RENDERING WEB-CONTENT TEXT SIGNALS THROUGH ADVANCED TEXT-TO-SPEECH

Georgios Kouroupetroglou
National and Kapodistrian University of Athens, Greece
koupe@di.uoa.gr

Dimitrios Tsonos
National and Kapodistrian University of Athens, Greece
dtsonos@di.uoa.gr

ABSTRACT

Web-content accessibility plays an important role in internet based learning, including e-learning, distance learning and mobile learning. Current Text-to-Speech systems, commonly used for web-content accessibility, do not include an effective and standard acoustic provision of the semantics and the cognitive aspects of the visual (such as typography) and non-visual (such as logical structure) knowledge embedded in the web-content. In this work we first introduce an appropriate architecture for the structure of web text documents. Then, we analyze the web-content text signals along with their semantics. By following a design-for-all (i.e. universal) methodology, we present the Document-to-Audio approach for the universal rendering of web-content text signals to auditory modality. Finally, for the case of font size in the typographic layer, we present the results of two quantitative approaches (direct mapping and emotional-based mapping) for rendering web-content text signals through advanced Text-to-Speech systems.

Keyword: Text-to-Speech, Web-content accessibility, Expressive Speech Synthesis, Document-to-Audio

1. INTRODUCTION

Web-content is the information in a web page or a web application. It includes text, images, sounds, videos as well as electronic documents in the form of common file formats (such as .doc, .ppt, .pdf). Web-content plays an important role in internet based learning, including e-learning, distance learning and mobile learning. Web browsers, media players, assistive technologies and other user agents constitute common ways for the internet user to receive and interact with the web-content. Software or hardware Assistive technologies (AT) allow persons with a physical, mental or learning disability to interact with a computer device and thus with the web-content. Common AT includes screen readers, Text-to-Speech applications, software screen magnifiers, on-screen keyboards and speech recognition.

Web-content accessibility enables an electronic document to be used effectively, efficiently and satisfactory by more users in more situations or context of use. It concerns all the aspects of document functionality that includes browsing, searching, navigation and reading [1]. Basic accessibility of documents is provided by

Text-to-Speech (TtS) systems [2]. Accessibility of web-content is very important not only for the print-disabled readers, i.e. those with vision impairment (blindness, low vision, color blindness, dyschromatopsia, etc.), a learning disability (including dyslexia) or a motor disability (such as loss of dexterity that prevents the physical handling of a document), but also for those with an occasional or situational "disability". A typical example is a user who requires accessing the web-content while driving a car, i.e. while his/her eyes and hands are busy.

Currently, TtS systems with synthetic speech of very high quality exist not only for desktop and laptop computers but also for smart phones and tablets. Modern TtS systems have the possibility to set the prosody parameters of the synthetic speech (pitch, volume or speech rate) or to select among different voices or include sounds (e.g. earcons, auditory icons or spearcons) to render some semantic, structural or other information of a rich text web document, such as the typographic attributes (bold, italics, underline, etc.). But these mappings cover a limited number of document's attributes, do not support all the kinds of the structural schema [3], in most cases have to be customized by the user, they support only tagged information (i.e. they do not include a knowledge-based extraction of non-content semantics) and finally they are proprietary or not universal.

Thus, current TtS systems do not include an effective and standard acoustic provision of the semantics and the cognitive aspects of the visual (such as typography) and non-visual (such as logical structure) knowledge embedded in the web-content. Thereby, a blind person or one with busy eyes (i.e. a car driver) misses important information when they access a rich text web document through TtS.

In this work, first we introduce an appropriate architecture for the structure of web text documents. Then we analyze the web-content text signals along with their semantics. Then, by following a design-for-all methodology, we present the Document-to-Audio approach for the universal rendering of web-content text signals to auditory modality.

2. WEB-CONTENT TEXT SIGNALS

1.1 Structure of web text documents

The content of a web document includes mainly text and images (i.e. figures, drawings, graphs, pictures, charts, diagrams, schematic illustrations, maps, photos, etc.). With the term text-document we refer to the textual content only of web-content. A text-document contains a number of presentation elements or attributes that arrange the content on the page and apply design glyphs or typographic elements, i.e., visual representation of letters and characters in a specific font and style (Figure 1). For example, the title of a chapter can be recognized by placing it at the top of the page and in larger font size than the body of the text. Moreover, text color or the bold font style can be used to indicate emphasis in a specific part of a text document.

