# Communicating text structure to blind people with Text-To-Speech

Laurent Sorin, Julie Lemarié, Nathalie Aussenac-Gilles, Mustapha Mojahid, Bernard Oriola

# Communicating Text Structure to Blind People with Text-to-Speech

Laurent Sorin[1], Julie Lemarié[2], Nathalie Aussenac-Gilles[1], Mustapha Mojahid[1], and Bernard Oriola[1]

[1] Université de Toulouse; UPS & CNRS; IRIT; ELIPSE et MELODI, Toulouse, France
{sorin,aussenac,mojahid,oriola}@irit.fr
[2] Université de Toulouse; Laboratoire CLLE, 5 Allées Antonio Machado, Toulouse, France
lemarie@univ-tlse2.fr

**Abstract.** This paper presents the results of an experiment conducted with nine blind subjects for the evaluation of two audio restitution methods for headings, using Text-To-Speech. We used specialized audio and two voices to demarcate headings. This work is part of a research project which focuses on structural information accessibility for the blind in digital documents.

**Keywords:** Accessibility of Digital Documents, Blind People, Document Structure, Text-to-Speech, Specialized Audio.

## 1 Introduction

Accessibility of information contained in digital documents is a crucial challenge for visually impaired people, especially for blind users. Indeed, blind users should be in the center of design issues since Internet and new technologies are an unprecedented opportunity for them to perform tasks that they can hardly do without [1]. Even though there has been much effort on designing assistive technologies and accessible information, the situation often remains frustrating for blind users [2], [3]. Indeed, digital documents in general (web pages, text-documents, spreadsheets, etc.) are primarily designed to be visually displayed, so that the expressive means offered by a spatial layout are often intensively used to create complex objects like tables, graphs, outlines, menus, etc. In this context, our project aims at allowing blind users to access a document's visual properties and logical structure and at designing new reading tools.

In the frame of our project, we focus on documents "textual objects", i.e. every block or portion of the text visually distinct from the rest via its disposition and typography, which we could call the "contrast" principle. The Textual Architecture Model [4], a linguistics model, states that every salient block (or textual object) in a document was created by the author in order to structure his message. The different types of textual objects in a document each have different properties and relationships between them. This structural information coming from the visual properties of a

document is crucial for the understanding of its content from a sighted reader point of view (see [5] for a review).

Most of the digital documents have a very rich layout and many visual properties. Yet, only few existing projects try to restore part of this structural information to blind users: for instance [6] focuses on HTML tables and frames, whereas [7] focuses on enumerations (lists of items in a document) and [8] focuses on hierarchical structures in general. Our global approach in this research project is to annotate the different textual objects along with their properties (as described by the Textual Architecture Model and SARA), and to restore them during document presentation with a Text-To-Speech software. The aim is to restore the multi-level logical structure of the documents, for instance local emphasis and global structure.

In this paper we present our results about the restitution of headings properties to blind people using Text-To-Speech, in order to validate our global approach. We made the hypothesis that providing blind users with headings properties will help them build a better mental representation of the document.

In the first part we describe what structural information is and why we want to make it accessible. The second part describes the methodology used to test our hypothesis. Finally, we present and discuss the obtained results.

## 2 Structural Information Restitution to Blind People

As mentioned before, logical information about documents structure conveyed by the text formatting and layout is crucial in order to comprehend the text content. We chose to focus on the restitution of headings because of their important role in text comprehension. Indeed, according to [9], headings help (sighted) readers to build a global representation of the text topic structure, which improves memorization in general and also activates the reader's relevant prior knowledge.

