Assessing how vision benefits speech-in-noise perception and the impact of ageing and hearing loss on audiovisual benefits

Chint Lida Alampounti

A thesis submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

March 2024

I, Chint Lida Alampounti, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Acknowledgements

I would like to thank all the people who influenced me in their own way towards achieving my goal of doing a PhD. While words fall short of capturing the full extent of my gratitude, I offer this attempt to acknowledge their impact.

I would like to thank, first and foremost, my supervisors Professor Jennifer Bizley and Dr Hannah Cooper. The wealth of their knowledge, insightful advice, and unwavering support – including during the difficult times of the COVID-19 pandemic – have been invaluable. Their dedication to nurturing my academic growth and empowering me to shape this PhD based on my own scientific thought only strengthened my zeal to learn more about the beautiful world of auditory neuroscience and inspired me to push my boundaries. This PhD has been one of the most enriching and transformative journeys of my life; I thank them for entrusting me with it.

I am also grateful to all my fellow members of the Bizley lab, past and present, Dr Joe Sollini, Dr Stephen Town, Dr Huriye Atilgan, Carla Griffiths, Katarina Poole, Bhavisha Parmar, Dr Soraya Dunn, Dr Rebecca Norris, Amit Khandhadia, Dr Andrea Caso, Jules Lebert and Louay Madanat, as well as my colleagues at the UCL Ear Institute more generally, for creating an ideal working environment. My time with them has been a continuous source of enjoyment talking science and life.

I would like to thank my thesis committee members Professor Mairéad MacSweeney, Professor Stuart Rosen and Professor Maria Chait for their instrumental feedback throughout my PhD, their support and their contagious enthusiasm for research. I would like to thank Professor Mairéad MacSweeney and her colleagues for the creation of the Test for Adult Speechreading, which I chose to include in my test battery. I am very grateful to Professor Stuart Rosen for sharing with me his code for the *Children's Coordinate Response Measure* (*CCRM*) which I adapted to develop the audiovisual speech in noise test *vCCRMn*. Finally, I am grateful to Professor Maria Chait, as well as her team, with special thanks to Claudia Contadini-Wright, Dr Alice Milne, Mert Huviyetli, Kaho Magami and Mingyue Hu, for their support of many forms during my participant testing phase at the UCL Ear Institute's human labs.

An important part of this journey would have not been possible without our technician Gordon Mills, who provided technical support towards moving online for participant testing due to the COVID-19 pandemic, as well as his advice on recording and processing my vCCRMn stimuli at UCL's anechoic chamber.

Special thanks must go to Dr Tim Green and Dr Gwijde Maegherman for training me in using UCL's anechoic chamber.

I am also very grateful the talkers of my speech in noise test. They all handled their difficult task with so much care and dedication. It was a great pleasure to work with them.

A big "thank you" to all my participants, without whom this project would have not been possible. Their honest contribution, spirit of teamwork, their sharing of their experiences living with and without hearing loss and their great zeal to contribute to the neuroscientific field made the participant testing phase of my PhD one of its most motivating and beautiful aspects.

I would also like to thank the National Institute for Health Research (NIHR), University College London Hospitals (UCLH) Biomedical Research Centre (BRC) and the UCL Ear Institute for funding my work.

To my friends Mariarenia Vitorou, Christos Kanellopoulos, Ioanna Anagnou, Cadi-Hazel Davies, Maggie Kumar, Ai Fukunaga, Benjamin Tubiana, Simon Busch-Moreno, Xiao Fu and Giorgos-Miltos Maragkakis. Thank you for your friendship over the years, for always believing in me and for your continuous support in following my dreams.

Words fall short in expressing my gratitude to Marcos for everything he has done. It was through his example of being incredibly kind-hearted, generous and a brilliant scientist personifying discipline, hard work and ability to achieve the impossible with excellence that I came closer to whom I want to become. I also thank him for volunteering to be recorded as a practice talker for my speech in noise test vCCRMn at UCL's anechoic chamber – I owe to him the successful multitasking required for the recording of stimuli.

I am also most grateful to my brother Alexandros for being an example of an excellent person and scientist. There is no other little sister that is as proud of her brother as me. His kindness, generosity, unending thirst for knowledge, hard work and resilience towards achieving his goals in excellence are nothing short of inspiring. I have been, and always will be, in awe of his brilliance, evident from his childhood years till this day. I thank him for his selfless love, his guidance and support in the face of any difficulty I encountered in life, and his protection throughout the years. I owe my path to the sciences to him, from the moment he asked me: "Lida, why not do a master's in science?". The journey has been wonderful.

Finally, I would like to wholeheartedly thank my parents who have led by example both as people and as academics. I would not have been writing these acknowledgements if it was not for them and for every part of their heart and soul they put towards my personal and academic growth. I thank them for always believing in me, for their unwavering support and selfless love. I devote this thesis to both of them.

Abstract

Investigations of the role of audio-visual integration in speech-in-noise perception have, in the previous decades, largely focused on the benefit provided by lipreading cues. A growing body of evidence however suggests that additional audio-visual processes exist, and that these have the capacity to influence auditory scene analysis. Specifically, audio-visual temporal coherence – the extraction, and integration, of the temporally correlated information of the amplitude envelope of speech and the opening and closing of the mouth – is a potential candidate. Whether audio-visual temporal coherence can aid speech-in-noise, and to what extent, relative to lipreading, remains however largely unknown. The current work examines, across 125 individuals spanning the ages 19 to 85, and with both typical and impaired hearing, the contributions of these two mechanisms to the visual enhancement of speech-in-noise perception.

An audio-visual speech-in-noise task (the vCCRMn) was developed, designed to capture both lipreading, and audio-visual temporal coherence-related enhancements of listeners' auditory performance. The vCCRMn task was employed, along with a battery of accompanying tests, in three experimental groups: younger participants with normal hearing, older participants with normal hearing, and older participants with hearing loss.

vCCRMn presented participants with sentence stimuli, containing two key words which participants had to report. These target sentences were presented in the background of two competing (masker) talkers. In audio-visual (AV) and interrupted (Inter) conditions of the task, participants were also presented with a video of a talker. This talker could either match the speaker reading the target sentences (target-coherent conditions), or one of the talkers reading the masker sentences (masker-coherent conditions). In the Inter condition the video of the talker would freeze during the key words – thus, removing the possibility of participants exploiting lipreading cues for identification of the target words. A static-image-with-audio (A) condition was also employed as a control, and baseline for comparing against the video conditions to compute visual enhancements. The image displayed in the condition could also be target- or masker-coherent.

Among the target-coherent conditions, participants performed the best in the AV, followed by the Inter, and then the A condition. Thus, participants were gaining visual enhancements from both the AV and the Inter video conditions, compared to the static image condition, with the visual enhancement obtained from the AV condition being the largest. This finding suggested that participants' auditory performances were at the same time enhanced by both lipreading, and audio-visual temporal coherence cues. Additionally, participant performances were better in the target coherent video conditions, compared to the masker video conditions. Performances in the AV masker video condition were, in fact, even worse than those in the static-image-with-audio condition, suggesting that obligatory dynamic visual cues can impair the listening experience, as well as support it. Finally, participant speech-in-noise performance, and audio-visual benefit were negatively impacted by both ageing, and hearing loss.

Impact statement

Hearing loss is among the most prevalent sensory pathologies, impacting the communication abilities of millions of people worldwide, in everyday environments often filled with background noise. While the use of hearing aids, and cochlear implants have proven relatively effective in improving the lives of clinical populations, the use of complementary senses in the enhancement of the auditory experience remains understudied, and underexploited. The auditory modality has long been known to enjoy influences from vision. Audio-visual illusions, such as the McGurk effect and the Ventriloquist illusion, and the tendency of the brain to draw correspondences across the two senses (e.g. high pitch with small size) are but a few examples of such influences.

Closer to the speech perception problem, it has long been known that lipreading, the extraction of linguistic information from mouth movements, offers a visual enhancement (an "audio-visual benefit") of audition when listening in noise. More recently, it has been suggested that additional, and language independent audio-visual integration mechanisms, may also have the capacity to provide such audio-visual benefits. Specifically, one such promising mechanism is that derived from the temporal coherence between the amplitude envelope of speech, and the mouth opening and closing of the speaker.

Although both lipreading, and audio-visual temporal coherence have been studied independently in their contributions to the audio-visual benefit, it remains unknown whether both mechanisms can provide simultaneous visual enhancements in the context of speech-in-noise perception. Further, the conditions under which these mechanisms operate (including listener demographics), and the relative contributions of each to the audio-visual benefit of listeners are not well-understood.

The current body of work advances our understanding of audio-visual integration in the context of speech-in-noise perception, shedding light on the complex interplay between visual cues provided by the mouth movements produced during speech. Specifically, with the development and employment of the vCCRMn, an innovative speech-in-noise task tailored to assess audio-visual speech perception, this research demonstrates the simultaneous contributions of lipreading and audio-visual temporal coherence. Furthermore, with the development of statistical models applied to a robust sample size of 125 participants, spanning the ages 19 to 85 years, and including individuals having both typical, and impaired hearing, this research was able to provide a nuanced understanding of how lipreading, and audio-visual temporal coherence contribute to listeners' audio-visual benefits.

Importantly, these findings extend implications beyond academic circles, and into the clinical realm. This research shows, for example, that listeners have the capacity to bolster their visual enhancements with practice – a promising finding, and relevant to the design of clinical audio-visual training regimes for the enhancement of patients' living conditions. Furthermore, this work underscores the advantage of individualised training regimes, tailored to prioritise lipreading or audio-visual temporal coherence training based on individual patient demographics. Such training regimes are likely to yield the most beneficial outcomes for the patients.

Ultimately, the current work emphasises the importance of vision in our daily auditory activities, highlighting the necessity of adopting an audio-visual mentality when addressing hearing challenges.

Contents

List of Figures

List of Tables

List of abbreviations

1U1D	1-up-1-down
1U2D	1-up-2-down
1U3D	1-up-3-down
ASA	Auditory scene analysis
AV	Audio-visual
BKB	Bamford-Kowal-Bench
CCRM	Children's Coordinate Response Measure
CRM	Coordinate Response Measure
CST	Connected Speech Test
DoF	Degrees of freedom
FFT	Fast Fourier Transform
HINT	Hearing in Noise Test
HL	Hearing loss
IAC	Industrial Acoustics Company
IEEE	Institute of Electrical and Electronics Engineers
LMEM	Linear mixed effects model
MoCA	Montreal Cognitive Assessment
NH	Normal hearing
ORN	Object-related negativity
PTA	Pure-tone audiometry
RMS	Root mean square
SD	Standard deviation
SIN	Speech-in-noise
SNR	Signal-to-noise ratio
SPIN	Speech Perception in Noise
SRT	Speech reception threshold
TAS	Test of Adult Speechreading
U3A	University of 3rd Age
VIF	Variance inflation factor
fMRI	functional magnetic resonance imaging
vCCRM	video version of the Children's Coordinate Response Measure
vCCRMn	video version of the Children's Coordinate Response Measure with nouns at the end

Chapter 1: Literature review

1.1. Introduction

Listening in noise is difficult, so much so that our ability to cope with interactions in noisy settings has interested researchers since Cherry (1953) defined his famous “cocktail party problem”. Yet one needs not attend a cocktail party to find oneself in a noisy environment: most human interactions do, in fact, occur in the presence of background noise (Walden et al., 2004). And the difficulty in coping with this noise becomes even more apparent when living with hearing loss – the most common complaint of hearing aid users being that their device is insufficient in noisy settings (Kochkin, 2000).

On the other hand, it has also been known for decades now that our auditory experiences are not only supported by our auditory systems, but also by our vision. For example, it was only a year after Cherry (1953) had defined the cocktail party problem, that Sumby and Pollack, (1954) published their seminal speech-in-noise work, showing how looking at the talker’s face augmented their participants’ speech perception. More recent work, points to similar directions, with findings that the audio-visual condition improves both the signal-to-noise ratio at which speech can be perceived (Bernstein et al., 2004; Grant & Seitz, 2000), and the percentage of correct stimuli identification (Tye-Murray et al., 2007), compared to the audio-only condition. Other advantages of audio-visual speech perception (compared to audio-only), include better comprehension of stories (Arnold & Hill, 2001), more efficient shadowing of spoken passages (Reisberg et al., 1987), reduced cognitive effort to perform the auditory task at hand (Anderson & Gagné, 2011), and improved ability to learn degraded speech (Wayne & Johnsrude, 2012).

All these studies are describing a “visual enhancement” of audition, or, what I herein call an audio-visual benefit. The audio-visual benefit conferred by vision when looking at the talker’s face while listening to them in noisy settings, is the topic I investigated in this thesis by developing a speech-in-noise task that allowed me to measure it in experimental participants. More specifically, my research was focused on the audio-visual benefit related to movements of the talker’s mouth, including investigations of the factors that influence, and contribute to it.

The contribution of two broader mechanisms is explored in my work: That of lipreading (i.e. language-dependent) factors, and that of language-independent audio-visual factors that also involve looking at the talker’s lips. The latter constitute low-level audio-visual (mouth opening, and auditory) sensory cues that are temporally coherent with each other and can be interpreted as resulting in a so-called “audio-visual binding” (Bizley et al., 2016; Lee et al., 2019). Lipreading, on the other hand, always refers to visually-conveyed (through place and manner of articulation) phonetic information in this work. Contributions from both lipreading, and temporal coherence mechanisms could, in principle, provide an enhancement to audition by influencing the process of auditory scene analysis (ASA) – the process by which our auditory system segregates and groups sound elements such that they can be attributed to individual sources (Bregman, 1990, Chapter 1, pages 2 – 4).

I was particularly interested in investigating whether audio-visual temporal coherence mechanisms could facilitate listening in complex environments. Although the contributions of lipreading to speech-in-noise perception have remained relatively uncontroversial (since, in fact, Sumby and Pollack, (1954)), the contributions of temporal coherence mechanisms, such as audio-visual binding, have been largely neglected, or, rather, have not been differentiated from those of lipreading in previous research.

Thus, among my primary goals for this chapter were to outline, and make the distinction between the two mechanisms, and to discuss the experimental difficulties involved in isolating the effect of audio-visual binding from that of lipreading when measuring audio-visual benefits. To set the stage, I begin the chapter with a discussion of ASA and auditory object formation. With this, I provide an outline of the specific challenges the auditory system faces, and the foundation onto which visual enhancements can be considered. Then, I discuss the mechanisms of lipreading and temporal coherence, with an emphasis on audio-visual binding, and how they support ASA and provide individual audio-visual benefits when listening in noise. Next, I discuss speech-in-noise tests, their application for the measurement of audio-visual benefits, and how they must be designed in order to allow for measurement of both lipreading, and audio-visual binding, contributions to audio-visual benefits. Finally, I present the aims of my thesis, and a brief outline of the chapters to follow.

1.2. Object formation in audition: Auditory scene analysis

The listener lives in an acoustically busy world. Numerous sounds of varying spectral content and intensities constantly appear and arrive at the ear as a complex mixture. So, how can any one of those sound sources be distinguished from the rest? Most, if not all of our human interactions occur in the presence of some noisy background (Walden et al., 2004). How are we, in the midst of all this noise, able to perceive what our interlocutor is saying?

Albert Bregman (1990) (Chapter 1, pages 2 – 4) referred to people's efforts to solve these everyday cocktail party-like problems as *auditory scene analysis* (ASA): The listener's ability to separate the different sounds arising from different sound sources, into distinct mental representations. These mental representations are called *auditory objects* (Bizley & Cohen, 2013). For example, while sitting at a restaurant, we might hear glasses clinging, the door opening, or the sound of cutlery, and each of these almost "discrete" (in time) sounds can be described as auditory objects.

Usually, however, sounds don't occur as isolated events; they unfold in time. A sequence of auditory objects unravelling in time, such as our friend talking, or the footsteps of the waiter at the restaurant, form what is called an *auditory stream* (Bizley & Cohen, 2013). And a stream can itself be thought of as an auditory object (Schnupp et al., 2019) – we perceive the footsteps of the waiter as an "entity" not each one of them as an independent event.

Each of the sounds arriving at the ear is associated with a set of physical acoustic features, which the brain encodes into perceptual features. To form distinct auditory objects from these sounds, the brain must then "decide" which of these perceptual features are similar enough to form part of the same sound. In doing so, it conducts what is called *auditory scene segregation* (also referred to as *streaming*): It groups similar features together, and thus segregates the unfolding auditory scene into distinct streams. Stream segregation is what helps a listener separate a specific voice of interest from the complex mixture.

There are two broad principles by which the brain groups acoustic features together into perceptual objects: One is through exploitation of the so-called *simultaneous grouping cues*, and the second is through the use of *sequential grouping cues* (Bizley & Cohen, 2013; Bregman, 1990). Further, Bregman (1990) postulated that this segregation of the auditory scene is largely unconscious, bottom-up-driven (or "primitive" as he called it) but can be influenced by top-down processes such as attention and prior knowledge (in what he called "schema-based processing") (Bregman, 1990, Chapter 1, page 38). I expand on these topics in the next sections, beginning with a discussion on the use of simultaneous and sequential grouping cues in auditory stream formation.

1.2.1. Simultaneous and sequential grouping cues for auditory object formation

As mentioned just above, the formation of auditory streams depends on sequential and simultaneous grouping processes. The two processes are not exactly mutually exclusive, but it is useful to treat them as such in their description. Auditory data arrive from the senses in the form of a complex sound mixture: The ear receives a single pressure wave, consisting of the sum of every distinct sound happening in our vicinity. Simultaneous grouping is the process of selecting, from the simultaneously arriving data, the acoustic features that are likely parts of the same sound source (i.e. of the same auditory object). The sequential grouping process is the connection, into streams, of these sensory data across time.

I begin with examples of simultaneous cues (see Bizley and Cohen, (2013) and Darwin and Carlyon, (1995) for reviews). Certain sounds, such as human voices, are "harmonic", in that they consist of a set of frequencies, all of which are integer multiples of their so-called fundamental frequency. If the auditory system detects frequencies that are integer multiples of some common fundamental in the soundscape, it is likely to group them together as belonging to the same sound (see Figure 1.1. for an illustration). This "harmonicity" is a crucial cue for separating a friend's voice from a mixture.

Another example of an important simultaneous cue is the synchrony of onsets and offsets of a sound's components: When our friend talks, all the components of her voice start together, and stop when she stops talking. Furthermore, the spatial location of the sound source may serve as a simultaneous, albeit weak (Darwin & Hukin, 1997), grouping cue. Other common simultaneous cues are loudness patterns of a sound's components, and how close two components are in frequency. Worth noting is that simultaneous cues are also sometimes referred to as "instantaneous" cues (see e.g. Shamma et al., 2011), from the fact that the auditory system extracts these acoustic features in a matter of a few tens of milliseconds (i.e. almost instantaneously). This is contrary to the sequential grouping which applies over longer timescales.

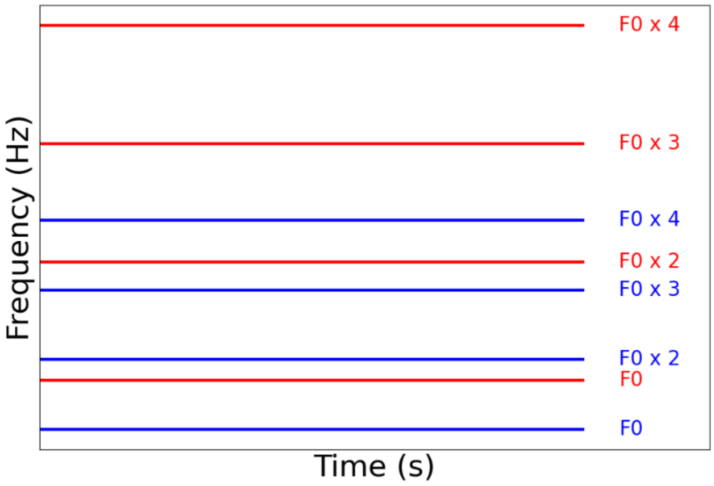


***Figure 1.1**: Simultaneous grouping cues: Grouping two sounds based on their harmonic content. Two distinct harmonic sounds (blue and red), along with some of their harmonics. The fundamentals are denoted with F0, and the harmonics are integer multiples of the fundamentals. The auditory system groups the blue components and the red components into two distinct auditory objects.*

The formation of auditory streams, by linking the individual auditory pieces across time is achieved through sequential processes, applied over timescales of the order of hundreds of milliseconds, or even seconds (Shinn-Cunningham et al., 2017). These processes are best illustrated with the so-called "ABA" experimental paradigm (see Noorden, 1975 for an example). During an ABA experiment, participants are presented with two pure tones of different frequencies, A, and B. These are played one after the other, in a repeating cycle, such that the participant is exposed to A, then B, then A, then B, and so on. The experiment has two potential outcomes as far as stream formation is concerned: Either the participant perceives the two tones as members of the same stream (and thus hears a "galloping" sequence of sound), or he perceives the two tones as members of two separate streams (the A stream, and the B stream) (these outcomes are illustrated in Figure 1.2).

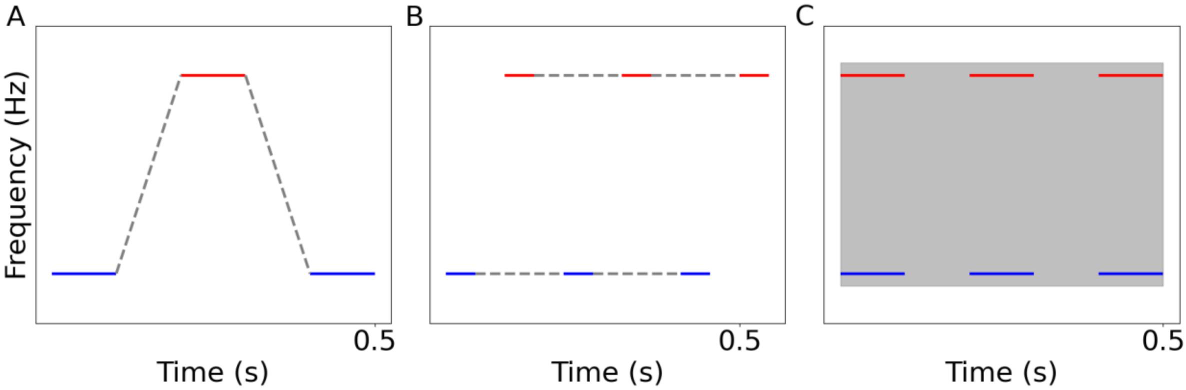


Figure 1.2: *Sequential grouping cues illustrated with the ABA paradigm. If the two tones, A and B, are separated enough in terms of their frequency, the following apply: (A) When the presentation rate is slow, A and B will be perceived as belonging to the same stream (indicated with the connecting grey lines). (B) When the presentation rate is fast, the two tones will be segregated into, and perceived as two streams, the stream for tone A and the stream for B. (C) When the tones are presented concurrently, they will be perceived as members of the same stream (indicated with the grey background). See Noorden (1975) for more information on the ABA paradigm.*

Whether streams will be segregated, or not, depends on two factors: The tones' separation in frequency, and the separation of same frequency tones in time (i.e. the separation in time of one A to the next A, and from one B to the next B). If tones are well-separated in frequency, but the ABA paradigm is presented with a slow rate, then the As and Bs will be perceived as part of the same stream (Figure 1.2 (A)). If tones are well-separated in frequency and the presentation is fast, then they will be segregated into two streams, the stream of As, and the stream of Bs (Figure 1.2 (B)). There is, also, an intermediary frequency separation setting (called "bistable") where the listener alternates from stream segregation to stream integration.

Generally, it is useful to think of these two factors (frequency and temporal separations) as competing with each other in how they influence the brain's decision to split the streams in two or fuse them. In the first example, where one stream is perceived, it is because the "perceptual distance" of the frequency separation of A and B is smaller than the perceptual distance of the temporal separation of any one A from the next A, and any one B from the next B. Thus, the sequence of tones is perceived as part of the same "melody". On the other hand, in the second example, where the streams are segregated, the perceptual distance of the temporal separation of any one A from the next A, and any one B from the next B, is smaller than the perceptual distance of the frequency separation between A and B. Thus, the two are grouped accordingly into different streams.

Furthermore, when the sounds are more complex, other factors, such as timbre, or the manner with which transitions from A to B occur (e.g. abrupt vs smooth), also contribute to this perceptual distance (see e.g. Bregman, 1990, Chapter 1, page 52), and the brain must compute, using the totality of these contributions, a measure of what is closer to what, such that it can segregate or fuse streams accordingly (Bizley & Cohen, 2013). To this end, one physical feature of the components of sound arising from the same source that elegantly links them all together, and thus helps the brain group the sound elements together into streams, is the fact that these elements occur concurrently, or are temporally coherent with each other, when they arise from the same source. This is illustrated by panel (C) of Figure 1.2, where the two tones are perceived as part of the same stream if they are presented concurrently.

It has been shown that temporal coherence between the A and B tones results in them being perceived as one stream even when they are separated by more than one octave (Elhilali et al., 2009), suggesting it must be an important factor influencing stream formation. In fact, Shamma et al., (2011) went as far as postulating that temporal coherence (in terms of their co-occurrence) between different acoustic features of a sound (e.g. pitch, timbre, loudness and so on), is *the* connecting link that helps the brain assign the features to a particular stream and keep them differentiated from the features of other sounds.

Having outlined the grouping principles, in the next section, I shift the discussion to bottom-up and top-down views for auditory object formation and segregation.

1.2.2. Primitive versus schema-based processing: Bottom-up and top-down views of auditory scene analysis

1.2.2.1. Primitive processing

Bregman, (1990) (Chapters 2 and 3), suggested that the simultaneous and sequential grouping processes described in the previous section are automatic and bottom-up. He used the word “primitive” to express that the processes depend on the features of the incoming acoustic data and are innate – evolutionarily conserved. It has been, since then, shown that these phenomena are indeed “innate”, in that they have been observed in infants (McAdams & Bertoncini, 1997). Also, they are not restricted to humans: Wisniewski & Hulse, 1997 have shown, for example, that European starlings are capable of discriminating the segments coming from the songs of their conspecifics, even when presented in the background of other birds’ songs (see also Izumi, 2002, 2003 for work on monkeys).

Furthermore, human studies combining psychophysical paradigms with the measurement of event related potentials (ERPs) have shown that these grouping processes are independent of top-down attention (Alain, 2007). One such study was that of (Alain et al., 2001), where the authors investigated their participants’ ability to discriminate between presentations of tuned, or mistuned stimuli. In this study, tuned stimuli consisted of the presentation of a fundamental frequency, along with its harmonics, and mistuned stimuli consisted of the same stimuli, but with one of the harmonics having a frequency that was not exactly an integer multiple of the fundamental. The authors showed that tuned stimuli were perceived as one sound, whereas mistuned as two.

The presence of the so-called *object-related negativity* (ORN) signal, measured as an ERP, reflected whether stimulus segregation into two tones was successful, in the mistuned case. Importantly, the ORN was present even when participants were pre-occupied with another task (e.g. reading a book), when the stimuli were presented, indicating that the grouping process operated independent of top-down attentional controls. Other studies (see Kujala et al., 2007 for a review), assessing the detection of deviant (rare) sounds incorporated into otherwise standardised patterns of sound streams through mismatched-negativity ERP components, also provide evidence that these grouping mechanisms occur automatically, and without the need for top-down attention. These suggestions are not, however, uncontested, especially when considering the more complex sequential grouping processes of stream segregation. There, the question of bottom-up versus top-down becomes more controversial.

1.2.2.2. Schema-based processing

It is worth noting that Bregman himself believed that ASA is not a purely unconscious phenomenon, immune to the influences of higher cognitive processes (Bregman, 1990, Chapter 1, page 38). He coined the term schema-based processing to encapsulate top-down influences such as attention and prior knowledge. Indeed, given the complexities of real-world sensory experiences, it is likely that ASA will not be independent of the listener's attention, prior knowledge, and goals (Shinn-Cunningham, 2008).

Sussman and colleagues (Sussman et al., 2007), ran a series of experiments that led to the conclusion that attention is not always needed for stream segregation. On the other hand, van Noorden, (1975), using the ABA paradigm described in section 1.2.1., observed that stream segregation was sometimes influenced by the participants' active engagement. Indeed, it has been suggested that directing attention to the target auditory stimulus enhances its neural representations in the cortex (Tiitinen et al., 1993). On the other side of the argument, Shamma et al., 2011, 2013 are of the opinion that attention is necessary for stream formation to occur in the first place.

Nonetheless, and regardless of whether attention is required for auditory object formation, an important principle of object-based attentional theories is that when a listener attends to a specific feature of an auditory object (e.g. the pitch of a voice), the listener's sensitivity to all of its remaining features will also be enhanced (Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008) – i.e. attention seems to operate on the auditory object as a whole.

The effects of prior knowledge are more easily exemplified through the example of speech perception. Take for example contextual cues. We are often able to conclude the sentence of a talker based on contextual cues provided by the talker, which are based on prior knowledge we have about that person, or previous conversations. And, if we have already concluded that something might be said, then we are more likely to be able to hear it once it's been said. This is indeed what Pichora-Fuller et al., (1995) have shown in their study. Their participants were presented with sentences in a multi-talker babble noise background and were tasked with the identification of the sentences' final words. The authors showed that when the sentences were high in contextual cues, participants were more likely to detect the final words, compared to when the sentences were low in contextual cues.

1.2.3. Auditory scene analysis: A schematic model

In the preceding sections (1.2.1 and 1.2.2), I provided descriptions of how the auditory system exploits simultaneous and sequential grouping cues to segregate the auditory scene. I also presented evidence to support that although segregation can occur automatically, it can also be influenced by top-down cognitive processes. Here, I conclude section 1.2. with a schematic model of the workings of ASA, applied to the extraction of a voice from a noisy background. Based on the schematic, ASA is a complex procedure and includes interdependencies between the bottom-up and the top-down processes.

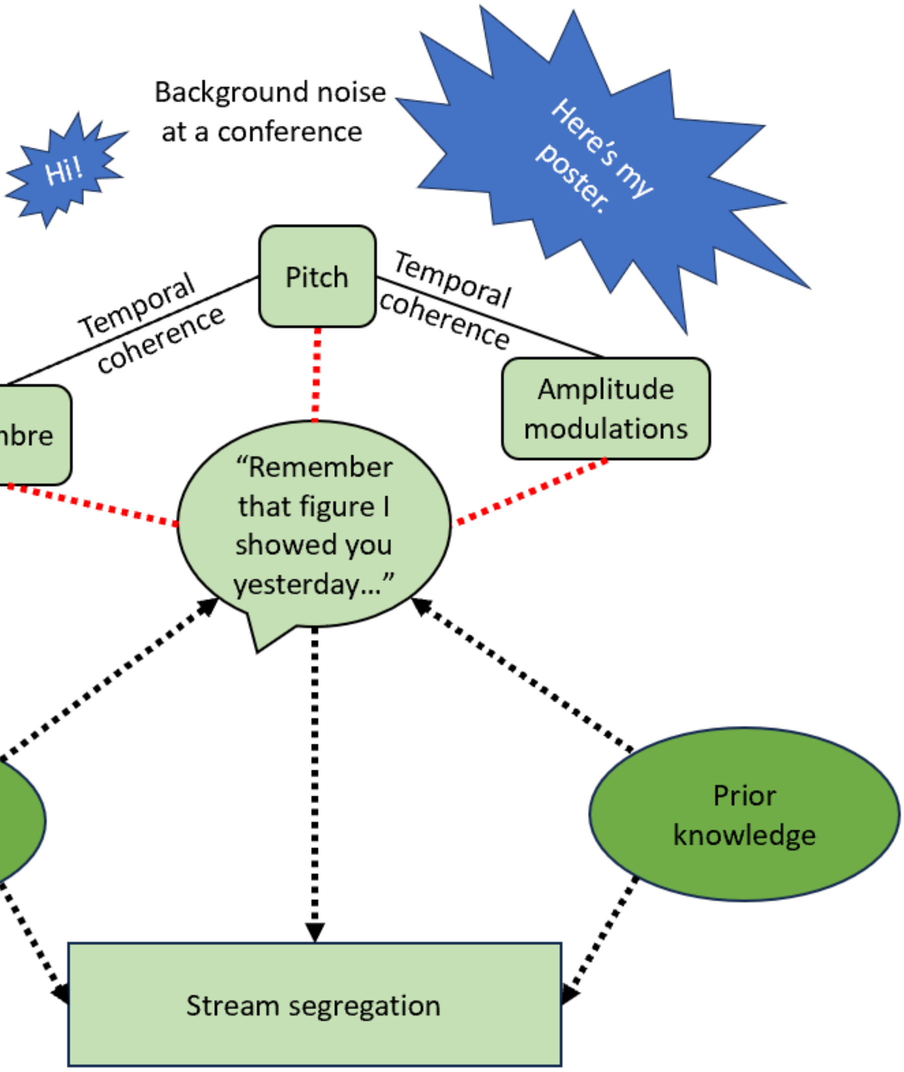


Figure 1.3: *Extracting a voice at a science conference: A schematic model for auditory scene analysis. The voice of the talker utters the sentence "Remember that figure I showed you yesterday…", which has associated with it acoustic features such as timbre, pitch, and amplitude modulations (red dotted line connections). These features are temporally coherent with one another (solid black line connections), in the sense that they co-occur, and they help promote stream segregation in a bottom-up manner (black dotted line connection). Top-down attentional, and prior knowledge, schema-based processes, influence both the voice's feature representations and stream segregation (black dotted line connections).*

Up to this point, I have been focusing on audition, and how the auditory system performs the difficult task of analysing the complex acoustic landscape. In doing so, I sought to provide a foundation for the sections to follow, which introduce the modality of vision, and discuss how it helps with the auditory scene analysis problem. I begin with a discussion of lipreading, which has long been known to provide a boost in the ability to listen in noise, commonly measured as an audio-visual benefit.

1.3. Lipreading: A language-based mechanism for audio-visual benefit

This section discusses lipreading, as one of the primary contributors to the audio-visual benefit.

1.3.1. General description

The existing literature investigating speech perception in noise has drawn a great deal of attention to the role of lipreading in enhancing the listening experience. Lipreading refers to one's ability to perceive speech solely by viewing the talker's mouth movements, including the lips, tongue, and teeth (Auer & Bernstein, 2007; Summerfield, 1992). It is here distinguished from *speechreading*, which is a broader term that involves the interpretation of mouth movements but also additional cues such as facial expressions and body language (Arnold, 1997). In this thesis, the term lipreading is preferred due its specific focus on the visual cues that can be derived from mouth movements, on which I focus on in my work.

Lipreading could be described as a language-based mechanism contributing to the audio-visual benefit: it helps the perceiver extract, and even predict phonetic information (Auer, 2010; Murthy, 2020). It does so through a learnt process of matching mouth movements with speech sounds. This process is, in fact, an integral part of language acquisition and even infants are aware of these matchings soon after birth (Woodhouse et al., 2009). Exposure to mouth movement-phoneme associations, for example, teach the young child that a lip closure represents a bilabial sound (/p/, /b, /m/) rather than a velar sound (/k/, /g/).

For more information on how speech is produced, its visible aspects, and their relationship to lipreadability, the interested reader is referred to Appendix A. For the purposes of this section, the important message is that adults have learnt the association between such visual cues with certain phonemes, syllables, and words from years of experience with speech and language. Thus, lipreading is a language-based AV integration mechanism. And, since it constitutes prior-knowledge, it falls under the umbrella of the schema-based processes (section 1.2.2.2.) contributing to auditory scene analysis.

The distinction is made here, preliminarily, that lipreading operates in the absence of sound, contrary to audio-visual binding, and other temporal coherence mechanisms, which rely on the presence of sound (as discussed in later sections). In the next section, I outline evidence showing that lipreading indeed provides an audio-visual benefit, when listening in noise. I begin the discussion with signal detection theoretical considerations of how, through providing lipreading cues, vision might influence a person's auditory performance. Then, I discuss evidence from published literature that lipreading indeed improves speech-in-noise performance.

1.3.2. The importance of lipreading when listening in noise

1.3.3.1. Lipreading biases the listener's decision of what was heard

In understanding the way by which lipreading enhances the auditory experience, it is useful to look at it through a signal detection theoretical lens (see Wickens, 2001 for a textbook on signal detection theory). Consider, for example, the words cap and cat. Articulating these words, we immediately see how a person looking at us could use the visual information of the positioning of our articulators to make the distinction between the consonants /p/ and /t/. Regardless, the words cap and cat are, to a large extent, similar to each other and are thus, from a purely auditory perspective, easily mixed up.

To express this more formally, using signal detection theory terms, we represent each of the two words with a normal distribution reflecting the conditional probabilities of perceiving each respectively (Figure 1.4; hypothetical example, using artificial data). On such plots, the horizontal axis usually represents the person's perceptual evidence for having heard the word that was presented to them. Furthermore, since the words cap and cat are similar, there is substantial overlap between their distributions, reflecting that they can often be perceptually mixed up.

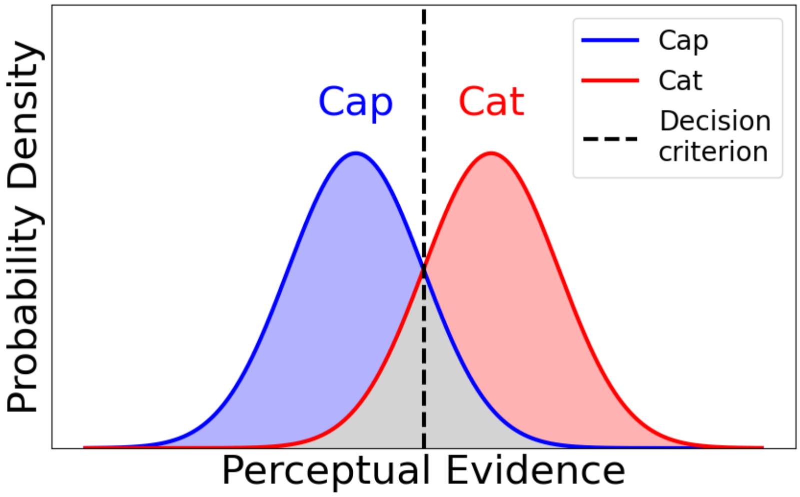


Figure 1.4*: Signal detection theoretical depiction for the perception of the words cap and cat. Each word is represented with its own normal distribution (blue for cap, red for cat), reflecting the probability that it will be perceived, given that the listener was presented with it. The horizontal axis represents the listener's perceptual evidence for having heard a word. The overlap between the sounds of the two words is reflected by the overlap in their respective distributions (grey highlighting). The vertical line between the two distributions depicts the listener's decision criterion. Artificial data were used for the creation of this figure. See Wickens (2001) for more information on signal detection theory.*

In between the two distributions, and through their point of intersection, I have placed a vertical line. This line represents the listening person's so-called *decision criterion*. The decision criterion serves the following function: If the listener was presented with the word cap, and her perceptual evidence for it fell to the left of her decision criterion, then she will have reported hearing the word cap. On the other hand, if her perceptual evidence for it fell to the right of the decision criterion, she will have reported hearing the word cat. Conversely, if the listener was presented with the word cat, and her perceptual evidence for it was to the right of her decision criterion, she will have reported hearing the word cat. Had the perceptual evidence been to the left of the decision criterion she would have reported hearing the word cap. Thus, the decision criterion reflects the point, on the perceptual axis, past which the listener decides to switch from reporting one word, to reporting the other.

There are two ways in which visual cues, and thus audio-visual integration, can influence the listener's chances of reporting one word over the other: One is by imposing a perceptual (bottom-up) change on the listener; this would cause a shift of the distributions, but would not influence the position of the decision criterion. The second is by biasing the listener's decision of what she has heard; this would leave the position of distributions unaltered but would cause a shift of the decision criterion (this is top-down). Lipreading cues operate via the latter mechanism, as illustrated in Figure 1.5 (which also makes use of artificial data for the sake of this example).

Consider the audio-only case from Figure 1.5 (A). The listener was presented with the word cap but, due to the words' high confusability, her perceptual evidence for it fell to the right of her decision criterion, and she reported hearing the word cat. Now consider the audio-visual case from Figure 1.5 (B). The listener was once more presented with the word cap. Her auditory perceptual evidence for hearing the word once again fell to the right of the decision criterion, but once seeing her interlocutor's articulators for /p/, she shifted her decision criterion to the right. With this shift, her auditory perceptual evidence now falls to the left of the (shifted) decision criterion, and she reports hearing the word cap.

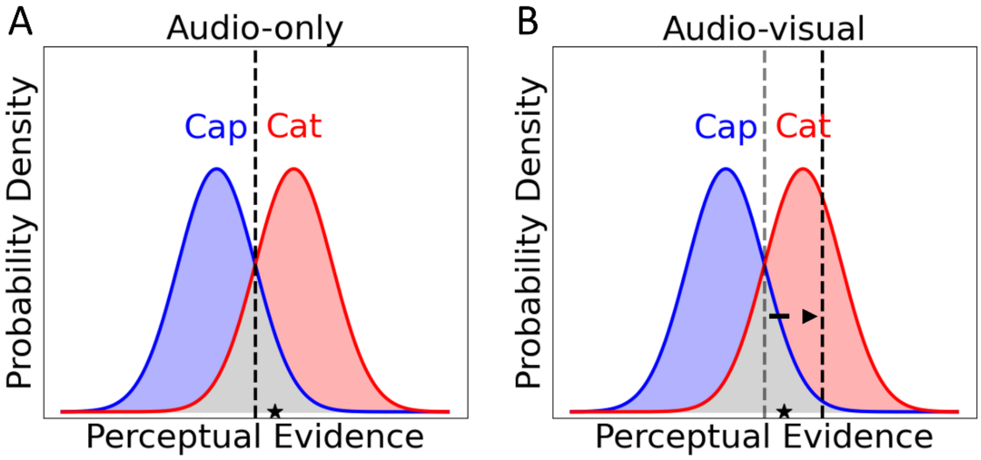


***Figure 1.5**: The effect of lipreading cues on word identification. Colour-coding is the same as in Figure 1.4. (A) Audio-only condition: The listener was presented with the word cap, but the* auditory *perceptual evidence (star symbol) for having heard it fell to the right of the decision criterion, leading to a report of the word cat instead. (B) Audio-visual condition: The listener was presented with the word cap once more, and, although the* auditory *perceptual evidence remained at the same point, lipreading cues promoted a rightward shift in the decision criterion, such that the evidence now falls to its left. Thus, the listener reports hearing the word cap. Artificial data were used for the creation of this figure. See also Bizley et al. (2016) for more information on the theoretical concepts on which this graph is based on.*

Having discussed a potential mechanism via which lipreading could be offering a listening benefit, in the next section I provide evidence from publications showing that it indeed does provide one, especially when listening in noise.

1.3.2.2. Lipreading provides an audio-visual benefit

In clear listening conditions, normal hearing listeners find it relatively easy to follow the speaker without the need to see their face. The importance of lipreading becomes more evident in situations where the acoustic signal is degraded, such as in noisy environments or when a person has hearing loss. As discussed in the previous section, looking at the talker's mouth movements can complement the degraded auditory signal by providing information about the auditory signal's content or by helping disambiguate it. In such cases, if the talker's face is available, most listeners will make use of lipreading cues from it to gain an audio-visual benefit (Erber, 1975).

How can we be certain that it is specifically (or even just partly) lipreading cues that contributed to these reported listening benefits? As I make the case in this thesis that lipreading cues are not the only cues available to extract by looking at a talker's mouth, I should make a clarification regarding the studies I cite here. Namely, I sought to provide speech-in-noise studies that made use of more "lipreading-friendly" stimuli, where the effect of lipreading would be expected to play a clear role in boosting participant performance. Word stimuli are one such example, not only because they are relatively simple to lipread, but also because they are short in duration and allow little time for longer non-lipreading related temporal effects to take place.

Take for example the study of Sumby and Pollack, (1954) mentioned at the beginning of this chapter. In their study, the authors presented their participants with bisyllabic words of spondaic stress pattern (e.g. cupcake, baseball), in audio-only and audio-visual conditions. They reported percentage differences as great as 80% in the performances participants exhibited in the two conditions, especially at low signal-to-noise ratios (SNRs). Further, the audio-visual scores of their participants were always greater than their audio-only scores. Erber, (1975) summarises several other studies similar to Sumby and Pollack (1954), showing similar results. Picou et al., (2011) showed too that listening effort in noise, as measured by a word-pair recall task, was reduced in audio-visual conditions, when participants were good lipreaders.

Grant and Walden, (1996) employed even more elementary stimuli than words. They presented their participants with consonants surrounded by the vowel /a/ (vowel-consonant-vowel stimuli, e.g. /aca/) and tasked them with identifying the consonant presented. In doing so, in addition to showing an audio-visual benefit conferred by lipreading cues when listening in noise, they were also interested in investigating the relative influences of articulatory phonetic features on this benefit. They compared their participants' performances between audio-only and audio-visual conditions, and reported that vision provided almost zero voicing information, significant manner of articulation information, and near perfect place-of-articulation information, across a range of SNRs. Thus, the authors demonstrated some of the complexities underlying lipreading. Given their results, they concluded that the most reliable information provided by lipreading are the surface features of the lips and tip of the tongue.

Other researchers used more complex stimuli to quantify the audio-visual benefit of listening in noise, but at the same time separately measured their participants' lipreading ability. Then they exploited this information to provide evidence that lipreading was indeed contributing to their participants audio-visual performance. One such study is that of Macleod and Summerfield, (1987). In this study the authors made use of sentence stimuli, and rank ordered from easy to hard in terms of the difficulty with which they could be lipread. They presented these sentences to their participants in audio-only, and audio-visual settings, and used the ascending method of limits procedure to determine their participants 50% speech-reception thresholds – the SNR values at which they were identifying 50% of the trials correctly. They also included a visual-only condition to assess their participants lipreading ability. On average, the authors showed that participant performances were greater in the audio-visual condition compared to the audio-only, by 11 dB SNR. Importantly, this audio-visual benefit correlated strongly with their participants' lipreading ability. And, further, it correlated substantially with the lipreading difficulty of the sentences, with easier-to-lipread sentences rendering greater benefits.

The studies cited above provide clear evidence that lipreading does provide a boost in speech-in-noise performance. But how much does lipreading ability vary from person to person, and what are the factors that might contribute to this variance? Further, if lipreading skill does vary, is there potential for it to be trainable? I expand on these topics in the next section.

1.3.2.3. Individual differences in lipreading ability and training

1.3.2.3.1. Individual differences

In the previous section I cited a study reporting that participant performance in dB SNR was greater in audio-visual settings, compared to the audio-only, by 11 dB SNR on average (Macleod & Summerfield, 1987). This performance, however, ranged from 6 to 15 dB SNR among different subjects, and from 3 to 22 dB SNR across sentences of different lipreading difficulty. These large ranges suggest that the ability to exploit lipreading cues varies substantially across individuals.

Similar conclusions were drawn by Middelweerd & Plomp, (1987). In this speech-in-noise study, the authors confirmed an audio-visual benefit conferred by lipreading cues. However, they also showed that participant performance variance increased in the audio-visual condition, compared to the audio-only condition. Namely, the standard deviations reported increased from 1 dB SNR for the audio-only condition, to 2.4 dB SNR for the audio-visual. The authors interpreted these results as indicative that some individuals were substantially better than others in exploiting lipreading cues.

Large variations in the lipreading-based audio-visual benefit have also been reported by (Grant et al., 1998), considering participants with hearing loss. The authors reported an audio-visual benefit (measured as a percentage difference between audio-visual and audio-only conditions), that ranged from 8.5% to 83% across different subjects. Furthermore, the study of Sommers et al., (2005) showed that lipreading-based audio-visual benefits vary with participant age.

These studies suggest that there is substantial variation in the audio-visual benefit individuals can obtain by using lipreading cues. The variation, however, could be reflective of either participants' capacity to integrate the audio-visual information, or of participants' lipreading abilities (or both). Multi-sensory integration has been known to be impacted negatively by age (Pepper & Nuttall, 2023). Age is, however, more generally associated with impaired cognitive ability, including reductions in processing speed and working memory (Nettelbeck & Burns, 2010), and these effects might encapsulate lipreading ability as well. Indeed, Middelweerd & Plomp (1987), and Tye-Murray et al., (2016), found differences in lipreading (visual-only tasks) performances between young and older individuals. The latter study reported a range of percentage correct values from 10% to 90%, across the age groups. Nonetheless, lipreading ability has been shown to vary substantially across individuals within the same age group too, including adult individuals with both normal, and impaired hearing (Campbell et al., 2003), as well as children (Lyxell & Holmberg, 2000).

1.3.2.3.2. Training

Despite individual differences in lipreading skill, evidence suggests that all sighted individuals from every culture use lipreading information (Rosenblum, 2008). And, given its contributions to listening in noise, it would constitute a skill worth cultivating, especially for people with hearing loss. It remains unclear however whether training regimes can reliably make a difference in real-word lipreading performance (Campbell & Mohammed, 2010), although a recent meta-analysis of training experiments suggested that lipreading training can be successful (Bernstein et al., 2022). Campbell and Mohammed, (2010) outline a list of "traits of good lipreaders", and, albeit only indirectly related to the trainability question, I found the list insightful. I reproduce it below.

Traits that make a good lipreader:

- Exposure to, and experience with lipreading (e.g. being deaf, or living with a deaf person).

- Familiarity with the talker, including their accent and speech style.
- Good knowledge of the spoken language, and especially of its vocabulary – also, good reading skills.
- Visual acuity.
- Good verbal short term memory.
- Personality traits rendering the person more likely to attempt to lipread (e.g. risk-taking traits).

1.4. Audio-visual binding: A language-independent mechanism for audio-visual benefit

In the chapter's introductory section (1.1), I stated that one of my goals for the current chapter is to distinguish between two contributors to the audio-visual benefit. I briefly mentioned that one of them is lipreading, and the other includes low-level processes that involve temporal coherence between audio-visual sensory cues. In section 1.3, I discussed lipreading, and provided evidence that it does indeed contribute to the audio-visual benefit: it does so through providing visually-conveyed phonetic information. For that reason, it was dubbed a language-based mechanism for audio-visual benefit. I also illustrated (section 1.3.3.1), through signal detection theoretical considerations, how lipreading cues might bias a listener's decision of what they heard, without altering their perceptual representations of what they heard. In these ways, I concluded, lipreading constitutes a schema-based process for influencing auditory scene analysis (section 1.2.2.2).

In the following sections, I shift the attention to the mechanisms of temporal coherence, with an emphasis on that of audio-visual binding. Contrary to lipreading, audio-visual binding, although still reliant on visual cues from the talker's mouth, is a language-independent mechanism for the audio-visual benefit. It is also a bottom-up (or primitive, as Bregman would call it), mechanism of audio-visual integration, which operates early on to change the perceptual representation of the auditory objects formed.

Namely, in the context of speech perception, audio-visual binding exploits the temporal coherence between the amplitude envelope of speech, and that of the opening and closing of the mouth, to combine the auditory object of the speech sound, with the visual object of the mouth into one, unified, and enhanced audio-visual object (Bizley et al., 2016). Thus, audio-visual binding operates on larger temporal windows and is encompassed within the scope of the aforementioned sequential cues for auditory stream segregation (section 1.2.1).

It is thought that it is easier to segregate the voice of a talker from the background noise when it is part of an enhanced audio-visual object. Nonetheless, as we will see, it has proven difficult to isolate its effect from that of top-down audio-visual integration mechanisms, although there has been growing evidence supporting its existence (Atilgan et al., 2018; Atilgan & Bizley, 2021; Maddox et al., 2015).

I have talked about auditory objects, and auditory scene analysis, in section 1.2. To bring visual objects into the framework, before addressing audio-visual objects, I begin the current discussion with a brief section on the process of vision segmentation, which is the visual analogue of auditory scene analysis. Next, as with lipreading, I take a signal detection theoretical approach to show how audio-visual binding might influence auditory perception. I then outline the experimental difficulties and considerations surrounding its isolation from top-down, decision-biasing forms of audio-visual integration. I follow with a discussion of human psychophysical studies, and make a distinction between the ones that failed, and the ones that were successful in isolating audio-visual binding. I continue with a discussion of neurophysiological studies providing additional evidence for its existence. Finally, I conclude with speech-in-noise studies providing potential evidence for audio-visual binding contributing to the audio-visual benefit.

1.4.1. Object formation in vision: Vision segmentation

Previously, I discussed object formation in audition through the processes of auditory scene analysis (section 1.2). In the current section, I discuss how the visual system forms visual objects through the processes of vision (or image) segmentation. This research topic is, however, broad, and complex, and it is beyond the scope of this thesis to provide a thorough description of it. Instead, I opted to provide here a brief description of its principles, following the textbook *Visual Science, by Palmer, (1999)*.

The rules of visual perceptual organisation have been inspired by studies conducted by Gestalt psychologists in the early 20th century. Gestalt (German word for “shape” or “form”) psychology focused on understanding how people perceive patterns and objects in their visual environment, and on identifying our innate tendencies to organise visual sensory information in particular ways. According to these studies, and much like auditory objects, visual objects arise in perception by the grouping of similar visual sensory features, and segregation of disparate ones. The principles behind this perceptual organisation are the following:

- *Proximity*: Elements that are close to each other in space, or time, are likely to be perceived as part of the same group.
- *Similarity*: Elements with a common characteristic, like brightness, are inclined to be seen as belonging together.
- *Connectedness*: A continuous area, such as a blackboard, is usually perceived as a unified piece.
- *Memory*: Elements associated with a common memory tend to be grouped together in perception.

These principles are not dissimilar to those of auditory scene analysis, and Bregman was, in fact, inspired by the Gestalt principles of vision organisation when devising his theory of auditory scene analysis (Bregman, 1990, e.g. Chapter 1, page 24). Nonetheless, using these principles, the visual system innately segments the visual scene into its component objects (see, for example Pasupathy, 2015, for relevant work on monkeys).

Going back to the topic of listening in noise, when visual cues of a talker’s mouth are available, our visual system will form a visual object of that mouth. And our auditory system will form an auditory object of the talker’s voice. The next section discusses, through the lens of signal detection theory, how the two may be combined perceptually, via audio-visual binding, to form an audio-visual object that influences a listener’s auditory responses.

1.4.2. Audio-visual binding influences the listener’s perception

When discussing the workings of auditory scene analysis, I mentioned that temporal coherence between the acoustic features of a sound (e.g. timbre, amplitude modulations), is potentially the connecting link that binds these features together and helps the brain assign them to a particular auditory stream (and segregate that stream from others). Shamma et al., (2011) extended this concept to perceptual features coming from other senses, postulating that temporal coherence between any such features, may lead to their binding into unified objects.

Considering the audio-visual framework, what might these temporally coherent features be? Research on human speech has shown that the amplitude envelope of an individual's voice is temporally correlated with the size of their mouth opening (Chandrasekaran et al., 2009; Grant, 2001). This temporal coherence is believed to be the foundation for the mechanism of audio-visual binding for audio-visual object formation (Bizley et al., 2016; Lee et al., 2019).

As mentioned in previous sections, auditory, visual, and now by extension, audio-visual, objects are perceptual objects. That is, they arise through primitive, bottom-up, and unconscious procedures. Thus, contrary to lipreading cues, which bias the listener's response about what was said, without altering their perception of what was said (section 1.3.3.1), audio-visual binding does influence the listener's perception. Furthermore, since audio-visual binding operates on longer time scales, different examples of stimuli must be used here for the signal detection theoretical illustration of this perceptual change.

When I discussed, in section 1.3.3.1, how visual cues from lipreading might influence a listener's auditory experience, I used the word stimuli "cap" and "cat" in my illustration. For the illustration of audio-visual binding influencing perception, to make the stimuli longer, I simply incorporated them into sentences: "I wish I owned a cap", and "I wish I owned a cat". Below I describe an example where the listener is presented with the former sentence.

Figure 1.6 (A) (created using artificial data) illustrates how the two stimuli are represented perceptually, in the absence of visual cues, as shown before (Figure 1.4) for the simpler stimuli "cap" and "cat". The principles explained in section 1.3.3.1 are the same here. Once again, it can be seen that stimuli overlap in auditory perception and are thus easily confused with one another. Also, as before, the vertical line represents the listener's decision criterion. In this example, having been presented with the sentence "I wish I owned a cap", the listener will report having heard the sentence if the auditory perceptual evidence for having heard it falls to the left of the decision criterion. The chance of this happening is equal to the area under the blue distribution, and to the left of the decision criterion.

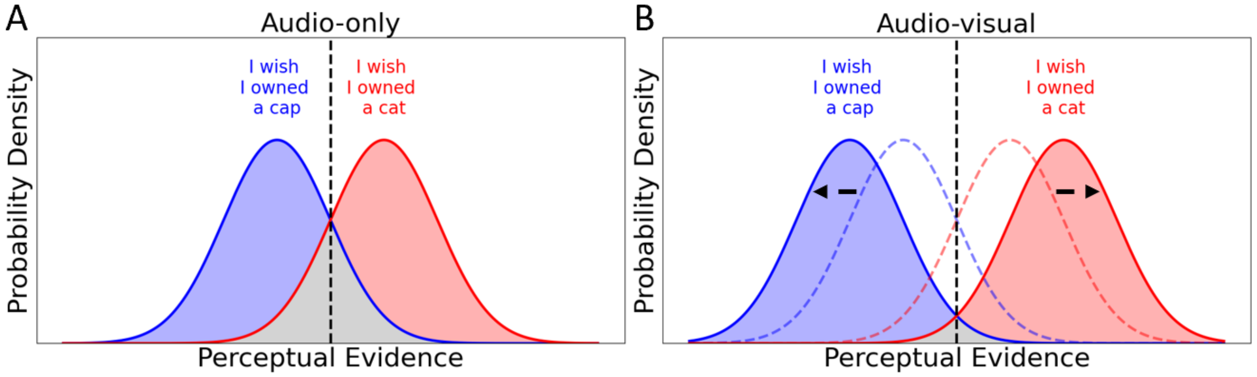


Figure 1.6*: The effect of audio-visual binding on auditory stimulus perception. Colour-coding is the same as in Figure 1.4. (A) Audio-only condition: The listener is presented with the sentence "I wish I owned a cap", the probability of reporting having heard the presented sentence is equal to the area under the blue distribution, and to the left of the decision criterion. (B) Audio-visual condition: The listener is once more presented with the sentence "I wish I owned a cap". Once again, the probability of*

the listener reporting having heard the sentence presented is equal to the area under the blue distribution, and to the left of the decision criterion. In the audio-visual case, the formation of the audio-visual object has caused a left-ward shift of the blue distribution and an equivalent right-ward shift of the red, while leaving the decision criterion unaffected. This results in an increase in the area of the blue distribution to the left of the decision criterion, and thus an increase in the probability of the listener reporting to have heard the presented sentence. Artificial data were used for the creation of this figure. See also Bizley et al. (2016) for more information on the theoretical concepts on which this graph is based on.

Figure 1.6 (B) introduces visual cues, namely the talker's mouth, into the sensory mix. When visual cues are made available, audio-visual binding effects start building up and, by the end of the sentence, an audio-visual object has formed, enhancing the perceptual representation of the otherwise unimodal audio-object in the listener's brain. This can be schematically demonstrated with a left-ward shift of the blue perceptual distribution, and a concurrent rightward shift in the red (the discriminability between the two stimuli has increased). Notice, however, that contrary to the case of lipreading cues (Figure 1.5 (B)), the decision criterion has been left unaltered here. Nonetheless, this shifting of the distributions, results in an increase in the probability of the listener's perceptual evidence falling to the left of the decision criterion, and thus of the probability of him reporting having heard the sentence presented.

Of course, the argument could be made that visual cues for the stimuli used in this hypothetical scenario contribute both lipreading, and temporal information. And, thus, it would be impossible to know whether it was just the perceptual distributions that were influenced, and not the decision criterion. Conversely, when we are conducting psychophysical experiments of speech-in-noise perception, and observe audio-visual improvements in participants' performances, how can we be certain the improvements we see are attributable to audio-visual binding? The answer is, we can't, unless we carefully design our experiment to control for these confounds. In the next section, I outline the experimental considerations that would allow, in principle, for the isolation of the effects of audio-visual binding.

1.4.3. Experimental considerations for the isolation of audio-visual binding effects

Bizley et al., (2016) argue that there are two key factors to consider when designing a psychophysical experiment to demonstrate binding effects. The first is to design the experiment such that it tests participants on a feature of the stimulus that is *orthogonal* to (i.e. independent of) the features that are creating the binding. The second is to make the experiment sufficiently hard, by introducing competition in the stimuli. I expand on the former factor first.

1.4.3.1. Orthogonal features

The expectation here is that there will be an enhancement in participant performance in the experimental task in the audio-visual condition, when an audio-visual object has been formed. But the goal is to ensure that the stimulus features involved in the creation of the audio-visual object do not convey task-relevant information that would bias participant responses (i.e. by influencing the participants' decision criteria; see Micheyl & Oxenham, 2010 for an example of bias-free assessment of unimodal auditory objects). Then, the enhancement measured can be attributed to audio-visual binding-driven perceptual changes.

Orthogonality between the task-relevant stimulus feature and the features creating the binding ensures that such decision biasing influences are ruled out. As mentioned in the previous section, the stimulus features that are involved in creating the binding within the context of audio-visual speech, are the amplitude envelope of the voice, and the opening and closing of the mouth of the talker. These are the features that are temporally coherent with one another, and are, effectively, measures of sound loudness, and geometric size. Auditory and visual examples of physical features orthogonal to these are frequency, and brightness, respectively.

Inspired by these natural statistics of speech, to experimentally demonstrate audio-visual binding, one could design an experiment where the participants are presented, in some form of background noise, with a simple pure tone, whose loudness is changing coherently with the size of a geometric shape, and the task is to identify brief frequency modulations of the tone (see for example Maddox et al., 2015, also discussed in detail in section 1.4.4.2). Then, participant performance differences in the audio-visual case, compared to, say, an audio-only case where visual stimulus is omitted, would be reflective of the effects of binding. Notably, performance in the audio-visual case would be expected to be enhanced: As described in section 1.2.2.2, one of the axioms of object-based attention theories is that attention operates on the object as a whole. In the experiment described above, the tone and the size-varying geometric shape are both components of the same audio-visual object. Thus, when the listeners attend the size-varying geometric shape while listening to the tone, the frequency features of that object should also be enhanced perceptually, compared to the audio-only case.

1.4.3.2. Stimulus competition

The second consideration suggested by Bizley et al., (2016) for experimental demonstration of audio-visual binding was the introduction of stimulus competition. “Competition” here means the exploitation of *maskers* (or distractors), as a form of background noise, and the incorporation of a so-called masker-coherent condition into the experimental pipeline. Firstly, Bizley et al., (2016) argue that introduction of stimulus competition in the form of maskers makes the task more naturalistic and taxing. This would increase the perceptual benefit of audio-visual binding, thus making it easier to measure experimentally. The incorporation of a masker condition in the task helps detecting audio-visual binding effects in an additional way, described below.

Consider an experimental task where the participants must make a judgment on a specific auditory stimulus (the target) presented among several other similar auditory stimuli (the maskers). At the same time participants are provided with a video of a visual stimulus that can either be coherent with the target auditory stimulus (target-coherent condition), or with one of the maskers (masker-coherent). They are asked to attend the video, regardless of coherence type, while making judgments on the target audio. Then, target-coherent performances are compared with the masker-coherent ones.

Object-based attention theory states that, when attention is divided between two perceptual objects, with the task-relevant object among the two, it results in a decrease in performance compared to when attention is focused on just the task-relevant object (Bizley et al., 2016). In the aforementioned scenario, participants would be attending a single audio-visual object – the result of binding – in the target-coherent condition, whereas they would be dividing their attention between two objects (the target audio, and the masker video) in the masker-coherent condition. Hence the difference in performance between the two conditions would constitute a measure of audio-visual binding.

On the topic of attention, there is some debate as to whether it can actually be "divided" between different talkers (Best et al., 2006; Broadbent, 1954), or not (Cherry, 1953). If it cannot, then it would simply shift from one object to the other (Shinn-Cunningham et al., 2017). Nonetheless, this does not weaken the usefulness of the target- and masker-coherent conditions. If attention shifts, instead of being divided, then the performance costs of the masker-coherent conditions would be due to constant shifting between the masker-related object and the target sentence. In fact, the masker-related object, would actually consist of a prominent audio-visual object, resulting from the binding of the masker video and respective audio. Thus, being a more salient object, it would be very difficult to ignore (Talsma et al., 2010), and volitionally steer one's attention away from it and towards the target sentence.

1.4.4. Examples of studies failing to or succeeding in isolating audio-visual binding effects

Visual cues supplementing audition provide a benefit to listeners; some of these benefits are due to decision-biasing, and top-down processes such as lipreading (see also section 1.3.3.1), and some are due to processes altering early auditory perception, such as audiovisual binding. In section 1.4.2 I explained that it is difficult to experimentally isolate the audio-visual binding effects, from top-down effects. Then, in section 1.4.3, I outlined two principles, orthogonality of features, and stimulus competition, that are believed to constitute control measures that would enable the experimental demonstration of the binding effects. Using these considerations as my foundation, I now present examples of psychophysical studies that have failed or succeeded in doing so.

1.4.4.1. Studies that have failed to isolate the effects of audio-visual binding

I first consider the study of Rahne et al., (2007). In this study, the authors presented their participants with a variation of the ABA auditory paradigm (described in section 1.2.1), alongside a visual stimulus (Figure 1.7). Namely, the auditory stimulus consisted of two sets of frequencies, one low set and one high set, which were separated enough in frequency with each other so that they fell in the "bistable" range of stream segregation. As mentioned in section 1.2.1, stream segregation is ambiguous within this range of separation, and the participants alternate between stream segregation and stream integration.

The low frequency set consisted of three tones, L1, L2, and L3, of increasing frequency (i.e. L2 had a higher frequency than L1, and L3 higher than L2). These were normally presented in order of increasing frequency (i.e. L1 then L2 then L3), with the exception of a deviant presentation where they were presented in decreasing frequency order. The high frequency set also consisted of three tones with different frequencies. These were presented in a random order. Further, every third tone within each set was presented at a higher loudness than the preceding, and succeeding, tones (indicated with blue squares in Figure 1.7).

The visual stimulus included two conditions. In one, a square was presented whose size increased and decreased following the frequencies of the low frequency set of tones (i.e. it would start with a given size when L1 was presented, and grow larger with presentations of L2, and L3). In the second visual condition, a circle was presented whose size would alternate in synchrony with the higher intensity tone presentations. Participants were asked to attend the visual stimuli, and ignore the auditory, regardless of condition. The authors showed, using mismatched-negativity components of event-related potentials, that the first visual condition resulted in enhanced auditory stream segregation, whereas the second visual condition resulted in auditory stream integration.

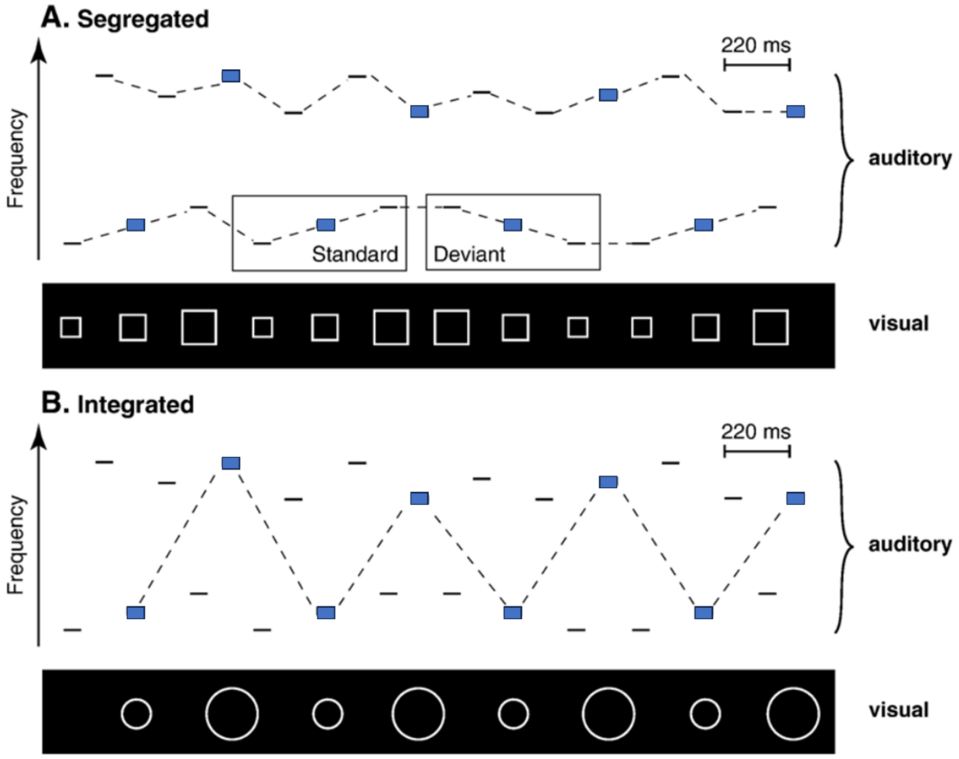


***Figure 1.7**: Audio-visual stimuli used in the study of Rahne et al., (2007). Participants were presented with two sets of tones: A low frequency set, and a high frequency set. The tones of the low frequency set would be presented in increasing order of frequency, with the exception of deviant presentations where the tones would be presented in decreasing order of frequency. The tones of the high frequency also differed in their frequencies but would be presented in random order. For both low and high sets, every third tone would be presented at a higher loudness than the tones preceding and succeeding it (blue rectangles). Visual stimuli, either squares or circles of varying sizes would accompany the tone presentations, depending on the condition. (A) In the stream segregation condition, the visual stimulus was a set of squares which increased and decreased in size in accordance with the frequencies of the tones of the low frequency set. This promoted low and high frequency set stream segregation. (B) In the stream integration condition, the visual stimulus was a set of circles, changing in size in synchrony with the loud tone presentations. This visual cue promoted low and high frequency set stream integration.*

Given these findings, the authors concluded that they provided evidence of an early level (bottom-up) audio-visual interaction (along the lines of audio-visual binding). I argue, however, that although their results may also include the effects of audio-visual binding, they have not succeeded in isolating the binding effect from decision-biasing effects. I conclude that on the basis of their visual stimuli providing information on the organisation of their experimental auditory scene – i.e. information on which stream to follow.

Another audio-visual phenomenon that is often suggested to underline audio-visual binding is that of the ventriloquist illusion. Ventriloquism is the art of speaking without moving the lips, and ventriloquists, by moving the mouth of a puppet while they speak, create the illusion that their voice is coming from the puppet. As the auditory perception is influenced by the visual cues, the ventriloquist illusion is endorsed as proof for audio-visual binding.

Alais and Burr, (2004), however, showed that observers of the illusion compute the location of the voice according to the relative reliability of the auditory and visual cues. Namely, the authors showed that rendering the visual cues unreliable by blurring the visual stimuli is sufficient to undo the illusion. Further, when observers of the illusion are explicitly asked to provide the location of both the sound and the visual cue, the illusion seems to become weaker (Alais & Burr, 2019). These findings suggest that the audio and visual information provided by the ventriloquist are processed independently, and integrated together at a later, decision-making stage. Thus, the findings speak against early-stage integration at the perceptual level. Similar arguments have been made against other illusions reflecting audio-visual binding effects. See, for example Bizley et al., (2016) and Lee et al., (2019) where the authors discuss the cases of the McGurk effect, and the sound-induced flash illusion.

In the following section, I present the study of Maddox et al., (2015) as a good example of a psychophysical study that successfully managed to isolate audio-visual binding effects.

1.4.4.2. A study that has successfully isolated the effects of audio-visual binding

The goal of Maddox et al., (2015) was to isolate the effects of audio-visual binding in an auditory selective attention paradigm. The authors designed a psychophysical experiment that incorporated both the elements of stimulus competition, and orthogonality of features discussed in section 1.4.3. They employed two variants of the same experimental task, one where they assessed their participants' ability to detect pitch modulations in pure tone stimuli, and another where the assessment was on timbre modulations in artificial vowels, both of which produced similar results. I outline the pitch-task below (also in Figure 1.8):

The authors presented their participants with two pure tones, one intended to be the target tone, with a frequency of 440 Hz, and the other intended to be the masker, with a frequency of 565 Hz (Figure 1.8 (C)). The two tones were independently amplitude modulated at a rate varying between 0 and 7 Hz (Figure 1.8 (A)). A visual stimulus, in the form of a circle accompanied the audio. This circle would vary in size, following either the target audio amplitude envelope (target-coherent), or the masker (masker-coherent), or changing independently of the two (independent) (Figure 1.8 (B)). This temporal coherence, the authors conjectured, would lead to the formation of audio-visual objects (between the target audio and circle in the target-coherent, and between the masker audio and circle in the masker coherent).

The authors also introduced brief frequency modulations at random time-points in both the target and masker tones (Figure 1.8 (C)). Their participants' task was to look at the circle video, and to follow the target tone and report when it exhibited modulations in frequency.

Participant performance, the authors report, was significantly greater in the target-coherent condition, compared to the masker-coherent, in terms of both discriminability and hit rate measures. The authors attributed this difference in performance to audio-visual binding effects between the target audio and the visual stimulus: they showed that when auditory amplitude was coherent to visual size, the perceptual representation of frequency was also enhanced as all three features are elements of the same audio-visual object. Notably, they explain, the features of the stimulus creating the binding (i.e. the amplitude envelope of the tone, and the size of the circle), were orthogonal to the changes in frequency that the participants had to report – ruling out decision-bias influences.

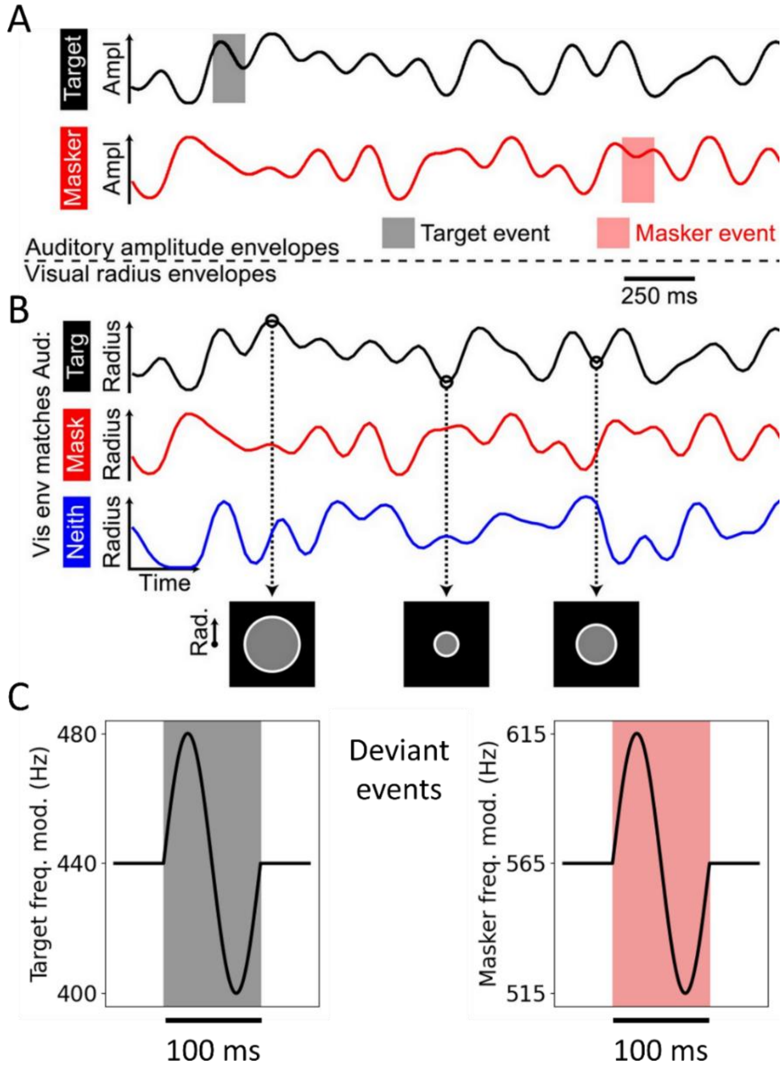


***Figure 1.8**: Illustration of the pitch-task* variant of the *experimental paradigm of Maddox et al., 2015. (A) The authors used two auditory stimuli, a target (black trace) and a masker (red trace), whose amplitude envelope (shown here) was modulated independently. The target audio was a pure tone at 440 Hz, and the masker audio was a pure tone at 565 Hz. (B) Visual cues: A disk would vary in size, following the amplitude envelope of either the target audio (black trace), or the masker audio (red trace), or independently (blue trace). (C) Target (left) and masker (right) deviant events (highlighted in (A)). The frequency of the target and masker tones would exhibit, random in time, 100 ms-long modulations as shown. Participants were tasked with detecting when these modulations occurred in the target audio.*

The effect sizes captured by the authors in the comparisons across the three conditions were, however, generally small. And, unfortunately, the authors were not able to show a significant difference between the target-coherent, and independent condition, although measures of discriminability and hit rate performances were, on average, lower for the independent condition compared to target (and higher compared to the masker-coherent condition). Thus, the lack of statistical significance there might be reflective of a lack of statistical power. Finally, the authors did not include an audio-only condition in their experiments, which would be necessary to uncover an audio-visual binding-specific audio-visual benefit, by comparison with the target-coherent audio-visual condition. Nonetheless, the report of Maddox et al., (2015), was among the first papers to have provided strong evidence of existence of audio-visual binding through a psychophysical experiment.

The neural substrate for audio-visual binding has been hinted to through the results of neurophysiological studies. These studies also provided additional evidence for the existence of this integration mechanism. They form the next topic of discussion.

1.4.5. Evidence of the existence of audio-visual binding from neurophysiological studies

The auditory cortex is believed to constitute the neurological substrate of auditory scene analysis (Nelken, 2004; Nelken & Bar-Yosef, 2008). Supporting this claim, (Micheyl et al., 2005) showed that the primary features of auditory stream segregation observed in human psychophysical experiments can also be quantitatively captured via single-neuron responses measured from the primary auditory cortex of monkeys listening to the ABA paradigm. Further, Fishman et al., (2016), showed that monkeys are able to discriminate between concurrently presented vowel stimuli, at the level of the primary auditory cortex.

Generally, however, the auditory cortex is not considered to be as "neurologically isolated" as it was once thought to be. On the contrary several publications have shown that it receives inputs from the visual brain areas (Bizley et al., 2007; Budinger et al., 2006; Calvert et al., 1997; Falchier et al., 2002; Kayser et al., 2008). For example, Calvert et al., (1997) showed, with fMRI measurements, that silent lipreading cues are sufficient to activate auditory cortical areas in humans. Further, Bizley et al., (2007), working with ferrets, reported that approximately 15% of single unit recordings within the primary auditory cortex responded to visual stimuli; this percentage increased up to ~50% for secondary areas.

Thus, auditory scene analysis occurs in the auditory cortex, and the auditory cortex receives several inputs, and processes information from, visual areas. Further, audio-visual binding influences auditory scene analysis. These accounts formed the rationale for the authors of Atilgan et al., (2018) to search for a neural substrate of audio-visual binding within the auditory cortex.

The authors employed a variation of the experiment of Maddox et al., (2015), and combined it with neurophysiological recordings from the ferret auditory cortex. Their stimuli consisted of two artificial vowels they called A1 and A2 (A1 = [u], with fundamental frequency at 175 Hz; A2= [a], with fundamental frequency at 195 Hz), and their visual cue was a full-field visual stimulus. Similarly to Maddox et al., (2015), the luminance of the visual stimulus would change, either temporally coherently with the amplitude envelope changes of the auditory stimuli, or independently. The authors also used auditory-only and visual-only stimuli.

