How Sighted And Blind Students Perceive Relational Similarity Between Font-Size And Loudness In Text-To-Speech

Philippos Katsoulis
National and Kapodistrian University of Athens,
Graduate Program in Basic and Applied Cognitive Science
phikats@phs.uoa.gr

Georgios Kourpupetoglou
National and Kapodistrian University of Athens,
Department of Informatics and Telecommunications
koupe@di.uoa.gr

ABSTRACT

Font-size variations constitute text signals that help readers to create an organizational framework for the coding of a text. The current study investigates the relational similarity between font-size and the voice loudness in Text-to-Speech (TtS) as perceived by sighted and congenitally blind students in primary and secondary education. We conducted two experiments with 25 blind and 26 sighted participants. In the first experiment, we have explored how a polar value of one dimension maps to the polar value of the other dimension. In the second experiment we have studied how the notions of the participants about one dimension map on to the perceived poles of the other dimension. The results confirm the hypothesis that all participants demonstrated a high consistency of polarity choices and relational similarity between font-size and loudness in TtS.

INTRODUCTION

Printed or electronic textbooks constitute the main class of documents in the domain of education. The content of a document includes mainly the text and the images. The term *text-document* refers to the textual content only of a document. Besides its content, a printed or an electronic document contains a number of presentation elements or attributes (Kouroupetroglou & Tsonos, 2008) that apply on its text content: a) design glyphs or typographic elements (i.e. visual representation of letters and characters in a specific font and style) and b) arrangement of the content on the page and the document as a whole. Presentation elements include: a) font (type, size, color, background color, etc.) and b) font-style, such as bold, italics, underline. In contrast to the rich text, the term plain text indicates text in any unique font type and size, but without font style.

Rich-text documents use writing devices which intend to highlight the important information which is in the text as well the text structure (Kintsch, & van Dijk, 1978). The term text signal has been proposed (Lorch, 1989) as the writing device that emphasizes aspects of a text's content or structure carrying semantic information over and above the content. It attempts to pre-announce or emphasize content and/or reveal content relationship (Lemari, Eyrolle, & Cellier, 2006). Headings or titles in text-documents are considered as signals (Lorch, Chen, & Lemari, 2012).

The font-size of a chapter heading in a textbook is usually bigger than the font-size of the main text. It is also bigger than the one of the subheadings, if there are any in the text (Steno & Retti, 2003). Thus, the different font-sizes used in textbooks aim mainly to differentiate the headings and the footnotes from the main text as well as to accomplish a hierarchy among different level headings. It has been experimentally proved that font-size and font-type influence memorization and comprehension. Both of them help the readers in the creation of an organizational framework for the coding of a text (Smith & Sera, 1992; Spyridakis, 1989a), which facilitates them to maintain and to recover information (Spyridakis, 1989a; Spyridakis, 1989b; Sanchez, Lorch & Lorch, 2001). Font-size represents the main characteristic that reveals the text macrostructure to the reader (Kintsch & Yarbrough, 1982).

Nowadays, Text-to-Speech (TtS) software systems (Freitas & Kouroupetroglou, 2008), combined with screen readers (Asakawa & Leporini, 2009), constitute the main alternative for the blind and partially sighted students to access the content of schoolbooks and other educational resources. Moreover, TtS represent an emerging technology in teaching and learning of non-disabled students (Rughooputh & Santally, 2009), as well as in the practice of Universal Design for Learning (Gordon, Proctor & Dalton, 2012).

Most of the current TtS systems treat the content as plain text and do not support an effective audio provision of the presentation elements or text signals of a document, such as font (type, size, color, background color, etc.) and font style (Fellbaum & Kouroupetroglou, 2008). As a consequence, blind students or learners who use the audio channel only to access educational content through TtS lose important information incorporated in a rich text document and they are at a disadvantage respectively to the typical readers who use their vision to access the same content. Recently, there has been an effort towards Document-to-Audio (DtA) synthesis (Kouroupetroglou, 2013), which essentially represent the next generation in TtS. DtA supports the efficient acoustic representation of typography and text formatting through modelling the prosodic parameters of the synthesized speech signal.

