20*(4), 773–782. https://doi.org/10.1093/cercor/bhp142

Peelle, J. E. (2014). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *35*(5), 561–566. https://doi.org/10.1097/AUD.0000000000000099

Peelle, J. E. (2017). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *38*(1), 1S–11S. https://doi.org/10.1097/AUD.0000000000000389

Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear & Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Peelle, J. E., Troiani, V., Grossman, M., & Wingfield, A. (2011). Hearing loss in older adults affects neural systems supporting speech comprehension. *Journal of Neuroscience*, *31*(35), 12638–12643. https://doi.org/10.1523/JNEUROSCI.2559-11.2011

Peng, Z. E., & Wang, L. M. (2019). Listening Effort by Native and Nonnative Listeners Due to Noise, Reverberation, and Talker Foreign Accent During English Speech Perception. *Journal of Speech, Language, and Hearing Research*, *62*(4), 1068–1081. https://doi.org/10.1044/2018_JSLHR-H-17-0423

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73–89.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, *110*(11), 1842–1857. https://doi.org/10.1016/s1388-2457(99)00141-8

Pichora-Fuller, M. K. (2017). Audition and cognition: What audiologists need to know about listening. *Hearing Care for Adults*, 71–85.

Pichora-Fuller, M. K., Alain, C., & Schneider, B. A. (2016a). Older adults at the cocktail party. *Attention, Perception, & Psychophysics*, *79*(5), 1426–1428. https://doi.org/10.3758/s13414-017-1329-2

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016b). Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel). *Ear and Hearing*, *37*(1), 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Stenfelt, S., Rudner, M., Rönnberg, J., & Wingfield, A. (2016c). A framework for understanding effortful listening. *Ear and Hearing*, *37*(Supplement 1), 5S–27S.

Pickles, J. O. (2012). *An introduction to the physiology of hearing* (4th). Brill.

Picou, E. M., & Ricketts, T. A. (2016). The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear and Hearing*, *37*(5), 547–558. https://doi.org/10.1097/AUD.0000000000000313

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2011). Visual Cues and Listening Effort: Individual Variability. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1416–1430. https://doi.org/10.1044/1092-4388(2011/10-0154)

Plack, C. J., Barker, D., & Prendergast, G. (2014). Perceptual consequences of "hidden" hearing loss. *Trends in Hearing*, *18*, 1–11. https://doi.org/10.1177/2331216514550621

Plack, C. J. (2018). *The sense of hearing* (3rd). Routledge. https://doi.org/10.4324/9781315208145

Polich, J. (2007). Updating p300: An integrative theory of p3a and p3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

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

Rance, G. (2008). Auditory neuropathy/dys-synchrony and its perceptual consequences. *Trends in Amplification*, *12*(2), 103–120. https://doi.org/10.1177/1084713808317062

Rance, G., Saunders, K., Carew, P., Johansson, M., & Tan, J. (2014). The use of listening devices to ameliorate auditory deficit in children with autism. *The Journal of Pediatrics*, *164*(2), 352–357. https://doi.org/10.1016/j.jpeds.2013.09.041

Reinten, I., De Ronde-Brons, I., Houben, R., & Dreschler, W. (2021). Measuring the Influence of Noise Reduction on Listening Effort in Hearing-Impaired Listeners Using Response Times to an Arithmetic Task in Noise. *Trends in Hearing*, *25*, 233121652110144. https://doi.org/10.1177/23312165211014437

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *Journal of the Acoustical Society of America*, *136*(5), 2642–2653. https://doi.org/10.1121/1.4897398

Richter, M., Gendolla, G., & Wright, R. (2016a). Three Decades of Research on Motivational Intensity Theory. In *Advances in Motivation Science* (pp. 149–186, Vol. 3). Elsevier. https://doi.org/10.1016/bs.adms.2016.02.001

Richter, M., Schmiedel, S., Tina andgarnier, & Strauss, D. J. (2016b). Monitoring listening effort from respiration: Sternocleidomastoid muscle activity reflects task demand in competing talker scenarios. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *24*(12), 1325–1335. https://doi.org/10.1109/TNSRE.2016.2532915

