Improving the understanding and diagnosis of Auditory Processing Disorder (APD) in Children

Shiran Koifman

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at University College London

March 15 2021

For Yihui Xie

Acknowledgements

This is where you will normally thank your advisor, colleagues, family and friends, as well as funding and institutional support. In our case, we will give our praises to the people who developed the ideas and tools that allow us to push open science a little step forward by writing plain-text, transparent, and reproducible theses in R Markdown.

We must be grateful to John Gruber for inventing the original version of Markdown, to John MacFarlane for creating Pandoc (http://pandoc.org) which converts Markdown to a large number of output formats, and to Yihui Xie for creating `knitr` which introduced R Markdown as a way of embedding code in Markdown documents, and `bookdown` which added tools for technical and longer-form writing.

Special thanks to Chester Ismay, who created the `thesisdown` package that helped many a PhD student write their theses in R Markdown. And a very special tahnks to John McManigle, whose adaption of Sam Evans' adaptation of Keith Gillow's original maths template for writing an Oxford University DPhil thesis in LaTeX provided the template that I adapted for R Markdown.

Finally, profuse thanks to JJ Allaire, the founder and CEO of RStudio, and Hadley Wickham, the mastermind of the tidyverse without whom we'd all just given up and done data science in Python instead. Thanks for making data science easier, more accessible, and more fun for us all.

Ulrik Lyngs
Linacre College, Oxford
2 December 2018

Abstract

This *R Markdown* template is for writing an Oxford University thesis. The template is built using Yihui Xie's `bookdown` package, with heavy inspiration from Chester Ismay's `thesisdown` and the `OxThesis` LaTeX template (most recently adapted by John McManigle).

This template's sample content include illustrations of how to write a thesis in R Markdown, and largely follows the structure from this R Markdown workshop.

Congratulations for taking a step further into the lands of open, reproducible science by writing your thesis using a tool that allows you to transparently include tables and dynamically generated plots directly from the underlying data. Hip hooray!

Contents

List of Figures

List of Tables

List of Abbreviations

1-D, 2-D . . . One- or two-dimensional, referring in this thesis to spatial dimensions in an image.

Otter . . . . . One of the finest of water mammals.

Hedgehog . . . Quite a nice prickly friend.

Introduction

Welcome to the *R Markdown* Oxford University thesis template. This sample content is adapted from `thesisdown` and the formatting of PDF output is adapted from the OxThesis LaTeX template. Hopefully, writing your thesis in R Markdown will provide a nicer interface to the OxThesis template if you haven't used TeX or LaTeX before. More importantly, using *R Markdown* allows you to embed chunks of code directly into your thesis and generate plots and tables directly from the underlying data, avoiding copy-paste steps. This will get you into the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build upon your results down the road.

Using LaTeX together with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may never have had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities.

Speech-in-noise in children

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* allows you to read in your data, analyze it and to visualize it using **R**, **Python** or other languages, and provide documentation and commentary on the results of your project.
Further, it allows for results of code output to be passed inline to the commentary of your results. You'll see more on this later, focusing on **R**. If you are more into

Python or something else, you can still use *R Markdown* - see 'Other language engines' in Yihui Xie's *R Markdown: The Definitive Guide*.

APD definition

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about reproducibility in research can benefit from using *R Markdown*. If you are working in 'softer' fields, the user-friendly nature of the *Markdown* syntax and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should still make it of great benefit to your thesis project.

Diagnosis

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about reproducibility in research can benefit from using *R Markdown*. If you are working in 'softer' fields, the user-friendly nature of the *Markdown* syntax and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should still make it of great benefit to your thesis project.

Binaural and spatial listening in APD

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about reproducibility in research can benefit from using *R Markdown*. If you are working in 'softer' fields, the user-friendly nature of the *Markdown* syntax and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should still make it of great benefit to your thesis project.

Summary

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about reproducibility in research can benefit from using *R Markdown*. If you are working in 'softer' fields, the user-friendly nature of the *Markdown* syntax and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should still make it of great benefit to your thesis project.

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit...

There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain...

— Cicero's *de Finibus Bonorum et Malorum.*

1

Binaural listening: interripted and alternated speech-in-noise in adults

Contents

1.1 Influence of distractor type on IM

1.1.1 Introduction

Communication in adverse listening situations where the target speech is incomplete or distorted is a typical everyday occurrence. Often, the sound source of interest is

masked by nearby interfering sounds (e.g., traffic noise or competing talkers) or degraded (e.g., due to reverberations, transmission artefacts or filtering). Remarkably however, listeners can often maintain high speech intelligibility even when large portions of the speech signal are physically missing or entirely masked by other sounds (Başkent et al., 2016; Miller & Licklider, 1950). This phenomenon is, among other things, attributed to the redundant characteristics of speech in the spectral and the temporal domain, enabling the listener to piece together short glimpses of the target signal to achieve high speech perception (i.e., "glimpsing theory"; Cooke, 2006). The way our auditory system overcomes such impoverished listening conditions is not well understood. One of the main obstacles when trying to answer this question is the large variation in performance across listeners, in particular in more ecological listening scenarios with several competing talkers with different complex spectro-temporal properties (Surprenant & Watson, 2001). In many cases, such individual differences cannot be explained by hearing sensitivity as measured with pure-tone-audiogram (Humes & Dubno, 2010; Kidd & Humes, 2012). Individual differences may arise from variations in the listeners' auditory processing abilities or their abilities to make use of perceptual acoustic and linguistic information (Pichora-Fuller & Singh, 2006; Surprenant & Watson, 2001). In addition, there is an increasing amount of evidence suggesting that variability in speech perception may be in part attributed to variations in cognitive abilities, especially in adverse listening conditions where the distractor is speech or speech-like (see review by Akeroyd, 2008; Arlinger et al., 2009; Humes et al., 2013; Kidd & Humes, 2012; van Esch et al., 2013). Understanding what causes certain groups of listeners to experience listening difficulties under challenging listening situations can help us finding better intervention plans or treatments that fit to their individual needs. Moreover, we can use this knowledge to improve currently used speech recognition and speech enhancement techniques. However, isolating and quantifying the contribution of the different mechanisms involved throughout the auditory system is challenging.

The present paper aims to investigate the utility of a novel speech-on-speech listening task that appears to demand higher-level cognitive aspects of listening and may aid in disentangling the reasons why different groups of people experience difficulty in listening in noisy situations. In the task, target speech is interrupted and segmented at a fixed rate. The segments are then alternated between the two ears out-of-phase with an interrupted distractor which is alternated in a similar way, resulting in alternated segments of both signals between the two ears, with only one stimulus present in each ear at any given time. The task necessitates the listeners' ability to switch and sustain their attention on the target speech, while inhibiting the distractor segments, and to integrate the short-term auditory information between the two ears. A preliminary study (unpublished BSc thesis Akinseye, 2015) compared performance in the task across young (mean age: 24, range: 20 - 33 years old) and older adults (mean age: 63, range: 50-72 years old) with audiometrically normal hearing up to 4 kHz. Normal cognitive skills were controlled for the older group using a standard screening test (MoCA; Nasreddine et al., 2005). Interestingly, while no significant difference in speech intelligibility was found between the young and older adults for a "standard" speech-in-noise test, there was a highly significant difference in performance between the groups, with older adults showing poorer intelligibility for the switching task when presented with connected speech as a distractor. These results suggest that the switching task may demand some higher-level cognitive aspects of listening that are not probed by more simple listening tasks. The objective here is to investigate different aspects of the task across normal hearing young adults. This includes examining the effect of distractor types (speech vs. non-speech); intelligibility of the speech distractors; and similarity between the target and the distractor, for same- and opposite-sex distractor talker configurations on the listeners' speech perception. In addition, test-retest reliability and reproducibility of the task's score is evaluated. To set the context, it is beneficial to review some aspects involved in speech perception in a 'Cocktail-party'-like environment (Cherry, 1953) as an effect of distractor

interference, interruption, and alternation.

i. Distractor interference

A considerable amount of literature was published supporting the idea that a distractor interference consists of at least two separate mechanisms, originating roughly at different physiological levels: "peripheral" and "central" (for an overview see Moore, 2012). Peripheral masking is equated to a distractor interference taking place at the basilar membrane and at the auditory nerve. Probably the most researched peripheral masking is often called *energetic masking* (EM; see Moore, 2012; Rosen et al., 2013). This is because EM has its origin from interactions of energy in the target and distractor signals at the same frequency bands, causing reduced audibility of the target signal. Another recently proposed type of peripheral interference is related to the distractor's amplitude modulations as opposed to its energy, hindering the detection of information-carrying amplitude modulations in the target signal due to within-frequency band interference (i.e., *modulation masking*, MM; Stone et al., 2012). Central masking is often referred as interference that cannot be attributed to EM or MM (as in spectro-temporal overlap between the target and the distractor), and is broadly termed *informational masking* (IM; Durlach et al., 2003; Kidd et al., 2002; Moore, 2012). IM reflects insufficient or non-optimal processing of the target information beyond the hearing organ, despite a sufficient audibility at the peripheral level.

The conceptualisation of IM can be drawn from attention theories and the auditory scene analysis model of auditory perception (ASA; Bergman, 1990). The term 'auditory object' refers to perceptual entity that is perceived as originating from a single physical sound source. When a listener tries to hear out a target speech from a mixture of competing talkers, the auditory system is thought to perform two tasks: segmenting the elements of the target from the competing speech (*segregation*) and integrating these elements across time into an elementary auditory

object (*streaming*). Auditory objects are parsed over time by grouping mechanisms, based on attributes such as similarity, proximity, and continuity of higher-level acoustic features such as pitch, timbre, spectral and temporal modulations, spatial location, syntax and semantic content. IM is linked by many psychoacoustic studies to perceptual *similarity* and *uncertainty* of the target with the distractor signal (e.g., Durlach et al., 2003; Kidd et al., 2002; Shinn-Cunningham, 2008; Watson, 1987). Based on object-formation theories (e.g., ASA; Bergman, 1990), Shinn-Cunningham (2008) posited a conceptual theory that takes into account both bottom-up processes (i.e., attributes that contribute to strength of the sound source) and top-down attention-related processes. It distinguishes between two types of IM, caused by failure of either (1) object formation, or (2) object selection. Failing object formation can occur by EM or MM interference or similarities between the target and the distractor, preventing bottom-up streaming and thus, resulting in a confusion between the two signals, e.g., when the target and the distractor originates from the same-sex talker with similar voice characteristics. On the other hand, failure of object selection can take place even when auditory objects were successfully formed and the different sources were successfully streamed. This can occur due to similarities between the target and the distractor or uncertainty as to which object is the target stimulus that the listener should attend to. Failing to attend to the target object can also occur due to external factors that involuntarily pull away attention from the target, e.g., when a competing talker says your name (Moray, 1959).

Studies that look into the role of auditory grouping cues in speech-on-speech listening tasks often indicate the importance of voice characteristics (such as voice pitch or fundamental frequency; Bergman, 1990; Brungart et al., 2001; Darwin et al., 2003; Leclère et al., 2017; Scheffers, 1983; Shen & Souza, 2017), spatial separation (Best et al., 2011; Freyman et al., 1999), temporal fine structure (TFS; Moore, 2008), and semantic content (Brouwer et al., 2012; Calandruccio et al., 2010;

Van Engen & Bradlow, 2007) on speech intelligibility.

Pitch is generally defined as an attribute of auditory sensation that can be scaled from low to high (Moore, 2012). In complex harmonic tones (i.e., a series of sinusoids whose frequency is an integer multiple of the lowest frequency component—the 'fundamental'), the pitch corresponds to the frequency of the fundamental component and is typically termed as fundamental frequency (F0). Brokx and Nooteboom (1982) have shown that a difference of as little as 6% in F0 of two simultaneous vowels can considerably improve identification as opposed to when F0 is identical. In natural speech, pitch is dynamic and changes over time, arising from periodic vibration of the vocal cords which forms voiced speech sounds. These dynamic changes in F0 were shown to facilitate speech perception in noise (Binns & Culling, 2007; Laures & Weismer, 1999; Miller et al., 2010). Periodicity of a distractor was also shown to aid speech intelligibility when compared with an aperiodic distractor of vocoded speech (Steinmetzger & Rosen, 2015). Pitch varies fairly slowly during a course of a spoken sentence, independently for the target and the distractor signal. Pitch can help the listener to easily latch onto the target signal after being "lost" by the distractor or by occurrence of an unvoiced speech sound.

The perceptual advantage or 'release from masking' (MR) of normal hearing listeners for speech in the presence of a temporally fluctuating distractor (in amplitude) is believed to arise from the auditory system's sensitivity to temporal changes, enabling the listener to detect 'glimpses' or 'multiple looks' of the target speech from the mixed signal by making use of the distractor's temporal dips or gaps with favourable signal-to-noise ratio, SNR (Cooke, 2006; Howard-Jones & Rosen, 1993; Miller & Licklider, 1950; Moore, 2008; Shafiro et al., 2011; Shafiro et al., 2015; Stuart, 2008). The use of glimpses is believed to take place at both peripheral and central level where they work together rather than independently (Cooke, 2006; Moore, 2003). At the periphery (cochlea), spectro-temporal features are being used

to segregate and group sound sources in multiple-source environments (cf. ASA model by Bergman, 1990). In the time domain, an incoming sound is decomposed into rapidly changing TFS, following variations in formants and/or voice F0, and to slowly varying envelopes following the stimulus amplitude within frequency bands. (Moore, 2012; Pichora-Fuller & Souza, 2003). Several studies suggested that TFS cues play an important role in speech perception in a fluctuating noise, aiding "dip listening" (Hopkins & Moore, 2010; Moore, 2008). At a more central level, beyond the hearing organ, the pieces of glimpsed signal information are integrated into perceptual categories. This involves the use of different cognitive processing such as attention, working memory, executive language, and language skills.

Spatial separation between the target and the distractor can influence the effectiveness of a distractor, resulting in improved intelligibility, or spatial release from masking (SRM) of up to 16 dB (Freyman et al., 1999). This spatial advantage is attributed to both physical (EM) and perceptual (IM) factors. A speech spectrum noise (SSN) is often assumed to produce mainly EM and is therefore considered as a "pure" form of EM (Brungart et al., 2001)[1], producing a SRM between 5 to 10 dB when the target speech is presented to one ear and the noise is presented to the opposite ear, or when the noise is placed 90° azimuth away from the target talker on the horizontal plane (see Best et al., 2011). This improvement in intelligibility is attributed to binaural processing of interaural time differences (ITDs) and monaural better-ear effects that give rise to SNR advantages due to the acoustic head-shadowing effect. IM interference on the other hand elicits considerably larger SRM, ranging from 6 to circa 18 dB (cf. Best et al., 2011). This SRM benefit may not necessarily arise from monaural cues, but rather from binaural cues, that may aid in segregation of the sound sources. Nonetheless, quantifying the contribution of these two cues is difficult. Freyman et al. (1999) devised a clever way to separate IM processing while minimising better-ear (monaural) cues. Using the *precedence*

[1]However, recent work by Stone et al. (2012) and Stone and Moore (2014), suggests that most of the peripheral masking in SSN is caused by MM and not EM.).

effect (i.e., the use of early reflections for sound source localisation; Hirsh, 1950) they created a perceptual impression of spatial separation between the competing talker and the target speech, resulting in a significant improvement in intelligibility, without changing EM. In a series of experiments, Freyman and colleagues showed that this perceived spatial separation facilitated release from central (IM) processing for speech, no matter whether the competing speech was intelligible or not (e.g., reversed or unfamiliar speech), while listeners obtained only a negligible masking release for other non-speech distractors (e.g., SSN or amplitude modulated SSN; Freyman et al., 2001, 2004; Freyman et al., 1999). This perceptual separation is in part attributed to higher-level cognitive processing (rather than simple SNR advantage) that enables the listeners to segregate and focus their attention on the target talker. Moreover, Brungart and Iyer (2012) have investigated the mechanisms involved in a rather more complex listening situation where the competing talkers are symmetrically located at either side of the target. Based on the glimpsing model theory, Brungart and Iyer have demonstrated that the improved perception of the target signal may be explained by the listeners ability to make use of short-term glimpses that vary quickly across frequencies and switches rapidly across the two ears (so called 'better-ear glimpses'). Hence, this benefit in spatial separation appears to be ascribed to higher-level cognitive processes and may not be directly accredited to spatial processing at all.

Masking release from a distractor spoken in a language that is unfamiliar to the listeners is well documented in simple listening tasks where a mixture of the target and the competing talker is presented binaurally (e.g., Calandruccio et al., 2010; Freyman et al., 2001; Rhebergen et al., 2005). Although the magnitude of EM may differ between distractors spoken in a different language (due to language-related characteristics differences, such as phoneme frequency distribution), most of the masking release can be attributed to central IM processing, driven by the meaning or semantic content of the familiar speech. Nonetheless, the amount of masking release may differ depending on the origin of the linguistic interference (e.g., lexical,

sublexical, and/or prosodic level) and the task's difficulty (Brouwer et al., 2012; Calandruccio et al., 2014; Calandruccio et al., 2010; Van Engen & Bradlow, 2007). Isolating the different IM components in more adverse speech-on-speech listening situations that involves binaural or spatial processing can be challenging. Listeners intelligibility is typically unaffected by contralateral competing speech (e.g., Cherry, 1953; Drullman & Bronkhorst, 2000; Moray, 1959). This is because of strong spatial separation cues which facilitate IM release.

Freyman et al. (1999) showed that listeners' benefit from spatial separation even when the distractor's semantic content is eliminated (e.g., unfamiliar language), whereas they showed no masking release for non-speech SSN. Later studies proposed a clever way to break down this beneficiary masking release effect in dichotic listening by presenting an additional distractor in the ipsilateral target ear (Brungart & Simpson, 2002; Carlile & Corkhill, 2015). Brungart and Simpson (2002) showed that masking release from a contralateral distractor can fail when there is a high uncertainty between the distractor and the target streams in the ipsilateral (target) ear. Brungart and Simpson's task required the listeners ability to segregate the target and the distractor streams in the ipsilateral ear and so, in case uncertainty between the two streams is high, the listeners could reach the limit of their attentional resources. At the same time, if the contralateral distractor is "speechy" enough, this could potentially interfere with the listeners ability to use binaural cues, which consequently will impair their ability to ignore the contralateral distractor and thus result in poorer intelligibility.

Carlile and Corkhill (2015) used a similar paradigm that involves perception of a target talker in two competing talkers. By manipulating the binaural and spatial properties of the stimuli they tried to tease apart the involvement of different masking processing (EM, MM & IM). They also investigated the effect of non-speech distractors by replacing one of the competing talkers with unintelligible "garbled" speech with speech-like amplitude modulations or a SSN distractor. Carlile

and Crokhill's results revealed that both the competing speech and the garbled speech produced a large amount of non-energetic masking, while the portion of such masking effect for SSN was negligible. Their findings further support the peripheral MM processing theory proposed by Stone and colleagues (Stone et al., 2012; Stone & Moore, 2014), suggesting that the distractor amplitude modulations as in the garbled speech, interfered with the detection of information-carrying amplitude modulations in the target signal. A comparison of the magnitude of this effect for the garbled speech and the original speech distractor revealed that a substantial amount of the non-energetic masking in the speech distractor (5.4 dB) is produced by peripheral MM rather than central attention or semantic processing.

ii. Interrupted speech

In many ways, perception of interrupted speech is very similar to the perception of speech in fluctuating noise and performance in these two listening conditions was shown to correlate (Buss et al., 2009; Grose et al., 2016). Likewise, glimpsing-based speech recognition models adequately predict speech recognition in both stationary and fluctuating noise (Cooke, 2006; Rhebergen et al., 2006). In view of the glimpsing model, the perception of interrupted speech involves the integration of temporally distributed segments of acoustic information of the original speech and the need of perceptual integration of these fragmented segments into existing auditory representations. Similarly to modulated noise, several studies also support the involvement of both higher-level cognitive factors (e.g., working memory and attention), and linguistic factors (e.g., semantic and context) as well as lower-level auditory factors in perception of interrupted speech (Başkent et al., 2016; Kidd & Humes, 2012; Miller & Licklider, 1950). In their pioneering study, Miller and Licklider (1950) showed that listeners were able to retain high intelligibility when segments of speech were periodically removed and replaced with silent intervals, even when only 25% to 50% of the original speech was available, as long as interruption rate was fast enough ($\sim \geq$ 10 Hz). The intelligibility of interrupted speech is

typically manipulated using two basic variables: (1) the number of interruptions per second, ips, or the frequency of interruption (typically referred as *gating*, or *interruption rate*, in Hz); (2) the relative duration of the signal 'on' and 'off' times within each interruption cycle, referred to as *duty cycle* (DC).

Miller and Licklider (1950) investigated the effect of interruption rate and the amount of the available target information (i.e., DC) on speech intelligibility in silence or in added noise. They found that performance for monosyllabic words (when DC is held fixed at 50%) is generally poor at low rates (< 10 Hz) with poorest performance at 1 Hz, and broadly high between 10 to 100 Hz. It is worth noting that susceptibility to the interruption rate may differ, depending on the temporal characteristics of the speech material at hand. For instance, the monosyllabic words Miller and Licklider used were on average 600 ms long. Hence, a 1 Hz interruption rate with a 50% DC resulted in an interruption cycle 500 ms long. Such duration is almost as long as an entire word and can potentially obliterate the word if the interruption cycle is in phase with the onset of the word. In the same study, the authors also explored the performance for noise by replacing the silent gaps with noise in varying SNR levels. Miller and Licklider found that the added noise made the interrupted speech sound continuous, in what they referred to as the 'picket fence' effect. This was an analogy to seeing a landscape through a picket fence, where the pickets hide the view at regular intervals, but the landscape is perceived as continuing behind the pickets. Interchangeably, this effect is also frequently called the phonemic restoration effect, coined by Warren (1970). Interestingly, performance was nearly the same for interrupted speech with or without noise (for rates up to ~ 10 Hz) irrespective of the SNR level, while the decline in performance for higher rates was dependent on the SNR levels. In other words, although by filling the silent gaps with noise the speech was perceived as more continuous and natural, no actual improvement in intelligibility was found. Nonetheless, later studies suggested that the benefit of phonemic restoration is more prominent when the target speech contains sufficient contextual information, e.g., for speech material

consisting of sentences as opposed to single words (Bashford et al., 1992) and is believed to aid in top-down grouping processing (Saija et al., 2014).

A number of studies have found age-related decline in perception for interrupted speech (e.g., Bergman et al., 1976; Saija et al., 2014). Saija et al. (2014)} for instance, found that the performance of older normal hearing adults was significantly poorer than their younger adult counterparts at interruption rates 2.5 and 5 Hz, with a DC of 50%. Similarly, the older listeners showed poorer performance for interrupted speech in noise, but the difference in performance was not significant. The authors also investigated the listeners' ability to make use of phonemic restoration, by filling the silent gaps with noise. Interestingly, the older listeners benefited more from phonemic restoration than the younger listeners. The latter findings suggest that older listeners may benefit from training of specific listening strategies to improve speech perception in difficult listening situations. Some of the findings suggest that the age related decline in performance is in part related to the interruption rates, and seems to be most disruptive for older listeners at rates between 2.5 to 5 Hz (Shafiro et al., 2015). Nonetheless, Bergman (1980) showed that older listeners, aged 55 years and above, performed substantially poorer than younger adults also at a higher interruption rate (8 Hz) at various DCs, ranging from 30 to 70%. Kidd and Humes (2012) investigated the effect of age, hearing loss and sentence context on perception of interrupted words, presented either separately or inserted at the end of sentences with low or high semantic context. They found that younger normal hearing listeners performed better than older (normal hearing and hearing impaired) listeners. Nonetheless, the ability to make use of additional top-down contextual information was similar across the listeners, irrespective of age or hearing loss. Conversely, Kidd and Humes (2012) postulated that the most dominant factor that affects interrupted speech performance is the proportion of an utterance that is available to the listener, while changes in DC and interruption rate have comparatively little effect on speech perception performance.

Perception of interrupted speech may be useful in disentangling the reasons why different groups of listeners experience difficulties in noisy situations. Nonetheless, measuring speech perception in a non-adaptive way is not always clinically viable due to time constrains, and have several other drawbacks such as a possible floor/ceiling effects if different components of the test haven't been appropriately selected, or audibility limitations at low SNRs which may reduce the expected effect on performance. Mair (2013) has suggested a new method to estimate perception of interrupted speech using an adaptive method, similar to measurements of SRT, whereby the varying variable is the amount of DC that yields 50% of key words correct in sentences (SRdT). A fixed 4 Hz interruption rate was applied to the target sentences (equivalent to 250 ms long cycles of the speech signal per second), presented dichotically in silent gaps or with a SSN replacing the silent gaps. Mair found no significant difference in performance with silent gaps or with noise across neurotypical normal hearing listeners, with SRdTs of circa 0.45 (DC) on average. Overall Mair's test method produced comparable results with data from the literature (cf. Fig. 6 in Nelson & Jin, 2004) and psychometric functions fitted for the data were reported to show no evidence of a non-monotonicity. This is of particular interest for the current paper, since the test paradigm that will be used is based on Mair's adaptive procedure to estimate the listeners SRdTs.

iii. Alternated speech

More ecological listening situations often involve the need to switch our attention between competing sound sources and/or locations (Bronkhorst, 2015). One way to introduce such target uncertainty is by applying interaural alternations, where the stimulus is periodically switched from one ear to the other, whilst fully preserving the stimulus information when combining the alternating segments coming from each ear. In their seminal work, Cherry and Taylor (1954) were interested in the effect of periodically alternated speech on speech perception using an electronic switch to quickly alternate the signal between the ears via headphones. Speech intelligibility

was measured for varying alternation rates, determined by the number of switching cycles per second (cps). In theory, it seems sensible to assume that performance for alternated speech shouldn't be impaired, since the stimulus information is fully preserved. Cherry and Taylor showed that this is indeed the case at both low and high alternation rates (0.1 cps and $>$ 6 cps, respectively). Interestingly however, performance was noticeably reduced for alternation rates between 3 to 5 cps (corresponding to about 167 to 100 ms long speech segments per ear in a cycle, respectively), resulting in a V-shaped intelligibility function. Furthermore, at higher alternation rates ($>$ 6 cps), localisation of the incoming sound source direction was disturbed, resulting in a rather diffused sound image, where the sounds are perceived to be located more centrally in the listener's head (Hoffman & Levitt, 1978).

The cause of poorer intelligibility at low alternation (2 - 3 cps) has been a source of debate amongst researchers throughout the years. Cherry and Taylor (1954) attributed the loss in intelligibility to the existence of a lag in reaction time of the auditory system to switch attention from one ear to the other in what they called 'mental switching'. They postulated that at a critical rate the switched signal and the mental switching are out-of-phase, thus, making perception impossible. Another explanation to this phenomenon was suggested by Huggins (1964). Huggins demonstrated that the critical rate of alternation could be shifted when speech rate was increased, arguing that this suggests that poor performance is attributed to the duration of the syllables in the speech signal, rather than to a delay in reaction time.

Perception of alternated speech may arise from the listeners' ability to switch their attention between the ears and to attend to a particular sound source. Stemming from the glimpsing model theory (Cooke, 2006), Brungart and Iyer (2012) posited that perception of speech in challenging conditions is based on the ability to make use of better-ear glimpses. Schubert and Parker (1955) compared the effect of alternated speech passages with gaps of silence or with a white noise in the contralateral ear. They found that replacing the silent gaps with noise resulted in an improved speech

intelligibility at the critical alternation rates. Their findings speak in favour of what they described as "contralaterally-inhibitive off-effect" when a speech segment is switched abruptly to silence, rather than to a lag in reaction time of the auditory system to the switched segments as Cherry and Taylor (1954) postulated.

Hoffman and Levitt (1978) have proposed to use alternated speech in noise as a way to tease apart central (IM) and peripheral (EM) interference, using *simultaneous* and *interleaved* masking conditions. In simultaneous masking, the alternating cycles of both signals are in phase, i.e., they are presented at the same ear at the same time. This type of masking is thought to take place at both the peripheral level, in the cochlea, and at the central level, following binaural integration. In interleaved masking on the other hand, cycles of the target and the noise are alternated synchronously to the opposite ear, i.e., only one stream (target/noise) is presented in each ear at any given time. It therefore enables us to isolate central masking (IM) by eliminating peripheral masking introduced by interaction of the noise and the target energy. Hoffman and Levitt (1978) were particularly interested in perception at higher alternation rates ($>$ 6 cps), where lateralisation cues are hindered, resulting in ambiguous spatial perception of the competing streams. Their results revealed a benefit in MR for interleaved noise of circa 20 dB as opposed to simultaneous noise. This MR was reported by the authors to be much higher than binaural MR of similar speech material of 3 to 6 dB, which suggests that IM results in a greater MR when EM was controlled for.

Akinseye (unpublished BSc thesis, 2015) used a novel speech-in-noise task (referred to as the 'switching task') which involved perception of interrupted speech in noise, presented dichotically either without switching (i.e., a target in one ear and a distractor in the other ear), or switched between the left and the right ear several times throughout a sentence, i.e., interleaved noise as in Hoffman and Levitt (1978). In the task, the speech signal was interrupted at a fixed rate (5 Hz) while adaptively varying the speech DC to track the listener's SRdT as in Mair (2013),

with the signals presented at a fixed 0 dB SNR. The segments of the interrupted speech were then presented alternately to the two ears, yet only in one ear at a time. The task's key advantage is drawn from the use of an interleaved distractor, which eliminates peripheral masking (EM), while obtaining high IM, which is enabled by the relatively fast switching rate which reduces lateralisation causing a more diffused spatial percept of the competing streams. Moreover, using derived measures, by comparing for example performance with and without switching, enables the determination of the relative change in performance while controlling for variability in the cognitive skills involved (e.g., verbal working memory, attention, linguistic knowledge, and/or auditory closure skills that aid in filling in the missing pieces of degraded information). Akinseye compared performance in the switching task across younger (mean age: 24, range: 20-33 years old) and older (mean age: 63, range: 50-72 years old) adults with audiometrically normal hearing up to 4 kHz. Normal cognitive skills were controlled for the older participants using a standard screening test. Performance was compared with SRTs measured using a standard speech-in-noise test with two distractor types: SSN, and a harmonic complex, dynamically changing F0, with F0 contours extracted from speech recordings of an adult male voice reading connected speech. F0 contours were interpolated through periods of silence and voicelessness (for more details about the distractor see the Methods section or Green & Rosen, 2013). Both distractors had the same long-term average spectrum as the target speech (LTASS). The target speech was the same in both tasks and comprised of everyday sentences (ASL; MacLeod & Summerfield, 1990), spoken by a male talker, whereby the distractor used in the switching task was connected speech spoken by a single female talker. Interestingly, while no significant difference in SRTs was found between groups for the speech-in-noise test, there was a highly significant difference in performance between the groups for the switching task. In the latter task, older listeners performed considerably poorer only when the stimuli switched between the ears. Akinseye's data suggests that the switching condition demands some higher-order cognitive

aspects of listening that is not probed by more simple speech-in-noise listening tasks.

The aim of the present paper was to unravel the contribution of IM on perception of speech with a contralateral distractor, presented dichotically with streams of the two signals switching rapidly between the two ears. In the first experiment, we evaluated the amount of IM induced by different types of speech and non-speech distractors, with or without talker-sex agreement between the target and the distractor. The speech distractor comprised of unrelated connected speech, spoken by a talker from the same/opposite sex to the target talker. The non-speech distractors were derived from specific speech features that were extracted from the original speech distractors. They were selected to have different amount of speech-like characteristics, and thus were expected to differ in the magnitude of IM they produce. A speech-spectrum-shaped-noise modulated with the speech distractors envelope (AMSSN), preserving the slowly varying wide-band amplitude envelope of the speech distractor, representing a more rudimentary distractor and was expected to reflect a small IM effect. The second non-speech distractor was single-band vocoded speech with a natural mix of periodicity and aperiodicity (FxNx), preserving the original speech temporal fine structure (TFS) associated with periodicity and aperiodicity and was expected to produce a larger IM. We hypothesised that introduction of a distractor will result in a decrement in performance, and that the magnitude of the decrement will be moderated by the distractor type, with speech distractors eliciting the largest IM. We expected to get, little to no IM for AMSSN, while maintaining the natural speech periodicity and aperiodicity in the FxNx distractor was expected to produce a larger IM. Finally, as seen in other studies (e.g., Brungart et al., 2001; Festen & Plomp, 1990), we expected that an increase in similarity between the target and the distractor, as in the presentation of a same-sex distractor talker, will elicit further decrement in performance (i.e., increased IM) for FxNx and speech distractors.

Findings in the first experiment demonstrated that performance in the task was uniquely affected when speech distractors were presented, whereas none of the nonspeech distractors exerted any IM. To extend these findings, in the second experiment we investigated specific aspects of the speech distractor that may contribute to the IM effect in the task. We examined the contribution of familiarity with the spoken language, and similarity-related features such as pitch, by comparing performance for speech distractors spoken in a familiar (English) or in an unfamiliar language (Mandarin), spoken by talkers either from the same- or the opposite-sex to the target talker. To expand the generalisation of our findings, instead of using the same single speech passage spoken by a single talker in every trial as in the first experiment; the speech distractors in the second experiment comprised of forty different passages spoken by forty different talkers (twenty for each language, with an even number of male and female talkers). Finally, we evaluated some aspects concerning the applicability of the task for future clinical use. We examined the test-retest reliability within a single session and the reproducibility of the task's measure by comparing between performance measured in the first and second experiment.

1.1.2 Experiment I: speech vs. non-speech distractors

Methods

Participants

Sixteen young adults who were native British English speakers participated in the first experiment (mean age 25.5 ± 5.3 years, ranging from 18 to 34 years, 8 females). All the participants were tested to have normal hearing acuity, defined by air conduction pure tone audiometric thresholds ≤ 25 dB HL for frequencies ranging from 0.25 to 8 kHz. On one occasion, a threshold of 30 dB HL at 2 kHz was accepted. Nonetheless, all the participants had a PTA_4 below 25 dB HL, averaged across the frequencies 0.25, 1, 2, and 4 kHz (World Health Organisation, 1998), in the left (3.6

± 3.6 dB HL) and the right ear (4.5 ± 5.8 dB HL). The listeners' thresholds for the left and the right ear are plotted in Fig. 1.1. The shaded grey area represents the range of audiometric thresholds at each frequency, while the white line represents the mean of the participants at each frequency. The dashed line represents the threshold criteria. None of the participants reported a history of ear or hearing problems or language or other cognitive impairment. The Study was approved by the UCL Research Ethics Committee (Project ID Number 0544/006) and testing commenced once an informed consent was given. Participants were recruited from the UCL psychology subject pool and were paid for their participation.

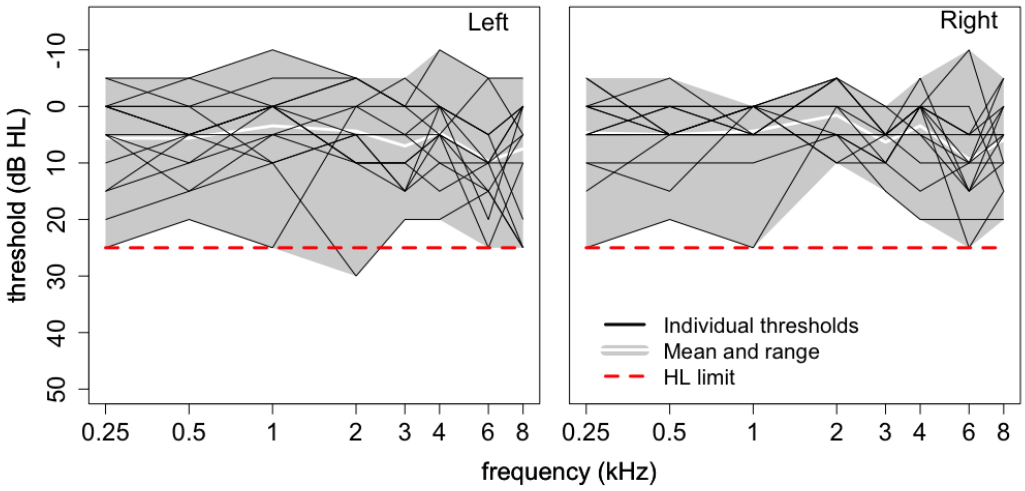

Figure 1.1: Individual pure-tone-audiogram thresholds plotted separately for the right and the left ear (in black). The shaded grey area represents the range of the audiometric thresholds and the white line represents the mean at each frequency across the listeners. The red dashed line represents the threshold criteria of hearing level $\leq$ 25 dB HL.

Stimuli

The target stimuli were taken from the Adaptive Sentence List corpus (ASL; MacLeod & Summerfield, 1990), comprising 270 sentences spoken by an adult male talker with a standard southern British English accent (sampled at 22.05 kHz with 16 bits per sample, low-pass filtered at 10 kHz). The speech material is based closely on the BKB sentences (Bench et al., 1979), comprising simple "everyday" sentences of five words on average (range: 4-6 words) with three keywords each. The sentences

are suitable for testing listeners with a wide range of speech perception abilities from children to adults. A loose keyword scoring method was used, whereby errors of case or declension were considered as correct responses. For example, as in a repetition of the keywords '<clown**s**> <funny> <face**s**>' to the stimulus 'The <clown> had a <funny> <face>'. Six different distractors were used in the first experiment and can be grouped into two types: speech- and non-speech distractors, with different degrees of acoustic similarity to speech. The speech distractors consisted of two short unrelated conversational passages (each 5-6 sentences long) with durations roughly ranging between 15 to 30 s. They were taken from a large corpus of passages spoken by native speakers of Southern standard British English (EUROM corpus; Chan et al., 1995). Out of the two selected passages, one was spoken by a male talker, i.e., a talker of the same sex as the target talker ($\mathrm{ENG}_{same-sex}$), while the second passage was spoken by a female talker ($\mathrm{ENG}_{opposite-sex}$). The male talker used for the same-sex distractor was different from the one used for the target sentences. However they had similar speech rate and fundamental frequency.

The non-speech distractors were derived from the original speech distractors, separately for same- and opposite-sex talker, and varied in their amount of "speech-like" characteristics from high to low, respectively. The first one is thought to preserve the original speech temporal fine structure (TFS) associated with the speech periodicity and aperiodicity (but not that associated with overall spectral shape), and comprised of single-band vocoded speech with natural mix of periodicity and aperiodicity (FxNx; also described in Steinmetzger & Rosen, 2015). The second non-speech distractor was an amplitude modulated speech-shaped-noise, with the same long-term spectrum, and modulation envelope as the speech distractors (AMSSN), preserving the original speech slowly varying wide-band amplitude envelope. Exemplary waveforms and spectrograms of the different distractor types are shown in Fig. 1.2. The distractors were generated in MATLAB (Version R2017b, Mathworks, Natick, Massachusetts) using a channel vocoder (described in Green & Rosen, 2013; Steinmetzger & Rosen, 2015). First, the speech distractors were

bandpass filtered into a single band using zero-phase-shift 6th-order Butterworth filter (frequency range: 70 Hz - 10 kHz). The amplitude envelope was then extracted by applying full-wave rectification of the filter output and a low-pass filtering at 30 Hz (zero-phase shift, 8th-order Butterworth filter) to remove any modulations arising from voice fundamental frequency. For the generation of the AMSSN, the envelope of the single channel was multiplied with a wide-band noise carrier and the resulting waveform was low-pass filtered at 10 kHz using 6th-order elliptic filter. Next, the output signal was scaled to the RMS level of the original speech signal. FxNx was generated by multiplication of the single-band envelope with either a white noise carrier for unvoiced speech segments in the original speech, or with the fundamental frequency contour of the original signal when speech was voiced. F0 contours were extracted in PRAAT (Version 6.0.19; Boersma, 2001) using ProsodyPro (Version 5.7.2; Xu, 2013), and subsequently manually corrected. Next, F0 contours were sampled at 1 kHz and interpolated through periods of voiceless and silent segments using piecewise cubic Hermite interpolation in logarithmic frequency. The start and end of each pitch contour were anchored to the signal's median frequency, resulting in a carrier with the same length as the original signal. Finally, filtering was applied to the vocoded AMSSN and FxNx signals to have the same LTASS as the original speech signals.

The switching task

The listening task was developed locally in MATLAB, and involves perception of target speech which is interrupted and alternated between the ears out-of-phase with an interrupted distractor, resulting in alternated segments of both signals between the two ears, with only one stimulus present in each ear at any given time. Interruption is applied by gating the signal at a fixed modulation rate of 5 Hz, i.e., a period of 200 ms (with 5 ms rise/fall times), and varying the duty-cycle (DC), which is the proportion of time the signal is present in each modulation period. As

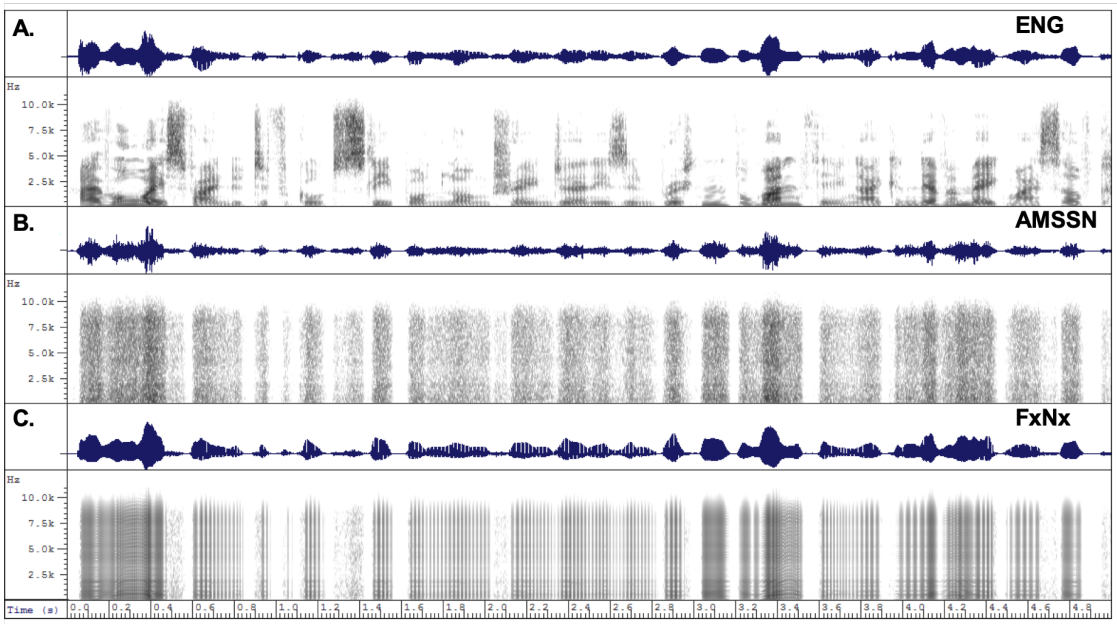

Figure 1.2: Waveforms and broadband spectrograms of a short segment of the speech distractor spoken by a female talker, $\text{ENG}_{opposite-sex}$ (A.), and the two non-speech distractors, generated from features extracted from the original speech distractors: amplitude modulated speech spectrum noise, AMSSN (B.), and single-band vocoded speech with natural mix of periodicity and aperiodicity, FxNx (C.).

illustrated in Fig. **??**, DC ranged between 0.1, where signal is nearly completely 'off' (left figures), to 0.9, where the signal is almost entirely 'on' (right figures).

