Text-to-Speeches: Evaluating the Perception of Concurrent Speech by Blind People

João Guerreiro, Daniel Gonçalves
Instituto Superior Técnico, Universidade de Lisboa / INESC-ID
Rua Alves Redol 9, 1000-029, Lisboa, Portugal
joao.p.guerreiro@ist.utl.pt, daniel.goncalves@inesc-id.pt

ABSTRACT

Over the years, screen readers have been an essential tool for assisting blind users in accessing digital information. Yet, its sequential nature undermines blind people's ability to efficiently find relevant information, despite the browsing strategies they have developed. 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, oppositely to one sequential speech channel, we hypothesize that blind people can leverage concurrent speech channels to quickly get the gist of digital information. In this paper, we present an experiment with 23 participants, which aims to understand blind people's ability to search for relevant content listening to two, three or four concurrent speech channels. Our results suggest that it is easy to identify the relevant source with two and three concurrent talkers. Moreover, both two and three sources may be used to understand the relevant source content depending on the task intelligibility demands and user characteristics.

Categories and Subject Descriptors

H.5. [**Information Interfaces and Presentation (e.g. HCI)**]: Multimedia Information Systems – Audio Input/Output.

Keywords

Cocktail party effect; Screen reader; blind; visually impaired; skimming; scanning; concurrent speech.

1. INTRODUCTION

Screen readers have a central role in providing access to digital information to visually impaired users. Making this information accessible is crucial and has been the object of intensive research (e.g. [5]). A different challenge arises from the need to process potentially useful (and accessible) information more efficiently. Sighted users are able to quickly sift through a document or web page by looking through visually prominent content or diagonal reading. These skills enable to get a general idea of the content – *skimming* – or to find specific information – *scanning* [1].

Although blind people lack this visual ability, they have developed browsing strategies [7, 29, 30], such as navigating through headings and increasing the speech rate, which help them to mitigate this limitation. Nevertheless, comparisons between sighted and blind users browsing the web highlight significant differences in prejudice of the latter [6, 26]. Unlike the visual presentation on screen that depicts a lot of information at a time, screen readers rely on a sequential channel that impairs a quick overview of the content.

The sequential characteristic of screen readers, 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 the attention on a single talker among several conversations and background noise [12]. Moreover, one may detect interesting content in the background (e.g. own name or favourite subject) and shift the attention to another talker.

In addition, there is evidence that our brain's ability to segregate simultaneous speech depends on characteristics such as the number of concurrent talkers [10], their differences in spatial locations [9, 10], or voice characteristics [10, 14, 28], among others. In fact, a good configuration of these characteristics enhances the speech intelligibility for both selective [10, 14] and divided attention [24, 27] tasks. In the former, one focuses the attention on a specific talker, whilst in the latter the attention is divided amongst several speech sources. It is important to note that most experiments that focus on speech use small phrases, wherein the participants have to identify all words. We believe that with longer sentences people will be able 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 [12] pioneer study, which reported one's ability to perceive an entire *cliché* by hearing just a few words.

A central tenet of our approach relies on the fact that blind people have enhanced capabilities to segregate speech signals [19]. This fact is due to the process of *Neuro-Plasticity*. In the particular case of blindness, it states that a blind person's brain is reorganized so part of their visual cortex is used in auditory processing [11]. Harper highlighted such advantage [18] when suggesting the use of simultaneous audio sources to convey web information faster to visually impaired users.

In this paper, we argue that screen reader users can leverage the *Cocktail Party Effect* to scan for relevant information more efficiently. As a use case scenario, while exploring news sites one may be targeting specific subjects to pay further attention to. Instead of listening to all headings sequentially, one could listen to two or three simultaneously to detect the relevant ones. We believe that the use of concurrent speech enables blind people to listen to several unrelated information items (e.g. articles in news sites, search results and social media posts), get the gist of the information and identify the ones that deserve further attention.

ASSETS '14, October 20 - 22 2014, Rochester, NY, USA

ACM 978-1-4503-2720-6/14/10…$15.00.
http://dx.doi.org/10.1145/2661334.2661367

We present an experiment with 23 visually impaired people that aims to evaluate the perception of concurrent speech whilst scanning for relevant information. In particular, we address the following questions: 1) How many voices can blind users listen to, and still be able to identify the one with relevant content? 2) And to keep track of its content? 3) Do differences in voice characteristics enhance both identification and selective attention?

