Scanning for Digital Content: How Blind and Sighted People Perceive Concurrent Speech

JOÃO GUERREIRO and DANIEL GONÇALVES, Instituto Superior Técnico, Universidade de Lisboa/INESC-ID

The widespread availability of digital media has changed the way that people consume information and has impacted the consumption of auditory information. Despite this recent popularity among sighted people, the use of auditory feedback to access digital information is not new for visually impaired users. However, its sequential nature undermines both blind and sighted people's ability to efficiently find relevant information in the midst of several potentially useful items. We propose taking advantage of the *Cocktail Party Effect*, which states that people are able to focus on a single speech source among several conversations, but still identify relevant content in the background. Therefore, in contrast to one sequential speech channel, we hypothesize that people can leverage concurrent speech channels to quickly get the gist of digital information. In this article, we present an experiment with 46 (23 blind, 23 sighted) participants, which aims to understand people's ability to search for relevant content listening to two, three, or four concurrent speech channels. Our results suggest that both blind and sighted people are able to process concurrent speech in scanning scenarios. In particular, the use of two concurrent sources may be used both to identify and understand the content of the relevant sentence. Moreover, three sources may be used for most people depending on the task intelligibility demands and user characteristics. Contrasting with related work, the use of different voices did not affect the perception of concurrent speech but was highly preferred by participants. To complement the analysis, we propose a set of scenarios that may benefit from the use of concurrent speech sources, for both blind and sighted people, toward a *Design for All* paradigm.

CCS Concepts: • **Human-centered computing** → **Sound-based input/output**; **Auditory feedback**; **Empirical studies in accessibility**; *Accessibility systems and tools;*

Additional Key Words and Phrases: Cocktail party effect, screen reader, blind, visually impaired, sighted, skimming, scanning, concurrent speech, spatial audio, web browsing, accessibility, auditory feedback, simultaneous speech

ACM Reference Format:
João Guerreiro and Daniel Gonçalves. 2016. Scanning for digital content: How blind and sighted people perceive concurrent speech. ACM Trans. Access. Comput. 8, 1, Article 2 (January 2016), 28 pages.
DOI: http://dx.doi.org/10.1145/2822910

1. INTRODUCTION

We live in a time in which the widespread availability of digital media in a myriad of devices has changed people's daily lives. It has changed not only the amount of information people consume, but also how people consume information. The rising popularity of auditory media [Peoples and Tilley 2011] is a good example of how media consumption has changed. People *listen to* books (and podcasts) when performing tasks that require the use of the visual channel, such as running or driving, to name a few. For the case of

This work is supported by the Portuguese Foundation for Science and Technology, under grant nos. UID/CEC/50021/2013 and INCENTIVO/EEI/LA0021/2014.
Authors' addresses: J. Guerreiro and D. Gonçalves, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal; emails: joao.p.guerreiro@ist.utl.pt, daniel.goncalves@inesc-id.pt.

ACM 1936-7228/2016/01-ART2 $15.00
DOI: http://dx.doi.org/10.1145/2822910

visually impaired people, the auditory feedback of screen readers has a central role in providing access to digital information. However, unlike the visual spatial presentation on screen that may depict a lot of information at a time, the traditional auditory feedback relies on a sequential channel that impairs a quick overview of the content. For example, sighted users are able to quickly sift through a document or website by looking through visually prominent content or by catching separate phrases and slowing down depending on their needs (so-called *diagonal reading* [Backhaus and Tuor 2015]). These skills allow the individual to get a general idea of the content—*skimming*—or to find specific information—*scanning* [Ahmed et al. 2012b].

Previous research on blind people's browsing behaviors has shown that they develop their own strategies [Borodin and Bigham 2010; Vigo and Harper 2013; Watanabe 2007], such as navigating through headings and increasing the speech rate, which help them to mitigate the lack of visual feedback. However, the sequential characteristic of auditory feedback impairs a quicker overview of the content when compared to the visual presentation on screen. This results in less efficiency when searching for potentially relevant information [Borodin and Bigham 2010], particularly when compared to sighted people's browsing techniques [Bigham et al. 2007; Takagi et al. 2007].

Traditional auditory feedback, however, does not take advantage of the human ability to process concurrent, parallel speech channels. The *Cocktail Party Effect* states the human ability to focus attention on a single talker among several conversations and background noise [Cherry 1953]. Moreover, one may detect interesting content in the background (e.g., one's own name or favorite subject) and shift the attention to another talker.

There is evidence that the human brain's ability to segregate simultaneous speech depends on characteristics such as the number of concurrent talkers [Brungart and Simpson 2005a], their differences in spatial locations [Brungart and Simpson 2005a, 2005b], or voice characteristics [Brungart and Simpson 2005a; Darwin et al. 2003; Vestergaard et al. 2009], among others. In fact, a good configuration of these characteristics enhances the speech intelligibility for both selective [Brungart and Simpson 2005a; Darwin et al. 2003] and divided attention [Shinn-Cunningham and Ihlefeld 2004; Vazquez-alvarez and Brewster 2011] tasks. In the former, one focuses the attention on a specific talker; in the latter, the attention is divided among several speech sources. Furthermore, most experiments that focus on speech intelligibility use small phrases, wherein the participants have to identify all words. However, with longer sentences, people may have enough time to achieve a basic understanding of the text, and therefore perform scanning and/or skimming tasks more efficiently. This hypothesis is supported by Cherry's [1953] pioneer study, which reported one's ability to perceive an entire *cliché* by hearing just a few words.

Previous research stated that blind people, particularly early-blind, have enhanced capabilities to segregate speech signals [Niemeyer and Starlinger 1981; Hugdahl et al. 2004] due to the process of *neuroplasticity*. In the particular case of blindness, it states that a blind person's brain is reorganized so that part of one's visual cortex is used for other purposes, including auditory processing [Burton 2003]. Harper highlighted this advantage of visually impaired people [2012] when suggesting the use of simultaneous audio sources to convey web information faster to visually impaired users.

In this article, we argue that both blind and sighted people can leverage the Cocktail Party Effect to scan for relevant information more efficiently. Moreover, we identify the differences between these populations, as previous comparisons focus on very small speech signals (e.g., syllables [Hugdahl et al. 2004]). As a use case scenario, while exploring news sites, one may be targeting specific subjects to which to pay further attention. Instead of listening to all headings sequentially, one could listen to two or three simultaneously to detect the relevant ones. We argue that the use of concurrent

speech enables blind and sighted people to listen to several unrelated information items (e.g., articles in news sites, podcast summaries, search results, and social media posts), get the gist of the information and identify the ones that deserve further attention.

We report on an experiment with 46 people (23 blind, 23 sighted) designed to evaluate people's perception of concurrent speech while scanning for relevant information. In particular, we address the following questions: (1) How many voices can blind and sighted people listen to and still be able to identify the one with relevant content? (2) How many voices can blind and sighted people listen to and understand the relevant one? (3) Do differences in voice characteristics enhance both identification and selective attention? (4) Do blind and sighted people perform differently?

A main result from this experiment was that blind and sighted participants appear to be performing in the same way, suggesting that both groups may take advantage of concurrent speech in scanning tasks. The absence of significant differences between these user groups promotes new approaches and interfaces that target wider audiences, rather than very specific solutions focused exclusively on blind people. In particular, our results suggest that the identification of the relevant source is a straightforward task when listening to two simultaneous talkers, and most participants were still able to identify it with three talkers. Moreover, both two and three simultaneous voices may be used to understand the relevant source's content depending on speech intelligibility demands and user characteristics.

This article is an expanded version of a conference paper in the *16th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)* [Guerreiro and Gonçalves 2014], which presents an evaluation of the perception of concurrent speech by blind people. We extend this experiment with the inclusion of sighted participants and consequent analysis of their performance. We report a comparison between both populations and discuss the scenarios in which the use of concurrent speech may benefit both sighted and blind people. Moreover, we provide a qualitative analysis in greater detail.

2. BACKGROUND ON THE COCKTAIL PARTY EFFECT

The *Cocktail Party Effect* states the human ability to focus the attention on a single talker among several conversations and background noise [Cherry 1953]. Moreover, one may detect interesting content in the background (e.g., one's own name or favorite subject) and shift attention accordingly. Several researchers investigated how concurrent speech intelligibility can be maximized. In humans, *Auditory Scene Analysis* (ASA) is the process in which the auditory system segregates a mixture of sounds into different streams [Bregman 1990], and relies on features such as sound spatial locations, frequencies, and synchrony (e.g., Carlyon [2004], Turgeon et al. [2005], and Brungart and Simpson [2005a]).

Although intelligibility decreases with the increase in competing talkers, the separation of speech signals between ears (dichotic speech) outperforms the use of mixed signals (monaural speech) [Cherry 1953] and is surpassed only by spatial audio [Brungart and Simpson 2005a; Drullman and Bronkhorst 2000]. The localization of sound relies mainly on two interaural cues: Interaural Level Differences (ILDs) and Interaural Time Differences (ITDs). ILDs refer to the differences in the sound level/intensity at each ear, and are used mainly at frequencies above 1.5kHz. At frequencies below this threshold, such as speech, people are more sensitive to differences in the arrival time of the sound at the two ears (ITDs) [Feddersen et al. 1957; Wright and Fitzgerald 2001], which contribute to the segregation of sound into different streams (e.g., Darwin [1997], Bronkhorst [2000], and Schwartz et al. [2012]).

Many experiments (e.g., Brungart and Simpson [2005a] and Darwin et al. [2003]) measured concurrent speech intelligibility using the Coordinate Response Measure

(CRM), a test originally used for military communications [Moore 1995]. In this test, the listener hears simultaneous phrases of the type "*Ready, (call sign), go to (color) (number) now.*" Each concurrent phrase has one of eight call signs ("Baron," "Charlie," "Ringo," and so on), one of four colors (red, blue, green, white) and one of eight numbers (1–8). The participants have to listen to the target phrase containing a specific call sign (usually "*Baron*") and report the corresponding color and number. Using this measure, Brungart and Simpson [2005a] showed that spatial separation enhanced the perception of a target speech signal by 25% with one interfering talker and near 50% with two and three when compared with dichotic speech. In most experiments, including Brungart's, the speech sources were equally spaced in the frontal horizontal plane, because former research suggested that identifying a speech signal location is more difficult when varying the vertical position and distance [Wenzel et al. 1993; Yost 1998]. Additionally, a probable reason behind the use of the frontal horizontal plane is to avoid the *front-back confusion* found in previous studies on sound localization [Wenzel et al. 1993]. This confusion refers to the localization of a sound source in the front that should be in the back, and the other way around.

Other researchers compared the identification and/or intelligibility of speech using different settings in which the sound sources may have different distances to the listener or have differently spaced locations (e.g., Brungart and Simpson [2005b]). Although some of these settings provided better results overall, results were more disparate among the different locations. For instance, people were able to understand better the sound sources that were closer, but that reduced the intelligibility of the ones that were furthest away.

The impact of spatial separation can be minimized for people with spatial hearing loss, however (more frequent as people age), as it may influence the ability to understand speech in multitalker contexts [Byrne and Noble 1998; Litovsky et al. 2009]. To cite one example, children with spatial hearing loss have more listening difficulties in the classroom [Cameron and Dillon 2008]. In addition, more often than not, experiments about ear dominance in the ability to process concurrent speech report a right ear advantage (e.g., Dirks [1964] and Shankweiler and Studdert-Kennedy [1967]), which is also suggested by Brungart and Simpson [2005b].