Both awake, and anesthetised ferrets were presented with the stimuli, to control for attentional effects potentially confounding audio-visual binding. The authors reported three main findings, applicable to both awake and anesthetised animals: 1) Visual stimuli significantly influenced the phase of the local field potentials recorded in the auditory cortex (irrespective of the type of temporal coherence). 2) The synchronisation of auditory and visual stimuli affected how mixed sounds were processed in the auditory cortex, leading to a more consistent neural representation of the audio-visual stimuli that were temporally coherent. 3) The spiking response to variations in auditory timbre (an attribute not involved in the creation of the binding) was found to be stronger in audio-visual stimuli that were temporally coherent.

Since the visual stimuli influenced the neural representation of non-binding features, the results of Atilgan et al., (2018) were consistent with the formation of audio-visual objects in the auditory cortex. Furthermore, since these effects were applicable to both awake and anesthetised animals, the authors have provided evidence that audio-visual binding can occur independently of top-down attentional influences. They did, however, report a difference between awake and anesthetised animals. In awake animals only, they observed a significant increase in the reliability of the local field potential phase (at 10.5 – 12.5 Hz), when the visual stimuli were temporally coherent to the auditory stimuli, compared to when the two were independent.

This state-dependent finding supports the implication of an attentional (or other top-down) signal serving to enhance the neural representations in the awake state. Translating this to speech perception in humans, it provides an alternative mechanism, to that of audio-visual binding, for how temporal coherence cues may offer a visual enhancement. Namely, it is possible that a talker’s mouth, being temporally coherent with the auditory amplitude envelope of the talker’s voice, provides temporal cues that hint the listener to *when* the voice’s loudness will change. And thereby directing the listener’s attention to this feature of the talker’s voice (see also Ding & Simon, 2012 and Peelle & Sommers, 2015). Since attention operates on objects (Shinn-Cunningham, 2008), when the listener attends the loudness of the voice, the rest of its features will also be enhanced, including the talker’s pitch and so on, helping them segregate it from the sound mix. Notably this attention-based mechanism does not necessitate the formation of an audio-visual object in the listener’s brain – the visual cue simply hints to the auditory loudness, which then hints to the remaining features of the voice – although it would still be independent of linguistic cues.

Nonetheless, since the authors have shown that audio-visual objects were indeed being formed in the auditory cortex of the anesthetised animal as well, it is likely that the two mechanisms (attention, and audio-visual binding) both make independent (and possibly interacting) contributions. Overall, given these results, the authors concluded that the role of early integration of audio-visual information in the auditory cortex is to support auditory scene analysis.

1.4.6. Audio-visual temporal coherence contributes to the audio-visual benefit when listening in noise, and is trainable

1.4.6.1. Evidence of audio-visual temporal coherence contributing to the audio-visual benefit from speech-in-noise studies

So far, in section 1.4, I have outlined the concept of audio-visual binding, and explained how it might influence listeners on a perceptual level (section 1.4.2). Then I discussed the experimental difficulties underlying its isolation from top-down influences (section 1.4.2) and described a set of experimental considerations (orthogonality of features, and stimulus competition) that together constitute the strong test for the presence of audio-visual binding (section 1.4.3). Based on these considerations, I then described studies that failed to illustrate binding, and concluded with the study of Maddox et al., (2015) which succeeded in doing so (section 1.4.4). Finally, in section 1.4.5, I provided evidence from neurophysiological studies, with Atilgan et al., (2018) at the forefront, that audio-visual objects are indeed formed, their neural substrate is the auditory cortex, and they seem to support auditory scene analysis.

What I haven’t shown yet, is whether the mechanism of audio-visual binding, or more generally, language-independent, temporal coherence-related mechanisms, are able to provide an audio-visual benefit to listening in noise, as I have showed for lipreading in section 1.3.2.2. The problem with isolating the contributions of such mechanisms to the audio-visual benefit when the auditory stimuli are speech stimuli, is that they are inherently complex. How can we know for certain how speech interacts with a visual stimulus, and how the brain interprets these interactions – does the visual stimulus convey phonetic cues, in addition to temporal cues, and do temporal cues involve both attentional mechanisms in addition to audio-visual object formation? As discussed in earlier sections (1.3.3.1 and 1.4.2) it is difficult to isolate these effects. Nonetheless, two studies, Yuan et al., (2020) and (2021), have provided evidence towards the direction of temporal cues contributing to the audio-visual benefit.

Yuan and colleagues presented their participants with speech sentences in the background of multi-talker babble noise, at several different signal-to-noise ratios (SNRs). Their experiments included both audio-only, and audio-visual conditions. For the audio-visual conditions, participants were provided with the video of a sphere, whose size would change, following the amplitude envelope of the target sentence. In the audio-visual condition, and at certain SNR ratios, the performance was significantly better than in the audio-only condition. For example, at an SNR of -3 dB, performance in the audio-visual condition improved by up to 63%, and at an SNR of -1 dB, it improved by up to 87%, compared to the audio-only performance (Yuan et al., 2021).

Arguably, the sphere size changes could have been conveying phonetic information to the listeners in these experiments (each phoneme might be associated with a different mouth opening size). However, the experimenters showed that the audio-visual benefit was independent of whether the sphere was growing with increasing speech-amplitude envelope, or shrinking with increasing amplitude envelope, suggesting that the effect was likely related to the temporal cues. Thus, Yuan and colleagues have provided potential evidence to support that audio-visual temporal coherence contributes to the audio-visual benefit.

1.4.6.2. Evidence of audio-visual temporal coherence being trainable

As with lipreading, the benefits of temporal coherence cues to speech-in-noise perception would not be of great value to people who lack the ability to use them, unless this ability is trainable. Atilgan and Bizley, (2021) showed that such training is possible.

The authors presented their participants with the timbre-variant stimuli from Maddox et al., (2015) (the pitch-variant of which is described in section 1.4.4). Briefly, these included two artificial vowels that would vary in loudness, and a visual stimulus of a disk that would change in size coherently with the amplitude envelope of one of the vowels, or independently. They had their participants take the test twice, once prior to, and once subsequent of receiving a so-called audio-visual temporal coherence detection training.

The training consisted of presentations of pairs of audio-visual stimuli: In one of the pairs, the audio-visual stimuli were temporally coherent, and in the other they weren't. Participants had to report which of the two pairs included the temporally coherent stimuli. Further, they received five short training sessions (over the time period of 2 weeks) prior to re-running the task.

Results of participant performances prior to training resembled those of Maddox et al., (2015). After having received training, the participants of Atilgan and Bizley (2021), showed improvements in all three conditions (target-coherent, masker-coherent, and independent), compared to pre-training scores. Notably, post-training scores in the task were significantly enhanced compared to pre-training scores. Further, post-training performances in the target-coherent were significantly increased, compared to performances in the independent condition – something that Maddox et al., (2015) had failed to show.

These findings, the authors suggest, illustrate that the exploitation of temporal cues, for the formation of audio-visual objects is a trainable skill. Their conclusion offers hope, and scope for the application of temporal coherence training regimes in the clinic for the enhancement of the auditory experience of hearing loss populations.

1.5. Assessment of the audio-visual benefit and its contributors

For the contributors of the audio-visual benefit to be examined, the audio-visual benefit conferred by looking at a talker's mouth movements when listening to them, must first be assessed. As discussed on several occasions in this chapter, this benefit is most relevant when listening in noise, and, accordingly, has been assessed with speech-in-noise tests. However, based on the discussion so far, we might also conclude that it would be challenging to capture both the listening benefit contributions of audio-visual binding (or temporal cues more generally), and lipreading cues, with a speech-in-noise test. To begin with, the speech-in-noise tests employed in the citations we have seen so far were not specifically tailored such that their audio-visual condition stroke a balance between lipreading and temporal coherence components. Yuan et al., (2020) and (2021), for example (arguably) excluded lipreading cues from their stimuli; on the other hand, reports like Sumby and Pollack (1954), using simple word stimuli, were likely capturing only lipreading cues.

Thus, in order for the speech-in-noise test to capture both lipreading and audio-visual binding/temporal cue contributions, it must be specifically designed to do so. Furthermore, the speech-in-noise task should be suitable for use in the clinic, in addition to the research sector, for the science to have a translational impact. In anticipation of my own speech-in-noise test's development (Chapter 2), for which I sought to apply these design features, in the current section I discuss the different "flavours" of speech-in-noise tasks. Additionally, I discuss what the ideal properties would be, for a speech-in-noise test that measures both lipreading and temporal coherence effects.

Next, I make the argument that clinicians would benefit from inclusion of tests assessing visual contributions to listening to their diagnostic toolset. These include speech-in-noise tests that collectively capture these visual contributions, and tests that independently measure the patients' visual contributors, for targeted training guidance. To this end, I follow with a discussion of lipreading, and temporal coherence assessors. Finally, I provide some examples of audio-only speech-in-noise tests that have been employed in the clinic for the assessment of hearing loss. I conclude with a brief mention of the audio-only speech-in-noise task of Messaoud-Galusi et al., (2011) as a segway to the aims, and main chapters of my thesis.

1.5.1. Speech-in-noise test variations

There are various types of speech-in-noise tests that could be used to measure an audio-visual benefit, deriving from different combinations of their features. These features include the different conditions they include, the types of background noise they employ, the way in which the signal-to-noise ratio varies (or not) during the test, the type of stimuli used, and the internal structure of these stimuli. In this section, I expand on each of these parameters and on how they can be exploited to construct a speech-in-noise task that strikes a balance between capturing both lipreading, and temporal cues. To avoid repetition, such a speech-in-noise test is, in the following sections, referred to as a "comprehensive speech-in-noise test".

1.5.1.1. Types of background noise

There are several different types of noise that speech-in-noise tests can use as background. I outline some commonly used ones below:

- White noise: White noise contains all the frequencies at equal intensity, thus producing a hissing sound that is consistent across the human audibility spectrum. It is often used to provide a neutral background noise in the masking of other sounds.
- Pink noise: Pink noise has intensity that varies across the frequency spectrum. Namely, it has equal energy per octave, with an intensity decrease by 3 dB SPL per octave as frequency increases. Its frequency spectrum more closely resembles that of human speech, making it useful for evaluating hearing and speech understanding in an environment that simulates natural listening conditions more closely than white noise.
- Speech noise: This noise mimics the frequency spectrum of human speech, but without containing intelligible words, creating a continuous sound that resembles talking without discernible language. Speech noise, thus, is employed to challenge the listener's ability to discern speech from a background that has similar acoustic properties to speech. It's particularly useful in speech audiometry to assess speech recognition thresholds and speech intelligibility in a controlled manner.
- Multi-talker babble: This type of noise is generated by blending the voices of multiple speakers, creating a simulation of an unintelligible, crowded environment sound. It is ideal for assessing speech perception in crowded social settings. Multi-talker babble noise can vary in the number of background talkers it uses, in the stimuli they produce, and in the gender of these talkers.
- Competing talkers: Competing talker noise involves background speech that is more focused, usually coming from one or a few talkers, and can be at least partially intelligible. This type of noise creates a scenario where the listener must distinguish the target speech from other speech sources that are also meaningful. Competing talker noise is particularly challenging because the brain needs to process multiple auditory streams and select which one to attend to.

Of the listed, the optimal type of noise to use in a comprehensive speech-in-noise test would be the competing talker background. This is true for three reasons: First, it provides a perceptually taxing framework where visual contributors to listening become more apparent and measurable. Second, it provides the ground (masker talkers) for exploiting the element of stimulus competition in capturing audio-visual binding (section 1.4.3.2; see also next section on test conditions). Third, and this is especially true if masker and target talkers use similar stimuli (see also section 1.5.1.4), competing talker background provides energetic and informational masking of the target stimulus, requiring thus that the listener allocates greater reliance on vision, for the identification of the target (Helfer & Freyman, 2005).

A comprehensive speech-in-noise test ideally includes two background competing talkers. Studies have shown that performance in speech-in-noise tasks decreases when the number of background talkers is increased from one to two, and then improves again with additional talkers (Freyman et al., 2001, 2004; Rosen et al., 2013). Thus, two background talkers should be used, to avoid making the task too easy (in which case listeners may not require the visual cues to a substantial extent). Finally, the two background talkers should ideally be a male and a female, such that one of them always has the same gender as the target talker. This would further add to the challenge of the task by contributing further informational masking to the mix (Kidd & Colburn, 2017).

1.5.1.2. Task conditions

Clearly, in order for a comprehensive speech-in-noise task to be able to measure an audio-visual benefit, it should include both audio-visual, and audio-only conditions. The latter, however, would benefit from the incorporation of target-coherent, and masker-coherent sub-conditions. In the target-coherent sub-condition, the talker presented in the video would match to the voice uttering the target stimulus, whereas in the masker-coherent sub-condition, the talker would match the voice of one of the maskers. The rationale behind this is the same as that discussed previously (section 1.4). Briefly, the contrast between target-coherent and masker-coherent performances, and their respective comparisons to audio-only performances, could provide insights into the nature of the audio-visual benefit being captured.

1.5.1.3. Signal-to-noise ratio variation procedures

As far as signal-to-noise (SNR) variations go, there are two broad types of speech-in-noise tests: The fixed SNR tests, and the adaptive SNR tests (Taylor, 2003). In fixed SNR tests, experimenters simply choose a constant SNR, present their stimuli at that SNR, and measure their participants' performances as a percentage-correct, for that SNR. Then, they might change to different SNRs and repeat the process (Yuan et al., 2020, 2021, for example, used fixed SNR tests).

Adaptive SNR tests, on the other hand, measure participant responses as the SNR is varied. This can be done by varying the intensity of both the signal, the noise, or both. As a performance metric, instead of percentage-correct responses, adaptive SNR tests usually output the so-called speech-reception thresholds (SRTs) of participants. An SRT is the SNR at which the participant is expected to be able to identify a certain percentage of the stimuli correctly. For example, an SRT_{50} is the SNR at which the participant is expected to be able to identify 50% of the stimuli correctly.

These SRTs constitute different points on what is called the participant's psychometric curves, which, in the context of speech-in-noise tasks, graphically represent the relationship between SNR and the participant's performance (in percentage-correct responses) on each condition of the task (Figure 1.9). Several adaptive SNR speech-in-noise tests employ adaptive staircase procedure algorithms in varying their SNR. Levitt, (1971) provides a comprehensive review of adaptive staircase procedures, and the SRTs they output, and I provide a detailed explanation of these in Chapter 2, alongside the discussion of my own speech-in-noise task.

Both fixed SNR and adaptive SNR speech-in-noise methods have their merits and disadvantages. Fixed SNR tests, for example, allow the experimenter to sample several different fixed SNR values, and thus collect multiple points on their participants' psychometric functions. However, the disadvantage of this method is that the experimenter cannot know ad-hoc, which percentage point on the psychometric function each SNR matches to, and this may prove to be problematic experimentally when comparing performances between audio-visual and audio-only conditions. This is illustrated in Figure 1.9.

Figure 1.9. shows two psychometric curves of a hypothetical participant for two conditions of a speech-in-noise task: The audio-only (blue curve), and the audio-visual (red-curve). Clearly, as the provision of visual cues enhances participant performance, the participant's psychometric curve for the audio-visual condition would be shifted to the left, compared to his psychometric curve for the audio-only condition. Focusing, for example, on the SRT$_{50s}$ (marked point on each curve), we see that it is lower for the audio-visual condition compared to the respective SNR for the audio-only condition. The difference between the two would provide a measure of the audio-visual benefit (marked with a dotted line between the two points).

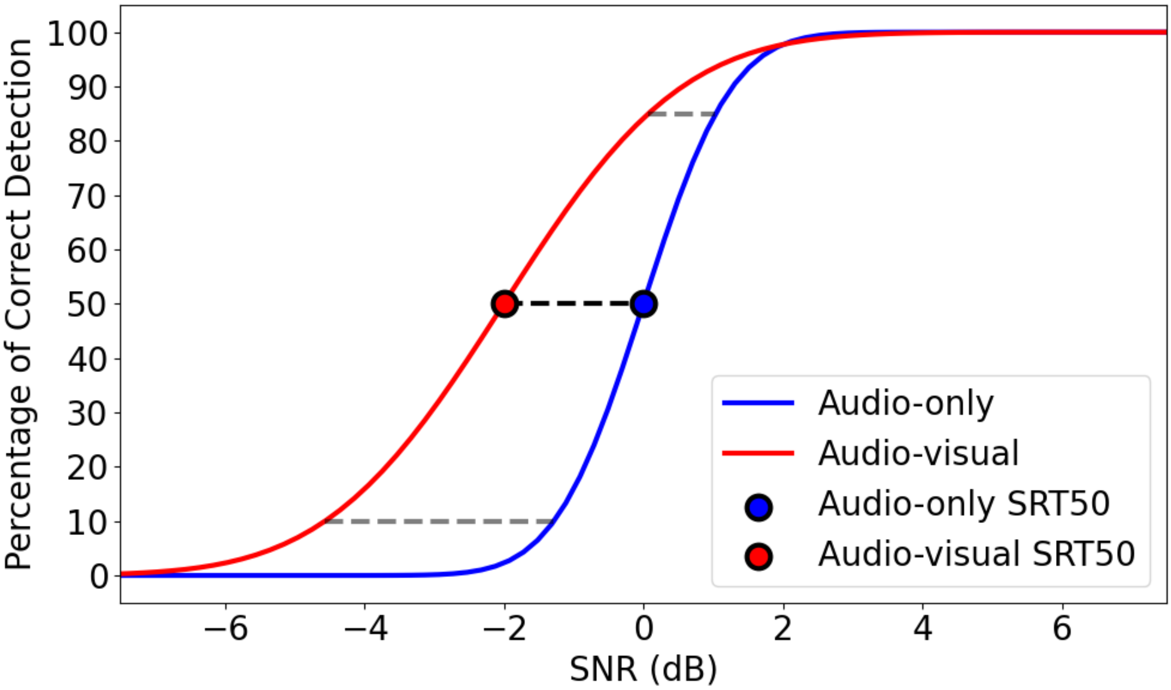


***Figure 1.9**: Illustration of psychometric curves obtained from a hypothetical speech-in-noise task. The task includes two conditions: An audio-only (blue curve), and an audio-visual (red curve) condition. The audio-visual curve is shifted to the left compared to the audio-only curve, to illustrate the boost in performance provided by the added visual cues. SRT$_{50}$ points for each condition are depicted with markers. The difference between the points of the two curves (depicted with dotted lines) reflect the audio-visual benefit the experimenter would measure from the two conditions. This audio-visual benefit is not constant along the range of SNRs (grey dotted lines). See also Blackburn et al. (2019), Levitt (1971), and Sumby and Pollack (1954) for examples and more information related to the topics on which this graph is based on.*

This difference should ideally be substantial, otherwise the speech-in-noise test will fail to measure a substantial audio-visual benefit. The difference is, however, unlikely to be constant for the two curves throughout the range of SNRs. As illustrated in Figure 1.9, at greater SNRs, the difference is smaller, and at lower SNRs, the difference is larger – visual cues are not very useful when the auditory signal is relatively clean but are relied upon almost exclusively when the auditory signal is very noisy. If the experimenter unknowingly samples SNRs that correspond to high percentage points, they will fail to capture a substantial audio-visual benefit. On the other hand, if they sample very low SNRs, they are more likely to capture audio-visual benefits that mostly reflect lipreading-dependent boosts in performance – temporal coherence mechanisms such as audio-visual binding require that the audio signal is reasonably audible, otherwise auditory objects cannot form. Ideally sampling somewhere in the middle is desired for a comprehensive speech-in-noise task.

To achieve this with fixed SNR tests, the experimenter might have to sample several different SNRs, which would be time-consuming – and thus also not appropriate for use in the clinical setting. Adaptive SNR methods on the other hand, which output a specific point on the psychometric curve that the experimenter can select ad-hoc, are ideal for use in a comprehensive speech-in-task. They also have the added advantage of achieving convergence to the SRTs relatively quickly (~10 minutes) and would thus be easier to use by clinicians as part of their diagnostic toolset.

1.5.1.4. Types of stimuli, and stimulus structures

Another important feature of a speech-in-noise task are the types of stimuli it employs. We saw for example in section 1.3.2.2 that several reports assessing the audio-visual benefit provided by lipreading used simple speech stimuli, such as words, or even consonants. Others, such as Yuan et al., (2020), (2021), used sentence stimuli for both the target and masker – although the maskers must have been unintelligible in these reports; the authors used an eight-person multi-talker babble as noise.

Temporal coherence effects such as those of audio-visual binding require relatively long timescales to take place (from hundreds of milliseconds to seconds; Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008). Thus, for a comprehensive speech-in-noise task that accounts for both the effects of lipreading and those of audio-visual temporal coherence, sentence stimuli should be used, for both target and masker talkers. Sentence stimuli also render the task more naturalistic, and hence what it measures more generalisable to real-world scenarios.

The structure of the sentences is also important, especially the comparative structures between target and masker sentences. It would be ideal, for example, if target and masker sentences had similar structure, as this would increase the difficulty of segregating between them (and thus increasing reliance on lipreading). The types of target words the participants are tasked with identifying within the sentence stimuli are also something to consider. Ideally these words should not be too easily lipreadable (to increase reliance on temporal cues). Furthermore, their position within the sentence is also important: For maximal exploitation of mechanisms that build over longer time scales, the words should be positioned towards the end of the sentence stimuli. Finally, the sentence stimuli of a comprehensive speech-in-noise task should be free of contextual cues that would enable participants to solve the task by guessing the target words (see for example Pichora-Fuller et al., 1995).

1.5.2. Speech-in-noise tests used in the clinical setting

It is relatively common practise for clinicians to employ speech-in-noise tests in assessing their patients' hearing loss (Sharma et al., 2016), but it isn't common to consider the effects of vision on the patients' ability to comprehend speech. According to a survey by (Walden et al., 2004), in most social situations, the talker of interest is positioned in front of the listener, thus making visual cues from the talker's face available. And, given what we have discussed so far about the positive influence of vision on hearing, I would argue that an adaptation, in the clinical setting, of a more "multisensory mindset" would be beneficial for the patients.

In the ideal scenario that I am considering, health-care professionals are equipped with speech-in-noise tests that include both audio-only, and audio-visual conditions so that they are able to assess their patients' audio-visual benefits. Further, they are provided with tools they can employ to measure the individual contributors, such as lipreading, and audio-visual binding, to these audio-visual benefits.

Then, based on these results, they are able to provide targeted training guidance to the patients, to help them learn to make use of their vision to aid their hearing.

I begin this section with a discussion on lipreading, and audio-visual temporal coherence assessments. Then I provide examples of audio-only speech-in-noise tasks that have also been used, or have the potential to be used in the clinic. These tests form a foundation onto which audio-visual settings can be integrated.

1.5.2.1. Tests for assessing lipreading and audio-visual binding

There are several silent lipreading assessment tests, which vary on a few dimensions including, importantly, their test material, target populations, availability of normative data, and suitability for use in both the academic and the clinical sectors (see Campbell and Mohammed, 2010 for a comprehensive list). In my own work, I selected the Test of Adult Speechreading (TAS; Campbell et al., 2003; Mohammed et al., 2006), as a test providing an optimal option along these dimensions. I describe TAS in detail in Chapter 3 of this thesis. Briefly, it offered the following set of advantages over other lipreading tests:

1. TAS assesses lipreading ability on six different levels, including at the levels of words, sentences, and short stories. Thus, it offers a broad range of possibilities for estimating an individual's lipreading skills.

2. It was developed to be suitable for the assessment of lipreading ability of adults in the United Kingdom, and does not target a specific population, in terms of hearing status (contrary to some tests which are, for example, targeted towards deaf populations).

3. It has normative data, against which individual results can be compared, for both deaf (ages 21 – 60), and hearing (ages 16 – 76) populations.

4. It is suitable for use in both research, and the clinic.

Contrary to lipreading, there are no readily available tests for the assessment of the ability of listeners to benefit from audio-visual temporal coherence cues. Experimental paradigms such as the one developed by Maddox et al., (2015) (described in section 1.4.4.2), could in principle serve this purpose, by inclusion of an audio-only condition (Maddox and colleagues did not) against which the target-coherent condition would be compared to provide an audio-visual binding-specific audio-visual benefit. Arguably, however, as this test uses artificial stimuli, what it measures may deviate from listeners' ability to make use of temporal cues in more naturalistic speech-in-noise situations.

An alternative way of doing so would be to incorporate, within a comprehensive speech-in-noise task, an audio-visual condition that limits the availability of lipreading cues (e.g. by blurring of the talkers' mouths). Comparing this condition to the participants' audio-only condition performances would then help isolate temporal cue effects.

1.5.2.2. Examples of audio-only speech-in-noise tests that have been used in the clinic

In this section I outline some examples of audio-only speech-in-noise tests. I have sought to provide descriptions of tests that have been applied in, or have the potential to be applicable to, the clinical setting (see Billings et al., 2024 and Taylor, 2003 for reviews). I begin with fixed SNR speech-in-noise test examples, and then move on to describe adaptive SNR tests (see also section 1.5.1.3).

Two examples of a fixed SNR speech-in-noise tests are those of the Connected Speech Test (CST) (Cox et al., 1987), and the Speech Perception in Noise test (SPIN) (Bilger et al., 1984). As its name suggests, the CST has been used to measure individuals' ability to understand continuous or "connected" speech, in noisy environments. Unlike tests that use single words or phonemes, the CST uses longer passages of speech, 9 to 10 sentences in length, which more closely resemble normal conversational speech-patterns. These can be excerpts from everyday conversations and are presented in a multi-talker babble background. Each passage contains 25 target words, out of which a percentage correct score is computed, at the SNR selected by the test's administrator. The SPIN test incorporates sentences with varying degrees of contextual predictability, to assess the listeners ability to use linguistic context to understand speech under challenging listening conditions. Half its stimuli have low predictability, and the other half have high predictability; percentage correct scores are computed for each independently.

Examples of popular adaptive SNR speech-in-noise tests are the Hearing in Noise Test (HINT) (Nilsson et al., 1994) and the Quick Speech in Noise (SIN) (Killion et al., 2004) tests. The HINT uses a series of phonemically balanced sentences modified from the Bamford-Kowal-Bench (BKB) lists (Bench et al., 1979). These sentences are presented in speech-shaped noise, in groups of 10, and participants must identify all key words of a sentence to score a point. The SNR is varied depending on the participant's responses, by varying the intensity of the signal in 2 dB steps (noise intensity is kept constant). HINT outputs the participant's SRT_{50} in 5 to 10 minutes.

Quick SIN is even quicker, lasting only 2 to 3 minutes. This test uses IEEE sentences including 5 keywords per sentence, which are scored independently. These sentences are delivered in a 4-person multi-talker babble background. It begins with a high SNR of 25 dB SNR, which declines (or increases) adaptively in 5 dB SNR steps, depending on participant responses. Instead of an SRT, the outcome of Quick SIN is something called "SNR loss", which quantifies the decrease in performance in noise compared to normative data of individuals with normal hearing. The SNR loss is computed based on the number of correct words identified by the participant, using a formula provided in the Quick SIN test materials. An SNR loss of 0 dB SNR suggests excellent performance, while higher values indicate increasing difficulty with speech-in-noise comprehension.

In the previous paragraphs I have outlined examples of speech-in-noise tests that have also been used in the clinical sector. These diagnostic tests only include auditory stimuli: The measurement of patients' audio-visual benefits when listening in noise has not been part of standard diagnostics (there are exceptions; for example, Boothroyd et al., 1985 also assessed lipreading ability). I hereby argue, however, that an assessment of audio-visual benefit can be relatively easily integrated into these audio-only speech-in-noise tests, with the inclusion of an audio-visual condition. The difficulty derives from requiring that the audio-visual speech-in-noise test captures both lipreading, and audio-visual temporal coherence patient abilities. And, based on the considerations discussed in section 1.5, the aforementioned speech-in-noise tests are likely lacking on this aspect.

The audio-only speech-in-noise test employed by Messaoud-Galusi et al., (2011), however, seemed promising in this aspect. And, as I discuss in the next section, and in detail in Chapter 2 of this thesis, I selected it as my starting point for the development of the speech-in-noise task that forms the foundation of this thesis.

1.6. Summary, aims of this thesis, and chapters to follow

The current chapter has delved into a variety of subjects pertaining to the perception of speech in noisy environments. First, I described auditory scene analysis, the process by which our auditory system organizes the complex and noisy acoustic landscape that inhabits our daily lives, into perceptually meaningful elements, the so-called auditory objects. Then, I made the case that vision helps audition in this process, especially when it comes to speech-perception in noise. It does so in two ways, I concluded. Firstly, it does so by providing language-dependent, lipreading cues: Lipreading has been shown repeatedly to provide an audio-visual benefit in understanding speech-in-noise. Secondly, by providing language-independent temporal cues, vision enhances processes such as audio-visual binding, and audio-visual object formation. These have the potential to assist auditory scene analysis in a bottom-up manner. I explained, nonetheless, that it has proven difficult to isolate the effects of audio-visual binding from those of top-down processes, and provided the theoretical considerations that would allow experimentalists to do so. Namely, Maddox et al., (2015) and Atilgan et al., (2018), following these considerations have provided evidence for the existence of this mechanism. Yuan et al., (2020), (2021) have also shown that it likely contributes to the audio-visual benefit of understanding speech in noise.

Despite the fact that vision clearly benefits speech perception, audio-visual benefit assessments, I wrote next, have not been part of standard diagnostic procedures explored by clinicians in aiding patients with hearing loss improve their living conditions. In particular, an audio-visual speech-in-noise test that could capture both the abilities of patients to exploit lipreading, and temporal cues, would be ideal for use in both the clinic, and in research. For this purpose, I outlined considerations of what properties such test must have and provided examples of commonly used speech-in-noise tests. With these themes in mind, I now present the aims of my thesis:

1. Aim 1 (Chapter 2): To develop a speech-in-noise test that is capable of measuring individuals' audio-visual benefit provided by looking at a talker's mouth when listening to them in noise. It was my further goal to develop the test such that it is capable of capturing both lipreading, and temporal cue contributions to the audio-visual benefit.
2. Aim 2 (Chapters 4, 5, and 6): To test the hypothesis that both lipreading and audio-visual temporal coherence cues contribute to the audio-visual benefit simultaneously, in speech-in-noise perception, and provide a measure of their relative contributions.
3. Aim 3 (Chapters 4, 5, and 6): To assess potential factors that influence the audio-visual benefit. In addition to lipreading, and audiovisual temporal coherence cues, these factors include primarily age and hearing loss, but also gender, and ability to match the face to the voice of a talker (or "talker familiarity").
4. Aim 4 (Chapter 6): To use the participant data collected in the experiments of this thesis to develop two statistical models: One to model the speech-in-noise performance of individuals, and one to model their audio-visual benefit. This aim is complementary to, and supplements aims 2 and 3 above.

In Chapter 2, I detail the steps towards the development of my speech-in-noise task, which I hereby call the video version of the Children's Coordinate Response Measure with nouns at the end, or vCCRMn for short. The original Children's Coordinate Response Measure (CCRM), test, by Messaoud-Galusi et al., (2011) was used as a starting point for my task's development. Next, in Chapter 3, I outline the general methods used in this work, beyond the task development and the statistical modelling (the methods for which are included in Chapters 2 and 6 respectively). These methods include participant recruitment, experimental procedures, and analytical and statistical methods used in the analyses of participant data. Chapter 4 showcases the results obtained from three experimental groups: The younger participants with normal hearing, the older participants with normal hearing, and the older participants with hearing loss. Chapter 5 pools the datasets from the three experiments together and explores global analyses of the factors measured. Chapter 6 outlines the statistical models developed to describe participants' speech-in-noise performances, and audio-visual benefit measures. Finally, in Chapter 7, I provide an overall discussion of the findings of my work.

As a final note to this chapter, and to supplement the aims outlined above, I would like to add that throughout my PhD, I approached my research with the broader ambition in mind that the work conducted herein would be, or would have the capacity to become, of use to the clinic for the improvement of the quality of life of clinical populations.

Chapter 2: Development of a speech-in-noise test for measuring the audio-visual benefit

2.1. Introduction

Among the primary aims of my study was to develop a speech-in-noise test capable of measuring participants' AV benefit. Crucially, all other aims relating to the AV benefit (as outlined in section 1.6) depended on this test. Generally, to capture an AV benefit is relatively trivial: A test that does so would have to include a condition where both audio and video cues are made available to listeners, to be compared with a condition where only audio cues are available. Turn down the signal-to-noise ratio (SNR), and, below some threshold, the contribution of visual cues become apparent in the performance differences between audio-video and audio-only (see for example Bernstein and Grant, 2009).

It was not my sole goal to measure any AV benefit. That a visual condition has the capacity to aid participant performances via the provision of lipreading cues is a well-documented phenomenon (since, Sumby and Pollack, 1956). I wanted to cast a broader net on the audio-visual integration underlying this AV benefit: As per aim 2 of this project, my goal was to test the hypothesis that both lipreading, and temporal coherence mechanisms contributed to the AV benefits of participants.

My speech-in-noise task had to be designed such that the AV benefit it captured supported this goal. To briefly remind the reader, audio-visual temporal coherence cues between visual and auditory objects may involve two non-linguistic, AV benefit contributing mechanisms: That of auditory and visual object binding (Bizley et al., 2016; Lee et al., 2019), and that of directed attention towards the auditory object (see also section 1.4). Thus, for these mechanisms to take place, access to audition is necessary to such that the auditory object can be formed.

The implications of these for the design of the speech-in-noise task are the following: The task should not be one that could be solved merely by means of lipreading, instead, performance should also rely on accessing sound. Simultaneously, the stimuli should not be audible to the extend where the visual cues become redundant.

This chapter details the developmental steps that led to my speech-in-noise task, including the steps taken to meet the aforementioned requirements. It gives details about the motivation for choosing the CCRM speech in noise test as a starting point and walks through the process of iterating and piloting adaptations to it to meet the study's purposes.

2.2. Starting point: The Children's Coordinate Response Measure (CCRM)

The audio-only Children's Coordinate Response Measure speech-in-noise task (CCRM; Messaoud-Galusi et al., (2011) see also (Bolia et al., 2000; Brungart et al., 2001) for the Coordinate Response Measure, or CRM, corpus) was deemed a promising starting point towards the development of the speech-in-noise task.

The CCRM sentence stimuli take the form "Show the <animal> where the <colour> <number> is.", where an animal word can be any of [dog / cat / cow / duck / pig / sheep], colour any of [black / blue /

green / pink / red / white] and number any of [1 / 2 / 3 / 4 / 5 / 6 / 8 / 9]. The target sentences always contain the animal “dog” (i.e. “Show the dog where the <colour> <number> is”). The rest can be used as background noise, in the form of a multi-talker babble, or as background competing talkers. Notably, animals and colour-number combinations used in masker sentences for a trial are different from those used in the target sentence, and from each other (if using more than one masker). The participants’ task is to identify the colour and number mentioned in the target sentence.

The task provided the following advantages, also in line with the broader thesis goals: Firstly, the CCRM (or its CRM predecessor), has been widely used in research, in both adults and children, and has the capacity to be used in the clinic (Bianco et al., 2021; Billings et al., 2024; Messaoud-Galusi et al., 2011; Saleh et al., 2023) . Secondly, its stimuli and methodological manipulations formed a foundation on which this study’s hypotheses could be explored.

For example, its use of sentences with target words at the end would allow the probing of AV benefit mechanisms exhibiting accumulative effects over time, such as auditory object-based ones (Carlyon et al., 2001, Cussack et al., 2004). On the same point, the nature of the target words it makes use of (monosyllabic and of similar lengths to each other), provide relatively little lipreading cues, aiding thus the potential measurement of these object-based mechanisms. Further, the structure of its sentences would likely, when competing talkers (maskers) are used as background noise, result in energetic and informational masking on the auditory level, and thus would require a greater reliance on vision for identification of the target words (Helfer & Freyman, 2005). This would, in turn, manifest in an AV benefit when audio-visual and audio-only conditions are compared. Finally, the CCRM stimuli do not provide contextual cues for identifying the target words.

In the following section I outline the various parameter options available to CCRM users and discuss the reasons for making specific choices for the purposes of my experiments. Although these options were specific to the CCRM, the reasons for selecting these also applied to subsequent versions of the CCRM developed herein to achieve the goals of this thesis.

2.2.1. Parameter choices for the CCRM speech-in-noise task

The CCRM interface allowed the experimenter to set various parameters to characterise the speech-in-noise task. These included options on the preferred target talker(s), the types of background noise and, if applicable, the number of background speakers. Further, being an adaptive SNR test, it included several options related to the way the SNR varied during an experimental run. I outline these options below, beginning with a discussion of background noise.

2.2.1.1. Background noise type

CCRM offered several options of background noise types to choose from. These included speech-shaped noise, multi-talker babble, and multiple competing background talkers, as well as the option to choose the number of the background talkers. Among the different types of everyday background noise experience, multiple-talker situations are likely among the most common – as Miller, (1947) said “The best place to hide a voice is among other voices”.

As mentioned in section 2.2, exploiting the sentence structure of CCRM stimuli, to use these sentences as background noise in the form of competing talkers, would likely, through auditory masking, result in an enhanced AV benefit. Furthermore, masker talkers uttering sentences of similar format to the target, would, in an audio-visual setting, allow the probing of temporal coherence-based contributions to the AV benefit (see also sections 1.4.3.2 and 1.4.4.2 for further details) – later sections discuss the addition of talker videos to the CCRM. For these reasons, I chose the competing talkers option for background noise.

With regards to the number of background talkers, studies have shown decreased performance in speech in noise tasks with two masker talkers compared to when using a single masker and found improved performance with more than two talkers (Freyman et al., 2001, 2004; Rosen et al., 2013). Rosen et al., (2013) found that this was likely due to an inability to disentangle similar streams of information. Thus, to make the task relatively difficult, two background masker talkers were chosen.

2.2.1.2. Gender of background talkers

In addition to the number of background talkers, the CCRM also allowed for the selection of the background talkers' gender. To make the task more challenging on an informational masking level, one of the two talkers was set to always be of the same gender as the target talker (Kidd & Colburn, 2017).

The total number of options of male and female voices available for selection at each trial were three for each, from which a male and a female were randomly selected.

2.2.1.3. The transformed up-down method for dynamic SNR manipulation

Another parameter option of the CCRM was the rule with which the SNR of the stimulus trials varied given the participants' responses. In this study, the CCRM was set to use an adaptive transformed up-down staircase procedure (also touched on in section 1.5.1.3). The procedure was adaptive in that it adjusted the SNR at which each trial was presented, based on the outcome(s) of the preceding trial(s). These procedures, depending on their rule, modulate the SNR of at each trial such that it quickly converges to one of the participant's speech reception thresholds (SRTs) (see Levitt, (1971) for a detailed description of transformed up-down staircase procedures). An SRT_X being the SNR at which a participant is able to correctly identify X% of the trials. Given the interest in potentially extending the use of the speech-in-noise task in the clinical setting, it was considered advantageous that the test could be swiftly administered to provide participants with an SRT.

Each SRT_X constitutes a point on the participant's psychometric function (see section 1.5.1.3 for more details). The simplest up-down method is one-up-one-down (1U1D), that is, if the participant correctly identified the target of the current trial, in the next trial the SNR will be reduced by one step (one-down), otherwise the SNR will be increased by a step of equivalent magnitude (one-up). The SNR at convergence in the case of the 1U1D procedure, is SRT_{50}. Other commonly used transformed procedures are the one-up-two-down (1U2D), and one-up-three-down (1U3D), which converge to SRT_{70} and SRT_{79} respectively (Levitt, 1971).

All three procedures were available as options within the CCRM, and all were exploited, as detailed in subsequent sections, in the process of developing the speech-in-noise task. One notable reminder, from section 1.5.1.3, is that the higher the "X" of SRT_X, the higher the SNR. The higher the SNR the easier it generally is for a participant to identify the target words in the sentences with audition alone. When that is the case, the comparison between audio-visual, and audio-only conditions will likely not result in a large difference. In exploring these adaptive rules here, I sought to strike a balance signal audibility, and impact of visual information, in the speech-in-noise performances measured.

2.3. Development of the speech-in-noise task: An outline of all pilot stages

As mentioned before, CCRM is an audio-only task. Since the thesis' goals all revolved around audio-visual integration, there was a need for the creation of video material to compliment the CCRM's stimuli. The recording, and incorporation of audio-visual material to the CCRM, resulted in a modified CCRM version, hereby dubbed the video-version of the CCRM, or vCCRM for short. The process of development, from the first version of the vCCRM, to the final version of the speech-in-noise task, included six piloting/adaptation stages. These are briefly described below in the current section, and in more detail in their respective sections within this chapter. The stages are also summarised in Table 2.1.

Due to COVID pandemic regulations prohibiting in-person human research, Pilots 1-5 were conducted online. Pilot 6 (the final pilot) was carried out in person at UCL owing to the easing of COVID restrictions. Although the aim was to include only native speakers of British English in the piloting stages, some piloting stages included participants who spoke English as a second language (see Table 2.1 for details). This was not considered problematic, as non-native participants, where applicable, reported being proficient in English; this was deemed sufficient for the piloting stages of the test's development. All participants were paid for their time and gave written informed consent to take part, approved by the Research Ethics Committee of University College London (ref: 3866/003).

Summary of the six piloting stages:

1. In Pilot 1, the audio-only CCRM speech-in-noise task was adapted to include video material (vCCRM). Results from piloting the vCCRM led to the hypothesis that some participants were potentially able to solve the task via lipreading alone (i.e. without accessing sound). This was indicated by observations of correct responses at SNR levels at which the target talker should be inaudible.

2. Pilot 2 was run to reject or confirm the hypothesis from Pilot 1. New video material was recorded to run a silent version of the target sentences of the vCCRM (Silent vCCRM). Results from the Silent vCCRM revealed that some participants could solve the task via lipreading alone, strengthening the hypothesis formed from the results of Pilot 1.

3. Pilot 3 used the same silent material used in Pilot 2 but included an extended response panel. It was run to confirm that the fewer number of options on the response panel of Pilot 2 was not the reason behind participants' high scores. The results of Pilot 3 strengthened the hypothesis that the stimuli could be lipread, making the task too easy to solve, even in difficult listening conditions.

4. Pilot 4 ran a newly developed version of the silent vCCRM, called Silent vCCRM Nouns Larger set. At this stage one of the two sets of target words were replaced with noun words that were selected to be confusable with each other from a lipreading point of view. A large set of confusable noun stimuli were deliberately included, with the aim of identifying an appropriate subset for use in subsequent pilots. This subset was identified with confusion matrix analyses of participant responses.

5. Pilot 5 was also run as a silent task, using the identified subset of noun stimuli from Pilot 4 (Silent vCCRMn Reduced set). It helped confirm the suitability of this subset for use as stimuli in a speech-in-noise task that potentially captured both lipreading, and non-linguistic contributions to the AV benefit.

6. Pilot 6 was the final pilot in developing the speech in noise task and used sentence stimuli containing the noun words identified in Pilots 4 and 5. Audio-visual masker and target sentence stimuli

were recorded at UCL's anechoic chamber for the creation of the speech-in-noise version of the task, hereby called vCCRMn. The pilot further confirmed that the stimuli were suitable for the project aims.

Pilot	**Task name**	**Task**	**Task type**	**N of talkers recorded**	**N of recorded stimuli**	**Stimuli recording location**	**Target talker presentations**	**Masker talker presentations**	**N of participants tested**	**Testing location**
1	vCCRM	a) Identify and report 2 target words (<colour>, <number>) in the sentence using the animal "dog". b) Ignore the 2 masker talkers	Speech in noise	1 fluent in English (trained in English linguistics)	48 target sentence videos	Recorded at home due to COVID-19 restrictions	Female	1 male and 1 female (pre-existing in the CCRM; randomly selected from a pool of 3 males and 3 females)	3 (2 native speakers of British English, 1 fluent in English)	Online (through sharing of task code with participants)
2	Silent vCCRM	Lipread the talker presented to identify and report 2 target words (<colour>, <number>)	Silent	3 native speakers of British English	144 target sentence videos	Recorded at home due to COVID-19 restrictions	Either male or female (randomly selected from a pool of 2 females and 1 male)	NA	27 (13 native speakers of British English, 14 fluent in English)	Online via Prolific and GORILLA
3	Extended response panel vCCRM	Same as Pilot 2	Silent	Same as Pilot 2	Same as Pilot 2	Recorded at home due to COVID-19 restrictions	Same as Pilot 2	NA	5 (all native speakers of British English)	Online via Prolific and GORILLA
4	Silent vCCRMn Larger set	Lipread the talker to identify and report 1 target word (<noun>)	Silent	3 native speakers of British English	105 target sentence videos	Recorded at home due to COVID-19 restrictions	Female (random selection from a pool of 3)	NA	13 (all native speakers of British English)	Online via Prolific and GORILLA
5	Silent vCCRMn Reduced set	Same as Pilot 4	Silent	Same as Pilot 4	57 target sentence videos taken from Pilot 4	Recorded at home due to COVID-19 restrictions	Same as Pilot 4	NA	9 (all native speakers of British English)	Online via Prolific and GORILLA
6	vCCRMn	a) Identify and report 2 target words (<colour>, <noun>) in the sentence that uses the animal "dog". b) Ignore the 2 masker talkers	Speech in noise	4 native speakers of British English	2304 sentences (576 per talker, including both target and masker stimuli)	In person at UCL's anechoic chamber	Either male or female (randomly selected from a pool of 2 males and 2 females)	1 male and 1 female (randomly selected from a pool of 2 males and 2 females, excluding the option selected as the target talker)	6 (1 native speaker of British English, 5 fluent in English)	In person at the UCL Ear Institute human labs

***Table 2.1**: Summary of the piloting stages that led to the development of the vCCRMn task.*

In the following sections I provide thorough descriptions of each of the 6 piloting stages.

2.4. Pilot Experiment 1: Incorporation of video material to the Children's Coordinate Response Measure

The main aims of this thesis concern audio-visual integration, and in specific, the audio-visual benefit conferred by looking at a talker's face while listening to their voice in noise. Thus, as a first step towards that direction, video material had to be incorporated into the otherwise audio-only CCRM. To this end, video and audio material of the original CCRM target talker stimuli were recorded and included in the task. To distinguish the original CCRM version from the audio-visual CCRM, the latter version will be referred to as the video-version of the CCRM, or vCCRM.

In the current pilot, the vCCRM was tested as a speech-in-noise task with the goal of determining its broader suitability for the purposes of this thesis. This pilot thus served as an exploratory step. Among the major considerations explored included the following: Will performance in the speech in noise task indeed be better in the audio-visual condition compared to the audio-only, as expected? If performance is indeed better in the audio-visual condition compared to the audio-only, is it better because the task can be solved with vision alone? How does the rule for presenting stimuli (staircase procedures, described in section 2.2.1.4) affect performance in the task's conditions?

2.4.1. Participants

Three participants (members of the Bizley lab; native speakers of British English N=1, Age: 36; speakers fluent in English N=2, Ages: 26 and 28; all female) ran the experiment at its first piloting stage. The experiment was run on a Desktop device using MATLAB and required the use of headphones. Participants had no history or diagnosis of hearing loss, and all had normal/corrected-to-normal vision.

2.4.2. Stimuli and procedure

As UCL was shut, due to the COVID-19 pandemic, stimuli were recorded at home. Forty-eight videos of target stimuli were recorded in total, read by me. Recordings were made using a high-definition camera at 25 frames per second with auditory recording using an internal microphone. The video frame captured the target talker from the shoulders upwards, as illustrated in Figure 2.1 (A). The masker sentences used were taken from the CCRM corpus and consisted of two talkers (one male, one female chosen randomly from a pool of three female and three male talkers). Target and masker stimuli had the form described in section 2.2. All trials were 3 seconds long.

The vCCRM was piloted testing both 1U2D and 1U3D staircase rules, which result in SNR convergence to SRT_{70} and SRT_{79} respectively (Levitt, 1971). The rule 1U1D, was omitted here on the basis that 1U2D or 1U3D might suffice for the purposes of the experiment, following the example of previous reports (e.g. Grant & Seitz, (2000), who had used the 1U3D procedure). Further, I tested both 1U2D, and 1U3D, (instead of just 1U3D) to explore which one was most appropriate for capturing the participants' AV benefit. As mentioned in section 2.2.1.3, SRT_{79} would be expected to render a smaller AV benefit.

The overall sound intensity of the stimuli was kept constant at 65 dB SPL, for both staircase procedures, and throughout the sequence of trials; the SNR was varied by adjusting of both signal and noise. Starting with an initial value of SNR at 20 dB SNR with each correctly identified trial, the SNR was decreased by 9 dB SNR until the first reversal. Then, the step size was reduced to 7 dB SNR to the second reversal, 5 dB SNR to third, and 3 dB SNR for the rest. The sequence was set to stop after four runs of the final step size had been concluded or when a total of thirty trials had been completed. The speech reception threshold was estimated by taking the mean of the last four reversal trials.

The vCCRM stimuli were also presented in both AV and audio-only fashion. In the audio-only condition, the target sentences were presented without the video; instead, a still frame of the target talker was displayed on the screen. For both AV and audio-only conditions, the response panel screen was immediately initiated after stimulus presentation. This consisted of an image of a dog and an array including all the possible colour-number combination options (Figure 2.1 (B)).

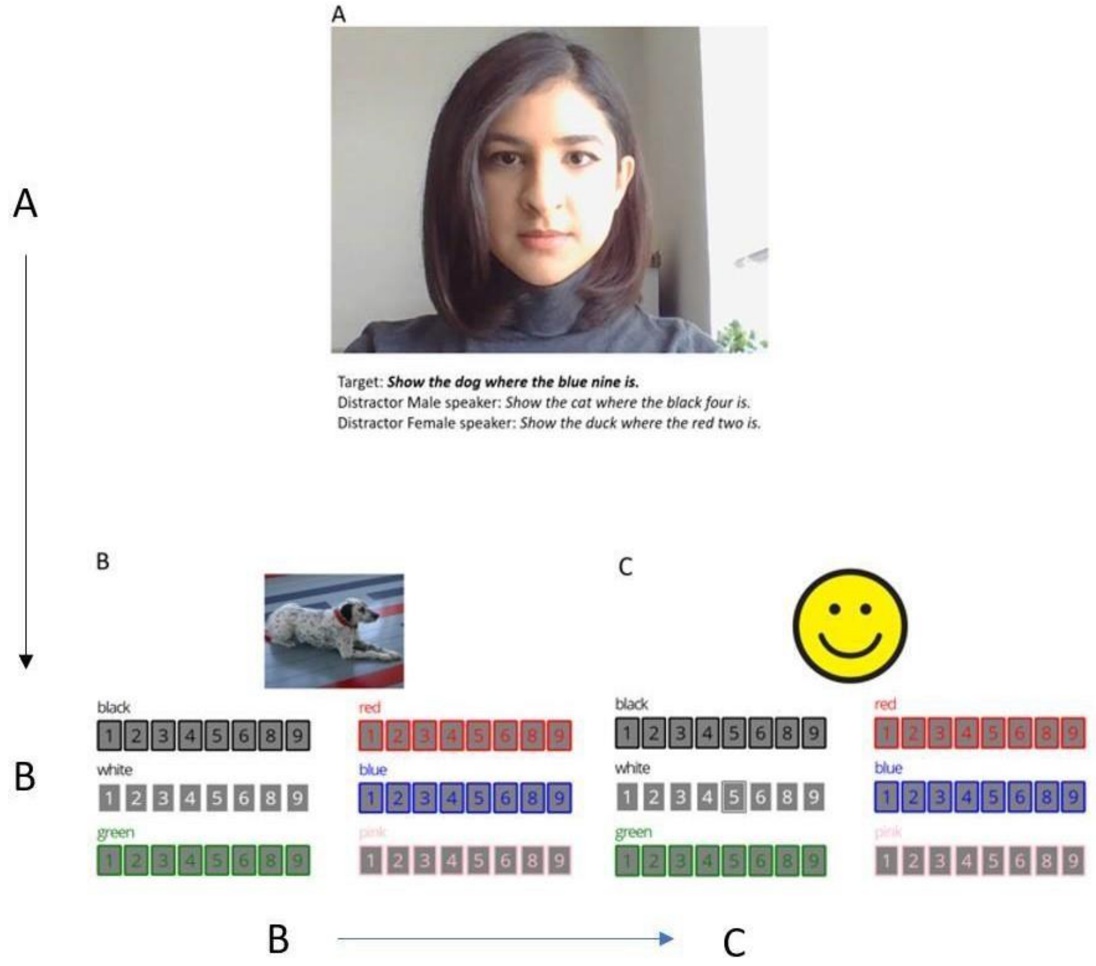

Target: ***Show the dog where the blue nine is.***
Distractor Male speaker: *Show the cat where the black four is.*
Distractor Female speaker: *Show the duck where the red two is.*

***Figure 2.1**: Pilot 1: vCCRM task schematic. (A) An example of an audio-visual stimulus. The talker in the video always read the target sentences "Show the dog…" and maskers were presented at the same time on the background. A still frame of this talker was used for the audio-only condition. (B) The response panel displayed after the audio-visual or auditory only stimulus was presented. The panel consisted of all 6 colours and all 8 numbers included in the sentence stimuli. Participants had to select the option they believed they had heard. (C) An example of immediate feedback (positive) for the response given post-stimulus presentation.*

Each of the participants completed 3 runs of audio-only and 3 of AV for each of the two aforementioned staircase rules. The order with which the staircase rules were employed, and the type of stimulus presented (AV or audio-only) were randomized to average out practice and/or fatigue effects. Mean SRT (that is, mean of the three runs ran for each condition) was calculated for each participant in all four conditions (AV with 1U2D, audio-only with 1U2D, AV with 1U3D, audio-only with 1U3D).

Participants received immediate feedback after every trial on whether their response was correct or not via a smiley/frowny face (Figure 2.1 (C)).

2.4.3. Pilot Experiment 1: Results

Figure 2.2 summarises the mean SRT for each participant in each condition, and participant AV benefits can be gauged by the difference in performances between respective audio-only and AV conditions. It was expected that AV conditions' SRTs would be lower than the audio-only ones, indicating that visual cues were benefiting speech-in-noise performance. Whereas participant 1 did not show an AV benefit, and generally showed similar performances across all four different conditions, participants 2 and 3 did show substantial AV benefits. Notably, these were in the excess of 10 dB SNR in the 1U2D condition for participants 2 and 3, and in the 1U3D condition for participant 2. These magnitudes seemed quite large, compared to previously published values (see e.g. Grant & Bernstein, (2019) for examples), to reflect realistic AV benefits obtained in natural speech-in-noise scenarios.

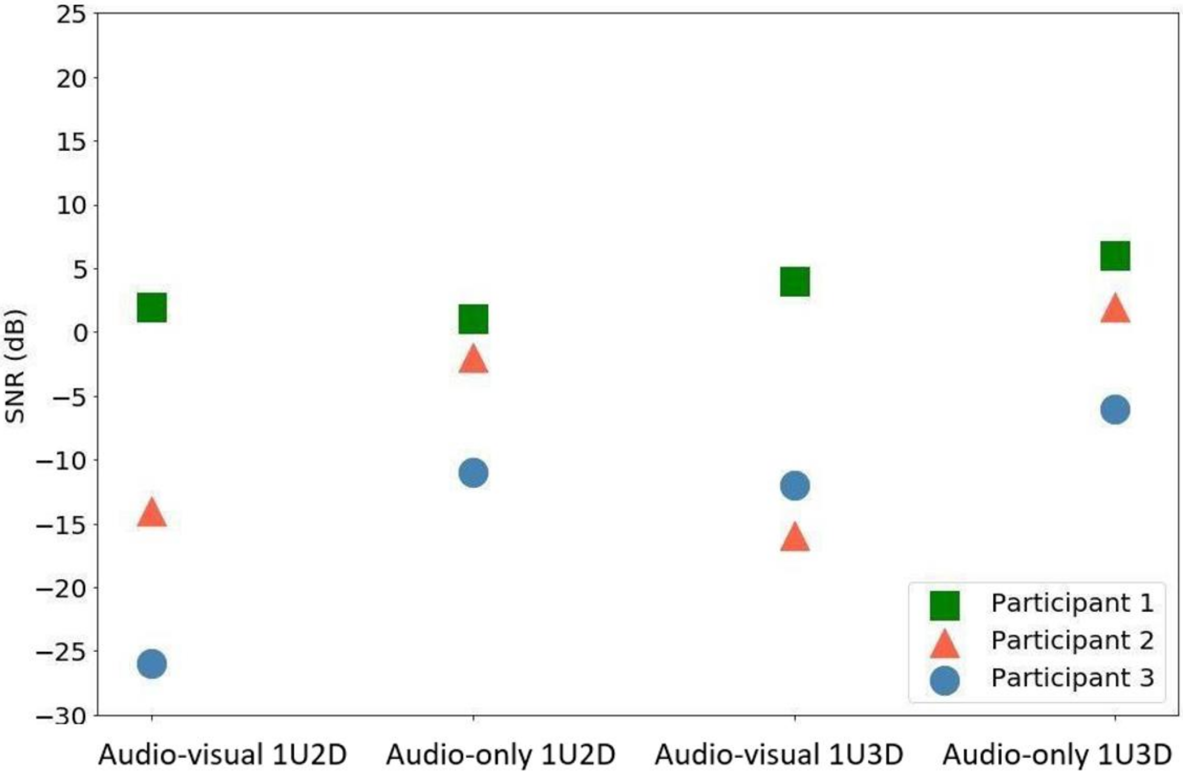


Figure 2.2: *Mean speech reception thresholds of vCCRM participants. Each dot represents an average of three runs completed by a participant for the indicated speech in noise condition and transformed procedure.*

Furthermore, examining the performance of participant 3 across the staircase procedures' trials (Figure 2.3), it was evident that the participant was correctly identifying target words at SNRs as low as -35 dB SNR. Correct identification of speech at such low SNRs would be implausible to achieve through auditory means.

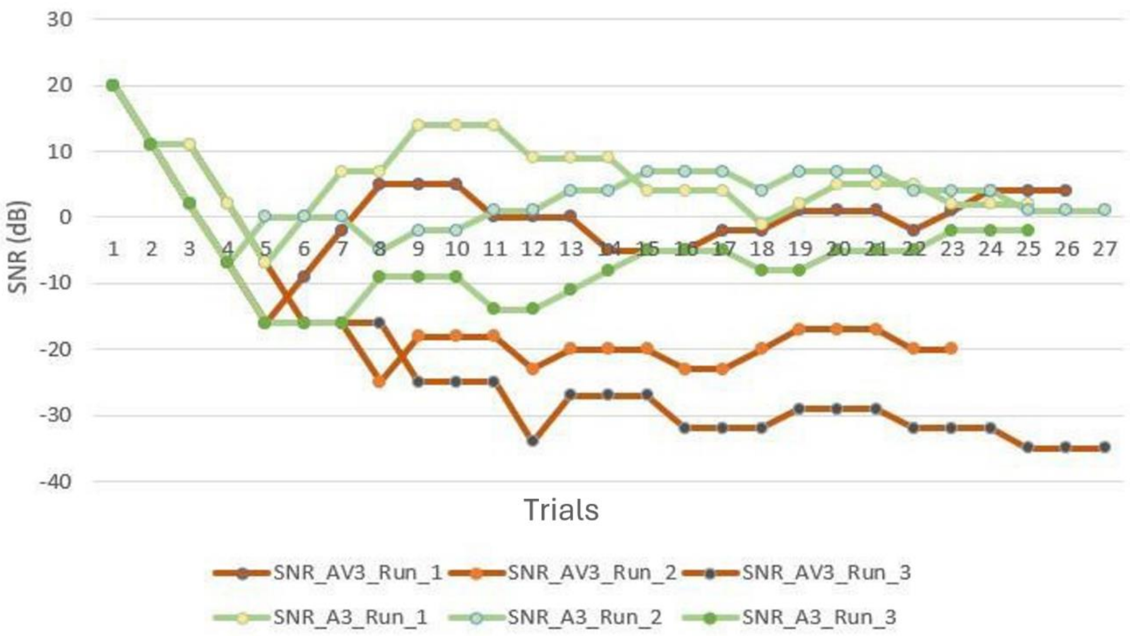


***Figure 2.3**: Staircase procedure examples of 1U3D rule, from participant 3 of Figure 2.2. SNR outputs are shown for all six runs (three runs per audio-only and audio-visual conditions, indicated as Run 1-3). Lines in green represent runs of audio-only (A) and lines in brown represent runs of audio-visual (AV) stimuli conditions.*

2.4.4. Pilot Experiment 2: Conclusions and next steps

The results of Pilot 1 suggested that the task was easily solvable when video material was included to the CCRM. Specifically, given the results it was hypothesised that, at its current stage the vCCRM could be solved by lipreading alone. This would, if true, interfere with the goals of this thesis, and the test would thus require modifications to mitigate the issue. To test this hypothesis, the vCCRM was run as a silent (visual-only) task in Pilot 2.

2.5. Pilot Experiment 2: Silent vCCRM task

Pilot 2 was run as a silent version of the vCCRM, using newly recorded video material of the target stimuli. Since COVID-19 regulations were in place, halting in-person human research, the stimuli were recorded at home and the testing was completed via online platforms used for human experiments.

2.5.1. Participants

Thirty participants were initially recruited via the online platform Prolific (https://www.prolific.com/). They were then directed to the Gorilla Experiment Builder (https://gorilla.sc/) where they ran the test. Inclusion criteria were having British English as a first language, no history of hearing problems, and normal/corrected- to-normal vision. Due to a problem relating to Prolific's filters, however, almost half of the tested participants were not native speakers of British English. Further prerequisites for participation included that participants owned a pair of headphones to wear during testing, and a desktop device with the Google Chrome web-browser installed.

Data collected from three participants who faced technical issues while taking the test were not used in subsequent analyses. Of the remaining twenty-seven, thirteen were native English speakers (Age: 18-30, Mean: 23 years, SD 4.56 years, 7 females, 6 males) and fourteen were non-native but fluent in English (Age: 18-30, Mean: 21.93 years, SD 3.15 years, 7 females, 7 males). Average proficiency of English of the non-native participants was 9.07/10, as reported via self-rating.

2.5.2. Stimuli and procedure

New, and improved (compared to the first pilot) stimuli were recorded for the Silent vCCRM experiment, including three different talkers. Videos of three native speakers of British English (2 female, 1 male) were recorded reading only the target sentences "Show the dog…" (Figure 2.4 (A)). Each talker recorded themselves at home, reading all unique sentences in one long recording session. Each long video recording was manually processed to remove audio and cut (with Avidemux; https://avidemux.sourceforge.net/) to produce multiple 3 seconds-long videos with equal silence at their beginnings and ends. Each video segment represented a unique silent vCCRM stimulus sentence. All 3 talkers read all possible colour-number combinations, resulting in a total of 48 unique sentences per talker (6 colours x 8 numbers). All of these were included as trials in the task, resulting in 144 trials. Each unique stimulus per speaker was only displayed once to avoid introducing a learning effect.

In the current pilot, the participants' task was to attempt to lipread the target words (colour and number) in each trial sentence and select the correct option from the stimuli panel (Figure 2.4 (B)). A practice session with audio was created to precede the main silent task. The practice session used videos with audio to quickly familiarise the participants with the stimulus types, and as a check that that the participant had clearly understood the task at hand.

Participants were first guided through a set of checks that ensured the experiment would run smoothly, including checks that video could be played without issues, and that audio (presented during the practice session) was of sufficient and comfortable loudness. Specifically, participants were asked to adjust their volume at ~20% and to use a sample audio to judge whether further adjustment of the loudness was necessary.

The practice session consisted of six trials. Each trial presented the audio and video of one of the three talkers reading a target CCRM sentence without the presentation of background noise. Participants were instructed to watch the talker uttering the sentence and to report the colour and number words they heard. After each video the participants were presented with the response panel (same one used in Pilot 1, also shown in Figure 2.4 (B)) at which point they were to provide their response. Feedback was not given on the correctness of their responses.

Participants who correctly identified a minimum of five out of six practice trials were directed to the main, Silent vCCRM task. They were given two opportunities to pass the practice session. Unsuccessful completion of the practice session led participants to a rejection screen, where testing was halted. All 27 participants recruited for Pilot 2 passed the practice session.

The task (including practice, and main sessions) lasted 20 minutes on average (out of a maximum allowed time of 1 hour). Participant performances were assessed in terms of a percentage correct identification across all trials. In addition to correct trial identification – that is, correct identification of both colour, and number of a given trial; hereby referred to as Total correct – performance on just colours (Colours correct), and numbers (Numbers correct) were also considered, independently.

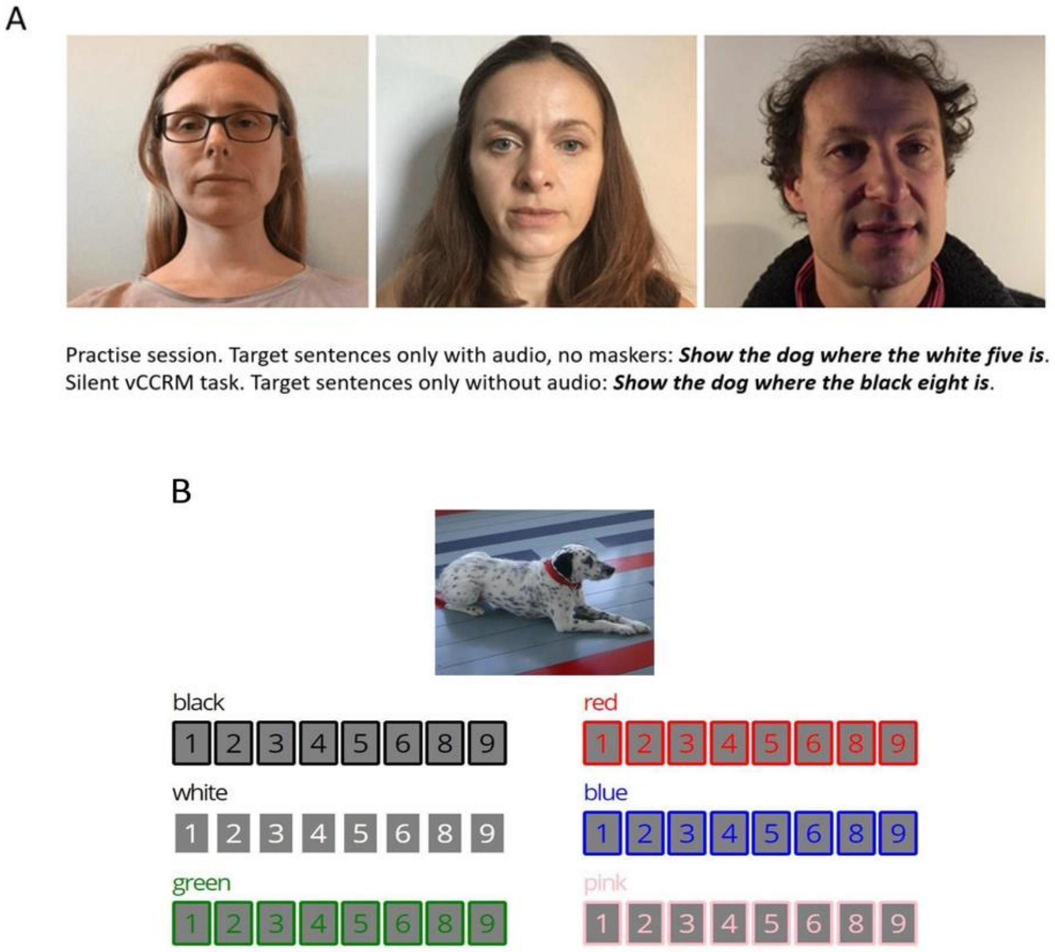

A
Practise session. Target sentences only with audio, no maskers: ***Show the dog where the white five is***.
Silent vCCRM task. Target sentences only without audio: ***Show the dog where the black eight is***.
B
black
1 2 3 4 5 6 8 9
red
1 2 3 4 5 6 8 9
white
1 2 3 4 5 6 8 9
blue
1 2 3 4 5 6 8 9
green
1 2 3 4 5 6 8 9
pink
1 2 3 4 5 6 8 9

Figure 2.4*: Pilot 2: Silent vCCRM task schematic. (A) The 3 talkers of the audio-visual, and visual-only stimuli included in the Silent vCCRM task. Only one of the 3 talkers appeared on the screen on each trial. (B) The response panel presented after each audio-visual trial of the practice session and each visual-only trial of the Silent vCCRM task.*

2.5.3. Pilot Experiment 2: Results

The results of Pilot 2 are shown in boxplot form in Figure 2.5. Mean participant performance for Total Correct was 34.75% (SD = 22.98%), whereas for Colours Correct it was 53.19% (SD = 20.86%), and for Numbers Correct 53.27% (SD = 23.86%). The red line in Figure 2.5 highlights the 70% percentage correct point, the proportion of correct answers for which the 1U2D adaptive procedure used in Pilot 1 outputs an SRT. It was evident from the results that this point fell within two standard deviations from the mean performance of participants (here, this comment only considers performance in the Total Correct variable, as the adaptive procedure treats a trial as correct only if both colour and number are correctly identified).

Based on this, it was reasoned that performance for some participants on the Silent vCCRM Total Correct was not substantially different to the performances measured in Pilot 1 (where both audio and video were included in the stimuli). Furthermore, when the variables Colours Correct and Numbers Correct are considered individually, then it becomes apparent that most colours and most numbers were indeed identifiable via lipreading by a large proportion of the participants.

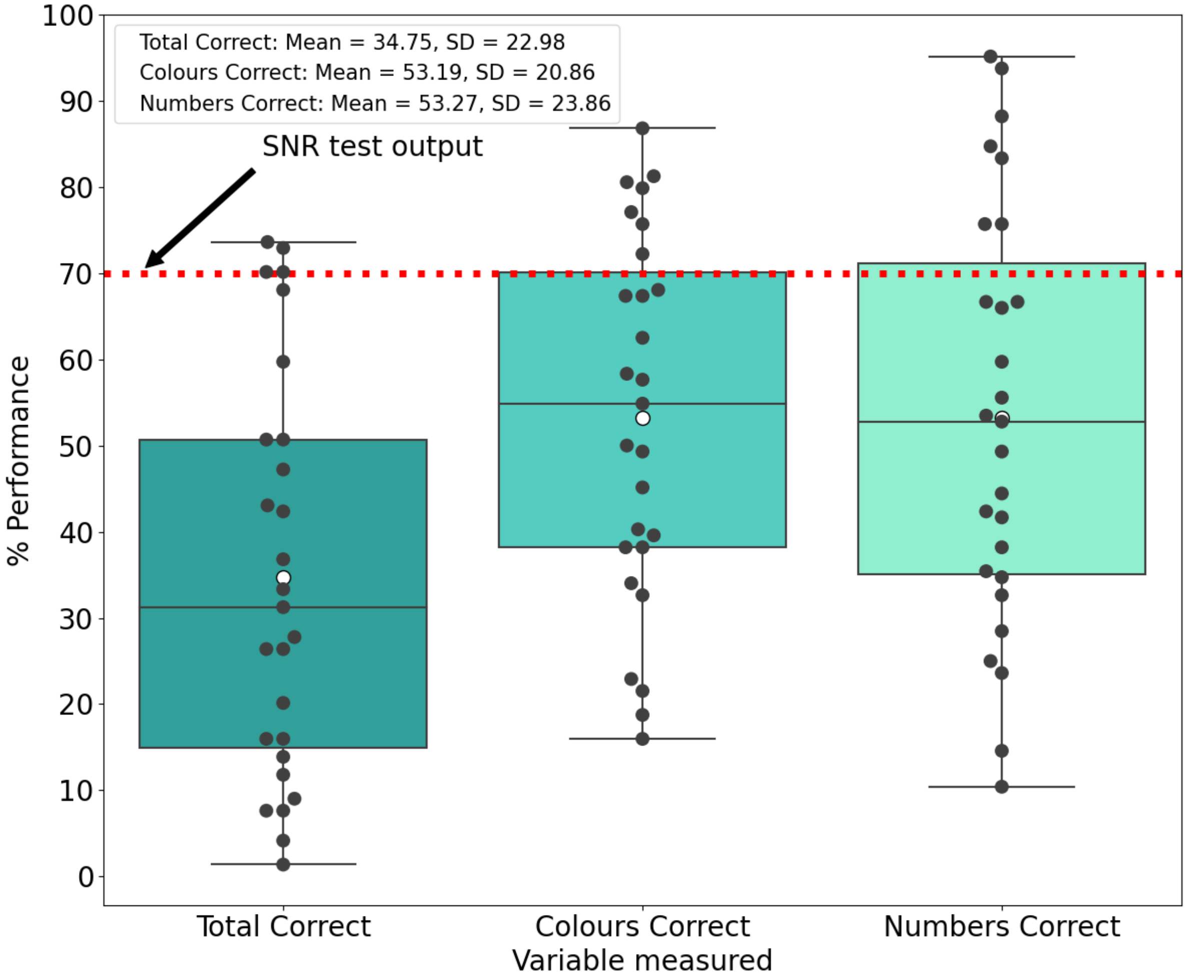


***Figure 2.5**: Silent vCCRM pilot task results shown in boxplots. The boxplots are showing the distribution of Total Correct (%), Colours Correct (%), and Numbers Correct (%) performance. White dots represent the group means; boxplot lines represent the medians. Red dotted line depicts the 70% point for which the 1U2D procedure outputs an SRT.*

To further confirm that participants were able to lipread the words during the task, each of the three variables measured were modelled with Binomial distributions representing participants selecting their responses at random (the distributions were the following: Total Correct, B(n = 144, p = 1/48); Colours Correct, B(n = 144, p = 1/6); Numbers Correct, B(n = 144, p = 1/8)). Participant performances were then compared against these distributions. Based on these comparisons, the majority of participants scored significantly better than chance in identifying all three measured variables correctly: the total number of participants that performed better than chance for Total Correct (both colour and number correct) was 25/27 (92.59%), for the Colours Correct 23/27 (85.19%) participants and for the Numbers Correct it was 25/27 (92.59 %) (results shown in Table 2.2).

Participant	Total Correct successful trials (out of 144)	Colours Correct successful trials (out of 144)	Numbers Correct successful trials (out of 144)
1	29	56	60
2	38	83	60
3	22	54	51
4	20	49	49
5	105	110	137
6	62	90	85
7	38	71	63
8	6	33	20
9	45	83	71
10	2	22	15
11	47	78	76
12	85	125	96
13	60	103	80
14	13	47	33
15	11	31	36
16	11	27	47
17	22	54	54
18	53	65	94

19	105	117	126
20	72	96	108
21	40	72	76
22	72	96	108
23	98	116	119
24	101	108	135
25	101	114	121
26	67	98	96
27	17	58	40
***Proportion of participants who performed over chance**	**25/27 (92.59 %)**	**23/27(85.19 %)**	**25/27 (92.59 %)**

Table 2.2: *Summary of participants' performances and the proportions of participants who performed above chance in the Silent vCCRM task. The total number of participants that performed better than chance for Total Correct was 25/27 (92.59%), for the Colours Correct it was 23/27 (85.19%), and for the Numbers Correct 25/27 (92.59 %). Highlighted are the instances where participants did not perform above chance (Binomial test, p > 0.05); p-values were adjusted for multiple-comparisons with the Bonferroni method.*

2.5.4. Pilot Experiment 2: Conclusions and next steps

The results of Pilot 1, where the vCCRM was run as a speech-in-noise task, hinted to the possibility that the task was too easy for participants, as they were exhibiting very large AV benefits, and could perform at SNR levels at which the signal was inaudible. This led to the hypothesis that participants were, for the most part, possibly lipreading the target words. To test this hypothesis, the silent version of the vCCRM was run in the current pilot. The results implied that it was indeed possible that participants were, to a substantial extent, lipreading the target words.

It was hypothesised that the high performance observed by participants could be partially attributed to a limited number of possible response options to choose from. Pilot 3, described in the next section, investigated this question.

2.6. Pilot Experiment 3: Extended response panel Silent vCCRM task

To test the hypothesis that increasing the available response options would make lipreading the stimuli more ambiguous, the response panel of the Silent vCCRM task from Pilot 2 was extended to include more response options.

2.6.1. Participants

Participants were recruited via the online platform Prolific and were then directed to the Gorilla Experiment builder. Native speakers of British English (N=5, Age: 18-30, Mean: 24.2 years, SD 4.66 years, 2 females, 3 males) who had not participated in Pilots 1 and/or 2 were recruited in this task, using the inclusion criteria as mentioned in the Silent vCCRM task (section 2.5.1).

2.6.2. Stimuli and procedure

Two colours and two numbers were added to the pre-existing sets, increasing the number of response panel options from the previous 48 (6 colours x 8 numbers) to 80 (8 colours x 10 numbers). In particular, the panel included the only two monosyllabic numbers available that were not already in the test (ten and twelve) as well as the two monosyllabic colours grey and brown (Figure 2.6). The experiment used the same stimuli and followed the same experimental procedure as the Silent vCCRM task run in Pilot 2.

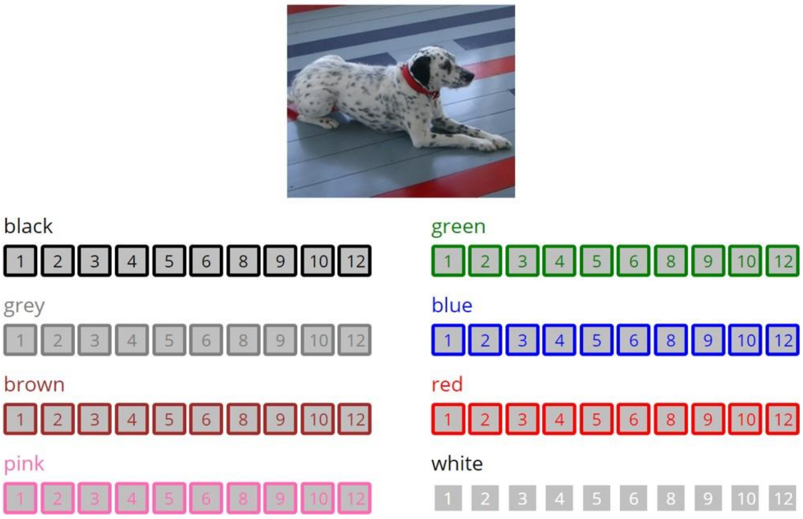


Figure 2.6: *Pilot 3: Extended response panel Silent vCCRM schematic. The figure illustrates the extended response panel displayed to participants, which included two additional colours (grey, brown) and two additional digits (10,12).*

2.6.3. Pilot Experiment 3: Results

To test whether the added ambiguity of the increased number of response panel options would increase the uncertainty in the participants' responses, participant performances across the three measured variables were once more compared to Binomial chance distributions. These were the following: Total Correct, B(n = 144, p = 1/80); Colours Correct, B(n = 144, p = 1/8); Numbers Correct, B(n = 144, p = 1/10)). The Binomial tests confirmed that all participants were performing above chance, across all three variables (results are also shown in Table 2.3).

Participant	**Total Correct successful** **trials (out of 144)**	**Colours Correct successful** **trials (out of 144)**	**Numbers Correct successful trials** **(out of 144)**
1	107	125	123
2	65	84	104
3	78	99	106
4	37	69	72
5	49	76	90
***Proportion of participants who performed over chance**	**5/5 (100 %)**	**5/5 (100 %)**	**5/5 (100 %)**

Table 2.3*: Summary of participants' performances and the proportions of participants who performed above chance in the Extended response panel Silent vCCRM. All participants performed above chance, across all three variables measured, based on Binomial tests. p-values were adjusted for multiple-comparisons with the Bonferroni method.*

In fact, participants were performing substantially better than what would be expected given random responses. Based on the stated Binomial distributions, the expected number of correct trials given random responses, would be ~2 for Total Correct, 18 for Colours Correct, and ~14 for Numbers Correct. Table 2.3 shows that all 5 participants very much exceeded these values.