The elements of a text document can be classified in three layers [4]:

- **Logical layer**: it associates content with structural elements such as headings, titles/subtitles, chapters, paragraphs, tables, lists, footnotes, and appendices.
- **Layout layer**: it associates content with architectural elements relating to the arrangement on pages and areas within pages, such as margins, columns, alignment and orientation.

- **Typography layer**: it includes font (type, size, color, background color, etc.) and font style such as bold, italics, underline. In contrast to the rich text, the term plain text indicates text in a unique font type and size, but without font style.

The above three layers are complementary and not independent. Typography can be applied to both the logical and the layout layers of a document. Moreover, typography can be applied to the main body of the text directly, for example, a word in bold can be used either for the introduction of a new term or to indicate a person's name. Also, a heading can be arranged in the center of a line (layout layer).

Rich-text web-content (Figure 2) is a text document that preserves all its presentation elements. In contrast, a plain-text web document ignores the presentation elemnts. Most of the current TtS systems essentially use web-content as plain text.

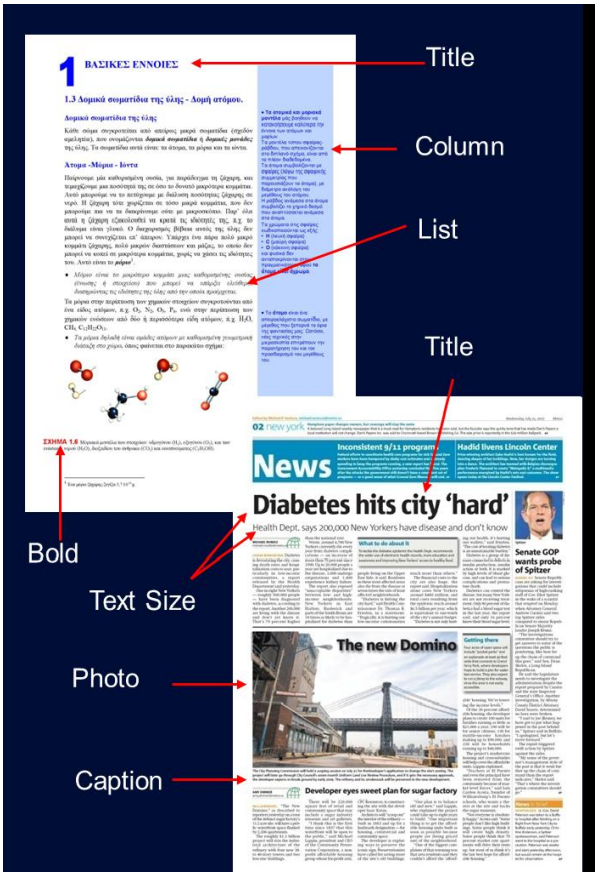


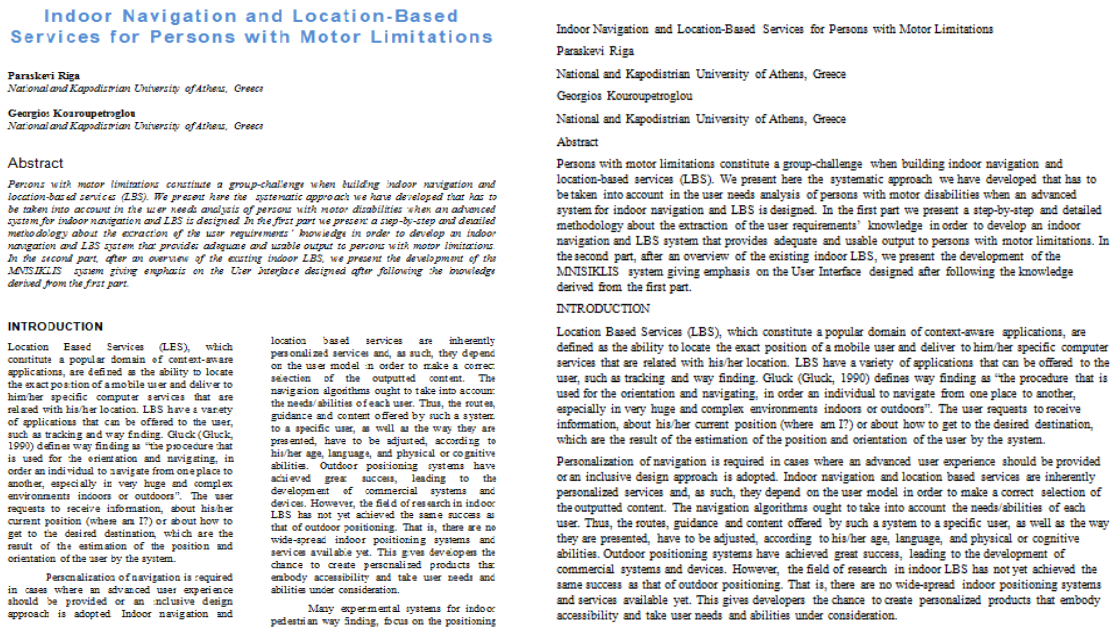
Figure 1. Presentation elements of web-documents

1.2 Text-signals in web-content

The term "signal" is introduced as "*the writing device that emphasize aspects of a text's content or structure without adding to the content of the text*" [5]. It attempts to pre-announce or emphasize content and/or reveal content relationship [6-10]. The title or heading typographic cues are considered as signals. Also, "*input enhancement*" is an operation whereby the saliency of linguistic features is augmented through e.g. textual enhancement for visual input (i.e. bold) and phonological manipulations for aural input (i.e. oral repetition) [11].

All the text signalling devices, either mentioned as signals or layers: a) share the goal for directing the reader's attention during reading, b) facilitate specific cognitive process occurring during reading, c) ultimate comprehension of text information, d) may influence memory on text and e) direct selective access between and within texts [5].

The primary goal of utilizing text signal or presentation elements in text documents is to distinguish parts of the text and to create a well-formed presentation of the content in order, for instance, to augment the reading performance or attract the reader.



Indoor Navigation and Location-Based Services for Persons with Motor Limitations

Paraskevi Riga
National and Kapodistrian University of Athens, Greece

Georgios Kouroupetroglou
