Acoustic Rendering of Data Tables Using Earcons and Prosody for Document Accessibility

Dimitris Spiliotopoulos, Panagiota Stavropoulou, and Georgios Kouroupetroglou

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
{dspiliot, pepis, koupe}@di.uoa.gr

Abstract. Earlier works show that using a prosody specification that is derived from natural human spoken rendition, increases the naturalness and overall acceptance of speech synthesised complex visual structures by conveying to audio certain semantic information hidden in the visual structure. However, prosody alone, although exhibits significant improvement, cannot perform adequately in the cases of very large complex data tables browsed in a linear manner. This work reports on the use of earcons and spearcons combined with prosodically enriched aural rendition of simple and complex tables. Three spoken combinations *earcons+prosody*, *spearcons+prosody*, and *prosody* were evaluated in order to examine how the resulting acoustic output would improve the document-to-audio semantic correlation throughput from the visual modality. The results show that the use of non-speech sounds can further improve certain qualities, such as listening effort, a crucial parameter when vocalising any complex visual structure contained in a document.

Keywords: document-to-audio, data tables, earcons, prosody, Text-to-Speech, ToBI, document accessibility.

1 Introduction

Electronic documents have by far exceeded the volume and variability of printed material for everyday use for quite some time. These electronic documents are browsed by the reader. Browsing and navigation may be done using web-browsers and pointing devices for most users. However, the document content cannot be accessed that way by a number of users, such as people that have a print disability [1] (those with partial or total vision loss, cognitive limitations, limited dexterity), the elderly or the moving user. In most of those cases, the document content is rendered to the acoustic modality through text-to-speech systems [4].

One major problem that arises is that the document content is source authored and optimised for visual presentation. This means that apart from the primary content – text – there are visual components and structures that are either applied onto the existing text or embedded in the document. Electronic documents is a general term used to denote the range of documents that include ebooks, web-pages, digital printed

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2009, LNCS 5616, pp. 587–596, 2009.
© Springer-Verlag Berlin Heidelberg 2009

documents (pdf) and a lot more similar types. Although the document types exhibit vast variability, they are in fact visual documents containing text enriched with visual attributes and structures. The complexity of accessing visual documents lies within the successful transfer of the semantic meaning of the visual components to the acoustic modality. The World Wide Web Consortium (W3C) works actively towards identifying and providing a wide range of recommendations and guidelines to make the content that is written for the web accessible [37].

Although speech-only user interfaces can utilize prosody control to convey the underlying semantic representation of complex visual structures, further assistance can be looked upon non-speech sounds such as earcons [33]. Voice Browsers provide an abundance of parameters for producing consistent auditory output when reading spoken format description using several attributes, such as speaking voice, prosodic parameterization and non-speech sounds.

In the case of document structures, earcons can provide a means of further improvement of semantically correct aural rendering of data tables. However, the use of non-speech sounds certainly decreases the naturalness of the rendition. There is, therefore, special interest in the exploration of the magnitude of improvement of aural rendition and how the tradeoff of the understanding against naturalness is received by the human listener.

This work reports on the psychoacoustic experimentation on spoken format of tables utilising earcons to assist a high-level prosody specification, focusing on well-formed HTML tables that, according to the W3C Web Accessibility Initiative (WAI) guidelines. The listener feedback would provide the means to understand on what level the use of non-speech sounds can aid the acoustic rendering data tables. Moreover, the analysis and evaluation of those results may form the basis for future improvement especially by further linguistic analysis of the data table structural properties and respective experimentation on the targeted use of non-speech sounds and other Document-to-Audio (DtA) methods.

2 Problem Formulation

Successful acoustic rendering of tables to speech presents a two-fold problem. Tables must first be accessed according to their logical representation and then rendered accordingly in a different modality in a way that stays true to their semantic relations derived from the underlying logical representation. Complex visual structures are used for the semantic grouping of text tokens and bear a distinct association between the physical layout and the underlying logical structure [5]. The columns and rows of a table represent the logical connections [6]. These connections constitute a relation matrix between “organized hierarchical concepts” [7]. A de-compilation process must be employed when rendered to speech because most of the semantic meaning of their enclosed text is implicit to the visual structure.

