Faster Text-to-Speeches: Enhancing Blind People's Information Scanning with Faster Concurrent Speech

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

Blind people rely mostly on the auditory feedback of screen readers to consume digital information. Still, how fast can information be processed remains a major problem. The use of faster speech rates is one of the main techniques to speed-up the consumption of digital information. Moreover, recent experiments have suggested the use of concurrent speech as a valid alternative when scanning for relevant information. In this paper, we present an experiment with 30 visually impaired participants, where we compare the use of faster speech rates against the use of concurrent speech. Moreover, we combine these two approaches by gradually increasing the speech rate with one, two and three voices. Results show 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, *One-Voice* requires larger speech rate increments, which cause a considerable loss in performance. Overall, results suggest that the best compromise between efficiency and the ability to understand each sentence is the use of *Two-Voices* with a rate of 1.75**default-rate* (approximately 278 WPM).

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation (e.g. HCI)**]: Multimedia Information Systems – *Audio Input/Output*.

General Terms

Design, Experimentation, Human Factors.

Keywords

Concurrent Speech; Speech Rate; Screen Reader; Text-to-Speech; Blind; Visually Impaired; Scanning; Skimming; Auditory Perception; Cocktail Party Effect.

1. INTRODUCTION

Blind people rely mostly on the auditory feedback of screen readers to consume digital information. Efficiency is a problem especially in situations where relevant information must be recognized among large amounts of irrelevant one. Sighted users use *scanning* as a strategy to achieve this goal, by glancing at all content expecting to identify information of interest to be subsequently analyzed with further care. Increasing the speech rate is one of the main techniques that blind people use for this purpose and may be used either alone or combined with other strategies (e.g. navigate through heading or links) [5].

Although researchers have reported a gradual decrease in intelligibility and comprehension as the speech rate increases [27], they have also shown its ability to accelerate information consumption substantially. For instance, two distinct experiments [3] and [27] reported blind users' ability to understand at least 50% of the content using rates nearly three times faster than the default rate (500 Words per Minute – WPM).

Despite the benefit provided by faster speech rates, we have argued before that the screen readers' sequential auditory channel is impairing a quicker overview of the content, when compared to the visual presentation on screen [15]. We proposed the use of concurrent speech to accelerate the scanning for relevant content, by taking advantage of the *Cocktail Party Effect*. It describes the ability to focus the attention on a single voice among several conversations and background noise, but still be able to detect interesting content in the background [8]. Moreover, the identification and intelligibility of the concurrent voices may be enhanced with the use of different spatial locations and different gender voices (e.g. [6]).

In a previous experiment [15], we have reported that blind users are able to identify and understand one particular sentence when listening to two or three simultaneous sentences (news snippets), suggesting that current screen readers can be imposing limitations on the way auditory feedback is being provided. While these results point out concurrent speech as a proper alternative to faster speech rates, only a direct comparison can determine their relative benefits and limitations. Moreover, in that experiment users had to identify and understand one relevant sentence. However, searching for relevant content does not require users to understand the whole content, but it asks for a decision of which information items (e.g. news) are relevant and deserve further attention.

In this paper, we compare the use of faster speech rates against the use of concurrent speech when scanning for relevant digital information. We combine these two approaches by gradually increasing the speech rate with one, two and three voices. In order to guarantee a fair comparison, we assign different speech rates depending on the number of *Voices* (*One-Voice*, *Two-Voices* or *Three-Voices*), so that they take the exact same time to read the same number of sentences. We refer to it as *Information Bandwidth* condition; the higher the bandwidth, the lesser time it takes to read the sentences (independently of the number of *Voices*). We present an experiment with 30 visually impaired participants, where in each condition they had to listen to news snippets and identify all that belong to a specific subject. In particular, we aim to answer the following research questions: 1) What *Voice* condition enables the best content comprehension for

ASSETS '15, October 26 - 28, 2015, Lisbon, Portugal

ACM 978-1-4503-3400-6/15/10$15.00
DOI: http://dx.doi.org/10.1145/2700648.2809840

each *Information Bandwidth*?; 2) How does increasing the speech rate affect scanning for relevant content with one, two and three voices?; and 3) What combination of speech rate and number of voices enable the fastest scanning for relevant content, while maintaining the basic understanding of the content?

Results show that the *One-Voice* condition performs significantly worse than concurrent speech as the *Information Bandwidth* increases. Moreover, both *Two* and *Three-Voices* maintain their high performances in the first conditions. This means that smaller increments in the concurrent voices' speech rate enables faster scanning for relevant content, maintaining the comprehension of the sentences. In contrast, to keep-up with concurrent speech timings, *One-Voice* requires larger speech rate increments, which cause a considerable loss in performance. Overall, results suggest that the best compromise between *Information Bandwidth* and the ability to understand each sentence is the use of *Two-Voices* with a rate 1.75**default-rate* (~278.4 WPM).

2. RELATED WORK

The access to digital information, for instance on the web, imposes several barriers for visual impaired users, as web pages are usually designed with visual interaction in mind [13]. A great effort has been made to assure the accessibility of digital content to visually impaired people. Studies have been conducted in order to understand how visually impaired users cope with their difficulties to access digital information (e.g. [31]). Moreover, the strategies that they use to accelerate their browsing experience, such as navigating through headings or increasing the speech rate [5] help them browse more efficiently ([5] [28]). Researchers have also suggested the use of summarization in order to reduce the amount of information to read while providing only a gist of the content [1] [18].

The speed to process digital information is still a major problem for visual impaired people [5] [17], particularly when compared with sighted peers [4] [28]. In this section, we review work related with the two approaches compared in this article: the usage of faster speech rates and concurrent speech.

2.1 Using Faster Speech Rates

Mainstream screen readers such as *JAWS*, *NVDA*, *VoiceOver* or *Talkback* soon started to enable the manipulation of their voices speech rates. The need for an easy and interactive way to change the speech rate with immediate response [3], resulted in faster ways to control it, such as using keyboard shortcuts instead of navigating to the respective menu. This interactivity allows users to control the speech rate depending on the content intelligibility requirements and allows them to deal with the exhaustion caused by the use of very high speech rates [29].

