M2 ATIAM - FINAL REPORT

August 21, 2025

# Improving Speech-in-Noise perception by training

**Internship** conducted from 3rd March to 3rd August 2025

**Azal LE BAGOUSSE**

*Under the supervision of **Professor Timothy D Griffiths FMedSci***

*[ tim.griffiths@newcastle.ac.uk ]*

Auditory Cognition Group, Newcastle University Medical School (NE2 4HH, UK)

**Public internship**

# Abstract

This project investigates the effects of training in auditory figure–ground segregation on Speech-in-Noise (SiN) perception and listening effort. Auditory figure–ground, the ability to detect and track coherent patterns of sounds (figures) embedded within a background of random, overlapping elements (ground), is a process used to further dissect the perception of speech in complex acoustic scenes. The training experiment, previously developed by the host laboratory, consists of a pre-training session, a 7-day training period, and a follow-up post-training session to compare performance before and after training. Pupillometry, the measurement of pupil size changes, was used as a physiological indicator of listening effort, based on evidence that pupil dilation increases with cognitive load.

The original 7-day training was delivered through a Matlab standalone application, which in this project was adapted into a web-based version to enable broader accessibility. This involved the development of a new interface, an updated database, and redesigned game characters to improve user engagement. In parallel of the website implementation, data collection was carried out with 12 new participants, bringing the total study sample to 23 participants (11 tested prior to this project).

This work therefore combines behavioural and physiological testing with software development. It illustrates the positive effects of training in auditory grouping on SiN perception and listening effort and the practical implementation of a flexible training platform to further carry out this study.

**Keywords** Speech perception, Speech-in-noise, Training, Auditory figure-ground, Pupillometry, Listening effort

---

*[French]* **Résumé** : Ce projet étudie les effets de l'entraînement à la séparation figure-fond (figure-ground) auditive sur la perception de la parole dans le bruit (Speech-in-Noise, SiN) et sur l'effort d'écoute. La séparation figure-fond auditive désigne la capacité à détecter et à suivre des motifs sonores cohérents (figures) intégrés dans un fond d'éléments sonores aléatoires et superposés (fond). Ces paradigmes sont utilisés pour mieux comprendre et analyser la perception de la parole dans des environnements acoustiques complexes. Le paradigme d'entraînement, développé plus tôt par le laboratoire d'accueil, se compose d'une session pré-entraînement, d'une période d'entraînement de 7 jours, et d'une session post-entraînement permettant de comparer les performances avant et après. La pupillométrie, la mesure des variations de diamètre de la pupille, a été utilisée dans cette étude comme indicateur physiologique de l'effort d'écoute, sur la base de travaux montrant que la dilatation de la pupille augmente avec la charge cognitive.

La version originelle du paradigme d'entraînement sur 7 jours était réalisée avec une application développée sur Matlab. Dans ce projet, l'application a été adaptée en version web afin de la rendre plus accessible. Cette adaptation a nécessité le développement d'une nouvelle interface, d'une nouvelle base de données, ainsi que la création de nouveaux designs pour renforcer l'engagement des utilisateurs. En parallèle de la mise en place du site web, des données ont été recueillies auprès de 12 nouveaux participants, amenant l'échantillon total de l'étude à 23 participants (dont 11 testés auparavant).

Ce travail associe ainsi des mesures comportementales et physiologiques au développement d'un site internet pour des tâches psychoacoustiques. Il met en évidence les effets positifs d'un entraînement sur la perception de la parole dans le bruit et sur l'effort d'écoute, tout en mentionnant le travail effectué au niveau de la plateforme d'entraînement pour élargir l'étendue de cette étude.

**Mots-clé** Compréhension de la parole, Parole dans le bruit, Entraînement, Figure-ground auditives, Pupillométrie, Effort d'écoute

# Acknowledgements

*I would like to express my gratitude to:*

**Prof. Tim Griffiths**, *for his guidance, assistance, and overall support throughout this project;*
**Ester, Hasan, Xiaoxuan, Kate and all the Newcastle group**, *for their valuable assistance and for contributing to a pleasant workplace atmosphere;*
**Claudia, Jakub and all the London group**, *for teaching me how to conduct pupillometry experiments and for their insights on my project;*
**The undergrad interns**, *for enlivening the lab and being the best students I could hope for;*
**The teaching staff (Département Pédagogie) at Ircam**, *for being a reliable pillar throughout this internship in a new country and different environment.*

# Contents

# Chapter 1
# Introduction

This internship was conducted at the Auditory Cognition Group Newcastle laboratory, under the supervision of Professor Timothy D. Griffiths.
The Auditory Cognition Group is a group of 3 laboratories studying normal perception of complex sound relevant to the analysis of speech, music and environmental sounds and the related brain bases [1]. The group is formed of the following entities : the Newcastle lab, the main branch supervised by Prof. Griffiths at the Newcastle University Medical School in Newcastle-upon-Tyne (UK) – the London lab, at the Wellcome Centre for Human NeuroImaging at University College London (UK) – and the Human Brain Research Laboratory at the University of Iowa (US). In this internship, most of the work was carried at the Newcastle lab, however a day was spent in the London lab to observe and carry out diverse experiments.
During this internship, a few projects were studied, but the main project carried out was about training in fundamental sound grouping ability to improve daily life Speech-in-Noise perception.
The internship proceeded as follows:

- Month 1 : Exploring several ongoing projects to understand the different research directions and determine the main focus for this internship. Different projects like the study of boundary cells [1] during auditory transitions (like a glide) as well as the study of microsaccades [2] during speech-in-noise listening (see Section 2.1) and even pupillometry [3] experiments were explored. During this month, a work trip to the Ear Institute (UCLondon, UK) was planned to observe and learn how to conduct a pupillometry experiment with speech-in-noise stimuli. Literature [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18] covering multiple topics was read to gain an overview of the fields explored across these projects.

- Month 2 : Writing a literature review about pupillometry and microsaccades and defining the internship project. The study for this internship was selected based on the lab's needs within the scope of ongoing research. Scripts were reviewed and studied, training to administer the experiment was completed, and recruitment adverts were posted. The objectives were defined as testing participants and coding to make the study more accessible.

- Months 3-5 : Contacting potential participants, testing individuals, coding an app (Matlab standalone app, the MacOS version) and a website version of the app for the training paradigm, training undergrad interns to enable them to carry out the study, taking part in other lab experiments...

[1] Neurons found in the hippocampal formation that respond to the presence of an environmental boundary at a particular distance and direction from an animal [*Wikipedia*]

[2] Tiny, involuntary eye twitches occurring several times per second, signalling rapid shifts of attention

[3] The study of pupil dilation during effortful listening

# Chapter 2
# State of the art

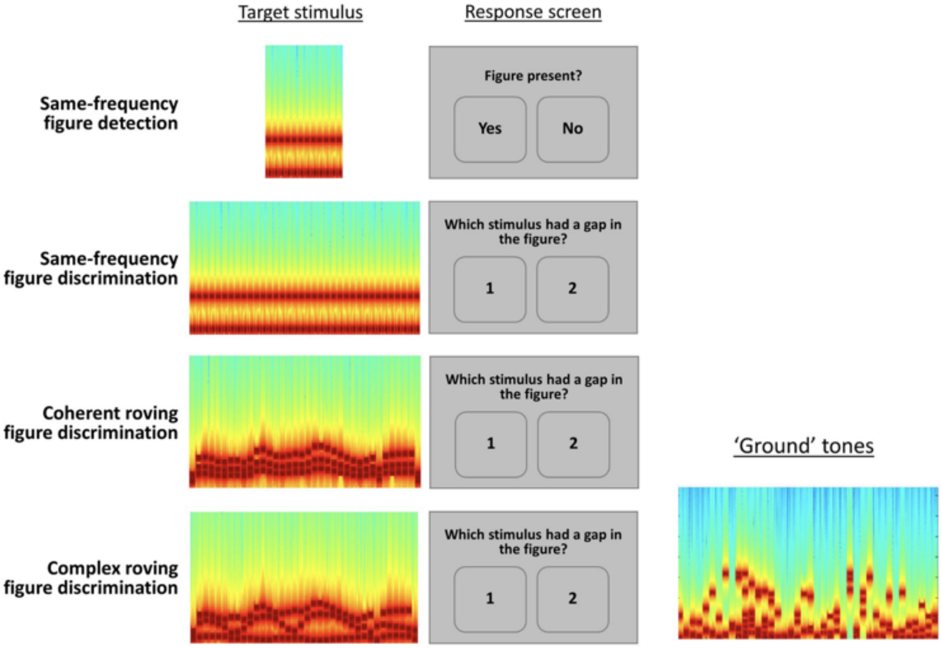
## 2.1 Speech-in-Noise

*Listening is an active process that involves much more than just our ears. It is fundamental for researchers to invetigate the mental gymnastics we do every time we listen to speech.*
Every day, people carry on conversations while surrounded by the noise and babble of others talking. Whether you're listening to the server announce the specials in a busy restaurant, exchanging a few words with a seatmate during the morning commute, or trying to follow your friend's story at a crowded party, you cannot escape what is called the cocktail-party problem. This is the challenge of focusing on a single voice amid background noise, making Speech-in-Noise (SiN) perception one of the most essential functions of human hearing.

Hearing aid and cochlear implant users often experience difficulties with speech-in-noise (SiN) perception. Even normal-hearing individuals may struggle with SiN, as it cannot be fully explained by peripheral hearing sensitivity alone [[5]]. Instead, it also relies on central auditory processing and higher-level cortical mechanisms for grouping and attention [2, 5]. The brain has to continuously isolate a relevant speech signal from overlapping sound sources, with different parts of the auditory system contributing to this process. Traditional hearing assessments, such as the pure-tone audiogram (PTA, see Fig.A.1-2), fail to capture the complexity of these central auditory abilities, often underestimating the difficulties individuals may face in noisy environments, even when they appear to have normal hearing in everyday situations.

To tackle the issue of SiN perception and investigate auditory segregation, Teki et al. (2013, [7]) synthesized a new kind of stimulus to examine the detection of coherent patterns ('figures') within a tone cloud, with randomised frequency and time ('background'). This design was based on the hypothesis that automatic auditory segregation mechanisms exist and are highly sensitive to temporal and spectral correlations. The data from this article support the role of temporal coherence as an organizational principle underlying auditory segregation. The groundwork for this study came from Teki et al. (2011, [6])'s paper about the distinct neural substrates of duration-based and beat-based auditory timing. It demonstrated that the human brain employs distinct timing mechanisms for processing absolute (interval-based) and relative (beat-based) durations. These findings implied that the brain integrates information across time depending on temporal structure of an auditory stimulus, which is important to integrate in order to understand segregation of coherent sources from a noisy background. Building on this, the 2013 article [7] introduced the Stochastic Figure-Ground (SFG) task. In this paradigm, listeners are presented with sequences of short tone "beeps" coherent across time, within a ground. A few years later, Teki al. (2016, [19]) defined the concept of the Static Auditory Figure Ground (Static AFG/SFG), based on their 2013 findings. These paradigms, using stimuli with a static 'figure' and a 'ground' as mentioned prior, offer a non-verbal, non-semantic, and linguistically neutral way to assess central auditory grouping abilities and are useful for identifying SiN difficulties that are not caused by peripheral hearing loss alone.
The relevance of AFG to shape speech perception became clearer with the work of Holmes and Griffiths (2019, [5]), who used modified AFG tasks to investigate individual differences in SiN performance. Their study introduced variants of the original figure-ground paradigm, including same-frequency

figure detection (figures with fixed tones), roving-frequency detection (figures with frequency-changing tones), and figure gap detection, which required listeners to detect brief silence moments (gaps) within the figure stream.