Data tables are classified into simple and complex based on the levels of logical row or column headers. In complex tables, header and data cells can be expanded to encompass more than one row or column forming nested tables. In that respect, complex tables are considered to be three-dimensional structures [8] on the logical level, compared to the two-dimensional simple data tables, the third dimension of the semantic structure embedded inside the two dimensional visual structure (Fig. 1).

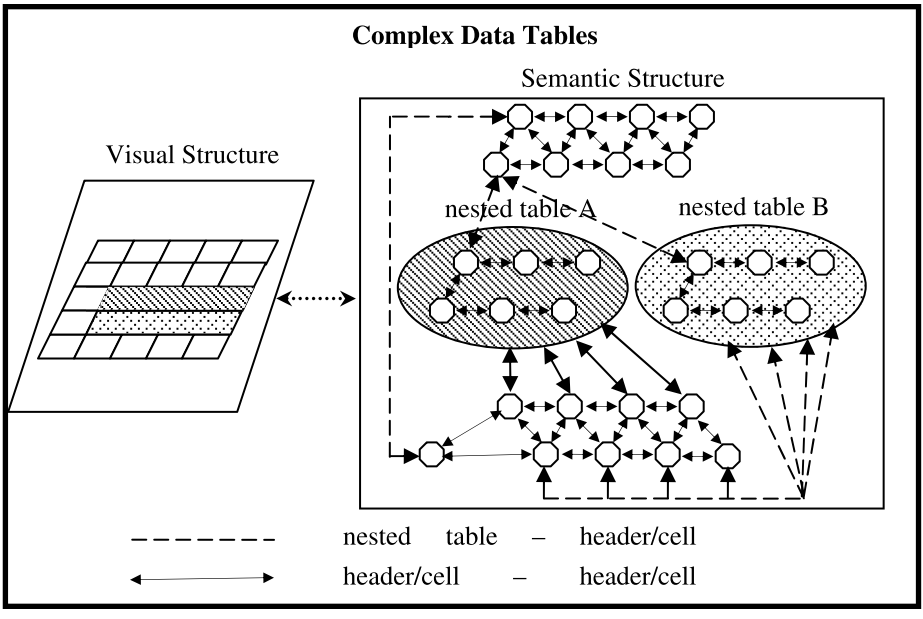

Fig. 1. Nested tables, visual and logical representation

Tables can be processed by identifying their dimension, which is directly proportional to the complexity, and therefore deriving the logical grid [9]. The important meta-information hidden in tables is reconstructed in order to provide a means for readers to comprehend the representation of tables. This can be done by constructing a "semantic description" of the tables either automatically or manually [10]. In this case, the first part of the problem is addressed on the visual level, the second and harder part being the successful acoustic rendering.

The aural rendition of tables requires accessing the semantic information under the visual structure using existing techniques from earlier work. Information about the semantic structure of HTML tables can be used to aid navigation and browsing of such visual components [11]. Works on simpler visual structures, such as lists, reveal the inherent hierarchy manifested in nested bulleting and how that must be taken into consideration between the levels of the structure [12]. Appropriate markup can be used to assign logical structure arrangement to table cells [13], while navigation can be improved by additional markup annotation to add context to existing tables [14]. Other suggestions include automated approaches for retrieval of hierarchical data from HTML tables [15]. Smart browsers are used to access critical information for use in reading tables as well as linearization techniques are employed for transforming tables into a more easily readable form by screen readers [16]. Table browsing techniques include the use of Conceptual Graphs for the classification of header and data cells using Hidden Markov Models for identification [17] as well as systems that decompile tables into discrete HTML documents using an HTML index for navigation [18].

Earlier experimentation used the natural language human renderer paradigm to derive a basic prosodic model for table-to-speech [2]. This diction-based specification has successfully rendered both simple and complex tables to audio. The resulting evaluation showed that, for linear browsing, the approach failed to achieve adequate scoring for complex tables. That means that prosody alone was not sufficient when no spatial information was passed along during browsing. The embedded nested structures needed increased listening effort by the human ear to recognise. Prosody manipulation, however, was very successful on providing the means for header/data cell distinction and row start/end within a nested structure [3].