Previous research has explored the intelligibility and comprehension of synthesized speech at higher speech rates, observing a gradual decrease as the speech rate increases (e.g. [27] [30]). Furthermore, other experiments have shown that visually impaired people are able to listen and understand synthesized speech at higher speech rates than sighted people (e.g. [29]). However, this ability depends on factors such as the person's age, being a native speaker and familiarity with both the synthesizer and the voice [27].

Asakawa and colleagues [3] reported that blind advanced users were able to listen to sentences in Japanese at speech rates 2.8 times faster (approximately 500 WPM) than the default rate of the screen reader and still understand at least 50% of the information. Moreover, novice users were able to listen to a screen reader 1.6 times faster than the default rate and be able to understand the entire content. Stent and colleagues tested speech rates up to 550 WPM (in English), reporting a comprehension above 50% with a 500 WPM rate for all synthesizers tested [27].

2.2 Using Concurrent Speech

The human ability to process concurrent speech is supported by the *Cocktail Party Effect* [8]. In the middle of conversations and background noise, one is capable to focus the attention on a single voice and still detect interesting content in the background and shift the attention accordingly. Literature reviews have documented several features that increase concurrent speech intelligibility. Such features include the use of dichotic speech (separate the sound sources between ears) when using two competing voices [6] [8] and the use of spatial audio with three or more voices [6]. Another example is the use of different gender voices [6] [10], due to the human brain's ability to segregate different sound frequencies. The benefits of its usage are more moderate when already using spatial audio, but it is clearly preferred to the use of the same or similar voices [15]. Some evidence supports that blind people, in particular early-blind, show enhanced sensitivity to discriminate speech sounds [20] due to the process of *neuro-plasticity* – the reorganization of unused areas of the brain for different purposes [7].

In line with experiments about the *Cocktail Party Effect* and with the increasing use of speech in the interaction with computers, researchers have tried to leverage this phenomenon to convey information more efficiently, particularly to screen reader users. *Sasayaki* [23] is a web browser that augments the standard auditory channel with a whispering voice that, among other things, may locate the screen reader position in the web page or provide important contextual information. Goose and Moller [14] also use two voices, but added spatial audio to indicate the location in the web page. *SpatialTouch* [16] is a non-visual multitouch QWERTY keyboard for tablet devices that supports two-handed input through spatial and simultaneous audio feedback. It relies on male and female voices, which spatial location depend on each character position in the keyboard.

Both *Spatial Speaker [26]* and *AudioStreamer* [25] use simultaneous speech to present different sentences in different spatial locations. The first is able to present text files read in pre-defined spatial locations, while the latter reads three audio news programs and detects head movement to select the current focus of interest. On the other hand, SpeechSkimmer [2] divides the content of the same sentence between the two ears to present it faster, by selecting the most important segments to an ear and the remaining to the other.

We have conducted an earlier experiment with visually impaired people to understand the limits of concurrent speech when scanning for relevant content [15]. Participants had to listen to two, three or four simultaneous news snippets while trying to identify the relevant one and understand its content. We found that such task is easy to perform when listening to two simultaneous voices and for most people with three. However, four voices considerably decreases both the identification and intelligibility of the relevant sentence.

The studies and projects described in this section support the use of concurrent speech to accelerate the consumption of digital

Table 1. The speech rate as the number of times it is faster than the default-rate (and mean WPM) for all IB X Voice conditions

Voices	Information Bandwidth (IB) Condition					
	2	3	3.5	4	5	6
1	2 (325.6 wpm)	3 (488.4 wpm)	3.5 (569.8 wpm)	4 (651.2 wpm)		
2	1 (159.1 wpm)	1.5 (238.7 wpm)	1.75 (278.4 wpm)	2 (318.2 wpm)	2.5 (397.8 wpm)	3 (477.3 wpm)
3		1 (161.7 wpm)	1.167 (188.7 wpm)	1.333 (215.5 wpm)	1.667 (269.6 wpm	2 (323.4 wpm)

information. Similarly to the use of faster speech rates, our former experiment [15] suggests it could be leveraged by screen reader users when scanning for relevant content.

3. EXPERIMENTAL SETTING

In this experiment, we investigate two approaches that intend to accelerate blind people's scanning for relevant digital content: 1) the use of faster speech rates; and 2) the use of concurrent speech. In addition, we combine these two approaches and compare the identification of relevant content with one, two and three voices, using different speech rates. An important concept in this investigation is that of *Information Bandwidth* (*IB*), which refers to how fast the information is transmitted in comparison to a baseline condition – one voice with the default speech rate (*default-rate*). For instance, an *IB* of 2 means the content is transmitted two times faster, which can be through the use of *Two-Voices* with the *default-rate* or *One-Voice* twice as fast. We defined 6 *IB* conditions, where we assigned different speech rates to each *Voice* condition so that they take the same time to complete. In Table 1, are depicted all *IB X Voice* conditions (detailed in *Section 3.2.2*). We aim to answer the following research questions:

1. What *Voice* condition enables the best content comprehension for each *Information Bandwidth*?
2. How does increasing the speech rate affect scanning for relevant content with one, two and three voices?
3. What combination of speech rate and number of voices enable the fastest scanning for relevant content, while maintaining the basic understanding of the content?

3.1 Auditory Feedback

Current *Text-to-Speech* (TTS) software restricts the auditory feedback to a unique, sequential auditory channel. Therefore, we relied on the *Text-to-Speeches* framework[1], the same way we did in our previous concurrent speech experiment [15]. This framework is able to position several simultaneous pre-recorded audio files in a 3D space and supports the use of digital filters (*Head Related Transfer Functions)* that simulate the acoustic cues used for spatial localization [33].

