Auditory Accessibility of Metadata in Books: A Design for All Approach

Dimitrios Tsonos, Gerasimos Xydas, and Georgios Kouroupetroglou

University of Athens,
Departement of Informatics and Telecommunications,
Panepistimiopolis, Ilisia, GR-15784 Athens, Greece
{ea02534,gxydas,koupe}@di.uoa.gr

Abstract. There are two issues that are challenging in the life-cycle of Digital Talking Books (DTB): the automatic labeling of text formatting meta-data in documents and the multimodal representation of the text formatting semantics. We propose an augmented design-for-all approach for both the production and the reading processes of DAISY compliant DTBs. This approach incorporates a methodology for the real-time extraction and the semantic labeling of text formatting meta-data. Furthermore, it includes a unified approach for the multimodal rendering of text formatting, structure and layout meta-data by utilizing a Document-to-Audio platform to render the acoustic modality.

Keywords: e-books, digital talking books, document accessibility, document-to-audio, auditory accessibility of text formatting, DAISY.

1 Introduction

Nowadays, e-books have exceeded printed books by providing browsing, navigation, searching, highlighting and multimedia facilities to the reader. The content is being rendered based on embedded meta-data of structural and layout properties that affect the visual stimuli. The meta-data and the way it is combined and represented in visual modality, has specific meaning for each reader in different documents [1][2] and the way the document is read [3][4]. There are also studies on how the combination of colors on a page and the different type of style on text, affects narrator's emotional state and so the readability of the document [5].

Access to an e-book requires a two-way communication between the user and the e-book. One common format specification is the Digital Talking Book (DTB) [15]. DTB provides multimodal presentation of the content (audio, visual and haptic [16]) and also introduce additional special tagging to e-books in order to facilitate navigation and audio synchronization with the text. DTB can be also used as Digital Audio Books and the content is reproduced (due pre-recorded speech) through common devices like a CD player.

The complexity of accessing e-books in a user-friendly way has led to the proposal of several guidelines and specifications in the field of electronic content accessibility. The NISO/ANSI [17] standard has been proposed for the creation of Digital Talking

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 436–445, 2007.

Books. This standard is supported by the DAISY consortium [16] and specifies a faithful access on the representation of a published work to visually impaired and print-disabled readers like to readers of the original printed publication. The Open Document Format (ODF) [18] is an open XML standard to be used for documents containing text, spreadsheets, charts, and graphical elements. W3C also provides recommendations for making web content accessible for the disabled, through Web Content Accessibility Guidelines [19].

In acoustic modality, variations and differences of speech characteristics can have great impact to the perception of the speech. Depending on the listeners and the content that is read, there are different expressive styles of speech [6][7]. Pre-recorded speech and decorative sounds are most commonly used in e-books. This makes the production of an e-book a very time consuming and expensive process but also prevents the dynamic creation of e-books from any available content. Nowadays, synthetic speech has become very natural in quality by the introduction of corpus-based Text-to-Speech (TtS) techniques, but speech synthesis is mono-dimensional and only deals with plain typographic texts. E-books are currently support TtS functionality but the semantics of the original visual format of the book are not successfully transferred to the listener.

Previous works have shown the weaknesses and the errors in the comprehension of the acoustic version of visual-oriented documents when stripping the visual meta-data and synthesizing only the content text [20][21]. Recent works have demonstrated encouraging results in the transformation of the visual stimuli to the acoustic modality. There is a systematic work [22] to develop the Document-to-Audio open platform for vocalizing meta-data along with the content text, producing a two-dimensional audio stream, by: (a) combining alternative text insertion in the document's text stream, (b) altering the prosody, (c) switching voices and (d) inserting other sounds in the waveform stream, according to the class of meta-data provided in the document. A recent study has modeled speech styles when vocalizing tables using synthetic speech [8]. In [9], it is shown how the variations of speech and its characteristics affect the listener's state of emotions (emotional prosody). For the optimization of the acoustic representation, there are efforts that combine speech synthesis with special sounds like earcons [10] and auditory icons [11]. For example, earcons can be used for navigation in a menu [12][13]. Another approach is using 3D Audio for the acoustic representation of the content [14].

In this work we propose an augmented design-for-all approach for both the production and the reading process of DAISY/NISO compliant DTBs. This approach incorporates a methodology for the real-time extraction and the semantic labeling of text formatting meta-data. Furthermore, it includes a unified approach for the multimodal rendering of text formatting, text structure, text layout and non-textual meta-data. Specifically for the acoustic modality, we present a prototype for the auditory presentation of e-book contents.

2 The Proposed Architecture

We introduce an open real-time XML-based architecture allowing a multimodal access to books. The architecture fulfils the needs of two major tasks: a) the

production of compliant DTBs and b) the presentation of their content and its attributes provided by the meta-data. The proposed system is able to process any book, in printed or electronic format.