Although most experiments focus on selective attention tasks, for which people have to pay attention to a single speech signal while neglecting the others, the use of spatial audio is also valuable in divided attention tasks, in which the attention is divided between two speech signals [Shinn-Cunningham and Ihlefeld 2004; Vazquez-alvarez and Brewster 2011].

Another important feature in sound segregation is the use of voices with different frequencies (pitch and formant frequencies). Brungart and Simpson [2005a], Darwin et al. [2003], and Vestergaard [2009] showed the advantage in using different gender talkers. The ability to hear frequencies separately—*frequency selectivity*—occurs in the basilar membrane in the cochlea (in the inner ear). It has a tonotopic organization in which different frequencies are processed at different places along the membrane [Bear et al. 2006; Talavage et al. 2004]. Each place on the membrane behaves like an auditory filter that responds to the frequencies within its range [Moore and Gockel 2011].

The increasing use of speech and sound in the interaction of humans with computers may leverage this phenomenon to provide information more efficiently and/or effectively. Actually, a few studies report that blind people, in particular early-blind, are more capable in discriminating speech than sighted people are, due to the process of *neuroplasticity* [Hugdahl et al. 2004; Niemeyer and Starlinger 1981]. The concept of neuroplasticity states that areas of the brain that are not used (in this case, the visual cortex) are reorganized for different purposes [Burton 2003]. To cite one example, an experiment in which two speech stimuli were presented at the same time to each ear

(dichotic) evinced the greater capabilities of blind people to understand and report speech [Hugdahl et al. 2004]. Therein, users had to identify consonant-vowel (CV) syllables in three conditions: no specific conditions regarded attention (Nonforced, NF); pay attention to the syllables on the right ear (Forced Right, FR); and pay attention to the syllables on the left ear (Forced Left, FL). The blind participants had significantly more correct answers from the right ear in the NF condition. They were also better in the forced conditions, mainly in the FL. Although, in this case, *neuroplasticity* is usually associated with congenital or early-blind people, other researchers support that it occurs also in late-blind people's brains (e.g., Kujala et al. [1997] and Théoret et al. [2004]).

3. RELATED WORK

The work reviewed in this section is twofold: first, we look into the research and techniques that aim to accelerate blind people's textual scanning; second, we present research that takes advantage of simultaneous speech feedback both for blind and sighted people.

3.1. Fast-Reading Techniques

The aforementioned browsing behaviors developed by blind people are crucial to overcome accessibility limitations, but also to enable more efficient browsing and information scanning. For instance, a proper use of the *Heading* elements and skip links can significantly speed up web browsing, particularly for scanning tasks [Takagi et al. 2007; Watanabe 2007].

Moreover, previous research has shown that visually impaired people are able to listen to and understand synthesized speech at higher speech rates than sighted people [Papadopoulos 2010; Trouvain 2007]. However, this ability depends on factors such as the person's age and familiarity with a synthesizer and voice [Stent et al. 2011]. One experiment that tried to understand blind people's maximum listening speeds has shown that advanced users were able to listen to a screen reader 2.8 times faster than the default rate and still understand at least 50 percent of the information [Asakawa et al. 2003]. Moreover, novice users were able to listen to a screen reader 1.6 times faster than the default rate and be able to understand all the content. At that time, the authors reported the need for an easy and interactive way to change the speech rate with immediate response [Asakawa et al. 2003]. This feature is now supported through keyboard shortcuts by mainstream screen readers, which may help avoid the exhaustion caused by very high speech rates [Trouvain 2007].

A frequent approach to surpassing the lack of efficiency in browsing digital content is summarization (e.g., Ahmed et al. [2012a] and Harper and Patel [2005]). Ahmed et al. [2012a] stand out by creating a method that eases changing between the summaries and the original text. Other researchers try to provide a privileged access and navigation to key page sections and content (e.g., Yesilada et al. [2007], Lunn et al. [2011], Mahmud et al. [2007], and Gadde and Bolchini [2014]). Another approach is the use of *sonification*, which can be defined as the *"use of non-speech audio to convey information"* [Kramer et al. 2010], such as *auditory icons* [Gaver 1986], *earcons* [Brewster et al. 1993] and *spearcons* [Walker et al. 2006]. For example, *BlindSight* [Li et al. 2008] uses *earcons* to inform the availability of a calendar slot during a phone conversation. Although such strategies and approaches enable a more efficient consumption of information, "*the biggest problem in non-visual browsing remains the speed of information processing*" [Borodin and Bigham 2010]. Therefore, other approaches, combined or not with current browsing techniques, are needed to accelerate blind people's information processing.

3.2. Simultaneous Sound Applications

The insights provided by the experiments inspired in the *Cocktail Party Effect*'s paradigm led to applications that try to take advantage of concurrent speech to present larger amounts of information more efficiently. *Sasayaki* [Sato et al. 2011] provides the output of a standard auditory browser, augmented with a whispering voice channel used, for example, to locate the screen reader position in the web page or to provide important contextual information. Other authors introduced spatial audio to map the current position in a web page, while a different voice provided other information [Crispien et al. 1996; Goose and Möller 1999].

Another example is *Clique* [Parente 2006], which places 4 assistants with distinct voices around the user in a virtual sound space. Each assistant has a role involving tasks or events (e.g., email, calendar, and browser activity) and is able to use conversation features such as referencing, pacing, and turn taking.

AudioStreamer [Schmandt and Mullins 1995] uses 3 speech sources from audio news programs in the frontal horizontal plane (1 ahead and others 60 degrees on both sides) and enhances the signal of the one that is the current focus of interest. To select the current focus, it captures the gesture of turning the face to the sound's direction. Similarly, Sodnik and Tomažič [2009] present different files (two or three) in different spatial locations. Participants were able to keep track of two simultaneous files; yet, when three were presented, they were able to focus on only a single file.

Aoki et al. [2003] presented a social audio space supporting multiple simultaneous conversations. They monitored the participants' behavior to identify conversational floors as they emerge and to modify the audio delivered to each participant, enhancing the signals of interest. *SpeechSkimmer* [Arons 1997] tries to present recorded speech faster by presenting the most important segments to one ear and the discarded material to the other. *SpatialTouch* [Guerreiro et al. 2015b] is a nonvisual multitouch *QWERTY* keyboard for tablet devices that enables blind users to leverage previous experience in physical keyboards, by supporting two-handed input through spatial and simultaneous audio feedback. It relies on male and female voices, for which spatial location on the frontal horizontal plane depend on each character position in the keyboard. Similarly, we supported two-handed exploration of large touch surfaces using simultaneous spatial audio feedback [Guerreiro et al. 2015a]. In this solution, each hand was mapped to a specific voice (male or female) and location (left or right ear) to aid interaction feedback distinction.

These applications are valuable contributions for their scenarios and tasks; yet, there are no guarantees that they are suitable when scanning for relevant content.

4. EXPERIMENTAL SETTING

In Guerreiro and Gonçalves [2014], we investigated blind people's ability to cope with simultaneous speech in fast-reading tasks such as scanning for relevant content. Afterwards, we conducted the exact same experiment with sighted people. This experiment with both blind and sighted users intend to answer the following research questions: (1) How many voices can blind and sighted people listen to, and still be able to identify the one with relevant content? (2) How many voices can blind and sighted people listen to, and understand the relevant one? (3) Do differences in voice characteristics enhance both identification and selective attention? (4) Do sighted and blind people perform differently?

4.1. Text-to-Speeches

In order to evaluate the perception of concurrent speech by both blind and sighted people, we built the *Text-to-Speeches* framework. This framework was first described in Guerreiro and Gonçalves [2014].

Text-to-Speeches is able to position several prerecorded audio files in a 3D space simultaneously. We built a Java framework on top of Paul Lamb's 3D Sound System [Lamb 2015], using the LightWeight Java Game Library [LWJGL 2.0 2015] binding of OpenAL Soft 1.15.1 [OpenAL Soft 2014]. This setting supports the use of digital filters called Head-Related Transfer Functions (HRTFs), which simulate the acoustic cues used for spatial localization [Wenzel et al. 1993]. These filters are able to reproduce both times of arrival and intensity differences of both ears to spatially locate the sources around a person's head. The HRTFs are based on measurements influenced by the listener's head and ears. Like most experiments (e.g., Brungart and Simpson [2005a] and Shinn-Cunningham and Ihlefeld [2004]) and for simplicity purposes, we used nonindividualized measurements from a KEMAR manikin (in this case, from MIT [Gardner and Martin 2000]).

Current *Text-to-Speech* software demands a unique, sequential auditory channel. Therefore, we prerecorded all sentences to .wav files, using DIXI [Paulo et al. 2008], a TTS developed by INESC-ID's Spoken Language Systems Laboratory [L2F 2015] and now commercialized by Voice Interaction [Voice Interaction 2014] (Vicente's male voice). These audio files are then placed at different positions in the 3D audio space.

To guarantee different and controlled voices, we manipulated our original voice's pitch (Glottal Pulse Rate, GPR) and formant frequencies (Vocal Tract Length, VTL), using the *Praat* software [Boersma and Weenink 2014] the same way that Darwin et al. [2003] did. Moreover, we measured the sound intensity levels of all sound files, in decibels. For each voice, we calculated a mean intensity value based on all levels from the 103 snippets. Then, we used *Praat* to adjust the voice intensities so that all voices had the same mean intensity level.

4.2. Methodology

The experiment methodology was the same for blind [Guerreiro and Gonçalves 2014] and sighted people. The multitalker environment was set up based on previous work in which the Cocktail Party Effect was investigated. In addition, in this experiment, all sound sources are equally important as all of them may have the information one is searching for. Hence, the selected configurations were designed to avoid overbalancing any of the sources. For instance, we decided not to use a different onset time and volume for each voice, as it would benefit some voices over others. In what follows, we describe our setting regarding the number of talkers, their spatial location, and voice characteristics.

4.2.1. Number of Talkers and their Location. Our main research questions focus on the number of simultaneous talkers that blind and sighted people can listen to and still identify and understand the content of the relevant one. The related work identified a constant decrease in performance as the number of talkers increased, whereas results are nearly 50% of success with four speech sources [Brungart and Simpson 2005a]. Although these results focus on different tasks, they were a good indicator for the number of sources that we should consider. We decided to conduct the experiment with two, three, and four simultaneous talkers.

The sound source locations took inspiration from several experiments that use equally spaced positions in the frontal horizontal plane (e.g., Brungart and Simpson [2005b] and Drullman and Bronkhorst [2000]). Although other spatial configurations (e.g., with differently spaced locations) were proposed and provided better results overall [Brungart and Simpson 2005b], they ended up sacrificing specific locations that dropped their results significantly. Figure 1 shows our spatial setting. The sound sources are separated by 180°, 90°, and 60°, for two, three, and four talkers, respectively.

4.2.2. Voice Conditions. Most experiments on simultaneous speech segregation focus on pitch variations. Yet, the best results are achieved when varying the two main

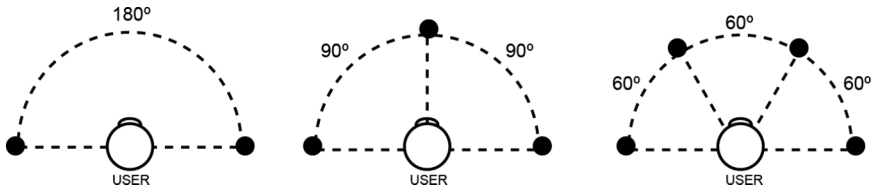

Fig. 1. The sound source spatial positioning in the user's frontal horizontal plane for two, three, and four talkers.

characteristics that influence male and female voices: the pitch (GPR) and formant frequencies (VTL) [Brungart and Simpson 2005a; Darwin et al. 2003].