However, blind people almost never have access to this information while using a Text-To-Speech (TTS) software (for instance via a screen-reader) on a computer. In fact, TTS still struggle to render text objects like headings [10]. Indeed, with a screen-reader a heading is signaled with the sentence "Heading Level N", either on a braille terminal or orally with a TTS. This restitution method doesn't emphasize the heading over the rest of the content, so that the headings are not as distinct from the rest of the text in the audio modality as compared to the printed text. It may consequently hamper the identification of the different text headings by the blind people, what could impair text processing. Indeed, [10] showed that it is easier for sighted readers to catch the text structure when a text containing headings is printed than in an auditory presentation via a TTS. In their second experiment, they also showed that it is possible to improve text structure processing in an auditory presentation by systematically restoring the headings information functions. Our goal in the present study is to assess the efficiency of different restitution methods with blind people. We conducted an experiment with blind volunteers who were instructed to listen to a text oralized by a TTS and then, to answer questions about the text structure.

In order to make the headings more salient, we compare three different restitution methods. The first is the basic restitution provided by a TTS. The second is to use two different voices, one to enunciate the headings and the other one for the rest of the text. The third chosen method is to use spatialized audio to simulate one enunciation location on the left of the participant and another on his right, at head level. We used the left location to enunciate the headings and the right location to enunciate the rest of the text. The principle was to enrich the text presentation by restoring structural information through voice modifications without adding any discursive content to avoid cognitive overload for the listeners. .

The main hypothesis was that enriching the auditory presentation with voice modification to signal headings results in a better comprehension of the text structure than the basic restitution.

# 3 Methodology

The global principle of the experimentation was to present several documents to participants using different restitution methods, and to measure the general comprehension of the documents contents along with the outline retention.

## 3.1 Experimental Design

Nine legally blind volunteers participated in our study, without particular hearing problems, all using synthetic voices on everyday basis. We used a synthetic voice reading text at about 175 words per minute, which is far slower than maximum listening speed for blind people using Text-To-Speech [11]. Three different conditions were defined for documents reading.

The first was the control condition which was equivalent to what a screen-reader would read of a web page containing headings and raw text, that is to say a discursive segment indicating “Heading level N” before the heading oralization. The text in this condition was read using a male voice. The second condition used a female voice to enunciate the headings and a male voice for the documents contents. Note that we used free voices from the MRBOLA Project.[1] Finally, the third condition used spatialized audio and defined two “reading” locations: one on the left in front of the listener and one on the right, both at head level. We chose those particular locations since it appears that the left/right arc in front of the listener is where the locations discrimination works best [12], [13]. In every condition headings were announced by saying “Heading level N” (and “Main title” for the first title of each document). Subjects had the possibility to pause the reading using the space-bar, but couldn’t play it back so that each subject listened one time to every part of every text.

We used five documents in total. Three of them contained about 875 words (5 minutes of listening) and were used to test each of the three conditions. They were expository texts which topics were chose so the subjects would have enough basic

[1] Mbrola page : `http://tcts.fpms.ac.be/synthesis/mbrola.htmll`

knowledge to understand them, but were unlikely experts of the concerned fields; the topics of those three documents were energy problems, firefighting and energy solutions. Each of those three texts had 3 levels of headings: main title, headings level 1 and headings level 2. Two other short documents containing about 450 words (3minutes of listening) were used as distractors during the experiment, and dealt with random topics, namely Brazil and French Louisiana.

In order to avoid a possible rank bias, each condition and each text were played in total three times at each rank (i.e. three times in first position, three times in second and three times in third position during the different tests).

The procedure was the following one: after signing a consent form, we asked subjects few questions about themselves. Then, the three long texts were read, each in a different condition, either in control, dual-voices or spatialized audio condition. After each text we asked the subject to recall the outline of the text, ideally the headings with their level in the hierarchy, along with a self-evaluation of the difficulty they had to comprehend and memorize it, on a scale ranging from 1 to 7, 7 being very difficult and 1 very easy. After that, comprehension questions about the text were asked, each question being related to one or several text sections.

The two "distracting" texts were read after text 1 and text 2. Once the subject heard them, we asked them three questions about details of each text. The aim of the distractors was to prevent subjects from over-focusing on the text headings and to try to understand the whole texts.

Finally, after all the text readings, subjects would give us feedback about the experiment.

