UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

# Concurrent Speech Feedback for Blind People on Touchscreens

Pedro André Mendes Francisco

**Mestrado em Engenharia Informática**

Dissertação orientada por:
Prof. Doutor João Pedro Vieira Guerreiro
Prof. Doutor André Filipe Pereira Rodrigues

2022

# Acknowledgments

I am deeply thankful to my advisors, *João Guerreiro*, and *André Rodrigues*, for their critical help and suggestions throughout this work. Without them, none of this would be possible.

I also want to thank *Fundação Raquel e Martin Sain* for their support during the study and all the participants who volunteered to participate.

Finally, I want to thank my family and friends who encouraged me and never gave up on me. A special word to my girlfriend Inês, who was my main supporter throughout this work.

*To a more accessible world.*

# Resumo

O uso de *smartphones* é especialmente exigente. Estes dispositivos para além de possuírem uma grande escala de aplicações, cada uma com o seu objetivo e comportamento, também têm um conjunto de elementos físicos limitados, sendo na sua maioria compostos por um único ecrã tátil. Os utilizadores cegos têm que recorrer a outros métodos para interagir e navegar no seu *smartphone*. Os *smartphones* modernos já incluem um conjunto de serviços que tornam a interação com o *smartphone* acessível. Sendo que, os utilizadores cegos e com deficiências visuais remetem para os leitores de ecrã para conseguirem interagir com o seu dispositivo. Os leitores de ecrã disponibilizam um conjunto de gestos que permite aos utilizadores navegar no seu telemóvel e, leem em voz alta o conteúdo focado no ecrã. Apesar de a experiência para os utilizadores ser acessível e positiva, algumas tarefas podem ser consideradas ineficientes e incómodas. Nomeadamente, o facto de quando uma notificação é recebida, o conteúdo, que era anunciado é interrompido em favor de ler o conteúdo da notificação. Quando o conteúdo da notificação acaba de ser lido, o utilizador perde o progresso que tinha feito no conteúdo original. De outra forma, notificações não prioritárias não interrompem a tarefa atual, no entanto, levam a que o utilizador tenha que remeter para a barra de notificações caso queriam recuperar o conteúdo da notificação recebida. Outra das limitações dos leitores de ecrã é o facto de o consumo de informação estar limitado, associado ao único canal de áudio utilizado para ler o conteúdo aos utilizadores. Estudos anteriores [35] [36] também indicam que os gestos utilizados para navegar no telemóvel apresentam uma longa curva de aprendizagem, muito devido ao facto dos tutoriais usados serem confusos e à falta de *feedback* quando os gestos são executados e quando estes não são corretamente identificados pelo sistema. Este trabalho explora diferentes formas de potenciar a maneira como a informação é transmitida nos smartphones, através da utilização de: diferentes canais de áudio reproduzidos em simultâneo; áudio espacial; e/ou adaptação da velocidade de leitura, diferentes vozes e outras características conforme o contexto/aplicação do utilizador.

Neste trabalho exploramos 5 cenários diferentes:

1. **Interrupção de tarefas.** Dois canais de áudio são utilizados para reproduzir em simultâneo uma notificação sem interromper a leitura do conteúdo que era anunciado.

2. **Aumento do consumo de informação.** Exploramos o uso de diferentes canais de áudio em simultâneo para apresentar diferentes tipos de conteúdo aos participantes.

3. **Propriedades de texto.** As propriedades de texto como o negrito ou itálico são anunciadas junto do conteúdo que as acompanha. Estas propriedades são apresentas de diferentes formas: introduzindo uma pausa, onde a propriedade de texto é lida antes de continuar com o conteúdo original; ler a propriedade em simultâneo com a palavra ou conjunto de palavras a que diz respeito; introduzindo uma pausa onde para além da propriedade ser anunciada, um som correspondente também é reproduzido; reproduzir um som em simultâneo com a respetiva palavra que apresenta uma determinada propriedade.

4. **Mapa.** Exploramos o som espacial como forma de informar o utilizador acerca do quão perto ou distante este está de uma localização em específico. Aqui, quanto mais perto estiver da localização, mais será nitidamente ouvida a localização, enquanto à medida que se afasta, esta leitura desvanece.

5. **Interação com o *smartphone.*** Cada gesto utilizado para navegar no smartphone tem um som correspondente. Adicionalmente, a leitura dos elementos presentes no ecrã, como, por exemplo, um botão, foi substituída por um som correspondente reproduzido quando o utilizador interage com o elemento.

De forma a avaliar os cenários que pretendíamos explorar, conduzimos um estudo com 10 participantes cegos cuja experiência com smartphones varia entre novato e especialista. Durante o estudo, para cada um dos cenários explorados, perguntamos aos participantes para descrever a sua experiência, o que poderia ser melhorado ou em que situações as funcionalidades exploradas poderiam ser úteis. Os resultados obtidos indicam que a utilização de vários canais de áudio em simultâneo é benéfico. Apesar de os participantes expressarem alguma dificuldade em relembrar o conteúdo que ouviram, conseguiram na sua maioria identificar corretamente o tópico dos conteúdos que estavam a ouvir, especialmente quando apenas eram reproduzidos dois conteúdos em simultâneo. Por outro lado, todos os participantes mencionaram que a utilização de diferentes canais de áudio em simultâneo funciona melhor quando o segundo canal de áudio é utilizado apenas para mensagens curtas, caso das notificações. Desta forma, a mensagem não retira o foco do conteúdo que estavam a ouvir, com o valor acrescentado de a tarefa atual não ter que ser interrompida. Aqui, os participantes também realçaram a necessidade do sinal reproduzido antes de uma notificação ser recebida. Sendo que, com este sinal, conseguem mudar o foco, caso contrário necessitavam de "estar num estado permanente de alerta" como indica um dos participantes. Durante os cenários, para além de canais de áudio em simultâneo, também exploramos a espacialização do som. Numa primeira fase, como forma de distinguir os diferentes conteúdos, ao posicionar cada canal numa posição específica. 8 dos participantes indicou preferir ouvir áudio em simultâneo recorrendo à espacialização do som, indicando ser mais natural e aproximando-se daquilo que experienciam durante o dia a dia. Adicionalmente, para os diferentes canais de áudio foram utilizadas diferentes características. Nomeadamente, diferente timbre, tom ou diferentes tipos de voz, masculino ou feminino. Os resultados indicam ser crucial distinguir os diferentes conteúdos ao utilizar vozes diferentes. Os participantes indicaram que só assim não confundiam as notícias que

ouviam. Apesar da nossa tentativa em utilizar vozes diferentes, alguns participantes indicaram que algumas delas eram muito parecidas. Esta situação foi mais evidente, quando reproduzidos 3 ou 4 conteúdos diferentes em simultâneo e, ao utilizar som espacial, quando no mesmo lado eram reproduzidas duas vozes masculinas ou femininas. Por outro lado, diferentes participantes expressaram diferentes preferências no que toca às vozes utilizadas, reforçando ainda mais a necessidade de cada utilizador poder adaptar as características de cada voz à sua preferência. Noutro cenário, o do mapa, os participantes expressaram a sua satisfação com o facto de obterem um *feedback* extra sobre o quão distantes estão de uma determinada localização. Os participantes indicaram que a utilização da espacialização do som como forma de indicar a distância como sendo uma mais-valia, mencionando que de nenhuma forma impacta negativamente aquilo que utilizam no seu dia a dia e que não haveria problema se esta opção estivesse permanentemente ativa. No cenário das propriedades de texto os participantes indicaram ser valioso poder saber a formatação do texto que estão a ler, alguns até mencionando que já o fazem quando utilizam os leitores de ecrã do computador. De entre as opções utilizadas para apresentar estas propriedades, a opção com uma pausa, seguida da leitura da propriedade e de um sinal sonoro quando a respetiva palavra ou conjunto de palavras termina foi a preferida entre os participantes. A justificação é que com a combinação da pausa e do sinal sonoro, conseguem inequivocamente identificar que palavras estão destacadas com a respetiva palavra e através da pausa o conteúdo não se confunde com a propriedade. Por outro lado, os participantes indicaram preferir ouvir a propriedade e não utilizar um som para o efeito, justificando com o facto de assim não precisarem de saber o que cada som representa. No entanto, alguns participantes admitem que com o tempo esta seria talvez a opção utilizada, sendo que tornaria o texto menos pesado. Para o último cenário explorado, os participantes indicaram o quão importante é obter feedback imediato acerca das ações feitas no *smartphone*, neste caso, após cada gesto efetuado. Muitos dos participantes já são considerados especialistas na utilização do seu telemóvel, não obstante, indicam que a presença deste *feedback* continua a ser útil, na medida que teriam sempre a certeza daquilo que o sistema assumiu com a ação realizada e com o facto de já estarem habituados a este tipo de *feedback* sonoro durante a utilização do *smartphone*, este simplesmente não acontece sempre.

Os nossos resultados indicam que as soluções exploradas neste trabalho são úteis nas atividades do dia a dia e, na sua maioria, não impactaram negativamente a utilização do smartphone. No entanto, é importante cada utilizador poder ativar as opções dependendo da sua preferência. Como exemplo, um participante indicou que ao ler livros teria sempre a leitura das propriedades de texto desativada, enquanto que ao realizar um trabalho académico a opção estaria sempre ativa. Adicionalmente, os resultados obtidos durante este estudo permitiram-nos perceber como utilizadores cegos interagem com os seus *smartphones*, que funcionalidades estão em falta e de que forma podemos apresentar soluções alternativas para tornar a interação com *smartphones* mais eficiente.

**Palavras-chave:** Acessibilidade, Saída de voz, Sonificação, Deficiência Visual, Concurrent speech

# Abstract

Smartphone interactions are demanding. Most smartphones come with limited physical buttons, so users can not rely on touch to guide them. Smartphones come with built-in accessibility mechanisms, for example, screen readers, that make the interaction accessible for blind users. However, some tasks are still inefficient or cumbersome. Namely, when scanning through a document, users are limited by the single sequential audio channel provided by screen readers. Or when tasks are interrupted in the presence of other actions.

In this work, we explored alternatives to optimize smartphone interaction by blind people by leveraging simultaneous audio feedback with different configurations, such as different voices and spatialization. We researched 5 scenarios: Task interruption, where we use concurrent speech to reproduce a notification without interrupting the current task; Faster information consumption, where we leverage concurrent speech to announce up to 4 different contents simultaneously; Text properties, where the textual formatting is announced; The map scenario, where spatialization provides feedback on how close or distant a user is from a particular location; And smartphone interactions scenario, where there is a corresponding sound for each gesture, and instead of reading the screen elements (e.g., button), a corresponding sound is played. We conducted a study with 10 blind participants whose smartphone usage experience ranges from novice to expert. During the study, we asked participants' perceptions and preferences for each scenario, what could be improved, and in what situations these extra capabilities are valuable to them.

Our results suggest that these extra capabilities we presented are helpful for users, especially if these can be turned on and off according to the user's needs and situation. Moreover, we find that using concurrent speech works best when announcing short messages to the user while listening to longer content and not so much to have lengthy content announced simultaneously.

**Keywords:** Mobile accessibility, Speech output, Sonification, Blindness, Concurrent speech

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Mobile device interaction is visually demanding and, as such, poses several challenges for people with visual impairments. With smartphones mainly consisting of a touchscreen with little to no physical buttons, this group of users must rely on types of feedback other than touch. Auditory feedback is the chosen substitute most of the time, with screen readers playing a significant part in its adoption. They are already preinstalled on most user devices and work well with little to no configuration. This software enables reading information out loud and interacting with the device exclusively through gestures, without the need for visual aid. Moreover, even if it is primarily used by the visually impaired, every user can take advantage of its features. Sighted users, for instance, can use them in particular situations, such as walking, where they cannot focus on the device and must pay attention to their surroundings.

Screen readers are an ideal solution for consuming information. However, despite their adoption rate, users have reported several challenges in using it [36, 35]. For starters, to provide the best experience possible, it relies on developers following accessibility guidelines, such as alternative text for images or identification of headers that screen readers can use to allow users to skip between sections, which are often ignored. Additionally, some characteristics are lost when depending exclusively on audio. For instance, when exploring the device by swiping, the element's position on the screen is not transmitted to the user. Another area affected by the usage of screen readers is the task of scanning for information in a large document, especially unexplored ones. The single sequential audio channel is a bottleneck for the number of information users can consume. While sighted users can analyze a document by quickly glancing at its structure or text characteristics, namely text font or color, people with visual impairments must sequentially go through every document's element. As a workaround, some users, especially those experienced with screen readers, playback the audio at speeds multiple times the regular rate to achieve a similar experience to sighted people. Users also experience interruptions when receiving notifications or when they need to consult sporadic information such as current time or smartphone battery. With all this, smartphone interaction is still a problem for blind people, making the help of sighted people still needed in some situations, such as when the user is unfamiliar with the application or the content has changed unexpectedly. This difficulty is especially notorious with novice users. There are still some scenarios left to be explored with current smartphone accessibility solutions to make its interaction more efficient. In particular, the viability of transmitting information using multiple voices simultaneously while allowing the user to configure its characteristics, such as type of voice. Current smartphone screen readers rely on a single sequential auditory channel to provide feedback to the user. In contrast, studies show this approach might not be efficient as users can comprehend multiple voices simultaneously, even when these are played at faster rates [13], without compromising the intelligibility of the information. This work explores the viability of augmenting how information is transmitted in smartphones through the use of: concurrent audio streams; spatial audio; and/or adaptation of speech rates, voices, and other features according to user context/application.

This work explores the viability of augmenting how information is transmitted in smartphones through the use of: concurrent audio streams; spatial audio; and/or adaptation of speech rates, voices, and other features according to user context/application.

Our work was divided into 4 phases: 1) Analysis; 2) Development; 3) User study; 4) Evaluation.

During the analysis phase we started by doing a literature review to understand how blind users use their smartphone and what issues do they normally face. Given the limitations of current approaches to audio feedback on mobile screen readers and the potential of concurrent speech, spatial audio and augmenting screen reader interactions suggested by prior literature, we identified and developed a set of scenarios to optimize information consumption and interaction awareness. We explore how to augment 1) Notifications, 2) Document Skimming, 3) Readability of Text Properties, 4) Touchscreen Gesture, and 5) Map Awareness.

After concluding the analysis phase, the next step was to develop an Android application where the different scenarios could be explored. To validate the different scenarios in our work, we recruited 10 blind participants with different levels of expertise to participate in a user study. Using an Android smartphone (SM-F926B with Android 12) with the developed app pre-installed and the headphones provided, participants explored each of the five scenarios developed and reflected on the benefits, disadvantages, utility and improvements.

After the study, we transcribed the participants' feedback which was then analyzed using an inductive coding approach. This involved examining the transcripts, through several rereads, and identifying a total of 42 codes related with user profile, perception, preferences, smartphone behaviour, concurrency and spatialization. We coded and organized data into categories, making notes of any interesting observations or comments that were made by the participants. We then reviewed the findings and discussed them between us.

Finally, we reflected on the findings and discussed the results which are presented in the discussion section of this work

## 1.1 Motivation

This dissertation aims to understand how we can make the experience of interacting with smartphones more efficient. The main goal is to explore how the auditory channel can be more efficient by employing different configurations or simultaneous auditory feedback. By using concurrent speech with different audio characteristics such as timbre, speech rate or pitch, and spatial audio, we aim to improve or provide an alternative for handling different scenarios detailed in the next paragraph. We also intend to introduce users to areas not as explored in today's default smartphone accessibility mechanisms, as is the case with feedback about text formatting or using spatialization to provide additional feedback on how close or distant they are from a specific place.