Robertson, C. E., & Baron-Cohen, S. (2020). Sensory perception in autism. *Nature Reviews Neuroscience*, *21*, 431–444. https://doi.org/10.1038/s41583-020-0333-z

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

Rossing, T. D., Moore, F. R., & Wheeler, P. A. (2002). *The science of sound* (3rd ed.). Addison Wesley.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working Memory Capacity May Influence Perceived Effort during Aided Speech Recognition in Noise. *Journal of the American Academy of Audiology*, *23*(08), 577–589. https://doi.org/10.3766/jaaa.23.7.7

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech,*

Language, and Hearing Research, *52*(5), 1230–1240. https://doi.org/10.1044/1092-4388(2009/08-0111)

Schmidt, C., Collette, F., Cajochen, C., & Peigneux, P. (2007). A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology*, *24*(7), 755–789. https://doi.org/10.1080/02643290701754158

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

Sjouwerman, R., & Lonsdorf, T. B. (2018). Latency of skin conductance responses across stimulus modalities. *Psychophysiology*, *56*(4). https://doi.org/10.1111/psyp.13307

Sommers, M. S., & Huff, L. M. (2015). Aging, context processing, and comprehension. *Encyclopedia of Geropsychology*, 1–8. https://doi.org/10.1007/978-981-287-080-3_112-1

Souza, P., & Arehart, K. (2015). Robust relationship between reading span and speech recognition in noise. *International Journal of Audiology*, *54*(10), 705–713. https://doi.org/10.3109/14992027.2015.1043062

Stamper, G. C., & Johnson, T. A. (2015). Auditory function in normal-hearing, noise-exposed human ears. *Ear and Hearing*, *36*(2), 172–184. https://doi.org/10.1097/AUD.0000000000000107

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, *153*(3736), 652–654. https://doi.org/10.1126/science.153.3736.652

Strauß, A., Henry, M. J., Scharinger, M., & Obleser, J. (2014). Alpha phase synchronization reflects listening effort in noise. *Journal of Neuroscience*, *34*(46), 15688–15695.

Strauss, A., Wöstmann, M., & Obleser, J. (2013). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience*, *7*, 851. https://doi.org/10.3389/fnhum.2013.00851

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *Journal of the Acoustical Society of America*, *125*(5), 3328–3345. https://doi.org/10.1121/1.3097469

Suga, N., & Ma, X. (2008). Multiparametric corticofugal modulation and plasticity in the auditory system. *Nature Reviews Neuroscience*, *9*(1), 74–85. https://doi.org/10.1038/nrn2273

Szabadi, E. (2011). Functional organization of the sympathetic pathways controlling the pupil: Light-inhibited and light-stimulated pathways. *Frontiers in Neurology*, *2*, 1–10. https://doi.org/10.3389/fneur.2011.00020

Tallal, P., Miller, S., & Fitch, R. H. (1993). Neurobiological basis of speech: A case for the preeminence of temporal processing. *Annals of the New York Academy of Sciences*, *682*(1), 27–47.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Tiippana, K. (2014). What is the mcgurk effect? *Frontiers in Psychology*, *5*, 725. https://doi.org/10.3389/fpsyg.2014.00725

Tremblay, K. L., & Backer, K. C. (2015). Listening and learning: Cognitive contributions to the rehabilitation of older adults with and without audiometrically defined hearing loss. *Ear and Hearing*, *36*, 1S–2S. https://doi.org/10.1097/AUD.0000000000000224

Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology*, *114*(7), 1332–1343. https://doi.org/10.1016/S1388-2457(03)00059-2

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*, *24*(3), 761–766. https://doi.org/10.1037/a0014802

Veronesi, L., & Mauney, D. (2022). Equal loudness contours: A primer. *Audio Engineering Society*.