### 3.2 Measures

We measured mainly two variables: outline recall and comprehension (through the questions asked after each reading).

Three scores were extracted from the outline recall. First we evaluated the number of topics recalled over the total number of topics, each topic corresponding to a heading, which gave us a score between 0 (no topic recalled) and 1 (all the topics recalled). A second score concerned the recalled hierarchy: we computed the distance between the recalled outline and the original outline, using the absolute difference between the level of recalled headings and the level of the corresponding headings in the documents. Score ranged between 0 (recalled hierarchy identical to the original) and 1 (recalled hierarchy completely different). Finally, we computed a third score corresponding to the correlation between the recalled order of the headings and the original order (normalized between 0 and 1).

Lastly, we rated each of the answers to comprehension questions. The constructed questions dealt with the text macrostructure and the correct answers were the headings contents (e.g. what are the consequences of dwindling fuel resources? correct answer: hazardous productions methods, increasing costs of fuel resources). For each expected topic, we calculated a correctness score to the total of the question, each score ranging between 0 and 4 (0: topic not recalled, 1-3: more or less semantic

equivalent, 4: exact literal topic). For instance, if a question dealt with three different headings, the answer rate ranged between 0 and 12.

# 4 Results

The first graph shows the results regarding outline recall for each tested condition with the three measures we performed on this data (topic recall, hierarchy recall and order recall).

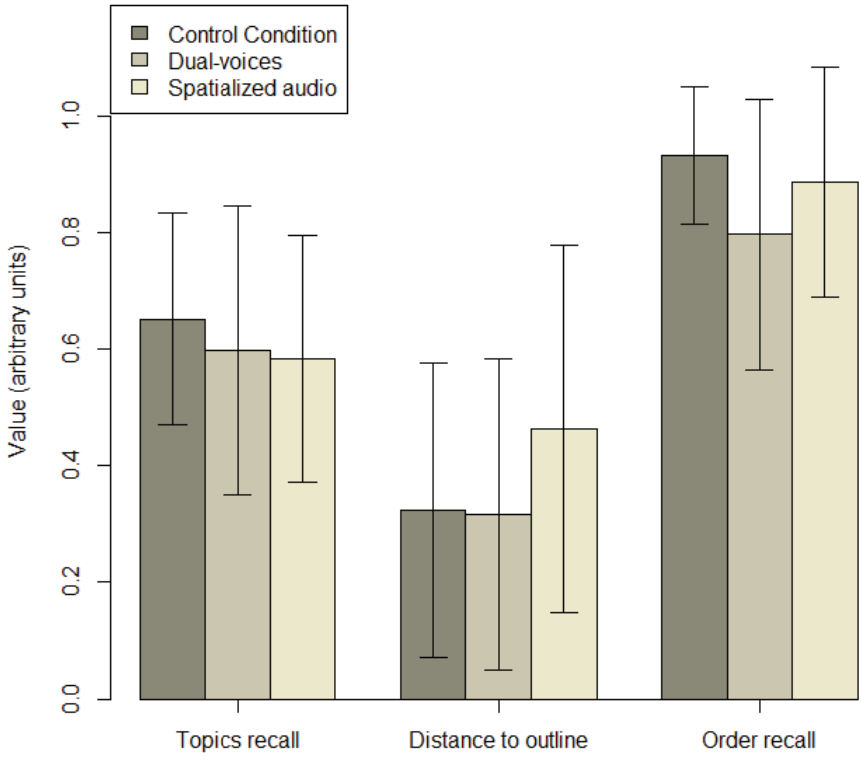


**Fig. 1.** Outline recall scores

The second graph shows the results regarding comprehension scores for each tested condition.

There were no statistically significant differences between the conditions, either in terms of outline recall, comprehension score and mental effort. As our sample is small, this absence of difference might be due to a lack of statistical power. Looking at the data with a descriptive approach shows that the differences observed for the outlining task and reported mental effort are not compatible with our hypothesis. However, the differences observed for the comprehension task, albeit small, are in line with our predictions: both restitution methods entail better comprehension scores than the control condition.