Performance was estimated using a 1-up/1-down adaptive staircase procedure (e.g., Levitt, 1971), whereby the speech level or signal-to-noise-ratio (SNR) is fixed, while DC varies depending on the listener's response on a trial by trial basis. The Speech Reception duty-cycle Threshold (SRdT) was estimated, which is the DC ratio at which 50% of the keywords were repeated correctly. A correct repetition of 50% or more of the keywords (i.e., two keywords or more), meant that the DC ratio of the next trial decreased (i.e., got more difficult), whereas a correct repetition of less than 50% of the keywords (i.e., up to one keyword), meant that the DC ratio of the next trial increased (i.e., got easier). The points at which the specified DC changes direction are called transition reversals. The outcome measure, SRdT, is then determined by averaging the test reversals that followed three practice reversals. In case of an odd number of test reversals, the first test reversal was ignored.

Next, the switching of the interrupted stimuli was applied. As illustrated in Fig. 1.3, the interrupted target signal was multiplied with a modulation carrier (grey carrier), separately for the left (blue) and the right ear (red). The modulation carrier in one of the ears was time-shifted, resulting in alternated segments of the signal between the two ears, but only in one ear at each given time (middle figures). The same step was also applied to the distractor, by inverting the modulation carriers used for the target signal. For presentation of the target speech in quiet, the distractor's segments were replaced with silence. The carrier had a fixed modulation rate of 5 Hz, which was found in several studies to significantly impair speech perception in adults and was shown to be slow enough to be able to perceive the switched speech segments between the two ears (Cherry & Taylor, 1954).

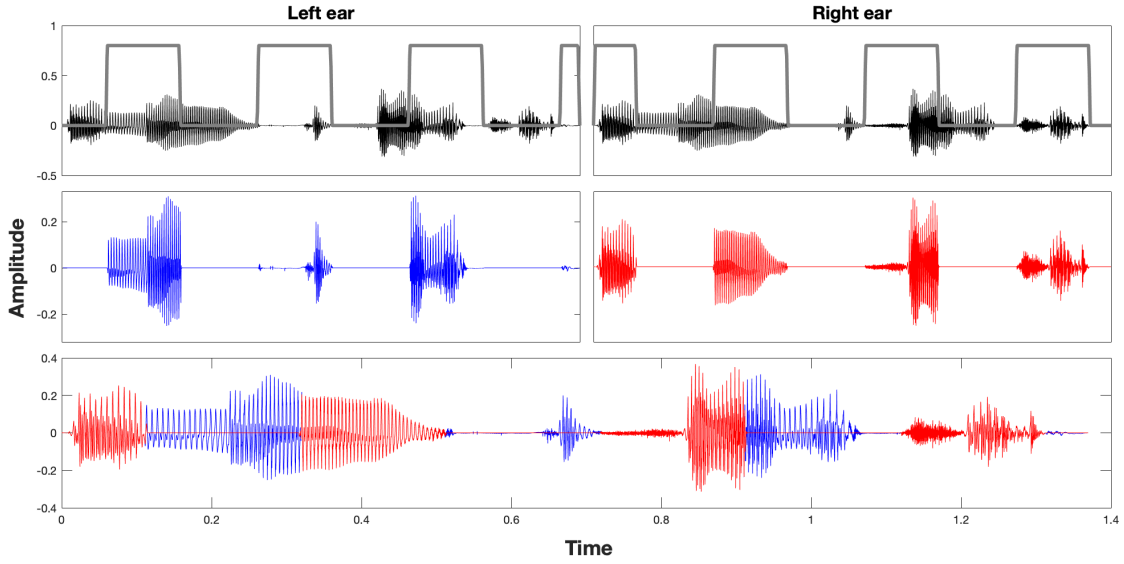

Figure 1.3: Illustration of an alternated speech signal with a duty-cycle (DC) of 0.5 and a modulation rate of 5 Hz (i.e., 200 ms periods). Upper and middle figures shows multiplication of a modulation carrier (grey) for the left (blue) and the right (red) ear. Note that the phase of the modulation envelope is selected by random in each trial. The lower figure illustrates the alternated speech signal, achieved by adding together the left and the right channels.

Listeners were presented with two listening conditions, with or without a distractor (see Fig. 1.4). A listening condition with a distractor is depicted in the right side of the figure, where segments of interrupted target signal (black

bars) and segments of the distractor signal (grey bars) are alternated out-of-phase between the left and the right ear. Similarly, a reference condition where the target signal is presented without a distractor is shown in the left half of the figure, by replacing the distractor segments with silence.

Figure 1.4: Schematic of the switching task listening conditions. The target speech and the distractor are represented by the black and grey bars, respectively. The stimuli presented in the left ear are depicted in the upper part of the figure as a function of time, whereas the stimuli presented in the right ear are depicted in the lower part.

Procedure

A single experimental session with a maximal duration of 2 hours (including breaks) took place in a sound attenuated chamber. Stimulus presentation and scoring were carried out using a locally developed MATLAB script via a MacBook Pro 13 laptop (macOS High Sierra 10.13.4) connected via USB to an RME Babyface soundcard (Audio AG, Haimhausen, Germany). The test signals were presented through Sennheiser HD-25 headphones (Wedemark, Germany) at a fixed output level of circa 70 dB SPL, measured using an artificial ear (Brüel & Kjær 4153, Sound and Vibration Measurements A/S, Nærum, Denmark) over a frequency range of 100 Hz to 10 kHz. A 30 ms long cosine onset ramp was applied to the segmented target signal to avoid the stimulus from sounding abrupt. For conditions with a distractor, the target onset was 1 s after the distractor to avoid uncertainty to which signal the listener should attend to. In each trial, a distractor segment was randomly selected from the long signal to match the length of the target sentence (plus 1 s onset time). The starting DC was 0.97 (i.e., signal is almost entirely present). Subsequently, the DC varied depending on the listeners response, with an initial step-size of 0.12 which decreased gradually over the first three (practice) reversals to 0.05. Nonetheless, examination of pilot data suggested that the psychometric

functions of speech distractors are shallower, thus it was decided to set the minimum DC step-size speech distractors to 0.1. The starting ear of the switched segments was randomised in each trial.

In experiment I, a self-scoring method was used via a graphical user interface (GUI), whereby the listeners were instructed to transcribe the sentence using a keyboard and press the 'OK' button once completed using a computer mouse. The response was thereafter recorded and could not be altered any more. Next, the listeners were asked to select the correctly recalled keywords from the options shown on the screen, based on their displayed transcription. Pressing again the 'OK' button prompted the presentation of the next trial. Feedback was given following each trial only for the practice phase where both the non-degraded target sentence and the test stimuli were presented. Prior to the beginning of the data collection, listeners were familiarised with the task by responding to a set of five practice runs in the following fixed order: Quiet, AMSSN, FxNx with 5 trials each, and ENG (same- and opposite-sex) with 15 trials each. The presentation order was set to reflect the expected decline in listeners' score caused by increased masking interference. Due to the limited number of ASL sentences, the target sentences in the training phase were taken from the BKB corpus (Bench et al., 1979) which are very similar in structure to the ASL sentences. In addition, a short practice run was given during the testing phase at the beginning of each run, whereas no feedback was given in order to reduce testing time.

In total, seven test conditions were recorded in the testing phase, originating from the following factorial design: 3 distractor types (ENG, FxNx, AMSSN) x 2 distractor talker-sex (same-/opposite-sex), and a reference condition, where the interrupted target signal was presented without a distractor (Quiet). Listeners were presented only once with each test condition. Each condition consisted of 19 ASL target sentences. The order of the test conditions and target sentence lists was

quasi-randomised to account for order or fatigue effects.

Statistical methods

The listeners SRdTs was assessed using a model comparison approach in *R* environment (RStudio Team, 2019). Linear mixed-effects regression models (LMEMs) were fitted by maximum likelihood (ML) using the *lmer()* function (*lme4* package in Bates et al., 2014). The first model examined the overall effect of distractor type using 1x7 LMEM with the seven test conditions as fixed factors (3 distractor types x 2 distractor talker-sex configuration and Quiet condition), with the Quiet condition set as a reference level, and subjects included as by-subject random intercept. The second model assessed differences in performance between speech and nonspeech distractors and the effect of talker-sex using 3x2 LMEM with distractor type (ENG, FxNx, & AMSSN) and distractor talker-sex (same/opposite) as fixed factors and again random intercepts for subjects (reference levels: distractor type = AMSSN; distractor talker-sex = opposite). Note that observations for the Quiet condition were excluded from the second model. LMEM assumptions of homogeneity and normal distribution were fulfilled, tested with Levene's test (Fox & Weisberg, 2011) and Shapiro-Wilk test (R Core Team, 2018). The initial saturated model included by-subject random intercepts and slopes. However, because the model did not converge, it was simplified to a model that would converge by including only random intercepts. We used backward model selection (cf. Barr et al., 2013), by removing fixed terms that did not significantly degrade the model's fit (significance level $\alpha = 0.05$) using likelihood ratio test (χ^2). Independent post-hoc t-test comparison was performed on the fitted model and included adjusted least-squared-mean for the random intercepts (subjects) using *lsmeans()* (lsmeans package; Lenth, 2016). The p-values were Bonferroni-adjusted.

Results

Descriptive statistics of the listeners performance (in SRdTs) for the different test conditions is given in Tab. 1.1. In total, seven SRdTs were recorded for each participant across four background conditions: Quiet, and the distractors AMSSN, FxNx, and ENG, whereby distractors originated from either opposite- or same-sex talker. Boxplots of the SRdTs are shown in Fig. 1.5. The results reveal that the non-speech distractors elicited little to no interference with the target speech, with similar SRdTs as for the reference Quiet condition, while the speech distractors showed a large interference effect, resulting in increased SRdTs (i.e., poorer performance) for opposite-sex and same-sex talkers.

To put these results in what might be a more understandable context, the SRdT reflects the amount of speech information (glimpses) required by the listeners to understand 50% of the sentence correctly. An SRdT of roughly 0.34 obtained for the non-speech distractors and the reference condition Quiet (at a 5 Hz modulation rate) is equivalent to five 68 ms audible glimpses of the target sentence per second, each preceded and followed by 132 ms of silence. For the speech distractors on the other hand, in order to understand 50% of the sentence correctly, the listeners needed more than double the duration of audible target glimpses per period (164 ms) for the same-sex distractor and about 56% longer (106 ms) for the opposite-sex distractor.

The effect of distractor type in general on the listeners' performance, was tested by a comparison of the SRdTs with the reference (Quiet) condition included using 1x7 LMEM (see Tab. 1.2 for the model coefficients and p-values). Model comparison showed a highly significant main effect of background [$\chi^2(6)$=178.76, p <0.001]. The results revealed that speech distractors significantly impaired the listeners performance, for both opposite- and same-sex talker [b=0.19, t(96)=7.08, p <0.001 and b=0.48, t(96)=17.66, p <0.001, respectively]. On the other hand, no difference

in performance between the non-speech distractors (AMSSN & FxNx) and the reference condition was found (all $p's$ <0.05).

Table 1.1: Descriptive statistics for the SRdTs obtained in experiment I across the different test conditions.

Background type	Grand mean	Distractor talker-sex	
		Opposite	Same
	M (SD)	M (SD)	M (SD)
Quiet	0.34 (0.07)	-	-
AMSSN	0.35 (0.09)	0.34 (0.08)	0.35 (0.09)
FxNx	0.37 (0.09)	0.37 (0.08)	0.36 (0.10)
ENG	0.67 (0.18)	0.53 (0.13)	0.82 (0.09)

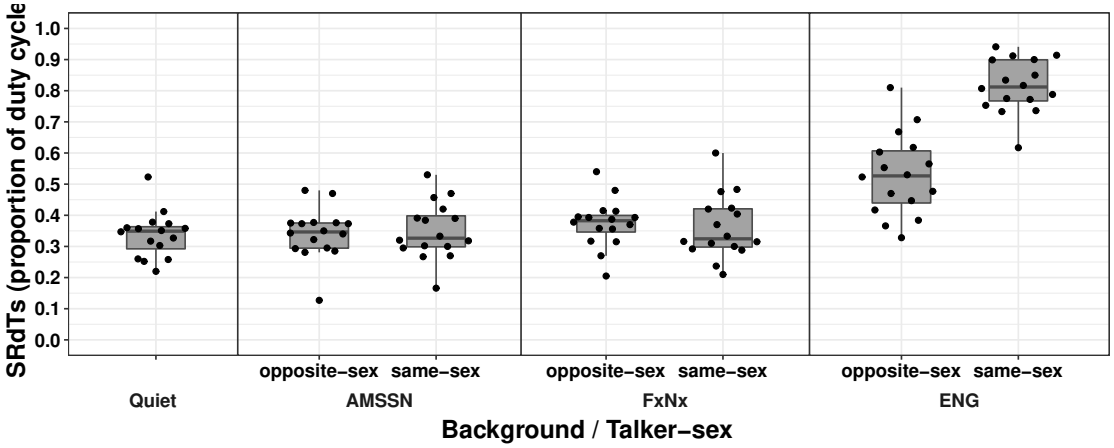

Figure 1.5: Boxplots of the SRTds measured in experiment 1 for the baseline condition Quiet and the distractor conditions AMSSN, FxNx and ENG with the same- and opposit-sex talker. Individual scores are represented by the black circles.

A separate model without observations measured with the Quiet condition examined whether there was a difference in performance between the speech and nonspeech distractors, **as well as the effect of distractor talker-sex** using 3x2 LMEM (see Tab. 1.3). Model comparison showed a significant main effect

Table 1.2: 1x7 mixed-effects model for SRdTs measured in experiment I across all subjects (N observations = 112; N Subjects = 16). Reference level = Quiet condition. Significant p-values are marked as bold.

SRdT ~ BackgroundType + (1 | Subjects)			
Main effects	Df	χ^2	p
BackgroundType	6	178.76	<**0.001**
Fixed effects	Estimated mean difference	SE	95 % CI
intercept	0.34	0.02	0.29 – 0.38
$\text{AMSSN}_{opposite-sex}$	0.00	0.03	-0.05 – 0.06
$\text{AMSSN}_{same-sex}$	0.01	0.03	-0.04 – 0.07
$\text{FxNx}_{opposite-sex}$	0.04	0.03	-0.02 – 0.09
$\text{FxNx}_{same-sex}$	0.02	0.03	-0.03 – 0.08
$\text{ENG}_{opposite-sex}$	0.19	0.03	0.14 – 0.24
$\text{ENG}_{same-sex}$	0.48	0.03	0.43 – 0.53

for distractor type [$\chi^2(5)$=159.45, p <0.001], distractor talker-sex [$\chi^2(3)$=72.08, p <0.001] and their interaction [$\chi^2(2)$=54.96, p <0.001]. Next a post-hoc t-test comparison showed no significant difference in SRdTs between the two non-speech distractors FxNx and AMSSN [t(85.33)=1.11, $p = 0.808$], and a highly significant difference between the non-speech and speech distractors [AMSSN vs. ENG: t(85.3)=-16.85; FxNx vs. ENG: t(85.3)=-15.73, $p < 0.001$] . Moreover, differences in performance due to distractor talker-sex (two-way interaction) was significant (and highly so) only for the speech distractors [t(85.3)=-10.46, $p < 0.0001$].

Table 1.3: 3x2 mixed-effects model for SRdTs measured in experiment I across all subjects (N observations = 96; N Subjects = 16. Reference levels: distractor type = AMSSN; distractor talker-sex = opposite. Significant p-values are marked as bold.

SRdT ~ DistrType + DistrTlkrSex + DistrType * DistrTlkrSex + (1 | Subjects)			
Main effects	Df	χ^2	p
DistrType	5	159.45	<**0.001**
DistrTlkrSex	3	72.08	<**0.001**
DistrType x DistrTlkrSex	2	54.96	<**0.001**
Fixed effects	Estimated mean difference	SE	95 % CI
intercept	0.34	0.02	0.30 – 0.39
DistrType (FxNx)	0.03	0.03	-0.02 – 0.08
DistrType (ENG)	0.19	0.03	0.14 – 0.24
DistrTlkrSex (same)	0.01	0.03	-0.04 – 0.06
DistrType (FxNx) x DistrTlkrSex (same)	-0.02	0.04	-0.10 – 0.05
DistrType (ENG) x DistrTlkrSex (same)	0.28	0.04	0.20 – 0.35

Discussion

The objective of the first experiment was to evaluate the amount of IM induced by different types of speech and non-speech distractors with or without talker-sex agreement between the target and the distractor. To tease apart the key factors that contribute to IM, speech intelligibility was measured for three types of distractors. In addition, the listeners' baseline performance was measured for the switched target

with silent intervals replacing the distractor (Quiet condition).

The SRdTs measured in the reference Quiet condition (0.34 ± 0.07) is in line with Akinseye (unpublished BSc thesis, 2015) preliminary study, and is in accordance with the literature for interrupted speech (e.g., Kidd & Humes, 2012; Miller & Licklider, 1950) and alternated speech (e.g., Stuart, 2008). Different distractor types affect performance differently. We hypothesised that performance will get poorer (i.e., higher DC) by introducing a distractor and that the decline in speech perception (or the increase in IM) will be moderated by the type of the distractor, with speech distractors potentially producing the largest IM. Moreover, we hypothesised that introducing more speech-like features into the non-speech distractors would result in increased similarity and uncertainty between the target and the the distractor, which consequently will result in a larger interference effect for FxNx as opposed to AMSSN. We therefore expected FxNx to introduce similar IM as the speech distractor. The outcomes of the study showed that speech distractor (ENG) resulted in the largest IM. In fact, only the speech distractor showed a significant difference in performance, while performance for the non-speech distractors was the same as for the target sentences in quiet.

Informational masking can be attributed to both bottom-up processes, as in signal characteristics that support streaming of a sound source (i.e., object formation) and top-down attention-related processes that support attending to the target signal (i.e., object selection; Shinn-Cunningham, 2008). Increased target-distractor similarity and uncertainty increases IM. The present study revealed that only the speech distractor produced IM. Due to the complex nature of speech signals, trying to disentangle the different contributing factors that produced this exclusive IM effect for speech distractors is not straight forward. Although some properties of the stimuli (i.e., speech distractors and their derived nonspeech distractors) we used were to some extent controlled for, due to the variable nature of speech, some differences between the stimuli are still possible (e.g., sentence structure, semantic content,

vocabulary, speech rate, vocal-tract length, F0, or generally different speaking style), and could have had an effect on the amount of IM that is produced. Nevertheless, one obvious factor that had a large effect on the amount of IM was the distractor talker-sex. Performance for speech distractors spoken by a same-sex talker was significantly poorer (i.e., larger DC) than for a distractor spoken by a talker from the opposite sex. In the present study we chose a same-sex distractor talker with a similar median F0 as the target talker. This may add an element of uncertainty with the target signal, resulting in a combination of bottom-up failure in object formation in addition to the impaired top-down object selection as seen for the opposite-sex distractor talker. Nonetheless, the stimuli used in the present study originated from single talkers and did not change from trial to trial. Thus, one should be cautious when trying to draw more general conclusions about the effect of the talker-sex agreement between the target and the distractor on the performance.

Another possible contributing factor is semantic content. The speech signals in the present study originated from different talkers and differed in their semantic content: ASL sentences (target) vs. unrelated connected speech (distractor). Nonetheless, similarity between the target and speech distractors at the word-level, or more likely at the phoneme-level are short enough to be conveyed within the 200 ms long switching signal segments, and could potentially cause attentional uncertainty, resulting in failure of top-down processing in attending to the target signal. The lack of IM interference for FxNx may suggest that semantic content is weighted as a more reliable cue in the process of auditory stream segregation in adverse listening conditions (such as here), and may have been prioritised over other cues such as F0 and TFS. The unaffected performance for amplitude modulated speech shaped noise was expected and is in line with other studies demonstrating that typically neurotypical normal hearing adults can maintain high intelligibility for speech in amplitude modulated noise when presented dichotically (e.g., Brungart et al., 2013).

Overall, these results suggests the important role of semantic content in IM in the switching task. However, further research should be done to investigate this more closely. One possible way to look into the contribution of meaning of the speech distractors is to include speech distractors spoken in a language that the listeners are not familiar with, thus preserving the natural spectrotemporal characteristics of speech, while eliminating the influence of semantic content.

The present study used an automated self-scoring method to record the listeners performance. All the participants were able to adequately use the scoring method with no particular problems. This was supported by an inspection of the listeners' transcription and selected keywords. Automated scoring methods in speech perception tasks are mostly used for closed set speech material such as the matrix sentences (Kollmeier et al., 2015) or the coordinate response measure (CRM; Bolia et al., 2000). The main advantage of the scoring method used in the current study is that it enables a fully automated testing for open set speech material. Thus, it excludes the need for the examiner to manually select the listener's verbal response and eliminates the need of the examiner to speak the language spoken in the task. Selecting the listener's correct answer based on their verbal response in some cases can introduce bias to the measurement (e.g., when the listener has pronunciation difficulties). Therefore, this method avoids such bias and has the potential to reduce the scoring error rate. Nonetheless, it has two major disadvantages which probably makes this method most likely not suitable for children and elderly listeners, nor for use in the clinic listeners or clinically viable. First, it requires adequate typing and spelling skills and working memory may possibly affect the listeners' performance, especially in adverse listening conditions. Secondly, it substantially increase the testing time, and testing times vary greatly depending on the listeners typing skills.

1.1.3 Experiment II: speech distractors spoken in a familiar vs. unfamiliar language

Findings in the first experiment demonstrated that performance in the task is specifically affected when speech distractors are used, and that this IM effect did not occur for the non-speech distractors. To extend these findings, in the second experiment we examined the contributions to IM of familiarity with the spoken language of the distractor (English vs. Mandarin), and similarity-related features as in voice characteristics of the talkers (same-sex vs. opposite-sex talkers). Furthermore, the applicability of the proposed task for future clinical and research use was examined.

Methods

Participants

The data in the second experiment was taken from a larger study which aimed to compare performance in the task between two groups of young and older adults, native British English speakers with 20 listeners in each group (Huang, 2018). None of the participants were familiar with Mandarin. Here we present only the data collected with the younger group. To enable a better comparison of listeners scores between experiment I and II, the same inclusion criteria were employed. Thus, only listeners with an age $\leq$ 35 years old were included, resulting in a total of 15 listeners. Next, inspecting for outliers (more than 2 s.d.'s from the mean), revealed that one listener was indicated as a possible outlier 9 times out of 10 with an over all poor performance, and was therefore removed. The remaining 14 listeners mean age was 25.1 $\pm$ 4.2 (range: 19-35 years, 11 females) and were tested to have normal hearing acuity based on the same criteria as in the previous experiment (right ear $PTA_4 = 3.6 \pm 2.6$ dB HL, left ear $PTA_4 = 4.1 \pm 3.2$ dB HL; see Fig. 1.6). Participants were recruited from the UCL psychology subject pool and from the Speech and Language Therapy MSc programme at City, University of London and were paid for their participation. The Study was approved by the UCL research

Ethics Committee (Project ID Number 0544/006).

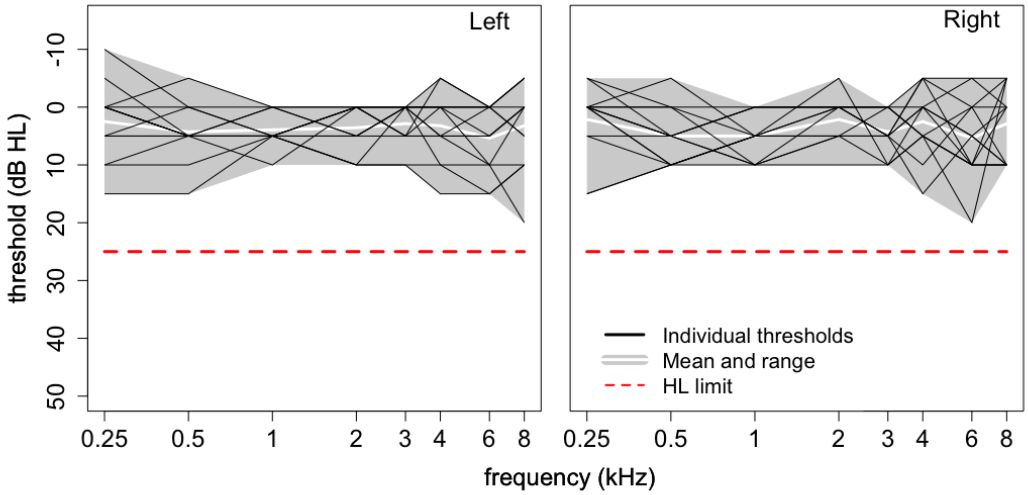

Figure 1.6: Individual pure-tone-audiogram thresholds plotted separately for the right and the left ear (in black). The shaded grey area represents the range of the audiometric thresholds and the white line represents the mean at each frequency across the listeners. The dashed line represents the threshold criteria of hearing level $\leq$ 25 dB HL.

Stimuli

The same ASL target sentences were used in the second experiment. Nonetheless, unlike the first experiment where the frequency range of the stimuli extended up to 10 kHz, the stimuli in the current experiment were low-pass filtered at 4 kHz. This was carried out in order to minimise the effects of any possible high-frequency hearing loss in the older-adult group, which is known to increase in prevalence with age (e.g., "Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging", 1990). As in the previous experiment, several speech and non-speech distractors were used. However, only data for speech distractors will be discussed here. The speech distractors in experiment II consisted of either familiar English passages (ENG), originating as before from the EUROM corpus, or unfamiliar Mandarin passages (MDR), spoken by native Mandarin Chinese adult speakers. The Mandarin passages were recorded in the Department of Speech, Hearing, and Phonetics Sciences, University College London (UCL) in an anechoic

chamber and followed similar recording and editing steps as in the EUROM passages (Chan et al., 1995). Each of the speech distractors (ENG and MDR) comprised twenty different talkers (10 same-sex and 10 opposite-sex), with a total of forty different speech passages.

Procedure

A similar experimental design was employed in the second experiment with a few exceptions. Instead of a self-scoring method, listeners were asked to verbally repeat the target sentences to the experimenter who was situated alongside the participant in the sound treated chamber. The experimenter scored the response by selecting the correctly repeated keywords on the screen. Listeners were encouraged to guess if unsure and no feedback was given at any time. Additionally, while in the first experiment the same passage was used throughout the testing, here, a distractor passage was selected at random out of the ten different passages in each trial. Finally, each test condition was measured twice with no repetition of the target sentences. The order of the test conditions was pseudo-randomised.

Results

In the second experiment, listeners were presented with the target sentences without a distractor (Quiet), and with a speech distractor spoken either in a familiar or unfamiliar language (ENG and MDR, respectively) spoken by either same-sex or opposite-sex distractor talkers than the target talker. Each participant was presented with two runs for each test condition with a total of 10 runs (5 conditions x 2 runs).

Within-session test-retest reliability

Descriptive statistics of the listeners performance (in SRdTs) for the different test conditions is given in Tab. 1.4. A comparison between the test runs is depicted in Fig. 1.7, with the SRdTs obtained in the first run (x-axis) plotted as a function of the second run (y-axis). The figure reveals that most observations are fairly close to or on the diagonal line across the different test conditions, which represents an identical performance between the first and the second run.

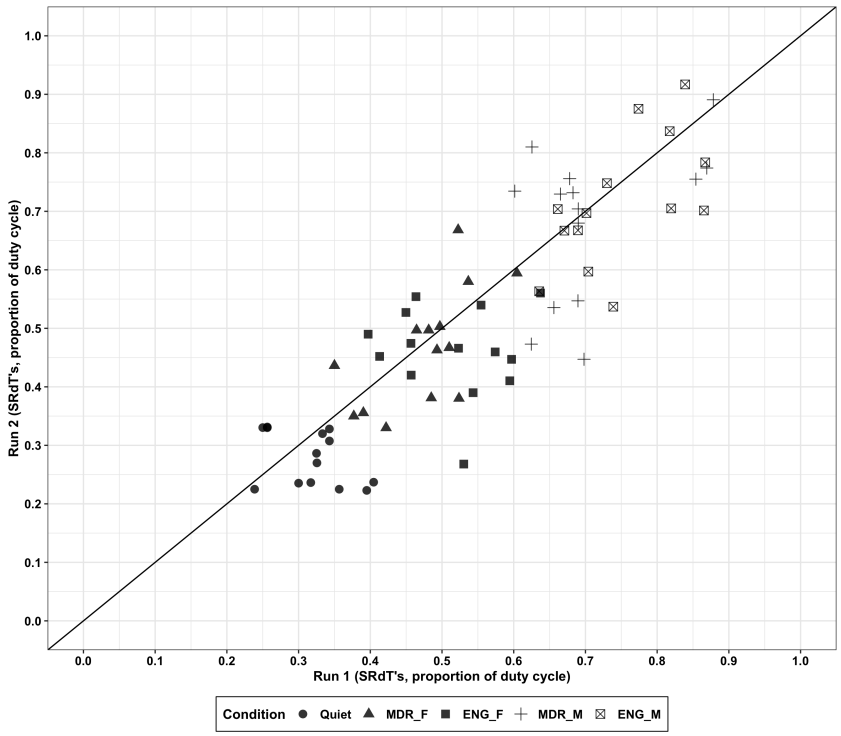

Figure 1.7: Test-retest SRdTs obtained in experiment II for the test conditions Quiet, $\text{ENG}_{opposite-sex}$ and $\text{ENG}_{same-sex}$. Individual scores are represented by the different shapes corresponding to the test condition, whereby the diagonal line represents an optimal agreement between run 1 and 2.

To evaluate the test-retest reliability between run 1 and 2 across the different test conditions, we first calculated the intraclass correlation coefficients (ICCs) using *icc()* in *irr* R package (Gamer et al., 2019). We used the ICC(1) formula for a two-way mixed effects model, with absolute agreement and single measures (cf. Koo & Li, 2016). The ICC is ".. an index of reliability representing the ratio

Table 1.4: Descriptive statistics for SRdTs obtained in experiment II with M indicates the mean and SD for the listeners SRdTs, whereas the grand mean indicates the aggregated data across both experiments.

	Distractor talker-sex					
	Opposite M (SD)			Same M (SD)		
Background type	Run 1	Run 2	Grand mean	Run 1	Run 2	Grand mean
ENG	0.51 (0.07)	0.46 (0.08)	0.49 (0.08)	0.75 (0.08)	0.71 (0.11)	0.73 (0.10)
MDR	0.48 (0.07)	0.46 (0.10)	0.47 (0.09)	0.71 (0.09)	0.68 (0.13)	0.70 (0.11)
	Run 1 M (SD)		Run 2 M (SD)		Grand mean (SD)	
Quiet	0.30 (0.05)		0.32 (0.05)		0.28 (0.05)	

of the between-subject variability to the total variability in the data" (Leensen & Dreschler, 2013, p. 458). An ICC of 1 stands for high reliability and an ICC of 0 stands for no relationship at all. Despite the small between- and within-subjects differences in scores across the two runs, all the calculated ICCs were negative. A negative ICC is typically considered as unreliable and thus considered as an ICC of zero (e.g., Matheson, 2019; Qin et al., 2019). Negatives ICC can arise from several factors such as a small between-subject variance and a small sample size. Since test-retest reliability was not the main objective of the study, it was decided to use a less conservative approach to quantify the difference between the two runs among the different listeners. For this, the null hypothesis that the mean difference between the runs is zero was tested using a paired t-test (*t.test()*, stats package; R Core Team, 2020a). The data met the test assumptions for normal distribution (Shapiro-Wilk test; R Core Team, 2018) and homogeneity of variance (Levene's test; Fox & Weisberg, 2011). The tests results are shown in Tab. 1.5 , where there was no significant difference found between the first and the second run across all conditions (all $p's > 0.05$), thus for further analysis the individual averaged scores were used.

Table 1.5: SRdTs test-retest reliability analysis: paired t-test using *t.test()* function (stats package; R Core Team, 2020).

	Estimated mean difference	95% CI	p-value
Quiet	0.040	-0.007 - 0.087	0.091
$\text{ENG}_{same-sex}$	0.037	-0.015 - 0.089	0.150
$\text{ENG}_{opposite-sex}$	0.052	-0.012 - 0.116	0.100
$\text{MDR}_{same-sex}$	0.024	-0.047 - 0.095	0.480
$\text{MDR}_{opposite-sex}$	0.011	-0.033 - 0.055	0.596

Score reproducibility — a comparison between experiment I and II

Next, the reproducibility of the test scores was examined by comparison of the SRdTs obtained in experiment I (dark gray) and II (light gray) for Quiet and ENG speech distractor for same- and opposite-sex distractor talker(s) (see Fig. 1.8). No listener participated in both experiments. Overall, the averaged SRdT scores in the two experiments were fairly similar across the different condition, with mean SRdTs of roughly 0.32, 0.51 and 0.78, respectively. Nonetheless, there is a small but noticeable tendency for increased SRdTs (i.e., poorer performance) in the first experiment and for a larger variance when compared with the results in the second experiment.

The assumption of normal distribution was fulfilled (Shapiro-Wilk test), however, the assumption of homogeneity of the variance (Levene's test) for the interaction between the two experiments and test conditions Quiet, $\text{ENG}_{same-sex}$ and $\text{ENG}_{opposite-sex}$) was not met ($F(5, 84) = 4.86, p < 0.0001$). Thus, a nonparametric approach using *nparLD()* function (nparLD package; Noguchi et al., 2012) was applied to examine the differences in SRdTs between experiments. The function offers a robust rank-based ANOVA-type statistic test (ATS) for analysis of skewed

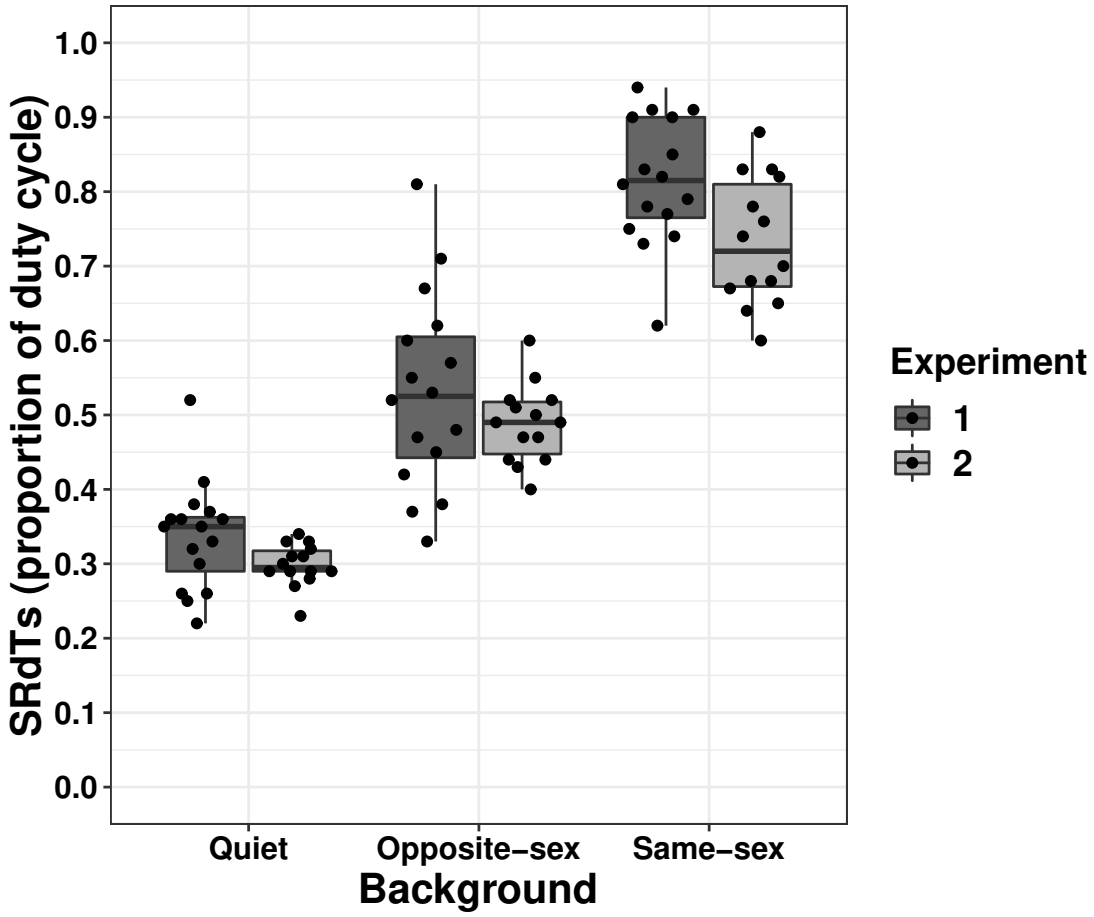

Figure 1.8: Boxplots of the SRdTs obtained in experiment I (dark gray) and experiment II (light gray) for the reference condition Quiet and ENG speech distractor with the same- and opposite-sex talker(s). Individual scores are represented by the black circles.

data or for data with outliers or from a small sample size (see Feys, 2016, for a good introduction on robust nonparametric techniques). The analysis was based on a f1-ld-f1 design ATS test, which refers to an experimental design with a single between-subjects factor (Experiment: I & II) and a single within-subject factor (Condition: Quiet, $\text{ENG}_{opposite-sex}$, & $\text{ENG}_{same-sex}$). There was no significant interaction between Experiment x Condition (Statistic = 0.412, df = 1.74, $p = 0.634$), indicating that performance in the two experiments did not differ between conditions. Whereas there was a highly significant main effect of Condition (Statistic = 271.580, df=1.74, $p < 0.001$) and a significant main effect of Experiment (Statistic = 8.260, df = 1.00, $p < 0.01$). Nevertheless, the effect-size for Experiment was small with a Cohen's d of 0.264 (95%-CI: -0.158 - 0.686), whereas the effect-size of condition was

large with d ranging between -2.280 to -5.850 (*effsize::cohen.d()*; Torchiano, 2020).

Effects of the distractor's language familiarity and talker-sex on IM

A comparison between the listeners' SRdTs measured with the familiar speech distractor (ENG) and the unfamiliar speech distractor (MDR), for same- and opposite-sex distractor talkers, is shown in Fig. 1.9. As before, the diagonal line represents identical performance for the two distractors. The scores were on average very similar in the two distractor-talker configurations, with a DC of roughly 0.5 for opposite-sex and 0.7 for same-sex distractor talkers.

The effect of familiarity of the speech distractor was tested using an 2x2x2 factorial design LMEM with repeated measures, with speech distractors as fixed factor (DistrType: ENG & MDR), distractor talker-sex (DistrTlkrSex: same- and opposite-sex), and the run's order (Order: 1 & 2) as fixed factors, and subjects as random intercepts (reference levels: $\text{ENG}_{opposite-sex}$, Order=1). The model coefficients and p-values are given in Tab. 1.6. A backward model selection, starting from a fully saturated model with three-way interaction for the fixed factors (DistrTlkrSex x DistrType x Order), revealed no significant interaction. The final model did not include interaction terms. Model comparison revealed a highly significant main effect of distractor talker sex ($p < 0.001$) and a significant effect for familiarity with the language of the speech distractor ($p = 0.029$), although, the estimated mean difference (0.03) is very small. Similarly, there was a significant main effect of Order ($p = 0.014$), whilst the overall DC improvement in the second run was again very small (-0.03). The lack of interaction between Order and the other predictors implies that the main effect of Order was the same across the predictors with an overall improvement in the second run. The effect size, Cohen's d, for Order (d = 0.205) was small. The effect size for language was considered "negligible" (d = -0.181) and is much smaller than that for the talker-sex (d =

-2.494, “large”).

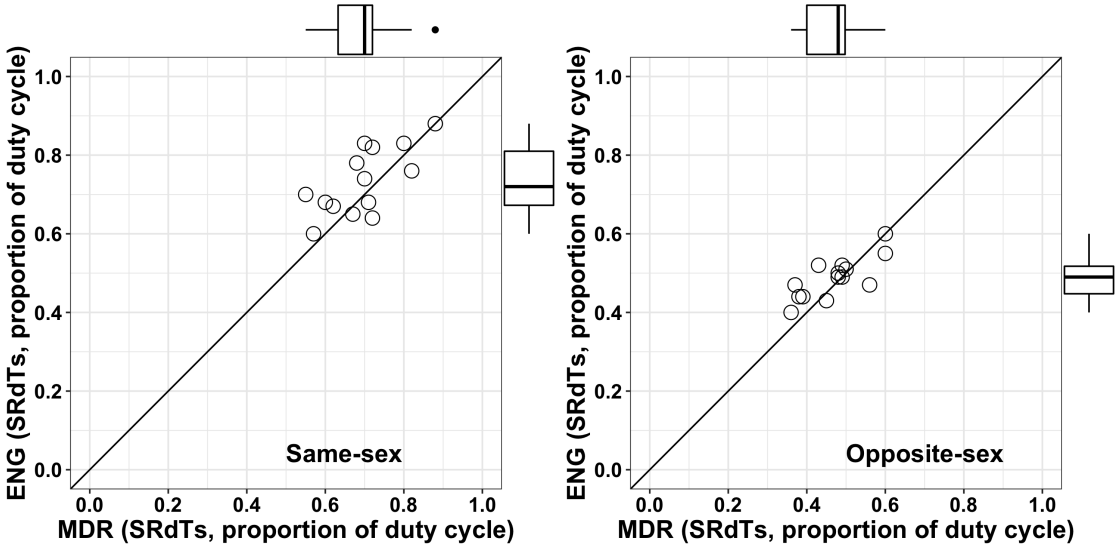

Figure 1.9: SRdTs obtained in experiment II for connected-speech distractors spoken in a familiar language (English, ENG), and an unfamiliar language (Mandarin, MDR) for both same-sex and opposite-sex target/distractor talker configurations. Individual scores are represented by the black circles. The diagonal line represents identical performance for the two speech distractors in the respective distractor talker-sex configuration.

Discussion

Within-session test-retest reliability

Reliability of the outcome measure is an important requirement for both research and clinical use. Reliability reflects the degree to which a test measure is reproducible when measured by the same listener at different points in time. Low reliability negatively affects the test sensitivity, thus making it difficult to detect difference in scores across different test conditions and/or to distinguish whether the listener’s score falls within the normal range (Cameron & Dillon, 2007). Test-retest reliability analysis of the listeners SRdTs showed no significant difference between the first and the second run across the different test conditions with estimated mean difference ranging between 0.014 to 0.047. Thus, the switching task appears to provide reliable

Table 1.6: 2x2x2 mixed-effects model for SRdTs measured in experiment II across all subjects (N observations = 112; N Subjects = 13). Significant p-values are marked as bold.

SRdT ~ DistrTlkrSex + DistrType + Order + (1 | Subjects)			
Main effects	Df	χ^2	p
DistrTlkrSex	1	151.26	**<0.001**
DistrType	1	4.76	**0.029**
Order	1	6.06	**0.014**
Fixed effects	Estimated mean difference	SE	95 % CI
intercept	0.48	0.02	0.44 – 0.52
DistrTlkrSex (same-sex)	0.24	0.01	0.21 – 0.26
DistrType (ENG)	0.03	0.01	0.00 – 0.05
Order (2)	-0.03	0.01	-0.06 – -0.01

and reproducible results which is an important requirement for a clinical tool.

Score reproducibility — a comparison between experiment I and II

Overall, there was a fairly good agreement in SRdTs obtained in experiments I and II across the different test conditions, whereby both experiments showed the same trend of decline in performance when a speech distractor was introduced, with a further decline in performance when the distractor talker was the same sex as of the target talker. Nonetheless, Fig. 1.8 reveal a small but noticeable positive shift in SRdTs (i.e., poorer performance) as well as a larger variance in the first experiment than in the second experiment. Furthermore, a statistical analysis revealed a significant difference in performance averaged across conditions ($p <$

0.01), albeit the effect-size (Cohen's d = 0.264) is considered small.