Earcons and auditory icons have been used extensively in auditory human-computer interfaces [19]. Earcons are used to communicate context such as hierarchies [20], depict visual categorisation [21], for menu navigation in graphical user interfaces [22], telephone interfaces [23] and mobile phones [24]. Their scalability allows for creation of intermittent sounds that are used to create cues [25].

Spearcons are small sequences created from natural or synthesised speech which can be processed according to designer specifications. They can be used to replace earcons by providing a more familiar and sometimes contextual sound in cases such as menu navigation [26].

This work utilises earcons in order to bridge the semantic gap between nested sub-table structures for complex tables. For simple tables, a clearer distinction of row-end markers would be the expected result for carefully places non-text audio. The primary hypothesis was that the combined use of earcons and prosody assistance would relieve the listening effort parameter of the listeners, mostly in the linear browsing cases. A secondary hypothesis was that, based on the actual manner of information feed from the acoustic rendering of data tables, the response time of listeners would in most cases improve when identifying desired information during listening. The aim of this work was to set the basis for experimentation on the use of standalone and combinatory methods for improved acoustic rendering of complex visual structures.

3 Experiment Description and Setup

Two sets of example simple complex tables were used for the acoustic experiment, fully compliant to the W3C recommendations as illustrated in [2, 3]. The data tables were marked according to their properties, for complexity, browsing, and size, as simple/complex, linear/intelligent, medium/large, respectively. For the experiment, sets of simple and complex data tables were synthesized using the "DEMOSTHeNES" Text-to-Speech advanced platform [27] according to the prosodic specification already derived from earlier work.

Analysis of both phonological and phonetic properties of the utterances showed that there is a clear correlation between the hierarchical grouping of the semantic content of the tables (hierarchical discourse structure of data tables) and the prosody of the resulting discourse segments and sub-segments. In particular:

A. *Qualitative specification of phonological parameters: Tonal Events (Phrase Accents and Boundary Tones) Distribution.* Discourse structure of tables is primarily conveyed by alternation of high (H-H%) and low (L-L%) boundary tones. High boundary tones indicate that the utterance forms an interpretive unit – (sub-)segment

with the subsequent phrase, whilst low boundary tones signal the end of the respective unit. Consequently, different browsing techniques, which reflect different segmentations of the semantic content, result in different distributions of tonal events (Fig. 2).

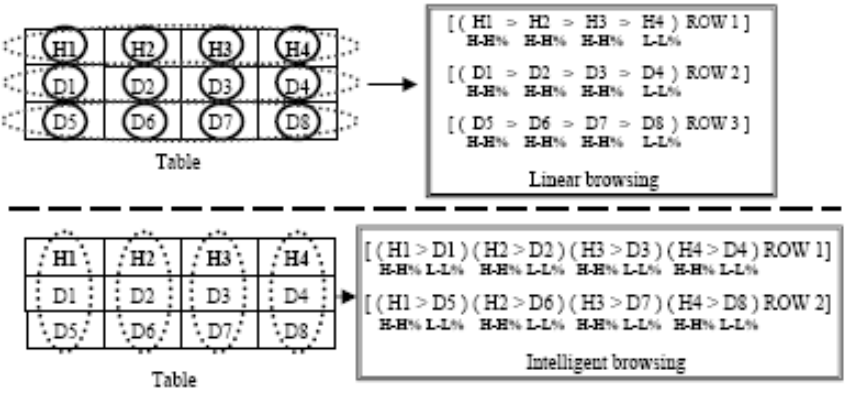

Fig. 2. Linear and Intelligent rendition of simple tables (Speaker 02 [3])

Row finality is consistently marked by an L-L% boundary tone in both cases. In the linear rendition however single cells are interpreted within a single row and are thus followed by a high boundary tone. In the intelligent rendition the basic interpretive unit is the header-data pair, and thus data cells coinciding with the end of each pair are marked by a low tone instead. Note that the listeners' preferred linear rendition ([3]) was not the one depicted in this figure, but the one closest to the intelligent rendition, whereas data cells were followed by an L-L% tone representing a single interpretive unit.