We wanted to validate these differences for longer speech signals. We used a single voice whose characteristics were manipulated to obtain different voices. Similar to Vestergaard et al. [2009], this central voice (an androgynous talker), was obtained by manipulating a male's voice. Such variations enabled us to measure the effects of pitch and formant frequencies together while excluding other factors such as intonation or prosody. Moreover, this option favored a consistent variation toward both male and female talkers, rather than the predominance of one gender in the experiments. The analysis of previous research resulted in three conditions:

—*Same*. In this baseline condition, all talkers have the same central voice previously mentioned. This voice has a mean pitch of 155Hz and VTL of 147mm.
—*Large*. This condition aimed at the larger known separation that could still provide an improvement in performance, for both pitch and VTL variations [Darwin et al. 2003]. In this condition, each voice differs from the subsequent voice in a distance of 7.4 semitones (a ratio of 1.53) in pitch, and a 0.88 ratio in VTL. For instance, with two voices, the mean pitch values were approximately 125.6 and 192.2Hz, while with three voices they were 155Hz (the central voice), 237.2Hz, and 100.8 Hz. The central, androgynous voice was manipulated to obtain all the others. This rather large separation between voices, when listening to four talkers, results in two extreme voices similar to Darwin's super male and super female, which deviate from normal human voices [Darwin et al. 2003].
—*Small*. This condition has half the variation (3.7 semitones in pitch and a 0.945 ratio in VTL) than the previous condition. This option guaranteed the use of human-like voices for all talkers (including with four). Moreover, these values are very close to the larger separation in the study by Vestergaard [2009].

5. DATA COLLECTION

People search daily for information among search engine results, posts, tweets, mail messages, or news. The process of deciding which pieces of information are relevant deserves further attention. We centered our task in this frequent need: *Relevance Scanning*. Among some distractors, the participants have to identify the relevant message and try to understand its content.

In this experiment, the dataset consists of 103 news snippets from a Portuguese news site. The snippets contain only raw text and have consistent sizes, so that all sources stop emitting the information about the same time.

In a first phase, a larger amount of snippets were randomly selected from the news site. Afterwards, we selected all the ones (103) that held the following constraints: contained only Portuguese words, correctly pronounced by the TTS; all resulting audio files had durations between 10s and 11s. Moreover, we changed names, places, and any other element that could benefit from previous knowledge of particular news or subjects. During the experiment, the 103 snippets were chosen randomly such that

none was presented twice per participant. The topics of the snippets include sports, politics, celebrities, television, and science.

The task consisted of finding the relevant snippet among the presented snippets at each trial (there could be two, three, or four simultaneous sources) and trying to understand its content. Before the trial, one relevant snippet was randomly selected and the researcher provided a set of cues (consistent across participants), which worked as a hint, to simulate the search for relevant information.

5.1. Procedure

The experiment consisted of two phases that were conducted in the same session: one to assess the participants' profiles and a second to investigate the perception of concurrent speech. It was conducted at a training center for blind and visually impaired people (*Fundação Raquel e Martin Sain*) and at our lab (*INESC-ID*). The characterization session took approximately 15min and included an oral questionnaire about demographic data and screen reader usage, as well as a working memory assessment. To measure the working memory, the subtest Digit Span of the revised Wechsler Adult Intelligence Scale (WAIS-R) [Wechsler 1981] was used. In a first phase, the participant must repeat increasingly long series of digits presented orally; in a second stage, the participant must repeat additional sets of numbers, but backwards. Such tasks allow the calculation of a grade that is known to correlate with the participant's working memory, as they capture a part of people's attention mechanisms and executive function.

At the beginning of the evaluation phase, participants were told that the overall purpose of the experiment was to investigate the perception of concurrent speech for its potential use in future technological solutions. We then explained the experimental setup and adjusted the headphones' volume to a level comfortable for each participant, using two trials with a single speech signal.

The evaluation consisted of one practice trial and six test trials for each possible number of talkers (two, three, and four talkers) and had a fixed ascending order. We based this decision on our objective to investigate the maximum number of simultaneous talkers, instead of a fair comparison between them. This option takes advantage of the previous trials, with fewer talkers, as practice. Moreover, we did not complete the condition with four talkers, to avoid participants' fatigue and/or frustration, when the participant missed more than half the questions with three talkers; or when the participant was not able to identify the first three with four talkers. Fifteen blind and 18 sighted participants completed the condition with four talkers.

The six trials followed a completely randomized order and consisted of two trials for each voice characteristics condition (*same*, *large*, and *small*). We ensured that both the voice (except in same condition) and the location of the relevant source were different for those two repeated trials. Each trial consisted of the following five phases:

—*Hint given by the researcher*. The researcher gives a hint about which news/sentence the participant should pay attention to. This hint consists of the three most important and defining words in the beginning of the sentence (in the first five words, excluding prepositions and connectors). It enabled the participants to understand the sentence subject and provided a clear distinction between news. This procedure is similar to the one performed in Brungart and Simpson [2005a] and Darwin et al. [2003], but they use only one word due to their smaller sentences (CRM).
—*Play simultaneous speech*. The simultaneous sentences start to play at the exact same time. The participant tries to identify the relevant sentence and understand as much content as possible.
—*Participant's Report*. Participants report the content of the relevant sentence. They are encouraged to reveal everything that they heard and remember, using the same or

different words. Related experiments [Brungart and Simpson 2005a; Darwin et al. 2003] ask participants to report the exact same words. In this case, we want to understand if people can get the gist of the information, independently of the words perceived.

—*Question*. The researcher asks a question about the relevant sentence only if the participant did not provide the answer in the previous report. This question is used to help recall some of the previously heard content. All sentences have a predefined question whose answer is not in the first three seconds or in the final two seconds.

—*Identification*. The researcher asks whether the participant was able to identify the relevant source and to describe which of them it was. Participants could use the location, voice, or every other way to describe the sound source. We intended to assess the easiest way to define a specific sound source.

After the six trials per number of talkers, we asked for participants' feedback. After completing all trials, participants were asked to fill out a questionnaire in which they rated a set of sentences using Likert-type items with a five-point scale. These sentences included rating their ability to understand the relevant sentence, their comfort listening to the N number of voices, and the effect of spatial location and voice differences. The evaluation procedure took on average 45min.

5.2. Apparatus

The *Text-To-Speeches* framework, previously described in Section 4.1, was used in the experiment. Participants used AKG K540 headphones that were connected to an audio interface—Saffire Focusrite PRO 40—to enhance audio quality. The researcher controlled the experiment through a Java application. The researcher registered the participants' answers and sound was recorded during the whole session for further analysis.

5.3. Participants

Forty-six participants took part in the experiment. First, it included 23 visually impaired people, 17 male and 6 female, with ages ranging from 22 to 62 ($M = 40.74$, $SD = 12.36$) years old. Nine participants had a congenital visual impairment (two of them) or their onset age preceded 18 years old (six fully blind; three partially sighted), while 14 had later onset ages (11 fully blind). All participants reported using screen readers daily (including partially sighted participants). To assess their self-reported experience with screen readers, we asked them to rate it using Likert-type items with a five-point scale (from 1 being not experienced to 5 being highly experienced). Most participants (16) rated themselves as *somewhat experienced* (rated 3) or experienced (rated 4) ($M = 3.39$, $SD = 1.12$), while four rated themselves as highly experienced. Only six participants browse websites less than once per week. All participants were recruited from a training center for visually impaired people. Later, we recruited 23 sighted people, 14 male and 9 female, with ages ranging from 23 to 62 ($M = 39.70$, $SD = 14.74$). All sighted participants browse websites on a daily basis. No participant reported having severe hearing impairments. Moreover, sighted and blind participants were balanced in terms of age ($U = 232{,}00$, $z = -0.477$, $p = 0.633$) and digit span (working memory assessment) scores ($t(43) = -0.58$, $p = 0.954$).

5.4. Design and Analysis

We used a 3x3x2 within-subjects design in which participants tested each combination of number of sources level (two, three, or four) and source separation (*same*, *small*, and *large*) two times. In each of these two repetitions, the frequency of the voice and the location of the relevant news item within the number of available sources was

completely randomized (but avoiding repetitions in the same voice condition). First, we analyzed the data for sighted and blind participants separately in order to assess their ability to cope with concurrent speech. Then, we performed a between-subjects comparison between both populations to assess their differences.

This design resulted in 366 trials in total for blind participants; 15 completed all conditions (18 trials), while the remaining 8 did not complete the condition with 4 voices (12 trials). Sighted participants performed 384 trials; 18 completed all conditions. We performed Shapiro-Wilkinson normality tests to observed values in all continuous dependent variables. Parametric tests were applied for normally distributed variables; nonparametric tests were applied otherwise. *Bonferroni* corrections were applied in post-hoc comparisons.

To analyze the participants' performance, we used the following metrics:

—*Identification of the Relevant Sentence*. After each trial, participants were asked to identify which sound source contained the relevant sentence. This metric shows the identification *success rate*, where a correct identification means that the participant described the exact location or voice of the relevant sentence. When we refer to *side identification* (in the four-voice condition) it means that participants were able to identify the correct side, but were not able to distinguish between the lateral and diagonal locations.

—*Completeness of participants' descriptions*. After listening to the concurrent news, participants were asked to report everything that they recalled about the relevant sentence. This measure reflects the percentage of relevant content that is reported by participants. To this effect, we considered all the parts of speech considered as content words (versus function words) in the snippets: verbs, nouns, adjectives, and adverbs [Winkler 2008]. Each snippet had between 14 and 20 content words ($M = 16.80$, $SD = 1.47$). To obtain a percentage for each description, we counted the content words that were reported, either using or not using the exact words. This assessment was performed by the authors, using a thesaurus and their contextual knowledge when needed (for instance, when replacing "*our country*" with "*Portugal"* or "*the Lisbon rivals"* with "*Benfica and Sporting*" (football/soccer teams)).

—*Answers Correctness*. After the participants' description, we asked them a specific (predetermined) question about the relevant sentence. This metric shows this question success rate. In the case that the participants had already mentioned the response in their descriptions, it is marked as correct without asking the question.

6. ANALYSIS OF THE PERCEPTION OF CONCURRENT SPEECH

Our goal was to understand how people cope with simultaneous information items in a *Scanning* task. In this evaluation, we analyze and compare blind and sighted people's ability to identify the item of interest and to focus their attention on it. Moreover, we compare voice conditions and the effect of working memory.

6.1. Identification of the Relevant Sentence

In order to identify the relevant sentence, participants needed to describe its location or voice characteristics. Overall, and including side identifications, out of 750 trials, participants were able to identify the correct speech source in 638 (85.1%).

6.1.1. Blind Participants. In 366 trials, blind participants correctly identified 289 news (79%), which increases to 301 if we include side identifications (82%). When asked about the source, participants identified the locations more times (298) than the talker's voice (16).

Figure 2 presents the success rate in the identification of the audio source. Table I—*B values*—shows that the voice conditions alone did not affect the identification of the

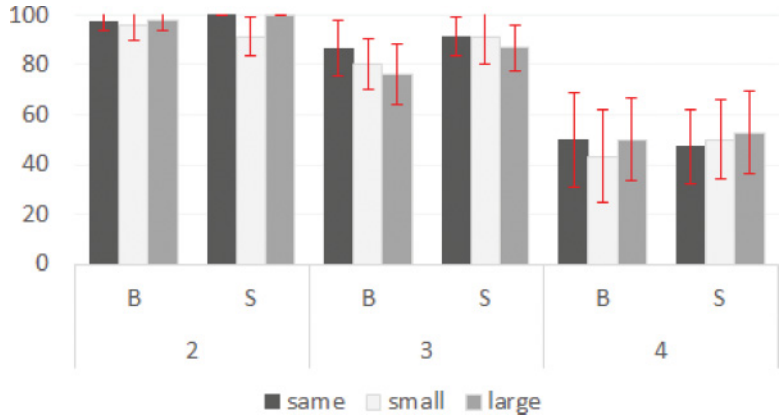

Fig. 2. The success rate (y-axis) for the identification of the relevant audio source, per number of sources and voice conditions, for both blind (B) and sighted (S) participants. Error bars denote 95% confidence intervals.