There are several factors that may have contributed to the observed differences in scores. The smaller variability in the SRdTs in experiment II may have been partially as a result of averaging the listeners scores across the two runs, reducing their variability. In experiment I on the other hand, the listeners were presented only once with each test condition. Another, less likely contributing element stems from the different ways the listeners' response was recorded. Typically, in listening tasks that use (non-matrix) everyday sentences, the examiner records the listeners' verbal response. This method was used in experiment II. The self-scoring method we used in the first experiment was deemed lengthy and may have increased the testing error by imposing fatigue and decline in motivation which may explain the overall small trend of poorer SRdTs in experiment I.

Nonetheless, probably the most influential factor responsible for the difference in scores may be due to differences in the distractor stimuli. In the first experiment, the speech distractor consisted of a random segment taken from a long passage recorded by a single talker. To maximise the similarity between the target and the distractor, the male talker was chosen to have similar voice characteristics as for the target male voice. In experiment II however, each distractor originated from ten different talkers with a varying voice characteristics, from which a short segment was selected at random every trial. The good agreement in performance between the two experiments in the opposite-sex condition (see Fig. 1.8) suggests that when reliable differences in F0 were available, variations in voice characteristics had only a negligible effect on the listener's performance. The IM effect in the opposite-sex distractor talker(s) is likely to be dominated by top-down attentional processing of object-selection, related to target-distractor uncertainty, and may be supported by cues such as phonological cues, semantic content and spatial separation. Such masking interference can take place even when the target and the distractor signals are well formed. The magnitude of the distractor interference also depends on

similarity between the two streams in terms of their voice characteristics. Listeners are able to use F0 differences as little as 6% to considerably improve identification of two simultaneous vowels (Brokx & Nooteboom, 1982). F0 cues are known to facilitate speech perception in noise (e.g., Binns & Culling, 2007; Miller et al., 2010), helping the listener to easily latch onto the target signal after being "lost" by the distractor or by occurrence of an unvoiced speech sound. As for same-sex condition, IM is most likely to be attributed to bottom-up processing, driven by target-distractor similarities (e.g., pitch and prosody) that hinder object formation. One possible explanation for the improved intelligibility in the second experiment may be assigned to the larger set of talkers, resulting in larger variation in talker voice characteristics than in the first experiment which consisted of only a single talker. It is possible that in the second experiment some talkers were more similar to the target talker than others, and that talkers that had less in common with the target talker significantly improved performance when trials were averaged together.

Effects of distractor's language familiarity and talker sex on IM

One of the main objectives of the second experiment was to examine the role of the semantic content of a distractor on IM in the switching task. The distractor's semantic content was controlled by having distractors spoken in a language that the listeners are or are not familiar with.

To our knowledge, no other study has attempted to investigate the components of IM involved in a speech-on-speech listening as presented here; where the target and the distractor signals are interrupted and periodically switched between the two ears out-of-phase with one another. Perhaps the most striking outcome of the first experiment was that only speech distractors impaired task performance. In the absence of a noticeable masking effect for the non-speech distractors, one possible explanation to this is that the ability to ignore a competing talker and to focus on the target talker is hindered by the distractor's semantic content. We therefore

hypothesised that the unfamiliar speech distractor in the second experiment will produce smaller masking interference, resulting in better performance than for the familiar speech distractor. However, in contradiction to our expectation, the listeners did not display a masking release when the target speech was presented with an unfamiliar speech distractor (MDR), with only small difference in performance between the two speech types (ENG vs. MDR). In addition, the non-significant interaction between the distractor type (familiar/unfamiliar) and distractor talker-sex (same/different), indicates that the effect of distractor's talker sex was the same in both distractor types.

The findings in the present study corroborate earlier studies (Brungart & Simpson, 2002; Carlile & Corkhill, 2015; Freyman et al., 2001; Summers & Roberts, 2020), and further support the idea that in some more challenging listening tasks, non-energetic/central masking can also be produced for unfamiliar (i.e., non-intelligible) competing speech. The results further confirm the involvement of other factors than semantic content in masking such as MM and attention. Furthermore, although the use of FxNx speech-like distractor in experiment I did not produce a similar masking effect, it would be interesting to see if we can get a similar masker interference in the task using the garbled speech distractor as used by Carlile and Corkhill (2015) or an unintelligible three-formant buzz-excited vocoded speech as proposed by Summers and Roberts (2020).

1.1.4 General discussion and conclusion

The results in the first experiment showed that perception of switched speech presented with an interleaved speech distractor taps into an aspect of IM that is highly specific, and not probed by non-speech distractors. The results in the present study were comparable to those obtained by Akinseye (unpublished BSc thesis, 2015) for Quiet and $\text{ENG}_{opposite-sex}$ conditions, and are in accordance with other

studies that used interrupted or alternated speech.

We did not observe IM for non-speech distractors, not even for the most "speechy" one (FxNx) and with no other obvious explanation for the lack of IM, we speculated this may be due to the lack of semantic and linguistic information in the nonspeech distractors. Presumably, higher level perceptual cues of lexical and prosodic speech information were prioritised by the listeners over more fine-grained lower-level of acoustic segmentation cues (such as F0 and TFS). Nonetheless, the results of the second experiment speak against this explanation, where we found no or minimal masking release for a speech distractor spoken in an unfamiliar language (MDR). The small difference in IM due to language familiarity could also arise from differences between the talkers. Nevertheless, this is likely to be a less of a factor because several talkers were used and not just one. The remaining burning question is what feature(s) in the MDR distractor facilitated this large target interference?

Moreover, in corroboration with other studies (e.g., Brungart et al., 2001; Festen & Plomp, 1990), the results of the present study demonstrate that similarity between the target and distractor has a large influence on the amount of IM that is produced. A distractor talker of the same sex as the target talker was found to elicit significantly more IM (i.e., poorer performance or larger DC) than a distractor spoken by a talker from the opposite sex to the target talker. Nevertheless, this was only the case for speech distractors, no matter if they were intelligible (ENG) or not (MDR). No IM was found for the non-speech distractors, despite being generated from features extracted from the original speech distractors. The increase in IM for same-sex distractor talker is likely to be caused by a combination of bottom-up failure in object formation in addition to the impaired top-down object selection elicited by an opposite-sex speech distractor.

The amount of IM produced by a speech distractor can vary depending on various voice characteristics of the distractor talker and it's similarity to the target talker voice. While the distractors used in the first experiment originated from one

realisation spoken by a single talker, in the second experiment, each of the speech distractors (ENG and MDR) comprised of different speech passages, spoken by twenty different talkers (10 same-sex and 10 opposite-sex), with a total of forty different speech passages. A comparison with the listeners performance in both experiments showed a fairly good agreement, indicating that listeners' ability to use voice characteristics as cues to segregate sound streams is robust to variations in voice characteristics across talkers.

In conclusion, the present study investigated the utility of a novel speech-on-speech listening task that involves perception of interrupted speech that is switched between the two ears out-of-phase with an interrupted distractor. The proposed paradigm enables us to eliminate peripheral (EM) masking, while maintaining high IM for speech distractors. Providing this "purer" measure of IM may aid in disentangling the reasons why different groups of people experience difficulties in adverse noisy listening situations. One such group is children with developmental auditory processing disorder (APD). APD children typically express difficulties in understanding speech in noisy environments (e.g., a classroom), despite having normal peripheral hearing. There is a growing notion that APD arises from higher-level cognitive deficits [e.g., Moore et al. (2010); DeWit2018]. Since the switching task taps into attentional or other cognitive aspects, it may be useful in better understanding the underlying causes of APD.

More research is required to further understand the underlying mechanisms involved in the switching task. For example, the extent to which listeners are able to obtain information from both ears, as opposed to attending to one ear only, cannot be drawn from the present results and is yet to be examined. The underlying assumption is that the task necessitates sustained and selective attention functions in order to attend to the target signal and to integrate the short-term binaural glimpses of auditory information across the two ears. Nonetheless, determining whether the listeners are attending both ears or only one ear while they carry

out the task may be challenging to confidently estimate. Future studies could for instance compare the listeners' performance with an additional monotic listening configuration, where only the information from either the left or the right ear is presented (i.e., presentation of a single channel out of the binaural stimuli), as opposed to a binaural configuration in which the stimuli are fully preserved when the switched segments are combined across the two ears. Another interesting direction could be to investigate the influence of the speech material (as in its structure and complexity) on performance. Future studies will explore the feasibility of a test version that uses CRM-type sentences (Bolia et al., 2000), e.g., 'Show the <animal> where the <colour> <digit> is'. Furthermore, the ability to attend to the target speech while ignoring a competing distractor can be estimated using a distractor with the exact same structure as the target sentence. Several studies used this technique to estimate the distractor-related response error in CRM sentences (e.g., Brungart et al., 2001). The 'distractor error' reflects the distractor's intrusion, indicated by a response that corresponds with the distractor word rather than the target word. A distractor error reflects attentional aspects of IM, meaning that the listener attended to the wrong stimulus. Such a test version may have several other advantages. It reduces the role of language skills due to the fixed and simpler sentence structure, thus making the task more suitable for both children and adults and potentially for non-native speakers. It also eliminates the need to verbally recall the keywords and enables an automatic testing, negating the need of the examiner to manually score the listeners' responses.

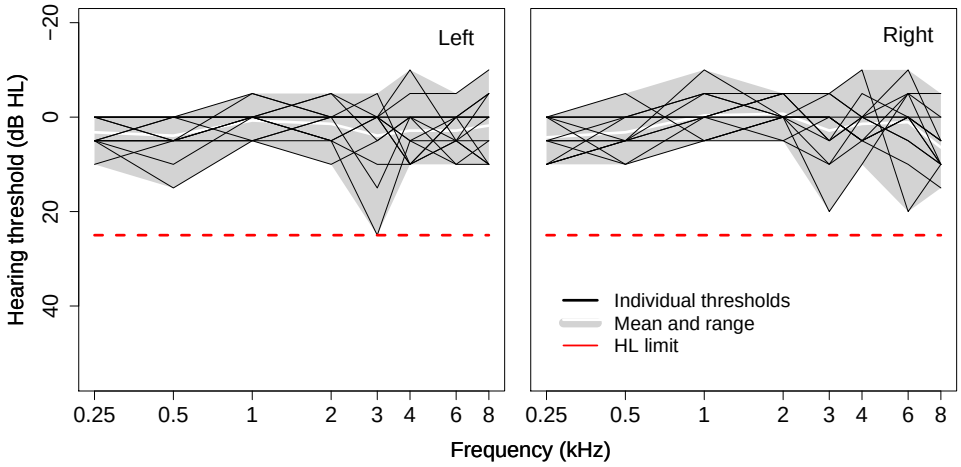

Figure 1.10: Individual pure-tone detection thresholds plotted seperately for the left and the right ear (black lines). The shaded grey area represents the range of the audiometric threholds and the white line represents the mean at each frequency across the listeners. The dashed red line represents the threshold criteria of hearing level $\leq$ 25 dB HL.

1.2 Dichotic vs. monotic presentation and the influence of speech material

1.2.1 Introduction

1.2.2 Methods

Participants

Stimuli

The target stimuli comprised of sentences taken from two type of speech material. The first type was the ASL sentences (MacLeod & Summerfield, 1990) as used in Experiment I and II spoken by a single male talker, which are simple "everyday" sentences presented as an open-set (e.g., *'the clown had a funny face'*). The sentences were scored by the experimenter as described in experiment II using loose keyword scoring method (i.e., errors of case or declension were considered as correct responses). The second type of sentences were taken from the Children's Coordinate Response Measure (CCRM) corpus, which is a locally developed children's friendly version of the Coordinate Response Measure corpus (CRM; Bolia et al., 2000), yet is equally suitable for adults. The CCRM sentences are matrix-based with a

simple fixed syntax of a carrier phrase 'show the where the is', with a set of six animals ('cat', 'cow', 'dog', 'duck', 'pig', and 'sheep'), six colours ('black', 'blue', 'green', 'pink', 'red', and 'white'), and eight digits (from 'one' to 'nine', excluding 'seven' as it is bisyllabic and thus may be more recognisable). The target sentences always started with the animal 'dog', whereas colour and digit words were varied randomly across trials. A closed-set scoring procedure was used, where the listeners were instructed to select the colour-digit combination they heard from an array of six coloured grids, each containing eight possible digits, which was displayed on the screen (see Figure 1.11). Although the number of test items is small, the speech material has low semantic predictability with a guessing rate of only about 2% (1/48). This is because a response was counted as correct only when both the target colour and digit were selected correctly. The listeners were instructed to guess if unsure as only by selecting a digit from one of the colour grids prompted the presentation of the next trial. Visual feedback was given for a correct/incorrect response by displaying a smiling/sad image of a bear instead of the dog picture. No feedback was given for the ASL sentences.

Various talker acoustic characteristics such as F0, accent or speech rate are known to serve as important cues, and can have a positive as well as negative effect on the listeners intelligibility. Something like: "the rate of speech can affect the performance for different duty cycle.."

While using stimulus spoken by a single talker can be advantages as it minimises performance variability, it is also possible that the listeners' used specific cues that may be relevant only to the particular talker used in the study and thus the results may not be directly generalised. Therefore, in the present study, the CCRM target sentences were spoken by three different male talkers which varied at random on a trial-by-trial basis.

Both the ASL and the CCRM target sentences were presented on their own without a distractor (Quiet) and with a competing speech distractor as it showed in the earlier experiments to exert the largest IM. The target sentences were presented

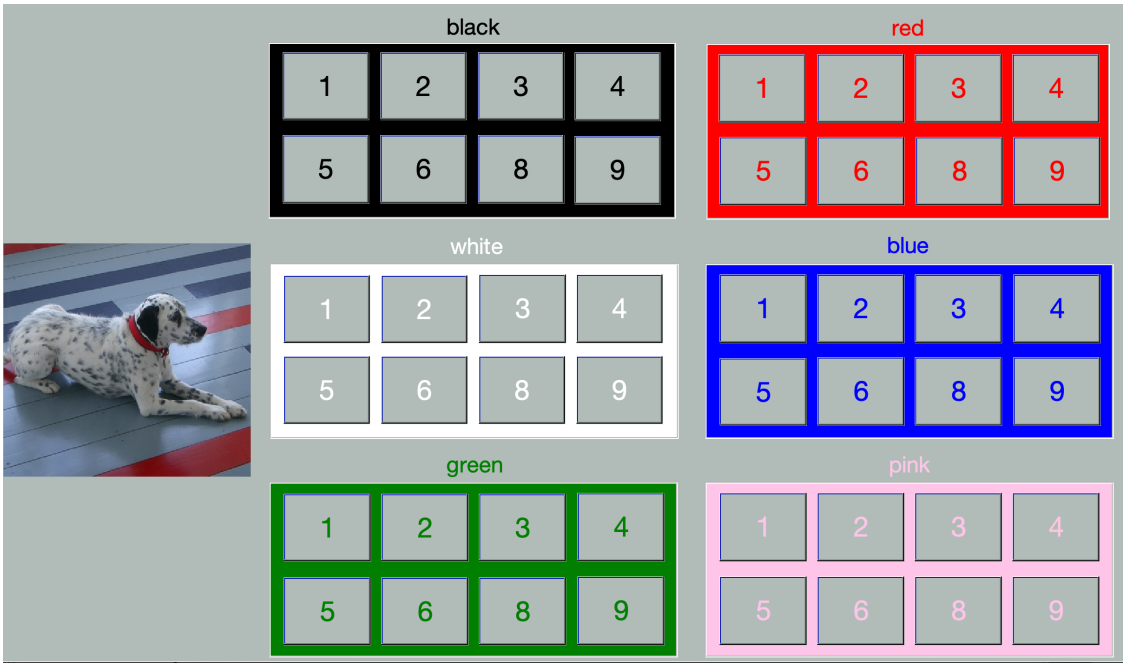

Figure 1.11: The CCRM self-scoring response array which was displayed on the screen during the testing.

with the same English unrelated connected-speech passages (ENG) as described in experiment II, spoken by ten different female talkers (opposite-sex) which were selected at random in each trial. Nonetheless, since the syntax of the CCRM sentences is relatively simple and fixed, presenting them with a competing unrelated speech is likely to exert smaller IM, making it relatively easier for the listener to attend the target speech while ignoring the distractor than for the ASL sentences. To investigate this, the CCRM target sentence were also presented with a CCRM-type sentences spoken by three different female talkers (opposite-sex) with a different animal, colour and digit, chosen at random. The listeners were instructed to listen to any male talker starting with the priming animal 'dog', while ignoring the female talker starting with any other animal. The CCRM distractor started together with the target sentence. Because the individual sentence varied in length, it was possible that in certain trials the distractor was shorter than the target sentence, leaving it's end unmasked. To minimise such cases, the duration of the sentences was equalised across the different talkers to be of a similar length using the 'respeed' feature in the SFS software (version: SFSWin 1.9 Huckvale, 2013). The program

changes the speaking rate without change in pitch and is based on the Synchronised Overlap-Add (SOLA) algorithm of Roucos and Wilgur (Roucos & Wilgus, 1986). The change in speed was employed by a relative rate change factor, where a factor of 2 means changing the signal's rate to be twice as fast, a factor of 0.5 for half as fast, and 1 for unchanged speed. The rate change factor was calculated by subtracting the desired duration (median duration of all sentences of 2.17 s) from the duration of each sentence. As a final step, each sentence was manually corrected to ensure natural sounding and avoiding artefacts. The median duration of the final sentences was 2.17 with a maximal difference in length of circa 0.4 s, which is conveniently about the length of the ending phrase 'is', thus reducing the possibility that one of the target words were left unmasked.

The target (Tar) and distractor (Dstr) segments were presented in three listening configurations (see Figure 1.12): (1) *binaural* (*TarB+DstrB*), in which the stimuli are fully preserved when the segments of the stimuli from both ears are combined, (2) *monaural* (*TarM+DstrM*), where only information in one ear is presented, with only half of the stimuli available to the listener, and (3) *loosely monaural* (*TarM+DstrB*) where the target segments are presented only in one ear, while the distractor segments are fully preserved.

Procedure

The testing comprised of two sessions with a total duration of circa 1.5 hours which took place at a sound attenuated chamber at the SHaPS, UCL laboratory. The stimulus was presented at a fixed output level of circa 70 dB SPL via headphones using the exact same equipment and software as described in Experiment I & II. Similarly, the exact same adaptive procedure was used for speech materials used in the previous experiments. The initial DC was 0.97, with an initial step-size of 0.12 which was gradually decreased over the first three practice reversals until reaching 0.05. A run was completed following 3 practice and 4 test reversals or

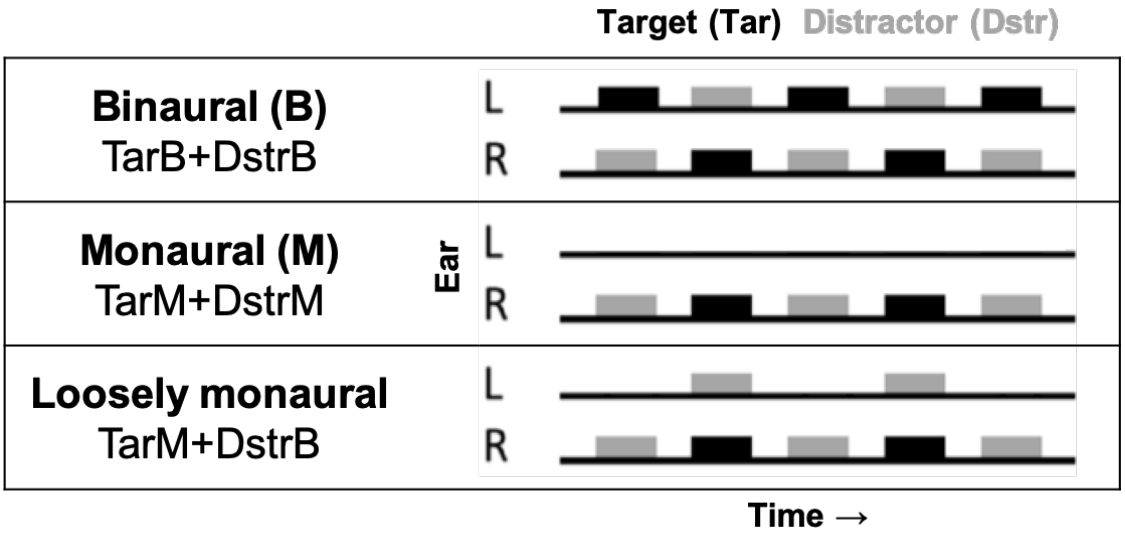

Figure 1.12: Schematic of the switching task listening configurations: Binaural (TarB+DstrB), Monaural (TarM+DstrM), and Loosely monaural (TarM+DstrB). The target speech (Tar) and the distractor (Dstr) segments are represented by the black and grey bars respectively. The ear of presentation (left/right) is given on the y-axis as a function of time.

ended if the maximal number of trials of 30 was reached. The first session comprised of ten conditions (x4 ASL; x6 CCRM), where the target speech was presented with each background stimulus (Quiet, $\text{ENG}_{opposite-sex}$, and $\text{CCRM}_{opposit-sex}$ for CCRM material) in the binaural and the monaural listening configuration. The second session comprised of six conditions (x2 ASL; x4 CCRM), where the target speech was presented with each background stimulus in the monaural and the loosely monaural listening configurations.

Each session started with a practice phase, comprising of a run with five trials for each of the test condition. The initial DC was set to 0.75 in order to familiarise the listeners with the adaptive procedure. Due to low semantic predictability of the CCRM sentences and since the keywords are randomly combined, the same test sentences were used for practice. Whereas, as before, since the number of ASL sentences is limited and because they cannot be re-used within a session due to their high semantic predictability, we used the BKB sentences instead for the practice phase. In addition, each test run started with a short practice of three trials to familiar the listener with the test condition that is about to be presented.

The listener's test ear for the monaural condition was assigned at random and was counter-balanced across the listeners. The test ear was the same in both speech materials and was fixed across the sessions. The order of the test conditions was pseudo-randomised using *Mix()* utility (van Casteren & Davis, 2006) to ensure a fairly balanced frequency of condition per order. In addition, to account for order or fatigue effects, the order of presentation of the two test materials was counter-balanced across the listeners, where about half of the listeners started the session with the ASL sentences, while the other half started with the CCRM sentences. In total, 14 test conditions were recorded per listener where listeners were presented only once with each test condition, except for conditions measured with the monaural listening configuration which was tested twice, once in each session.

ASL list was picked at random from list 1 to 9.

Statistical methods

1.2.3 Results

Test-retest

First, since listeners were presented with the MM configuration twice, once in each session, differences in performance between the two sessions were examined (test-retest).The listeners' SRdTs recorded in the first and the second session are plotted in Figure 1.13. The scatterplot reveals that most observations are relatively close or on the diagonal line which represents an identical performance in session one and two. It is also noticeable that variance in performance was smaller for the ASL than for the CCRM speech material.

Inspection of the data for parametric assumptions revealed that while homoscedasticity of variance was met, the assumption of normal distribution was rejected for the ASL and the CCRM data measured with the ENG_F (different material) distractor. Therefore, differences between sessions were tested using a

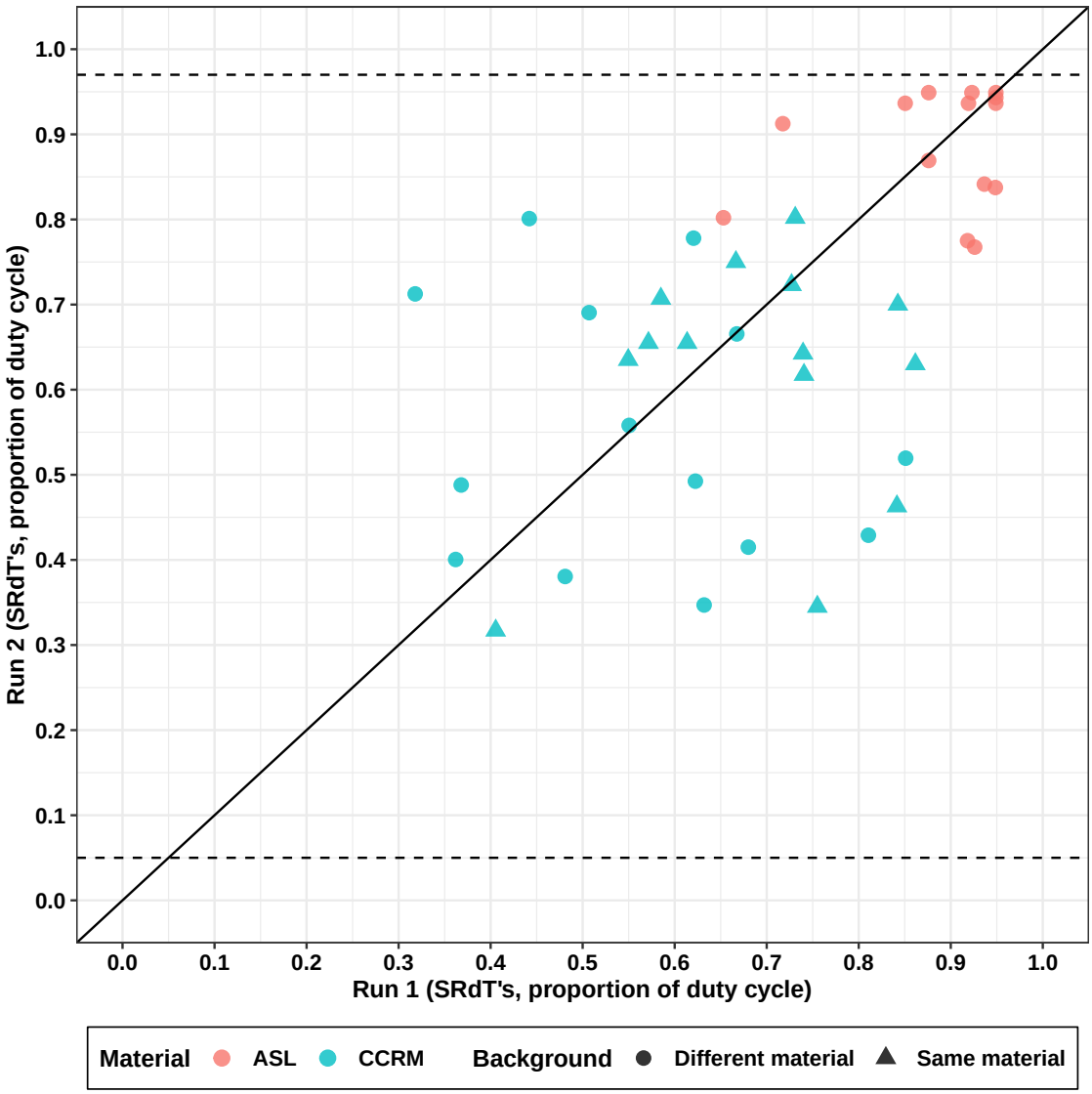

Figure 1.13: Test-retest: SRdTs obtained for 'TarM+DstrM' listening configuration in session 1 (x-axis) and session 2 (y-axis). Indivdual SRdTs are represented by the different shapes and colours corresponding to different material (ENG_F, circles) and same material (CCRM_F, triangle) distractors presented with the ASL (red) or CCRM speech material (cyan). The diagonal line represents the same score in both sessions. The dashed lines represents the task's lower and upper DC limit of 0.05 and 0.97.

noneparametric method using Wilcoxon rank-sum test with permutation (N=999999) (*wilcox_test()* function, coin package, REF). Descriptives and test results are shown in Table **??**. There was no significant difference between SRdTs in the first and the second session across test material and distractor type (all p's > 0.05). Therefore, since only MM condition was tested twice while the rest of the test conditions were only measured once, the individuals' average score across the two sessions was used for further analysis.

Binaural benefit (differences between BB & MM and BB & MB)

Boxplots of the listeners SRdTs split by speech material, background type (Quiet, ENG_F & CCRM_F) and listening configuration are shown in Figure 1.14. It is apparent from the plot that binaural listening greatly improved performance in all conditions across the two speech materials. Furthermore, there is a clear trend for monaural target with binaural distractor (TarM+DstrB) to further degrade performance, especially for the CCRM material. However, there is a clear ceiling effect for ASL performance for a masker in the monaural configurations. Thus making it difficult to determine whether the two monaural configurations in the ASL material differ or not. Nonetheless, this experimental limitation does not hinder the examination of the present stud main research question which was to examine whether the BB and MB or BB and MM significantly differ.

Furthermore, as expected, the CCRM was more intelligible than the ASL material. However, both materials showed the same trend in performance. Finally, as predicted, intelligibility for CCRM-like distractors (CCRM_F, same material) was poorer than for the unrelated connected speech (ENG_F, different material).

Next, difference between the listening configurations and between test materials were statistically analysed with linear mixed-effects regression models (LMEMs) using *lmer()* function (REF). Parametric methods assumption of homoscedasticity of variance was met, while the assumption of normal distribution was rejected for the ASL data for ENG_F condition. This is likely due to the cieling effect seen

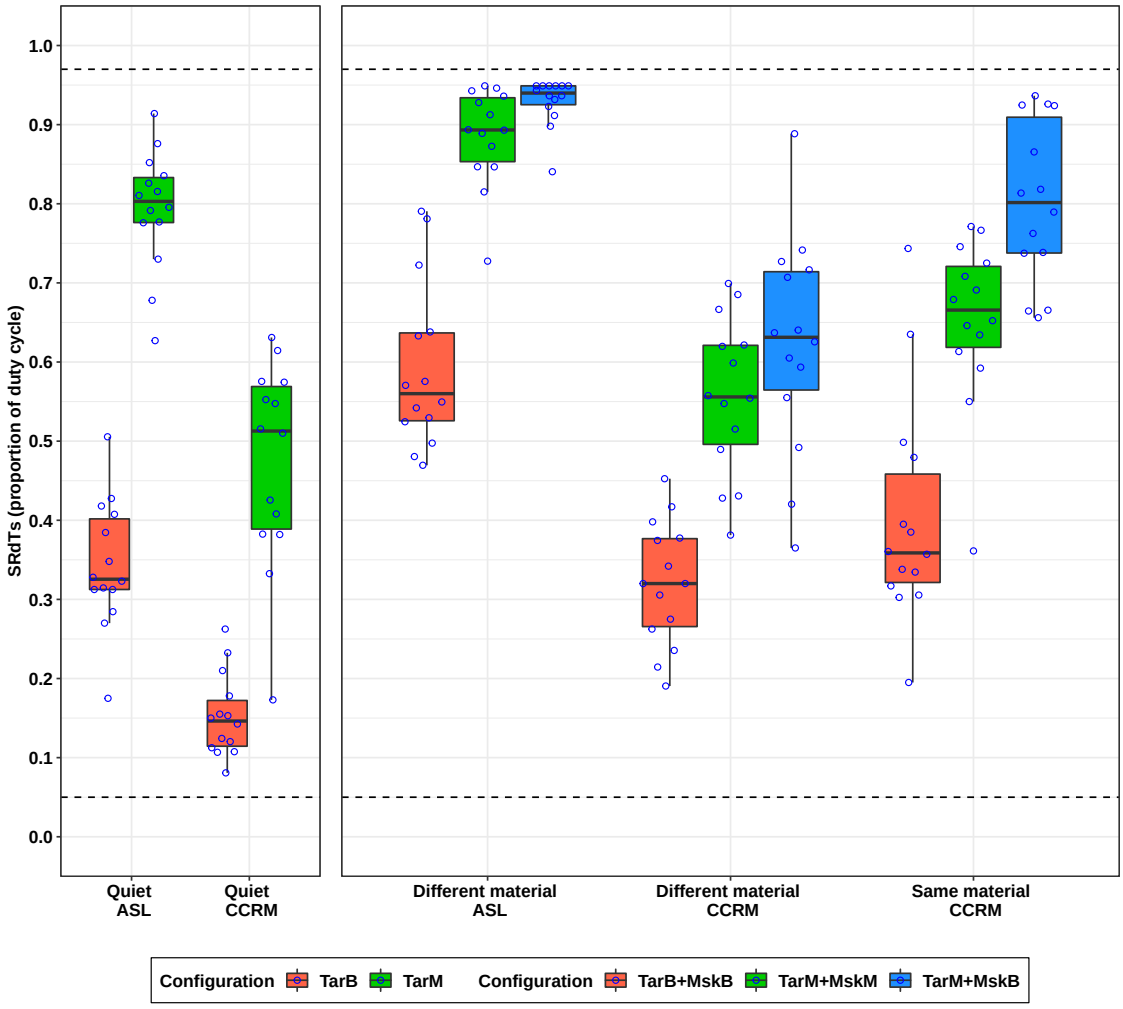

Figure 1.14: Boxplots of the listeners SRdTs split by speech material (ASL/CCRM), background type (Quiet, ENG_F and CCRM_F) and listening configuration (binaural: TarB and TarB+DstrB; monaural: TarM and TarM+DstrM; loosely monaural: TarM+DstrB). The boxes are colour-coded, with red, green, and blue, marking the binaural, monaural and loosely monaural condition, respectively. The dashed lines represents the task's lower and upper DC limit of 0.05 and 0.97.

for the monaural configurations. However, since noneparametric tests gave similar results, it was decided to report here only the outcomes of the parametric method.

The first 2x2 LMEM model examined the effect of speech material (ASL / CCRM) and listening configuration (TarB / TarM) as fixed factors on the listeners SRdTs measured in the reference condition, Quiet, and random intercepts for subjects. There was a significant interaction between material and configuration [$\chi^2(1) = 9.044$, $p = 0.003$]. Thus indicating that performance for target sentences presented monaurally and binaurally is affected differently across the speech materials.

Quiet Model: uRevsM ~ Material · Configuration + (1 | Listener)

- There was a significant Background x Connfiguration interaction
- Post-hoc tests: all Material x Configuration pairs were significant. –> add effect size Cohen's d!

ASL model (ENG_F): uRevsM ~ Configuration + (1 | Listener)

- Data is not normally distributed (due to ceiling effect for MM condition..), yet results were the same when tested with a noneparametric test [nparLD()].
- There was a highly significant main effect of Configuration.
- Post hoc-tests using lsmeans: sig. difference btw BB vs. MM and BB vs. MB. No sig difference btw MM vs. MB -> ceiling effect!

CCRM model (ENG_F,CCRM_F): uRevsM ~ Background + Configuration + (1 | Listener)

- significant main effect of both Background and Configuration.
- Post hoc test for Configuration revealed a highy sig. difference between BB and the 2 monaural configurations. As well as a strong significant difference between MM and MB.

Differences between materials for ENG_F only: uRevsM ~ Material + Configuration + (1 | Listener)

- Data is not normally distributed (due to ceiling effect for MM condition..), yet results were the same when tested with a noneparametric test [nparLD()].
- There was a highly significant main effect of Material and Configuration.
- Post hoc tests for Configuration found a highly significant difference between BB and the two MM conditions. While no significant difference was found between MM and MB (again ceiling...).

The significant Material x Configuration tells us that performance for target sentences presented monaurally (TarM) and binaurally (TarB) is affected differently across speech materials. Listeners show to be affected more by the monaural configuration for the ASL than for the CCRM speech material.

Why? -> Due to the close nature of the responses in the CCRM material.. Why? The CCRM is a closed-set material with a fixed syntax and a small set of test items (8 digits and 6 colours). In addition, the test items differe in their phonemes combinations, thus resulting in a relatively low confusion between test items within a set. For example, the digit 'six' has a unique sounding when compared with the remaining digits or for example the pairs of digits 'four' and 'five', or 'two' and 'eight'. In addition,... it is enough to hear only a short snippet of the item to recognise it from the rest.Thus, while the theoretical guess rate is only about 2%, the effective guess rate is probably higher.

listners are less affected by reducing the target information in half in the monaural configuration

Material effect

1.2.4 Discussion

- *R Markdown: The Definitive Guide* - https://bookdown.org/yihui/rmarkdown/
- *R for Data Science* - https://r4ds.had.co.nz

1.2.5 Conclusion

2

Spatial listening: development and normalisation of a children's spatialised speech-in-noise test

Contents

The magic of R Markdown is that we can add code within our document to make it dynamic.

We do this either as *code chunks* (generally used for loading libraries and data, performing calculations, and adding images, plots, and tables), or *inline code* (generally used for dynamically reporting results within our text).

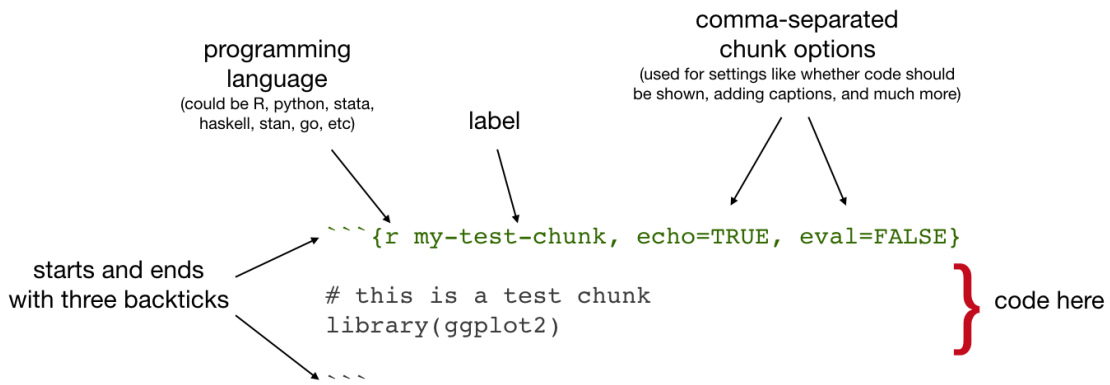

Figure 2.1: Code chunk syntax

2.1 Introduction

2.2 Methods

2.3 Discussion

2.4 Conclusion

The syntax of a code chunk is shown in Figure 2.1.

Common chunk options include (see e.g. bookdown.org):

- `echo`: whether or not to display code in knitted output
- `eval`: whether or to to run the code in the chunk when knitting
- `include`: wheter to include anything from the from a code chunk in the output document
- `fig.cap`: figure caption
- `fig.scap`: short figure caption, which will be used in the 'List of Figures' in the PDF front matter

IMPORTANT: Do *not* use underscoores in your chunk labels - if you do, you are likely to get an error in PDF output saying something like "! Package caption Error: \caption outside float".

3

APD study

Contents

3.1 Introduction

- APD definition: “unexplained idiopathic (spontanous) listening difficulty (LiD) is often termed auditory processing disorder (APD) in children who have symptoms of difficulty hearing and understanding speech, and abnormal results on more complex auditory tests, despite having normal pure-tone hearing sensitivity (Jerger & Musiek 2000; Musiek et al., 2017)” [Hunter et al., 2020]

- Prevalance of LiD ~ 10% (Sharma et al., 2009). Prevalence of LiD complaints with measured NH, complying with APD definition is estimated at ~0.5 to 1% of the general population (Hind et al., 2011; Halliday et al., 2017)

- Association with other developmental disorders and lack of understanding of the underlying auditory deficits of APD.

- “Hearing involves both”bottom-up" (ear to brain) and “top-down” (cortical to subcortical) pathways through simultaneous and sequatial processing (Moore & Hunter, 2013)" [Hunter et al., 2020]

- Two general mechanistic hypotheses of APD:

  (1) **Sensory processing difficulties (bottom-up)**: involving the central auditory nervous system, are based on animal and human lesion studies (Snow et al., 1997). Supporters of this hypothesis suggested this can be assessed using low-redundancy (simple) speech tests (e.g., using added noise, filtering, rapid speech,..) to “stress” the highly redundant central auditory pathways to reveal deficits (Keith 1995, 2000; Cameron et al., 2014).

  (2) **Higher-level cognition or attention (top-down)**: especially in children with language disorders (Rees, 1973; Moore et al., 2010).

Individuals may have a combination of both.

- There is no accepted consensus or gold standard diagnosis of APD (Wilson & Arnott 2013)
- Possible link between OME (+ grommets) or EHF HL and APD in a sub-group of children.
- OME related HL has been shown to persist after recovery at frequencies above 4 kHz (Hunter et al., 2020; REFs..)
- OME or EHF HL can potentially be a basis for poorer speech perception, especially in noise. Findings are not consistent. Studies that tested both TD and APD with OME or EHF HL found that they are predictors of measurable peripheral damage in both groups.
  - Besser et al. (2015) and Levy et al. (2015) found that better thresholds between 6 to 12.5 kHz were associated with better reception of speech in noise (adult studies).
  - Motlagh Zadeh et al. (2019): impairment in higher frequency regions could negatively impact speech perception.

  Conductive loss results in impaired spatial processing (Cameron et al., 2014) and binaural interaction (Hall et al., 1995; Hogan et al., 1996)

3.2 Methods

3.2.1 Participants

Forty-four primary school children who were native British English speakers with normal hearing acuity participated in the study. Amongst them twenty-one belonged to the APD clinical group (5 females) with an average age of 11.0 ± 1.4 years (range: 7.8 - 12.9 years). The remaining twenty-three (12 females) comprised of typically developing control children (TD) with no reported concerns or diagnosis

of an auditory, language or other cognitive developmental disorder. The TD group average age was 9.5 ± 1.6 years and ranged between 7.0 to 12.1 years. Since not all the measurement equipment was easily portable and in order to maintain the same environment during the assessment across the complete sample, the children and their caregivers were required to travel to central London for the testing. In order to maximise the number of children taking part in the study, 8 out of the 23 TD children (35%) had an APD sibling (TD_{sib}) which took part in a parallel study that took place on the same day of testing. All the children who participated in the study were required to have normal hearing acuity, defined as thresholds $\leq$ 25 dB HL at the octave frequency bands between 0.25 to 8 kHz and their eardrum had to be visible, healthy and intact in both ears following otoscopic inspection. One APD participant was excluded from the analysis due to raised thresholds predominantly in the right ear, ranging between 30 to 45 dB HL (PTA_{Right} = 36.25 dB HL; PTA_{Left} = 13.75 dB HL), thus resulting in a final APD group size of twenty[1]. Otoscopic inspection of the child's ear canal revealed a large accumulation of cerumen in both ears with an occluded right ear. Two additional children (x1 APD, x1 TD) had slightly raised thresholds at 8 kHz in one ear of 35 and 30 dB HL, respectively. However, since thresholds at all other frequency bands were well within the $\leq$ 25 dB HL criteria they were not excluded.

APD children were recruited in two ways. Children diagnosed with APD at Great Ormond Street Hospital (GOSH) or at the London Hearing and Balance Centre (LHBC), London, UK, were identified based on their clinical records and were contacted by a clinical team member. The caregivers were provided with information about the study and means of contact to express interest in participation. Others, including the TD group were recruited by advertisements on social networks (e.g., APD Support UK Facebook group), science events, local information boards and UCL staff newsletter email, where parents were requested to fill-out an online interest

[1]PTAs were calculated by averaging the individual's thresholds at the frequencies 0.5, 1, 2 and 4 kHz separately for the right and left ear (PTA_{Right}, PTA_{Left}).

form with short screening questions to ensure that the child met the participation requirements. Most of the children in the APD group (85%, 17/20) were reported to undergo an APD assessment at GOSH, about a third were directly recruited from the clinic. The remaining three were reported to be assessed at the LHBC, at the University of Southampton Auditory Implant Service or the Chime Audiology Royal Devon & Exeter Hospital (screening only).