All sentences were pre-recorded to .wav files, using three different voices (two male and one female) from *DIXI* [22], a TTS developed by *INESC-ID's Spoken Language Systems Laboratory*[2] and now commercialized by *Voice Interaction*[3]. These voices had the following mean pitch and Words Per Minute (WPM) rate using the experiment dataset:

Female (*Violeta*)**.** Mean pitch of 186.2 Hz and 162.8 WPM

Male1 (*Vicente*)**.** Mean pitch of 111.3 Hz and 155.3 WPM

Male2 (*Viriato*)**.** Mean pitch of 98.0 Hz and 166.9 WPM

The voices selected were based on studies that suggest the use different gender voices to enhance speech segregation [6] [10]. We only used voices with similar pitches (*male1* and *male2*) in the condition with three voices, but in very far spatial locations (right and left ears, respectively). Moreover, the *female* voice was placed between them (in a central position), to maximize the segregation of the three voices. Such spatial and frequency differences assured the segregation of the voices, without the need to manipulate their frequencies, which would reduce their quality without guaranteed benefits on speech segregation [15].

In order to obtain the desired speech rate variations, we manipulated the audio files using the *Praat* software[4] with a linear time compression algorithm (PSOLA [21]). Small gains could be accomplished with non-linear methods that use, for instance, silence-cut, but with a significant increase in system complexity [19]. Furthermore, we measured the sound intensity levels of all sound files, obtaining a mean intensity value in decibels. Then, we used *Praat* to adjust the voice intensities so that all voices had the same mean intensity level.

3.2 Methodology

We based our decisions on the results of previous experiments, regarding both the speech rates used and the concurrent speech setting. Moreover, we performed a pilot study with 5 people, where we gradually increased the speech rate with one, two and three voices in order to determine their maximum values.

3.2.1 Concurrent Speech

Previous research about the *Cocktail Party Effect* has shown that speech intelligibility decreases as the number of voices increase (e.g. [6]). In particular, we have reported a greater decrease with four voices [15]. As a result, besides a one-voice condition, we only included two and three concurrent voices.

In order to enhance speech segregation, we used different, spatially separated voices (Figure 1). The spatial separation was inspired on experiments that use equally spaced positions in the users' frontal horizontal plane (e.g. [6] [11] [15]). In the *One-Voice* condition, we use a single channel with a *Female* voice. In *Two-Voices*, the speech sources were separated by 180º, where the *Female* and *Male1* voices talked to the users' right and left ear, respectively. With *Three-Voices*, the speech sources were separated by 90º, where the two *Male* voices talked to the users' right and left ears, with the *Female* voice in a central position.

[1] https://bitbucket.org/sound/texttospeeches (Visited in 07/2015)

[2] http://www.l2f.inesc-id.pt/

[3] http://www.voiceinteraction.eu/

[4] http://www.fon.hum.uva.nl/praat/ (Visited in 07/2015)

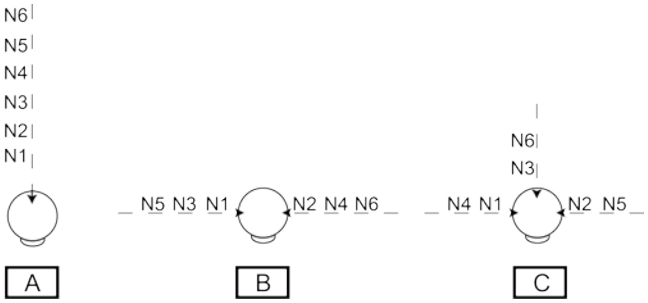

Figure 1. The voices spatial positions in the frontal horizontal plane for A) One, B) Two and C) Three voices. It shows a sequence of 6 sentences (*News*) for all conditions.

3.2.2 Information Bandwidth (IB) Conditions: Voices and Speech Rates

A direct comparison between the use of faster speech rates and the use of concurrent speech, can only be achieved if both approaches take the same time to read the same amount of information. Table 1 shows all the *IB X Voice* conditions and the speech rate increment comparatively to the *default-rate* (and the mean WPM rates of the voices used).

For instance, an *IB* of 2 requires the use of two concurrent voices with the *default-rate* (condition *2 X 2*) or a single voice twice as fast (*2 X 1*). Likewise, three *default-rate* voices (*3 X 3*) match the use of one voice three times faster (*3 X 1*) or two voices 1.5 times faster (*3 X 2*). These first two conditions correspond to a final time to complete reading the content of half and one third the *default-rate* reading time with one voice, respectively. In the subsequent conditions, we reduced gradually the time needed to read the entire content, which resulted in faster speech rates either with one, two or three voices. The pilot study helped us to determine the maximum speech rates for each number of voices in order to avoid to frustrate and to overwhelm the participants.

In order to enable the direct comparison among the use of one, two or three voices, we used six sentences in all conditions. This means that in the conditions with one voice, the participants received the six sentences sequentially. In contrast, with two voices they listened to three sequences of two simultaneous sentences, and two sequences of three simultaneous sentences with three voices (exemplified in Figure 1). In all conditions, a short *beep* was played between news to indicate the end of the current sentence(s) and to prepare the following.

4. DATA COLLECTION

We referred to *Relevance Scanning* as the process of assessing "*which pieces of information are relevant and deserve further attention*" [15]. In our prior experiment, we restricted the number of relevant sources to one and asked the user to understand its content. In contrast, herein we focused exclusively on the relevance assessment, but forcing the users to make such decision for all sentences instead of focusing on a single one. With this goal in mind, we were able to measure the participants' ability to obtain a basic understanding of all sentences in order to determine their relevance.

We gathered a dataset of 200 news snippets from a Portuguese news site archive, which included only raw text and had similar sizes and durations (between 11 and 14 seconds). The news were divided in three groups, so that their durations could differ the maximum of one second from each other. This is most important in the concurrent speech conditions, since larger periods with one voice would benefit them. In order to overcome that issue, we stopped all concurrent voices at the same time (when the first one ended).