Vlemincx, E., Abelson, J. L., Lehrer, P. M., Davenport, P. W., Van Diest, I., & Van den Bergh, O. (2013). Respiratory variability and sighing: A psychophysiological reset model. *Biological Psychology*, *93*(1), 24–32. https://doi.org/10.1016/j.biopsycho.2012.12.001

Wagener, K., Kühnel, V., & Kollmeier, B. (1999). Development and evaluation of a german sentence test i: Design of the oldenburg sentence test. *Zeitschrift Fur Audiologie*, *38*, 4–15.

Weisz, N., Hartmann, T., Müller, N., Lorenz, I., & Obleser, J. (2011). Alpha rhythms in audition: Cognitive and clinical perspectives. *Frontiers in Psychology*, *2*, 73. https://doi.org/10.3389/fpsyg.2011.00073

Wendt, D., Hietkamp, R. K., & Lunner, T. (2016). Impact of noise and hearing loss on processing speed and cognitive workload during speech perception. *Frontiers in Psychology*, *7*, 1893. https://doi.org/10.3389/fpsyg.2016.01893

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

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience*, *32*(40), 14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, *36*(4), e153–e165. https://doi.org/10.1097/AUD.0000000000000145

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2016). Pupil responses reveal cognitive load during speech recognition in noise for listeners with hearing loss. *Ear and Hearing*, *37*(1), 1–13.

Winn, M. B., & Moore, A. N. (2018). Pupillometry shows the effort of auditory attention switching. *The Journal of the Acoustical Society of America*, *144*(3), 1876–1877.

Wiśniewski, M. G., Ásgeirsdóttir, F., Møller, C., Ragert, P., & Ødegaard, M. (2017). Frontal theta oscillations indicate level of cognitive control during auditory selective attention. *NeuroImage*, *162*, 222–230. https://doi.org/10.1016/j.neuroimage.2017.09.005

Wong, P. C. M., Jin, J. X., Gunasekera, G. M., Abel, R., Lee, E. R., & Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia*, *47*(3), 693–703. https://doi.org/10.1016/j.neuropsychologia.2008.11.032

World Health Organization. (2018). *Environmental noise guidelines for the european region* (tech. rep.). World Health Organization. Copenhagen.

World Health Organization. (2021). Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Wöstmann, M., Obleser, J., & Strauß, A. (2015). Multitasking and working memory load modulate distractor gating in the human auditory cortex. *Journal of Neuroscience*, *35*(30), 10754–10762. https://doi.org/10.1523/JNEUROSCI.1060-15.2015

Wu, J. (2012). Cluster analysis and k-means clustering: An introduction. In *Advances in k-means clustering: A data mining thinking* (pp. 1–16). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29807-3_1

Wu, Y. H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing*, *37*(6), 660–670. https://doi.org/10.1097/AUD.0000000000000335

Wu, Y.-H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., & Oleson, J. (2015). Characterizing listening effort: Ecological momentary assessment study. *JMIR mHealth and uHealth*, *3*(4), e4001.

Yip, M. (2002). *Tone* (1st). Cambridge University Press.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Heslenfeld, D. J., & Rönnberg, J. (2012). Behavioral and fmri evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. *Brain and Language*, *122*(2), 103–113. https://doi.org/10.1016/j.bandl.2012.05.006

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry: Processing load across a wide range of listening conditions. *Psychophysiology*, *51*(3), 277–284. https://doi.org/10.1111/psyp.12151

Zekveld, A. A., & Kramer, S. E. (2019). Cognitive processing load across the lifespan: Effects of hearing loss and task demands on the pupillary response during listening. *Biological Psychology*, *148*, 107771. https://doi.org/10.1016/j.biopsycho.2019.107771

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

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive Load During Speech Perception in Noise: The Influence of Age, Hearing Loss, and Cognition on the Pupil Response. *Ear & Hearing*, *32*(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2018). The relationship between cognitive performance and pupil dilation during a speech-in-noise task. *Ear and Hearing*, *39*(3), 478–488.

Ziegel, E. R. (2000). Handbook of chemometrics and qualimetrics, part b.