Based on the explored related work, we understand that the unique sequential audio channel provided by mobile phone screen readers can be limiting. We will study how we can provide a more efficient way of consuming information using multiple concurrent voices. Moreover, by combining spatial audio, we can also make listening to concurrent sounds easier.

Furthermore, we are also aware that, on the one hand, our solutions should be easy to integrate into modern screen readers. However, on the other hand, if that is not the case, they must not affect the user's everyday usage if they decide to use both our solutions and traditional screen readers.

## 1.2 Contributions

Our main contributions with this dissertation are:

- Literature review on smartphone accessibility and different ways of interacting with it, through which we found some limitations.

- Identification of different scenarios where concurrent speech and spatialization can provide alternative ways of consuming information and interacting with a smartphone.

- Exploration of alternative ways of consuming information in smartphones through concurrent speech and additional features currently not present in smartphone screen readers, such as announcing textual formatting or using spatialization to provide feedback about the distance.

- Validation of the explored scenarios in a study with 10 visually impaired participants. The results indicate that users are open to new features to be included in their smartphones, considering that they are easy to turn on and off. Additionally, our results suggest that concurrent speech works well to deliver short messages while listening to longer content. Alternatively, concurrent speech can also be used to augment information consumption, but the results indicate that it provides better results when limited to 2 concurrent sources.

## 1.3 Document's structure

- **Chapter 2 - Related Work.** We provide an overview of the literature review. We start by giving a background overview of smartphone accessibility. Then, several solutions present in the literature are explored to show what areas are still lacking and how our work can help improve blind people's smartphone efficiency.

- **Chapter 3 - Design.** We present the explored scenarios in this work. We start by explaining the design motivations and the use case scenarios that motivate their implementation. Then, we describe what each scenario accomplishes and in what ways they can be valuable.

- **Chapter 4 - Implementation.** We start by displaying an overview of the system. Then, we go into how we accomplished each solution and our thought process when developing them.

- **Chapter 5 - Evaluation.** We conducted a user study with 10 visually impaired participants with different levels of smartphone expertise, from novice to expert. During the study, we asked participants their thoughts for each of the explored scenarios, concretely in what situations they could be helpful if they would and when they would activate them, and how we could improve the presented solutions. We present the results.

- **Chapter 6 - Conclusion.** Our final thoughts on the work and prospects for forthcoming work.

# Chapter 2

# Related work

## 2.1 Understanding Smartphone Accessibility

Smartphone usage is demanding. There are thousands of applications, each with its purpose and unique user interface. Long gone are the days when it was only used for communication. It has now become a full-fledged device where several activities can be performed. The amount of options available makes users feel lost and even consider these devices inaccessible [36, 35, 11]. Mobile device usage is especially hard for visually impaired users since they cannot rely on their vision to access the screen's content or logical structure. To make smartphones more accessible and help this group of people, smartphone accessibility has been the target of multiple studies and, as a result, evolved over the years. Before modern screen readers existed, touchscreen interaction was not accessible for visually impaired users. This group of people struggled with item selection as well as a lack of audio and haptic feedback. Slide Rule [20] was one of the pioneers in touchscreen interaction, introducing a group of gestures used to interact with touchscreens and providing feedback over the items on the screen. In this project, users can navigate through screen content, such as lists, by swiping with one finger, and as the items are being transversed, their content is announced. Other gestures introduced include using double-taps anywhere on the screen to select a particular item or using an L-shape gesture to browse multilevel content, such as browsing a list of artists by swiping down on the left and their songs by moving to the right. The possibility of tapping anywhere on the screen makes users not worry about being within item bounds, as they do not need to press on a particular position to select that particular element. Now, current screen readers, such as TalkBack [1] for Android or VoiceOver [2] for iOS, are now preinstalled on current devices and use some of the concepts introduced in Slide Rule to make the smartphone interaction more accessible. Screen readers translate the screen content into an accessible audio format which can then be customized regarding pitch, speed, or even voice gender. There are two main ways of interacting with the device using these screen readers. The first is a defined set of gestures to explore the screen's content and perform specific actions on the device. These actions include moving from paragraph to paragraph, scrolling the screen, or copying and pasting content. The

[1] Google TalkBack. https://github.com/google/talkback, (Last visited on October 15th, 2021)

[2] iOS VoiceOver. https://www.apple.com/accessibility, (Last visited on October 15th, 2021)

second approach screen users provide exploration by touch, where, as the user drags his finger on the screen, every item on the screen is announced. When the user pauses at a specific position, the screen reader might suggest a particular action related to the hovered item, such as how to view or activate the content. Other accessibility features, also built-in into most modern devices, include speech recognition, used to control the smartphone with verbal commands, switch access supporting interaction with the device without pressing the touch screen, or even content magnification to increase the size of the information displayed. Blind users can use these accessibility services to perform their daily activities more independently without relying on sighted users.

Despite the evolution of smartphone accessibility and their respective tools, this field remains an important research topic as users continue to face several challenges. For one, applications do not share the same structure, making it hard to apply the same strategies to new applications. The mental model formed before is not valid anymore, increasing user cognitive overload. This situation is aggravated when dynamic content is involved, invaliding any assumptions the user might have developed earlier. Sometimes the challenge is not even inherent to the task at hand but to the user's lack of knowledge on what he can use to complete it successfully. The tutorials used to teach the visually impaired users the available gestures, integrated with TalkBack or VoiceOver, are seen as non-intuitive [35], sometimes going as far as demotivating the user of using such gestures. To correctly perform these gestures, the user needs to complete them according to the defined speed and location while also missing negligible feedback after performing such gestures. To mitigate the user's lack of knowledge, [11] endorsed the creation of a recommendation system that would automatically suggest or even apply accessibility features based on some criteria, such as setting the volume abnormally high, which might indicate hearing problems. Four prototypes were created to assess the solution viability. An example was the creation of a font size recommender which would suggest enabling the font size modifier if the phone was too close to the user's face. The researchers also mentioned the concept of group recommendations which would suggest features based on others already active on the user device, for example, the suggestion of larger text if the bold text option is enabled. Some participants, both blind and sighted, had the chance to go through these prototypes, which they found extremely useful. One user even stated that he thought accessibility features were exclusive to impaired people, so he never considered exploring them for his needs. Sonification can also be leveraged to support users when learning gestures. In the work of [29], two techniques are explored to give feedback while performing gestures on a touchscreen device. The first uses audio cues to describe the gesture and then provides corrective feedback on what the user should do differently for the gesture to be correctly recognized by the system. The last consists of simple messages such as "draw faster/ slower" or "try drawing the gesture narrower/taller.". The second technique, used in the research study, provided richer details to the participant using sonification, based on the combination of pitch and stereo. A sonified preview of the gesture is given to the user together with a text description before he tries to draw the gesture on the screen. When the gesture is performed, an audio cue is replayed when the system fails to recognize the gesture or when it successfully does so, providing an efficient form of feedback. The study conducted by the researchers found that participants usually preferred verbal input since it was seen as more precise. Despite their preference, sonification was viewed as a complement to verbal feedback and as a clear advantage for providing feedback about aspects such as speed.

Despite mobile accessibility guidelines being constantly worked on by several entities, such as the Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C), they are not as defined as their Web content accessibility counterpart. Developers continue to not give enough focus on developing accessible mobile apps. Analysis done on the accessibility of several android apps [9] found many issues. These include missing text alternative to an image, small items, or lack of text language characterization (ie. application in portuguese and specific word in english resulting in the wrong intonation by the screen reader). Privacy is also seen as a topic of extreme importance. Users are afraid bystanders will hear their conversations or what they are doing. They also have to often rely on sighted people's assistance, which makes them entrust their mobile device to another person posing a security risk. Another security concern is the fact that visually impaired users usually do not protect their smartphone with a password because they consider it an inconvenience [6] [4]. If they choose to have a password, the screen reader will read every character. On the other hand, if the screen reader is not active, the user will not have the necessary feedback to know what he is currently typing or if he made an error. Browsing new documents is also seen as intimidating, especially when dealing with large documents. The sequential nature of the screen reader makes it cumbersome for the user to scan for information, as he has to sit through all of it before reaching the section of interest. With current solutions, the user has to move his finger on nearly every content before understanding what is important to him or apply other strategies such as navigating from heading to heading, taking the risk of missing important information.

Other than the built-in screen readers in Android and iOS, other screen readers for smartphones have been built to provide an alternative way of interacting with the smartphone. One such case is ShinePlus [3]. These custom screen readers allow users to label the screen elements according to their needs. This is especially important when there are elements that developers did not label. ShinePlus also provides other features and things done differently than traditional screen readers. However, at the time of writing, ShinePlus is not available in the play store and is not easily accessible.

All in all, accessibility on smartphones has been rapidly improving, with many valuable features already being built into today's smartphones. For instance, Talkback has recently been revamped and can now be configured to only read the document headings as the user swipes on the screen. This new feature makes looking for information in a document easier since the user can first go into the desired section and then read the content from there. Despite the several accessibility features smartphones offer out of the box, other custom solutions keep being developed to mitigate existing gaps. There is always room for improvement. Blind users still do not have the same smartphone experience as sighted people. There are still issues like tasks that are interrupted

[3] Shineplus. http://www.atlab.biz/en/html/platform.html, (Last visited on September 20th, 2022)

when listening to content with a screen reader, namely when receiving a notification. Other problems include cumbersome processes when dealing with large pieces of data where the user must sit through all of it to reach the desired information.

## 2.2 Audio based interaction and feedback

Visually impaired users mostly rely on audio modalities to enable actions and consume information on their smartphones. Screen readers remain the most exploited feature for visually impaired users' interaction, justified by their ease of use and the fact that modern smartphones already come with them out of the box. These have come a long way since they were implemented and can now easily be configured to present the speech in different languages and configure properties like pitch or speed. Despite answering many user needs, there are still some gaps in its features which pushed the community to explore other solutions. For instance, the text is read in a linear fashion rather than in a way that represents the spatial position currently presented to the user. This way of giving the information is considered slow as the user must sit by every item until he reaches the relevant section [14] [15] [27]. For example, when reading an email, users must first listen to the header, and only then can they hear the main content, which can be cumbersome. Screen readers allow an area of the screen to be clicked, which then moves the reader's focus to this specific section from which the content can be read. However, since blind users lack the spatial awareness provided by vision, the probability of focusing on the wrong place is high, often leading to different workarounds. More experienced users will increase the speech rate of their screen reader to its limit. In contrast, novice users will keep the default configuration, afraid they will miss any critical information [17] [33]. Another problem is that users cannot listen to secondary information while simultaneously hearing their primary audio source. In the session conducted in [21], participants mentioned that while connected to a meeting, they could not read missed calls or notifications nor check the time without losing focus.

Due to the sequential nature of screen readers, several solutions were explored to reduce the slow output and lack of spatial representation. To provide a dimensional model of the screen to visually impaired users, [27] introduced the concept of tag thunder to separate web pages sections into logical groups. The goal was to provide the ability to mimic the strategy sighted people used in skipping irrelevant page segments, such as ads. The page would be split into different sections, each with its own keyword to summarize the division's content. Each piece would then be played concurrently to the user while preserving the location of the content, as seen on the respective page, by employing spatial audio. This solution allowed users to quickly grasp the page content despite feeling that the task required extra concentration. The work done in [2] investigated the use of summarization techniques to shorten the content present on the screen. Nevertheless, the authors argued that it was important for users to use a shortcut to switch between the summarized and full content view quickly. The session carried out by the researchers found that participants were able to answer most of the questions asked about the data presented to them. However, some found the task hard since certain information was lacking in the summary, and there was no easy way to further explore a specific section to understand more about it. Other solutions investigated concurrent speech to present the information faster to the user while still maintaining the expected comprehensibility level for the respective content. [15] evaluated the performance and efficiency of the user's ability in identifying the relevant audio source while multiple concurrent audio sources are being played. Inspired by the *Cocktail Party's Effect's* phenomenon, which states people can focus on one primary audio source amidst other noise. The results show that participants can quickly identify the relevant font when using two concurrent voices. Furthermore, while the performance dropped when using three voices, the participants could still complete the task successfully. However, it was predominant that the use of four or five voices was exaggerated, despite results showing that the percentage of completeness can still be within acceptable boundaries if some loss of information is satisfactory for the task at hand. While the conclusions were positive, the participants' capacity to retain the consumed information decreased significantly, indicating that this solution might not suit these types of tasks. In [14] the authors compared the use of concurrent audio sources versus increased speech rate for the task of scanning for information and identifying the relevant detail. To provide a fair comparison between the two approaches, the efficiency of both is always compared taking into account the same amount of information. Meaning a single voice would have to be played twice as fast to accomplish the same result as two voices with the default rate. The results show that using concurrent speech really shines when dealing with large amounts of information. While with one voice, the speed the content has to be played to match the concurrent speech output, makes the content indistinguishable. With the use of 2 or 3 simultaneous voices, users can maintain a lower speed while still correctly identifying the relevant snippet of information.

For text entry, visually impaired users can leverage the screen reader to announce each character present on the keyboard as the user moves over it. As the characters are swiped, the user can stop at any time and select the desired character once he listens to it. However, it has been concluded that this type of approach results in slower typing rates while also having higher levels of errors [30] [5]. A common way to deal with errors is the introduction of spellcheckers, where an alternative word is suggested. With screen readers, these suggestions interrupt the flow of the user's typing. The user must stop typing and manually explore the given suggestions read to him individually. Whereas if he continues typing, he will not leverage the benefits of the word suggestion mechanism. To take advantage of the ability that users have in interacting with a mobile device using both hands or multiple fingers, SpatialTouch [16] enables keyboard interaction using two fingers simultaneously. Each hovered character is read to the user, and the spatial representation of the keyboard key is preserved by having sound emitted from the corresponding positions on the 3D audio space. However, while the solution is promising, it did not improve the typing rate of the participants when compared to traditional solutions. In [28] the authors took advantage of concurrent speech to play word suggestions to the user while he is typing, thus not interrupting its flow.

Screen readers also have a hard time transmitting graphical information, such as images, due to their complexity. Thus, screen readers must rely on the alternative text associated with the graphical information, which is not always present. EdgeSonic [41] was developed to convey image shape through touch, resembling the perception by feel, visually impaired users use to identify a figure. In this solution, sound is generated when an image edge is touched, or no edge is present in the area. Different sounds are played based on the distance to the nearest edge. Results showed that users were mostly able to determine the image presented to them, especially after training and exploring the framework. To provide information about maps, [32] produced a solution that adds an extra overlay to the smartphone navigation system, which the user can exploit to gain feedback about roads extension. While the user touches a road on the touchscreen, both vibration, and speech stating the road's name. With this, the user gains a perception of the road's length, helping him navigate. [22] explored the ability of sonification properties such as volume, tempo, or type to help in identifying different scenarios with the corresponding sound. Multiple sounds were played concurrently in the sessions used to measure these audio characteristics, each from a different position in the 3D space. Auditory segregation regarding frequency, pitch, or change in amplitude was described as one of the main reasons participants were able to easily recognize the different components, even in the presence of simultaneous audio sources.