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 sources may be used to understand the relevant source's content depending on speech intelligibility demands and user characteristics.

2. RELATED WORK

The related work reviewed in this section is three-fold: first, we look into the research and techniques that aim to accelerate blind people's textual scanning; second, we provide a background on speech segregation using multiple sound sources; third, we present example applications that make use of simultaneous speech feedback.

2.1 Fast-Reading Techniques

Screen reader users develop several browsing strategies in order to overcome accessibility and usability limitations. For instance, web users may re-check their actions, increase the speech rate or navigate through HTML heading elements to obtain an overview of the website [7, 29]. These strategies may indeed help them browse more efficiently. For instance, a proper use of *Heading* elements can significantly speed up web browsing, particularly for scanning tasks [26, 30]. Yet, information overload remains a heavy load, as "*the biggest problem in non-visual browsing remains the speed of information processing*" [7].

A frequent approach to surpass this challenge is summarization (e.g. [1, 17]. Yet, other approaches are needed to, by themselves, or together with current browsing techniques, accelerate blind people's information processing.

2.2 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 [12]. Moreover, one may detect interesting content in the background (e.g. own name or favourite subject) and shift the attention to another talker.

Several researchers investigated how concurrent speech intelligibility can be maximized. Although intelligibility decreases with the increase of competing talkers, the separation of speech signals between ears (dichotic speech) outperforms the use of mixed signals (monaural speech) [12] and is only surpassed by spatial audio [10]. In fact, the use of spatial audio is also valuable in divided attention tasks, where people have to pay attention to two speech signals [24, 27]. Regarding voice characteristics, Brungart, Darwin and colleagues [10, 14] showed the advantage in using different gender talkers, as it makes use of the human brain's ability to segregate sound frequencies [8]. Moreover, alike the use of increasing speech rates with practice, there is also evidence that even short-term training improves sound segregation and identification [2].

The increasing use of speech and sound in the interaction with computers may leverage this phenomenon to provide information more efficiently and/or effectively. Actually, blind people, in particular early-blind, are more capable to discriminate speech than sighted people are, due to the process of *neuro-plasticity* [19]. It states that areas of the brain that are not used (in this case, the visual cortex) are reorganized for different purposes [11].

2.3 Simultaneous Sound Applications

The insights provided by the aforementioned experiments led to applications that try to take advantage of concurrent speech to present larger amounts of information more efficiently. Sasayaki [22] 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 providing important contextual information. Other authors introduced spatial audio to map [13, 16] the current position in a web page, while a different voice provided other information.

Another example is Clique [20], which places 4 assistants with distinct voices around the user in a virtual sound space. Therein, 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 [23] 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 colleagues [25] 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 only able to focus in a single file.

Aoki and colleagues [3] presented a social audio space supporting multiple simultaneous conversations. They monitored the participants' behaviour to identify conversational floors as they emerge and to modify the audio delivered to each participant enhancing the signals of interest. SpeechSkimmer [4] tries to present recorded speech faster by presenting the most important segments to an ear and the discarded material to the other.

These applications are valuable contributions for their scenarios and tasks; yet, there are no guarantees that they are suitable when *scanning* for relevant content. We intend to leverage the knowledge of the previous section and assess if similar conclusions can be drawn to the use of longer sentences, when *scanning* for relevant information.

3. TEXT-TO-SPEECHES

Our main goal was to evaluate the perception of concurrent speech by blind people, in order to leverage this ability to accelerate their access to information. In this section, we describe the framework, Text-to-Speeches, which enabled such evaluation.