Target stimulus
Response screen
Same-frequency figure detection
Figure present?
Yes
No
Same-frequency figure discrimination
Which stimulus had a gap in the figure?
1
2
Coherent roving figure discrimination
Which stimulus had a gap in the figure?
1
2
Complex roving figure discrimination
Which stimulus had a gap in the figure?
1
2
'Ground' tones

AFG Figure design

AFG Ground design

Figure 2.1: *From Holmes 2019* - AFG paradigms

These tasks were designed to simulate the perceptual demands of real-world SiN listening, where the listener has to group and track acoustic elements amid background noise. Furthermore, according to collected data, different figure-ground tasks explained different variance in speech-in-noise, even after accounting for audiometric thresholds [5]. This shows that an AFG paradigm can be used as a measure of sound grouping to assess SiN perception.
Further literature has introduced pitch-dynamic AFG tasks as additional tools for investigating auditory segregation on the sentence-level [2, 3]. As mentioned earlier, in static AFG tasks, the figure consists of pure-tone components at fixed frequencies that repeat over time within a background of random varying tones. On the other hand, dynamic AFG tasks involve figures with varying frequency patterns, following the pitch contour of natural speech. For example, one task uses a figure whose frequency follows the pitch contour of real sentences (the fundamental frequency (F0) with a harmonic structure), embedded in a background of randomly varying tones. Listeners can be asked to compare the figure patterns across different AFG stimuli, requiring a continuous tracking of the target in a complex noisy background. These dynamic tasks aim to simulate more realistic listening conditions and have been shown to predict SiN perception more strongly (showing better performance) and in addition to static AFG tasks [2]. Finally, both static and dynamic tasks are free from semantic and linguistic content, making them useful to test central auditory processing abilities across diverse populations.

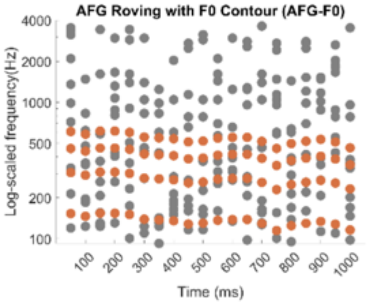
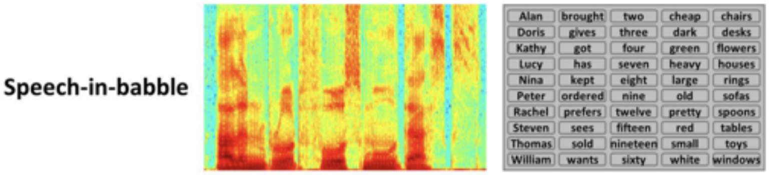


Figure 2.2: *From Guo 2025* - Pitch-dynamic AFG

## 2.2 Eye movements

Speech-in-Noise (SiN) comprehension is usually assessed using behavioural measures like accuracy or threshold scores in sentence- and word-in-babble tests (SiB and WiB tests). Sentence- and word-in-babble tests are tests using different voices in the background talking at the same time (babble), with a unique voice saying a target sentence or word to track amid the background. The participants have to select the words that make up the sentence (for SiB) using a word matrix or the 1 word pronounced (for WiB).

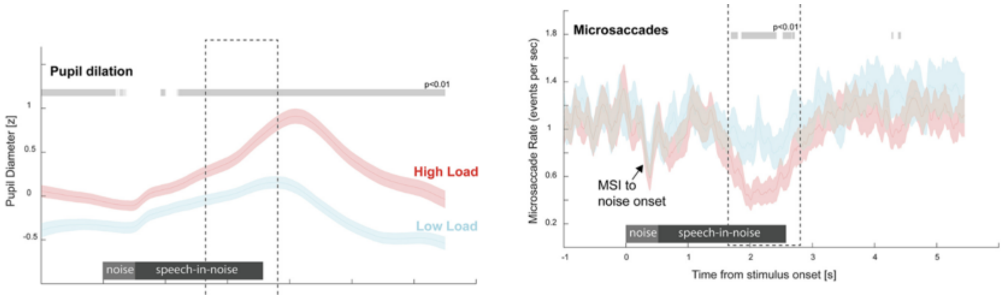


| | | | | |
|---|---|---|---|---|
| Alan | brought | two | cheap | chairs |
| Doris | gives | three | dark | desks |
| Kathy | got | four | green | flowers |
| Lucy | has | seven | heavy | houses |
| Nina | kept | eight | large | rings |
| Peter | ordered | nine | old | sofas |
| Rachel | prefers | twelve | pretty | spoons |
| Steven | sees | fifteen | red | tables |
| Thomas | sold | nineteen | small | toys |
| William | wants | sixty | white | windows |

Figure 2.3: *From Holmes 2019* - SiB test

While the behavioural results offer great insight into speech perception, they mainly reflect final performance and don't capture the full cognitive effort involved.
To gain a deeper understanding of SiN, researchers can also explore physiological measures of listening effort. This project focuses in particular on a specific metric: pupil dilation (PD). Pupil dilation increases in proportion to listening effort, reflecting overall arousal and cognitive load. Pupillometry is thus widely talked about in the literature as a reliable indicator of listening effort, reflecting cognitive resource allocation during auditory tasks and being sensitive to variations in intelligibility, linguistic complexity, and memory load [10, 14, 17, 13, 20, 4, 9]. However, it is not specific when it comes to distinguishing between arousal and attention, as it reflects a combined measure of these 2 components. As a result, it may not clearly isolate attention-specific effects from broader changes in arousal. Additionally, the accuracy and stability of pupil-based measures have been questioned due to individual variability, such as differences in cognitive resource allocation, fatigue, and sensitivity to environmental factors like lighting conditions [14, 9].
Thus, microsaccade rate (MS rate) has been proposed as a more specific marker of auditory attention in the literature. Microsaccades are tiny, involuntary eye twitches occurring several times per second. They briefly slow down when listeners hear new or difficult sounds, signaling rapid shifts of attention. This phenomenon is known as microsaccadic inhibition (MSI). [4, 11, 18]. MSI provides higher temporal precision than pupillary responses [4, 17]. It is also particularly pronounced when attention is actively directed toward auditory stimuli [11], making MS rate a reliable and important tool for tracking engagement during SiN tasks. This metric was originally meant to be analysed in this project but

will be in future works.



Pupil dilation

Microsaccade inhibition

Figure 2.4: *From Contadini-Wright 2023* - Pupillometry and microsaccades study. Pupil size shrinks with lower listening effort and MSI is stronger with higher effort.

In addition to these phasic responses, research also mentions broader eye-movement patterns, such as how long listeners fixate on a single point, how widely their gaze drifts and how slow their blinking becomes, changing as speech becomes harder to follow [18, 9, 8].

Together, pupil dilation (PD), microsaccade (MS) rate, and other eye-movement metrics reveal the real-time allocation of cognitive resources by the brain during listening, offer sensitive measures of listening effort beyond traditional tests, and can be used to improve hearing aid designs for example. These measures are also relevant in the context of SiN perception, as listening effort is expected to decrease when SiN perception is improved. As participants become more efficient at segregating speech from noise, their listening effort is expected to decrease, which should be reflected in a reduction in pupil dilation and a more stable microsaccade rate (i.e. fewer or less pronounced MSI events). These changes, when observed at the same time as improved behavioural performance, can provide stronger evidence that comprehension has improved on the perceptual level and in terms of reduced cognitive load during listening. Studying MS rate alongside PD measures and behavioural performance thus can help reveal whether listeners are processing speech more efficiently and if improvement in SiN comprehension can be noticed.

## 2.3 Training

### 2.3.1 Paradigm

Building on the evidence that central auditory grouping mechanisms, particularly those captured by auditory figure-ground (AFG) tasks, play a critical role in speech-in-noise (SiN) perception, the Auditory Cognition Group Newcastle lab has developed a training paradigm designed to directly target this ability, in hopes to improve SiN perception. The goal of the training is to strengthen fundamental sound grouping skills through daily exposure to structured figure-ground discrimination tasks. This approach rests on the hypothesis that enhancing auditory object formation at a basic, non-linguistic level can lead to measurable improvements in SiN comprehension (with behavioural measures) and a reduction in listening effort (with eye tracking measures, here mostly pupil dilation), independently of peripheral hearing sensitivity. The paradigm consists of two complementary AFG tasks: a dynamic AFG pattern discrimination task, where participants identify the odd stimulus out based on differences in frequency contour (the pitch trajectory differs), and a static AFG gap detection task, where participants identify which stimulus contains a brief temporal gap (silence) embedded in the figure. The training paradigm is carried out by participants in daily 20-minute sessions over 7 days. Both tasks in the paradigm use a three-alternative forced-choice design and are presented in an

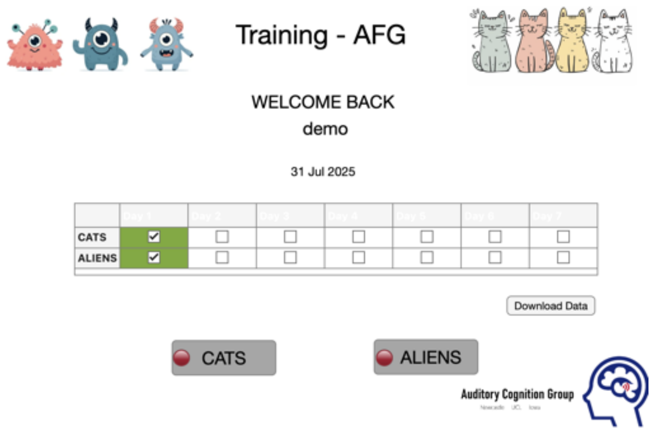
engaging format, with cartoon cats and aliens shown aside the different stimuli played.
The full experimental protocol is split into three phases (see Section 3.1 for details).:

- Pre-training lab session (Day 1): Participants complete 2 SiN tests (sentence- and word-in-babble), 2 AFG assessments (in a two-alternative forced-choice format), and have their pupil responses recorded during the word-in-babble test to assess listening effort. They also complete a first SSQ (Speech, Spatial, and Qualities of Hearing Scale) questionnaire[1].
- Training phase (Day 1-7): Participants use a standalone Matlab application installed on their laptop to complete the 2 AFG tasks for 20 minutes per day, over 7 consecutive days.
- Post-training lab session (Day 8): Participants return for the same battery of tests conducted pre-training, allowing for a direct comparison of behavioural and physiological outcomes. They also complete a general intelligence test (see Section 3.1) and a second SSQ test.

A pilot study involving 11 participants demonstrated significant improvement in sentence-in-noise perception (p[2] = 0.007, Cohen's d[3] = 0.899) and reduced listening effort (p = 0.001, Cohen's d = 1.307), as reflected by pupil dilation. These preliminary results support the study's ongoing development for wider use across diverse populations.

### 2.3.2 App

To implement home-based auditory training, the overseeing research team developed a standalone desktop application in Matlab 2021, referred to here as *the training app*. This tool is the backbone of the experimental training phase mentioned above and is designed to test and improve participants' auditory figure-ground (AFG) abilities using interactive tasks. The app was developed using Matlab App Designer and is structured around a graphical user interface (.mlapp file) that allows the participants to launch the static and dynamic tasks. It shows the participant ID (given by the experimenter), buttons for each task, a download data button and a table corresponding to the days on which the tasks were completed to track progress. For 7 days, daily, the participants do both tasks; once completed, the corresponding boxes will be ticked.



Training - AFG

WELCOME BACK
demo

31 Jul 2025

| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Day 6 | Day 7 |
|---|---|---|---|---|---|---|---|
| CATS | ☑ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| ALIENS | ☑ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

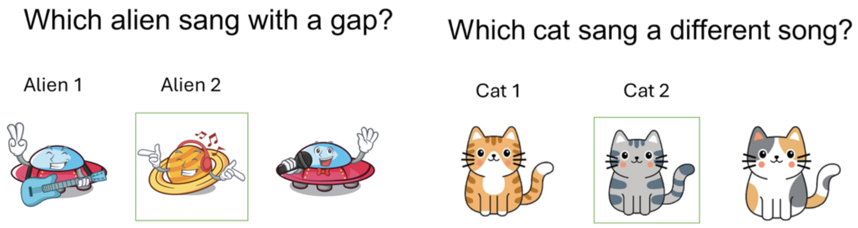
Figure 2.5: Training app interface

[1] The SSQ asks about aspects of one's ability and experience hearing and listening in different situations

[2] The p-value (probability value) shows how likely it is that the observed results could have occurred by chance. Results are considered statistically significant if $p < 0.05$.

[3] Cohen's d is a measure of effect size, it quantifies how big the effect was, in standard deviation units. It shows whether a statistically significant result is also meaningful in size. If $d > 0.8$, it means a large effect.