Despite the advancements in audio based interaction and feedback in mobile devices, some areas can still be further improved. For instance, users should be able to consume information faster depending on their use case when skimming or scanning information. The location of the elements present on the screen is one thing that still requires users to remember it between smartphone usage sessions. They do not have the tools to quickly create a mental model of the element's spatial location. Some users employ strategies to mitigate this problem, such as grouping their home screen items in rows according to a specific category (e.g., the first row has news apps, the second entertainment, and the last accessibility applications) [17]. Whereas this can work for system icons (smartphone home screen), application specific elements cannot be rearranged to a location that makes sense to the user. So, the tools to efficiently interact with the smartphone and receive feedback can still be refined, still being the target of active development in the research community.

## 2.3 Voice based interaction

Voice commands are the second most used accessibility feature, right after screen readers. These commands allow a hands-free interaction with the smartphone through speech. By employing voice commands, the user can easily navigate through the screen's content or even write a text message. Users do not have to worry about navigating to a section, finding the desired target, and clicking on it. A simple spoken command is enough to perform the appropriate action.

Today's mobile devices already allow voice interaction out of the box, under Siri for iOS and Google Assistant for Android. These are called voice assistants (VA's), and besides interpreting voice commands, they can also be used to consume information. Through artificial intelligence, these frameworks can speak in a more human-like manner. When VA's were first introduced, they were restricted to commands that performed system-level actions, such as calling a contact or turning bluetooth, not being tailored for visually impaired users' needs. This group of people requires a hands-free interaction not only for system actions but also for any specific application they might use. So, particular VA types were created, especially suited for impaired users, such as JustSpeak [42]. This solution extends voice commands to any android application by taking advantage of the application metadata, such as labels name. Issued commands are first transformed into text using Google Automatic Speech Recognition services, parsed by the framework's processor, and then matched with the objects on the screen. As an additional feature, *JustSpeak* can interpret multiple commands, resulting in different actions, using a single speech command, such as *"Open Gmail then refresh"*. The work done in [10] is another example of an application created to enable smartphone interaction through speech. It divides the user's screen into a square grid where each cell represents a position on the screen. To interact with the smartphone, the user must announce a grid position, for example, B3, and then the framework will automatically click on the element in this position. Despite its usefulness, it still requires the user to know what is present in a particular space, so it might not be a good fit for blind users. Nowadays, current solutions already allow interactions with all kinds of applications. For instance, Google Assistant has the concept of app shortcuts and actions that the users can issue to quickly access any application functionality, such as *"Order pizza from Domino's"*. Despite this impressive ability, Google Assistant still requires some work on the developer's side. The development team has to implement a mechanism called built-in intents which the framework then uses, together with its natural processing language mechanism, to interpret the command and execute the user's request on the specific application.

Other works in the community extended voice command features and used them for purposes other than interacting with the device. In particular, *Hint' Me* [34] was created as an accessibility service for Android, which blind users, particularly, can leverage to ask any question about a specific application. Users interact with the framework by first clicking on the overlay made available, indicating they want to ask a question, and then stating their problems or needs as they usually would with a friend. Once this is done, *Hint' Me* saves this question, alongside the context needed (name of the application, a screenshot of the user's screen, metadata about the screen's elements, among other things), in a shared knowledge base. In turn, a group of users, called volunteers, can answer these questions after they are validated. When answers are submitted, the author of the question receives a notification which he can further inspect to see the details of the answer. The framework also allows users to consult previously asked questions and their respective answers for a given app or a particular element within it if they highlighted it previously. So, the solution viability will keep improving as the number of gathered answers increases, minimizing or even replacing blind users' need to request their friends' assistance. Another application that eliminates the need for nearby assistance and provides autonomy to the non-sighted user is *Be My Eyes*. This solution connects blind users with sighted volunteers via a video call. The first group can ask for help with any problem they might have, such as checking if the television is on or navigating through new surroundings. However, this application is not suited for in-app questions like *Hint' Me* since it only uses the smartphone's rear camera to provide the volunteer with the information they need to answer the user's demands.

Voice interaction is used to interact with the smartphone in a hands-free manner, remaining a fundamental feature for the visually impaired and sighted users, especially in circumstances where physical interaction is not possible. Despite its utility, it is mainly used to input information and not consume it.

## 2.4 Audio and speech perception

Screen readers use synthetic speech to transmit information, and as such, they must be able to clearly deliver messages to the listener. Speech rate, pitch, or timbre are some metrics that can influence the message intelligibility. Most speech synthesis technologies now come with built-in mechanisms that support the configuration of these voice characteristics. Voices can be arranged with different accents and languages, and some can even be made to look more human-like. However, one thing that's still lacking is automatically adapting the voices to the different environments. The audio should match the user's expectations. For instance, when reading the news, a more expressive and slow tone is expected, while in a navigation application or in a more noisy location, the communication should be sufficiently loud. [40] details some of these scenarios stating that there is no golden standard for text-to-speech voices. What works for one application is not guaranteed to translate well into another. The listener's specific needs should always be taken into account. The user should have full ownership of the voice's configuration according to what he feels works best. Tech-savvy users also use applications such as *Auto TTS* to automatically switch text-to-speech language according to the context they are currently in [17]. For instance, reading news online in english while reading their friend's messages on *WhatsApp* in their native language. Without the aid of third-party applications, users must manually change this configuration.

Sighted and visually impaired users differ in their ability to understand speech at faster rates. For once, blind users have more practice dealing with voice synthesizers knowing how they differ from daily conversations, such as having the punctuation marks announced, which might make the speech incomprehensible. Moreover, it is believed that blind users can perform better in hearing activities since they do not use their neural capacities to process visual aspects. The study conducted in [38] shows that sighted users are not able to comprehend speech at rates faster than ten syllables per second, while blind users, who are proficient in screen readers usage, maintain their performance level at 18 syllables per second. However, the performance of both groups was affected when playing the voices in a more natural way, contrary to what was observed in other studies [31]. The authors justified this finding in the fact that humanlike voices are not well suited for fast speech rates since information will be greatly condensed and word pronunciation will differ significantly from the user's daily experience. The intelligibility is also affected by other factors such as age or listening ability. [19] reports that the user's age plays a significant role in the drop of performance in understanding speech at a faster rate. The study participants had to spend more time processing the incoming information justified by hearing problems and the cognitive changes that come with age, such as the processing speed, memory, or attention deficit.

Factors such as age or level of expertise also affect the intelligibility of the synthesized speech [37]. For instance, younger people (under 25) are more accurate when transcribing speech played at faster rates. The same applies to users who use synthesized speech daily compared to novice users.

All things considered, users must be able to configure voice synthesis characteristics, such as speech rate, timber, or pitch, so that they can adapt it to their needs or expectations. For instance, older people might need a slower speech rate or higher volume so the cognitive overload is not too much and the voice can be clearly heard. Expert users might also be more acquainted with screen readers and can take advantage of faster rates or multiple voices played simultaneously to consume information faster. As stated before, no rule works for every situation and person, so these configurations can never be static.

## 2.5 Haptics

Haptics has been used to convey meaningful information on smartphones. They are mostly seen as a complement to other non-visual solutions such as audio. Haptic feedback has evolved from only being able to notify users of simple information, such as notifications, to allow the mapping of any vibration pattern to contacts, calls, or even different applications. Users can customize these patterns to their preference according to what makes sense to them (e.g., the vibration that resembles knocking on the door assigned to a neighbor contact [26]). This type of feedback is extremely useful in loud environments where we cannot rely on audio. It is also an advantage for privacy purposes, where only the user interacting with the device will feel the *stimuli*, protecting him from an overseer. One of the most typical scenarios is using vibration to notify the user that he has received a notification without disrupting his current smartphone usage. More advanced haptic technology, such as Taptic Engine, can even be used to create custom haptic patterns representing complex scenarios.

Even though sighted and blind users use haptics, it is clearly seen as a critical advantage for the latter. With the aid of vibrotactile feedback, touch can compensate for the lack of sight while providing a richer experience, closing the gap between them and sighted users. For instance, this type of feedback can be used to divide the screen into logical UI partitions, each with its vibration pattern, replacing the visual channel typically used to grasp the page structure. This approach is explored in [7] where custom elements were developed, able to reproduce both audio and vibration feedback, which can be inserted into any application by its developer. These components serve as cues that should be distributed along the UI to provide awareness to the user. As a demo, the researchers used an open-source email client where the list of emails was delimited by placing an element on the top and the bottom of the email list, and another between the smartphone keyboard and the area being edited to minimize the probability of the user clicking on the wrong area. The user can also use these cues to reposition himself on the screen whenever he feels lost since the element position will not change. Despite the visually impaired participants seeing the solution as valuable, its testing was limited to only two apps. It remained open to how it would perform in other more complex areas. Another challenge in this approach is relying on the developer to implement these cues and position them in a meaningful place. Other studies, such as [8], used vibration together with audio feedback to aid users when inserting text into the smartphone. The screen is divided into different groups of keys, matching the European Telecommunications Standards Institute telephone keypads standard. Each keypad produces a different sound and vibration. To insert a character, the user must first swipe the screen until he reaches the desired keypad and consequently tap the screen as many times as needed until the wanted character is read, which he can then select. This research also introduces some gestures which can be used to switch from numeric to letter insertion, delete one or all characters or even insert a new line.

Despite the great advantage of haptics usage, the user must understand what the *stimuli* provided means. So, it must always be accompanied by significance. The concept of these meaningful messages is called *haptic icons*, which are used to deliver information in a non-auditory way. These can be used to improve the feedback about a particular icon and its location on the screen, as explored in [12]. In this study, the use of simple vibration patterns is compared to more complex ones to determine if there is an improvement in the user's capability to remember the location of the icons. The session conducted in the study determined that richer vibration increased the recognition rate of both blind and visually impaired participants. It is important to note that the application used in the study has a practice section that can be used to reinforce the learning process. Despite their findings, the research was limited to 16 icons and occurred in the same time frame. The experience may differ in real-world scenarios where the user's smartphone contains dozens of applications, each with a corresponding pattern, which the user has to recall between each device usage. However, the fact remains that customizing these patterns is really important. One potential scenario is the user's ability to customize only their most used applications, significantly reducing his cognitive overload as the pattern will be familiar and will not change.

Other research in the community employs the use of external apparatus together with the smartphone to provide more reliable haptics to the user. Researchers argue that the simple form of vibrotactile feedback that smartphones can offer is not enough and does not use the potential of the user's hand to the max. Such is the work developed in [18], where external actuators were added to the participant's smartphone to convey extra information about their applications. Examples include adding two buttons at the top of the mobile device to allow the user to move right or left when playing a game. As another usage example, these actuators are used in a reading application to represent the user's progress on a particular book. As the user reads the book, the actuator accompanies its progress and moves accordingly (e.g., in the beginning, the actuator will be on top, whereas at the end, it will be placed on the bottom of the device). The work observed in [23] used a smartwatch and a smartphone to aid the user in navigating indoor locations. In the presence of an intersection, the smartwatch or the mobile device vibrates to notify the user to turn left or right. Using the developed solution, participants successfully navigated a shopping center despite some reporting that they sometimes felt lost since the system did not provide enough information while navigating.

Even though smartphones provide users the tools to customize vibration patterns and associate them with system actions, such as notifications or calls, the same is not always possible for user applications. Typically, users must rely on the developers to provide these cues and trust their knowledge in providing a meaningful vibration to the vibration. Associating a vibration pattern with a meaning is hard since different people associate actions with different *stimuli*. So, when the developer is responsible for providing this meaning and does not allow the user to change it, it is not guaranteed to be meaningful for everyone, making the solution less effective. Several solutions, as seen previously, also rely on external or custom devices, despite the appearance of modern haptic frameworks, which provide more robust mechanisms that, in some cases, can replace these devices. Some work done in the community still relies on custom devices. While most of them are inexpensive, it can be cumbersome to carry them around while also introducing additional aspects the user must know to leverage the solution to its fullest.

In the end, relying on haptic feedback to convey information can be challenging, especially when the user cannot fully customize the vibration patterns according to their preference or needs or if external devices are involved. Moreover, this type of solution works best when used with other forms of feedback, such as audio, and as such, it should not be seen as a one size fits all solution.

## 2.6 Summary

Through the continuous effort of the community, accessibility on smartphones has come a long way. Screen readers are now present on most modern smartphones by default. Users no longer have to download external software or do extra configurations. The fact that these solutions already come with mobile devices out of the box has contributed to visually impaired users' adoption of smartphones by making their experience more accessible. Screen readers work well most of the time but fail in some cases. The most prominent situation is when developers have not correctly adhered to the accessibility guidelines, mainly when no alternative text is provided for images, the screen reader cannot provide the respective information since he does not know anything about it and is not able to provide feedback by other means other than audio. Other sporadic situations include users' difficulty when dealing with large documents and scanning for information. This difficulty comes from the fact that audio is provided sequentially and thus limiting the amount of information the user can take at the time. Users will have to sit through all of the content before reaching a section of interest. Experienced screen reader users will typically employ specific strategies to mitigate this problem, such as increasing the speech rate and enabling other modern features, such as reading from headline to headline. However, most novice users are unaware of these capabilities, and even if they know them, they are afraid they will lose important information or misconfigure their device when enabling them. Blind users also face problems in understanding the screen's content in regards to the element's position and receiving information about the text characteristics, namely the text color or size. Built-in solutions are still not able to easily provide these sorts of feedback to the users. The research community has developed custom solutions to provide better feedback about the screen content properties, for example, using sonification to provide an understanding of the screen layout or dividing the screen into sections and then replaying them to the user. Other works have explored external devices to provide an extra layer of information. However, these require first that the device is purchased and second that the user also knows how to correctly configure it and use it correctly.

All in all, accessibility in smartphones can still be improved, and tasks could be done more efficiently, namely when consuming information. For this, the right tools must exist in their smartphones in order to fully customize and optimize their experience according to their needs. Users still face problems where their task is interrupted amidst the presence of other events, such as notifications, in which case their progress is lost when returning to their task. It is also essential to consider users' different needs, from novice to expert or young to old user groups.

# Chapter 3

# Concurrent speech in smartphone-based interaction

## 3.1 Design rationale

Given the limitations of sequential audio feedback and the potential of concurrent speech suggested by prior literature, we identified a set of scenarios where we believe concurrent speech can be leveraged to optimize information consumption. Moreover, we also identified scenarios that include additional capabilities currently not present on smartphone screen readers. For each scenario, we introduce what is currently available in traditional smartphone accessibility services, such as Google Talkback. Then, we describe a limitation observed in a real-world scenario when using traditional smartphone accessibility tools. Finally, we describe in more detail how our solution aims to solve that limitation or how it can be used as an alternative to existing solutions.

### 3.1.1 Task interruption

Currently, some smartphone actions interrupt the user's current task. Namely, a received notification with high priority will interrupt the content currently being read by the screen reader and will instead announce the notification's content. In contrast, notifications with lower priority will not interrupt the user's task but will lead to a different problem where the user misses information he might consider essential. This low priority will only be read if the currently read content finishes and the user does not trigger other actions. Moreover, since no cue is given to indicate the presence of a notification, this information might be lost until the user checks his notifications which may take some time. Another problem users face is that after the screen reader content is interrupted, the reading would start from the beginning if the user were to return to the same document. Talkback has recently introduced the concept of landmarks which allow the user to mark a position on the screen that he can use to return to, providing an improvement for the problem mentioned before. Nevertheless, this feature requires user interaction, first in enabling it since it is not active by default, and second in correctly marking the intended position and then accessing the respective menu to return to the specified mark.