Table I. Statistical Analysis of the Comparison of Voice Characteristics Within Number of Voice Conditions: Results of the Identification of the Relevant Sentence for Both Blind (B) and Sighted (S) People

		Two Voices			Three Voices	Four Voices
		large vs small	large vs same	small vs same	large vs small vs same	large vs small vs same
B	χ^2		.667		3.045	.412
	Z					
	p		.717		.218	.814
S	χ^2		8.000 **(p=.018)**		.350	.047
	Z	-2.236	.000	-1.890		
	p	.025	1.000	.059	.839	.977

Note: The chi-square (χ^2) row corresponds to the comparison between the three related groups (large, small, and same) using a *Friedman test*. When this test shows significant differences, we run the post-hocs (*Wilcoxon signed rank tests*) and present the Z- and p-values. Otherwise, the p-value corresponds to the Friedman test analysis.

relevant source. In the case that we consider also side identifications with four voices, the rate of correct identifications increases (from 50%, 43.3% and 50%) to 63.3%, 53.3%, and 66.7% for the *same*, *small*, and *large* conditions (still with no significant differences). The absence of an effect of voice conditions contradicts the related work (e.g., Brungart and Simpson [2005a]), but may be explained by the length of our sentences (nearly 10s), which provides more time to explore the audio space.

In contrast, the number of sources has a significant effect on sound source identification for all voice characteristics (Table II—B values), mainly between two and four talkers. Moreover, results also differ when comparing between two and three sources, mostly in the *large* condition; however, the *same* and *small* conditions also suggest an effect of the number of talkers from two to three. The difference between three and four talkers is more evident for both the *same* and *small* conditions. Although the data shows a tendency in the *large* condition, there is not a significant difference if we also consider side identifications ($p = 0.132$). A deeper insight on this matter is provided by the participants' comments. They reported that, even though very high-pitched or deep voices are somehow annoying, they are easier to distinguish in the midst of several other voices.

These results show that blind people are able to identify the relevant source when there are two simultaneous talkers. In fact, 20 (out of 23) participants were able to identify the relevant source in all six trials with two voices. Moreover, eight participants were able to keep this record with three simultaneous talkers, while seven missed

Table II. Statistical Analysis of the Comparison of Number of Voices Within Voice Characteristics Conditions: Results of the Identification of the Relevant Sentence for Both Blind (B) and Sighted (S) People

		Same			Small			Large		
		2 vs 3	2 vs 4	3 vs 4	2 vs 3	2 vs 4	3 vs 4	2 vs 3	2 vs 4	3 vs 4
B	χ^2	18.216 (**p=.000**)			15.116 (**p=.001**)			15.500 (**p=.000**)		
	Z	-1.667	-3.035	-2.804	-2.804	-2.950	-2.810	-2.887	-3.071	-2.332
	p	.096	**0.002**	**.005**	.035	**.003**	**.005**	**.004**	**.002**	.020
S	χ^2	23.625 (**p=.000**)			17.637 (**p=.000**)			20.837 (**p=.000**)		
	Z	-2.000	-3.578	-3.087	-.577	-3.087	-3.095	-2.449	-3.314	-2.952
	p	.046	**.000**	**.002**	.564	**.002**	**.002**	**.014**	**.001**	**.003**

Note: The chi-square (χ^2) row corresponds to the comparison between the three related groups (2, 3, and 4 voices) using a *Friedman test*. As all such tests showed significant differences, we run the post-hocs (*Wilcoxon signed rank tests*) and present the Z- and p-values. Highlighted p-values correspond to significant differences with Bonferroni corrections.

only one trial. On the other hand, with four talkers, no participant identified the relevant source in the six trials. Specifically, one participant was able to identify the relevant source in five trials, which increases to three participants if we consider side identifications.

6.1.2. Sighted Participants. In the replicated experiment, sighted participants correctly identified the relevant sentence in 311 out of 384 trials (80.7%). This number increases to 337 (87.8%) if we consider side identifications. Similar to blind participants, they identified the source more often using its location (322 trials) than the talker's voice (14 trials).

Figure 2 shows the success rate for the correct identification of the audio source. Yet, if we consider side identifications with four voices, the rate of correct identifications increases (from 47.2%, 50%, and 52.8%) to 63.9%, 69.4%, and 75.0% for the *same*, *small*, and *large* conditions, respectively. Similar to blind participants' results, the voice variations alone did not affect the identification of the relevant source when considering three and four simultaneous talkers. However, there were significant differences in voice variations with two talkers (Table I—*S values*). The post-hoc analysis has shown a slight disadvantage for the *small* condition when compared to *large* and *same* conditions because, in these two conditions, all participants were able to identify the correct source in all trials.

As with blind participants, the number of sources has a significant effect on sound source identification in all voice characteristics conditions (Table II—*S values*), particularly between two and four talkers. The comparison between two and three sources revealed greater differences in the *large* condition and a smaller tendency in the *same* condition. However, there were no differences in the *small* condition, mainly due to its results with two voices, which were significantly worse than the other two voice conditions. The difference between three and four talkers is clear for all voice conditions.

These results show that sighted users are able to identify the relevant source when there are two simultaneous talkers. In this case, 19 (out of 23) participants were able to identify the relevant source in all six trials. Moreover, 10 participants were able to keep this record with three simultaneous talkers, while 11 missed only one trial. On the other hand, with four talkers, no participant identified the relevant source in the six trials (2–8, if we consider side identifications, were able to identify it in five trials).

6.1.3. Comparison. The separate analysis of these groups has shown some differences between conditions within each group, which ended up not being verified in the other (e.g., the *small* condition with two voices was worse than *large* and *same* only for sighted people). However, the comparison between blind and sighted participants' ability to identify the correct sound source has shown no significant differences between groups

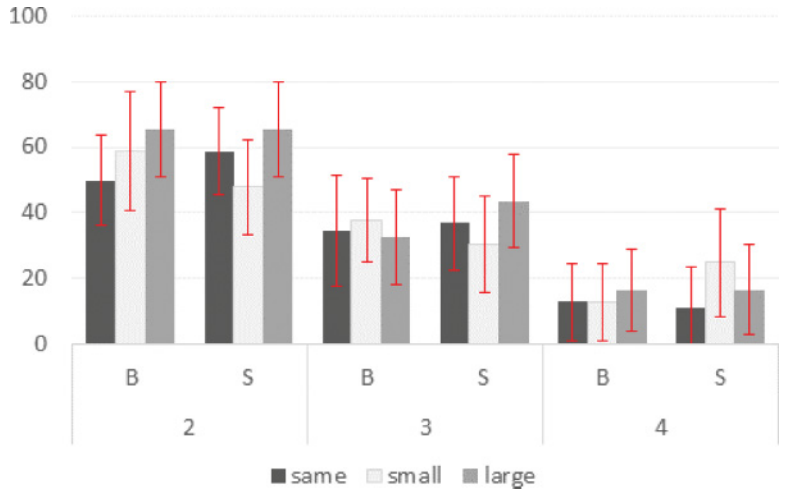

Fig. 3. The success rate (y axis) for correct answers to the predetermined question, per voice characteristics and number of sources, for both blind (B) and sighted (S) participants.

($p < 0.05$) in all nine conditions (Number of voices * Voice Condition, 3*3). The absence of differences between blind and sighted people suggests that both groups may have a similar ability to identify the correct source under the same conditions. In particular, this analysis and Figure 2 show that this is a trivial task with two concurrent talkers and, for most participants, with three.

6.2. Intelligibility and Report

To assess speech intelligibility, we relied on two methods: **completeness of participants' descriptions** and **answer correctness**. First, we analyze the participants' answer correctness, which provides some indications of whether the participants understood the sentence or not, but cannot be used to assess the comprehension of the entire sentence. It might be the case that participants missed, or forgot, the specific part that is important to answer the question. Therefore, we then analyze the completeness of participants' descriptions to assess the comprehension of the entire sentence (Figure 4 presents the mean values for each condition).

6.2.1. Blind Participants. An analysis of answer correctness supports that the voice characteristics did not affect the comprehension of the relevant sentence, as no significant differences were found within each number-of-talkers condition ($p > 0.05$ in all comparisons). However, there is a decreasing tendency of speech intelligibility when the number of talkers increases (Figure 3 and Table III—B values). Pairwise comparisons have shown differences between the use of two and four voices for all voice characteristics (but only a minor effect in the same voice condition). However, no significant differences were found between two and three voices, unless in the large-separations condition due to its better results with two voices. Specifically, in this condition, participants answered correctly 65% of the questions (76% if we consider incomplete answers). Although there seems to be a minor effect of the number of voices between three and four talkers in all voice conditions, no significant differences were found. This result is explained by the eight participants that did not complete the condition with four voices, thus are not included in the analysis. Furthermore, the differences in the number of talkers seem to be consistent among users as seven participants were able to answer

Table III. Statistical Analysis of the Comparison of Number of Voices Within Voice Characteristics Conditions: Results of the Correctness of Participants' Answers for both Blind (B) and Sighted (S) People

		Same			Small			Large		
		2 vs 3	2 vs 4	3 vs 4	2 vs 3	2 vs 4	3 vs 4	2 vs 3	2 vs 4	3 vs 4
B	χ^2	5.421 (p=.067)			10.042 **(p=.007)**			11.488 **(p=.003)**		
	Z	-1.226	-2.309	-1.994	-1.799	-2.863	-2.111	-2.696	-2.910	-1.732
	p	.220	.021	.046	.072	**.004**	.035	**.007**	**.004**	.083
S	χ^2	17.213 **(p=.000)**			6.565 **(p=.038)**			12.218 **(p=.002)**		
	Z	-2.134	-3.343	-2.387	-1.547	-2.496	-.879	-1.978	-3.106	-2.041
	p	.033	**.001**	**.017**	.122	**.013**	.380	.048	**.002**	.041

Note: The chi-square (χ^2) row corresponds to the comparison between the three related groups (2, 3, and 4 voices) using a *Friedman test*. We ran the post-hocs (*Wilcoxon signed rank tests*) and present the Z- and p-values. Highlighted p-values correspond to significant differences with Bonferroni corrections.

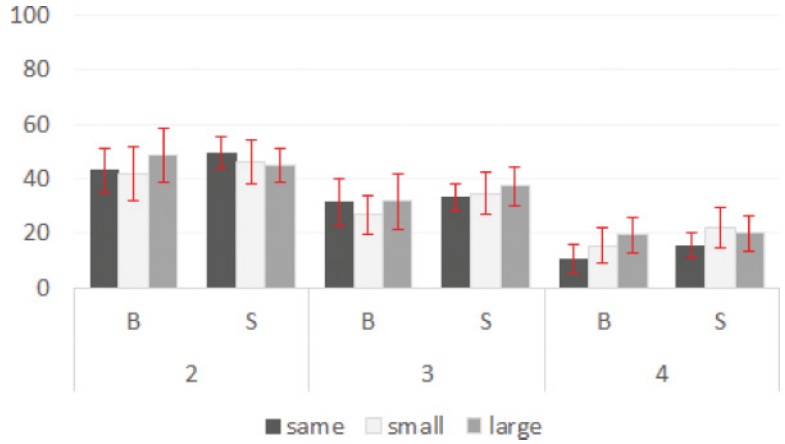

Fig. 4. Mean values (y axis) of user descriptions—percentage of content words reported—per number of sources and voice condition, for both blind (B) and sighted (S) participants. Error bars denote 95% confidence intervals.

correctly to at least five trials with two talkers, but none achieved that result with three or four talkers.