Since intelligent browsing techniques are closer to the intended semantic-discourse structure of tables, it came as no surprise that listeners showed a clear preference for the corresponding prosodic rendition [3]. Once the distribution of boundary tones is decided upon, choice of Pitch Accents (PAs) straightforwardly follows, as the choice of boundary tone greatly determines the choice of PA [28, 29]. In particular high edge tones are associated with L* nuclear PAs, while low edge tones are associated with H* nuclear PAs. The resulting contours correspond to the typical continuation rise and final lowering contours respectively.

B. *Quantitative specification of phonetic parameters (Pauses, Speech Rate, Duration, Pitch and Intensity).* Differences in discourse structure are reflected upon the length of pause breaks as well. Shorter breaks are associated with lower level subsegments, while longer breaks follow the end of hierarchically higher units. For example, for complex linear tables the edge of rows was marked by longer pauses than the edge of data cells within. Accordingly, for complex intelligent tables, single cells, header-data cell pairs, rows and nested tables were marked by increasingly longer pauses reflecting each unit's status in the hierarchical organization of discourse (Table 1, Figure 3). Pausing, on the other hand, could not differentiate between initial header cell and subsequent data cells within a single row.

Table 1. Pause Breaks in seconds. Degree of pausing reflects the hierarchical organization of the semantic content within tables (see also Fig. 3).

Complex Int. Table (Medium)	
Placement	**Pause**
Data cell	0.68
Header-Data Cell Pair	2.83
Row	3.76
Nested Table	5.29

Association of pausing with different levels of discourse structure has already been reported for plain texts [30, 32, 34]. Discourse structure is in general accepted to associate with various aspects of prosody. As previous works [29, 30, 31, 35] show, parameters such as pitch range, max and min F0, pre-boundary lengthening, max and min intensity serve as strong indicators of discourse relations. These parameters were not assessed in this experiment, since one needs to control for the segmental content of (sub-)segments. In other words, same phrases should be examined in different positions for the results to be comparable. The assessment, thus, of these parameters remains as direction for future research. Finally, speech rate didn't play an important role being roughly constant throughout segments of different levels, headers and data cells (approximately 6.5 syl/sec for simple tables and 5.3 syl/sec for complex tables). Again caution should be taken in the interpretation of this result, since each segment consists of syllables with inherently varying duration.

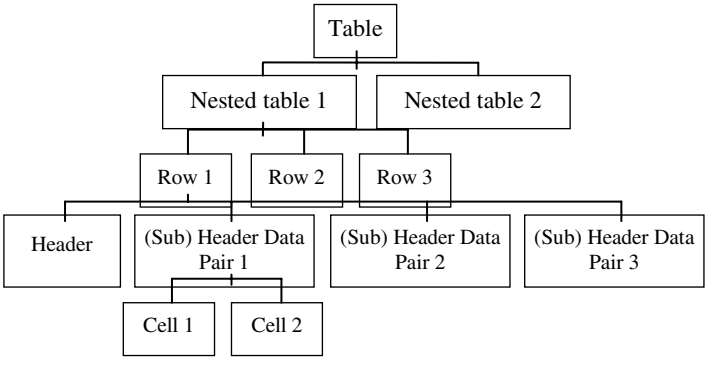

Fig. 3. Discourse segmentation for complex tables (intel). Degree of pausing (Table 1) increases as we move higher in the tree structure.

The earcons were constructed according to guidelines [36] and familiar short-length sounds from several internet soundfont databases were explored by linguistics experts following pilot experiments of non-combined use of these in earlier stages and for pairing with free-form text from the same speech synthesizer used in this experiment. The earcons were placed at the end of individual rows and nested tables. In effect for linear browsing, simple tables depended on two distinct earcons to show the transition between header cells and data cells. For complex tables earcons were used

in between header-data cell tokens as well as for end of rows. For intelligent browsing, earcon sounds were used to denote the transition between nested-tables (complex tables) and end-of-row (semantic) for header-data cell pairs. Three distinct sounds were used for complex tables and only one needed for simple tables.

