In Human-Computer Interaction, Theory and Practice Vol. 2, Ed. C. Stephanidis and J. Jacko, Lawrence Erlbaum Associates, 2003, pp 806-810, Proceedings of the 10th International Conference on Human - Computer Interaction (HCII2003), Crete, Greece, June 22-27 2003

Modelling Emphatic Events from Non-Speech Aware Documents in Speech Based User Interfaces

Gerasimos Xydas, Dimitris Spiliotopoulos and Georgios Kouroupetroglou

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
Tel: +30 210 7275305, Fax: +30 210 6018677
{gxydas, dspiliot, koupe}@di.uoa.gr

Abstract

Most of the every day documents we come across have been composed without any information of how to be rendered in a speech-based user interface. As a result, visual formations that might imply emphasis are being ignored by Text-to-Speech systems or text-adapting applications (screen readers) and furthermore, complex structures, such as tables, are usually being vocalized in a rough linearized form, which leads to a confusing provision of information. In this work we accommodate both cases, by altering segments of emphasis in the content text, leaving a prosodic space for the vocalization of meta-information as well. We present a model for locating emphatic events and assigning to them a custom prosodic behaviour. Events are being divided in implicit and explicit ones. We concluded that the latter requires insertions of text to the linear form of structures in order to be properly realized. A script-based framework (e-TSA Composer) that supports the manipulation of prosodic elements in response of specific meta-information has been used. Finally, a model of table vocalization using our approach shows the significant improvement of the information provision compared to commercial applications.

1 Introduction

The majority of the electronic documents currently composed and viewed are non-speech aware, in the sense that they do not contain information about how to be appropriately vocalized by Text-to-Speech (TtS) systems. This is more conceivable when elements of visual structures have to be spoken in cases of speech-based user interfaces. Even tools for visual impaired people (e.g. screen readers) fail sometimes to deliver a meaningful speech representation of visual structures. Commercial systems do not properly construct a speech format of tables, as cells are parsed in a row, and character formation (bold, italics) are also ignored thought it should cause some prosodic alteration. This problem propagates to all cases of auditory-only interfaces (e.g. telephone Web access, directory services).

The Aural Cascaded Style Sheets (W3C, Aural style sheets) (Lilley & Raman, 1999) is a recommendation of the World Wide Web Consortium (W3C) that concerns the transfer of speech information (mainly prosodic) along with documents. In our work we deal with cases of, by any means, non-speech aware documents. Attempts to deal with the problem of the speech generation of documents have been also made in the past. Raman has developed a system to provide an audio format of (LA)TEX documents, focusing on the vocalisation of complex mathematical formulas (Raman, 1992). To achieve this, he assigned non-speech sounds (to indicate formulas) and

prosodic features (to group elements in formulas) to math meta-information. W3C now provides recommendations on speech formatting of mathematics (Ausbrooks et al., 2002) (Kowaliw, 2001).

Realising emphatic events in the speech format of documents serves an augmented and usually meaningful auditory representation of them. Emphasis is a use of language to mark importance or significance, through either intensity of expression or linguistic features such as stress and intonation. Here we focus on the exploitation of speech emphasis in order to achieve an augmented auditory representation of documents with visual meta-information. Section 2 presents the modelling of the so-called *emphatic events*. An application of the model for table vocalization is shown in section 3, while a sort discussion is followed.

2 Modelling emphatic events

In this work we are interested in modelling emphatic events in documents to be used in speech-based user interfaces, as this way we manage to accentuate and, thus, distinguish the actual information (text), while we are able to vocally represent the visual format by using non-emphatic speech elements. Therefore, in order to properly vocalize visual structures, the hierarchy that they represent should be also retained in their speech format. Hierarchies might be in one out of two forms: *list* and *tree*. The first one is almost straightforward to vocalize as all the emphatic events occur in a sequence, without affecting each other. An example of a list is given in Figure 1. This figure presents a Heterogeneous Relation Graph (Taylor et al., 2001) with two relations: the phrase, which carries syntax information about the text on the left, and the cluster, which carries visual meta-information, as has been described in the e-TSA Composer (Xydas & Kouroupetroglou, 2001a), (Xydas & Kouroupetroglou, 2001b). Usually, TtS systems or even screen reader applications parse the phrase relation, hiding any structural or visual information provided by the cluster one (which represents the visual format of the document).

