An Experimental Approach in Recognizing Synthesized Auditory Components in a Non-Visual Interaction with Documents

Gerasimos Xydas[1]*, Vasilios Argyropoulos*[2]*,*
Theodora Karakosta[1] *and Georgios Kouroupetroglou*[1]

[1]University of Athens, Department of Informatics and Telecommunications
{gxydas, koupe}@di.uoa.gr

[2]University of Thessaly, Department of Special Education
vassargi@sed.uth.gr

Abstract

The auditory formation of visual-oriented documents is a process that enables the delivery of a more representative acoustic image of documents via speech interfaces. We have set up an experimental environment for conducting a series of complex psycho-acoustic experiments to evaluate users' performance in recognizing synthesized auditory components that represent visual structures. For our purposes, we exploit an open, XML-based platform that drives a Voice Browser and transforms documents' visual meta-information to speech synthesis markup formalism. A user-friendly graphical interface allows the investigator to build acoustic variants of documents. First, a source de-compilation method builds a logical layer that abstractly classifies visual meta-information. Then, the investigator can define distinctive sound fonts in a free way by assigning combined prosodic parameters and non-speech audio sounds to the logical elements. Four blind and four sighted student subjects were asked about their distinctive ability in response to several controlled versions of auditory components, as well as their opinion on the quality of the selected features. The results provided evidence to discuss the general hypothesis that the mean scores for blind and sighted students in recognizing differentiated auditory components are equal. Also, we extracted some preliminary data for the evaluation of the appropriateness of several sound fonts in the auditory representation of visual components in cases of WWW documents.

1 Introduction

Web documents (e.g. HTML) are mainly concerned with visual modality, though recommendations are being developed (mainly by the W3C) for enabling other modalities to be delivered as well through the Web (Burnett et al., 2004)(Raggett et al., 2004). The visual formation of Web documents is guided by meta-data that accommodate the actual text. Examples of the utilization of meta-information by the Web Browsers in the visual domain are the optical effects of "bold" or "italics" letters, or the structured layout of tables. However, when setting such documents in auditory environments, the Text-to-Speech conversion traditionally strips out any meta-information from the source document. This results in less effective aural presentation that would be the case if the document structure was retained. For example, "bold" letters usually imply emphasis, which is not delivered, for example, through a screen reading software or a telephone Web browser.

Related works in the field of the representation of documents in auditory-only interfaces have pointed out the need for retaining part of the meta-information through out the TtS procedure. Raman in (Raman, 1992) has developed a series of systems basically for providing an audio format of complex mathematical formulas in LaTeX documents. He used non-speech audio sounds to indicate the formula and some pitch modifications in order to group elements within the formula. The importance of combining speech and non-speech signals to support the presentation of visual components and structures has shown in (Hakulinen et al., 1999), while other works focus on the prosodic variations and speaker-style changes (Shriver et al., 2000). Earcons, i.e. structured sounds (the aural counterparts of icons (Blattner et al., 1989), and Auditory Icons have been used in the human-machine interfaces (Gorny, 2000).

Sound Fonts are also another approach that utilizes prosodic modifications of synthetic speech in order to deliver visual components in a speech modality. In (Truillet et al., 2000), the use of Sound Fonts is studied for the comprehension and memorization of bold letters. Blind and sighted objects were used and two cases were examined: (a) insertion of a pitch modified phrase “in bold” before the bold words and (b) a pitch modification that applied directly on the corresponding bold words, which was finally the preferred Sound Font.

The recent efforts in the accessibility of Web documents focus on the understanding of the semantic specification of HTML structures. For example, Hidden Markov Models were trained in (Kottapally et al., 2003) in order to identify cells and headers in tables.

This work is part of a bigger effort to provide some means for the auditory specification of Web documents, thus we focus on presentation issues. We introduce an open-platform capable of (a) parsing one- and two-dimensional visual components and (b) transforming them into speech markup elements with the combination of prosodic and audio features. Since there are not enough data for the standardization of earcons and prosodic sound fonts (we use the term *auditory scripts* for both cases), we also take advantage of this platform to perform a pilot study on the acoustical effect of different auditory scripts to blind and sighted students. The scripts provide an auditory definition of document components that otherwise should produce visual effect (bold letters, italic letters and bulletin list).