Spearcons were constructed using the Acapela Greek Male voice to record "table start" and "table end" sequences. The same voice and speech rate was used for all instances throughout this experiment. The sequences were post-processed and time compressed to about 40% of their length.

Twelve first-time listeners (20-29 years of age) listened to pairs of simple/complex, linear/intelligent, medium/large synthesized tables that were evenly distributed among them so that everyone was presented with 4 renditions, one for each distinct table. This ensured that no memory effect had taken precedence over the measured evaluation qualities since no content was played more than once for each participant. The spoken renditions were distributed according to the following combinations for voice specifications: (1) prosody-only, (2) prosody + earcons, and (3) prosody + spearcons. The prosody specification used as input by the speech synthesis platform was derived from human natural spoken renditions, as presented above.

The listeners were asked to listen to the aurally rendered tables and present feedback and evaluation of several criteria all pertaining to semantic content and visual-to-aural rendition. The listeners effectively evaluated the use of any parameters or combinations in the aural renditions of simple and complex visual structures with the sole aim to show proof of preference. Preference markings were assigned to places of key table elements (such as nested table start/end, etc.) showing importance, semantic correlation, naturalness, and overall impression for each of the aforementioned voice parameters.

4 Evaluation

The experimental results were evaluated in order to classify all the above parameter combination renditions in terms of semantic awareness and naturalness for the full data tables. More importantly, following user preferences, we assembled a detailed breakdown of voice rendition quality improvement as described by specific parameter placements, such as preference of earcons instead of spearcons for identification of nested tables. The presented results are the derivation of the associated characteristics in a description that could be considered as part of a full specification for aural rendition of visual-oriented documents for a Document-to-Audio system.

Our secondary hypothesis was confirmed. The response time of the listeners when answering data control questions while listening to the acoustic renditions was improved by more than 10% when earcons/spearcons were employed. This elevates the certainty of the listener feedback and, along with the slightly improved precision, concludes that the listening effort was relieved, confirming the main hypothesis.

Moreover, that was explicitly the case in the subjective evaluation where the participants' proclaimed improvement was 14-16% for listening effort as shown in Figure. 4. Table 2 shows the precision and recall metrics calculated from the data question feedback.

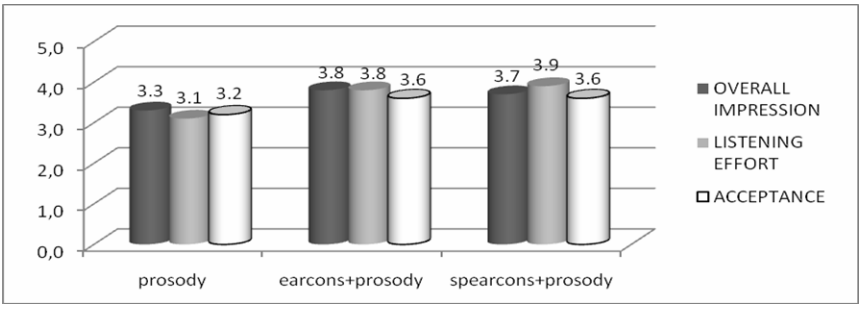

Fig. 4. Overall impression, listening effort and acceptance (higher=better)

Table 2. Evaluation overview for all tables (objective evaluation)

		prosody	earcon+prosody	spearcon+prosody
Simple (linear)	Precision	0.98	0.99	0.99
	Recall	0.93	0.94	0.94
	f-measure	0.95	0.96	0.96
Simple (intelligent)	Precision	0.99	0.99	0.99
	Recall	0.97	0.97	0.97
	f-measure	0.98	0.98	0.98
complex (linear)	Precision	0.92	0.92	0.94
	Recall	0.50	0.61	0.72
	f-measure	0.65	0.73	0.82
complex (intelligent)	Precision	0.98	0.99	0.98
	Recall	0.82	0.85	0.83
	f-measure	0.89	0.91	0.90

5 Conclusion

This work is concerned with complex data table aural rendition using earcons and spearcons to improve upon high-level prosody parameterization. As an overall assessment of the results from these experiments, it can be deducted that the use of earcons provides a sizable assistance mostly on the listening effort. However, although non-speech sounds may be thought as a step back for the naturalness of the acoustic rendering, in the case of complex visual structures there is an improvement in the acceptance and precision of data recognition. The listeners reported higher precision and recall effectively distinguishing between types of data, such as header and data cells better without any cost on the overall impression.