As a complementary visual, participant performances across the three variables were plotted side-by-side in Figure 2.7, for both the current, and the previous pilot results.

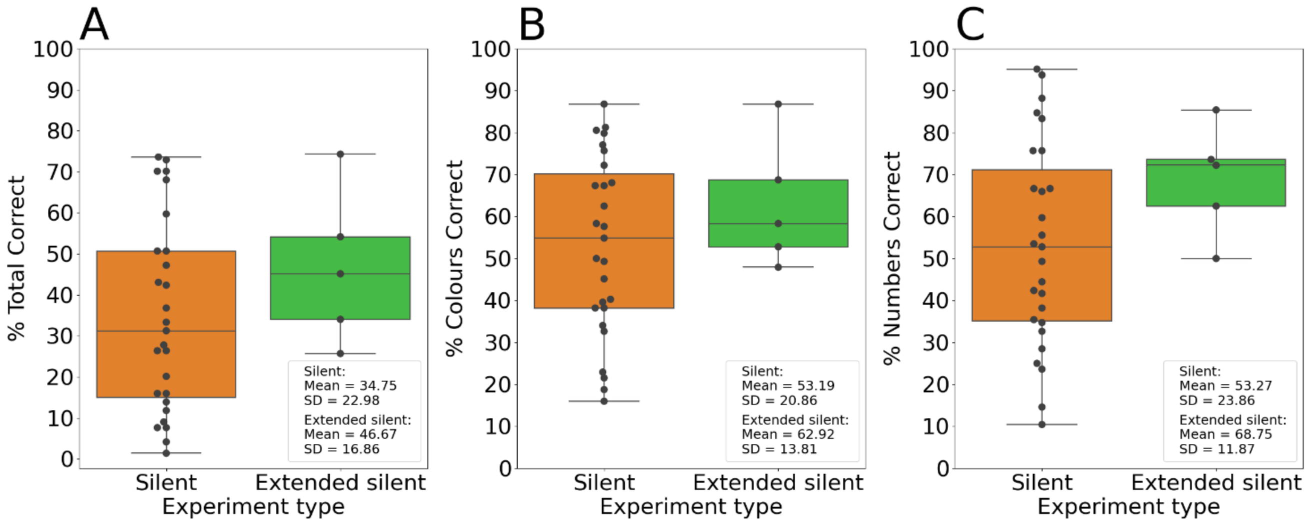


Figure 2.7*: Side-by-side boxplots of the original and extended response panel Silent vCCRM tasks. Boxplots (A) showing the distribution of Total Correct (%), (B) of Colours Correct (%) and (C) Numbers Correct (%) performances in each group.*

2.6.4. Pilot Experiment 3: Conclusions and next steps

Participant performances in the extended response panel Silent vCCRM task indicated that increasing the number of available options did not make the task more ambiguous, with participants performing substantially better than what would be dictated by mere chance. These results confirm the conclusion from Pilot 2, that participants were, to a large extent, lipread the CCRM's target stimuli from the talker videos.

It was thus concluded that the stimuli of the CCRM required an adaptation to reduce their lipreadability. To this end, a new version of the vCCRM was developed that used colours and nouns (instead of colours and numbers), and was run as another silent task, as detailed in the next sections (Pilot 4).

2.7. Pilot Experiment 4: Silent vCCRM Nouns Larger set

In Pilots 1,2, and 3 it was confirmed that the colour and number words could be correctly identified by many participants via lipreading alone. In the current pilot's sections, I discuss how the vCCRM stimuli were modified to reduce their lipreadability. Briefly, a large set of nouns was identified, including words that were easily confusable with one another in terms of the lipreading cues they provided to participants. These were used to create new stimuli that were employed in this pilot.

In addition to assessing the reduction in lipreadability of the selected nouns, the large set of nouns used here served to inform the selection of a smaller subset of nouns, destined to replace the number words in the CCRM stimuli.

2.7.1. Participants

Participants (N=13, 8 females, 5 males; Age: 18-35, Mean: 27.08 years, SD 3.40; 12/13 participants provided age information) were recruited via the online platform Prolific and ran the experiment via Gorilla Experiment Builder. Inclusion criteria were the same as in the previous pilots, with the addition that no participant had taken part in any of the previous versions of the vCCRM.

2.7.2. Stimuli and procedure

The nouns used in Pilot 4 was larger than necessary and served a two-fold purpose: a) To gauge the level of difficulty of the task and, more importantly b) to provide a broader option range for choosing the most appropriate (i.e. most confusable with each other) nouns for further assessment in the subsequent, and final silent pilot (here Pilot 5). The ambition here, in creating this confusability set was that participant performances in the AV conditions would, at least partly, reflect audio-visual mechanisms that depended on audition too (e.g. audio-visual binding), and not solely on lipreading.

The set included 30 monosyllabic nouns, selected such that they could not be unambiguously identifiable through lipreading cues alone. They were chosen for being monosyllabic, to retain the characteristic of CCRM words they were destined to replace. Further, they were selected to be highly visualisable (in terms of "thinking in pictures"), as this would be a useful feature to have, in potential children-friendly future versions of the test.

For the selection, the nouns' initial and final mouth movements were considered, ensuring that across the set, no single noun had unique mouth movements associated with it. In this manner, it was ensured that each noun in the set was visually (in terms of lipreading cues) confusable with two other nouns in the set: It shared a common beginning with one, and a common ending with the other. The phonetic transcription of each word as it appeared in the Oxford English Dictionary (Hornby, 2000) was also considered in this process.

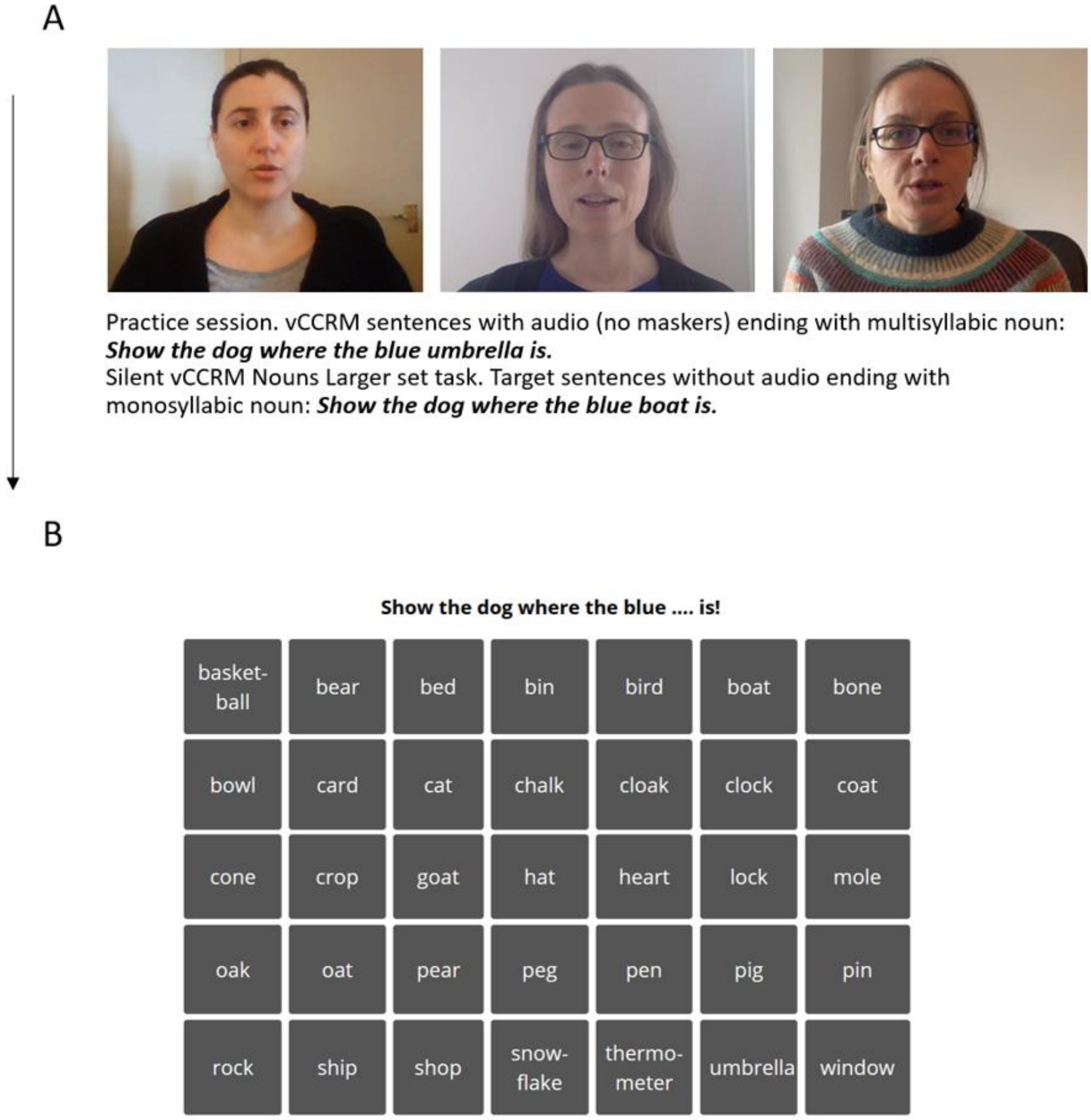
The task was expected to be difficult. Considering this foreseen difficulty, I wanted to ensure that participants would not find the task discouraging, resorting to dishonest and/or random responses. To this end, 5 multisyllabic, easily lipreadable, nouns were added to the set of 30 monosyllabic (details below).

To create the stimuli, each of the 3 female native speakers of British English (Figure 2.8 (A)), separately recorded themselves over a single session reading target sentences of the form "Show the dog where the blue <noun> is". These always included the colour "blue" (chosen randomly from all colours), and each contained a different noun from the 35 shown in Figure 2.8 (B), resulting in a total of 105 sentences.

The software Avidemux was used to manually cut each of the 3 audio-visual recordings to produce 3 seconds-long videos with one sentence in each. Avidemux was also used to remove the audio from all monosyllabic noun sentences. Two versions of the multisyllabic noun sentences were produced: a version with audio to be used in the practice session and as encouragement trials, and another without audio for use in catch trials (participant attentiveness checks).

Participants were recruited via Prolific and ran the task in Gorilla Experiment Builder. They were provided with instructions on the task, which included that the nouns would be presented in alphabetical order in the response panel, from left to right, and from top to bottom. They first ran the same series of audio and video checks described in Pilot 3. Then, they were led to the practice session, which consisted of 5 trials with audio. Each trial presented a video of one of the talkers reading a sentence ending with a multisyllabic noun. Participants then selected the noun they heard from a response panel. They had to identify at least 4 out of 5 trials correctly to proceed to the main task.

Each trial of the main task consisted of a silent video of one of the 3 recorded talkers (randomly chosen) reading a target sentence ending with a monosyllabic noun. Participants were tasked with identifying, are reporting, the noun. They were occasionally exposed to a catch trial (silent sentences using a multisyllabic noun at the end) or an encouragement trial (same as the catch trials but with audio). Every trial was followed by a response panel with all nouns including the multisyllabic catch/encouragement trial nouns (Figure 2.8 (B)). All 105 recorded sentences were presented to each participant running the task.



Practice session. vCCRM sentences with audio (no maskers) ending with multisyllabic noun: ***Show the dog where the blue umbrella is.***
Silent vCCRM Nouns Larger set task. Target sentences without audio ending with monosyllabic noun: ***Show the dog where the blue boat is.***

Show the dog where the blue .... is!

basket-ball	bear	bed	bin	bird	boat	bone
bowl	card	cat	chalk	cloak	clock	coat
cone	crop	goat	hat	heart	lock	mole
oak	oat	pear	peg	pen	pig	pin
rock	ship	shop	snow-flake	thermo-meter	umbrella	window

Figure 2.8*: Pilot 4: Silent vCCRM Nouns Larger set schematic. (A) Each trial presented a silent video of only one talker from the pool of 3 talkers reading a target sentence. The participants' task was to identify only the noun word at the end of the sentence, this could be any of the thirty monosyllabic nouns presented in (B) of this figure. Participants were sometimes exposed to vCCRM sentences of the same task with a multisyllabic noun at the end, either silently as a catch trial, or with sound as an encouragement trial (these were basketball, snowflake, thermometer, umbrella, and window). (B) Every trial was followed by a response panel with all thirty nouns and five multisyllabic catch/encouragement trial nouns.*

2.7.3. Pilot Experiment 4: Results

As a first step in analysing the data, mean performances were computed of the participant percentages of correctly identified nouns. This was 24.77% (SD = 4.96%), reflecting a drop in performance as compared to the performances reported for Pilot 2. Namely, in Pilot 2 the performances for the variables Colours Correct and Numbers Correct were 53.19% (SD = 20.86 %) and 53.27% (SD = 23.86 %) respectively.

Confusion matrix analyses of the data were also conducted and are shown in Figure 2.9. The figure's colour bar indicates the probability of giving a specific participant response (y-axis) given the word in the sentence presented (x-axis). The diagonal entries of the matrix represent the probabilities of correctly identifying the word that was presented. Notably, if all words were always identified by the participants, then, a diagonal across the matrix would be coloured in yellow. This was not the case here.

It is evident from the plot, that the multisyllabic words served their purpose as more easily lipreadable words for catch trials, leading to an identification probability of >80%. However, contrary to the multisyllabic words, probabilities of correct identification of the remaining, monosyllabic, words were generally substantially lower. Rather, for these words, the confusion matrix analysis pointed to the existence of clusters of confusability. These clusters indicated which sets of nouns were the most similar to each other on a mouth movement level. Prominent clusters included: 1) boat, bone, bowl, 2) cloak, clock, coat, cone, oat, lock, goat, oak, 3) pear, peg, pen, pig, pin, 4) bear, pear, bed, bin, pen, pin.

From these 4 prominent word-clusters all words were selected and kept for use in the subsequent piloting stage, with the exceptions of bear, goat, and pig. It was concluded, upon revision of the nouns in the clusters, that introducing a second animal word in the sentences could add unnecessary complexity to it. In addition to the clustered monosyllabic words, the three multisyllabic words exhibiting the highest probability of being identified correctly, basketball, thermometer, and window, were also selected for use in the final pilot stage (as catch and encouragement trials).

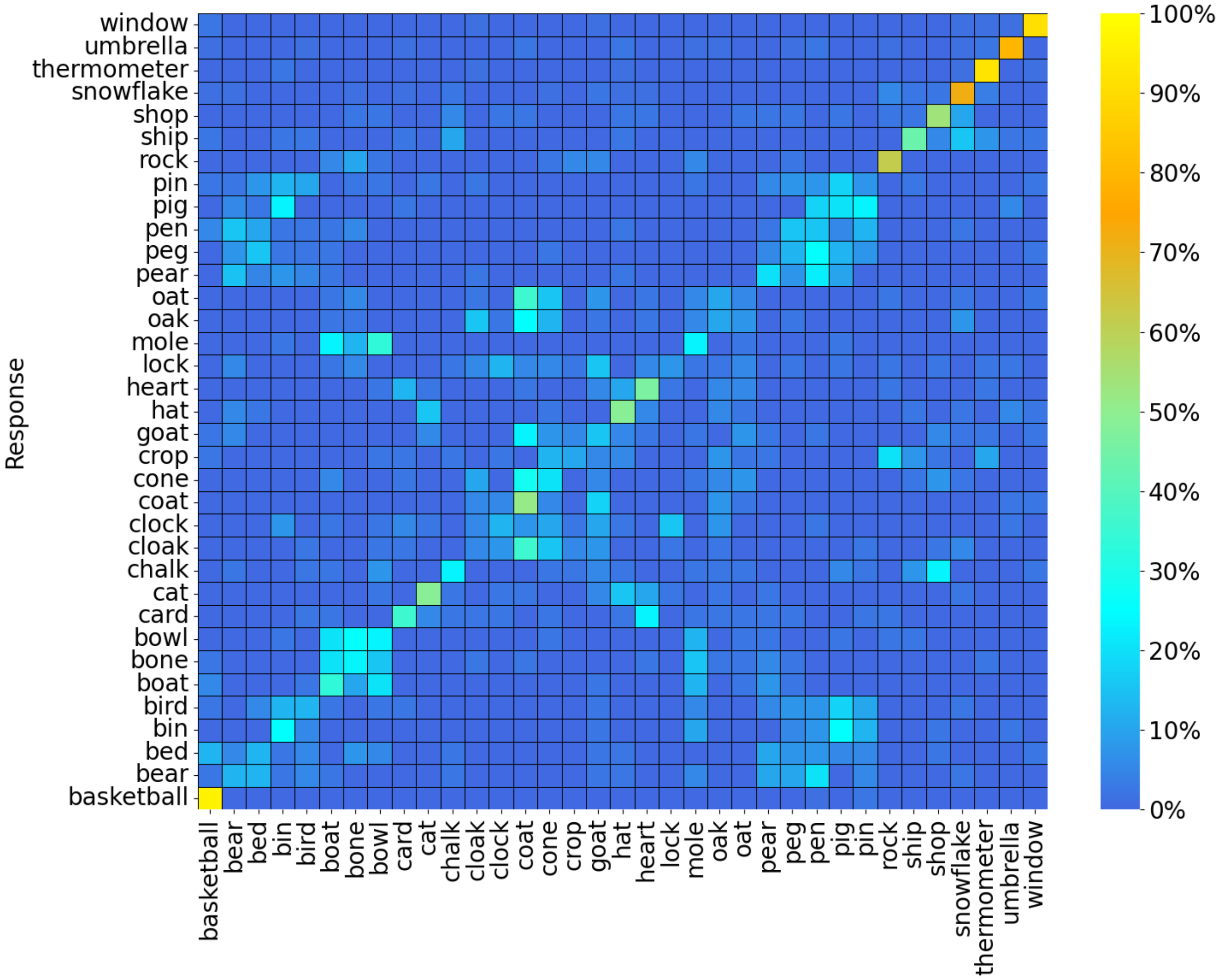


***Figure 2.9**: Confusion Matrix for the Silent vCCRM Nouns Larger set test with 30 monosyllabic nouns and 5 multisyllabic catch/encouragement trial nouns. The colour bar indicates the probability of giving a specific response (y-axis) given the word presented each time (x-axis). Clusters of confusability are evident as blobs of light-coloured cells in the vicinity of the matrix diagonal. Orange to yellow cells indicate high success rate for the catch words basketball, snowflake, thermometer, umbrella and window.*

2.7.4 Pilot Experiment 4: Conclusions and next steps

The main aims of Pilot 4 were to assess the lipreadability of the newly recorded noun stimuli, and to select a subset of nouns from the larger set recorded, for use in the next stages of the task's development. The nouns where hereby selected on the basis of how difficult it was for participants to lipread them, with a preference for the selected words being at least somewhat difficult to lipread. The rationale was that, if the nouns were difficult enough to lipread, their correct identification would depend on accessing sound. By extension, the AV benefit the test would capture would reflect not just lipreading, but potentially also temporal coherence-based mechanisms.

The confusion matrix analysis of Pilot 4 revealed clusters of nouns that were confusable with each other on a lipreading level, confirming that they were promising for use in a task that cannot be solved via lipreading alone. Some of these nouns were, nonetheless, lipreadable up to ~50% of the time, which also adds to their suitability for the speech-in-noise task, as lipreading cues contributing to the

AV benefit. At this stage, however, these results were interpreted with some caution, and instead of proceeding with directly incorporating the words selected in the speech-in-noise task, it was decided to first validate their suitability with a final silent task. The reasons for this were the following:

Although participants were attentive to the task, as indicated by good performances in the catch trials, it was observed that they often gave a response noun that shared neither a common beginning nor a common ending with the presented noun, as if they were choosing their responses at random. Further, although they were encouraged to read each row of the response panel carefully during the practice session and to take into consideration that the nouns are listed in alphabetical order, there was a possibility that their performance may have been affected by the larger set of nouns. It could have also been the case, as it is common in participants running experiments online just to obtain the reward provided, that they were not careful in their responses and were simply trying to finish the experiment as swiftly as they could to move on to a different one.

Considering these potential difficulties, a reduced silent task pilot version, including the nouns identified in the current pilot, was deemed necessary in confirming their suitability. The reduced and final version of the Silent vCCRM Nouns test is described in Pilot 5 below.

2.8. Pilot Experiment 5: Silent vCCRM Nouns Reduced set

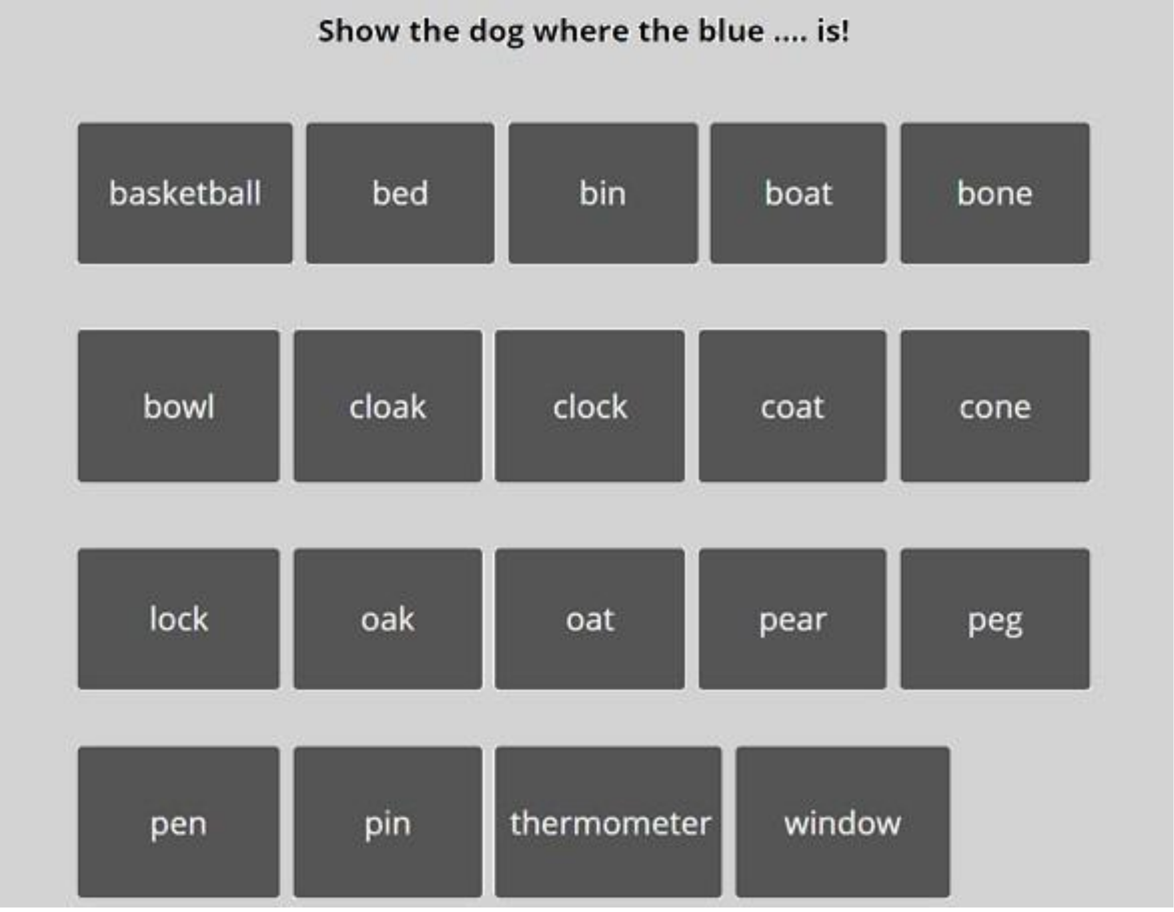
In the previous pilot (Pilot 4), a subset of nouns was selected from the larger set described to be tested as promising candidates for the speech-in-noise version of the vCCRM. Pilot 5, described here, was the final silent version of the vCCRM and aimed to test whether this subset of nouns retained their promising features as candidates (despite being half the size of the larger set of nouns used in Pilot 4).

2.8.1. Participants

Nine participants were recruited (N = 9, 5 females, 4 males; Age: 18-35, Mean: 27.25, SD: 5.55; 8/9 participants provided age information), using the same inclusion criteria as the ones used in Pilot 4. These participants further met the criterion of not having participated in any of the previous pilot stages.

2.8.2. Stimuli and procedure

In making the final version of the silent vCCRM Nouns (Silent vCCRM Nouns Reduced set), the set of thirty monosyllabic nouns plus five easily lipreadable multisyllabic nouns was reduced to sixteen monosyllabic plus three multisyllabic words. These are outlined in Figure 2.10, below. The task was run in the same fashion as Pilot 4 and using the same recording material.



Show the dog where the blue .... is!

***Figure 2.10**: Response panel of the testing stage of the silent vCCRM Nouns Reduced set version. Basketball, thermometer, and window served as catch or encouragement trials; the remaining words constituted the set of monosyllabic nouns words.*

2.8.3. Pilot Experiment 5: Results

As with the previous pilot, mean participant performance was calculated of the total percentage of nouns they identified correctly during the task. This was almost identical to the one observed in Pilot 4 (mean performance there was 24.77%), with participants scoring a mean performance of 27.54% (SD = 8.55%). This finding confirmed, at a first stage, that the selected subset of nouns retained their desirable features, despite the reduction of the response panel. To explore the extent of confusability, and thus the difficulty in lipreadability, confusion matrix analysis was once more applied to the participants' responses (Figure 2.11).

As before, all catch-trial words (basketball, thermometer, and window) served their roles as attentional checks, exhibiting identification probabilities of >90%. Furthermore, bright green/light blue colourings (indicative of ~50% correct identification) along the diagonal of the confusion matrix indicated that participants were, as before, to some extent, able to lipread these words. At the same time, distinct clusters of confusability were once more present. Consistently with Pilot 4, a similar set of four such clusters of nouns were present. These were 1) boat, bone, bowl, 2) clock, coat, cone, lock, oak, oat, 3) pear, peg, pen, pin, 4) bed, bin, pen, pin. Notably, and contrary to Pilot 4, the word "coat" was clearly more easily identifiable than the rest of the monosyllabic nouns in the current pilot. It was, nonetheless, still confusable with other nouns within its cluster. Given these outcomes, I concluded that the subset of words selected was indeed appropriate for use in the speech-in-noise task.

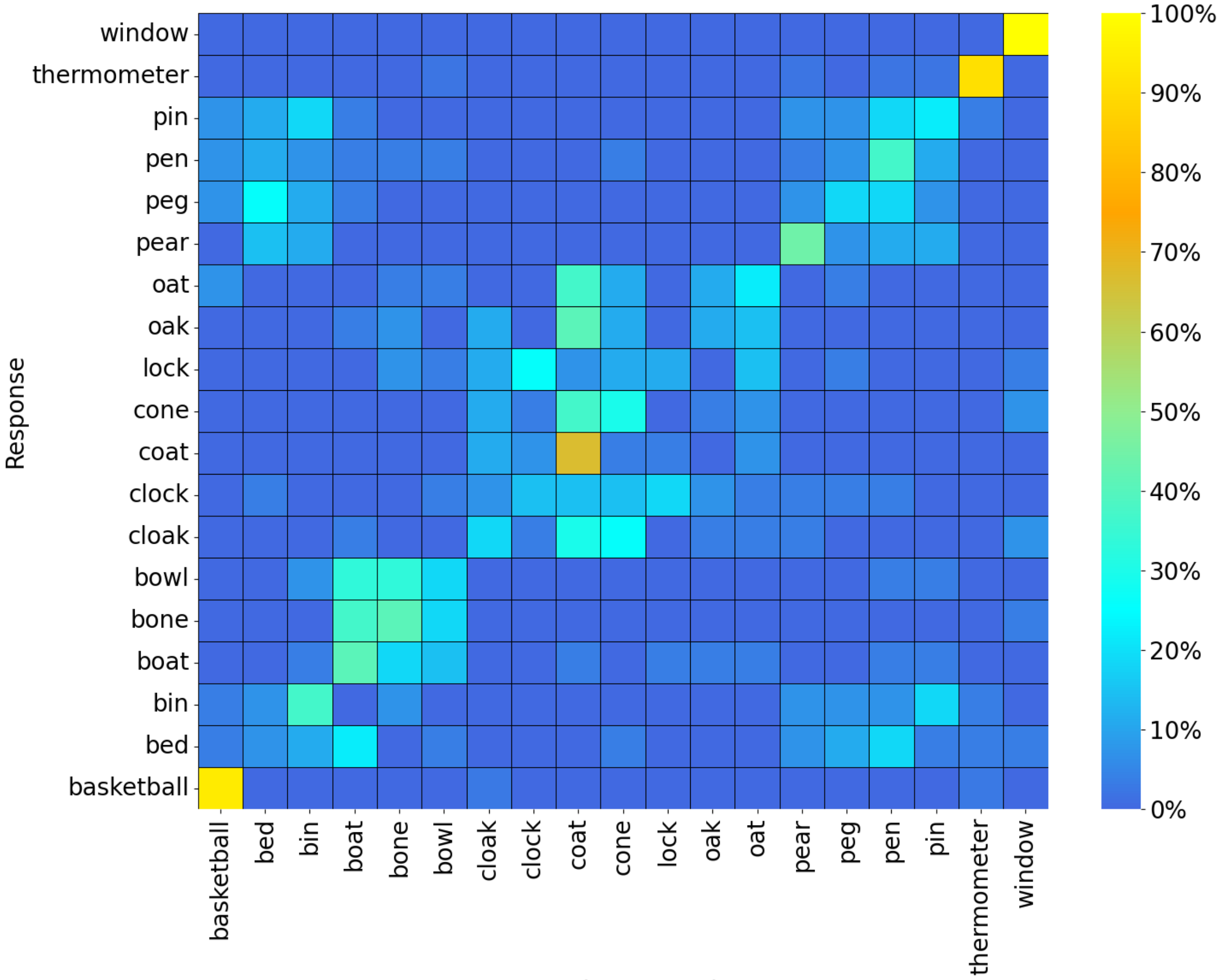


***Figure 2.11**: Confusion matrix analysis of the results of the vCCRM Nouns Reduced set version. The colour bar indicates the probability of giving a specific response (y-axis) given the word presented each time (x-axis). Clusters of confusability are evident as blobs of light-coloured cells in the vicinity of the matrix diagonal. Orange to yellow cells indicate high success rate for the catch words basketball, thermometer, and window.*

2.8.4. Pilot Experiment 5: Conclusions and next steps

Pilot 5's results confirmed that the candidate nouns maintained their confusability, in terms of lipreading, when the task options were reduced. These results were deemed promising, and in line with the thesis goals. As mentioned several times before, these goals included the assessment of the existence of additional, language-independent, mechanisms by which vision may contribute to the AV benefit. In light of the outcome of Pilot 5, the decision was made to proceed with the development of the final, speech-in-noise version of the task, which I dubbed vCCRMn (video version of the Children's Coordinate Response Measure with nouns at the end).

Thankfully, COVID-19 measures were eased by the time Pilot 5 was finished, allowing human research to resume once more, in-person, within the academic setting. In the next section, I detail the steps undertaken to prepare the vCCRMn speech-in-noise task.

2.9. Making the vCCRMn speech-in-noise test: From anechoic chamber recording preparations to the final pilot

In the previous sections, I detailed the piloting stages 1-5, which were taken to ensure that the stimuli of the vCCRM were appropriate for use in my speech-in-noise task. Importantly, the last two Pilots 4 and 5 confirmed that the vCCRMn stimuli were promising for use in the speech-in-noise task, vCCRMn.

Since COVID-19 regulations were eased after the last pilot (Pilot 5), all steps from this point on happened in-person, following strict COVID-19 safety measures for the protection of both participants and experimenter.

The remaining sections of this chapter detail all the steps that were taken to put the vCCRMn task together. Briefly these included: 1) talker selection for the recording of the audio-visual sentence material, 2) sentence material recordings, 3) recorded stimuli processing steps, and 4) experimental design decisions. Finally, this chapter concludes with the first (and final) pilot stage of the vCCRMn as a speech-in-noise task before its official launch.

2.9.1. vCCRMn talker selection for the sentence stimuli recordings

As soon as governmental COVID-19 measures had eased, allowing human research to resume, an advertisement was circulated at the UCL Ear Institute for the selection of talkers to take part in the vCCRMn speech-in-noise task. Eight native British English speakers (four male, four female) with a Standard Southern British English accent, no reported language-related difficulties, and ability to sit still for long recording sessions, were recruited to record the material for the task at UCL's anechoic chamber (21 Gordon square, Bloomsbury campus). Personal meetings were arranged with every individual who expressed interest to ensure suitability for the recordings before selection.

The anechoic chamber featured a design that incorporated a "room within a room" structure. It was constructed with external walls measuring 330 mm in thickness, while the interior space was lined with glass-fibre wedges. Additionally, the inner chamber consisted of metal acoustic panels that were installed on a floor designed to float. I trained in using the anechoic chamber, and allocated a week to practice testing volunteers, initially in the presence of an acoustic-electronics technician, and subsequently independently, before completing the recordings of my first four talkers.

Each of the talker recording sessions were one-day long, including breaks. Unfortunately, due to a technical issue of the recording apparatus, the recordings of these first four selected talkers (two male, two female) were deemed unusable and had to be discarded. A subsequent recording of one of the four talkers took place after a solution to the technical issue was suggested, however the solution led to a suboptimal recording which was also deemed unusable and discarded.

Shortly after that, another call was advertised for the talker role, this time across UCL, seeking 2 male and 2 female talkers (inclusion criteria same as above). A total of forty candidate talkers expressed interest, all of whom were interviewed. Four talkers were selected following the completion of the interview stage. Recordings of those four talkers were successful via the implementation of a different solution to the unresolved anechoic chamber problem. All talkers recorded in the anechoic chamber gave written consent and were paid for their time.

The recording procedures are outlined in the sections below.

2.9.2. Anechoic chamber recordings

2.9.2.1. Preparations

Each of the four talkers (Ages: 23-29, Mean = 25.75, SD = 2.17) was allocated a one-on-one recording session. To ensure the optimal recording of audio-visual material talkers were asked to: a) Not dress in highly patterned shirts and pinstripe, thus ensuring that the camera autofocus would not create unwanted effects (e.g. image buzz or moiré), b) not wear a red top, to avoid causing a halo effect, and c) instead opt for a plain-coloured top, as this would offer some, but not too much contrast against the dark blue/purple background of the anechoic chamber, d) not wear clothing with visible logos or statements, thus ensuring that the viewer would not find these distracting, e) if applicable, wear no or minimal make-up and finally, e) remove all jewellery to prevent reflections from the anechoic chamber's lights. Talkers confirmed ability to follow those instructions before their appointment date.

In addition to the anechoic chamber, the anechoic room space included a control room that contained associated recording and computing equipment. Before the arrival of each talker, the control room and anechoic chamber were prepared. As recordings happened soon after the COVID-19 pandemic, even though regulations were eased to allow human research again, COVID-19-related safety measures had to be taken for the protection of both experimenter and talkers. To this end, the rooms were ventilated, all surfaces were disinfected, both before and after recording sessions, and, where possible, face masks were worn.

Following room preparation, all equipment in both rooms were set up and checked for the recording session. Namely, the Speech Prompt and Record System program (ProRec 2.4., https://www.phon.ucl.ac.uk/resource/prorec/) was set up in the control room for the projection of the experiment's sentences to be read by the talkers on the teleprompter in the anechoic chamber. Further, the software Audacity (https://www.audacityteam.org/) was used to confirm sound was being picked up, via clapping from within the control room, with the anechoic chamber's door open. Finally, the Fireface sound-recording equipment (https://www.rme-audio.de/fireface-ucx.html) was switched on, and pre-prepared stimuli recording settings were loaded to its respective software on the computer.

These preparation steps were completed before the arrival of each talker.

2.9.2.2. Recording of audio-visual material

Upon arrival, each talker was welcomed and provided with an oral and written explanation of their task and of the recording process. They gave written consent to proceed. They were, then, presented a short video of an example talker reading some of the sentences that they were to read in the anechoic chamber. Prior to the video presentation, they were asked to follow the example talker in terms of a) their pace of reading the sentences (not too slow and not too fast), b) the lack of putting emphasis on any part of the sentence read, and c) their lack of facial expressions. They were asked to read the sentences as they would normally, and at a loudness level they would use in everyday conversation.

The talkers sat on a comfortable chair that was firmly fixed to the anechoic chamber's floor to ensure the position of the talker remained the same throughout the recording session. The anechoic chamber was then set up based on the talker's characteristics. The chair was adjusted such that the teleprompter's screen was at an eye's level both, ensuring that the talkers could comfortably read the displayed sentences, and establishing a natural and relatable perspective for the viewer of the experiment. The talker's head, neck and shoulders were directly facing the video-camera used for the video recordings (Canon Legria, HFG30 HD camcorder, recording at 25 frames per second), and were centred on the screen (Figure 2.12). Their eyes were framed using the 'rule of thirds' technique according to which the frame is divided into a grid of nine equal parts with two horizontal and two vertical lines. The talker's eyes were centred and positioned on the top horizontal line of the grid for a balanced shot.

Three battery-operated, light projectors in the anechoic chamber were adjusted to reduce shadows on the talker's face. After the adjustments, a still frame of the talker was taken and used as a reference to ensure that their position was consistently the same after every recording initiation (e.g. following a break). Finally, the loudness level at which the talker was speaking was checked with a sound-level meter (Bruel and Kjaer, 2231) to ensure it was between 60 – 70 dB SPL, and the talkers were instructed to speak louder, or quieter accordingly.

The anechoic chamber door was then closed, from which point on the talker was monitored through a screen in the control room that projected what the camera in the anechoic chamber was recording. Communication was possible through a microphone in the control room connected to speakers in the anechoic chamber. The communication microphone in the control room was always turned off during recordings, to avoid interference with the recorded material.

The session started with a practice recording serving the dual purpose of allowing the talkers to feel more comfortable reading the sentences off the teleprompter screen as well as a quality check for the recorded material. Talkers were provided with feedback on their sentence-reading during the practice session. Their voice was being recorded via a high-quality microphone (Bruel and Kjaer, 2231) positioned underneath the video camera in the anechoic chamber. The software Audacity allowed the listening through the sentences read and checking that loudness peaks were always between -20 dB FS and -12 dB FS, ensuring maintained signal strength but at the same time avoiding clipping effects. It also allowed the visualization of sentence waveforms when played to confirm they were within the ideal limit of being comfortably below 0.5 on the normalised amplitude scale. If the audio was too quiet or too loud, the gain was increased or decreased respectively using Fireface. Following the first practice session, a second practice session was then conducted as an additional quality check, and to ensure the talker felt comfortable with reading the stimuli.

The main recording session started as soon as the talker confirmed they were comfortable to proceed, and the sound checks were completed. The recording session was broken into small, manageable chunks with breaks in between. During the recordings, several checks and steps had to occur simultaneously which included the following:

a) Each read sentence was compared against a printed list containing all the sentences to confirm that every word in each sentence was correctly read. Every incorrectly read sentence was marked and was recorded again in a separate session.

b) The loudness level of the read sentences was continuously monitored to ensure it stayed within appropriate limits.

c) The battery-operated lights illuminating the talker's face were monitored to replace their batteries if they were dimmed or had run out.

d) Each recorded talker was asked to silently count to three before and after each sentence they read, to ensure sufficient separation between stimuli, for post-processing reasons. Silent counting was also checked to ensure it was always followed.

The stimuli read had the following format: "Show the <animal> where the <colour> <noun> is." where animal could be any of [cat / cow / dog / duck / pig / sheep], colour any of [black / blue / green / pink / red / white] and noun any of [bed / bin / boat / bone / bowl / cloak / clock / coat / cone / lock / oak / oat / pear / peg / pen / pin] (the set of sixteen monosyllabic nouns selected in pilot stages 4 and 5). Thus, the recordings yielded a total of 2304 sentences (576 sentences from each talker).

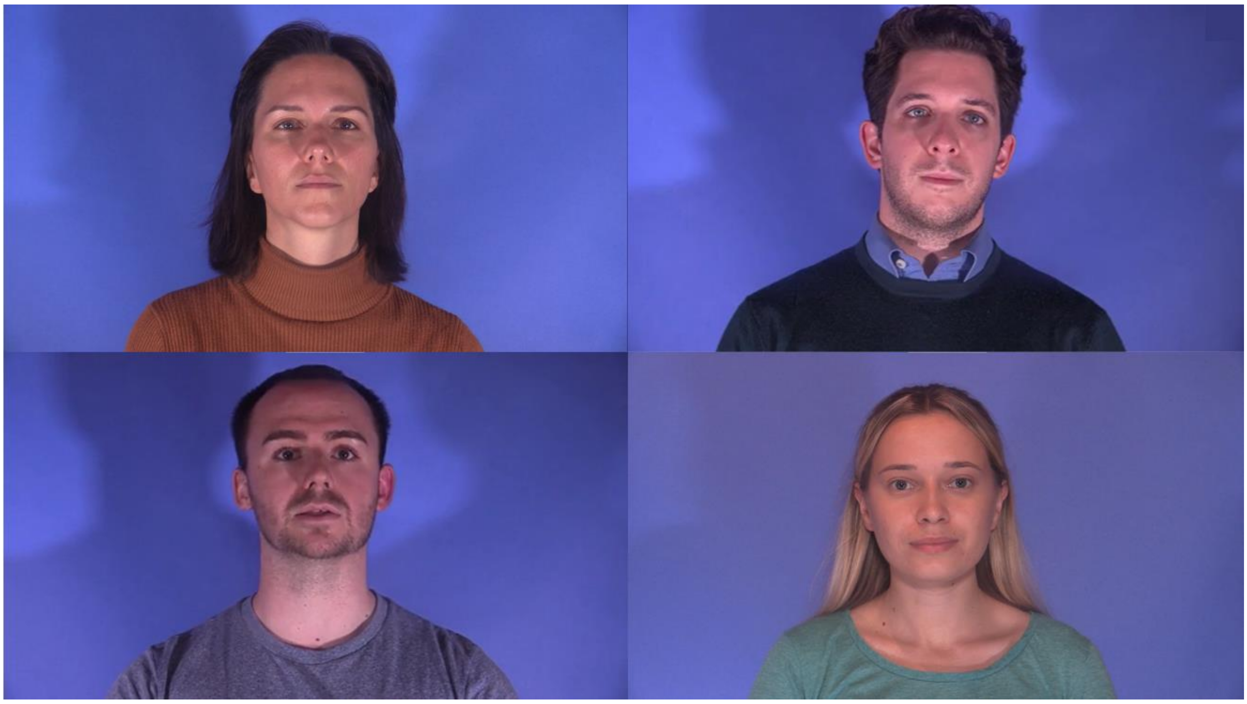


***Figure 2.12**: The talkers of the vCCRMn.*

2.9.2.3. Post-processing of recorded audio-visual stimuli

The recorded material consisted of audio recorded from the microphone, and video recorded from the video-camera (camera recorded audio was not used in the stimuli). The audio and video of each recording were first synced with Adobe Premiere Pro (https://www.adobe.com/uk/products/premiere.html), and subsequently segmented into individual-sentence stimuli with MATLAB and the software suite for handling multimedia data FFmpeg (https://ffmpeg.org/). The waveforms of the segmented sentences were independently visualised to ensure they were centred. As mentioned before, during recordings, the talkers were asked to count to three prior to, and after reading each sentence. At this stage, this enabled segmentation of the recordings into stimuli within which the sentences were roughly centred. Where that was not the case, silence was added to the beginnings and/or ends of the stimuli using MATLAB to ensure the sentences occurred in the middle and were 3 seconds long.

Having centred the stimuli, and confirmed they had equal lengths, an additional post-processing step was conducted in MATLAB to identify the time points at which the <colour> <noun> pair began and ended in each sentence (details in section 2.9.3.2.). These time points were stored and were also exploited during experiments to ensure the colour and noun words of all sentences presented per trial (both target and maskers) overlapped in time (thus ensuring the maskers indeed masked the target). This was achieved by adjusting the starting time of the masker sentences presentation within a trial.

Following some final checks, that ensured all modifications of the stimuli produced the expected outcomes, the material was ready for experimental use. The next sections outline the different conditions employed in the speech-in-noise task experiments.

2.9.3. Conditions of the vCCRMn

As mentioned on several occasions throughout this thesis, among my main aims for this project, was to assess whether the AV benefit enjoys contributions from both lipreading, and temporal coherence-related mechanisms. The stimuli developed herein, as described in Pilots 1 to 5, were tailored to, but were not the only measure taken towards this goal. The conditions of the vCCRMn speech-in-noise task were also designed with that aim in mind.

The primary conditions employed within the task were three, and each was further divided into two sub-conditions. The three main conditions were the following: 1) An audio-visual or "AV" condition, 2) an interrupted or "Inter" condition that removed lipreading cues during target word presentation and, 3) an audio-with-static-image, or "A" condition, that served as the audio-only baseline condition for computing AV benefits. The sub-conditions included a target-coherent, and a masker-coherent version of each of the main three. Target-coherency implied that the talker presented in the video, was also the talker uttering the target word sentence. Masker-coherency implied that the talker presented in the video, was the talker uttering one of the two masker sentences.

These conditions formed a solid basis for investigating the questions of this thesis. An audio-visual benefit provided by the "Inter" condition would, for example be reflective of contributions from mechanisms beyond lipreading, and likely relating to temporal coherence. Similarly, differences between performances in the target- and masker-coherent sub-conditions, would, based on object-based attentional theories, point in the same direction (Bizley et al., 2016; Lee et al., 2019; Maddox et al., 2015; Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008).

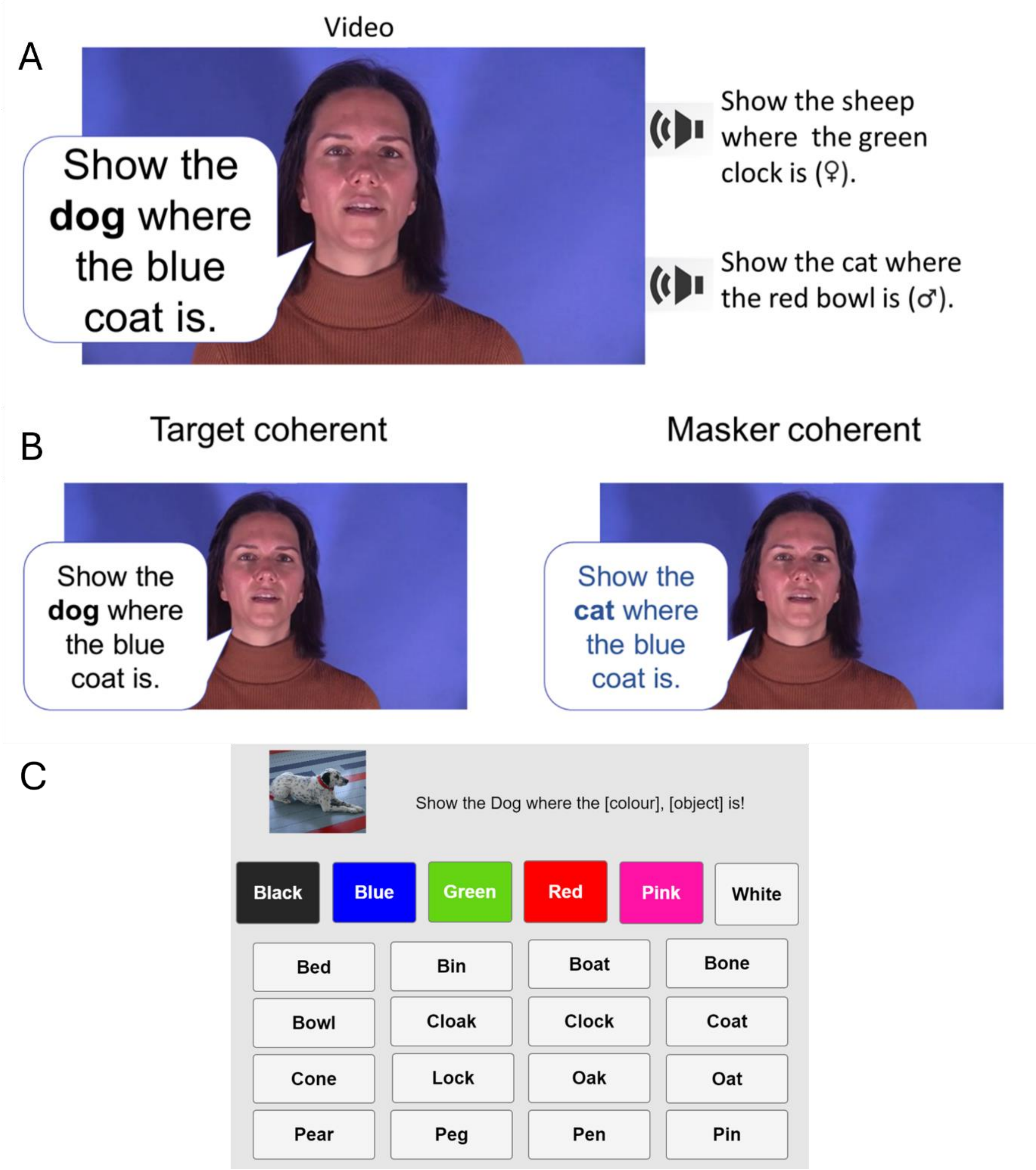
Nonetheless, regardless of condition, the participants' task during the vCCRMn was always the same. It was: 1) To follow the voice that said the sentence containing the animal "dog" in it to report the <colour> <noun> words at the end of that sentence, while simultaneously ignoring the voices of the two masker speakers in the background each saying a sentence of the same form as the target sentence but with different animal. 2) To direct their eye-gaze to the lips of the talker on the screen without effortfully attempting to lipread them, so that their scores in the task reflected a more "natural" performance.

The task randomly selected a target talker, which could be any of the 4 talkers recorded. The two masker talkers were also randomly chosen; one was always male, and the other was always female, as long as they were different from each other and from the target talker. Finally, in each of the visual conditions, only one talker appeared on the screen.

2.9.3.1. Audio-visual (AV) condition

The audio-visual, or AV condition, could be described as the "naturalistic" condition of the vCCRMn. It served as a simulation of more realistic, everyday life listening in noise conditions. The AV condition provided the full video of a speaker, thus making available mouth movements conveying both lipreading and timing information to the participant. As explained before, it consisted of both a target-coherent sub-condition (AV target), where the talker of the video provided matched the voice of the target talker, and a masker sub-condition (AV masker), where the video provided matched the voice of one of the masker talkers. The AV target condition was expected to capture the participants' total AV benefit.

An example of the AV condition, including the target- and masker-coherent variations, and the task's response panel are shown in Figure 2.13.


A
Video
Show the dog where the blue coat is.
Show the sheep where the green clock is (♀).
Show the cat where the red bowl is (♂).
B
Target coherent
Masker coherent
Show the dog where the blue coat is.
Show the cat where the blue coat is.
C
Black
Blue
Green
Red
Pink
White
Bed
Bin
Boat
Bone
Bowl
Cloak
Clock
Coat
Cone
Lock
Oak
Oat
Pear
Peg
Pen
Pin

Show the Dog where the [colour], [object] is!

***Figure 2.13**: An example of a trial presentation of the audio-visual, or AV condition of the vCCRMn. (A) Participants were tasked with listening out for the sentence that included the word "dog" while ignoring the voice of two (one male, one female) masker talkers in the background. (B) Example target-coherent (left) and masker-coherent (right) stimuli. (C) Response panel of the task.*

2.9.3.2. Interrupted (Inter) condition

The Interrupted, or Inter condition, was the same as the AV condition described in the previous section. Their only difference was that in the Inter condition, the entire screen (and hence, the lip movements) froze from the definite article ‘the’ before the first target word <colour> and until the end of the <noun> word, thus, eliminating the possibility that participants were able to access lipreading cues in identifying the target words during the condition.

The freezing of the screen was achieved as follows: Section 2.9.2.3, explained that the time points at which the <colour> <noun> pair began and ended in each sentence were identified and stored. This information was exploited in the creation of material for the Inter condition of the task. MATLAB was hereby used to freeze the video of the stimuli for the time between the identified time points of <colour> <noun> pair presentations. The time point identifications were conducted manually for each sentence. To do so, sentence waveforms were visualised, and audio played; then time point 1 (beginning of screen-freeze) and timepoint 2 (ending of screen-freeze) were identified on the waveform. While both <colour> <noun> words had to be correctly identified for a successful trial, additional measures were taken to ensure mouth movements preceding and following the <colour> <noun> pair did not bias participant responses. More specifically, time point 1 for the video freezing was selected to occur during the presentation of the definite article “the” (/ðə/) that preceded the <colour> (“Show the <animal> where *the* <colour> <noun> is”). Namely, it was chosen to be the time point right after the “th” sound (/ð/) of the article, so that the “e” (/ə/) sound would also be part of the frozen <colour> <noun> presentation. The reasoning for this choice was that the lip movements in the /ə/ sound of the definite article, could change depending on the first letter of the <colour> word that followed it. Freezing the video just after the “th” (/ð/) sound of “the” (/ðə/) eliminated every possibility that the altered mouth movement during the article could be helpful to the participant in identifying the <colour> word. Similarly, timepoint 2 for the end of the screen-freeze was as close to the final word “is” as possible.

As with the AV condition, the Inter condition included both a target-coherent (Inter target) and a masker-coherent (Inter masker) sub-condition. During a masker-coherent trial presentation, the video-freezing method was applied to the masker video.

2.9.3.3. Static image with audio (A) condition

In the static image with audio, or A condition, participants were tasked with identifying the target words contained in the sentence that included the animal “dog”. The visual background of this condition was, however, not dynamic. Instead, for the target-coherent trials (A target), a still image of the target talker was presented, whereas for the masker-coherent trials (A masker), a still image of one of the maskers was presented. This condition served as a control, confirming that presentation of non-dynamic visual stimuli was not influencing participant performances. Additionally, it served as a baseline condition, against which the visual benefits conferred to the participants by the video conditions (AV and Inter), were computed.

2.9.4. Conclusions and next steps

Section 2.9 described the process of putting together the speech-in-noise task vCCRMn, from talker selection and anechoic chamber recordings to stimuli processing and the experimental design of the final speech-in-noise task’s version. Having completed these, the next step was to run the first, and final, pilot of the vCCRMn (described in Pilot 6, below) before its official release and participant testing phases (described in subsequent chapters).

2.10. Pilot Experiment 6: Final piloting of the vCCRMn

Pilot 6 served to confirm that the sets of nouns chosen in Pilot 5 in the final form of the vCCRM, were appropriate for use in the speech-in-noise vCCRMn. That is, it served to confirm that the current stimuli overcame the issues met in previous piloting stages. It also helped confirm that the task ran smoothly, no bugs or errors were present, and that the data were being stored in a manner that facilitated subsequent analyses. Further Pilot 6 served as a preliminary, but informative, investigation of how vision impacted speech-in-noise performance across the different conditions of the task.

2.10.1. Participants

As COVID-19 regulations were, at this stage, eased to allow in-lab testing, the current experiment took place at the UCL Ear Institute's human labs. A total of six participants (five non-native but fluent in English UCL PhD students, and one native speaker of English, Ages: 31-41, Mean = 34.16, SD = 4.89, 3 female and 3 male) with no intrusive tinnitus and no known hearing loss gave informed consent to participate in the speech in noise task. They were all paid for their participation.

2.10.2. Stimuli and procedure

Contrary to experimental Pilot 1, which varied SNR via a 1U2D and 1U3D procedures, for the current speech-in-noise task, SNR was set to vary using a 1U1D staircase procedure, tracing participants' SRT_{50}. The reasons for changing to a 1U1D procedure were two-fold: Firstly, the 1U1D staircase rule led to faster convergence to its respective SRT than 1U2D and 1U3D. Faster convergence implied the test would be more easily adaptable for use in the clinical setting, should this option ever materialise in the future. Further, given the large number of tests I planned to use on participants, being able to swiftly assess their speech-in-noise performances would be helpful. Secondly, and perhaps more importantly, a convergence on SRT_{50} would create a large enough gap between the audio-visual, and audio-only, conditions, to extract substantial AV benefits from participants (see also section 1.5.1.3). Given the difficulty of the task, hinted to in Pilots 4 and 5, it was deemed fitting to aim for a staircase procedure that better showcased the AV benefit.

Each main condition was run 3 times and in a randomized order across all participants. Within each run, the target-coherent and masker-coherent conditions were presented randomly. The output of the experiment provided 18 SRTs, one for each run and for its equivalent target-coherent and masker-coherent sub-conditions (3 runs x 3 conditions x 2 sub-conditions per condition). Each SRT was computed as the mean SNR of the final four reversals of its respective run.

For the experiment, participants were asked to sit in front of a laptop in a sound-proof booth and wore closed-back, around the ear headphones (Sennheiser, HD 280 Pro Dynamic HiFi Stereo). They were instructed to direct their eye-gaze to the mouth of the talker on the screen. A camera sitting on top of the laptop screen was used as a confirmatory check that they were always attending to the display screen, during trial presentations.

2.10.3. Pilot Experiment 6: Results

The results of the piloting phase of the vCCRMn for all six participants are presented in boxplot form, in Figure 2.14 below. SNR scores presented here were computed as the mean of the three SRTs output for each of the three runs participants had to undertake for each condition. The horizontal dotted lines in the figure indicate the mean participant performances for the AV target (Mean = -3.52 dB, SD = 1.73 dB), Inter target (Mean = -0.75 dB, SD = 2.26 dB), and A target (Mean = -0.97 dB, SD = 2.26 dB) conditions. These were deemed to be promising, as discussed in the following sections.

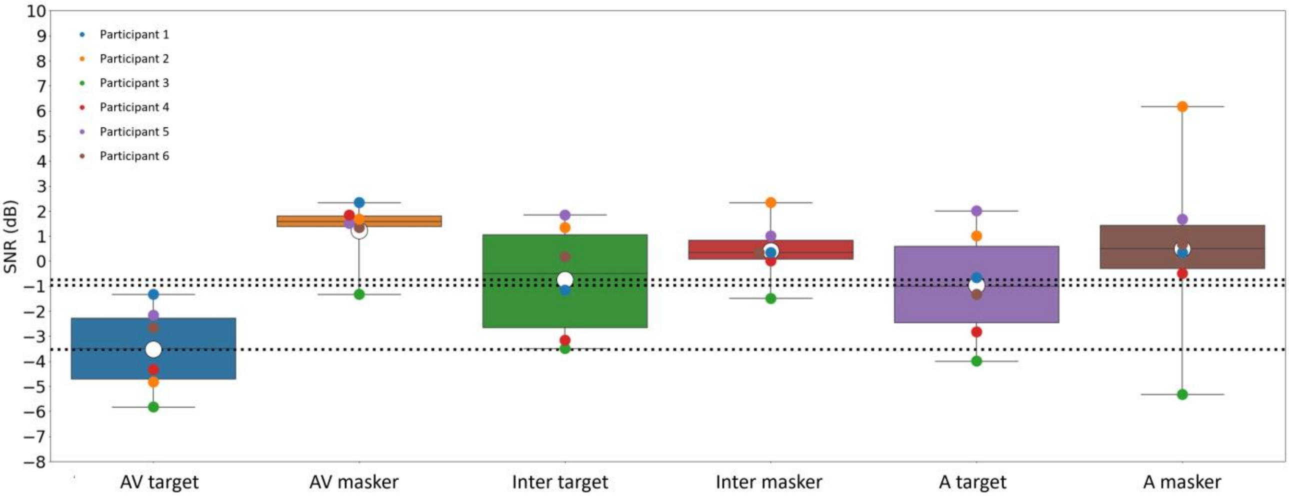


***Figure 2.14**: SNR scores for each of the six conditions of the vCCRMn for the participants of Pilot 6. Dotted lines cross the mean performance (white dots) of each of the target conditions.*

2.10.4. Pilot Experiment 6: Conclusions and next steps

As alluded to in the former section, the results from Pilot 6 were promising, as, in the least, they indicated that performances in the AV target condition were distinguishable, and better, than those of the rest. Further, these performances were not reaching implausibly low SNR levels, as was the case with the speech-in-noise task of Pilot 1. Thus, it was concluded that the vCCRMn speech-in-noise task was ready for its official launch, and official participant testing phases.

Chapter 3: General methods

3.1. Introduction

In the previous chapter, I outlined the piloting stages, methods, processing steps, and general considerations I took to develop the vCCRMn, the speech-in-noise task that forms the foundation of this project. I concluded the chapter with experimental Pilot 6, where I confirmed the suitability of the task for use in this research. In the current chapter, I detail out the methods used in what I refer to as the “official testing phase”. The official testing refers to the testing that started after (and excluding) Pilot 6 and involves participant recruitment for in-person testing in the lab, and the experimental procedures, from participant welcoming to the end of their experimental session. I also include detailed descriptions of all the tests (except for the already described speech-in-noise test vCCRMn described in Chapter 2) that I employed in my experiments.

Thus, this chapter serves as a reference chapter of all the information relevant to the experiments conducted during the official testing phase. In the same spirit, this chapter also summarises all, towards the end, all the statistical and analytical methods used in the analyses of the data collected during this experimental phase. These analyses constitute the main bodies of Chapters 4 and 5. The methods employed in the development of Linear Mixed Effects Models (LMEMs), which were also applied on data collected during the official experimental phase, are only briefly touched on here. It was deemed more appropriate to include these in detail in Chapter 6, which forms a self-contained chapter concerned with the development of the LMEMs.

3.2. Participant recruitment

To assess the potential effects of ageing, and age-related hearing loss on audio-visual integration, the official testing phase was divided into three experimental groups. The first experimental group included the younger (aged ≤ 35 years) normal hearing participants. The second experimental group consisted of older (aged > 35 years) normal hearing participants, and the third of older (aged > 35 years) participants with hearing loss.

In accordance with UCL ethical guidelines, all participants gave informed written consent before their testing session. They were also paid for their participation. The study was approved by the Research Ethics Committee of University College London (ref: 3866/003). In total, the experimental pool consisted of 125 participants (Age range 19 – 86) that were tested in-person at the UCL Ear Institute human labs.

To achieve this, several different platforms and forms of advertising were exploited for recruitment. Younger participants were contacted through the UCL Ear Institute and the UCL SONA subject pools and were mainly students. Older participants were recruited via 14 groups of the University of the Third Age (U3A) (https://www.u3a.org.uk/) in the UK. Recruitment methods included all of mailing lists, work calls and flyer advertisements.

Over 250 participants responded to these advertisements, and further correspondence, and/or phone calls were arranged with each confirmed at a first stage, whether they were suited for participation. Generally, to participate in the study, a participant had to:

1. Be a native speaker of English.

2. Be 18 years old, or older.
3. Have normal, or corrected-to-normal vision, and normal colour vision.
4. Have no neurological, psychiatric, or developmental conditions.
5. Have not received training in lipreading.
6. Have no intrusive tinnitus.
7. Have no language-related difficulties.
8. Based on audiological standards (British Society of Audiology, 2018), have normal hearing, if younger (< 35 years old), and, if older, have either normal hearing or mild and/or up to and inclusive of moderate hearing loss.
9. Be able to sit still and maintain focus for the duration of the study (with brief breaks every 5-10 minutes).
10. Be able to avoid exposure to loud sounds in the 24 hours preceding their participation to the study.

These criteria were listed in all advertisements of the study, and were re-shared, in more detail, with participants that had expressed interest in participating in the study – the criteria were also included in an information sheet that participants were provided with on the day of their participation. Participants who had confirmed they met all criteria, were offered a conditional testing session. Ability to complete the practice session (outlined in section 3.6.3), and an audiometric test (section 3.6.4), formed the final checks, before they proceeded with the main experimental session.

3.3. Experimental procedure overview

Participants undertook an array of tests, during the testing session. These included completing the vCCRMn speech-in-noise task (as well as related demo, and practice sessions), a pure-tone audiometry (PTA) test, a silent lipreading test, a cognitive ability test (if over the age of 50), and a talker identification task. Figure 3.1 below serves as a reference figure, outlining the study's experimental procedure.

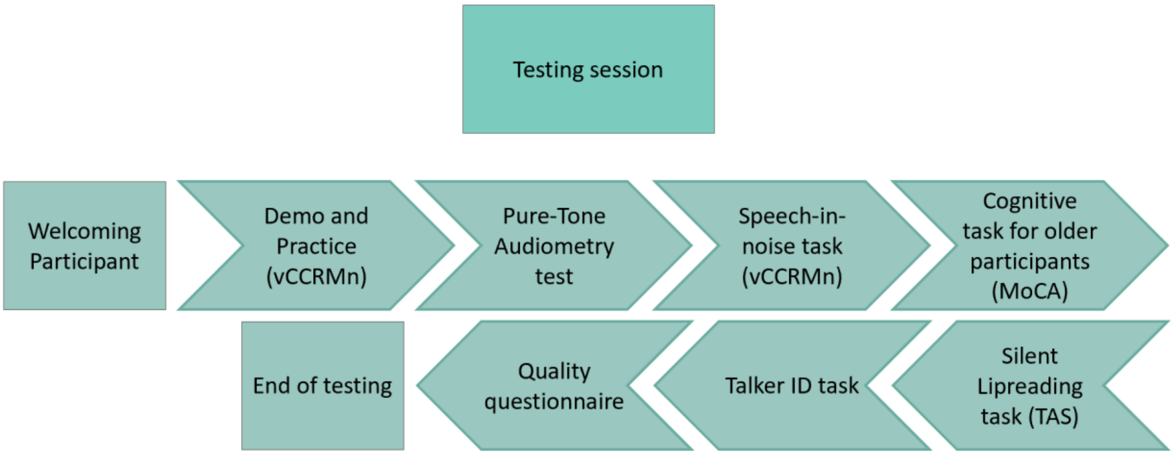

Testing session
Welcoming Participant
Demo and Practice (vCCRMn)
Pure-Tone Audiometry test
Speech-in-noise task (vCCRMn)
Cognitive task for older participants (MoCA)
Silent Lipreading task (TAS)
Talker ID task
Quality questionnaire
End of testing

Figure 3.1: *Official testing phase experimental procedure: From participant welcoming to the end of their testing session.*

3.4. Equipment and apparatus

Unless otherwise specified, all testing procedures were completed within a triple walled Industrial Acoustics Company (IAC) soundproof booth. Tasks were run on a Dell latitude laptop, and participants wore closed-back, around the ear headphones (Sennheiser, HD 280 Pro Dynamic HiFi Stereo). The headphones were tested and calibrated using an artificial ear (Bruel and Kjaer) and a measuring amplifier (Bruel and Kjaer, 3110-003), such that sound level output was at 65 dB SPL (RMS normalised).

Participants were instructed to maintain their eye-gaze on the task display screen during task presentation. A camera positioned on top of the screen and connected to a computer in the experimenter's control room, was used to ensure that this instruction was being followed. Participants could at all times communicate with the experimenter via a microphone set up within the booth.

Participant hearing status was assessed with a calibrated audiometer (Interacoustics, AS629, with Radioear DD450 headphones).

3.5. vCCRMn stimuli

Detailed descriptions of the sentence stimuli recorded for the vCCRMn task were provided in Chapter 2. Briefly, 2304 sentence stimuli were recorded in total, 576 for each of two male and two female talkers, and each sentence was 3 seconds long. The recordings' sampling rate was 44100 Hz. Each stimulation trial of the vCCRMn was comprised of three simultaneously presented sentences – one target and two maskers. Trials included either a male, or female target talker, and both male and female maskers. They were presented at a variable SNR, following a 1U1D staircase procedure.

The sentences had the form "Show the <animal> where the <colour> <noun> is". Specifically, the target sentence always contained the animal "dog", and different animals, colours, and nouns were used in each sentence. Animal options included any of [cat / cow / dog / duck / pig / sheep], colour any of [black / blue / green / pink / red / white] and noun any of [bed / bin / boat / bone / bowl / cloak / clock / coat / cone / lock / oak / oat / pear / peg / pen / pin]. Figure 3.2 below illustrates mean Fast Fourier Transform (FFT) spectra calculated from the sets of sentences recorded from each talker.

During each trial presentation, the sentence stimuli were accompanied by either a video of a talker (AV and Inter conditions), or a static image of a talker (A condition). These would either match with the target sentence (target-coherent conditions), or with one of the masker sentences (masker-coherent conditions). Participants were tasked with identifying the colour and noun words of the target sentence at each trial.

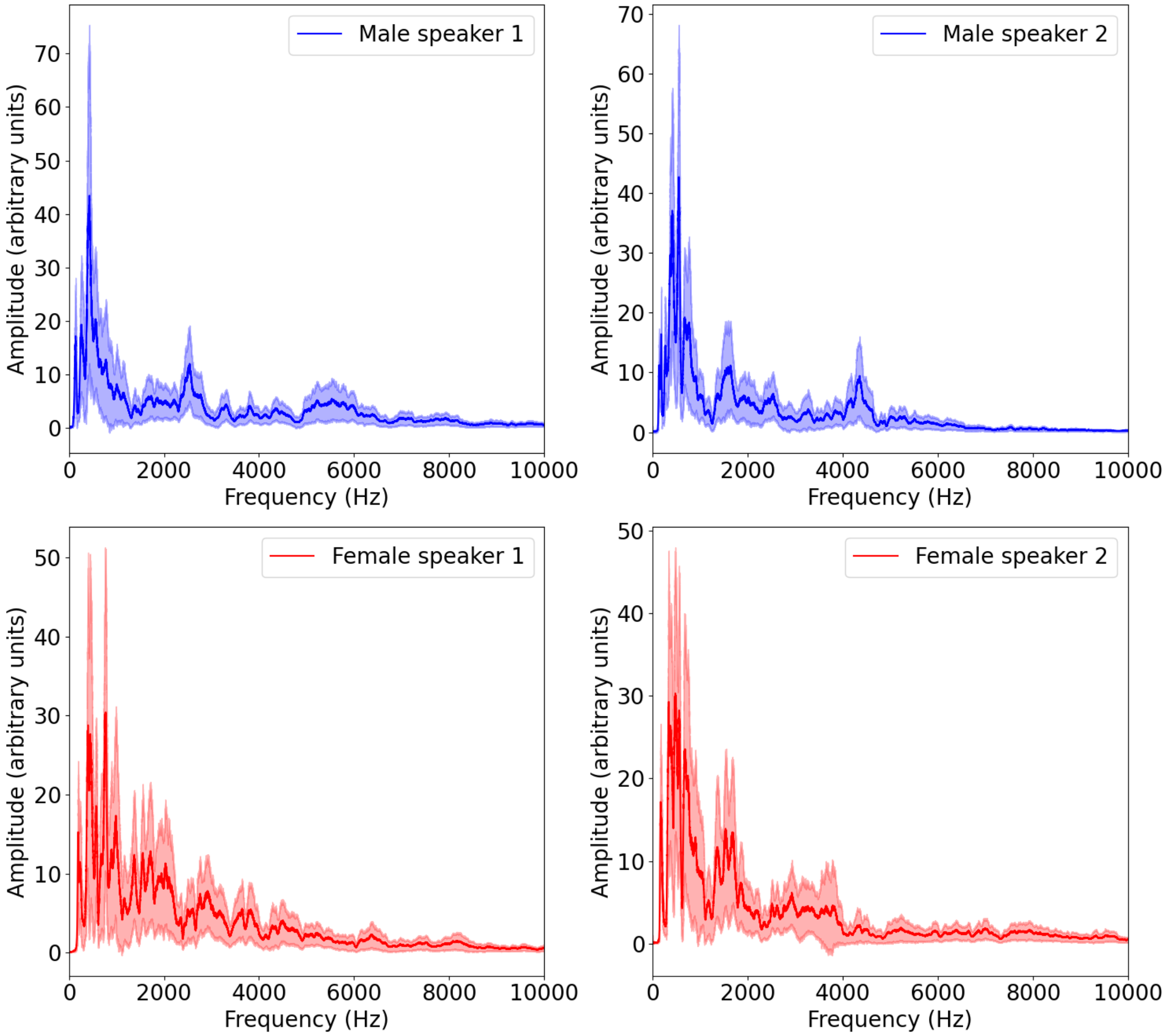


Figure 3.2*: Mean FFT spectra for the sentences read by each of the four speakers of the vCCRMn task. Each panel includes the mean of 576 sentences; errors are standard deviations. The top panels depict the mean spectra from the sentences read by the two male talkers while the bottom panels show the mean spectra from the sentences read by the two female talkers.*

3.6. Experimental procedures: Detailed descriptions

3.6.1. Participant arrival

The main testing phase of this project started soon after COVID-19 measures were eased, allowing human research to resume; thus COVID-19 safety measures were still in place. Participants confirmed the day prior to their appointment as well as upon their arrival at the UCL Ear Institute that they did not present with any of the NHS-listed potential COVID-19 symptoms. I always wore a facemask during testing, and, participants were, upon arrival, offered a mask to wear. They could remove it while running tasks in the soundproof booth or if they were uncomfortable wearing it. I removed the mask if the participant asked for its removal to assist their hearing and kept it if participants confirmed they were unaffected by it.

Upon arrival, participants were led to a quiet room at the Ear Institute's Human Labs area where they were briefly introduced to the study and the tasks involved. From the start, they were encouraged to provide feedback on aspects of communication; I adjusted my behaviour accordingly, to accommodate participants' hearing status (e.g. by changing voice loudness, or pace) or general circumstances (e.g. balance difficulties often co-occurring with hearing loss). To assist communication and ensure comprehension of the tasks, where applicable, participants kept their hearing aids on during instructions (but removed them during tasks).

Participants were provided with instructions for the vCCRMn, in preparation for the practice session they had to pass in order to participate in the main study. These instructions were also complemented by a demonstration (or "demo") of the vCCRMn, described in the next section.

3.6.2. Demonstration task of the vCCRMn

Participants were familiarised with the vCCRMn speech-in-noise task orally and in practice using a demonstration session, which lasted approximately 10 minutes. The demonstration task served 3 purposes: a) as a warm-up to the task and familiarisation of the reporting of the answers, b) as a check that the instructions and task were clear, and c) to give the opportunity for further questions before the practice session.

The interactive demonstration session was conducted in the same quiet room where instructions were given. It was comprised of an oral description of the vCCRMn and the participants' task. The conditions, stimuli, and procedure of the task were as described in sections 2.9 and 2.10 (and briefly restated in section 3.5). Notably, it was emphasised to participants that directing their gaze to the talker's mouth on the screen, during trial presentation, and irrespective of whether the talker was target- or masker-coherent, or a still picture, constituted an inclusion criterion throughout the entire time running the vCCRMn task.

Examples of all audio-visual conditions and two example SNR staircase changes were presented during the demonstration task, including a total of 6 example trials. The first trial presented the AV target-coherent condition, the second the AV masker-coherent (described in section 2.9.3.1), the third the Inter target-coherent and the fourth the Inter masker-coherent (described in section 2.9.3.2). The fifth presented the A target condition and the sixth the A masker condition (described in section 2.9.3.3). Trials 1-4 were presented at an easier SNR (20 dB SNR) to allow the participant to 'warm up' faster. The subsequent two trials were presented at gradually reduced SNRs after explaining to the participant that this change constituted an additional feature of the task. The fifth trial was presented at 10 dB SNR and the sixth at 0 dB SNR. Sentences for this presentation were randomly selected from the database of stimuli recorded for the vCCRMn task.

3.6.3. Practice session of the vCCRMn

The practice session of the vCCRMn lasted up to 5 minutes. It consisted of 8 trials, each of which presented a randomly selected vCCRMn sentence at 20 dB SNR, spoken by any of the 4 speakers. The sentence could be either target- or masker-coherent. Participants passed the practice session if they had correctly identified the target words in at least 7 out of the 8 trials correctly. If they scored below 7 on their first attempt, they were given a second chance to pass it. Participants who had passed the practice session were eligible to take part in the study. The final criterion for participation, and the next step, following the practice session was the pure-tone audiometry test.

3.6.4. Pure-tone audiometry test

Following the successful completion of the practice session, participants' hearing status was assessed via pure-tone audiometry (PTA) (air-conduction). As per the inclusion criteria mentioned in section 3.2., to take part in the study, participants had to have normal hearing, if younger (aged ≤ 35 years), and, if older (aged > 35 years), have either normal hearing or mild and/or up to and inclusive of moderate hearing loss. To this end, an audiogram was obtained for each participant following the British Society of Audiology's guidelines (British Society of Audiology, 2018).

Their hearing threshold was measured for each for the frequencies 1000 Hz, 2000 Hz, 4000 Hz, 8000 Hz, 500Hz, and 250 Hz in the order provided here. As a quick metric for the assessment of hearing status, mean PTA score was computed on the spot, of the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz, of the participant's better ear. This value was compared against the audiological ranges detailed in Table 3.1.

Description	Mean PTA score (dB HL)
Normal hearing	≤ 20
Mild hearing loss	21 – 40
Moderate hearing loss	41 – 70
Severe hearing loss	71 – 95
Profound hearing loss	> 95

Table 3.1*: British Society of Audiology audiometric descriptors.*

Following the completion of the PTA, participants who had successfully fulfilled the audiological criterion for participation proceeded with undertaking the vCCRMn speech-in-noise task.

3.6.5. vCCRMn speech-in-noise task

Prior to starting the experiment, participants were provided with a brief refresher of their task for the vCCRMn (as described in sections 2.9, 2.10, and 3.5). Briefly, the task was comprised of 9 runs, each lasting an average of 5 minutes, including the participant's response time. That is, each of the three main conditions (AV, Inter, and A) was run three times. The order with which each condition (and its sub-conditions) was presented was randomised across participants. Participants were asked to take a break after each session and were offered a longer break after they had finished the task.

3.6.6. Montreal Cognitive Assessment (MoCA) task for older participants

Following completion of the speech-in-noise task, participants younger than 50 years old proceeded with the silent lipreading task (described in following section); participants aged 50 and older, first had to complete the Montreal Cognitive Assessment test (MoCA, Nasreddine et al., 2005). This screening test of cognitive impairment included a one-page 30-point test, lasting 15 minutes, with a score below 26 points being indicative of mild cognitive impairment. It assessed memory (delayed recall) (5 points), visuospatial/executive function (5 points), attention, concentration, working memory, mental subtraction (6 points), language (3 points), abstraction (2 points) and orientation to time and place (6 points), and naming of objects (3 points).

Participants with a score lower than 26 were kept in this study to inform individual variability in further analyses.

3.6.7. Test of Adult Speechreading – the silent lipreading task

Both younger (after finishing the vCCRMn task), and older (after finishing the MoCA test), participants had their lipreading ability measured via the silent Test of Adult Speechreading (Campbell et al., 2003).

Broadly speaking, the test consisted of video clips of silent speech, produced by a male or a female speaker. Participants had to select a picture of what they thought was said from an array of options. More specifically, the test was comprised of three core tests of speechreading ability at the following levels: Words, Sentences and Short stories. The test also included three additional subtests which measured speechreading ability at the level of minimal pairs (e.g. chair/bear), sensitivity to sentence stress (i.e. which word is stressed in a sentence) and question or statement differentiation ability. Participants ran all core and subtest tasks included.

A practice session was first run and could be repeated (with the same stimuli and the same order) as many times as the participant needed until they felt comfortable with the task. Participants ran individual practice sessions of, and prior to, each test and sub-test included in the TAS. All practice sessions provided feedback after each response with a green tick mark to indicate a correct response or a red “X” mark to indicate an incorrect response. If the response was incorrect, feedback also pointed to the correct answer.

The Words core test was comprised of 4 practice words and 15 test words. A participant watched either the male or the female speaker say a single word and were then presented with an array of 6 pictures and selected one as their response. This test was scored out of 15.

Similarly, the Sentences core test was comprised of 3 practice sentences and 15 test sentences. The participant watched either the male or the female speaker say a sentence, after which they were presented with 6 pictures to choose their response from. This test was also scored out of 15.