National and Kapodistrian University of Athens, Greece

Abstract

Persons with motor limitations constitute a group-challenge when building indoor navigation and location-based services (LBS). We present here the systematic approach we have developed that has to be taken into account in the user needs analysis of persons with motor disabilities when an advanced system for indoor navigation and LBS is designed. In the first part we present a step-by-step and detailed methodology about the extraction of the user requirements' knowledge in order to develop an indoor navigation and LBS system that provides adequate and usable output to persons with motor limitations. In the second part, after an overview of the existing indoor LBS, we present the development of the MNISIKLIS system giving emphasis on the User Interface designed after following the knowledge derived from the first part.

INTRODUCTION

Location Based Services (LBS), which constitute a popular domain of context-aware applications, are defined as the ability to locate the exact position of a mobile user and deliver to him/her specific computer services that are related with his/her location. LBS have a variety of applications that can be offered to the user, such as tracking and way finding. Gluck (Gluck, 1990) defines way finding as "the procedure that is used for the orientation and navigating, in order an individual to navigate from one place to another, especially in very huge and complex environments indoors or outdoors". The user requests to receive information, about his/her current position (where am I?) or about how to get to the desired destination, which are the result of the estimation of the position and orientation of the user by the system.

Personalization of navigation is required in cases where an advanced user experience should be provided or an inclusive design approach is adopted. Indoor navigation and location based services are inherently personalized services and, as such, they depend on the user model in order to make a correct selection of the outputted content. The navigation algorithms ought to take into account the needs/abilities of each user. Thus, the routes, guidance and content offered by such a system to a specific user, as well as the way they are presented, have to be adjusted, according to his/her age, language, and physical or cognitive abilities. Outdoor positioning systems have achieved great success, leading to the development of commercial systems and devices. However, the field of research in indoor LBS has not yet achieved the same success as that of outdoor positioning. That is, there are no wide-spread indoor positioning systems and services available yet. This gives developers the chance to create personalized products that embody accessibility and take user needs and abilities under consideration.