In the next section we introduce the Document-to-Audio (DtA) platform and the functionality of a graphical tool for experimenting with auditory scripts. We then measure the rate of the acoustical distinction of visual components achieved by both blind and sighted students and we present their reactions on the proposed approach.

2 The Document-to-Audio (DtA) platform

We have dealt with the problem of the auditory formation of documents by introducing the e-TSA Composer (Xydas and Kouroupetroglou, 2001a). This modular architecture enabled the transformation of of documents’ meta-information to prosodic realizations and other audio mappings. That is done through successive transformations of the documents structure using XML for knowledge representation and XSTL scripts for the transformations. Thus, each document is split in the so-called *clusters* of information. A cluster consists of the textual data (e.g. the text to be spoken) and the meta-information that accommodates it (e.g. visual format). Within this process, the meta-information is mapped to auditory properties, allowing prosody modifications, voice switch and inserted audio sounds to take place in the aural formation of the document.

We have now extended the original design and the new DtA platform adds an abstract logical layer to the e-TSA architecture that holds a free of any presentation details transformation of the meta-information of the source. This layer allows the presentation of the information to user interfaces of any modality (Figure 1). Knowledge representation is now carried out by Heterogeneous Relation Graphs (Taylor et al., 2001), in a similar way as the Conceptual Graphs used in (Kottapally et al., 2003). HRG allows objects of different types to be interconnected via list or tree relations, forming bigger trees, while any item can be reached from any other one through a formalism that access relations.

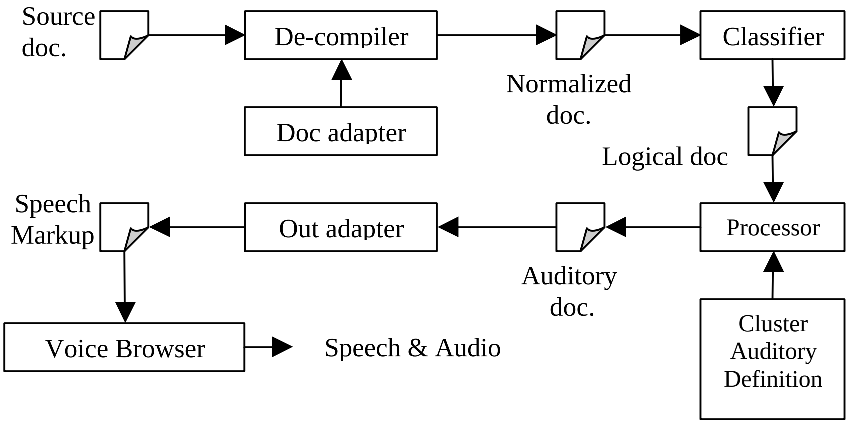
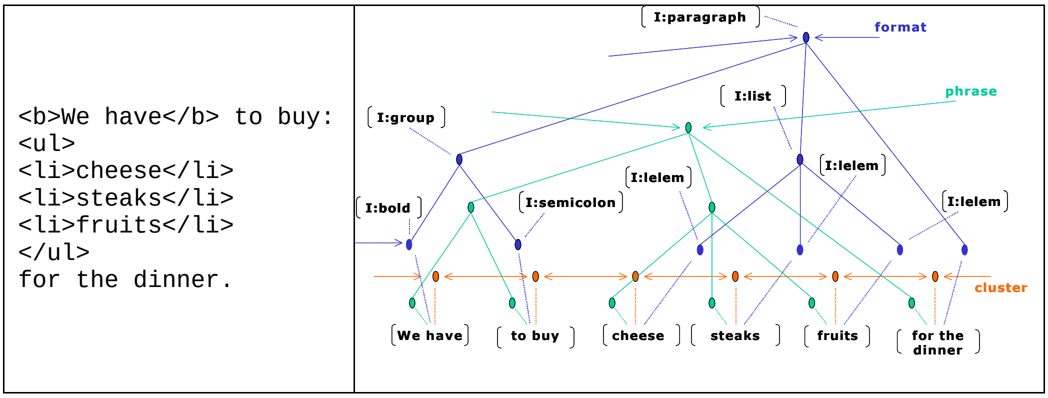


Figure 1: The DtA platform with the added logical stage in e-TSA Composer.