The Short stories test included 1 practice story and 5 test stories. The participant watched either the male or the female speaker tell a short story and were then asked 3 questions about the story. After each question, the participant was presented with 6 pictures to choose their answer from. The score was once more scored out of 15 (3 questions per story, 1 point per correct answer).

Following the presentation of the core tests, the sub-tests were initiated: Firstly, participants ran the minimal pairs sub-test, which included 6 practice words and 60 test words. Participants watched either the male or the female speaker say a word and were given a choice of 2 pictures to select their response. The test was scored out of 60.

Then, participants ran the sentence stress sub-test. This was comprised of 1 practice sentence used five times, and each time stressed on a different word, and 10 test sentences. During the test, the participant was first shown a set of pictures that matched each element of the sentence to be spoken by the speaker of the silent video that followed. This set of pictures also included the words in text form, under the pictures representing the main words in the sentence spoken. Then, participants were presented with either the male, or the female, speaker saying the sentence they had just seen in picture and text form. The participant then had to make a judgement as to where the stress was in the sentence. To this end, they were once more presented with the aforementioned set of pictures and were asked to select the picture of the word they thought was emphasized in the sentence. The test was scored out of 10.

Finally, participants ran the question or statement sub-test which included 4 practice sentences and 20 test sentences. During this subtest, the participant watched the male, or the female speaker say a sentence either as a statement or as a question. They did not have to have understood the sentence. Instead, their task was to identify whether the sentence spoken was spoken as a statement or a question. They were given a choice to select between a full stop picture (statement), or a question mark (question). This test was scored out of 20.

For all core tests, and sub-tests of the TAS, a single mark was awarded for each correct answer.

3.6.8. Talker ID task

The Talker ID task was developed to be run as the final step of the entire testing pipeline (Figure 3.1). This test consisted of 12 audio sentences, borrowed from the vCCRMn recording database. To compile the 12 sentences, 3 sentences were randomly selected from the stimuli recorded spoken by each of the 4 speakers the participants were exposed to during the vCCRMn task. A trial consisted of the presentation of one of the audio sentences, and the display of a response panel that included a still frame of each of the four speakers. The participants were instructed to ignore the content of the sentence, and simply focus on the voice of the speaker and then select, from the panel, the speaker whose voice they thought they had just heard.

The task was scored out of 12 and served as a metric of whether participants had learnt to match the face to the voice of each talker in the vCCRMn.

3.6.9. Quality questionnaire

At the end of the experimental process, participants completed an informal questionnaire, that was used as personal feedback for the experimental procedures they undertook. It included the following set of questions:

1. Did you experience any lagging between the speech-in-noise task's video and its audio?
2. Did you find any of the conditions of the speech-in-noise task distracting, or confusing?
3. Did you find any of the speakers of the speech-in-noise task distracting?
4. Were there any problems in the loading of the videos of the lipreading task?
5. Were you, at any point during the experiments, distracted by either potential sounds coming from nearby rooms, or by pop-up windows from the laptop?
6. Did you know any of the talkers of the vCCRMn?

As no problems were reported by the participants, this served as an on-the-spot check that no confounding factors listed in the questions affected performance throughout the testing session.

3.7. Statistical and analytical methods

In the previous sections, I outlined the steps undertaken towards the recruitment of participants and detailed the settings and tests that comprised the experimental procedures. In the current section, I detail the statistical, and analytical methods employed in the data analyses of Chapter 4 and 5. A brief section on the LMEMs is also provided at the end (detailed descriptions of the LMEMs are included in Chapter 6).

3.7.1. Participant audiometric profiling

A general audiometric score, indicative of the extent of their hearing loss, was calculated for each participant as per international audiometric standards (see e.g. Humes, 2019). This was computed as the mean of the better-ear scores, of the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz tested in the PTA. This score was compared against the PTA hearing loss guidelines shown in Table 3.1 (i.e. PTA scores ≤ 20 dB HL constitute normal hearing, and ≥ 21 dB HL constitute hearing loss).

3.7.2. Low and high frequency audiometric thresholds

Participant audiometric scores were divided into two categories: 1) Low frequency PTA scores, which were computed as the mean PTA threshold of both ears for the frequencies 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz, and 2) high frequency PTA scores, computed as the mean PTA threshold of both ears for the frequencies 4000 Hz and 8000 Hz. These two categories constituted average measures of low frequency hearing and high frequency hearing and were used in the analyses of Chapter 5 and 6.

3.7.3. Speech Reception Threshold estimation

As described in section 2.10 of Chapter 2, each of the main conditions (AV, Inter, and A), of the vCCRMn speech-in-noise task were run 3 times by each participant. And, within each run, both target-, and masker-coherent stimuli were presented. Further, each run output two speech reception thresholds (namely, SRT_{50s}), one for the target-coherent and one for the masker-coherent stimuli. Each of these two SRTs were computed as the mean of the SNRs at the last four reversals of their respective coherence-type track. Thus, for each participant, the speech-in-noise task would output a total of 3 SRT_{50s} for each of AV-target, AV-masker, Inter-target, Inter-masker, A-target, and A-masker conditions. The mean of the three values, for each of the conditions, was computed and used as the participant's SRT_{50} for the respective condition in all the analyses of Chapters 4 and 5.

Figure 3.3 below displays an example the three runs completed by one participant for the AV target-coherent condition of the vCCRMn. In the figure are highlighted the last four reversals of each run, which were used to compute the three SRT_{50s} for that condition.

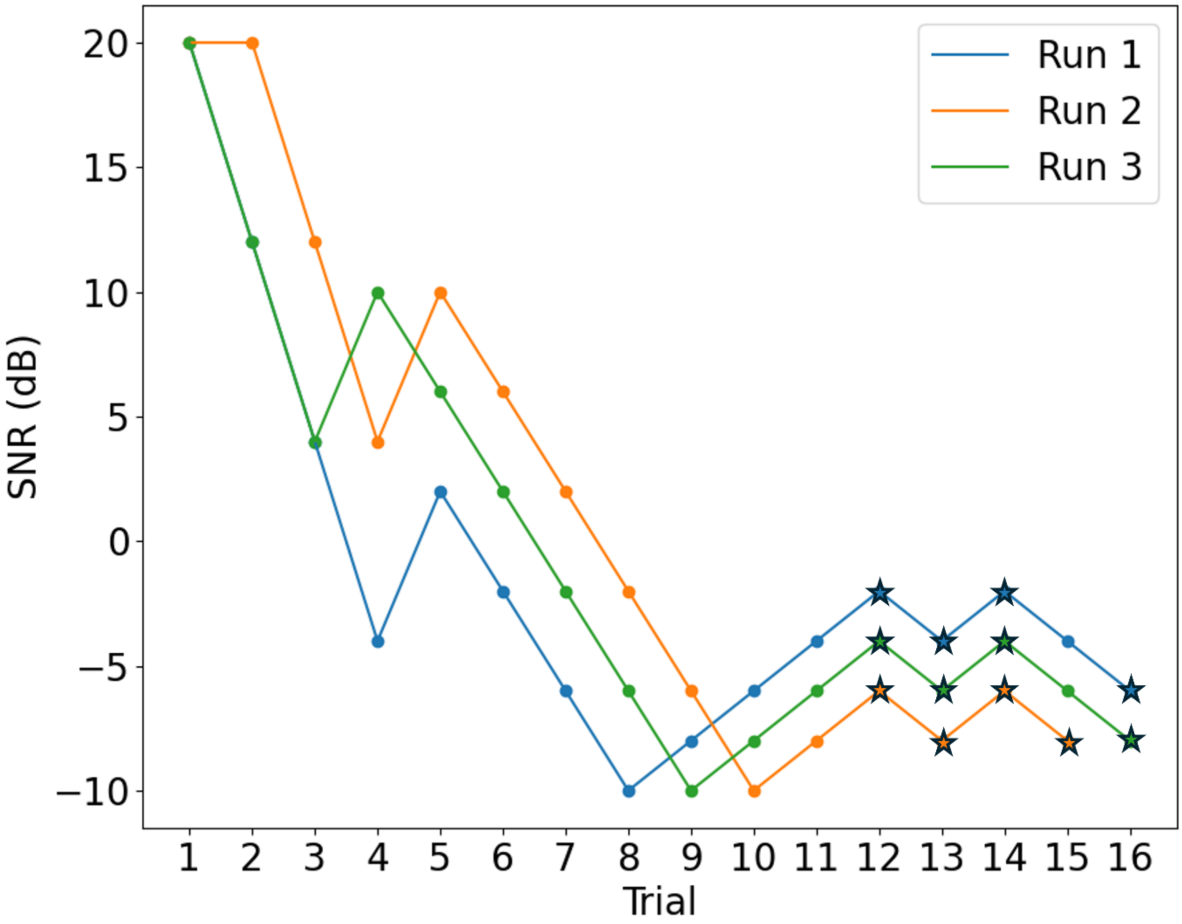


***Figure 3.3**: An illustration of the three experimental runs completed by a participant for the AV target-coherent condition of the vCCRMn. Each run is depicted with a different colour. The last four reversals of each run are highlighted with star symbols. The mean SNR at the last four reversals constituted the SRT$_{50}$ computed for this participant for each run.*

3.7.4. Silent lipreading scores

In section 3.6.7, I described the six levels of lipreading assessed by the TAS core tests and sub-tests. I also mentioned that each was of the tests was scored out of a possible total number of correct responses, with a single point allocated to each correct response provided by the participant. For the analyses conducted herein, these scores were converted to percentage correct performances and were computed independently for each of the six TAS tests.

3.7.5. TAS test score selection

To decide which of the six lipreading scores output by the TAS was to be used as a representative metric of the participants' lipreading ability in the analyses of Chapters 4, 5, and 6, the following procedures were taken: First, to get a general understanding of how participant performances faired across the different conditions of TAS, a bar-plot of mean performances for each subtest was plotted (Figure 3.4). Based on this figure, it was concluded that the TAS subtests minimal pairs, sentences stress and question or statement were relatively easy to solve across all participants (Mean scores:77.55%, 71.6%, and 64.68%, respectively; SDs: 8.43%, 20.14%, and 10.37%, respectively). Further, these sub-tests assessed specific aspects of lipreading ability, and not general ability. Thus, it was decided that these metrics were not suitable for use in downstream analyses and focus was shifted to the core TAS conditions.

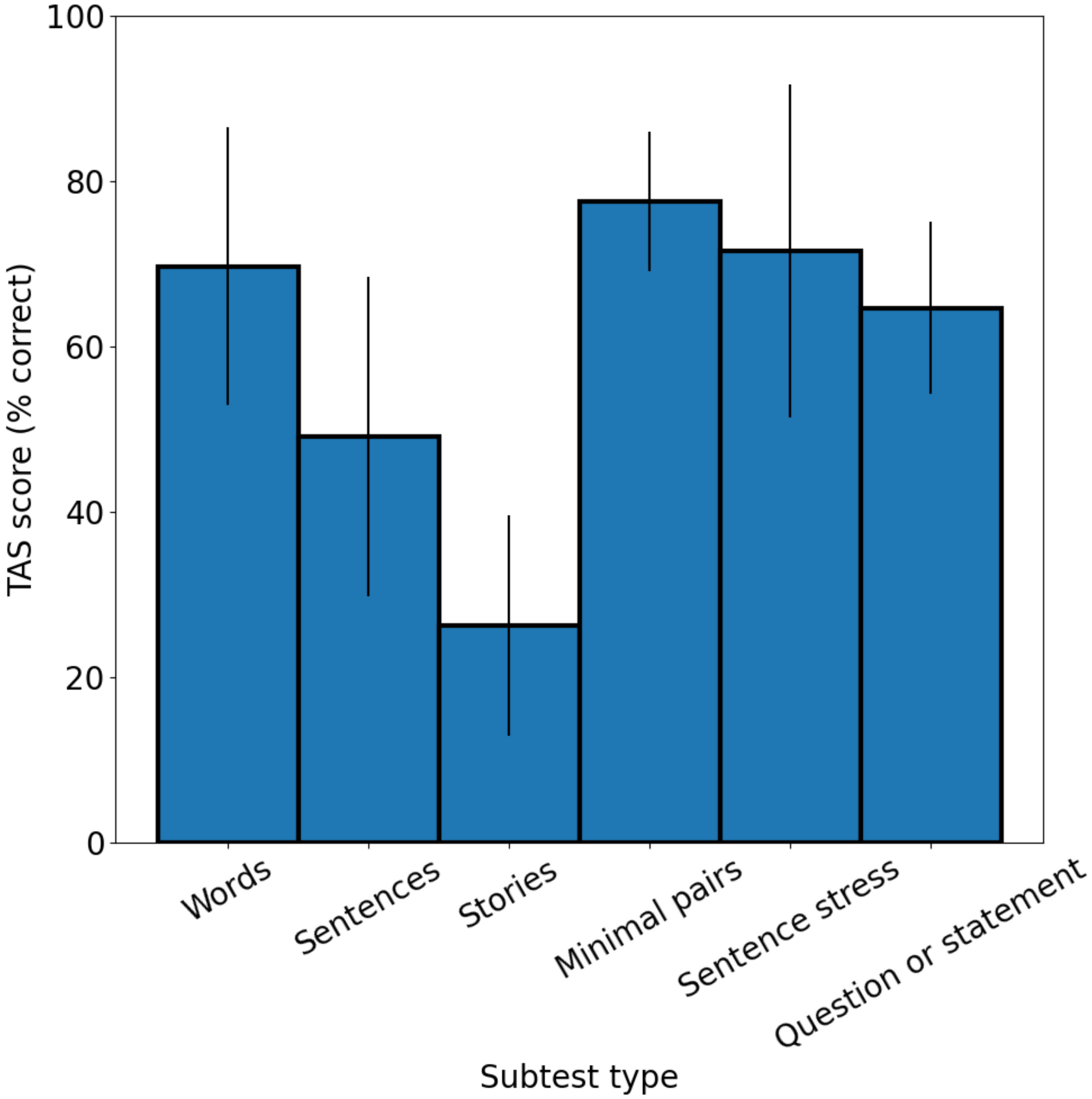


Figure 3.4*: Bar-plots of the mean performances of all 125 participants for each of the six subtests of TAS. Error bars are standard deviations.*

Performance at the level of words also yielded relatively high scores (Mean = 69.71%, SD = 16.79%), compared to performances at the levels of sentences and stories (Mean: 49.12%, and 26.19% respectively; SD: 19.26%, and 13.34% respectively). To further assess these differences, a one-way repeated measures ANOVA was performed, across the three core tests, and yielded significant results (F(2, 369) = 72.39, p < 0.001). Post-hoc paired t-test comparisons revealed that participant performance was significantly better in the words test, compared to the sentences and stories (words vs sentences: t(124) = 14.55, p < 0.001; words vs stories: t(124) = 29.34, p < 0.001). Further, participants performed better in the sentences test compared to the stories test (t(124) = 13.48, p < 0.001). These results are also shown in Figure 3.5.

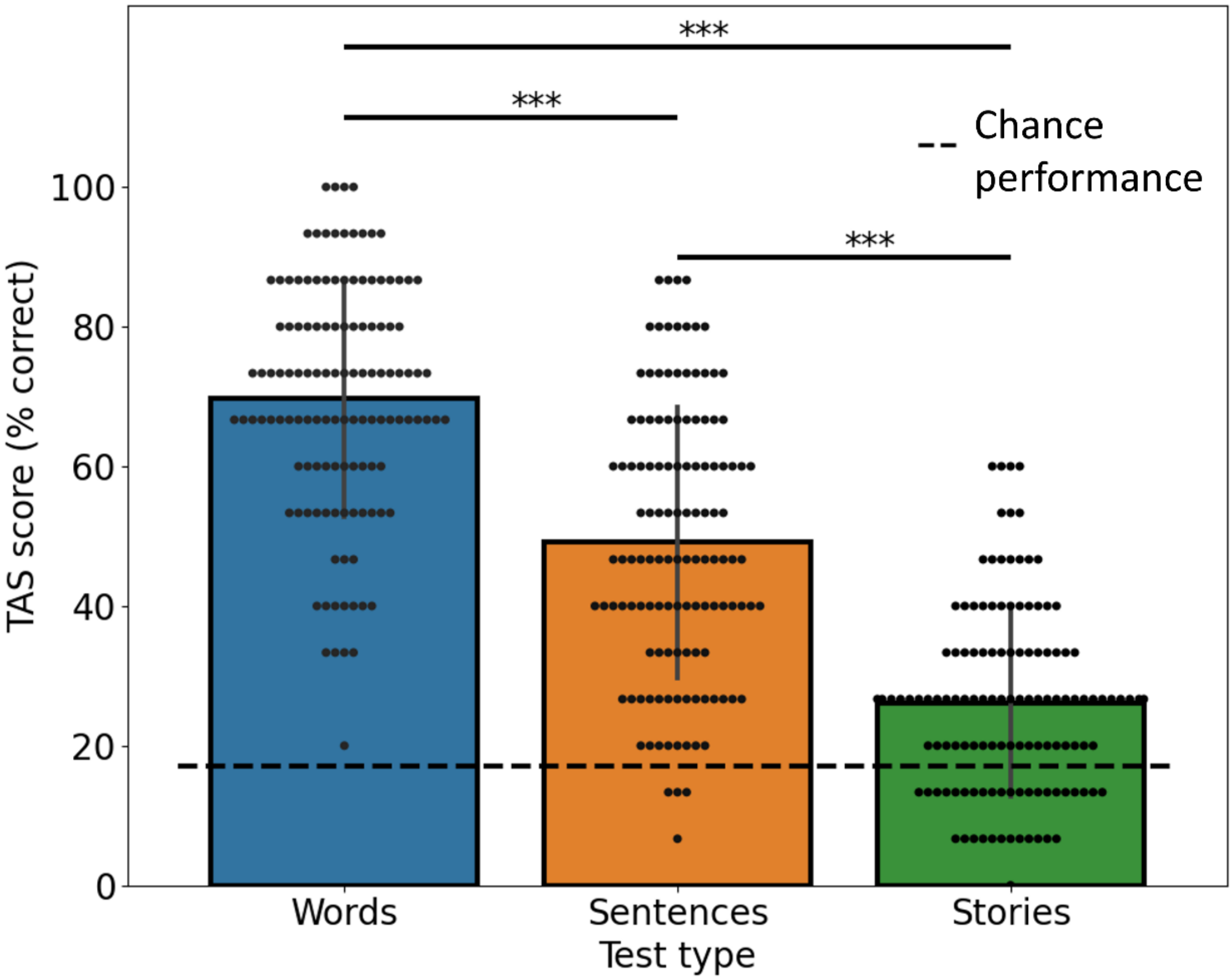


Figure 3.5: *Mean participant performances (n = 125) across the three core tests of TAS. Pairwise comparisons are indicated with horizontal lines over the relevant conditions (*** denotes significance with p < 0.001; errors are standard deviations). The dotted line at ~17% indicates chance performance.*

Notably, it can be seen from the figure that for a substantial number of participants, performance in the stories condition fell below the line denoting chance performance (~17%, estimated based on the 6 option panels of the tests). Thereby participants were often selecting at their responses at random for the stories task – the task was too difficult. On the other hand, participants often reached ceiling, or near-ceiling performance in the words test, indicating that the task was too easy. Given these results, it was decided that the sentences scores were the most suitable among the TAS assessment levels for use as representative measures of the participants' lipreading abilities.

3.7.6. Talker ID scores

Percentage performances were computed for each participant from their total number of correct identifications, out of the possible 12.

3.7.7. MoCA scores

Participants were assigned a score using the modified scoring procedure 'MoCA-H1' from Al-Yawer et al. (2019). Briefly, as described in the cited paper, this scoring procedure controls for the potential negative effects of hearing loss by removing all MoCA subtests that depend on hearing to be completed. MoCA-H1 scores range from 0 to 20, with scores below 16 being indicative of mild-cognitive impairment.

3.7.8. Statistical methods

All statistical analyses were conducted in R.

3.7.8.1. Omnibus comparisons and post-hoc tests

ANOVAs were used on several occasions in the analyses of Chapters 4 and 5, to compare participant performances across different conditions. These omnibus tests were often followed by post-hoc pairwise t-tests. The Bonferroni method was used to adjust the p-values obtained from these post-hoc comparisons. Furthermore, Cohen's d, and mean differences, were included alongside each of the pairwise comparisons as measures of its effect size.

3.7.8.2. Other statistical comparisons

Where data were normally distributed, t-tests were used in statistical comparisons performed between groups of variables outside the omnibus cased. Where data were not normally distributed, Mann-Whitney U tests were used. Normality was assessed via Shapiro-Wilk tests.

3.7.8.3. Correlation analyses

Correlational analyses were employed on several instances throughout Chapters 4, 5, and 6. Pearson correlations were used, where data were normally distributed, and Spearman otherwise. Normality was assessed with Shapiro-Wilk tests.

3.7.8.4. Power analyses

Audio-visual temporal coherence effects have been previously reported to be subtle (see e.g. Maddox et al. 2015; Cappelloni et al. 2023). In the current work, audio-visual temporal coherence effects are reflected in participant performance differences between the Interrupted ('Inter') and Audio with static image ('A') conditions of the vCCRMn. A power analysis was conducted to determine the sample size required to detect the relatively subtle differences between these two conditions. This used an effect size of 1 dB SNR, a standard deviation of 2 dB SNR, an α (type I error probability) of 0.05, and a power of 0.80. The minimum sample size of each experimental group was thus estimated to be 33.

3.7.8.5. Linear mixed effects model: Brief methods section

Two LMEM models were developed in this thesis: One for participant speech-in-noise performance in the vCCRMn task, and one for their measured AV benefits. The general procedure followed included a model with random effects, two-way interactions, and main effects, and reducing it term at a time, using likelihood ratio tests to inform each reduction step.

The modelling analyses were conducted using the R package lmerTest.

Chapter 4: Audio-visual benefit: Insights from three experiments

4.1. Introduction

In Chapter 2 I discussed the development of the vCCRMn speech-in-noise task, and in Chapter 3 I provided a detailed analysis of participant selection, experimental procedures, and the battery of tests I made use of in my research. As explained in Chapter 3, participant testing was divided into three experiments: The younger participant with normal hearing, the older participant with normal hearing, and the older participant with hearing loss experiments.

In the current chapter I delve into the data collected from these three experiments to begin to address the first three of the primary aims of my thesis. These were, namely: 1) the development of a speech-in-noise task that proved capable of capturing audio-visual (AV) benefit, 2) the provision of evidence for the contributions of both lipreading and audio-visual temporal coherence to the AV benefit, including a quantification of their relative contribution, and 3) the assessment of age, and hearing loss, as factors potentially influencing the AV benefit.

To this end, with the analyses that follow I sought to answer the following questions:

- Was the vCCRMn task able to capture an AV benefit?
- Was there a visual benefit in the interrupted condition compared to the static image condition?
- Was there a difference in performance between the target-coherent, and masker-coherent conditions of the task?
- Are the visual benefits associated with lipreading ability?
- Are there differences in the visual benefits measured, across the three experimental groups?

For quick reference I provide here, before proceeding with the analyses sections, a brief reminder of the vCCRMn task and its conditions: In the vCCRMn task, participants were tasked with identifying a target sentence in the presence of background noise – which was presented in the form of two maskers uttering sentences of similar structure to the target sentence. It included presentations under several different conditions, accompanied by different visual stimulus types of a talker:

a) A full naturalistic video, also referred to as “Audio-visual” or “AV” condition that matched the audio of either the target talker (target-coherent), or one of the two maskers (masker-coherent).

b) A video referred to as “Interrupted” condition or “Inter”. This also matched either the target or masker audio.

c) A static image condition, referred to as “Audio-only” or “A”, that too either matched the target talker or one of the maskers.

I begin with the presentation of the younger participant with normal hearing experimental group results, followed by the results of the older participants with normal hearing, and then the older participants with hearing loss.

4.2. Experiment 1: Results from the younger participants with normal hearing

A total of 39 younger participants with normal hearing (Ages 19-35, Mean = 25.92, SD = 4.33) constituted this dataset.

4.2.1. Participant audiograms

Hearing thresholds were measured for each participant and their audiograms are presented in Figure 4.1. All participants had normal hearing based on the hearing status classification criterion described in Chapter 3.

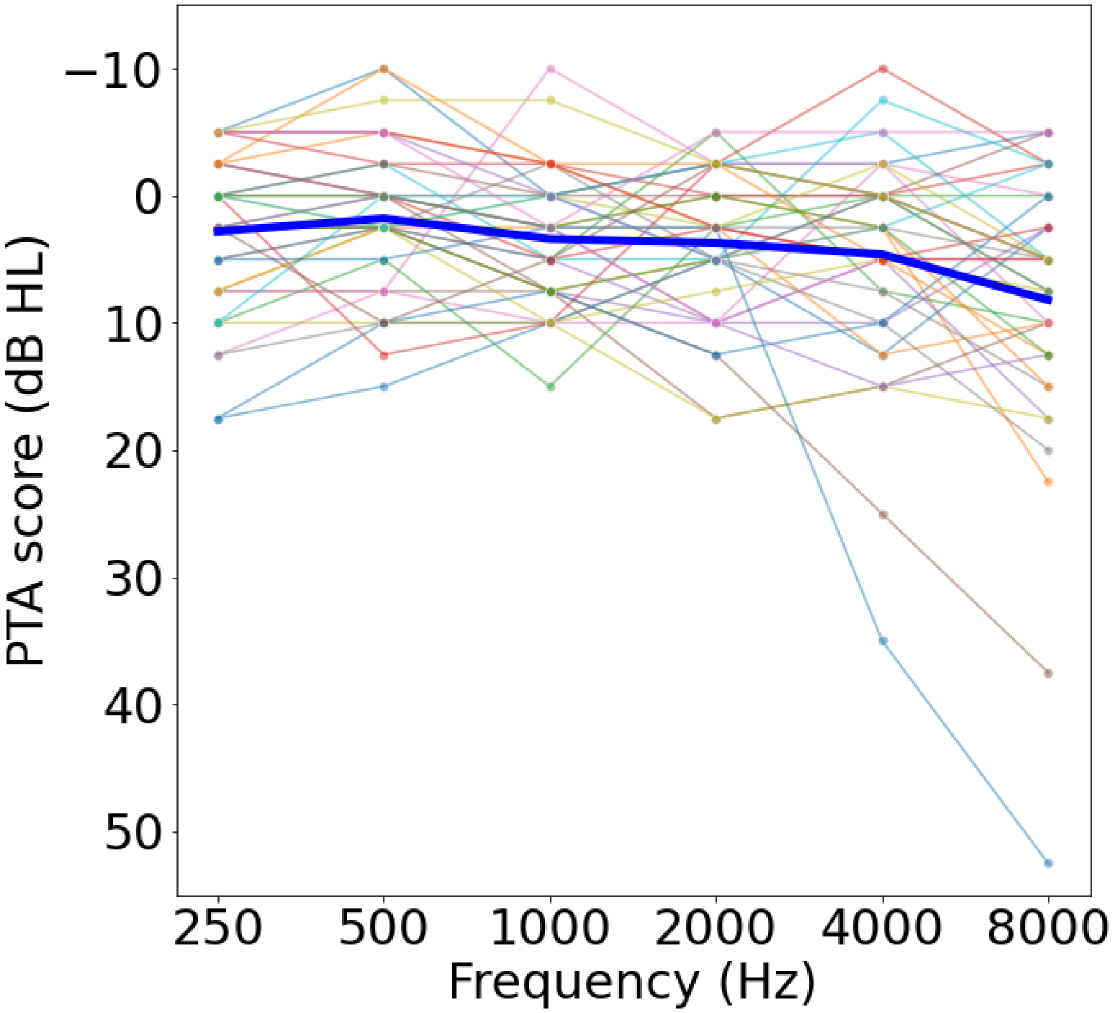


Figure 4.1*: Mean pure tone threshold across both ears as a function of frequency in younger participants with normal hearing. Each line represents a participant. The thick central line depicts the mean performance across all participants included in this group.*

4.2.2. Speech-in-noise performance varies across the conditions of the task

The analysis in this section investigates the impact of seeing the talker's face in perceiving speech in noise. To determine whether visual inputs (encapsulated by the vCCRMn conditions AV, Inter, and A, and target coherent or masker coherent) influenced participants' speech reception thresholds, these thresholds were plotted against the conditions of the vCCRMn task (Figure 4.2).

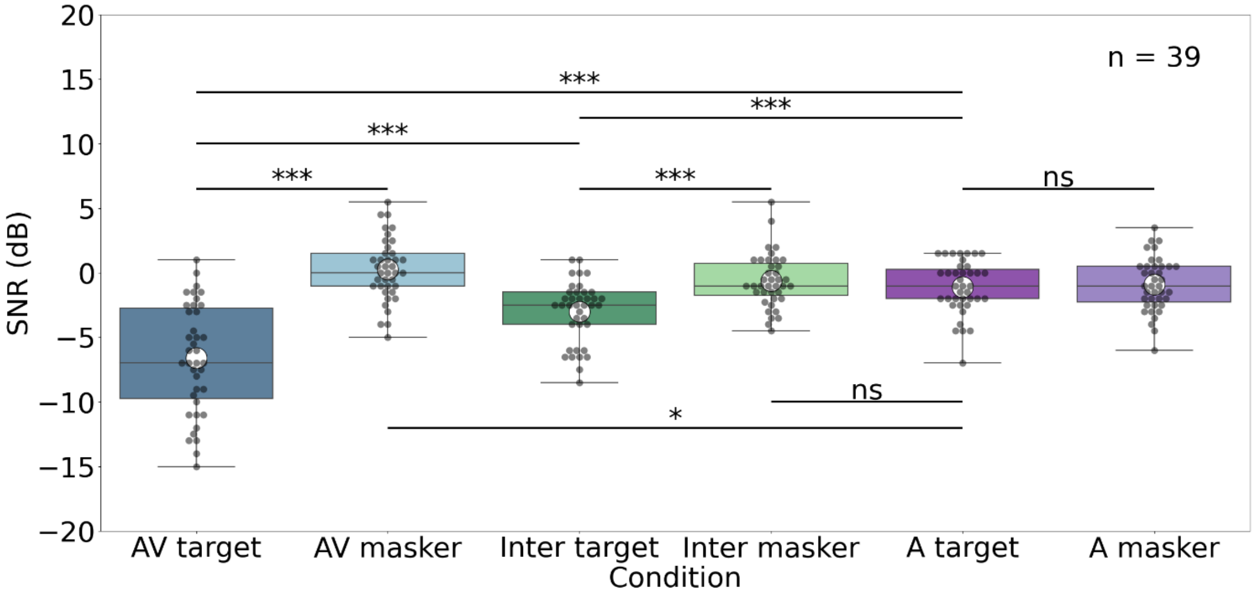


Figure 4.2*: SNR scores for each of the six conditions (target-coherent and masker-coherent for AV, Inter and A) of the vCCRMn speech-in-noise task for the younger participant* with *normal hearing experimental group. White dots depict mean performances. Pairwise comparisons are indicated with horizontal lines over the relevant conditions (* and *** denote significance with p < 0.05 and < 0.001 respectively; ns = non-significant).*

It is evident from the differences in the distributions displayed in Figure 4.2 that the different visual conditions influenced participants' speech reception thresholds. To confirm this observation, a three by two repeated measures ANOVA was conducted to examine the effects of visual condition (AV, Inter, and A), visual coherence type (Target, Masker), and the interaction between the two, on participants' thresholds. The test yielded significant results for both the main effects of visual condition (F(2, 222) = 6.459, p < 0. 01) and visual coherence type (F(1, 222) = 27.414, p < 0.001), as well as for the interaction between the two (F(2, 222) = 15.244, p < 0.001). Thus, participant performance was influenced by both the visual condition and the visual coherence type. Further, the interaction term significance suggested that the effect of visual condition on performance depended on visual coherence type (and vice versa). Post-hoc pairwise paired t-tests were used to explore these results in more detail, these are shown in Table 4.1. Specifically, Table 4.1 shows the comparisons between the visual conditions, taking into account the significance of the visual condition main effect, and the comparisons between all possible sub-conditions, taking into account the significance of the interaction term – the latter are also displayed in Figure 4.2. The table does not include comparisons between the visual coherence types as these were only two and already deemed significant based on the ANOVA test output. Table

4.1 also includes headers indicating the relevance of the comparisons to the chapter's research questions (4.1).

Comparison type	**Condition 1**	**Condition 2**	**t-statistic**	**DoF**	**Bonferroni corrected p-value**	**Cohen's d**	**Mean difference (dB SNR)** **(Condition 1 – Condition 2)**
Visual condition comparisons	AV	Inter	-2.97	38	0.012 *	-0.34	-1.36
	AV	A	-4.02	38	< 0.001 ***	-0.45	-2.18
	Inter	A	-2.83	38	0.018 *	-0.32	-0.82
Visual benefits	AV target	A target	-8.11	38	< 0.001 ***	-1.30	-5.51
	Inter target	A target	-4.72	38	< 0.001 ***	-0.76	-1.91
	AV target	Inter target	-5.27	38	< 0.001 ***	-0.84	-3.60
Visual detriments	AV masker	A target	3.24	38	0.037 *	0.52	1.36
	Inter masker	A target	1.28	38	1	0.21	0.48
	AV masker	Inter masker	2.56	38	0.22	0.41	0.88
Visual coherence	AV target	AV masker	-9.35	38	< 0.001 ***	-1.50	-6.87
	Inter target	Inter masker	-5.90	38	< 0.001 ***	-0.95	-2.39
	A target	A masker	-0.60	38	1	-0.10	-0.20
Additional	AV target	Inter masker	-8.96	38	< 0.001 ***	-1.44	-5.99
	AV target	A masker	-9.16	38	< 0.001 ***	-1.47	-5.72
	AV masker	Inter target	6.71	38	< 0.001 ***	1.07	3.27
	AV masker	A masker	3.02	38	0.07	0.48	1.15
	Inter target	A masker	-5.39	38	< 0.001 ***	-0.86	-2.11
	Inter masker	A masker	0.83	38	1	0.13	0.28

Table 4.1*: Summary of the post-hoc pairwise paired t-test comparisons conducted for the young*er *participant with normal hearing experimental group. The comparisons' header column indicates the relevance of these comparisons to the research questions of the chapter. Cohen's d and the mean differences between the conditions being compared are also included as measures of effect size. Star symbols in the p-value column denote statistical significance (* and *** indicate significance with p < 0.05 and < 0.001 respectively).*

These comparisons formed the basis for further exploration of these questions. In the next section I discuss the conditions involved under the "visual benefits" and "visual detriments" types of comparisons (Table 4.1). Then, in the subsequent section I discuss the conditions involved in the "visual coherence" set. All remaining possible pairwise comparisons are listed in Table 4.1, under the "additional" comparisons set.

4.2.3. Video conditions enhance participant speech-in-noise performance

From the three target-coherent conditions, on average, participants performed the best in the AV target condition (Mean = -6.62 dB SNR, SD = 4.20 dB SNR), followed by Inter target (Mean = -3.01 dB SNR, SD = 2.34 dB SNR), and A target (Mean = -1.10 dB SNR, SD = 2.02 dB SNR). To gauge whether participants indeed benefited from having a target-coherent visual stimulus, over just the static image, I looked at the results of the comparisons between the AV target and A target, as well as between the Inter target and A target conditions, from Table 4.1. Performance in both AV target and Inter target was significantly better than that in the A target condition (AV target vs A target: t(38), = -8.11, p < 0.001; Inter target vs A target: t(38) = -4.72, p < 0.001).

The AV benefit of participants was computed as the difference in performances between the A target and AV target conditions: AV benefit = A target – AV target. Since lower SNR scores indicated greater performance, computed as described, the AV benefit was positive when participants performed better at the AV target condition, compared to the A target. On average, younger participants with normal hearing gained 5.51 dB SNR of AV benefit. This finding confirmed that the vCCRMn task was indeed capable of capturing an AV benefit. The Inter benefit (= A target – Inter target), that is, the average benefit conferred by the Inter target condition relative to the A target, was 1.91 dB SNR.

Notably, in addition to providing an AV benefit, the visual condition was also able to impair listening in noise: In the comparison between AV masker and A target (t(38) = 3.24, p = 0.037) we see that participants performed better in the static image condition than the visual condition, by 1.36 dB SNR. The visual detriment however, exhibited by the Inter masker condition compared to the A target condition, was not statistically significant (t(38) = 1.28, p = 1).

As an alternative, and complementary way of visualising the data, they were also plotted as difference histograms (Figure 4.3). Namely, these histograms displayed the distributions of the variables A-target – AV-target and A-target – Inter-target (the two visual benefits). In both 4.3 (A), depicting the AV-benefit, and (B) depicting the Inter-benefit, most values are to the right of zero, suggesting that for most participants, the video conditions boosted auditory performance. Notably, the distribution of AV benefit has a considerably longer right tail than that of Inter benefit. This suggests that certain participants were able to gain disproportionately larger visual benefits from the AV target condition, compared to the Inter target. Further, the variability of the AV benefit (SD = 4.19 dB SNR) was larger than that of the Inter benefit (SD = 2.49 dB SNR).

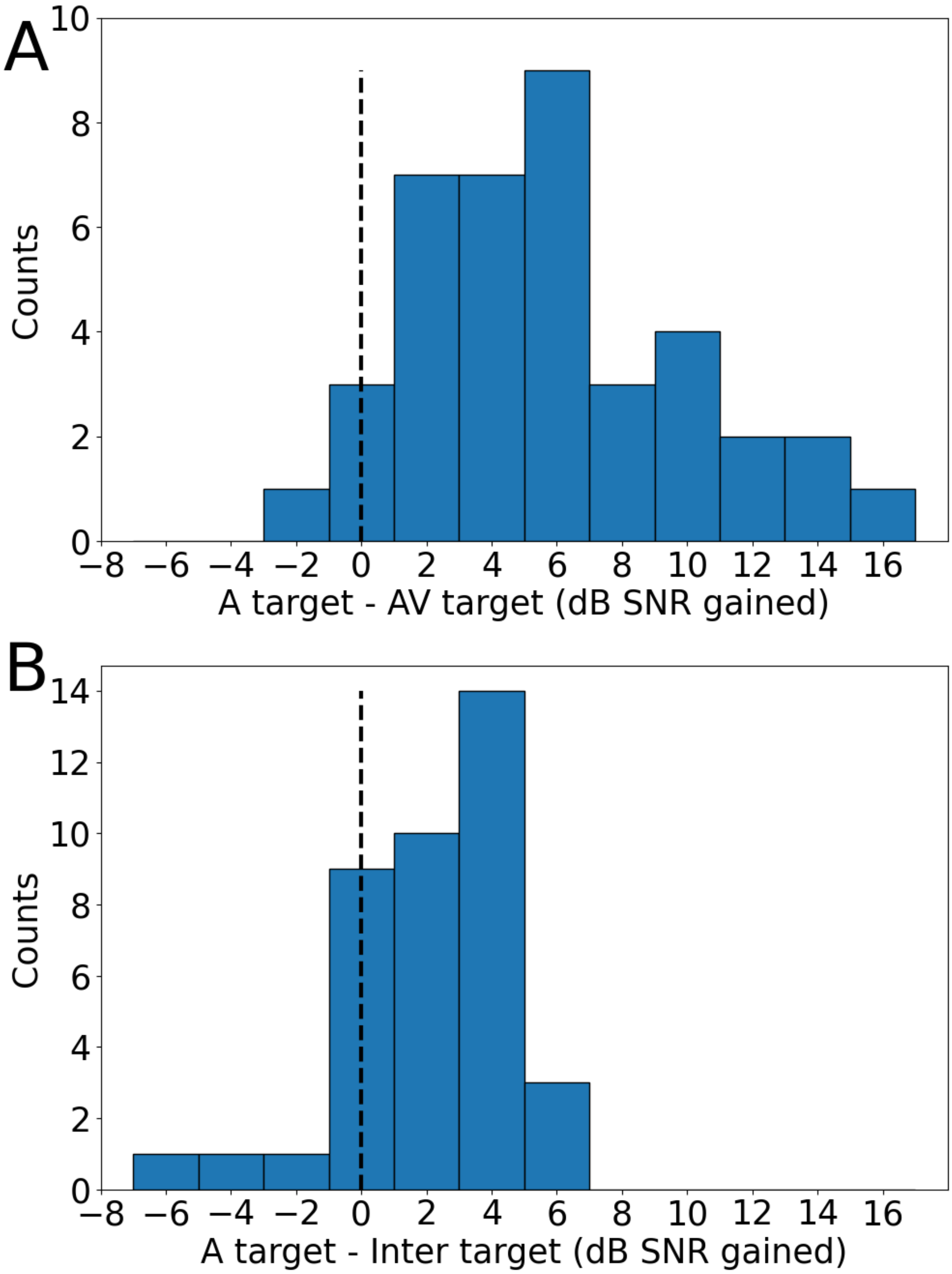


Figure 4.3*: Histograms of visual benefits for the younger participant with normal hearing experimental group. (A) Distribution of the performance differences (the AV benefits) between the full naturalistic (AV target) video and the static image (A target) conditions. (B) Distribution of the performance differences (the Inter benefits) between the interrupted (Inter target) and the static image (A target) conditions. All values at 0 indicate no difference in the performance in the video conditions compared to the static image. Values to the right of 0 demonstrate better performance in each video condition compared to the static image.*

Finally, as far as the pairwise comparisons go for the current section, I asked whether the full target-coherent audio-visual stimulus offered an advantage over the interrupted one. I also considered whether the full masker-coherent audio-visual stimulus presented an added AV detriment to the performances observed in the interrupted masker-coherent condition. To this end, I looked into the comparison between the AV target and Inter target conditions, and that between the AV masker and Inter masker. I found that the former pair were significantly different (t(38) = -5.27, p < 0.001), and the latter were not (t(38) = 2.56, p = 0.22).

Namely, participants exhibited greater performance, by 3.6 dB SNR, in the AV target condition compared to the Inter target, potentially reflecting the added benefits attributable to lipreading cues in the former. Given this result, and since the Inter target condition conferred a 1.91 dB SNR boost to performance compared to the A target condition, there seems to have been a stepwise increase in performance in moving from the audio-only, to the interrupted video, and finally to the full video condition.

Given the nature of the two conditions, it was expected that the visual benefit conferred by the Inter target condition constituted part of the visual benefit provided by the AV target condition (the temporal coherence part). Further, both conditions used the same auditory stimuli. Therefore, it was expected that the performances of participants on the two conditions would be correlated with each other. To test this hypothesis, a Pearson correlation was conducted between AV target and Inter target (Figure 4.4). Nonetheless, the correlation did not yield significant results (Pearson’s r = 0.27, p = 0.095), possibly due to lack of statistical power.

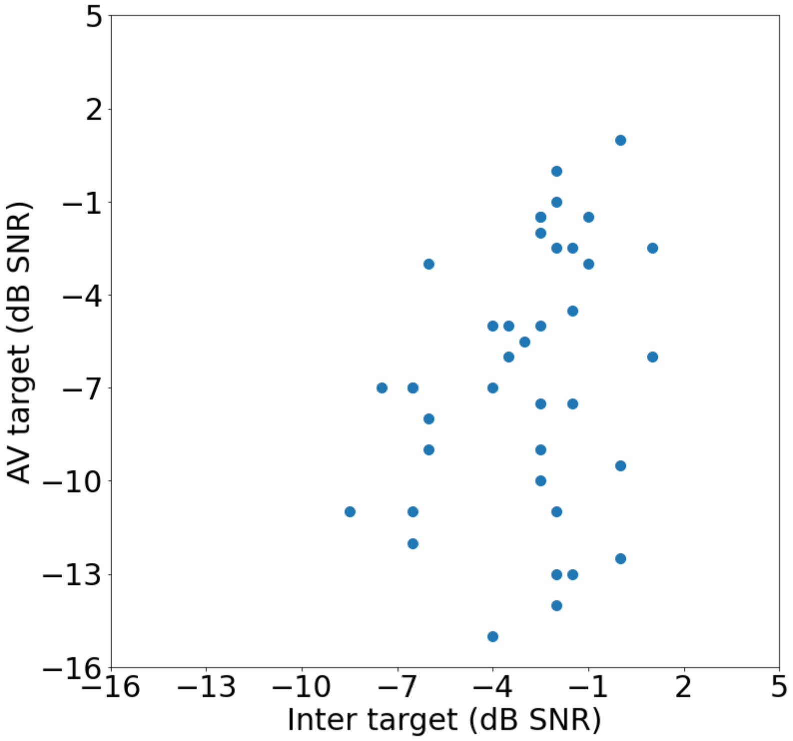


Figure 4.4*: Illustration of the relationship between the AV and Inter conditions for the younger participant with normal hearing experimental group. A Pearson correlation was conducted to assess the relationship and was not significant.*

4.2.4. Visual coherence influences participant performance

Next, to see if there was a performance difference when the visual stimulus matched the target, compared to when it matched the masker talkers of the vCCRMn, I looked at the AV target vs AV masker, Inter target vs Inter masker, and A target vs A masker comparisons from Table 4.1. These revealed better performance across the AV target and Inter target video conditions, compared to their masker counterparts, but not in the A target (static image) condition (AV target vs AV masker: t(38), = -9.35, p < 0.001; Inter target vs Inter masker: t(38), = -5.90, p < 0.001; A target vs A masker: t(38), = -0.10, p = 1). Also, for reference, the conditions' respective means, standard deviations, and mean differences were the following:

- AV target vs AV masker:
    - AV target: Mean = -6.62 dB SNR, SD = 4.20 dB SNR
    - AV masker: Mean = 0.26 dB SNR, SD = 2.36 dB SNR
    - AV masker – AV target = 6.87 dB SNR, SD = 4.53 dB SNR
- Inter target vs Inter masker:
    - Inter target: Mean = -3.01 dB SNR, SD = 2.34 dB SNR
    - Inter masker: Mean = -0.62 dB SNR, SD = 2.04 dB SNR
    - Inter masker – Inter target = 2.39 dB SNR, SD = 2.50 dB SNR
- A target vs A masker:
    - A target: Mean = -1.10 dB SNR, SD = 2.02 dB SNR
    - A masker: Mean = -0.90 dB SNR, SD = 2.05 dB SNR
    - A masker – A target = 0.20 dB SNR, SD = 2.12 dB SNR

As done before, for the visual benefits comparisons, I also visualised the current datasets with histograms depicting the distributions of the differences between the three pairs of masker and target conditions (Figure 4.5). Namely, Figure 4.5 (A) shows the distribution of the variable AV masker – AV target, (B) shows the distribution of Inter masker – Inter target, and (C) of A masker – A target. Participants at zero (dashed line in figure 4.5) performed equally well in both conditions. Notably, all values of (A) (AV conditions) and most of (B) (Inter conditions) are to right of zero, indicating better performance in the target-coherent rather than the masker-coherent video conditions. This was not the case for the A (static image) condition where the performance differences are roughly symmetrically distributed about zero, confirming that performances were similar across the two conditions. Further, and as was the case with the visual benefits of the previous section, the variability in the AV conditions' differences was greater than that observed in the Inter and A conditions' differences (AV masker – AV target: SD = 4.53 dB SNR; Inter masker – Inter target: SD = 2.50 dB SNR; A masker – A target: SD = 2.12 dB SNR).

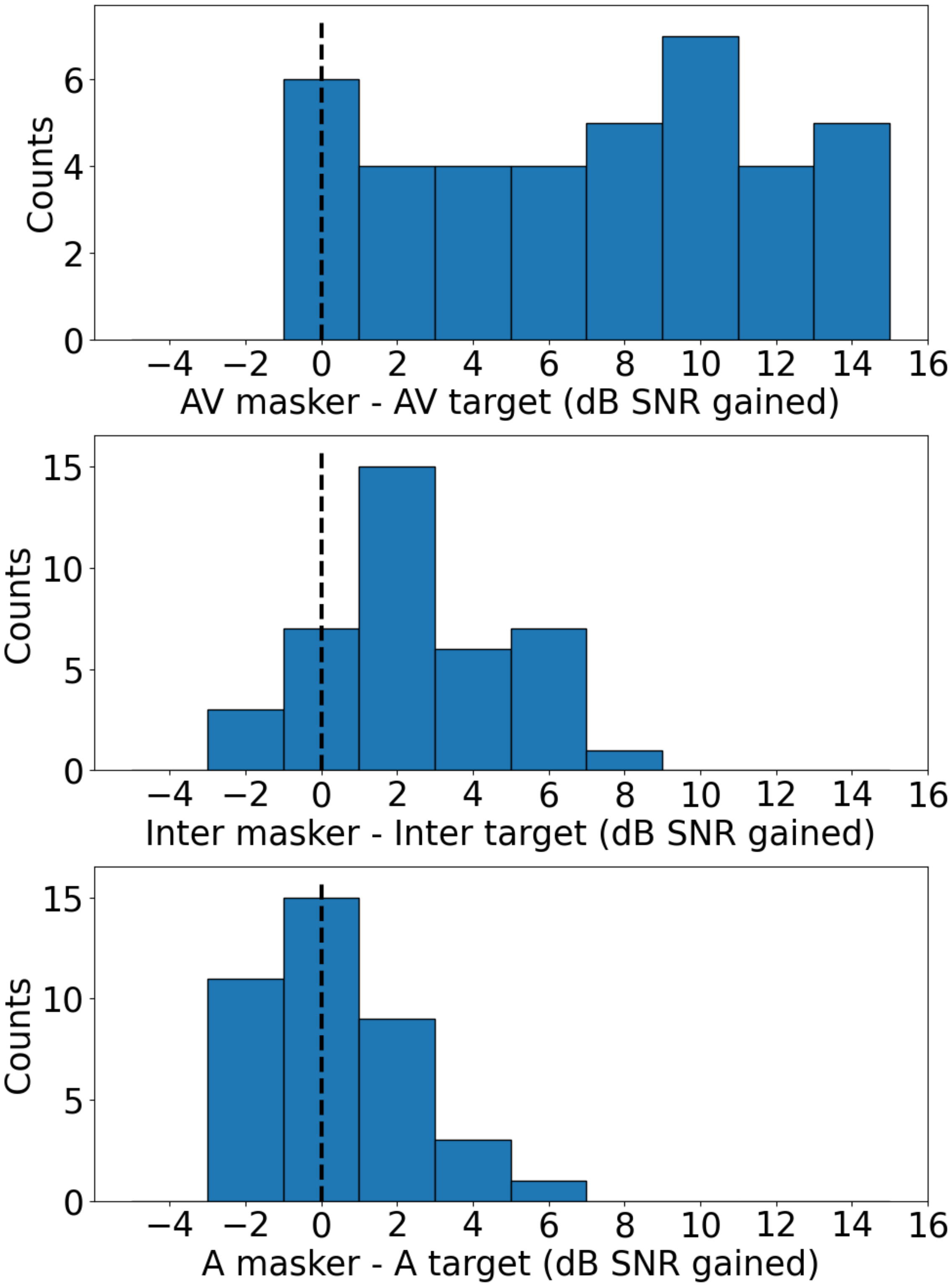


***Figure 4.5**: Distribution of the performance differences between the masker- and target-coherent conditions of the vCCRMn for the younger participant with normal hearing experimental group. (A) Distribution of the AV masker – AV target. (B) Distribution of Inter masker – Inter target. (C) Distribution of A masker – A target. All values at 0 indicate no difference in performance between the masker-coherent and target-coherent conditions. Values to the right of 0 demonstrate better performance in the target-coherent condition.*

4.2.5. AV benefit, but not Inter benefit correlates with lipreading ability

In section 4.2.3, it was shown that both the AV target, and the Inter target conditions provided younger participants with normal hearing with visual benefits over the static image condition. These were respectively called AV benefit and Inter benefit. As the vCCRMn was designed with this specific goal in mind, the AV benefit was expected to comprise of visual contributions from both temporal coherence, and lipreading cues, from the talker's mouth. In accordance with this, performance in the AV target condition was better than that in the Inter target condition. The Inter benefit, on the other hand, was designed to capture just temporal coherence contributions. Hence, the AV benefit was expected to be positively associated with lipreading skill, whereas the Inter benefit was not. To test this hypothesis, the "Sentences" sub-test of TAS was used here (for reasons discussed in Chapter 3), to conduct correlations with the two variables.

The TAS sentence scores were found to be significantly correlated with AV-benefit, albeit the association was weak-to-moderate (Pearson's r = 0.37, p = 0.02) (Figure 4.6). Therefore, at first sight, lipreading contributions to the measured AV benefit do not appear to be evidently substantial. This is consistent with the fact that the stimuli of the vCCRMn were made difficult to lipread by design.

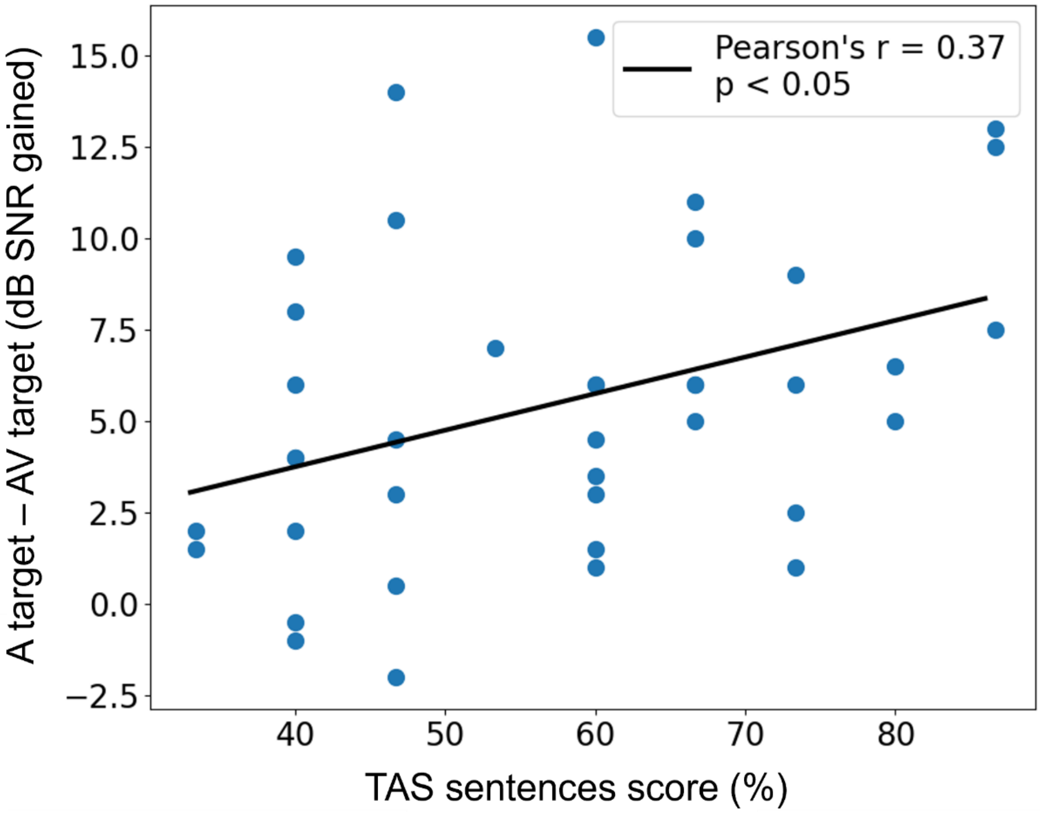


Figure 4.6*: Illustration of the relationship between AV benefit and TAS sentences scores for the younger participant with normal hearing experimental group. Pearson correlation analysis results are indicated in the top right corner of the figure.*

The correlation between Inter benefit and TAS sentences scores was not significant (Pearson's r = 0.20, p = 0.21) (Figure 4.7A). And, while caution should be used in the interpretation of negative results, lack of correlation here is consistent with the Inter benefit measuring a listening advantage attributable to temporal coherence, rather than lipreading. An alternative way of assessing the correlation between Inter benefit and lipreading ability is to correlate the latter with the difference Inter target – AV target. This difference captures the added, and presumably lipreading-related, benefit participants gained from the AV target condition compared to the Inter target condition which lacked lipreading cues during the target word presentations. This correlation is shown in Figure 4.7B and was not significant either (Pearson's r = -0.31, p = 0.16).

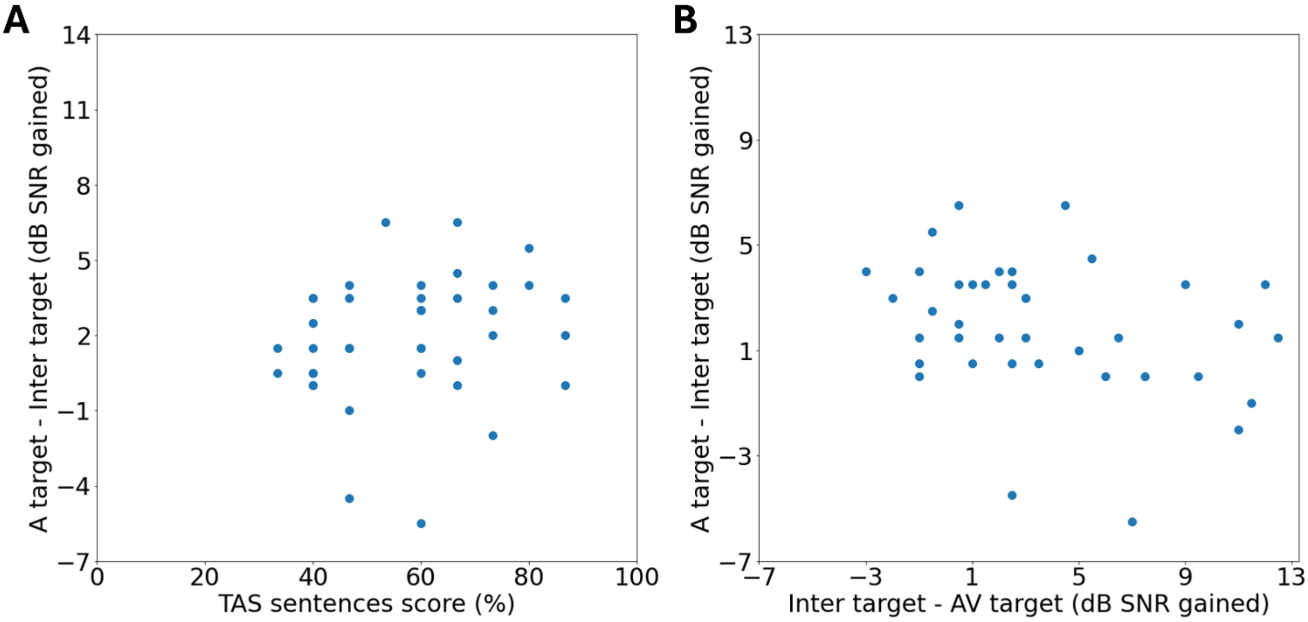


***Figure 4.7**: Illustration of the relationship between Inter benefit and lipreading ability for the younger participant with normal hearing experimental group. (A) Inter benefit plotted against TAS sentences score. A Pearson correlation was conducted to assess the variables' association and was not significant. (B) Inter benefit plotted against the difference between the performances in the Inter target and AV target conditions (i.e. the lipreading-related gains in the latter compared to the former). A Pearson correlation was conducted to assess the variables' association and was not significant.*

4.3. Experiment 2: Results from the older participants with normal hearing

In section 4.2, I discussed the results from the analysis of the younger participant with normal hearing experimental group. There, it was found that participants were gaining visual benefits from both the AV, and the Inter target conditions, compared to the A target. Further, the AV benefits, but not the Inter benefits of participants correlated with their lipreading performances. Finally, younger normal hearing participant performances in the target coherent video conditions were greater than those in their respective masker-coherent video conditions.

In the current section, I continue the discussion with the results from experiment 2, which included data collected from the older participants with normal hearing. A total of 49 older participants with normal hearing (Ages 37-82, Mean = 62.76, SD = 10.30, with 45 out of 49 participants over the age of 50) constituted this dataset.

4.3.1. Participant audiograms

Audiograms of the older participants with normal hearing are shown in Figure 4.8. As with younger participants with normal hearing, the current group were classified normal hearing based on the criterion outlined in Chapter 3. Briefly, the criterion was that participants had to have a mean audiometric threshold of less than or equal to 20 dB HL, for their better ear, over the mid-range of frequencies assessed (i.e. over the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz). Nonetheless, and as evident from Figure 4.8, several of these participants had hearing loss at higher frequencies. Further, it can be seen from Figure 4.8, where the mean of both ears was considered for each frequency assessed (instead of just the better ear), that some participants had hearing loss at lower frequencies as well.

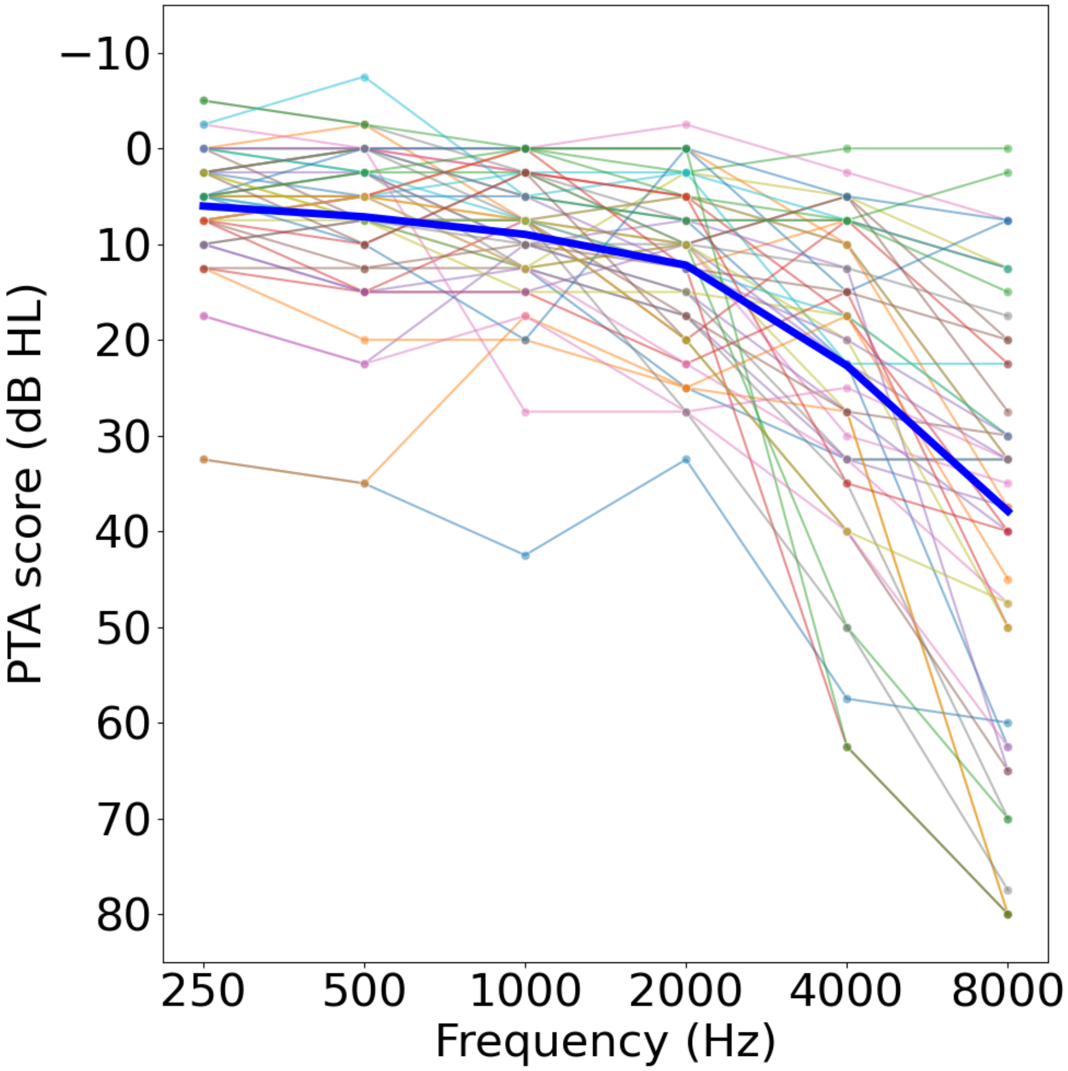


Figure 4.8*: Mean pure tone threshold across both ears as a function of frequency in the older participants with normal hearing. Each line represents a participant. The thick central line depicts the mean performance across all participants included in this group.*

For visual comparison, I also plotted side-by-side boxplots of the better ear mid-frequency range mean PTA thresholds computed for the participants of experiments 1 and 2 (Figure 4.9). It can be seen that, although older participants with normal hearing had higher mean PTA thresholds than those of the younger group, their thresholds were within the normal hearing range.

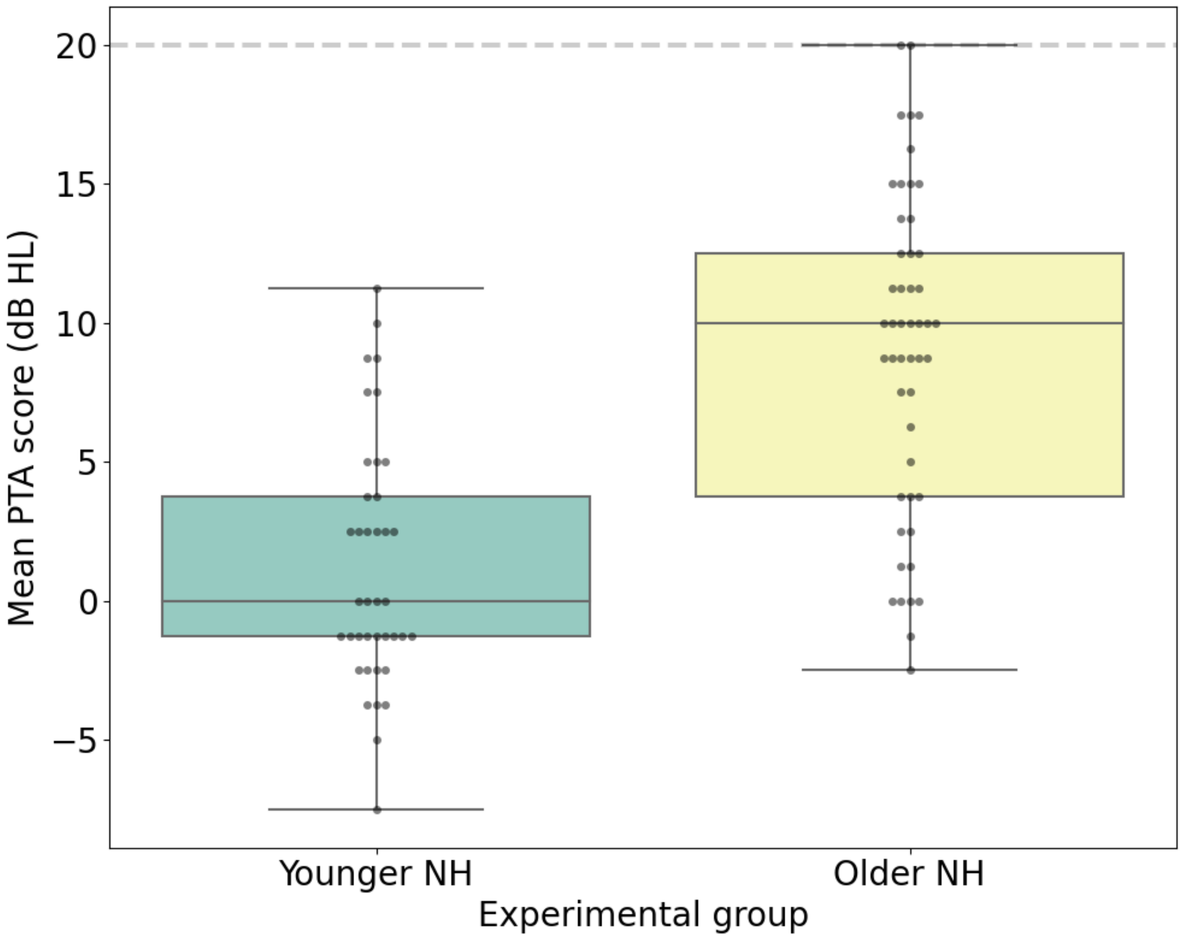


Figure 4.9*: Side-by-side boxplots of mean PTA thresholds for the younger participant with normal hearing and older participant with normal hearing experimental* groups*. Mean PTA thresholds here were computed for the participants' better ear, over the range of frequencies including 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. The horizontal dotted line indicates the threshold (20 dB HL) above which participants are classified to have hearing loss.*

4.3.2. Speech-in-noise performance varies across the conditions of the task

Previously, we saw that younger participant with normal hearing performances varied across the conditions of the vCCRMn. To determine the influence of these visual inputs on the older participant with normal hearing performances, speech reception thresholds calculated for this experimental group were once more plotted against the conditions of the task (Figure 4.10).

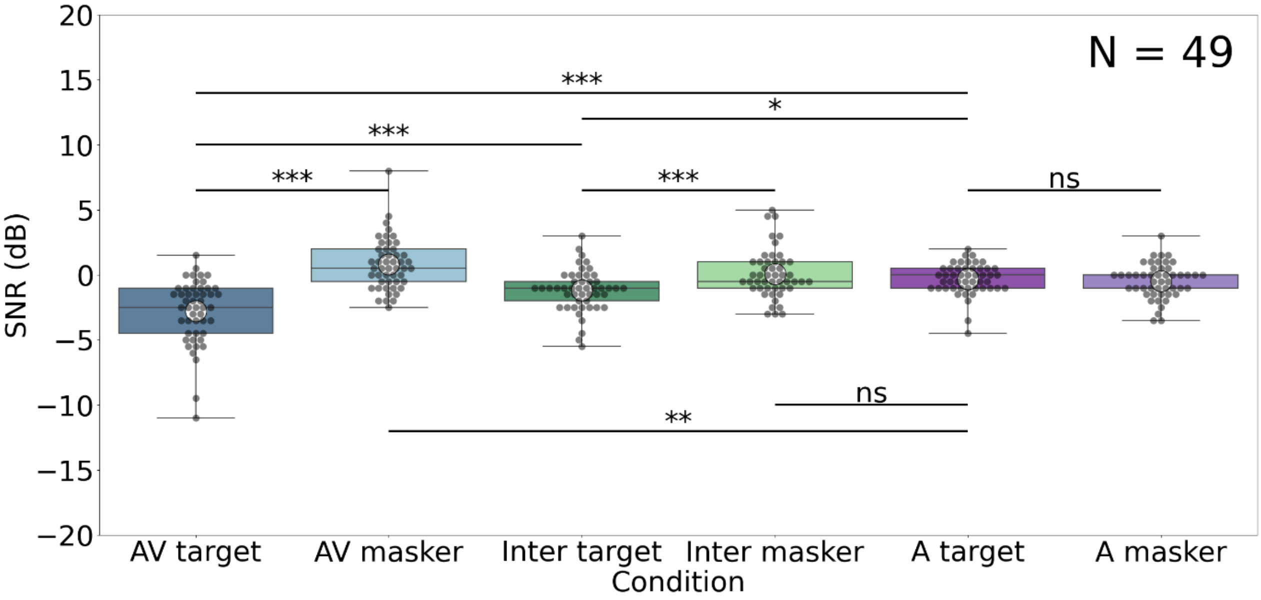


Figure 4.10*: SNR scores for each of the six conditions (target-coherent and masker-coherent for AV, Inter and A) of the vCCRMn speech-in-noise task for the older participant with normal hearing experimental group. White dots depict mean performances. Pairwise comparisons are indicated with horizontal lines over the relevant conditions (*, **, and *** denote significance with p < 0.05, < 0.01, and < 0.001 respectively; ns = non-significant).*

As with the younger normal hearing group, I conducted a three by two repeated measures ANOVA to examine the effects of visual condition (AV, Inter, and A), visual coherence type (Target, Masker), and the interaction between the two, on participants' thresholds. The test yielded non-significant results for the main effect of visual condition (F(2, 282) = 0.130, p = 0.88) and significant results for the main effect of visual coherence type (F(1, 282) = 17.869, p = 0.019), as well as for the interaction between the two (F(2, 282) = 10.494, p = 0.040). Thus, participant performance was influenced by the visual coherence type, but not by the visual condition. Further, the interaction term significance suggested, as with the younger normal hearing group, that the effect of visual condition on performance depended on visual coherence type. Post-hoc pairwise paired t-tests were used to explore these results in more detail, these are shown in Table 4.2 (and Figure 4.2 above).

Comparison type	Condition 1	Condition 2	t-statistic	DoF	Bonferroni corrected p-value	Cohen's d	Mean difference (dB) (Condition 1 – Condition 2)
Visual benefits	AV target	A target	-7.25	48	< 0.001 ***	-1.03	-2.50
	Inter target	A target	-3.40	48	0.020 *	-0.49	-0.87
	AV target	Inter target	-5.13	48	< 0.001 ***	-0.73	-1.59
Visual detriments	AV masker	A target	4.36	48	< 0.001 ***	0.62	1.10
	Inter masker	A target	1.44	48	1	0.21	0.36
	AV masker	Inter masker	3.16	48	0.041 *	0.45	0.73
Visual coherence	AV target	AV masker	-8.83	48	< 0.001 ***	-1.26	-3.56
	Inter target	Inter masker	-3.91	48	0.004 **	-0.56	-1.23
	A target	A masker	0.74	48	1	0.11	0.15
Additional	AV target	Inter masker	-7.92	48	< 0.001 ***	-1.13	-2.82
	AV target	A target	-7.25	48	< 0.001 ***	-1.03	-2.50
	AV target	A masker	-6.74	48	< 0.001 ***	-0.96	-2.31
	AV masker	Inter target	6.63	48	< 0.001 ***	0.95	1.97
	AV masker	A masker	5.06	48	< 0.001 ***	0.72	1.26
	Inter target	A masker	-2.78	48	0.12	-0.40	-0.71
	Inter masker	A masker	2.16	48	0.53	0.31	0.52

Table 4.2*: Summary of the post-hoc pairwise paired t-test comparisons conducted for the older participant with normal hearing experimental group. The comparisons' header column indicates the relevance of these comparisons to the research questions of the chapter. Cohen's d and mean differences between the conditions being compared are also included as measures of effect size. Star symbols in the p-value column denote statistical significance (*, ** and *** indicate significance with p < 0.05, < 0.01, and < 0.001 respectively).*

As for the younger participant with normal hearing experimental group, these results formed the basis for further exploration of the chapter's research questions. In the next sections the conditions under the comparison headers "visual benefits" and "visual detriments" are discussed first, followed by a discussion of the conditions under the comparison header "visual coherence".

4.3.3. Video conditions enhance participant speech-in-noise performance

Once more, for the target-coherent conditions, on average participants performed best in the AV target condition (Mean = -2.77 dB SNR, SD = 2.43 dB SNR), followed by the Inter target (Mean = -1.17 dB SNR, SD = 1.61 dB SNR), and the A target conditions (Mean = -0.31 dB SNR, SD = 1.17 dB SNR). Indeed, as with the younger participant with normal hearing group, the comparisons between the conditions were significant (AV target vs A target: t(48) = -7.25, p < 0.001; Inter target vs A target: t(48) = -3.40, p = 0.02; AV target vs Inter target: t(48) = -5.13, p < 0.001). Participants gained, on average, 2.50 dB SNR more in the AV target condition compared to the A target, 1.59 dB SNR more in the AV target condition compared to the Inter target, and 0.87 dB SNR more in the Inter target condition compared to the A target. Thus, older participants with normal hearing, much like younger participants with normal hearing, exhibit a stepwise increase in performance in moving from the static image to the interrupted video, and finally to the full video condition. Further, older participants with normal hearing also exhibited an AV detriment in the AV masker condition compared to the A target condition (t(48) = 4.36, p < 0.001), with performance in the latter exceeding performance in the former by 1.10 dB SNR. Notably, older participants with normal hearing exhibited an added AV detriment in the AV masker condition compared to the Inter masker (t(48) = 3.16, p = 0.041), but performances did not differ between the Inter masker and A target condition (t(48) = 1.44, p = 1).

Difference histograms for the AV, and Inter benefits are illustrated for the current group in Figure 4.11. In both 4.11 (A), depicting the AV benefit, and (B), depicting the Inter benefit, most values are to the right of zero, confirming that most participants performed better in the video, compared to the static image conditions. As with the younger group, the distribution of AV benefit for the older participant with normal hearing group is more right-skewed than that of the Inter benefit, albeit to a lesser extent than that observed in the younger group. Furthermore, the spread of the AV benefit distribution is greater than the Inter benefit (AV benefit: SD = 2.35 dB SNR; Inter benefit: SD = 1.77 dB SNR).

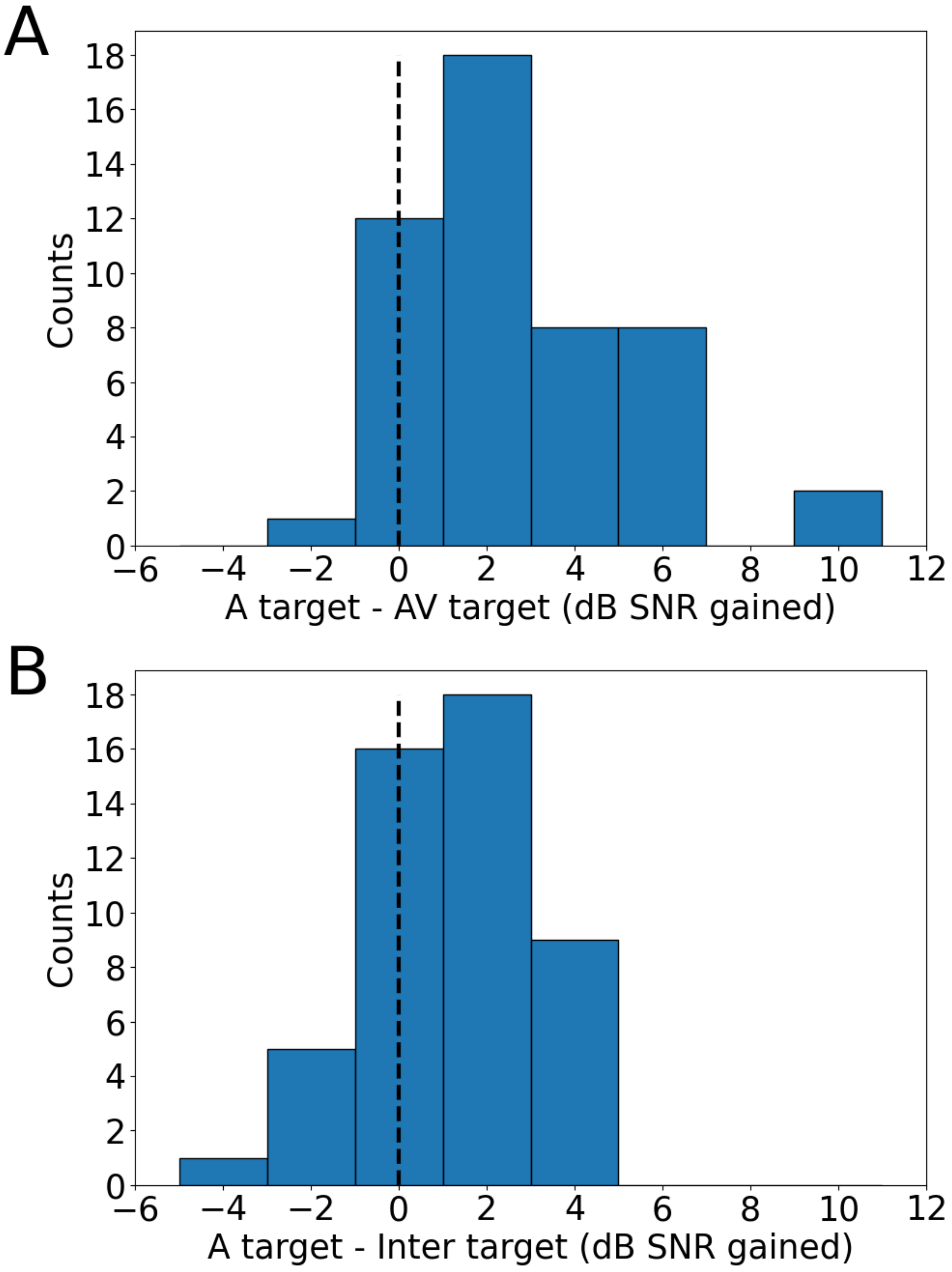


Figure 4.11*: Histograms of visual benefits for the older participant with normal hearing experimental group. (A) Distribution of the performance differences (AV benefits) between the full naturalistic (AV target) video and the static image (A target) conditions. (B) Distribution of the performance differences (Inter benefits) between the interrupted (Inter target) and the static image (A target) conditions. All values at 0 indicate no difference in the performance in the video conditions compared to the static image. Values to the right of 0 demonstrate better performance in each video condition compared to the static image.*

Following my previous approach, I also performed a correlation, investigating the relationship between the performances of the video conditions. Previously, I argued that such correlation would be expected, but nonetheless were not significant for the younger groups. The results are shown for the current group in Figure 4.12. The correlation was significant, and moderate in strength for the current group (AV target and Inter target: Pearson's r = 0.49, p < 0.001), potentially supporting the suggestion that their absence in the previous group was due to insufficient statistical power.