Our initial aim was to take a conservative stance on inclusion criteria by including only those who met a clinical APD criteria (2 SD below the norms on two or more tests during the assessment). Moreover, being aware of the high prevalence of APD children with additional co-occurring developmental disorders, we strived to recruit children who displayed a "pure" form of APD without reported diagnosis or concerns for additional developmental disorder/s. However, very few APD children met these strict criteria. Only 75% (15/20) met the clinical criteria of APD, out of which 60% (9/15) were diagnosed with spatial processing disorder (SPD) due to abnormal SRM in the LiSN-S task (see Table 3.1 for descriptives of the APD group). Of the remaining children in the APD group, four did not meet the diagnostic criteria for various reasons (e.g., young age, lack of psychological educational evaluation report and the need to exclude other deficits), however their assessment report acknowledged some "auditory processing difficulties", whereas the fifth child awaited an APD assessment following an APD screening. Due to the small sample-size these children were included in the APD group for the analysis. Nevertheless, they were subdivided as children with Listening Difficulties (LiD) and differences in performance LiD and APD children were later explored. Furthermore, half of the APD group (10/20) were reported for being diagnosed with one or more secondary developmental disorder/s (x6 Dyslexia, x3 HF-ASD, x3 DLD, x1 ADHD, x1 ADD, x1 Dyspraxia, x1 visual stress, x1 sensory integration disorder, and x1 poor short-term working-memory). Nonetheless, several caregivers reported that their motivation for seeking additional diagnosis was to get more help from the school, rather than a real concern, after feeling that their support for their child's

Table 3.1: APD group demographics and APD-related history background.

School type	85% (17/20) Mainstream (1 child in a special ASD unit, 2 in a private school), 15% (3/20) non-mainstream school
Assessment location	85% (17/20) GOSH, 15% (3/20) other
APD Diagnosis	75% (15/20) APD, 25% (5/20) LiD
SPD subtype	60% (9/15) SPD
Additional disorder (diagnosed)	50% (10/20) secondery developmental disorder/s
Additional disorder (undergoing assessment)	25% (5/20)
MEHx	60% (12/20)
PET history	25% (5/20)
FM-device usage	55% (11/20)
Auditory training	35% (7/20)

MEHx: History of middle ear problem
PET: Pressure equalisation tube

APD was lacking.

Caregivers from both groups completed a comprehensive background questionnaire, similar to the one that is typically given prior to an APD assessment, concerning the caregiver/s educational level, child and family history of hearing, listening problems and developmental disorders, child history of otitis media with effusion (OME), pressure-equalisation tubes (PET / grommets), pregnancy-related questions (e.g., complications, prematurity, etc.), APD-related (e.g., date of diagnosis, location, use of FM device and auditory training), any diagnosis or concerns regarding the child's speech, language, educational and/or cognitive skills, speech and language therapy, medication taken, musical training and the type of school the child attends.

Children in the APD group were on average 1.5 years older than children in the TD group. Difference in age between the two groups was tested with a one-way ANOVA using *anova()* function (parametric assumption of normal distribution and homoscedasticity were met). The test revealed a significant difference in age between the groups [$F(1,41) = 11.58$, $p < 0.01$]. Nonetheless, since age is often reported as a strong predictor for performance in other similar behavioural studies, analysis of the results obtained in the current study was conducted for age-independent scaled scores and should not affect the comparison between the two groups. The project was approved by the UCL Research Ethics Committee (Project ID Number 0544/006) and the NHS Health Research Authority (REC reference: 18/LO/0250). The testing commenced once an informed consent was given by both the caregiver and the child.

3.2.2 Measurements

The test battery used in the present study is described in the following section and summarised in Table 3.2.

Auditory evaluation

Standard & extended high-frequency (EHF) audiometry

Otoscopic inspection was performed prior to the audiometric test to ensure the ear was clear from cerumen and to avoid harming the eardrum when inserting the ear probe. Both standard and extended high-frequency (EHF) audiometry thresholds were measured using the Hughson-Westlake manual procedure, starting from 1 kHz. Standard air conduction pure-tone audiometry was carried out at six octave frequency bands ranging between 0.25 to 8 kHz using a standard clinical manual audiometer via headphones.

Extended high-frequency pure-tone detection thresholds were manually measured at four frequencies 8, 11, 16, & 20 kHz using locally written MATLAB based software which generated the stimuli and recorded the data. Target tones were pulsed (3

Table 3.2: Summary of the study test battery.

Task	Information	Measure
Standard & extended high-frequency (EHF) audiometry	Pure-tone detection thresholds measured at the octave frequencies between 0.25 and 8 kHz (standard), and 8 to 20 kHz (EHF).	Detection threshold in dB HL
Switching task (ST)	Adaptive speech-on-speech listening task that involves perception of interrupted and periodically segmented speech that is switched between the two ears out-of-phase with an interrupted distractor. ST assesses the ability to switch attention and integration of binaural information.	Proportion of speech required to understand 50% of the keywords, Speech Reception duty cycle Threshold (SRdT)
Listening in Spatialised Noise Sentences UK (LiSNS-UK)	Locally developed version of the LiSN-S (Cameron & Dillon, 2007), an adaptive speech-on-speech listening task that assesses the ability to use spatial release from masking (SRM), measured as the difference in perception between collocated and separated speech distractors.	Signal-to-noise-ratio (SNR) yielding 50% speech intelligibility, Speech Reception Threshold (SRT)
Speech-spectrum-noise (SSN)	Conventional adaptive speech in noise task that asseses speech perception of ASL sentences (MacLeod & Summerfield, 1990) in a speech-spectrum-noise with a spectrum matched to the ASL material.	SRT
The Environmental Auditory Scene Analysis task, ENVASA (Leech et al., 2009)	Non-linguistic self-administered task that involves detection of everyday environmental sounds presented in naturalistic auditory scenes and can be used to asses IM effects as well as sustained selective auditory attention skills.	%-correct
Recalling sentences, CELF-RS (Wiig et al., 2017)	A subtest from the Clinical Evaluation of Language Fundamentals UK 5^{th} edition (CELF-5-UK) which assesses expressive language skills, measured by the ability to repeat in verbatim sentences with varying length and complexity. Standardised for children aged 5 to 16 years.	Age-corrected scaled scores
The Evaluation of Children's Listening and Processing Skills, ECLiPS (Barry & Moore, 2014)	Standardised questionnaire comprised of 38 statements grouped into five categories designed to identify listening and communication difficulties in children aged 6 to 11 years. Respondent agreement is expressed using a five-point Likert scale ("*strongly agree*" - "*strongly disagree*").	Age-corrected scaled scores
The Children's Communication Checklist 2^{nd} edition, CCC-2 (Bishop, 2003)	Standardised questionnaire comprising 70 items designed to screen language and/or communication problems in children aged 4 to 16 years. Items consist of a behaviour statement (e.g., "*Mixes up words of similar meaning*") with respondents asked to judge how often the behaviours occur using a four-point Likert scale (0-3).	Age-corrected scaled scores

repetitions) with a duration of 700 ms and 50 ms rise/fall time. EHF measurements took place in a sound attenuated chamber with the child sitting in the chamber while the examiner was situated outside. Communication during the testing was carried out via a video-audio intercom system. The child was instructed to raise his/her hand each time s/he heard a tone. The MATLAB script was executed using a Windows PC which which was connected via USB to an RME FireFace UC sound card (Audio AG, Haimhausen Germany) and an ER10X Extended-Bandwidth Acoustic Probe System (Etym$\bar{o}$tic Research, Elk Grove Village, IL, USA). Stimuli were presented via an otoacoustic emission probe with silicon tips in variable sizes (between 8 to 13 mm), depending on the size of the child's ear.

Standing waves in the ear canal produces spatially non-uniform sound pressure at frequencies above 2-3 kHz, introducing calibration errors when estimating the sound pressure level arriving at the eardrum (Lee et al., 2012; Richmond et al., 2011; Siegel, 1994). Together with other factors such as individual variations in the ear canal length and differences in depth in which the ear probe is inserted into the ear canal, these factors can introduce up to 20 dB calibration error (Siegel, 1994). To account for that, in-situ forward-pressure-level (FPL) calibration was applied using ARLas MATLAB-based software package (Goodman, n.d.). Thereby improving the accuracy of the threshold estimates, especially at high frequencies (Lewis et al., 2009). The target stimulus was converted from dB SPL to dB HL following minimal audible pressure values measured across 84 NH listeners aged 10 to 21 years (see Table 1 in Lee et al., 2012). Frequency-specific weighting factors were estimated using a logistic function on a log frequency scale with a cut-off of 2 kHz with a 2 octave wide transition frequency. A fixed maximum presentation level of 50 dB HL was set to ensure that the listeners are not exposed to potentially harmful sound levels, especially at higher frequencies.

Switching task (ST)

Estimating the effect of IM while minimising peripheral EM on speech perception was measured using the switching task (ST) which is believed to assess the listeners ability to switch attention and integrate of binaural information. The same test procedure and equipment was used as described in Chapter 1. Listeners were presented with both test versions using the ASL and the CCRM speech material. As for the stimuli, the ASL target sentences, spoken by a single male talker, were taken from the final sentences selected following the normalisation study. In addition, a level correction was applied to each sentence using the sentence-specific weighing factors estimated in the normalisation study (see Chapter 2). The first five test lists out of the eight phonetically-balanced normalised test lists (25 sentences each) were used, whereby their order was quasi-randomised to account for order, masker combinations, and fatigue effects. The target CCRM sentences were the same as described in Chapter 1, spoken by three different male talkers. These were selected at random every trial and always began with the priming animal 'dog'. The target speech material was presented either without a distractor (Quiet), with and without switching (NoAlt / Alt) or with a distractor. A selection of four distractors were used (see Chapter 1 for detailed description): English (ENG_F) and Mandarin (MDR_F) unrelated connected-speech, each spoken by ten different female talkers, and a non-speech amplitude-modulated speech-spectrum-noise (AMSSN) with the envelope of a single talker out of 40 talkers (20 females). The fourth distractor was presented only with the CCRM speech material and comprised of CCRM target-like sentences (CCRM_F) with a different priming animal, colour and digit, spoken by ten different female talkers. Each participant was presented with a total of 11 runs, one for each test condition, with 5 conditions for the ASL (Quiet-NoAlt,Quiet-Alt, MDR_F-Alt, ENG_F-Alt), and 6 for the CCRM (with the additional CCRM_F-Alt condition).

The starting DC was 0.97 (i.e., signal is almost entirely present) which is the SRdT upper limit. Subsequently, the DC varied depending on the listeners response, with an initial step-size of 0.12 which decreased gradually over the first three (practice) reversals to 0.05. Nonetheless, as in the adult studies (see Chapter 1),

the minimum step-size for the ASL conditions ENG_F and MDR_F was set to 0.1. This is because of a pilot data which suggested that the psychometric functions of these conditions were shallower.

Testing started following a practice phase, where four trials of each of the eleven test conditions were presented. Practice runs started at an easy-to-moderate DC of 0.8 in order to expose the listeners to the adaptive procedure. In addition, every test run started with two practice sentences (initial DC = 0.97) to orient the listeners to the test condition that was about to be presented.

Listening in Spatialised Noise Sentences UK (LiSNS-UK)

The locally developed Listening in Spatialised Noise Sentences UK (LiSNS-UK) assesses the ability to use binaural cues in speech-on-speech listening conditions. The test development, speech material normalisation, and norms standardisation followed Cameron and Dillon (2007) and are described in detail in Chapter 2. The test uses virtualisation techniques to create a spatial distribution of sound sources in space for headphone presentation where target sentences (ASL; MacLeod & Summerfield, 1990) are presented in two simultaneous speech distractors (unrelated children's stories spoken by the target talker). The LiSNS-UK comprises two main listening conditions, differing in their availability of spatial cues. The target sentences are configured to always appear in front of the listener at 0° azimuth on the horizontal plane, with the two streams of speech distractors either collocated in space with the target (S0N0), resulting in relatively poor speech perception, or offset in space, with one distractor to either side of the target at ± 90°. The spatial separation in the latter condition results in an improvement in speech perception of circa 13 dB (Cameron et al., 2011), typically termed as spatial release from masking (SRM). This SRM advantage is calculated by taking the difference between performance in the collocated and the separated condition.

Speech distractors were presented continuously throughout a run at a fixed 65 dB SPL output level and comprised of a combination of two out of three available passages. A 1-up/1-down adaptive procedure was used, varying the level of the target talker relative to the distractors depending on the listener's response to measure their speech reception threshold (SRT), i.e., the signal-to-noise-ratio (SNR) yielding 50% speech intelligibility. A 200 ms long reference cue (1 kHz pure-tone) was presented 500 ms before the target sentence onset at 65 dB SPL. The initial target output level was 75 dB SPL for the collocated condition and 70 dB SPL for the separated condition with an initial step-size of 4 dB SNR. The step-size was reduced after the first three reversals, reaching a minimum step-size of 2 dB SNR. The adaptive procedure ended once all 25 test trials were presented and stopped in case a maximal output level of 89 dB SPL was reached more than three times. Nonetheless, such an event did not occur in the present study. Since each listener was only presented once with each condition, it was decided not to introduce any other stopping rules that could have expedited the testing time but may as well have introduced an estimation error for the SRTs in some cases. The SRT was calculated by averaging the test reversals SNRs, whereby test reversals were defined as any reversals following three practice reversals.

The order of the listening condition, test lists, sentences within a run, and distractor combinations was fixed across all the participants and started with the collocated condition. Each test list consisted of 25 sentences taken from the 8-phonetically-balanced ASL test lists which were constructed following the normalisation study with a sentence-specific level corrections (see Chapter 2). Spatialisation was applied by convolving each stimuli with head-related transfer functions (HRTFs) at the corresponding azimuthal direction separately for the left and the right channel. The HRTFs were measured with a Knowles Electronics Manikin for Acoustic Research (KEMAR) with a small pinnae taken from the

CIPIC HRTF database[2] (see Algazi et al., 2001, "special" HRTF data). A post-equalisation step was applied in order to flatten the magnitude of the headphone frequency response. Headphone-to-ear Transfer Functions (HpTFs) measured with a KEMAR manikin for HD-25 supraaural headphones were extracted from the Wierstorf et al. (2011) HRTF database. The final mixed stimulus was filtered with the inverse HpTFs separately for the left and the right channel before being combined together as a final step. Every participant was presented with two runs, one for each listening condition (collocated / separated). Testing started following a practice phase of two runs, one for each of the test conditions with five BKB sentences each (Bench et al., 1979). Listeners were instructed to verbally repeat the target sentences to the experimenter who was situated alongside in a sound treated chamber. The experimenter scored the response by selecting the correctly repeated keywords on the screen. Listeners were encouraged to guess if unsure while no feedback was given at any time. A loose keyword scoring method was used, whereby errors of case or declension were considered as correct responses, e.g., a repetition of the keywords '<clown**s**> <funny> <face**s**>' to the stimulus 'The <clown> had a <funny> <face>'.

Speech-spectrum-noise (SSN)

A speech-in-noise test was used as a more conventional listening task that is widely used in the clinic as opposed to the more complex listening conditions measured by the ST or the LiSNS-UK. The normalised ASL sentences were presented in a speech-spectrum-noise (SSN) with spectrum matched to the ASL corpus. The SSN onset was 500 ms before the target sentence began. The same adaptive procedure as for the LiSNS-UK was used with the same stopping-rules and SRT calculation. Each listener was presented with a single run of 25 sentences following a practice phase with seven BKB sentences. The same test list and sentences order was used across all the listeners.

[2]The database is available online in: https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/

The Environmental Auditory Scene Analysis task (ENVASA)

In analogy to the classic 'cocktail-party' scenario, ENVASA is a non-linguistic paradigm (Leech et al., 2009) that measures detection of everyday environmental sounds presented in naturalistic auditory scenes and can be used to asses IM effects as well as sustained selective auditory attention skills. In the task, short environmental target sounds (e.g., a dog's bark, a door knock, or a bouncing ball) were presented in a dichotic background scene (i.e., the target sound is presented only in one ear), consisting of either a single background scene, presented in both ears, or two background scenes, each presented in a different ear. The number of targets, the onset time and the ear of presentation varied across trials. Four SNRs were employed split into two categories 'low' (-6 and -3 dB) and 'high' (0 and +3 dB). Target-background contextual agreement was manipulated by embedding the target sound in a *congruent* background scene that is in agreement with the listener's expectations (e.g., a cow's 'moo' in a farmyard scene) or in an *incongruent* background scene which violate these expectations (e.g., a cow's 'moo' in a traffic scene). A schematic illustration of a single test sequence is shown in Figure 3.1.

The experiment was carried out using the original setup as described by Leech et al. (2009). Sounds were presented via Sennheiser HD-25 headphones (Wedemark, Germany) and the participants response was recorded using a USB-wired gamepad (Saitek Rumble P3200). The output level was adjusted to a comfortable level before the test started. The participants were situated in front of a laptop and were instructed to hold the gamepad. Prior to the test, the listeners were presented with a short child-friendly demonstration video with audio instructions. Next, a short recap was given verbally by the examiner and an exemplary trial was simulated together with the child to ensure that the child fully understood the task's instructions. The task began with three short practice trials with provided feedback, while no further feedback was given during the test phase.

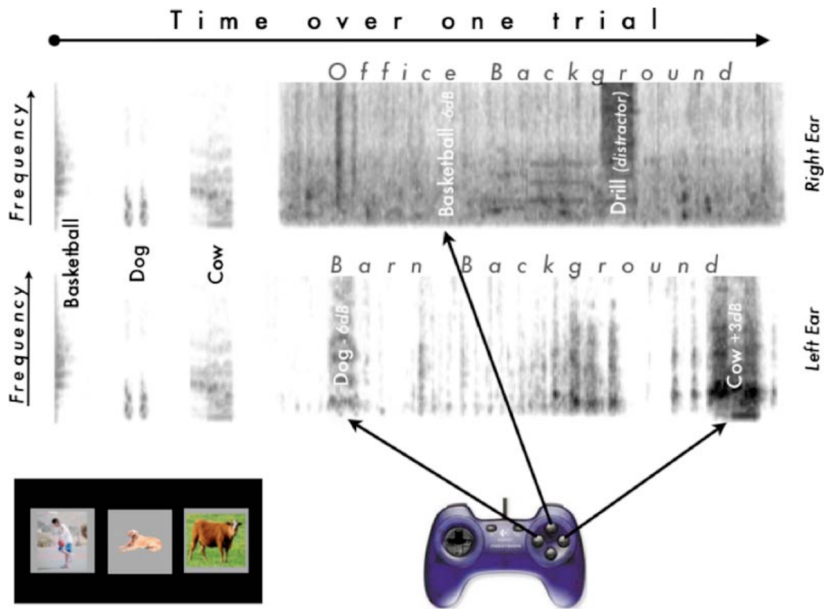

Figure 3.1: Schematic of the ENVASA experimental paradigm (taken from Leech et al., 2009)

Every trial was made of two parts, starting with a target audio and visual familiarisation phase before the main target detection phase. Target identification was recorded by pressing one of the three buttons on the gamepad which corresponded to the location of the target objects on the screen. A response was counted as correct only if the participants pushed the corresponding button within 2 seconds, 300 ms after the target onset. The outcome measure was calculated as the percentage of target sounds correctly identified within a condition (%-correct). In total there were 115 target sounds presented over 40 trials, where 46 target sounds were presented in a single background condition and another 46 in a dual-background condition. The 23 remaining target sounds served as foil items which were played at 0 dB SNR without a corresponding picture on the screen. The order of the foil items was quasi-randomised and was used to estimate the quality of the participants performance.

CELF-RS

The Recalling Sentences (RS) sub-test of the Clinical Evaluation of Language Fundamentals UK fifth edition (CELF-5-UK; Wiig et al., 2017) was administered

to assess the listeners expressive language skills, measuring the ability to repeat in verbatim sentences with varying length and complexity. Standardised norms are available for children aged 5 to 16 years. The CELF-RS is simple and quick to administer and has been shown to be a good psycholinguistic marker for children with Developmental Language Disorder (DLD) and to provide high levels of sensitivity and specificity (Conti-Ramsden et al., 2001), thus making it a good screening tool. Scoring was done by hand by the examiner as instructed by the test manual. The sentences were presented using a local MATLAB program via headphones using the same experimental equipment as listed above at a comfortable output level of 70 dB HL. The sentences were spoken by a female speaker with a standard southern British English accent and were recorded in a sound-treated recording booth at the Speech, Hearing and Phonetics Sciences (SHaPS), UCL laboratory, London. The task began with two practice sentences while the number of test items varied depending on the child's age and performance. No repetitions or feedback were given during the testing and the test was discontinued in case the child failed to score any points for four consecutive items. Age-scaled scores were calculated based on the test norms with a mean score of 10 and SD of 3. Scaled scores within $\pm$ 1 SD from the norms mean (between 7 to 13) are classified as average scores, whereas performance beyond $\pm$ 1 SD are classified as above / below the average score, with scaled-scores < 7 considered as abnormally poor.

Questionnaires

The Evaluation of Children's Listening and Processing Skills (ECLiPS)

The ECLiPS questionnaire (Barry & Moore, 2014) comprises 38 items, where the respondents are asked to express their agreement on simple statements about the child's listening and other related skills or behaviours using a five-point Likert scale (from "strongly agree" to "strongly disagree"). The ECLiPS was designed to identify listening and communication difficulties in children aged 6 to 11 years. Nonetheless, in their evaluation study, Barry and Moore (2014) found little to

no age effect in many of the scale items, suggesting that testing age could be extended below and beyond the population used for the development. Based on factor analysis the items were grouped into five subcategories: 1. Speech & Auditory Processing (SAP), assessing ability to interpret speech and non-speech input, 2. Environmental & Auditory Sensitivity (EAS), estimating the ability to cope with environmentally challenging conditions, 3. Language, literacy & laterality (L/L/L), assessing different abilities that are known to be coupled with language and literacy difficulties, 4. Memory & Attention (M&A), covering short-term and serial memory as well as attention, 5. Pragmatic & Social skills (PSS), assessing pragmatic language or non-normative social behaviours. Aggregated measures were calculated for *Listening* (SAP, M&A, & PSS), *Language* (L/L/L & M&A), *Social* (PSS & EAS), and a *Total* aggregate, calculated by taking the mean of scores across all the sub-scales. Individual age- and sex-scaled scores were computed using the test Excel-based scorer. A score below the 10^{th} percentile (corresponding to a scale score of circa 6) is generally considered clinically significant.

The Children's Communication Checklist 2^{nd} edition (CCC-2)

Communication abilities were assessed using the Children's Communication Checklist second edition questionnaire (CCC-2; Bishop, 2003) which is designed to screen communication problems in children aged 4 to 16 years. It comprises 70 checklist items each consisting of a behaviour statement, like "*Mixes up words of similar meaning*". The respondents are asked to judge how often the behaviours occur using a four-point Likert rating scale: 0. *less than once a week (or never)*, 1. *at least once a week, but not every day*, 2. *once or twice a day*, 3. *several times (more than twice) a day (or always)*. The items are grouped into ten sub-scales of behaviours tapping into different skills (A. Speech, B. Syntax, C. Semantics, D. Coherence, E. Inappropriate initiation, F. Stereotyped language, G. Use of context, H. Non-verbal communication, I. Social relations, J. Interests). Taking the sum of scores for the sub-scales A to H are used to derive the General Communication Composite

(GCC) which is used to identify clinically abnormal communication competence. A GCC score < 55 was found to distinguish well between control and clinical groups, using a criterion of scores in the bottom 10% (Norbury & Bishop, 2005). Another proposed composite is the SIDC (Social-Interaction Deviance Composite) which is calculated by taking the difference in the sum of subscales E, H, I, and J (tapping into pragmatic language and social skills) from the sum of scales of A to D (describing structural language skills). Abnormal GCC (< 55) combined with a negative SIDC score has been shown to be indicative of an autistic spectrum disorder profile (Bishop, 2003). The CCC-2 scaled and composite scores were computed using the test scorer.

3.2.3 Procedure

Testing took place at the SHaPS laboratory (UCL, London) in a sound-attenuated chamber. Unfortunately, since many of the APD children had to travel from outside London and because of difficulties in recruitment, all the testing had to be completed in a single session, lasting in total circa 2.5 to 3 hours (including breaks). To minimise possible fatigue effect, the session was carefully designed to ensure several planned and unplanned breaks The participants were encouraged to request a break between test runs whenever they required and were observed for any signs of fatigue by the examiner. The different tasks were gathered into short blocks and different measures were scattered throughout the session to keep the session fun and engaging for the child. At the end of the session, each child received a certificate and an Amazon voucher as a token of appreciation for taking part in the study. Travel costs of the family were reimbursed.

Participants from both the TD and the APD group completed the same test battery in the below listed order (see Table 3.3). The ECliPS, CCC-2 and the locally compiled background questionnaire were completed by the caregiver during the testing day. The session started with a standard pure-tone audiogram and

Table 3.3: Experimental design and measurements order.

Order	Group A	Group B	Group C	Group D
1	Otoscopy	Otoscopy	Otoscopy	Otoscopy
2	Standard audiometry	Standard audiometry	Standard audiometry	Standard audiometry
3	ST-ASL	ST-ASL	ST-CCRM	ST-CCRM
4	CELF-RS	SSN	CELF-RS	SSN
5	ST-CCRM	ST-CCRM	ST-ASL	ST-ASL
6	SSN	CELF-RS	SSN	CELF-RS
7	EHF audiometry	EHF audiometry	EHF audiometry	EHF audiometry
8	ENVASA	ENVASA	ENVASA	ENVASA
9	LiSNS-UK	LiSNS-UK	LiSNS-UK	LiSNS-UK

otoscopy to ensure that detection thresholds fulfilled the study criteria and that there were no abnormalities in the ear canal and the eardrum. Next, the switching task was conducted. Since performance in the task was one of the main focuses in the study, and because little is known about any possible learning effect in the task, presentation of the two speech materials (ASL and CCRM) was counterbalanced within each group, where about half of the children started with the ASL and the other half with the CCRM speech material. In between the two ST versions, each child completed the CELF-RS and the SSN task, whereby again, the order of presentation was counterbalanced within each group. Since both CELF-RS and SSN test duration are relatively short, they served as a short informal break between the ST test versions and kept the child engaged. Next, about half-way through the session, with a fixed order, all the participants were presented with the EHF audiometry, and the ENVASA task. The session was concluded with the LiSNS-UK, in-line with typical clinical assessment where the test is often presented last.

3.2.4 Data Analysis

All the data extraction, management and analysis in the present study was computed in an R environment (Version 4.0.3; R Core Team, 2020b) using RStudio (Version 1.4.938; RStudio Team, 2019).

Age-scaled scores

Age-independent scores were estimated using a linear regression model. The model was fitted per condition separately for each measure (ST-ASL, ST-CCRM, LiSNS-UK, SSN, & ENVASA) and was based on the control group data only with the respective test raw scores (e.g., SRdT, SRT or %-correct) as a dependent variable and age as a predictor. A two-step model comparison was performed to test the assumption that performance displayed a monotonic linear relationship with age versus a non-monotonic (segmented) linear relationship, implying an asymptote in performance with age (e.g., for a task in which children did not improve after a given age). Extreme outliers were initially trimmed from the TD group to reduce noise in the data and to improve the model fit. In the first step, both models were computed and the best model was selected based on an F-statistic model comparison based on analysis of variance ANOVA, using the *anova()* function. Standardised residuals were next calculated for each TD listener, based on the selected model prediction. Since age was included in the model, the standardised residuals are age-independent and are comparable to z-scores for data with a normal distribution, with a mean and SD of approximately 0 and 1. Since the main goal of the study was to find a measure that is able to well separate between the APD group and the typically developing control group, individual differences and group differences were explored using a deviance analysis procedure proposed by Ramus et al. (2003). Abnormal scores were defined by a two-tailed deviance cut-off of ± 1.96 SD from the TD group mean. Thus, circa 95% of the normal population residuals are expected to be within the deviance range of ± 1.96. Occasional occurrence of abnormal scores in the normal population is not unusual in behavioural measures. Therefore, since the prediction of the residuals is based on the control data, such outliers may skew the TD group true mean or SD and thus may introduce an error in the model prediction. Therefore, in the second step, additional TD outliers (with standardised residuals below/above TD mean ± 1.96) were trimmed from the data and the two models were refitted and compared again. Finally, the model with the best fit was

selected and was used to calculate the standardised residuals for all the listeners, including the trimmed TD observations and the APD group.

Statistical analyses

Residual analysis was performed separately for each measure to determine whether the data fulfils parametric methods assumptions of normal distribution using the Shapiro-Wilk test (*shapiro.test()*, R Core Team, 2020b) and homogeneity of variance using Levene's test (*leveneTest()*; Fox & Weisberg, 2019). Consequently, statistical analyses for factorial design data that met these requirement was performed using linear mixed-effects regression models (LMEMs). LMEMs was fitted using the *lmer()* function (lme4 package; Bates et al., 2015). A Backward model selection procedure was applied to find the model that gives the best fit using a likelihood ratio test (χ^2). Main effects and interaction terms were tested by comparing predictions of the full model to a reduced model where each fixed term was separately removed, starting with the interaction terms. When applicable, post-hoc paired-comparison t-tests were performed on the fitted model and included adjusted least-squared-mean for the random intercepts (subjects) using the *lsmeans()* function from the emmeans R package (Lenth, 2020). In addition, group differences for a single parametric measure such as in the CELF-RS and the CCC-2 total score were examined using a one-way analysis of variance using the *anova()* function. Post-hoc pairwise comparison t-tests with Bonferroni correction were computed using the *pairwise_t_test()* function (rstatix package; Kassambara, 2021).

Nonparametric data were analysed using the *nparLD()* function (nparLD package; Noguchi et al., 2012) which is a robust rank-based method for analysis of skewed data or for data with outliers or from a small sample size (see Feys, 2016, for a good introduction to robust nonparametric techniques). This function enables different types of nonparametric tests for factorial design data with repeated measures with variable between-/within-subjects factors. The results reported in the present study were based on the ANOVA-type statistic test (ATS) output. Inspection

of the ENVASA task age-independent z-scores revealed that the assumption of sphericity (Mauchly's test) was violated. Therefore, analysis was performed using the *npIntFactRep* package (Feys, 2015), which is another robust aligned-rank technique that enables sphericity correction (Greenhouse-Geisser). When applicable, post-hoc pairwise comparisons were computed using a Wilcoxon rank-sum test which is a t-test equivalent for nonparametric data using the *wilcox_test()* function either from the rstatix package (Kassambara, 2021) or the coin package which also enables permutation (Hothorn et al., 2006). Group differences for the ECLiPS total score were examined using a robust one-way ANOVA with trimming means (20%) and bootstrapping (N = 2000) using the *t1waybt()* function from the WRS2 package (Mair & Wilcox, 2020), followed with a corresponding post-hoc tests with the same trimming and bootstrapping using the *mcppb20()* function from the same package.

Perhaps the way I use *wilcox_test()* needs a further examination.

- I can't get *coin::wilcox_test()* function to run for groups with 3 levels, so used *rstatix::wilcox_test()* instead.
- On the other hand, only *coin::wilcox_test()* worked for 2-way interaction.

3.3 Results

3.3.1 Standard audiometry

The listeners' detection thresholds for the left and the right ear are plotted in Figure 3.2. The shaded grey area represents the TD group's range of thresholds and the white line represents the group mean at each frequency. The black lines mark the individual thresholds in the APD group and the group mean is marked by the bold black line. The dashed line indicates the maximal threshold criterion of $\leq$ 25 dB HL for participation in the study.

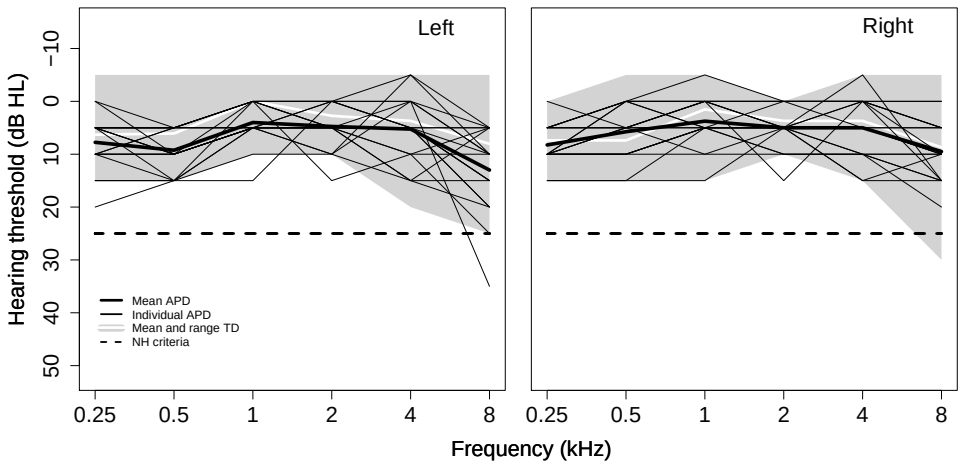

Figure 3.2: Standard audiometry: APD participants pure-tone detection thresholds plotted seperately for the left and the right ear (black lines). The shaded grey area represents the TD group thresholds range and the white line represents the TD group mean at each frequency. The dashed line represents the threshold criterion of hearing level $\leq$ 25 dB HL.

Boxplots of listeners' pure-tone detection thresholds measured at six frequencies between 0.25 to 8 kHz and their corresponding pure-tone-average (PTA) are shown in Figure 3.3 A-B. Individual PTAs were calculated by averaging thresholds at the frequencies 0.5, 1, 2 and 4 kHz separately for the left and right ear (PTA_{Left}, PTA_{Right}) and by taking the grand mean for thresholds in both ears (denoted as PTA), whereas the listeners' PTA at the better-ear is denoted as BE. Threshold descriptives by frequency and ear split by the two groups are given in Table 3.4, as well as Table **??** for PTAs and BE with additional statistics.

Differences between groups (APD, and TD children with/without an APD sibling) for detection thresholds across frequencies and ears were statistically tested with a three-way 6 x 2 x 3 factorial design with repeated measures. Inspection of the data for linear model residuals revealed that the assumption of normality and homoscedasticity were violated. Therefore, a non-parametric approach was adopted, using a rank-based ANOVA-type statistic test (ATS) with the *nparLD()* function (nparLD package; Noguchi et al., 2012). The ATS test results are given in Table 3.5. There was no significant three-way or two-way interaction between the three predictors, nor a significant main effect of Ear or Group (all p's > 0.05),

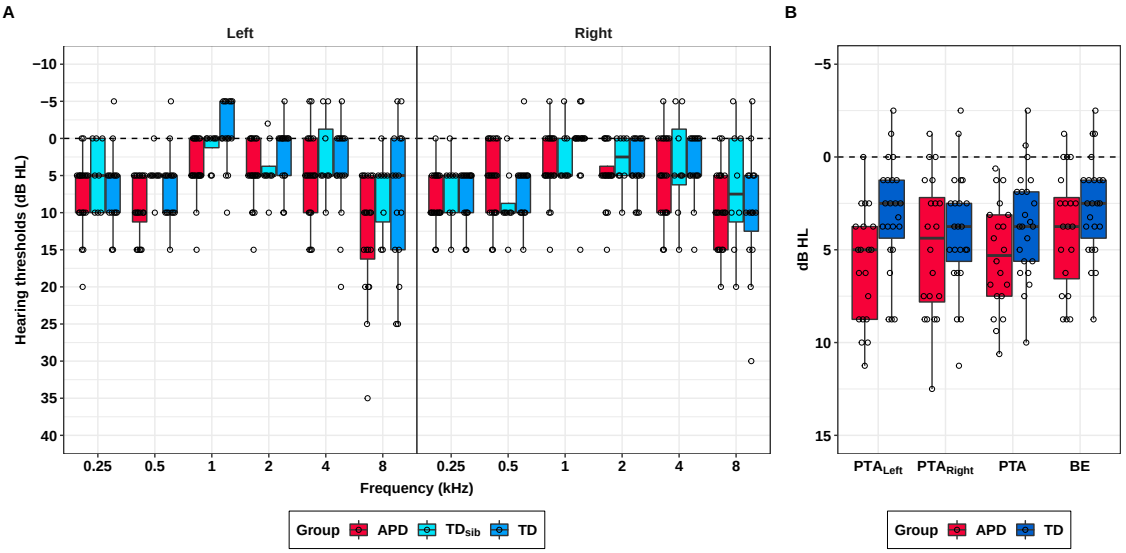

Figure 3.3: Standard audiometry: Pure-tone detection thresholds by frequency between 0.25 to 8 kHz (A), and averaged thresholds (B). Individual scores are indicated by circles. The boxes show the data interquartile range (25th-75th percentile) and the horizontal line indicate the median (i.e., 50th percentile). Values that fall within 1.5 times the interquartile range are indicated by the whiskers.

Table 3.4: Standard audiometry: Descriptives for pure-tone detection thresholds (dB HL) by frequency (kHz) and ear split by the two groups.

		APD					TD				
	Ear	N	median	sd	min	max	N	median	sd	min	max
Frequency											
0.25	L	20	5.00	4.99	0.00	20.00	23	5.00	5.05	-5.0	15.00
0.5	L	20	10.00	4.06	5.00	15.00	23	5.00	4.25	-5.0	15.00
1	L	20	5.00	3.84	0.00	15.00	23	0.00	3.99	-5.0	10.00
2	L	20	5.00	4.13	0.00	15.00	23	0.00	4.03	-5.0	10.00
4	L	20	5.00	5.95	-5.00	15.00	23	5.00	6.07	-5.0	20.00
8	L	20	10.00	8.01	5.00	35.00	23	5.00	8.49	-5.0	25.00
0.25	R	20	10.00	3.73	0.00	15.00	23	5.00	3.95	0.0	15.00
0.5	R	20	5.00	4.94	0.00	15.00	23	10.00	4.49	-5.0	15.00
1	R	20	5.00	4.55	-5.00	15.00	23	0.00	4.38	-5.0	15.00
2	R	20	5.00	3.97	0.00	15.00	23	5.00	3.76	0.0	10.00
4	R	20	5.00	5.62	-5.00	15.00	23	5.00	5.27	-5.0	15.00
8	R	20	10.00	5.36	0.00	20.00	23	10.00	8.29	-5.0	30.00
PTAs and better-ear											
PTA_{Left}	L	20	5.00	3.04	0.00	11.25	23	2.50	3.01	-2.5	8.75
PTA_{Right}	R	20	4.38	3.78	-1.25	12.50	23	3.75	3.16	-2.5	11.25
PTA		20	5.31	2.92	0.62	10.62	23	3.75	2.87	-2.5	10.00
BE		20	3.75	3.17	-1.25	8.75	23	2.50	2.68	-2.5	8.75

PTA: average detection threshold at 0.5, 1, 2, & 4 kHz.
BE: PTA at the better ear.

Table 3.5: Standard audiometry: Statistical analysis for the effects of Frequency (0.25 - 8 kHz), Ear (left/right) and Group (APD, and TD with/without an APD sibling) and their interaction (6x2x3 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f1-ld-f2 design ANOVA-type statistic (ATS) test, whereby f1 refers to an experimental design with a single between-subjects factor (Group) and f2 refers to two within-subjects factors (Frequency and Ear).

	Statistic	df	p-value
Group	2.126	1.836	*0.124*
Frequency	18.505	2.861	***< 0.001***
Ear	0.855	1.000	*0.355*
Group:Frequency	0.555	3.900	*0.691*
Ear:Frequency	0.400	3.767	*0.798*
Group:Ear	1.747	1.759	*0.179*
Group:Frequency:Ear	1.659	5.855	*0.128*

* significant p-values ($p < 0.05$) are shown in bold.

whereas there was a highly significant main effect of Frequency ($p < 0.001$).

Group differences for PTAs measured in the left and the right ear were examined using a 2 x 2 LMEM model (parametric model assumptions were met). Ear (Left/Right) and Group (APD/TD) were set as fixed factors (reference levels: Ear = Left; Group = APD) and PTA (in dB HL) as dependent variable, as well as random intercepts for subjects. Note that the TD children were treated as a single group, since there was no significant difference in thresholds across the TD children with or without an APD sibling. A model with an interaction term was found to give the best fit, showing a significant interaction between the tested ear and group [$\chi^2(1) = 4.32$, $p = 0.038$]. Post-hoc paired-comparison t-tests based on the fitted model were computed using the *lsmeans()* function [*emmeans* package, Lenth (2020); see Table **??**] which revealed a significant difference between the groups for PTA measured in the left ear ($p = 0.01$). However, a group difference of 2.5 dB is rather small and clinically negligible, and is likely to occur due to sampling error. No significant difference was found between the two groups for PTA measured in the right ear ($p > 0.05$). Therefore, the listeners average PTA across the two ears (PTA) was used for later analysis. Separate models for PTA across and BE using

Table 3.6: Standard audiometry: Post-hoc paired-comparison t-tests for Group x Ear interaction. The test was performed on the fitted LMEM model and included adjusted least-squared-mean for the random intercepts (subjects) using lsmeans package (emmeans package; Lenth, 2020).

Ear	Contrast	Estimate	SE	Df	t-value	p-value	CI	d	magnitude
PTA_{Left}	APD - TD	2.68	0.99	67.17	2.7	***0.01***	0.7 - 4.67	0.89	large
PTA_{Right}	APD - TD	0.80	0.99	67.17	0.8	*0.42*	-1.18 - 2.78	0.23	small

* significant p-values ($p < 0.05$) are shown in bold.
PTA: average detection threshold (dB HL) at 0.5, 1, 2, & 4 kHz.
BE: PTA at the better ear.

one-way ANOVA tests found no significant difference between the two groups [PTA: $F(1,41) = 3.867$, $p = 0.056$; BE: $F(1,41) = 1.938$, $p = 0.171$].

3.3.2 EHF audiometry

The listeners pure-tone detection thresholds measured at the frequencies of 8, 11 and 16 kHz are plotted in Figure 3.4 separately for the left and the right ear. In many cases it was not possible to recorded a response for thresholds measured at 20 kHz, resulting in a large portion of missing data points in both groups. This was because the maximal presentation level of the equipment was reached for thresholds above circa 10 dB HL at 20 kHz. Therefore, thresholds measured at 20 kHz were not included in the analysis. A comparison of the group means reveals relatively small differences in thresholds between the groups, with a relatively larger difference in the left ear, where APD thresholds at 11 and 16 kHz were on average 5 dB higher (i.e., poorer). Boxplots of the listeners thresholds by frequency and ear as well as their calculated PTAs and BE are shown in Figure 3.5 A-B. Descriptives of the groups' detection thresholds are given in Table 3.7.

Difference in thresholds across group (APD, and TD children with/without an APD sibling), frequencies (8, 11, & 16 kHz) and ears (left/right) were examined for a 3 x 2 x 3 repeated measures factorial design. Inspection of parametric model assumptions revealed that the assumptions of normality and homoscedasticity

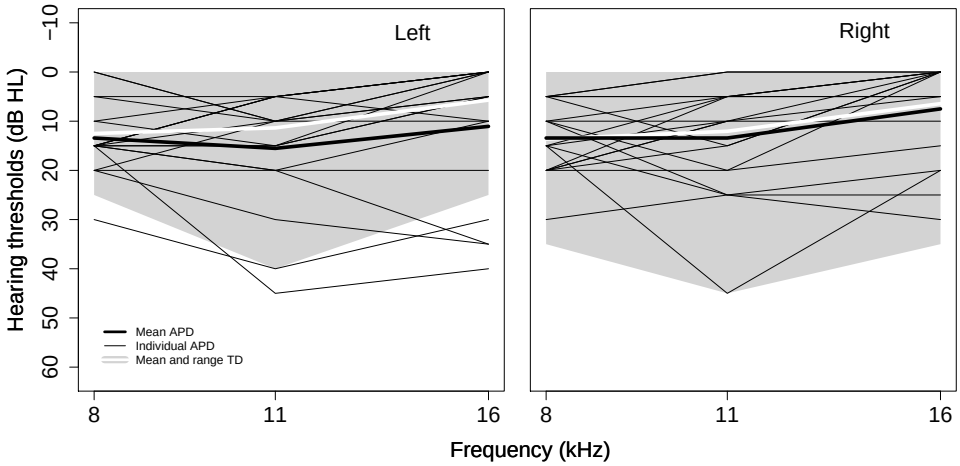

Figure 3.4: EHF audiometry: Pure-tone detection thresholds for the extended high-frequencies measured in the left and the right ear. The thin black lines represent the individual thresholds in the APD group and the group mean is marked by the bold black line. The shaded grey area represents the TD group threshold range and the white line represents the TD group mean at each frequency.