Concerning the user preferences, 5 subjects preferred the dual-voices condition re-gardless of the text content, 2 preferred the spatialized audio condition, 1 liked them both, while the last subject had no particular preference.

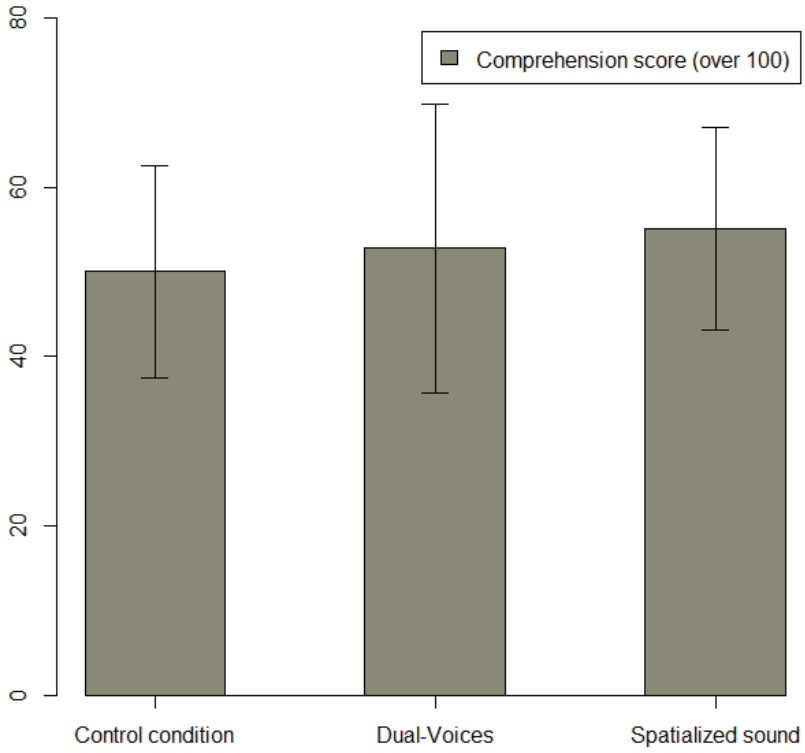


**Fig. 2.** Comprehension scores

## 5 Discussion

This experiment conducted to assess the efficiency of different methods to restore headings structural information gave mixed results. We note a very important variability in all our measures, which may be explained by a heterogeneous sample.

It is possible that subjects mainly relied on the discursive segment "Heading level N" to process the text structure, just as they usually do when reading web pages with their screen readers, all of them being everyday users of screen readers with a Text-To-Speech software. Consequently, our manipulation had no effect on the topic structure identification. However, text comprehension was slightly improved by our restitution methods. Moreover, the post-test results showed some individual preferences. A deeper analysis of data showed that subjects who preferred the spatialized audio condition performed better in this condition than in the control condition, and at the same time had worse performance in the dual-voices condition than in the control condition. The same trend occurred for 3 of the 5 subjects who preferred the dual-voices condition (they performed better in this condition than in the control condition, and worse in the spatialized audio condition than in the control condition). This analysis is also consistent with the post-test questions which showed that subjects who preferred spatialized audio disliked the dual-voices condition. At the same time, subjects who preferred the dual-voices condition disliked the spatialized audio condition. This could mean that blind users have preferences on how to contrast synthesized speech. This fact might also explain why we haven't any global trend on the overall

data, since some of the subjects preferred one tested method over the other and performed better in the preferred method.

We also had positive qualitative feedback from all the subjects who reported that "I felt the demarcation" (subject 5), "it draws attention" (subject 6), "I had an idea of what the next section will deal with" (subject 7), "I can see the structure well" (subject 8), "it is an interesting idea" (subject 9), etc.