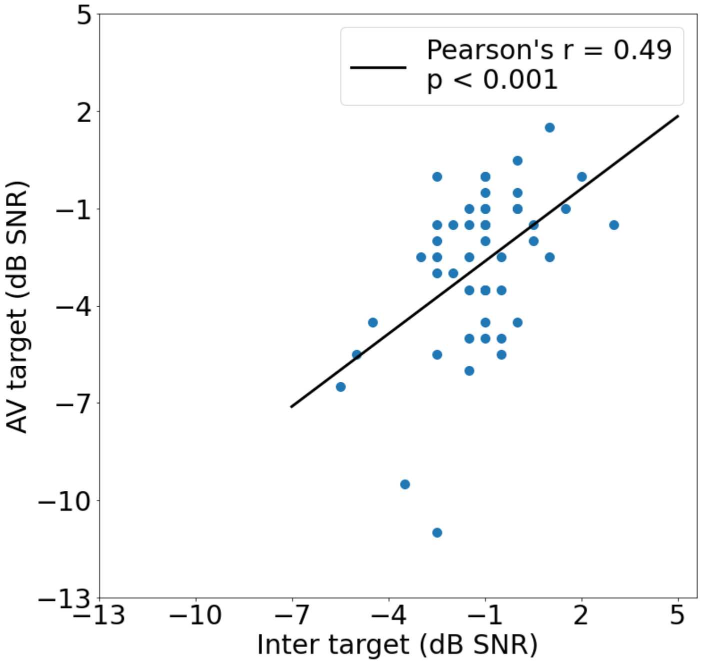


Figure 4.12: *Illustration of the relationship between the AV and Inter conditions for the older participant with normal hearing experimental group. The result of a Pearson correlation conducted between the pair of variables is indicated in the top right corner of the panel.*

4.3.4. Visual coherence influences participant performance

To see whether visual coherence influenced participant performance for the older participant with normal hearing experimental group, I investigated the comparisons under the header "visual coherence" from Table 4.2. The results of these comparisons were similar to those observed for the previous group. Namely, participants performed better in the target-coherent conditions, compared to their respective maskers, for both of the video conditions (AV target vs AV masker: t(48) = -8.83, p < 0.001; Inter target vs Inter masker: t(48) = -3.91, p < 0.01). Further, the comparison between A target and A masker was not significant (t(48) = 0.74, p = 1). Below I outline the conditions' respective means, standard deviations, and mean differences, for reference:

- AV target vs AV masker:
  - AV target: Mean = -2.76 dB SNR, SD = 2.43 dB SNR
  - AV masker: Mean = 0.80 dB SNR, SD = 1.94 dB SNR
  - AV masker – AV target = 3.56 dB SNR, SD = 2.79 dB SNR
- Inter target vs Inter masker:
  - Inter target: Mean = -1.17 dB SNR, SD = 1.61 dB SNR

    - Inter masker: Mean = 0.06 dB SNR, SD = 1.82 dB SNR
    - Inter masker – Inter target = 1.23 dB SNR, SD = 2.18 dB SNR
- A target vs A masker:
    - A target: Mean = -0.31 dB SNR, SD = 1.17 dB SNR
    - A masker: Mean = -0.46 dB SNR, SD = 1.31 dB SNR
    - A masker – A target = -0.15 dB SNR, SD = 1.42 dB SNR

The data were also visualised with histograms, plotting the distributions of the differences between the masker and target conditions (Figure 4.13). In particular, Figure 4.13 (A) depicts the distribution of AV masker – AV target, (B) the distribution of Inter masker – Inter target, and (C) the distribution of A masker – A target. Most values of (A) and most of (B) are to right of zero, indicating better performance in the target-coherent than the masker-coherent video conditions, in accordance with the results observed for the younger participant with normal hearing group. Further, the distribution of the static image differences was roughly symmetrical about zero, confirming that performances were similar across the two conditions. The AV conditions' difference distribution also had the longest right tail, among the three. Finally, and as was the case with the visual benefits, the variability in the AV conditions' differences was greater than that observed in the Inter and A conditions' differences, although comparable to that of the Inter conditions' differences (AV masker – AV target: SD = 2.79 dB SNR; Inter masker – Inter target: SD = 2.18 dB SNR; A masker – A target: SD = 1.42 dB SNR).

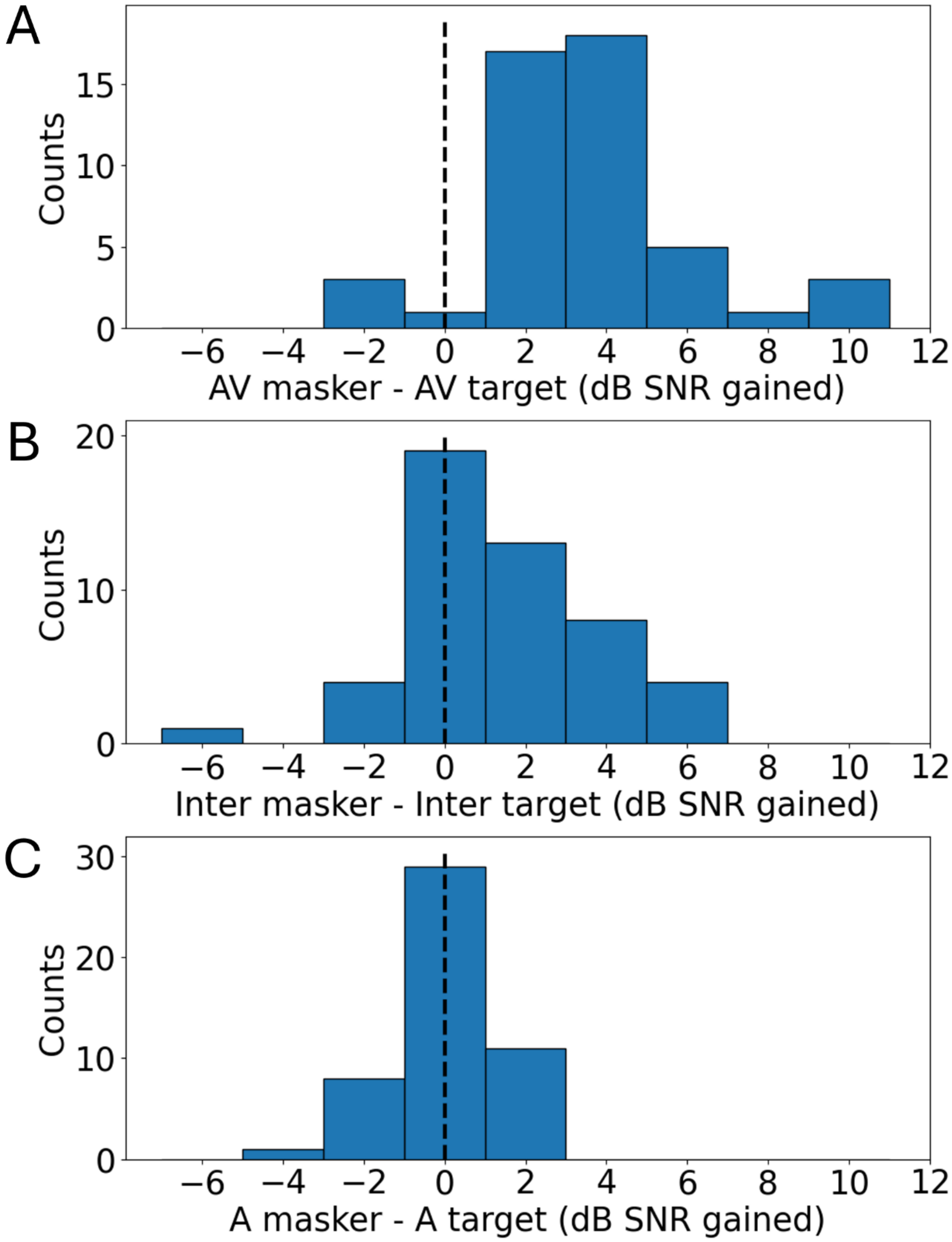


***Figure 4.13**: Distribution of the performance differences between the masker- and target-coherent conditions of the vCCRMn for the older participant with normal hearing experimental group. (A) Distribution of the AV masker – AV target. (B) Distribution of Inter masker – Inter target. (C) Distribution of A masker – A target. All values at 0 indicate no difference in performance between the masker-coherent and target-coherent conditions. Values to the right of 0 demonstrate better performance in the target-coherent condition.*

4.3.5. Visual benefits do not correlate with lipreading ability

We saw in section 4.2.5 that lipreading ability, measured via the TAS sentences scores, correlated with the AV benefits, but not the Inter benefits of younger participants with normal hearing. On the other hand, the Inter benefit of younger participants with normal hearing correlated negatively with the lipreading-related gains obtained from the AV target condition compared to the Inter target (Inter target – AV target variable). The same analyses were conducted herein to assess the relationship between the older participant with normal hearing group's benefits and lipreading abilities. Results of the correlations between AV benefit and TAS sentences scores and Inter benefit and TAS sentences scores are depicted in Figures 4.14 and 4.15A, respectively. Once more, the expectation was that lipreading ability would be correlated with the AV, but not the Inter benefit. Nonetheless, neither of the two analyses were significant (AV benefit and TAS sentences score: Pearson's r = 0.14, p = 0.33; Inter benefit and TAS sentences score: Pearson's r = 0.17, p = 0.24). To conduct the Inter benefit with Inter target – AV target correlation for the current group, the influence of PTA scores of participants (same values that were used in their hearing status classification) were controlled for by conducting a partial correlation. As with the younger normal hearing group, the correlation between Inter benefit and Inter target – AV target was not significant (Pearson's r = -0.29, p = 0.088) (Figure 4.15B).

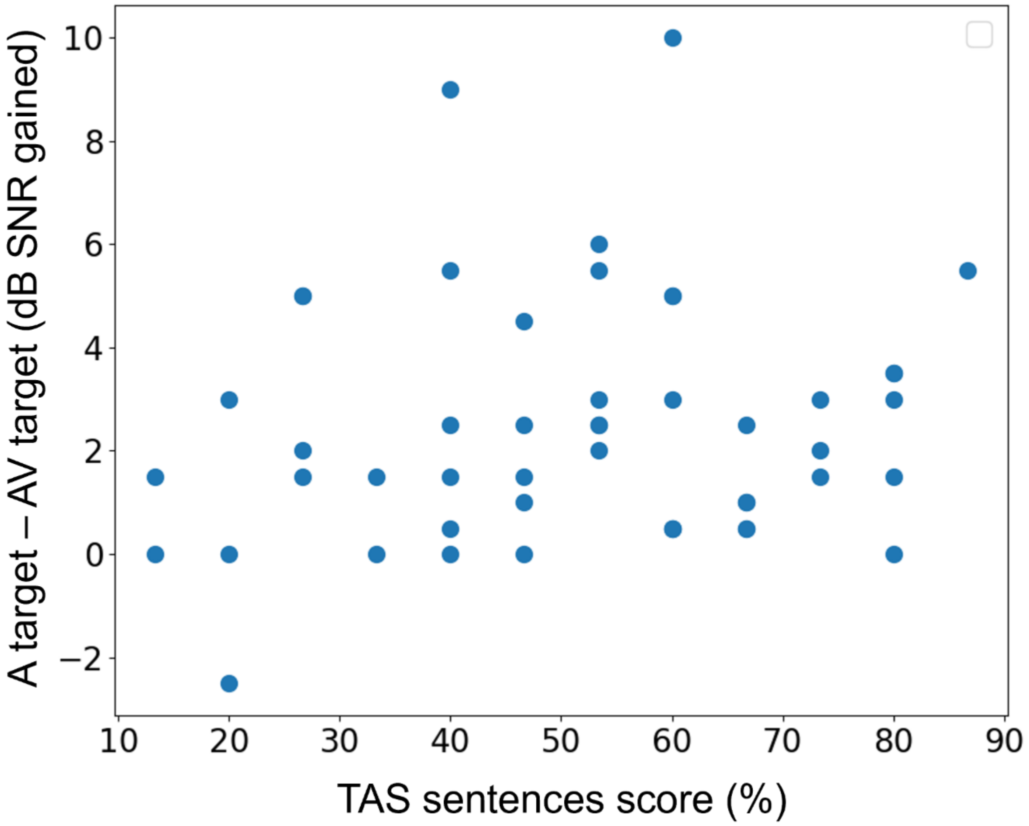


Figure 4.14*: Illustration of the relationship between AV benefit and TAS sentences scores for the older participant with normal hearing experimental group. A Pearson correlation analysis was performed for the two variables and was not significant.*

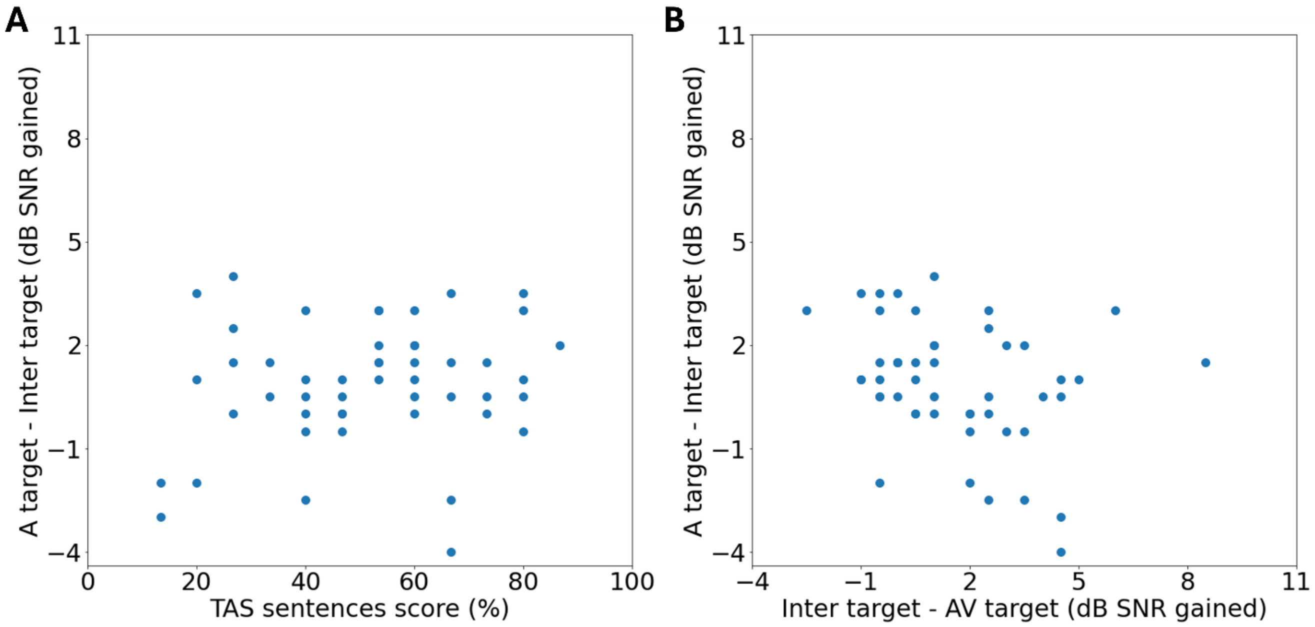


***Figure 4.15**: Illustration of the relationship between Inter benefit and lipreading ability for the older participant with normal hearing experimental group. (A) Inter benefit plotted against TAS sentences scores. A Pearson correlation was conducted to assess the variables' association and was not significant. (B) Inter benefit plotted against the difference between the performances in the Inter target and AV target conditions (i.e. the lipreading-related gains in the latter compared to the former). A partial Pearson correlation was conducted (controlling for PTA) to assess the variables' association and was not significant.*

4.4. Experiment 3: Results from the older participants with hearing loss

Sections 4.2 and 4.3 discussed the analyses of experiments 1 and 2. Generally, both experimental groups showed consistent results, with participants gaining visual benefits from both the AV and the Inter target conditions, compared to the A target. Further, target-coherence in the video conditions resulted in better performances than masker coherence, in both groups.

In the following sections, I analyse the data from the third experiment I conducted, which included testing of the older participants with hearing loss. A total of 37 participants constituted this dataset (Ages 46-85, Mean = 70.32, SD = 7.65, with 36 out of 37 participants over the age of 50).

4.4.1. Participant audiograms

Audiograms for the older participants with hearing loss are displayed in Figure 4.16. All participants herein were classified as having hearing loss, following the same classification criterion as that used for the hearing status determination of the previous groups. Namely, the participants of this experiment had better ear mean PTA thresholds of over 20 dB HL over the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. Specifically, 36 out of the 37 older participants with hearing loss had mild hearing loss (i.e. 20 dB HL < mean threshold ≤ 40 dB HL), and 1 had moderate hearing loss (i.e. 40 dB HL < mean threshold ≤ 70 dB HL).

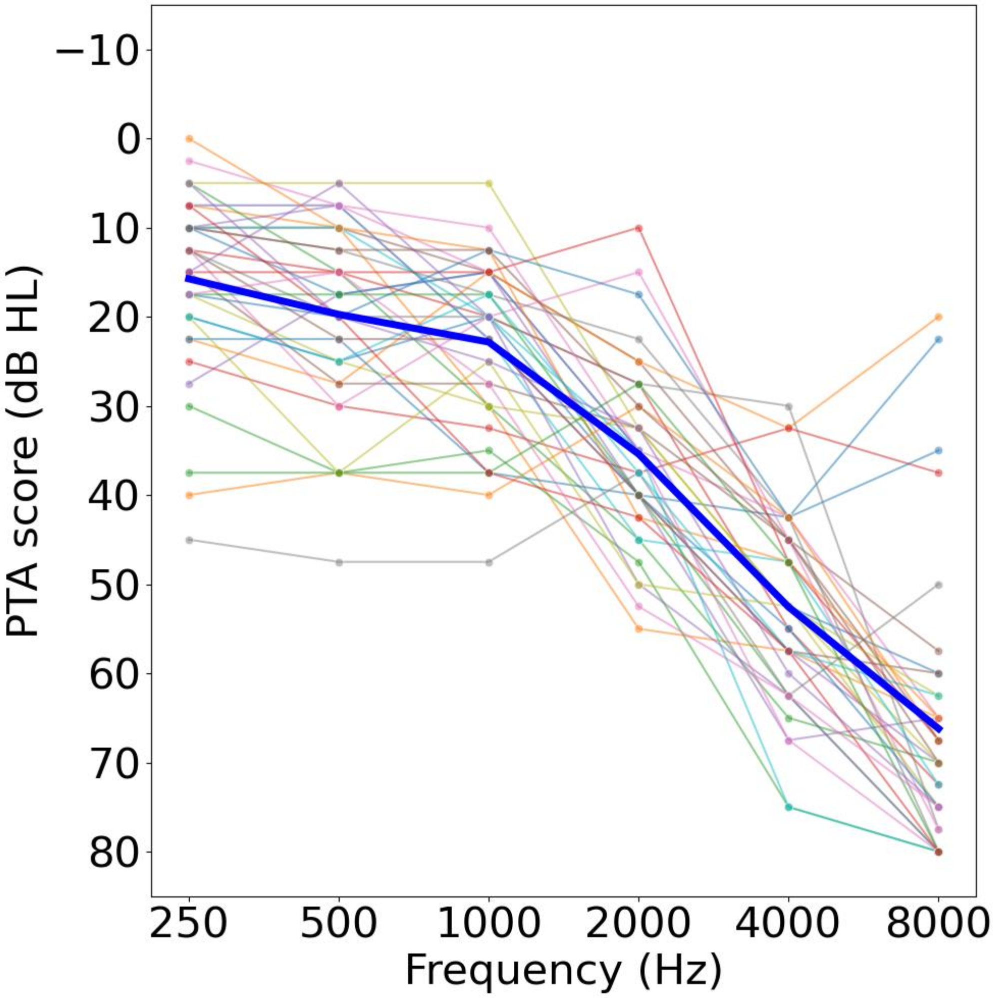


Figure 4.16*: Mean pure tone threshold across both ears as a function of frequency in older participants with hearing loss. Each line represents a participant. The thick central line depicts the mean performance across all participants included in this group.*

Additionally, as I did for in section 4.3.1, I plotted side-by-side boxplots of the mean thresholds (based on the aforementioned criterion), of the three experimental groups (Figure 4.17). For reference, the respective group means and standard deviations were:

- Younger participants with normal hearing: Mean = 1.18 dB HL, SD = 4.36 dB HL
- Older participants with normal hearing: Mean = 9.03 dB HL, SD = 5.79 dB HL
- Older participants with hearing loss: Mean = 29.09 dB HL, SD = 6.09 dB HL

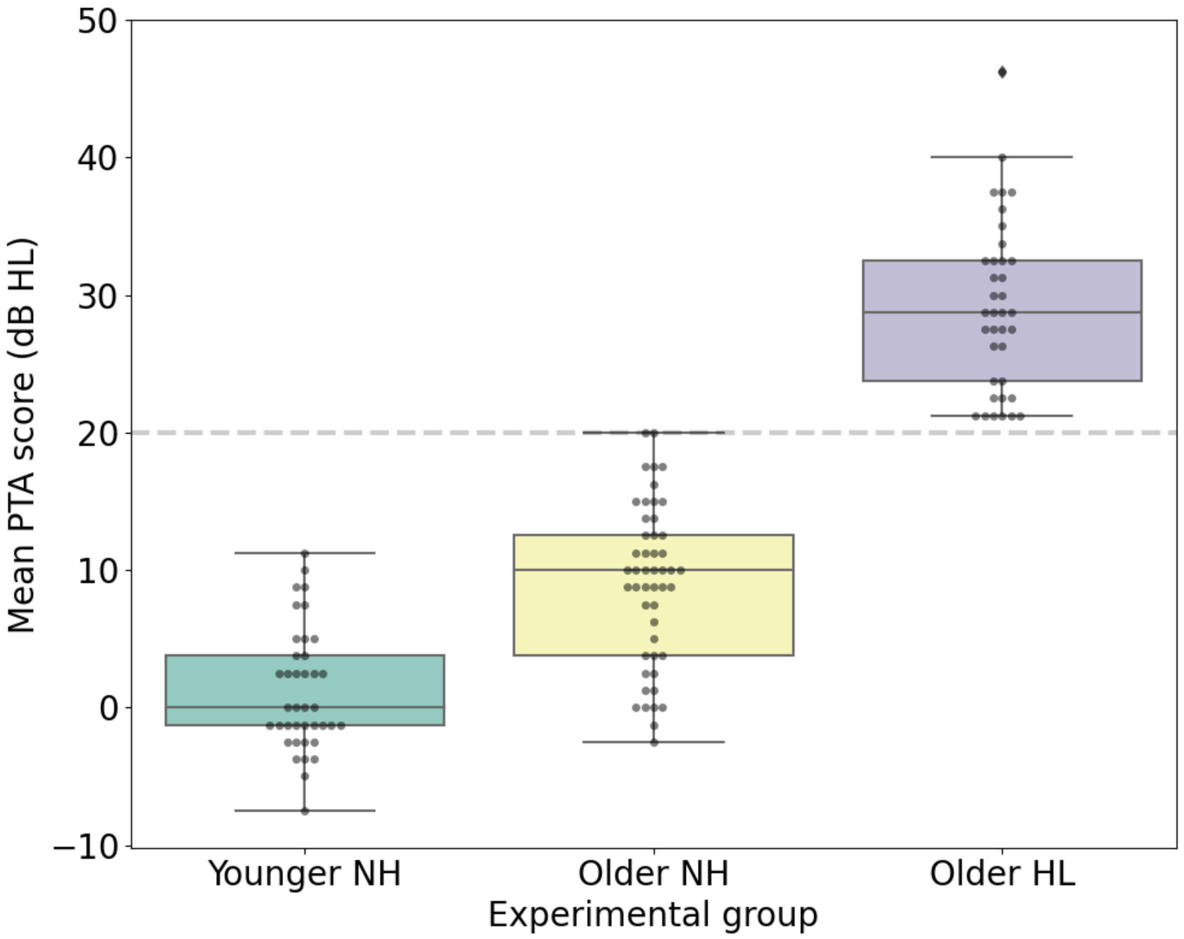


Figure 4.17*: Side-by-side boxplots of mean PTA thresholds for the younger participant with normal hearing, older participant with normal hearing, and older participant with hearing loss experimental groups. Mean PTA thresholds here were computed for the participants' better ear, over the range of frequencies including 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. The horizontal dotted line indicates the threshold (20 dB HL) above which participants* were *classified to have hearing loss.*

4.4.2. Speech-in-noise performance varies across the conditions of the task

In the younger, and older, normal hearing experiments' analyses, we saw that participants' speech-in-noise performances were boosted when looking at the talker's video. Both the AV and the Inter video conditions seemed to provide a benefit over the static image condition, and performance was better when the video displayed the face of the talker uttering the target sentence, than when displaying the face of one of the talkers uttering the masker sentences. To investigate the effects of visual conditions of the current group's performances, I once more plotted participant performances across the conditions of the vCCRMn (Figure 4.18).

Looking at the distributions of performances across the boxplots, the difference between the current and previous groups is evident: The boxplots for the last four plotted conditions (Inter target, and masker, and A target and masker) are greatly overlapping, indicating similar performances across these conditions; the only performance clearly standing out, with better scores than the rest, is the one for the AV target condition. Based on further visual inpection, the AV masker condition performances appear to be the worst, albeit not substantially, compared to the last four boxplots.

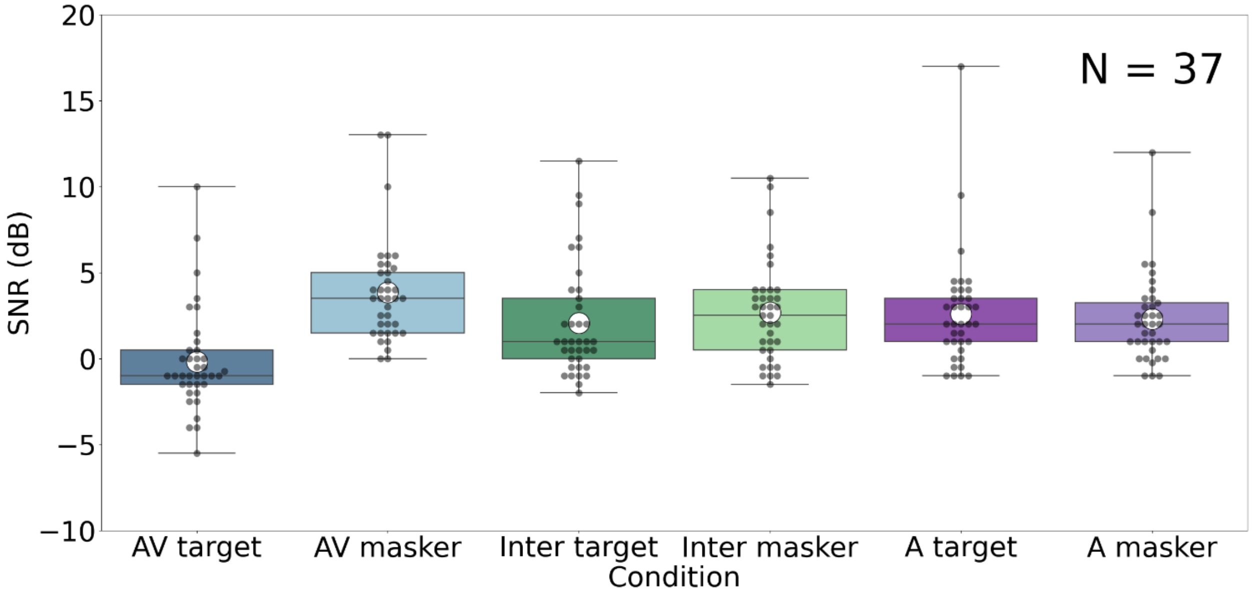


Figure 4.18*: SNR scores for each of the six conditions (target-coherent and masker-coherent for AV, Inter and A) of the vCCRMn speech-in-noise task for the older participant with hearing loss experimental group. White dots depict mean performances. A one way-repeated measured ANOVA was run to compared performance across the conditions and was not significant.*

A three by two repeated measures ANOVA was performed, to assess the effects of visual condition, visual coherence type, and their interaction, as was done for the previous groups. None of the effects were significant (Visual condition: F(2, 210) = 0.442, p = 0.64; Visual coherence type: F(1, 210) = 2.77, p = 0.10; Interaction: F(2, 210) = 2.80, p = 0.07). Given this outcome, I did not proceed with conducting all possible post-hoc pairwise comparisons between the conditions. I did, nonetheless, continue with further explorations of the conditions involved in visual benefits, and visual coherence, as I did with the previous groups.

4.4.3. Only the AV video condition enhances participant speech-in-noise performance

As before, the visual benefits obtained by participants in the AV and Inter target conditions compared to the static image condition were plotted as histograms (Figure 4.19). Summary statistics of participant performances and visual benefits are also provided below:

- AV target: Mean = -0.18 dB SNR, SD = 2.94 dB SNR
- Inter target: Mean = 2.08 dB SNR, SD = 3.24 dB SNR
- A target: Mean = 2.60 dB SNR, SD = 3.25 dB SNR
- AV benefit: Mean = 2.78 dB SNR, SD = 2.66 dB SNR
- Inter benefit: Mean = 0.52 dB SNR, SD = 2.65 dB SNR

Looking at the histograms, and considering the above summary statistics, we see that participants seem to have performed better in the AV target condition, compared to the A target (Figure 4.19 (A)), and thus were likely to have gained an AV benefit. Indeed, a paired t-test comparison between AV target and A target was significant (t(36) = -6.27, p < 0.001), with participants gaining an average AV benefit of 2.78 dB SNR. Similarly, a comparison between AV target and Inter target yielded significant results (t(36) = -3.10, p < 0.01), with participants gaining a performance boost of 2.26 dB SNR in the AV target condition compared to the Inter target. Further, and as with the previous two groups, in addition to the observed AV benefit, older participants with hearing loss also received an AV detriment from the AV masker condition (Mean = 3.84 dB SNR, SD = 2.98 dB SNR), compared to the A target (t(36) = 2.64, p = 0.01).

The distribution of the Inter benefits was roughly symmetrical about zero, suggesting that, on average, older participants with hearing loss were not benefiting from the Inter target condition relative to the A target condition. Similarly, the mean performances of the two conditions stated above were close to each other (mean difference = 0.52 dB SNR). A paired t-test comparison confirmed this, yielding non-significant results (t(36) = -0.68, p = 0.50).

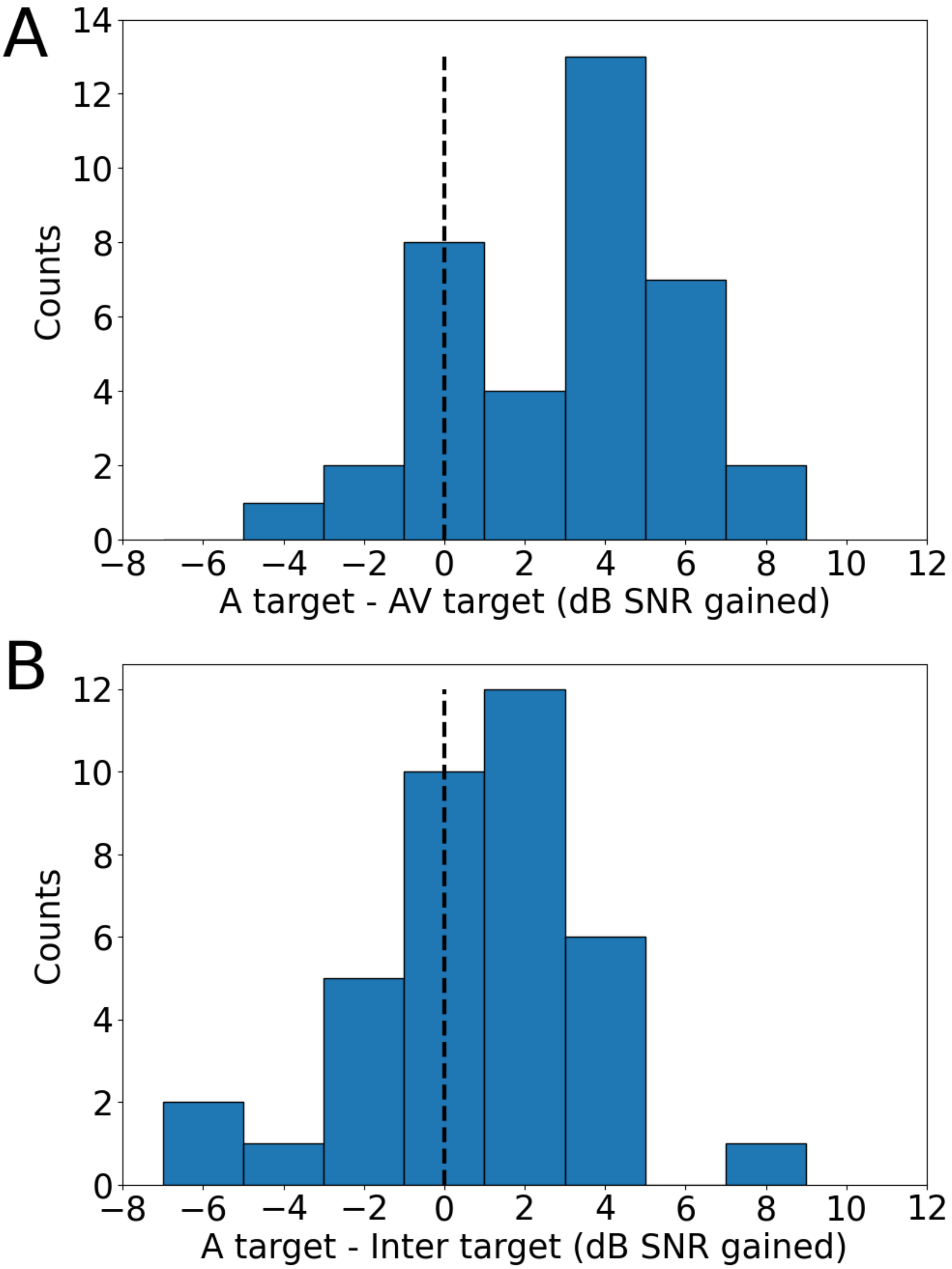


***Figure 4.19**: Histograms of visual benefits for the older participant with hearing loss experimental group. (A) Distribution of the performance differences (AV benefits) between the full naturalistic (AV target) video and the static image (A target) conditions. (B) Distribution of the performance differences (Inter benefits) between the interrupted (Inter target) and the static image (A target) conditions. All values at 0 indicate no difference in the performance in the video conditions compared to the static image. Values to the right of 0 demonstrate better performance in each video condition compared to the static image.*

Having computed the visual benefits for all three experimental groups, I observed that they appeared to decrease, on average, going from the younger participant with normal hearing group (AV benefit mean = 5.51 dB SNR, SD = 4.19 dB SNR; Inter benefit mean = 1.91 dB SNR, SD = 2.49 dB SNR), to the older groups (Older participants with normal hearing: AV benefit mean = 2.50 dB SNR, SD = 2.35 dB SNR, Inter benefit mean = 0.87 dB SNR, SD = 1.77 dB SNR; Older participants with hearing loss: AV benefit mean = 2.78 dB SNR, SD = 2.66 dB SNR, Inter benefit mean = 0.52 dB SNR, SD = 2.65 dB SNR). Inter benefit also appeared to decrease in the older participant with hearing loss group, compared to the older participant with normal hearing group. Side-by-side boxplots of the AV benefits and Inter benefits for each group are displayed in Figures 4.20 and 4.21 respectively.

To formally assess the effect of "group", I conducted two separate one-way ANOVAs: One for the AV benefits, and one for the Inter benefits, both of which were significant (AV benefits: F(2, 122) = 11.52, p < 0.001; Inter benefits: F(2, 122) = 3.786, p = 0.025). Post-hoc pairwise t-test comparisons are included in Table 4.3.

Comparison type	**Condition 1**	**Condition 2**	**t-statistic**	**DoF**	**Bonferroni corrected p-value**	**Cohen's d**	**Mean difference (dB)** **(Condition 1 – Condition 2)**
AV benefits	Younger NH	Older NH	4.02	56.51	< 0.001 ***	0.64	3.12
	Younger NH	Older HL	3.36	64.84	< 0.01 **	0.58	2.61
	Older NH	Older HL	-0.58	71.97	1.000	-0.16	-0.54
Inter benefits	Younger NH	Older NH	2.18	65.96	0.098	0.32	0.97
	Younger NH	Older HL	2.32	73.03	0.070	0.41	1.37
	Older NH	Older HL	0.68	59.05	1.000	0.07	0.26

***Table 4.3**: Summary of the post-hoc pairwise t-test comparisons conducted for the visual benefits across the three experimental groups. Younger NH = younger participants with normal hearing; Older NH = older participants with normal hearing; Older HL = older participants with hearing loss. Star symbols in the p-value column denote statistical significance (** and *** indicate significance with p < 0.01 and < 0.001 respectively).*

Based on these comparisons, younger participants with normal hearing gained more AV benefit than both older participants with normal hearing (t(56.51) = 4.02, p < 0.001, mean difference = 3.12 dB SNR), and older participants with hearing loss (t(64.84) = 3.36, p < 0.01, mean difference = 2.61 dB SNR). The AV benefits of older participants with normal hearing and older participants with hearing loss were, however, not statistically different (t(71.97) = -0.58, p = 1). Furthermore, although the ANOVA detected a significant difference between the three groups for the Inter benefits, the pair-wise comparisons produced non-significant results (younger participants with normal hearing vs older participants with normal hearing: t(65.96) = 2.18, p = 0.098; younger participants with normal hearing vs older participants with hearing loss: t(73.03 = 2.32, p = 0.070); older participants with normal hearing vs older participants with hearing loss: t(59.05, p = 1). Thus, in summary, the AV benefit decreased in the older groups compared to the younger group but were similar across the two older groups.

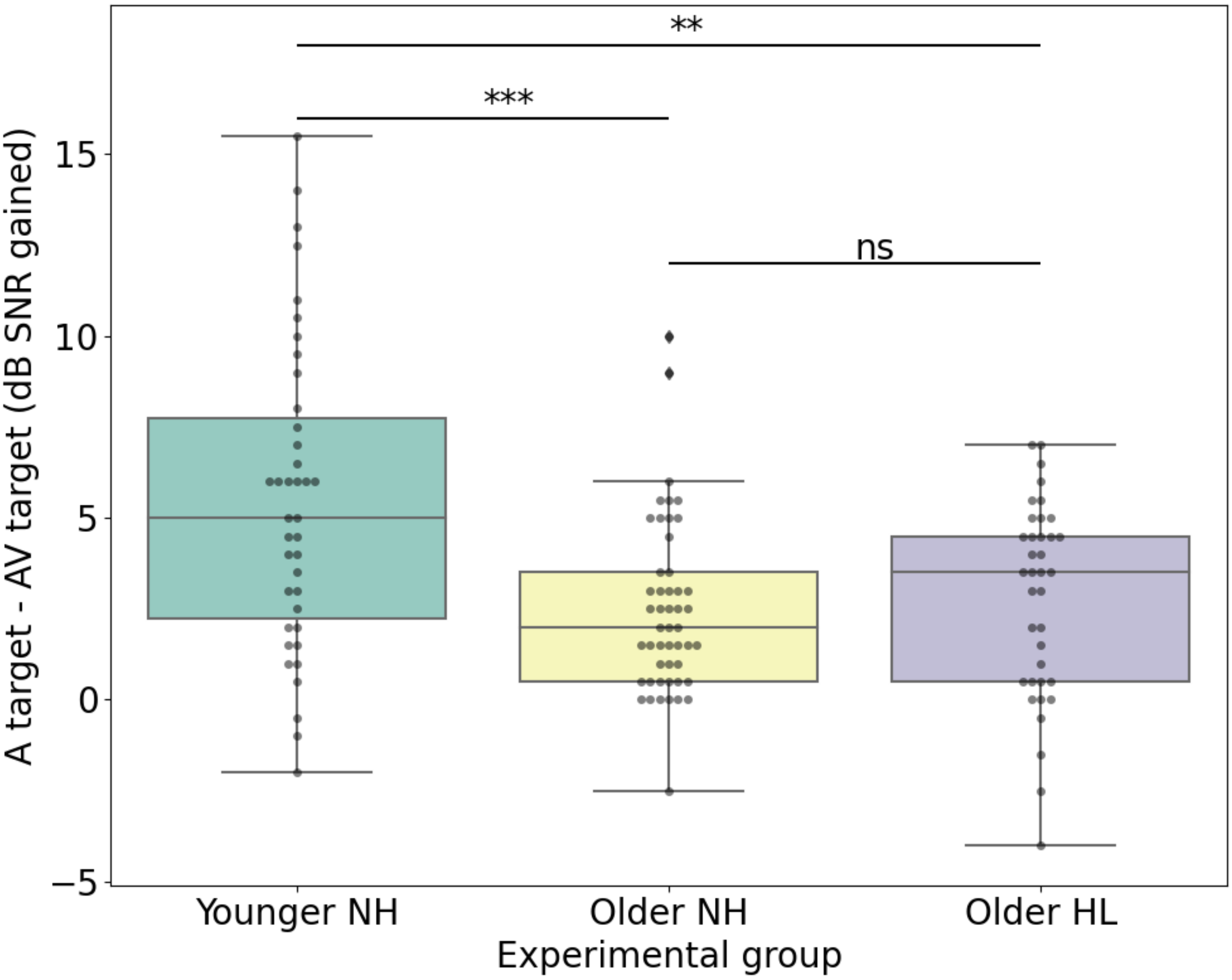


Figure 4.20*: Boxplots of the AV benefits obtained from participants in each of the three experimental groups. Pairwise comparisons are indicated with horizontal lines over the relevant groups (**, *** denote significance with p < 0.01 and < 0.001 respectively; ns = non-significant). Younger NH = younger participants with normal hearing; Older NH = older participants with normal hearing; Older HL = older participants with hearing loss.*

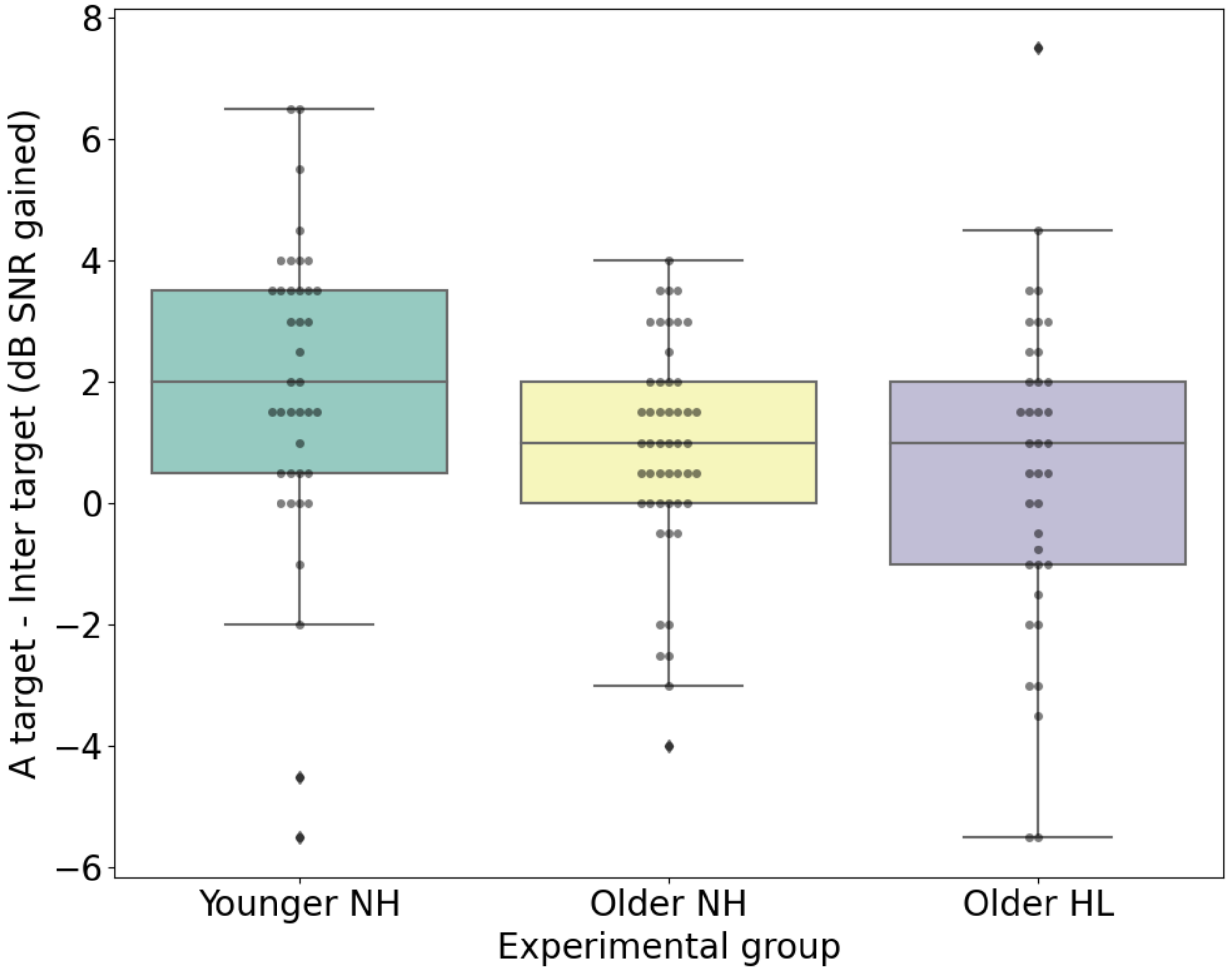


Figure 4.21*: Boxplots of the Inter benefits obtained from participants in each of the three experimental groups. Pairwise t-test comparisons were run and were not significant. Younger NH = younger participants with normal hearing; Older NH = older participants with normal hearing; Older HL = older participants with hearing loss.*

Finally, and as previously done for the other groups, I concluded this section's investigations with a correlational analysis, examining the association between the video conditions for the older participant with hearing loss group (Figure 4.22). Similarly to the older, and as was expected for the younger participant with normal hearing groups, the current group showed a significant correlation between both the AV target and Inter target conditions (Pearson's r = 0.63, p < 0.001).

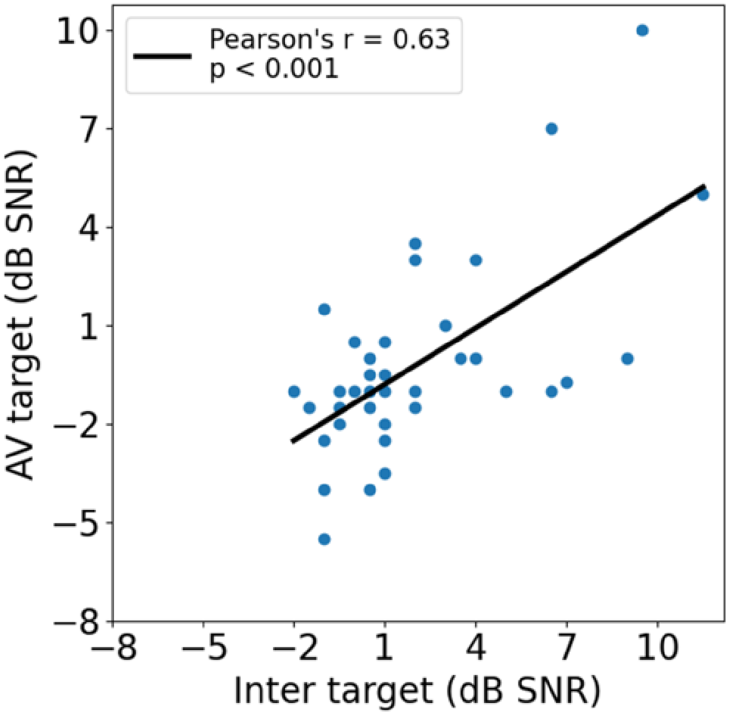


***Figure 4.22**: Illustration of the relationship between the AV and Inter conditions for the older participant with hearing loss experimental group. (A) AV target plotted against Inter target. (B) AV benefit plotted against Inter benefit. The results of Pearson correlations conducted between the two pairs of variables are indicated in the top left, and bottom right corners of their respective panels.*

4.4.4. Visual coherence in the AV conditions influences participant performance

Next, the distributions of the differences in performance between masker and target were plotted as histograms for the older participant with hearing loss group (Figure 4.23). Namely, Figure 4.23 (A) displays the variable AV masker – AV target, (B) the variable Inter masker – Inter target, and (C) A masker – A target. Summary statistics of participant performances for the relevant conditions and the listed differences are also provided below for reference:

- AV target vs AV masker:
  - AV target: Mean = -0.18 dB SNR, SD = 2.94 dB SNR
  - AV masker: Mean = 3.84 dB SNR, SD = 2.98 dB SNR
  - AV masker – AV target: Mean = 4.03 dB SNR, SD = 2.95 dB SNR
- Inter target vs Inter masker:
  - Inter target: Mean = 2.08 dB SNR, SD = 3.24 dB SNR
  - Inter masker: Mean = 2.69 dB SNR, SD = 2.89 dB SNR
  - Inter masker – Inter target: Mean = 0.61 dB SNR, SD = 2.40 dB SNR

- A target vs A masker:
  - A target: Mean = 2.60 dB SNR, SD = 3.25 dB SNR
  - A masker: Mean = 2.31 dB SNR, SD = 2.59 dB SNR
  - A masker – A target: Mean = -0.29 dB SNR, SD = 2.39 dB SNR

Based on the histograms of Figure 4.23, and the summary statistics above, it was deemed likely that older participants with hearing loss performed better in the AV target condition compared to the AV masker. A paired t-test comparison confirmed this hypothesis (t(36) = -8.18, p < 0.001). However, coherency-type does not seem to have differentially influenced the performances in the remaining two comparisons: The histograms of Figure 4.23 (B) and (C) are roughly symmetrical about zero, and the mean performances of the respective conditions involved in the distributions are similar to each other (based on the summary statistics reference list above). Indeed, paired t-test comparisons for the Inter masker/Inter target and A masker/A target pairs were not significant (Inter target vs Inter masker: t(36) = -1.52, p = 0.14; A target vs A masker: t(36) = 0.73, p = 0.47). Thus, coherency only influenced older participant with hearing loss performances within the AV conditions, and not within the Inter, and A conditions.

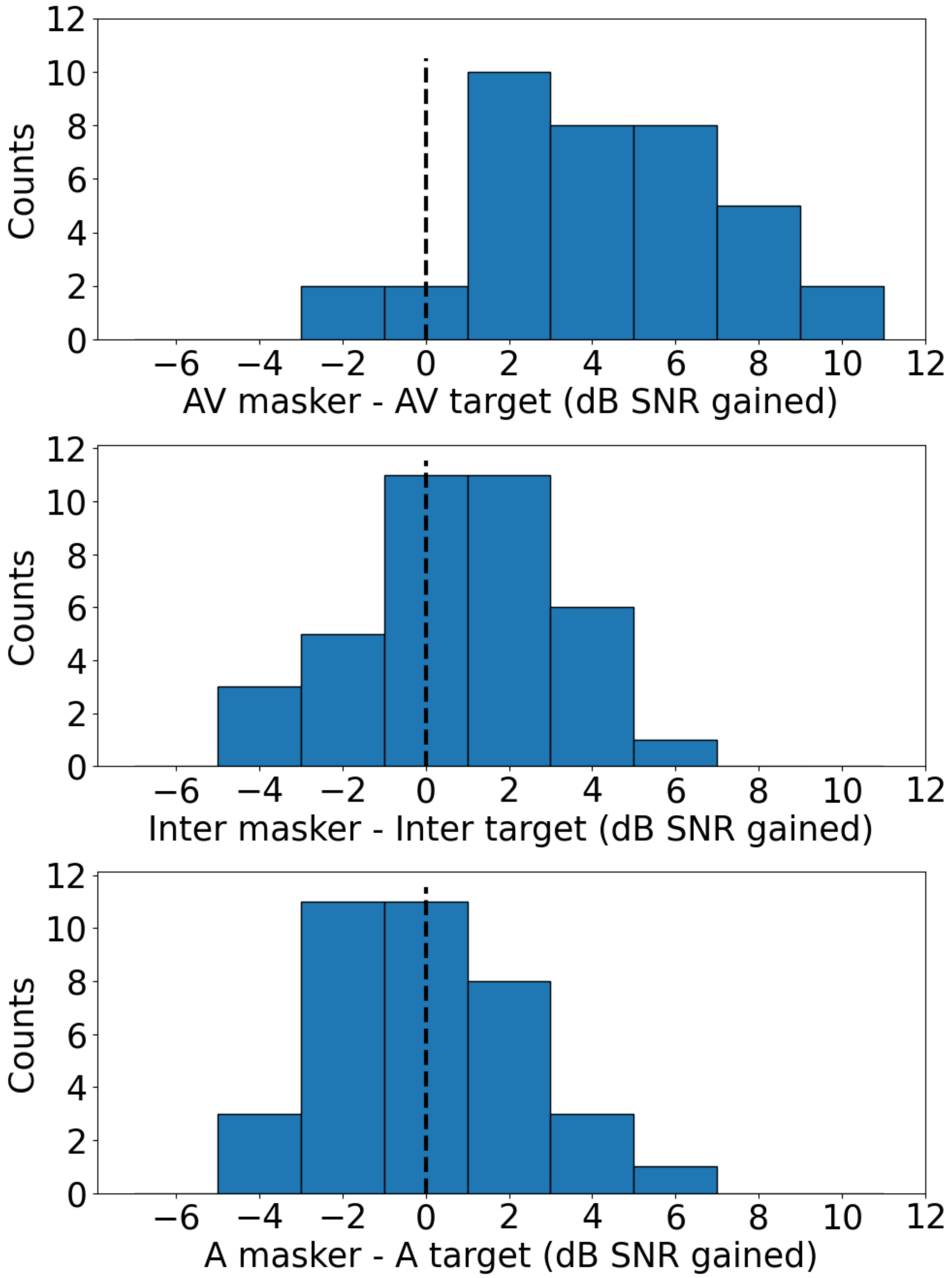


***Figure 4.23**: Distribution of the performance differences between the masker- and target-coherent conditions of the vCCRMn for the older participant with hearing loss experimental group. (A) Distribution of the AV masker – AV target. (B) Distribution of Inter masker – Inter target. (C) Distribution of A masker – A target. All values at 0 indicate no difference in performance between the masker-coherent and target-coherent conditions. Values to the right of 0 demonstrate better performance in the target-coherent condition.*

4.4.5. Visual benefits correlations with lipreading ability

In the younger participant with normal hearing experimental group, we saw that participant lipreading ability correlated with their AV benefit, but not with their Inter benefit. In the older participant with normal hearing group, neither of these two correlations were signficant. These analyses were once more repeated for the current group. Namely, the AV, and Inter benefits of older participants with hearing loss were correlated with their TAS sentence lipreading scores (Figures 4.24, and 4.25A respectively). Additionally, as with the previous groups, the Inter benefits of the current group were correlated with the lipreading related gains obtained in the AV target condition compared to the Inter target condition (i.e. with Inter target – AV target) (Figure 4.25B). As done for the previous older group, PTA was controlled for in the current group's correlational analysis.

Similarly to the older participant with normal hearing group, neither of the correlations with TAS sentences scores yielded significant results in the current group (AV benefit correlation: Pearson's r = 0.22, p = 0.19; Inter benefit correlation: Pearson's r = 0.0002, p = 1.00). Thus, based on the current analyses, it appears that performance on the TAS sentences test was not associated with the visual benefits observed in older individuals. Contrary to the previous two groups, however, the current group exhibited a moderate, negative, correlation between the variables Inter benefit and Inter target – AV target (Pearson's r = -0.53, p < 0.001). This suggests that, potentially, the worse the current participants were at one of the two visual mechanisms, the more they relied on the other for their visual gains.

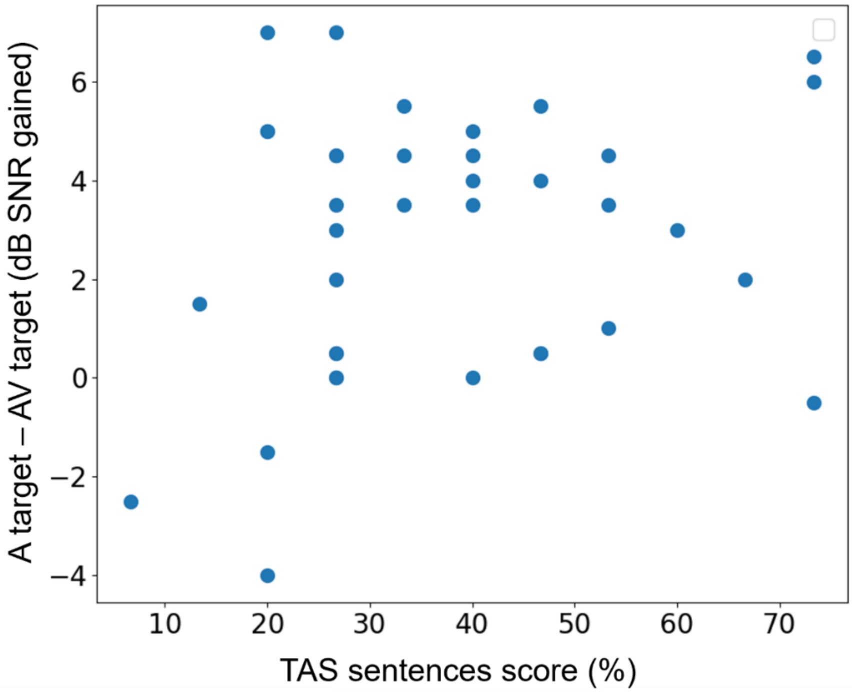


Figure 4.24: *Illustration of the relationship between AV benefit and TAS sentences scores for the older participant with hearing loss experimental group. A Pearson correlation analysis was performed for the two variables and was not significant.*

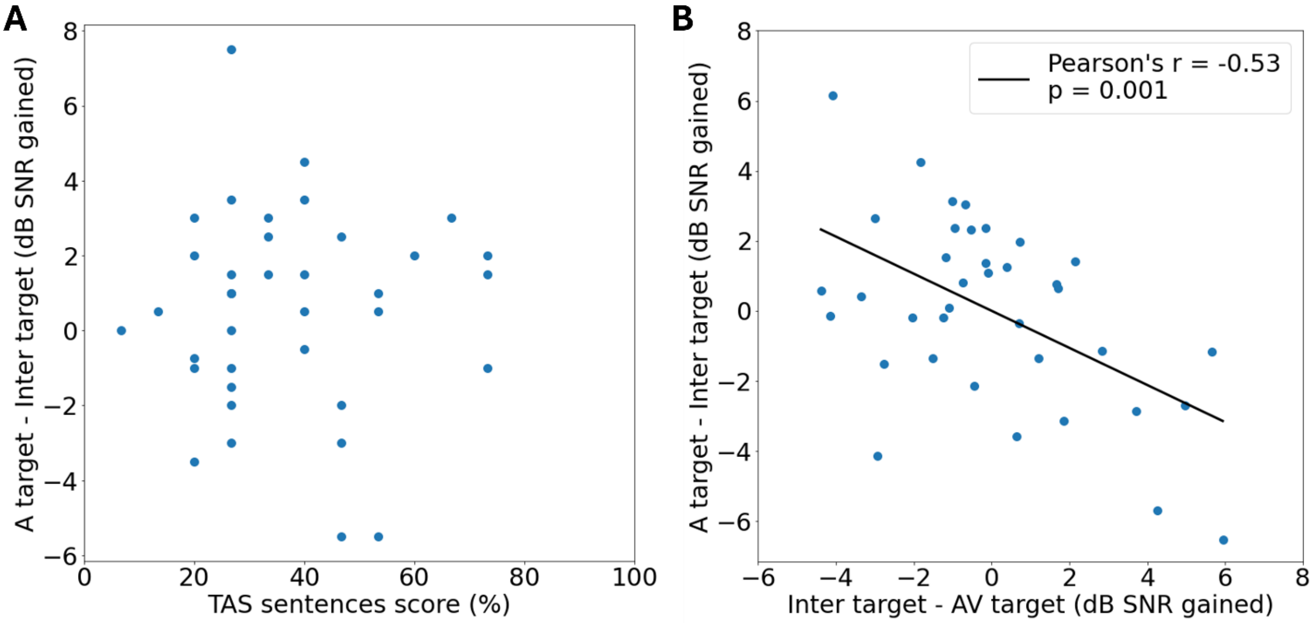


***Figure 4.25**: Illustration of the relationship between Inter benefit and lipreading ability for the older participant with hearing loss experimental group. (A) Inter benefit plotted against TAS sentences scores. A Pearson correlation was conducted to assess the variables' association and was not significant. (B) Inter benefit plotted against the difference between the performances in the Inter target and AV target conditions (i.e. the lipreading-related gains in the latter compared to the former). A partial Pearson correlation was conducted (controlling for PTA) to assess the variables' association and was significant (Pearson's r = -0.53, p < 0.001).*

As a last step in my investigation, I was interested in examining whether the lipreading capabilities differed between the three experimental groups. To this end, I plotted participants TAS sentence scores in side-by-side boxplots (Figure 4.26), across the three groups, and conducted an ANOVA to compare the effect of group. The ANOVA was significant (F(2, 122) = 13.07, p < 0.001), and the results of post-hoc t-test comparisons are indicated with horizontal lines over the relevant boxplots in Figure 4.26, and are also listed in Table 4.4 below.

Based on these comparisons, the lipreading ability of the normal hearing groups was greater than that of the older participant with hearing loss group (younger participants with normal hearing vs older participants with hearing loss: t(72.53) = 5.37, p < 0.001, mean difference = 19.31%; older participants with normal hearing vs older participants with hearing loss: t(82.01) = 3.52, p < 0.01, mean difference = 14.42 %). The comparison between the two normal hearing groups was, however, not significant (t(85.99) = 1.68, p = 0.29).

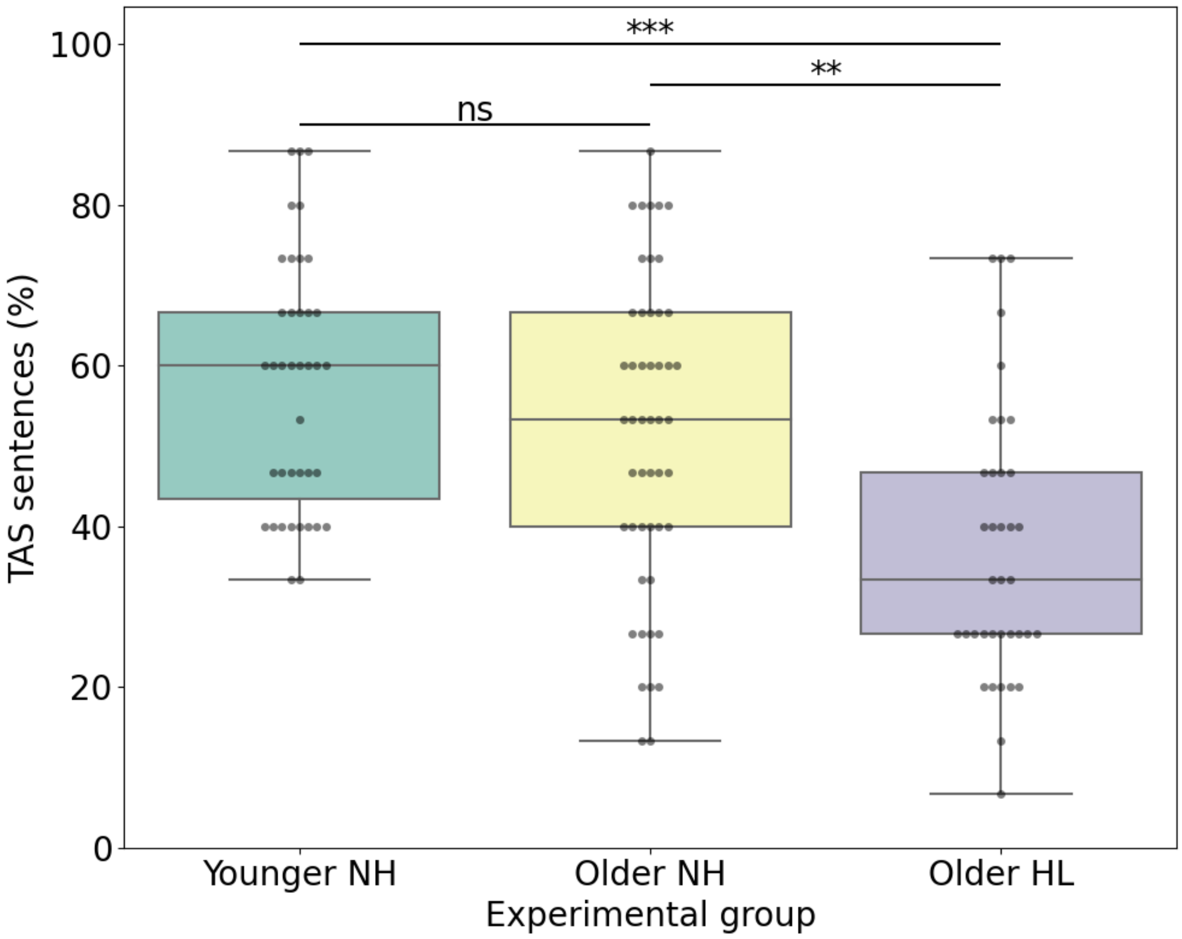


***Figure 4.26**: Side-by-side boxplots of TAS sentence scores of participants across the three experimental groups. Pairwise comparisons are indicated with horizontal lines over the relevant groups (**, and *** denote significance with p < 0.01 and < 0.001 respectively; ns = non-significant). Younger NH = younger participants with normal hearing; Older NH = older participants with normal hearing; Older HL = older participants with hearing loss.*

Condition 1	Condition 2	t-statistic	DoF	Bonferroni corrected p-value	Cohen's d	Mean difference (dB) (Condition 1 – Condition 2)
Younger NH	Older NH	1.68	85.99	0.29	0.26	5.99
Younger NH	Older HL	5.37	72.53	< 0.001 ***	0.96	19.32
Older NH	Older HL	3.52	82.01	< 0.01 **	0.55	14.42

***Table 4.4**: Summary of the post-hoc pairwise t-test comparisons conducted for the lipreading abilities across the three experimental groups. Younger NH = younger participants with normal hearing; Older NH = older participants with normal hearing; Older HL = older participants with hearing loss. Star symbols in the p-value column denote statistical significance (** and *** indicate significance with p < 0.01 and < 0.001 respectively).*

4.5. Summary of chapter and discussion

In the current chapter I conducted analyses of the data collected from the three experimental groups I tested: The younger participants with normal hearing, the older participants with normal hearing, and the older participants with hearing loss. In performing these analyses, I began to address two of the primary aims of my thesis, relating to: 1) Establishing that the speech-in-noise task developed herein was capable of capturing the AV benefit of individuals, and 2) providing evidence that the AV benefit measured included contributions from both lipreading, and audio-visual temporal coherence cues.

In the preceding sections, I showed that, indeed, the vCCRMn was able to capture the AV benefit of participants, with all three experimental groups exhibiting an increase in performance in the AV condition, compared to the audio-with-static image (A target) condition. Moreover, these findings were in accordance with the multitude of studies that have reported AV benefits for participants listening in noise (Bernstein & Grant, 2009; Grant & Bernstein, 2019; Sumby & Pollack, 1954; Yuan et al., 2020, 2021).

In my analyses, I also found that younger participants with normal hearing gained more AV benefit than the older groups' participants, potentially indicating a decline in audio-visual integration in the older groups (Pepper & Nuttall, 2023). Lipreading performance, measured by participant performance in the TAS sentence scores, was significantly decreased in the older participant with hearing loss group compared to the younger participant with normal hearing group, potentially also contributing to the observed decline in the AV benefit for the older group. A decrease in lipreading ability in older participants was also in line with previous findings (Tye-Murray et al., 2007), although the comparison between younger participants with normal hearing and older participants with normal hearing was not significant here.

It was not just the AV condition that conferred visual benefits to participants. Younger, and older participants with normal hearing performed significantly better in the Inter target condition of the vCCRMn task, compared to the A target condition. Since the Inter condition did not provide participants with lipreading cues of the target words, these Inter benefits are believed to underline the audio-visual temporal coherence cue-related aspects of the AV benefit, and not the linguistic. Indeed, the fact that there was no significant Inter benefit in the older participant with hearing loss experiment, but there was in the normal hearing experimental groups, suggests that the underlying mechanism measured by this benefit was related to auditory object formation. It is also possible that a lower visual acuity exhibited by older participants underlined this result (Morris et al., 2012). Although having normal/corrected to normal vision was a criterion for participation in my experiments, I did not measure the visual acuity of my participants. A test of visual acuity would be an important one to include in future experiments assessing audio-visual speech perception. Nonetheless, were significant, the Inter benefits measured may involve both attentional and/or audio-visual binding effects.

As discussed in Chapter 1, section 1.4, audio-visual binding is a bottom-up, audio-visual integration process, which is believed to enhance auditory stream segregation through the formation of audio-visual objects (Atilgan et al., 2018; Atilgan & Bizley, 2021; Bizley et al., 2016; Lee et al., 2019; Maddox et al., 2015). The audio-visual object, within the context of the vCCRMn, would be an object encompassing both the target talker's voice, and their mouth, and is believed to constitute the result of the temporal coherence between the amplitude envelope of the latter, and the magnitude of the opening and closing of the former (Chandrasekaran et al., 2009). Importantly, successful auditory segregation of the voices of the task, including target, and maskers, would be a prerequisite of successful identification of the target talker's voice (and thereby of enhanced performance in the nonlipreadable, Inter target condition). Audio-visual objects are believed to be more salient than unimodal auditory objects, leading to enhanced auditory stream segregation (Bizley et al., 2016), measured herein as an Inter benefit.

In addition to this early-stage improvement of auditory scene analysis, an audio-visual object between the target talker's voice and mouth would help the listener attend to the target talker's voice. Object-based attention theories state that attention operates on "whole-objects"; that is, if one feature of the object is attended, then the others will be enhanced as well (Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008). Participants in this experiment were asked to attend the talker's mouths during the task; thus, when a participant formed an audio-visual object, attending the mouth would also enhance their mental representation of the talker's voice, which could manifest in the Inter benefits measured in the experiments.

It might be argued however, that the Inter benefit did not involve the formation of an audio-visual object. For one thing, Maddox and colleagues, who had initially established the foundations for isolating audio-visual binding through temporal coherence in Maddox et al., 2015 (see also section 1.4.4.2), reported recently that temporal coherence is not an important factor in the binding process (Cappelloni et al., 2023). Arguably, however, it is a problematic experimental design that led to their results. Cappelloni et al., 2023 presented participants with two competing sounds, each alternating between two artificial vowels (e.g. alternating [u] and [a] vowels: [uauaua…]), one having the pitch of a female talker and the other of a male talker. At the same time, their participants were presented with a photo of one of the talkers, where the mouth of the talker was computer-edited to open and close, following the amplitude envelope of one of the two artificial vowel streams. The participants' task was to detect brief pitch modulations in the target stream, and the authors reported that temporal coherence of the target stream with the target talker's mouth did not provide participants with a benefit. On the other hand, the gender of the talker's photo displayed was enough to help listeners select the correct audio stream and enhance performance.

There are three weaknesses in the study of Cappelloni et al., (2023): First, the artificiality, or unnaturalness, of the stimuli may have impacted participant performances. Maddox et al., 2015 reported that the visual enhancements in performance attributable to temporal coherence were generally small. This was the case with the (arguably) even less "distracting" stimuli that were used in Maddox et al., 2015, compared to the ones used in Capelloni et al., 2023. Thus, the added artificiality of their current stimuli might have been enough to tip the scale in the unfavourable direction. Second, segregation, and subsequent selection between a male and a female "voice" is a relatively easy task for the auditory system to perform even without the added visual modality. It would be difficult to isolate audio-visual temporal-coherence related enhancements when the task is fully solvable with audition alone. Third, the fact that the authors were only using a single male, and a single female sound streams, and hinting to one by displaying a male, or female face on the screen, suggests that their results were confounded by top-down decision-biasing processes. Nonetheless, assuming that target-face and voice congruence, instead of temporal coherence, is the important cue for successful audio-visual binding, then in my experiments, participant performances would differ between the two static image conditions (A target, and A masker). As we saw, however, this was not the case.

Further arguments against audio-visual object formation include the fact that my study used speech sentences as stimuli, and thus could not have controlled for object feature orthogonality in the way described in section 1.4.3.1 (and Maddox et al., 2015). Thus, although the freezing of the mouth during the target word presentation in the Inter condition certainly excluded the possibility of participants lipreading the target words – as also confirmed by the lack of correlation between the measured Inter benefits and lipreading abilities – it may not have excluded other non-binding processes.

For example, as it is possible that mouth movements preceded the target talker's voice (Grant & Bernstein, 2019), attending to the mouth may have provided participants with a salient clue as to *when* the target voice's loudness would change, since the two are temporally coherent (see also Ding & Simon, 2012; Peelle & Sommers, 2015). Then regardless of whether an audio-visual object had been formed in their perception, participants would potentially be better able to track the loudness feature of the auditory object that constituted the target talker's voice. And, since attention operates on objects as a whole (Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008), them attending to the loudness would also enhance the representation of the remaining features of the target talker's voice, potentially leading to the Inter benefits observed.

On the other hand, the observed influence of coherence-type on participant performance may constitute evidence that the AV benefit captured by the vCCRMn task indeed included contributions from audio-visual binding processes. In particular, I argue that the difference in performance observed between the target-coherent, and masker-coherent Inter conditions points in the direction of audio-visual binding being a contributing factor. Object-based attention theory states that, when attention is divided between two perceptual objects, with the task-relevant object among the two, participant performance decreases, compared to when attention is focused on just the task-relevant object (Bizley et al., 2016). In my experiments, I instructed participants to look at the talker's lips, irrespective of whether the condition was target- or masker-coherent. The participants had to at the same time attend to the target sentence. Thus, if we assume that cross-sensory binding does not exist as a mechanism, it follows that in both target-coherent, and masker-coherent conditions, participants would be dividing their attention between two objects: the talker's voice and mouth. We would also then expect that performance in the two conditions would be similar. Performance was, however, greater in the target-coherent condition. This would be the result, according to binding theory, if participants were attending to a single, audio-visual object in the target-coherent condition, but were dividing their attention between two objects in the masker condition. As stated earlier, however, an attentional process without the necessity of audio-visual object formation cannot be excluded in explaining the performance differences observed here.

Continuing on the topic of coherence-type, it was also observed that participants across all three experimental groups performed substantially better in the AV target-coherent conditions compared to the AV masker-coherent conditions. It was the case too that performance in the AV masker condition was even worse than that in the A target condition. In the older participant with normal hearing, participants also showed decreased performances in the Inter masker compared to the A target. These findings show that vision not only has the capacity to provide a visual benefit, but also to impart a visual detriment, and impair the listening experience – a hypothesis that was postulated by, but was not shown in Maddox et al., 2015.

Participant lipreading ability, hereby measured as their score in the TAS sentences sub-test, was expected to be a contributor to their AV benefit (e.g. Sumby & Pollack, 1954). In line with this, it was observed that performance in the AV target conditions was significantly better than that in the Inter condition across all three experimental groups. Since the AV target condition captures both audio-visual temporal coherence effects, and lipreading, whereas the Inter target condition captures only the former, this boost in performance is believed to reflect the contributions of lipreading.

Along the same lines, I also showed that the AV benefits of younger participants with normal hearing correlated with their lipreading scores, albeit the effect was weak-to-moderate. There was a lack of correlation, however, in the respective analyses conducted for the older groups. Conversely, participant AV benefits were not correlated with their Inter benefits in the younger participant with normal hearing group but had moderate correlations with the Inter benefits in the older groups.

These inconsistencies could be attributed to lack of statistical power, or the underlying variation in the participants' AV benefits, and their respective contributing factors. As we saw, participant AV benefits, as well their performance differences due to coherence in the AV conditions seemed to vary substantially. Indeed, assessing the associations between the AV benefit and both the Inter benefit and lipreading skill proves challenging due to the influence of numerous underlying confounders. Moreover, it is difficult to evaluate their relative contributions to the AV benefit through simple, independent correlational analyses. Finally, while the independent analyses of data from the three experiments provided valuable insights into the effects of aging and, potentially, hearing loss on the measured visual benefits, they overlooked a crucial aspect: these variables are continuous. For hearing loss in particular it is clear that the division of listeners into "normal" and "with hearing loss" categories is problematic, as the "normal hearing" older group have markedly worse hearing than the "normal hearing" younger group and the hearing thresholds in the older listeners show a continuous spread of listening abilities.

To address these weaknesses, in the following chapters (5 and 6), I adopted a holistic approach by analysing the three experimental datasets collectively, pooling them together to form a robust sample of 125 participants. To supplement my analyses, I also drew upon the remaining factors that I collected during the experiments.

Chapter 5: Global analyses

5.1. Introduction

In Chapter 4, I conducted independent analyses for the data collected in each of the three experiments I performed during the participant testing period. In the current chapter, I pooled the datasets together to perform global analyses of the relationships between the different factors measured. These included analyses between speech-in-noise performances and audio-visual benefit measurements and their potential influencing factors. The latter also informed the inclusion of predictor variables in the statistical models developed in Chapter 6, for speech-in-noise performance and audio-visual benefit.

Unless stated otherwise, all analyses conducted herein include data from all 125 participants recruited in my experiments.

5.2. Results of global analyses

First, I present the results of the analyses relating to the participants' performances on the speech-in-noise task. Next, I examine lipreading performance at the sentences level (for reasons explained in Chapter 3) as measured by the Test of Adult Speechreading (TAS) and its relationship with other measured factors. Then scores on the Talker ID task are presented, followed by analyses related to hearing loss, which in turn are followed by analyses related to MoCA scores. Finally, I discuss the results of analyses associated with audio-visual benefits.

5.2.1. Speech-in-noise performances

5.2.1.1. Speech-in-noise performance varies across conditions

In Chapter 4 we saw how participant performance varied across the conditions of the vCCRMn groups. With the exception of the across-group visual benefits comparisons, analyses related to the vCCRMn conditions were conducted independently for each of the three experimental groups by independent use of three by two ANOVAs. To formally examine the effect of group on speech-in-noise performance, the current section introduces a three by two by three repeated measures ANOVA analysis, adding the main effect of Group to the independent ANOVAs used in previous sections. The results for the main effects and interactions are summarised in Table 5.1 below.