4.1 Relevance Scanning Task

All news were categorized in three different topics by the researchers: "*sports*", "*politics and economy*" and "*television, arts and celebrities*". We validated this categorization by asking 4 people to categorize all news. All news that could fit different categories were excluded.

In all trials, participants were told one of these topics and had to identify which news were related to that specific topic. Such task demanded the user to try to understand the topic of all news in order to assess its relevance. The news were chosen randomly such that none was presented twice per participant. We guaranteed the existence of two to four relevant news in all trials, but participants were unaware of this fact, believing there could be zero to six (all relevant) news.

4.2 Procedure

The experiment took approximately 50 minutes per participant and was conducted in a single session in a training center for blind and visually impaired people. First, we informed the participants that the overall purpose of the experiment was to investigate two approaches to accelerate the scanning for relevant digital information: faster speech rates and concurrent speech. Then, we performed an oral questionnaire about demographic data, screen reader usage and conducted a working memory assessment using the *Digit Span* of the revised *Wechsler Adult Intelligence Scale* (WAIS-R) [32].

Afterwards, we explained the experimental setting, which included the existence of conditions with one, two and three voices and how they work; the task of identifying the relevant news among six news and how to make such identification. In all conditions, the participants were in a sit position with the hands on top of a table. To mark one sentence as relevant, participants needed to tap the table in front of them:

1. **One-Voice.** Tap any of the hands to select the current sentence as relevant;
2. **Two-Voices.** Tap with the right or left hand to select the current sentence of the right or left ear, respectively (should tap both if both are relevant, but not necessarily tap at the same time)
3. **Three-Voices**. Same as Two-Voices for lateral sources, but should tap on top of a book between the two hands (placed there for the purpose) to select the central voice.

The three possible topics were thoroughly described in order to clarify their meaning and avoid interpretation doubts. Finally, we referred that participants could always correct a selection, during (e.g. withdraw a selection after understanding the tap was incorrect) and after a trial (e.g. "*I didn't tap the first on the left, but at the end I understood that it was relevant as well*").

Participants performed three practice trials before starting the evaluation, one for each number of voices in their first available *IB* condition. They performed every *IB X Voice* condition presented in Table 1 twice, resulting in 30 trials per participant. We randomized the order of the *Voice* conditions within every *IB* to enable a fair direct comparison between conditions that take the

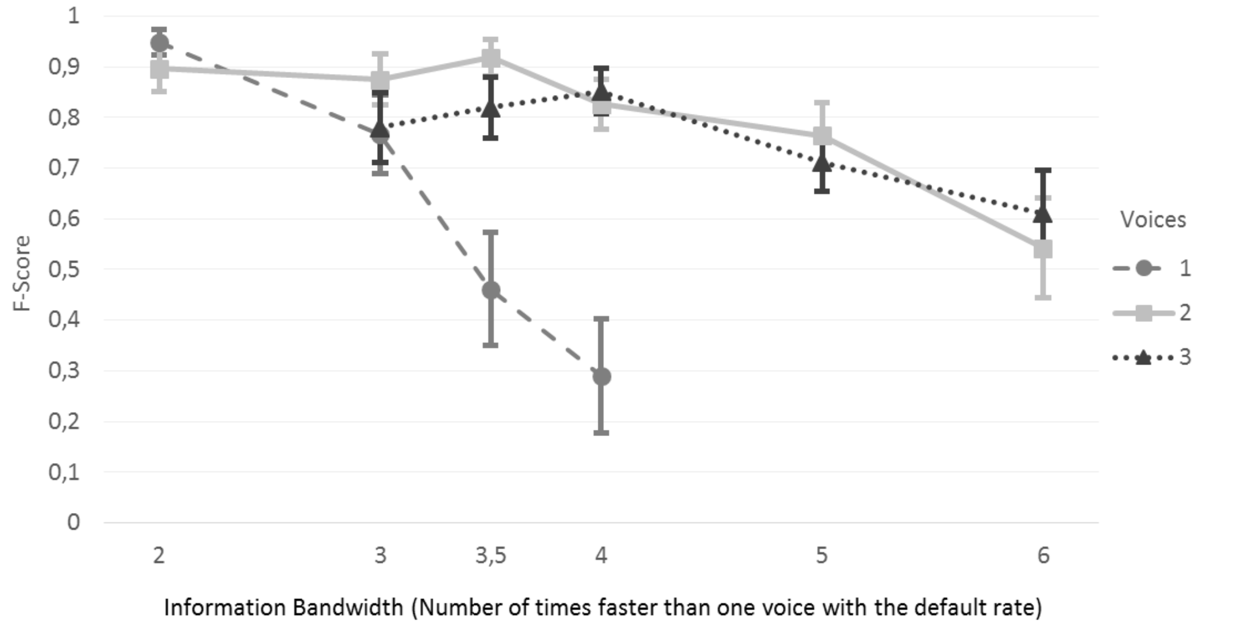

Figure 2. F-Score performance for all *Information Bandwidth X Voice* condition. Error bars denote 95% confidence intervals.

same time to complete. The *IB* conditions had a fixed ascending order, as we wanted to take advantage of the little practice of previous trials to understand the speech rate limits with one, two and three voices, rather than understanding which *IB* condition presents better results for novice users.

In each trial, we first indicated the *Voice* condition and the relevant topic that participants needed to identify (e.g. *sports*); then, we played the news and participants identified the relevant ones as they listened to them. After the consecutive two trials with the same *IB X Voice* condition, participants were asked the *Single Ease Question* [24], which required them to rate the task easiness for that condition, using a 7-point *Likert Scale* (from *Very Difficult* – 1 – to *Very Easy* – 7) .

4.3 Apparatus

The auditory feedback was provided by the aforementioned *Text-to-Speeches* framework. Participants used *AKG K540 Headphones* that were connected to the researcher laptop and were able to adjust the volume to a comfortable rate. The researcher controlled the whole experiment through a Java application and registered the participants' answers as they selected the relevant news. Both sound and video were collected during the whole session for further analysis.