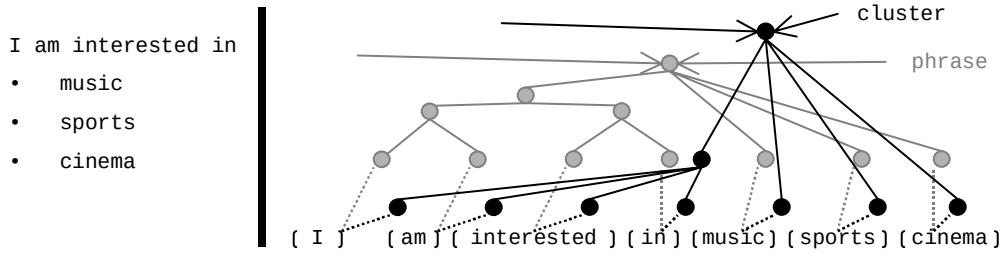

Figure 1

The tree form can be very complicated in vocalization, as it should be clear to the user that the spoken text is part of a specific hierarchy. An example of a tree form is shown in Figure 2. Testing the table on the left with state of the art commercial systems, we got two cases for speech:
A' case : "Table with four rows three columns. Model CC HP. 2000 1800 120 1990 2000 90. 1998 1900 130 table end"
B' case : "Model CC HP 2000 1800 120 1990 2000 90 1998 1900 130."
This however, leads to a misunderstanding of the content of the table.

These examples indicate the need for a more efficient handling of visual structures in Text-to-Speech process. We deal with this by introducing segments of emphasis in the text. We call the

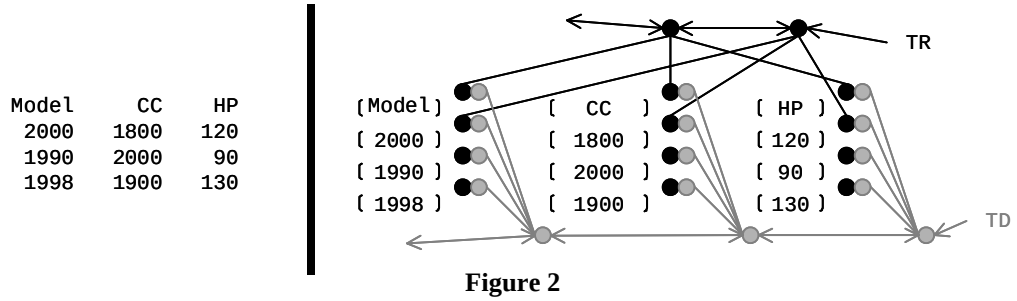

Figure 2

locations and the types of emphasis inside documents *emphatic events*, and these can be divided into two major classes:

- *Explicit events*, which are usually denoted by character special formation (e.g. bold, italics) where the level of emphasis depends on the actual format.
- *Implicit events*, which need to be identified and accessed from the special structures of the document for the structure meaning to be conveyed fully and correctly.

2.1 Explicit emphatic events

Explicit emphatic events are encountered in list form and should accentuate the corresponding cluster of text where they occur. They are conveyed by using text formation such as bold and italics. According to W3C (W3C, Information type elements) (W3C, Document structure) (W3C, 1999), italics is used to denote emphasis, whereas bold is used to denote strong emphasis. Two phrase elements, `<em>`, which indicates emphasis, and `<strong>`, which indicates stronger emphasis, are generally presented by visual agents as italics and bold text respectively. For non-speech aware documents there may be several ways to show explicit emphasis since more than the above may be needed. We classified the cases of explicit emphasis in HTML as:

- 1 or low (italics)
- 5 or medium (bold)
- 10 or high (italics & bold)
- or a value between 1-10 in cases of letters size or other formation.

2.2 Implicit emphatic events

Implicit events are encountered in either list or tree forms. In case of tree, we are using emphasis to denote the text content of structures, by distinguishing text from structural information that should also be vocalized. Implicit emphasis can be identified from structural objects inside documents, for example a row of cells in a table, bulleting, paragraph marking in conjunction with certain headings or capital lettering. In such cases emphatic events can be modelled by processing the structures individually. For example, in case of bullets it is generally required to emphasize the starting word or words of each bullet, and return to normal speech after a comma, if any exists. Nested bulleting reveals a hierarchy that should be taken into consideration, varying the levels of emphasis between the levels of nested bullets (Pitt & Edwards, 1997). Tables can also contain levels of hierarchy, but even when they don't, their complexity is still very high.

2.3 Emphatic events and prosody definition

To identify and classify visual formats of the source document, we use an XSLT-based HTML adapter in the e-TSA framework. This allows us to build the HRG presented in Figure 1 and Figure 2 and combine a hierarchy of visual directives with the traditional linguistic processing of the text. For the proper vocalisation of the documents, anything that is followed by structural

meta-information is marked to be emphasised during synthesis, while inserted text representing the structural meta-information is rendered de-emphasised.

In speech, emphasis is delivered through prosody, by raising the tone, making a stressed syllable longer and increasing the loudness. Alternatively changing the prosodical characteristics of function words against content words also emphasizes certain point in sentences. The way that emphasis should be realised relies on the preferences of the user. The e-TSA framework provides a custom pool of Cluster Auditory Definitions (CAD scripts) that can vary the prosodic behaviour of the system, depending on the type of the emphatic event.