Many experimental systems for indoor pedestrian way finding, focus on the positioning

Indoor Navigation and Location-Based Services for Persons with Motor Limitations

Paraskevi Riga

National and Kapodistrian University of Athens, Greece

Georgios Kouroupetroglou

National and Kapodistrian University of Athens, Greece

Abstract

Persons with motor limitations constitute a group-challenge when building indoor navigation and location-based services (LBS). We present here the systematic approach we have developed that has to be taken into account in the user needs analysis of persons with motor disabilities when an advanced system for indoor navigation and LBS is designed. In the first part we present a step-by-step and detailed methodology about the extraction of the user requirements' knowledge in order to develop an indoor navigation and LBS system that provides adequate and usable output to persons with motor limitations. In the second part, after an overview of the existing indoor LBS, we present the development of the MNISIKLIS system giving emphasis on the User Interface designed after following the knowledge derived from the first part.

INTRODUCTION

Location Based Services (LBS), which constitute a popular domain of context-aware applications, are defined as the ability to locate the exact position of a mobile user and deliver to him/her specific computer services that are related with his/her location. LBS have a variety of applications that can be offered to the user, such as tracking and way finding. Gluck (Gluck, 1990) defines way finding as "the procedure that is used for the orientation and navigating, in order an individual to navigate from one place to another, especially in very huge and complex environments indoors or outdoors". The user requests to receive information, about his/her current position (where am I?) or about how to get to the desired destination, which are the result of the estimation of the position and orientation of the user by the system.

Personalization of navigation is required in cases where an advanced user experience should be provided or an inclusive design approach is adopted. Indoor navigation and location based services are inherently personalized services and, as such, they depend on the user model in order to make a correct selection of the outputted content. The navigation algorithms ought to take into account the needs/abilities of each user. Thus, the routes, guidance and content offered by such a system to a specific user, as well as the way they are presented, have to be adjusted, according to his/her age, language, and physical or cognitive abilities. Outdoor positioning systems have achieved great success, leading to the development of commercial systems and devices. However, the field of research in indoor LBS has not yet achieved the same success as that of outdoor positioning. That is, there are no wide-spread indoor positioning systems and services available yet. This gives developers the chance to create personalized products that embody accessibility and take user needs and abilities under consideration.

Figure 2. Rich-text web-content (left) and plain-text web-content (right)

Authors of web-content use typography and layout in a specific way, e.g. there are "strict" typographic rules for the documents to be e-published in a scientific journal. But in e-newspapers and e-books the page designer (or the page manager), and not the author, has the primary responsibility for applying the typography and layout layers. Traditional factors that play a leading role in documents formation and presentation include readability (the gauge that measures how easily words, phrases, and blocks of copy can be read) and legibility (the measure of easiness to distinguish one letter from another in a particular typeface). More recent factors include visual aesthetics and accessibility. On the other hand, it seems that there is a plethora of semantics in applying the typographic layer. For example, in contrast to the tags introduced by the W3C for the bold and italics font styles [12], we have identified [13] eight different "labels" that the readers seem to use most frequently in order to semantically characterize text in "bold" and "italics (a total of 2.927 entities, of which 1.866 were occurrences of "bold" and 1.061 of "italics" were manually labelled in a corpus of 2.000 articles of a Greek e-newspaper):

- **Emphasis**: A word or phrase that is considered significant and needs to be stressed out.
- **Important / Salient**: A word or phrase, which is near or is part of a piece of information that is considered important and should be noticeable. A word or phrase that "catches the eye" of the reader.

- **Basic Block**: A block of text, which introduces or summarizes the main content of the article.
- **Quotation**: A piece of text corresponding to a fragment of written or oral expression of a person other than the writer of the article.
- **Note**: A piece of text serving at providing additional information or explanation related to a part or to the whole of the article.
- **Title**: A piece of text corresponding to the name of a movie, play, book and so on or the title of a newspaper, television channel or journal.
- **List / Numeration Category**: A word of phrase that is part of a list or a numeration and serves as a new "instance" indicator.
- **Interview / Dialogue**: A piece of text that is part of an interview (the question or the answer) or that corresponds to a dialogue between two persons.