4.4 Participants

We recruited for this experiment 30 visually impaired participants, 9 female and 21 male. Their ages ranged from 23 to 64 (M=43.43, SD=12.07) years old. None participant reported having neither severe nor moderate hearing impairments. There were 12 fully blind (light perception at most) and 18 low vision participants. Moreover, 28 participants used screen readers to interact with their devices on a daily basis. Nineteen (19) participants had a congenital visual impairment or acquired it before turning 18 years old.

4.5 Design and Analysis

We used a within-subjects design where participants performed each *IB X Voice* condition (in Table 1) twice, resulting in a total of 30 trials. One participant did not perform the last two conditions (*6X2* and *6X3*) due to fatigue, while another did not perform one condition (*6X2*) due to time restrictions. Overall, this experiment resulted in 894 trials.

In order to evaluate the participants' accuracy identifying the relevant news, we relied on the *F-Score* as it accounts both for *Precision* and *Recall* (their harmonic mean). *Precision* refers to the fraction of identified news that were in fact relevant, while *Recall* refers to the fraction of relevant news that were identified. This measure accounts both for erroneous identifications and missing identifications. When no news are identified, the *Precision* value is undefined and the *F-Score* equals the *Recall*, which is zero. *Shapiro-Wilkinson* tests were applied to the dependent variable *F-Score* in all *IB X Voice* conditions. As they were not normally distributed, we applied the *Friedman* test when comparing three or more groups and the *Wilcoxon Signed Rank* test to perform the *post-hocs* between pairs of samples (with *Bonferroni corrections* to deal with multiple comparisons).

5. RESULTS

Our main goal was to understand how the use of faster speech rates and concurrent speech affect the performance of a *Relevance Scanning* task. In this evaluation, participants were required to pay attention to six news and identify the ones referring to a particular topic. First, we compare the use of different voices within *Information Bandwidth* (*IB*) conditions. Furthermore, we analyze how each *Voice* condition evolves as the *IB* increases. Figure 2 shows the *F-Score* for each *IB X Voice* condition. Finally, we analyze the effect of users' characteristics on their performance and their subjective opinions and ratings to all *IB X Voice* conditions.

5.1 F-Score Within *Information Bandwidth*

The first *IB* condition – **2** – comprises only two *Voice* conditions (*One-Voice* and *Two-Voices*), because the speech rate would have been slower than the default one in the *Three-Voices* one. In this condition, a *Wilcoxon Signed Rank* test (Z=-1.940, p=.052) showed no significant differences, but a slightly superior *F-score* for the *One-Voice* condition (M=0.949, SD=0.070) than the *Two-Voices* one (M=0.897, SD=0.130).

The following three *IB* conditions – **3**, **3.5** and **4** – comprise the three *Voice* conditions. *Friedman* tests showed significant differences among *Voice* conditions, within each *IB* (p<.01 in all comparisons). *Post-hoc* analysis of *IB* **3** revealed a significant advantage for the *Two-Voices* (M=0.875, SD=0.142) condition over the *One-Voice* (M=0.766, SD=0.216) and *Three-Voices* (M=0.780, SD=0.192) (p<.01 in both comparisons). Moreover, no significant differences were found between *One* and *Three-Voices* conditions (Z= -0.157, p=.875).

IB **3.5** presented clearer differences among the *Voice* conditions (p<.001 in all post-hoc comparisons). Again, the *Two-Voices* condition presented a significantly higher *F-score* (M=0.918, SD=0.101), followed by *Three-Voices* (M=0.820, SD=0.170) and at last by *One-Voice* (M=0.461, SD=0.312).

In the *IB* **4** condition, *Two-Voices* (M=0.826, SD=0.138) and *Three-Voices* (M=0.852, SD=0.126) obtained clearly higher *F-Scores* than the *One-Voice* (M=0.289, SD=0.315) condition (p<.001 in both comparisons). However, in this case *Two-Voices* and *Three-Voices* scores presented no significant differences (Z=-1.067, p=.286). *IB* **4** was the last condition with *One-Voice*. This decision was based on literature reviews and a pilot study, but it is reinforced by these results as they showed a clear disadvantage in comparison to its alternatives.

IB **5** presented no significant differences between the use of *Two-Voices* (M=0.763, SD=0.183) and *Three-Voices* (M=0.712, SD=0.162) (Z=-1.320, p=.187). Although with lower *F-Scores*, *IB* **6** presented no significant differences (Z=-0.913, p=.361) between the two conditions as well (*Two-Voices*: M=0.542, SD=0.265; *Three-Voices*: M=0.61, SD=0.232).

5.2 F-Score Within *Voice* Conditions

In this section, we analyze the effect that the increase of *IB* (and speech rate) has on each *Voice* condition. Figure 2 depicts a clear and consistent decrease in performance in the ***One-Voice*** condition as the *IB* increases (p<.001 in all consecutive comparisons). This result is explained by the very fast speech rates that this condition reaches. To cite one example, *IB 4* results in a speech rate four times faster than the *default-rate*, which is approximately 651.2 *WPM*.

As the *IB* increases, the speech rates of both *Two-Voices* and *Three-Voices* conditions increase at a lower rate than with *One-Voice*. In particular, *Two-Voices* and *Three-Voices* speech rates are half and one-third the rate of the *One-Voice* condition, respectively. These smaller rate increments across *IB* conditions resulted in smoother differences overall.

The ***Two-Voices*** analysis revealed a very similar performance in the first three *IB* conditions (2, 3 and 3.5). In fact, the mean performance slightly increased (non-significant: Z=-1.605, p=.109) from *IB* 3 to 3.5 (from 0.875 to 0.918), probably due to the slighter increase in speech rate and a minor practice effect from previous conditions. *IB* 4 showed a significant difference in comparison to *3.5* (Z=-3.061, p<.005) and was the first to present a considerable difference when comparing with the first condition (*IB* 2) (Z=-2.328, p=.020). In this condition, the speech rate of both voices equals 2**default-rate* and even though its performance decreased, it reached a mean *F-Score* of 0.826. Furthermore, 7 participants identified correctly all news in both trials in this condition (14 in the first condition), suggesting it may also be used for *Relevance Scanning*. Although results did not show a significant difference between *IB* 4 and 5 (Z=-1.061, p=.289), they showed a difference between 5 and 6 (Z=-3.725, p<.001).