Consider the following use case, which describes the problem at hand.

*João is expecting a message from his mother*

João is having a Skype call with his boss about his company's current affairs. In the meantime, he is expecting a message from his mother, which he does not want to miss. However, he must not miss any information from his boss as this might hinder his evaluation in the company. Unfortunately, as he has the Skype call, his mother's message arrives, which interrupts the call and makes him lose the information about the day he should hand in the report to his company. So, he has to ask his boss for it again, which is not ideal for him.

João has been using a custom accessibility service for some time now that uses a secondary audio channel to provide feedback when receiving notifications without interrupting the current task. So, João is not worried about missing either his mother's message or the boss's information.

Our approach uses an alternative audio channel to conciliate the main content and other sporadic tasks such as notifications. In this, the main content would continuously be read while the notification content is announced in the concurrent audio stream. Moreover, through internal discussions, we believe that a user might be too ingrained in their current activity (e.g., reading a blog post) and might not be ready to have a notification popup at random. Thus, we have introduced a small earcon a few moments before the notification is read, indicating that a notification is coming up. We believe this can help users not miss important information since they can redirect their attention to the notification.

### 3.1.2 Information consumption

As observed in the related work section, the task of scanning new documents can be cumbersome in smartphones due to the sequential nature of the screen reader, which limits the amount of information the user can ingest. Experienced users will significantly increase the speech rate as a workaround to find specific information in the document faster. However, less proficient users will usually default to the default playback speed, which depending on the document size, can lead to a time-consuming task.

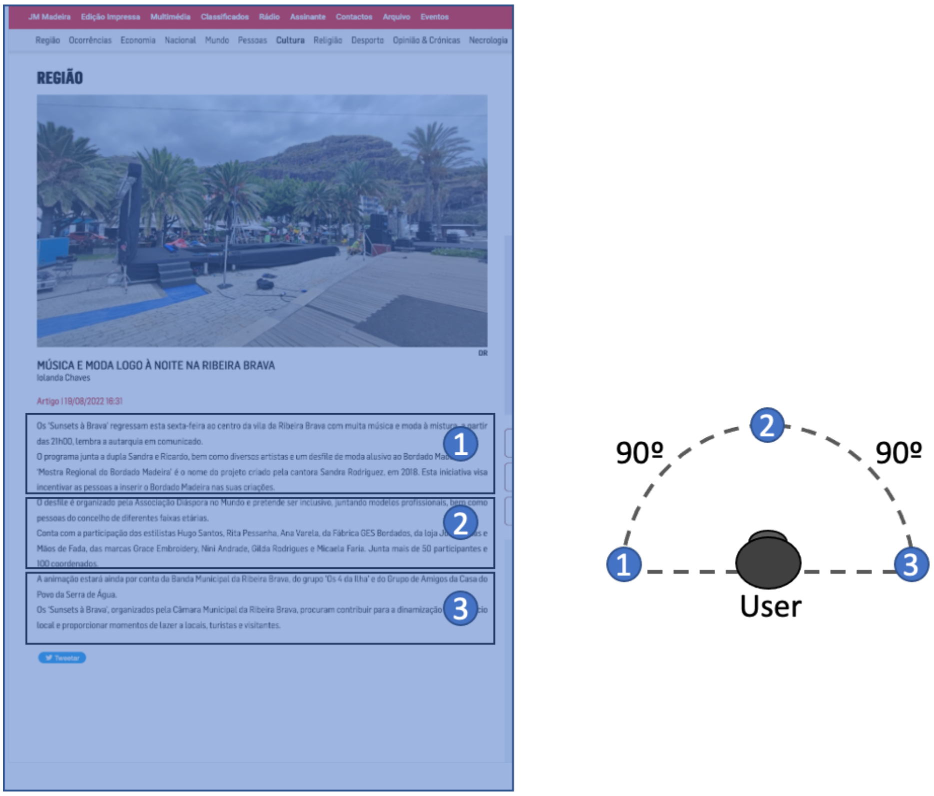
Consider the following use case, which describes the problem at hand.

*Tomás is in the middle of a meeting and wants to find some information quickly*

Tomás is at the office in a meeting, and someone asks him for some information. He knows this information is in one of the documents on his phone. Unfortunately, he possesses several documents on his phone and is unsure which of them contains the wanted information. Since he uses a traditional accessibility service, such as Talkback, he listens to each document sequentially until he finally finds the wanted information.

Tomás has recently installed a custom accessibility service that allows him to read multiple documents simultaneously. Tomás configures the service to read three documents concurrently, enabling him to grasp each document's contents. He then knows which of the documents contains the needed information. So he changes the focus to the document of interest and uses the custom talkback to hear multiple sections of the selected document simultaneously. He is then able to provide the information to his peers.

In our prototype, we divide the same content (e.g., a document) into different parts, each being simultaneously announced to the user (Figure 3.1). Concurrent speech can also be leveraged to announce different topics simultaneously to the user. For instance, when reading a news site, it can be used to read some of the news titles to the user concurrently, which he can use to quickly determine if he is interested in any of the available news.



REGIÃO

MÚSICA E MODA LOGO À NOITE NA RIBEIRA BRAVA
Iolanda Chaves

Artigo | 19/08/2022 16:31

Os 'Sunsets à Brava' regressam esta sexta-feira ao centro da vila da Ribeira Brava com muita música e moda à mistura, a partir das 21h00, lembra a autarquia em comunicado.

O programa junta a dupla Sandra e Ricardo, bem como diversos artistas e um desfile de moda alusivo ao Bordado Ma

'Mostra Regional do Bordado Madeira' é o nome do projeto criado pela cantora Sandra Rodriguez, em 2018. Esta iniciativa visa incentivar as pessoas a inserir o Bordado Madeira nas suas criações.

O desfile é organizado pela Associação Diáspora no Mundo e pretende ser inclusivo, juntando modelos profissionais, bem como pessoas do concelho de diferentes faixas etárias.

Conta com a participação dos estilistas Hugo Santos, Rita Pessanha, Ana Varela, da Fábrica GES Bordados, da loja J as e Mãos de Fada, das marcas Grace Embroidery, Nini Andrade, Gilda Rodrigues e Micaela Faria. Junta mais de 50 participantes e 100 coordenados.

A animação estará ainda por conta da Banda Municipal da Ribeira Brava, do grupo 'Os 4 da Ilha' e do Grupo de Amigos da Casa do Povo da Serra de Água.

Os 'Sunsets à Brava', organizados pela Câmara Municipal da Ribeira Brava, procuram contribuir para a dinamização cio local e proporcionar momentos de lazer a locais, turistas e visitantes.

Figure 3.1: The news page is divided into three different speech sources, and the user listens to each part in the corresponding spatial locations

### 3.1.3 Text properties

Concurrent speech can also be leveraged to provide a richer experience to the user. For example, one problem blind users face is that they can not observe the text properties present on the screen, such as text color, fonts, size, or the type of element (link, paragraph, header, among others). Sighted users usually take advantage of these properties as a strategy to scan important content. For instance, if a text is bold, it probably means it is worth highlighting and thus noteworthy, so it should be given adequate attention. Alternatively, if the text is a header, the context of the topic has probably changed, so it is safe to assume the user can start reading from that point onward without listening to previous information. The user can utilize this information to skip sections he is not interested in. Modern screen readers, like TalkBack, can jump between headers or paragraphs with the aid of defined gestures, thus providing an easy way to skip pieces of information. Nevertheless, these rely on developers correctly marking the respective text as a header, and specific gestures must be manually enabled since they come disabled by default.

Consider the following use case, which describes the problem at hand.

*Rosa is studying for an exam and wants to focus on important information*

Rosa is scanning through the professor's slides and is trying to understand the sections most important to him. However, since the elements present in the professor's notes, like headings, are not labeled with accessibility services in mind, combined with the lack of feedback informing her of highlighted areas, she has to resort to her friend's help or instead focus on all the content.

Rosa's friend Miguel has recently introduced her to a new custom accessibility service that provides audio cues during highlighted text with bold. She uses this to understand what sections are considered essential to the professor, thus giving these parts a higher level of attention.

Our application uses a secondary audio channel to provide feedback about these text characteristics while the main audio channel continues reading the main content as it usually would (Figure 3.2). We have different approaches to informing the user about different text properties through the use of audio:

- **Pauses before the text property:** before the text (with the respective text property is read), we introduce a small break, where the respective text property is announced.
- **Text properties and primary content are simultaneously read using spatial audio:** both the text properties and the respective text are read simultaneously. However, in order for the user to better understand what is said, we reproduce the main content and the text property on the left and right ear, respectively, thus avoiding a mix of audio that might otherwise not be understood entirely.
- **Pauses before the text property together with earcons:** same as in the first approach, we introduce a small break before the text property is read. However, in this scenario, we introduce an earcon that is played simultaneously as the text property is read.
- **Pauses before the text property only with earcons:** we replace the text property being announced with an earcon. Here, only a sound is played before the text with the respective property.

One common thing between all the scenarios is reproducing a sound when the text with the respective property is finished being announced. Since the property might be directed to more than one word, this lets the user know when the text stops having the respective property.

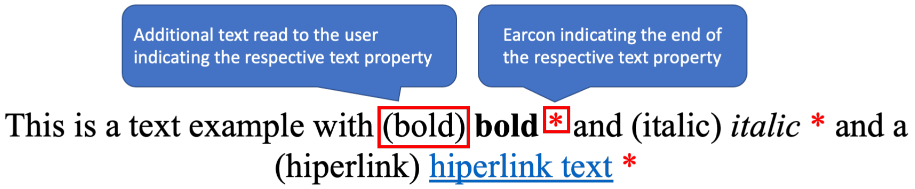

Additional text read to the user indicating the respective text property
Earcon indicating the end of the respective text property
This is a text example with (bold) **bold** * and (italic) *italic* * and a (hiperlink) hiperlink text *

Figure 3.2: Example of the text properties read to the user with the corresponding earcon in the end (indicating that the text property no longer applies, especially useful when the text property extends into more than one word)

### 3.1.4 Map

Current digital maps simply announce the nearby locations to the user without providing any kind of depth to the user. Consider the following use case, which describes the problem at hand.

*André is walking through a mall that he has never been in*

André is looking for a present for his girlfriend, and he knows that a nearby mall he has never visited has the item in stock. After arriving at the mall, he tries to understand where the store he is headed is, so he tries to find a map. After finding it, he learns that the store is on the floor he is currently at. However, he cannot know how far the store is since every store name was read to him in the same tone. So, he has to resort to a mall assistant.

In our prototype, we introduce spatial audio to provide a better perception of the user's surroundings. With this approach, the user has a different way of perceiving how distant the surrounding locations are from him and in which direction they can be found. So, for close locations, it would seem they are read from a close distance, whereas distant locations seem to be read farther away. In this scenario, we also inform the user how many meters they are from a particular location (e.g., *"College at 400 meters"*). Moreover, the user will hear each location from a different sound position, depending on his position regarding the respective location (Figure 3.3).

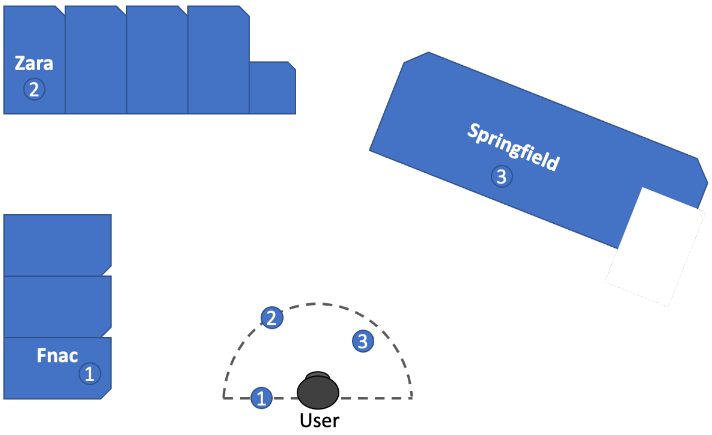


Figure 3.3: The shopping mall stores are read to the user in different spatial locations, depending on their location and distance.

### 3.1.5 Touchscreen gestures

Smartphone interaction by blind people usually happens through gestures provided by accessibility services such as Talkback. To first learn them, users usually go through a set of tutorials that goes over the most common ones and how they affect the smartphone. Related work [39][24][36] describes how this learning process can be complicated for some users. Some of the reasons include the system assuming different gestures other than the one the user is trying to do, or in some cases, no gesture is recognized. This interpretation by the system might happen for several reasons, for instance, the user not performing the gesture fast enough or not drawing it with the proper curvature. After going through the learning process of the gestures provided by Talkback, users can navigate the device in two ways. The first is through swiping, where the user performs swipes to move between screen items. The second is by touch, where the user slowly drags his finger through the screen, and when a particular element is focused, Talkback announces its content and type. In exploration by swipe mode, there is no feedback since the action of swiping itself has no feedback. However, with exploration mode, haptic feedback (a vibration) happens as the items are being hovered and when they are selected through a double tap.

Consider the following use case, which describes the problem at hand.

*Rosa, which has been using a smartphone for six months and knows most of the accessibility service-provided gestures, wants to check her notifications*

Rosa has received an email that she wants to read. She knows that to access the email faster, she can go into the notifications menu and open it directly. To open it, she swipes her finger right and then down. However, the phone is not doing anything whenever she tries to perform the gesture. Frustrated, she defaults to the old way of navigating the phone through *Explore by swipe* to reach the email application and check the received email.

In our approach, we leverage audio to provide an extra layer of feedback when navigating the smartphone with Talkback or when performing the gestures provided by the accessibility service. Every time the system recognizes one of the available gestures, we play a different sound to the user that should be meaningful to him. Moreover, Talkback reads the element content and its type after it (e.g., *"Click me, button, double tap to activate"*). As an alternative, we reproduce earcons to inform the user of the focused element, thus eliminating the need for Talkback to read the type. We want to understand the impact of adding this feedback when learning or performing gestures and explore if there are preferred ways of informing the user about things on the screen (such as buttons or icons).

# Chapter 4

# Implementation

## 4.1 Overview

In order to make a testable solution for different users, we have developed an Android application that includes several small prototypes, one for each scenario described in the previous section. In this section, we overview the implemented architecture for the developed Android application, which components are part of it, what they do and how they allow us to accomplish the proposed solutions. Figure 4.1 provides an overview of how the system behaves for different options. For scenarios that use spatial audio, we first start by preloading an audio file generated beforehand; we then set its sound position and play it using GvrAudioEngine. On the other hand, for scenarios where non-spatialized audio was used, we start by loading the audio properties defined for the respective scenario (these can be freely changed), generating an audio file using Amazon Polly, and then play the final audio file using Android native Media Player.

