Proc. Conference ACOUSTICS 2004, Thessalonica, 27-28 Sept. 2004

An Open Platform for Conducting Psycho-Acoustic Experiments in the Auditory Representation of Web 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

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία περιγράφει ένα ανοικτό, βασισμένο σε XML, γραφικό περιβάλλον για την πραγματοποίηση σύνθετων πειραμάτων ακουστικής απόδοσης της πληροφορίας οπτικής μορφοποίησης εγγράφων, όπως αυτά δημοσιεύονται στον παγκόσμιο ιστό. Ο χρήστης έχει την δυνατότητα (α) να απο-μεταγλωττίσει την πηγαία δομή των εγγράφων του Web δημιουργώντας ένα λογικό επίπεδο αναπαράστασης της οπτικής μετα-πληροφορίας που φέρουν και (β) να δημιουργήσει αυθαίρετες διακριτές ηχο-σειρές αντιστοίχησής της με συνδυασμό προσωδιακών παραμέτρων και ηχητικών, μη-φωνητικών, στοιχείων Χρησιμοποιώντας αυτή την πλατφόρμα, εκτελέσαμε ένα σύνολο από ψυχο-ακουστικά πειράματα με τυφλούς και βλέποντες για την αξιολόγηση της καταλληλότητας και της απόδοσης διαφόρων ηχο-σειρών στην ακουστική αναπαράσταση των οπτικών στοιχείων σε έγγραφα τύπου HTML. Τα αποτελέσματα δείχνουν ότι καταρχήν οι ηχο-σειρές που επιλέξαμε για τα "bold", "italics" και "bullets" γίνονται διακριτές σε ποσοστό μέχρι και 84.1%.

ABSTRACT

This work presents an open, XML-based, graphical environment, for conducting complex experiments in the field of the auditory representation of visual format of documents like the ones that are published in the World Wide Web. We allow the user to (a) de-compile the source structure of Web documents generating a logical layer that abstractly represents of the visual meta-information the document carry and (b) to create distinctive sound fonts in an arbitrary way that map of the meta-information to prosodic parameters and non-speech audio elements. Using this platform, we performed a set of psycho-acoustic experiments with blind and sighted students for the evaluation of the appropriateness and performance of several sound fonts in the auditory representation of visual components in cases of HTML

documents. The results show that the sound fonts we chose for "bold", "italics" and "bullets" are being distinctive up to 84.1% of the cases.

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 [1][2]. The visual formation of Web documents comes from data that accommodate the textual data, i.e. meta-information. Examples of utilization of the meta-information by Web Browsers in the visual domain are the effects caused by "bold" or "italics" letters, or the structured layout of tables. However, when setting such documents in an auditory environment, the aural presentation traditionally occurs by stripping out any meta-information from the source document prior to the Text-to-Speech (TtS) conversion. This result in less effective 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 [3] has developed a series of systems basically for providing an audio format of complex mathematical formulas in (LA)TEX 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 [4], while other works focus on the prosodic variations and speaker-style changes [5]. Earcons, i.e. structured sounds (the aural counterparts of icons [6]), and Auditory Icons have been used in the human-machine interfaces [7].

Sound Fonts are also another approach that utilizes prosodic modifications of synthetic speech in order to deliver visual components in a speech modality. In [8], 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.

This work is part of a bigger effort to provide some means for the auditory accessibility of the Web. We present 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 earcons and prosodic sound fonts. We then illustrate the usefulness of DtA by measuring the rate of the

acoustical distinction of visual components from the students used in the experiments.

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 [9]. This enabled the transformation of any kind of meta-information inside documents to prosodic realizations or audio insertions. We have now slightly modified the original architecture, as shown in Figure 1 and we have accommodated it with an easy-to-use graphical user interface. The DtA platform adds an abstract logical layer to e-TSA that is free of any presentation details. From this layer we are able to drive user interfaces of any modality.

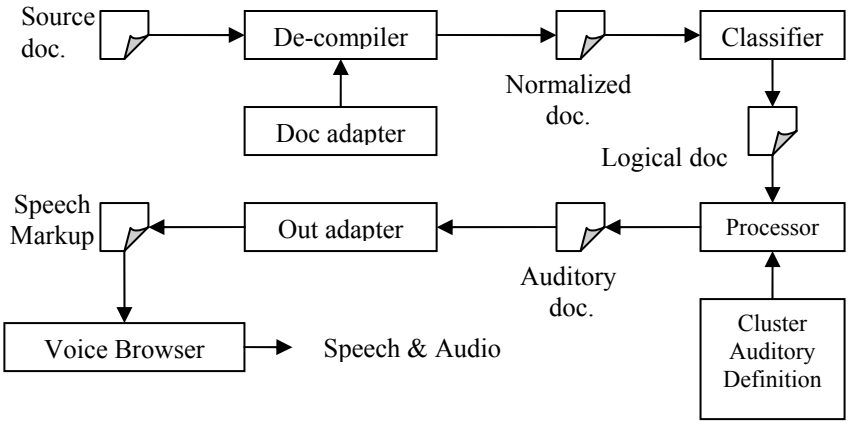

Figure 1: The DtA platform with the added logical stage in e-TSA Composer.

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).