The ***Three-Voices*** condition started in the second *IB* condition (*IB* **3**) with a mean *F-Score* of 0.780. The speech rate increases very smoothly with three concurrent voices, which explains the absence of significant differences among *IB* 3, 3.5 and 4. The mean performance ended up reaching the maximum of 0.852 in *IB* 4, possibly due to the brief practice time in the previous conditions. The analysis revealed significant differences between *IB* 4 and 5 (Z=-3.328, p<.001). Finally, the comparison between *IB* 5 and 6 also suggested a small difference in F-Score performances (Z=-2.250, p=.024). In these two conditions, only 3 participants were able to answer correctly to both trials.

In all *Voice* conditions, performance was affected by *Recall* as *IB* increased. Most participants were still able to identify correctly the relevant snippets (high *Precision* values), but failed to identify other relevant snippets in higher *IB* conditions. For instance, the mean *Precision* values for both *Two-Voices* and *Three-Voices* conditions were always above 0.88 (between 0.970 and 0.882). The *One-Voice* condition reached lower, but still high values in the 3.5 and 4 *IB* conditions (0.831 and 0.758, respectively). In contrast, the mean *Recall* reached 0.413 and 0.256 in the same *IB* conditions with *One-Voice*. In particular, in the *IB 4* condition, 13 participants did not select any snippet as relevant, resulting in undefined *Precision*, zero *Recall*, and zero *F-Score*.

The mean *Recall* values for both *Two-Voices* and *Three-Voices* conditions also decreased as the *IB* increased. *Two-Voices* reached 0.771, 0.742 and 0.478 in the *IB* 4, 5 and 6 conditions, respectively, showing a greater decrease in the last one (the speech rate is 3**default-rate*). In the same *IB* conditions, *Three-Voices* reached 0.797, 0.639 and 0.532, respectively.

5.3 User Characteristics

Mann-Whitney U tests revealed no significant differences between fully blind and low vision participants for all *IB X Voice* conditions. Likewise, no differences were found between people that acquired their visual impairment early or late in life.

Medium to large negative correlations (from *rho*=-.411 to *rho*=-.591, p<.05) were found between the participants' ages and their *F-Scores* in all conditions where the speech rate is 2.5 times the *default-rate* or higher (*One-Voice* with *IB* 3, 3.5 and 4; and *Two-Voices* with *IB* 5 and 6). Such correlations are supported by previous research that demonstrated a decline in speech perception with higher speech rates as people age (e.g. [12]). Furthermore, a negative medium correlation was found between age and the *IB* 3.5 *X Two-Voice* condition (*rho*=-.445, p<.05), but we found no explanation to such correlation.

The working memory is known to affect the ability to block out distracting information in concurrent speech scenarios [9]. In line

Table 2. Results of the Single Ease Question for all Information Bandwidth X Voice conditions.

Information Bandwidth	Voices	Median	Q1	Q3	IQR
2	1	6	5	7	2
	2	5.5	4.25	6	1.75
3	1	4	2	5	3
	2	5	5	6	1
	3	5	4	6	2
3.5	1	2.5	1	4	3
	2	5	4	6	2
	3	5	4	5.75	1.75
4	1	1.5	1	3	2
	2	5	4	6	2
	3	5	4	6	2
5	2	5	3	5	2
	3	5	3.25	5	1.75
6	2	3.5	2	5	3
	3	4	2	4	2

with this relation, we found medium to high positive correlations between the participants' digit span scores and all *Three-Voices* conditions (*rho*=.390 to *rho*=.543, p<.05), but none with *One-Voice* and *Two-Voices*. This suggests that the use of *Three-Voices* may hamper the ability to scan for relevant content, for people with lower working memory. However, a long-term analysis could help to determine if they could benefit from practice and overcome such disadvantage as they become more proficient.

5.4 Subjective Feedback

Table 2 shows the *Median*, *Quartiles 1* and *3*, and *Inter-Quartile ranges (IQR)* of participants' answers to the *Single Ease Question* (1 to 7), for all *IB X Voice* conditions. *Friedman* tests revealed significant differences both within *IB* and within *Voice* conditions (p<.01 in all tests). The condition *IB* 2 was the only one where participants rated *One-Voice* as easier (Z=-2.690, p<.01). Participants commented they felt more comfortable listening to a single voice because they use it daily and do not need to divide their attention.

In the subsequent *IB* conditions (3, 3.5 and 4), *One-Voice* was rated as significantly more difficult than both *Two-Voices* and *Three-Voices* (p<.001 in all *post-hocs*), which supports the performance *F-Scores* previously presented. In contrast, when comparing the ratings given to *Two* and *Three-Voices* there seems to be a small difference in prejudice of the latter only in *IB* 3 (Z=-1.969, p=.049), and no significant differences in the others. This can be explained by the first impact when listening to *Three-Voices*. However, the very slight increase in speech rate did not increase the difficulty in the subsequent *IB* conditions (3.5 and 4). *Three-Voices* started to be considered marginally more difficult in *IB* 5 (Z=-1.964, p=.054 in comparison to *IB* 3). Yet, the difficulty increased significantly from 5 to 6 (Z=-3.214, p<.001).

Two-Voices ratings evolved very similarly, as the first marginal difference was found between *IB* 2 and 4 (Z=-1.941, p=.052), but was amplified in the next conditions (Z=-2.441, p=.015 between *IB* 4 and 5; and Z=-3.611, p<.001 between 5 and 6). On the other hand, the *One-Voice* condition was considerably more difficult as the speech rate (and *IB*) increases (p<.001 in all *post-hocs*).