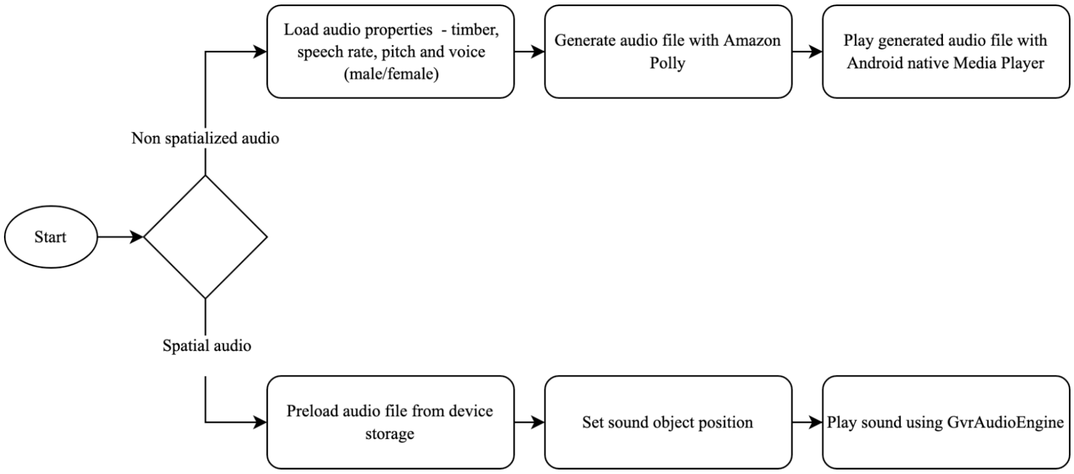

Load audio properties - timber, speech rate, pitch and voice (male/female)
Generate audio file with Amazon Polly
Play generated audio file with Android native Media Player
Non spatialized audio
Start
Spatial audio
Preload audio file from device storage
Set sound object position
Play sound using GvrAudioEngine

Figure 4.1: Application overview.

### 4.1.1 Architecture

Nowadays, Android applications are developed using several architectural patterns. These patterns bring many benefits. For once, they enable better separation of concerns, where each layer is responsible for one thing. Moreover, new features can easily be added, with little to no changes, since things are well divided. Also, newly added features can reuse most of the already implemented logic, which typically exists separated from the feature itself. For example, once a logger responsible for logging events in a separate service is developed, all the existing and new features can use it. Hence, significantly reducing the time spent on development. Another advantage of these patterns is the developers' familiarity, which enables applications to be extended much more quickly since they know where each thing is located. One of the most common Android patterns is the Model View ViewModel (MVVM). The Model contains all of the application data and business logic of the app, which several features can reuse. The View is responsible for displaying the User Interface (UI), which the user sees on the screen, registering user events, and passing them along to the other layers. The ViewModel exposes the information to the View and can also apply specific logic destined for a given screen.

In addition to the developed Android application, we also needed to customize Talkback, which we described in previous sections. This necessity came from the fact that some user events are restricted to accessibility services and can not be consumed by user applications. However, when some of these events could be consumed, Talkback would misbehave, which was not our intention. So, to fully keep the features of Talkback and accomplish our intended solution for each scenario, we used a custom version of Talkback and the developed Android application for our studies.

### 4.1.2 Spatialized Audio

Across our application, we have several scenarios that take advantage of spatial audio. Using this, we can reproduce sounds from different positions in a three-dimensional environment (along the x, y, and z axis). This approach mimics what people hear in the real world, where sounds come from different places. Furthermore, this allows for a better separation of different sound sources, which is especially useful when dealing with multiple concurrent sounds.

We have used two libraries to provide spatial audio in our application, each with its benefits, as we will explain below.

#### Resonance Audio

Resonance Audio[1] is a spatial audio Software Development Kit (SDK) that provides a simple way of reproducing spatial audio on different platforms, including Android. This library can reproduce audio files from a specified position, for example, (0, 0, 0). For that, we need to generate the appropriate files and then define from where they should be reproduced, all done programmatically. This SDK was used in scenarios where the reproduced sound only used earcons or a few words.

[1] Resonance Audio. https://resonance-audio.github.io/resonance-audio/, (Last visited on May 15th, 2022)

### ExoPlayer

ExoPlayer [2] is a media player for Android, which abstracts some complex low-level API's provided by the native Android MediaPlayer while also providing additional features. Recently, ExoPlayer introduced an update that provides spatial audio out of the box. However, here we do not have such fine-grained control over the position where each audio file should be reproduced. Contrary to what was possible with the previously mentioned solution. So, with this approach, we need to manually generate the audio files with an external tool (see section Generating Audio 4.1.6) in the corresponding position *a priori* and only then reproduce them with ExoPlayer. Through our testing, we have experienced better results, sound-wise, with ExoPlayer than using Resonance Audio.

## 4.1.3 Non spatialized Audio

While spatialized audio provides some added benefits, some users might prefer non-spatialized audio, especially when in the presence of a reduced number of concurrent sounds. To achieve this, we combined two steps. First, we use Amazon Polly [3] to generate the audio files to be reproduced. This service turns text into speech, which can be further reproduced as an audio file whenever needed. Amazon Polly also supports SSML Tags, which provide ways of customizing said speech regarding different voices (male or female), timber, speed, or pitch. These configurations are saved in the user device through a data structure called AudioChannelProperties. This structure reflects the changes to the audio configurations used as the input for the audio file rendering. After the audio file is generated, we reproduce it using the native Android media player. Android allows us to reproduce several media files simultaneously, thus achieving our goal of reproducing concurrent audio.

## 4.1.4 Track user events

We added a way of tracking user events and metrics to provide a better overview of the study participants' actions while going through the different scenarios. In essence, we needed a way of knowing precisely what screen elements were clicked and how long each participant experimented with a different scenario. With this, we could focus on supporting the participants without worrying about manually tracking their activities.

Several alternatives exist to track user events in Android, including Firebase Analytics or DataDog. However, despite Firebase Analytics being used more in Android development, this tool does not track user events automatically. So, if, for example, a button click needs to be tracked, this event needs to be specified programmatically. As a result, we chose DataDog, which automatically tracks all user events and time spent on each screen with no added configurations. However, despite user events being automatically tracked by DataDog, our Android application also needs

[2] ExoPlayer. https://exoplayer.dev/, (Last visited on May 15th, 2022)

[3] Amazon Polly. https://aws.amazon.com/polly/, (Last visited on May 15th, 2022)

to track other gestures not covered automatically by Datadog. For this purpose, we can manually set events on code, which will be triggered when that flow is executed.

### 4.1.5 Gestures tracking

Our application needed to track user gestures, such as swipes or taps, to reproduce specific sounds when the user does them. Moreover, we also needed information about the screen elements, such as buttons or text fields, where the gestures are made. For these requirements, we have combined three approaches.

#### Accessibility delegate

*Accessibility delegate* is a class that can be registered in a screen View, which listens to all events done by the user. These events contain information about what element triggered the event. So, for instance, if a user tapped on a button, we would get the following structure (some fields were omitted for brevity).

EventType: TYPE_VIEW_HOVER_ENTER;
ClassName: android.widget.Button;

Here we can observe the event type, which states that the user entered the area of the element. This is important because other event types can indicate that the user is no longer interacting with the element, and we are not interested in listening to those. The class name can also be found, which indicates the name of the clicked element, in this case, a button. With this, we know when the user interacts with a specific element and which one, so we can use this information to reproduce the corresponding earcon.

#### SimpleOnGestureListener

In Android, every interaction on the touch screen is reported as a MotionEvent object. This object describes what action was performed on the screen or the position of the touch. So, we can define a gesture as a series of MotionEvents. To abstract away the complexity of correctly identifying the gesture by interpreting a sequence of events, there is an Android API that does this work for us, *SimpleOnGestureListener*. This API provides an easy way to listen to simple gestures, such as taps or double taps. However, manual calculations are needed to detect more advanced gestures, such as swipes. *SimpleOnGestureListener* only listens to a onFling method, which provides the speed at which the fling was done but did not necessarily indicate a swipe. So, to detect a swipe, we first need to validate if the speed of the movement is fast enough. Otherwise, we might only be in the presence of a drag action. Secondly, we need to check on which axis (x or y) the fling occurred with more velocity. If the velocity in the x-axis direction is higher than the one on the y-axis, then the fling either indicates a swipe right or left. If not, it is a swipe up or down. Finally, to differentiate between the two directions (right or left and up or down), we need to check if the velocity on the x-axis was positive or negative, right or left swipe, respectively. The same happens for the other direction, but on the y-axis, if the velocity done on it was positive, it indicates an up swipe. Else, it is a down swipe. Despite *SimpleOnGestureListener* ability to correctly identify these gestures, when accessibility services are active, some of them might not be recognized. Thus, we needed an alternative when these are on.

#### Custom talkback

In Android, when accessibility services are active, like Talkback, some gestures are reserved for them. Listening to these gestures might make Talkback misbehave since they are not consumed on Talkback but in the user application instead. To fulfill our requirements while also guaranteeing that Talkback works as intended, we took advantage of the fact that Talkback is open-sourced and customized it to our needs. For this, we added some changes to two classes. The first one is the *TalkbackService*, which provides an onGesture method, where we can listen to swipes typically reserved for Talkback's explore by swiping mode. Also, as a side note, if we were to listen to these gestures in a user application, we would override the same onGesture method (on the application). However, this would make Talkback not receive the event since it would have to be consumed by the application. The second change is done in *ManualScrollInterpreter*, where we can check for up and down swipes. In these modifications, we introduced the reproduction of earcons whenever one of these gestures is performed without impacting Talkback.

### 4.1.6 Audio generation

Generating spatial audio required extra work since the library we used (GvrAudioEngine) cannot use an audio stream as an input. Therefore, to work with audio generated on runtime, we first had to manually generate the audio file, save it on the smartphone storage, track the saved file path, load the sound object, and run it on GvrAudioEngine. Since our main objective was exploring multiple sound techniques and not for the solution to work system-wide (more on this in the limitations section), we decided to generate the needed audio files for the scenario with external tools. Thus, we started by manually generating the audio file, embedding it on the app, and running it on the spatial audio engine or ExoPlayer, depending on the scenario. To convert text into an audio file, we used FreeTTS [4]. This website converts a given text into an audio file with the possibility of further customizing the generated speech. Despite the several configuration options, in our scenarios, we only needed to use the ability to read the text in different voices, such as male or female. For scenarios such as task interruption or faster scanning, no further configurations were needed for the generated audio file. However, this was not the case for the text properties or map scenarios, as we will see below. So, the next step after having the audio file with the text converted into speech was to set it up as per the requirements for each scenario. To edit the audio files, we used two tools, Audacity [5] and REAPER [6].

[4] FreeTTS. https://freetts.com/, (Last visited on May 15th, 2022)

[5] Audacity. https://www.audacityteam.org/, (Last visited on May 15th, 2022)

[6] REAPER. https://www.reaper.fm/, (Last visited on May 15th, 2022)

### Audacity

Audacity is a free audio editor used to manipulate audio files. We used its features in some of the scenarios we had. For example, in our text properties scenario, we used Audacity to insert pauses between our text and interpolate the text properties' announcement into the original text. Additionally, instead of programmatically handling the announcement of the text property and the respective text, when playing them simultaneously, we have also used Audacity to insert the announcement of the text property at the wanted timeframe. Another scenario we used Audacity was to decrease the clip volume, namely in the smartphone interaction scenario, where the exploration sound we have should not muffle other announcements, such as gesture earcons.

### REAPER

Reaper is a digital audio application that, beyond enabling the edition of audio files, also provides extra processing capabilities needed for our requirements. For our map scenario, we have used one of REAPER's capabilities in offering the possibility of configuring an audio track into a multichannel stream. This multichannel stream can include multiple sounds coming from different directions in a three-dimensional plane, which was needed in our scenario. So, first, we have to configure how many track channels will be on the final audio file. Since, for our scenario, we had a maximum of four locations announced simultaneously, we configured the audio tracks to be four. Then, we insert the needed audio tracks, which in our case, correspond to the announcement of the location and the distance to it. After, we use a plugin called *FOA Encode Planewave*, which assigns a position in a three-dimensional pane to the audio track. Finally, after assigning a position to each needed audio track, we exported the audio file with four channels and set it up as a multichannel file. Then, ExoPlayer, which we use in our application, can automatically play this audio file in the corresponding positions for each audio track without any extra setup.

# Chapter 5

# User study

To understand how our solution impacts users and how they feel about it, we conducted a user study with visually impaired users with different levels of expertise. The main goal was to determine in which scenarios our solution could provide a more efficient way of consuming information.

## 5.1 Research questions

- *RQ1:* Which scenarios benefit the most from the use of concurrent speech?
- *RQ2:* How can concurrent speech be leveraged in such scenarios?
- *RQ3:* What are the benefits, disadvantages, and improvements?

## 5.2 Methodology

In the study, the participants experimented with all the scenarios available in our solution. In some of them, the participant had to complete a task, which we used to assess how accessible our solution was. At the end of each scenario, we gathered the participant's feedback through open-ended questions. The aim was to understand in what scenarios the solution could be used, what they liked about it, things to be improved, or negative aspects. Additionally, after each scenario, following the Single Ease Question (SEQ) we asked participants how difficult a task was, measured on a 7-point Likert scale, from very difficult to very easy (Table 5.1).

Table 5.1: List of scenarios and SEQ Assessment

| **No** | **Task** | | **Assessment** | |
|---|---|---|---|---|
| **1** | Notifications with spatialization | Very difficult | 1 2 3 4 5 6 7 | Very easy |
| **2** | Notifications without spatialization | | 1 2 3 4 5 6 7 | |
| **3** | Skimming with 2 concurrent voices | | 1 2 3 4 5 6 7 | |
| **4** | Skimming with 3 concurrent voices | | 1 2 3 4 5 6 7 | |
| **5** | Skimming with 4 concurrent voices | | 1 2 3 4 5 6 7 | |
| **6** | Concurrent map | | 1 2 3 4 5 6 7 | |
| **7** | Sequential map | | 1 2 3 4 5 6 7 | |
| **8** | Text properties with pauses | | 1 2 3 4 5 6 7 | |
| **9** | Text properties concurrent with main content | | 1 2 3 4 5 6 7 | |
| **10** | Text properties with pauses and earcons | | 1 2 3 4 5 6 7 | |
| **11** | Text properties only through earcons | | 1 2 3 4 5 6 7 | |
| **12** | Smartphone interaction | | 1 2 3 4 5 6 7 | |

All the interactions made by the participant in the test smartphone were tracked, allowing us to track metrics like time spent on a task, how often they have replayed a scenario, or what gestures were done during the study. In addition, a unique ID was assigned to each session to distinguish the participant's sessions.

After the study, we transcribed the participants' feedback which was then analyzed using an inductive coding approach. We started by reading all the transcripts to fully grasp what was said, and then through several re-reads, we identified a total of 42 codes. The codes are available in A.

## 5.3 Participants

We recruited 10 blind participants, of which 3 were female, aged between 33 and 63 (M=48.7; SD=9.49). The participants have several years of smartphone experience (3 to 10), except for one, which has only been using a smartphone for 5 months. Most of the participants can do several tasks on their smartphones without the help of others. These tasks include calling someone, accepting other people's calls, sending messages, listening to music, browsing online, and installing new applications. The level of expertise of the participants was determined by themselves. At the end of the study, we asked the participants to rate themselves from 1 to 5 on their smartphone usage expertise. We attributed an expertise level to each of the ratings on the scale. 1 represents a novice user, 2 an advanced beginner, 3 a competent user, 4 a proficient user, and 5 an expert user.