Text-to-Speeches is able to position several pre-recorded audio files in a 3D space simultaneously. We built a java framework on top of Paul Lamb's 3D Sound System[1], using the LightWeight Java Game Library (LWJGL[2]) binding of OpenAL Soft 1.15.1[3]. This setting supports the use of digital filters called Head Related Transfer Functions (HRTFs), which simulate the acoustic cues used for spatial localization [32]. The HRTFs are based on

[1] http://www.paulscode.com/forum/index.php?topic=4.0

[2] http://lwjgl.org/

[3] http://kcat.strangesoft.net/openal.html

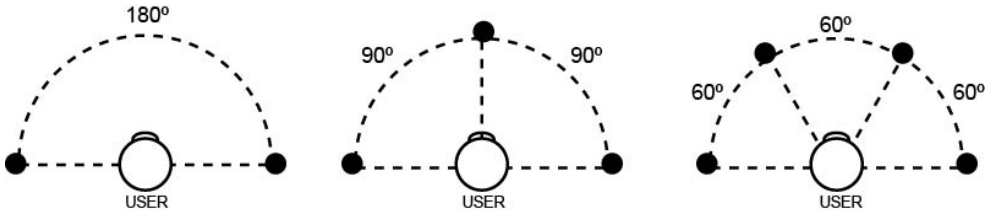

Figure 1. The sound source spatial positioning, in the user's frontal horizontal plane, for two, three and four talkers.

measurements influenced by the listener's head and ears. Alike most experiments (e.g. [10, 24]) and for simplicity purposes, we used non-individualized measurements from a KEMAR manikin (in this case, from MIT[4]).

Current Text-to-Speech software demands a unique, sequential auditory channel. Therefore, we pre-recorded all sentences to *.wav* files, using *DIXI* [21], a TTS developed by *INESC-ID's Spoken Language Systems Laboratory*[5] and now commercialized by *Voice Interaction*[6] (Vicente's voice – male). These audio files are then placed at different positions in the 3D audio space.

To guarantee different, controlled voices we manipulated our original voice's pitch (Glottal Pulse Rate - GPR) and formant frequencies (Vocal Tract Length – VTL), using the Praat software[7] the same way Darwin did [14]. Furthermore, we assured that all voices had the same mean intensity.

4. EVALUATION

The main purpose of this experiment was to investigate blind people's ability to cope with simultaneous speech, to perform fast-reading tasks such as scanning for relevant information items. In detail, we intend to answer the following research questions: 1) How many voices can blind users listen to, and still be able to identify the one with relevant content? 2) And to keep track of its content? 3) Do differences in voice characteristics enhance both identification and selective attention?

4.1 Methodology

In this experiment, the multi-talker 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 not overbalance 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.

4.1.1 Number of Talkers and their Location

Our main research questions focus in the number of simultaneous talkers that a blind user 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 increase, whereas results are nearly 50% of success with four speech sources [10]. Although these results focus on different tasks, they were a good indicator for the number of sources we should consider. We have decided to conduct the experiment with two, three and four simultaneous talkers.

The sound sources locations took inspiration from several experiments that use equally spaced positions in the frontal horizontal plane (e.g. [9, 15]). Although other spatial configurations were proposed and provided better results overall, they ended up sacrificing specific locations that dropped their results significantly [9]. Figure 1 shows our spatial setting. The sound sources are separated by 180º, 90º and 60º, for two, three and four talkers, respectively.

4.1.2 Voice Characteristics

Most experiments on simultaneous speech segregation focus on pitch variations. Yet, the best results are achieved when varying the two main characteristics that influence male and female voices - the pitch (GPR) and formant frequencies (VTL) [10, 14].

We wanted to validate these differences for longer speech signals. We resorted to a single voice whose characteristics were manipulated to obtain different voices. Similarly to Vestergaard and colleagues [28], 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 towards both male and female talkers, rather than the predominance of one gender in the experiments. The analysis of previous research resulted in three conditions:

1. **Same Voice.** 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.
2. **Large Separation.** This condition aimed at the larger known separation that could still provide an improvement in performance, for both pitch and VTL variations [14]. In this condition, each voice differs from the subsequent 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 192.2 Hz, while with three voices they were 155 (the central voice), 237.2 and 100.8 Hz. The central, androgynous voice was manipulated to obtain all the others. This rather large separation between voices, when resorting to four talkers, results in voices similar to Darwin's *super male* and *super female*, which deviate from normal human voices [14].
3. **Small Separation.** 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

4 http://sound.media.mit.edu/resources/KEMAR.html

5 http://www.l2f.inesc-id.pt/

6 http://www.voiceinteraction.eu/

7 http://www.fon.hum.uva.nl/praat/

voices for all talkers (including with 4). Moreover, these values are very close to the larger separation in Vestergaard's study [28]

4.2 Task and Dataset

Daily, people search for information among search engine results, posts, tweets, mail messages or news. Therein, lies a decision of which pieces of information are relevant and deserve 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 in 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.