Two distinct AFG training modules are embedded in the app and are executed through dedicated scripts: trainingCAT.m (dynamic AFG task with cat doodles) and trainingALIEN.m (static AFG gap detection with alien doodles).


Which alien sang with a gap?
Alien 1
Alien 2
Which cat sang a different song?
Cat 1
Cat 2

Alien task - Gap detection

Cat task - Pattern discrimination

Figure 2.6: Training app tasks

Each of these scripts implements a three-alternative forced-choice design in which participants listened to three brief auditory stimuli presented sequentially, and had to choose the one that differs in either temporal structure or frequency contour. The tasks are adaptive in difficulty, with the signal-to-noise ratio (SNR) dynamically adjusted based on participant performance using a staircase procedure. The volume of the figure would go 2dB up (the background stays stable) if the participant got the answer wrong and 2dB down if they got it right, on 5 reversals[4] out of the total 10 for each task. After that, the step size would go to 1dB. The tasks would end after 15 minutes if the 10 reversals weren't achieved, and a break screen appears around half of the task for every participant. Furthermore, both tasks showed feedback to the participant for each trial, so they would know if they got it wrong, right or if they missed (didn't click in time or clicked a wrong button).

Trainingcat.m governs the dynamic pattern discrimination task using cat characters. Participants are presented with 3 figure-ground sequences, 2 of them being identical repeating patterns, and 1 having a distinct frequency contour. The participant must identify the odd one out, the task being 15 minutes long. In this script, stimuli are drawn from pre-generated figure and ground audio files. Trials are randomized after following a certain order, and a fixed inter-stimulus interval ensures consistency. Participants respond using keyboard keys 1, 2, or 3 according to the cat they think "sings a different song". The script handles visual presentation of cartoon cats, sound playback, response logging, reaction time measurement, and adaptive SNR adjustment. The volume for each trial is logged into an excel sheet, as well as the end SNR calculated over the 6 last reversals (originally started at 15dB). This SNR would be the one used for the next task the next day during that 7-day training.

Trainingalien.m governs the static gap detection task using alien characters. Participants are presented with 3 figure-ground sequences, 1 of them having a gap in the figure sequence (selected with keyboard keys 1, 2, or 3). The participant must identify the odd one out. Like the cat task, this task is 15 minutes long, and its stimuli are drawn from pre-generated figure and ground audio files. Gap positions are randomized across trials, and the gap and non-gap figures are embedded in stochastic ground. Like for the cats, the task is also adaptive: the SNR (starting at -10dB) is adjusted in real time depending on whether the participant answers correctly or not. The same results as the cat task are calculated and saved for each day also.

Each daily session outputs the participant ID and timestamp, the task type and SNR trajectory, the trial-by-trial responses, accuracy, and timing and the estimated end SNR. The app is compiled and set through Matlab and can be installed on any Windows operated computer for personal use, even without Matlab installed on the participant's laptop.

[4] A reversal occurs when the participant's response pattern changes direction, making a switch in the adaptive staircase

# Chapter 3
# Methods

## 3.1 Testing

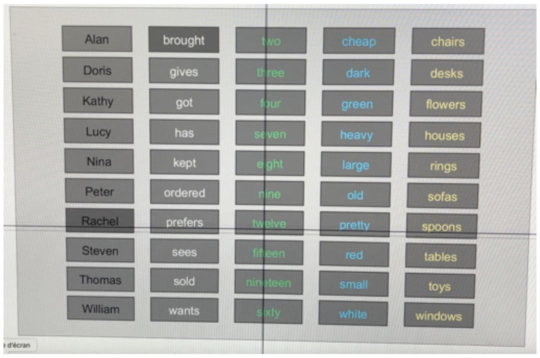
Recruiting participants started in mid April 2025, through different websites and lists like Yammer (the University's platform), Voice (`https://voice-global.org/opportunities/`), and the Newcastle Uni Neuroscience Department's Volunteering list. The inclusion criteria was:

- Being aged between 25 and 70 years old
- Being a native English speaker (English should be the home country's first language)
- Being able to read the screen from 0.5 meters away without glasses (required for eye tracking)
- Having no history of diagnosed hearing disorders (e.g. no worse than mild hearing loss, tinnitus) and diagnosed cognitive disorders, and not taking any psychotropic drugs

An information sheet was sent to each potential participant answering one of the online adverts. Participants were selected according to the inclusion criteria and the limited budget for this experiment, with a preference for ages 40-65 (as SiN difficulty is more common despite normal or near-normal hearing), no contacts (some of the participants' contacts reflected the light too much for the tracker to follow, but some data was still kept even when contacts were worn) and possession of a laptop (lab-owned laptops were handed out if needed but availability was limited). Testing was done in a audiometric booth (soundproof), with a screen on top of a desk 0.5m away from a chair, a keyboard and an eye-tracker right below the screen. The screen was piloted (mirrored) from the exterior of the booth, with an external system (computer, soundcard, keyboard and mouse) (see setup here).

**DAY 1 / PRE-TESTING :** Each participant first completed a 2-hour pre-training session. This session first included a briefing about the study and the tasks to perform, followed by installing the app on the participant's laptop (or lending the lab laptop if declared needed) whilst the participant signed the required forms. The participant then had to fill out an SSQ form, as mentioned in [2.3.1], of 12 questions to subjectively assess their SiN perception (extract shown in Fig.A.5). Finally, followed an audiogram assessment [Fig.A.1–2], which was succeeded by the 4 pre-training tasks (see 2.3.1):

**Task 1** - Sentence-in-babble (SiB) task :

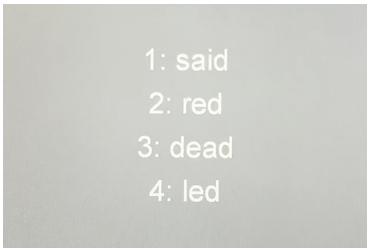


| | | | | |
|---|---|---|---|---|
| Alan | brought | two | cheap | chairs |
| Doris | gives | three | dark | desks |
| Kathy | got | four | green | flowers |
| Lucy | has | seven | heavy | houses |
| Nina | kept | eight | large | rings |
| Peter | ordered | nine | old | sofas |
| Rachel | prefers | twelve | pretty | spoons |
| Steven | sees | fifteen | red | tables |
| Thomas | sold | nineteen | small | toys |
| William | wants | sixty | white | windows |

Figure 3.1: Task 1 (SiB)

During this task, the participant listens to 144 stimuli composed of babble noise (random talking from different voice sources) as the background and one distinctive sentence as the target. The task has an adaptative signal-to-noise Ratio (SNR), which starts at 0dB (target as loud as the babble) and staircases with a 1-up 1-down procedure (target gets 1dB lower if the participant answers correctly, and 1dB higher if their answer is wrong). Sliding the mouse, the participant has to choose by clicking on the words one by one from a $5 * 10$ word-matrix that made up the sentence they heard. They have to click on 5 words in order for the next trial to start. At the end of this task, the final SNR in dB (calculated on the 6 last trials) is downloaded in a .mat file.

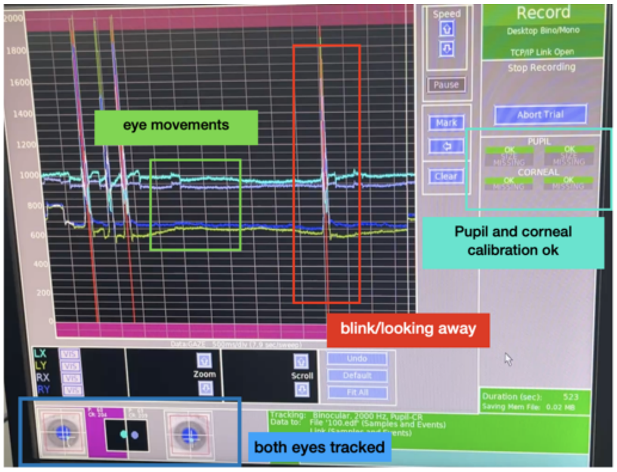
**Task 2** - Word-in-babble (WiB) task :



1: said
2: red
3: dead
4: led

Figure 3.2: Task 2 (WiB)

During this task, the participant listens to 60 stimuli composed of babble noise as the background and one distinctive word as the target (SNR = -2dB). This task is proportion correct based so the SNR stays fixed. 4 words appear on the screen after each stimulus is played, containing 4 similar sounding words. With the keyboard, the participant has to choose by hitting the 1, 2, 3 or 4 key according to the word they think they heard and the screen order. At the end of this task, the proportion correct (portion/percentage of words correctly identified by the participant for the whole task) is calculated and saved in another .mat file.
During this task, an eye-tracker (*Eyelink 1000* tracker) is used to track eye-movements and get the pupil dilation (related to listening effort, see Section [2.2]). A chin rest is installed in the booth on the desk in front of the participant to avoid involuntary movements. The participant rests his head against it and the eye-tracker starts functioning. Calibration and validation for the eye-tracker is done in order to maximise data quality and avoid loss. Once these are done, the task starts on the participant's side, whilst the experimenter also has access to a second screen with the eye recording (see below).



eye movements

Pupil and corneal calibration ok

blink/looking away

both eyes tracked

Figure 3.3: Task 2 - Eye tracker screen

This picture shows depicts screen that appears on the eye-tracker computer during the WiB task after calibration and validation. Each of the 4 coloured lines represents a real-time tracking of gaze/pupil data (x-y positions for the left and right eye). Peaks represent sudden gaze shifts : if a peak is high and red, it means the participant is blinking or quickly looking away. If a big portion of the graph display becomes red, it means the participant stopped looking at the task screen. On this picture, both eyes are tracked (binocular mode) : this is important for microsaccades analysis, which will follow this study, but monocular (1-eye) tracking can also be used if just calculating difference in pupil dilation for listening effort. The pupil data is stored in the same folder as the .mat file, in .asc form.

**Task 3 - Dynamic figure-ground (DFG) task** :

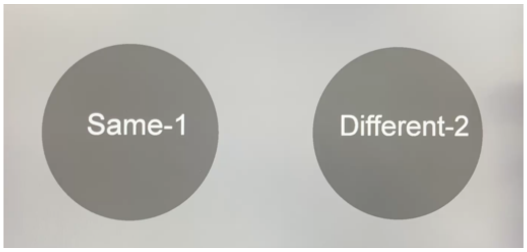


Figure 3.4: Task 3 (DFG)

This task is a pattern recognition task. The participant listens to 200 dynamic AFG stimuli (see Section 2.1), played in pairs. For each trial, they have to identify whether the roving pattern of the figure in each of the 2 AFG are the same or different. This task requires using the keyboard, offering a two-alternative forced-choice (using the 1 or 2 key). It uses an adaptative SNR, which starts at 10dB (target louder than the babble) and staircases with a 1-up 2-down procedure. The mean SNR calculated on the last 6 results is also saved into its own .mat file.

**Task 4 - Static figure-ground (SFG) task** :

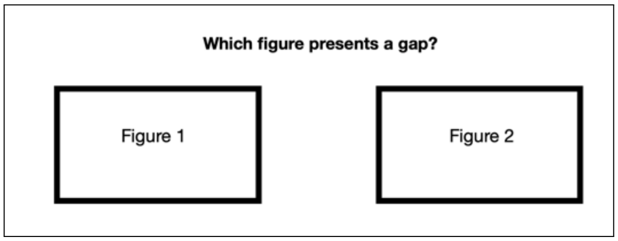

Which figure presents a gap?

Figure 3.5: Task 4 (SFG)

During this gap detection task, the participant listens to 114 static AFG stimuli (see Section 2.1). Stimuli are played in pairs, and the participant has to identify for each trial which figure has a temporal gap in it. This task requires using the keyboard, offering a two-alternative forced-choice (using the 1 or 2 key). It uses an adaptative SNR, which starts at 6dB (target louder than the babble) and staircases with a 1-up 1-down procedure.

Task 3 (Dynamic) and 4 (Static) are almost equivalent to the Cats/DFG and Aliens/SFG tasks of the training paradigm, mentioned in Section 2.3.2. Task 3 and 4 are 2-choice tasks and the Cats and Aliens are 3-choice tasks, meaning the participant has to choose the odd one out. This is useful because the participants understood the training AFG tasks more easily if associated with the tasks they did pre-training, and the effect on those tasks could be more noticeable post-training.