Table 5.2: Demographic Information and smartphone usage expertise of participants

| Participant | Gender | Age | Smartphone adoption | Age adquired blindness | Expertise |
|---|---|---|---|---|---|
| **P1** | Male | 63 | 3 years | 3 | Competent |
| **P2** | Male | 39 | 9 years | Born | Competent |
| **P3** | Male | 50 | 3 years | 15 | Competent |
| **P4** | Male | 57 | 3 years | 50 | Novice |
| **P5** | Female | 48 | 10 years | 23 | Expert |
| **P6** | Female | 33 | 7 years | Born | Expert |
| **P7** | Male | 37 | 15 years | 10 | Competent |
| **P8** | Male | 58 | 1 year | 47 | Expert |
| **P9** | Female | 56 | 5 months | 28 | Proficient |
| **P10** | Male | 46 | 7 years | 32 | Proficient |

## 5.4 Procedure

In our study, the participants explored 5 distinct scenarios that used a combination of concurrent speech and spatialization or provided additional feedback typically not given by traditional screen readers. For each scenario, we first would describe what the participants could explore and what they would need to do. If the participant was displaying some difficulty after exploring the scenario, we asked them if they wanted to repeat it since we believed that after the first time, participants would not be as confused as they were previously. Below we describe each scenario and the task the participants had to do, if applicable.

### 5.4.1 Task interruption

In this task, while the participant was listening to a news excerpt, he received a message concurrently without stopping any of the audio streams. We split this scenario into two. First, we used spatial audio to reproduce both audio streams. We played the news excerpt on the left ear and the message received on the right ear. Then, after listening to the scenario with spatial audio, we played the same scenario without it, so the participant listened to everything together as he usually would. After listening to the news and the message, the participant tried to identify the news content and what the message he received said.

### 5.4.2 Skimming

For the skimming scenario, the participant listened to different news excerpts simultaneously. The number of news the participants had to listen to concurrently ranged from 2 to 4. This scenario was played exclusively with spatial audio. Therefore, each audio source had a defined spatial position. The objective was for the participants to identify each news's main theme or topic. The participants were informed beforehand that they did not have to fully grasp what the news mentioned. They only had to try to understand if the information was, for example, about sports, politics, or science.

### 5.4.3 Text properties

For this scenario, participants listened to a phrase that contained different text properties, such as bold, italic, or hyperlink. This extra information was played to the participants in different ways: by introducing pauses where the text property is announced before continuing with the main content; reading the text property at the same time as the respective word; introducing a pause where the text property is announced together with a corresponding earcon; playing an earcon before the corresponding text property.

### 5.4.4 Map

This scenario was split into two. First, the participants listened to two different locations played simultaneously, a university and a pizzeria. We announced the place's name and how many meters there were until she reached it. To allow the participants to explore this scenario, we play 4 different steps to the participant, emulating a person walking on the street. As she was walking, she would listen to each location more clearly or hazy depending on whether they were closer or farther away from the place respectively. For the second part of the scenario, instead of playing the locations simultaneously, we played them one after the other, from the closest to the one farther away, always announcing how many meters there were until the location. For this option, we mimicked a shopping mall disposition. Participants listened to 4 different steps, representing 4 separate places inside the shopping mall. On each step, participants would listen to different shops. Shops that are too distant from the place played in the step are not announced to the participant.

### 5.4.5 Smartphone interaction

Participants interacted with the prototype application we developed using our research smartphone. For this scenario, we designed two pages with a layout similar to what would be presented on a smartphone home screen (Figure 5.1). This layout contained several elements, such as buttons or text fields, and a set of tabs where the user can swipe left and right to change between pages. Depending on the element the participant was interacting with, a different sound would be reproduced to him. Moreover, different states of elements, for instance, a checked/unchecked checkbox, have different sounds. The elements with sound included:

- Interacting with a checked/unchecked checkbox
- Interacting with a checked/unchecked switch
- Interacting with an input text field
- Interacting with a text field
- Interacting with an image

Furthermore, we also played different sounds for gestures usually used to interact with smartphones. These gestures include:

- Exploration mode
- One tap on the screen
- Two taps on the screen (i.e., selecting a button)
- Swiping left and right using one finger (i.e., changing between items)
- Swiping up and down using one finger (i.e., changing between items)
- Swiping up and down using two fingers (i.e., scrolling a list)
- Swiping left and right using two fingers (i.e., changing tabs or going back)

The participants were instructed to go through the pages presented. With this, they interacted with different elements and made different gestures, such as tapping or swiping. We also helped participants who either were not exploring all of the page and its different elements or were not trying all of the gestures with a corresponding sound.

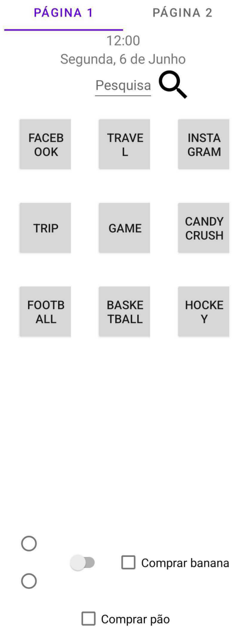


Figure 5.1: Layout resembling a typical smartphone home screen with several buttons, tabs and different elements.

## 5.5 Findings

This section presents the key findings identified through the quantitative analysis of the data gathered during the study, followed by the findings obtained from the qualitative analysis collected during the study.

### 5.5.1 Scenarios tasks

All participants explored every scenario. However, only in two were there concrete tasks where we could gather quantitative data.

In the first scenario, where participants listened to a notification concurrently with a piece of news, participants were asked to give an answer based on the news they had just heard and to identify what the message they received said. This task was requested for the option where spatialization was used and where it was not. All 10 participants correctly identified what the message they received said, both when using spatialization and not. As for answering a question regarding the news they had just heard, 8 out of 10 participants correctly answered the question in the option that used spatialization and 6 when it was not. The difference in the number of successful responses when spatialization was used and when it was not could also come from the fact that the question they answered was different in complexity. For the spatialization option, participants were asked to determine what would be mandatory to happen for small and medium electronic devices. However, for the non-spatialization option, they were asked to identify the result of a football game which some failed to identify correctly.

The other scenario that included tasks for the participants to do was the skimming scenario. In this scenario, participants were asked to identify the main topic of the news they had just heard. They heard three options: one for two simultaneous news, one for three, and one for four, so, in total, they were asked to do 3 tasks in this scenario. However, it is important to note that participants were not always asked to complete the 3 tasks. If they were uncomfortable for any reason, they could withdraw and skip exploring one of the options. All the participants explored hearing the option where 2 news were simultaneously read. Nine of them correctly identified the main topic of each source. One participant correctly identified only one of the two news they heard. For the option where 3 simultaneous pieces of news were played. 2 participants correctly identified the main topic of all the news they had just heard. 4 participants determined the theme of 2 of the 3 news they had just heard. One participant identified one correct main topic. While 3 participants could not understand anything and thus could not identify any topics. For the option where 4 news were concurrently played, 4 participants stated that it was too much and did not want to explore it. 2 of the participants identified some of the words present on the news but could not understand what the topic was and what did the news talk about. Three participants could not understand anything from what they had just heard. While 1 participant correctly identified 3 out of the 4 topics he heard.

Table 5.3 presents the results of the SEQ analysis of the 10 participants for each task.

### 5.5.2 Scenarios exploration

In this section, we present the findings gathered during the exploration of the scenarios by the participants. For each scenario, participants were asked about their opinions, improvements, aspects they disliked, and suggestions on how they would improve a scenario.

Table 5.3: Results of SEQ assessment

| Tasks | Count N | Mean | Standard Deviation |
|---|---|---|---|
| **Notifications with spatialization** | 10 | 5.1 | 1.76 |
| **Notifications without spatialization** | 10 | 4.7 | 1.73 |
| **Skimming with 2 concurrent voices** | 10 | 5.9 | 1.45 |
| **Skimming with 3 concurrent voices** | 10 | 3.2 | 1.4 |
| **Skimming with 4 concurrent voices** | 6 | 1.33 | 0.47 |
| **Concurrent map** | 10 | 5.7 | 1.35 |
| **Sequential map** | 10 | 6 | 1.41 |
| **Text properties with pauses** | 10 | 5.8 | 1.33 |
| **Text properties concurrent with main content** | 10 | 4.3 | 1.95 |
| **Text properties with pauses and earcons** | 10 | 4.9 | 1.51 |
| **Text properties only through earcons** | 10 | 3.9 | 1.81 |
| **Smartphone interaction** | 10 | 6.2 | 0.75 |

### Concurrent speech

Participants had conflicting opinions regarding listening to concurrent speech. On the one hand, they all agreed that listening to different audio sources simultaneously requires a higher degree of attention and often leads to information loss. However, despite their emphasized difficulties, 7 participants noted how useful it was to have immediate feedback when receiving notifications and not have to stop what they were previously doing. The consensus is that concurrent speech is more accessible when used to listen to short messages, such as notifications. Regarding lengthier content, participants agreed that 2 audio sources simultaneously are manageable and valuable when the objective is to have a general idea of the content but not so much when trying to know all the details. In opposition, more than 2 simultaneous audio sources is seen as too confusing and attention-demanding.

**Immediate feedback** Most participants (7) mentioned that it was positive to listen to messages immediately without stopping what they were listening to or having to shift focus. Additionally, they did not have trouble understanding the message and content they were listening to. They found that listening to a small transcript concurrently with lengthier content would not impact their experience in any way.

*"If they are shorter messages like this, you can understand them well and not lose much of the news. Now, if they are longer, I think it would be more complicated. If it is a short message, even if a person is paying attention to the news, if the message comes in, the person will not lose track of the news. If it is a short message, it is immediately understood."* - P10

However, some participants (3) said that this was not particularly useful for their daily usage since they do not use a smartphone to read lengthier content, such as books, so they rarely experience messages interrupting what they are doing. Despite this, they said if the option was present, they might activate it since it does not pose any confusion to them and would not negatively affect their smartphone experience.

**Faster information consumption** Several participants mentioned that listening to more than one piece of news at the same time was helpful as a way to speed up their search for something.

*”This is how I usually do in the mornings. I have an application that has all the newspaper’s covers. I go through them and only read the titles. If I want to know more about something, I then ask a friend, a news expert”* - P2.

Having explored listening to 2, 3, and in some cases, 4 pieces of news played concurrently, they have stated that 2 pieces of news are ideal if they want to grasp the news content fully.

**Information loss** All participants mentioned that it was hard to understand the news content when concurrently listening to 3 or more voices. For example, when trying to hear 3 pieces of news together, most participants could only identify the theme of 2 of them correctly.

*“I could not understand much because there were so many of them. Two is good, three is not. You cannot pay attention to all of them. You absorb the content of one. You absorb a small amount of the content of the other but nothing from the third one.”* - P2

Some participants also stated that trying to fully understand everything was very demanding, mentioning that even when they were able to understand something, they quickly forgot it while trying to identify the rest of the news.

*“And then it’s one of those things where you either take note right away or you wonder what they were talking about”* - P3

**Differentiate audio sources** Three participants mentioned that, in some cases, it was hard to understand some voices used to play the news concurrently. One participant noted that some of the voices used were not as appealing as the ones he listens to daily.

*”The voice in the left ear was better. The voice you guys put in the left ear was Google’s voice, and it draws more attention.”* - P2

Another thing that 2 participants noted was that, in some cases, the voices blurred together when played in the same position (i.e., left or right).

*”A slightly different voice so as not to confuse the two sources and distinguish between them.”* - P8.

*”The one on the left side is more complicated to understand. Either because it speaks a little faster or because the voices are different, i.e., they are male and female voices. And you understand the male voice better. It overlaps and is more paused.”* - P10.

*”There were two voices on the left side, and they are both male voices. And on the right side, a female voice would come through. So on the right side, it’s a little bit confusing with two male voices. It is a little bit harder to understand. The two voices are interfering with each other. Because they are speaking simultaneously, there are parts that you cannot really understand.”* - P10.

**Change of focus** Several participants mentioned that the sound we have added before receiving a notification is required when listening to concurrent audio sources. Since if this were not the case, it would require the user to be constantly on the lookout for a notification.

*”It is necessary to have this signal in advance so that the person has time to activate the attention distribution because otherwise it would not be possible and would force the person to not be permanently on alert/charged.”* - P1.

*”Being focused on the news if the click does not appear, you end up not paying attention to the message, but with this, the person pays attention to what is coming in.”* - P10.

### Spatialization

Spatialization was used throughout the scenarios to reproduce the audio that participants listened to. In the first scenario, the notifications one, participants explored the scenario with and without spatialization. Only 1 out of 10 participants stated that he preferred listening to the scenario without spatialization. However, one participant noted that while preferring to listen to different things in different positions, since he had a bit of hearing loss in his left ear, he would opt to listen to everything together. Participants mostly agreed that spatialization is mandatory when listening to concurrent speech, and without it, different audio sources would get confused. Besides using spatialization to differentiate between different audio sources, we also used it to communicate how close or distant they were from a particular location. It was unanimous that this capability would be valuable when navigating and would not negatively impact their everyday usage.

**Understanding concurrent audio sources** Most participants considered listening to concurrent audio sources using spatial audio was better. For instance, when listening to the piece of news on the left ear and the notification on the right ear, they noted that it enabled them to pay more attention to each thing, otherwise it would confuse them.

*"If they were separate, it would be much more worthwhile for the listener. The two together are a little confusing. You cannot really hear one or the other."* - P8.

**Navigation** All participants found that communicating how close or distant a place is through spatialization was interesting and a valuable extra layer of feedback on what usually exists on GPS systems:

*"For me, the one with the map is very interesting and is very useful. Both in places I know and places I do not know to get somewhere."* - P6.

*"It is actually really good because the distance is proportional to the volume, isn't it? The closer you are, the louder the volume is, which is very good!"* - P3.

Also, having explored both locations being announced simultaneously or in sequence, they preferred in sequence as it provides them with more details on what place is closer or farther when compared to each.

*"Then one can better identify oneself. That is, it is not misleading information"* - P10.

### Improving smartphone interaction

Our scenarios brought different capabilities not currently available in traditional screen readers like VoiceOver or Talkback. The main ones pointed out during the study was the possibility of receiving immediate feedback using concurrent speech, as described in a previous finding, receiving extra information about the content they are hearing, and having sound feedback while navigating through their smartphone.

**Sound feedback** Nine participants thought having sound feedback while performing gestures on their smartphones was valuable since it would help them understand what is being done on their devices.

*"For example I'm on the bus and I want to make a search to the right side. But in the meantime there is a stop and it runs off to the left. With the different sound I automatically know that it ran off to the left side and I won't continue."* - P8.

*"They are for us to know where we are. And we then get used to knowing the phone and it's a plus for us."* - P4.

Several participants, 4, also mentioned that if a different sound were reproduced depending on the element they were interacting with, it would eliminate the need to read the element.

*"People then get used to the sound and already know where they are."* - P4.

*"You could not even say anything, and from the sound of it, you knew it was a button."* - P2.

**Extra information** Nine participants considered it important to receive extra information about the text they were reading.

*"For example, we are reading the newspaper and suppose there is some information that should be highlighted or underlined. I think it would be good."* - P7.

*"I think so. Since we do not see, it is a way for us to get some extra information. In school, I was very fond of underlining everything I thought was important."* - P8.

*"Ah, then I can already tell you that that would be ideal because bold or something does not make much difference to me. However, that hyperlink makes all the difference. Because when you click on it, you know that you are on that subject. In that respect, that one makes sense."* - P5.