The 103 snippets were randomly selected and held the following constraints: contained only Portuguese words, correctly pronounced by the TTS; all resulting audio files have durations between 10 and 11 seconds; and we changed names, places and any other element that could benefit the previous knowledge of particular news or subjects. Moreover, the sentences were chosen randomly such that none was presented twice per participant.

The task consisted in finding the relevant source among the presented snippets at each trial (there could be 2, 3 or 4 simultaneous sources) and try to understand its content. Before the trial, the researcher provides a set of cues (consistent across participants), which work as a hint, to simulate the search for relevant information.

4.3 Procedure

The experiment comprised 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 in a training centre for blind and visually-impaired people. The characterization session took approximately 15 minutes and included an oral questionnaire about demographic data and screen reader usage and a working memory assessment. To measure the working memory, the subtest *Digit Span* of the revised *Wechsler Adult Intelligence Scale* (WAIS-R) was used [31]. In a first phase, the participant must repeat increasingly long series of digits presented orally, and on a second stage, repeat additional sets of numbers but backwards. Such tasks allow the calculation of a grade to the participant's working memory.

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 technologic 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 in one practice trial and six test trials for each possible number of talkers (2, 3 and 4 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 s/he was not able to identify the first 3 with four talkers. Fifteen participants completed the condition with four talkers.

The six trials followed a randomized order and consisted in two trials for each voice characteristics condition (same voice, large and small separation). We assured that both the voice (except in *same voice* condition) and the location of the relevant source were different for those two repeated trials.

Each trial consisted on the following five phases:

1. **Hint given by the researcher.** The researcher gives a hint about which news/sentence the participant should pay attention. 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 [10, 14], but they use only one word due to their smaller sentences (5 words).
2. **Play simultaneous speech.** The simultaneous sentences started to play at the exact same time. The participant tries to identify the relevant sentence and understand as much content as possible.
3. **Participant's Report.** Participants report the content of the relevant sentence. They are encouraged to reveal everything they heard and remember, using the same or different words. Related experiments [10, 14] ask participants to report the exact same words. Herein, we want to understand if people can get the gist of the information, independently of the words perceived.
4. **Question.** The researcher asks a question about the relevant sentence, only if the participant did not refer the answer in the previous report. This question is used to help recalling some of the previously heard content. All sentences have a pre-defined question whose answer is not in the first three seconds nor in the final two seconds.
5. **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 6 trials per number of talkers, we asked for participants' feedback. The evaluation procedure took on average 45 minutes.

4.4 Apparatus

The Text-To-Speeches framework, previously described in Section 3, 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.

4.5 Participants

Twenty-three (23) visually impaired participants, 17 male and 6 female, took part in the experiment. Their ages ranged from 22 to 62 (M=40.74, SD=12.36) years old. Nine (9) participants had a congenital visual impairment or their onset age preceded the 18 years old (6 of them are fully blind), while 14 had later onset ages (11 fully blind). They were recruited from a training centre for visually impaired people. No participant reporteded having severe hearing impairments and only 2 reported low experience with screen readers.

4.6 Design and Analysis

We resorted to a 3x3x2 within-subjects design where participants tested each combination of *Number of Sources* level (2,3, or 4) and Source Separation (small, medium and large) two times. Furthermore, 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 randomized. This design resulted in 366 trials, whereas 15 participants completed all conditions (18 trials) whilst the remaining 8 didn't complete the condition with 4 voices (12 trials). We performed Shapiro-Wilkinson tests of the observed values for our continuous dependent variables (relevant source identification error rate, description completeness). These showed to be not normally distributed; we applied non-parametric statistical tests to assess differences (Friedman test was used to compare 3 groups while Wilcoxon signed rank tests were used to perform post-hoc comparisons between pairs of samples (Bonferroni corrections were applied). Spearman test was used to assess correlations of non-normal data or ordinal data.

5. RESULTS

Our goal was to understand how blind people cope with simultaneous information items in a *Relevance Scanning* task. In this evaluation, we analyze blind 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.

5.1 Identification

After each trial, participants were asked to identify which sound source contained the relevant sentence. In 366 trials, participants were able to identify the correct speech source in 301 of them (82%). In detail, participants mentioned the source location in 298 trials, whilst the talker's voice was mentioned in 16 trials.