The DtA platform preserves the XML-based initiative of the e-TSA Composer, in the manner that a series of structured documents travels through the processing chain, as illustrated in the above figure. Thus, the source document is firstly de-compiled in order to classify meta-information into one- (e.g. font formation) and two- (e.g. tables) dimensional relations and align it with the corresponding textual data. Each pair of meta-information along with its related text is identified as a “cluster” in the source document (Figure 2). Current implementation uses XSL scripts for document de-compilation and classification.



```
<b>We have</b> to buy:
<ul>
<li>cheese</li>
<li>steaks</li>
<li>fruits</li>
</ul>
for the dinner.
```

Figure 2: an HRG representation of HTML code in the DtA. Three relations are shown: phrase, format and cluster.

The auditory transformation is then carried out by utilizing a library of Cluster Auditory Definition (CAD) scripts that define the desired prosodic behavior, as well as other audio insertions in response to the clusters’ class. Figure 3 and 4 illustrates examples of CAD scripts, using the XSL-based prototype. This results to a Speech Markup document to be fed in a Voice Brower. Voice Browsers differ from traditional text-to-speech systems in that the former are capable of parsing texts with speech and audio annotations rather than plain ones.

```
<xsl:template match = "emphasis">
<prosody pitch = ″+20%″ rate=″0.85″ volume = ″130″>
<xsl:apply-templates/>
</prosody>
```

```
<prosody pitch = ”default” rate = ”default” volume = ”default”/>
</xsl:template>
```

Figure 3: a Cluster Auditory Definition script of the prosodic sound font “emphasis”.

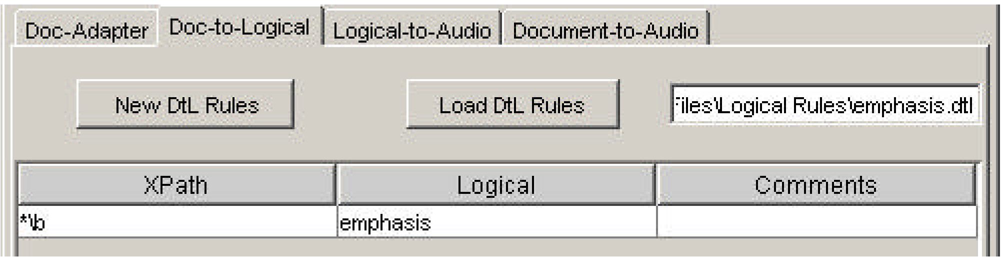
```
<xsl:template match = "ul">
<audio src = ”330Hz”/><audio src = ”440Hz”/>
<xsl:apply-templates/>
<audio src = ”440Hz”/>
<xsl:template match = "ul/li">
<audio src = ”330Hz”/> <ssml:audio src = ”220Hz”/>
<xsl:apply-templates/>
</xsl:template>
```

Figure 4: a CAD script of a bulletin’s earcon. A sequence sound of 2 tones forms an intro (played before the list), an outro tone sequence announces the end of the list, while a high tone (440Hz) points each list item.

2.1 Implementation and user interface

The DtA platform formed the basis of an experimental environment in order to perform psycho-acoustic experiments with students regarding the appropriateness of selected prosodic parameters and audio features in representing visual clusters.

As the original e-TSA Composer required an experienced XML programmer to use it, we further built on top of the DtA platform a user-friendly graphical user interface to be used by researchers not experienced with programming, as well as to facilitate the repetitive nature of the experiments: try some scripts, correct them and start over. The “Doc to Audio Tool”, written in Java, hides all the complexity of the underlying platform. The user is able to (a) create and test Prosodic Sound Fonts and Earcons and (b) to write the Cluster Auditory Definition scripts in an abstract manner. The tool is organized into four groups: the Adaptation of the input file, the Composition and Compilation of Logical rules, the Composition and Compilation of Audio rules and the Transformation of the input file, according to the selected rules, into a Speech Markup document, to be read out by the Voice Browser (Figure 5):