Main effects	**Degrees of Freedom**	**F statistics**	**p-values**
Visual condition	2	11.72	< 0.001 ***
Visual coherence	1	71.19	< 0.001 ***
Group	2	105.25	< 0.001 ***
Visual condition * Visual coherence	2	31.26	< 0.001 ***
Visual condition * Group	4	0.95	0.43
Visual coherence * Group	2	3.39	0.03 *
Visual condition * Visual coherence * Group	4	0.79	0.53

Table 5.1: *Three by two by three ANOVA table summary. The test included the main effects of Visual condition, Visual coherence, and experimental Group, as well as all possible two-way and three-way interactions.*

As can be seen from Table 5.1, all of Visual condition, Visual coherence, and Group significantly influence participant performance. Thus, Visual condition influences performance regardless of Visual coherence and Group, Visual coherence influences performances regardless of Visual condition and Group, and lastly, Group influences performances regardless of Visual condition and Visual coherence. Furthermore, of the two-way interactions, Visual condition by Visual coherence and Visual coherence by Group were significant. Thus, the effect of Visual condition on performance depends on Visual coherence, as we also saw in the three by two ANOVAs that were conducted independently for each group. Also, the effect of Visual coherence type, that is whether the focus is target, or masker, depends on the Group. Post-hoc pairwise comparisons were conducted to explore these results in more detail; they are included in Table 5.2 below.

Comparison type	**Condition 1**	**Condition 2**	**t-test type**	**t-statistic**	**DoF**	**Bonferroni corrected p-value**	**Cohen's d**	**Mean difference (dB)** **(Condition 1 – Condition 2)**
Related to Visual condition	AV	Inter	Paired	-3.74	124	< 0.001 ***	-0.24	-0.76
	AV	A	Paired	-4.64	124	< 0.001 ***	-0.29	-1.10
	Inter	A	Paired	-2.38	124	0.054 .	-0.15	-0.34
Related to Group	YNH	ONH	Indep.	-5.26	86	< 0.001 ***	-0.49	-1.36
	YNH	OHL	Indep.	-13.42	74	< 0.001 ***	-1.08	-4.22
	ONH	OHL	Indep.	-11.51	84	< 0.001 ***	-1.25	-2.87
Related to the Visual condition by Visual coherence interaction	AV target	A target	Paired	-11.48	124	< 0.001 ***	-1.03	-3.51
	Inter target	A target	Paired	-5.14	124	< 0.001 ***	-0.46	-1.09
	AV masker	Inter target	Paired	9.90	124	< 0.001 ***	0.89	2.31
	AV masker	A target	Paired	5.72	124	< 0.001 ***	0.51	1.22
	Inter masker	A target	Paired	1.57	124	1	0.14	0.32
	AV masker	Inter masker	Paired	4.65	124	< 0.001 ***	0.42	0.90
	AV target	AV masker	Paired	-14.01	124	< 0.001 ***	-1.25	-4.73
	Inter target	Inter masker	Paired	-6.40	124	< 0.001 ***	-0.57	-1.41
	A target	A masker	Paired	0.46	124	1	0.04	0.08
	AV target	Inter target	Paired	-8.44	124	< 0.001 ***	-0.75	-2.42
	AV target	Inter masker	Paired	-12.74	124	< 0.001 ***	-1.14	-3.83
	AV target	A masker	Paired	-11.49	124	< 0.001 ***	-1.03	-3.43
	AV masker	A masker	Paired	6.56	124	< 0.001 ***	0.59	1.31
	Inter target	A masker	Paired	-5.03	124	< 0.001 ***	-0.45	-1.01
	Inter masker	A masker	Paired	2.31	124	0.34	0.21	0.40
Related to the Visual coherence by Group interaction	Target (YNH)	Masker (YNH)	Paired	-7.99	38	< 0.001 ***	-0.74	-3.16
	Target (ONH)	Masker (ONH)	Paired	-6.96	48	< 0.001 ***	-0.57	-1.55
	Target (OHL)	Masker (OHL)	Paired	-4.76	36	< 0.001 ***	-0.45	-1.45
	YNH (target)	ONH (target)	Indep.	-5.53	86	< 0.001 ***	-0.73	-2.16

	YNH (target)	OHL (target)	Indep.	-10.67	74	< 0.001 ***	-1.41	-5.08
	ONH (target)	OHL (target)	Indep.	-7.99	84	< 0.001 ***	-1.07	-2.91
	YNH (masker)	ONH (masker)	Indep.	-2.19	86	0.27	-0.28	-0.55
	YNH (masker)	OHL (masker)	Indep.	-9.78	74	< 0.001 ***	-1.31	-3.37
	ONH (masker)	OHL (masker)	Indep.	-8.98	84	< 0.001 ***	-1.20	-2.82

Table 5.2**: Summary of post-hoc pairwise t-test comparisons conducted following the three by two by three ANOVA test. Cohen's d and mean differences between the conditions being compared are also included as measures of effect size. The type of t-test used for each comparison is indicated in the column 't-test type'. Star symbols in the p-value column denote statistical significance (indicates significance with p < 0.001). YNH = younger participants with normal hearing; ONH = older participants with normal hearing; OHL = older participants with hearing loss.*

Based on the post-hoc tests of Table 5.2, we see that, overall, performance in the AV condition was better than the Inter (t(124) = -3.74, p < 0.001) and the A (t(124) = -4.64, p < 0.001) conditions. We also see that all group performances differed from each other with younger normal hearing participants performing the best, followed by older normal hearing participants and finally older participants with hearing loss (YNH vs ONH: t(86) = -5.26, p < 0.001; YNH vs OHL: t(74) = -13.42, p < 0.001; ONH vs OHL: t(84) = -11.51, p < 0.001).

From post-hoc tests related to the Visual condition by Visual coherence interaction, we see similar results to those that were previously observed in Chapter 4; for example, performance in the AV target condition is greater than that in the A target condition (t(124) = -11.48, p < 0.001) and greater than that in the Inter target condition (t(124) = -8.44, p < 0.001). Performance in the Inter target condition is also greater than that in the A target condition (t(124) = -5.14, p < 0.001). As was shown previously, performance was greater in the target video conditions (AV target vs AV masker: t(124) = -14.01, p < 0.001; Inter target vs Inter masker: t(124) = -6.40, p < 0.001), but not for the static image conditions (A target vs A masker: t(124) = 0.46, p = 1). Finally, we also see the deleterious effects of AV masker compared to A target (t(124) = 5.72, p < 0.001). These comparisons are also included in the form of boxplots, in Figure 5.1 below.

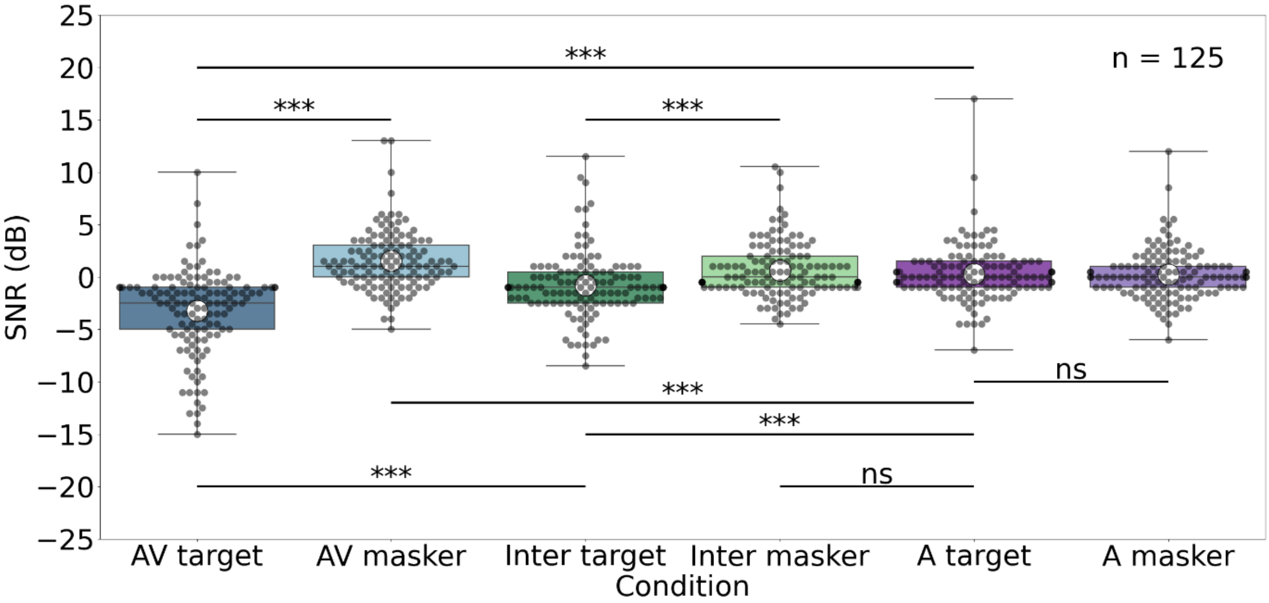


Figure 5.1**: SNR scores for each of the six conditions of the vCCRMn the speech-in-noise task, conducted on a dataset including all 125 participants. White dots depict mean performances. Pairwise comparisons are indicated with horizontal lines over the relevant conditions (denotes significance with p < 0.001 respectively; ns = non-significant).*

Since the Visual coherence type by Group interaction was significant, Table 5.2 also includes its relevant post-hoc tests. We see that within each group, performance at the target condition was overall greater than that at the masker (Target (YNH) vs Masker (YNH): t(38) = -7.99, p < 0.001; Target (ONH) vs Masker (ONH): t(48) = -6.96, p < 0.001; Target (OHL) vs Masker (OHL): t(36) = -4.76, p < 0.001). Further, we have the following across-group comparisons: The younger normal hearing group performed better than the older groups in the target conditions (YNH (target) vs ONH (target): t(86) = -5.53, p < 0.001; YNH (target) vs OHL (target): t(74) = -10.67, p < 0.001), and better than the older with hearing loss group in the masker condition (YNH (masker) vs OHL (masker): t(74) = -9.78, p < 0.001). The older normal hearing group overall performed better than the older group with hearing loss in both the target (ONH (target) vs OHL (target): t(84) = -7.99, p < 0.001) and the masker (ONH (masker) vs OHL (masker): t(84) = -8.98, p < 0.001) conditions.

Finally, since the three-way interaction Visual condition by Visual coherence by Group was not significant in the conducted tests, it can be concluded that the pattern of interaction between visual condition and visual coherence type remains similar across the three groups. That is, the motif observed in the boxplots indicating the performance across the different sub-conditions (e.g. Figure 5.1) does not vary significantly across the three groups.

5.2.1.2. Speech-in-noise performance correlates with age

Next, the relationship between speech-in-noise performance and age was assessed. A Spearman correlation was run between the mean speech-in-noise score and age, with mean speech-in-noise score computed as the mean participant performance across the six conditions of the vCCRMn task. To control for the effects of hearing loss, PTA score (specifically the PTA metric detailed in section3.6.4) was controlled for in the correlation. The correlation was marginally significant, positive, and weak in effect size (Spearman’s r = 0.17, p = 0.060), suggesting that age was negatively associated with speech-in-noise perception.

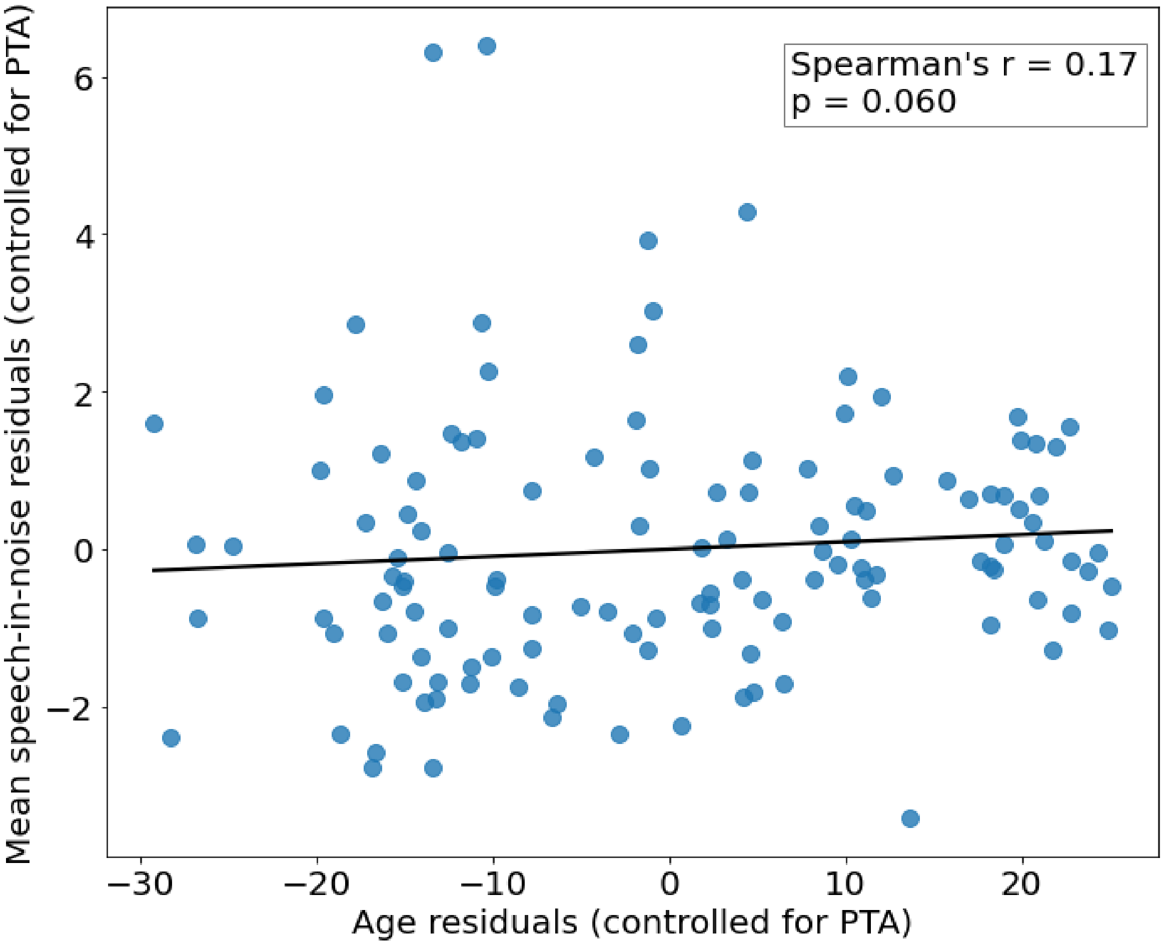


Figure 5.2: *Relationship between the residuals of mean speech-in-noise performance and age, controlling for PTA. Mean speech-in-noise performance was calculated as the mean participant performance across the six conditions of the speech-in-noise task. Partial Spearman correlation results, controlling for PTA score, are indicated at the top right corner.*

5.2.1.3. Speech-in-noise performance correlates with hearing loss

To assess whether speech perception was associated with hearing loss, further correlations were conducted between the mean speech-in-noise score, and audiogram PTA scores (Figure 5.3). The PTA scores were divided into two categories, and partial Spearman correlations, controlling for age, were computed for each independently: The PTA score low frequencies category included the mean thresholds of participants across the assessed frequencies 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz, for both ears. Similarly, the PTA score high frequencies included the mean score of the frequencies 4000 Hz, and 8000 Hz, for both ears. Both correlations were significant (Figure 5.3 (A) for low, and (B) for high frequency averages), with Spearman coefficients indicating moderately strong relationships (Spearman’s r = 0.51, p < 0.001 for low frequencies and Spearman’s r = 0.44, p < 0.001 for high frequencies). These results suggested that speech-in-noise performance was negatively associated with hearing loss, at both low, and high frequencies.

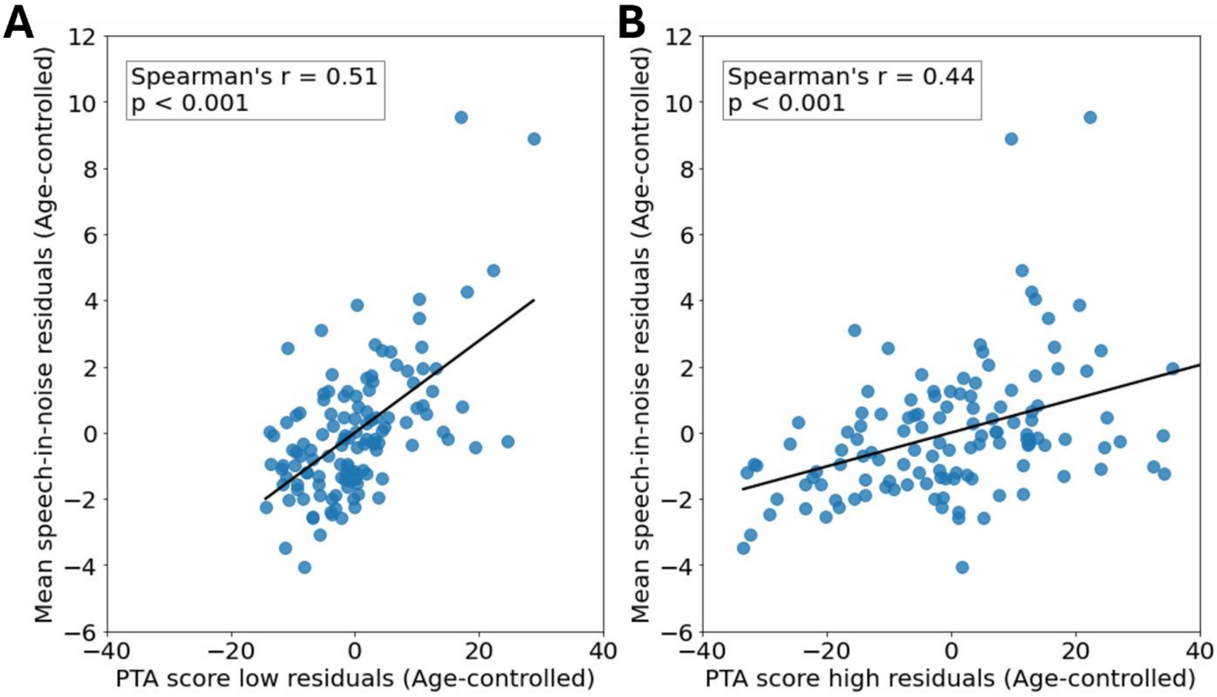


***Figure 5.3**: Relationships between the residuals (after controlling for age) of mean speech-in-noise performances and hearing loss, as measured with pure tone audiometry. The latter were divided into two categories: (A) The PTA score low frequencies category was calculated as the mean of PTA thresholds obtained from participants at frequencies 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz. (B) The PTA score high frequencies category was calculated as the mean of PTA thresholds obtained from participants at frequencies 4000 Hz and 8000 Hz. Respective partial Spearman correlation results, controlling for age, are indicated in the top left corner of each panel.*

5.2.1.4. Speech-in-noise performance is not influenced by gender

Gender has been shown to influence lipreading ability, with female participants reporting greater scores than male (Dancer et al., 1994). To assess this relationship within the context of participants' average ability to perform in speech-in-noise, I divided the participants by gender into two groups and compared the groups with a t-test. The result was not significant (t(123) = 1.48, p = 0.14), however, the mean score for the females was slightly lower than the males (Female mean = -0.5 dB, n = 71; Male mean = 0.16 dB, n = 54), indicating a potential trend for better average performance by female participants (see also Figure 5.4). Nonetheless, if indeed females are better lipreaders, but otherwise similar performers, the non-significant result would be expected, as the mean speech-in-noise performance metric across all vCCRMn conditions includes several conditions where lipreading was not possible.

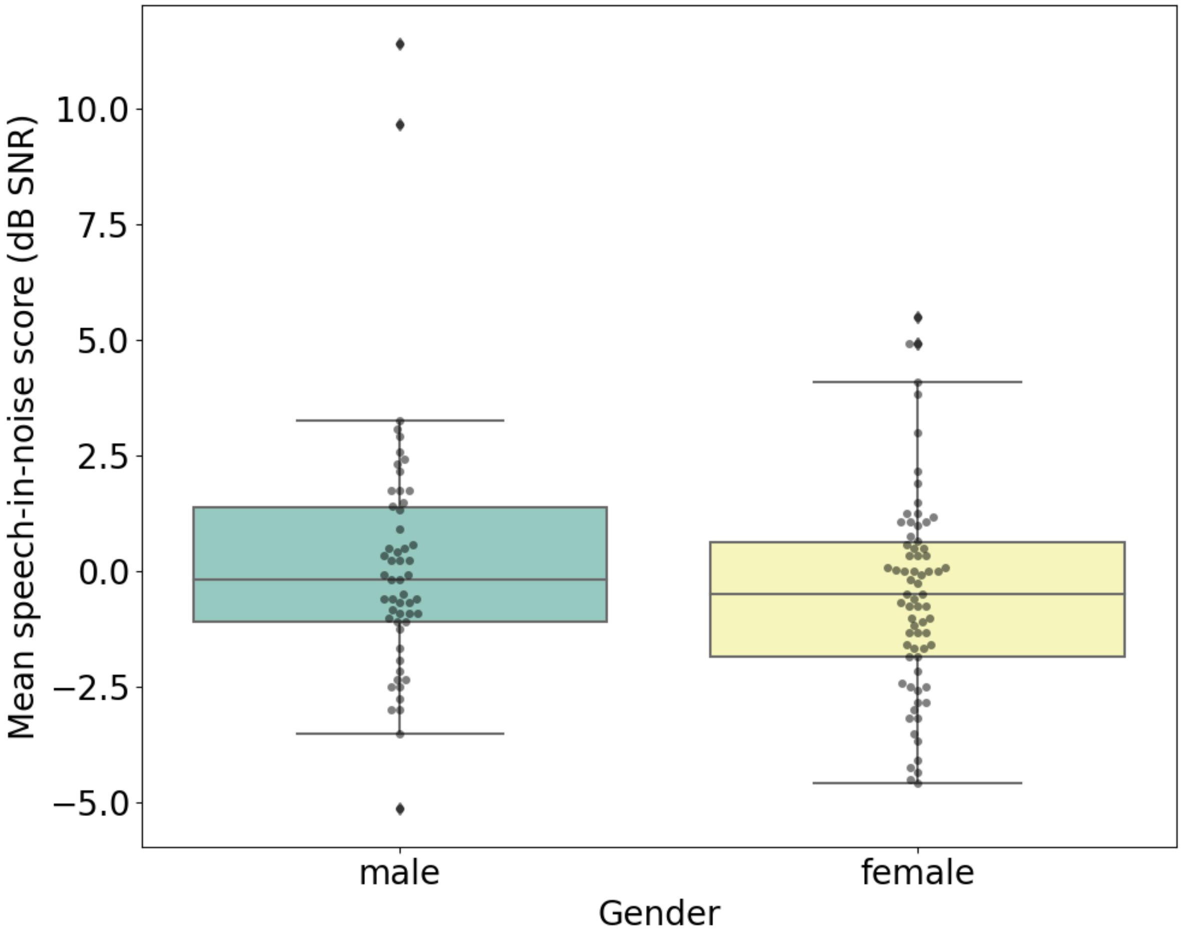


***Figure 5.4**: Mean speech-in-noise performance plotted against participant gender (male n = 54, female n = 71). A t-test was used to compare the two groups, and the result was not significant.*

5.2.1.5. Speech-in-noise performance is not influenced by vCCRMn run number

Participants completed three experimental runs for each of the six conditions of the vCCRMn task. To examine whether there was an effect of run number on speech-in-noise perception in the test, mean speech-in-noise performance was plotted against run (Figure 5.5). Looking at the medians of the three boxplots depicted in the figure, it might be said that there exists a trend in the mean performances, with speech-in-noise performance improving with run. Nonetheless, a one-way repeated measures ANOVA was run to compare the three groupings, and was not significant (F(2, 369) = 0.929, p = 0.40).

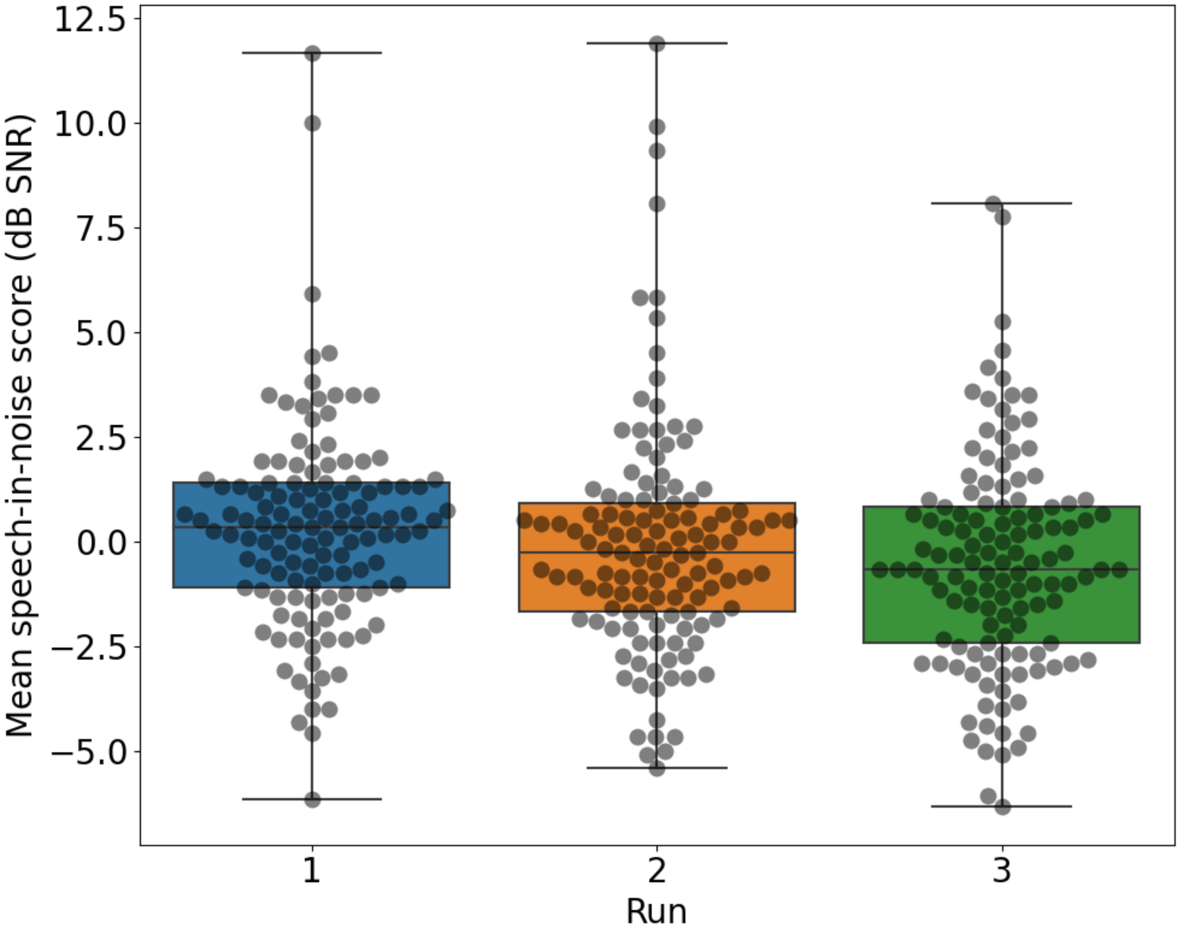


Figure 5.5*: Boxplots of mean speech-in-noise performance plotted against the experimental run number of the vCCRMn task. Each condition of the vCCRMn was run three times by each participant. To produce this plot, the mean performance of participants was computed across all conditions for each run independently. A one-way repeated measures ANOVA was run to assess the effect of run number on mean speech-in-noise scores and was not significant.*

5.2.1.6. Speech-in-noise performance correlates with lipreading performance

It was shown in Chapter 4, that lipreading ability for the younger participant with normal hearing group was positively associated with their audio-visual benefit. Lipreading ability is, in general, expected to enhance an individual's speech-in-noise comprehension. To assess this hypothesis, a partial Pearson correlation was conducted between the measures mean speech-in-noise performance and TAS sentences scores, controlling for the effects of age (Figure 5.6). The association was significant, weak-to-moderate in strength, and negative (Pearson's r = -0.39, p < 0.001).

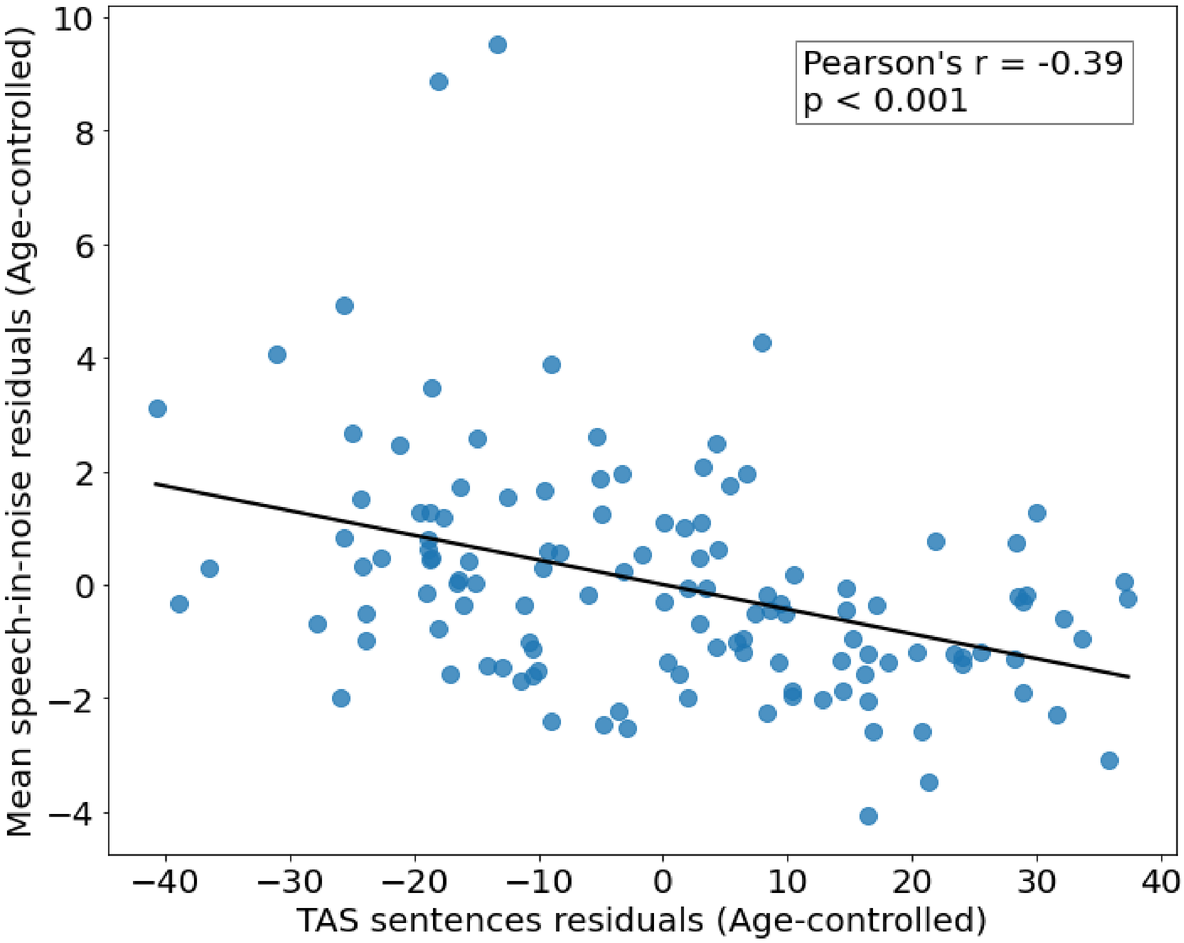


***Figure 5.6**: Illustration of the relationship between mean speech-in-noise performance and lipreading ability, controlling for age. Mean speech-in-noise scores residuals plotted against the residuals of TAS sentences scores after controlling for age. Partial Pearson correlation results are indicated in the top right corner of the figure.*

5.2.2. Lipreading performances

I proceeded with investigations of lipreading performances in the TAS sentences test and their relationships with several other factors measured in the experiments.

5.2.2.1. Lipreading performance does not correlate with age

Lipreading ability has been reported to decline with age (Sommers et al., 2005). To examine whether the data collected herein support this finding, I performed a partial Spearman correlation, controlling for PTA score (specifically the PTA metric detailed in section 3.6.4), between the TAS sentences scores and age (Figure 5.7). The analysis yielded non-significant results (Spearman's r = -0.08, p = 0.34).

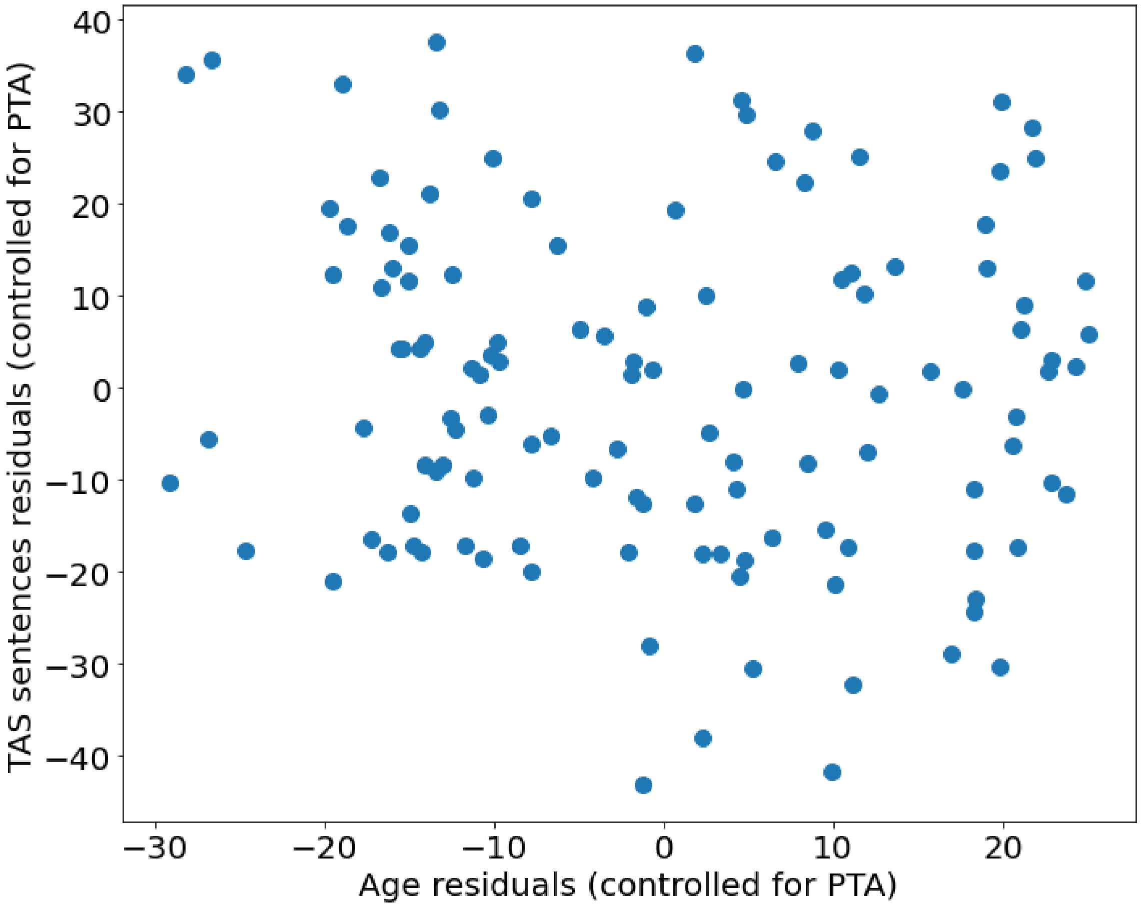


***Figure 5.7**: Illustration of the relationship between lipreading performance, and age, controlling for PTA score. TAS sentences scores residuals (after controlling for PTA) are plotted against the residuals (after controlling for PTA) of age.*

5.2.2.2. Lipreading performance correlates with hearing loss

Grant et al., (1998) reported that individuals with hearing loss exhibit substantial variation in their lipreading-related audio-visual benefits. Further, (Campbell et al., 2003) showed that hearing impaired participants displayed substantial variation in their lipreading abilities. Following these studies, to assess whether hearing loss is associated with lipreading performance, I performed partial Spearman correlations (controlling for age) between a) participants' TAS sentences scores and low frequency range mean PTA scores, and b) TAS sentences scores and high frequency range mean PTA scores (Figure 5.8). As in section 5.2.1.3, the low frequencies category included audiometric thresholds measured at 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz; the high frequencies included thresholds measured at 4000 Hz and 8000 Hz.

The association with low frequencies PTA scores yielded significant negative relations, of weak effect size (Spearman's r = -0.17, p 0.05). The association with high frequencies was not significant (Spearman's r = -0.08, p = 0.38). These results suggest that hearing loss at low frequencies might influence lipreading ability, with greater hearing loss leading to a decline in lipreading skill.

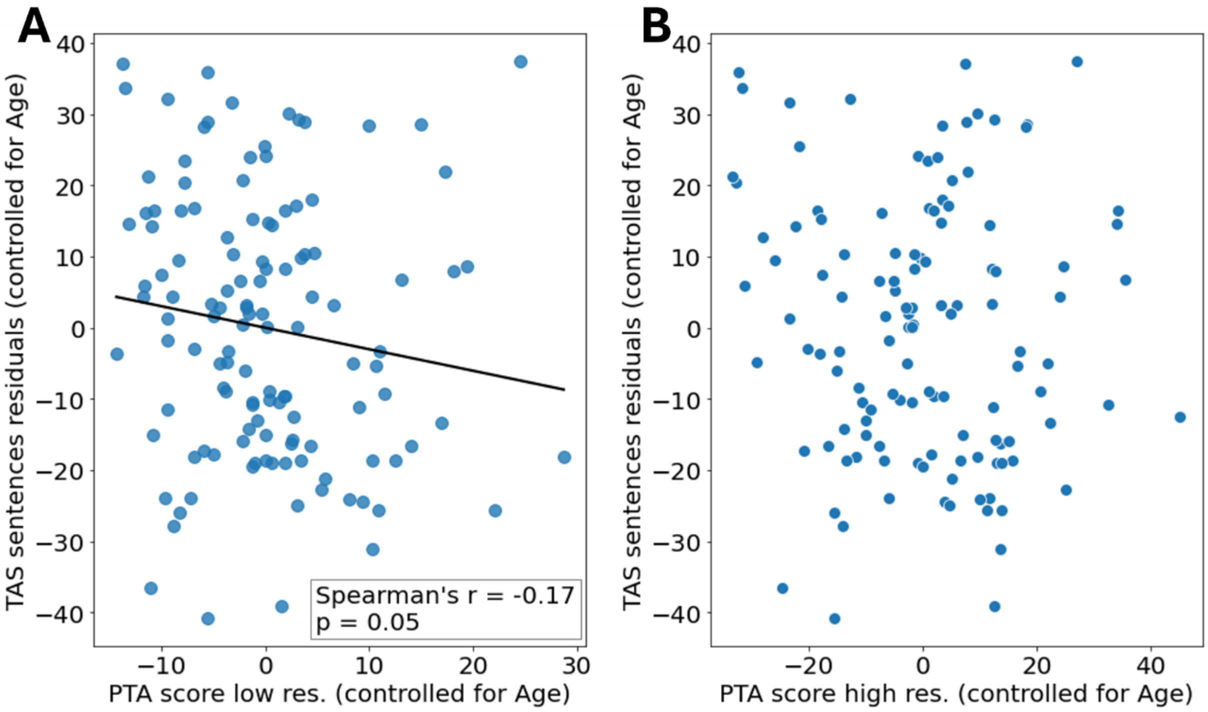


***Figure 5.8**: Illustrations of the relationship between lipreading* scores *and hearing loss, controlling for age. (A) TAS sentences scores residuals (after controlling for age) are plotted against the mean PTA thresholds residuals (after controlling for age) computed for the frequencies 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz (low). (B) TAS sentences scores residuals are plotted against the mean PTA thresholds residuals computed for the frequencies 4000 Hz and 8000 Hz (high). Partial (controlling for age) Spearman correlation results are indicated in the bottom right corner of the left panel; the correlation was not significant for the right panel. Res. = residuals.*

5.2.2.3. Lipreading performance does not correlate with Talker ID performance

Familiarity with a conversational partner has been shown to result in higher lipreading performances (Lander & Davies, 2008). To examine this finding within my own datasets, I conducted a partial Spearman correlation between participants' TAS sentences scores and their performances at the Talker ID task, controlling for participant age (Figure 5.9). The latter measured participants' ability to match the voice of the speech-in-noise task talkers to their person, and thus served as a proxy measure of their ability to familiarise with their interlocutors. The analysis yielded non-significant results (Spearman's r = 0.15, p < 0.087).

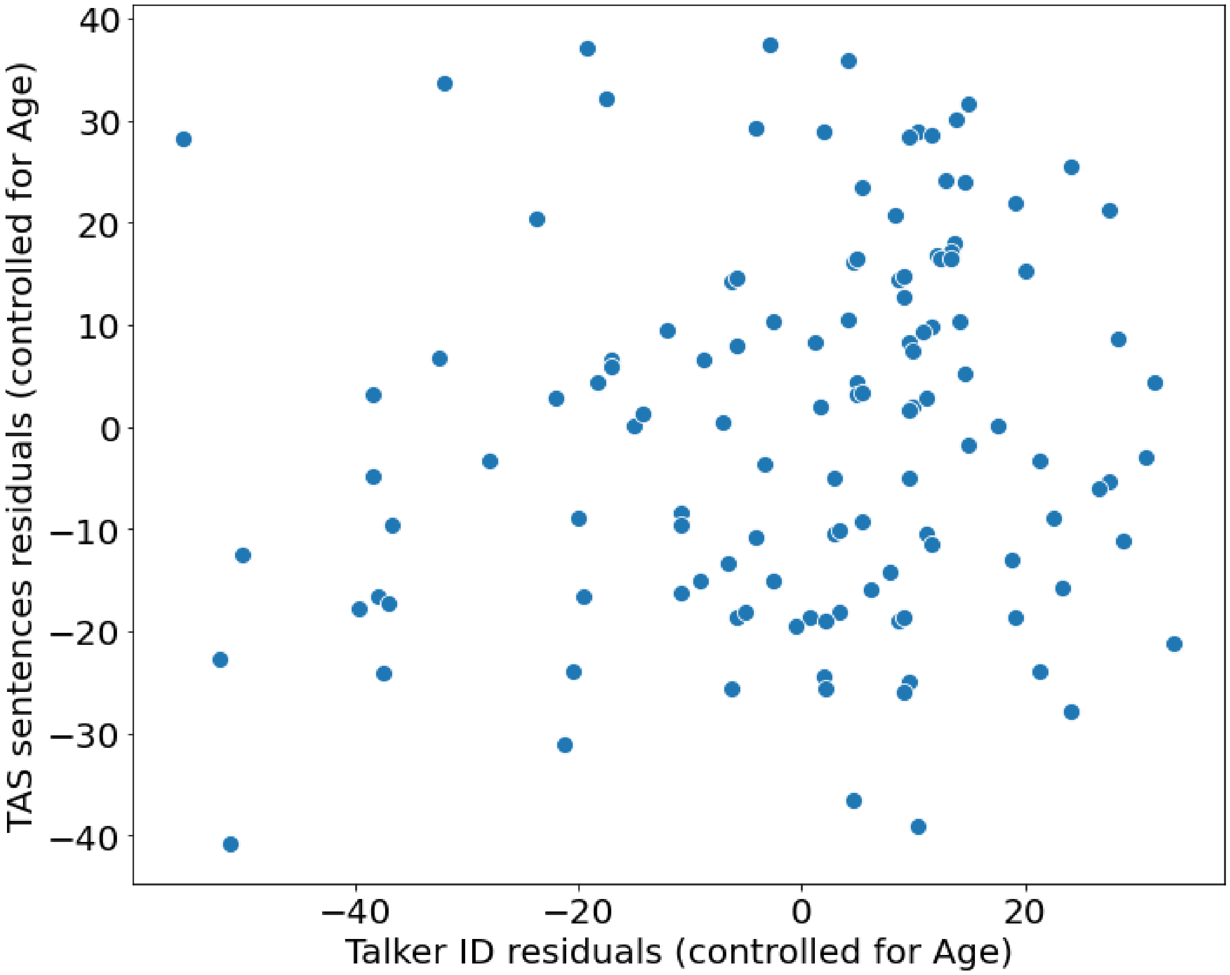


Figure 5.9*: Illustration of the relationship between TAS sentences scores and performance in the Talker ID task, controlling for participant age. TAS sentences scores residuals (after controlling for age) are plotted against the residuals of Talker ID scores (after controlling for age). Spearman correlation results are indicated in the top left corner.*

5.2.2.4. Lipreading performance is influenced by gender

The TAS sentences scores were also compared between male and female participants (Figure 5.10). The comparison was conducted via a t-test and was significant (t(123) = -2.46, p = 0.015), with female participants outperforming male ones (Female n = 74, mean = 52.79%, SD = 19.33%; Male n = 54, mean performance = 44.32%, SD = 18.15%). This result was in line with previously published findings (Dancer et al., 1994) of females being better lipreaders than males.

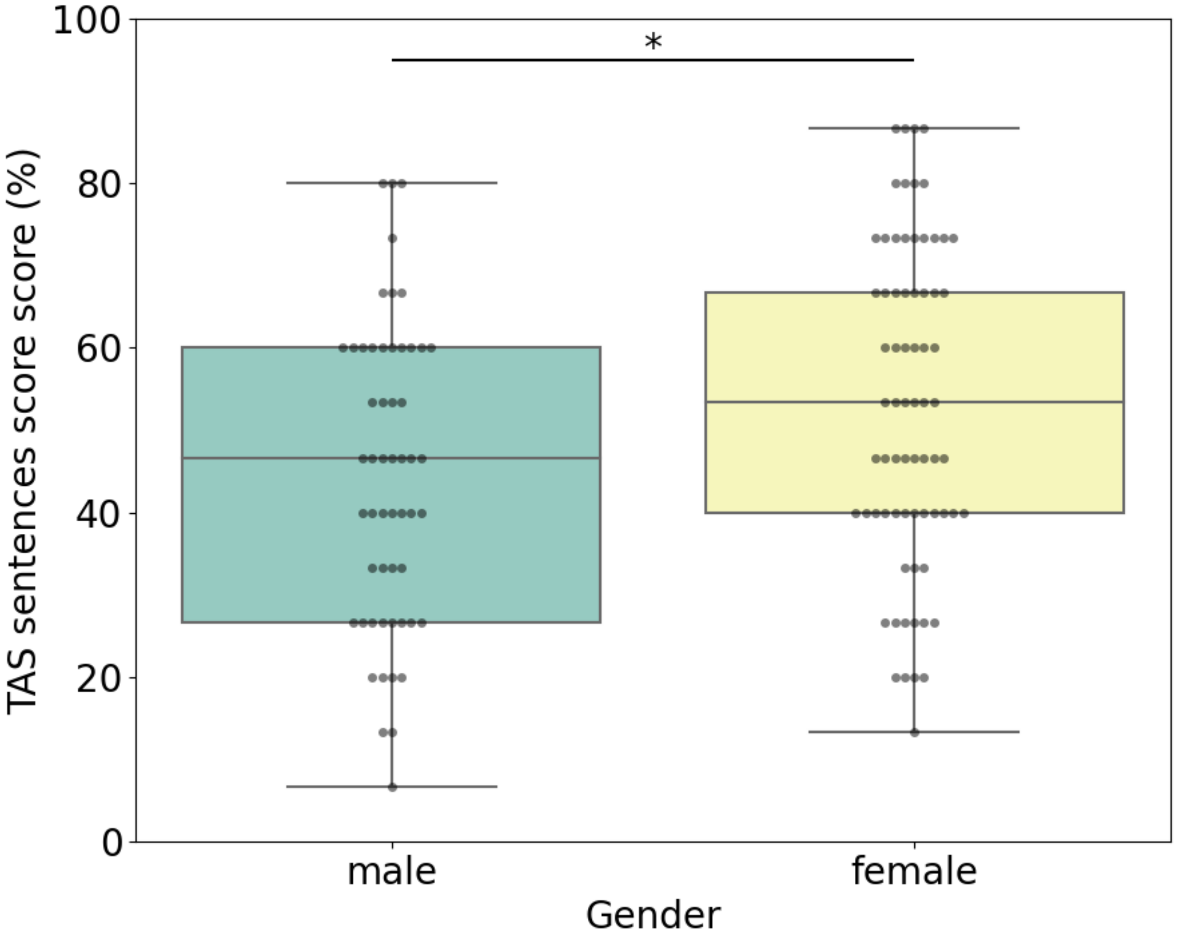


***Figure 5.10**: TAS sentences score comparison between male and female participants. The comparison was conducted via a t-test (male n = 54, female n = 71; * denotes significance with p < 0.05).*

5.2.2.5. Lipreading performance does not correlate with Inter benefit

In Chapter 4 it was shown that the Inter benefits of participants across all three experimental groups were not associated with their lipreading skill. This outcome was expected, as the possibility of lipreading the target words was removed from the Inter condition, from which the Inter benefit was derived. To see if this result persisted in the collective dataset, a partial Pearson correlation was run between the TAS sentences scores and Inter benefits of participants, controlling for participant age (Figure 5.11). As for the three experimental groups, this yielded non-significant results (Pearson's r = 0.14, p = 0.11).

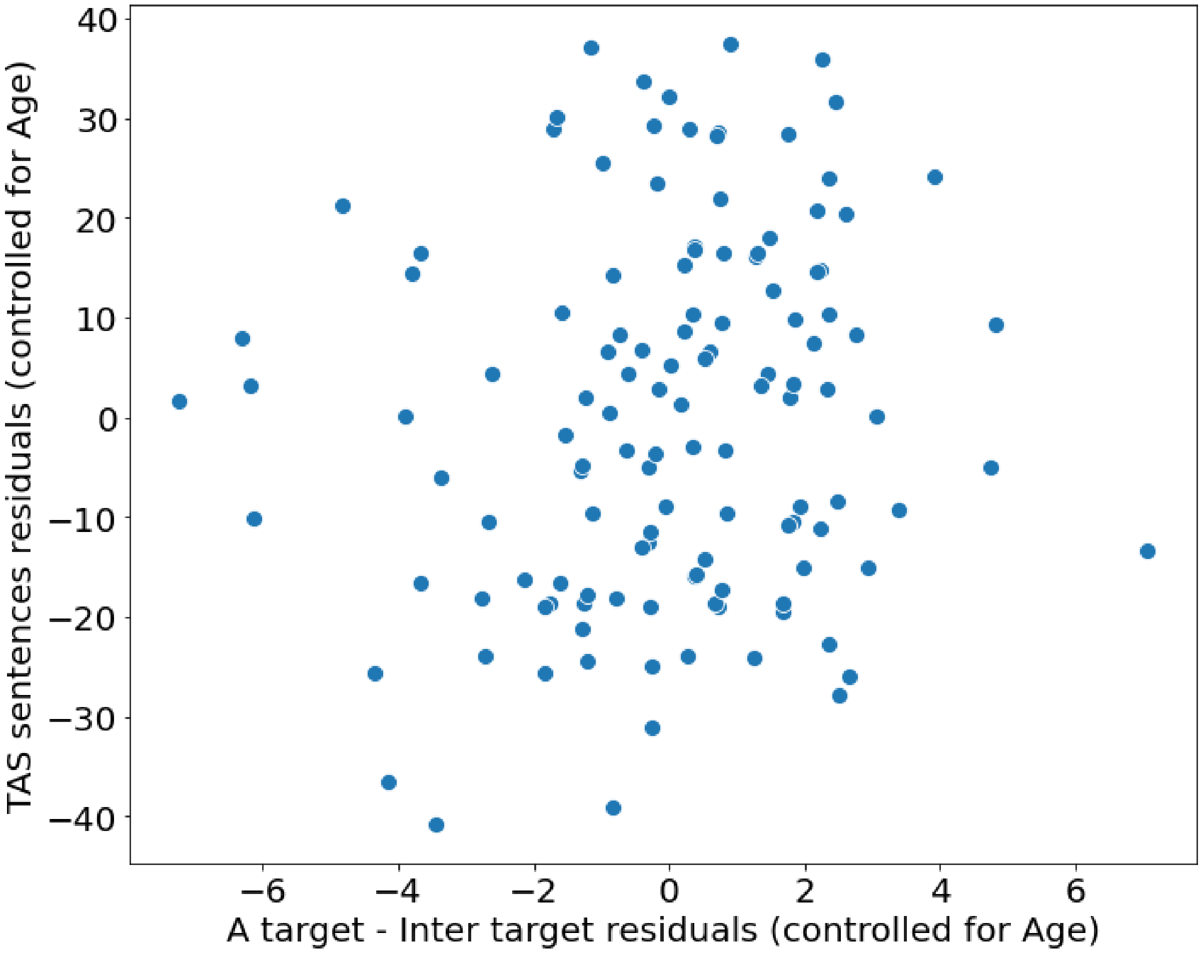


***Figure 5.11**: Illustration of the relationship between TAS sentences scores and the Inter benefits of participants, controlled for age. TAS sentences scores residuals (after controlling for age) are plotted against the residuals of Inter benefit (after controlling for age).*

5.2.3. Talker ID performances

Next, I investigated data relating to my participants' Talker ID scores (this time, with Talker ID scores as the dependent-variable). As alluded to in section 5.2.2.3, where lipreading was shown to be associated with the Talker ID metric, Talker ID scores served as a proxy measure of participants' ability to familiarise themselves with their interlocutors.

5.2.3.1. Talker ID performances overall view

As a first step, a histogram of the variable was plotted (Figure 5.12). The distribution was highly skewed towards the left, indicating that most participants had accurately learned to associate faces and voices through the course of the vCCRMn task. The median participant score is also indicated in Figure 5.12 with a vertical line at 83.33%.

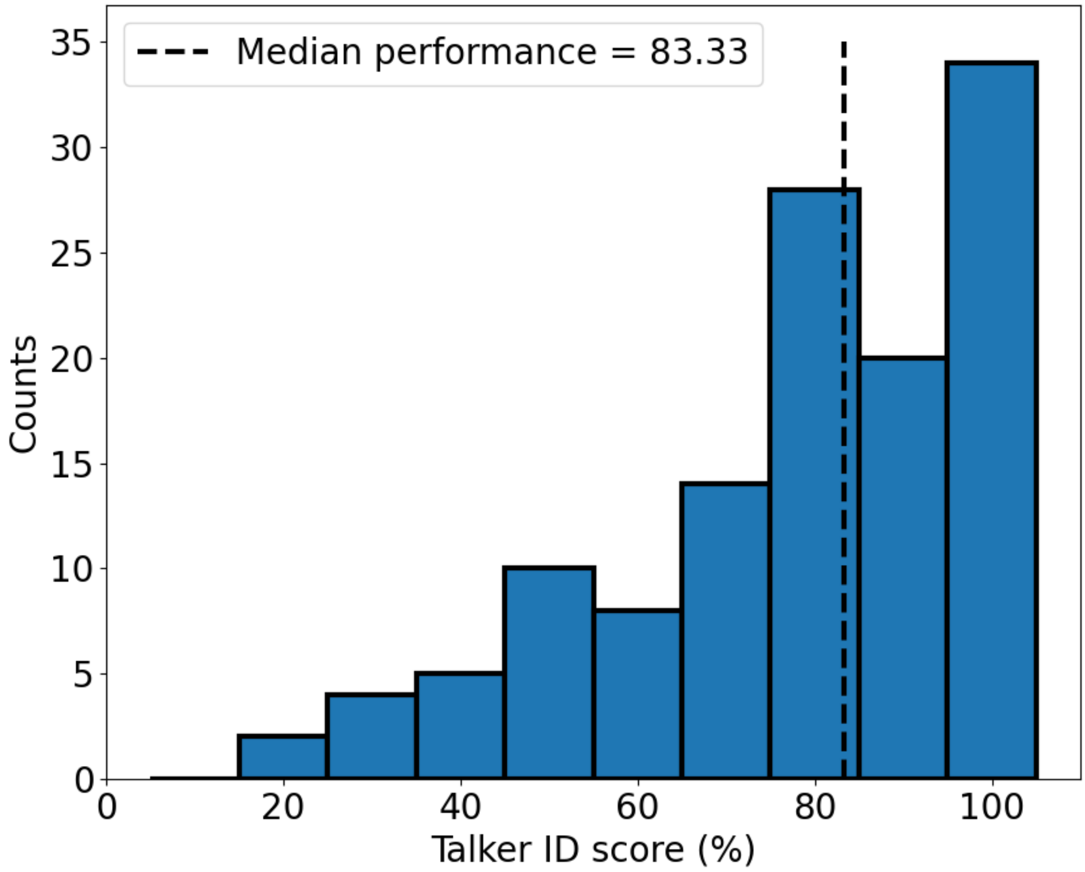


***Figure 5.12**: Distribution of Talker ID scores. The distribution is left skewed, with most participants scoring high in the task. Median performance shown with a vertical line.*

5.2.3.2. Talker ID performance correlates with age

As we have seen in previous sections, ageing was negatively associated with performances in most of the tasks participants undertook. To assess whether this effect was true for Talker ID scores as well, a partial Spearman correlation was conducted, controlling for PTA score (with PTA score as defined in section 3.6.4) (Figure 5.13). Indeed, this yielded a significant, and negative association, with moderate effect size (Spearman's r = -0.31, p < 0.001).

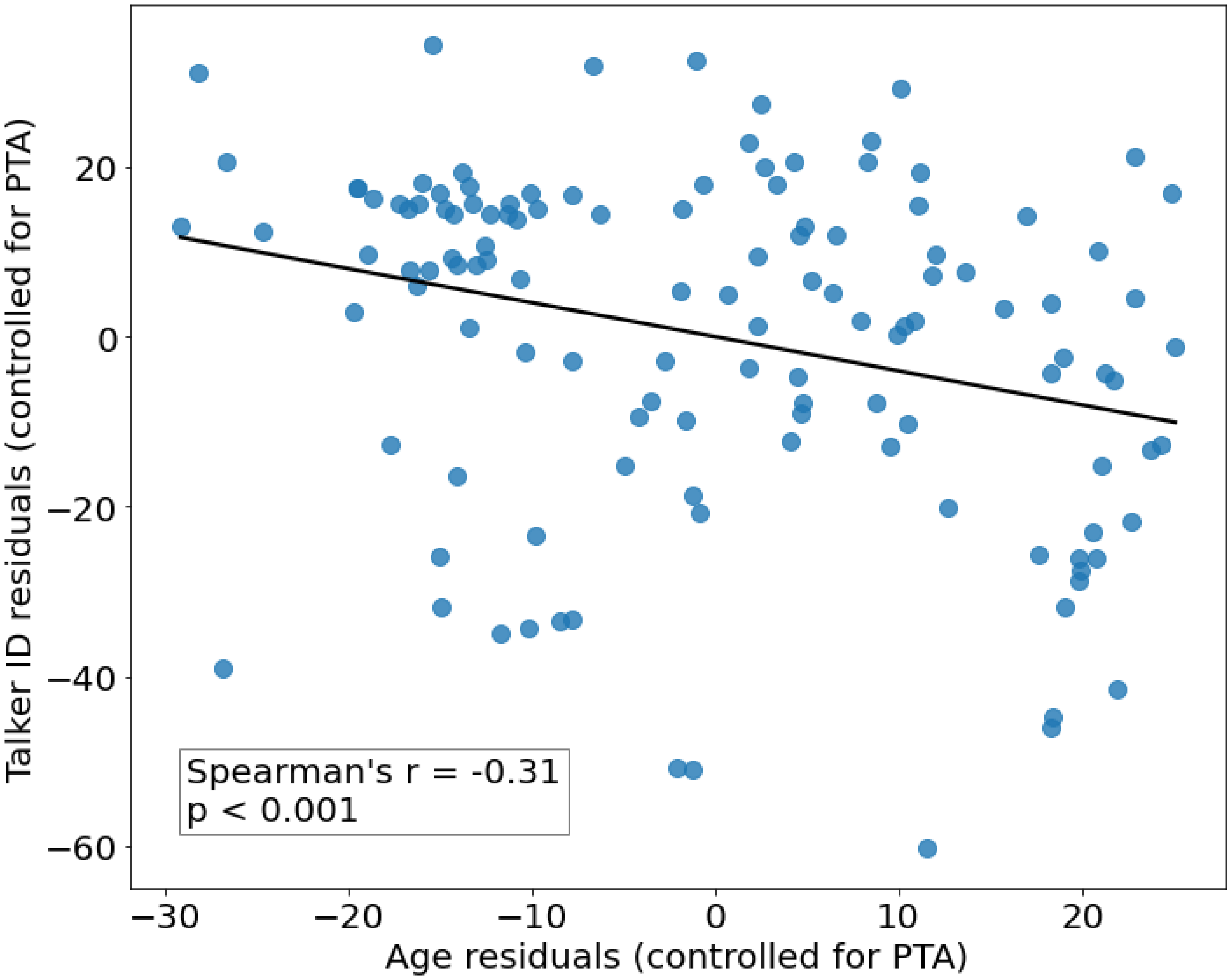


***Figure 5.13**: Illustration of the relationship between Talker ID task performance and age. Talker ID scores residuals are plotted against the residuals of age, after controlling for PTA score. Partial Spearman correlation results indicated in the bottom left corner.*

5.2.3.3. Talker ID performance is not influenced by gender

Mean lipreading scores were found to be higher in females, compared to males (section 5.2.2.4), and mean speech-in-noise performance scores showed a similar trend (albeit non-significant), with the female average slightly higher than the male (section 5.2.1.4). To examine whether this trend persisted for Talker ID scores, the data were divided once more into male and female and displayed in boxplots (Figure 5.14).

The distributions appeared to be substantially overlapping for the two groups. A Mann-Whitney U test confirmed that the comparison was not significant (Male n = 54, Female n = 71; U-statistic = 1747, p = 0.39).

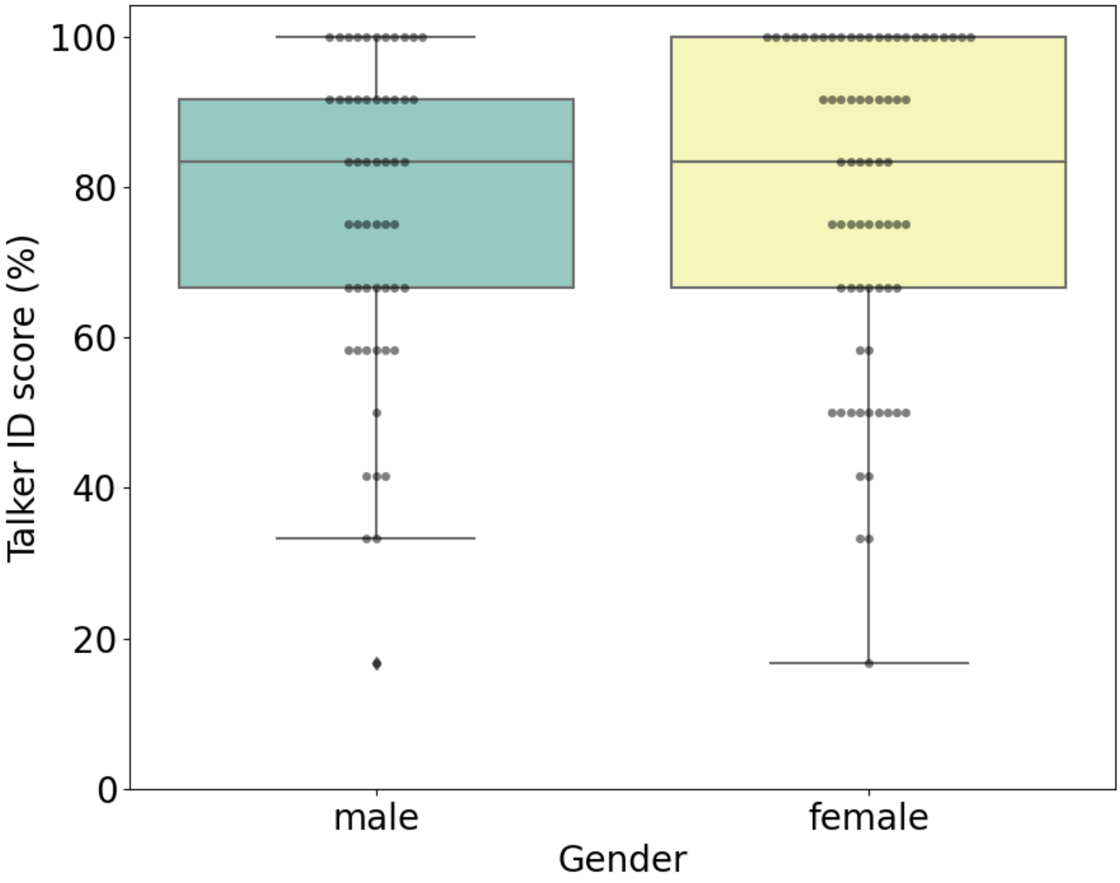


Figure 5.14*: Boxplots of Talker ID scores for male and female participants (male n = 54, female n = 71). The two groups were compared with a Mann-Whitney U test and yielded non-significant results.*

5.2.4. Hearing loss

In previous sections, I showed that hearing loss is associated with several other factors measured within the context of this thesis. These included negative influences on speech-in-noise performance, lipreading performance, and Talker ID task performance. In the current section, I switch viewpoint and examine the effect of other factors on hearing loss.

5.2.4.1. Hearing loss correlates with age

As mentioned in section 5.2.2.2, ageing is known to exert a negative effect on hearing (Jayakody et al., 2018). To confirm that this was true in the current datasets, Spearman correlations were run between mean PTA threshold over low frequencies (250 Hz, 500 Hz, 1000 Hz, 2000 Hz) as well as mean PTA threshold over high frequencies (4000 Hz, 8000 Hz), with participant age (Figure 5.15). Indeed, both analyses revealed positive, and strong associations (Low frequencies: Spearman's r = 0.63, p < 0.001; High frequencies: Spearman's r = 0.80, p < 0.001), confirming the known fact that ageing is associated with increases in hearing loss.

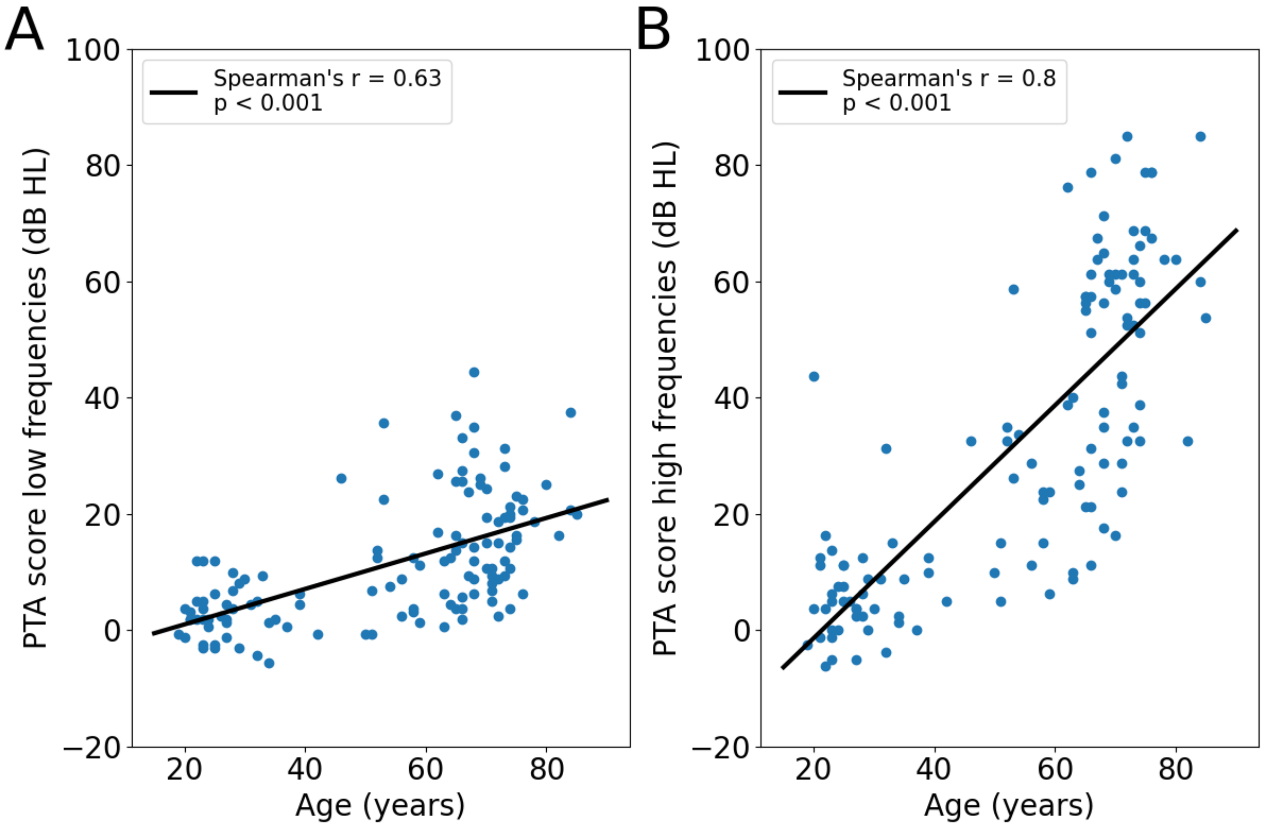


***Figure 5.15**: Relationships between hearing loss and age. (A) Mean PTA thresholds for low frequencies (250 Hz, 500 Hz, 1000 Hz, 2000 Hz), were plotted against age. (B) Mean PTA thresholds for high frequencies (4000 Hz, 8000 Hz) were plotted against age. Respective Spearman correlation analyses are indicated at the top left corner of each panel.*

5.2.4.2. Hearing loss is not influenced by gender

Gender was not expected to affect hearing loss. Thus, a significant difference between male and female PTA thresholds within the datasets collected herein would likely reflect uneven sampling from the two populations. To confirm that this was not the case, mean low, and high frequency PTA threshold were plotted against participant gender (Figure 5.16), and compared with Mann-Whitney U tests. Indeed, the comparisons for both the low frequency and the high frequency scores across males and females were not significant (Male n = 54, Female n = 71; Low frequencies: U-statistic = 1825, p = 0.65; High frequencies: U-statistic = 2139, p = 0.27). This finding suggests that even sampling was achieved in the collection of data from male and female participants.

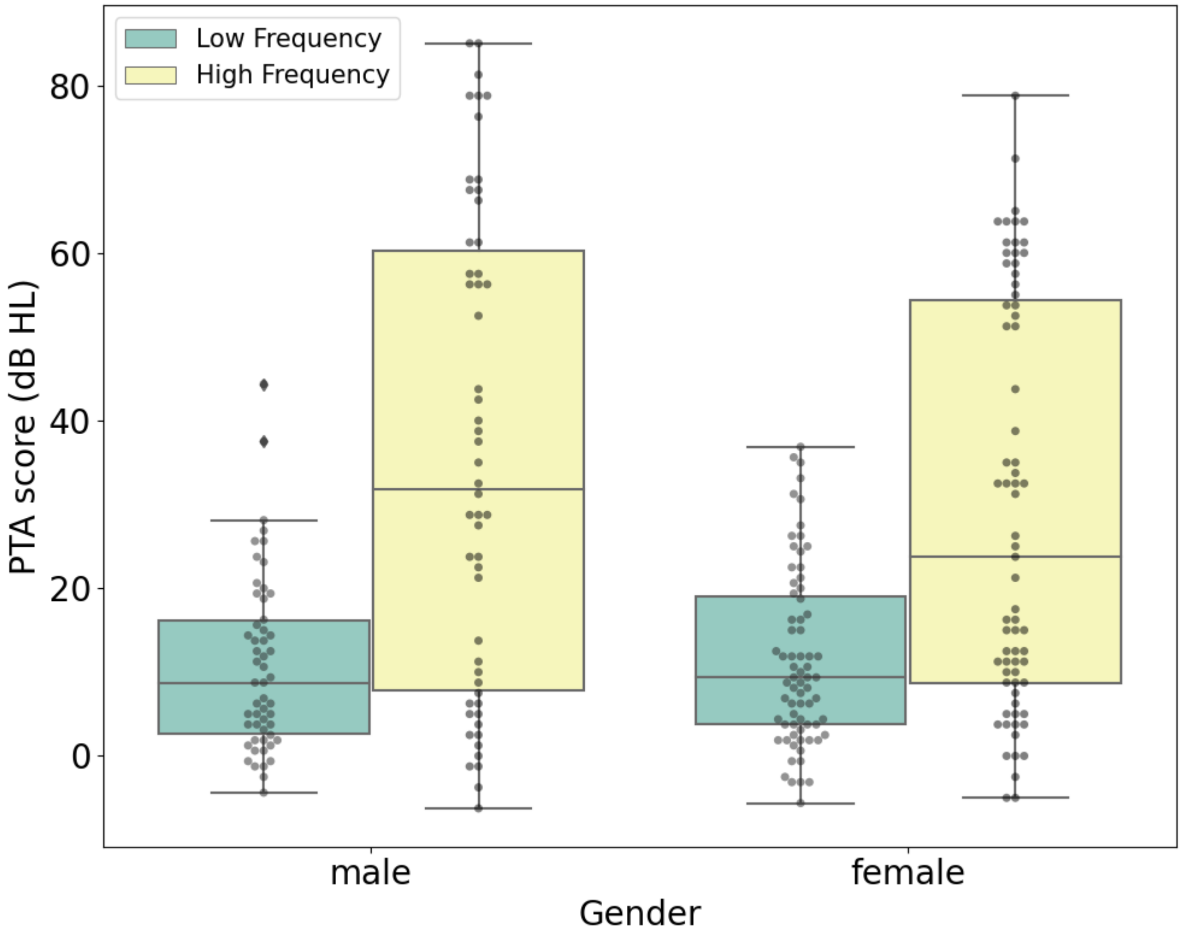


***Figure 5.16**: Boxplots of PTA scores plotted against participant gender. Mean low frequency scores (green boxplots) and mean high frequency scores (yellow boxplots) are plotted separately. Mann-Whitney U tests comparing the scores across the two genders yielded non-significant results.*

5.2.5. MoCA test scores

The Montreal Cognitive Assessment (MoCA) was used in this work for the detection of mild cognitive impairment in participants over the age of 50 (n = 81). As mentioned in section 3.7.7. of Chapter 3, the modified scoring version 'MoCA-H1' from Al-Yawer et al. (2019) was used here, considering only subtests of MoCA that were not dependent on having normal hearing, and thus controls for the effect of hearing loss on the scores. All participants tested had scores greater than 16/20 and were, thus, assumed to not have cognitive impairment.

5.2.6. Audio-visual benefits

The audio-visual (AV) benefits of participants were calculated as described in Chapter 4. Briefly, participant performance in the speech-in-noise task for the AV target condition was subtracted from their performance in the A target condition. These AV benefits formed a summary metric of the extent with which a participant benefited from the availability of visual cues, when listening in noise. In the next sections, I investigated how the measured AV benefits related to, and were influenced by, other measured variables.

5.2.6.1. Audio-visual benefit correlates with Inter benefit

As a brief reminder, similarly to the AV benefit, the Inter benefit was calculated by subtracting participant performance for the Inter target condition from their performance on the A target condition. The AV benefit was expected to correlate with Inter benefit, as the latter constitutes part of what the former is measuring. Further, such correlations were shown to be significant in Chapter 4, in both older participant experimental groups.

To assess this hypothesis using the collective data from all participants, a partial Pearson correlation was conducted between the two variables, controlling for the effects of age (Figure 5.17). Indeed, this yielded significant results of a positive, and moderate in strength association (Pearson's r = 0.40, p < 0.001).

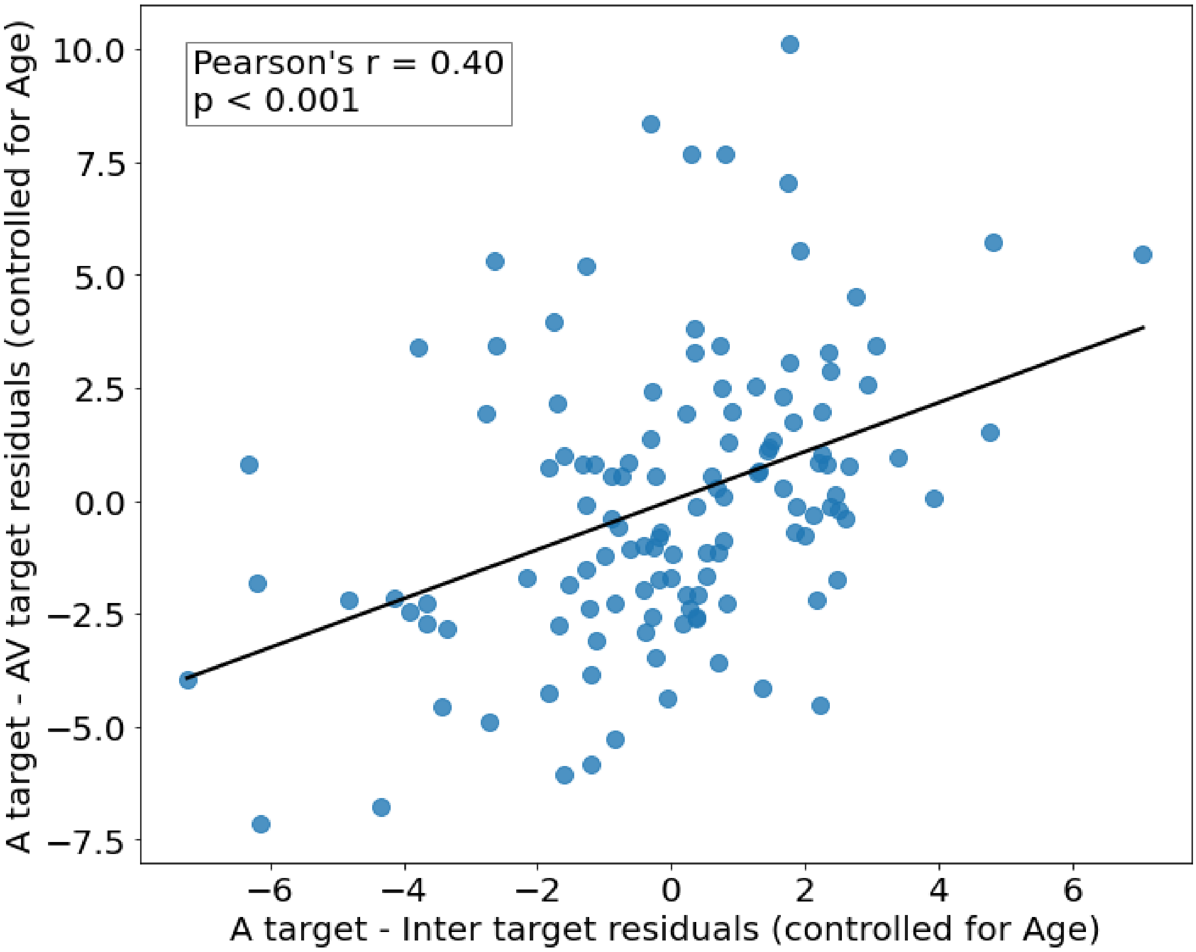


***Figure 5.17**: Illustration of the relationship between the AV benefit and the Inter benefit, controlling for age. The former was computed as the difference between the A target and the AV target conditions of the vCCRMn, whereas the latter was computed as the difference between the A target and the Inter target conditions. Here, AV benefit residuals (after controlling for age) are plotted against the residuals of Inter benefit (after controlling for age). Pearson correlation results are indicated at the top left corner of the figure.*

5.2.6.2. Audio-visual benefit is not influenced by vCCRMn run

A similar analysis to that conducted for mean speech-in-noise performance (section 5.2.1.5), was conducted in the current section to see if there was an effect of vCCRMn run number on the AV benefit of participants. Namely, a one-way repeated measures ANOVA was performed to examine the effect of run number of the AV benefits. This was no significant however (F(2,369) = 2.49, p = 0.08), and thus post-hoc pairwise comparisons were not run. The data are also displayed in Figure 5.18.