Only four participants expressed their preference for *One-Voice* overall, but recognized the greater difficulty to understand the content in the last two *One-Voice* conditions (*IB* 3.5 and 4). One participant referred that: *"with such high speech rates I was only able to capture a few keywords, but was not able to get a deeper understanding of the content"*. In contrast, in the same *IB* conditions, *Two-Voices* and *Three-Voices* enabled a "*greater understanding of the content, since it is straightforward to focus on a particular voice and also to switch the attention to another*". The breaking point was in the last *IB* conditions, where *Two* and *Three-Voices* reach faster speech rates. One participant stated that "*the problem is that with faster speech rates I do not have the time to roam through all voices, but I can understand them!*"

In general, as the *IB* increased, participants preferred the use of *Two-Voices* (until IB 3.5 or 4). However, only three participants clearly stated that the use of *Three-Voices* was excessive. In contrast with previous experiments where the central voice was found more difficult to understand, herein 21 participants (70%) noted an equal ability to understand all sound sources. This may be related to the greater contrast between the central voice and the others. One participant commented that he "*could easily focus his attention on the lateral sources due to their locations, but also in the central voice because it was so different from the others*".

After completing all trials, four participants reported their preference for specific topics (e.g. *sports* or *politics*). They noted that it is easier to capture the topic if they are indeed interested and know more information that can help in that process. One participant stated: "*I like soccer, so if I listen a player's name I immediately relate it to sports. However, if I listen a television series or an actor's name it is more difficult as I do not know many of them*".

6. DISCUSSION

Based on the results described in the previous section, we present the main *take-home* messages by revisiting our research questions.

It is better to use *Two-Voices* or *Three-Voices* with slightly faster speech rates than *One-Voice* with very fast speech rates. Within *Information Bandwidth* (IB) comparisons showed an explicit advantage for concurrent speech conditions, unless for the first condition where *One-Voice* has a speech rate of 2**default-rate*. In the subsequent conditions, the *One-Voice* speech rate has to increase substantially in order to keep up with *Two-Voices* and *Three-Voices* completion times. The use of speech rates faster than 3**default-rate* markedly decreased the ability to identify the relevant sentences with *One-Voice*, when compared with the alternatives within *IB* conditions. For example, with a bandwidth of 3.5, the *One-Voice F-Score* performance averaged 0.461, while Two-Voices and Three-Voices averaged 0.918 and 0.820, respectively. In this condition, 1, 12 and 8 participants identified correctly all news with *One*, *Two* and *Three-Voices*, respectively. In line with this results, *Two-Voices* was significantly better than *Three-Voices* in the first conditions (up to *IB* 3.5), but their values became very similar as the speech rate (and *IB*) increased.

Gradually increasing the speech rate motivate a major performance loss with *One-Voice*, but a smoother evolution with *Two* and *Three-Voices*. Reducing the time needed to read the six sentences implies a speech rate increment, but at different rates for *One*, *Two* and *Three-Voices* conditions. Results revealed that the larger increments with *One-Voice* lead to a consistent decrease in the ability to identify the relevant sentences (F-Score averages decreased from 0.949 to 0.766, 0.461 and 0.289). In contrast, both *Two-Voices* and *Three-Voices* maintained a very similar performance until the *IB* conditions 3.5 and 4, respectively. This means that the smaller speech rate increments in concurrent speech conditions alongside with a very brief practice from preceding trials enabled participants to maintain their performance, while reducing the time needed to complete the task.

Using *Two-Voices* with 1.75default-rate* or 2**default-rate* is the best compromise between a basic comprehension of the sentences and the speed to process them**. More often than not, participants felt uncertain about the use of concurrent speech before starting the experiment. However, in addition to a clear advantage in performance, participants also preferred the concurrent speech conditions as the *IB* increased. In particular, most participants referred to *Two-Voices* in *IB* 3.5 or 4 as the best compromise between the speed to process all the information and the comprehension of the news topics. These conditions reached a performance of 0.913 and 0.826, respectively. Similar results were also achieved by *Three-Voices* in *IB* conditions 3.5 and 4. Although these values did not show an always perfect identification of the relevant sentences, participants reported an advantage when identifying news of topics that they are genuinely interested and therefore have more knowledge. Overall, the results suggest that the best compromise between information consumption speed and comprehension is the use of *Two-Voices* with a rate 1.75**default-rate* (~278.4 WPM). However, some participants were able to maintain their performance with faster rates. Similarly to what occurs with the use of a single auditory channel, we believe that experience will enable users to gradually increase the concurrent voices rate.

7. CONCLUSIONS

Both the use of faster speech rates and concurrent speech can accelerate blind people's scanning for relevant digital information. We have presented an experiment that compares these two approaches and combines them by gradually increasing the speech rate with one, two and three voices. Results show that *Two* and *Three-Voices* with speech rates slightly faster than the *default-rate*, enable a significantly faster scanning for relevant information, while maintaining its comprehension. In contrast, to keep-up with concurrent speech completion times, *One-Voice* requires larger speech rate increments, which cause a greater loss in performance. Overall, the use of *Two-Voices* with a rate 1.75**default-rate* (~278.4 WPM) enables the most efficient *Relevance Scanning* without a loss in performance. However, several participants were also able to maintain a basic understanding of the sentences in the *IB 4* condition with both *Two-Voices* (2**default-rate*) and *Three-Voices* (1.333**default-rate*). Furthermore, when analyzing the effect of user characteristics, we found that *Age* correlates negatively with their performance when speech rates are higher than *2.5*default-rate*. Moreover, the participants' working memory correlated with their performance with *Three-Voices*, suggesting an effect on the ability to ignore distracting information. A long-term experiment may help to understand if practice can mitigate such effects.