iles\Logical Rules\emphasis.dtl

XPath	Logical	Comments
*\b	emphasis	

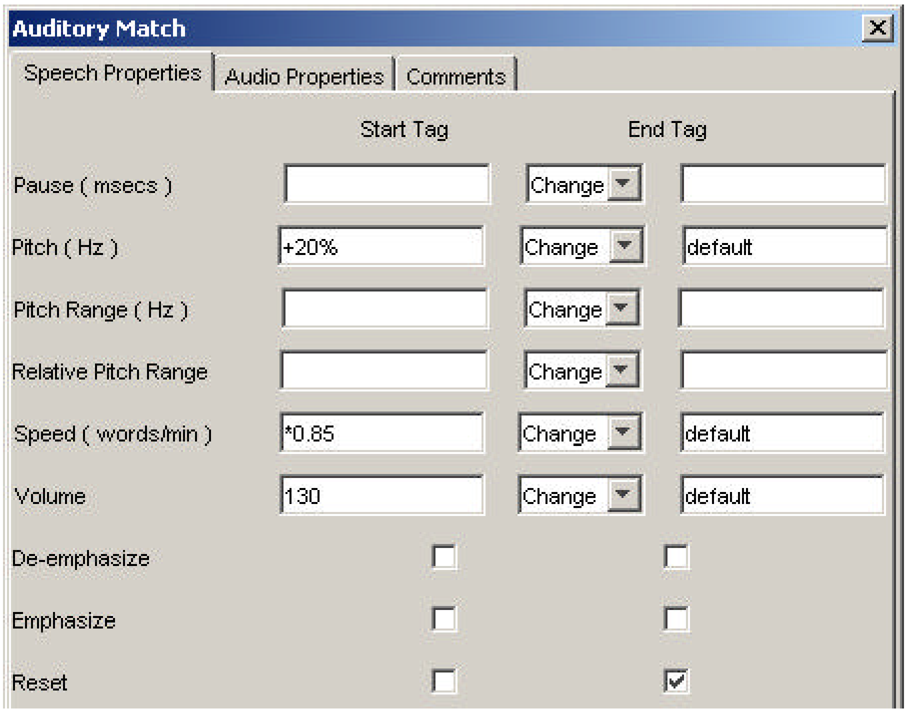
Figure 5: The 4 functionality groups of the Doc to Audio Tool. This example illustrates how the bold visual component is assigned to the logical emphasis. This is a simplified example as one can additionally define several degrees of emphasis.

In respect to the DtA platform of Figure 1, the Doc to Audio Tool consists of:

- ? he Doc-Adapter that utilizes the JTidy tool for the conversion of HTML documents to XML ones (*De-compiler*).
- The Doc-to-Logical (*Classifier*) converter that interacts with the XSLT files that deal with the formation of the logical layer. The XSLT implementation is hidden

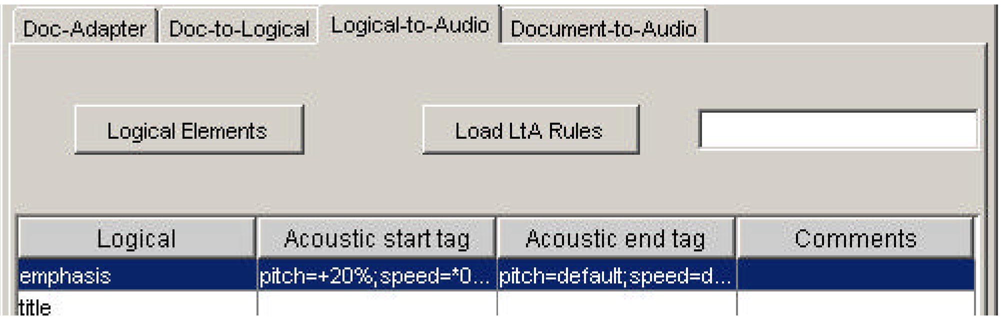
under a set of user-friendly commands, such as “map the bold element to emphasis”.

- The Logical-to-Audio converter (*Processor*) that provides a set of speech and audio controls that the user can modify in order to built auditory scripts. More specifically, the user can select prosodic attributes, such pitch, rate and volume and also generate audio files and signals using a hidden interface with the PRAAT tool (Boersma, 2001) (Figure 6). After testing the script, the user can assign it to a logical element (Figure 7).
- The last stage, the DtA, performs a batch processing of all the above-mentioned steps. The user is able to apply any stored rule for Doc-to-Logical and Logical-to-Audio transformations, generating SSML Documents that are provided to the Voice Browser (in our Greek experiments we used DEMOSTHeNES Speech Composer (Xydas and Kouroupetroglou, 2001b)).