**DAY 1-7 / TRAINING :** For 7 days daily, starting on the day of the pre-testing, the participant had to complete both the Cats/DFG task and the Aliens/SFG task. At home using the training app (see Section 2.3.2), training was done for a week to improve Speech-in-Noise (SiN) perception.

**DAY 8 / POST-SESSION :** On the 8th day, participants came back to the lab for a post-training session, with their laptop. Training app data was downloaded onto the lab's computers to keep to correlate it with the pre/post-sessions results. For 2 hours, the same tasks as the pre-testing session were performed by the participant (without the audiogram assessment). The SSQ had to be filled again to compare the participant's views on their training and subjectively quantify their improvement. Then, tasks 1-4 were performed in order to compare behavioural (SiB-end SNR, WiB-proportion correct) and eye data (pupil dilation) results. Finally, a general intelligence test (a matrix reasoning "MTX" test — like a puzzle IQ test — to qualify fluid intelligence[1]) was given to the participants. The test consists of 26 puzzles, each lasting 30 seconds, with difficulty increasing progressively across trials. An example of a puzzle is provided in [Fig.A.7]. The results could be able to explain the difference between behavioural results, checking whether people with a higher MTX score were better at training. After those 7 days of training, the participant was in any case expected to feel improvement on their perception of SiN (quantified with the SSQ) and to get better results on the post-session tasks.

## 3.2 Training online

The training app used during this study was already computed by the team previously in charge of the study. However, some changes needed to be made in order to expand the project reach and simplify the training process. The app previously computed only worked on laptops with Windows as their operating system. During the second month of this internship, a MacOS version was developed, based on the Windows scripts. Functions were defined or changed accordingly, and the app was able to run on a MacOS laptop. However, MacOS can a very unstable environment combined with standalone apps computed on Matlab. As a result, the app was only compatible with certain versions of the operating system and was not yet fully deployable across all MacOS platforms.

Therefore, an a an alternative solution was envisioned in the form of a web-based version of the application. This would make the training paradigm easier to access for all participants, on any operating system and without needing to bring a laptop during the pre and post sessions. The initial phase of this project started during testing, and involved learning the fundamentals of JavaScript, HTML, and CSS simultaneously. As this stage began in the third month of the internship, the learning process is still ongoing beyond the end of the internship. A redesign of the cats and aliens drawings was also requested by the lab for copyright reasons. The resulting drawings are presented in [Fig A.6].

The web project within this internship began by adapting the Matlab aliens script (the easier to compute, static figure-ground task) to a JavaScript version; a sfg_gap.js script.
This script is constructed around core variables like a trial counter (sfgTrialNo), a list of current volume levels used per trial (sfgTrialVols), the position of the alien with the gap (sfgFigPos), the SNR threshold (sfgThreshold), an adaptative step[2] size array (sfgVolSteps), a binary tracker of correct/incorrect responses

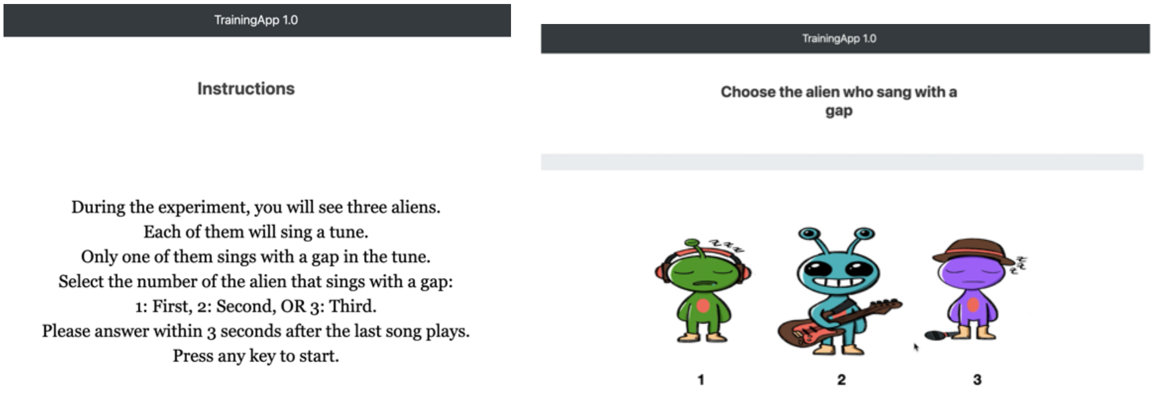
[1] Fluid intelligence ($g_f$) and crystallized intelligence ($g_c$) were introduced in 1943 by the psychologist R. Cattell. Fluid intelligence is the ability to solve novel reasoning problems. Crystallized intelligence is the ability to deduce secondary relational abstractions using learned primary relational abstractions [*Wikipedia*]

[2] The increment (+$X$ dB) or decrement (-$X$ dB) of the target sound applied according to the participant's response

(sfgVerdict) and the total user responses (sfgTrialClicks).

The trial flow goes as follows: 3 stimuli are generated with different figures composed of ”beeps” (chords) at SNR=0dB. The parameters are defined as $s_{rate} = 48000$, $\text{chord}_{duration} = 50\text{ms}$, $\text{trial}_{duration} = 3.5\text{s}$, $\text{gap}_{length/nbChords} = 6$, $\text{figure}_{length/nbChords} = 42$. The background is generated by selecting a random set of frequencies (5–14 chords per time bin) from a quarter-tone-spaced pool[3] around 440 Hz, independently for each time bin. The figure is coherent, formed by 3 fixed frequencies that repeat across a contiguous 42-chord window within the stimulus, creating an object against the background. A temporal gap (300 ms = 6 missing chords) is introduced, at a random location, into the figure for one of the three aliens in each trial, with its onset constrained to avoid edge effects.
A text prompt shows ”Choose the alien who sang with a gap”, on top of 3 alien images. When an alien ”sings”, his active form is triggered by the script. His passive form stays on otherwise (see Fig.A.6). The participant uses the keyboard to identify which of the 3 aliens had the gap according to what they heard. The system records whether the choice was correct.

An adaptive SNR staircase updates the stimulus volume (the difficulty), starting from a step value of 4dB until the 4th reversal, then a step value of 2dB until the 10th reversal then a step value of 1dB for the last reversal. Once 10 reversals are collected, the task stops. If 10 reversals haven't been completed in 15 minutes, the task is stopped automatically. The final SNR threshold is computed using the mean volume of the last 6 reversals. A break is offered on a pop-up screen mid-task (7 minutes in) or if the participant misses on more than 4 trials in a row.
The alien task is fully computed and integrated into the web code infrastructure.



TrainingApp 1.0

Instructions

During the experiment, you will see three aliens.
Each of them will sing a tune.
Only one of them sings with a gap in the tune.
Select the number of the alien that sings with a gap:
1: First, 2: Second, OR 3: Third.
Please answer within 3 seconds after the last song plays.
Press any key to start.

TrainingApp 1.0

Choose the alien who sang with a gap

Instructions screen

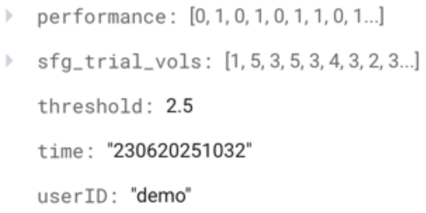
Trial example (2nd alien is ”singing”)

Figure 3.6: Aliens/SFG task

The second phase of the project involved developing the user interface (UI) and managing the database to stock all the participants' training data. This was previously done using an Excel sheet within the training app.
A database was first created using Google Cloud Firestore[4]. This choice was made so that the final version of the website can later be deployed on the Google servers easily. Linking this new database to the code into main.js [Fig.A.8] allows for chosen variables to be saved permanently and the participants' training data to be retrieved by the experimenters.

[3] Each adjacent frequency is one quarter-tone apart ($2^{1/24}$)

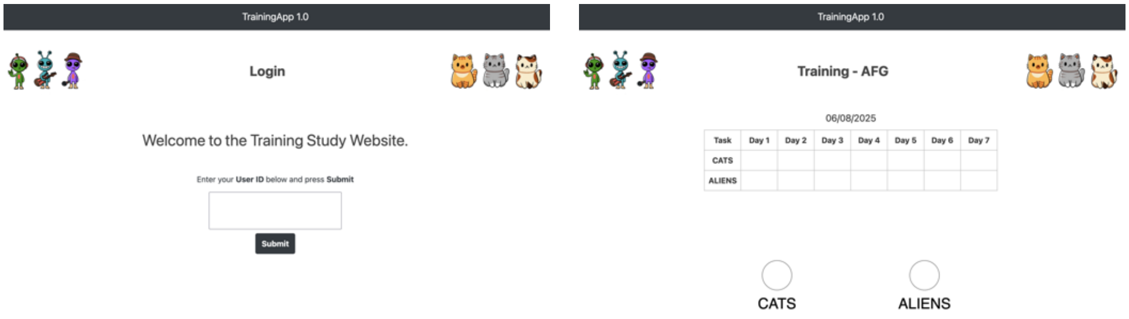
[4] Flexible, scalable NoSQL cloud database, built on Google Cloud infrastructure, to store and sync data for client- and server-side development (https://firebase.google.com/docs/firestore)



```
performance: [0, 1, 0, 1, 0, 1, 1, 0, 1...]
sfg_trial_vols: [1, 5, 3, 5, 3, 4, 3, 2, 3...]
threshold: 2.5
time: "230620251032"
userID: "demo"
```

Figure 3.7: Variables saved in the database after completing the Aliens task once. *Performance=right/wrong answer, trial_vols=adapted SNR for each trial, threshold=end SNR (6 reversals)*

The UI contains different elements set up to match the training app. First, the main app screen is designed close to the app one. However, it is separated in 2 following screens.



Welcome to the Training Study Website.

Enter your **User ID** below and press **Submit**

06/08/2025

| Task | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Day 6 | Day 7 |
|---|---|---|---|---|---|---|---|
| CATS | | | | | | | |
| ALIENS | | | | | | | |

First window

Second window

Figure 3.8: Website setup

The second screen remains incomplete; logos are missing, and a data download feature still needs to be implemented.

The webpage is set up as follows. In the first screen, the participant ID is loaded using a HTML text-type user input variable. In the second screen, the tracking table is replicated within the main.js, main.css and index.html scripts using JavaScript's *Date()* function and an offset to count days. The buttons ”CATS” and ”ALIENS” use a block/unblock division logic within a self-made *ChooseTask()* function in a controller.js script to allow the right task to start. Then, the user uses the same controls (also defined in controller.js) as in the Matlab app : mainly the keyboard during the tasks, and the mouse to select the task on the main app screen. The drawings used to make the app interactive are the newly designed ones [Fig.A.6]. All of these website commands are spread out between an index.js, controller.js, main.js, index.html and a main.css file.

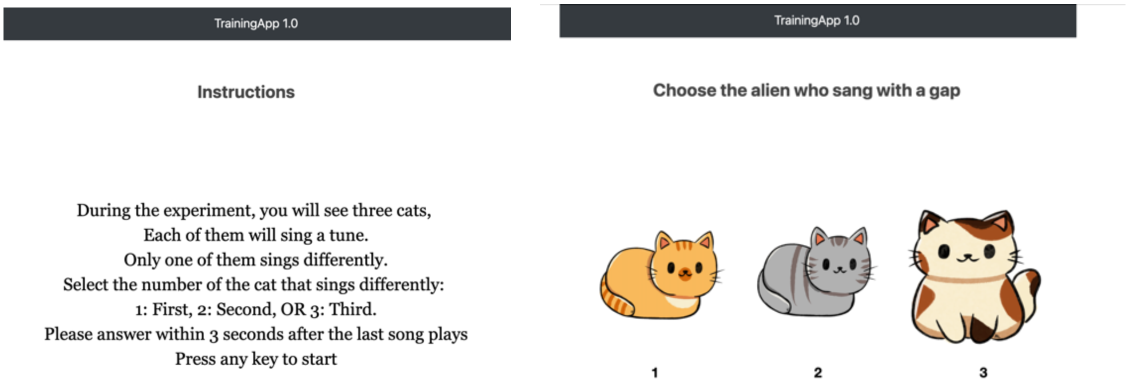
The last phase of the project, before online deployment, was the Cats task. Due to the use of roving figure patterns and harmonics in dynamic figure-ground (DFG) tasks, this task has proved more challenging to code and has since been taken over by the rest of the team for further development. However, initial work began during the last month of the internship. The task was adapted into a ”dfg.js” file.