Most of the participants commented on how they preferred to receive information through text as it does not imply any memorization.

*"Because while with sounds we have to be with that cognitive load of having in memory the meaning of the sounds and then be able to attribute them, with the designation made by reading we have no doubt what it is and we do not have to worry about it."* - P1.

Despite only exploring receiving information about text properties, such as bold, italic, and hyperlinks, several participants mentioned that they would also like to be informed of other things, especially in books. For example, they mentioned that they would like an easy way to understand when a new chapter is starting since sometimes chapters are only indicated through numbers, which do not have any particular feedback.

*"Sometimes, a person does not realize where another chapter begins and ends. If it has words that identify it, fine, but if it is just so by numbers, no. (...) I am now reading a book, and that book does not have exactly chapters, it has 4, 3, 4, and if I just look for numbers, 3 and 4 are in many places, then it will end up in places I do not want"* - P1.

#### Suggestions

During the exploration of the scenarios, participants often stated in what situations they would like to activate specific capabilities, how they would modify the scenarios or suggestions regarding what configurations should be possible or additional features.

**Focus on an audio source of interest** Two persons mentioned that it would be great to be able to focus on a specific audio source once they find what they are interested in and then stop hearing the rest of the sources.

*"Once you get the news you are interested in, you can stop listening to the others"* - P9.

**Spatialization with GPS** Several people mentioned that spatialization should be integrated into current GPS systems as it is a really valuable piece of information they typically do not get - "Yes, I do. In fact, Google Maps should do just that by now." - P6. Additionally, participants saw spatialization in navigation systems as being more used to navigate to a specific place and not as much as a way to get to know the area's layout. Moreover, in addition to using spatialization, they would also like to be informed about the direction that the place is in (e.g., left, center or right):

*"So I am not interested in the other stores... I am only interested in the one I am looking for."* - P5.

*"Or we could have a possibility to put the name of the stores that we want to visit and then follow this system. First, there is this one, and then there is that one. That is very useful."* - P8.

*"This must work like a GPS. For example, FNAC on the left 100 meters."* - P3.

**Customizable** Several participants mentioned that the sounds they hear and their correspondence (i.e., gestures or text property) should be fully configurable, to allow for a sound that is more meaningful to them. Additionally, they stated that they would like to be able to activate certain options, such as receive text properties, only in some scenarios.

*"For example in books I would turn off the words. Because when we're reading a book I don't think it's necessary to know that kind of thing. Because for me reading has to be pleasurable. But now for other things, I wouldn't turn off(...) It must be according to the need."* - P9.

*"Yes, yes exactly. For example, I would activate the hyperlink to know that I could go somewhere else."* - P5.

*"The person can also choose the type of sound and which one they want. There I don't think it is complicated."* - P8.

# Chapter 6

# Discussion

Our study explored ways to provide a more efficient smartphone interaction and other types of feedback currently unavailable in traditional smartphone screen readers. The added features in sound feedback, information about the text properties, and the use of concurrent speech were seen as helpful by the participants, especially if these can be easily toggleable since these might not make sense in every daily scenario. The feedback gathered during the study is super valuable for the research going forward.

Below we reflect on our findings and describe the possible reasons for some of them.

## 6.1 Augmenting smartphone information consumption efficiency

Throughout our study, we explored different scenarios where we leveraged concurrent speech, spatialization, and different audio characteristics to provide a more efficient or alternative way of consuming information on smartphones. Traditional screen readers are limited by their sequential way of providing information, leaving the user with only the option of increasing the playback speed if they want to consume information faster. We have explored several scenarios where we use concurrent speech to increase the efficiency of information consumption. We note that hearing several voices can be complex and demanding since we must pay attention to multiple things simultaneously, often contributing to the loss of information. This also is the case for other studies done in the matter [14][15]. However, our results suggest that hearing short sentences with lengthy content poses no problem in understanding the main content and the short message. Some participants state this after exploring hearing lengthy content concurrently, stating that concurrent speech is valuable but only for shorter messages as these are easy to understand amidst other content, as is the case of messages. This is contrary to other studies in the literature [1], where intermittent content is seen as distracting instead of having longer content fed concurrently, where participants could easily distinguish each audio source. Despite some confusion when hearing concurrent speech, our results indicate that using concurrent speech with 2 simultaneous audio sources poses no problem when identifying the topic of the respective sources. João Guerreiro and Daniel Gonçalves [14] also noticed this and stated that the best compromise for basic comprehension of sentences and the speed to process them is two voices with 1.75x the default value.

Regarding 3 simultaneous voices, we find that only 2 of the 3 sources can correctly be identified. However, this could also be attributed to a lack of practice with this kind of feedback since current smartphone screen readers announce the information sequentially. When participants expressed their desire to retry an option, they displayed better results, for instance, correctly identifying all news topics. However, this is hard to conclude since some participants said that the second time they were only paying attention to a specific audio source (i.e., the one they could not identify the first time). Despite concurrent speech being valuable as an alternative and more efficient way of consuming information, our results indicate that it often leads to information loss and difficulties recalling information from what they just heard. Other studies also faced this issue [14][15]. Ultimately, we believe that concurrent speech can be used with positive results for skimming, where the detail of the content is not the most important thing to capture, improving frustrations reported in other studies [25][3]. Moreover, our results also show it is valuable for receiving intermittent situations, such as messages, avoiding having the current task interrupted, and as a way to have immediate feedback.

Spatialization was used throughout our scenarios for several purposes. One of the reasons was precisely to play concurrent audio, which, as our study suggests, helps in listening to simultaneous audio sources, making the experience more accessible when each source is individually placed around the listener. This separation of the content allows for a better distribution of attention, and it avoids different sources getting confused with one another. However, this may not be the same for everyone, and it should be possible to configure this by each user and not have it enabled for everyone. For example, people with hearing loss in one of the ears might opt to turn this option off.

Our study is aligned with what we saw during the literature review regarding using different voices (e.g., male or female or different timbre) [14] for different audio sources when exploring concurrent speech. During the study, participants noted how some voices were harder to understand and not as attention grabby. In some cases, different voices got confused, giving the wrong perception that they were talking about the same content. The fact that the voices were harder to understand could be because one of the voices overlapped the other. For example, when exploring 3 concurrent news, two audio sources used a male voice with a different timbre and pitch, while the other used a female voice. Despite the male voices being different, since they are both males, it might be harder to separate them since they can be similar. This could also justify why participants could usually only detect 2 of the 3 news themes. The content read by the female voice was only not identified once, whereas for the other two themes, in most cases, only one was identified, and not always the same. However, we noted that spatialization helps with this separation of concerns and should be used together with concurrent speech and different voices for users to distinguish between different contents clearly.

## 6.2 Additional screen readers capabilities

Our study explored different sets of features not available in smartphone screen readers.

One of the extra features we explored was the use of spatialization to convey extra information. This was the case with the map scenario, where spatialization was used to understand how close or distant a particular destination was. We believe this can be a valuable feature in navigation systems as it conveys an extra piece of information to the user and, as our study suggests, does not negatively impact user navigation. However, we understand that the content might be hard to understand if spatialized. This is the case when dealing with locations that are too distant, where the sound is placed far away and is low, which might indicate that not all the information should be spatialized. For example, the remaining meters to reach a location should be said in the default way. Alternatively, only locations till a certain distance should be announced so as not to have the locations be read so low that the users cannot fully understand them. Another option could be only to use spatialization to reach a particular destination. Here the user would already know the information about the place he is trying to reach and would only use spatialization to help him understand how far away he is from reaching it.

Additionally, we researched the consensus on announcing different text formatting, such as bold, italic, or hyperlink. Some participants are used to this while using their preferred desktop screen readers. Our study indicates that this information is valuable to have. However, the participants mentioned that this might not be the case in every scenario, and it should be possible to activate and deactivate according to the use case. In some situations, it can affect the leisure or comprehension of the text they are reading. This is the case in books for leisure reading, where knowing how certain words or phrases are formatted is not essential, whereas this kind of information would only get in the way. Some participants compare it to the deactivation of punctuation, which they often do when it comes to this kind of content. The comments done by participants in our study suggest that this might be more suited for other kinds of content, for example, administrative work, where it is essential to guarantee that the formatting is correct. Alternatively, in situations where it is valuable to have additional feedback about a piece of content, for example, a text highlighted by a professor indicating that it can be on the exam. So this feature should be toggleable according to the user's needs, once again adding to the need for a customizable solution accessible by everyone. Moreover, the results of our study indicate that this kind of information works better when given through explicit words (i.e., bold or italic) instead of sounds. However, this can be hard to conclude since the participant's comments about these focused on the fact that words are straightforward and do not require memorization about what they mean. In contrast, sounds require a continuous learning journey until users know what they represent. Moreover, participants preferred the option where there was a pause before the word where the text property begins, used to announce the property, and only then continue with the phrase. However, we also believe that giving users the flexibility of choosing their sounds would help in remembering what it represents and for the text not to be as crowded and not add to the reading time compared to reading both the text and the respective property. Furthermore, announcing the text properties with words required an added audio cue at the end of the word where the respective property ends. This was done since more than one word can have the same property after the pause. Combining sounds and concurrent speech, we could reproduce a continuous sound while the text containing the property is being read.

Our study indicates that having these extra features in smartphone screen readers is beneficial and gives a more flexible experience in smartphone usage. When spatialization was used to convey distance to the user during navigation, there was no mention of this negatively impacting the user experience. Thus we believe this could always be turned on, eliminating the need for users to know how to activate this feature. Regarding text formatting, we recognize that it might not be applicable and valuable in every situation, so it should be possible to turn it off and on easily.

## 6.3 Increasing smartphone interaction accessibility

Smartphone interactions using screen readers have a steep learning curve. Users must start by learning the gestures used to navigate the smartphone, discover the corresponding action to each gesture and perhaps customize them in a meaningful way. Several studies found that the process of learning and afterward using gestures is hard [36][35]. Either because the gesture does not recognize the user's intended gesture, the system misinterpreted the gesture, or because tutorials are hard to understand. In an attempt to minimize this complexity and give users a safer way of interacting with their smartphones (i.e., not performing unwanted actions), we have added sound feedback for smartphone interactions, more precisely for gestures. This feedback serves as a way to confirm actions done on the smartphone and, more importantly, if the intended action by the user was done or if the system assumed something else. We understand that this kind of feedback is more relevant for novice users who are still learning the intrinsics around gestures and are unsure if what they did was what they intended. However, despite most of the participants in our study being considered experts, our study indicates that having sound feedback for their actions would not negatively impact their usage. This feedback is something they are used to when using their smartphone screen readers, for example, a sound being reproduced when a button is pressed. So, we believe having this extra sound feedback can be helpful for both novice and expert users, first, because it is not something that impacts their smartphone interaction. Moreover, even though they are considered expert users, there are some gestures and actions that they might not be used to, for example, more advanced gestures like the L-shaped, which might help them confirm if they made the intended gesture. We also believe this sound feedback could be used for tutorials when learning new gestures. This was something that was pointed out by one participant, who stated he would only activate this feedback for this kind of situation and turn it off otherwise.

In our study, we also explored what the effect of replacing the announcement of the screen elements (i.e., the screen reader saying ”button”) with corresponding sounds for different elements and states is (e.g., a checked/unchecked switch). Some participants in our study indicated that they would make the switch. However, some mention that since they are already familiarized with their smartphones, they do not need to know if they are interacting with a button or a text field. They already have this knowledge beforehand. Most of the time, they say that the screen reader does not even announce the element because they do their activities so quickly that the screen reader does not have the time to announce the element's info.

## 6.4 Desktop vs. smartphone screen readers

Desktop screen readers currently provide a broader set of features when compared to smartphone screen readers, such as Talkback or VoiceOver. For instance, JAWS [1] and NVDA [2], popular desktop screen readers, report textual formatting such as font style (e.g., bold or italic) or size. In previous NVDA versions (up to 2015.4; now in version 2022.3.2 at the time of writing), this textual formatting would be enabled by default and read to the user sequentially (i.e., before the formatted word). However, this resulted in several user complaints and has thus been deactivated by default. In the current version, if users want this information, they either have to request it on demand or enable the feature to always have it announced. During our study, the participants also noted that textual formatting is useful only in certain situations and should not be permanently active. For example, when listening to a book, they would turn it off as the primary goal would be to have fewer distractions to make the reading as pleasant as possible. Nevertheless, like NVDA, we found that this option gives users more flexibility as they could turn it on or off depending on the scenario. However, this would require users to know how to turn these options on and off. These desktop screen readers also include the ability to inform the user about possible spelling errors. For example, if enabled, NVDA will play a short buzzer when a typed word contains an error. On the other hand, when requested, JAWS will find the words with spelling errors and state they are not found in the dictionary, spelling them along the way. This feature is typically enabled when users are composing documents. However, we understand it is rarely done using smartphones as users will generally refer to other means, like a desktop or a braille writing machine. So, this feature may not make sense for smartphones.

We comprehend that the user's usage differs when using a smartphone or a desktop; they serve different purposes, and some of these features might not make sense to include. Nevertheless, we believe the added features would significantly increase smartphone users' flexibility, especially if these options can be toggled or used on demand according to the user's needs.

[1] JAWS. https://www.freedomscientific.com/products/software/jaws/, (Last visited on November 28th, 2022)

[2] NVDA. https://www.nvaccess.org/, (Last visited on November 28th, 2022)

# Chapter 7

# Conclusion

Smartphone interaction is visually demanding, with multiple elements on the screen and different actions you can do. Additionally, not every application has the same disposition (e.g., one might have a footer or bottom navigation, whereas another does not) or reacts the same to an interaction (e.g., tapping on a button might not navigate anywhere but open a dialog instead). As a result, visually impaired users must rely on other forms of interaction and feedback. Screen readers are typically the ones providing these alternative ways of interacting with the smartphone, which is the case in VoiceOver for iOS or Google Talkback for Android. However, despite screen readers providing an accessible way of interacting with a smartphone, there are still several challenges, such as gestures being hard to learn or tasks being interrupted by others.

In this work, we explored the viability of augmenting how smartphone information is transmitted through concurrent audio streams and spatial audio. Moreover, we also studied the impact and how we can transmit extra information, such as text properties like bold or italic, to the user.

We performed a study to understand what works well and poorly in our solutions and what we can do to improve them. Moreover, we also tried to analyze scenarios other than the ones that were explored where concurrent speech, spatial audio, or extra information could be used and needed based on suggestions by our participants.

Participants saw our solutions as helpful for their daily activities. For most of them, it would not negatively impact their smartphone usage. However, as noted, it is crucial to be able to toggle the features quickly and customize them according to the user's needs.

The results acquired during the study allowed us to understand how visually impaired users interact with their smartphones, what features they need and are missing, and how we can provide an alternative way for them to interact with their smartphones.

## 7.1 Limitations

It is important to note that the solution we have implemented in this work is a prototype. The developed features are contained in the application we have created and are not available to be used system-wide. We made this decision because we needed more control over our scenarios. For example, suppose we were to develop something that could be used in the whole system, then we would need to create an accessibility service. However, accessibility services come with some limitations, one being that we would not be able to use concurrent speech since accessibility services are currently limited to a sequential audio channel. Since our objective was to explore different scenarios, we needed to control how they would behave. Ideally, screen readers should be the ones providing these capabilities.