	Start Tag		End Tag
Pause (msecs)		Change	
Pitch (Hz)	+20%	Change	default
Pitch Range (Hz)		Change	
Relative Pitch Range		Change	
Speed (words/min)	*0.85	Change	default
Volume	130	Change	default
De-emphasize	□		□
Emphasize	□		□
Reset	□		☑

Figure 6: The prosodic options for a logical element.



Logical	Acoustic start tag	Acoustic end tag	Comments
emphasis	pitch=+20%;speed=*0...	pitch=default;speed=d...	
title			

Figure 7: After testing the auditory scenario, the user can assign it to a logical element.

3 Experiments

Eight participants took part in this pilot study; four blind students (two males & two females) and four sighted ones (two males & two females) in the ages 22-25. We experimented with three visual components: (a) "bold", (b) "italic" and (c) "bulletin". These were selected within the frame of a research which has been conducted for all textbooks used in Greek high school in terms of the usage of one- and two-dimensional visual constructs. The chosen text (Stimulus Material – Aural Presentation Text- see Figure 2) was extracted from the textbook "Home economics" (high school).

The default synthetic voice that we used featured trained prosodic models (Xydas et al., 2004) and the Mbrola synthesizer (Dutoit, 1997) with the Greek diphone database gr2 (Xydas and Kouroupetroglou, 2001b). The prosodic baselines used were: pitch=110Hz, speed=140 words per minute and volume=100. In line with common practice, literature review (Truillet et al., 2000)(Brewster, 1991) (Kallinen, 2003) and internal tests within the research team we arrived at the following auditory definitions for the selected visual components (Table 1)[1].

Table 1: Qualitative auditory specification for the prosodic characteristics in all 4 versions. Earcons' auditory definition is shown in Figure 4.

Version	Bold	Italics	Bullets
1	Pitch = 132 Hz (+20%)	Speed = 161 wpm (+15%)	earcons
2	Volume = 130 (+ 30%)	Speed = 161 wpm (+ 15%)	earcons
3	Volume = 130 (+ 30%)	Pitch = 94 (-15%)	earcons

4	Pitch = 132 Hz (+20%)
	Volume = 130 (+ 30%)
	Speed = 119 wpm (-15%)
		Pitch = 94 (-15%)
		Speed = 161 wpm (+ 15%)
			earcons

Before running the tasks, the students listened to plain synthetic speech from DEMOSTHeNES in order to get used to its voice. During the experiments the participants wore stereophonic headphones to isolate all external aural components and were given 10 minutes to familiarize themselves by listening to a range of eight different levels of pitch, volume and speed. The stimulus material was read out first in a flat (plain) version, followed by the 4 alternate ones.

More analytically, after the student has heard the flat version (version 0) subsequently he/she was presented aurally version 1. After hearing version 1 the subject was invited to identify the aurally modified words or phrases, which he/she has noticed when comparing the two versions of the same stimulus material (versions 0 & 1). Both blind and sighted students were given handouts with the text in use but in plain formation (in Braille for the blind) to note any differentiation they noticed when listening to the two versions. The combination of versions 1, 2, 3 & 4 was chosen randomly.

4 Results

The results from the experiments addressed comparisons between the performances of blind and sighted students mainly on their auditory distinctions towards the neutral version and the enriched versions of DtA as an effort in evaluating different "sound fonts".

[1] These samples can be accessed via http://www.di.uoa.gr/~gxydas/doc2audio_eval1.shtml

4.1 Auditory Distinction (AD)

Firstly, we performed some measurements concerning the Auditory Distinction (AD), i.e. the ability of the students to identify the different visual components in the auditory versions (Figure 8). The following figure indicates that blind students performed more accurately than their sighted peers in recognizing differentiated auditory components (2nd, 3rd and 4th version – see Table 1). The biggest difference was noticed in the 4th version (84,09% average for the blind toward 70,45% for the sighted). In general, the performances of all students were at high level of distinction.