3 Vocalizing tables

One of the most common document types, which the on-line community uses in an every day basis, is HTML, which provides visualization meta-information about the text data. We model here the vocalization of one of its most common and quite complex structures; the table. Special recommendations to promote accessibility containing guidelines on how to make the web content accessible to people with disabilities are also provided by the W3C (Chisholm, Vanderheiden & Jacobs, 1999). According to these, the use of `<TH>` (for headers) and `<TD>` (for data cells) is mandatory. The use of `<THEAD>`, `<TFOOT>`, and `<TBODY>` to group rows and `<COL>` and `<COLGROUP>` to group columns is also required to associate data and header cells. This experiment also assumes (according to the guidelines as well) that tables are not used for layout purposes, unless it makes sense when linearised.

Our approach provides a scalable three way modelling of tables in non-speech aware documents, depending on which of the recommendations are present, as well as the content of the tables. In each case, an example of a table that inherits from the one of Figure 2 is provided.

<u>High compliance</u>
Any accessibility oriented text provided is being exploited. The summary for the table is uttered first. That informs the hearer of the information that is going to follow; by all means it is a title. If additional text is provided on the `<TH>` elements it is treated as the starting text for each sentence. Then the value of the corresponding `<TD>` is added. That way one sentence is constructed for each row, and text and cell values added column by column. This would be an ideal situation, however it is not common. The values contained in the data cells are marked as the emphatic parts of the sentences, with high level of emphasis. Example: “This table provides information about cars. Model ***2000***, CC ***1800***, HP ***120***. Model ***1990***, CC ***2000***, HP ***90***. Model ***1998***, CC ***1900***, HP ***130***.”

<u>Medium compliance</u>
If the above accessibility oriented text is not provided, the utterances are constructed by alternating the headers and the data values for each row connecting using phrasal patterns like: “header *has* data” or “*for* header *the value is* data”. In this case, each pair of header-data is uttered separately. This approach sometimes fails when the tables contain nested tables or type of values with which the added generic text results in unintelligible meaning. Data values are assigned to high level of emphasis, while headers are also emphasized to distinguish from the inserted phrasal pattern. Example: “For *Model* the value is ***2000***, for *CC* the value is ***1800*** and for *HP* the value is ***120***. For *Model* the value is ***1990***, for *CC* the value is ***2000*** and for *HP* the value is ***90***. For *Model* the value is ***1998***, for *CC* the value is ***1900*** and for *HP* the value is ***130***”

<u>Low compliance</u>

The most generic approach is to model each row as a list of header-data pairs marking the data cell values with high level of emphasis. Example: "Model ***2000***, CC ***1800***, HP ***120***. Model ***1990***, CC ***2000***, HP ***90***. Model ***1998***, CC ***1900***, HP ***130***."

4 Conclusions

We presented a methodology for providing a more understandable speech format of visual elements. We used the e-TSA framework for the identification and classification of the visual meta-information along with emphatic alterations in order to accentuate content text against structured text. In all cases, we achieved a meaningful vocalization of tables compared to that of commercial applications. However, the scalable model presented needs to be supported by a language model in order to provide better phrasal patterns.

References

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). Web Content Accessibility Guidelines 1.0, W3C Recommendation, 5 May 1999, *http://www.w3.org/TR/1999/WAI-WEBCONTENT-19990505/*

Lilley, C., & Raman, T.V. (1999). Aural Cascading Style Sheets (ACSS), W3C Working Draft 2 September 1999, *http://www.w3.org/TR/WD-acss*

Kowaliw, T. (2001) Accessible Mathematics on the Web: Towards an audio representation of MATHML Technology and Persons with Disabilities Conference, California North State University Northridge, 2001.

Pitt, I., Edwards, A. (1997). An Improved Auditory Interface for the Exploration of Lists. ACM Multimedia 1997, pp. 51-61

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

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

W3C, Aural style sheets, W3C description, *http://www.w3.org/TR/REC-CSS2/aural.html*

W3C, Document structure, W3C description, http://www.w3.org/MarkUp/html-spec/html-spec_5.html

W3C, Information type elements, W3C description, *http://www.w3.org/MarkUp/html3/logical.html*

W3C (1999), HTML 4.01 Specification, W3C Recommendation, *http://www.w3.org/TR/REC-html40*

Ausbrooks, R., Buswell, S., Carlisle, D., Dalmas, S., Devitt, S., Diaz, S., Hunter, R., Ion, P., Miner, R., Poppelier, N., Smith, B., Soiffer, N., Sutor, R., & Watt, S. (2002). Mathematical Markup Language (MathML) Version 2.0 (2nd Edition), W3C Working Draft 19 December 2002, http://www.w3.org/TR/2002/WD-MathML2-20021219/

Xydas, G. & Kouroupetroglou, G. (2001a). Augmented Auditory Representation of e-Texts for Text-to-Speech Systems, in Proceedings of the 4th International Conference on Text, Speech and Dialogue, TSD 2001, Plzen (Pilsen), Czech Republic, September 2001, pp. 134-141

Xydas, G. & Kouroupetrolgou, G. (2001b). Text-to-Speech Scripting Interface for Appropriate Vocalisation of e-Texts, in Proceedings of the 7th European Conference on Speech Communication and Technology, EUROSPEECH 2001, Aalborg, Denmark, September 2001, pp. 2247-2250