This script is constructed around core variables such as a trial counter (dfgTrialNo), the current figure sound level in each trial (dfgTrialVols), the position of the cat with the different figure pattern (dfgFigPos), the adaptive step size array (dfgVolSteps), a binary tracker of correct/incorrect responses (dfgVerdict), and the total number of participant responses (dfgTrialClicks). The main parameters are the same as the Aliens task.

The trial flow is as follows: 3 stimuli are generated, each lasting 3.5s. The background in each interval is formed independently by selecting a random set of frequencies (5–14 chords per 50 ms time bin) from a quarter-tone–spaced pool around 220 Hz. The figure is composed of 3 simultaneous tone components whose frequencies change over time according to a pre-defined pattern loaded from a set of dynamic trajectories (DFG_figfreq). The harmonics trajectories are also calculated and then added to the figure. In two of the stimuli (”same-pattern” stimuli), the figure follows the same trajectory, while in the third (”different-pattern” stimulus) the figure follows a different trajectory. The time–frequency coherence in the figure contrasts with the stochastic background.

A text prompt shows “Choose the cat who sang differently”, above 3 cat images. When a given cat “sings” (its interval is playing), the script swaps its displayed image to an active form; otherwise, it shows the passive form. The participant uses the keyboard to indicate which of the three cats had the different figure pattern. The system records whether the response was correct.

An adaptive staircase procedure updates the figure level (difficulty) based on performance. The step sizes in dfgVolSteps start at 2 dB until the 4th reversal, then decrease to 1 dB thereafter, using a 1-up/1-down rule. The final threshold is computed as the mean figure level over the last 6 reversals.



Instructions

During the experiment, you will see three cats,
Each of them will sing a tune.
Only one of them sings differently.
Select the number of the cat that sings differently:
1: First, 2: Second, OR 3: Third.
Please answer within 3 seconds after the last song plays
Press any key to start

Choose the alien who sang with a gap

Instructions screen

Trial example (3rd cat is ”singing”)

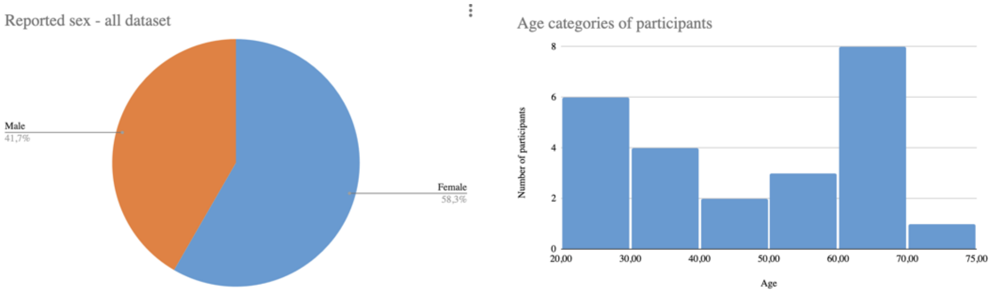
Figure 3.9: Cats/DFG task

This task is still under development as of now. The sounds have not yet been generated, with only the figure patterns being loaded. The final version should also terminate after 10 reversals, or automatically after 15 minutes if this criterion is not met. A break screen should also be displayed midway through the session (7 minutes) or sooner if the participant misses more than four consecutive trials.

The full, finished version of the website could not be deployed online due to time constraints related to the end of this internship. However, the rest of the team is expected to continue working on the website until it is fully deployed for use in the remainder of the training study.

# Chapter 4
# Results

This section presents the behavioural and physiological (pupillometry) results obtained during the training study. Results are reported for a certain portion of all participants tested to date in this study: 11 kept from prior to this internship and 12 from during this internship (out of 15 tested), making it 23 in total. Presenting the complete set of results, rather than only the ones obtained during the internship, provides a clearer overview of the study's scope and facilitates interpretation.
During the study, those 23 participants with normal hearing to mild loss were tested. The demographic profile of the sample is as follows:



Participants' age

Participants' sex

Figure 4.1: Demographics

The sample included close proportions of male and female participants, with a majority of females. Ages ranged from 21 to 72 years old in accordance with the inclusion criteria (the age criterion was modified between the start of the study at the lab and the internship, going from 20-75 to 25-70 years old for better data accuracy). Most were between 40 and 65 years old, which was the preferred age range as noted in Section 3.1. The second age range with the most participants was the 20-30 years old.

## 4.1 Behavioural

All 23 participants' data were considered for the behavioural results, with no exclusion criteria. One possible criterion for exclusion would have been to remove individuals with high baseline scores in the pre-session (e.g., more than 80% correct on the WiN task), but this was not implemented.

### 4.1.1 Training

Shown below are the training app results for a randomly selected participant from this study.

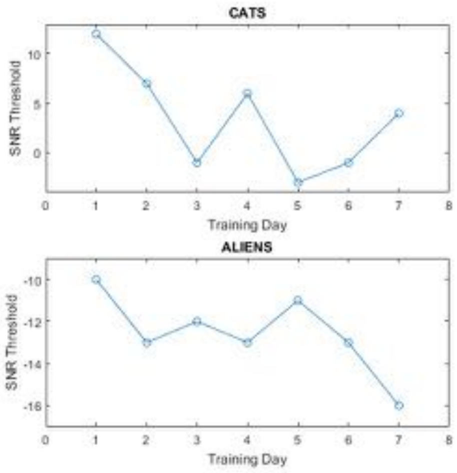


Figure 4.2: Training app results for a random participant

This figure shows the week-long training progression for a single participant using the Training app (see Subsection 2.3.2), with separate plots for the two tasks: the Cats (dynamic figure–ground (DFG)/pattern discrimination) and Aliens (static figure–ground (SFG)/gap detection) tasks.

The top figure's y-axis represents the SNR threshold (in dB) at which the participant could perform the DFG task, estimated from the adaptive staircase at the end of each day's session and calculated over the last 6 reversals in every session (thus the starting SNR of each day is the calculated threshold of the previous day). The x-axis represents each day the participant has trained using the app. The 15 dB starting point on the vertical axis corresponds to the initial SNR of the task. Higher thresholds indicate that the participant required a more prominent figure relative to the background to perform accurately. Over the week, the participant's threshold fluctuates (day 4 and day 7 being the highest peaks) but shows an overall improvement compared to the first day. In 3 days, the participant goes from 15dB to a -2dB performance. The peak observed on day 4 could be due to the SNR ending too low on day 3, as the task each day begins with the SNR from previous day's performance. The same explanation may apply to day 7. A minimum can be noticed around day 5, showing better performance towards the end nevertheless. Day 7's performance also stays -10dB better in SNR than day 1, indicating a strong effect within the training task itself.

On the bottom figure, the axes are the same as the previous plot, but the graph shows progress within the SFG training task. The SNR threshold starts at -10dB, as this is how this easier, static task is designed. The overall trend shows gradual improvement over the week, with the lowest threshold on day 7 indicating the best performance at the very end. A slight increase in SNR is observed on day 5, but the overall downward trend remains consistent.

These subplots suggest that the participant adapted differently to the two tasks. In the Cats task, performance varied more day-to-day, possibly due to the greater difficulty, variability of tracking dynamic figure patterns, and the need for higher levels of attention. In the Aliens task, performance improved more steadily, consistent with it being an easier, static figure–ground task. Lower thresholds in both subplots correspond to better auditory segregation ability in noise. These trends were seen in most participants' training app results.

### 4.1.2 Lab sessions

For the first three tasks performed pre- and post-session, results have been plotted side-by-side for each participant using box plots for clearer comparison.

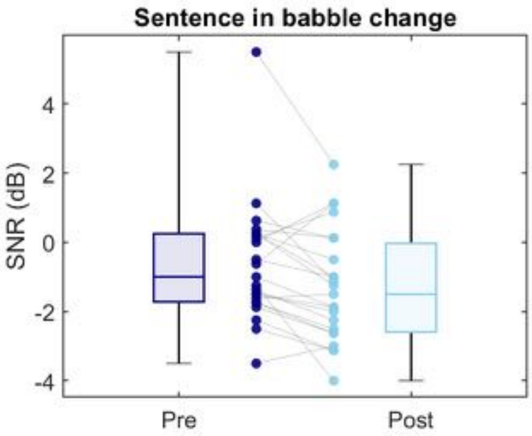


Figure 4.3: Task 1 (SiB) results - mean SNR

Figure 4.3 shows the change in speech-in-noise performance on the first task of the experiment, the Sentence in Babble (SiB) task, measured pre- and post- training (x-axis). The y-axis represents the SNR threshold in dB, calculated from the final six reversals, like for all tasks. Lower SNR values indicate better performance on the task and reflect the ability to understand speech in more challenging noise conditions.

Each participant's data is represented by paired dots connected by a line to show their progress between pre and post-training, with dark blue indicating pre-training scores and light blue indicating post-training ones.

The box plots provide a visual summary of the distribution across participants, showing the spread and central tendency of the data. It highlights key statistics like the median (the middle value), the interquartile range[1], and potential outliers. The box itself spans the interquartile range, while the bars over and under it extend to the minimum and maximum values. Here, the box spans lower SNR values in the post-session, and the median is set almost 1dB lower than the pre-session. The interquartile range is however larger in the post-session, indicating less consistent results in this range. Nevertheless, the bars surrounding the box span less values overall, indicating a more consistent performance overall post-training.

This shows that, following training, most participants demonstrated a reduction in SNR thresholds. The participants showing an increased SNR in post-session were participants with an already low threshold from the pre-session (between 0 and -4dB). Since their initial performance was already strong, there may have been less room for noticeable improvement, even with a strong training effect. Overall, a significantly improvement in performance is seen on the figure, reflected in the downward shift of the median and a general, significant trend toward improved performance in lower thresholds.

[1] The IQR represents the range within which the middle 50% of values fall

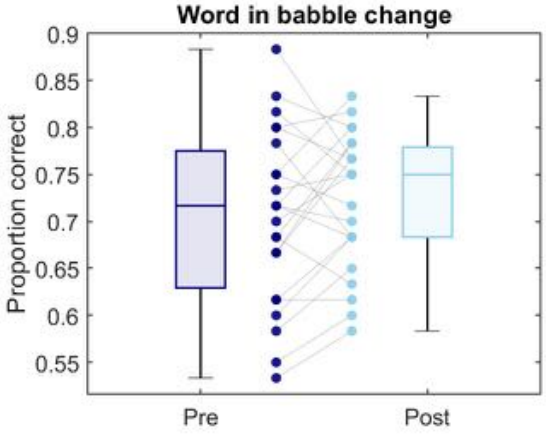


Figure 4.4: Task 2 (WiB) results - proportion correct

Figure 4.4 shows the proportion of correct responses in the second task, the Word-in-Babble (WiB) task, pre and post session. Like in the SiB plot, each dot represents a singular participant, and lines connecting them indicate within-subject change over time. The x-axis is the same as the previous figure, while the y-axis represents the proportion correct score.
The box plots indicate a general trend of improvement in WiB performance following training. Median performance increased from the pre to post conditions (of approximately 0.04%), and the interquartile range appears slightly reduced, suggesting a more consistent performance post-training and less variability in performance across participants. Some individual variability is still present, which may be explained in the same way as for the SiB task: participants scoring above 70% in the pre-session may have found it difficult to improve on an already high score, even after training.
Overall, the upward shift in the group suggests a positive effect of training on word-in-noise perception, with most participants showing higher accuracy in the task after training.