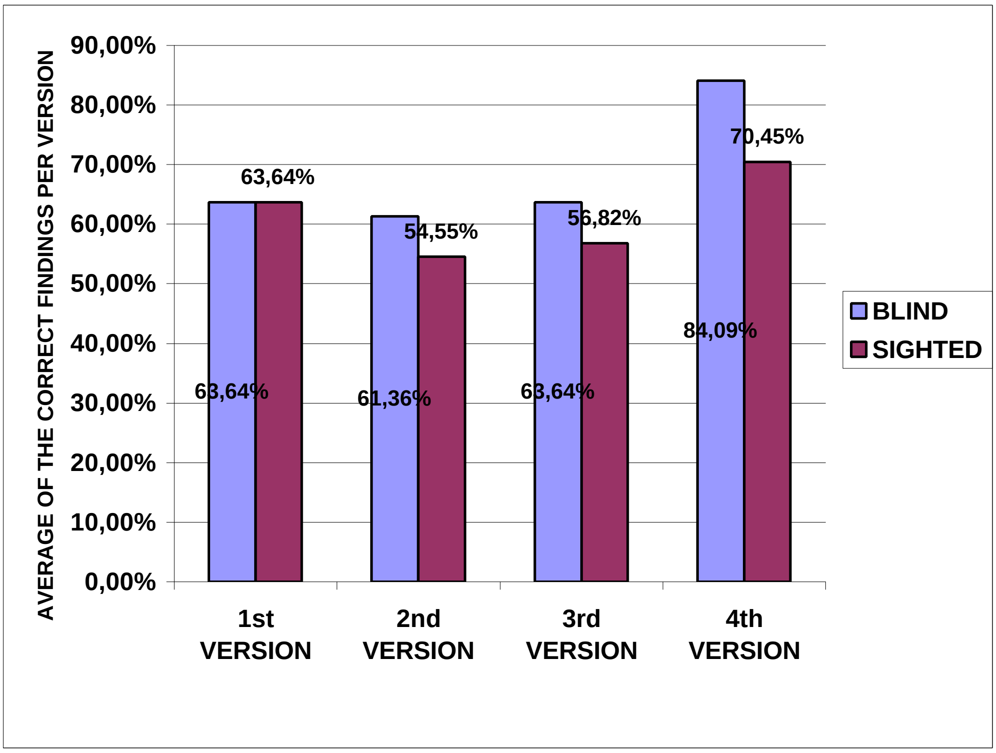


Figure 8: Comparative table in recognizing differentiated auditory components per version.

Table 2 contains statistical information (although 8 students are close to being too few to motivate statistical tests). The total number of the prosodic components in the stimulus material was 11 (5 for the bold, 5 for the italic and 1 for the bullet). The statistical analysis with the means and standard deviations drew some inferences in terms of the comparisons between the performances of blind and sighted students.

Table 2: Mean and stdev scores in correct distinction of the 11 visual components. (BS = blind students, SS = sighted students).

Version	1	2	3	4
BS Mean	7	6, 75	7	9,25
BS Std. Deviation	2,94	4, 27	2,45	0,96
SS Mean	7	6	6,25	7,75
SS Std. Deviation	1,63	0,82	0,96	1,26

It is worth mentioning the significant difference between the values of std. deviation in the version 2 between the performances of blind and sighted students. The attributes of the prosodic characteristics in version 2 were modified only in terms of speed and volume and not at all in pitch (Table 1). In total, variances are bigger in the cases of the blind students than in the sighted ones. This may be happened due to the fact that blind students were users of a different, formant-based speech synthesizer for Greek (along with screen reader software) for a long time. The range of the standard deviation is decreased significantly in the complex version 4.

4.2 Qualitative Auditory Distinction (AD)

Secondly, we performed measurements concerning the Qualitative Auditory Distinction (QAD) i.e. the ability of the students to identify the quality of the recognized differentiated auditory components in terms of changes in pitch, volume or speed (Figure 9).