Για να διατηρήσει την καλή κατάσταση της υγείας του ο ανθρώπινος οργανισμός, *όπως αναφέρεται στο μάθημα της οικιακής οικονομίας*, χρειάζεται να παίρνει καθημερινά **θρεπτικά συστατικά**. Αυτά χρησιμοποιούνται ως δομικά συστατικά των **ιστών**, για την παραγωγή ενέργειας, *δηλαδή σαν καύσιμη ύλη για την κίνηση και την εργασία* και τέλος, για τη ρύθμιση όλων των πολύπλοκων λειτουργιών που επιτελεί ο οργανισμός. **Ισορροπημένη διατροφή σημαίνει ποικιλία τροφίμων**. Τα θρεπτικά συστατικά λαμβάνονται με την τροφή. Τα βασικά από αυτά είναι

Figure 2: Examples of "bold" and "italic" clusters.

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 audio insertions in response to the clusters' class. Figure 3 and 4 illustrates

examples of CAD scripts, using an XSL-based implementation. This transformation is depicted in a Speech Markup document that is being used to drive 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.

This 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 built on top of the DtA platform an 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 Document-to-Audio Application is a Java-based user interface that hides all the complexity of the DtA platform and allows the user to (a) create and test Prosodic Sound Fonts and Earcons and (b) to write the Cluster Auditory Definition scripts in an abstract manner. The system is organized into four groups: the Adaption 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:

- The Doc-Adapter utilizes the JTidy tool for the conversion of HTML documents to XML ones.

- The Doc-to-Logical part 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 function 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 using a hidden interface with the PRAAT [15] tool to generate signals. After testing the script, the user can assign it to a logical element.
- At the last stage, the DtA, the user is able to apply any stored rule for Doc-to-Logical and Logical-to-Audio transformations, generating an SSML Document [1] that is fed to the Voice Browser (in our Greek experiments we used DEMOSTHeNES Speech Composer [10]).

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) was extracted from the book "Home economics" (high school).

The default synthetic voice that we used featured trained prosodic models [11] and the Mbrola synthesizer [12] with the Greek diphone database gr2 [10]. The prosodic baselines used were: pitch=110Hz, speed=140 words per minute and volume=100. In line with common practice, literature review [8, 13, 14] and internal tests within the research team we arrived at the following auditory definitions for the selected visual components (Table 1).

Table 1: Qualitative auditory specification for the prosodic characteristics in all 4 versions. Earcons' 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. All subjects 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. First, 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 (Table 2). In general, the performances of all students were at high level of distinction and the fact that blind students surpassed their peers might be happened due to the experience of the former in listening to a big variety of pre-recorded study materials in conjunction with speech synthesizers. Table 3 contains some statistical information (although 8 students are close to being too few to motivate statistical tests). The total number of the visual components in the stimulus material was 11 (5 "bold", 5 "italic" and 1 bulletin list).

Table 2: Average of correct auditory differentiations per version.

Version	Blind	Sighted
1	63.6%	63.6%
2	61.4%	54.6%
3	63.6%	56.8%
4	84.1%	70.5%

Table 3: 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 to mention 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 the version 2 were adapted only in 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.

Table 4: Qualitative auditory distinctions per version.

Version	Blind	Sighted
1	89.4%	70.9%
2	92.5%	78.3%
3	91.0%	77.1%
4	100.0%	97.2%

The second part of the analysis refers to the specification of the quality of the recognised differentiated auditory components (Qualitative Auditory Distinction – QAD) (i.e. changes in the pitch, volume or speed). Table 4 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 5 tabulates mean and std deviations.

Table 5: 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 to mention 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.

Summing up the results, nearly all students (apart from two) agreed that version 3 (volume=+ 30%, pitch=-15%, compound earcons) appeared to be more natural, more distinctive. Nevertheless, if we look at Tables 3 & 5 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 which did not motivate the students was version 2 (volume=+30%, speed=+15%). This version was the only one which 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 conceptualize 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.

5. Conclusions

We presented an open-platform, namely DtA, for transforming visual oriented documents to speech and audio. An easy-to-use graphical user interface facilitates the performance of psycho-acoustics experiments. The purpose of these experiments was to evaluate mappings of each logical layer element to an arbitrary set of

acoustical presentations. While extreme prosodic renderings sound unnatural, they also lead 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 behaviour of the three selected visual components (bold, italic & bullet) leading to an integrated and enriched auditory web accessibility.

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

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

[2] "Cascading Style Sheets", http://www.w3.org/Style/CSS/"

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

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

[5] Shriver, S., Black, A. and Rosenfeld, R. "Audio Signals in Speech Interfaces", In Proceedings of International Conference on Spoken Language Processing, Beijing, China, 2000

[6] Blattner, M.M., Sumikawa, D.A. and Greenberg, R.M. "Earcons and Icons: Their Structure and Common Design Principles", in Human Computer Interaction, 1989, Vol 4, pp. 11-14.

[7] Gorny, P. "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, 2000, pp. 17-21.

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

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

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

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

[12] Dutoit, T. "An Introduction to Text-to-Speech Synthesis", Kluwer Academic Publishers, 1997

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

[14] Kallinen, K. “Using sounds to present and manage information in computers”, Center for Knowledge and Innovation Research Helsinki School of Economics, Finland, 2003

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