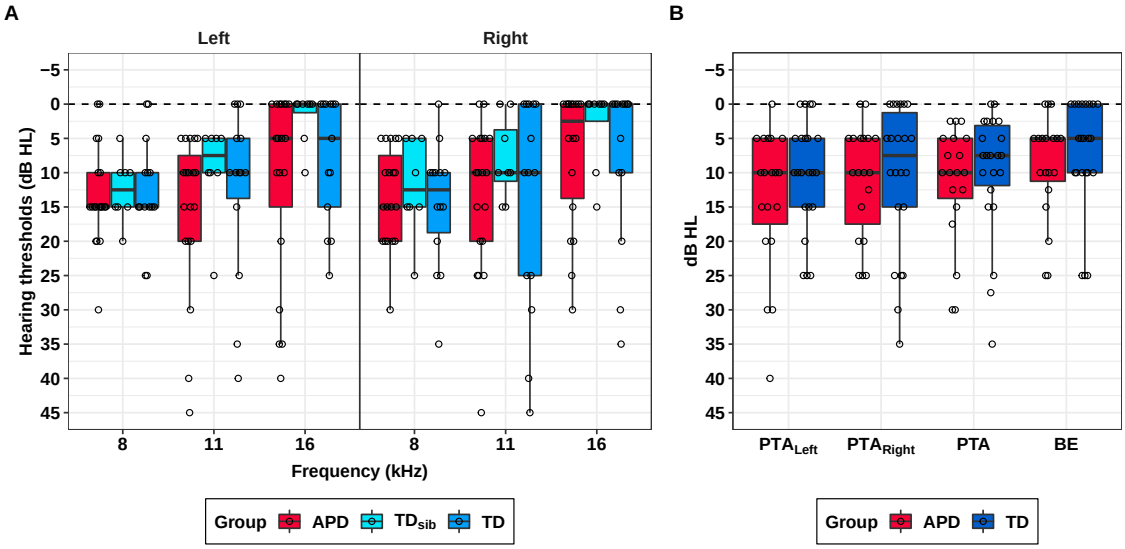

Figure 3.5: EHF audiometry: Boxplots for pure-tone detection thresholds measured at the extended high-frequencies split by ear and groups (A). Boxplots of the groups averaged PTAs and better-ear BE thresholds are depicted in figure B. Individual scores are indicated by circles.

Table 3.7: EHF audiometry: Descriptive for pure-tone detection thresholds (dB HL) by extended-high frequencies (kHz) split by ear and group.

		APD					TD				
	Ear	N	median	sd	min	max	N	median	sd	min	max
Frequency											
8	L	19	15.0	7.27	0.0	30	22	15.0	6.50	0	25
11	L	19	10.0	11.65	5.0	45	22	10.0	10.71	0	40
16	L	19	5.0	13.80	0.0	40	22	0.0	8.11	0	25
8	R	19	15.0	7.08	5.0	30	22	12.5	8.34	0	35
11	R	19	10.0	11.19	0.0	45	22	10.0	13.24	0	45
16	R	19	2.5	10.04	0.0	30	22	0.0	10.51	0	35
PTAs and better-ear											
PTA_{Left}	L	19	10.0	10.39	0.0	40	22	10.0	8.09	0	25
PTA_{Right}	R	19	10.0	8.27	0.0	25	22	7.5	10.83	0	35
PTA		19	10.0	8.59	2.5	30	22	7.5	9.05	0	35
BE		19	5.0	7.60	0.0	25	22	5.0	8.27	0	25

PTA: average detection threshold at 8, 11, & 16 kHz.
BE: PTA at the better ear.

were violated. Therefore, the same nonparametric procedure as used for standard audiometry was performed using nparLD package. The ATS ANOVA-type test (given in Table 3.8) found no significant three-way nor two way interaction between the different predictors. There was however a highly significant difference in thresholds between the three frequency bands ($p < 0.001$), whereas no significant main effect for Group or Ear was found.

As in the standard audiometry, group differences for PTAs (PTA_{Left}, PTA_{Right}, and PTA) and BE were examined using multiple comparisons tests. As before, the TD group was treated as a single group since no significant difference was found between TD children with or without an APD sibling. Parametric model assumption of normal distribution was rejected (Shapiro-Wilk test; $p < 0.05$), while the assumption of homoscedasticity was met. Therefore, Wilcoxon rank-sum test was used with Bonferroni correction (see Table 3.9). There was no significant difference between the groups for any of the PTAs (all p's > 0.05).

Table 3.8: EHF audiometry: statistical analysis for the effects of Frequency (8, 11, & 16 kHz), Ear (left/right) and Group (APD, and TD with/without an APD sibling) as well as their interaction (3x2x3 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f1-ld-f2 design ANOVA-type statistic (ATS) test, whereby f1 refers to an experimental design with a single between-subjects factor (Group) and f2 refers to two within-subjects factors (Frequency and Ear).

	Statistic	df	p-value
Group	1.124	1.911	*0.323*
Frequency	29.793	1.992	***< 0.001***
Ear	0.226	1.000	*0.635*
Group:Frequency	1.924	3.564	*0.112*
Ear:Frequency	0.150	1.940	*0.855*
Group:Ear	0.167	1.998	*0.846*
Group:Frequency:Ear	0.716	3.638	*0.568*

* significant p-values ($p < 0.05$) are shown in bold.

Table 3.9: EHF audiometry: Paired-comparison t-tests for PTA x Group. Bonferroni correction was applied for multiple comparisons

	Estimate	statistic	p-value	95%-CI	d	magnitude
PTA_{Left}	5.0	247.0	*1*	-5 - 10	0.37	small
PTA_{Right}	0.0	229.5	*1*	-5 - 5	0.05	negligible
PTA	2.5	243.5	*1*	-2.5 - 5	0.19	negligible
BE	0.0	247.5	*1*	0 - 5	0.23	small

* significant p-values ($p < 0.05$) are shown in bold.
PTA: average detection threshold (dB HL) at 8, 11, & 16 kHz.
BE: PTA at the better ear.

3.3.3 ST

Outliers & missing data

As a first step, the listeners adaptive tracks and psychometric functions were manually inspected for abnormalities. Since the SRdT is limited to a DC of 0.97, the adaptive procedure may not be able to present trials that are easy enough for performance levels to reach 50%-correct of key words in sentences. This could have potentially occurred in the more challenging test conditions with speech distractors. Thus, the proportion of correct keywords within the final test trials (LevsPC) was

calculated as a measure describing the success of the adaptive procedure, whereby a successful procedure is expected to have a LevsPC at approximately 50%. A binomial statistical test was applied to identify observations that significantly differ from 50%. Observations with LevsPC $\leq$ 35% were labelled as possible outliers and were further inspected (see Figure 3.6). Interestingly, most of the outliers belonged to the CCRM material with 29 observations from a total of 258 (6 conditions x 43 listeners), whereas only 3 observations out of a total of 215 (5 conditions x 43 listeners) were labelled as outliers for data measured with the ASL speech material.

As expected, most of the identified cases in both materials were for observations measured with the more demanding conditions with speech distractors. In five cases (2 ASL; 3 CCRM) we were able to confidently determine that the listener's true score was near to ceiling, and thus these observations were set to the maximal DC in the task (0.97). In other cases it was not possible to confidently determine the true SRdT, either because the procedure ended after reaching the maximum number of trials before a minimum number of test reversals was obtained (x1 CCRM, x2 ASL), or due to aberrant adaptive tracks (x5 CCRM). Since all these cases belonged to more challenging test conditions with speech distractors, it is very likely that the children's true score is at or beyond the upper DC limit. Thus, to account for that, rather than removing these observations, which will consequently reduce the statistical power and may not represent the true performance in the group, they were set to a DC of 1, which is above the task's upper DC limit of 0.97.

SRdTs by age

Since the present study sample comprised young children of different ages from circa 7 to 13 years, a developmental age effect was expected, whereby performance was expected to improve with increasing age. This is illustrated by the scatterplots and linear regression lines plotted in Figure 3.7 A-B split by groups for the listeners' SRdTs obtained across the different test conditions and speech material (ASL /

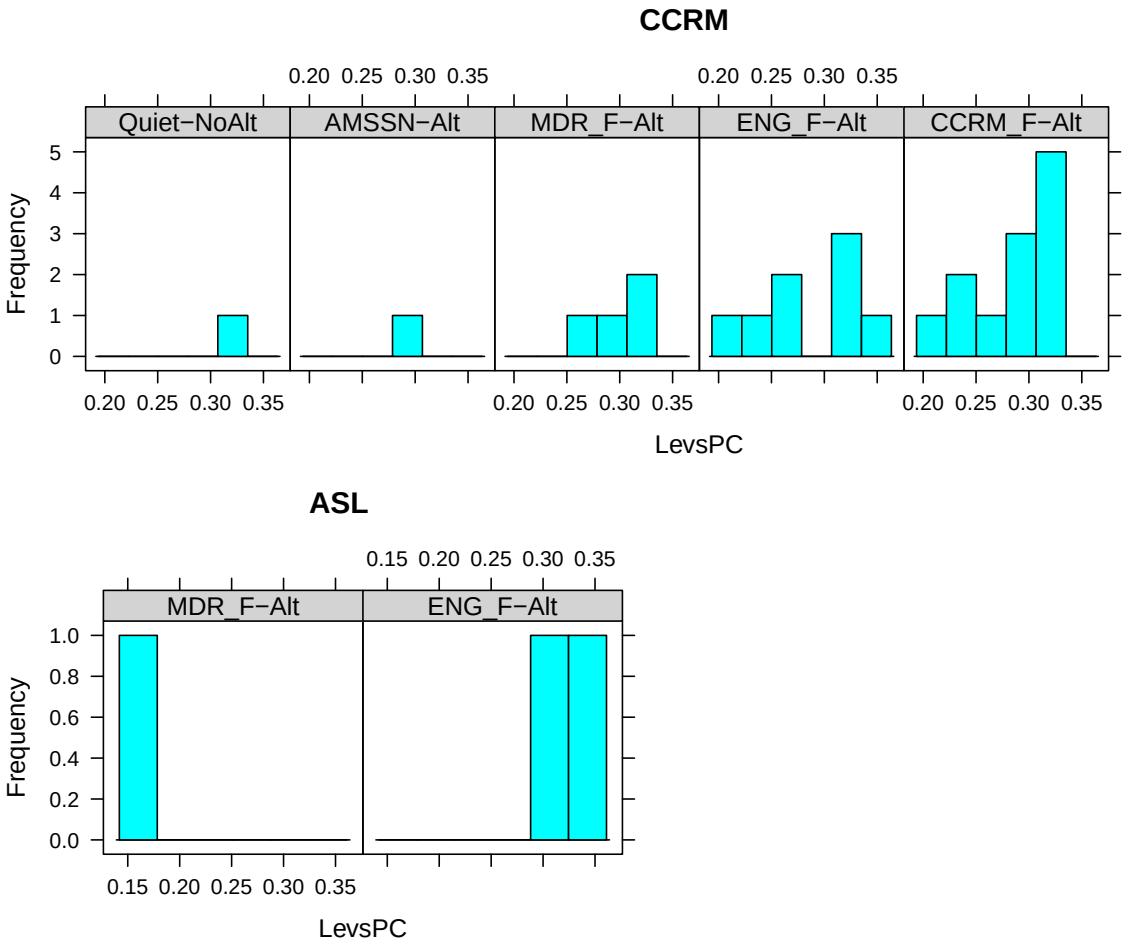

Figure 3.6: ST raw data: Frequency of potential outliers with LevsPC $\leq$ 35%. LevsPC denotes the proportion of correct keywords within the final test trials.

CCRM) as a function of age. Note that smaller SRdTs indicate better performance. The age effect was tested against the TD group alone because this group is more homogeneous and thus expected do display smaller variability than the APD group. Also, developmental changes may well be different in the APD group. Nonetheless, despite the larger spread in the APD group, this group showed a similar trend in performance, albeit shifted towards higher SRdTs (i.e., poorer performance). The TD regression lines were determined based on a model comparison and outlier trimming procedure to improve model prediction (described in Section 3.2.4). Simple regression lines were found to be the most suitable in describing the relationship between the TD children's performance and age in all test conditions but the MDR_F condition for the ASL material, where a segmented line was found to give the best fit. The MDR_F segmented line indicated that DC improved with age by

circa 0.1 per year until reaching a plateau at the age of 9.5 years.

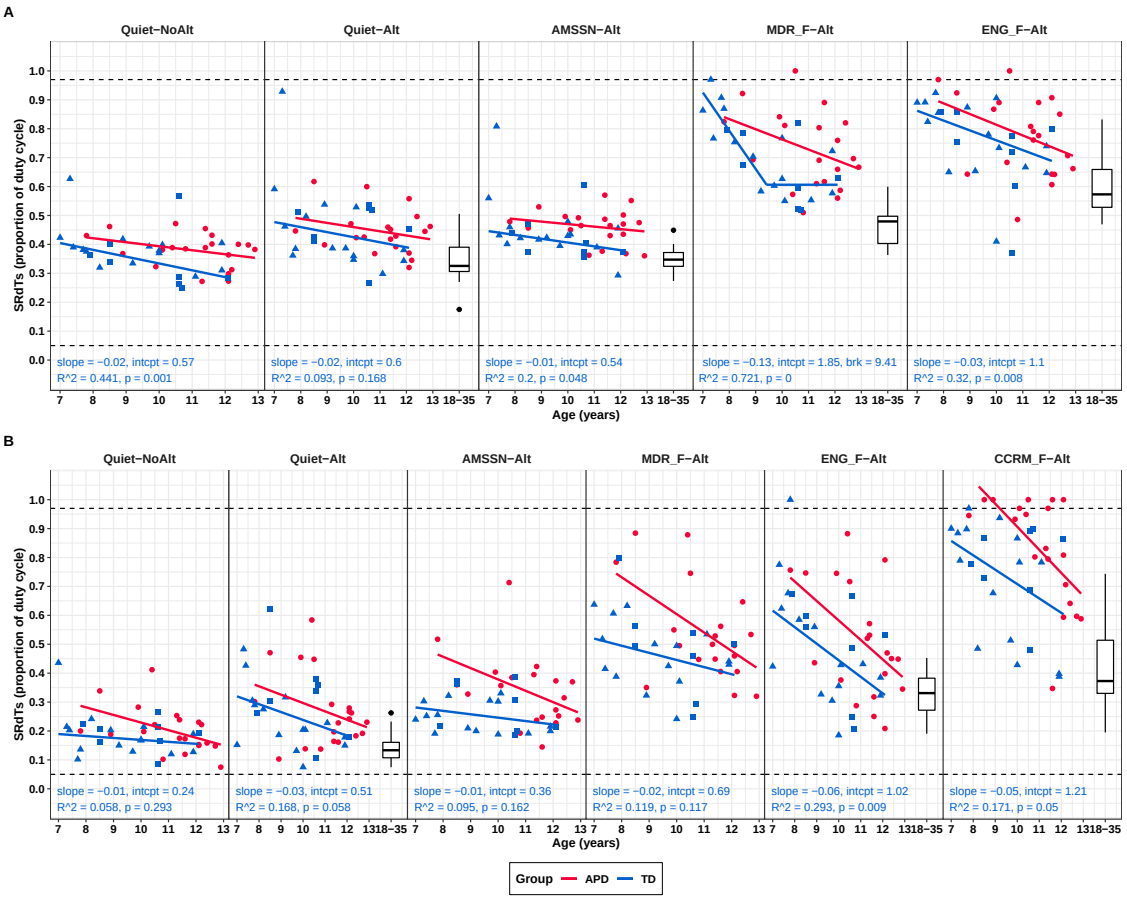

Figure 3.7: ST: Scatterplot and linear regression lines for the listeners SRdTs measured with the ASL (A) and the CCRM speech material (B) as a function of age. Corresponding regression coefficients and statistics are provided for the TD group only. Red indicates data from the APD group and cyan indicates data from the TD control group (square shapes: TD children with an APD sibling; triangle shapes: the remaining TD children). Data for normal hearing adults taken from Chapter 2 is shown in the boxplots as a reference. The dashed lines represents the task lower and upper DC limit of 0.05 and 0.97, respectively.

Looking at Figure 3.7 A-B, it is noticeable that children in both groups showed a larger decrement in performance when presented with speech distractors. The regression lines indicate that the improvement in performance by age was more prominent for speech distractors, with relatively steeper slopes (at least twice as steep) than for the non-speech distractor (AMSSN) or for conditions without a distractor. Furthermore, as expected, CCRM sentences were more intelligible, with

performance shifted towards lower DC values relative to performance for the ASL speech material. The lower DC meant that the children were able to understand 50% of the sentences with larger portions of the speech information missing.

A closer look at the regression lines shows several interesting trends. The non-speech AMSSN distractor had little-to-no effect on performance, at least in the TD group, where performance was fairly similar to performance in the Quiet conditions. Introducing alternations (as in Quiet-Alt vs. Quiet-NoAlt), seems to hinder intelligibility in both groups. However the effect is relatively small and may not be significant due to the large spread in the APD group. Furthermore, when comparing the regression lines, there appears to be a relatively larger separation between the groups for SRdTs measured with the CCRM material, especially for AMSSN, but also for the speech distractors. However, it is possible that the APD regression lines do not reflect the true population due to the large spread in performance and the small sample size and thus any interpretation should be taken with caution. Another interesting observation is that the children showed little-to-no *masking-release* for speech spoken in an unfamiliar language (MDR_F) when compared with a distractor spoken in English (ENG_F). This is in agreement with findings in the adults' study in Chapter 2. Lastly, it is apparent from the figure that performance for the CCRM_F distractor was near-to-ceiling for some children, mostly among the APD group.

Next, the age effect was tested using an LMEM model, with Condition (Quiet-NoAlt, Quiet-Alt, AMSSN, MDR_F, & ENG_F), Material (ASL / CCRM), Age, and Sibling (TD children with/without an APD sibling, TD/TD_{sib}) as fixed factors, SRdT as the dependent variable and random intercepts for subjects (reference levels: Condition = Quiet-NoAlt; Material = ASL, Sibling = TD). Note that data for CCRM_F was excluded from the model since it was only measured for the CCRM

Table 3.10: ST: Age effect analysis using LMEM for SRdTs measured across condition, speech material, age and children from the TD group with/without an APD sibling (Sibling: TD/TD$_{sib}$) as fixed factors and random intercepts for subjects. Reference levels: Condition = Quiet-NoAlt, Material = ASL, Sibling = TD. Note: only data for the control group (TD) following outlier trimming was included.

SRdT ~ Condition + Material + Age + Sibling + Condition:Material + Condition:Age + Material:Age + Condition:Sibling + Material:Sibling + (1 | Subjects)			
Effects	Df	χ^2	p
Condition:Material	4	10.073	***0.039***
Condition:Age	4	15.948	***0.003***
Material:Age	1	2.073	*0.150*
Condition:Sibling	4	2.724	*0.605*
Material:Sibling	1	4.927	***0.026***

* significant p-values ($p < 0.05$) are shown in bold.

material[3]. The final LMEM model that gave the best fit and main effects are given in Table 3.10. Inspection of parametric assumptions based on the model's residuals confirmed that both the assumption of normal distribution and homogeneity of variance were met. Model comparison revealed a significant two-way interaction between Condition x Age, between Condition x Material, and between Material x Sibling (all p's < 0.05).

The significant Condition x Age interaction supports the observation in Figure 3.7 A-B, that the effect of age was different across the test conditions. These findings raises the following questions – do all the conditions show a significant age effect? Moreover, since the effect of age is not the same across the test conditions, which conditions showed the largest age effect? One possible way to tackle these questions is to compare the separate regression models using F-statistics. Nonetheless, due to the small sample-size and the large number of paired comparisons, such test lacks a statistical power and the results may not reflect the true effect in a larger sample. The TD group regression models R^2 and p-values are

[3]A separate model for the CCRM data with CCRM_F-Alt condition showed similar results, with a strong significant Condition x Age interaction ($p < 0.001$) and no main effect of Sibling ($p > 0.05$).

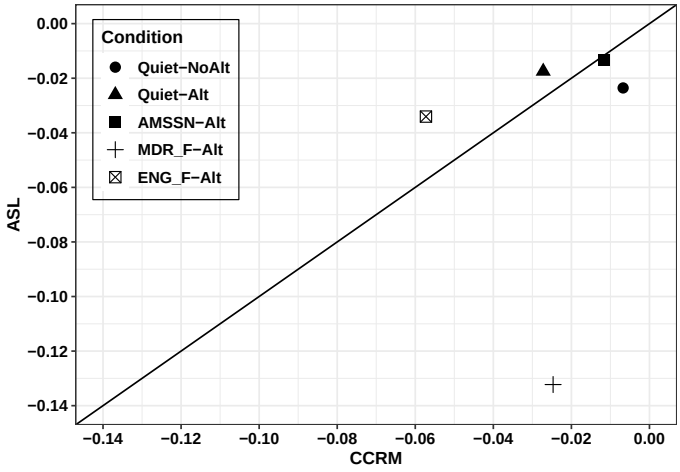

Figure 3.8: ST: Age effect: a comparison beteween the regression line slopes fitted for the CCRM (x-axis) and ASL speech material (y-axis). Test conditions are represented by the different symbols. The diagonal line represents an optimal agreement between the speech materials. Observations falling below the line indicate a steeper slope for the ASL material than for the CCRM material.

given at the bottom part of Figure A and B. The ASL models p-values indicated a highly significant age effect for ENG_F, MDR_F and Quiet-NoAlt condition as well as a marginal effect for AMSSN ($p = 0.048$), whereas no significant age effect was found for Quiet-Alt ($p = 0.168$). As for the CCRM material, there was a highly significant age effect for ENG_F and a marginal effect for the Quiet-Alt condition ($p = 0.058$) and for CCRM_F condition ($p = 0.05$) which was not included in the LMEM model, whilst there was no significant age effect found for Quiet-NoAlt, AMSSN and MDR_F conditions. Furthermore, age was found to be a better predictor (i.e., accounting for larger variance in SRdT) for conditions with speech distractors, with R^2 ranging between 32% to 72% for the ASL material and about 12% to 29% for the CCRM. A comparison between the test conditions regression line slopes split by test material is depicted in Figure 3.8. A possible pattern emerges from the figure, where slopes for the quiet and non-speech conditions are fairly similar across the two speech material (indicated by their proximity to the diagonal line), while, differences between the slopes are relatively larger for speech distractors, in particularly for MDR_F where the slope for the ASL material (-0.13) is about six times steeper than the slope for the CCRM material (-0.02).

Furthermore, the significant Condition x Material interaction implies that the difference between the two speech materials depends upon condition. A model-based post-hoc pairwise t-tests comparison of the test conditions by material are given in Table 3.11. The tests revealed a similar trend in both speech materials, but with a larger effect size in some conditions than others. Thus, suggesting there is no crossover interaction. Differences in means across the conditions were all significant, excluding Quiet-Alt - AMSSN-Alt, and MDR_F - ENG_F, which were not significant. The insignificant difference in performance between the speech distractors MDR_F and ENG_F suggests that the TD children did not benefit from a release from masking for a speech distractor spoken in an unfamiliar language (MDR_F) as opposed to a familiar speech spoken in English (ENG_F). This sits well with our previous findings with adults where adults showed no benefit for the MDR_F speech masker (see Chapter 1).

Lastly, the significant interaction between Material and Sibling was examined with a model-based post-hoc t-tests comparison (see Table 3.12). There was no significant difference in SRdTs between the two TD groups for both speech materials (both p's > 0.05). Comparison of the groups overall performance by material revealed an opposite direction of performance (see also Figure 3.9). While the performance of the TD_{sib} children was on average 0.06 better than their TD peers for the ASL material, their performance was 0.04 poorer for the CCRM material. These results are in contradiction to our expectations if any differences arose. In such a case, we would have predicted a larger decrement in performance (i.e., poorer score) for the ASL sentences which are more linguistically challenging than for the CCRM sentences. This disagreement and the very small estimated mean difference for the interaction between Material and Group in the full model (0.052) suggests that the differences picked up by the model are due to sampling error. The results suggests that the differences between the two groups depends upon the material type. This is not surprising considering the large differences between the two material types,

Table 3.11: ST: Age-effect: post-hoc paired-comparison t-tests for Condition x Material interaction. The test was performed on the fitted LMEM model and included adjusted least-squared-mean for the random intercepts (subjects) using lsmeans package (emmeans package; Lenth, 2020). P-values were adjusted for multiple-comparisons using Bonferroni correction.

Contrast	Estimate	SE	Df	t-value	p-value	95%-CI
ASL						
Q-NoAlt - Q-Alt	-0.09	0.03	221.97	-3.49	***0.01***	-0.16 - -0.02
Q-NoAlt - AMSSN-Alt	-0.07	0.03	221.94	-2.90	***0.04***	-0.15 - 0
Q-NoAlt - MDR_F-Alt	-0.33	0.03	221.97	-13.00	***< 0.001***	-0.4 - -0.26
Q-NoAlt - ENG_F-Alt	-0.38	0.03	221.96	-14.97	***< 0.001***	-0.45 - -0.31
Q-Alt - AMSSN-Alt	0.02	0.03	221.97	0.59	*1*	-0.06 - 0.09
Q-Alt - MDR_F-Alt	-0.24	0.03	221.97	-9.51	***< 0.001***	-0.31 - -0.17
Q-Alt - ENG_F-Alt	-0.29	0.03	221.96	-11.48	***< 0.001***	-0.36 - -0.22
AMSSN-Alt - MDR_F-Alt	-0.26	0.03	221.97	-10.10	***< 0.001***	-0.33 - -0.18
AMSSN-Alt - ENG_F-Alt	-0.31	0.03	221.95	-12.07	***< 0.001***	-0.38 - -0.23
MDR_F-Alt - ENG_F-Alt	-0.05	0.03	221.96	-1.96	*0.51*	-0.12 - 0.02
CCRM						
Q-NoAlt - Q-Alt	-0.08	0.03	222.77	-3.13	***0.02***	-0.16 - -0.01
Q-NoAlt - AMSSN-Alt	-0.08	0.03	222.09	-3.14	***0.02***	-0.16 - -0.01
Q-NoAlt - MDR_F-Alt	-0.28	0.03	222.79	-10.57	***< 0.001***	-0.36 - -0.21
Q-NoAlt - ENG_F-Alt	-0.30	0.03	222.60	-11.35	***< 0.001***	-0.37 - -0.22
Q-Alt - AMSSN-Alt	0.00	0.03	222.60	0.00	*1*	-0.07 - 0.07
Q-Alt - MDR_F-Alt	-0.20	0.03	222.60	-7.51	***< 0.001***	-0.27 - -0.12
Q-Alt - ENG_F-Alt	-0.22	0.03	222.38	-8.27	***< 0.001***	-0.29 - -0.14
AMSSN-Alt - MDR_F-Alt	-0.20	0.03	222.60	-7.52	***< 0.001***	-0.27 - -0.12
AMSSN-Alt - ENG_F-Alt	-0.22	0.03	222.40	-8.28	***< 0.001***	-0.29 - -0.14
MDR_F-Alt - ENG_F-Alt	-0.02	0.03	222.39	-0.68	*1*	-0.09 - 0.06

* significant p-values ($p < 0.05$) are shown in bold.

Table 3.12: ST: Age-effect: post-hoc paired-comparison t-tests for Material (ASL/CCRM) x Sibling (TD/TD_{sib}) interaction. The test was performed on the fitted LMEM model and included adjusted least-squared-mean for the random intercepts (subjects) using lsmeans package (emmeans package; Lenth, 2020).

contrast	material	estimate	SE	df	t.ratio	p.value	95%-CI	d	magnitude
TD - TD_{sib}	ASL	0.01	0.03	35.16	*0.21*	0.84	-0.06 - 0.07	0.18	negligible
TD - TD_{sib}	CCRM	-0.04	0.03	36.93	*-1.36*	0.18	-0.11 - 0.02	-0.13	negligible

* significant p-values ($p < 0.05$) are shown in bold.

where the CCRM sentences are expected to be more intelligible.

Where should it best go?

Simple correlation for the listeners SRdTs between conditions is given in

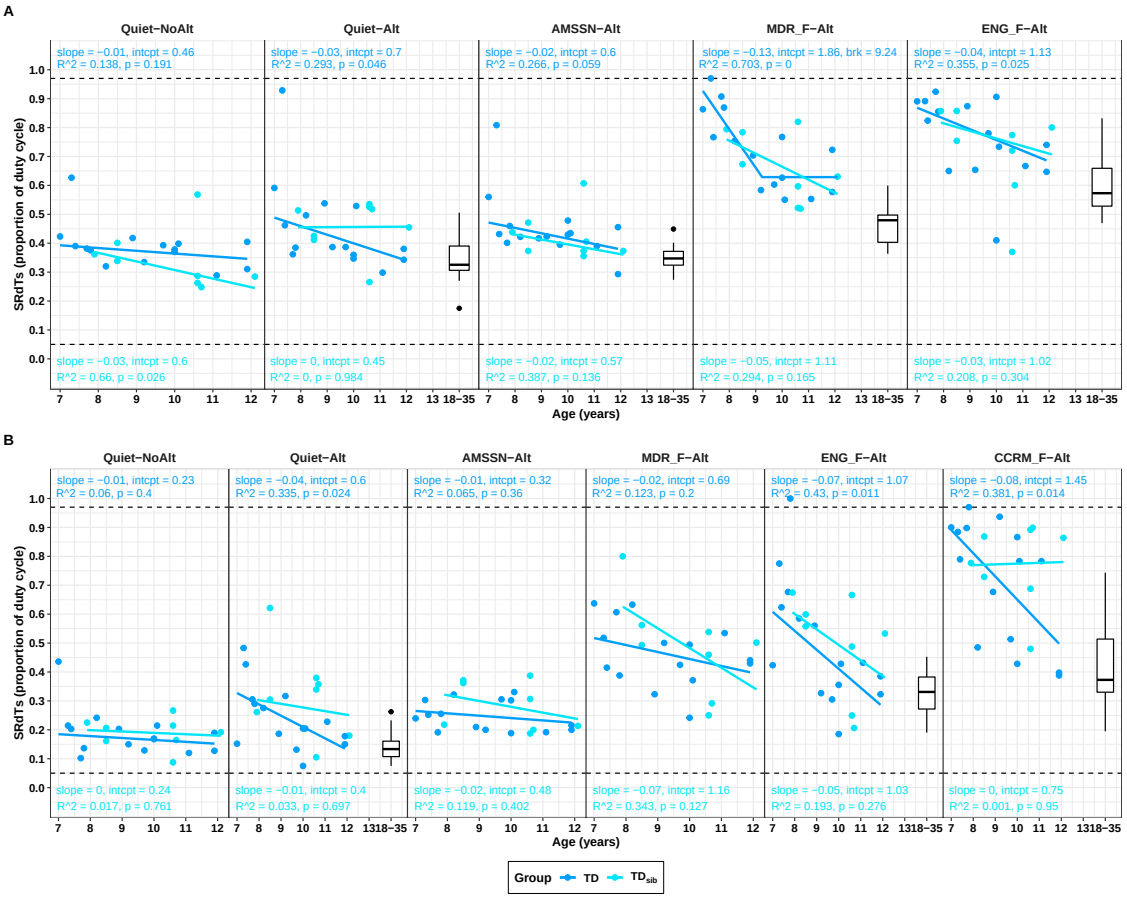

Figure 3.9: ST: sibling and age effect: Scatterplot and linear regression lines for the TD listeners SRdTs measured with the ASL (A) and the CCRM speech material (B) as a function of age.

appendices, separately for the ASL (Figure C.1) and CCRM material (Figure C.2).

Age-independent z-scores

Age-independent standardised residuals (z-scores) were calculated based on a model prediction for the TD group data using a multiple-case study approach [Ramus et al. (2003); or see section 3.2.4 for more details]. Descriptive statistics for the listeners' z-scores are given in Table 3.13. Additional boxplots are shown in Figure 3.10 A-B, for the ASL and CCRM speech material respectively. Scores were calculated separately for each test condition, with better performance indicated by lower z-scores. The grey area marks scores in the 'normal' region (| score | $\pm$ 1.96), where about 95% of the normal population is expected to lay within. Overall, APD

Table 3.13: ST: Descriptives for standardised residuals (z-scores) calculated for data measured with the ASL and CCRM speech material.

	APD						TD					
	N	median	sd	min	max	abnormal	N	median	sd	min	max	abnormal
ASL												
Quiet-NoAlt	20	1.81	1.39	-0.76	3.74	45.00%	23	0.00	1.96	-1.69	6.05	13.04%
Quiet-Alt	20	0.29	0.87	-0.79	2.12	10.00%	23	-0.13	1.46	-1.72	5.27	4.35%
AMSSN-Alt	20	1.79	1.45	-0.82	4.50	50.00%	23	0.10	2.35	-2.18	9.04	13.04%
MDR_F-Alt	20	0.99	1.75	-1.31	5.37	40.00%	23	-0.13	1.11	-1.44	2.91	8.70%
ENG_F-Alt	20	0.90	1.53	-2.96	3.09	20.00%	23	0.12	1.55	-4.44	1.75	0.00%
CCRM												
Quiet-NoAlt	20	0.36	1.75	-1.73	5.70	20.00%	23	0.38	1.57	-1.92	5.72	8.70%
Quiet-Alt	20	0.47	1.23	-1.66	3.58	15.00%	23	-0.08	1.19	-1.68	3.44	4.35%
AMSSN-Alt	20	1.62	2.03	-1.39	7.95	40.00%	23	-0.28	1.09	-1.38	2.50	4.35%
MDR_F-Alt	20	0.86	1.40	-1.12	4.06	25.00%	23	0.28	1.11	-1.87	2.77	4.35%
ENG_F-Alt	20	1.05	1.22	-0.80	3.25	20.00%	23	0.26	1.14	-1.80	2.99	4.35%
CCRM_F-Alt	20	1.11	0.89	-1.59	2.24	10.00%	23	0.24	0.98	-1.76	1.47	0.00%

abnormal: defined as the percentage of abnormal z-score > 1.96.

children's performances were noticeably poorer for both test materials, with higher median z-scores compared with the TD children. The next paragraphs will cover inspection of the data and statistical analysis of group differences separately for each type of speech material.

Where should it best go?

Again, simple correlation for the listeners z-scores between conditions is given in appendices, separately for the ASL (Figure C.3) and CCRM material (Figure C.4).

ASL speech material

Surprisingly, a comparison of the groups averaged z-score reveals that the non-switched quiet condition (Quiet-NoAlt) and the switched condition with the nonspeech distractor (AMSSN) yielded the largest separation between the groups, with APD median z-scores of 1.81 and 1.79, respectively, laying just within the norms upper limit. Performance of the APD children was also noticeably poorer for conditions with speech distractors (MDR_F and ENG_F), each with a median z-score of circa 1, whereas performance for Quiet-Alt condition was fairly similar between the groups.

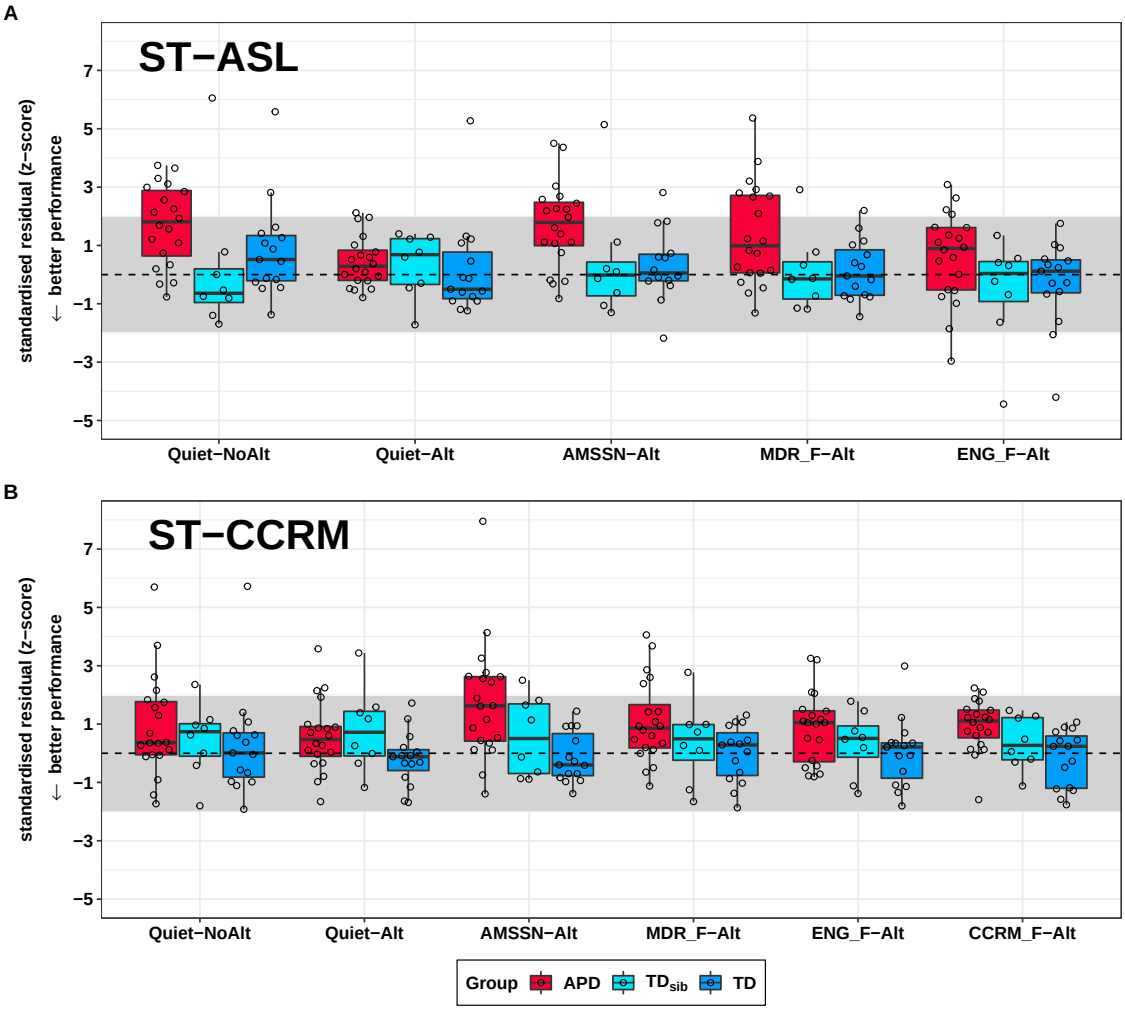

Figure 3.10: ST: Boxplots of the listeners age-independent standardised residuals for data measured with the ASL (A) and the CCRM speech material (B). Residuals were calculated separately for each condition and are based on a model prediction for the TD group only. The grey area represents scores in the 'normal' region, where about 95% of the normal population is expected to lay within. To each side of the grey area, deviance scores were defined as z-scores below or above 1.96, which represents the upper and bottom 2.5% in the TD group. The dashed line represents the theoretical TD group mean (z = 0). Individual scores are indicated by circles.

Within the APD group AMSSN, Quiet-NoAlt and MDR_F resulted in the highest proportion of abnormal scores[4]. Surprisingly, the AMSSN distractor yielded the highest proportion of abnormal scores, where half of the APD children fell outside the norm (20/10, 50%). This was followed by the non-switched condition

[4]With the aim to develop a clinically applicable test that exhibits good sensitivity and specificity, we were only interested in identifying children with clinically poor performance. Thus, an abnormal score was defined as a one-tailed deviance cut-off of z-score > 1.96, within which circa 97.5% of the normal population is expected to lay.

Quiet-NoAlt, where paradoxically and against our expectation 45% of the APD group (9/20) had abnormally poor scores, whereas only 10% (2/20) had abnormal scores in the switched condition Quiet-Alt. Moreover, while the overall performance was similar for the two speech distractors, the percentage of abnormal scores was twice as large for the MDR_F condition (8/20, 40%) than for the ENG_F condition (4/20, 20%). The proportion of abnormal scores amongst the TD group ranged between 0% to 13% (M = 7.8%), which is relatively higher than expected in the normal population.

CCRM speech material

Figure 3.10 B reveals a similar trend for the CCRM sentences, nonetheless with more modest differences between the two groups. Again, AMSSN yielded the largest separation between the groups, where 40% (8/20) of the APD children obtained abnormal scores and with a median score of 1.62, which is relatively close to the +1.96 upper deviance cut-off. In comparison, only 4.3% of the TD children (1/23) had abnormal performance for the AMSSN condition. The APD group median score for the speech distractors was approximately 1 (range: 0.86 - 1.11), however the proportion of abnormal APD children was noticeably smaller than seen for the AMSSN, with 25% (5/20) for the MDR_F, 20% (4/20) for the ENG_F, and only 10% (2/20) for the CCRM_F distractor. Lastly, in contrast to the ASL material, performance for the CCRM sentences presented in quiet were relatively better without switching (NoAlt) than with switching (Alt). Nonetheless, the spread in performance for the non-switched condition was larger. The percentage of abnormal scores in the TD group were relatively low, ranging between 0 to 8.7% (M = 4.3%).

A three-way 3 x 2 x 5 factorial design model with repeated measures was used to test the main effects of Group (APD, TD, & TD_{sib}), Material (ASL / CCRM) and Condition as well as their interaction on performance in the task with z-scores as a dependent variable. Note that the model did not include the CCRM test

condition with CCRM-type sentences as distractor (CCRM_F) since there was no comparable condition in the ASL speech material. Inspection of parametric methods assumptions for the residuals of a linear model revealed that the assumption of a normal distribution was rejected, whereas the assumption of homogeneity of the variance was met. Since there are several obvious outliers in the data and due to the incomplete fulfilment of parametric assumptions a non-parametric approach was adopted. This was tested with a rank-based ANOVA-type statistic test (ATS) using the *nparLD()* function (nparLD package; Noguchi et al., 2012). The analysis was based on a f2-ld-f1 design ATS test, whereby f2 refers to an experimental design with two between-subjects factors (Group & Material) and f1 refers to a single within-subjects factor (Condition). The test results are given in Table 3.14. No significant three-way or two-way interaction were found, except for a Group x Material interaction ($p < 0.05$). In addition, there was no significant main effect of Condition.

To further examine the significant Group x Material interaction, we tested two separate f1.ld.f1 models per test material with Group and Condition as predictors. Both models found a significant effect of Group [ASL: Statistic = 4.099, Df = 1.720, $p < 0.05$; CCRM: Statistic = 3.699, Df = 1.605, $p < 0.05$], while there was no significant interaction or main effect of condition (all p's > 0.05). A post-hoc pairwise comparison using Wilcoxon rank-sum test was performed to examine the main effect of Group (see Table 3.15). There was a highly significant difference in performance between the APD and the TD children without an APD sibling in both test materials. Furthermore, the performance of APD children for the ASL material was (highly) significantly poorer than the performance of the TD children with an APD sibling. However, there was no significant difference in performance between the two groups when measured with the CCRM material. Moreover, while there was no significant difference between the TD groups when measured with the ASL material, TD children with an APD sibling performed significantly poorer

Table 3.14: ST: Statistical analysis for the effects of Group, Material, and Condition as well as their interaction (3x2x5 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f2-ld-f1 design ANOVA-type statistic (ATS) test, whereby f2 refers to an experimental design with two between-subjects factors (Group and Material) and f1 refers to a single within-subjects factor (Condition).