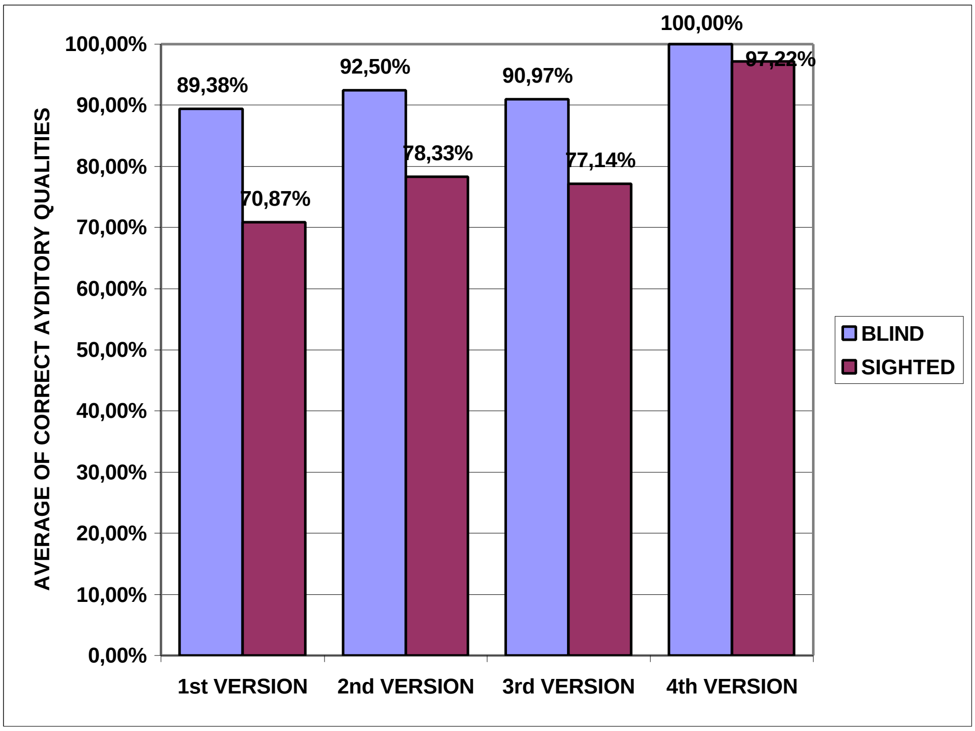


Figure 9: Comparative table in recognizing qualities in auditory differentiated components per version.

Figure 9 shows that blind students had higher distinctiveness compared with that of the sighted with respect to the qualitative clarification of the differentiated prosodic components. Version 4 (complex) bears special interest because of the 100% of accurate opinions of the blind students when they classified the qualities of the differentiated auditory prosodic elements. Table 3 tabulates mean and std deviations.

Table 3: Results from the QAD of the correct differentiated auditory prosodic components. (BS = blind student, SS = sighted student).

Version	1	2	3	4
BS mean	6,75	6,25	6,75	8,75
BS Std. Deviation	2,87	3,86	2,87	0,96
SS mean	5,25	4,75	4,75	7,50
SS Std. Deviation	2,87	1,71	1,50	1,00

In total, the absolute values of the variance are bigger in the performances of the blind students than in the sighted ones (not so much as in Table 3) whereas, the values of the means converge. It is worth mentioning here that the biggest value of std. deviation is when blind students perform in version 2 (std. deviation=3, 86). The fact that this version does not have any change in the pitch may bring some interesting issues for further research in the area of psycho-acoustics.

5 Conclusions

To sum up the results, nearly all students (apart from two) agreed that version 3 (volume=+ 30%, pitch=-15%, compound earcon) appeared to be more natural, more distinctive. Nevertheless, if we look at Tables 2 & 3 the students performed higher in version 4 rather than in version 3. According to them version 4 was a bit extreme to their ears when “overstretching” the qualities of pitch or volume. The usage of earcons for the starting and ending points of the list (bullets) made all students enthusiastic and could identify at once the presence of the prosodic component “bullet” in the stimulus material. The version that did not motivate the students was version 2 (volume=+30%, speed=+15%). This version was the only one that did not contain any modification of the pitch. Also it is very interesting to mention that all participants stressed out that they faced difficulties to realize modifications in speed of 15% of the default value. On the contrary, they were well disposed toward modifications of pitch in conjunction with modifications of volume. These characteristics are embedded in version 3 and it seems to be closer to the natural way of spoken language.

Additionally, the performances of the blind and the sighted students had neither a statistical difference in the AD (Auditory Distinction) nor in the QAD (Qualitative Auditory Distinction).