Regarding the completeness of the participants' descriptions (Figure 4), a Friedman test for each number-of-talkers condition showed no effect of voice condition in the sentence reports. Yet, the number of voices had a significant role in speech intelligibility in almost every comparison within voice characteristics ($p < 0.01$). The exceptions lie between three and four talkers, for both *large* and *small* conditions ($p = 0.041$ and $p = 0.026$, respectively), which also show a clear tendency dependent of the number of sources. Furthermore, there could be greater differences if all participants had performed the conditions with four voices, as those participants were the ones having more difficulties during the test. An average of sentence completeness within the six trials of each number of sources has shown that seven participants reported more than half of the sentence content when listening to two talkers, while three were able to keep that result with three talkers. If we consider an understanding of a quarter of the sentence, the numbers rise to 18 and 12 participants for two and three talkers, respectively.

6.2.2. Sighted Participants. For sighted participants, voice characteristics did not affect the comprehension of the relevant sentence ($p > 0.05$ in all comparisons within the number-of-talkers condition). Similar to the analysis of blind participants' results, there is a decreasing tendency in the answer correctness when the number of talkers

Table IV. Correlations Between Digit Span Scores and Description Completeness in all Conditions, for Both Blind and Sighted Participants

		2large	2same	2small	3large	3same	3small	4large	4same	4small
Blind	rho	.482	.349	.761	.633	.559	.447	-.111	.118	.339
	Sig	**.031**	.132	**.000**	**.003**	**.010**	**.048**	.719	.710	.257
	N	20	20	20	20	20	20	13	13	13
Sighted	rho	.173	.308	.263	.173	.213	.199	.341	.074	.030
	Sig	.430	.153	.226	.430	.329	.364	.166	.769	.094
	N	23	23	23	23	23	23	18	18	18

increase (Figure 3 and Table III—*S value*s). The *large* condition with two talkers also presents the best results (64.4%, reaching 80% if we consider incomplete answers). The comparisons between two and four voices have shown significant differences in all voice conditions, but the effect of the number of voices is smaller in the other pairwise comparisons. However, the differences between three and four voices could be greater if all participants had completed the condition with four voices. In general, three sighted participants were able to answer correctly to at least five trials with two talkers, but none with three or four talkers.

Regarding the completeness of the participants' descriptions (Figure 4), the number of sources had more impact in speech intelligibility than the voice conditions, which presented no significant differences. In contrast, speech intelligibility decreased as the number of sources increased in almost every comparison within voice conditions ($p < 0.01$). The exceptions lie between two and three talkers, for both *large* and *small* conditions ($p = 0.091$ and $p = 0.046$, respectively), which also suggest a minor decreasing tendency (although nonsignificant). An average of the six trials for each talker condition shows that nine participants reported more than half of the sentence content when listening to two talkers, while two were able to keep that result with three talkers. If we consider an understanding of a quarter of the sentence, the numbers rise to 22 and 21 participants for two and three talkers, respectively.

6.2.3. Comparison. Similar to the identification of the relevant source, the ability to understand and report the relevant content was not significantly different between groups. Figure 3 shows that blind and sighted participants' results are very similar in all conditions. They show minor differences on some conditions, but are somehow balanced between the user groups (Mann-Whitney U tests have shown no significant differences in all conditions). An analysis of the completeness of descriptions between sighted and blind participants has also shown no significant differences in all conditions ($p > 0.05$ in the nine comparisons). Again, these results suggest that both groups have a similar ability to understand the content of the relevant sentence. Moreover, their ability to report it depends on the number of simultaneous talkers. Figure 4 shows mean completeness scores between 49.6% and 42.1% with two voices, between 37.4% and 26.7% with three voices, and between 21.9% and 10.4% with four voices. In addition, 13 participants did not complete the condition with four voices, supporting a greater difficulty to identify and to understand the relevant sentence in comparison with the conditions with two and three voices.

6.3. Effect of Working Memory

Although being a cognitively demanding task, the previous results suggest that the use of simultaneous speech depends on the ratio of information that needs to be processed. Moreover, the person's cognitive abilities are also crucial to assess the usage of multiple talkers. Table IV presents the Spearman's rho correlation between Digit Span scores and sentence completeness for each condition, for both groups. Regarding blind participants, it shows medium to large correlations between digit span and all

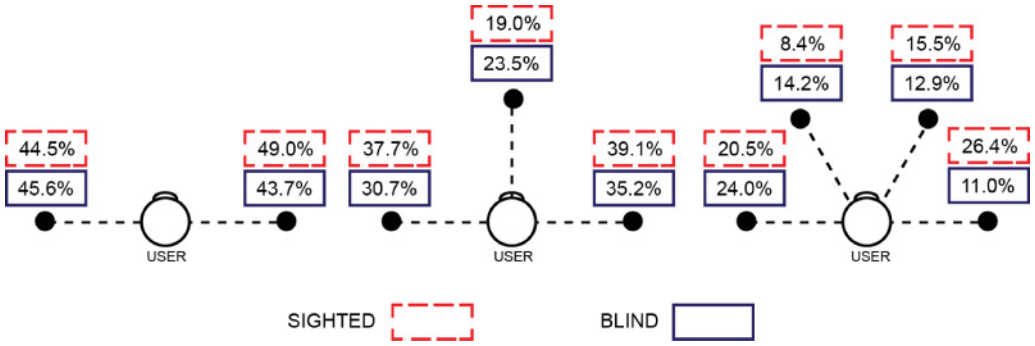

Fig. 5. Mean completeness of participants' descriptions for each location, depending on the number of talkers, for both sighted (dashed) and blind participants.

conditions with two and three talkers. In these conditions, participants have reported greater ease to identify and listen to the relevant sentence, but claimed having difficulties recalling the content that they have heard. These results and comments support that participants with lower working memory capacities (either related to attention and/or executive function) may have heard the sentences but, meanwhile, forgot the content.

In contrast, sighted participants' working memory scores did not correlate with sentence completeness for the none condition. This is a surprising result, but may be related to the fact that sighted participants' digit span scores ($M = 59.7$, $SD = 15.2$) were more homogeneous than those of blind participants ($M = 59.4$, $SD = 22.3$). Regardless of these differences, the need to recall what they had listened to was identified as the main challenge by most participants for two and three talkers. However, participants reported that it was easier to recall information about sentences in which they were genuinely interested.

6.4. Relevant Talker's Position and Voice

The talkers' spatial positions were fixed and established beforehand. Still, the relevant news could vary among them. Figure 5 shows the mean completeness of participants' descriptions for each location depending on the number of talkers, for both blind and sighted participants. Overall, there were no significant differences between the two groups ($p > 0.05$ in the respective comparisons) except for the lateral right position with four voices, for which the blind participants performed significantly worse than the sighted ones ($U = 49.5$, $Z = -2.36$, $p < 0.05$). However, we found no explanations for this difference.

The results with two talkers were very similar between the left and right locations, within both blind ($Z = -0.156$, $p = 0.875$) and sighted ($Z = -1.460$, $p = 0.144$) participants. However, with three talkers, the description completeness score was lower for the frontal position ($p < 0.05$ in the comparisons between lateral and frontal positions, except between left and frontal positions, but just for blind participants: $Z = -1.416$, $p = 0.157$). This difference is supported by 26 participants' comments that stated that it was more difficult to listen to and isolate the frontal voice. One blind participant stated:

> *When I want to focus my attention on a lateral source, I shut down the other ear and therefore I'm able to focus my attention on the ear of interest. However, for the frontal voice I cannot shut down any of the ears or I would listen to that lateral voice more clearly... so I really have to listen to the three sources, which augments the confusion.*

With four talkers, the lateral positions produced slightly better results (but not significantly different) and participants commented that the lateral audio sources were better perceived than the diagonal ones. These comments are consistent with previous work as the leftmost and rightmost locations are easier to identify due to the higher intensity arising from the proximity to the ears [Zurek 1993]. This difference occurs because, even though all voices have the same mean intensity, the sound intensity at each ear depends on the spatial position of the respective talker. Although the talkers' positions are equally distant to the user's head, the lateral ones are slightly closer to the user's ear (and directed straight to the ear). When trying to understand a lateral source, this very slight difference enabled the participants to focus their attention on the respective ear. In contrast, the diagonal talkers' voices are heard in both ears, but slightly lower at each ear than the lateral ones. Focusing the attention in both ears results in a greater confusion, as the four voices are heard simultaneously, while focusing on the respective ear diverts the attention to the lateral source instead of the diagonal one.

The variations of the relevant talker's voice ended up not having a noticeable effect. The only exception is with four talkers, for which the high-pitched voice held better results: 26.7% for blind and 26.8% for sighted participants in comparison to other voices; from 10.4% to 22.1% for the androgynous (for blind participants) and male (for sighted participants) voices, respectively. One blind participant noted that "*the high-pitched voice is irritating, but actually it is easier to distinguish it in the midst of several talkers*." This result is surprising as lower frequencies are less likely to be processed by the same auditory filter and therefore are easier to segregate than higher frequencies, which may be perceived as belonging to the same auditory stream [Darwin 1997]. However, some participants preferred listening to lower-pitched voices, as they found them more natural and easier to understand.

6.5. Neuroplasticity and Blindness Onset Age

Previous research about *neuroplasticity* and its effect on speech segregation led us to compare the blind participants' performance based on their blindness onset age. In a first attempt, described in Guerreiro and Gonçalves [2014], we analyzed the performance of participants with a congenital visual impairment or with an onset age that precedes 18 years, in comparison with participants with later onset ages. Mann-Whitney U tests revealed no significant differences in sentence completeness (neither source identification) between the two groups in all conditions ($p < 0.05$ in all conditions). However, the eight participants that were unable to complete the condition with four talkers were either late blind or had partial sight, which may suggest an effect of *neuroplasticity*.

Literature reviews are not consistent in the ages used to classify (and compare the results between) early- and late-blind people, nor there is an agreement on when people's *neuroplasticity* fades. In fact, there is evidence that this phenomenon also occurs in late-blind people's brains (e.g., Kujala et al. [1997] and Théoret et al. [2004]). We performed a correlation analysis between onset age and the completeness of blind participants' descriptions, which has shown negative correlations (earlier onset age, results in more complete descriptions), but only in three conditions (two voices with small separations, $rho = -0.543$, $p < .05$; three with large separations, $rho = -0.460$, $p < .05$; and four *same* voices, $rho = -0.511$, $p = .062$). Again, while these results may suggest a greater effect of *neuroplasticity* in participants with earlier-blindness onset, further research would be needed in order to better understand the effects of onset age and residual vision in this particular task.

Similarly, and although we could expect an advantage for blind people due to sensory compensation and/or *neuroplasticity*, results have shown no significant differences in

Table V. Median and Interquartile Range (IQR) Values of Questionnaire Ratings

	Blind		Sighted	
	Median	IQR	Median	IQR
Able to understand with two voices	5	0	5	0
Comfortable listening to two voices	4	1	4	1
Able to understand with three voices	3	1	3	1
Comfortable listening to three voices	3	0	3	0
Able to understand with four voices	2	0	2	0
Comfortable listening to four voices	1	.75	1	1
Effect of location	3	1	4	1
Effect of voice	4	0	4	1
Effect of difference among voices	3	1	5	0

comparison to sighted participants regarding the ability to both identify and understand the relevant content. These results may be explained by the small number of early fully blind participants in our experiment (six participants with onset ages prior to 18 years old), but may as well be related with the information scanning task and our experimental setting. In particular, the use of different spatial locations enhances speech segregation, which may have reduced the impact of *neuroplasticity*. Moreover, previous experiments about speech segregation and the effect of *neuroplasticity* use smaller utterances, while the news snippets enable people to spend more time trying to identify and understand the relevant sentence.