## 7.2 Future Work

In this work, we focused on exploring different scenarios to augment smartphone usage efficiency. As explained in the limitations section, we opted to implement a prototype where all planned scenarios could be explored. Using a prototype, we did not compromise the quality of the solution (i.e., the scenarios work the same way they would if we opted for a solution that could be plugged into any smartphone) while significantly reducing the development time. However, this is a limitation of our work, and the next step would be to implement a solution that every smartphone user could use. From our perspective, these solutions can be provided in one of two ways.

1. Implementing an accessibility service that users can install on their smartphones to have access to the options provided in our scenarios. We have all of the required tools for this implementation, and as such, it could be started immediately. Moreover, this development could be done in iterations. So, for example, in the first phase, we could start by providing users the ability to listen to different audio sources simultaneously.

2. The existing screen readers, such as Talkback for Android and VoiceOver, would be responsible for providing these capabilities. This approach could be made independently by their respective development team. Alternatively, in the case of Talkback, since it is open-sourced, we could implement these capabilities ourselves and then make it known to the Talkback team so they could evaluate if it makes sense for them to be integrated into their screen reader.

# Bibliography

[1] Muhammad Abu ul Fazal, Sam Ferguson, Shuaib Karim, and Andrew Johnston. Vinfomize: a framework for multiple voice-based information communication. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, pages 143–147, 2019.

[2] Faisal Ahmed, Yevgen Borodin, Yury Puzis, and IV Ramakrishnan. Why read if you can skim: towards enabling faster screen reading. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10, 2012.

[3] Faisal Ahmed, Yevgen Borodin, Andrii Soviak, Muhammad Islam, IV Ramakrishnan, and Terri Hedgpeth. Accessible skimming: faster screen reading of web pages. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 367–378, 2012.

[4] Shiri Azenkot, Kyle Rector, Richard Ladner, and Jacob Wobbrock. Passchords: secure multi-touch authentication for blind people. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 159–166, 2012.

[5] Shiri Azenkot, Jacob O Wobbrock, Sanjana Prasain, and Richard E Ladner. Input finger detection for nonvisual touch screen text entry in perkinput. In *Proceedings of graphics interface 2012*, pages 121–129. 2012.

[6] Noam Ben-Asher, Niklas Kirschnick, Hanul Sieger, Joachim Meyer, Asaf Ben-Oved, and Sebastian Möller. On the need for different security methods on mobile phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 465–473, 2011.

[7] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Maria Teresa Paratore. Vibro-tactile enrichment improves blind user interaction with mobile touchscreens. In *IFIP Conference on Human-Computer Interaction*, pages 641–648. Springer, 2013.

[8] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Amaury Trujillo. Designing a text entry multimodal keypad for blind users of touchscreen mobile phones. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 131–136, 2014.

[9] Sen Chen, Chunyang Chen, Lingling Fan, Mingming Fan, Xian Zhan, and Yang Liu. Accessible or not an empirical investigation of android app accessibility. *IEEE Transactions on Software Engineering*, 2021.

[10] Emerson Barbosa da Cunha, Daniella Dias Cavalcante da Silva, and César Rocha Vasconcelos. Enabling full interaction with the android system and applications through speech recognition. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 497–500, 2017.

[11] Rafael Jeferson Pezzuto Damaceno, Juliana Cristina Braga, and Jesús Pascual Mena-Chalco. Mobile device accessibility for the visually impaired: problems mapping and recommendations. *Universal Access in the Information Society*, 17(2):421–435, 2018.

[12] Francisco Javier González-Cañete, José Luís López-Rodríguez, Pedro María Galdón, and Antonio Diaz-Estrella. Improving the screen exploration of smartphones using haptic icons for visually impaired users. *Sensors*, 21(15):5024, 2021.

[13] João Guerreiro. The use of concurrent speech to enhance blind people's scanning for relevant information. *ACM SIGACCESS Accessibility and Computing*, (111):42–46, 2015.

[14] João Guerreiro and Daniel Gonçalves. Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech. In *Proceedings of the 17th international ACM SIGACCESS conference on computers & accessibility*, pages 3–11, 2015.

[15] João Guerreiro and Daniel Gonçalves. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)*, 8(1):1–28, 2016.

[16] João Guerreiro, André Rodrigues, Kyle Montague, Tiago Guerreiro, Hugo Nicolau, and Daniel Gonçalves. Tablets get physical: non-visual text entry on tablet devices. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 39–42, 2015.

[17] Mohit Jain, Nirmalendu Diwakar, and Manohar Swaminathan. Smartphone usage by expert blind users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[18] Sungjune Jang, Lawrence H Kim, Kesler Tanner, Hiroshi Ishii, and Sean Follmer. Haptic edge display for mobile tactile interaction. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3706–3716, 2016.

[19] Esther Janse. Processing of fast speech by elderly listeners. *The Journal of the Acoustical Society of America*, 125(4):2361–2373, 2009.

[20] Shaun K Kane, Jeffrey P Bigham, and Jacob O Wobbrock. Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 73–80, 2008.

[21] Ravi Kuber, Amanda Hastings, and Matthew Tretter. Determining the accessibility of mobile screen readers for blind users. *UMBC Faculty Collection*, 2020.

[22] Orly Lahav, Jihad Kittany, Sharona T Levy, and Miriam Furst. Perception of sonified representations of complex systems by people who are blind. *Assistive Technology*, pages 1–9, 2019.

[23] Hyunchul Lim, YoonKyong Cho, Wonjong Rhee, and Bongwon Suh. Vi-bros: Tactile feedback for indoor navigation with a smartphone and a smartwatch. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2115–2120, 2015.

[24] Vikas Luthra and Sanjay Ghosh. Understanding, evaluating and analyzing touch screen gestures for visually impaired users in mobile environment. In *International Conference on Universal Access in Human-Computer Interaction*, pages 25–36. Springer, 2015.

[25] Tonja Machulla, Mauro Avila, Pawel Wozniak, Dillon Montag, and Albrecht Schmidt. Skim-reading strategies in sighted and visually-impaired individuals: a comparative study. In *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference*, pages 170–177, 2018.

[26] Pedro Maria Galdon, R Ignacio Madrid, Ernesto J De La Rubia-Cuestas, Antonio Diaz-Estrella, and Lourdes Gonzalez. Enhancing mobile phones for people with visual impairments through haptic icons: the effect of learning processes. *Assistive technology*, 25(2):80–87, 2013.

[27] Fabrice Maurel, Gaël Dias, Stéphane Ferrari, Judith-Jeyafreeda Andrew, and Emmanuel Giguet. Concurrent speech synthesis to improve document first glance for the blind. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 3, pages 10–17. IEEE, 2019.

[28] Hugo Nicolau, André Rodrigues, André Santos, Tiago Guerreiro, Kyle Montague, and João Guerreiro. The design space of nonvisual word completion. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 249–261, 2019.

[29] Uran Oh, Shaun K Kane, and Leah Findlater. Follow that sound: using sonification and corrective verbal feedback to teach touchscreen gestures. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–8, 2013.

[30] João Oliveira, Tiago Guerreiro, Hugo Nicolau, Joaquim Jorge, and Daniel Gonçalves. Blind people and mobile touch-based text-entry: acknowledging the need for different flavors. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 179–186, 2011.

[31] Konstantinos Papadopoulos, Evangelia Katemidou, Athanasios Koutsoklenis, and Eirini Mouratidou. Differences among sighted individuals and individuals with visual impairments in word intelligibility presented via synthetic and natural speech. *Augmentative and Alternative Communication*, 26(4):278–288, 2010.

[32] Benjamin Poppinga, Charlotte Magnusson, Martin Pielot, and Kirsten Rassmus-Gröhn. Touchover map: audio-tactile exploration of interactive maps. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 545–550, 2011.

[33] Gisela Reyes-Cruz, Joel E Fischer, and Stuart Reeves. Reframing disability as competency: Unpacking everyday technology practices of people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[34] André Rodrigues, Kyle Montague, Hugo Nicolau, João Guerreiro, and Tiago Guerreiro. In-context q&a to support blind people using smartphones. In *Proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility*, pages 32–36, 2017.

[35] André Rodrigues, Kyle Montague, Hugo Nicolau, and Tiago Guerreiro. Getting smartphones to talkback: Understanding the smartphone adoption process of blind users. In *Proceedings of the 17th international acm sigaccess conference on computers & accessibility*, pages 23–32, 2015.

[36] André Rodrigues, Hugo Nicolau, Kyle Montague, João Guerreiro, and Tiago Guerreiro. Open challenges of blind people using smartphones. *International Journal of Human–Computer Interaction*, 36(17):1605–1622, 2020.

[37] Amanda Stent, Ann Syrdal, and Taniya Mishra. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 211–218, 2011.

[38] Jürgen Trouvain. On the comprehension of extremely fast synthetic speech. 2007.

[39] Radu-Daniel Vatavu. Visual impairments and mobile touchscreen interaction: state-of-the-art, causes of visual impairment, and design guidelines. *International Journal of Human–Computer Interaction*, 33(6):486–509, 2017.

[40] Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, et al. Speech

synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, 2019.

[41] Tsubasa Yoshida, Kris M Kitani, Hideki Koike, Serge Belongie, and Kevin Schlei. Edgesonic: image feature sonification for the visually impaired. In *Proceedings of the 2nd Augmented Human International Conference*, pages 1–4, 2011.

[42] Yu Zhong, TV Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P Bigham. Justspeak: enabling universal voice control on android. In *Proceedings of the 11th Web for All Conference*, pages 1–4, 2014.

# Appendix A

# User Study - Codebook

Table A.1: Study codebook

| Code/Label | Description | Example |
|---|---|---|
| **1. User Profile** | How the user profile affects the scenarios | “Here the user should be able to choose the sound” |
| 1.1 Memorization | The user should be familiared with the sounds; knowing what the sounds represents requires practice | “We first have to memorize it “ |
| 1.2 Hearing loss | How hearing loss affects the audio comprehension | “But I had more difficulty on the left because I am a little hard of hearing” |
| **2. Perception** | How the content is perceived | “sequentially is probably more logical and gives us more information I think” |
| 2.1 End signal | The need for an end signal after the text property is mentioned | “You get a good idea of when it starts and when it ends because that ”txe” sound helps a lot” |
| 2.2 End cue | The need for a cue to change focus between multiple audio sources is mentioned | “it is effectively necessary to give this signal beforehand so that the person has time to trigger the distribution of attention” |
| 2.3 Sequential | Participants mention they prefer to read content sequentially | “I probably prefer to listen sequentially” |
| **3. Preferences** | How would the participants prefer to explore a given scenario | “The sound each person has to adjust to his or her needs” |
| 3.1 Customization | The ability to fully customize an option is mentioned | “Here the user should be able to choose the sound.” |
| 3.2 Sonification | The substituion of the announcement of the element types with sound feedback is mentioned | “by the sound we knew right away that it was a button” |

| | | |
|---|---|---|
| 3.3 Word feedback | Announcing the text properties with words is mentioned | “designation made with words so that we have no doubt what it is about” |
| 3.4 Audio content | Hearing topics of interest helps in capturing the information is mentioned | “things I liked to hear, it would be easier “ |
| **4. Suggestions** | Suggestions made by participants | “I would prefer it to be more specific. For example, I wanted Zara or I wanted cinema” |
| 4.1 Source selection | The ability to select a an audio source of interest is mentioned | “Once you get the news that interests you, to be able to stop listening to the others.” |
| 4.2 Directions | The suggestion of adding directions to the map scenario is mentioned | “It would also be helpful to say, besides the meters, left, right, etc.” |
| 4.3 GPS integration | GPS integration is mentioned | “In fact, Google Maps should do just that by now.” |
| **5. Smartphone usage** | Aspects of interacting with a smartphone are mentioned | “it is useful to listen to two at a time to move faster in the news” |
| 5.1 Gesture confirmation | Sound feedback during gestures is mentioned | “With the different sound I automatically know that it has escaped to the left side and I will not continue” |
| 5.2 Audio properties | Exploration of text properties | “suppose there is some information that should be highlighted or underlined I think it would be good” |
| 5.2.1 Sound feedback | Indication of text properties with sound is mentioned as not demanding | “the sounds could help and it would be more enjoyable, not so dull” |
| 5.2.2 Difficulties | The difficulties while hearing the text properties | “And you know that 3 sounds or 4 is already a bit... And imagine that a link appears that is also bold and underlined” |
| 5.2.2.1 Overwhelming | Text properties are mentioned as overwhelming | “Yeah, it would get more tiring. When I’m on the computer reading I take everything out. Dots, quotation marks, commas... Listening to a book is not pleasant at all” |
| 5.2.2.2 Distracting | Text properties are mentioned as distracting | “lloses the reading of anything.” |
| 5.2.3 Advantages | Advantages of knowing the text properties | “when we are doing some work and need to know the formatting” |
| 5.2.3.1 Knowledge | Knowing extra information about the text is mentioned | “Since we do not see it is a way to get some extra information.” |

| | | |
|---|---|---|
| 5.3 Concurrent speech | Aspects of concurrent speech are mentioned | "I think two at the same time can be." |
| 5.3.1 Different voices | How different voices help understand different audio sources | "A slightly different voice so as not to confuse the two sources and distinguish" |
| 5.3.2 Short messages | How there is no difficulty in hearing short messages together with other lengthier content | "and I am reading a news article and get a short message I can read." |
| 5.3.3 Dificulties | Difficulties experienced while exploring scenarios where concurrent speech is used | "I was listening to one and then changed my focus to the other. So I could not understand it" |
| 5.3.3.1 Words mix up | Participants mention that when text properties are played at the same time as the text, they get confused | "without a pause one can't, at least I couldn't attribute exactly which word had the characteristic that was referred to" |
| 5.3.3.2 Tiring | Participants mention that hearing several things concurrently is tiring | "Here I think the reading would be exhausting." |
| 5.3.3.3 Confusing | Loss of information is mentioned | "Two is good, three is not. You can't absorb" |
| 5.3.3.4 Demanding | Participants mention that more attention is needed | "I end up forgetting the other one because I was focusing on one." |
| 5.3.4 Advantages | Advantages of being able to hear several things at the same time | "can be useful in that, I don't know, the notification can be a message that is more urgent" |
| 5.3.4.1 Efficiency | The ability to hear more things in less time is mentioned | "it's a matter of being faster instead of just reading one single piece of news" |
| 5.3.4.2 Immediate feedback | The ability to check information immediately without interrupting the task at hands is mentioned | "Any information that is received. For example if we are talking on the phone and something comes in, we should be informed" |
| **6. Spatialization** | The effects of spatialization on the scenarios | "I like to have things differentiated" |
| 6.1 Difficulties | Negative effects of spatialization | "I could only understand the Telepizza that was 50 meters away" |
| 6.1.1 Capturing information | Difficulties in understanding spatial audio is mentioned | "the only thing I didn't quite understand was the meters" |
| 6.2 Advantages | Advantages of using spatialization | "makes it easier because we can pay attention on one side to one thing and on the other side to another" |

| | | |
|---|---|---|
| 6.2.1 Listen to concurrent audio | Participants mention that hearing to different audio sources is easier with spatialization | “I prefer the news on the left side and the message in the right ear” |
| 6.2.2 Navigation | Participants mention that the use of spatialization to indicate proximity is useful | “And I can also tell it’s further away because of the sound.” |

# Appendix B

# Scenarios - Youtube playlist

We made the scenarios explored during our study available on Youtube. They can be consulted via the following Youtube Playlist link