	Statistic	df	p-value
Group	4.009	1.531	***0.028***
Material	0.047	1.000	*0.828*
Condition	1.394	3.251	*0.24*
Group:Material	3.767	1.952	***0.024***
Condition:Material	0.669	2.682	*0.554*
Group:Condition	1.594	5.261	*0.154*
Group:Material:Condition	0.660	4.294	*0.631*

* significant p-values ($p < 0.05$) are shown in bold.

Table 3.15: ST: Post-hoc paired-comparison tests (Wilcoxon rank-sum test) for Group differences in z-scores split by material type.

	ASL model							CCRM model						
contrast	n1	n2	estimate	95%-CI	p	r	magnitude	n1	n2	estimate	95%-CI	p	r	magnitude
APD - TD	100	75	0.89	0.46 - 1.34	***< 0.001***	0.30	moderate	100	75	0.96	0.58 - 1.33	***< 0.001***	0.37	moderate
APD - TD_{sib}	100	40	1.15	0.6 - 1.68	***< 0.001***	0.33	moderate	100	40	0.40	-0.1 - 0.94	*0.34*	0.13	small
TD - TD_{sib}	75	40	0.24	-0.23 - 0.7	*1*	0.09	small	75	40	-0.57	-1.03 - -0.1	*0.05*	0.22	small

* significant p-values ($p < 0.05$) are shown in bold.

than their TD peers when presented with the CCRM material.

An additional 3 x 6 model was computed for the full CCRM data, including the test condition with the CCRM-type distractor (CCRM_F-Alt). The model included Group and Condition as between- and within-subjects predictors, respectively, with z-scores as the dependent variable using nparLD ATS test (f1.ld.f1 design). The ATS test results were similar to those of the full model, with a significant main effect of Group (Statistic = 4.922, Df = 1.597, $p < 0.012$). However, there was no significant main effect for Condition nor a significant Group x Condition interaction (both p's > 0.05). A post-hoc pairwise-comparisons for groups (Wilcox rank-sum tests) found a significant difference between all three groups.

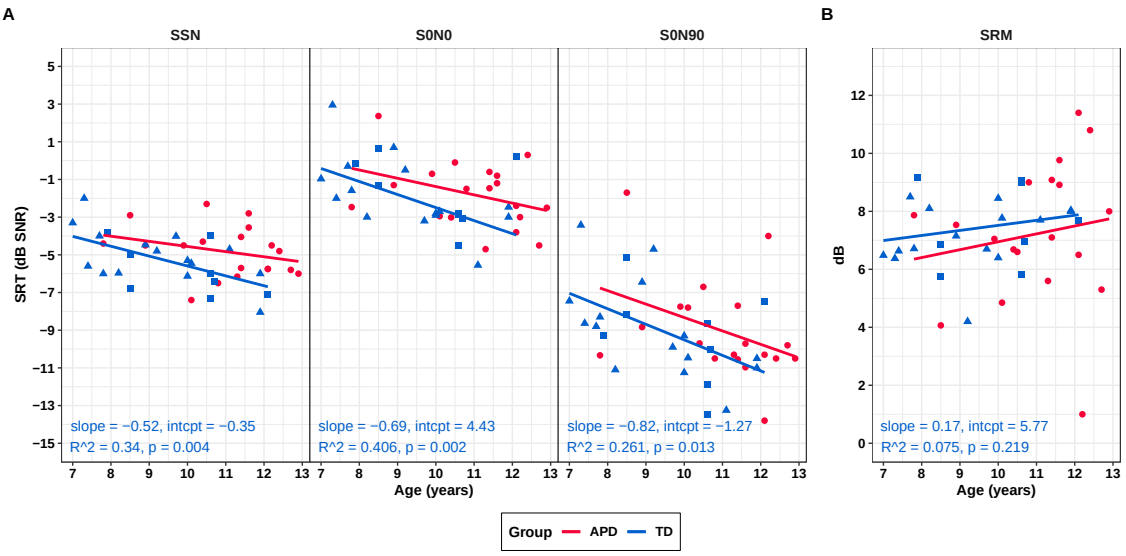

Figure 3.11: LiSNS-UK: Age-effect: scatterplot and linear regression lines for SRTs obtained for SSN and the spatialised conditions S0N0 (collocated) and S0N90 (separated) (A) and the derived measure SRM (B) as a function of the listeners age. Corresponding regression coefficients and statistics are provided for TD group only. Red indicates data from the APD group and cyan indicates data from the TD control group (square shapes: TD children with an APD sibling; triangle shapes: the remaining TD children).

3.3.4 LiSNS-UK

SRTs by age

Listener SRTs and their corresponding regression lines split by group are shown in Figure 3.11 A for the spatially- collocated (S0N0) and separated condition (S0N90), as well as for the non-spatialised condition where the ASL sentences were presented with a speech-shaped-noise (SSN). The listeners binaural advantage, calculated as the difference between the collocated and separated spatial conditions (SRM = S0N0 - S0N90) is shown in Figure 3.11 B. As in the switching task, the age effect was tested in the TD group only, where the regression lines for the TD group were estimated based on a model comparison and outlier trimming procedure to improve the model fits (model coefficients and statistic are given at the bottom of the figures).

As previously reported by other researchers that used a similar test paradigm in children from a similar age group (e.g., Cameron & Dillon, 2007; Murphy et al., 2019), the scatterplots show a clear developmental trend, with an overall improvement in

performance with an increase in age. The test conditions S0N90 and S0N0 showed the largest age effect, with near to 1 dB improvement in performance per 1 year increase (TD slope: -0.82 & -0.69, respectively). The slope for SSN was shallower, with roughly half a dB improvement in performance per 1 year increase, with a TD slope of -0.52. Differences in performance with age for the SRM were negligible, with a predicted improvement of circa 1 dB between the age of 7 to 13 years. There was a significant effect of age in all three test conditions ('moderate' effect size), with the largest effect for S0N0, accounting for circa 40% of variability in performance, followed by SSN with 34% and about 26% for S0N90. The linear regression for SRM showed no significant age effect ($R^2 = 0.075$, $p = 0.219$).

A factorial design model with repeated measures was used to test the main effects for Condition (SSN, SON0, S0N90), Age, and Sibling (TD children with/without an APD sibling, TD/TD_{sib}) with SRTs as a dependent variable and random intercepts for subjects. Note that also here the model included only data for the control group. Assumptions of normal distribution and homogeneity were met, and thus a parametric approach was applied using LMEM (reference levels: Condition = SSN). The model with the best fit and main effects are given in Table 3.16. The final model did not include the fixed factor Sibling or interaction terms, thus, suggesting that performance of the TD children with and without an APD sibling was the same. Since the effect of sibling was not significant, it was not further examined in this section. There was however a highly significant main effect of both Condition and Age ($p < 0.001$), hence indicating that age affected performance similarly across the three test conditions. A separate two-way ANOVA using the *anova()* function was performed to examine the effect of Age and Sibling on SRTs for the SRM data. There was no significant effect of Age [$F(1,18) = 1.48$, $p = 0.239$] or Sibling [$F(1,18) = 0.25$, $p = 0.648$], nor a significant interaction between the two terms [$F(1,18) = 0.20$, $p = 0.658$].

Table 3.16: LiSNS-UK: Age effect: LMEM model for SRT with Condition (SSN, S0N0, S0N90) and Age as fixed factors and random intercepts for subjects (reference level: SSN). Note: only data measured with the control group (TD) following outliers trimming was included.

SRT ~ Condition + Age + (1 | Subjects)			
Main effects	Df	χ^2	p
Condition	2	100.356	***< 0.001***
Age	1	13.364	***< 0.001***

* significant p-values ($p < 0.05$) are shown in bold.

How do I check the assumption of homogeneity here?

lm(uRevs ~ Age * APDsibling)

Age-independent z-scores

Boxplots of the listeners age-independent standardised residual z-scores (blue circles) collapsed across the different test conditions are shown in Figure 3.12, separately for the APD group (red) and TD group (cyan). The z-scores were calculated in the same way as for ST. Again, the dashed line indicates the theoretical TD group mean of zero, and the grey area indicates the lower and upper limit of the normal population (TD mean ± 1.96). Descriptive statistics collapsed by group and test conditions are given in Table 3.17. Overall, when compared with the control group, the APD children exhibited poorer performance across all three test conditions (i.e., higher z-scores) as well as for the derived SRM measure (i.e., lower z-scores).

S0N0 and SRM yielded the largest separation between the groups. However the spread in scores was relatively large and the percentage of abnormal performances in the APD group was rather small, with only circa 26% (5/19) in each condition. Only about 16% (3/19) and 10% (2/19) of the APD children had abnormal scores for SSN and S0N90, respectively. No abnormal performance was obtained in the TD group for SSN and S0N90, while two TD children (~9%) had abnormal scores for S0N0 and one child for SRM. Nonetheless, when excluding the TD outliers that

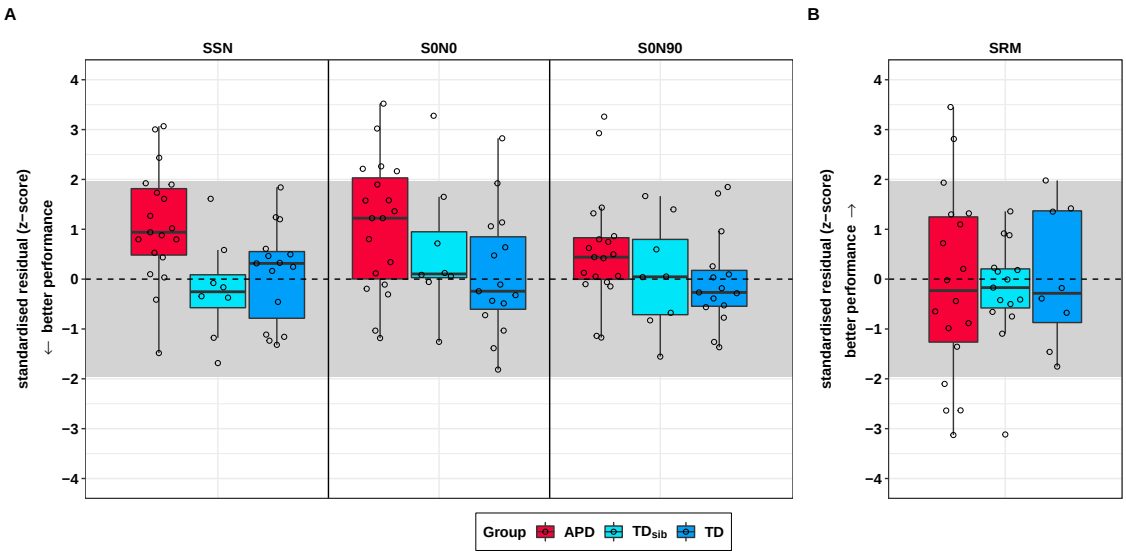

Figure 3.12: LiSNS-UK: Boxplots of the listeners age-independent standardised residuals (open circles) for data measured with LiSNS-UK task (A) and the derived measure SRM (B). Residuals were calculated separately for each condition and are based on a model prediction for the TD group only as plotted for the ST data.

Table 3.17: LiSNS-UK standard residuals (z-scores) descriptives by group. Abnormal: defined as the percentage of z-scores > 1.96 (SSN, S0N0, & S0N90) and z-scores < 1.96 (SRM).

	APD						TD					
	N	median	sd	min	max	abnormal	N	median	sd	min	max	abnormal
SSN	19	0.94	1.14	-1.48	3.07	15.79%	23	0.16	0.98	-1.68	1.84	0.00%
S0N0	19	1.22	1.31	-1.18	3.52	26.32%	23	0.06	1.28	-1.81	3.28	8.70%
S0N90	19	0.44	1.11	-1.17	3.26	10.53%	23	-0.18	0.98	-1.55	1.85	0.00%
SRM	19	-0.44	2.39	-6.77	3.45	26.32%	23	-0.18	1.15	-3.12	1.98	4.35%

were trimmed during the z-score calculation procedure, all the TD observations were within the norms.

Group differences between the APD and the TD group for the test conditions SSN, the spatialised conditions S0N0 and S0N90 were tested with an LMEM model with z-scores as a dependent variable and random intercepts for subjects (reference levels: Condition = SSN; Group = APD). Parametric methods assumptions of normal distribution and homogeneity of variance were met. The model which gave the best fit and main effects are given in Table 3.18. The final model did not incloude Condition x Group interaction, thus suggesting that the two groups

Table 3.18: LiSNS-UK: Group differences: LMEM model for the age-independent z-scores with Condition and Group as fixed factors (reference levels: Condition = SSN; Group = APD) and random intercepts for subjects.

z ~ Condition + Group + (1 | Subjects)			
Main effects	Df	χ^2	p
Condition	2	3.809	*0.149*
Group	1	8.673	***0.003***

* significant p-values ($p < 0.05$) are shown in bold.

behaved in a similar way in the three test conditions. There was a significant main effect of Group ($p = 0.003$), whereas there was no significant main effect of Condition ($p > 0.05$). Differences in z-scores between the two groups for the SRM data were examined with a one-way ANOVA test. The parametric assumption of homoscedasticity was violated (Levene's test; $p = 0.013$), while the assumption of a normal distribution was met. Nevertheless, since nonparametric methods gave similar results, for simplicity, it was decided to report here only the outcomes of the parametric method. The test found no significant difference in z-scores between the groups [$F(1,40) = 0.334$, $p = 0.566$].

3.3.5 ENVASA

Due to technical problems, observations for six listeners are missing (x2 TD; x4 APD), resulting in a total sample-size of 21 and 17 for the TD and the APD group, respectively. Initial inspection was performed to ensure that the task instructions were followed and well understood. Performance for the reference condition (single incongruent background at a high SNR), which is expected to least impact performance, was compared with a cut-off criterion of 56%, calculated as 2 SD from the TD group mean (84% ± 14%). Individuals with performance below the cut-off criterion were excluded from the analysis. One TD listener aged 7 years old scored 45 % and was thus excluded, resulting in a total of 20 listeners in the TD group.

%-correct by age

The ENVASA measurements followed the same factorial design as used by Leech et al. (2009), with 2 background types (single/dual) x 4 SNRs (low: -6, -3 dB; high: 0 +3 dB), resulting in a total of 92 responses (%-correct, PC) per listener or between 10 to 11 test items per background-SNR combination. Because of the small number of test items per condition, responses were averaged into three measures: 1. *single background*, 2. *dual backgrounds*, and 3. *combined background* which reflects the overall performance across the two background types.

The relationship between performance and age was inspected in the same way as carried out for the other auditory tasks, with the listeners average performance plotted as a function of age, with linear regression lines and model coefficients for the trimmed TD group (see Figure 3.13). The regression lines revealed a noticeable developmental trend in all three measures, where performance improved with increasing age. A single linear regression line with a monotonic increase in performance by age was found to best fit performance for a single background, with an increase of circa 3.5% in PC per year. Performance for dual backgrounds and the combined score on the other hand were best described using segmented linear regression models, with an increase of PC by circa 12% per year until the age of 9 years, where PC plateaued thereafter.

The effect of age was statistically tested using an LMEM model with PC as a dependent variable, and with background type (single / dual), the listeners' age, and Sibling (TD children with/without an APD sibling, TD / TD_{sib}) as fixed factors as well as random intercepts for subjects (reference levels: Background = single-background; Sibling = TD). Parametric assumptions of normal distribution and homogeneity were met. As before, the analysis included only data for the control group following outlier trimming procedure. A model without an interaction term was found to give the best fit (see Table 3.19). Model comparison revealed

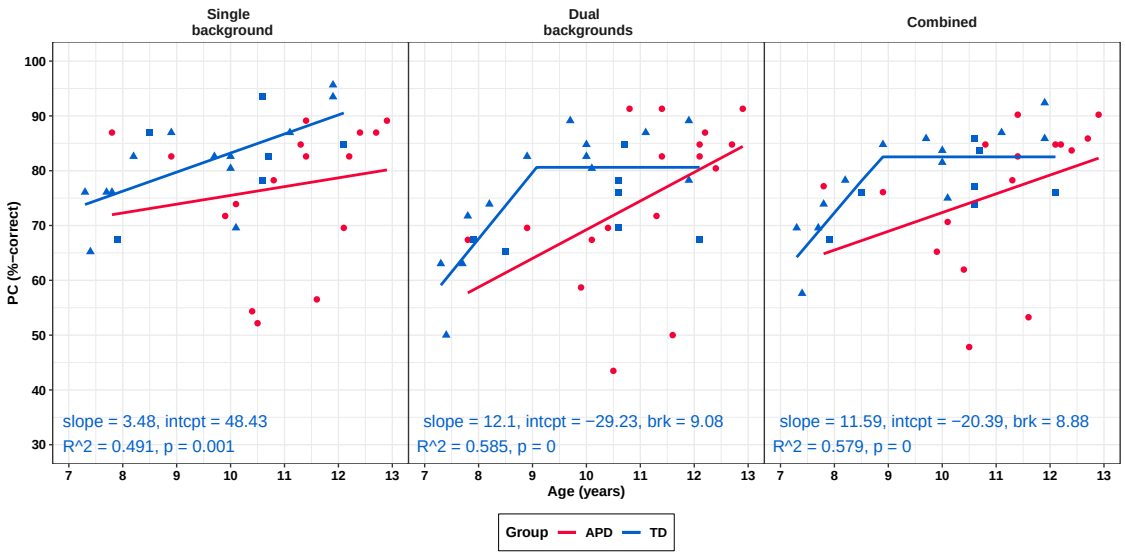

Figure 3.13: ENVASA: Scatterplot and linear regression lines for the listeners' PC (%-correct) as a function of age for single background, dual backgrounds and the combined measure. Red indicates data from the APD group and cyan indicates data from the TD control group (square shapes: TD children with an APD sibling; triangle shapes: the remaining TD children).

a highly significant main effect of Age and Background ($p < 0.001$). This is in agreement with Krishnan et al. (2013) where they found a strong developmental effect across normal-hearing typically-developing children in a similar age range to those measured in the present study. The listeners' PC score was significantly poorer for the dual-backgrounds than for the single-background condition with an estimated mean difference of 6.93 (SE = 1.83, 95%-CI = 3.12 - 10.73, Cohen's d = 0.72, 'medium' effect-size). In addition, the main effect of Sibling was found significant ($p < 0.05$), with an estimated mean difference of 5.47 (SE = 2.73, CI = -0.18 - 11.13) and a 'small' effect-size (Cohen's d = 0.21). A separate two-way ANOVA using the *anova()* function was performed to examine the effect of Age and Sibling on PC for the combined score across the two background types. There was a highly significant effect of Age [$F(1,16) = 19.759$, $p < 0.001$], whereas there was no significant effect of Sibling [$F(1,16) = 2.953$, $p = 0.105$], nor an Age x Sibling interaction [$F(1,16) = 1.748$, $p = 0.205$].

Table 3.19: ENVASA: Age effect: LMEM model for PC (%-correct) with Background (single / dual), Age, and Group (TD children with/without an APD sibling, TD / TD_{sib}), as fixed factors and random intercepts for subjects (reference levels: Background = single-background, Sibling = TD). Note: only data measured with the control group following outlier trimming was included.

PC ~ Background + Age + Sibling + (1 | Subjects)			
Main effects	Df	χ^2	p
Background	1	11.285	***< 0.001***
Age	1	17.802	***< 0.001***
Sibling	1	4.251	***0.04***

* significant p-values ($p < 0.05$) are shown in bold.

Age-independent z-scores

For further analysis, age was controlled for using the same multiple-case approach method described in Section 3.2.4. Descriptives of the listeners' z-scores collapsed by the three test conditions and groups are given in Table 3.20. Boxplots of the age-independent z-scores for the three ENVASA measures are shown in Figure 3.14, with larger z-score indicating better performance. Surprisingly, the less demanding condition with the single competing background yielded the largest separation between the APD and the TD group with a median z-score of roughly -1, while the median performance for dual backgrounds and the combined score was relatively similar to those in the control group, albeit with larger spread. The percentage of abnormal APD scores was relatively low, with circa 29% (5/17) for the combined score, 24% (4/17) for single background and 18% (3/17) for dual backgrounds condition. There was only one case of an abnormal score in the TD group for a single background (5%, 1/20) when trimmed TD outliers were included.

Group differences for z-scores measured with the two background conditions (single / dual) were examined with a 3x2 factorial design model with repeated measures. Since the previous model with PC as a dependent variable revealed a significant difference in performance between the TD children with and without an APD sibling, this was further investigated here with the predictor Group which

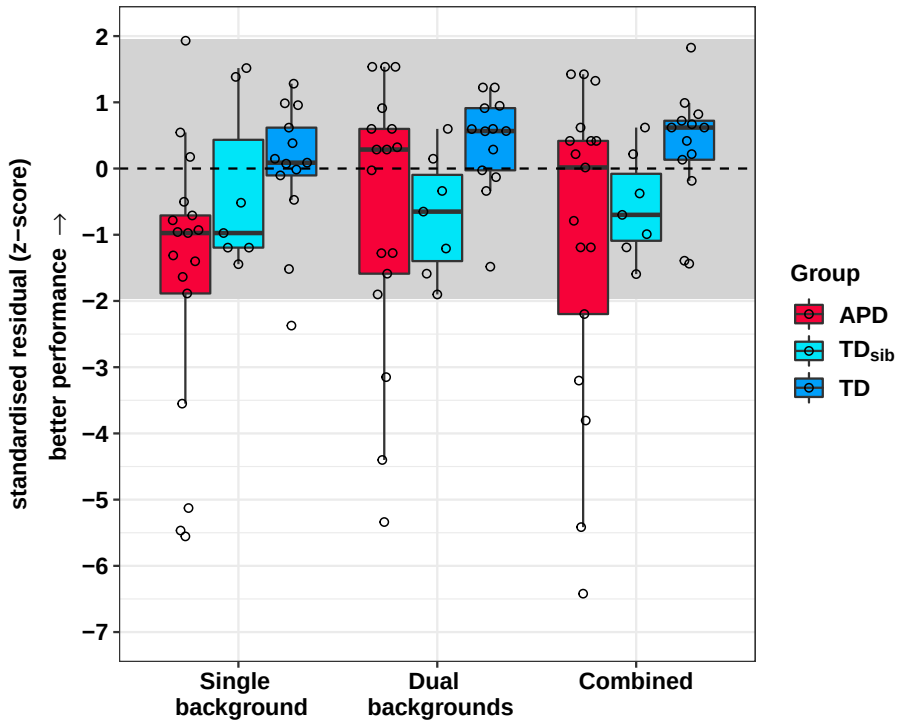

Figure 3.14: ENVASA: Listeners' age-independent standardised residuals for single background, dual backgrounds & the combined measure. Residuals were calculated separately for each condition and are based on a model prediction for the TD group only as plotted for the ST data.

comprised of three levels: APD, TD, & TD_{sib}. Parametric model assumption of normal distribution was rejected (Shapiro-Wilk test; p = 0.001), while the assumption of homoscedasticity of variance was met. Thus, a robust nonparametric rank-based ANOVA-type ATS test was computed using the *nparLD()* function with a f1.ld.f1 experimental design with Background as a within- and Group as a within-subjects factors. There was a significant main effect of Group (Statistic = 3.280, Df = 1.950, p = 0.039), whereas there was no significant effect of Background (Statistic = 1.047 , Df = 1.000, p = 0.306) or Group x Background interaction (Statistic = 1.859, Df = ,1.589 p = 0.164). Differences between the three groups were examined with nonparametric post-hoc pairwise-comparison Wilcoxon rank-sum tests [*wilcox_effsize()* function, rstatix package]. The test revealed a significant difference between the APD and TD group (p < 0.05, 'moderate' effect-size), whereas there was no significant difference between the APD or the TD group and the TD_{sib} group (all p's > 0.05; see Table 3.21).

Table 3.20: ENVASA: Descriptive and statistics of the listeners age-independent standard residuals (z-scores) split by groups and test measures.

	APD						TD					
background	N	median	sd	min	max	abnormal	N	median	sd	min	max	abnormal
Single	17	-0.97	2.11	-5.56	1.93	23.53%	20	0.03	1.08	-2.37	1.52	5.00%
Dual	17	0.29	2.07	-5.34	1.54	17.65%	20	0.22	0.95	-1.90	1.22	0.00%
Combined	17	0.02	2.39	-6.42	1.42	29.41%	20	0.22	0.95	-1.59	1.83	0.00%

Table 3.21: ENVASA: Post-hoc paired comparison tests with Bonferroni correction (Wilcoxon rank-sum test) for Group differences in z-scores.

contrast	n1	n2	estimate	95%-CI	p	r	magnitude
APD - TD	34	26	-1.07	-1.87 - -0.31	***0.02***	0.34	moderate
APD - TD_{sib}	34	14	-0.21	-1.34 - 0.63	*0.62*	0.07	small
TD - TD_{sib}	26	14	1.04	0.07 - 1.57	*0.08*	0.33	moderate

3.3.6 CELF-RS

The children's raw scores were converted into age-corrected scaled scores using the CELF-5 UK Recalling Sentences subtest standardised norms (M = 10, SD = 3). Boxplots of the children's scaled scores split by groups are given in Figure 3.15. The white area indicates the upper and lower limit among the normal population (± 1 SD). On average, performance was within the normal range in both the APD group (Mdn = 9) and the TD group, albeit laying within the upper limit (Mdn = 13). Thus, although the majority of the APD children had expressive language skills that were within the norms, the figure shows a clear difference in performance between the group, with the TD children performing noticeably better. Almost half of the TD children obtained a scaled score above the average (i.e., scaled score > 13) and none exhibited abnormal scores. On the other hand, only three APD children performed above the average and the performance of two children was considered abnormal (scaled score < 7).

From the boxplots, it is apperant that the TD_{sib} group poorer expressive language skills than their TD peers. While circa 67% (10/15) of the children's scaled score in the TD group was above the average norm (in other words, outside the grey area), this was the case only in one child out of eight in the TD_{sib} group.

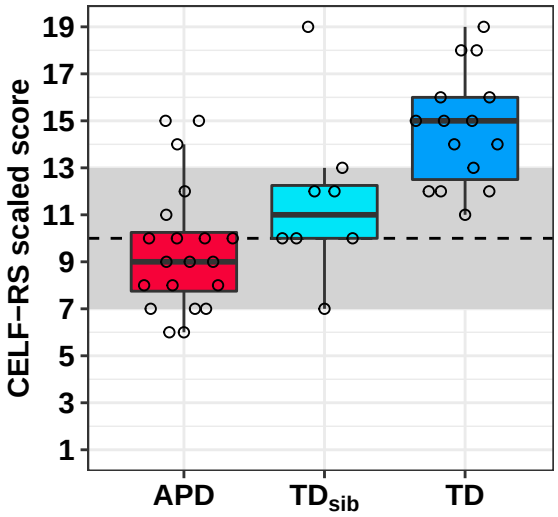

Figure 3.15: CELF-RS: Boxplots for CELF-5 UK Recall Sentences subtest scaled scores by groups. The grey area represents scores in the 'normal' region (± 1 SD) and the dashed line represents the average score within the normal population.

A one-way ANOVA was computed using the *anova()* function to compare the listeners scaled scores in the three groups (APD, and TD children with/without an APD sibling). The parametric assumption of homoscedasticity was met while the assumption of a normal distribution was marginally significant (Shapiro-Wilk test; $p = 0.041$). However, since nonparametric methods gave similar results, it was decided to report here only the outcomes of the parametric method. There was a highly significant difference in scaled scores between the groups [$F(2,40) = 14.476$, $p < 0.001$]. A post-hoc pairwise comparison t-tests with Bonferroni correction using the *pairwise_t_test()* function (rstatix package; Kassambara, 2021) found a highly significant difference between the APD group (Mdn = 9.0, SD = 2.7) and the TD group without an APD sibling [Mdn = 15, SD = 2.4, $t(31.9) = -5.84$, $p < 0.001$], whereas there was no significant difference between APD and TD children with an APD sibling [Mdn = 11, SD = 3.5, $t(10.6) = -1.50$, $p = 0.486$] or between the two TD groups [$t(10.7) = 2.19$, $p = 0.155$].

3.3.7 Questionnaires

CCC-2

Data for one TD listener was flagged as inconsistent using the test scorer and was thus removed from the analysis. The group descriptives for the parental reports in the different sub-scales as well as the GCC and SIDC composites are given in Table 3.22. GCC stands for general communication composite, calculated by taking the sum for scaled scores A to H. It is used to clinically identify abnormal communication skills, defined by a GCC < 55 (10^{th} percentile). The SIDC stands for social-interaction deviance composite [sum(E+H+I+J)-sum(A+B+C+D)], where in combination with abnormal GCC score, the SIDC can be used to identify the child's primary difficulty, whereby, a positive SIDC is indicative of a predominantly structural language deficit (referred to here as DLD), and a negative SIDC reflects social communication problems and is indicative of autistic spectrum disorder (ASD) traits (Bishop, 2003; Norbury, 2014).

Boxplots of the groups scaled scores in the ten sub-scales and a scatterplot depicting the relationship between GCC and SIDC are shown in Figure 3.16 A-B, respectively. A striking 90% of the APD children (18/20) obtained a scaled score below the 5th percentile two or more times, which has been found to indicate clinically significant communication problems (Bishop, 2003), whereas, only one such case (out of 22) was found in the TD group. The single-value GCC composite showed the exact same proportion of abnormal scores in both groups when a cut-off value of 55 was used, where only one TD child had abnormal communication skills (see Figure 3.16 B). Half of the APD children with an abnormal GCC score (45%, 9/20) exhibited a score pattern that is indicative of DLD, whereas the other half exhibited a negative SIDC, indicating social communication deficits as the primary difficulty. Interestingly, out of the nine APD children who fell within the latter category, three were reported by their parents to have HF-ASD diagnosis, and an additional two children were undergoing an ASD assessment at the time of testing

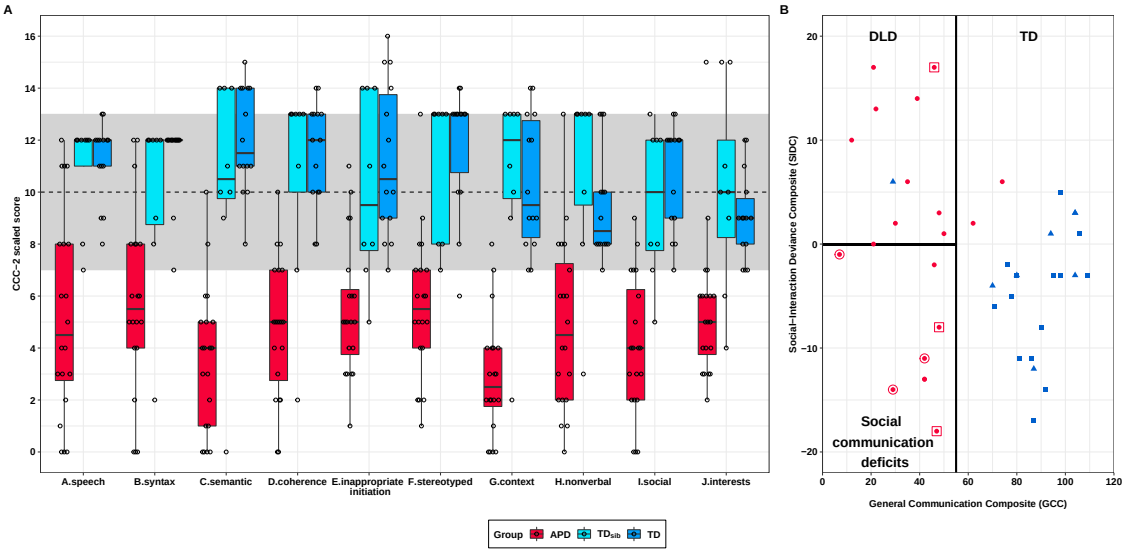

Figure 3.16: CCC-2 parental reports for the APD (red) and TD group (cyan). (A) Boxplots for scaled scores in the ten sub-scales. (B) Scatterplot for General Communication Composite (GCC) as a function of Social-Interaction Deviance Composite, (SIDC). Figure A: The grey area represents scores in the 'normal' region (± 1 SD) and the dashed line represents the average score within the normal population. Figure B: Red indicates data from the APD group and cyan indicates data from the TD control group (filled square shapes: TD children with an APD sibling; triangle shapes: the remaining TD children). APD children with diagnosed high-functioning Autism (HF-ASD) are denoted with open circles. APD children with undergoing ASD assessment on the day of testing are marked with open squares. The lines indicates the GCC cut-off criteria for typically developing children (TD) SIDC scores indicative of predominantly structural developmental language disorder (DLD) and more social communication deficits (cf. Norbury, 2013).

(see scores marked with open circles and squares in Figure 3.16 B). Differences in GCC between the three groups (APD, and TD with/without an APD sibling) were tested using a one-way ANOVA test using the *anova()* function. The parametric assumption of a normal distribution and homoscedasticity were met. There was a highly significant difference between the groups [$F(2,39) = 43.712$, $p < 0.001$]. Post-hoc pairwise comparison t-tests with Bonferroni correction (*pairwise_t_test()*, rstatix package) revealed that performance of the APD group (Mdn = 42.0, SD = 16.4) was significantly poorer than of the TD group with [Mdn = 90.5, SD = 25.5, $t(9.4) = -4.7$, $p < 0.01$] or without an APD sibling [Mdn = 88.5, SD = 11.3, $t(32.0) = -10.7$, $p < 0.001$]. Furthermore, there was no significant difference found between the two control groups [$t(8.61) = 0.53$, $p = 1.00$].

Table 3.22: CCC-2 subscales descriptives split by groups.

	APD					TD				
Measure	N	median	sd	min	max	N	median	sd	min	max
A.speech	20	4.5	3.96	0	12	22	12.0	1.72	7	13
B.syntax	20	5.5	3.61	0	12	22	12.0	2.49	2	12
C.semantic	20	4.0	2.78	0	10	22	11.0	3.23	0	15
D.coherence	20	5.0	2.68	0	10	22	12.5	2.87	2	14
E.inappropriate.initiation	20	5.0	2.61	1	11	22	10.5	3.17	5	16
F.stereotyped	20	5.5	2.82	1	13	22	13.0	2.52	6	14
G.use.of.context	20	2.5	2.28	0	8	22	10.5	2.97	2	14
H.nonverbal	20	4.5	3.31	0	13	22	10.0	2.75	3	13
I.social	20	4.0	2.68	0	9	22	12.0	2.41	5	13
J.interests	20	5.0	2.84	2	15	22	9.0	2.63	4	15
GCC	20	42.0	16.38	7	74	22	88.5	17.38	29	109
SIDC	20	1.5	10.70	-18	17	22	-3.0	6.14	-17	6

GCC, General Communication Composite sum(A+B+C+D+E+F+G+H);
SIDC, Social Interaction Deviance Composite sum(E+H+I+J) - sum(A+B+C+D)

ECLIPS

Descriptives of the ECLiPS parental report scaled scores for the different subscales and composite measures split by groups are given in Table 3.23 and depicted in Figure 3.17. A score below the 10^{th} percentile (corresponding to a scale score of circa 6) is generally considered to indicate clinically significant listening and processing difficulties (Barry & Moore, 2014). Overall, the ECLiPS was able to well separate between the two groups across all the different sub-scales. All APD children exhibited an abnormal Total score, whereas only two TD children (out of 22) did.

A closer look at the boxplots in Figure 3.17 reveals a clear difference in the distribution of the scaled scores across the two groups, with relatively larger spread for the TD group. Another interesting trend was that the parental reports for the TD_{sib} group was on average better (i.e., higher) than for the TD group. Inspection of the Total score by groups revealed that the APD group did not follow a normal distribution and that the assumption of homoscedasticity was violated ($p < 0.05$). Thus group differences for the listeners' Total score was examined using a robust one-way ANOVA with trimmed means (20%) and bootstrapping (N=2000) using

Table 3.23: ECLiPS descriptives split by groups and sub-scales.

	APD					TD				
Measure	N	median	sd	min	max	N	median	sd	min	max
SAP	20	1	0.93	0	3	23	10	2.77	4	14
L/L/L	20	2	1.55	0	5	23	10	2.78	3	15
M&A	20	3	1.10	2	6	23	11	2.69	7	16
PSS	20	4	1.53	2	8	23	10	3.37	3	15
EAS	20	3	1.64	1	8	23	9	3.52	4	14
Listening	20	2	0.93	1	4	23	11	2.69	5	15
Language	20	2	1.28	0	4	23	11	2.48	5	15
Social	20	3	1.52	1	7	23	10	3.15	5	15
Total	20	1	0.91	0	3	23	10	2.92	4	15

SAP = Speech & Auditory Processing; L/L/L = Language, Literacy & Laterality; M&A = Memory & Attention; PSS = Pragmatic & Social skills; EAS = Environmental & Auditory sensitivity; Listening = (SAP + PSS) / 2; Language = (L/L/L + M&A) / 2; Social = (PSS + EAS) / 2; Total = mean of all sub-scales

the *t1waybt()* function (WRS2 package; Mair & Wilcox, 2020). The test found a highly significant difference between the groups (F = 99.35, $p < 0.001$). A post-hoc pairwise comparison of groups with bootstrapping (N=2000) was computed using the *mcppb20()* function from the same package, whereby $\hat{\psi}$ denotes the pairwise trimmed difference (Mair & Wilcox, 2020). There was a highly significant difference ($p < 0.001$) between the APD group (Mdn = 1.0 , SD = 0.91) and both TD groups with (Mdn = 12.0, SD = 3.45, $\hat{\psi}$ = -11.08, 95%-CI = -13.08 - -7.42) or without an APD sibling (Mdn = 9.0, SD = 2.47, $\hat{\psi}$ = -8.47, 95%-CI = -10.03 - -6.86), whereas no significant difference was found between the TD groups ($\hat{\psi}$ = -2.61, 95%-CI = -5.05 - 1.28, p = 0.106).

3.4 Overall performance

An overview of the childrens' performance split by group is given in Figure 3.18, which provides a simple graphical display indicating which test scores fell outside norms for each participant (filled black cells). Abnormally poor performance for the listeners age-independent scores was defined using standardised norms for the CELF-RS, ECLiPS and the CCC-2 data or was defined as a one-tailed cut-off of

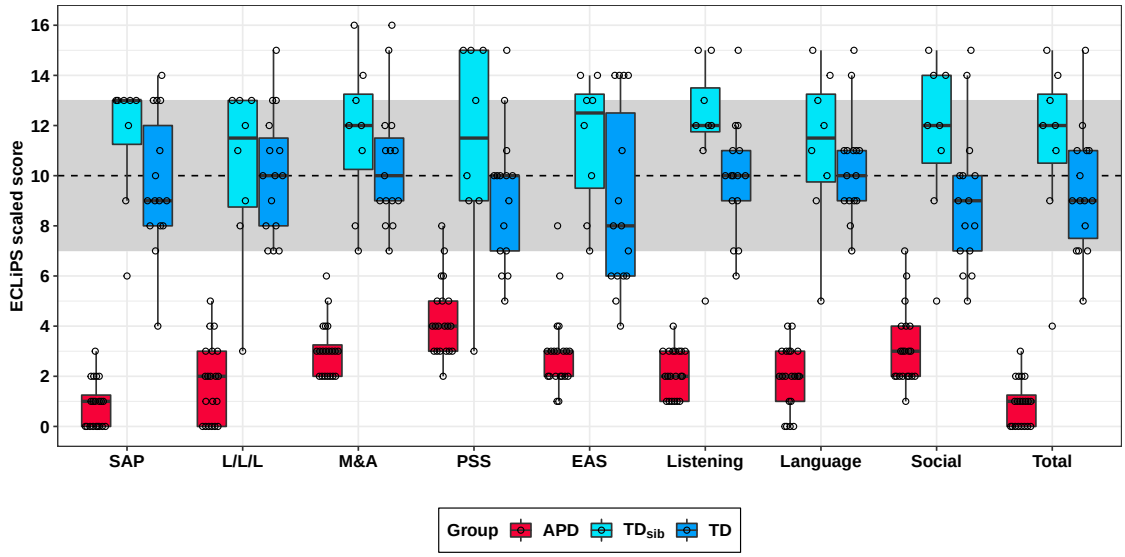

Figure 3.17: ECLiPS parental report scaled scores split by groups and sub-scales. The grey area represents scores in the 'normal' region (± 1 SD) and the dashed line represents the average score within the normal population.

± 1.96 (where circa 97.5% of the normal population is expected to lay within) for the rest of the tasks. Note that DLD and PLI were composed as a way to discriminate children with more structural versus pragmatic language deficit and were based on the CCC-2 data as a combination of abnormal GCC score (< 55) and the SIDC score. The DLD score (developmental language disorder) denotes a combination of abnormal GCC and a positive SIDC (≥ 0) which is expected to capture severe deficits in structural language in conjunction with only mild pragmatic difficulties. The PLI score (pragmatic language impairment), on the other hand, denotes a combination of abnormal GCC and a negative SIDC (≤ 0) which is expected to be a strong indicator for social communication problems with only mild structural language difficulties.

As seen in the figure, the proportion of abnormal scores across the APD group is substantially higher than in the TD group. The majority of the APD children (80%, 16/20) performed abnormally in at least two test conditions either in the ST or LiSNS-UK task, whereas there were only three cases (13%, 3/23) in the TD children. Another interesting observation is that apart from one TD child,

who experienced difficulties in various measures including the CCC-2, none of the other TD children experienced language difficulties. This is in contrast to the APD group where 90% (18/20) of the children experienced some kind of language deficit. Although the CELF-RS has been reported to be a good marker for children with DLD, nevertheless, the results of the present study suggests otherwise. While performance in the APD group was noticeably poorer than in the TD group, only two APD children obtained abnormally poor CELF-RS score, whereas nearly half of the APD children (45%, 9/20) exhibited a CCC-2 score indicative of DLD, and about the remaining half (40% 8/20) obtained a CCC-2 score indicative of pragmatic language and social communication deficit (PLI).

Potential experimental bias of reporters when recruited due to an informed group affiliation? email Courtenay!

The proportion of abnormal scores by measure or task split by group is shown in Figure 3.19. Both the ECLiPS total score and the CCC-2 GCC sum score resulted in the largest separation between the groups. Out of the auditory tasks, the tests conditions that resulted in the highest proportion of abnormal scores in the APD group were AMSSN (ASL: 50%, CCRM: 40%), Quiet-ASL-NoAlt (45%) and MDR_F-ASL (40%), whereas only 26% of the APD children had abnormal SRM score.

3.4.1 Switching task: effect-size

Effect-size [Cohen's d; *rstatix::cohens_d()*] was calculated for pairwise group comparisons by material & condition (see Table 3.24). Three test conditions resulted in a 'large' effect size: CCRM_F-Alt-CCRM (d = 1.01), AMSSN-Alt-CCRM (d = 1.00) and MDR_F-ASL-Alt (d = 0.85). Four other conditions resulted in a 'moderate' effect-size, with d ranging between 0.63 to 0.75. These conditions

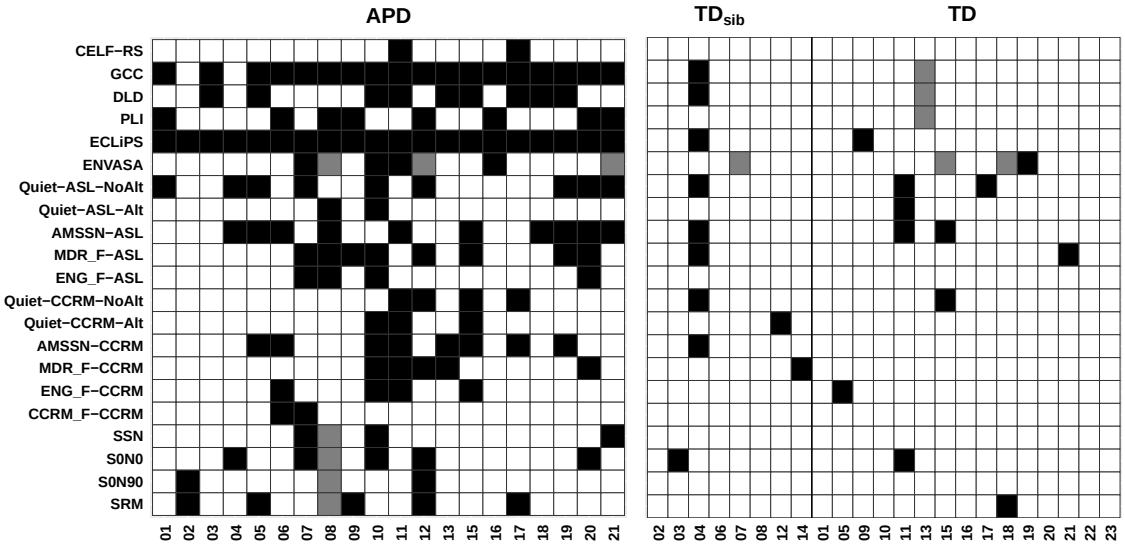

Figure 3.18: Overall performance: Abnormal (black cells) and normal (empty cells) performance in the test battery of individuals from the APD group (n=20) and the TD group (n=23). Missing data is marked by the grey cells. The vertical black line in the TD group separates between children with an APD sibling (TD$_{sib}$) from the remaining TD children.