Figure 2 presents the success rate in the identification of the audio source. It shows that voice variations alone did not affect the identification of the relevant source ($p>0.05$ for all comparisons within each set of sound sources – two, three and four). This can be explained by the length of our sentences (nearly 10 seconds), which provide more time to explore the audio space.

In contrast, the number of sources has a significant effect on sound source identification, mainly between two and four talkers ($p<0.001$ for all comparisons between each set of sound sources – two, three and four that match the identical voice separation). Moreover, results also differ when comparing between two and

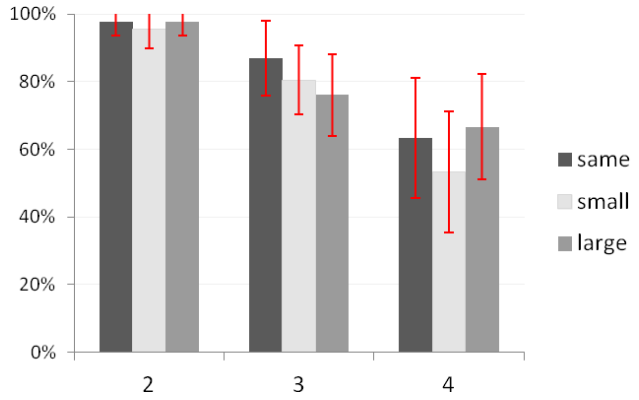

Figure 2. The success rate (y-axis) for the identification of the relevant audio source, per number of sources and voice conditions. Error bars denote 95% confidence intervals

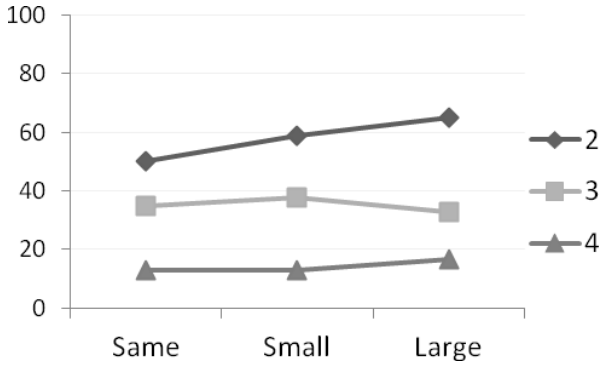

Figure 3. The success rate (y-axis) for correct answers to the pre-determined question, per voice characteristics and number of sources.

three sources, mostly in the large separation condition ($p<0.01$); however, the conditions with the same voice ($p=0.096$) and with small separations ($p=0.035$) also suggest an effect of the number of talkers from two to three. The difference between three and four talkers is also clear for both the same voice ($p<0.01$) and small separation ($p<0.01$) conditions. Still, there is not a significant difference in the large separation condition ($p=0.132$). A deeper insight on this matter is provided by the participants' comments: 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 users are able to identify the relevant source when there are two simultaneous talkers. In fact, 20 (from 23) participants were able to identify the relevant source in all six trials. Moreover, eight participants were able to keep this record with three simultaneous talkers, whilst seven missed only one trial. On the other hand, with four talkers no participant indentified the relevant source in the six trials (three were able to identify it in five trials).

5.2 Intelligibility and Report

To assess speech intelligibility we relied on two methods: first, participants reported everything they recalled about the relevant sentence; then, we asked them a specific question about it (if they have not answered it already). An analysis to the questions' correctness supports the decreasing tendency of speech intelligibility when the number of talkers increases (Figure 3).

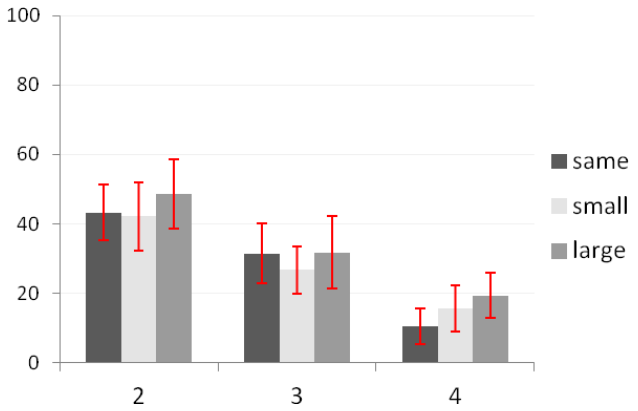

Figure 4. Average completeness (y-axis) of user descriptions – how much was reported - per number of sources and voice condition. Error bars denote 95% confidence intervals.