In the present study our effort is to discover relation similarities between dimensions that are perceptible through different senses, such as the font-size and the loudness in TtS. The semantics of quantitative dimensions, such as the size and the loudness, are often conceptualized with the significance of the named *poles*. One pole is the positive or differently *more* and the movement towards this pole is augmentative, while the other pole is negative or *less* and the movement in this direction is decreasing (Holyoak, 1978). A number of researchers have supported that the origin of these poles is found in our sensory system (Boring, 1993; Marks, Hammeal & Bomstein, 1987; Treisman & Gormican, 1988). According to Stevens (1957), the quantitative dimensions, such as size and loudness, have certain unitary and well-defined psychophysics attributes which he calls "prothetic". These psychophysics attributes are reflected in a common sensory physiology. Thus, size and loudness have a common sensory physiology and consequently the directions of psychological decrease or increase are specified by the physiology of the sensory system (Smith & Sera, 1992).

Smith & Sera (1992) found that the children that are older than 2 years begin to correspond to the dimensions of size and loudness with the significance *more* or *less*. Another factor which contributes to this cross-correlation is the natural structure of the world. The bigger objects tend to make more noise than the smaller ones. The results of Smith & Sera research show that the young children know this cross-correlation, which helps them to combine the notions *big* and *loud*. Under this view, Smith & Sera, propose that, apart from the predetermined sensory structure, other factors also exist, such as the language and the physical structure of the world that converge to this correlation.

As there is a lack of research on the relational similarity between font-size and loudness in TtS used by blind and sighted persons, this work aims to contribute towards this achievement, particularly in the domain of education. Our main hypothesis is that congenitally blind and sighted students in primary and secondary education perceive a linear relational similarity between the dimensions font-size and voice loudness in TtS.

METHOD

Of the 51 Greek students who took part in the study, 25 were congenitally blind or students who became blind during the first years of their life and the other 26 were sighted. Among them, 29 were females (15 blind and 14 sighted) and 22 males. The sighted students ranged in age from 10 to 17 and the blind students ranged in age from 10 to 18. In particular, 15 of the 25 blind participants were students of the secondary education and 10 of the primary education. Moreover, 16 of the sighted participants were students of the secondary education and 10 of the primary education.

In order to select the values for the font-size, we statistically analyzed a corpus of 72 textbooks (a mixture of all subjects): 36 of them use by the K-12 schools in Greece and 36 in the English language used by the K-12 American Community School in Athens, Greece. The results indicate that the text size has a range between 6 pts and 72 pts. With a view to design an experiment with duration of less than half an hour, the selected font-sizes were 12pts, 32pts and 56pts. In order to achieve a linear relationship between the two modalities ($y=0.6566x+45.7$, y=loudness in db and x=font-size in pts), we selected the values of loudness to be 53db, 68db and 82db.

The acoustic stimuli generated with the Document-to-Audio (DtA) software tool (Xydas, et all, 2005) along with the DEMOSTHéNES Greek TtS system (Xydas, & Kouroupetroglou, 2001). The optical stimuli were generated as MS-Power Point presentations. Blind participants had access to the presentation using the JAWS screen reader software Ver. 11.0 (Freedom Scientific, 2014).

All participants used a laptop (Acer Aspire 1314LC) with a screen of 15΄΄ (resolution 1024X768), MS-Windows Vista operating system and semi-open headphones (AKG K-66).

There were two tasks:

A) Percept-to-Percept (P-P) task: This task investigates how a value on one dimension maps onto the polar values of another dimension; the participant is presented with an exemplar stimulus of a value on one dimension and is asked which of two choice stimuli values on the other dimension is like the exemplar. Initially, a visual exemplar of the biggest or the smallest value of one dimension (e.g., the font size, as it appears in Figure 1) was presented to each sighted participant and the researcher raised the following question: «If you have to read the word you see, in such a way so that your schoolmates perceive the specific font size the word is written, which voice you will select between the two you will hear? » For a blind participant, the researcher first explained that:

«Contrary to the Braille writing, the letters used in the texts for sighted people do not always have the same size. Sometimes they are bigger than regular, such as in headings, and at other times they are smaller, as in footnotes. Imagine a Braille cell to be bigger in the titles than in the main text, and be smaller in the footnotes». Then, he raised the following question: «If you had to read the word "pyramid" that is written with very big letters, in such a way, so that your schoolmates perceive the font size the word is written, which voice you will select between the two you will hear? » Next, the P-P task was repeated with the addition of an intermediary value in each dimension. In the case the exemplar was auditory and the stimuli of choice visual (Figure 2), the researcher raised the following question to the sighted participants: «Which of the words you see matches better the voice you will hear?». For a blind participant, the researcher first explained that: «The word triangle has been written twice: first with very big letters and second with very small letters» and then he asked: «Which of the two words matches better the voice that you will hear?»