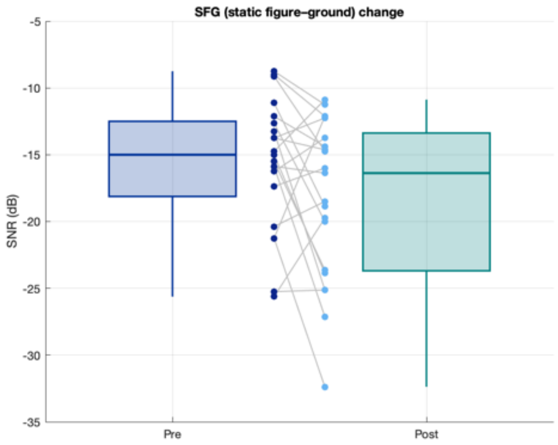


Figure 4.5: Task 3 (SFG) results - mean SNR

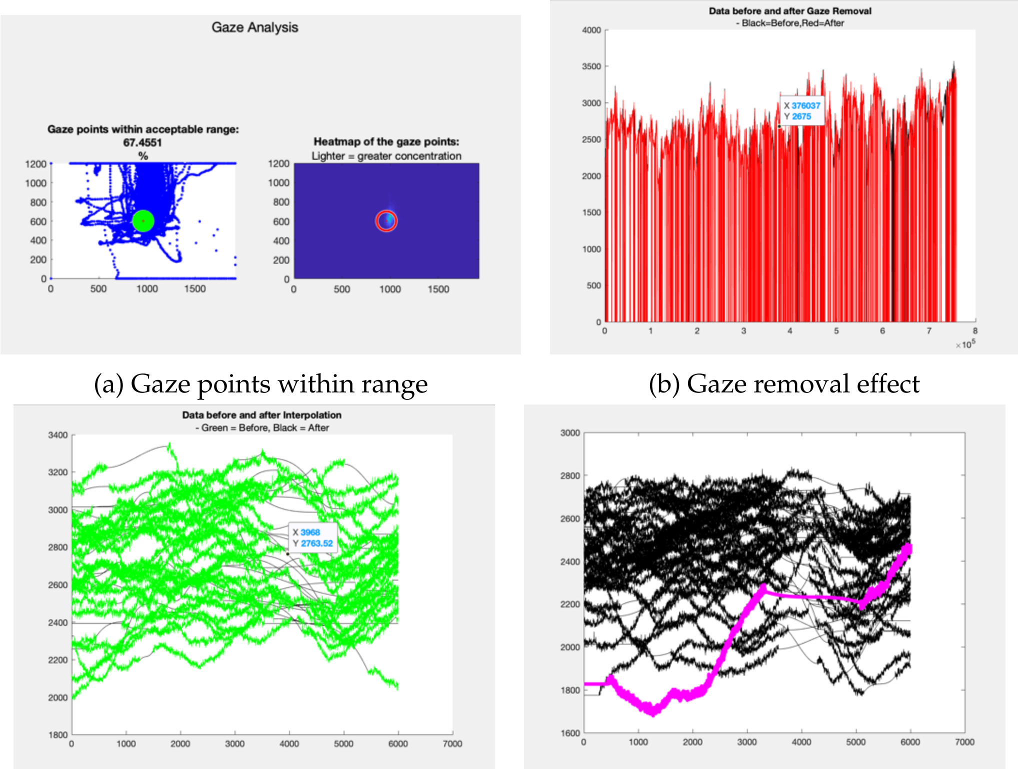
Figure 4.5 shows the mean signal-to-noise ratio (SNR) thresholds for the static figure–ground (SFG) task, measured pre and post training. The dots, lines, and axes represent the same elements as described for the SiB figure previously (Fig.4.3). Lower SNR values indicate that the participant was able to detect the figure against the background at a lower target-to-background ratio, thus indicating better performance.

The box plots show that the median SNR threshold decreased slightly from pre- to post-session, with a -1dB difference, suggesting improvement in SFG performance following training. Surprisingly, the interquartile range is wider in the post condition, indicating greater variability across participants after training, like for the SiB task (Fig.4.3). While many individuals' lines slope downwards to show improved performance, some show little change or even an increase in threshold. Overall, the pattern suggests that training may have led to a small group-level improvement in static figure–ground segregation ability, although individual responses to training varied considerably. Since this was a repeated daily task over 7 days within the training app, some participants may have been less engaged during post-testing, unbalancing the progress.

The dynamic plot (Task 4 results) was not generated due to lack of access to the full data, but the few data accessible also showed great improvement in performance across participants.

## 4.2 Eye data

For the eye-tracking data, only 15 participants were retained due to poor data quality for the other 8 participants. Data quality was assessed using a dedicated Matlab pre-processing script. The script loads the .asc files obtained during Task 2 in both the pre- and post-session using the ”asc2data” Matlab function. It then evaluates the proportion of the gaze stream within an acceptable on-screen region, removing samples outside this area from the data series. It detects blinks and gaze dispersion, removes the corresponding segments, and returns a cleaned pupil time series. Bad epochs are identified and excluded based on predefined criteria (e.g. too many missing points, excessive range...), with interpolation applied to fill short missing segments. The script also flags outliers (e.g. trials with abnormally large amplitudes) and asks the experimenter to decide whether to exclude them on the console. Finally, it saves the processed data as .mat files for following analysis.



(a) Gaze points within range

(b) Gaze removal effect

(c) Interpolation filling empty segments

(d) Outlier detection (in pink)

Figure 4.6: Example of eye data pre-processing (random participant, pre-session)

Fig.4.6 shows the results of this preprocessing for one participant, each subplot being important to determine data quality.
The first subplot (4.6.a) shows gaze analysis, with the left panel showing raw gaze samples over screen coordinates and the right panel showing a heatmap of gaze density. The axes for both are defined as horizontal and vertical gaze position in pixels. On the left panel, the green circle marks the acceptable region around fixation. On the heatmap, a light spot centred on the target indicates good fixation; the red/white circle shows the rectangular region of interest. It verifies that the participant actually looked near the fixation area during the task, which is important because off-screen gaze often correlates with noisy pupil baselines. The second subplot (4.6.b) shows continuous pupil trace before (black) and after (red) removing samples associated with bad gaze. X-axis corresponds to sample index and y-axis to pupil size, in arbitrary units defined by the Eyelink tracker. It demonstrates that off segments were cleaned before further analysis. The third subplot (4.6.c) corresponds to data before and after interpolation (the green lines show pre-interpolation and black segments show the result post operation). The axis are the same as the previous subplot. This plot shows that short missing data (like blinks) were filled to produce continuous epochs suited for averaging. Finally, the last subplot (4.6.d) corresponds to the traces in black, with the flagged outlier in pink. This plot helps to decide whether to remove outliers to avoid strong bias when averaging. The axes are the same as the previous subplots.

A participant's data was retained for full processing if fewer than 30 trials (out of 60) were rejected, ensuring that at least half of the trials remained for analysis. A trial was rejected if more than 50% of its data samples were missing, regardless of the cause (e.g. gaze points outside the acceptable range or excessive blinking). 8 participants had data flagged for too much missing trials and too little gaze

points in range, making the dataset size equal 15.
For this participant, in Fig.4.6, there is more than 67% of gaze points within acceptable range, meaning the participant mainly looked at the middle of the screen during the task, as requested by the experimenter. The heatmap of the gaze points shows greater concentration of gaze right in the middle of the screen also, supporting this observation. Then, the 2nd subplot of Fig.4.6 shows that most of the trials were kept for this participant, with the red plot being closely matched to the black plot. The interpolation managed to fill the empty segments accordingly, and 1 outlier was flagged for this participant. The outlier in this case was discarded because of very strong amplitude difference in the first samples.

After pre-processing the entire dataset, the main analysis stage was carried out using a second Matlab script. For each participant, all pre- and post-session data from the preprocessing stage were loaded. A baseline correction[2] was first applied. The script then z-scored[3] both conditions. For each condition, the script averaged across trials to produce a single time-course per participant.
A one-tailed paired t-test[4] was then performed on each participant's mean, capturing the sustained pupil dilation. The script saved the mean p and t values over this whole time window. Finally, it averaged across participants to generate group-level mean time-courses for pre-training and post-training, and annotated the graph with the t-test values.

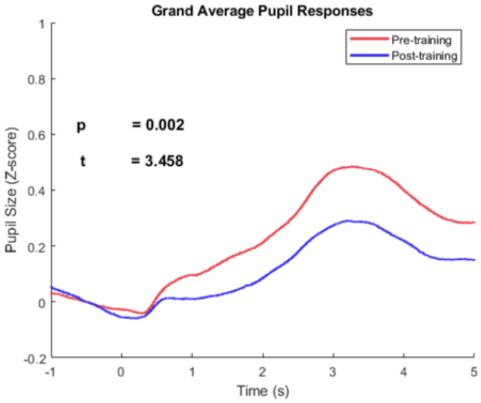


Figure 4.7: General pupil response change (z-scored)

This plot (Fig.4.7) shows the grand-average pupil dilation responses across the reduced dataset (15 participants) from the study, measured pre and post-training. X-axis corresponds to the time in seconds, with 0s marking the onset of the stimulus, and y-axis corresponds to the pupil size in z-score units (baseline-corrected to the −1s to 0s window as mentioned earlier). Pre and post-training data appear in curves of different colours.
The trend for both curves is the same : after an initial dip following stimulus onset, the pupil size increases, peaking around 3s before returning toward baseline. However, the blue curve lies below the red after stimulus onset, with a more noticeable difference within the 2-5 second period. This means pupil size is smaller post-training. The biggest difference appears around 3s, which represents

[2] Sets each trial's starting point to zero by subtracting the average pupil size in the period just before the sound onset

[3] Scales the data so that each value represents the number of standard deviations above or below the participant's mean, for comparison across participants

[4] A paired t-test is a statistical test comparing the means of two conditions on the same participant to assess whether the observed difference is likely due to chance. The t-value indicates the size of the difference relative to variability, while the p-value gives the probability of obtaining such a difference if no real effect exists. A one-tailed t-test is used when the hypothesis predicts a specific direction of change (e.g. performance improves post-training)

the word onset for each stimulus during the WiB task. This difference between pre- and post-training curves demonstrates smaller sustained pupil dilation after training.

## 4.3 Other results

This section reports complementary results from the study, such as the SSQ test results (see [2.3.1]) and the MTX/generalized intelligence test ones (see Section 3.1).

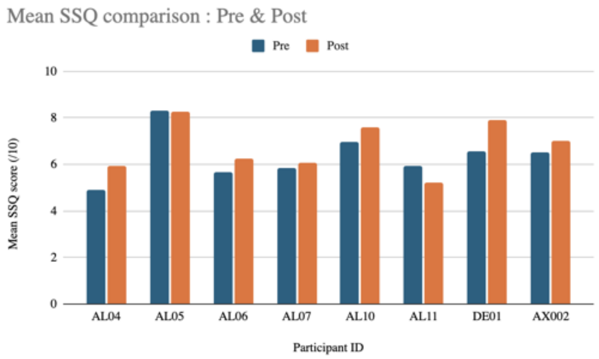


Figure 4.8: Examples of mean SSQ scores pre- and post- training

This chart shows mean SSQ test scores (rated on a 0–10 scale) for each participant before (blue) and after (orange) training. These mean scores have been obtained after averaging all 12 questions scores on the SSQ test (see Fig.A.5) for each participant for both the pre- and post-session. In this figure, only a few participants were selected due to data missing for previous participants. Since SSQ measures perceived speech, spatial, and quality of hearing, these changes reflect those participants' subjective improvement after training.

Most participants show an increase in mean SSQ score from pre- to post-training, suggesting an overall trend of improvement in self-reported hearing abilities, gaining from 0.25 to 1.34 points of difference between sessions. Some participants show almost no change in their qualification of hearing abilities, or even slight decrease.

Being subjective in nature, these results hold less weight in the context of the study, but are still interesting combined with other results. Unfortunately, plotting the SSQ score differences against the SNR data from the SiB task and the proportion correct scores from the WiB task revealed no meaningful correlations. This might be due to the limited availability of SSQ results for many participants, which prevents a reliable correlation analysis.