Table 1. *Spearman's rho correlation* between digit span scores and sentence completeness for each condition.

		2-large	2-same	2-small	3-large	3-same	3-small	4-large	4-same	4-small
Digit span	rho	0.482	0.349	0.761	0.633	0.559	0.447	-0.111	0.118	0.339
	Sig.	**0.031**	0.132	**0.000**	**0.003**	**0.010**	**0.048**	0.719	0.701	0.257
	N	20	20	20	20	20	20	13	13	13

Moreover, it seems to indicate an advantage for larger voice separations when listening to two concurrent talkers. Specifically, 65% of the answers were correct when listening to two talkers in the *large* condition (76% if we consider incomplete answers). Furthermore, these differences in the number of talkers seem to be consistent among users as seven participants were able to answer correctly to at least five trials with two talkers, but none of them achieved that result with three or four talkers. Although this measure provides us some indicators, it cannot be used to assess the intelligibility of the entire sentence. It might be the case that the participants missed, or forgot, that specific part.

The completeness of a participant's description derives from the percentage of relevant content (of the target sentence) that s/he reported (Figure 4 presents the average completeness). To measure their descriptions' completeness, we considered all verbs, nouns, adjectives and adverbs in the sentence. To establish a percentage for each description, we accounted those that were reported, either using or not the exact same words. These elements varied between 14 and 20 in the 103 news. A Friedman test for each number of talkers condition showed no effect of voice characteristics in the sentence reports. Yet, the number of sources 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 suggest a minor effect of the number of sources.

An average of the six trials for each talker condition shows that seven participants reported more than half sentence content when listening to two talkers, whilst three of them were able to keep that result with three talkers. If we consider an understanding of a quarter of the sentence, the numbers rise to eighteen and twelve participants for two and three talkers, respectively.

Although being a cognitively demanding task, these results suggest that the use of simultaneous speech depends on the ratio of information that needs to be processed. Moreover, the user's cognitive abilities are also crucial to assess the usage of multiple talkers. Table 1 presents the Spearman's rho correlation between *Digit Span* scores and sentence completeness for each condition. It shows medium to large correlations between digit span and all the conditions with two and three talkers, suggesting that the participants were able to hear the sentences, but meanwhile forgot the content. In fact, to recall what they had been listening to, was referred as the main challenge by most participants, for two and three talkers. Yet, an accurate identification could allow the user to select or go back to the relevant item for further analysis. Moreover, participants reported that it was easier to recall information about sentences that they were genuinely interested.

5.3 Relevant Talker's Position and Voice

The positions for each number of talker's conditions were fixed and established beforehand. Still, the relevant snippet could vary among them. Figure 5 shows the average report completeness for each location depending on the number of talkers. The results are very similar with two talkers, but with three talkers, the differences are larger. The lower score for the frontal position (23.8%, in comparison to 30.7% and 35.2% for left and right, respectively) is supported by ten participants' comment, which mentioned that it was more difficult to listen to the frontal voice. One 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 3 sources, which augments the confusion*". With four talkers, the lateral left position held the best results. Users commented that the lateral audio sources were best perceived than the diagonal ones; however, we found no explanations for the differences between left and right positions (24% and 11%, respectively).

The variations of the relevant talker's voice ended up not having a noticeable effect. The only exception is with four talkers, where the high-pitched voice held better results (26.7%) in comparison to the others (from 10.4% to 15,4% for the androgynous and woman's voices, respectively). One participant noted that "*the high-pitched voice is irritating, but actually it is easier to distinguish it in the midst of several talkers*".