In this experiment, we limited the news snippets durations, in order to guarantee that the concurrent speech sources ended at the same time. It would be interesting to understand how scanning for relevant content is affected by the length of the sentences. In particular, in the last conditions with *Three-Voices*, participants referred that the main problem was not having time to go through all voices, reporting no difficulties to understand their content.

Along with other experiments inspired in the *Cocktail Party Effect,* this experiment supports the use of concurrent speech to accelerate blind people's scanning for digital information. The use of faster, concurrent speech clearly outperformed the use of a single voice with very fast speech rates. While this experiment focused on the identification of relevant news snippets, it would be interesting to explore different contexts and tasks where this approach could be leveraged. Similar scenarios are the ones of websites and applications that comprise lists of items, where only part of them are relevant to the user (e.g. Social Networking Sites, search engine results, RSS feeds, e-mail). Another interesting scenario is the one of notifications that could use a secondary voice to avoid interrupting the user. In contrast, scenarios that require full attention and the comprehension of the whole text do not seem to be appropriate to the use of concurrent speech, nor very fast speech rates as supported by the greater decrease in *Recall*. In future work, we will investigate different scenarios and interaction methods to enable the consumption and exploration of digital information using simultaneous speech sources.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] Ahmed, F., Borodin, Y., Puzis, Y., and Ramakrishnan, I. V. 2012. Why Read if You Can Skim : Towards Enabling Faster Screen Reading. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A)*.

[2] Arons, B. 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.

[3] Asakawa, C., Takagi, H., Ino, S., and Ifukube, T. 2003 Maximum listening speeds for the blind. In *Proceedings of the International Community for Auditory Display (ICAD), pp. 276–279.*

[4] Bigham, J., Cavender, A., Brudvik, J., Wobbrock, J., and Ladner, R. 2007. WebinSitu: a comparative analysis of blind and sighted browsing behavior. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (ASSETS)*.

[5] Borodin, Y., and Bigham, J. 2010 More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A)*

[6] Brungart, D. S., and Simpson, B. D. 2005. Improving Multitalker Speech Communication with Advanced Audio Displays. *Air Force Research Lab Wright-Patterson AFB OH*.

[7] Burton, H. 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.

[8] Cherry, E. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*.

[9] Conway, M. 2001. Sensory-perceptual episodic memory and its context: autobiographical memory. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences 356*, 1413, 1375–84.

[10] 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*, 5, 2913.

[11] 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 107, 4, 2224-2235*.

[12] Fostick, L., Ben-Artzi, E., and Babkoff, H. 2013 Aging and speech perception: Beyond hearing threshold and cognitive ability. *Journal of basic and clinical physiology and pharmacology 24*, 3, 175–183.

[13] Goble, C., Harper, S., and Stevens, R. 2000. The travails of visually impaired web travellers. *Proceedings of the eleventh ACM on Hypertext and hypermedia - HYPERTEXT*, 1–10.

[14] Goose, S., and Moller, C. 1999. A 3D Audio Only Interactive Web Browser : Using Spatialization to Convey Hypermedia Document Structure. *Proceedings of the seventh ACM international conference on Multimedia, pp.* 363–371.

[15] Guerreiro, J., and Gonçalves, D. 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, pp. 169–176.

[16] Guerreiro, J., Rodrigues, A., Montague, K., Guerreiro, T., Nicolau, H., and Gonçalves, D. 2015. TabLETS Get Physical: Non-Visual Text Entry on Tablet Devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 39-42.

[17] Guerreiro, J., and Gonçalves, D. 2013 Blind People Interacting with Mobile Social Applications: Open Challenges. In *Mobile Accessibility Workshop at CHI.*

[18] Harper, S., and Patel, N. 2005 Gist Summaries for Visually Impaired Surfers. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pp. 90–97.

[19] He, L., and Gupta, A. 2001. Exploring benefits of non-linear time compression. In *Proceedings of the ninth ACM international conference on Multimedia*, ACM, pp. 382–391.

[20] Hugdahl, K., Ek, M., Takio, F., Rintee, T., Tuomainen, J., Haarala, C., and Hämäläinen, H. 2004. Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds. *Brain research. Cognitive brain research 19*, 1, 28–32.

[21] Moulines, E., and Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication 9*, 5, 453–467.

[22] Paulo, S., Oliveira, L. C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., and Moniz, H. 2008. Dixi–a generic text-to-speech system for european portuguese. *Computational Processing of the Portuguese Language*, 91–100.

[23] Sato, D., Zhu, S., Kobayashi, M., Takagi, H., and Asakawa, C. 2011. Sasayaki : Voice Augmented Web Browsing Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2769–2778.

[24] Sauro, J., and Dumas, J. S. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 1599–1608.

[25] Schmandt, C., and Mullins, A. 1995. AudioStreamer: exploiting simultaneity for listening. In *Conference companion on Human factors in Computing Systems (CHI'95)*, ACM, 218–219.

[26] Sodnik, J., and Tomažic, S. 2009. Spatial Speaker : 3D Java Text-to-Speech Converter. In *Proceedings of the World Congress on Engineering and Computer Science*, vol. II.

[27] Stent, A., Syrdal, A., and Mishra, T. 2011. On the Intelligibility of Fast Synthesized Speech for Individuals with Early-Onset Blindness. . In *Proceedings of the international ACM SIGACCESS conference on Computers & accessibility*, ACM, 211–218.

[28] Takagi, H., Saito, S., Fukuda, K., and Asakawa, C. 2007 Analysis of navigability of Web applications for improving blind usability. *ACM Transactions on Computer-Human Interaction 14*, 3, 13–es.

[29] Trouvain, J. 2007. On the comprehension of extremely fast synthetic speech.

[30] Valentini-Botinhao, C., Toman, M., Pucher, M., Schabus, D., and Yamagishi, J. 2014. Intelligibility analysis of fast synthesized speech. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[31] Vigo, M., and Harper, S. 2013. Coping tactics employed by visually disabled users on the web. *International Journal of Human-Computer Studies 71*, 11, 1013–1025.

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

[33] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America 94*, 1, 111–123.