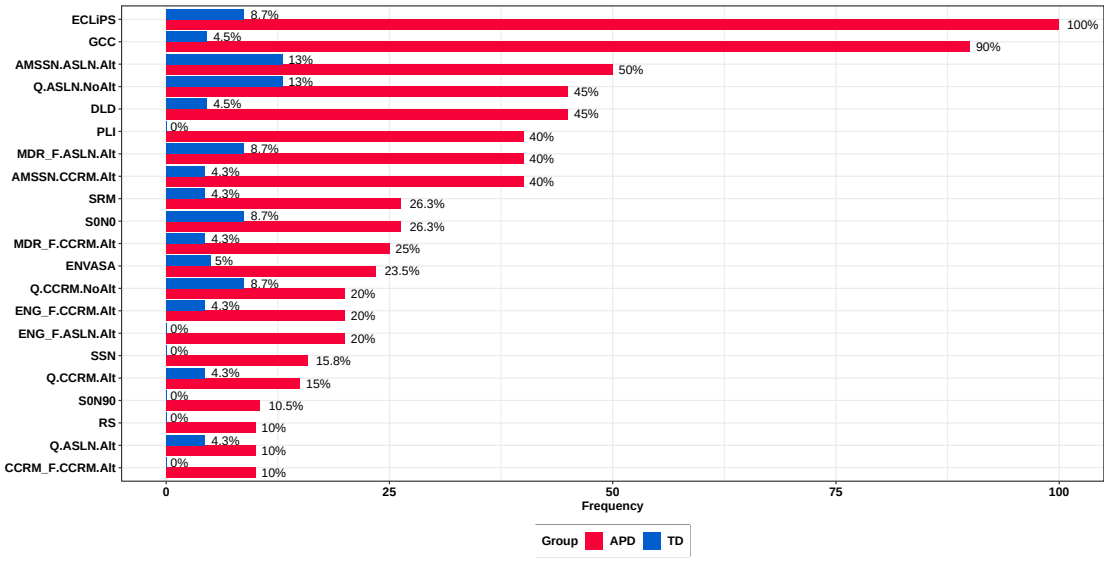

Figure 3.19: Overall perfromance: proportion of abnormal score per measure or task split by groups.

Table 3.24: Cohen's d by condition and material.

condition	material	N_{APD}	N_{TD}	95%-CI	Cohen's d	magnitude
Q-NoAlt	ASL	20	23	0.02 - 1.72	0.63	moderate
Q-Alt	ASL	20	23	-0.39 - 0.98	0.18	negligible
AMSSN-Alt	ASL	20	23	-0.15 - 1.57	0.49	small
MDR_F-Alt	ASL	20	23	0.2 - 1.54	0.85	large
ENG_F-Alt	ASL	20	23	0.05 - 1.31	0.64	moderate
Q-NoAlt	CCRM	20	23	-0.26 - 0.97	0.34	small
Q-Alt	CCRM	20	23	-0.27 - 1.07	0.35	small
AMSSN-Alt	CCRM	20	23	0.49 - 1.79	1.01	large
MDR_F-Alt	CCRM	20	23	0.2 - 1.38	0.76	moderate
ENG_F-Alt	CCRM	20	23	-0.01 - 1.36	0.64	moderate
CCRM_F-Alt	CCRM	20	23	0.43 - 1.86	1.01	large

comprised of speech distractors (ENG or MDR) from either material and Quiet-NoAlt for the ASL material. The test material which resulted in the largest effect-size was estimated by averaging d across conditions for each type of material. Both materials had a 'moderate' average effect-size, whereby the CCRM material had the largeer effect-size of 0.69, following with 0.56 for the ASL material.

3.4.2 Interaction between measures

The present study involved a large number of test conditions and various measures assessing different skills. For example, the ST data alone comprises 11 different conditions (x5 ASL, x6 CCRM speech material). Another set of measures consisting of the CELF-RS, ECLiPS and the CCC-2 taps into language and communication related skills, whereby the latter two consists of a sum of 15 different sub-scales which, however, have been shown to strongly correlate with one another (Barry & Moore, 2014). Examining the extent to which performance is explained by such a large number of measures will result in a very conservative significance level in order to minimise Type-I error (false positive), and could increase Type II error rate (false negative) (McDonald, 2014). Since the measures within the ST and within the language dataset are expected to strongly correlate, it was decided to use

an exploratory data analysis technique – Principal Components Analysis (PCA). PCA is a technique used to reduce a large number of correlated parameters into a smaller set of components that together explain a considerable amount of the variability in the larger dataset. Each of the PCA components is composed of a linear combination of the input parameters (James et al., 2013). PCA was performed separately for the ST and language data set using the FactoMineR package (Lê et al., 2008) with scaled units and will be discussed separately below.

ST

The PCA for the ST z-scores comprised 11 input variables and a sample size of 43. Sample size adequacy for PCA was verified using a Kaiser-Meyer-Olkin test (psych::KMO; Revelle, 2020), with an overall KMO of 0.76 ('good'; Field et al., 2012), and a KMO range between 0.66 to 0.85 across the conditions. Bartlett's sphericity test was significant [$\chi^2(55) = 190.36$, $p < 0.001$], indicating that the correlations between the different items were large enough for a PCA. Table 3.25 shows the variables loadings (no rotation was applied), their eigenvalues and percentage of variance explained. Loadings are indicators of substantive importance of a given variable to a given component (Field et al., 2012). The first three components were used, yielding eigenvalues > 1 (Kaiser's criterion), explaining together circa 67% of the variance in the data. The first component (PC.ST) accounted for the largest portion of spread in the data of 40.6% and was interpreted as an overall measure for performance in the switching task with relatively high loadings across all separate thresholds. The remaining components explained each circa 16% and 11% of the variance (ascending order). Figure 3.20 illustrates the different dimensions in the data captured by the three PCA components. Clustering in the second component (PC2.Material) reflected differences in performance across the two speech materials (ASL & CCRM). The third component (PC3.Nz) reflected the degree of distractability introduced by speech distractors (MDR_F, ENG_F, & CCRM_F) irrespective of the speech material used, resulting in decrements in performance when compared with non-speech distractors or target-only conditions (Quiet and

Table 3.25: Switching task PCA: Input variables loading.

Item	PC1.ST	PC2.Material	PC3.Nz
Q-ASLN-NoAlt	**0.59**	**0.60**	0.08
Q-ASLN-Alt	**0.61**	**0.42**	**0.43**
AMSSN-ASLN.Alt	**0.61**	**0.50**	**0.36**
MDR_F-ASLN-Alt	**0.68**	**0.36**	**-0.41**
ENG_F-ASLN-Alt	**0.69**	0.22	**-0.40**
Q-CCRM-NoAlt	**0.52**	**-0.35**	**0.56**
Q-CCRM-Alt	**0.59**	**-0.42**	0.09
AMSSN-CCRM-Alt	**0.67**	**-0.49**	0.17
MDR_F-CCRM-Alt	**0.72**	**-0.34**	-0.11
ENG_F-CCRM-Alt	**0.72**	**-0.34**	-0.16
CCRM_F-CCRM-Alt	**0.58**	-0.12	**-0.41**
eigenvalue	4.46	1.73	1.21
variance (%)	40.52	15.72	10.98
cumulative variance (%)	40.52	56.24	67.22

|loading| >0.3 are highlighted in bold.

AMSSN). Boxplots of the listeners weighted scores for the PCA components split by group are shown in Figure 3.21. PC1.ST shows a substantial separation between the two groups, with very little overlap in scores between the TD group and the majority of the APD children. Separation between the two groups in the remaining components is noticeably smaller.

Figure 3.22 illustrates the relationship between the listeners weighted scores based on the three PCA components (PC1.ST, PC2.Material and PC3.Nz) and three calculated composites composed from the listeners z-scores based on the interpretation stated above; where *ST* denoted the listeners' aggregated overall score across all ST conditions, and the two calculated discrepancy composites denoted as *Material* and *Nz*. The Material composite was calculated by subtracting the mean score of all CCRM conditions ($\overline{CCRM}$) from the mean score of all ASL conditions ($\overline{ASL}$), i.e., Material = $\overline{ASL}$ - $\overline{CCRM}$. The remaining composite, Nz, was calculated by subtracting the listeners performance averaged across conditions with speech distractors ($\overline{Spch}$) from the average performance taken across the nonspeech and quiet conditions ($\overline{NoSpch}$), i.e., Nz = $\overline{NoSpch}$ - $\overline{Spch}$. As can be seen in the figure,

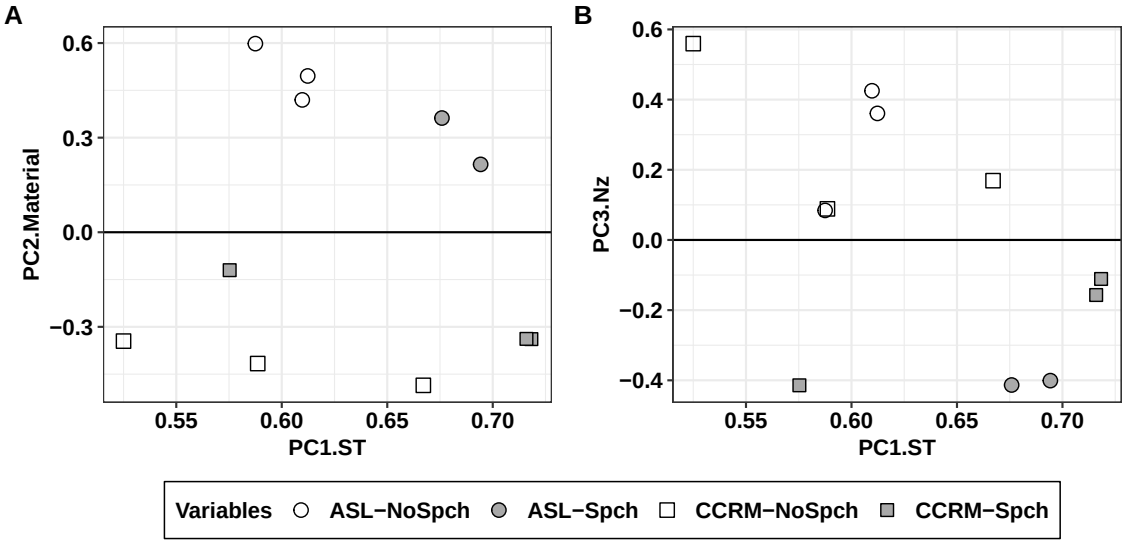

Figure 3.20: Switching task PCA: Scatterplot for the input variables as a function of PCA components: PC1.ST vs. PC2.Material (A), PC1.ST vs. PC3.Nz (B). Loadings for ASL conditions are indicated by circles and loadings for CCRM conditions are indicated by rectangles. Filled shapes denote conditions with speech distractors (Spch) and non-filled shapes denote nonspeech conditions (NoSpch).

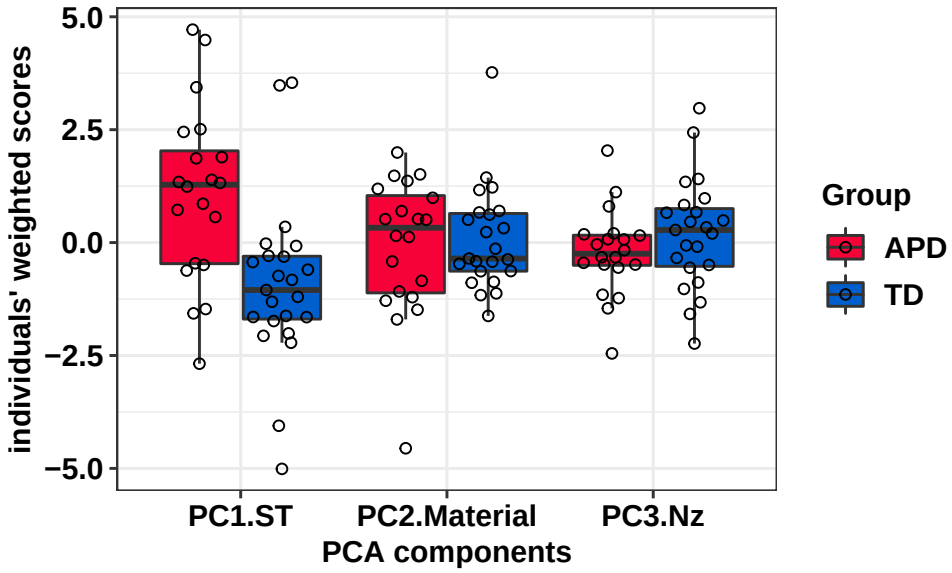

Figure 3.21: Switching task PCA: Listeners weighted scores split by components and group.

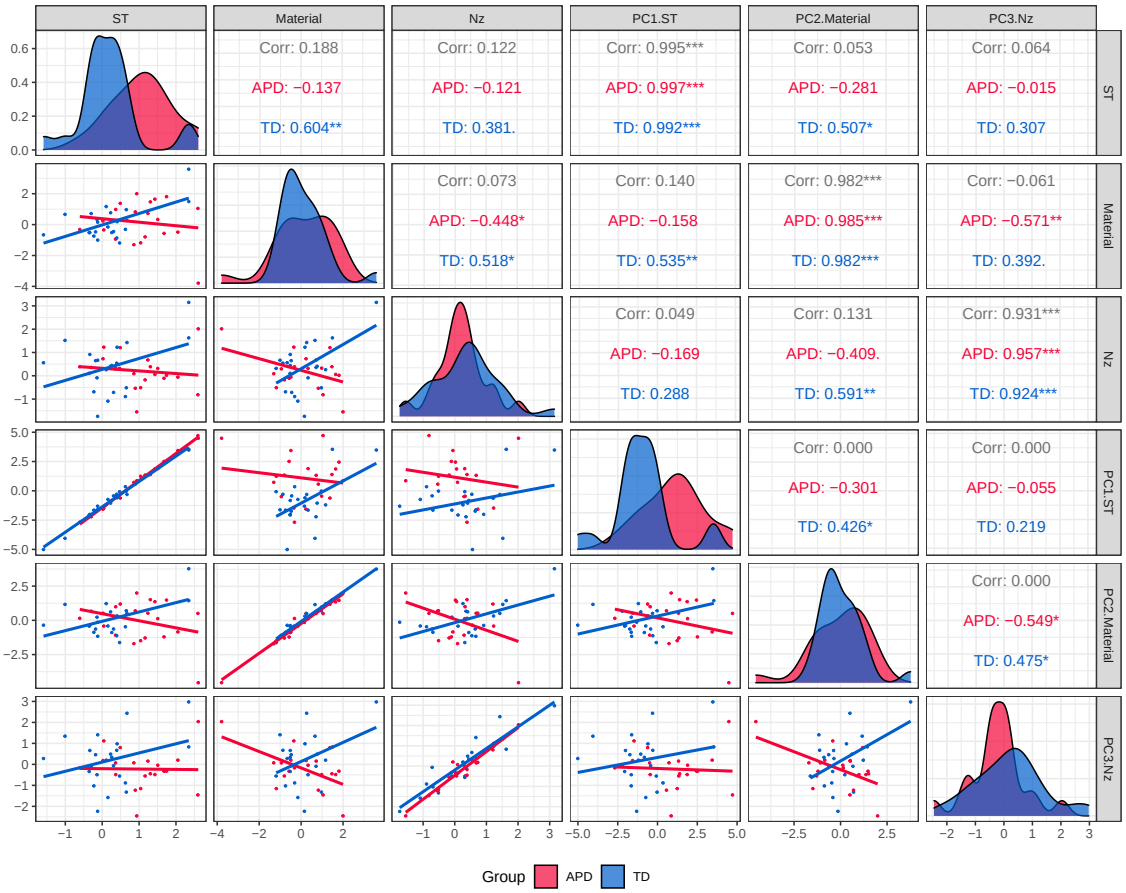

Figure 3.22: Switching task PCA: Comparison between PCA weighted scores and calculated measures: (1) ST = mean score across all ST data, (2) Material = $\overline{ASL}$ - $\overline{CCRM}$, (3) Nz = $\overline{NoSpch}$ - $\overline{Spch}$.

the PCA components highly correlated with the respective calculated composites (PC1.ST - ST, PC2.Material - Material, PC3.Nz - Nz), whereas none correlated with another composite, thus indicating that the components are independent from one another and that each describe different dimensions within the data.

Language measures

A PCA with three components was computed for the listeners' scaled scores obtained in the different language measures, comprising of 16 input variables (x1 CELF-RS, x5 ECLiPS, x10 CCC-2) with a sample size of 42. Data for one TD child was excluded from the analysis due to inconsistent CCC-2 responses. The Kaiser-Meyer-Olkin test for sample-size adequacy was 'superb' (Field et al., 2012) with an overall

KMO of 0.93 (range: 0.86 - 0.97) and the assumption of sphericity was verified using Bartlett's sphericity test [$\chi^2(120) = 787.52$, $p < 0.0001$]. The PCA variables loadings, eigenvalues and percentage of variance explained split by components is given in Table 3.26. The first component (PC1.Lang) yielded eigenvalue > 1, explaining circa 73% of the variance, reflecting an overall performance averaged across all the language measures. The remaining components had eigenvalues of just under 1 (0.95 & 0.85, respectively), each explaining circa 6% and 5% of the variance. The second component (PC2.Lang) reflected the discrepancy between expressive language skills, measured by the CELF-RS and listening and communication skills measured by the ECLiPS subscales. Interestingly, the third component (PC3.Lang) reflected once again a discrepancy, clustering together variables that taps into pragmatic language and social interaction skills such as the ECLiPS subscale PSS (pragmatic & social skills) and the CCC-2 subscales E, H, I & J, separating them from other variables that assess more structural language skills such as the CELF-RS and the CCC-2 subscales speech (A) and Syntax (B). Boxplots of the listeners weighted scores for the PCA components split by group is shown in Figure 3.23. As seen in the ST data, the first component (PC1.Lang) best separated the two groups, whereas separation between the two groups in the remaining components was noticeably smaller.

Despite the small proportion of variance explained by the latter two principal components, they do capture other aspects of language and communication skills that may be relevant in explaining the individual and group differences in the auditory tasks and were therefore included in the analysis. Nevertheless, interpretation of the relationship between these components with performance in the auditory tasks should be viewed with caution. Inspection of the individuals' scaled scores split by groups for loadings in PC1.Lang as a function of loadings in PC2.Lang and PC3.Lang shown in Figure 3.24 A-B revealed a linear relationship between PC1.Lang and PC2.Lang (APD group) and between PC1.Lang and PC3.Lang (TD group), thus indicating that they are not entirely independent from one another.

Table 3.26: Language measures PCA: Input variables loading.

Item	PC1.Lang	PC2.Lang	PC3.Lang
CELF-RS	**0.69**	**0.40**	**0.37**
ECLIPS.SAP	**0.91**	**-0.32**	0.14
ECLIPS.LLL	**0.92**	-0.14	0.11
ECLIPS.M.A	**0.88**	**-0.30**	0.14
ECLIPS.PSS	**0.83**	**-0.36**	-0.17
ECLIPS.EAS	**0.79**	**-0.52**	0.07
CCC2.A.speech	**0.78**	0.08	**0.35**
CCC2.B.syntax	**0.82**	0.19	0.27
CCC2.C.semantic	**0.92**	0.14	0.05
CCC2.D.coherence	**0.92**	0.09	0.05
CCC2.E.inappropriate.initiation	**0.82**	0.13	**-0.41**
CCC2.F.stereotyped	**0.89**	0.22	-0.03
CCC2.G.use.of.context	**0.93**	0.04	-0.08
CCC2.H.nonverbal	**0.84**	0.13	-0.26
CCC2.I.social	**0.88**	0.15	-0.16
CCC2.J.interests	**0.80**	0.11	**-0.40**
eigenvalue	11.67	0.95	0.85
variance (%)	72.96	5.95	5.3
cumulative variance (%)	72.96	78.91	84.21

|loading| >0.3 are highlighted in bold.

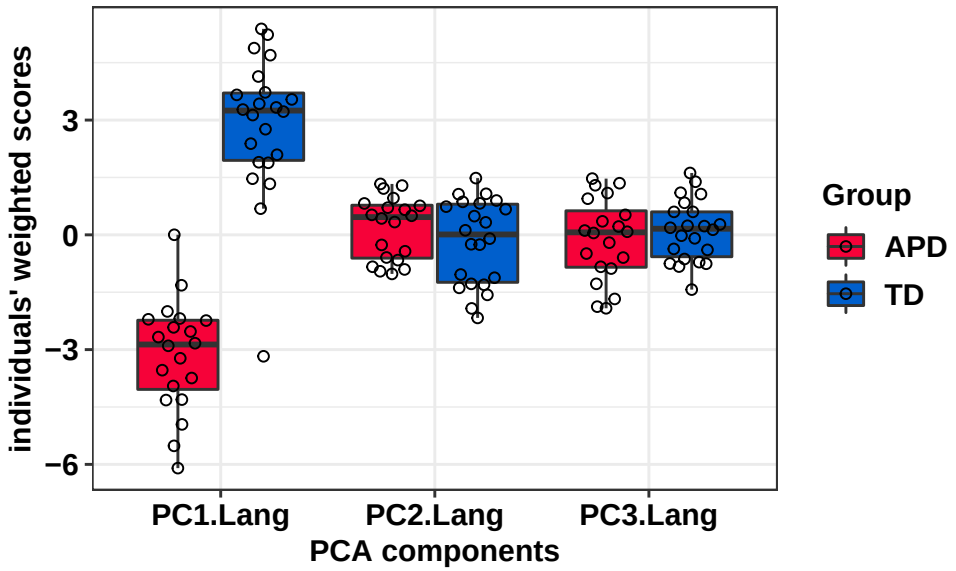

Figure 3.23: Language measures PCA: Listeners weighted scores split by components and group

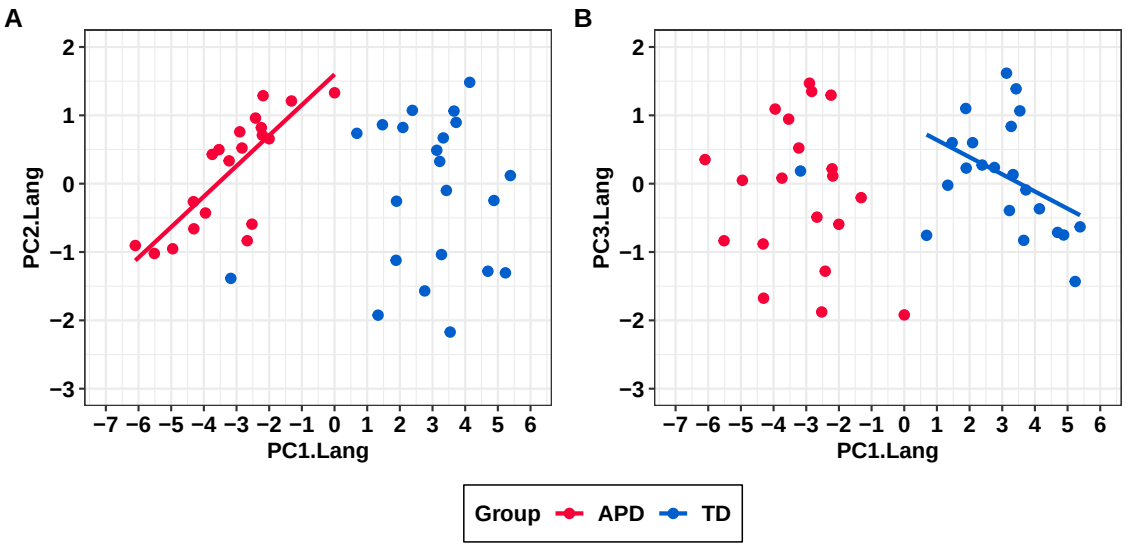

Figure 3.24: Language measures PCA: Individual scores split by groups for loadings in PC1.Lang as a function of scores for PC2.Lang (A), and PC3.Lang (B).

The partial lack of independence may be in part explained by the large difference in scores between the groups across the different input variables.

Again, the relationship between the PCA components (PC1.Lang, PC2.Lang and PC3.Lang) and the three calculated composites that reflects the component interpretations is illustrated in Figure 3.25. The calculated components were based on the listeners scaled scores, where *Lang1* represents the overall performance aggregated across all the language scores, *Lang2* represents discrepancy between expressive language skills (CELF-RS) and listening and communication skills (all ECLiPS subscales), and lastly, *Lang3* stands for discrepancy between structural and pragmatic & social skills. As seen in the figure, correlations were high between each PCA component and the corresponding calculated composites (range: 1 to 0.58).

Discussion(?): Which measures were best described by the PCA?

All the measures showed strong correlation with PC1.Lang, whereas the CCC-2 GCC score showed the largest correlation ($\rho = 0.98$, p <.0001). This was true not only for the data aggregated across groups, but also when correlations were examined

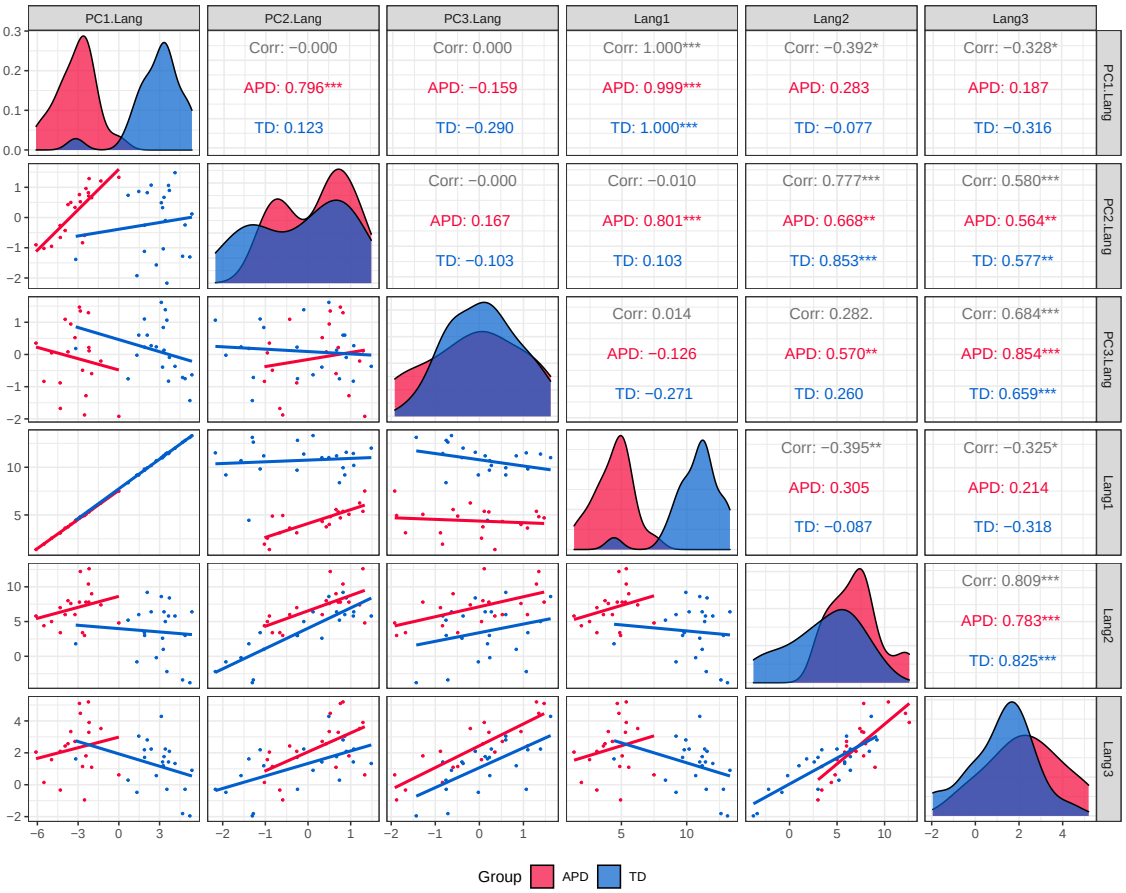

Figure 3.25: Language measures PCA: Comparison between the listeners weighted scores by components, PC1.Lang - PC3.Lang (A), and calculated measures, Lang1 - Lang3 (B).

separately in each group. Therefore, taking into account the short administration time and simplicity, the CCC-2 alone provides a good screening tool for children's language and communication skills with high levels of sensitivity and specificity. Nonetheless, children in the present study knowingly consented to take part in the study either as part of the clinical APD group or the control group, which may have introduced bias in the reporters' responses, and may have resulted in a larger separation between the two groups than one would expect across the true population.

Correlations

Next, the extent to which individual differences in speech perception could be explained by other measures was examined for the aggregated data across the two

groups with multiple Spearman's rho correlations using the *rcorr* function (Hmisc R package; Harrell Jr, 2020) between SSN scores, LiSNS-UK scores for the spatialised conditions and the derived score for spatial release from masking (S0N0, S0N90 & SRM), the principal components for the switching task PC1.ST, PC2.Material and PC3.NZ, and for the language measures PC1.Lang, PC2.Lang and PC3.Lang, average PTA at standard audiometry frequencies (0.5-4 kHz), average PTA at high-frequencies (PTA_{EHF}, at 8, 11 and 16 kHz), and ENVASA total score as a measure for sustained and selective-attention skills. The age effect was accounted for either by using standardised norms when available or by a regression model based z-score transformation. The correlation matrix outcomes are given in Table 3.27.

Table 3.27: Correlation matrix (Spearman) between the study test measures for aggregated data across the two groups.

	PTA	PTA_{EHF}	ENVASA	SSN	S0N0	S0N90	SRM	PC1.ST	PC2.Material	PC3.Nz	PC1.Lang	PC2.Lang
PTA												
PTA_{EHF}	0.31											
ENVASA	-0.10	-0.13										
SSN	0.26	0.01	-0.40*									
S0N0	0.26	0.02	-0.19	0.39*								
S0N90	0.45**	0.34*	-0.23	0.30	0.64****							
SRM	-0.39*	-0.36*	0.12	-0.07	0.08	-0.67****						
PC1.ST	0.46**	0.09	-0.27	0.46**	0.34*	0.44**	-0.23					
PC2.Material	-0.17	-0.14	0.01	0.30	0.34*	0.20	0.12	0.06				
PC3.Nz	-0.03	0.03	0.05	0.09	-0.10	-0.10	-0.03	-0.11	0.01			
PC1.Lang	-0.16	-0.07	0.46**	-0.51***	-0.19	-0.15	-0.02	-0.55***	-0.03	0.16		
PC2.Lang	0.07	0.07	0.12	-0.02	0.21	0.23	-0.14	-0.01	0.16	-0.04	0.08	
PC3.Lang	-0.10	-0.26	0.00	0.09	-0.03	-0.12	0.08	-0.02	0.09	-0.14	-0.05	-0.02

significant p-values: **** p < .0001, *** p < .001, ** p < .01, * p < .05

There was a significant correlation between the listeners overall performance in the switching task (PC1.ST) and their language skills (PC1.Lang; ρ = -0.55, p < 0.001), PTA ($\rho = 0.46$, p < 0.01), speech perception in noise (SSN; $\rho = 0.46$, p < 0.01), and the spatialised LiSNS-UK test conditions S0N0 ($\rho = 0.35$, p < 0.05) and S0N90 ($\rho = 0.45$, p < 0.01). The second ST principal component, PC2.Material, significantly correlated with S0N0 ($\rho = 0.33$, p < 0.05) and SRM ($\rho = 0.33$, p < 0.05), whereas no relationship was found between the third PC3.Nz and any of the study measures.

Performance in the LiSNS-UK exhibited the highest correlation coefficients, with highly significant correlations between S0N0 and S0N90, where better performance in

one condition was highly associated with better performance in the other ($\rho = 0.64$, $p < 0.0001$), and between S0N90 and SRM (ρ = -0.67, $p < 0.0001$), where better SRM was predicted by better performance for S0N90, whereas correlation between S0N0 and SRM was not significant ($\rho = 0.08$, $p = 0.62$). Note that a lower z-score in the spatialised conditions denotes better performance, whereas the opposite holds for SRM, with higher z-scores marking better performance, thus explaining the negative correlation between SRM and S0N90. A separate group-wise analysis gave similar results for the correlation between S0N90 and SRM, whereas correlations in the APD group between S0N0 and S0N90, and between S0N0 and SRM were smaller and not significant (ρ: 0.35 and 0.30, respectively). The non-significant correlation between SRM and S0N0 stands in contrast to our expectations, for a positive correlation, where listeners with poorer (i.e., higher) S0N0 scores were expected to have a larger (i.e., better) SRM. The insignificant and reduced correlation in the APD group is likely due to sampling error and due to the small sample size in the present study (correlation between the LiSNS-UK condition for the listeners SRT and z-scores are given in appendices in Figures C.5 and C.6).

The SSN score was found to be related to performance in the two spatialised LiSNS-UK test conditions with correlation coefficients of 0.30 (S0N90) and 0.39 (S0N0). However only the correlation for S0N0 was significant ($p < 0.05$), while the p-value for correlation with S0N90 was just above the significance level ($p = 0.055$). The listeners' S0N90 scores significantly correlated with hearing sensitivity thresholds measured at both standard (PTA; $\rho = 0.45$, $p < 0.01$) and extended frequencies (PTA_{EHF}; $\rho = 0.34$, $p < 0.05$). Moreover, none of the LiSNS-UK measures significantly correlated with the language principal components PC1.Lang - PC3.Lang or the attention measure ENVASA. Additional significant correlations were found between PC1.Lang and SSN (ρ = -0.51, $p < 0.0001$) and between PC1.Lang and the ENVASA task ($\rho = 0.46$, $p < 0.001$). No p-value Bonferroni correction for multiple comparisons was applied.

Exploratory predictors – APD group

Association of potential predictors with performance in the APD group was examined in the following section. Nevertheless, it is important to emphasise that this is an exploratory examination across a small sample size and thus the outcomes may not be generalised in a larger sample. Predictors were selected based on the caregivers response in the background questionnaire, where the APD children were subdivided into the following pair of groups: 1. APD diagnosis (APD vs. LiD), 2. SPD diagnosis (SPD vs. non-SPD), 3. Regular use of FM-device (FM vs. No FM), 4. History of middle ear problem (MEHx vs. No MEHx), 5. Pressure equalisation tube history (PET vs. No PET), and 6. Auditory training (Training vs. No training). The listeners performance subdivided by predictors is shown in Figure 3.26 for data measured with the ST task (PCA1.ST), the language composite (PCA1.Lang), SRM, and thresholds for standard audiometry (PTA) and EHF audiometry (EHF PTA). Individual observations are marked in circles, whereby observations of children diagnosed with APD are filled in dark blue, and LiD observations are filled in light blue. Individual data for the TD group is marked in black. From the boxplots, PET and MEHx emerges as the best predictors, explaining the largest portion of the within-group differences. A history of PET showed the highest association with poorer EHF PTA thresholds, and to a relatively smaller extent with PC1.ST (higher score indicates poorer performance) and with the SRM score (higher score indicates better performance). Consequently, it is not surprising that a related predictor – history of middle ear problem (MEHx) was also highly related to poorer EHF PTA thresholds. Nevertheless, the association between MEHx and the other measures was weak. Interestingly, there was no association between SRM score and a diagnosis of SPD, with only a small difference between APD children with or without an SPD diagnosis.

MEHx: is a composite calculated based on the caregivers indication of history of **Ear infection** & **Glue ear** in the background questionnaire. MEHx is 1 if

response was 'Yes' to at least one of these items, whereas MEHx is 0 if response was 'No' for both items.

Exploratory predictors – TD group

3.5 Discussion

3.5.1 EHF

Lee's thresholds for 10-21 yrs group: 8=16.35 (1.46-29.33); 11=22.99, 16=48 (20.01-91.35); 20=93.07 (48.57-105.00) all dB SPL

EHF in children: Read Schechter et al., 1986:

- 6-10 yrs: 10k=23, 12k=20, 16k=39 dB SPL
- 11-15 yrs: 10k=21, 12k=22, 16k=51 dB SPL

3.5.2 ST

APDsibling While the causes of APD are not fully understood, amongst others, studies have shown a strong association between APD and history of hearing problems (e.g., due to chronic OME) causing auditory deprivation, and Developmental Language Deficit (DLD). Moreover, several studies demonstrated that these factors are genetically influenced (Pennington & Bishop, 2009; Bishop, Adams, & Norbury, 2006), while others suggested that some APD-linked aspects (tone sequencing) are more environmentally driven (cf. Moore 2007; Bishop, 2002). Studying twins or siblings is a usefull way to examin both genetic and environmental factors. Although the inclusion of control children with APD sibling(s) was not part of the study design and was not carefully balanced for the control group sample size, it is possible that the two control groups would preform differently in the study measures. Whether the influencing factors were mostly heritable or accquiered by the child's environment, we hypothesised that the control children with APD sibling will perform poorer than the non-sibling control children.

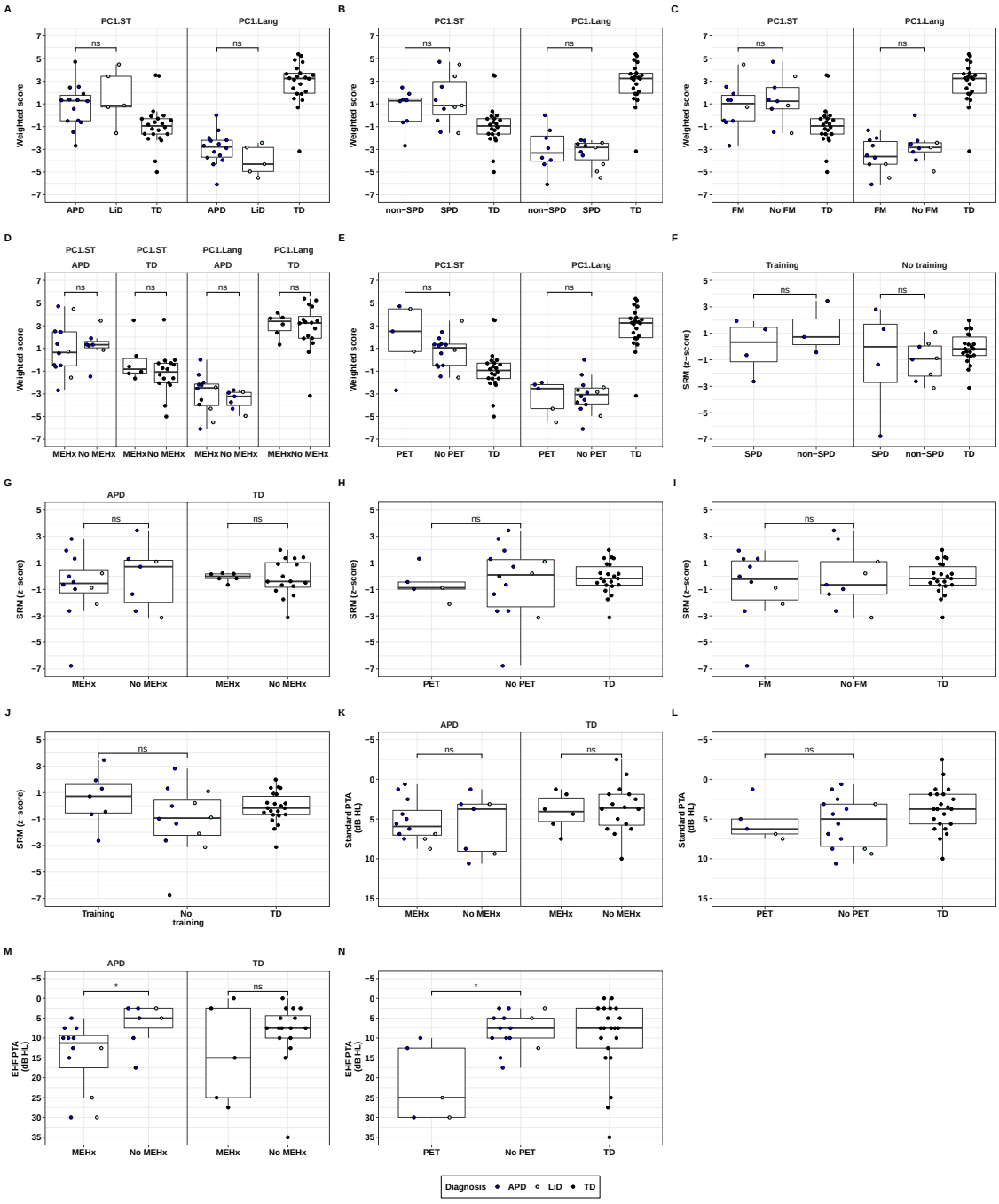

Figure 3.26: Association between predictors and performance in the APD group for the switching task composite (PC1.ST), language composite (PC1.Lang), SRM, standard and EHF PTA. Predictors included: 1. APD diagnosis (APD vs. LiD), 2. SPD diagnosis (SPD vs. non-SPD), 3. Regular use of FM-device (FM vs. No FM), 4. History of middle ear problem (MEHx vs. No MEHx), 5. Pressure equalisation tube history (PET vs. No PET), and 6. Auditory training (Training vs. No training). Individual observations are marked in circles. Observations of children diagnosed with APD are filled in dark blue, and LiD observations are filled in light blue. TD group observations are marked in black. Significant p-values for independent t-tests paired-comparisons are marked with asterisk ($p < 0.05$).