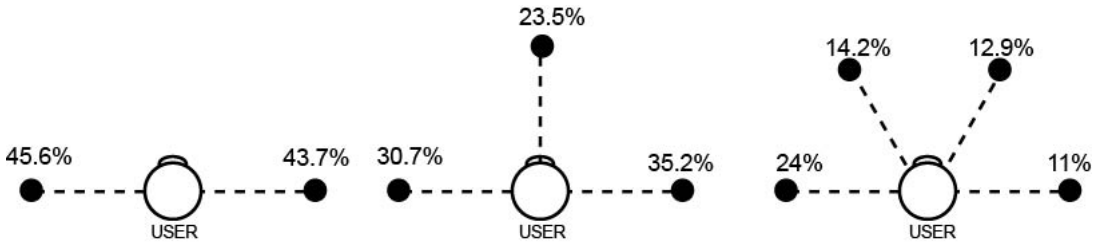

Figure 5. Average report completeness for each location, depending on the number of talkers.

5.4 Early and Late Blind Participants

The eight participants that were unable to complete the condition with four talkers were either late blind or had partial sight. Although this result suggests an effect of *neuro-plasticity* for early blind (congenital and onset prior to 18 years) participants, a Mann-Whitney U test revealed no significant differences in sentence completeness (neither source identification) between late and early blind participants, in all conditions. In contrast, a significant difference in a *two-talker* condition (*small* separation) was observed between early fully blind (six participants) and late or partially sighted participants. However, further research would be needed in order to understand the effects of onset age and residual vision, in this particular task.

6. DISCUSSION

After analyzing all data, we are now able to address the research goals by revisiting the aforementioned questions.

Two and three concurrent talkers enable identification. Results show that blind people are able to identify the relevant snippet when listening to two simultaneous talkers (Figure 2). In fact, 20 of 23 participants had a 100% success rate. Despite the fact that the identification rate reduces with three talkers, some users are still able to identify the relevant snippet. In particular, 15 participants identified the relevant snippet in at least five of the six trials. 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 news feed posts 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 snippet. 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 [13], but can also be applied to the usage of gestures in touch screens, specific keys in keyboards, among others.

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 in 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 participant suggested the use of "*three talkers in search engines, as the relevant result is usually among the first three*". In cases where the intelligibility demands are greater, the option should go to two talkers. Actually, one 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. *Digit Span* scores are highly correlated to the amount of information reported after a trial. Moreover, several users pointed out the difficulty to recall 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 can only take advantage of simultaneous speech in tasks where 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, whilst 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 neither for speech identification nor to intelligibility. Although the related work shows that using different frequencies enhance speech segregation, it also shows that each attribute provide a greater effect when varied alone [10]. Herein, 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 detail, 16 participants preferred listening to different voices, while only two preferred the same voice. One 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*.

7. CONCLUSIONS

Previous research concerning the *Cocktail Party Effect* supports the perception of simultaneous speech sources. We intended to leverage this knowledge and assess if similar conclusions could be drawn to the use of longer sentences in scanning tasks. Likewise the related work, 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, differences in voice characteristics did not provide a greater effect in neither speech identification nor 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 [2]. This experiment comprised a unique session with approximately 45 minutes. 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, some participants noted that the subject of the news influence their ability to recall and report what they have heard. In fact, 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 [10]. For instance, one could listen to sports content always on the right side, whilst economics on the left.

We provided useful guidelines to the use of concurrent speech in fast-exploration tasks. In future work, we aim to explore

interaction mechanisms to cope with the additional demands. Moreover, results suggested a slight advantage for early-blind participants. Still, further research is needed in order to assess the effect of *neuro-plasticity* in this *Relevance Scanning* task.

8. ACKNOWLEDGMENTS

We thank the *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. Work supported by national funds through *Fundação para a Ciência e Tecnologia*, under project PEst-OE/EEI/LA0021/2013.

9. REFERENCES

[1] Ahmed, F. et al 2012. Why Read if You Can Skim : Towards Enabling Faster Screen Reading. *In proc. of W4A.*

[2] Alain, C. 2007. Breaking the wave: effects of attention and learning on concurrent sound perception. Hearing research, 229(1), 225-236.

[3] Aoki, Paul M., et al. 2003 The mad hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations. Proceedings of CHI, ACM.

[4] Arons, B. 1997. SpeechSkimmer : A System for Interactively Skimming Recorded Speech. ACM TOCHI – Special issue on speech as data, 4(1):3-38.

[5] Asakawa, C., & Takagi, H. 2008. Transcoding. In Web Accessibility (pp. 231-260). Springer London.