Acknowledgements. The work described in this paper has been funded by the KAPODISTRIAS Programme of the Special Account for Research Grants, University of Athens.

References

1. Freitas, D., Kouroupetroglou, G.: Speech Technologies for Blind and Low Vision Persons. Technology and Disability 20(2), 135–156 (2008)
2. Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G.: Diction Based Prosody Modeling in Table-to-Speech Synthesis. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 294–301. Springer, Heidelberg (2005)
3. Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G., Argyropoulos, V.: Experimentation on Spoken Format of Tables in Auditory User Interfaces. In: Proc. 11th Int. Conf. on Human-Computer Interaction (HCII 2005), USA, July 22-27 (2005)
4. Fellbaum, K., Kouroupetroglou, G.: Principles of Electronic Speech Processing with Applications for People with Disabilities. Technology and Disability 20(2), 55–85 (2008)
5. Ramel, J.-Y., Crucianou, M., Vincent, N., Faure, C.: Detection, Extraction and Representation of Tables. In: Proc. 7th Int. Conf. Doc. Analysis and Recognition (ICDAR), pp. 374–378 (2003)
6. Silva, A.C., Jorge, A.M., Torgo, L.: Design of an end-to-end method to extract information from tables. Int. J. Document Analysis and Recognition, Special issue on detection and understanding of tables and forms for document processing applications, 8(2), 144–171 (2006)
7. Hurst, M.: Towards a theory of tables. Int. J. of Doc. Analysis 8(2-3), 123–131 (2006)
8. Pontelli, E., Xiong, W., Gupta, G., Karshmer, A.: A Domain Specific Language Framework for Non-visual Browsing of Complex HTML Structures. In: Proc. ACM Conf. Assistive Technologies (ASSETS), pp. 180–187 (2000)
9. Embley, D.W., Hurst, M., Lopresti, D.P., Nagy, G.: Table-processing paradigms: a research survey. Int. J. Document Analysis 8(2-3), 66–86 (2006)
10. Pontelli, E., Gillan, D.J., Gupta, G., Karshmer, A.I., Saad, E., Xiong, W.: Intelligent non-visual navigation of complex HTML structures. Universal Access in the Information Society 2(1), 56–69 (2002)
11. Pontelli, E., Gillan, D., Xiong, W., Saad, E., Gupta, G., Karshmer, A.: Navigation of HTML Tables, Frames, and XML Fragments. In: Proc. ACM Conf. on Assistive Technologies (ASSETS), pp. 25–32 (2002)
12. Pitt, I., Edwards, A.: An Improved Auditory Interface for the Exploration of Lists. In: ACM Multimedia 1997, pp. 51–61 (1997)
13. Hurst, M., Douglas, S.: Layout & Language: Preliminary Experiments in Assigning Logical Structure to Table Cells. In: Proc. 4th Int. Conf. Document Analysis and Recognition (ICDAR), pp. 1043–1047 (1997)
14. Filepp, R., Challenger, J., Rosu, D.: Improving the Accessibility of Aurally Rendered HTML Tables. In: Proc. ACM Conf. on Assistive Technologies (ASSETS), pp. 9–16 (2002)
15. Lim, S., Ng, Y.: An Automated Approach for Retrieving Hierarchical Data from HTML Tables. In: Proc. 8th ACM Int. Conf. Information and Knowledge Management (CIKM), pp. 466–474 (1999)
16. Yesilada, Y., Stevens, R., Goble, C., Hussein, S.: Rendering Tables in Audio: The Interaction of Structure and Reading Styles. In: Proc. ACM Conf. Assistive Technologies (ASSETS), pp. 16–23 (2004)
17. Kottapally, K., Ngo, C., Reddy, R., Pontelli, E., Son, T.C., Gillan, D.: Towards the Creation of Accessibility Agents for Non-visual Navigation of the Web. In: Proc. of the ACM Conf. on Universal Usability, Vancouver, Canada, pp. 134–141 (2003)