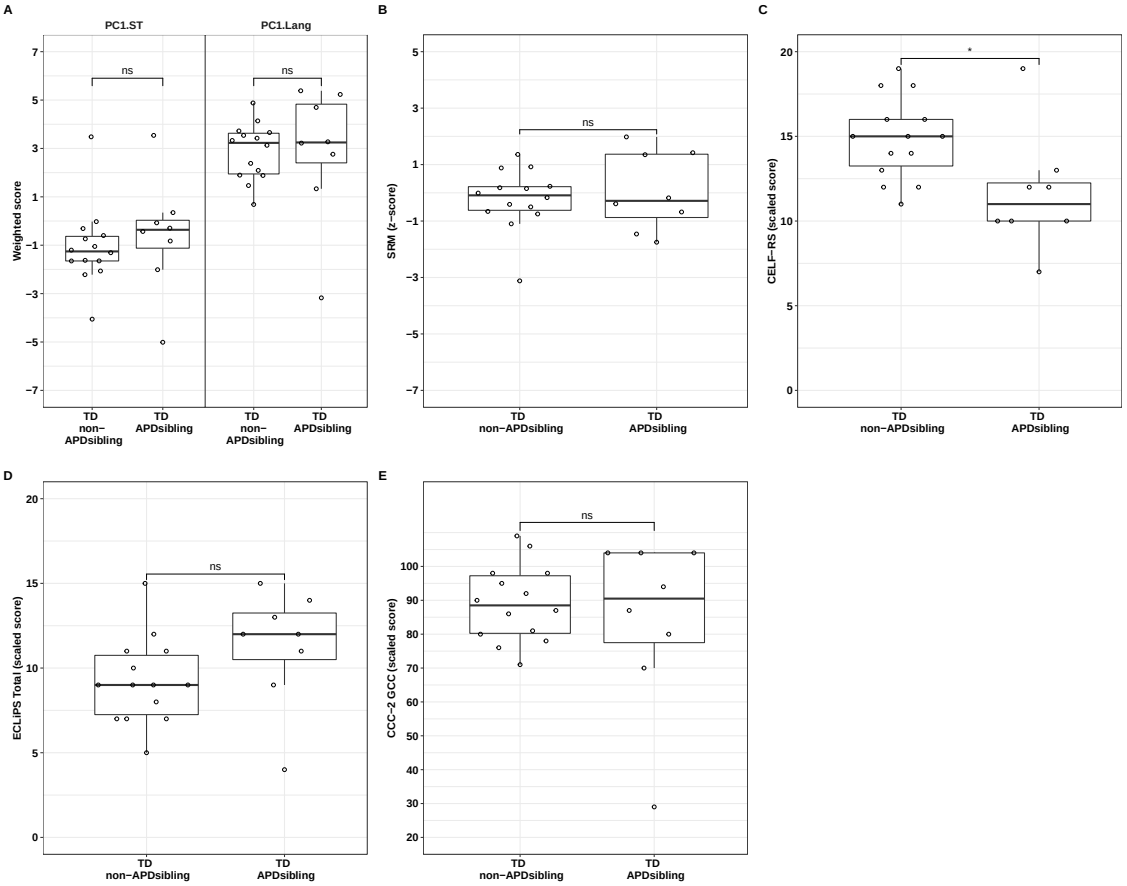

Figure 3.27: Add text here. Significant p-values for independent t-test comparison are marked with asterisk ($p < 0.05$).

Comparison with adults data (SRdTs!)

An exploratory comparison between the children's data measured in the present study with data measured across young NH adults collected in Chapter 2 further highlights the strong developmental trend, with SRdTs still not entirely "adult-like" even at the age of 13 years, especially for speech distractors (see boxplots in Figure 3.7 A-B). The children in both groups seem to be markedly susceptible to competing CCRM sentences and for familiar- or unfamiliar-speech presented with ASL sentences, with performance at the age of 12 years still largely differing from those obtained by the adults. On the other hand, by the age of 12 years, the TD children reached near to "adult-like" performance when CCRM target sentences were presented with ENG_F speech distractor or when ASL sentences were presented with AMSSN distractor.

Why CCRM performance is better

The improved intelligibility in the CCRM material is amongst others due to the more simple speech material, the reduced confusion between the target sentences and the connected speech distractors as well as the restricted alternative responses of the CCRM matrix-based sentences.

TD siblings with APD These results are in contradiction to our expectations if any differences arose. In such a case, we would have predicted a larger decrement in performance (i.e., poorer score) for the ASL sentences which are more linguistically challenging than for the CCRM sentences.

What we predicted and why

SR: Maybe you need to say that this supposition arises from the common finding that family members tend to be more alike, and that sibs with APD should be more APD-like, and esp insofar as there appear to be language deficits in the APD group, etc, etc,

Actually, this would be clearer to say after all the rest of the results are reported so more in the discussion?

z- scores by material: proportion of abnormal TD kids:

ASL: The proportion of abnormal scores amongst the TD group ranged between 0% to 13% (mean = 7.8%), which is relatively higher than expected in the normal population. Nonetheless, when taking into account TD observations that were trimmed during the z-score calculation procedure, the proportion of abnormal scores are smaller, ranging between 0% to 9.5% (mean = 3.8%), which corroborate fairly well with the theoretical probability of 2.5% (one-tailed).

CCRM: The percentage of abnormal scores in the TD group were relatively low ranging between 0 to 8.\7% (mean = 4.3%) and were at 0% across all conditions when TD observations that were trimmed as part of the z-score calculation procedure were accounted for.

Why there was no interaction between Group x Condition x Material? Discussion or here? The lack of significant interaction (Group x Condition or Group x Condition x Material), is somewhat surprising and do not reflect some of the differences seen in Figure 3.7 A-B between the two groups in some conditions or the overall difference in performance between the speech materials and may suggest that the model was under-powered to test these questions.

Points for age effect:

- Goldsworthy et al. 2018 found that age explained only a small portion of variability in speech perception performance (n.s.) for Quiet, SSN and 2-talker connected-speech distractors (children aged 5-17). See table 3.

Points for SSN:

- "Despite mature peripheral encoding, school-children have more difficulty understanding speech in noise compared with adults. For example, 5-7 year-old children require 3 to 6 dB more favourable SNR than adults to achieve comparable speech detection, word identification, or sentence recognition performance in a speech-shaped noise maker (e.g., Corbin et al., 2016)" [Leibold, Buss and Calandruccio, 2019, Acoustics today]. - "Speech recognition gradually improves until 9-10 years of age , after which mature performance is generally observed" [Leibold, Buss and Calandruccio, 2019, Acoustics today].

- SSN age effect in other studies are smaller

3.5.3 CCC-2

3.5.4 ECLiPS

Discussion: Correlation with CCC-2 sub-scales (Barry & Moore, 2014): Overall, all the EClIPS sub scales shows strong correlation with most of the CCRM 10 sub-scales. Interestingly, PSS strongly correlates with all 10 CCC-2 sub-scales, suggesting that both tests taps into similar abilities.

In the results: compare scores with scores obtained by: https://www.nature.com/articles/s41598-018-25316-9.pdf and Moore et al. 2020 (Listening Difficulties

in Children: Behaviour and Brain Activation Produced by Dichotic Listening of CV Syllables)

Discussion: - Compare data with Ferguson et al. 2011

3.6 Conclusion

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe (von Goethe, 1829)

General discussion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Summary of main findings

Conclusion

Appendices

A

The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

```
library(tidyverse)

knitr::include_graphics("figures/chunk-parts.png")
```

And here's another one from the same chapter, i.e. Chapter ??:

B

The Second Appendix

C
The Third Appendix

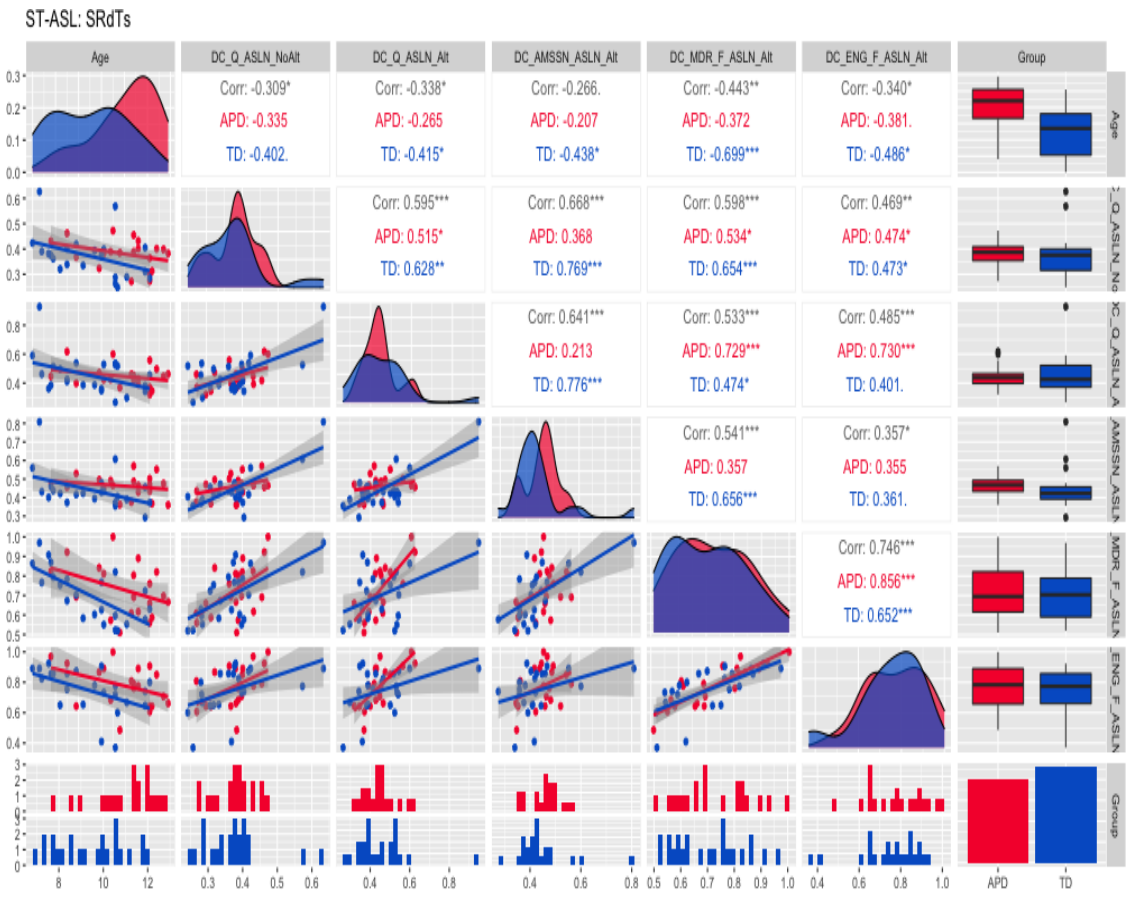

Figure C.1: Switching task: ASL speech material - correlations for listeners SRdTs (proportion of duty cycle).

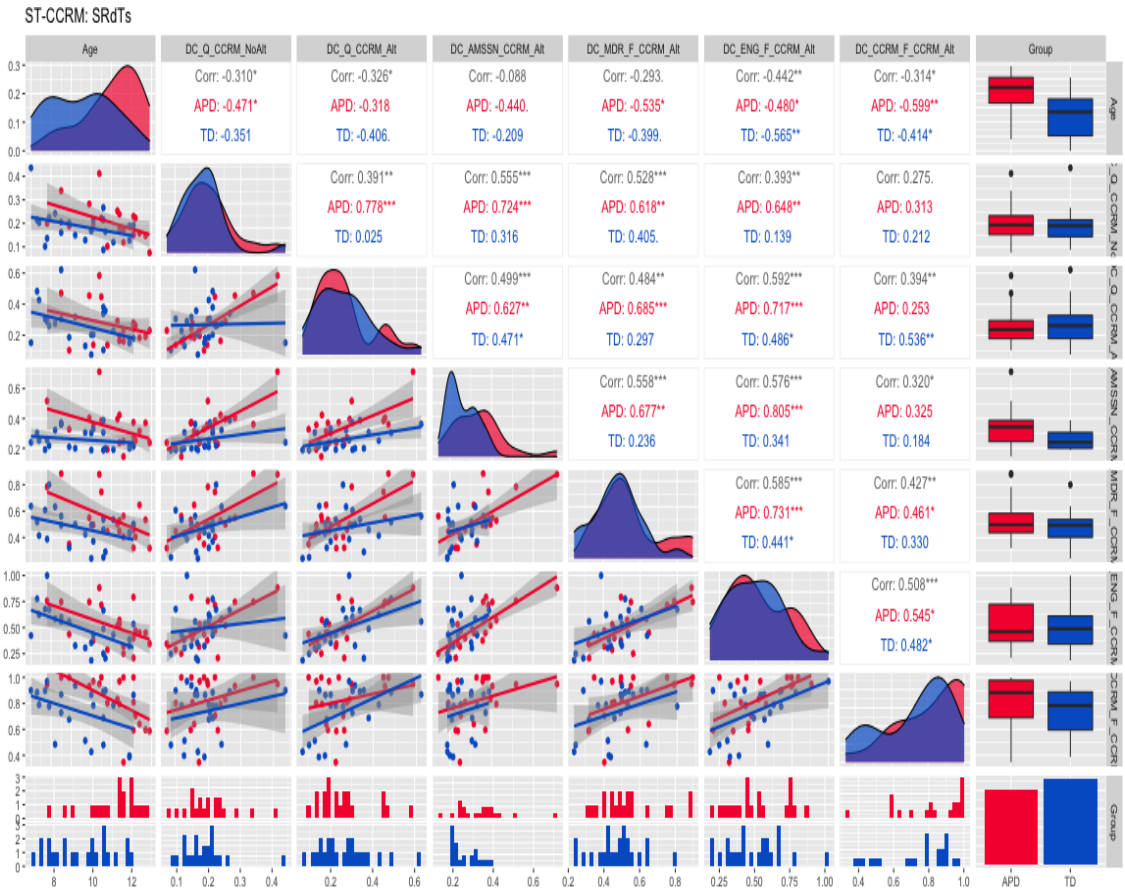

Figure C.2: Switching task: CCRM speech material - correlations for listeners SRdTs (proportion of duty cycle).

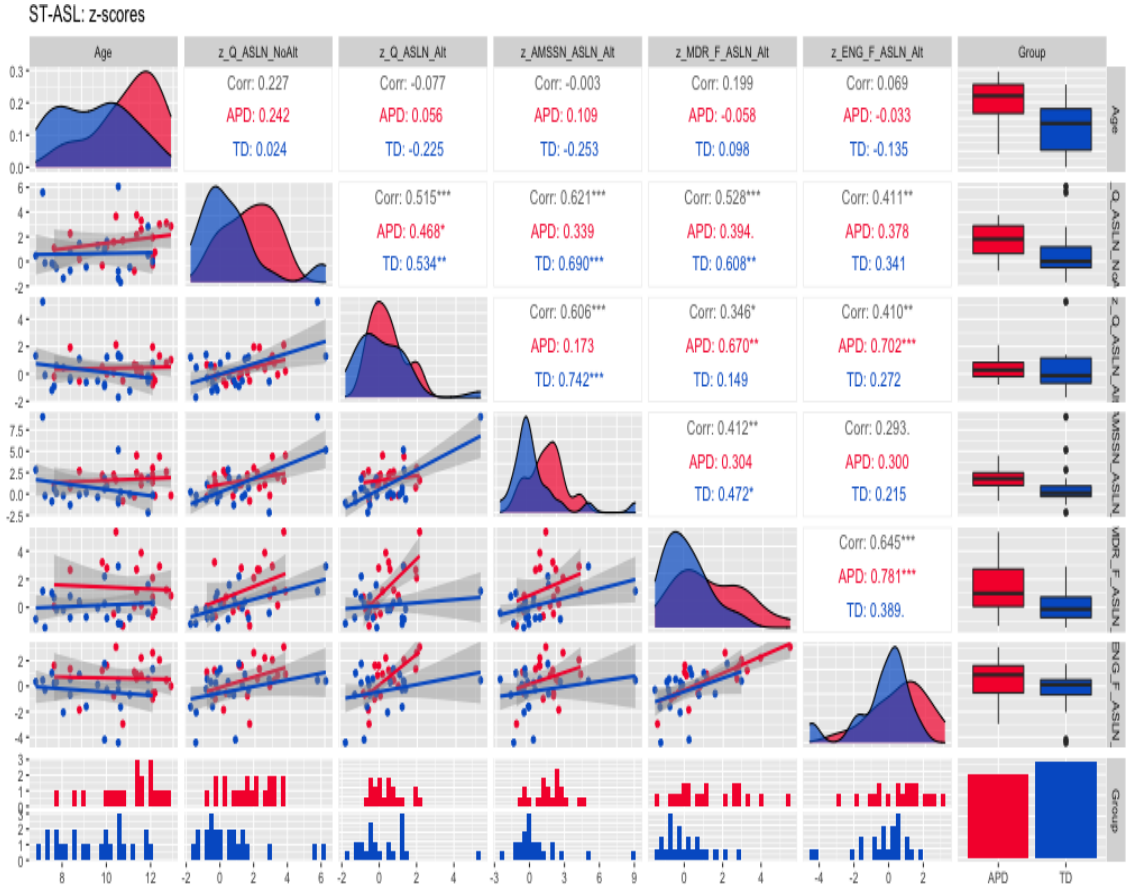

Figure C.3: Switching task: ASL speech material - correlations for listeners z-scores.

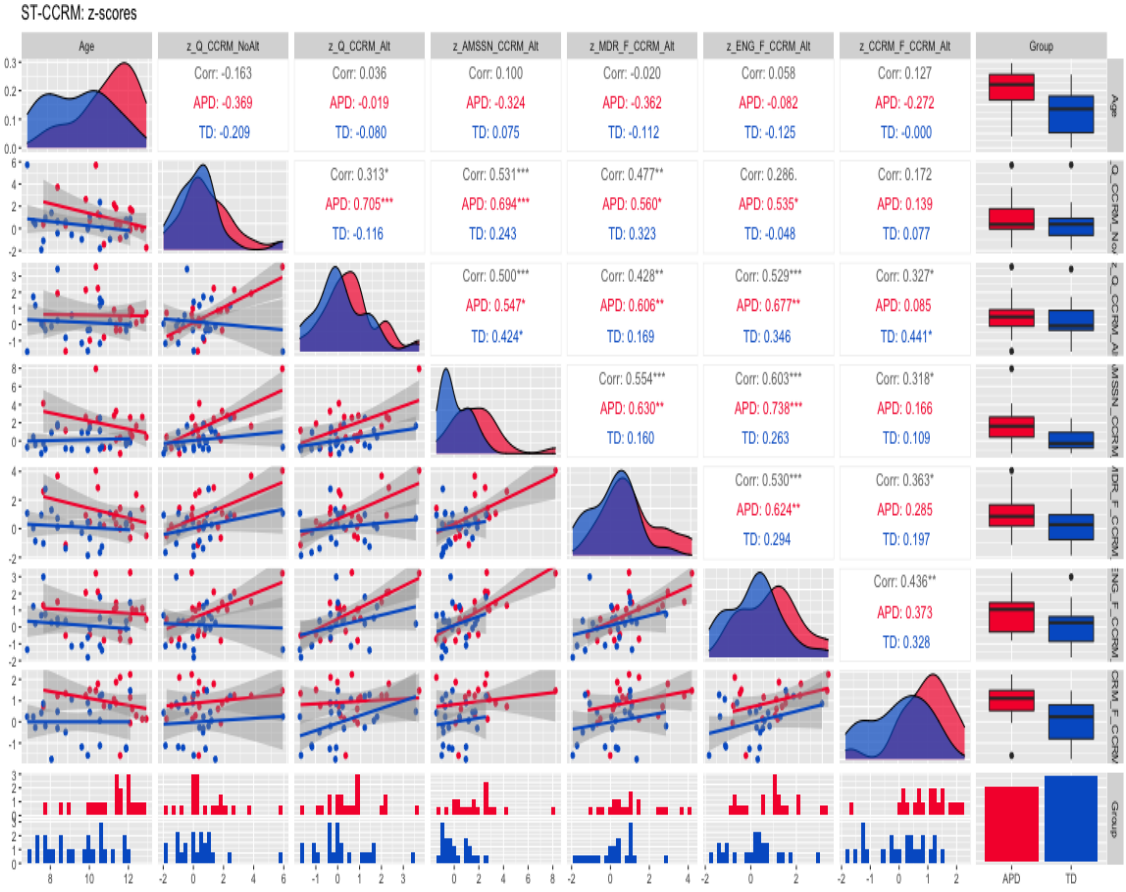

Figure C.4: Switching task: CCRM speech material - correlations for listeners z-scores.

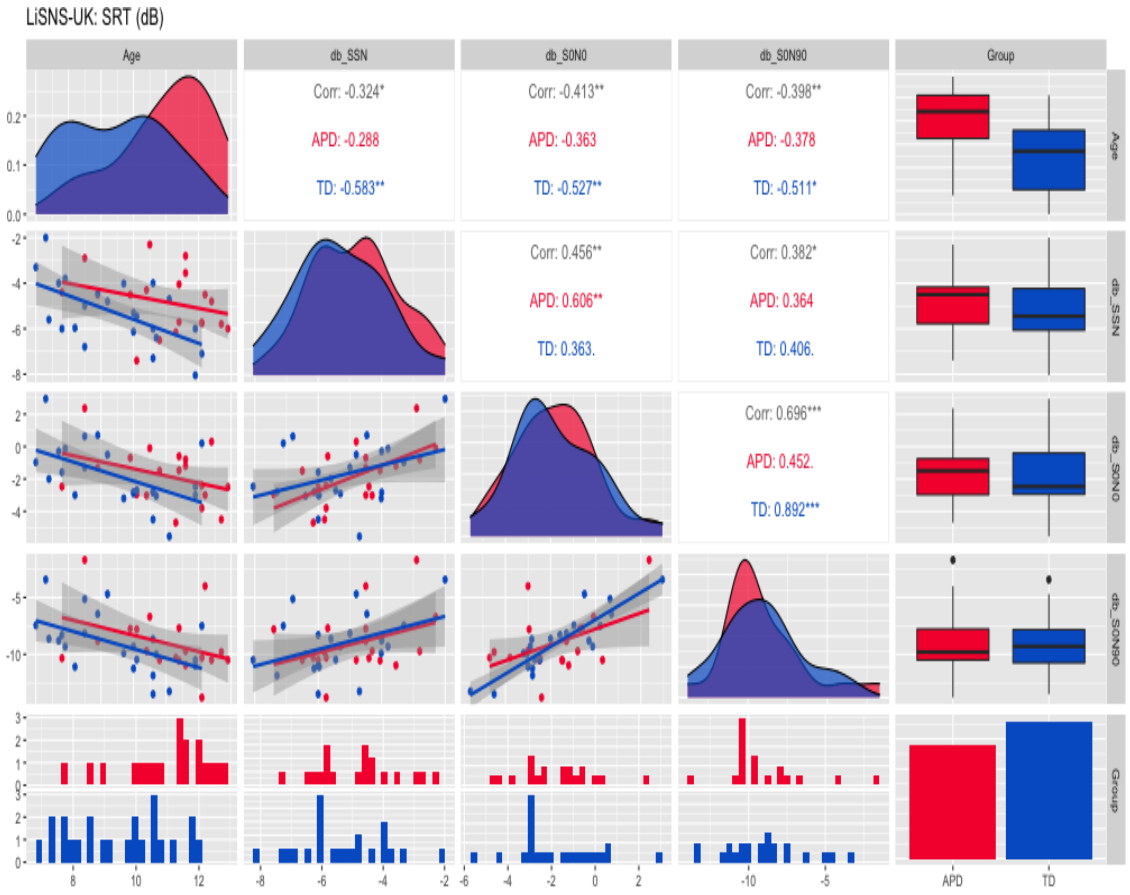

Figure C.5: LiSNS-UK: Correlations for listeners SRTs (dB SNR).

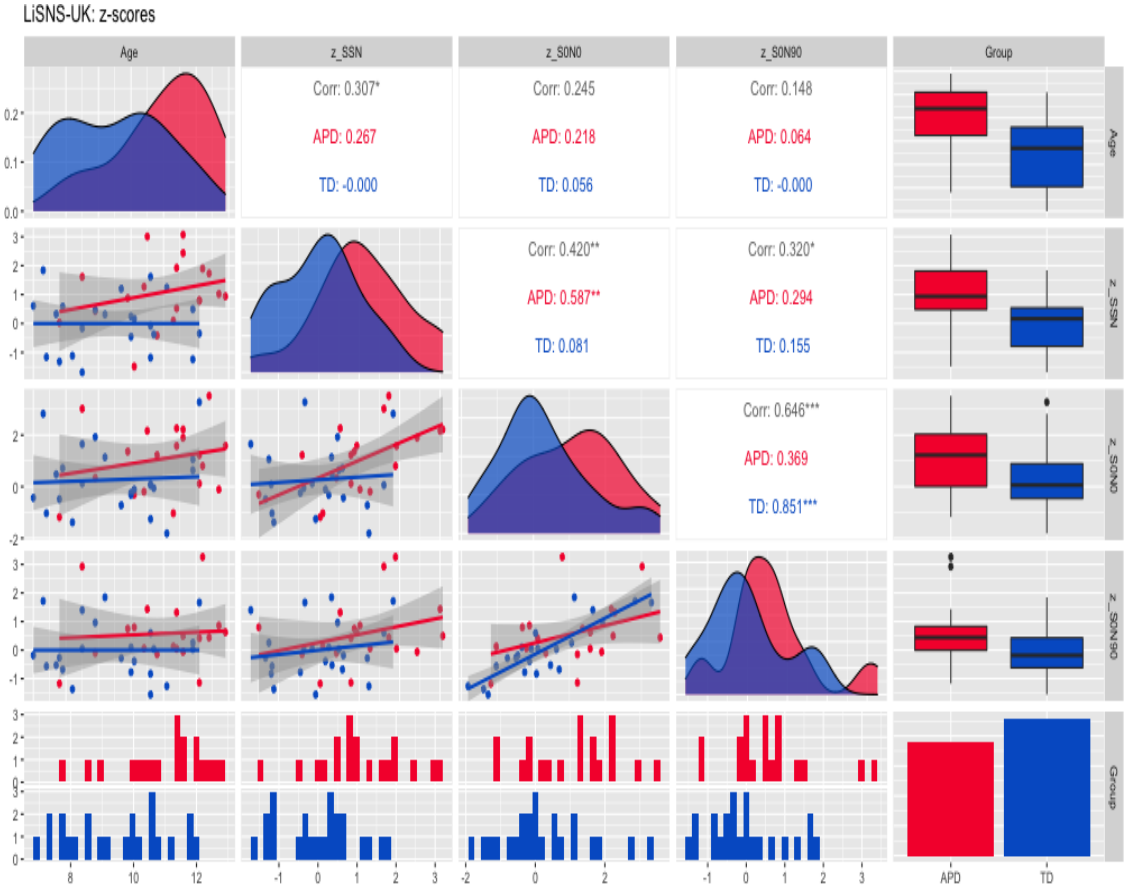

Figure C.6: LiSNS-UK: Correlations for listeners age-independent z-scores.

References

Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging. (1990). *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.399731

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(SUPPL. 2).

Akinseye, G. (2015). *The perception of interrupted and speech in older and younger adults with normal hearing.* (unpublished BSc thesis). University College London, UCL.

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. https://doi.org/10.1109/aspaa.2001.969552

Arlinger, S., Lunner, T., Lyxell, B., & Kathleen Pichora-Fuller, M. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology*, *50*(5), 371–384. https://doi.org/10.1111/j.1467-9450.2009.00753.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2012.11.001

Barry, J. G., & Moore, D. R. (2014). *Evaluation of Children's Listening and Processing Skills (ECLiPS)* (tech. rep.). MRC-T. London, United Kingdom.

Bashford, J. A., Riener, K. R., & Warren, R. M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Perception & Psychophysics*, *51*(3), 211–217. https://doi.org/10.3758/BF03212247

Başkent, D., Clarke, J., Pals, C., Benard, M. R., Bhargava, P., Saija, J., Sarampalis, A., Wagner, A., & Gaudrain, E. (2016). Cognitive Compensation of Speech Perception With Hearing Impairment, Cochlear Implants, and Aging: How and to What Degree Can It Be Achieved? *Trends in Hearing*, *20*. https://doi.org/10.1177/2331216516670279

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *67*(1). https://doi.org/10.18637/jss.v067.i01

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Bergman, A. S. (1990). *Auditory scene analysis : the perceptual organization of sound.* Cambridge, Massachusetts : The MIT Press
Includes bibliographical references (pages 737-761) and index. Includes bibliographical references and index.

Bergman, Blumenfeld, Cascardo, Dash, Levitt, & Margulies. (1976). Age-Related Decrement in Hearing for Speech. *Journal of Gerontology*, *31*(5), 533–538.

Bergman, M. (1980). *Aging and the perception of speech*. University Park Press.

Best, V., Mason, C. R., & Kidd, G. (2011). Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers. *The Journal of the Acoustical Society of America*, *129*(3), 1616–1625. https://doi.org/10.1121/1.3533733

Binns, C., & Culling, J. F. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *The Journal of the Acoustical Society of America*, *122*(3), 1765–1776. https://doi.org/10.1121/1.2751394

Bishop, D. V. M. (2003). *The Children's Communication Checklist, Version 2 (CCC-2)* (tech. rep.). The Psycho- logical Corporation. London, United Kingdom.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

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*, *107*(2), 1065–1066. https://doi.org/10.1121/1.428288

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneons voices. *Journal of Phonetics*, *10*, 23–36.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, *77*, 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.3675943

Brungart, D., Iyer, N., Thompson, E. R., Simpson, B. D., Gordon-Salant, S., Schurman, J., Vogel, C., & Grant, K. (2013). Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. *Proceedings of Meetings on Acoustics*, *19*(1), 60146. https://doi.org/10.1121/1.4800033 doi: 10.1121/1.4800033

Brungart, D. S., & Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America*, *132*(4), 2545–2556. https://doi.org/10.1121/1.4747005

Brungart, D. S., & Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *The Journal of the Acoustical Society of America*, *112*(6), 2985–2995. https://doi.org/10.1121/1.1512703

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*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Buss, E., Whittle, L. N., Grose, J. H., & Hall, J. W. (2009). Masking release for words in amplitude-modulated noise as a function of modulation rate and task. *The Journal of the Acoustical Society of America*, *126*(1), 269–280. https://doi.org/10.1121/1.3129506

Calandruccio, L., Bradlow, A. R., & Dhar, S. (2014). Speech-on-speech masking with variable access to the linguistic content of the masker speech for native and

nonnative English speakers. *Journal of the American Academy of Audiology*. https://doi.org/10.3766/jaaa.25.4.7

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.3458857

Cameron, S., & Dillon, H. (2007). Development of the Listening in Spatialized Noise-Sentences Test (LISN-S). *Ear and Hearing*, *28*(2), 196–211. https://doi.org/10.1097/AUD.0b013e318031267f

Cameron, S., Glyde, H., & Dillon, H. (2011). Listening in Spatialized Noise—Sentences Test (LiSN-S): Normative and Retest Reliability Data for Adolescents and Adults up to 60 Years of Age. *Journal of the American Academy of Audiology*, *22*(10), 697–709. https://doi.org/10.3766/jaaa.22.10.7

Carlile, S., & Corkhill, C. (2015). Selective spatial attention modulates bottom-up informational masking of speech. *Scientific Reports*. https://doi.org/10.1038/srep08662

Chan, D., Fourcin, A., Gibbon, D., Grandstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., in'T Veld, C., & Zeiliger, J. (1995). EUROM - A spoken language resource for the EU. *European Conference on Speech Communication and Technology*.

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

Cherry, E. C., & Taylor, W. K. (1954). Some Further Experiments upon the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *26*(4), 554–559. https://doi.org/10.1121/1.1907373

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry and Allied Disciplines*. https://doi.org/10.1111/1469-7610.00770

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913–2922. https://doi.org/10.1121/1.1616924

Drullman, R., & Bronkhorst, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America*, *107*(4), 2224–2235. https://doi.org/10.1121/1.428503

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *114*(1), 368–379. https://doi.org/10.1121/1.1577562

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

Feys, J. (2015). *Npintfactrep: Nonparametric interaction tests for factorial designs with repeated measures* [R package version 1.5]. https://CRAN.R-project.org/package=npIntFactRep

Feys, J. (2016). Nonparametric tests for the interaction in two-way factorial designs using R. *R Journal*. https://doi.org/10.32614/rj-2016-027

Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R - 17 Exploratory factor analysis. *Discovering statistics using r*.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second). Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

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

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588. https://doi.org/10.1121/1.428211

Gamer, M., Lemon, J., & <puspendra.pusp22@gmail.com>, I. F. P. S. (2019). *Irr: Various coefficients of interrater reliability and agreement* [R package version 0.84.1]. https://CRAN.R-project.org/package=irr

Goodman, A. S. (n.d.). Auditory research lab audio software (arlas). version 0.20.2, data 2017-04-11. [Accessed: 02-01-2021]. https://github.com/myKungFu/ARLas

Green, T., & Rosen, S. (2013). Phase effects on the masking of speech by harmonic complexes: Variations with level. *The Journal of the Acoustical Society of America*, *134*(4), 2876–2883. https://doi.org/10.1121/1.4820899

Grose, J. H., Porter, H. L., & Buss, E. (2016). Aging and Spectro-Temporal Integration of Speech. *Trends in Hearing*, *20*, 1–11. https://doi.org/10.1177/2331216516670388

Harrell Jr, F. E. (2020). *Hmisc: Harrell miscellaneous* [R package version 4.4-2]. https://CRAN.R-project.org/package=Hmisc

Hirsh, I. J. (1950). The Relation between Localization and Intelligibility. *The Journal of the Acoustical Society of America*, *22*(2), 196–200. https://doi.org/10.1121/1.1906588

Hoffman, I., & Levitt, H. (1978). A note on simultaneous and interleaved masking. https://doi.org/10.1016/0021-9924(78)90013-8

Hopkins, K., & Moore, B. C. J. (2010). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, *127*(3), 1595–1608. https://doi.org/10.1121/1.3293003

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2006). A Lego system for conditional inference. *The American Statistician*, *60*(3), 257–263. https://doi.org/10.1198/000313006X118430

Howard-Jones, P., & Rosen, S. (1993). The perception of speech in fluctuating noise. *Acta Acustica united with Acustica*, *78*(5), 258–272.

Huang, H. W. (2018). *The Effects of Different Types of Contralateral Distractors on Switching Attention for Speech in Elder & Younger Adults with Normal Hearing* (Master's thesis). University College London, UCL.

Huckvale, M. (2013). Speech filing system tools, sfswin (version: 1.9, data: 2013-04-18). https://www.phon.ucl.ac.uk/resource/sfs/

Huggins, A. W. F. (1964). Distortion of the Temporal Pattern of Speech: Interruption and Alternation. *The Journal of the Acoustical Society of America*, *36*(6), 1055–1064. https://doi.org/10.1121/1.1919151

Humes, L. E., & Dubno, J. R. (2010). Factors affecting speech understanding in older adults. In S. Gordon-Salant, R. Frisina, R. Fay, & A. Popper (Eds.), *The aging auditory system*. Springer-Verlag New York.

Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience*, *7*(October), 1–16. https://doi.org/10.3389/fnsys.2013.00055

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Vol. 103). Springer.

Kassambara, A. (2021). *Rstatix: Pipe-friendly framework for basic statistical tests* [R package version 0.6.0.999]. https://rpkgs.datanovia.com/rstatix/

Kidd, G. R., & Humes, L. E. (2012). Effects of age and hearing loss on the recognition of interrupted words in isolation and in sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1434–1448. https://doi.org/10.1121/1.3675975

Kidd, G., Mason, C. R., & Arbogast, T. L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, *111*(3), 1367–1376. https://doi.org/10.1121/1.1448342

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. https://doi.org/10.3109/14992027.2015.1020971

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research (2016/03/31). *Journal of chiropractic medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krishnan, S., Leech, R., Aydelott, J., & Dick, F. (2013). School-age children's environmental object identification in natural auditory scenes: Effects of masking and contextual congruence. *Hearing Research*, *300*, 46–55. https://doi.org/10.1016/j.heares.2013.03.003

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of speech, language, and hearing research : JSLHR*, *42*(5), 1148–1156. https://doi.org/10.1044/jslhr.4205.1148

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Leclère, T., Lavandier, M., & Deroche, M. L. (2017). The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location. *Hearing Research*, *350*, 1–10. https://doi.org/10.1016/j.heares.2017.03.012

Lee, J., Dhar, S., Abel, R., Banakis, R., Grolley, E., Lee, J., Zecker, S., & Siegel, J. (2012). Behavioral Hearing Thresholds between 0.125 and 20 kHz Using

Depth-Compensated Ear Simulator Calibration. *Ear and Hearing*. https://doi.org/10.1097/AUD.0b013e31823d7917

Leech, R., Gygi, B., Aydelott, J., & Dick, F. (2009). Informational factors in identifying environmental sounds in natural auditory scenes. *The Journal of the Acoustical Society of America*, *126*(6), 3147–3155. https://doi.org/10.1121/1.3238160

Leensen, M. C., & Dreschler, W. A. (2013). The applicability of a speech-in-noise screening test in occupational hearing conservation. *International Journal of Audiology*. https://doi.org/10.3109/14992027.2013.790565

Lenth, R. V. (2016). Least-Squares Means: The {R} Package {lsmeans}. *Journal of Statistical Software*, *69*, 1–33. https://doi.org/10.18637/jss.v069.i01

Lenth, R. V. (2020). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.5.3]. https://CRAN.R-project.org/package=emmeans

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Lewis, J. D., McCreery, R. W., Neely, S. T., & Stelmachowicz, P. G. (2009). Comparison of in-situ calibration methods for quantifying input to the middle ear. *The Journal of the Acoustical Society of America*, *126*(6), 3114–3124.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43. https://doi.org/10.3109/03005369009077840

Mair, K. R. (2013). *Speech Perception in Autism Spectrum Disorder: Susceptibility to Masking and Interference* (PhD dissertation March). University College London, UCL.

Mair, P., & Wilcox, R. (2020). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, *52*, 464–488.

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*. https://doi.org/10.7717/peerj.6918

McDonald, J. (2014). Multiple comparisons. *Handbook of biological statistics* (3rd ed., pp. 254–260). Sparky House Publishing.

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

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America*, *128*(1), 435–443. https://doi.org/10.1121/1.3397384

Moore, B. (2008). The role of temporal fine structure in normal and impaired hearing. *Auditory Signal Processing in Hearing-Impaired Listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007)*, (Isaar), 247–262.

Moore, B. C. J. (2012). *An introduction to the psychology of hearing* (6th ed.). Bingley : Emerald
Includes bibliographical references and index.

Moore, B. C. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, *31*(3-4), 563–574. https://doi.org/10.1016/S0095-4470(03)00011-1

Moore, D. R., Ferguson, M. A., Edmondson-Jones, A. M., Ratib, S., & Riley, A. (2010). Nature of Auditory Processing Disorder in Children. *PEDIATRICS*, *126*(2), e382–e390.

Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60.

Murphy, C. F., Hashim, E., Dillon, H., & Bamiou, D. E. (2019). British children's performance on the listening in spatialised noise-sentences test (LISN-S). *International Journal of Audiology*. https://doi.org/10.1080/14992027.2019.1627592

Nasreddine, Z., Phillips, N., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J., & Chertkow, H. (2005). The Montreal Cognitive Assessment , MoCA : A Brief Screening. *Journal of the American Geriatric Society*, *53*, 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nelson, P. B., & Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *115*(5), 2286–2294. https://doi.org/10.1121/1.1703538

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, *50*(12), 1–23. http://www.jstatsoft.org/v50/i12/

Norbury, C. F. (2014). Practitioner Review: Social (pragmatic) communication disorder conceptualization, evidence and clinical implications. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. https://doi.org/10.1111/jcpp.12154

Norbury, C. F., & Bishop, D. V. M. (2005). Children ' s Communication Checklist - 2 : a validation study. *Publie dans Revue Tranel*, *42*, 53–63.

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of Age on Auditory and Cognitive Processing: Implications for Hearing Aid Fitting and Audiologic Rehabilitation. *Trends in Amplification*, *10*(1), 29–59. https://doi.org/10.1177/108471380601000103

Pichora-Fuller, M. K., & Souza, P. E. (2003). Effects of aging on auditory processing of speech. *International Journal of Audiology*, *42*(sup2), 11–16. https://doi.org/10.3109/14992020309074638

Qin, S., Nelson, L., McLeod, L., Eremenco, S., & Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-2076-0

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2020a). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2020b). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain*, *126*(4), 841–865. https://doi.org/10.1093/brain/awg076

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.0.12]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.2000751

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America*, *120*(6), 3988–3997. https://doi.org/10.1121/1.2358008

Richmond, S. A., Kopun, J. G., Neely, S. T., Tan, H., & Gorga, M. P. (2011). Distribution of standing-wave errors in real-ear sound-level measurements. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.3569726

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*, *133*(4), 2431–2443. https://doi.org/10.1121/1.4794379

Roucos, S., & Wilgus, A. M. (1986). High quality time-scale modification for speech. *Behavior Research Methods*, *52*, 464–488.

RStudio Team. (2019). *Rstudio: Integrated development environment for r*. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Saija, J. D., Akyürek, E. G., Andringa, T. C., & Başkent, D. (2014). Perceptual restoration of degraded speech is preserved with advancing age. *JARO - Journal of the Association for Research in Otolaryngology*, *15*(1), 139–148. https://doi.org/10.1007/s10162-013-0422-z

Scheffers, M. T. M. (1983). *Sifting vowels. Auditory pitch analysis and sound segregation* (PhD dissertation). University of Groningen.

Schubert, E. D., & Parker, C. D. (1955). Addition to Cherry's findings on switching speech between the two ears. *Journal of the Acoustical Society of America*, *27*, 792–794. https://doi.org/10.1121/1.1908042

Shafiro, V., Sheft, S., & Risley, R. (2011). Perception of interrupted speech: Effects of dual-rate gating on the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *130*(4), 2076–2087. https://doi.org/10.1121/1.3631629

Shafiro, V., Sheft, S., Risley, R., & Gygi, B. (2015). Effects of age and hearing loss on the intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, *137*(2), 745–756. https://doi.org/10.1121/1.4906275

Shen, J., & Souza, P. E. (2017). The Effect of Dynamic Pitch on Speech Recognition in Temporally Modulated Noise. *Journal of Speech Language and Hearing Research*, *60*(September), 2725–2739. https://doi.org/10.1044/2017_JSLHR-H-16-0389

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

Siegel, J. H. (1994). Ear-canal standing waves and high-frequency sound calibration using otoacoustic emission probes. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.409829

Steinmetzger, K., & Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *The Journal of the Acoustical Society of America*, *138*(6), 3586–3599. https://doi.org/10.1121/1.4936945

Stone, M. A., Füllgrabe, C., & Moore, B. C. J. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, *132*(1), 317–326. https://doi.org/10.1121/1.4725766

Stone, M. A., & Moore, B. C. J. (2014). On the near non-existence of "pure" energetic masking release for speech. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.4868392

Stuart, A. (2008). Reception Thresholds for Sentences in Quiet, Continuous Noise, and Interrupted Noise in School-Age Children. *Journal of the American Academy of Audiology*, *19*(2), 135–146. https://doi.org/10.3766/jaaa.19.2.4

Summers, R. J., & Roberts, B. (2020). Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *The Journal of the Acoustical Society of America*, *147*(2), 1113–1125. https://doi.org/10.1121/10.0000688

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America*, *110*(4), 2085–2095. https://doi.org/10.1121/1.1404973

Torchiano, M. (2020). *Effsize: Efficient effect size computation* [R package version 0.8.1]. https://doi.org/10.5281/zenodo.1480624

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.2400666

van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, *38*(4), 584–589. https://doi.org/10.3758/BF03193889

van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., & Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International Journal of Audiology*, *52*(5), 305–321. https://doi.org/10.3109/14992027.2012.759665

von Goethe, J. W. (1829). *Wilhelm Meisters Wanderjahre oder die Entsagenden*. Cotta.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392 LP –393. http://science.sciencemag.org/content/167/3917/392.abstract

Watson, C. S. (1987). Uncertainty, informational masking and the capacity of immediate auditory memory. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 267–277). Hillsdale, N.J. : L. Erlbaum Associates Includes bibliographies and indexes.

Wierstorf, H., Geier, M., Raake, A., & Spors, S. (2011). A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. *AES130*.

Wiig, E., H, Semel, E., & Secord, W. (2017). *Clinical Evaluation of Language Fundamentals - Fifth Edition UK (CELF-5UK)* (tech. rep.). PsychCorp, Pearson Clinical Assessment.

World Health Organisation. (1998). *Occupational exposure to noise: evaluation, prevention and control* (tech. rep.).

Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7–10.