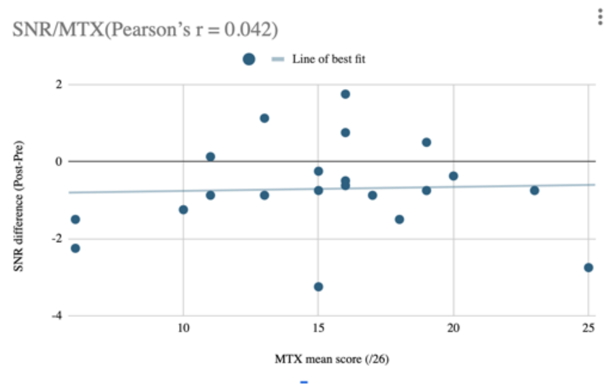


Figure 4.9: Relationship between mean MTX score and SiB improvement

The relationship between mean MTX scores (/26, x-axis) and improvement in SiB thresholds (SNR difference Post–Pre in dB, y-axis) was also examined. As shown in Figure 4.9, the correlation was very weak (r[5] = 0.042) and not statistically significant ($p = 0.858$). This indicates that fluid intelligence, as measured by the matrix reasoning test, did not reliably predict the amount of improvement participants gained from training. The line of best fit is almost flat, further confirming the absence of a systematic trend. While some participants with high MTX scores did show strong improvements, others showed little or even negative change, suggesting that factors other than fluid intelligence (e.g. baseline auditory ability, attention, or task engagement) may play a larger role in shaping training outcomes. In this sample, MTX scores appear more related to overall performance levels rather than to the magnitude of training-related gains. However, this non-effect is shown for only a small sample of participants and could appear within a bigger sample.

[5] Pearson's r (also called the Pearson correlation coefficient) is a number between –1 and +1 that tells how strongly two variables are linearly related. If close to 0, there is no linear correlation.

# Chapter 5
# Discussion

The project was set out to evaluate whether targeted auditory figure–ground (AFG) training could improve Speech-in-Noise (SiN) perception, using a combined approach of behavioural testing, pupillometry, subjective self-reports and assessment of fluid intelligence. This chapter will provide insight on the meaning of the results obtained in Chapter 4 and link them together, and validate the hypotheses put forward in Chapter 2.

**Behavioural results:**
The behavioural results provide evidence for a positive training effect. They align with the hypothesis outlined in Chapter 2: training in auditory grouping abilities improve Speech-in-Noise (SiN) perception.

The reduction in SNR thresholds observed in the Sentence-in-Babble (SiB) task [Fig.4.3] from pre- to post-training indicates that they were able to correctly identify sentences in babble under harder listening conditions, reflecting an improvement in their ability to segregate speech from background noise. Given that SiB is a more demanding measure of SiN perception, these gains suggest that the auditory grouping training effectively strengthened the perceptual and cognitive mechanisms needed for complex listening environments.

In the WiB task, proportion correct scores also increased generally from pre- to post-training [Fig.4.4], although the improvement was deemed smaller than in SiB. Several factors could explain this difference. First, WiB may be perceived as an easier task (only tracking 1 word), with many participants starting at relatively high scores pre-training, leaving little space for improvement. Second, participants may have been more tired by the time they reached WiB, as it followed SiB in the testing order. Finally, the need to maintain steady gaze (e.g. reduce blinking) for the eye tracker during WiB may have diverted attention away from the listening task, limiting performance gains.

For the SFG task [Fig.4.5], SNR thresholds were substantially lower post-training for most participants, indicating improved detection of static figure-ground in noise. The magnitude of improvement was generally larger than in SiB or WiB, which may be due to the strong similarity between the SFG task and the Aliens task within the training paradigm, bringing task-specific transfer. However, the post-training box plot is wider spread than the pre-training one, suggesting bigger variability between participants despite the strength of improvement for some participants. This could reflect differences in individual listening strategies or the extent to which each participant adapted after training.

When considered together, the results for SiB, WiB, and SFG provide converging evidence for the benefits of training. The stronger gains in SiB and in SFG suggest that the training is particularly effective for situations where listeners must segregate complex auditory scenes, and that task similarity to the training paradigm amplifies this effect. Overall, the pattern supports the idea that enhancing auditory figure–ground segregation skills can improve SiN perception, while also highlighting that the degree of improvement may depend on task difficulty, baseline performance, and cognitive load during testing.

After considering all those pre–post behavioural results, it is worth noting that the training application

itself showed evidence of a within-task learning effect [Fig.4.2].
The training curves for both tasks indicate a general movement toward lower SNR thresholds over the course of the week, consistent with improved perceptual sensitivity. While some fluctuations are noticeable over days, likely reflecting variability in attention, fatigue, or task engagement, the overall trajectory suggests that repeated exposure to the adaptive figure-ground challenges enhanced participants' ability to detect the target figures at progressively lower SNR thresholds. The more pronounced gains in the Aliens task may be related to its simpler, static structure, allowing faster learning, whereas the Cats task's dynamic figures could require longer exposure to consolidate perceptual strategies. These patterns align with the hypothesis that training in auditory segregation can bring strong improvement in SiN perception, even over a relatively short time window.

**Physiological results:**
The eye-tracking results [Fig.4.7] provide physiological evidence consistent with the hypothesis that reduced pupil dilation reflects lower cognitive load and listening effort during auditory tasks (see Section 2.2). In the 2–5s time window after stimulus onset, we observe a clear reduction in pupil size from pre- to post-training, suggesting that participants used less cognitive resources to perform the same task following training. This effect peaks around 3s, which corresponds to the onset of the target word in the WiB task. This timing strengthens the interpretation that the observed pupil-size reduction is directly linked to the improved processing of SiN, rather than reflecting a general arousal change.

The statistical analysis further supports the robustness of this finding: the paired-sample t-test provided a t value of $t = 3.458$, indicating that the magnitude of the reduction in pupil size is large relative to within-participant variability, and a p value of $p = 0.002$, well below the 0.05 threshold, confirms that this effect is highly unlikely to be due to chance.

When considered at the same time as the behavioural results, which showed measurable improvements in performance, the eye data suggest that training not only improved accuracy and reduced SNR thresholds but also made listening more efficient. Participants were thus able to achieve better SiN perception with less cognitive effort. This is also important in the context of reducing listening fatigue and improving communication ease in everyday noisy environments, particularly for older adults or individuals with hearing difficulties.

It should be noted that the eye-tracking dataset was limited in scope due to data quality issues in several participants, many of which were related to age. Common problems included pupil drift/astigmatism, bright reflections from thick contact lenses, and unstable gaze signals due to eye tremors. While these factors reduced the sample size for the analysis, the strength of the effect in the available data still provide strong evidence for a training-related reduction in listening effort.

**Other results:**
Beyond the main behavioural and physiological outcomes, several secondary factors were explored to better understand individual variability in training benefits. Age appeared to play a role in performance patterns: younger participants between 20-30, forming the largest pool of participants after the 60-70 category [Fig.4.1], tended to score highly on all tasks. This high baseline performance could reflect reduced effect of hidden auditory losses in younger listeners and generally faster perceptual processing. Then, their smaller post-training gains may be explained by either a ceiling effect, where little room for improvement remained because of high performance, or by the possibility that the training paradigm offered less benefit to listeners already performing near optimal levels.

Fluid intelligence, as measured by the MTX test, could also influence outcomes. Higher MTX scores could facilitate the integration of training, allowing participants to adapt more quickly and effectively to the figure–ground tasks. At the same time, MTX may simply reflect a general cognitive advantage that translates into better performance across all tasks, regardless of training. In this dataset, however, the effect size of MTX was not large enough to reveal a clear trend linking high MTX scores to greater training-related improvement [Fig.4.9]: while cognitive ability may contribute to general performance, its effect on the observed gains is uncertain, and even poor in this study.

While they were deemed improved for most participants [Fig.4.8], self-reported subjective listening abilities, measured with the SSQ test, showed little correlation with the objective outcomes. However, it could have been used to exclude participants according to whether they felt improvement or not (this choice was not made in this study to show the full results). It is to be noted that SSQ data were particularly limited, as this questionnaire was only introduced partway through the study and thus available for fewer participants than other measures. While it could be expected that participants reporting subjective improvement would also show stronger objective gains, it is also possible that the short training period (one week) was insufficient for noticeable perceptual changes to be consciously recognised and reported by participants.

Taken together, these additional analyses highlight that the impact of auditory training is shaped not only by the training paradigm itself but also by participant characteristics such as age, cognitive profile, and subjective perception. While the core findings support the hypothesis that training in auditory grouping improves SiN perception and reduces listening effort, these secondary factors help explain individual differences in the magnitude of improvement and suggest important considerations for future targeted experiments.

Finally, results from the web-based version of the training are not yet available, as the platform remains under development. Once completed, the outcome data should take a similar form to those obtained with the training app (Fig.4.2), since the web version replicates the same structure, but in a format that is easier to deploy and more accessible for large-scale implementation of the study.

**Considerations on engagement:**
The training results presented in this report (the Cats and Aliens trajectories in Fig.4.2) highlight issues directly relevant to online testing. The overall downward trends in SNR thresholds are consistent with learning, but some atypical fluctuations were observed as mentioned earlier. For example, in Fig.4.2, performance appeared to worsen at the final day of training after good improvement on the Cats task curve. Such deviations are unlikely to reflect a genuine decline in perceptual ability and are more likely the result of lapses in attention or drops in engagement during training. This highlights a general challenge in training and online studies: ensuring that participants remain motivated and consistently focused across multiple days of repetitive tasks. When monitoring is limited, as it is the case in unsupervised online contexts, the risk of disengagement becomes stronger and can directly impact data quality and future interpretation.

On another aspect, the implementation of a web-based platform also brings both opportunities and challenges.
On one hand, moving to a JavaScript-based platform substantially increases accessibility, scalability, and ease of deployment. Tasks can be run directly in a browser without installing specialised software (Matlab Runtime is needed for standalone apps, plus needing to install an app on itself makes it difficult when going through various firewalls), making large-scale studies and remote participation easier. This shift also supports a more flexible infrastructure, notably with improved database management,

redesigned user interfaces, and the integration of updated ”game” characters, all of which can help sustain user motivation.
On the other hand, the risks associated with limited supervision in online training need to be tackled. Without experimenters present, participants may not always follow instructions closely, may lose focus, or may approach the tasks with different degrees of effort. To help this, the web version could incorporate dedicated tools to monitor and maintain engagement. These could include attention-check trials appearing at random intervals, adaptive reminders when performance drops suddenly, or more present and interactive feedback mechanisms to reinforce task compliance. By combining accessibility with reminders for engagement, the web platform can deliver reliable data at scale while preserving the integrity of training outcomes.

**Limitations:**
While the findings presented here are encouraging, several limitations should be acknowledged. First, the absence of a control group makes it difficult to fully rule out alternative explanations for the observed improvements, such as increased task/test familiarity. Including a control group in future work would allow stronger causal conclusions about the specific contribution of the training paradigm. Second, the broad age range of participants may have influenced the results. Younger participants generally started with higher baseline scores, which likely limited their room for improvement and may have led to smaller training gains compared to older participants. On the other hand, older participants may have shown more pronounced improvements but were also more susceptible to eye-tracking data quality issues. These factors should be considered in the interpretation of the current results and addressed in future study designs.

# Chapter 6
# Conclusion

This study combined a large range of elements to investigate the impact of training in fundamental sound grouping ability through auditory figure-ground paradigms on Speech-in-Noise perception, attention, and listening effort. Beyond the usual testing itself, the work covered recruiting participants, training new interns for testing, developing and implementing a web-based training platform using JavaScript/CSS/HTML, and even the design of cartoon characters (to enhance participant motivation and maintain engagement across multiple training sessions).

Participants completed a battery of behavioural tests, including the Speech-in-Babble (SiB), Word-in-Babble (WiB), static and dynamic figure-ground (SFG/DFG) tasks, both before and after the training phase. They also completed 2 SSQ tests to subjectively qualify their hearing abilities and 1 matrix-reasoning (MTX) test to track their general intelligence. In parallel, physiological measures were collected through eye-tracking, enabling a quantification of changes in pupil size as an index of listening effort. The combination of behavioural and physiological data provided complementary insights: while behavioural tasks revealed measurable improvements in performance after training, the eye-tracking results suggested a reduction in pupil dilation over time in certain tasks, consistent with decreased cognitive load after training.

The web-based training application was thought to be a more flexible tool to offer adaptive psychoacoustic tasks remotely to participants. Although not yet complete, its implementation will be supporting running the training tasks in a browser without the need for local software installation, enabling easier participant access and centralised data collection. Starting the website implementation during this internship facilitated the team's progress toward a final usable version.