6.6. Subjective Feedback

After completing all trials, we asked participants to rate two sentences using Likert-type items with a five-point scale (1 being Strongly Disagree to 5 being Strongly Agree) for each number of sources (2, 3, and 4). They had to rate the ability to understand the relevant sentence ("*I am able to understand the content of the relevant sentence with X talkers*") and how comfortable they were ("*I am comfortable listening to X news simultaneously*"). Moreover, we asked the participants to rate the influence of the spatial location of the relevant sentence (*the spatial location of the relevant sentence affected its comprehension*), the influence of the voice of the relevant sentence (*the voice of the relevant sentence affected its comprehension*) and the effect of the difference between simultaneous voices (*the voice differences among the simultaneous sentences have an effect on the relevant sentence comprehension*).

Both the self-reported ability to understand the content and the comfort levels support the results reported in the previous sections (Table V). Both user groups were able to understand the relevant content and felt comfortable with two voices, but the ratings with three voices do not show a clear tendency either for or against its use. In line with the experiment's objective results, some participants were able to maintain a similar ability to understand the relevant content, but others considerably decreased their real and self-reported performance. Participants' ratings of four voices were consistent with their comments, as most reported that four voices are excessive as it is very difficult to identify the relevant source and, after a successful identification, segregate it among the others to understand its content.

The difficulty in dealing with four voices was many times accompanied by comments referring to the need to practice more often in order to achieve better results. One blind participant reported that "*the same happens with the use of faster speech rates. We start with the default rate, but as we become experts, we increase the speech rate. Probably the same would happen with the number of voices*." This ability to improve the segregation of sound with practice is supported by previous research and would be interesting to study in future work [Alain 2007].

Again, there is an absence of differences between user groups in all the aforementioned self-reported ratings (p-values between 0.731 and 0.863). Table V shows that both groups (no significant difference: $p = 0.128$) had mixed opinions regarding an effect of the voice of the relevant sentence (blind: $M = 3.1$, $SD = 1.4$; sighted: $M = 3.7$, $SD = 1.3$). However, the ones that argued in favor of voice importance relied on two contrary arguments. While some users preferred listening to deep voices, as they sounded softer and more natural, others claimed that high-pitched voices were "*easier to distinguish in the midst of several talkers*" even though they were irritating. One blind participant said that he "*would prefer listening to a normal voice when listening to default news, but use a high-pitched voice if there is a need to highlight a particular one*."

The influence of the voice location was corroborated by several participants, but more by sighted ($M = 4.1$, $SD = 1.2$) rather than blind ($M = 3$, $SD = 1.6$) participants ($p < .05$). Although most participants had no preference between the right and left sides, 10 blind and 16 sighted participants reported more difficulty understanding the central voices (both with three and four voices).

Regarding the effect of difference/similarity of the concurrent voices, it was more evident ($p < 0.05$) for sighted ($M = 4.0$, $SD = 1.4$) than for blind people ($M = 3.3$, $SD = 1.3$). However, most participants reported a preference for different voices. The main reason was that it was "*easier to separate the different news and isolate the relevant one*." This was "*more important as the number of voices increase. With two voices, it is not that relevant, but with three and mainly four, it gets very difficult when the voices are similar*." In contrast, the participants that found no advantages in the use of different voices claimed that the different locations are much more important and suffice to enable the discrimination between sound sources.

All participants referred to the condition with four voices as the most difficult. In this condition, it was harder to identify and keep track of the relevant sentence. The ability to recall what they have heard was a challenge transverse to all conditions, but it decreased as the number of sources increased. Furthermore, some participants noted that concurrent sentences with similar subjects hinder their ability to identify the relevant one and that the subject of the news influences their ability to recall and report what they have heard.

7. DISCUSSION

The analysis of this experiment enabled us to address our research goals. In this section, we present the main findings that may guide the design of future interfaces.

Blind and sighted people appear to be performing in the same way. The comparison between blind and sighted participants' performance has shown no significant differences between them. This includes the identification of the relevant source as well as the ability to understand its content. Moreover, blind and sighted people's self-reported ability to understand the relevant sentence is very similar and complies with their actual performance. The need to consume digital information in several contexts, including without visual feedback, approximates the design space of future solutions for both sighted and blind populations. The lack of significant differences between them in this experiment provides two important implications for future work. First, it shows that sighted people are also able to cope with simultaneous speech, encouraging new solutions based on this approach. Second, it diminishes the need for adaptations in future interfaces that may be used by both blind and sighted persons.

Two and three concurrent talkers enable identification. Results show that both blind and sighted people are able to identify the relevant snippet when listening to two simultaneous talkers (Figure 2). In fact, 39 of 46 participants had a 100% success rate. Despite the fact that the identification rate decreases with three talkers, some users are still able to identify the relevant snippet. In particular, 36 participants identified

the relevant snippet in at least five of the six trials. The use of four voices considerably affected the identification rate. Yet, an average result near 50% suggests that it may be suitable in scenarios that admit worst performances. These results support the usage of concurrent speech (two to three talkers) in tasks that require the selection of an item of interest. Articles from news sites, search results, or posts in *Social Networking Sites* (SNS) are good examples, as users may scan through the content to select the ones that deserve further attention.

Identification through location. Location was by far the preferred attribute to describe the relevant sentence. This finding can be leveraged for interaction purposes, for instance, to select or to increase one source's volume. It was previously done with head movements [Crispien et al. 1996], but can also be applied to the usage of gestures in touch screens, specific keys in keyboards, among other approaches.

Use two or three talkers, depending on intelligibility demands. The report task is demanding by itself and is aggravated by the presence of another talker, since intelligibility is clearly influenced by the number of simultaneous talkers. The decision to use two or three talkers should take into consideration the intelligibility demands. The use of three talkers may be used when one needs to obtain solely the gist of the sentence. To cite one example, one blind participant suggested the use of "*three talkers in search engines, as the relevant result is usually among the first three*." In cases in which the intelligibility demands are greater, the option should go to two talkers. Another blind participant stated: "*I usually listen to two news channels simultaneously (in the television and computer) and I am able to focus the attention on one of them when I identify relevant content*." These results show that not only concurrent speech can be used to identify the relevant content, but also to understand its content.

Working memory plays an important role. Blind participants' Digit Span scores are highly correlated with the amount of information reported after a trial. Although there were no such correlations for sighted participants, both user groups pointed out the difficulty of recalling what they had just heard. These scores should be used to determine the tasks that support the use of multiple sources. People with lower digit span scores, and therefore a greater difficulty to recall what they heard, can take advantage of simultaneous speech only in tasks for which the intelligibility demands are lower. In contrast, people with higher scores may perform (more) demanding tasks with both two and three talkers. Moreover, our results showed that identification and intelligibility can be attained with two or three sources, when this was done as the user's main task. The high correlations with digit span scores suggest that this could be harder to accomplish in more demanding settings (e.g., a blind person walking in the street).

Voice differences are not crucial, but preferred. Apart from very specific situations, voice differences did not provide an advantage either for speech identification or intelligibility. Although the related work shows that using different frequencies enhances speech segregation, it also shows that each attribute provides a greater effect when varied alone [Brungart and Simpson 2005a]. In our work, the use of different spatial locations seems to suffice for the task addressed. Nevertheless, the participants felt more confident when the voices were different. In particular, 33 participants preferred listening to different voices, while only two preferred the same voice. One blind participant stated, "*It is better to use different voices, because it requires less effort to follow the same sentence. This is particularly useful when listening to three or four talkers*."

8. USAGE SCENARIOS

The overall results open an avenue for different approaches and interfaces that target a much wider audience, instead of solutions to exclusively enhance blind people's scanning for relevant content. In this case, both groups' ability to process concurrent sentences eases and reduces the effort of *Design for All* (or *Universal Design*)

[Stephanidis 2001] approaches. Interface designers can and should consider proactive strategies in order to address different ranges of human abilities/disabilities.

The abundance of digital information, as well as people's constant need to consume it, result in ubiquitous consumption either in personal computers or smartphones. Using auditory feedback to convey digital information to sighted people is not new. This is supported by the rising popularity of podcasts and audio books [Peoples and Tilley 2011] and may be leveraged in scenarios in which it may be difficult, unpleasant or less appropriate to rely on visual feedback (e.g., hiking, running, commuting, and so on). For instance, *Capti-Narrator* [Borodin et al. 2014] enables users to listen to content instead of reading it, by creating a playlist with news, blog posts, documents, or e-books. Although it is very unlikely that sighted users would make use of concurrent auditory feedback for ordinary use, they could take advantage of concurrent speech to increase the efficiency of information consumption in specialized scenarios.

The growing auditory information consumption by both blind and sighted people stimulates the emergence of techniques to improve and accelerate this process. We present some scenarios in which concurrent speech may be used to improve the efficiency of processing auditory information.

8.1. Scanning for Relevant Information Items

Listening carefully to documents, news, or blog posts requires a person's complete attention; the use of concurrent speech would most likely hamper the full comprehension of the text. However, its use on a preliminary selection task, in which users assess the worthiness of an information item, does not require understanding the entire content. Among several podcasts, *Capti-Narrator* items [Borodin et al. 2014], search results, posts, or news lies a decision of which are relevant and deserve further attention. This preliminary assessment may be currently done visually by sighted users or via the sequential audio of screen readers by blind users. Yet, the use of concurrent speech could eliminate the need for visual feedback for the former in specific scenarios while trying to accelerate this task for the latter.

The web accommodates a multitude of platforms that comprise numerous summarized, or already small *per se*, information items that (try to) provide the gist of the content to help the user decide if they need further attention. These platforms may contain titles or small descriptions/snippets and include, for example, search engines, SNS such as *Facebook* and *Twitter*, blogs, RSS feeds, news sites, and e-mail platforms. During our experiment, one blind participant suggested the use of (three) concurrent talkers in search engines to improve efficiency, since the relevant result is usually among the first ones. Another blind user mentioned that he already listens to simultaneous news from television and his computer, in order to avoid listening to uninteresting content and focus on the ones of interest.

Scanning for relevant content may be seen as a task for immediate information consumption, but also to mark interesting content to process later. In the first, users may listen to simultaneous items and select the one of interest to listen to at that time. In the latter, participants may be presented with several information items, whereas they may mark the ones they would like to listen to in the future. A parallel may be found, for example, in adding items to a podcast or Capti-Narrator playlist. Furthermore, it may be found in mainstream platforms, where websites or posts in SNS are bookmarked or favorited and e-mails are starred.

8.2. Scanning for Specific Information

Websites and documents with a lot of text may make the search for specific content difficult when the user struggles to find a particular word or phrase to search for. The use of concurrent speech may be an alternative to higher speech rates or paragraph/heading

navigation. Instead, different paragraphs or sections could be read simultaneously, helping the user to find the actual content of interest, or at least its paragraph or section. In such scenarios, interaction mechanisms are very important to manage the simultaneous audio. Better interaction tools would help the user to easily discard material or save for further reading.