A few "logical" labels, such as "subtitle" or "footnote" were also mentioned by some readers, but these were not considered to be "semantic" labels and were therefore not of value for the purposes of this study.

3. Rendering Document Signals to Auditory Modality

We present here our design-for-all based efforts towards rendering web-content text signals to auditory modality. We have introduced the relative term Document-to-Audio (DtA) synthesis, which essentially constitutes the next generation Text-to-Speech systems that support the efficient acoustic rendition of the presentation elements of a text document (i.e. the typography and logical layers of a document).

First, we have applied analytics to large corpora of text documents in both English and Greek language (selected as a representative minor language with non-Latin alphabet) in order to extract knowledge for the logical, layout and typography layers embedded in text. Specifically, 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 in the case of text size has a range between 6 pts and 72 pts.

For the signals at the logical and layout layers in detail, we first extract the relative information from the document. Then, we model the parameters of the synthesized speech signal by: (a) combining alternative text insertion in the document text stream, (b) altering the prosody, (c) switching between voices, and/or (d) inserting non-speech audio (like earcons, auditory icons or spearcons) in the waveform stream, according to the class of metadata extracted from the document.

For the typography layer two approaches for rendering document signals to auditory modality are incorporated in DtA: a) direct mapping and b) emotional-based mapping.

3.1 Direct Mapping

Based on the relation similarity, each typographic cue is directly mapped into a respective acoustic cue. The principle of relational similarity explores two physical quantities with magnitudes that humans perceive by different senses in an analogous way. For example, the font size of a text and the volume of the speech signal when the text is vocalized comprise relational similarity in the case we perceive the change of their magnitudes in a proportional way.

51 Greek students, age 10-17 years, 29 females (15 blind and 14 sighted) and 22 males (10 blind and 12 sighted) were participating in the experiments.

The results show the following relation between the loudness (y) in db and font-size (x) in pts:

$$y=0.6566x+45.7 \quad (1)$$

3.2 Emotional-based Mapping

First, we measure and model the way emotional states are induced in the reader by the document typographic visual characteristics. Then, following a number of psychoacoustic experiments, we determine analogous prosodic cues that produce the same emotional states to the listener when he/she hears the acoustic rendition of the document by a TtS system. Recently [14], we have developed an automated reader's emotional state extraction process derived by the typographic cues, as well as an appropriate modeling of reader's emotional state response on document's typographic elements [15, 16] combined with the following expressive speech synthesis model [17]:

$$S = A * \Delta E + I \quad (2)$$

where:

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_n \end{bmatrix} \qquad A = \begin{bmatrix} a_1^P & a_1^A & a_1^D \\ a_2^P & a_2^A & a_2^D \\ \dots & \dots & \dots \\ a_n^P & a_n^A & a_n^D \end{bmatrix} \qquad I = \begin{bmatrix} I_1 \\ I_2 \\ \dots \\ I_n \end{bmatrix} \qquad \Delta E = \begin{bmatrix} P \\ A \\ D \end{bmatrix}$$

and:

S: the speech (prosodic) characteristics matrix

A: the factors matrix

I: the intercept (offset) matrix

P: Pleasure in [−100, 100]

A: Arousal in [−100, 100]

D: Dominance in [−100, 100]

We can model the PAD relationship with text size using the following equation along with the data in Table 1:

$$f_{\{d\}}(s) = B_3^{\{d\}} \cdot s^3 + B_2^{\{d\}} \cdot s^2 + B_1^{\{d\}} \cdot s + Intercept^{\{d\}} \quad (3)$$

where:

d: P, A or D
P: Pleasure in [-1, 1]
A: Arousal in [-1, 1]
D: Dominance in [-1, 1]
B_1, B_2 and B_3 are the polynomial coefficients
s: font size in px.