Figure 1 shows the overall architecture consisting of two parts: the Universal player (U-Player) and the Book Explorer. The later hosts the analysis of the books, as well as the processing of navigation issues. We have followed the client – server model (CS model) in order to serve any small-footprint specifications on the user's device. The U-Player is being installed on user's device (e.g. a personal computer, a mobile device etc). On the other hand, the resource demanding Book Explorer could be hosted on a powerful server machine that performs analysis tasks.

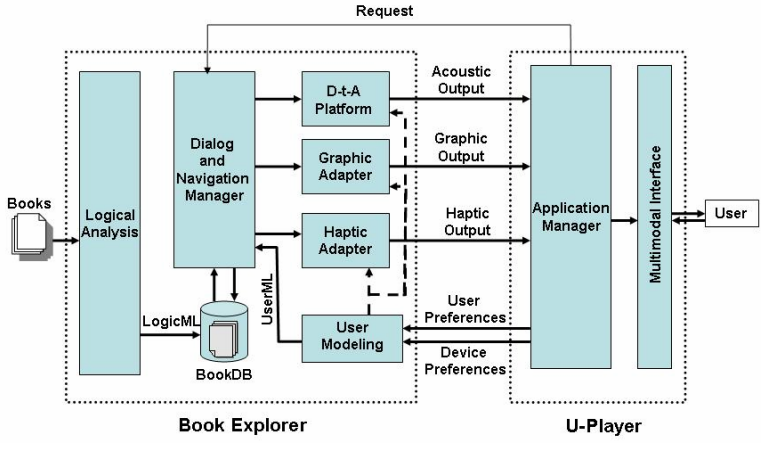

Fig. 1. The overall architecture

The CS model also allows the Server to serve simultaneously multiple clients of different modalities. A user can access the application through internet but another user through a mobile provider, selecting the appropriate gateway for their interface.

2.1 Production Task

The Book Explorer is the module that deals with the production of the e-book and handles the reading and exploration tasks. The Logic Analysis module parses the book (in printed or electronic form) and the produced output (DAISY/NISO standard compliant) is stored in the BooksDB database. The stored content and any additional information (meta-data) is accessed by the Dialog and Navigation Manager. The information retrieval from the BooksDB is done according the demands of the present under execution task. Furthermore, the Dialog and Navigation Manager handles the user and device preferences (UserML) produced by the User Modeling module. Depending on the user preferences it manages the way that content will be handled and delivered in each modality (acoustic, visual, haptic) through the Document-to-Audio platform (or/and the Graphic Adapter and/or the Haptic Adapter) to the Application Manager.

To serve the multimodal accessibility requirements, our methodology first creates a logical layer from the source book, free of presentation details. Then, these logical data can be then transformed to any modality. Though books are aiming at visual modality, this logical layer aims to record the reader's understanding of the book. We will show in detail how the original visual stimuli can affect the acoustic presentation via this layer.

2.2 Presentation Task

The presentation task is carried out at the U-Player component. This takes care for user interaction with the service (and consequently with the content) in multimodal way. The U-Player is divided into two major tasks: input and output.

In output task, the user receives the requested content or other system prompts. The output task can accomplish the type of DTBs with full audio through the Navigation Center (NCC or NCX). The structure is two-dimensional, providing both sequential and hierarchical navigation. In many cases, the structure in this type of DAISY DTB resembles the table of contents of its print source. Some of these productions provide page navigation. But depending on the user's preferences the functionality can change respectively.

In input task the architecture supports multimodal accessibility, depending on the user's needs. The input devices (or applications) vary, from depend on the device Automatic Speech Recognition (A.S.R.), keyboard, mouse or buttons.

The Application Manager handles the user preferences, which are used in Book Explorer or Output tasks. The input can be operation or navigation commands. The application manager creates Requests that contain information about: (a) User preferences, and (b) Navigation or Application Command.

2.3 User Modeling

The User Modeling Module handles the user profile. The profile is a collection of actions and preferences acquired by the interaction with the Application Manager. Such can be the user's needs, the interaction modality, the navigational factors, environmental factors etc.

2.4 Dialog and Navigation Manager

The navigation in DTBs supports several functionalities. As basic navigation is considered the movement through text and sophisticated movement is the hierarchical navigation in book (Using Navigation Control File that describes the hierarchy of the book). The hierarchy is provided separately from the text.

The Dialog - Navigation Manager Module (D.N.-M.) can handle commands from the user for navigation tasks and generally tasks that satisfies user's content and information retrieval. An adaptive dialog system [31] is used that corresponds to the two following functions [32]: a) Supporting System Use, taking over parts of routine tasks, adapting the interface, giving advice of system use, controlling a dialogue.

b) Supporting Information Acquisition, helping users to find information, tailoring information presentation.