[6] Bigham, J. P. et al 2007. WebinSitu: a comparative analysis of blind and sighted browsing behavior. In Proc. of ASSETS (pp. 51-58). ACM.

[7] Borodin, Y. et al. 2010 More than meets the eye: a survey of screen-reader browsing strategies. In Proc. of the 2010 W4A.

[8] Bregman, A. S. 1994 Auditory scene analysis: The perceptual organization of sound. MIT press.

[9] Brungart, D. S., and Simpson, B. D. 2005. Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception* (TAP), 2(4), 430-436.

[10] Brungart, D. S., & Simpson, B. D. (2005). Improving multitalker speech communication with advanced audio displays. Air Force Research Lab Wright Patterson AFN OH.

[11] Burton, H. 2003. Visual cortex activity in early and late blind people. The *Journal of neuroscience*, 23(10), 4005-4011.

[12] Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and with two ears. The *Journal of the acoustical society of America*, 25(5), 975-979.

[13] Crispien, K.et al 1996. A 3D-Auditory Environment for Hierarchical Navigation in Non-visual Interaction. Proceedings of ICAD.

[14] Darwin, C. J., Brungart, D. S., and Simpson, B. D. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. The Journal of the Acoustical Society of America, 114, 2913.

[15] Drullman, R. and Bronkhorst, A. 2000. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. The Journal of the Acoustical Society of America.

[16] Goose, S., & Möller, C. 1999. A 3D audio only interactive Web browser: using spatialization to convey hypermedia document structure. In Proc. the ACM international conference on Multimedia (Part 1) (pp. 363-371). ACM.

[17] Harper, S., and Patel, N. 2005. Gist summaries for visually impaired surfers. In Proc of ASSETS (pp. 90-97).

[18] Harper, S. (2012). Deep Accessibility: Adapting Interfaces to Suit Our Senses. Invited Talk - University of Lisbon. [online] Available at: http://www.slideshare.net/simon-harper/adapting-sensory-interfaces.

[19] Hugdahl, K.et al. 2004. Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds. Cognitive brain research, 19(1), 28-32.

[20] Parente, P. (2006) .Clique: a conversant, task-based audio display for GUI applications. ACM SIGACCESS Accessibility and Computing 84: 34-37.

[21] Paulo, Sérgio, et al. 2008, DIXI–a generic text-to-speech system for European Portuguese. Computational Processing of the Portuguese Language. Springer Berlin Heidelberg.

[22] Sato, D. et al. 2011. Sasayaki: augmented voice web browsing experience. In Proc of CHI , ACM.

[23] Schmandt, C. and Mullins, A. 1995. AudioStreamer: exploiting simultaneity for listening. Conference companion on Human factors in Computing Systems, pages 218-219.

[24] Shinn-Cunningham, B. G., & Ihlefeld, A. 2004. Selective and Divided Attention: Extracting Information from Simultaneous Sound Sources. In ICAD.

[25] Sodnik, J. et al. 2010 Enhanced synthesized text reader for visually impaired users. Advances in Computer-Human Interactions

[26] Takagi, H. et al. 2007. Analysis of navigability of Web applications for improving blind usability. ACM Transactions on Computer-Human Interaction, 14(3):13{es.

[27] Vazquez-Alvarez, Y., & Brewster, S. A. 2011. Eyes-free multitasking: the effect of cognitive load on mobile spatial audio interfaces. In Proceedings of CHI (pp. 2173-2176).

[28] Vestergaard, M. D. et al. 2009. The interaction of vocal characteristics and audibility in the recognition of concurrent syllablesa).The Journal of the Acoustical Society of America, 125(2), 1114-1124.

[29] Vigo, M., & Harper, S. (2013). Coping tactics employed by visually disabled users on the web. International Journal of Human-Computer Studies, 71(11), 1013-1025

[30] Watanabe, T. 2007. Experimental Evaluation of Usability and Accessibility of Heading Elements Components of Web Accessibility. Disability & Rehabilitation: Assistive Technology, pages 1-8.

[31] Wechsler, D. 1981. WAIS-R manual: Wechsler adult intelligence scale-revised. Psychological Corporation.

[32] Wenzel, E. M. et al (1993). Localization using non-individualized head related transfer functions. The Journal of the Acoustical Society of America, 94(1), 111-123.