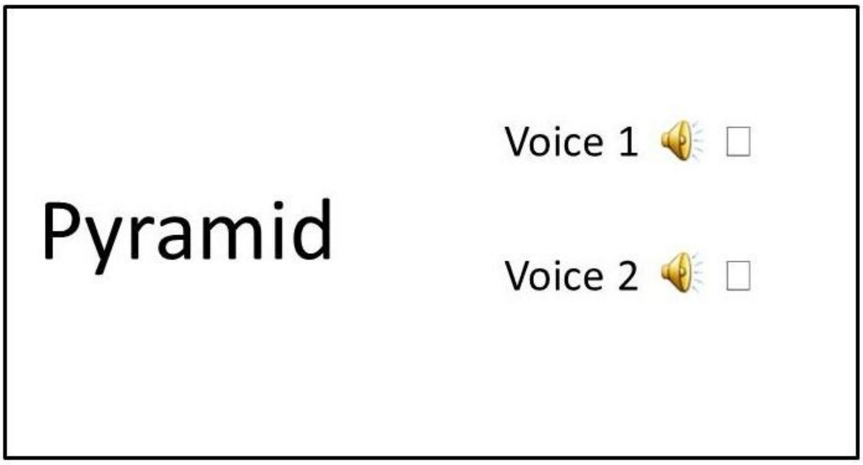

Figure 1: Visual exemplar with auditory choice stimuli.

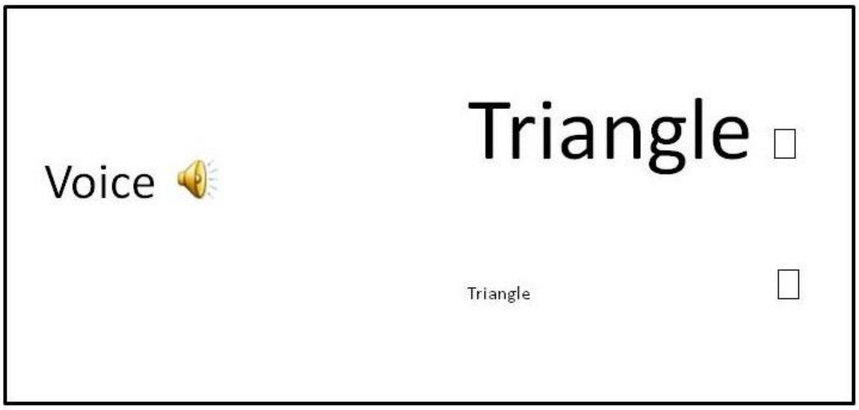

Figure 2: Auditory exemplar with visual choice stimuli.

B) Word-to-Percept (W-P) task: In this task we examine how the words on one dimension map to the perceived poles of the other dimension. A word expressing a value in one dimension was presented and the participants were asked to correlate it with the suitable choice stimulus of the other dimension. In the case of the blind participants, the researcher asked: a) for the word exemplar with the notion of font-size: «With which of the three voices that you will hear can you match the phrase big letters? » and b) for the word exemplar with the notion of voice loudness «We have written the word triangle three times. The first time with very big letters, the second with normal letters and the third time with very small. With which of these words can you match the word loud?» The words used as exemplars had always the same font-size.

In the Percept-to-Percept task (P-P) 20 questions were presented in total. In 10 of them, the exemplar was visual, i.e., a word with one of the three font-sizes and the choice stimuli were auditory. In the rest 10 questions the opposite was applied. In the Word-to-Percept task (W-P), 20 questions were also presented in total. In 10 of them, the exemplar was a word or phrase described one of the three font-sizes and the choice stimuli were auditory (3 voices with different loudness). In the rest 10 questions the exemplar was a word or phrase describing one of the three voice loudness and the choice stimuli were visual (words with three font-sizes). In both the P-P and the W-P tasks the questions were presented in a random order for each participant. Between the two tasks, there was a time interval of four weeks. The acoustic stimuli were repeated, if a participant hesitated to answer or asked to hear them again.

The participants were asked individually in a quiet room. Each participant was sitting in front of the computer desk. Prior to the above tasks they were familiarized with all the visual and auditory stimuli.