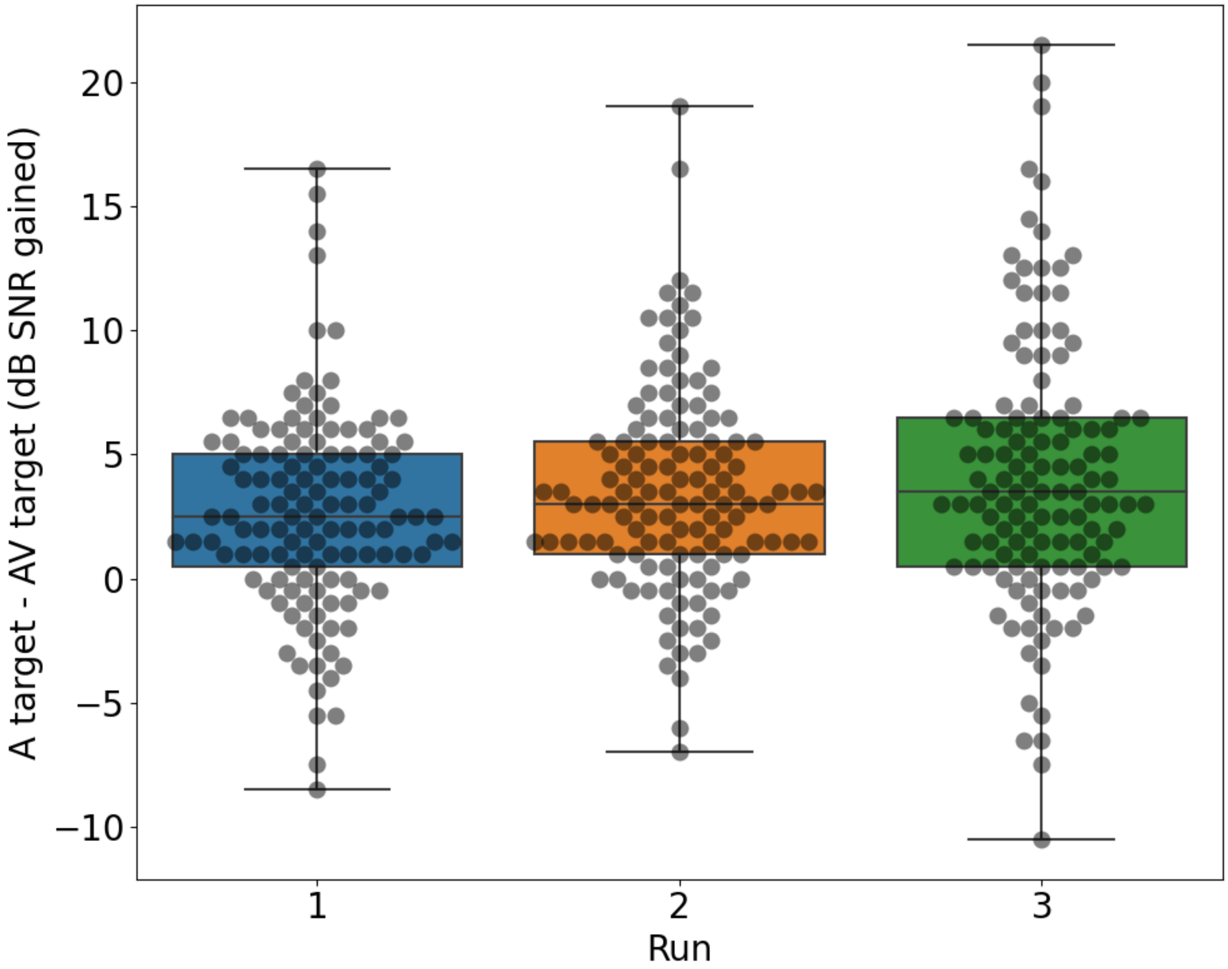


***Figure 5.18**: Boxplots of AV benefit plotted against the experimental run number of the vCCRMn task. A one-way repeated measures ANOVA was run to assess the effect of run number on the AV benefits and was not significant.*

5.2.6.3. Audio-visual benefit correlates with age

Age has been known to negatively impact audio-visual integration (Pepper & Nuttall, 2023), including lipreading contributions to it (Grant et al., 1998; Sommers et al., 2005). The AV benefit is a measure of audio-visual integration and was thus expected to decline with age. Further, the case was made in Chapter 4 that AV benefit indeed declines with age. To examine whether these findings were manifested in the collective dataset, the relationship between AV benefit and age was examined with a partial Spearman correlation, controlling for the effects of PTA (with PTA as defined in section 3.6.4) (Figure 5.19).

The analysis resulted in a significant, negative association, of weak strength (Spearman's r = -0.35, p < 0.01), suggesting that indeed, age negatively impacts AV benefit, in support of previous reports.

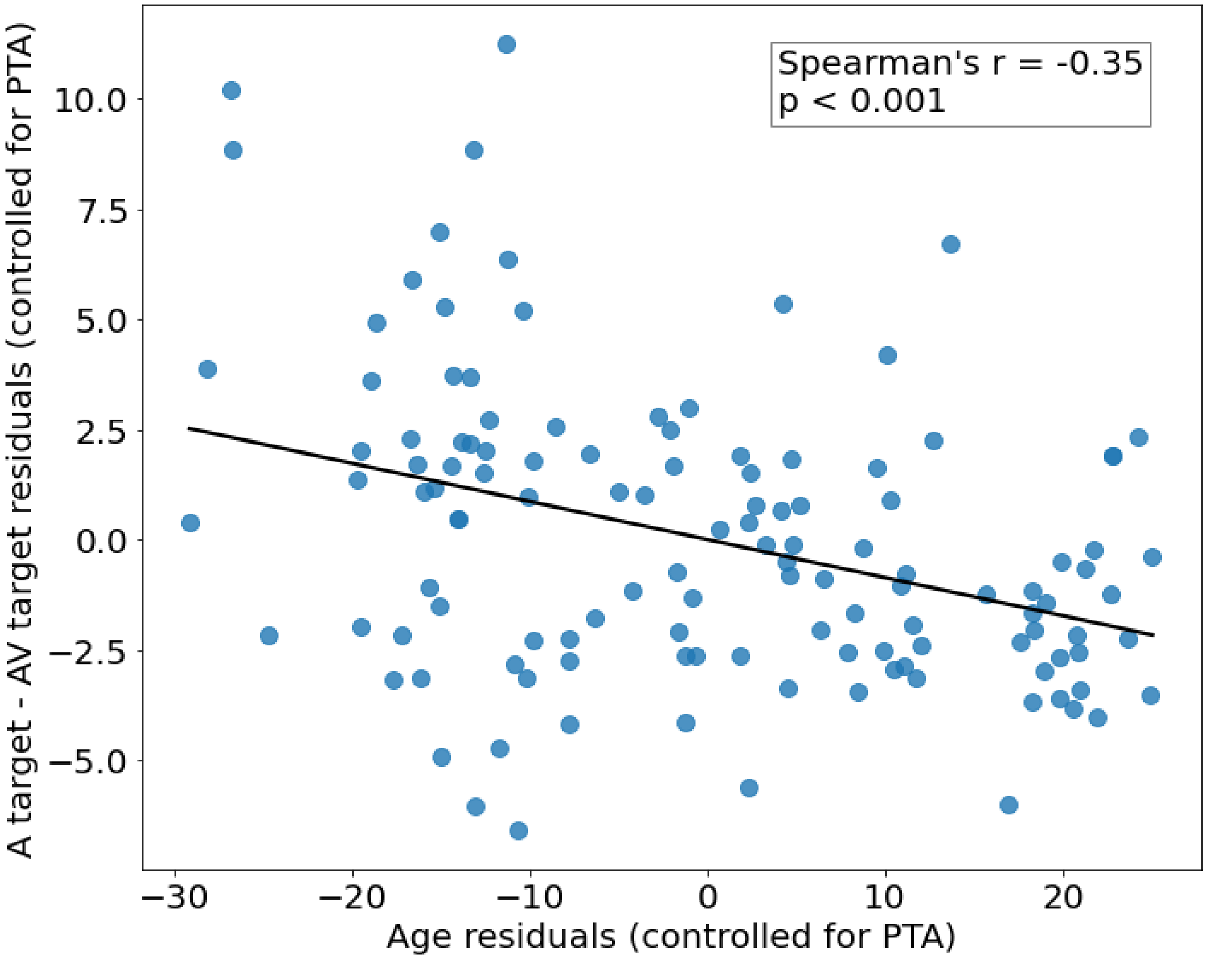


Figure 5.19*: Relationship between AV benefit and age. AV benefit residuals are plotted against the residuals of age, after controlling for PTA. Partial Spearman correlation results are indicated at the top right corner of the figure.*

5.2.6.4. Audio-visual benefit correlates with hearing loss

Next, I investigated the association between the AV benefit and hearing loss. In previous chapters, the argument was made that the AV benefit captured with the vCCRMn task included both lipreading and audio-visual temporal coherence-related contributions. Since the latter are dependent on the formation of auditory perceptual objects, which in turn would be affected by hearing status, it was expected that AV benefit would be negatively correlated with hearing loss.

Two independent partial correlations were conducted, one including the low frequency PTA scores, and another including the high frequency PTA scores (computed as described in previous sections), controlling for age (Figure 5.20). The former correlation was significant, and the latter was not (Low frequencies: Spearman's r = 0.20, p = 0.028; High frequencies: Spearman's r = 0.15, p = 0.085). These results suggest that hearing loss at the low frequencies is weakly, and positively associated with AV benefit.

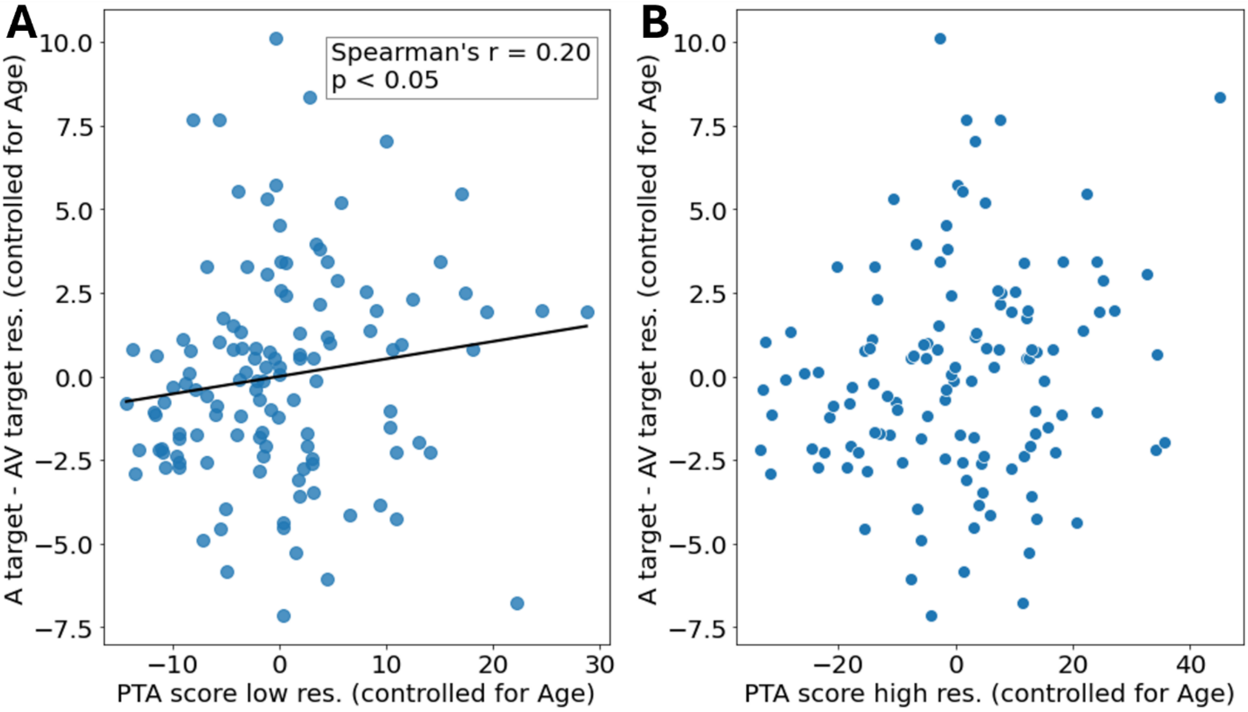


Figure 5.20*: Relationship between AV benefit and hearing loss, controlling for the effects of age. (A) AV benefit residuals (after controlling for age) plotted against mean PTA thresholds residuals of the low frequency range (250 Hz, 500 Hz, 1000 Hz, 2000 Hz). (B) AV benefit residuals plotted against mean PTA thresholds residuals of the high frequency range (4000 Hz, 8000 Hz). Partial Spearman correlations were conducted for both relationships and were significant for the association of AV benefit with low frequency thresholds (shown on the top right corner of the left panel).*

5.2.6.5. Audio-visual benefit is influenced by gender

We saw previously that gender influenced participant lipreading performances, with female participants scoring significantly higher scores than males. Since lipreading ability is a contributor to the AV benefit, it was expected that gender would impose a similar effect on AV benefit. To test this hypothesis, the data were split into male and female groups and compared with a t-test. The result was significant, with female participants obtaining, on average, higher AV benefits than male by 1.33 dB SNR (Male n = 54, Female n = 71, t(123) = -2.20, p = 0.03). The data are shown in Figure 5.21.

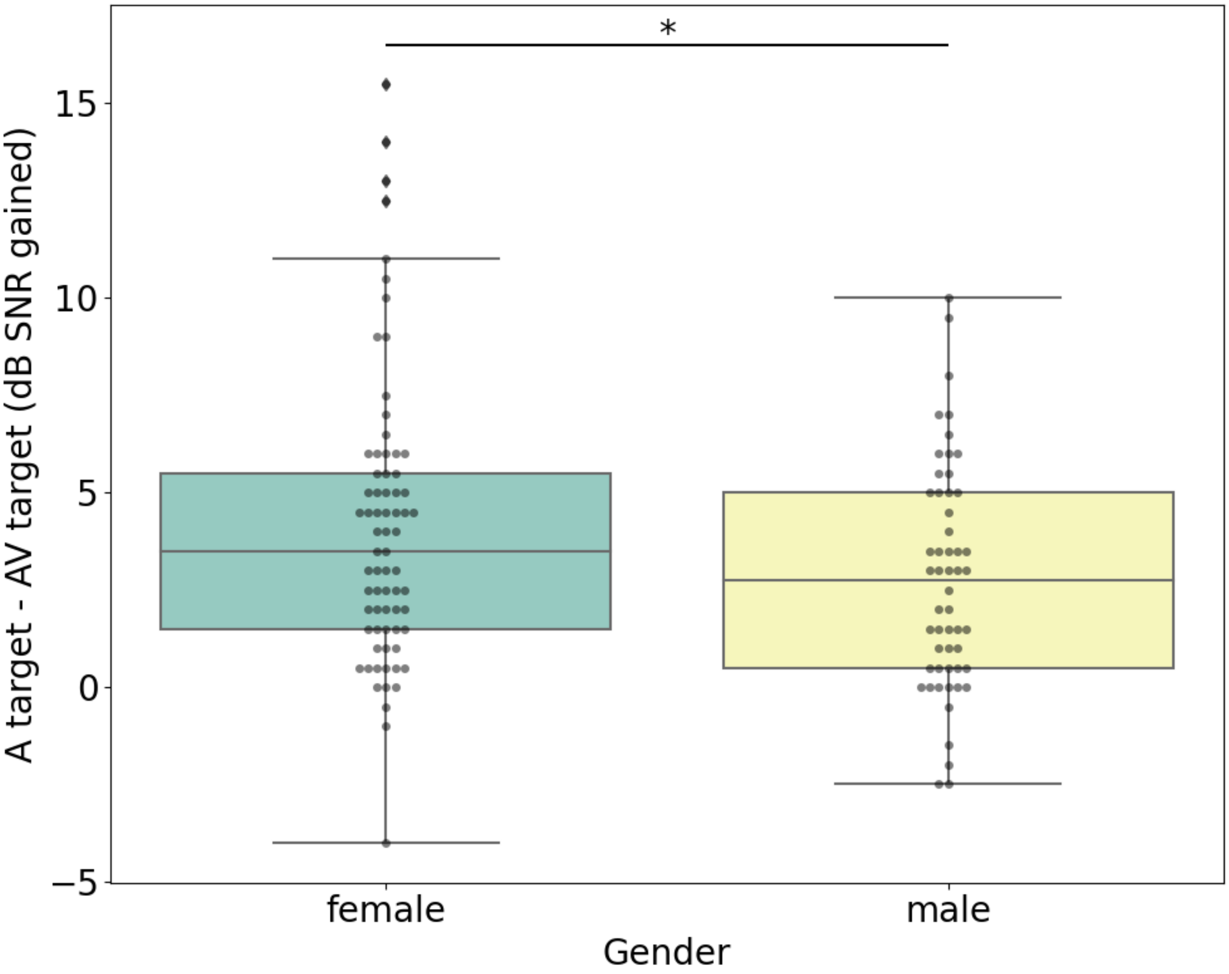


Figure 5.21*: Boxplots of AV benefits for male and female participants. The groups were compared with a t-test, and females were found to outperform males (male n = 54, female n = 71; * denotes significance with p < 0.05).*

5.2.6.6. Audio-visual benefit correlates with lipreading performance

To confirm that AV benefit was associated with mean lipreading performance (computed as described in previous sections), a partial Pearson correlation was run, controlling for age (Figure 5.22). This yielded significant results, indicating a positive, and weak association between the two variables (Pearson's r = 0.21, p = 0.021).

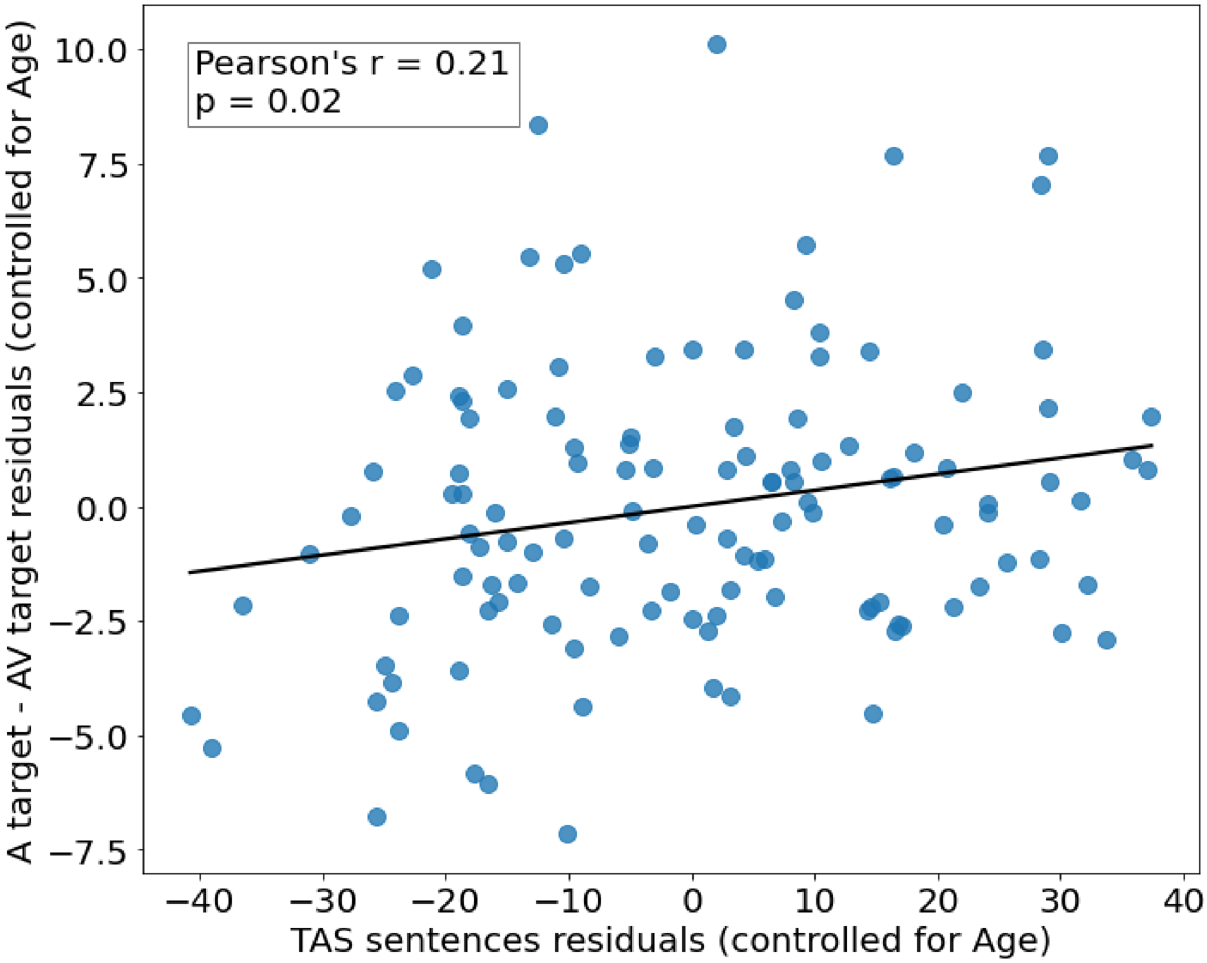


***Figure 5.22**: Illustration of the relationship between AV benefit and TAS sentences score, controlling for age. AV benefit residuals (after controlling for age) are plotted against the residuals of TAS sentences scores (after controlling for age). Partial* Pearson *correlation results are indicated at the top left corner of the figure.*

5.2.6.7. Audio-visual benefit correlates with Talker ID performance

In section 5.2.2.3, it was shown that mean lipreading participant scores were positively associated with Talker ID scores. To examine whether this finding was also true for AV benefit, a partial Spearman correlation was run, controlling for the effects of age (Figure 5.23). This yielded marginally significant, positive, and weak in strength results (Spearman's r = 0.17, p = 0.053). Thus, participants' ability to familiarise with their interlocutors seems to have, to a weak extent, positively influenced their AV benefit as well as their lipreading performances.

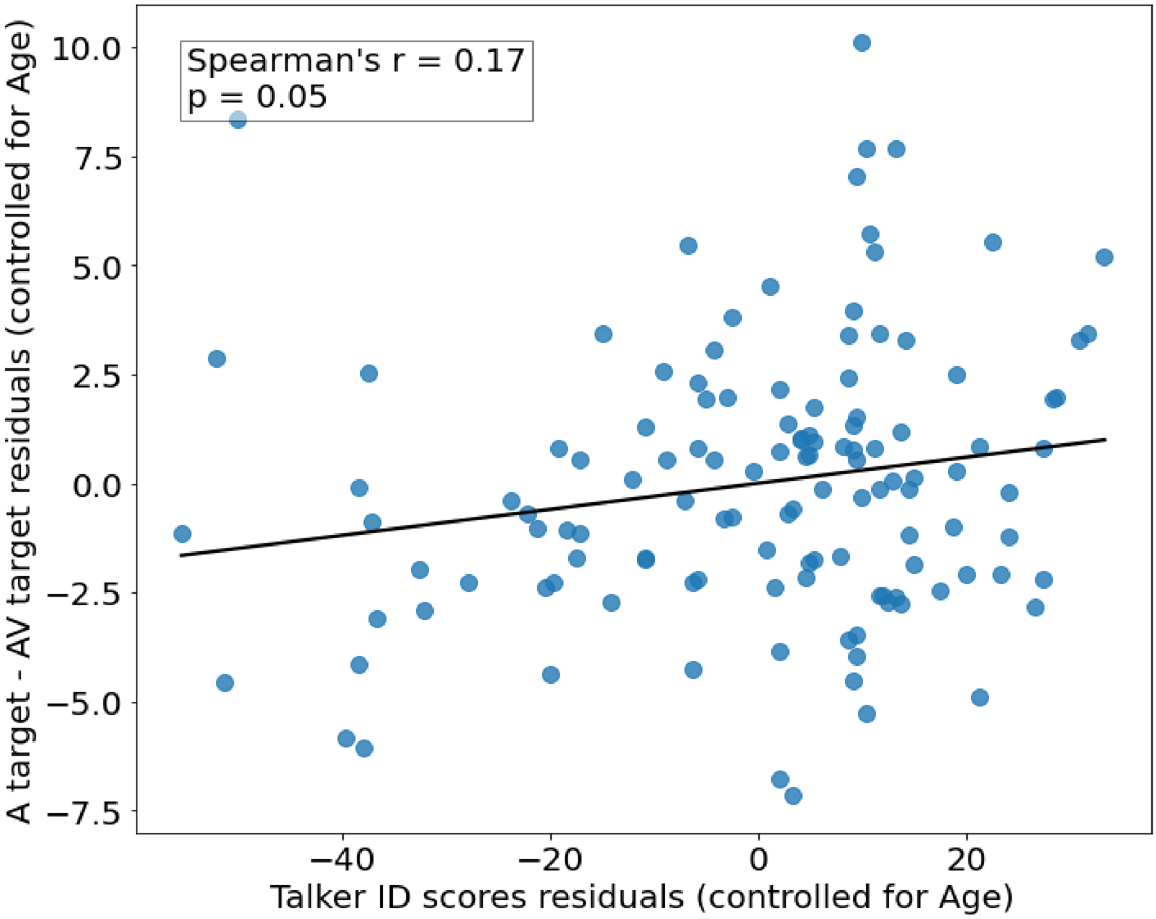


***Figure 5.23**: Illustration of the relationship between AV benefit and Talker ID score, controlling for age. AV benefit residuals (after controlling for age) are plotted against Talker ID scores residuals (after controlling for age). Partial Spearman correlation results are indicated at the top left corner of the figure.*

5.3. Summary of chapter and discussion

In the current chapter, participant data across the three experiments were pooled together to form a larger sample of 125 participants. With the collective dataset, I investigated participant performances across a range of metrics, including speech-in-noise measures, lipreading ability, Talker ID scores, hearing loss thresholds, MoCA scores, and AV benefits.

Speech-in-noise performances for the collective sample largely confirmed the findings of Chapter 4. Namely, speech-in-noise performance was here shown to vary significantly across the different conditions of the vCCRMn. Participants gained a visual benefit from the AV target, and the Inter target condition, compared to the A target, and their performance was significantly better in the target-coherent video conditions, compared to their masker-coherent counterparts. Further, performance was not different between the A target and A masker conditions, suggesting that the previously observed lack of significance was not due to a sample size restriction.

I also showed here that participant mean speech-in-noise performance was negatively associated with age, and hearing loss, and positively associated with lipreading ability. Further, although a trend existed that suggested a potential learning effect, mean speech-in-noise performance was not influenced by the vCCRMn task run number. It was also not influenced by the participants' gender.

Lipreading performance, was not correlated with age, when PTA was controlled for, and it was also not correlated with Talker ID scores, when age was controlled for. It was also not associated with Inter benefit, corroborating the findings of Chapter 4. It was, however, associated with mean hearing loss thresholds for low frequencies, and, further, female participants were better lipreaders than males (also in accordance with previous reports, e.g. Dancer et al., 1994).

Talker ID scores were, in general, high, with median performance being 83.33%. They were not influenced by participant gender, however, they were negatively correlated with age. Hearing loss measures were also negatively correlated with age, as expected (Jayakody et al., 2018), albeit only weakly. Nevertheless, the association of hearing loss with age, may have been a confounder in the correlation between lipreading and hearing loss. Participant hearing loss measures were not different between males and females, confirming that even sampling was achieved across the two populations in the experiments.

With regards to participant AV benefits, I showed that they were positively associated with both Inter benefit, and lipreading performances. This finding provides further evidence for audio-visual temporal cues and lipreading cues, respectively, contributing to the AV benefits measured. Age was only weakly associated with AV benefit (in accordance with what was concluded in Chapter 4, and with previous findings Pepper and Nuttall, 2023), as was mean PTA score at the low frequencies. The latter was positively correlated with AV benefit, suggesting that hearing loss might lead to greater reliance on visual cues and thereby greater AV benefits. These results confirmed, too, that a more complex modelling procedure, would be required to fully understand how the different measured factors influence these benefits (see Chapter 6). Finally, as with the lipreading scores, female participants received higher AV benefits than males, potentially due to the fact that they were better lipreaders.

The analyses conducted in this chapter, provided further, and more robust insights into the inter-relationships between the factors examined during participant testing. These investigations generally supported the conclusions made in Chapter 4, regarding participant speech-in-noise performances and AV benefits. They also formed the basis, and exploratory steps necessary to inform the development of the speech-in-noise and AV benefit models in the chapter that follows.

Chapter 6: Linear mixed effects modelling of speech-in-noise performance and audio-visual benefit

6.1. Introduction

The vCCRMn is a test that simulates the real-word, day-to-day environments that people find themselves in, and their hearing systems are faced with. Thus, the measured participant performances in the task served as a proxy of their general ability to listen in real-world environments. Further, and as discussed in Chapters 4 and 5, the controlled conditions of the vCCRMn, namely audio-visual and audio only, enabled the quantification of each participant's AV benefit.

Previous scientific studies have pointed to useful directions for exploring the factors that influence listening performance and the contributions of vision to it (e.g. Bernstein & Grant, 2009; Maddox et al., 2015; Sumby & Pollack, 1954; Yuan et al., 2020, 2021). However, these studies achieved varying degrees of success in providing clarity and confidence in their findings. Adding to these directions, in Chapter 4, I sought to investigate the factors contributing to the general audio-visual integration across three experimental groups: The younger participants with normal hearing, older participants with normal hearing, and older participants with hearing loss. I concluded that ageing, and potentially hearing loss negatively contributed to participant AV benefits. Then, in Chapter 5, I combined the data collected from the three experimental groups to form a collective dataset, and used it to further explore, through better-powered analyses, the associations between speech-in-noise performances and AV benefits with other factors.

Nonetheless, all my investigations have so far adopted a discrete approach, examining one factor at a time, and employed pairwise comparisons and simple correlations. This approach, while straightforward, is arguably susceptible to confounding, as observed at certain points, and lacks statistical flexibility.

To tackle this issue, the current chapter takes a robust and multifaceted approach to answering the questions of speech-in-noise performances and audio-visual benefits. Namely, two statistical models were developed here: The first was a model of the speech-in-noise performance of participants, as measured with the vCCRMn task. The second was more directly related to questions of audio-visual integration; it was a model of the AV benefit, once more, as measured from the data collected with the vCCRMn task. Thus, collectively these models provide a more robust and insightful synthesis of the hypotheses already touched on in Chapters 4 and 5.

The development of the two models comprised the fourth, and final aim I set out for in this work, and as described in Chapter 1 (section 1.6), it supplements all previous three aims. Namely, it helps establish that the vCCRMn indeed measured an AV benefit, and that both lipreading cues and audio-visual temporal coherence cues contributed to it. Further, the development of the models helps assess the relative contributions of the two. Finally, the models provide a quantitative method for assessing the influences of other factors, such as ageing, and hearing loss, on the speech-in-noise performances and AV benefits of participants.

6.2. Model development

6.2.1. Introductory comment

I exploited linear mixed effects modelling methods for the development of both models described herein. Upon starting this project, I had no experience in statistical modelling with linear mixed effects models. And, in developing the two models for this chapter, I had to first establish an understanding of their underlying mechanistic details, including their applicability to datasets, and modelling assumptions. It was my personal experience that these concepts can be involved, and that without an appreciation of the models' inner workings, the "danger" exists that the beauty and strengths of these models are lost behind statements of statistical significance.

I wanted to ensure that readers unfamiliar with the subject could appreciate not only the model output but also the modelling procedure itself. To this end, I created, for the interested reader, an appendix (Appendix B) where I provide a detailed description of the general principles of linear mixed effects modelling.

6.2.2. Brief description of the modelling procedure

As mentioned in the previous section, linear mixed effects models were deemed appropriate for use in the models constructed herein. This decision was made on the following basis (see also Appendix B): The model residuals met the modelling assumptions of linearity, homoscedasticity, and normality. Further, and importantly, linear mixed effects models supported the structure of the datasets used herein, where several datapoints were collected from each participant.

In the next paragraph I outline the systematic procedure (adapted after Zuur et al., (2009)) that I used for the development of the models:

1. Start with an overfitted model: The first step in model development was to include many predictors and random effects, to be subsequently trimmed down.

2. Gradually eliminate the random effects (in particular, the random slopes): Keeping the predictors unchanged, the next step was to trim down the random effects, removing one random slope at a time. With each random slope removal, the "full" model was compared to the "reduced" model using the Likelihood ratio test method in deciding whether to keep the removed random slope.

3. Gradually eliminate the fixed effects: Similar procedure to the one described in the point just above. Keeping the (now sorted) random effects unchanged, the next step was to trim down the number of predictors one at a time (beginning with interaction terms and following with the simpler terms) and comparing full and reduced models at each step with Likelihood ratio tests.

4. Present the final model.

In the next sections, I begin the discussion with the description of the development and results of the speech-in-noise model and conclude with the model for the audio-visual benefit.

6.2.3. Modelling participant speech-in-noise performance

In the sections to follow, I discuss the development of a model of the speech-in-noise performance of my participants. This performance, measured in signal-to-noise ratio (SNR), reflected the minimum

SNR a participant was able to perform at, within the context of the vCCRMn task. As alluded to in the introduction of this chapter, it also constituted a proxy measure of how adept participants were at listening in day-to-day, real-life, scenarios.

6.2.3.1. Model development

Table 6.1 depicts a subset of the datasheet used in the development of this model, illustrating the structure of the dataset on which the model was applied. The multitude of data points collected per participant, evident from the table, also forms part of the reason why a linear mixed effects method was appropriate for the modelling (see also Appendix B). To account for this interdependency across the within-participant datapoints, the variable "Participant" was included as the model's random intercept. Further, as can be seen, the variable "Condition" varied within each participant, and thus constituted a good candidate for inclusion in the model as a random slope.

SNR	Participant	Condition	Run	Age	Gender	TAS sentences task score
-5	1	AV target	1	32	Male	60
-8	1	AV target	2	32	Male	60
-7	1	AV target	3	32	Male	60
3.5	1	AV masker	1	32	Male	60
8	1	AV masker	2	32	Male	60
1.5	1	AV masker	3	32	Male	60
-4	1	Inter target	1	32	Male	60
0.5	1	Inter target	2	32	Male	60
-4.5	1	Inter target	3	32	Male	60
-3.5	1	Inter masker	1	32	Male	60
2.5	1	Inter masker	2	32	Male	60
1.5	1	Inter masker	3	32	Male	60
1	1	A target	1	32	Male	60
-2.5	1	A target	2	32	Male	60
-1	1	A target	3	32	Male	60
0.5	1	A masker	1	32	Male	60
0.5	1	A masker	2	32	Male	60
-2.5	1	A masker	3	32	Male	60
-4	2	AV target	1	21	Female	40

-2	2	AV target	2	21	Female	40
-2	2	AV target	3	21	Female	40

***Table 6.1**: Subset of the datasheet used in the development of the speech-in-noise performance model.*

Following the systematic procedure described in section 6.2.2, as a first step in the developmental process, the following model was fitted:

SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + (1 + Visual coherence | Participant).

Thus, SNR was modelled as a function of the fixed effects "Condition", "Age", "Low frequencies PTA score", and "TAS sentences score". The interactions "TAS sentences score" by "Age" and "Low frequencies PTA score" by "Age" were also included as the effect of lipreading and hearing loss on speech-in-noise performance could be influenced by participants' age. "Participant" was included as the random intercept and, instead of "Condition", "Visual coherence" was used as random slope (reason explained below). The choice of predictors was based on the exploratory analyses (conducted across all participants) performed in Chapter 5. Namely, it was shown there that speech-in-noise performance varied across the different conditions of the vCCRMn task (predictor "Condition"); additionally, it was correlated with age ("Age"), hearing loss ("Low frequencies PTA score"), and lipreading ability ("TAS sentences score"). Further information on the variables used in the model are provided in Table 6.2.

Variable name	**Type of variable**	**Units or categories**	**Explanation**
SNR	Continuous	dB SNR	Measure of performance in the vCCRMn task for a given condition and a given run. Lower SNR means better performance
Participant	Categorical	Participants 1 to 125	Participant IDs
Condition	Categorical	AV target, AV masker, Inter target, Inter Masker, A target, A masker	Condition of the vCCRMn task run
Visual coherence	Categorical	Target, Masker	Modification of the variable "Condition", where all "target" conditions were collectively renamed "Target", and all "masker" conditions "Masker". This

			modification proved useful in modelling random effects
Age	Numeric	Years	The age of the participant
Low frequencies PTA score	Continuous	dB HL	Mean audiogram score for frequencies 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz, across both ears
TAS sentence score	Continuous	% correct	Measure of performance in the TAS silent lipreading task, for the “sentences” level

***Table 6.2**: Descriptions of the predictors used in the model for speech-in-noise performance.*

One comment before proceeding: The variable “Visual coherence”, which was a modified version of the variable “Condition” (see Table 6.2), was included here as a random slope, instead of “Condition” itself. The reason was that “Condition” included four more levels than the “Visual coherence” (6 in total for “Condition” versus 2 for “Visual coherence”), and its larger number of parameters resulted in model parameter estimation issues.

The next step was to assess the usefulness of the inclusion of “Visual coherence” in the model with a Likelihood ratio test. Thus, the following “full” model:

SNR full including visual coherence ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant).

Was compared to the following “reduced” model:

SNR reduced ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 | Participant).

Based on the comparison, the random slope for “Visual coherence” influenced “SNR” performance ($\chi^2(2) = 15.53$, $p < 0.001$). Thus, it was decided to keep “Visual coherence” as a random slope effect. Further reductions to the model were deemed unnecessary, thereby the model SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant), comprised the final version in the developmental process.

The model also met all the necessary linear modelling assumptions. Namely, variable inflation factors (VIFs) estimated for all predictors were below 5 (Table 6.3), suggesting a non-concerning level of co-linearity among them. In addition, the residual plot for this model (shown in Figure 6.1), showed no clear violations of the criteria for linearity and homoscedasticity, and the histogram of the residuals (Figure 6.2) indicated that they were normally distributed. The model results are discussed in the next section.

Predictor	VIF
Condition	1.00
Age	3.83
Low frequencies PTA score	3.53
TAS sentences score	2.26
Low frequencies PTA score*Age	4.18
TAS sentences score*Age	3.31

Table 6.3*: VIF output for the model SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant).*

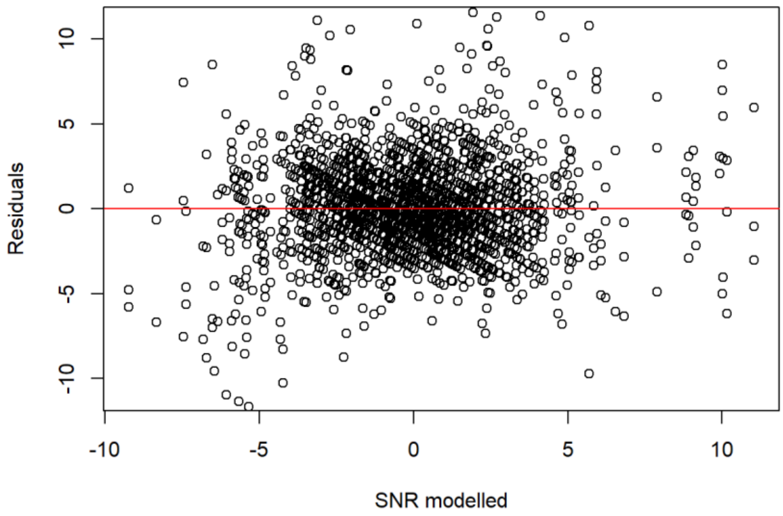


Figure 6.1*: Residual plot for the model SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant). Based on this plot, it was concluded that the modelling assumptions of linearity and homoscedasticity were met.*

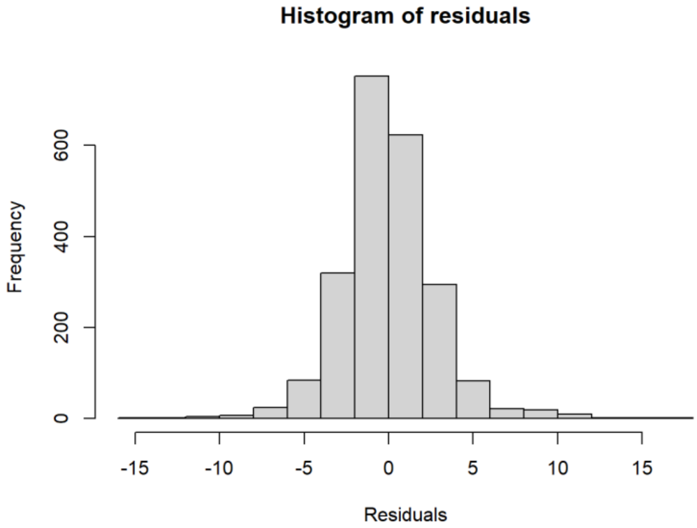


***Figure 6.2**: The distribution of residuals for the model SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant). Based on this plot, it was concluded that the modelling assumption of normality of residuals was met.*

6.2.3.2. Model results

In the previous section, the model SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant), was developed for participant speech-in-noise performance. The model output is summarised in Table 6.4 and visualised in Figure 6.3.

Term	**Estimate**	**Std error**	**t-statistic**	**DoF**	**p-value**
Intercept	1.99	0.92	2.172	124.74	0.0318 *
Condition A masker	0.1400	0.2321	0.603	565.0890	0.5466
Condition AV masker	1.2587	0.2321	5.423	565.0890	< 0.001 ***
Condition AV target	-3.4600	0.2112	-16.379	1996.0003	< 0.001 ***
Condition Inter masker	0.3840	0.2321	1.654	565.0885	0.0986
Condition Inter target	-0.9413	0.2112	-4.456	1996.0003	< 0.001 ***
Age	-0.0443	0.0238	-1.860	118.9979	0.0700
Low frequencies PTA score	0.0187	0.0429	0.436	118.9979	0.6639
TAS sentences score	-0.0665	0.0149	-4.450	118.9979	< 0.001 ***
Low frequencies PTA score*Age	0.0027	0.0009	2.278	118.9979	< 0.01 **
TAS sentences score*Age	0.0010	0.0003	2.637	118.9979	< 0.01 **

***Table 6.4**: Output for the model SNR ~ Condition + Age + Low frequencies PTA score + TAS sentences score + Low frequencies PTA score*Age + TAS sentences score*Age + (1 + Visual coherence | Participant). The model's R^2 was 0.492. Star symbols in the p-value column denote statistical*

significance (, ** and *** indicate statistical significance with p < 0.05, < 0.01, and < 0.001 respectively).*

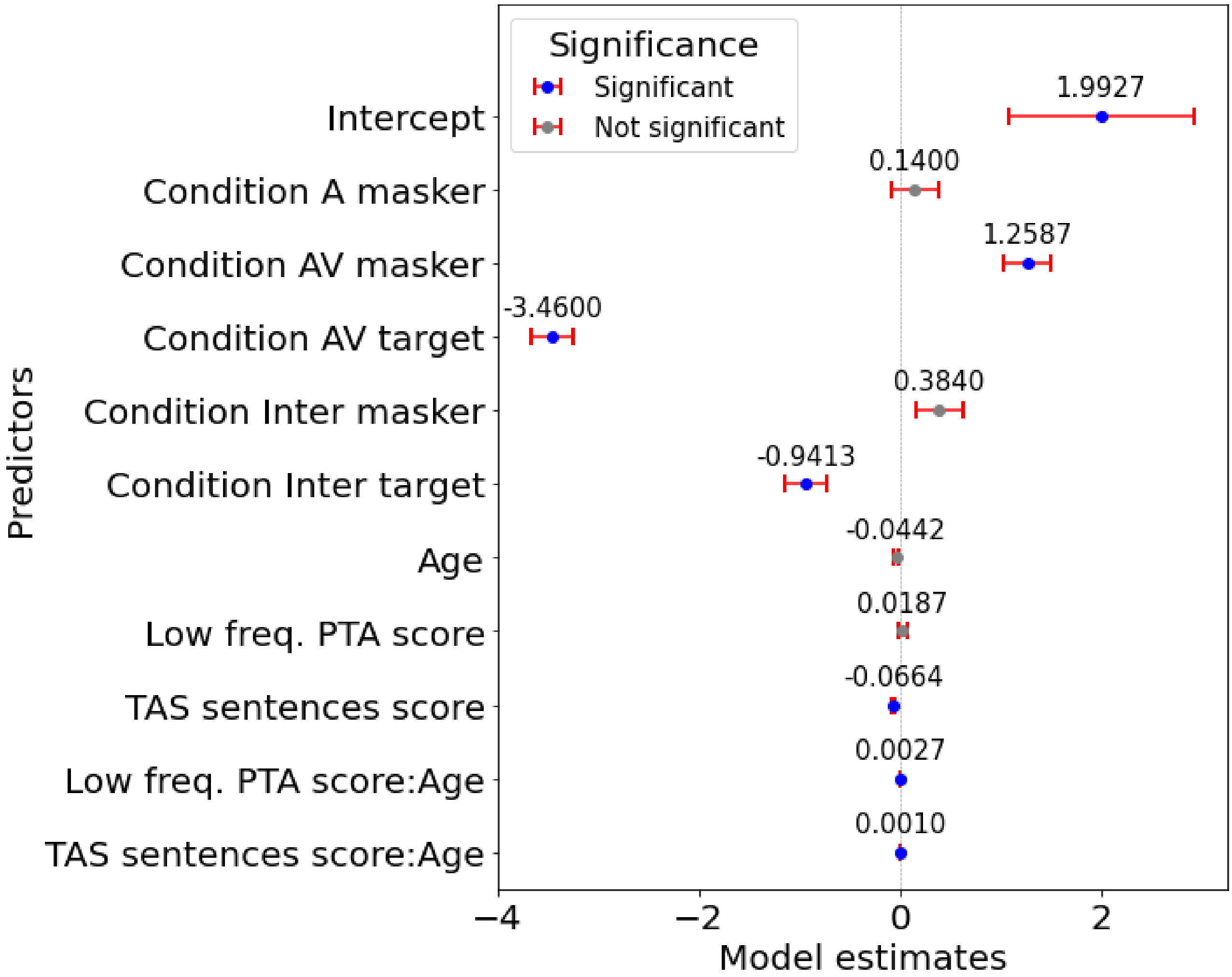


***Figure 6.3**: Coefficient plot of the speech-in-noise performance model's estimated intercept, and predictor slopes. Blue marker colour indicates the estimate was statistically significant and grey that it was not significant. Values of the estimates are also illustrated on top of their respective marker. Error bars are standard errors.*

Based on the model's output, on average, when all other predictors were kept constant, and at their baseline values, participants were performing at ~2 dB SNR (the intercept estimate). This value changed as the values of the model's predictors changed. The model output included all the subcategories of the categorical variable "Condition", except "A target". This is because "A target" was used as the model's reference value, with which it compared all the remaining categories of the "Condition" predictor, with each comparison listed individually. Generally, the results of the model were consistent with the conclusions drawn in Chapters 4 and 5: All subcategories of "Condition" had a significant coefficient, except from A masker and Inter masker. Thus, for all subcategories, except A masker and Inter masker, participant performance differed from their performance at the A target condition, by a value provided by their respective coefficient estimate. These were, 1.26, -3.46, and -0.94, for AV masker, AV target, and Inter target respectively. Therefore, participant performance worsened by 1.26 dB SNR in the AV masker condition, compared to the A target, and improved by 3.46 dB SNR and 0.94 dB SNR in the AV target and Inter target conditions respectively, compared to the A target.

The predictors “Age” and “Low frequencies PTA score” (i.e. hearing loss) did not have a significant effect on speech-in-noise performances as main effects, but were involved in significant interactions. Namely, the effect of “TAS sentences score” on the speech-in-noise performance, although otherwise positive (“TAS sentences score” coefficient = -0.06) declined with participant age (“TAS sentences score*Age” coefficient = 0.001). Similarly, the effect of “Low frequencies PTA score” on speech-in-noise performance was worse for older participants (“Low frequencies PTA score*Age” coefficient = 0.0027).

6.2.4. Modelling participant AV benefit

The second model developed is of the AV benefit received by participants, as measured with the vCCRMn test. As a brief reminder, this metric was calculated as the difference in performance between the “A target”, and the “AV target” conditions of the task. Specifically, to ensure more positive (or, more generally, larger) values implied *more* AV benefit, the AV benefit was computed as AV benefit = A target – AV target.

This AV benefit reflected the maximum auditory support a participant was able to derive from vision, within the context of the task. As discussed in previous chapters, the AV benefit measured with the vCCRMn was expected to include both linguistic (i.e. lipreading-based), and non-linguistic (i.e. audio-visual temporal coherence-based) visual helping components.

6.2.4.1. Model development

Table 6.5. depicts a subset of the datasheet used in the development of this model, illustrating the structure of the dataset on which the model was applied. Following a similar reasoning to that of the previous model for speech-in-noise performance, the variable “Participant” was included here as the model’s random intercept. Further, we see that the values of “Inter benefit”, and “Run” varied within each participant, and thus potentially formed good candidates for inclusion in the model as random slopes. The effects of “Inter benefit” on the “AV benefit”, however, were not expected to vary substantially from participant to participant and so the factor was not included as a random slope (also to avoid complicating the model further, see below).

AV benefit	**Participant**	**Inter benefit**	**Run**	**Age**	**Gender**	**TAS sentences task score**
6	1	5	1	32	Male	80
5.5	1	-3	2	32	Male	80
6	1	3.5	3	32	Male	80
4	2	2.5	1	21	Female	80
0	2	0.5	2	21	Female	80
5.5	2	4.5	3	21	Female	80

2	3	-2	1	35	Male	73.33
3.5	3	1.5	2	35	Male	73.33
4	3	1.5	3	35	Male	73.33

***Table 6.5**: Subset of the datasheet used in the development of the AV benefit model.*

Following the systematic procedure for model development described in section 6.2.2, the first model fit attempt was the following:

AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + Talker ID score + MoCA score + TAS sentences score*Age + TAS sentences score*Talker ID score + TAS sentences score*Gender + Low frequencies PTA score*Age + Inter benefit*Age + Inter benefit*TAS sentences score + Inter benefit*Low frequencies PTA score (1 + Run | Participant)

Thus, AV benefit was at a starting stage modelled as a function of the fixed effects "Inter benefit", "Run", "Age", "Low frequencies PTA score", "Gender", "TAS sentences score", "Talker ID score", "MoCA score", (and interactions), "TAS sentences score*Age", "TAS sentences score*Talker ID score", "TAS sentences score*Gender", "Low frequencies PTA score*Age", "Inter benefit*Age", "Inter benefit*TAS sentences score", "Inter benefit*Low frequencies PTA score", and included "Participant" as a random intercept and "Run" as random slope. Table 6.6 provides explanations for each of the variables used in the model.

Variable name	**Type of variable**	**Units or categories**	**Explanation**
AV benefit	Continuous	dB SNR	Measure of participants' audio-visual benefit. Computed as the difference in performance between the conditions "A target" and "AV target" of the vCCRMn task.
Inter benefit	Continuous	dB SNR	Similar to AV benefit. Measures of participants' audio-visual benefit attributable to the interrupted, "Inter target" condition of the vCCRMn. Computed as the difference between "A target" and "Inter target".
Participant	Categorical	Participants 1 to 125	Participant IDs

Run	Numeric	Runs	The run number that the participant ran for a given condition
Age	Numeric	Years	The age of the participant
Low frequencies PTA score	Continuous	dB HL	Mean audiogram score for frequencies 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz, across both ears.
Gender	Categorical	Male, Female	The gender of the participant
TAS sentence score	Continuous	% correct	Measure of performance in the TAS silent lipreading task, for the "sentences" level
Talker ID score	Continuous	% correct	Measure of participant's ability to match the vCCRMn task's talkers' voices to their faces
MoCA score	Continuous	% correct	Measure of older participant's cognitive function as measured by the MoCA task. Younger participants were given a full score for modelling purposes even though they had not undertaken this test.

***Table 6.6**: Descriptions of the predictors used in the model for the AV benefit.*

The predictors included in the model were selected based on their presumed relevance to the AV benefit, also influenced by the results of Chapter 4, and the global analyses of Chapter 5. Namely, "Inter benefit", the benefit gained from the interrupted condition was assumed to constitute the part of the AV benefit attributable to mechanisms other than lipreading the target words. There was a concern that Inter benefit and AV benefit were measuring similar psychophysical parameters, due to the similarity of the conditions AV target and Inter target from which the respective benefits were derived. If this were the case it would have been inappropriate to include Inter benefit as a predictor for the AV benefit.

Inter benefit was shown to be correlated with AV benefit in Chapter 5, with a moderate-in-strength coefficient. To further ensure that AV benefit and Inter benefit were not strongly correlated, within the context of the dataset going into the model, another correlation was conducted here that included the AV and Inter benefits obtained from each Run (see Table 6.5) separately – in Chapter 5, the means across the three runs were used in the correlation. The result was a non-concerning, weak-to-moderate correlation (Pearson's r = 0.35, p-value < 0.001; see also Schober et al., (2018) for a guide in interpreting correlation coefficients); thus "Inter benefit" was included in the model.

To capture the contributions of lipreading, the variable "TAS sentences score", was included. TAS sentences score was also used in correlational analyses between participant AV benefit and lipreading ability in Chapter 4, and was shown to be positively correlated with the AV benefits of the participants in the younger participant with normal hearing experimental group. Furthermore, it was shown in Chapter 5 that collectively, participant lipreading ability was positively associated with AV benefit.

In the collective analyses of Chapter 5 it was shown that the run number of the vCCRMn did not influence participant AV benefit. The variable ("Run") was nonetheless included as a predictor to see if the more powerful modelling procedure would uncover an effect. "Gender" was included in the model too, following the results of Chapter 5. Furthermore, it was shown in Chapter 5 that Talker ID scores ("Talker ID score") and age ("Age") were correlated with AV benefit; thus, they were also included as predictors in the model. On the other hand, no association was found between AV benefit and MoCA ("MoCA score"). The variable was included as a predictor, nonetheless, hypothesising it might have an effect on the AV benefit that the model would be better fitted to unravel. For similar reasons, although hearing loss was not found to be associated with the AV benefit in the analyses of Chapter 5, it was included as a predictor in the model ("Low freq. PTA score").

Some two-way interaction terms included between predictor pairs were deemed reasonable, while striving to avoid overcomplicating the model. Thus, it was considered that the effects of lipreading ability ("TAS sentences score"), hearing loss ("Low frequencies PTA score"), and "Inter benefit" on the AV benefit, could be influenced by "Age". The effect of lipreading ("TAS sentences score") could be influenced by ability to familiarise with the one's interlocutor ("Talker ID score"), and the "Gender" of participants. Additionally, a two-way interaction term between "Inter benefit" and "Low frequencies PTA score" was included to further assess the conclusion of Chapter 4 that the effect of Inter benefit on AV benefit depends on the formation of auditory objects, which would depend on a participant's ability to hear. Finally, an interaction was included between "Inter benefit" and "TAS sentences score" to assess the hypothesis that the mechanisms of audio-visual temporal coherence, and lipreading, synergistically contributed to the AV benefit.

To assess the effect of including "Run" as a random slope, the model was compared with a reduced version that included only the random intercept for "Participant" and excluded "Run" as a random slope. The comparison was significant ($\chi^2(2) = 9.86$, p = 0.007), suggesting that the random slope for "Run" did affect "AV benefit", and was thus kept in the model.

Next, I proceeded with the trimming down of predictors. Starting with interaction terms, I removed predictors one by one, comparing full and reduced models, and assessing how model output changed with each removal. The starting-stage model output (not shown here) was used as guide for selecting the first predictor to remove, and with each removal, the reduced models' outputs were used for further guidance. Considerations for removal included primarily a predictor's p-value, and whether that predictor was associated with terms that had high VIFs. Briefly, the following terms were removed, in the order provided: "Inter benefit*Age", "Inter benefit*Low frequencies PTA score", "Inter benefit*TAS sentences score", "TAS sentences score*Gender", "TAS sentences score*Talker ID score", "MoCA score", and "Talker ID score" (respective Likelihood ratio test results: $\chi(1) = 0.25$, p = 0.62; $\chi(1) = 0.15$, p = 0.70; $\chi(1) = 1.66$, p = 0.20; $\chi(1) = 2.40$, p = 0.12; $\chi(1) = 0.07$, p = 0.80; $\chi(1) = 0.01$, p = 0.92; $\chi(1) = 1.71$, p = 0.19).

The end result was the model AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + TAS sentences score*Age + Low frequencies PTA score*Age + (1 + Run | Participant). As done previously for the speech-in-noise model, co-linearity between the model's predictors was assessed with VIF estimates (Table 6.7). All VIFs were smaller than 5, suggesting non-concerning levels of co-linearity. Further, it was concluded based on the model's residual plot (Figure 6.4) and histogram (Figure 6.5), that the assumptions of linearity, homoscedasticity, and normality of residuals were met. Model output results are discussed in the following section.

Predictor	**VIF**
Inter benefit	1.01
Run	1.00
Age	3.84
Low frequencies PTA score	3.55
Gender	1.03
TAS sentences score	2.29
TAS sentences score*Age	3.33
Low frequencies PTA score*Age	4.19

Table 6.7: *VIF output for the model AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + TAS sentences score*Age + Low frequencies PTA score*Age + (1 + Run | Participant).*

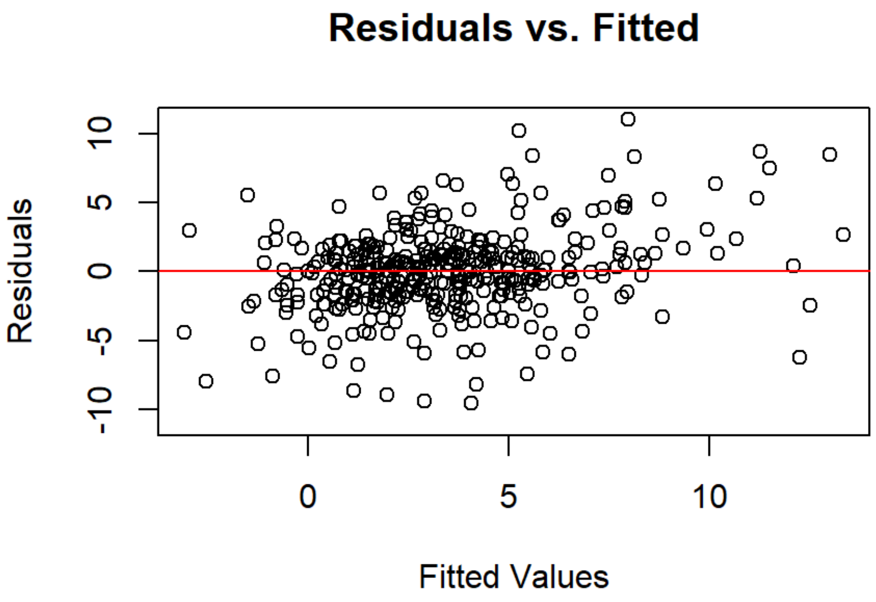


***Figure 6.4**: Residual plot for the model AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + TAS sentences score*Age + Low frequencies PTA score*Age + (1 + Run | Participant). Based on this plot, it was concluded that the modelling assumptions of linearity and homoscedasticity were met.*

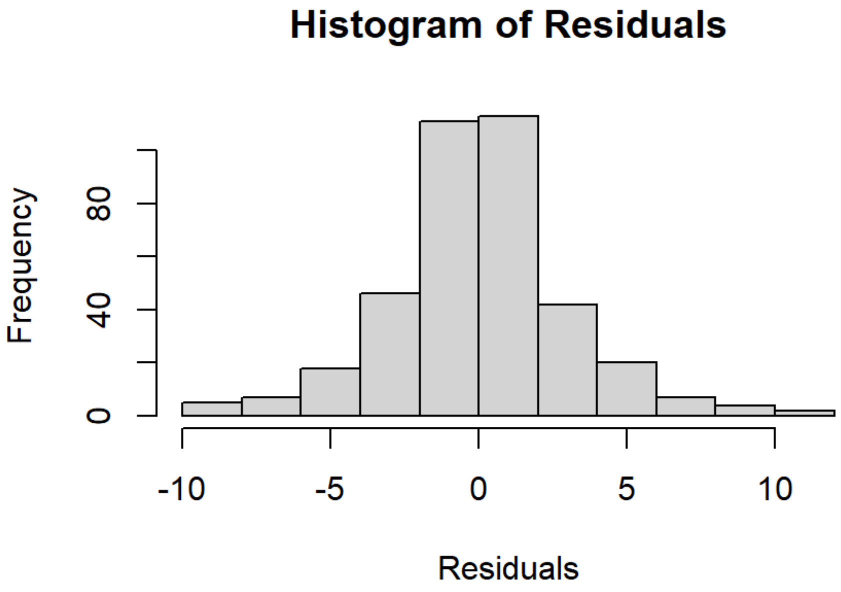


***Figure 6.5**: The distribution of residuals for the model AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + TAS sentences score*Age + Low frequencies PTA score*Age + (1 + Run | Participant). Based on this plot, it was concluded that the modelling assumption of normality of residuals was met.*

6.2.4.2. Model results

In the previous section, I detailed the procedures undertaken to develop the following model for the AV benefit: AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + TAS sentences score*Age + Low frequencies PTA score*Age + (1 + Run | Participant). In the current section, I discuss the final model results.

The model's output is shown in Table 6.8; this includes the estimate, standard error, and relevant statistical test parameters for the model's intercept, and each of its predictors. The intercept, and slope coefficient estimates, along with their standard errors are also visualised in Figure 6.6.

Term	Estimate	Std error	t-statistic	DoF	p-value
(Intercept)	1.0180	1.5666	0.650	119.8469	0.5171
Inter benefit	0.3778	0.0569	6.634	357.0711	< 0.001 ***
Run	0.6247	0.2736	2.283	123.8618	0.0241 *
Age	-0.0159	0.0398	-0.399	115.0240	0.6907
Low freq. PTA score	-0.1347	0.0721	-1.869	115.2627	0.0600 .
Gender Male	-0.9543	0.4446	-2.148	114.3329	0.0338 *
TAS sentences score	0.0794	0.0252	3.151	116.4936	0.0021 **
TAS sentences score:Age	-0.0014	0.0006	-2.227	115.4513	0.0279 *
Low freq. PTA score:Age	0.0043	0.0016	2.715	114.7321	0.0078 **

Table 6.8: *Output for the model AV benefit ~ Inter benefit + Run + Age + Low frequencies PTA score + Gender + TAS sentences score + TAS sentences score*Age + Low frequencies PTA score*Age + (1 + Run | Participant). The model's R^2 was 0.444. Star, and dot symbols in the p-value column indicate statistical significance (*, **, and *** indicate statistical significance with $p < 0.05$, < 0.01, and 0.001 respectively; the . symbol indicates marginal significance).*

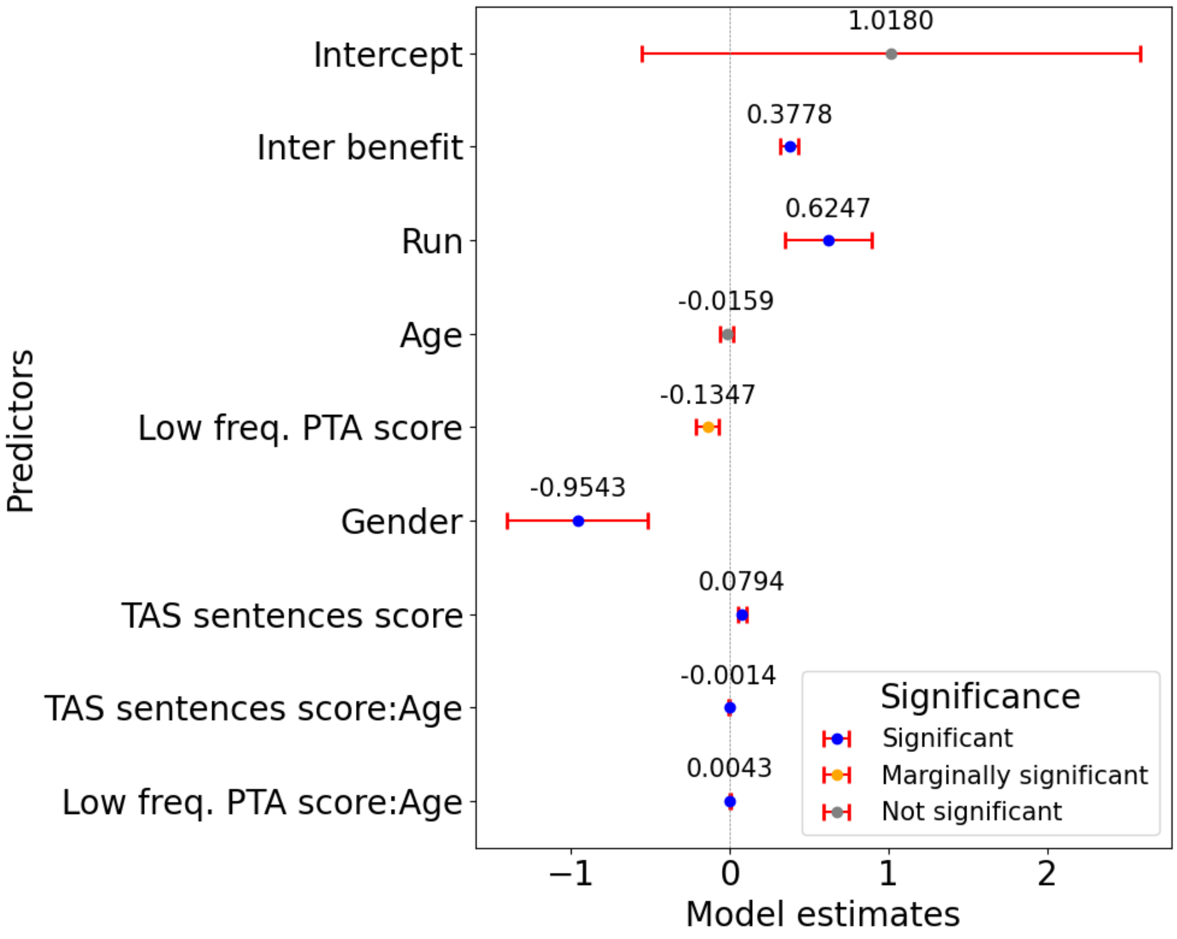


***Figure 6.6**: Coefficient plot of the AV benefit model's estimated intercept, and predictor slopes. Blue marker colour indicates the estimate was statistically significant, orange that it was marginal significant, and grey that it was not significant. Values of the estimates are also illustrated on top of their respective marker. Error bars are standard errors.*

Based on this model, the average participant had a baseline AV benefit of 0 dB SNR (the model intercept was not significant). This value changed, with changes in the values of the significant predictors from Table 6.8 and Figure 6.6.

The predictor "Gender", with the "Male" category displayed (Table 6.8), was significant, with a coefficient of -0.95. With the "Female" category serving as the reference, this result suggested that males gained, on average, 0.95 dB less AV benefit than females. The predictor "Run" was also significant, with a coefficient estimate of 0.62. Thus, the effect of a unitary increase of "Run" resulted in an increase of the AV benefit by 0.62 dB SNR. This result suggests there was a learning effect during the experiments, with participants gaining more AV benefit with each run of the experiment they ran. Similarly, the predictor "Inter benefit" was significant with a positive coefficient estimate of 0.38; thus, with every unitary increase in Inter benefit, participants gained 0.38 dB more AV benefit.

"Low frequency PTA score" had a marginally significant coefficient of -0.13 and was also involved in a significant interaction with "Age", which itself had a coefficient of 0.004. Thus, although the negative coefficient of Low frequency PTA score suggests that an increase in hearing loss reduces participant AV benefit, the positive coefficient of the interaction term suggests that this effect is attenuated when participants are older: Being older with hearing loss is less detrimental on the AV benefit than being younger with hearing loss. To get a better understanding of this interaction between hearing loss and age, I used the model to predict the AV benefit across a range of values for the two variables, and plotted the results as a heatmap, in Figure 6.7. More specifically, to produce the figure, I set all the predictors not involved in the interaction to zero and computed the AV benefit as Low frequency PTA score and Age varied (Ranges used: -5 to 45 dB HL, and 20 to 85 years, respectively – based on participant data).

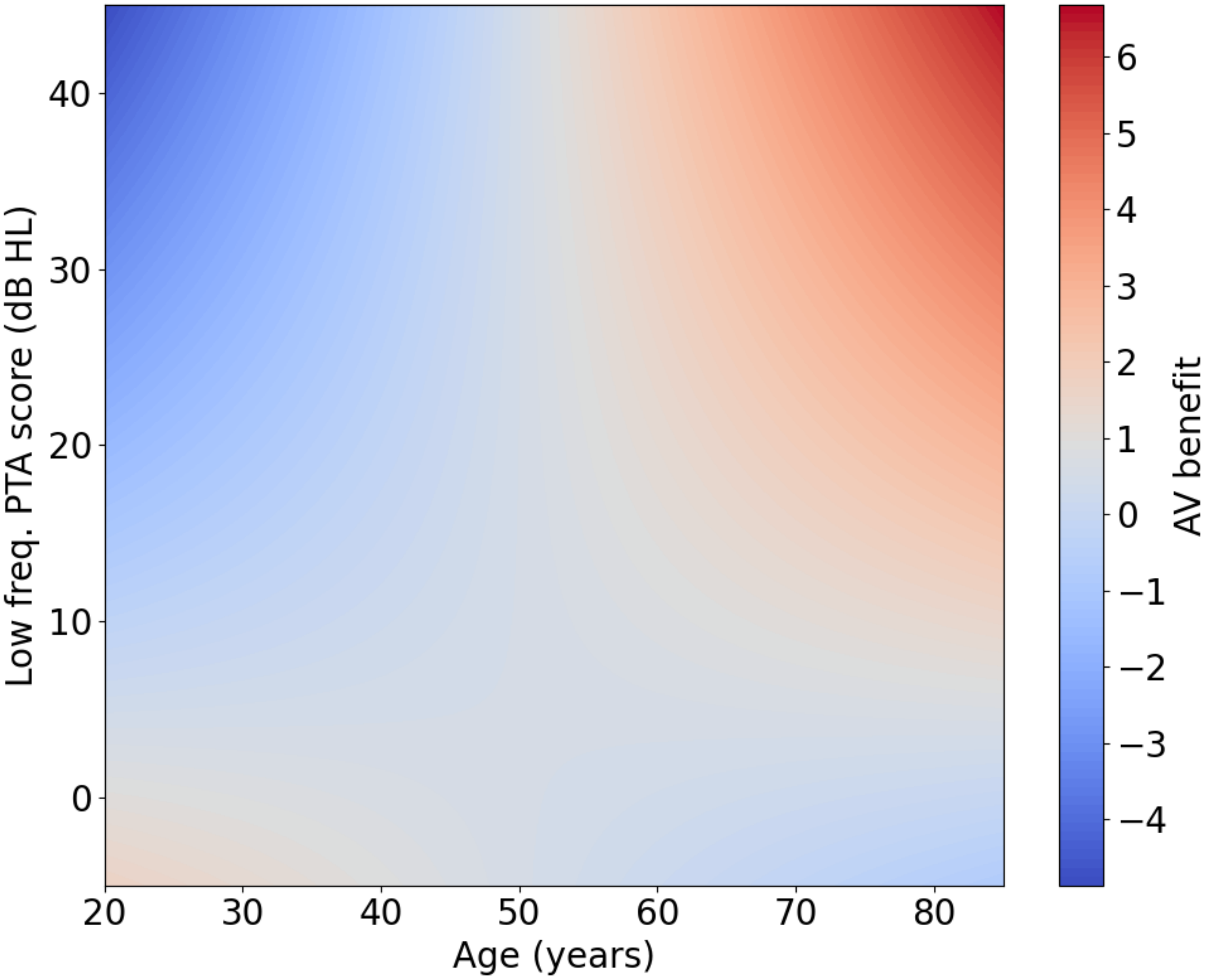


***Figure 6.7**: Visualisation of the Low frequency PTA score by Age interaction of the AV benefit model. To generate this plot, all predictors not involved in the interaction were set to zero, and AV benefits were predicted from the model across the range of values shown for the Low frequency PTA score (-5 to 45 dB HL) and Age (20 to 85 years).*

From the figure, we can see that over a PTA threshold of about 5 dB HL, for the same hearing loss, younger participants were gaining less AV benefit than older. Below the ~5 dB HL threshold however (i.e. when participants had excellent hearing), younger individuals were getting more AV benefit than older.

The predictor "TAS sentences score", was also significant, as was its interaction with "Age" ("TAS sentences score:Age"). "TAS sentences score" had a slope estimate of 0.08, and the interaction with age had a slope estimate of -0.001. Thus, although "TAS sentences score" generally contributed to the

AV benefit (positive slope), its influence was smaller for older participants compared to younger. Starting with 19-year-old participants (the baseline for these datasets), the AV benefit gained from a unitary increase of "TAS sentences score" was 0.08 dB SNR, and 0.001 dB SNR was subtracted from this with each additional year of age. Once more, to get a better appreciation of the interaction term, the model was used to predict AV benefits and plotted them as a heatmap (Figure 6.8), over a range of values for TAS sentences score and Age (ranges: 0 to 100 %, and 20 to 85 years, respectively).

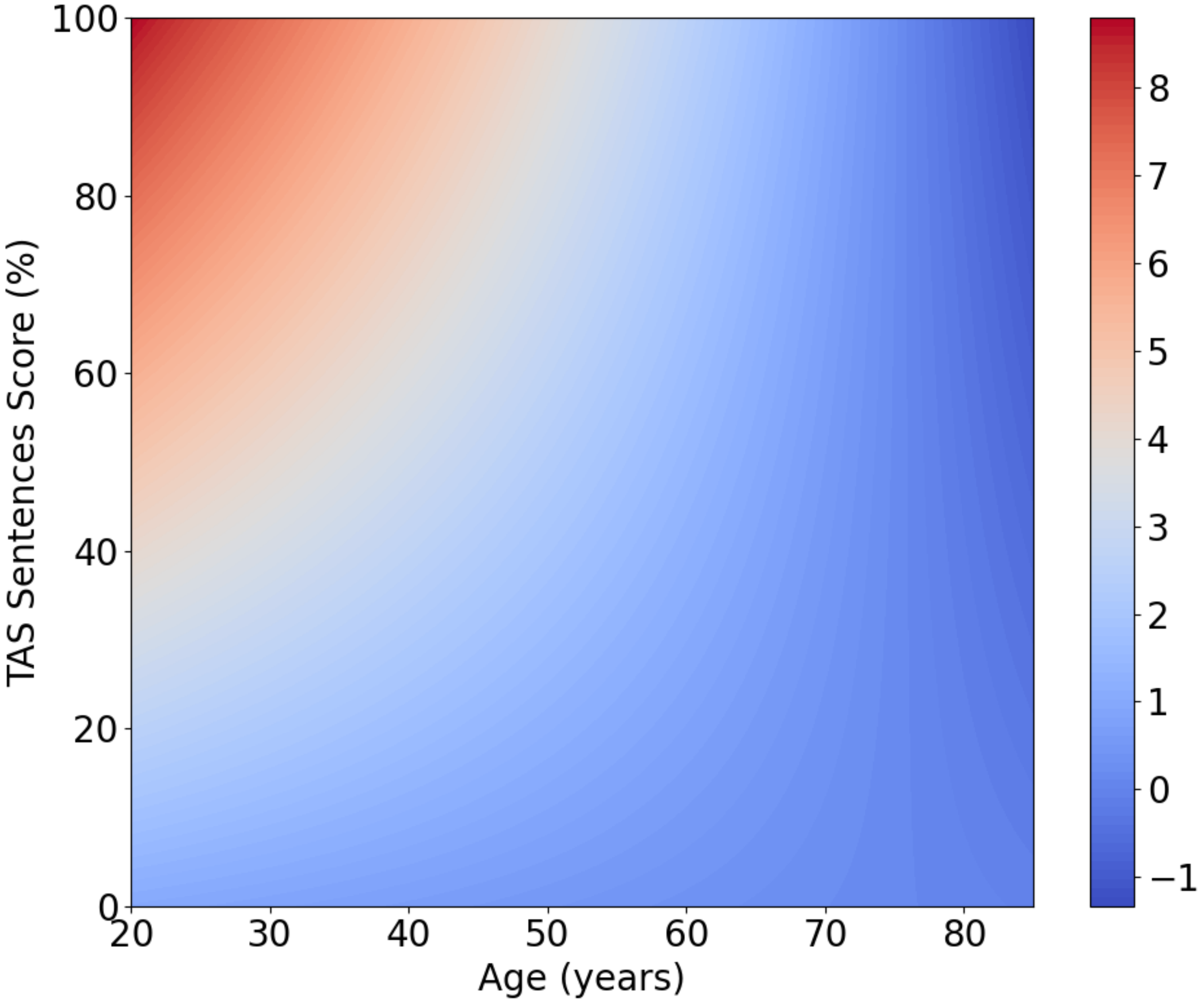


Figure 6.8*: Visualisation of the TAS sentences score by Age interaction for the AV benefit model. To generate this plot, all predictors not involved in the interaction were set to zero, and AV benefits were predicted from the model across the range of values shown for TAS sentences score (0 to 100 %) and Age (20 to 85 years).*

Looking at Figure 6.8, we see that generally, for the same lipreading ability, younger participants obtained higher AV benefits than older participants. Notably, over the age of approximately 80 years, and above a lipreading score of ~60%, participants were even obtaining negative AV benefits – that is, being good lipreaders caused them to perform worst in the AV target condition of the vCCRMn than they did in the static image (A target) condition.

Finally, the predictor “Age” itself did not result in a significant slope estimate, but, as just discussed, influenced performance via its interactions with hearing loss (“Low frequency PTA score”) and lipreading ability (“TAS sentences score”).

6.3. Summary of chapter and discussion

In the analyses of Chapter 5, we saw that speech-in-noise performance decreased with age and hearing loss, and improved upon seeing the talker’s face, and being a good lipreader. Here we see that the effects of age are mediated through an interaction with lipreading, and also that the effects of hearing loss interact with those of age to bring down participant performance. The model also confirmed the results of Chapter 4 and 5, showing that participant performance varied significantly across the different conditions of the vCCRMn task.

Specifically, using the A target condition as its baseline, the model compared all the remaining conditions against this baseline to produce the following results: Firstly, participant performance was better in the AV target condition. This confirmed the previous claim that the vCCRMn task was capable of capturing participants’ AV benefits. Second, it confirmed that participants were gaining a benefit from the Inter target condition (an Inter benefit). Third, it showed that the AV masker condition, impaired participant scores, resulting in what I previously called a “visual detriment” of their auditory performance. This impairment was shown in Chapters 4 and 5, as well. These results collectively suggest that obligatory visual cues were able to not only provide participants with a visual benefit, but also to impair their listening performance.

Similarly to the results of the previous chapters, participant performance did not differ here between the two static image conditions (A target and A masker). This control comparison here too confirms that it was not the identity of the talkers displayed in the video that provided participants with the measured visual benefits. As discussed in Chapter 4, this finding is contrary to that of Cappelloni et al., (2023) where it was suggested that the talker displayed on the screen was sufficient to improve participant performance, and audio-visual temporal cues were not relevant.