The purpose of these experiments was to evaluate mappings of visual components to an arbitrary set of acoustical presentations. While extreme prosodic renderings sounded unnatural, they also led to a high level of auditory distinction of the document components. The findings of this pilot study provided a rough assessment for the determination and specific auditory behavior of the three selected visual components (bold, italic & bullet) leading to an integrated and enriched auditory web accessibility. These issues should be addressed in future research to arrive at robust conclusions expanding the number of participants and testing more than three visual structures.

6 Acknowledgments

Part of this work has been funded under a postdoctoral grant scholarship to the second of the authors from the Hellenic State Scholarships Foundation (I.K.Y.).

7 References

Blattner, M.M., Sumikawa, D.A. and Greenberg, R.M. (1989). Earcons and Icons: Their Structure and Common Design Principles. *Human Computer Interaction*, Vol 4, pp. 11-14.

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International* 5(9/10), pp. 341-345.

Brewster, S. (1991). Providing a model for the use of sound in user interfaces. Department of Computer Science University of York, Heslington, pp 20-24, 35-40.

Burnett C.D., Walker R. M. and Hunt A, (2004). Speech Synthesis Markup Language (SSML) Version 1.0. *W3C Recommendation*, http://www.w3.org/TR/speech-synthesis

Dutoit, T. (1997). An Introduction to Text-to-Speech Synthesis. *Kluwer Academic Publishers*.

Gorny, P. (2000). Typographic semantics of Webpages Accessible for Visual Impaired Users, Mapping Layout and Interaction Objects to an Auditory Interaction Space. *International Conference on Computer Helping with Special Needs*, pp. 17-21.

Hakulinen, J., Turunen, M. and Raiha, K. (1999). The Use of Prosodic Features to Help Users Extract Information from Structured Elements in Spoken Dialogue Systems. *Proceedings of ESCA Tutorial and Research Workshop on Dialogue and Prosody*, Eindhoven, The Netherlands, pp. 65-70.

Kallinen, K. (2003). Using sounds to present and manage information in computers. Center for Knowledge and Innovation Research Helsinki School of Economics, Finland.

Kottapally, K., Ngo, C., Reddy, R. Pontelli, E., Son, T.C. and Gillan, D. (2003). Towards the Creation of Accessibility Agents for Non-visual Navigation of the Web. *Proceedings of the ACM Conference on Universal Usability*, Vancouver, Canada, pp. 134-141.

Raggett, D., Glazman, D. and Santambrogio, C. (2004). CSS3 Speech Module. W3C Working Draft, http://www.w3.org/TR/css3-speech/

Raman, T.V. (1992). An Audio View of (LA)TEX Documents. *TUGboat*, 13, Number 3, Proceedigns of the 1992 Anuual Meeting, pp. 372-379.

Shriver, S., Black, A. and Rosenfeld, R. (2000). Audio Signals in Speech Interfaces. *Proceedings of International Conference on Spoken Language Processing*, Beijing, China.

Taylor, P., Black, W. A. and Caley, R. (2001). Heterogeneous Relation Graphs as a Mechanism for Representing Linguistic Information. *Speech Communication*, vol.33, pp.153-174.

Truillet, P., Oriola, B., Nespoulous, J.L. and Vigoroux, N. (2000). Effect of Sound Fonts in an Aural Presentation. *6th ERCIM Workshop*, UI4ALL, pp. 135-144.

Xydas, G. and Kouroupetroglou, G. (2001a). Augmented Auditory Representation of e-Texts for Text-to-Speech Systems. *Lecture Notes in Artificial Intelligence*, Springer-Verlag Berlin Heidelberg, Vol. 2166, pp. 134-141.

Xydas, G. and Kouroupetroglou, G. (2001b). The DEMOSTHeNES Speech Composer, *Proceedings of the 4th ISCA Tutorial and Workshop on Speech Synthesis* (SSW4), pp. 167-172.

Xydas, G., Spiliotopoulos, D. and Kouroupetroglou, G. (2004). Modeling Prosodic Structures in Linguistically Enriched Environments. *Lecture Notes in Artificial Intelligence*, Springer-Verlag Berlin Heidelberg, Vol 3206, pp. 521-528.