Taken together, the results highlight the role of training in improving perceptual performance and reduce listening effort. Improvements observed in some behavioural measures, alongside changes in physiological indices, suggest that repeated exposure to controlled auditory tasks targeting auditory segregation, like AFG tasks, can enhance participants' ability to perceive speech in background noise and potentially reduce the effort required to do so. However, the magnitude of improvement varied between tasks and individuals, pointing to the influence of cognitive, perceptual, and possibly motivational factors.

This project also lays important basis for future research. A logical next step would be to explore neurofeedback-based auditory training, where participants could receive immediate feedback on their listening effort via EEG[1] measures, enabling a more targeted modulation of cognitive engagement. Such approach could deepen our understanding of the mechanisms underlying successful training and help identify the optimal conditions for improving listening in noise, both for the general population and for individuals with hearing difficulties. A neurofeedback component of this study has already been initiated and is being carried out by the team responsible for the training study in the Auditory Cognition Group.

[1] An electroencephalogram (EEG) is a recording of brain activity, obtained via electrodes on the scalp

# Bibliography

[1] Auditory Cognition Group. *Webpage*. https://www.auditorycognition.org/. Accessed: 2025-07-31. 2019.

[2] Xiaoxuan Guo et al. “Predicting speech-in-noise ability with static and dynamic auditory figure-ground analysis using structural equation modelling”. In: *Preprint* (Sept. 2024). DOI: https://doi.org/10.1101/2024.09.08.611859. URL: https://www.google.com/url?q=https://www.biorxiv.org/content/10.1101/2024.09.08.611859v1&source=gmail&ust=1741779158395000&usg=AOvVaw2eeAOFBhnJISqJFoIMWJTI.

[3] et al. Xiaoxuan Guo. “Manuscript : Neural tracking of Speech-in-Noise and Auditory Figure Ground”. unpublished, in-the-works; no official title yet. 2025.

[4] Claudia Contadini-Wright et al. “Pupil Dilation and Microsaccades Provide Complementary Insights into the Dynamics of Arousal and Instantaneous Attention during Effortful Listening”. In: *The Journal of Neuroscience* 43 (June 2023), pp. 4856–4866. DOI: https://doi.org/10.1523/JNEUROSCI.0242-23.2023. URL: https://www.jneurosci.org/content/43/26/4856.

[5] Emma Holmes and Timothy D. Griffiths. “‘Normal’ hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception”. In: *Scientific Reports* 9.1 (Nov. 2019). ISSN: 2045-2322. DOI: 10.1038/s41598-019-53353-5.

[6] Sundeep Teki et al. “Distinct Neural Substrates of Duration-Based and Beat-Based Auditory Timing”. In: *The Journal of Neuroscience* 31.10 (Mar. 2011), pp. 3805–3812. ISSN: 1529-2401. DOI: 10.1523/jneurosci.5561-10.2011.

[7] Sundeep Teki et al. “Segregation of complex acoustic scenes based on temporal coherence”. In: *eLife* 2 (July 2013). Ed. by Dora Angelaki, e00699. ISSN: 2050-084X. DOI: 10.7554/eLife.00699. URL: https://doi.org/10.7554/eLife.00699.

[8] Quirin Gehmacher et al. “Eye movements track prioritized auditory features in selective attention to natural speech”. In: *Nature Communications* 15.1 (May 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-48126-2.

[9] M. Eric Cui and Björn Herrmann. “Eye Movements Decrease during Effortful Speech Listening”. In: *The Journal of Neuroscience* 43.32 (July 2023), pp. 5856–5869. ISSN: 1529-2401. DOI: 10.1523/jneurosci.0240-23.2023.

[10] G Naylor et al. “The Application of Pupillometry in Hearing Science to Assess Listening Effort”. In: *Trends in Hearing* 22 (Jan. 2018). ISSN: 2331-2165. DOI: 10.1177/2331216518799437.

[11] Sijia Zhao, Claudia Contadini-Wright, and Maria Chait. “Cross-Modal Interactions Between Auditory Attention and Oculomotor Control”. In: *The Journal of Neuroscience* 44.11 (Feb. 2024), e1286232024. ISSN: 1529-2401. DOI: 10.1523/jneurosci.1286-23.2024.

[12] Sara Alhanbali et al. “Measures of Listening Effort Are Multidimensional”. In: *Ear and Hearing* 40.5 (2019), pp. 1084–1097. ISSN: 0196-0202. DOI: 10.1097/aud.0000000000000697.

[13] Matthew B. Winn et al. "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started". In: *Trends in Hearing* 22 (Jan. 2018). ISSN: 2331-2165. DOI: 10.1177/2331216518800869.

[14] Mihaela-Beatrice Neagu et al. "Investigating the Reliability of Pupillometry as a Measure of Individualized Listening Effort". In: *Trends in Hearing* 27 (Jan. 2023). ISSN: 2331-2165. DOI: 10.1177/23312165231153288.

[15] Francesca Yoshie Russo et al. "Pupillometry Assessment of Speech Recognition and Listening Experience in Adult Cochlear Implant Patients". In: *Frontiers in Neuroscience* 14 (Nov. 2020). ISSN: 1662-453X. DOI: 10.3389/fnins.2020.556675.

[16] Sophia E. Kramer Adriana A. Zekveld Thomas Koelewijn. "The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge". In: *Trends in Hearing* 22 (2018). DOI: https://doi.org/10.1177/2331216518777174.

[17] Yue Zhang et al. "Pupillometry reveals effects of pitch manipulation within and across words on listening effort and short-term memory". In: *Scientific Reports* 14.1 (Sept. 2024). ISSN: 2045-2322. DOI: 10.1038/s41598-024-73320-z.

[18] Junchao Hu and Petra Vetter. "How the eyes respond to sounds". In: *Annals of the New York Academy of Sciences* 1532.1 (Dec. 2023), pp. 18–36. ISSN: 1749-6632. DOI: 10.1111/nyas.15093.

[19] Sundeep Teki et al. "Neural Correlates of Auditory Figure-Ground Segregation Based on Temporal Coherence". In: *Cerebral Cortex* 26.9 (June 2016), pp. 3669–3680. ISSN: 1460-2199. DOI: 10.1093/cercor/bhw173.

[20] Sophia E. Kramer Adriana A. Zekveld Thomas Koelewijn. "The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge". In: *Trends in Hearing* 22 (2018). DOI: https://doi.org/10.1177/2331216518777174.

# Appendix A
# Appendix

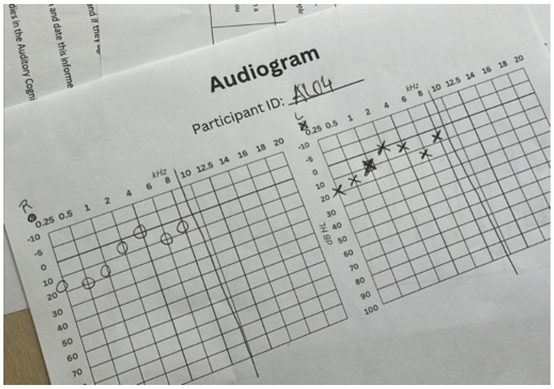


Figure A.1: Example of a pure-tone audiogram (PTA) for participant AL04

Figure A.2: Example of PTA testing (left: M Sinclair - right: A Le Bagousse/ALB)

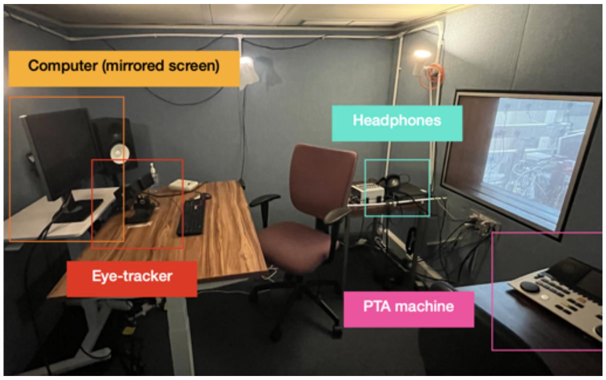
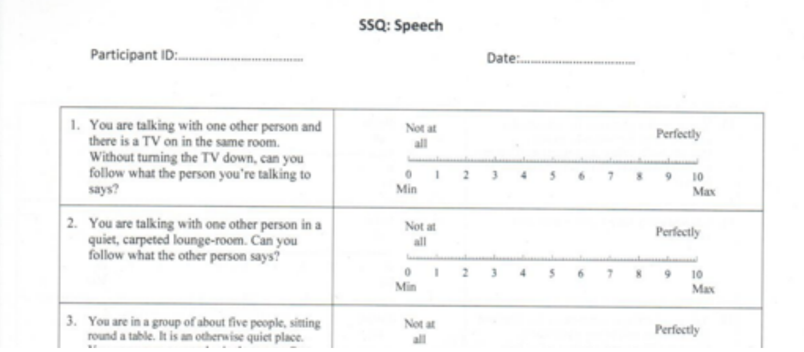


Figure A.3: Inside the booth (participant side)

Figure A.4: Outside the booth (experimenter side - ALB)



SSQ: Speech

Participant ID:..............................

Date:..............................

| | |
|---|---|
| 1. You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says? | Not at all Perfectly<br>0 1 2 3 4 5 6 7 8 9 10<br>Min Max |
| 2. You are talking with one other person in a quiet, carpeted lounge-room. Can you follow what the other person says? | Not at all Perfectly<br>0 1 2 3 4 5 6 7 8 9 10<br>Min Max |
| 3. You are in a group of about five people, sitting round a table. It is an otherwise quiet place. | Not at all Perfectly |

Figure A.5: Extract from the SSQ given in this study

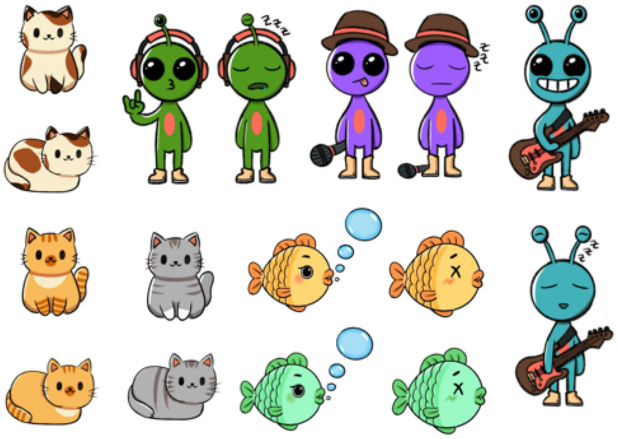


Figure A.6: New designs for the training app/website

*[Fig A.6] These are all the hand-drawn illustrations created for the training app and website. Each drawing has a double version that shows up when the character is inactive (e.g. the aliens sleep if they're not playing the related stimulus during a trial, and wake up when it's their turn to "sing"). The fish illustrations were designed for a control task not covered in this report, as its implementation is planned for a later stage beyond the scope of this internship.*

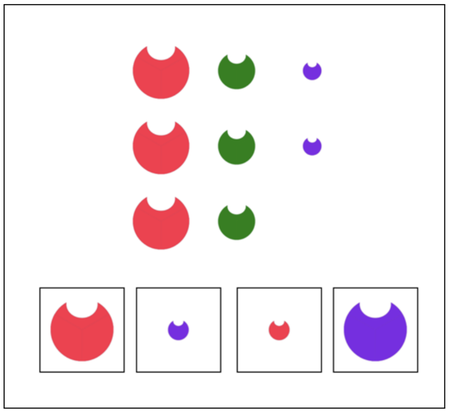
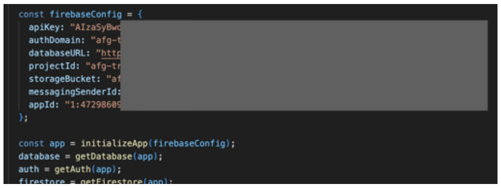


Figure A.7: 1st puzzle shown during the MTX reasoning test



```
const firebaseConfig = {
  apiKey: "AIzaSyBwc[redacted]
  authDomain: "afg-t
  databaseURL: "http
  projectId: "afg-tr
  storageBucket: "af
  messagingSenderId:
  appId: "1:47298609
};

const app = initializeApp(firebaseConfig);
database = getDatabase(app);
auth = getAuth(app);
firestore = getFirestore(app);
```

Figure A.8: main.js : Firestore log