One last interesting fact is that even though the two tested restitution methods don't induce more mental effort than the control condition, there is a trend in the spatialized audio condition showing an increase of mental effort. This increase of mental effort could be due to the lack of familiarity with spatialized audio.

## 6 Conclusion

Even though the results don't show statistical evidences that either the dual-voices method or the spatialized audio method has better performance than current TTS oralization, we found that blind users may have preferences on which methods to use. Here, those preferences may have impacted performance.

The feedback from the users encourages us to pursue our research. Future work will focus on creating a new reading system which combines the tested restitution methods (spatialized audio and voice change) with intra-document navigation techniques. This system should be able to take into account individuals preferences. We will also study other textual objects than headings, and possible ways for blind people to access their properties.

**Acknowledgements.** We would like to thank all the volunteers from the IJA (Institute for Blind youths) of Toulouse who took part in our study and especially Claude Griet; their contribution was greatly appreciated. We are also grateful to the "PRES Toulouse" and the "Région Midi-Pyrénées" for funding this work.

## References

1. Giraud, S., Uzan, G., Thérouanne, P.: L'accessibilité des interfaces informatiques pour les déficients visuels. In: Dinet, J., Bastien, C. (eds.) L'ergonomie des objets et environnements physiques et numériques. Hermes - Sciences Lavoisier, Paris (2011)
2. Lazar, J., Allen, A., Kleinman, J., Malarkey, C.: What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. Int. J. Hum. Comput. Interact. 22(3), 247–269 (2007)
3. Petit, G., Dufresne, A., Robert, J.: Introducing TactoWeb: A Tool to Spatially Explore Web Pages for Users with Visual Impairment. In: Stephanidis, C. (ed.) Universal Access in HCI, Part I, HCII 2011. LNCS, vol. 6765, pp. 276–284. Springer, Heidelberg (2011)
4. Pascual, E., Virbel, J.: Semantic and Layout Properties of Text Punctuation. In: Proceedings of the Association for Computational Linguistics Workshop on Punctuation, pp. 41–48 (1996)

5. Lemarié, J., Lorch, R.F., Eyrolle, H., Virbel, J.: SARA: A Text-Based and Reader-Based Theory of Signaling. Educ. Psychol. 43(1), 27–48 (2008)
6. Pontelli, E., Gillan, D., Xiong, W., Saad, E., Gupta, G., Karshmer, A.I.: Navigation of HTML tables, frames, and XML fragments. In: Proceedings of the Fifth International ACM Conference on Assistive Technologies, ASSETS 2002, pp. 25–32 (2002)
7. Maurel, F., Lemarié, J., Vigouroux, N., Virbel, J., Mojahid, M., Nespoulous, J.-L.: De l'adaptation de la présentation oralisée des textes aux difficultés perceptives et mnésiques du langage. Rev. Parol. 2004–29–30, 153–187 (2005)
8. Smith, A.C., Cook, J.S., Francioni, J.M., Hossain, A., Anwar, M., Rahman, M.F.: Nonvisual tool for navigating hierarchical structures. In: Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2004, pp. 133–139 (2004)
9. Lemarié, J., Lorch, R.F., Péry-Woodley, M.-P.: Understanding How Headings Influence Text Processing. Discours. Rev. Linguist. Psycholinguistique Informatique 10 (2012)
10. Lorch, R.F., Chen, H.-T., Lemarié, J.: Communicating headings and preview sentences in text and speech. J. Exp. Psychol. Appl. 18(3), 265–276 (2012)
11. Asakawa, C., Takagi, H.: Maximum Listening Speed For The Blind. In: Proceedings of the 2003 International Conference on Auditory Display, pp. 276–279 (2003)
12. Rumsey, F.: Spatial audio, pp. 1–233. Taylor & Francis (2001)
13. Goose, S., Möller, C.: A 3D Audio Only Interactive Web Browser: Using Spatialization to Convey Hypermedia Document Structure. In: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), pp. 363–371 (1999)