8.3. Notifications using a Secondary Audio Channel

The aforementioned scenarios that may use concurrent speech focus on scanning tasks that occur occasionally. The use of concurrent speech as the main mode to consume auditory information would be highly cognitively demanding and therefore somehow unrealistic. While the main exploration mode may still rely on a unique speech source, notifications do not need to be confined to uninformative alert sounds. While listening to a document, blog post, or the daily news, chat or e-mail notifications could include the subject or the sender's name, instead of a beep sound that may induce the user to interrupt one's current task. Another example is the one of SNS, in which a user may be listening to the news feed and simultaneously being informed about new notifications or chat alerts. In the particular case of blind people, a proper use of Accessible Rich Internet Applications specification (WAI-ARIA) could leverage a secondary speech channel to help deal with dynamic content, website refreshes and advanced interface functions developed with *Ajax*, *HTML5*, or *JavaScript*.

8.4. Multitouch to Multisound

Touchscreens have become pervasive, mostly due to the success of smartphones. Recently, these device dimensions have been increasing, tablet devices are growing in popularity, and large interactive tabletops are gaining space. As these devices are inherently visual and lack the tactile feedback of keypad phones and laptop keyboards, blind people rely largely on mainstream accessibility features such as Apple's *VoiceOver* or Android's *Talkback*, which allow them to explore and control the device by providing audio feedback for touch actions. While touchscreens support multitouch interaction, current screen readers are limited to a single, sequential auditory channel. However, the growing dimensions of touchscreen surfaces enables two-handed interaction and exploration of the screen. Blind people's demonstrated ability to understand simultaneous speech sources encourages new interaction methods for nonvisual access on touchscreens that can leverage their dimensions and multitouch capabilities. For instance, we have developed *SpatialTouch* [Guerreiro et al. 2015b], an input system for tablet devices that supports two-handed interaction through multitouch exploration and spatial, simultaneous audio feedback. It tries to mimic and leverage previous experience with physical QWERTY keyboards by enabling the user to rest one's fingers on the home keys to orient one's position within the keyboard and locate desired keys. In a different context, we supported two-handed exploration of large touch surfaces [Guerreiro et al. 2015a] using a similar approach. Each hand was mapped to a specific voice (male or female) and assigned a specific location (left or right ear). Similar approaches may rely on concurrent speech to leverage the multitouch capabilities of these devices.

9. CONCLUSIONS

Blind people rely mostly on the auditory channel to access digital information, but sighted people's auditory consumption of information is also increasing. In this article, we aimed to study the human ability to leverage the *Cocktail Party Effect* to scan for digital content using concurrent speech. The comparison between blind and sighted participants suggested that both user groups may take advantage of concurrent speech in scanning tasks. Moreover, it pointed toward the design of interfaces that target both

user groups without major adaptations, as their results were very similar. In line with previous research, in this experiment we found that both identification of the relevant source and speech intelligibility decrease with an increasing number of concurrent talkers. Our results show that identification of the relevant source is a straightforward task when listening to two talkers and, for most participants, it was also easy to identify with three. Moreover, both two and three simultaneous sources may be used to understand the relevant source's content depending on speech intelligibility demands and user characteristics (working memory). Unlike the related work, which deals with much smaller speech signals, differences in voice characteristics did not provide a greater effect in either speech identification or intelligibility. However, participants preferred, and felt more confident with, the use of concurrent talkers with different voices.

Similar to the use of faster speech rates, simultaneous speech segregation can benefit from practice [Alain 2007]. This experiment comprised a unique session lasting approximately 45min. We believe that the frequent use of simultaneous speech will improve both speech identification and intelligibility scores. Moreover, these were one-shot trials, wherein participants were not able to return to the relevant content. In realistic settings, interaction solutions should provide easy access to recently explored content. From this experiment, we have learned that the sound source location is the best mechanism to identify and therefore interact with such a concurrent sound source system.

A limitation of this experiment regards the number of relevant sources, which are restricted to one. Furthermore, we used news snippets because most of our participants explore news sites in a daily basis. Since the relevant snippets were not chosen by the participants in order to conduct a controlled experiment, the actual participants' interests and knowledge may have influenced their ability to report the snippets. This limitation was reported by the participants, which noted that the subject of the news influence their ability to recall and report what they have heard. However, in realist scenarios users would be focusing their attention on their favorite subjects, and therefore would be able to recall more information. In addition, in future interfaces if we prime the user with pre-defined subject locations, we can take advantage of *apriori* expectations [Brungart and Simpson 2005a]. For instance, one could listen to sports content always on the right side, while listening to economics on the left.

Some results suggested an effect of blindness onset age in the performance of blind participants, which may be related to *neuroplasticity*. For example, all visually impaired participants that did not complete the condition with four voices were either partially sighted or became blind after turning eighteen years old. If they had performed this last condition, results might had suggested a greater effect of onset age and *neuroplasticity* in the conditions with four voices. However, the comparison presented in Guerreiro and Gonçalves [2014] and the absence of correlations between onset age and descriptions completeness in most conditions, support that participants appeared to behave independently of their onset age. For that reason, we avoided subdividing the blind participants' group based on their onset age and/or residual vision in the remaining analysis, as it would create considerably smaller groups hampering a statistical analysis. We believe that further research is needed in order to understand the effects of onset age and residual vision in this particular task.

In Section 7 we attempted to lay out a set of guidelines to the use of concurrent speech in fast-exploration tasks. Moreover, we proposed four scenarios that may benefit from the use of concurrent speech sources. After submitting this paper and during the reviewing process, we compared the use of concurrent speech against the use of faster speech rates by blind people [Guerreiro and Gonçalves 2015]. Moreover, we combined these two approaches by gradually increasing the speech rate with one, two and three

voices. Results showed that concurrent voices with speech rates slightly faster than the default rate, enable a significantly faster scanning for relevant content, while maintaining its comprehension. In contrast, to keep-up with concurrent speech timings, a single voice requires larger speech rate increments, which cause a considerable loss in performance. In future work, we intend to explore interaction mechanisms that will enable users to cope with the additional demands of multiple feedback sources.

ACKNOWLEDGMENTS

The authors would like to thank Fundação Raquel e Martin Sain, Carlos Bastardo and all participants in this experiment. We also thank both Voice Interaction and INESC ID's Spoken Language Systems Laboratory. This work was supported by national funds through the Portuguese Foundation for Science and Technology, under grants UID/CEC/50021/2013 and INCENTIVO/EEI/LA0021/2014.

REFERENCES

Faisal Ahmed, Yevgen Borodin, Yury Puzis, and I. V. Ramakrishnan. 2012a. Why read if you can skim: Towards enabling faster screen reading. In *International Cross-Disciplinary Conference on Web Accessibility*.

F. Ahmed, Y. Borodin, A. Soviak, M. Islam, I. V. Ramakrishnan, and T. Hedgpeth. 2012b. Accessible skimming: Faster screen reading of web pages. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. http://dl.acm.org/citation.cfm?id=2380164

Claude Alain. 2007. Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research* 229, 1, 225–236.

Paul M. Aoki, Matthew Romaine, Margaret H. Szymanski, James D. Thornton, Daniel Wilson, and Allison Woodruff. 2003. The Mad Hatter's cocktail party: A social mobile audio space supporting multiple simultaneous conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 425–432.

Barry Arons. 1997. SpeechSkimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer– Human Interaction (TOCHI)—Special Issue on Speech as Data* 4, 1, 3–38.

C. Asakawa, H. Takagi, S. Ino, and T. Ifukube. 2003. Maximum listening speeds for the blind. *Proceedings of ICAD* (*ICAD'03*), 276–279. Retrieved December 22, 2015 from http://www.icad.org/Proceedings/2003/AsakawaTakagi2003.pdf.

Norman Backhaus and Rico Tuor. 2015. OLwA: Use of literature. Retrieved December 22, 2015 from http://www.geo.uzh.ch/microsite/olwa/olwa/en/html/unit3_kap33.html. (2015).

M. F. Bear, B. W. Connors, and M. A. Paradiso. 2006. *Neuroscience: Exploring the Brain*. Lippincott Williams & Wilkins, Philadelphia, PA.

J. P. Bigham, A. C. Cavender, J. T. Brudvik, J. O. Wobbrock, and R. E. Ladner. 2007. WebinSitu: A comparative analysis of blind and sighted browsing behavior. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*. http://dl.acm.org/citation.cfm?id=1296854.

Paul Boersma and David Weenink. 2014. Praat: doing phonetics by computer. Retrieved December 22, 2015 from http://www.fon.hum.uva.nl/praat/. (2014). Accessed in: 06-2015.

Y. Borodin and J. P. Bigham. 2010. More than meets the eye: A survey of screen-reader browsing strategies. *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A'10).* http://dl.acm.org/citation.cfm?id=1806005

Yevgen Borodin, Yuri Puzis, Andrii Soviak, James Bouker, Bo Feng, Richard Sicoli, Andrii Melnyk, Valentyn Melnyk, Vikas Ashok, Glenn Dausch, and I. V. Ramakrishnan. 2014. Listen to everything you want to read with Capti Narrator. In *Proceedings of the 11th Web for All Conference(W4A'14)*. ACM, New York, NY, Article 33, 2 pages. DOI: http://dx.doi.org/10.1145/2596695.2596728

A. S. Bregman. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Bradford Books, MIT Press, Cambridge, MA.

S. A. Brewster, P. C. Wright, and A. D. N. Edwards. 1993. An evaluation of earcons for use in auditory human–computer interfaces. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. 222–227. http://dl.acm.org/citation.cfm?id=169179

Adelbert W. Bronkhorst. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica* 86, 1, 117–128.

Douglas S. Brungart and Brian D. Simpson. 2005a. Improving multitalker speech communication with advanced audio displays. Air Force Research Lab Wright-Patterson AFB OH.

Douglas S. Brungart and Brian D. Simpson. 2005b. Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception* 2, 4, 430–436. DOI: http://dx.doi.org/10.1145/1101530.1101538

H. Burton. 2003. Visual cortex activity in early and late blind people. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 23, 10, 4005–11. http://www.ncbi.nlm.nih.gov/pubmed/12764085

Denis Byrne and William Noble. 1998. Optimizing sound localization with hearing aids. *Trends in Amplification* 3, 2, 51–73. DOI: http://dx.doi.org/10.1177/108471389800300202

Sharon Cameron and Harvey Dillon. 2008. The listening in spatialized noise-sentences test (LISN-S): Comparison to the prototype LISN and results from children with either a suspected (central) auditory processing disorder or a confirmed language disorder. *Journal of the American Academy of Audiology* 19, 5, 377–391. DOI: http://dx.doi.org/10.3766/jaaa.19.5.2

Robert P. Carlyon. 2004. How the brain separates sounds. *Trends in Cognitive Sciences* 8, 10, 465–71. DOI: http://dx.doi.org/10.1016/j.tics.2004.08.008

E. C. Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* 25, 5, 975–979.

Kai Crispien, Klaus Fellbaum, Anthony Savidis, and Constantine Stephanidis. 1996. A 3D-auditory environment for hierarchical navigation in non-visual interaction. *Proceedings of ICAD.*

C. J. Darwin. 1997. Auditory grouping. *Trends in Cognitive Sciences* 1, 9, 327–333. DOI: http://dx.doi.org/10.1016/S1364-6613(97)01097-8

Christopher J. Darwin, Douglas S. Brungart, and Brian D. Simpson. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America* 114, 5, 2913. DOI: http://dx.doi.org/10.1121/1.1616924

Donald Dirks. 1964. Perception of dichotic and monaural verbal material and cerebral dominance for speech. *Acta Oto-Laryngologica* 58, 1–6, 73–80. DOI: http://dx.doi.org/10.3109/00016486409121363

R. Drullman and A. W. Bronkhorst. 2000. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America* 107, 2224. http://link.aip.org/link/jasman/v107/i4/p2224/s1.