With regards to the AV benefits measured from my participants, in the discussion of Chapter 4 I argued that, in addition to audio-visual integration related to lipreading, audio-visual temporal coherence, measured via the participants’ Inter benefits, was also a likely contributor. I was not however able to satisfactorily quantify the relative contributions of each mechanism to the AV benefit in previous chapters.

The model confirmed that both lipreading, and Inter benefit were significant contributors, and revealed their relative contributions. Additionally, the model did not yield a significant interaction between Inter benefits and lipreading scores, suggesting that the two contributors acted independently in their contributions to the AV benefit. The contribution due to the Inter benefit was, in fact, substantial, with each unitary increase in Inter benefit resulting in a 0.38 dB SNR increase in the AV benefit. Lipreading ability-related improvements were more complex, with participant age influencing the contribution of lipreading through an interaction. Generally, although for younger ages lipreading skill improvements resulted in increases in the AV benefit, for older participants it reduced their AV benefit. This finding could be related to the fact that attentional control, the ability to direct attention to the relevant stimuli, in the presence of several options, declines with age (Pepper & Nuttall, 2023). Thus, older participants, albeit good lipreaders, potentially struggled to apply their skill in the presence of the multiple distracting stimuli of the vCCRMn tasks.

During my experiments I had confirmed, via a camera, that participants directed their gaze to the tasks' display screen, but no specific method of measuring their attention was employed otherwise. Simple measures like reporting colour changes to lips (see, for example Cappelloni et al., 2023 and Maddox et al., 2015), or even more complex measures involving pupillometry (Zhao et al., 2019), would be useful for future assessments. I believe, however, that most useful would be to include in the battery of tests a test that explicitly measures attention, such as the Test of Everyday Attention (TEA; Robertson et al., (1996)).

Beyond attentional control, the negative effect of ageing on lipreading's contribution to the AV benefit may reflect differences in cognitive processes, such as working memory and processing speeds (de Dieuleveult et al., 2017; Nettelbeck & Burns, 2010), between young and old, that were likely not captured by MoCA (as MoCA was not found to be a significant contributor to the AV benefit). Nonetheless, the finding that lipreading ability decreases older individuals' AV benefits may partly explain why there exists no reliable evidence that lipreading training regimes can make a difference in real-world situations (Campbell & Mohammed, 2010).

In Chapter 5 I observed, through simple correlational analyses, that hearing loss did not appear to exert a significant effect on the AV benefit. I argued there that this lack was likely due to the inability of simple analyses to capture the dependencies of multifaceted factors such as the AV benefit. Indeed, the AV benefit model showed that this was the case. The model confirmed that hearing loss did impose a detrimental effect on the AV benefit but revealed a more complex interaction with participant age. Namely, below a PTA threshold of ~5 dB HL, younger individuals were getting more AV benefit than older – as would be expected. That is, when participants had excellent hearing, being older resulted in lower AV benefits. The situation was, however, reversed for PTA thresholds above the ~5 dB HL threshold. Over this threshold, for the same hearing loss, younger participants were gaining less AV benefit than older. This result was potentially due to older individuals having developed adaptation mechanisms, after living with hearing loss for many years, including better exploitation of the visual modality. These results should, nonetheless, be interpreted with caution, as the datasets collected herein, and on which the AV benefit model was based, did not include younger participants with hearing loss. Future experiments including such participants would likely provide a clearer picture of the interaction of hearing loss with age.

Participant gender was previously (Chapter 5) found to significantly influence the AV benefit participants obtained. Namely, female participants were shown to gain more AV benefit than males (also in accordance with previous reports, e.g. Dancer et al., 1994). Confirming these results, the model showed that female participants gained ~ 1 dB SNR more than males. Experimental run number was not found to significantly influence the AV benefit in Chapter 5. Nonetheless, with the more powerful statistical approach of the model, the factors were shown to be a significant predictor of the AV benefit in the current chapter. Specifically, with each unitary increase in the task's run, participants gained 0.62 dB SNR more AV benefit. This result is potentially promising for clinical populations, as it supports that with training patients with hearing loss can, in principle, become better at using visual cues to aid their audition.

Finally, age did not by itself influence the AV benefit, but as explained in the previous paragraphs, the modelling revealed that it did so through interactions with hearing loss and lipreading ability. Specifically, the negative effects imparted on the AV benefit through the interaction of age with lipreading are in line with the known negative effect of age on AV integration and cognitive abilities (Nettelbeck & Burns, 2010; Pepper & Nuttall, 2023).

In the current chapter, I have shown that linear mixed effects models form a powerful and flexible method, suited for complementing, and enhancing the analyses conducted in the previous chapters. This modelling method was particularly useful in this work, as it enabled the investigation of the nuances of audio-visual integration across subjects, predictors, and predictor interactions. Nonetheless, it does not necessarily constitute the best, or, for that matter, only method with which one could approach this problem (see for example Blackburn et al., (2019) and Stacey et al., (2016) for signal detection theory-based models).

On a final note, I'd like to shift the attention towards the clinical sector. As I stated at the end of Chapter 1, throughout this PhD, I approached my research with the ambition that it has translational impact. In developing the model for the AV benefit, I had envisioned that it, along with the vCCRMn task, and a lipreading skill assessment task to accompany the model, would find a place in the hands of health-care professionals, where they could be used to improve the lives of people with hearing loss. Given these tools, a health-care professional could, for example, judge that her patient would benefit from receiving temporal coherence detection training (e.g. Atilgan & Bizley, (2021)), but that lipreading training might not be worthwhile. Ultimately, the findings reported here underline the importance of such audio-visual skill-related testing, and the adaptation of a “multi-sensory” mentality, in providing appropriate guidance that would optimally benefit the patients.

Chapter 7: General discussion

This PhD thesis was concerned with audio-visual integration, and specifically, with how looking at their interlocutor's mouth helps individuals listen in noise. It is a well-known fact that lipreading – the extraction of linguistic information via the learned matchings between mouth movements and speech sounds (Auer, 2010; Murthy, 2020) – provides a visual enhancement to listening (Erber, 1975; Grant & Walden, 1996; Macleod & Summerfield, 1987; Middelweerd & Plomp, 1987; Picou et al., 2011; Sumby & Pollack, 1954). This visual enhancement was hereby called the audio-visual (AV) benefit. In Chapter 1 of this work, it was argued that, in addition to lipreading, the mechanisms of audio-visual temporal coherence may also contribute to the AV benefit. These mechanisms, it was explained, rely on the fact that the amplitude of the speech envelope, and the opening and closing of the mouth during speech, are temporally correlated (Chandrasekaran et al., 2009; Grant, 2001). They are language-independent and depend on the formation of the so-called perceptual auditory objects in the listener's brain. Generally, the contribution of these mechanisms to the AV benefit has often been neglected or overshadowed by that of lipreading – with a couple of notable exceptions (Yuan et al., 2020, 2021).

Both lipreading, and audio-visual temporal coherence mechanisms, I argued in Chapter 1, have the potential to influence what Bregman (1990) called the problem of auditory scene analysis – which is, within the context of speech-in-noise perception, the isolation and extraction of a voice from a mixture. Previous experiments (cited in the paragraph above), however, were not specifically designed to test this hypothesis. Thus, some experimenters, that used, for example, words as stimuli, captured predominantly lipreading contributions (e.g. Sumby & Pollack, 1954), and others purely those of audio-visual temporal coherence (Yuan et al., 2020, 2021). Among the primary goals of my work, was to assess whether both of these could enhance the listeners' auditory experience simultaneously, and independently, in realistic speech-in-noise scenarios (Aim 2 of the thesis).

With this end in mind, in Chapter 2, I set out to develop an audio-visual speech-in-noise task considering design criteria that would allow for a) the capturing of an AV benefit (Aim 1 of the thesis) that also b) consisted of both lipreading, and temporal coherence contributions (Aim 2). Namely, the final version of the test I developed (the vCCRMn task) considered parameters including type of stimuli used, type of background noise, SNR variation procedures, and the different types of conditions it used.

To achieve the first two aims of the project, the vCCRMn made use of speech sentences for stimuli, with target words at their end, to provide sufficient temporal window for the mechanism of temporal coherence to build up (Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008). The target words it used were specifically designed to be sufficiently (but not overly) difficult to lipread, striving to strike a balance between the listener's reliance on linguistic and non-linguistic audio-visual mechanisms for solving the task. A background noise of two competing masker talkers was selected, one male, one female, reading sentences of the same format as that of the target talker. These served to provide an optimal amount of informational and energetic masking, making the task sufficiently difficult (Freyman et al., 2001, 2004; Kidd & Colburn, 2017; Rosen et al., 2013) for increased reliance of listeners on the visual modality and thus likely detection of AV benefits. The use of competing masker talkers, and in specific the masker-coherent conditions, also served to increase the likelihood of uncovering obligatory audio-visual interactions that were detrimental to the listener's performance (Bizley et al., 2016; Maddox et al., 2015).

In terms of SNR variation procedures, the vCCRMn made use of the 1U1D adaptive staircase procedure for SNR variation (Levitt, 1971). This procedure was selected for its swiftly outputting of listener SRT_{50}. As discussed in Chapter 2, the SRT_{50} point on the listeners' psychometric curves was hypothesised to be optimal for capturing an AV benefit comprised of both lipreading and non-linguistic contributors.

Finally, the vCCRMn task included three audio-visual main conditions, two with video (AV and Inter conditions) and one with a static image (A condition); all three main conditions included both target-, and masker-coherent sub-conditions. The two video conditions were similar and were designed to capture visual benefits when compared to the static image condition (more specifically, when their respective target-coherent sub-conditions were compared to target coherence static image condition). The AV condition served to capture the listeners' total AV benefit (including contributions from both lipreading and audio-visual temporal coherence). The Inter condition, on the other hand, was designed with Aim 2 in mind, such that the displayed talker video froze during the target word presentation, such that the condition captured only the non-linguistic, audio-visual temporal coherence contributions to the AV benefits – hereby called "Inter benefits".

I began the development of the vCCRMn using the Children's Coordinate Response Measure task (CCRM; Messaoud-Galusi et al., 2011) audio-only speech-in-noise task as my starting point. To the CCRM, video stimuli were added, converting it to an audio-visual speech-in-noise task (vCCRM). Following several piloting tests, where the vCCRM was first piloted as a speech-in-noise task (Pilot 1), and subsequently as a silent task (Pilots 2 and 3), it was concluded that the target word stimuli it used were lipreadable to the extent where several participants could rely solely on lipreading cues to solve the task. To amend this issue, the target words (the "number" words, more specifically) were replaced with confusable, with each other (from a lipreading point of view), noun words. Following several piloting stages (Pilots 4, 5 and 6), these noun words were deemed appropriate for use in the final version of the vCCRMn task. This task, along with several other complementary tests, which included the TAS lipreading assessment (Campbell et al., 2003), were employed during the official testing phase of participants.

As described in Chapter 3, the testing phase was divided into three main experiments: The collection of data from younger participants with typical hearing, older participants with typical hearing, and older participants with hearing loss. These experiments enabled the investigation of the question of whether both lipreading, and audio-visual temporal coherence contributed to the AV benefits measured via the vCCRMn. Additionally, they allowed the investigation of the effects of ageing, and hearing loss on these AV benefits.

The data collected from the three experiments were first examined in Chapter 4. There, I showed that across all three experimental groups, participants were gaining an AV benefit in the AV target condition, compared to the A target (audio-with static image) condition – confirming that the vCCRMn task was capable of measuring an AV benefit (in line with Aim 1). In addition to the AV benefit, participants with typical hearing also gained a visual benefit from the Inter target condition, compared to the A target (in line with Aim 2). Generally, participant performance in the target-coherent conditions was the best for the AV target, followed by the Inter target, followed in turn by the A target. It was argued in Chapter 4 that the Inter condition benefits reflected audio-visual temporal coherence cues, and that the additional gains conferred by the AV condition were likely due to the provision of lipreading cues. Thus, it was concluded that the vCCRMn was capturing both these contributions in the measured AV benefits of participants. Further evidence for this claim was provided through showing that participant AV benefits and lipreading scores (measured via the TAS sentences test), correlated positively with each other.

In Chapter 4 it was also shown that the coherence type made a difference in the performances of participants across the conditions of the vCCRMn. Namely, participants across all three experimental groups performed better in the AV target condition, compared to the AV masker. According to object-based attentional theories, it was argued, these performance differences could have been due to underlying audio-visual temporal coherence mechanisms, providing further evidence that these were captured by the AV benefits measured (Bizley et al., 2016; Maddox et al., 2015; Shinn-Cunningham et al., 2017; Shinn-Cunningham, 2008). Furthermore, in Chapter 4 (and also 5 and 6), it was shown that the AV masker condition was even capable of impairing the listening experience, with AV masker performances being worse than performances in the A target condition. Thus, the results herein not only showed that the visual modality has the capacity to enhance, but also to impair the auditory experience of listeners.

A further goal of this work was to assess the factors that influenced participants' visual benefits (Aim 3). Comparisons across the three experimental groups, analysed in Chapter 4, suggested that age, and potentially hearing loss imparted a negative effect on the visual benefits measured from participants, with younger participants gaining more AV benefits than both older participants with typical hearing, and older participants with hearing loss. This result was confirmed by the findings of Chapter 5, where it was shown that both participant speech-in-noise performance, and measured AV benefits, were negatively associated with age. In Chapter 5, the data collected from the three experimental groups were pooled together to form a statistically robust sample of 125 participants.

Generally, the analyses conducted in Chapter 5 served two purposes: First, they constituted investigations of the relationships between the various factors measured in the experiments conducted for this PhD – including factors additional to the speech-in-noise performances and AV benefits. These analyses showed, for example, that in addition to being negatively associated with speech-in-noise performance and the AV benefit of participants, age was generally negatively associated with most of the participant performances measured. These findings confirmed previous reports and included negative effects on lipreading (Sommers et al., 2005), hearing loss (Jayakody et al., 2018), and cognitive ability (Nettelbeck & Burns, 2010). Second, the collective data analyses of Chapter 5 helped explore which factors influenced participant speech-in-noise performances and AV benefits and informed the predictor variables included in the two models developed in Chapter 6 (Aim 3; and Aim 4, which was the development of the statistical models). These were a linear mixed effects model of participant speech-in-noise performances, and one of participant AV benefits. These models formed the final aim of this thesis and served to provide conclusive evidence for the findings reported in previous chapters. Further, the models allowed the quantification of the relative contributions of lipreading, and audio-visual temporal coherence contributions to the AV benefits measured – something I was not able to extract through the analyses of Chapters 4 and 5 (Aims 3 and 4). They also allowed for better exploration and unravelling of the complex interrelations between the modelled variables, and the various factors measured in my experiments.

The speech-in-noise model generally confirmed the results of Chapters 4 and 5. It showed that the speech-in-noise performance of participants was influenced by dynamic visual cues. Namely, their best performance was provided by the AV target condition, confirming that participants were gaining an AV benefit (Aim 1), followed by the Inter target condition, confirming that participants were gaining a visual benefit from cues other than lipreading as well (Aim 2). Further, participant performance was worse than the baseline static-image (A target) condition, in the AV masker condition. As mentioned earlier, these findings showed that obligatory visual cues can both benefit and impair the listening experience. The model also showed that performance was not different between the two static image conditions (A target and A masker), suggesting that the identity of the talker was not sufficient to provide participants with a benefit in the video conditions (contrary to what has been suggested by Cappelloni et al., (2023), as discussed in Chapter 4). Further, the model showed that increased lipreading ability generally improved speech-in-noise performance, while age and hearing loss worsened it.

The AV benefit model confirmed that both Inter benefit, reflecting audio-visual temporal coherence contributions, and lipreading ability, influenced the AV benefit of participants. The contribution of Inter benefit was substantial, with a coefficient of 0.38, suggesting that with every 1 dB SNR gained in Inter benefit, participants gained 0.38 dB SNR of AV benefit (Aims 2 and 3). Notably, the contributions of Inter benefit were independent of participant lipreading skill, suggesting that the mechanisms of audio-visual temporal coherence and lipreading, as measured by the current tests, provided independent enhancements to the participants listening experience (Aim 2). Inter benefit was also not interacting with age nor with hearing loss in its contribution to the AV benefit. The model revealed a more complex contribution from lipreading.

Although the fixed term included for lipreading yielded a positive coefficient, suggesting positive contributions to the AV benefit, it was also involved in an interaction with age that had a negative coefficient. Thus, the contributions of lipreading skill to the AV benefit were attenuated with participant age: Older participants would gain less AV benefit than younger participants with the same lipreading skill. As discussed in Chapter 6, this finding could be due to the decline in attentional control that has been reported for older participants (Pepper & Nuttall, 2023). It could also form part of the reason why there exists no reliable evidence that lipreading training regimes can help hearing loss populations listen better in real-world situations (Campbell & Mohammed, 2010).

In addition to lipreading, hearing loss was also found to be involved in an interaction with age, in the AV benefit model. The model showed that, although hearing loss as a fixed effect had a negative effect on the AV benefit (negative coefficient), its interaction term with age had a positive coefficient. Thus, older participants would gain more AV benefit than younger participants with the same hearing loss. As discussed in Chapter 6, however, these results are to be taken with caution, since in the experiments conducted in this work, and on whose datasets the model was applied, did not include younger participants with hearing loss.

Finally, the AV benefit model showed that female participants gained more AV benefit than males (as expected, based on previous reports; e.g. Bernstein, (2018); Dancer et al., (1994); Johnson et al., (1988)). It, also, further showed that with each run of the vCCRMn they ran, participants gained more AV benefit. The latter result could be due to a familiarity effect: It has been reported that generally, speech-in-noise performance increases with increased familiarity with the speaker (Kim et al., 2018). Nonetheless, Talker ID score was not found to be significant predictors of the AV benefit in the model. Alternatively, the increase in AV benefit with run number could be due to an audio-visual training effect, where participants learned to better exploit audio-visual cues with repeated runs. This result would be promising for the clinic, where training regimes might help patients enhance their skills of exploiting their vision to help their audition.

Limitations and future directions

In Chapter 1 (section 1.4.4.2), I outlined the experimental paradigm employed by Maddox et al., (2015), and argued that the task developed by the authors would be ideal for use as a tool for specific capturing of individuals’ ability to exploit audio-visual temporal coherence cues. The initial plan for this project, which was changed due to the COVID-19 pandemic, was the following: The speech-in-noise task would be used to measure participant AV benefit. Then, to assess the contributions of lipreading, and audio-visual temporal coherence effects to the AV benefit, the two factors would be independently measured using the TAS assessment, and one of the two variants of the task from Maddox et al., (2015).

Nonetheless, the task of Maddox et al., (2015) was proven difficult to employ online. Experiments in our lab performed by my supervisor suggested that the in-lab results of Maddox et al., (2015) could not be replicated online. Further investigations demonstrated that this was due to the fact that when the audio-visual stimuli were played in a browser, the variance in the latency of the audio and video elements was large. These results were problematic for this project, as testing had to be moved online during the piloting stages of the speech-in-noise task, due to the COVID-19 pandemic. Due to these issues, and time restrictions, the current work used the Inter condition of the vCCRMn task as an alternative, albeit less controlled method, to capture audio-visual temporal coherence effects. Importantly, since the vCCRMn used speech stimuli, which naturally carry audio-visual latencies and for which audio-visual integration can be successful even when such latencies are large (> hundreds of milliseconds) (Grant & Bernstein, 2019), the online testing issue was mitigated.

Nonetheless, the use of the Inter condition as an alternative to the task from Maddox et al., (2015) had its disadvantages. For example, although the Inter condition removed lipreading cues during key word presentation, it did not control for orthogonality (see also section 1.4.3.1) between the task-relevant stimulus features and the stimulus features potentially involved in audio-visual binding. Future research seeking to independently assess the contributions of lipreading and audio-visual temporal coherence to the AV benefit would ideally control for these variables.

Earlier in this chapter, and in the discussion of Chapter 6, I suggested that the interaction between lipreading and age in the AV benefit model – where older participants gained less AV benefit than younger participants with the same lipreading skill – could be reflective of a decline in attentional control in older participants (as per previous reports; e.g. Pepper and Nuttall, (2023)). A limitation of my study was that I did not measure participant attention during my experiments. As discussed in Chapter 6, future investigations of audio-visual speech perception should include attentional assessments (e.g. via pupillometric methods Zhao et al., (2019)), and ideally an independent test designed to directly measure attention (e.g. the Test of Everyday Attention by Robertson et al., (1996)).

Alternatively, the negative effect of age on the AV benefit provided by lipreading could reflect a decline in cognition that accompanies age (Nettelbeck & Burns, 2010). The MoCA test was employed in the experiments conducted here to detect cognitive impairment in older participants but was not a significant predictor of the AV benefit when included in the AV benefit model. It is possible, nonetheless, that such effects were too subtle to be captured by MoCA.

Another limitation of this work was that, although all participants reported to have normal or corrected-to-normal vision, their visual acuity was not formally assessed during the experiments. As briefly discussed in Chapter 4, visual acuity could impact audio-visual integration (Morris et al., 2012) and could have also been the reason for the non-significant Inter benefits exhibited by the older participants with hearing loss (section 4.4.3). Future work on audio-visual integration would, ideally, directly test the visual acuity of participants.

Finally, the current work assessed participant AV benefit at a single point of their auditory, and audio-visual psychometric curves (the SRT_{50}). Future work employing the vCCRMn could, instead of using adaptive staircases, sample participant AV benefits at various SNRs, to get a more holistic picture of how participant AV benefit varies with listening difficulty. However, assessing AV benefits across a range of listening difficulties would extend the test duration, potentially undermining the task's practicality for future clinical application.

Overall, and as was stated in Chapter 1 and discussed towards the end of Chapter 6, the broader ambition of this work was that it has translational impact. For this goal to materialise, the vCCRMn would have to be validated to ensure it meets several prerequisites – in addition to the requirement that the task is time-efficient. For example, multiple administrators (in addition to myself) should pilot the task to confirm it reliably yields consistent outcomes irrespective of the administrator. Further, the task should be standardised to include systematised administration procedures and usage instructions. There should be normative data available for it as well, to provide health-care professionals with a reference against which to compare their patients' results. These may include collecting more data across ages, hearing levels, and backgrounds.

Final comment

Overall, my work has confirmed that, indeed, our listening experiences are far from purely auditory; on the contrary, as I have shown, vision can both enhance and impair realistic speech-in-noise perception. Although the latter was known already (e.g. Sumby & Pollack, 1954), to my knowledge, the latter has not been shown previously. Further, I have shown that these visual influences are not purely linguistic, but can also include non-linguistic elements, such as the temporal coherence between the mouth opening and the loudness of speech, corroborating previous findings from Maddox et al. (2015) and Yuan et al. (2020, 2021). Notably however, I have shown here for the first time that both linguistic and non-linguistic components can provide simultaneous, and independent, contributions to our ability to understand speech in noisy environments. Thus, with these findings I have disentangled part of the complexity with which vision influences our hearing. It is my hope that with this work I have yielded deeper insight on the principles of audio-visual integration in its relation to speech-in-noise perception and offered useful directions for the future of the clinical sector.

Appendix A: Visible and invisible aspects of speech production

Lipreading can often become a difficult task, as speech is meant to primarily be heard, not to be viewed (Campbell and Mohammed, 2010). However, the visible aspects of lipreading can contribute greatly to the audio-visual benefit (see e.g. Sumby & Pollack, 1954). A few examples of visible and invisible aspects of speech are discussed here, following the textbook *The Sounds of English: Units and Patterns* (Antonopoulou & Pagoni-Tetlow, 2004).

Speech sound production is sometimes difficult, and sometimes visible on the segmental level. Phonemes, which distinguish different words from each other, are the smallest units of sound in speech. For example, /p/ and /b/ are different phonemes; they can be used to distinguish between pairs of words such as pin and bin. Phonemes are different from the letters of the alphabet, in that they represent sounds rather than written language. The English language (with focus placed on the accent closely associated with the Southeast of England called Received Pronunciation or RP here) consists of 44 phonemes (Table A1). And, although these units of speech sound do not themselves carry meaning, they form the building blocks of the higher-level, meaningful language units (McRoberts, 2008) that are stored in memory. Their various combinations constitute the thousands of words in common usage by the native speaker of English in everyday conversations.

Consonants	Vowels
/p/ - pin	/iː/ - eat
/b/ - bed	/ɪ/ - it
/t/ - tune	/e/ - elf
/d/ - deer	/æ/ - act
/k/ - cat	/ɑː/ - are
/g/ - gap	/ɒ/ - on
/f/ - fair	/ɔː/ - or
/v/ - vast	/ʊ/ - put
/θ/ - thin	/uː/ - cool
/ð/ - then	/ʌ/ - other
/s/ - sun	/ɜː/ - earn
/z/ - zeal	/ə/ - another
/ʃ/ - ship	/eɪ/ - eight
/ʒ/ - genre	/əʊ/ - oh
/h/ - hat	/aɪ/ - eye

/tʃ/ - chair	/aʊ/ - out
/dʒ/ - job	/ɔɪ/ - oyster
/m/ - man	/ɪə/ - ear
/n/ - new	/eə/ - air
/ŋ/ - king	/ʊə/ - poor
/l/ - lot	
/r/ - ray	
/j/ - yes	
/w/ - well	

***Table A1**: List of received pronunciation consonants and vowels.*

Generally, for any sound to be produced, two physical conditions must be met: Firstly, air must be set in motion, and secondly, the air flow must meet some form of obstruction. In the English language, all sounds are produced when air from the lungs moves outwards (what is called a "pulmonic egressive airstream"). The pressure with which air can be squeezed out of the lungs to produce a sound can vary. For example, a greater pressure is necessary to produce stressed syllables. On its way out of the lungs, the airstream moves through the bronchi, and into the trachea and larynx. At the larynx, the air meets its first obstruction: the vocal folds (also known as vocal cords).

The vocal folds are comprised of flexible muscle and connective tissue and can be moved to be closer together or further apart. The space between them, called the glottis, is pushed to open and close via the egressive airstream coming from the lungs. This continuous vibration is called *voice*. The so-called *voiced* RP consonant speech sounds, such as /z/ and /v/ are generated via this procedure. In contrast, when air passes freely through an open glottis, *voiceless* consonants are produced (e.g. /s/ and /f/). For the complete list of voiced and voiceless consonants see Table A2.

To feel the vibrational difference between voiceless and voiced consonants one can simply place the hand on the throat while alternating between /s/ and /z/, or /f/ and /v/. These examples also make it clear, that presence or absence of voicing is not visible, from a lipreading point of view. As a result, a person relying solely on lipreading, would not be able to distinguish /s/ from /z/ or /f/ from /v/.

Voiced	/b, d, g, v, ð, z, ʒ, dʒ, m, n, ŋ, l, r, j, w/
Voiceless	/p, t, k, f, θ, s, ʃ, h, tʃ/

***Table A2**: Voicing state of received pronunciation consonants.*

RP consonant phonemes are also described, in addition to their voicing features, by the so-called manner of articulation. This term is used in phonetics, to refer to how sounds are produced. For instance, when the sound is being produced, does the air escape through the nose, or the sides of the tongue? Is the mouth open or closed? There are some aspects of manner of articulation that are visible, and available to lipread. The lipreader can see, for example, that the talker completely closes the mouth to produce /p/ and /b/. Other aspects are not visible, however. It is impossible, for instance, to see that air escapes through the nose for the production of the phoneme /n/, or that it escapes through the sides of the tongue for the production of /l/.

Another characteristic of a sound is the place of articulation – that is, where the sound is produced. Depending on the sound, place of articulation can help or challenge the lipreader. For example, place of articulation is visible for some consonants (e.g. the lips are visible and involved in the production of /p/ and /b/), while invisible for others (e.g. /k/ and /g/ are created at the back of the oral cavity when the tongue presses against the palate). Further, even for the case of consonants that have visible place of articulation, some have the same place of articulation (are “homorganic”), and thus easily distinguishable (e.g. /p/ and /b/). These differ on the timing of the onset of vocal fold activity relative to the opening of the lips. For example, in the case of /pa/ versus /ba/, the onset of vocal fold vibration follows the /p/ sound by a 70 ms gap before accompanying the vowel /a/, whereas for /ba/, the vibration of the vocal folds follows /b/ with a shorter gap of 20 ms (Schnupp et al., 2019).

All English vowels are voiced. They also differ in terms of lip shape and tongue involvement. For example, the lips are to a small extent spread when the /e/ in bed is produced but are rounded for the production of the /uː/ in cool. Thus, lip shape can be helpful, from a lipreading point of view – /ə/, the most common RP vowel, however, has a neutral lip shape (e.g. ago, soda).

Appendix B: General linear modelling principles

The first step in the development of a statistical model is to decide what kind of model to use. In Chapter 6 of this thesis, I discussed the development of two statistical models of the linear mixed effects type. These were used to describe participant speech-in-noise performance, and audio-visual benefits. The discussion there delved into the modelling procedures without provision of sufficient background information for the unfamiliar reader to appreciate the suitability and inner workings of these models. The current appendix serves this purpose.

Broadly speaking, linear mixed effects models are similar to plain linear models (such as linear regression), but with an added twist that provides them with more power and flexibility in their applications. But, before these are discussed, some terminology would be helpful, including illustrations and explanations.

A linear model consists of a *predicted* (also called *response*, or *dependent*) variable, and one, or several, *predictor* (or *explanatory*, or *independent*) variables. Usually, the predicted variable is denoted with “Y”, and the predictors with “xs” – x_1, x_2,... and so on depending on how many there are. Consider having collected a few observations “Yobserved” of the response variable. In statistical modelling formula terminology, and assuming there is only one predictor x, we write, Yobserved ~ x, to symbolically show that the predicted variable Yobserved is modelled as a linear function of the predictor x. More accurately, an "error" term, ε, is usually added to this formula to indicate that there are other "factors", beyond our predictor x, and beyond our experimental control, that may explain (the variation in) our Yobserved. Hence, we write Yobserved ~ x + ε. On the right-hand side of the formula, the model for Yobserved is essentially divided into two parts: one part is the error, which has no structure, and is probabilistic and random, and the other part is the predictor, which is structural, or fixed – another name for predictors is, following this logic, *fixed effects*.

The fixed effects part of the formula is virtually the same for both standard linear models and linear mixed effects models. Where the two differ is in how they deal with the random error part. Standard linear models treat the random part as a broad, unstructured, error term, absorbing everything that could not have experimentally been controlled for. Contrary to this, linear mixed effects models offer the capacity to give some structure to this random term, and they do it via way of addition of the so-called random effects – the "mixed" effects part of their name comes from the fact that they use a combination of fixed and random effects.

More on random effects later; for now, let's underline some linear modelling restrictions. To be applicable to a dataset, linear mixed effects models, like plain linear models, must meet certain linear model criteria. The first three, linearity, constant variance (or homoscedasticity), and normality have to do with, and can be assessed through, the so-called model *residuals*. I explain through these three assumptions first.

The use of symbols Yobserved ~ x + ε implies that what is sought is to formulate a model with equation Ymodelled = α + βx (where α is the intercept, and β the slope), to constitute our regression line. Consequently, Yobserved = α + βx + ε, or Yobserved = Ymodelled + ε.

The random, or error term, ε, is now called the residual, as it represents the difference, or distance, between the observed Y and the model's prediction for Y. It is computed like this: ε = Yobserved – Ymodelled, or equivalently, ε = Yobserved – α + βx. Thus, for each pair of values Yobserved, and Ymodelled, a residual can be computed. Figure B1 bellow provides a visual for this result.

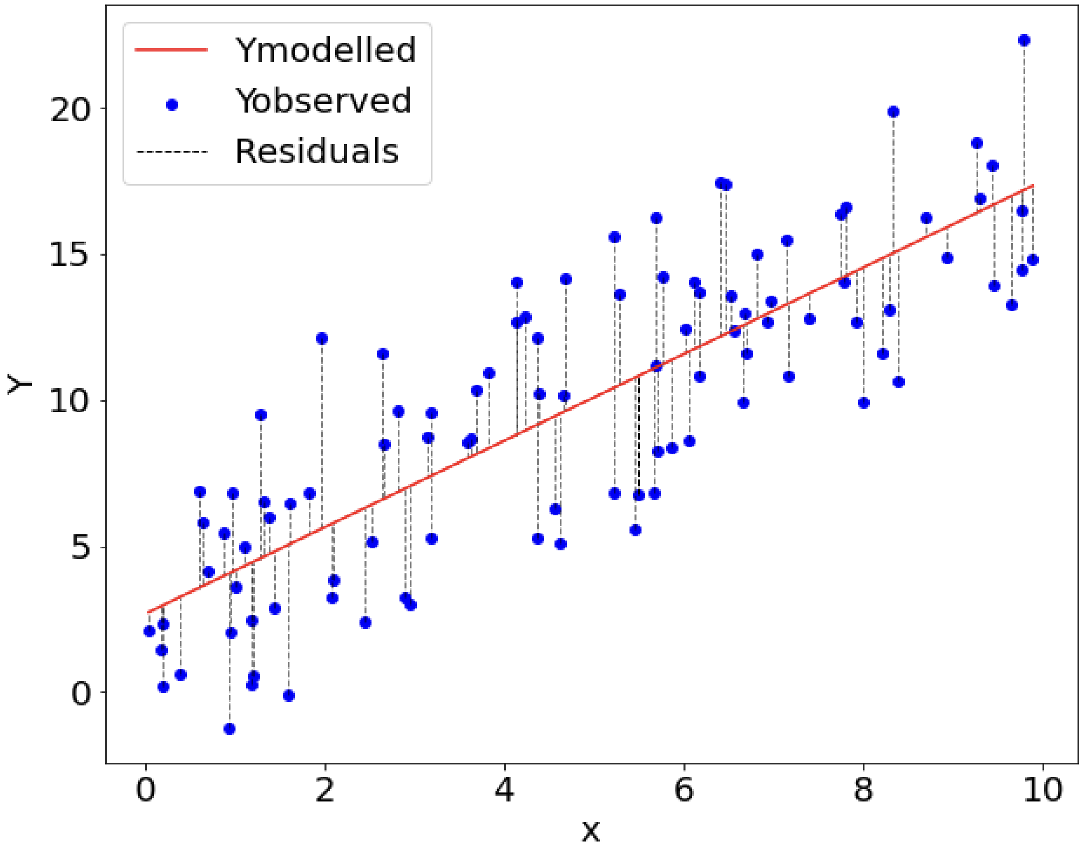


Figure B1*: An illustration of model residuals. The blue dots represent observations of the variable Y, here called Yobserved, for specific values of x. The red line represents the regression line obtained from fitting the formula Ymodel = α + β x to the dataset. The distances between Yobserved and Ymodelled (the blue dots and the red line), form the residuals of the model (grey vertical lines), and are calculated as Yobserved – Ymodelled.*

It is clear from this plot that the vertical grey lines, representing the residuals, are roughly similarly scattered around the modelled red line, throughout the range of the x-axis. This implies a linear relationship between Yobserved and the predictor x, and that the linear model used (Ymodel = α + βx) is potentially applicable to the data. This conclusion is usually assessed by computing the actual residuals (Yobserved – Ymodelled) and plotting them against the red line values (Ymodelled). If the data meet the criterion for linearity, it would be expected to see a similar scatter of the residuals around the horizontal line of zero, and across the range of Ymodelled values. This is shown for the above dataset, in Figure B2 (in what is called, the residual plot).

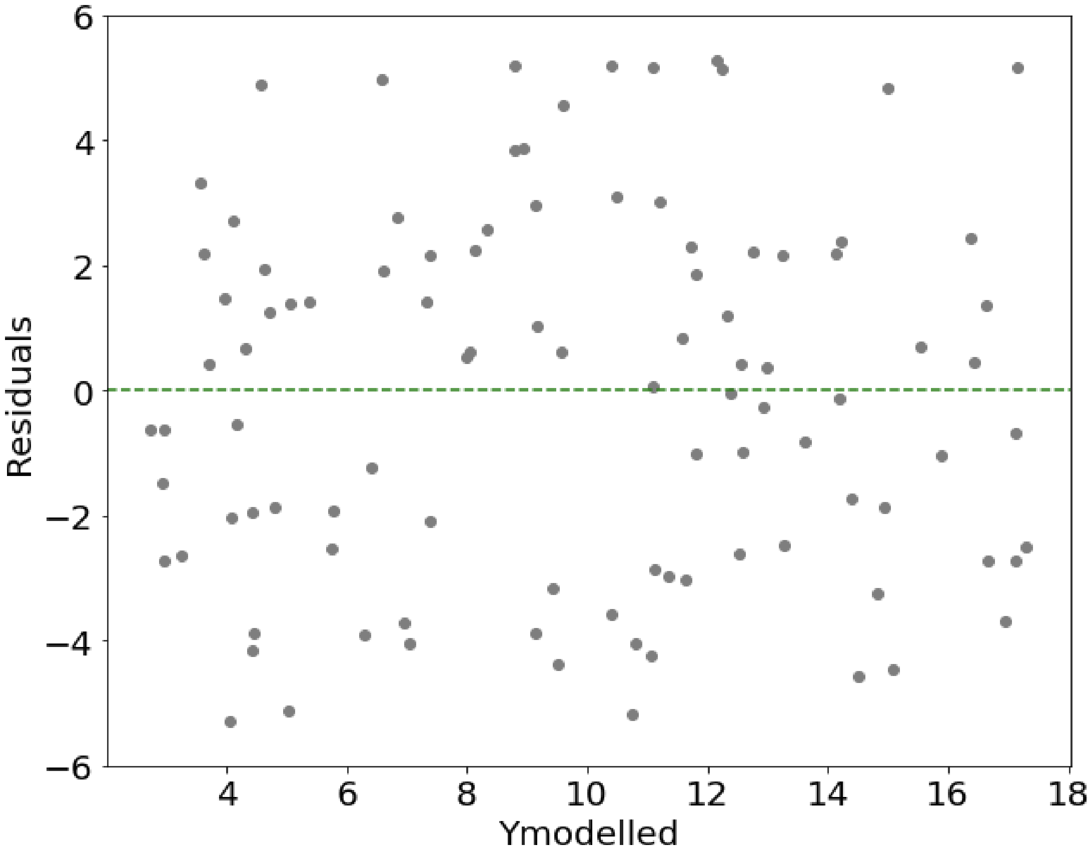


***Figure B2**: Residual plot for dataset shown in figure B1. Across the range of modelled values (Ymodelled), the residuals are similarly scattered around the horizontal line of zero, suggesting that the relationship between the predicted and predictor is linear.*

Had the relationship between Yobserved and x not been a linear one, as shown in Figure B3, a pattern would have been evident in the residual plot (Figure B4).

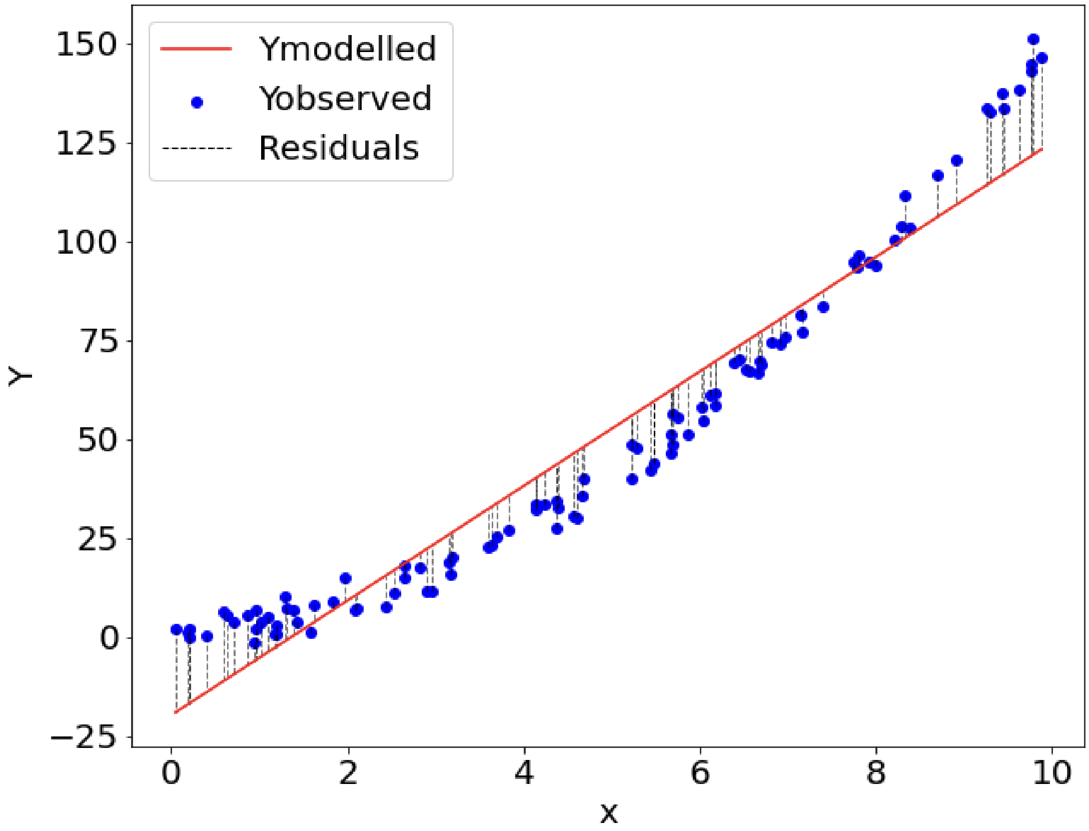


Figure B3*: A non-linear relationship between the variables Yobserved (shown in blue), and x is illustrated. A linear model, Ymodelled, shown in red, was forced upon the dataset. The residuals are, as before shown in grey vertical lines.*

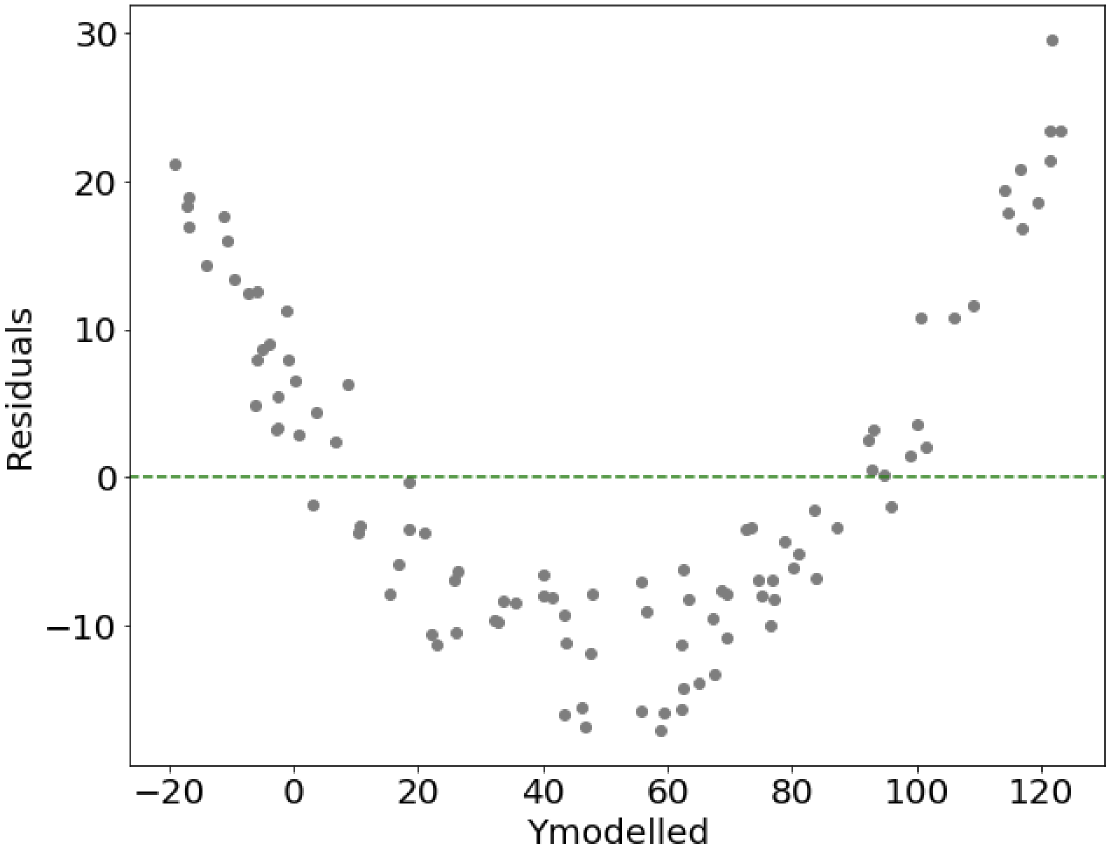


Figure B4*: Residual plot for the dataset shown in figure B3. The underlying non-linear nature of the data is clearly visible in this plot (compare to linear case of figure B2).*

The second linearity criterion, homoscedasticity, refers to the property of Yobserved values to be roughly equally scattered about the Ymodelled line across the range of x values. That is, the variance of Yobserved is constant – it does not change with increasing or decreasing values of x. Homoscedasticity also translates to the residual plot, and Figure B1 along with its corresponding residual plot in Figure B2, represent homoscedastic data. Heteroscedastic data, however, where the variance of Yobserved changes with changing values of the predictor x, result in patterned residual plots. See for example Figures B5, and B6 below, where the variance of Yobserved increases as x increases, resulting in a funnel-shaped residual plot.

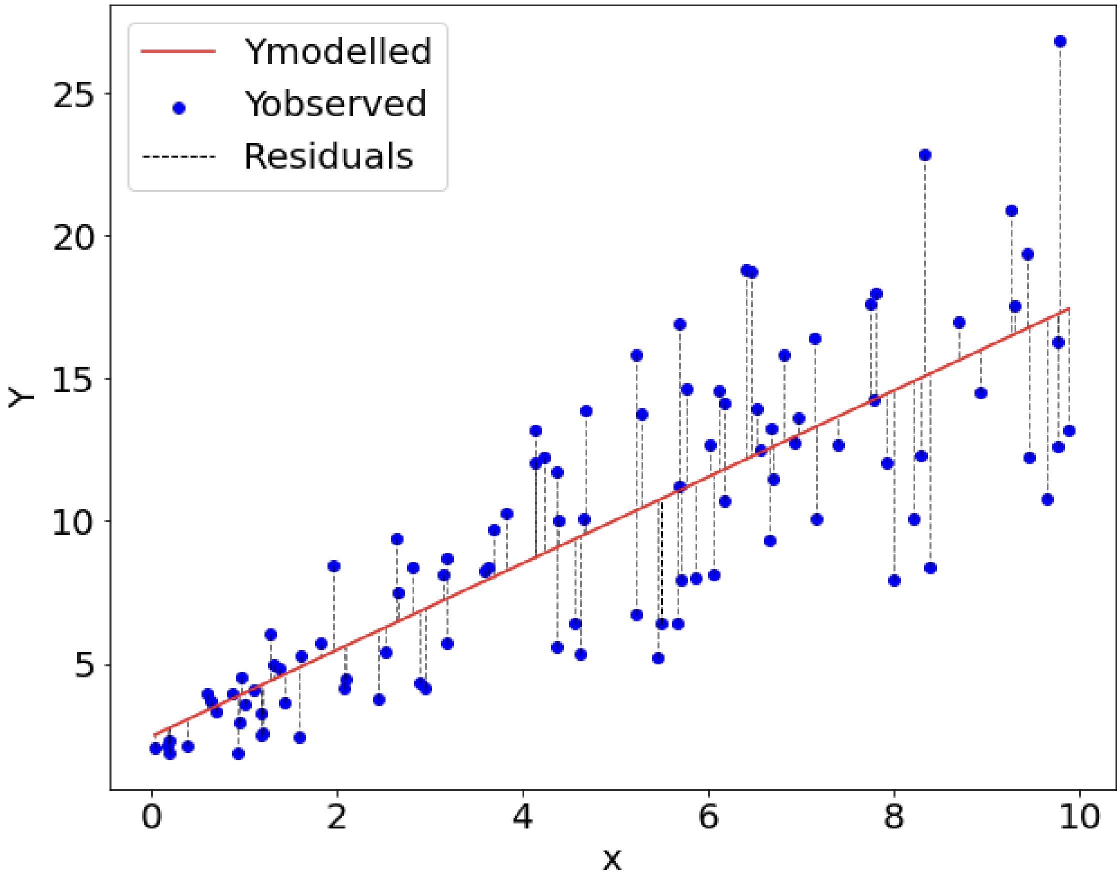


***Figure B5**: An illustration of a heteroscedastic dataset. The variance of Yobserved about the Ymodelled line, increases with increasing x.*

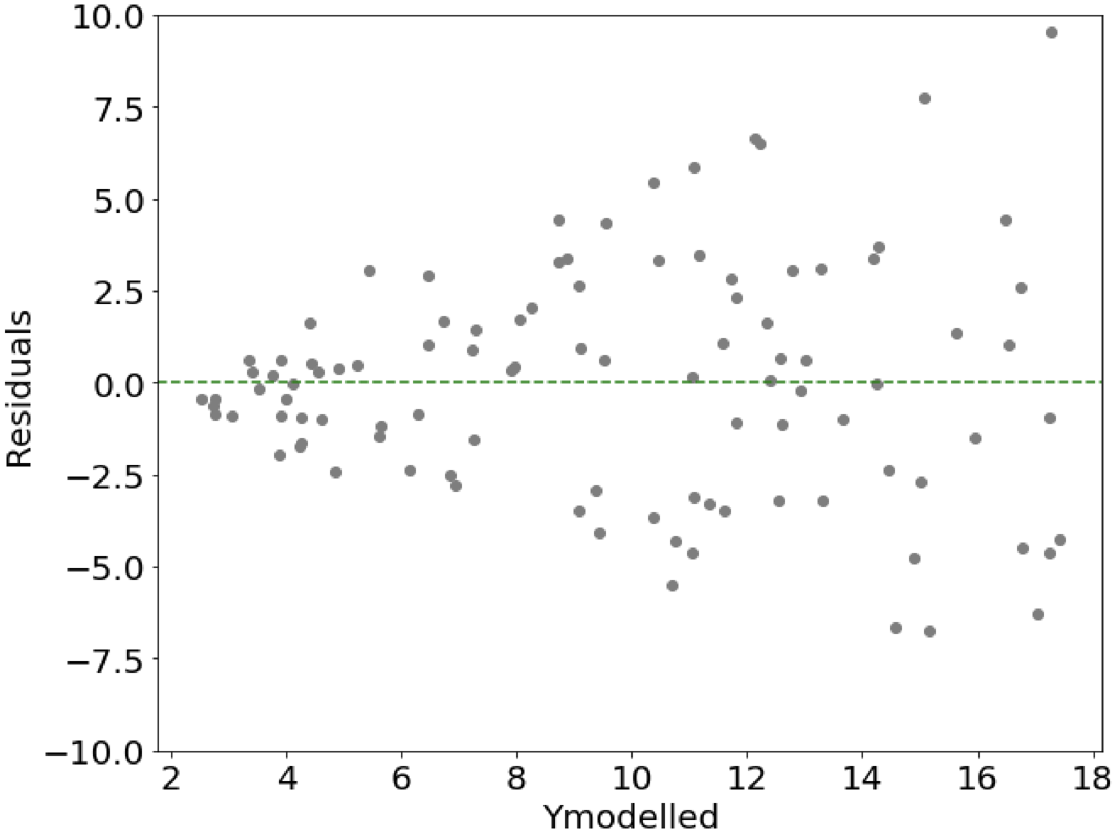


***Figure B6**: Residual plot for the dataset of figure B5. The heteroscedasticity of the dataset is evident here, resulting in a funnel-shaped plot.*

The last of the three criteria, normality, refers to the assumption that the model residuals are normally distributed. This is usually assessed with a visual inspection of a histogram of the residuals, or a so-called QQ-plot, (see figure B7, for data shown below for the residuals from figures B1 and B2). Normally distributed residuals result in a QQ-plot where the data roughly fall on a straight line.

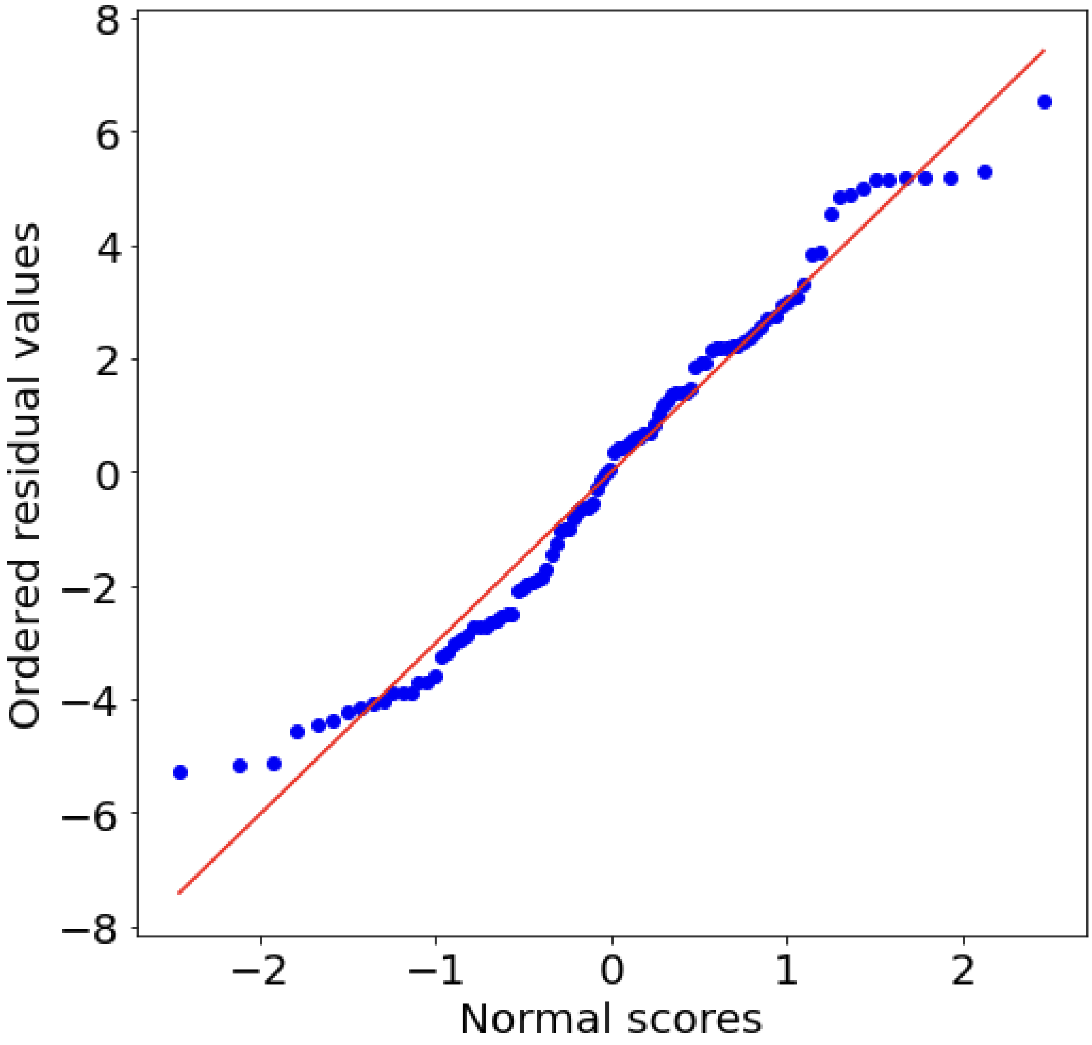


Figure B7*: QQ plot for the dataset of figure B1. Normally distributed residuals result in a QQ plot where the data (blue dots) fall roughly on the straight red line.*

The penultimate criterion for applicability of linear models is absence of co-linearity between predictors. This simply means that variables used as predictors should not be correlated (at least not substantially) to each other as, if that's the case, the correlated predictors "steal" each other's explanatory power and model interpretation becomes unreliable. Co-linearity between predictors can be easily assessed in R, after linear model fitting, via the variance inflation factor (VIF) function. The VIF of each predictor in a model is a measure of how much it correlates with all the other predictors used in the model. A VIF of 1 indicates no correlation, while a VIF between 1 and 5 indicates acceptable correlation. If the VIF is above 5, then it is generally advisable to remove the predictors responsible for it.

Finally, the last, and perhaps most important criterion for applicability of linear models is the criterion of independence between the observed data. In terms of the symbols used above, for a linear model to be applicable to the data, each Yobserved datum should be independent of the rest. That is, during the experimental procedure, one, and only one observation is allowed to be taken per participant, if a standard linear model is to be used. If multiple observations are taken from each participant, then observations taken from the same participant would not be independent of each other. This is because each participant has a slightly different, and personal, way with which they respond to the demands of the experimental tasks compared to the other participants. This individuality is reflected in all of a participant’s responses, violating the criterion of independence for all the data collected from them. This criterion of independence thus puts a significant restriction over the range of applicability of standard linear models. Thankfully, linear mixed effects models offer a nifty way for addressing this issue, and this brings the discussion back to the random effects.

It was mentioned earlier that the main difference between plain linear and linear mixed effects models lies in their treatment of the random term (the ε from Y ~ x + ε). It was stated that plain linear models assume an unstructured random term. Contrary to this, linear mixed effects models provide some structure to this term via the inclusion of the so-called random effects. It turns out that structuring the random term with random effects is also how linear mixed effects models address the issue of the independence criterion. Using the example of taking multiple observations from the same participants to illustrate this, the modeler can instruct the model that they expect that each participant is different in their baseline responses. For an example from my own experiments, if the modelled variable is the participants' AV benefit, and several observations of AV benefit have been taken from each participant, as it was the case in this study, a linear mixed effects model could be applied to the data and instructed that each participant is expected to have a different baseline AV benefit. This effectively allows the direct modelling, and resolution, of the confounding inter-dependency between data points collected from the same participants.

There are, actually, two flavours of random effects that can be included in a linear mixed effects model: Random intercepts, and random slopes. Usually, the first step in structuring the random effects of the model is to include the random intercept – this is effectively the modelling of different baseline responses for each of the participants (or any other form of grouping of the data), alluded to in the previous paragraph. Then, random slopes can be added. Using again the AV benefit example from above: If the random intercept instructs the model that each participant has a different baseline AV benefit, then the addition of random slopes instructs the model that the way in which AV benefit changes for each participant, when some predictor changes, is also different - i.e. the AV benefit of each participant starts at a different value, and changes differently too, with the change in some predictor. For example, in this study, participants ran three runs of each of the vCCRMn task’s conditions, and thus had three AV benefits – one computed for each run. The variable "Run" then, could be used as a predictor of AV benefit in the model, assuming that there was some sort of learning or detrimental effect which influenced a participant's AV benefit score with each new run they ran. But it might also be assumed that this learning or detrimental effect, the effect of "Run", differed for each participant, and the way to assess this would be to include Run as a random slope in addition to including it as a predictor.

It is straightforward to include several random effects in a model in R. This is done, essentially, by extending the Y ~ x + ε formula. For the AV benefit example, with Participants as a random intercept, and using Run as the sole predictor and sole random slope for simplicity, we have AV benefit ~ Run + (1 + Run | Participants). The terms within the brackets represent the random effects, and the terms outside the brackets represent the fixed effects. Within the brackets, and to the right of "|" is placed the random intercept, and to the left, the random slopes. Had we not included a random slope of Run and had just included the random intercept of Participant, the format would have been (1|Participant).

To include further random slopes, we simply add more terms to the "1" on the left side of "|"; e.g. (1 + Run + Other predictors | Participants).

Including complex random effects provide linear mixed effects models with power and flexibility to model a wide range of datasets but, unfortunately, they are not without cost. Building a complex model structure with several random effects would require the model to estimate a large number of parameters, which would require an even larger dataset. Further, with added complexity, the interpretability of the model goes down, and, finally, we may also run into the problem of overfitting the dataset (i.e. tailoring the model to such a great extent to the specific dataset that it fits the data almost perfectly but cannot be used to make any predictions beyond the dataset itself) and, thus, losing predictive power beyond it.

With regards to model complexity, one thing I haven't mentioned so far, that could be additionally incorporated into a linear mixed effects model (and standard linear model), and increase the model's complexity, are interactions between predictors. Interaction terms are also simple to include in R: Given a predicted variable Y and two predictors x1 and x2, to include the interaction term in the model, we write Y ~ x1 + x2 + x1*x2. Where the interaction term is x1*x2. Interpreting interactions is easier to show with an example. Borrowing the AV benefit example from above once more, and given two predictors, Lipreading ability and Age, the model without an interaction term would be AV benefit ~ Lipreading ability + Age. This model suggests that the variable Lipreading ability influences AV benefit equally, across all ages. But would that be a justifiable assumption to make? Potentially the effect of Lipreading ability is different, depending on whether the participant is younger, or older. Thus, the effect of Lipreading ability, on the AV benefit, could depend on the age of the participant. And this is exactly what adding an interaction term between Lipreading ability and Age would capture. Interaction terms are members of the set of fixed effects included in the model.

These are, in a nutshell, the principles behind the models employed in Chapter 6.

Bibliography

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research*, *229*(1–2), 225–236. https://doi.org/10.1016/j.heares.2007.01.011

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom–up and top–down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1072–1089. https://doi.org/10.1037/0096-1523.27.5.1072

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Alais, D., & Burr, D. (2019). *Cue Combination Within a Bayesian Framework* (pp. 9–31). https://doi.org/10.1007/978-3-030-10461-0_2

Anderson Gosselin, P., & Gagné, J.-P. (2011). Older Adults Expend More Listening Effort Than Young Adults Recognizing Speech in Noise. *Journal of Speech, Language, and Hearing Research*, *54*(3), 944–958. https://doi.org/10.1044/1092-4388(2010/10-0069)

Antonopoulou, E., & Pagoni-Tetlow, S. (2004). *The Sounds of English: Units and Patterns*. JRT Systems.

Arnold, P. (1997). The Structure and Optimization of Speechreading. *Journal of Deaf Studies and Deaf Education*, *2*(4). https://doi.org/10.1093/oxfordjournals.deafed.a014326

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*. https://doi.org/10.1348/000712601162220

Atilgan, H., & Bizley, J. K. (2021). Training enhances the ability of listeners to exploit visual information for auditory scene analysis. *Cognition*, *208*. https://doi.org/10.1016/J.COGNITION.2020.104529

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., & Bizley, J. K. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. *Neuron*, *97*(3), 640-655.e4. https://doi.org/10.1016/j.neuron.2017.12.034

Auer, E. T. (2010). Investigating Speechreading and Deafness. *Journal of the American Academy of Audiology*. https://doi.org/10.3766/jaaa.21.3.4

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/1092-4388(2007/080)

Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, *13*(3), 108–112. https://doi.org/10.3109/03005367909078884

Bernstein, J. G. W., & Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *125*(5), 3358–3372. https://doi.org/10.1121/1.3110132

Bernstein, L. E. (2018). Response Errors in Females' and Males' Sentence Lipreading Necessitate Structurally Different Models for Predicting Lipreading Accuracy. *Language Learning*, *68*(S1), 127–158. https://doi.org/10.1111/lang.12281

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1-4 SPEC. ISS.), 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Bernstein, L. E., Jordan, N., Auer, E. T., & Eberhardt, S. P. (2022). Lipreading: A Review of Its Continuing Importance for Speech Recognition With an Acquired Hearing Loss and Possibilities for Effective Training. *American Journal of Audiology*, *31*(2), 453–469. https://doi.org/10.1044/2021_AJA-21-00112

Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *The Journal of the Acoustical Society of America*, *120*(3), 1506–1516. https://doi.org/10.1121/1.2234849

Bianco, R., Mills, G., de Kerangal, M., Rosen, S., & Chait, M. (2021). Reward Enhances Online Participants' Engagement With a Demanding Auditory Task. *Trends in Hearing*, *25*, 233121652110259. https://doi.org/10.1177/23312165211025941

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a Test of Speech Perception in Noise. *Journal of Speech, Language, and Hearing Research*, *27*(1), 32–48. https://doi.org/10.1044/jshr.2701.32

Billings, C. J., Olsen, T. M., Charney, L., Madsen, B. M., & Holmes, C. E. (2024). Speech-in-Noise Testing: An Introduction for Audiologists. *Seminars in Hearing*, *45*(01), 055–082. https://doi.org/10.1055/s-0043-1770155

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. In *Nature Reviews Neuroscience* (Vol. 14, Issue 10, pp. 693–707). https://doi.org/10.1038/nrn3565

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. In *Trends in Neurosciences* (Vol. 39, Issue 2, pp. 74–85). Elsevier Ltd. https://doi.org/10.1016/j.tins.2015.12.007

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex*, *17*(9), 2172–2189. https://doi.org/10.1093/CERCOR/BHL128

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of Background Talkers. *Trends in Hearing*. https://doi.org/10.1177/2331216519837866

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.428288

Boothroyd, A., Hanin, L., & Hnath, T. (1985). *A sentence test of speech perception: reliability, set equivalence, and short term learning and short term learning*.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. https://doi.org/10.7551/MITPRESS/1486.001.0001

British Society of Audiology. (2018). *Pure-tone air-conduction and bone- conduction threshold audiometry with and without masking*. Https://Www.Thebsa.Org.Uk/Wp-Content/Uploads/2023/10/OD104-32-Recommended-Procedure-Pure-Tone-Audiometry-August-2018-FINAL-1.Pdf.

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, *47*(3), 191–196. https://doi.org/10.1037/h0054182

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.1408946

Budinger, E., Heil, P., Hess, A., & Scheich, H. (2006). Multisensory processing via early cortical stages: Connections of the primary auditory cortical field with other sensory systems. *Neuroscience*, *143*(4), 1065–1083. https://doi.org/10.1016/j.neuroscience.2006.08.035

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of Auditory Cortex During Silent Lipreading. *Science*, *276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

Campbell, R., Mohammed, T. E., & Macsweeney, M. (2003). *Developing the TAS: Individual differences in silent speechreading, reading and phonological awareness in deaf and hearing speechreaders*. https://www.researchgate.net/publication/252626157

Campbell, R., & Mohammed, T.-J. E. (2010). *Speechreading for information gathering: A survey of scientific sources 1*.

Cappelloni, M. S., Mateo, V. S., & Maddox, R. K. (2023). Performance in an Audiovisual Selective Attention Task Using Speech-Like Stimuli Depends on the Talker Identities, But Not Temporal Coherence. *Trends in Hearing*, *27*. https://doi.org/10.1177/23312165231207235

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7). https://doi.org/10.1371/journal.pcbi.1000436

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

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Development of the Connected Speech Test (CST). *Ear and Hearing*, *8* (SUPPLEMENT), 119s. https://doi.org/10.1097/00003446-198710001-00010

Dancer, J., Krain, M., Thompson, C., Davis, P., & al, et. (1994). A cross-sectional investigation of speechreading in adults: Effects of age, gender, practice, and education. *The Volta Review*, *96*(1), 31–40.

Darwin, C. J., & Carlyon, R. P. (1995). Auditory Grouping. In *Hearing*. https://doi.org/10.1016/b978-012505626-7/50013-3

Darwin, C. J., & Hukin, R. W. (1997). Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America*, *102*(4), 2316–2324. https://doi.org/10.1121/1.419641 de Dieuleveult, A. L., Siemonsma, P. C., van Erp, J. B. F., & Brouwer, A. M. (2017). Effects of aging in multisensory integration: A systematic review. In *Frontiers in Aging Neuroscience* (Vol. 9, Issue MAR). Frontiers Research Foundation. https://doi.org/10.3389/fnagi.2017.00080

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, *61*(2), 317–329. https://doi.org/10.1016/J.NEURON.2008.12.005

Ellis, T., MacSweeney, M., Dodd, B., & Campbell, R. (n.d.). *TAS: A NEW TEST OF ADULT SPEECHREADING. DEAF PEOPLE REALLY CAN BE BETTER SPEECHREADERS*.

Erber, N. P. (1975). Auditory visual perception of speech. *Journal of Speech and Hearing Disorders*. https://doi.org/10.1044/jshd.4004.481

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical Evidence of Multimodal Integration in Primate Striate Cortex. *The Journal of Neuroscience*, *22*(13), 5749–5759. https://doi.org/10.1523/JNEUROSCI.22-13-05749.2002

Fishman, Y. I., Micheyl, C., & Steinschneider, M. (2016). Neural Representation of Concurrent Vowels in Macaque Primary Auditory Cortex. *Eneuro*, *3*(3), ENEURO.0071-16.2016. https://doi.org/10.1523/ENEURO.0071-16.2016

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *109*(5), 2112–2122. https://doi.org/10.1121/1.1354984

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256. https://doi.org/10.1121/1.1689343

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America*, *109*(5), 2272–2275. https://doi.org/10.1121/1.1362687

Grant, K. W., & Bernstein, J. G. W. (2019). *Toward a Model of Auditory-Visual Speech Intelligibility* (pp. 33–57). https://doi.org/10.1007/978-3-030-10461-0_3

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.1288668

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory–visual consonant recognition. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.417950

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.422788

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.1836832

Hornby, A. (2000). *Oxford Advanced Learner's Dictionary* (6th edition). Oxford University Press.

Humes, L. E. (2019). Examining the Validity of the World Health Organization's Long-Standing Hearing Impairment Grading System for Unaided Communication in Age-Related Hearing Loss. *American Journal of Audiology*, *28*(3S), 810–818. https://doi.org/10.1044/2018_AJA-HEAL18-18-0155

Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. *Cognition*, *82*(3), B113–B122. https://doi.org/10.1016/S0010-0277(01)00161-5

Izumi, A. (2003). Effect of temporal separation on tone-sequence discrimination in monkeys. *Hearing Research*, *175*(1–2), 75–81. https://doi.org/10.1016/S0378-5955(02)00712-8

Jayakody, D. M. P., Friedland, P. L., Martins, R. N., & Sohrabi, H. R. (2018). Impact of Aging on the Auditory System and Related Cognitive Functions: A Narrative Review. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00125

Johnson, F. M., Hicks, L. H., Goldberg, T., & Myslobodsky, M. S. (1988). Sex differences in lipreading. *Bulletin of the Psychonomic Society*, *26*(2), 106–108. https://doi.org/10.3758/BF03334875

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual Modulation of Neurons in Auditory Cortex. *Cerebral Cortex*, *18*(7), 1560–1574. https://doi.org/10.1093/cercor/bhm187

Kidd, G., & Colburn, H. S. (2017). *Informational Masking in Speech Recognition* (pp. 75–109). https://doi.org/10.1007/978-3-319-51662-2_4

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *116*(4), 2395–2405. https://doi.org/10.1121/1.1784440

Kim, J., Karisma, S., Aubanel, V., & Davis, C. (2018). Investigating the Role of Familiar Face and Voice Cues in Speech Processing in Noise. *Interspeech 2018*, 2276–2279. https://doi.org/10.21437/Interspeech.2018-1812

Kochkin, S. (2000). MarkeTrak V: 'Why my hearing aids are in the drawer': The consumers' perspective. *The Hearing Journal*.

Kujala, T., Tervaniemi, M., & Schröger, E. (2007). The mismatch negativity in cognitive and clinical neuroscience: Theoretical and methodological considerations. *Biological Psychology*, *74*(1), 1–19. https://doi.org/10.1016/j.biopsycho.2006.06.001

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*(7), 961–967. https://doi.org/10.1080/17470210801908476

Lee, A. K. C., Maddox, R. K., & Bizley, J. K. (2019). *An Object-Based Interpretation of Audiovisual Processing* (pp. 59–83). https://doi.org/10.1007/978-3-030-10461-0_4

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.1912375

Loughrey, D. G., Kelly, M. E., Kelley, G. A., Brennan, S., & Lawlor, B. A. (2018). Association of Age-Related Hearing Loss With Cognitive Function, Cognitive Impairment, and Dementia. *JAMA Otolaryngology–Head & Neck Surgery*, *144*(2), 115. https://doi.org/10.1001/jamaoto.2017.2513

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11-14 years). *British Journal of Educational Psychology*, *70*(4), 505–518. https://doi.org/10.1348/000709900158272

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141. https://doi.org/10.3109/03005368709077786

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *ELife*, *4*. https://doi.org/10.7554/eLife.04995

McAdams, S., & Bertoncini, J. (1997). Organization and discrimination of repeating sound sequences by newborn infants. *The Journal of the Acoustical Society of America*, *102*(5), 2945–2953. https://doi.org/10.1121/1.420349

McRoberts, G. W. (2008). Speech Perception. *Encyclopedia of Infant and Early Childhood Development*, *1–3*, 244–253. https://doi.org/10.1016/B978-012370877-9.00154-7

Messaoud-Galusi, S., Hazan, V., & Rosen, S. (2011). Investigating Speech Perception in Children With Dyslexia: Is There Evidence of a Consistent Deficit in Individuals? *Journal of Speech, Language, and Hearing Research*, *54*(6), 1682–1701. https://doi.org/10.1044/1092-4388(2011/09-0261)

Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *JARO - Journal of the Association for Research in Otolaryngology*, *11*(4), 709–724. https://doi.org/10.1007/s10162-010-0227-2

Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques. *Neuron*, *48*(1), 139–148. https://doi.org/10.1016/j.neuron.2005.08.039

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. In *Journal of the Acoustical Society of America* (Vol. 82, Issue 6). https://doi.org/10.1121/1.395659

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105–129. https://doi.org/10.1037/h0055960

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clinical Linguistics & Phonetics*, *20*(7–8), 621–630. https://doi.org/10.1080/02699200500266745

Morris, N. L., Chaparro, A., Downs, D., & Wood, J. M. (2012). Effects of simulated cataracts on speech intelligibility. *Vision Research*, *66*, 49–54. https://doi.org/10.1016/j.visres.2012.06.003

Murthy, N. S. (2020). Lip-Reading Techniques: A Review. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *9*, 2. www.ijstr.org

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

NELKEN, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology*, *14*(4), 474–480. https://doi.org/10.1016/j.conb.2004.06.005

Nelken, I. (2008). Neurons and objects: the case of auditory cortex. *Frontiers in Neuroscience*, *2*(1), 107–114. https://doi.org/10.3389/neuro.01.009.2008

Nettelbeck, T., & Burns, N. R. (2010). Processing speed, working memory and reasoning ability from childhood to old age. *Personality and Individual Differences*, *48*(4), 379–384. https://doi.org/10.1016/j.paid.2009.10.032

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099. https://doi.org/10.1121/1.408469

Noorden, L. V. (1975). Temporal coherence in the perception of tone sequences.

Palmer E., S. (1999). *Vision Science: Photons to Phenomenology*.

Pasupathy, A. (2015). The neural basis of image segmentation in the primate brain. *Neuroscience*, *296*, 101–109. https://doi.org/10.1016/j.neuroscience.2014.09.051

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. In *Cortex*. https://doi.org/10.1016/j.cortex.2015.03.006

Pepper, J. L., & Nuttall, H. E. (2023). Age-Related Changes to Multisensory Integration and Audiovisual Speech Perception. In *Brain Sciences* (Vol. 13, Issue 8). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/brainsci13081126

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608. https://doi.org/10.1121/1.412282

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)

Rahne, T., Böckmann, M., von Specht, H., & Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research*, *1144*(1), 127–135. https://doi.org/10.1016/J.BRAINRES.2007.01.074

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In *Hearing by Eye: The Psychology of Lip-reading*.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1996). The structure of normal human attention: The Test of Everyday Attention. *Journal of the International Neuropsychological Society*, *2*(6), 525–534. https://doi.org/10.1017/S1355617700001697

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.4794379

Rosenblum, L. D. (2008). *Speech Perception as a Multimodal Phenomenon*.

Saleh, S. M., Saeed, S. R., & Vickers, D. (2023). Test-Retest Reliability of the Coordinate Response Measure in Adults with Normal Hearing or Cochlear Implants. *Audiology and Neurotology*, *28*(2), 84–93. https://doi.org/10.1159/000521466

Schnupp, J., Nelken, I., & King, A. J. (2019). Auditory Neuroscience. In *Auditory Neuroscience*. https://doi.org/10.7551/mitpress/7942.001.0001

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. In *Trends in Neurosciences* (Vol. 34, Issue 3, pp. 114–123). https://doi.org/10.1016/j.tins.2010.11.002

Shamma, S., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., Pressnitzer, D., Yin, P., & Xu, Y. (2013). *Temporal Coherence and the Streaming of Complex Sounds* (pp. 535–543). https://doi.org/10.1007/978-1-4614-1590-9_59

Sharma, S., Tripathy, R., & Saxena, U. (2016). Critical appraisal of speech in noise tests: a systematic review and survey. *International Journal of Research in Medical Sciences*, *5*(1), 13. https://doi.org/10.18203/2320-6012.ijrms20164525

Shinn-Cunningham, B., Best, V., & Lee, A. K. C. (2017). *Auditory Object Formation and Selection* (pp. 7–40). https://doi.org/10.1007/978-3-319-51662-2_2

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-Visual Speech Perception and Auditory-Visual Enhancement in Normal-Hearing Younger and Older Adults. *Ear and Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Stacey, P. C., Kitterick, P. T., Morris, S. D., & Sumner, C. J. (2016). The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure. *Hearing Research*. https://doi.org/10.1016/j.heares.2016.04.002

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. https://doi.org/10.1098/rstb.1992.0009

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, *69*(1), 136–152. https://doi.org/10.3758/BF03194460

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. In *Trends in Cognitive Sciences* (Vol. 14, Issue 9, pp. 400–410). https://doi.org/10.1016/j.tics.2010.06.008

Taylor, B. (2003). Speech-in-noise tests. *The Hearing Journal*, *56*(1), 40. https://doi.org/10.1097/01.HJ.0000293000.76300.ff

Tiitinen, H. T., Sinkkonen, J., Reinikainen, K., Alho, K., Lavikainen, J., & Näätänen, R. (1993). Selective attention enhances the auditory 40-Hz transient response in humans. *Nature*, *364*(6432), 59–60. https://doi.org/10.1038/364059a0

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual Integration and Lipreading Abilities of Older Adults with Normal and Impaired Hearing. *Ear & Hearing*, *28*(5), 656–668. https://doi.org/10.1097/AUD.0b013e31812f7185

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389. https://doi.org/10.1037/pag0000094

Walden, B. E., Surr, R. K., Cord, M. T., & Dyrlund, O. (2004). Predicting Hearing Aid Microphone Preference in Everyday Listening. In *J Am Acad Audiol* (Vol. 15).

Wayne, R. V., & Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *Journal of Experimental Psychology: Applied*, *18*(4), 419–435. https://doi.org/10.1037/a0031042

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195092509.001.0001

Wisniewski, A. B., & Hulse, S. H. (1997). Auditory scene analysis in European starlings (Sturnus vulgaris): Discrimination of song segments, their segregation from multiple and reversed conspecific songs, and evidence for conspecific song categorization. *Journal of Comparative Psychology*, *111*(4), 337–350. https://doi.org/10.1037/0735-7036.111.4.337

Woodhouse, L., Hickson, L., & Dodd, B. (2009). Review of visual speech perception by hearing and hearing-impaired people clinical implications. *International Journal of Language and Communication Disorders*, *44*(3), 253–270. https://doi.org/10.1080/13682820802090281

Yuan, Y., Lleo, Y., Daniel, R., White, A., & Oh, Y. (2021). The Impact of Temporally Coherent Visual Cues on Speech Perception in Complex Auditory Environments. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.678029

Yuan, Y., Wayland, R., & Oh, Y. (2020). Visual analog of the acoustic amplitude envelope benefits speech perception in noise. *The Journal of the Acoustical Society of America*, *147*(3), EL246–EL251. https://doi.org/10.1121/10.0000737

Zhao, S., Bury, G., Milne, A., & Chait, M. (2019). Pupillometry as an Objective Measure of Sustained Attention in Young and Older Listeners. *Trends in Hearing*, *23*, 233121651988781. https://doi.org/10.1177/2331216519887815

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer New York. https://doi.org/10.1007/978-0-387-87458-6