Table 1. Polynomial coefficients for modeling PAD with text size using equation (3) along with their standard errors

	Pleasure		**Arousal**		**Dominance**	
		SE		SE		SE
Intercept	-2.45287	0.30771	3.58459	0.73588	2.85307	0.77138
$\mathbf{B_1}$	0.24942	0.03299	-0.59264	0.12133	-0.41265	0.12828
$\mathbf{B_2}$	-0.00523	0.00077	0.02866	0.00632	0.01631	0.0067
$\mathbf{B_3}$	0	0	-0.000427	0.0001029	-0.00019986	0.0001093

4. CONCLUSIONS

In this paper we introduced an appropriate architecture for the structure of web text documents. Then, we analyzed the web-content text signals along with their semantics. By following a design-for-all (i.e. universal) methodology, we presented the Document-to-Audio approach for the universal rendering of web-content text signals to auditory modality. Finally, for the case of font size in typography, we have presented the results of two quantitative approaches (direct mapping and emotional-based mapping) for rendering the web-content text signals through advanced Text-to-Speech systems.

5. ACKNOWLEDGEMENTS

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) under the Research Funding Project: "THALIS-University of Macedonia- KAIKOS: Audio and Tactile Access to Knowledge for Individuals with Visual Impairments", MIS 380442.

6. REFERENCES

[1] D.S. Doermann, E. Rivlin and A. Rosenfeld, The function of documents. *Image Vision Computing*, 16(11), 799-814, 1998.

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

[3] M. Olhausen and C. Roller, The operation of text structure and content schemata in isolation and in interaction. *Reading Research Quarterly*, 23, 70–87, 1988.

[4] G. Kouroupetroglou and D. Tsonos, Multimodal Accessibility of Documents. In S. Pinder (Ed.), *Advances in Human-Computer Interaction* (pp. 451-470). Vienna: I-Tech Education and Publishing, 2008.

[5] R.F. Lorch, Text-Signaling Devices and Their Effects on Reading and Memory Processes. *Educational Psychology Review*, 1, 209–234, 1989.

[6] J.H. Spyridakis, Signaling effects: A review of the research—Part I. *Journal of Technical Writing and Communication*, 19, 227–240, 1989.

[7] J. Lemarié, H. Eyrolle and J.M. Cellier, Visual signals in text comprehension: How to restore them when oralizing a text via a speech synthesis? *Computers in Human Behavior*, 22, 1096–1115, 2006.
[8] J. Lemarié, R.F. Lorch, H. Eyrolle and J. Virbel, SARA: A Text-Based and Reader-Based Theory of Text Signaling. *Educational Psychologist,* 43, 27–48, 2008.
[9] R.F. Lorch, J. Lemarié and R.A. Grant, Signaling hierarchical and sequential organization in expository text. *Scientific Studies of Reading*, 15, 267–284, 2011.
[10] R.F. Lorch, H.T. Chen, and J. Lemarié, Communicating headings and preview sentences in text and speech. *Journal of Experimental Psychology: Applied*, 18, 265–276, 2012.
[11] Z.H. Han, E.S. Park and C. Combs, Textual Enhancement of Input: Issues and Possibilities, *Applied Linguistics*, 29, 597–618, 2008.
[12] World Wide Web Consortium (W3C) http://www.w3.org/International/ questions/qa-b-and-i-tags
[13] F. Fourli-Kartsouni, K. Slavakis, G. Kouroupetroglou, and S. Theodoridis, A Bayesian Network Approach to Semantic Labelling of Text Formatting in XML Corpora of Documents, *Lecture Notes in Computer Science*, 4556, 299–308, 2007.
[14] G. Kouroupetroglou, D. Tsonos, and E. Vlahos: DocEmoX, A System for the Typography-Derived Emotional Annotation of Documents. *Lecture Notes in Computer Science*, 5616, 550–558, 2009.
[15] D. Tsonos and G. Kouroupetroglou, Modeling reader's emotional state response on document's typographic elements. *Advances in Human-Computer Interaction 2011*, Article ID 206983, 1–18, 2011.
[16] D. Tsonos, G. Kouroupetroglou and D. Deligiorgi, Regression Modeling of Reader's Emotions Induced by Font Based Text Signals. *Lecture Notes in Computer Science*, 8010, 434-443, 2013.
[17] Schröder, M.: Expressing degree of activation in synthetic speech. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4), 1128–1136, 2006.