W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress. 1957. Localization of high-frequency tones. *The Journal of the Acoustical Society of America* 29, 9, 988–991. DOI: http://dx.doi.org/10.1121/1.1909356

Prathik Gadde and Davide Bolchini. 2014. From screen reading to aural glancing: Towards instant access to key page sections. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, New York, NY, 67–74. DOI: http://dx.doi.org/10.1145/2661334.2661363

Bill Gardner and Keith Martin. 2000. HRTF Measurements of a KEMAR Dummy-Head Microphone. Retrieved December 22, 2015 from http://sound.media.mit.edu/resources/KEMAR.html.

William Gaver. 1986. Auditory icons: Using sound in computer interfaces. *Human–Computer Interaction* 2, 2, 167–177. DOI: http://dx.doi.org/10.1207/s15327051hci0202_3

Stuart Goose and Carsten Möller. 1999. A 3D audio only interactive web browser: Using spatialization to convey hypermedia document structure. In *Proceedings of the 7th ACM International Conference on Multimedia (Part 1)*. ACM, New York, NY, 363–371.

João Guerreiro and Daniel Gonçalves. 2014. Text-to-speeches: Evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, New York, NY, 169–176.

João Guerreiro and Daniel Gonçalves. 2015. 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*. ACM, New York, NY.

João Guerreiro, André Rodrigues, Kyle Montague, Tiago Guerreiro, Hugo Nicolau, and Daniel Gonçalves. 2015b. TabLETS get physical: Non-visual text entry on tablet devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY.

Tiago Guerreiro, Kyle Montague, João Guerreiro, Rafael Nunes, Hugo Nicolau, and Daniel Gonçalves. 2015a. Blind people interacting with large touch surfaces: Strategies for one-handed and two-handed exploration. In *Proceedings of the 2015 ACM International Conference on Interactive Tabletops and Surfaces*. ACM, New York, NY.

Simon Harper. 2012. Deep accessibility: Adapting interfaces to suit our senses. *Invited Talk-Technical Superior Institute, LaSIGE, Lisbon, Portugal.*

Simon Harper and Neha Patel. 2005. Gist summaries for visually impaired surfers. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*. 90–97.

Kenneth Hugdahl, Maria Ek, Fiia Takio, Taija Rintee, Jyrki Tuomainen, Christian Haarala, and Heikki Hämäläinen. 2004. Blind individuals show enhanced perceptual and attentional sensitivity for

identification of speech sounds. *Brain Research. Cognitive Brain Research* 19, 1, 28–32. DOI:http://dx.doi.org/10.1016/j.cogbrainres.2003.10.015

Gregory Kramer, Bruce Walker, Terri Bonebright, Perry Cook, John H. Flowers, Nadine Miner, and John Neuhoff. 2010. Sonification Report: Status of the Field and Research Agenda. Faculty Publications, Department of Psychology, Paper 444.

T. Kujala, K. Alho, M. Huotilainen, R. J. Ilmoniemi, A. Lehtokoski, A. Leinonen, T. Rinne, O. C. Salonen, J. Sinkkonen, C.-G. Standertskjöuld-Nordenstam, and others. 1997. Electrophysiological evidence for cross-modal plasticity in humans with early- and late-onset blindness. *Psychophysiology* 34, 2, 213–216. DOI:http://dx.doi.org/10.1111/j.1469-8986.1997.tb02134.x

L2F 2015. INESC-ID's Spoken Language Systems Laboratory. Retrieved December 22, 2015 from http://www.l2f.inesc-id.pt/.

Paul Lamb. 2015. Paul Lamb's 3D Sound System. Retrieved December 22, 2015 from http://www.paulscode.com/forum/index.php?topic=4.0. (2015). Accessed in: 06-2015.

Kevin A. Li, Patrick Baudisch, and Ken Hinckley. 2008. BlindSight: Eyes-free access to mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1389–1398.

Ruth Y. Litovsky, Aaron Parkinson, and Jennifer Arcaroli. 2009. Spatial hearing and speech intelligibility in bilateral cochlear implant users. *Ear and Hearing* 30, 4, 419. DOI:http://dx.doi.org/10.1097/aud.0b013e3181a165be

Darren Lunn, Simon Harper, and Sean Bechhofer. 2011. Identifying behavioral strategies of visually impaired users to improve access to web content. *ACM Transactions on Accessible Computing* 3, 4, 13.

LWJGL 2.0. 2015. LightWeight Java Game Library. Retrieved December 22, 2015 from http://legacy.lwjgl.org/.

Jalal Mahmud, Yevgen Borodin, Dipanjan Das, and I. V. Ramakrishnan. 2007. Combating information overload in non-visual web access using context. *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI'07),* 341. DOI:http://dx.doi.org/10.1145/1216295.1216362

B. C. J. Moore. 1995. *Hearing*. Handbook of perception and cognition. Academic Press.

B. C. J. Moore and H. E. Gockel. 2011. Resolvability of components in complex tones and implications for theories of pitch perception. *Hearing Research* 276, 1–2, 88–97. http://www.sciencedirect.com/science/article/pii/S0378595511000062

W. Niemeyer and I. Starlinger. 1981. Do the blind hear better? Investigations on auditory processing in congenital or early acquired blindness II. Central functions. *International Journal of Audiology* 20, 6, 510–515. http://informahealthcare.com/doi/abs/10.3109/00206098109072719

OpenAL Soft 2014. OpenAL Soft 1.15.1. Retrieved December 22, 2015 from http://kcat.strangesoft.net/openal.html.

K. Papadopoulos. 2010. 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. http://informahealthcare.com/doi/abs/10.3109/07434618.2010.522200

Peter Parente. 2006. Clique: A conversant, task-based audio display for GUI applications. *SIGACCESS Accessibility and Computing* 84, 34–37. DOI:http://dx.doi.org/10.1145/1127564.1127571

Sérgio Paulo, Luís C. Oliveira, Carlos Mendes, Luís Figueira, Renato Cassaca, Céu Viana, and Helena Moniz. 2008. DIXI–A generic text-to-speech system for european portuguese. *Computational Processing of the Portuguese Language* 91–100.

Brock Peoples and Carol Tilley. 2011. Podcasts as an emerging information resource. *College & Undergraduate Libraries* 18, 1, 44–57. DOI:http://dx.doi.org/10.1080/10691316.2010.550529

Daisuke Sato, Shaojian Zhu, M. Kobayashi, Hironobu Takagi, and Chieko Asakawa. 2011. Sasayaki: Voice augmented web browsing experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2769–2778.

C. Schmandt and A. Mullins. 1995. AudioStreamer: Exploiting simultaneity for listening. *Conference Companion on Human Factors in Computing Systems* 218–219. http://dl.acm.org/citation.cfm?id=223355.223533

Andrew Schwartz, Josh H. McDermott, and Barbara Shinn-Cunningham. 2012. Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America* 132, 1, 357–368. DOI:http://dx.doi.org/10.1121/1.4718637

Donald Shankweiler and Michael Studdert-Kennedy. 1967. Identification of consonants and vowels presented to left and right ears. *The Quarterly Journal of Experimental Psychology* 19, 1, 59–63. DOI:http://dx.doi.org/10.1080/14640746708400069

B. Shinn-Cunningham and A. Ihlefeld. 2004. Selective and divided attention: Extracting information from simultaneous sound sources. (2004).

Jaka Sodnik and Sašo Tomažič. 2009. Spatial speaker : 3D Java text-to-speech converter. In *Proceedings of the World Congress on Engineering and Computer Science*, Vol. II.

Amanda Stent, Park Ave, Florham Park, Ann Syrdal, and Taniya Mishra. 2011. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. 211–218.

Constantine Stephanidis. 2001. User interfaces for all: New perspectives into human– computer interaction. *User Interfaces for All-Concepts, Methods, and Tools* 1, 3–17.

Hironobu Takagi, Shin Saito, Kentarou Fukuda, and Chieko Asakawa. 2007. Analysis of navigability of web applications for improving blind usability. *ACM Transactions on Computer– Human Interaction* 14, 3 13–es. DOI:http://dx.doi.org/10.1145/1279700.1279703

Thomas M. Talavage, Martin I. Sereno, Jennifer R. Melcher, Patrick J. Ledden, Bruce R. Rosen, and Anders M. Dale. 2004. Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. *Journal of Neurophysiology* 91, 3, 1282–1296. DOI:http://dx.doi.org/10.1152/jn.01125.2002

Hugo Théoret, Lotfi Merabet, and Alvaro Pascual-Leone. 2004. Behavioral and neuroplastic changes in the blind: Evidence for functionally relevant cross-modal interactions. *Journal of Physiology–Paris* 98, 1 221–233. DOI:http://dx.doi.org/10.1016/j.jphysparis.2004.03.009

J. Trouvain. 2007. On the comprehension of extremely fast synthetic speech. *Saarland Working Papers in Linguistics* 1, 5–13. http://scidok.sulb.uni-saarland.de/volltexte/2007/1176/

Martine Turgeon, Albert S. Bregman, and Brian Roberts. 2005. Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping. *Journal of Experimental Psychology. Human Perception and Performance* 31, 5, 939–953. DOI:http://dx.doi.org/10.1037/0096-1523.31.5.939

Yolanda Vazquez-Alvarez and Stephen A. Brewster. 2011. Eyes-free multitasking: The effect of cognitive load on mobile spatial audio interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2173–2176.

Martin D. Vestergaard, Nicholas R. C. Fyson, and Roy D. Patterson. 2009. The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America* 125, 2, 1114–1124.

Markel Vigo and Simon Harper. 2013. Coping tactics employed by visually disabled users on the web. *International Journal of Human–Computer Studies* 71, 11, 1013–1025.

Voice Interaction 2014. Voice Interaction. Retrieved December 22, 2015 from http://www.voiceinteraction.eu/.

B. N. Walker, A. Nance, and J. Lindsay. 2006. Spearcons: Speech-based earcons improve navigation performance in auditory menus. *Proceedings of ICAD.* 63–68. Retrieved December 22, 2015 from http://sonify.psych.gatech.edu/~walkerb/publications/pdfs/2006ICAD-WalkerNanceLindsay.pdf.

Takayuki Watanabe. 2007. Experimental evaluation of usability and accessibility of heading elements components of web accessibility. *Disability & Rehabilitation: Assistive Technology* 1–8.

David Wechsler. 1981. *WAIS-R Manual: Wechsler Adult Intelligence Scale-Revised*. Harcourt Brace Jovanovich, New York, NY.

Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1, 111–123.

Elizabeth Grace Winkler. 2008. Understanding language. *Continuum International.*

Beverly A. Wright and Matthew B. Fitzgerald. 2001. Different patterns of human discrimination learning for two interaural cues to sound-source location. *Proceedings of the National Academy of Sciences* 98, 21, 12307–12312. DOI:http://dx.doi.org/10.1073/pnas.211220498

Yeliz Yesilada, Robert Stevens, Simon Harper, and Carole Goble. 2007. Evaluating DANTE: Semantic transcoding for visually disabled users. *ACM Transactions on Computer–Human Interaction* 14, 3 14–es. DOI:http://dx.doi.org/10.1145/1279700.1279704

W. A. Yost. 1998. Spatial hearing: The psychophysics of human sound localization, revisited edition. *Ear and Hearing* 19, 2, 167.

P. M. Zurek. 1993. Binaural advantages and directional effects in speech intelligibility. *Acoustical Factors Affecting Hearing Aid Performance* 2, 255–275.

Received March 2015; revised June 2015; accepted September 2015