RESULTS AND DISCUSSION

The results for the means of the consisted choices are presented in Table 1: a) for the blind participants in the P-P task M=19.56 (97.8%) and M=20 (100%) for the W-P task, and b) for the sighted participants in the P-P task M=19.96 (99.8%) and M=19.85 (99.95%) for the W-P task respectively. Thus, we observed very high rates of consistent choices between font-size and voice loudness in both conditions of the research. The mixed design ANOVA, 2(visual condition of the participants)×2 (task), did not show significant differences, either between the participants, F(1.49)=0.478 p >0.05, Partial Eta-squared = 0.01, or within the participants, F(1.49)=2.384, p >0.05, Partial Eta-squared =0.046.

Table 1: The means of the consisted choices in the P-P and W-P tasks.

Task	Participants	Mean	%	Std. Deviation	N
P-P	Blind	19.56	97.8	1.64	25
	Sighted	19.96	99.8	0.20	26
	Total	**19.76**	**98.8**	**1.16**	**51**
W-P	Blind	20.00	100.0	0.00	25
	Sighted	19.85	99.2	0.78	26
	Total	**19.92**	**99.6**	**0.56**	**51**

The means of the valid answers between the blind and the sighted participants for the polar choices (Table 2) did not present important differences. Thus, the polarity does not seem to be influenced by the visual condition of the participants. The results of the mixed ANOVA design 2(vision condition)×3(pole) showed no statistically significant differences between the groups F(1.49)=0.478 p>0.05, Partial Eta-squared=0.01, and within groups F(1.49)=1.19 p>0.05, Partial Eta-squared=0.24.

Table 2: The polar consistent choices between the blind and the sighted participants.

Pole	Participants	Mean	%	Std. Deviation	N
MORE	Blind	15.80	98.75	0.82	25
	Sighted	16.00	00.0	0.00	26
	Total	**15.90**	**99.4**	**0.57**	**51**
MEDIUM	Blind	7.96	99.5	0.20	25
	Sighted	7.92	99.0	0.39	26
	Total	**7.94**	**99.3**	**0.31**	**51**
LESS	Blind	15.80	8.75	0.82	25
	Sighted	15.88	99.3	0.43	26
	Total	**15.84**	**99.0**	**0.64**	**51**

Table 3 presents the results of the consistent choices among the students of primary and secondary education when the exemplar was a word (with different font-sizes) or a voice (with different loudness). The mixed ANOVA design 2(education level)×2(exemplar) shows that there was no significant difference either between the groups F(1.49)=2.223 p=>0.05, Partial Eta-squared=0.043, or within the groups F(1.49)=1.164 p>0.05, Partial Eta-squared=0.23.

Table 3: The consistent choices among the students of primary and secondary education.

Exemplar	Education	Mean	%	Std. Deviation	N
FONT-SIZE	Primary	20.00	100.0	0.00	21
	Secondary	19.63	98.2	1.07	30
	Total	**19.78**	**98.9**	**0.83**	**51**
VOICE LOUDNESS	Primary	20.00	100.0	0.00	21
	Secondary	19.83	99.2	0.75	30
	Total	**19.90**	**99.5**	**0.57**	**51**

CONCLUSIONS

The results of this research study confirm our initial hypothesis that all participants demonstrated a very high consistency of polarity choices and relational similarity between font-size and loudness in TtS. Moreover, the results showed that important differences do not exist between students of primary and secondary education. Thus, the same mapping between the text font-size and the voice volume in TtS can be applied in both cases. In our future work, we will investigate the mapping between the font-type (e.g. bold, italic, and bold-italic) and the prosodic parameters in TtS as perceived by sighted and blind students.

ACKNOWLEDGMENT

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Project: THALIS-University of Macedonia- “KAIKOS: Audio and Tactile Access to Knowledge for Individuals with Visual Impairments” MIS 380442.

References

Asakawa, C. & Leporini, B. (2009). Screen readers. In C. Stephanidis (Ed.) The Universal Access Handbook. Chapter 28, CRC Press, Florida, USA, ISBN: 9780805862805

Bierswisch, M. (1970). On semantics. In Lyons J. (Ed.), New horizons in linguistics. London: Penguin. 164-184.

Boring, E. G. (1933). The physical dimensions of consciousness. New York: Century.

Fellbaum, K., & Kouroupetroglou, G. (2008). Principles of Electronic Speech Processing with Applications for People with Disabilities. Technology and Disability, 20(2), 55–85.