18. Oogane, T., Asakawa, C.: An Interactive Method for Accessing Tables in HTML. In: Proc. Intl. ACM Conf. on Assistive Technologies, pp. 126–128 (1998)
19. Brewster, S.A., Wright, P.C., Edwards, A.D.N.: An evaluation of earcons for use in auditory human-computer interfaces. In: SIGCHI Conf. on Human Factors in Computing Systems, Amsterdam (1993)
20. Lucas, P.: An evaluation of the communicative ability of auditory icons and earcons. In: Kramer, G. (ed.) Proc. ICAD 1994, Santa Fé Institute, Santa Fé, NM. Addison-Wesley, Reading (1994)
21. Lemmens, P.M.C., Bussemakers, M.P., de Haan, A.: Effects of Auditory Icons and Earcons on Visual Categorisation: The Bigger Picture. In: Proc. 2001 Int. Conf. on Auditory Display, Espoo, Finland, July 29-August 1 (2001)
22. Brewster, S., Raty, V.-P., Kortekangas, A.: Earcons as a method of providing navigational cues in a menu hierarchy. In: Proc. Int. Conf. Human Computer Interaction. Imperial College, London (1996)
23. Brewster, S.: Navigating telephone-based interfaces with earcons. In: BCS HCI 1997, UK (1997)
24. LePlâtre, G., Brewster, S.: Designing non-speech sounds to support navigation in mobile phone menus. In: Int. Conf. on Auditory Display (ICAD 2000), Atlanta, USA (1998)
25. Watson, M.: Scalable earcons: Bridging the gap between intermittent and continuous auditory displays. In: Proc. 12th Int. Conf. on Auditory Display, UK, June 20-23 (2006)
26. Walker, B.N., Nance, A., Lindsay, J.: Spearcons: speech-based earcons improve navigation performance in auditory menus. In: Proc. 12th Int. Conf. on Auditory Display (ICAD 2006), London, UK, June 20-23 (2006)
27. Xydas, G., Kouroupetrolgou, G.: Text-to-Speech Scripting Interface for Appropriate Vocalisation of E-Texts. In: Proc. 7th European Conf. Speech Communication and Technology, pp. 2247–2250 (2001)
28. Baltazani, M., Jun, S.-A.: Focus and topic intonation in Greek. In: Proc. 14th Int. Congress of Phonetic Sciences, vol. 2, pp. 1305–1308 (1999)
29. Dainora, A.: Does intonational meaning come from tones or tunes? evidence against a compositional approach. In: Proc. Speech Prosody 2002, pp. 235–238 (2002)
30. Herman, R.: Intonation and discourse structure in English: Phonological and phonetic markers of local and global discourse structure, Ph.D Thesis (1998)
31. Nakatani, C., Hirschberg, J., Grosz, B.: Discourse Structure in Spoken Language, Studies on Speech Corpora (1995)
32. Swerts, M.: Combining statistical and phonetic analyses of spontaneous discourse segmentation. In: Proc. 12th Int. Congress of Phonetic Sciences, Stockholm, August 1995, vol. 4, pp. 208–211 (1995)
33. Blattner, M.M., Sumikawa, D.A., Greenberg, R.M.: Earcons and Icons: Their Structure and Common Design Principles. Human-Computer Interaction 4(1), 11–44 (1989)
34. Van Donzel, M.: Prosodic Aspects of Information Structure in Discourse. LOT dissertations, vol. 23. Holland Academic Press, The Hague (1999)
35. Den Ouden, H., Noordman, L., Terken, J.: The prosodic realization of organizational features of texts. In: Proc. Speech Prosody 2002, pp. 543–546 (2002)
36. Brewster, S.A., Wright, P.C., Edwards, A.D.N.: Experimentally derived guidelines for the creation of earcons. In: Proc. HCI 1995, Huddersfield, UK (1995)
37. Caldwell, B., Cooper, M., Guarino Reid, L., Vanderheiden, G. (eds.): Web Content Accessibility Guidelines 2.0, W3C Candidate Recommendation, April 30 (2008), http://www.w3.org/TR/WCAG20/