Freedom Scientific (2014). JAWS, http://www.freedomscientific.com/Products/Blindness/Jaws

Freitas, D., & Kouroupetroglou, G. (2008). Speech Technologies for Blind and Low Vision Persons. Technology and Disability, 20(2), 135-156.

Gordon, D., Proctor, C. P., & Dalton, B. (2012). Reading strategy instruction, universal design for learning, and digital texts: Examples of an integrated approach. In T.E. Hall, A. Meyer, & D.H. Rose (Eds.). Universal design for learning in the classroom: Practical applications (pp. 25-37). New York: Guilford Press.

Holyoak, K. (1978). Comparative judgments with numerical reference points. Cognitive Psychology, 10, 203-243.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363–394.

Kintsch, W., & Yarbrough, C.J. (1982). Role of rhetorical structure in text comprehension. Journal of Educational Psychology, 74, 828-834.

Kouroupetroglou, G. (2013). Incorporating Typographic, Logical and Layout Knowledge of Documents into Text-to-Speech. In Encarnacao, P. et al. (Eds.), Assistive Technology: from Research to Practice. Vol. 33, pp. 708–713. Amsterdam: IOS Press.

Kouroupetroglou, G., & Tsonos, D. (2008). Multimodal Accessibility of Documents. In S. Pinder (Ed.) Advances in Human-Computer Interaction (pp. 451–470). Vienna: I-Tech Education and Publishing. DOI: 10.5772/5916

Lemari, J., Eyrolle, H., & Cellier, J. M. (2006). Visual signals in text comprehension: How to restore them when oralizing a text via a speech synthesis? Computers in Human Behavior, 22(6), 1096–1115. doi:10.1016/j.chb.2006.02.013

Lorch, R.F. (1989). Text-Signaling Devices and Their Effects on Reading and Memory Processes. Educational Psychology Review, 1(3), 209–234. DOI:10.1007/BF01320135

Lorch, R.F., Chen, H.T., & Lemari, J. (2012). Communicating Headings and Preview Sentences in Text and Speech. Journal of Experimental Psychology: Applied, 18(3), 265–276. DO:10.1037/a0029547 PMID:22866682

Marks, L.E., Hammeal, R.J., & Bomstein, M.H. (1987). Perceiving similarity and comprehending metaphor. Monographs of the Society for Research in Child Development, 51, (I, Serial No. 215).

Rughooputh, S., & Santally, M. (2009). Integrating Text-to-Speech Software into Pedagogically Sound Teaching and Learning Scenarios. Educational Technology Research and Development, 57(1), 131-145.

Sanchez, R.P., Lorch, E.P., & Lorch, R.F. (2001). Effects of Headings on Text Processing Strategies. Contemporary Educational Psychology, 26, 418-428.

Sax, L. (2010). Sex Differences in Hearing. Implications for best practice in the classroom. Advances in Gender and Education, 2:13-21.

Smith, L., & Sera, M. (1992). A developmental analysis of the polar structure of dimensions. Cognitive Psychology. 24, 99-142.

Spyridakis, J.H. (1989a). Signaling effects: A Review of the Research, part I. Journal of Technical Writing and Communication, 19(3), 227-240.

Spyridakis, J.H. (1989b). Signaling effects: A Review of the Research, part II. Journal of Technical Writing and Communication, 19(4), 395-415.

Stehno, B., & Retti, G. (2003). Modeling the logical structure of books and journals using augmented transition network grammars. Journal of Documentation, 59(1), 69-83.

Stevens, S.S. (1957). On the psychophysical law. Psychological Review, 64, 153-181.

Treisman, A. & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. Psychological Review, 95, 1548.

Xydas, G., & Kouroupetroglou, G. (2001). The DEMOSTHéNES Speech Composer, In *Proceedings of the 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW4)*, International Speech

Communication Association, Perthshire, Scotland, August 29 - September 1, 2001, pp. 167-172, DOI 10.13140/2.1.4992.0968

Xydas, G., Argyropoulos, V., Karakosta, T., & Kouroupetroglou, G. (2005). An experimental approach in recognizing synthesized auditory components in a non-visual interaction with documents. In Proceedings of the 11th International Conference on Human-Computer Interaction (HCII2005), Las Vegas, Vol. 3, pp. 411-420. Lawrence Erlbaum Associates, Inc (ISBN 0-8058-5807-5) DOI 10.13140/2.1.4566.1122