This module is agent-based, allowing every module to interact and to be capable of reasoning. This kind of architecture is very flexible and user-centred. The user can have full control of the dialog [33]. Also the proposed architecture is flexible and can support multimodal dialog input and not only spoken dialogs [34], while the navigation features are conforming to the DAISY/NISO standard, supporting Fast Forward and Fast Reverse, Reading at Variable Speed, Notes, Cross Reference Access, Index Navigation, Bookmarks, Highlighting, Excerpt Capability, Searching, Spell-Out Capability, Text Attributes and Punctuation, Tables, Nested Lists, Text Elements, Skipping User-Selected Text Elements, Location Information, Summary and Reporting Information, Science and Mathematics.

3 Logical Document Analysis

The proposed system targets to process all kind of books in printed or electronic format. The Markup Normalization Module converts all non-tagged books as well as tagged books not conforming with the DAISY/NISO format into tagged compliant to DAISY/NISO format. Printed books are scanned and parsed through an OCR system so to be digitized and exported in a tagged format. Documents being already in a tagged format include meta-data about the format and the structure of the text. E-books are also tagged digital books but have specific file format and tagging. All these have to be normalized to the required meta-data and file type (DAISY/NISO standard). This format is described in our methodology as BooksML.

In case the book has the required by our approach semantic annotation (described later), the logic analysis can be by-passed, and the e-book can be stored directly to systems database (BooksDB) (Figure 2).

3.1 Type of Elements

The e-books include meta-data that can be categorized in:

- Text Formatting meta-data
- Text Structure meta-data
- Text Layout meta-data
- Non-textual meta-data (Figures, Drawing, Pictures, Logos etc)

Text Formatting meta-data include the formation elements of the text, typesetting elements and font elements (like bold, italics, font size). Text Structure meta-data specify the attribute of a part of the document (chapter, title, paragraph e.t.c.), while the Text Layout meta-data concerns the formation of the text (like columns, headlines, borders). The text with the formation and structural metadata can be combined with other non-textual metadata, such as figures, drawing e.t.c.

One can notice that there is a relation between these elements. For example, the title (text structure) is in font size 16pt (text formatting) and it is placed in a column (text layout). Also, the subtitle has font size 14pt and it is in a column, but under the

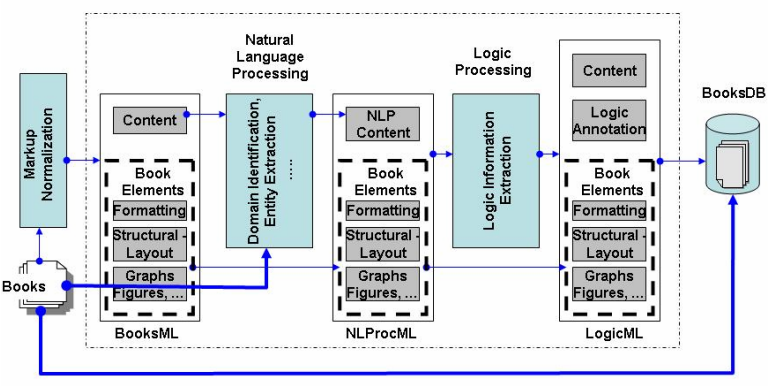

Fig. 2. The Logic Analysis Module. Each sub-system produces an XML file which contains the information that derived by its procedure.

title. We can distinguish that between these elements, the relation is M:N. But there are combinations that can be excluded through a statistical analysis.

The above elements should be included in the tagged book (BookML). The elements that are described in the DTBs standard [28] can be considered as a sub-group of the elements that can be processed by our proposed methodology.

3.2 Semantics of Meta-data

The elements of a document are set by the editor for different purposes. Stricter typographic rules are used in technical documents than in magazines. Not considering the domain of the document, we can see a few examples. A sentence written in bold, size 16pt and on the top of the page is the title of a chapter and gives to the reader the main concept that will be analyzed in the body of the text. A bold, numbered sentence under an image is the caption of the image explaining the image above it.

As we can see, the combination of elements has a semantic mapping. DAISY/NISO standard supports a plethora of semantic meta-data for representation or navigation. Though in acoustic modality these meta-data can be exploited, in this work we introduce additional semantics that are more useful for the acoustic rendering of the e-book. For example, several level of emphasis in synthetic speech are required to produce a more augmented acoustical image of the document compared to the emphasis and strong emphasis levels provided by the DAISY/NISO standard. Other semantic – logic meta-data includes “dialog” (“question” and “answer”), “definition of a word”, “new term” e.t.c.

Logic Information Extraction adds logic meta-data to the e-book using machine learning algorithms (e.g. Bayesian Network) to learn the mapping between visual and logic elements. The statistical analysis can be done through a series of experiments to allocate the mapping, which in some cases is 1:N (e.g. bold can be mapped as emphasis but also as strong emphasis). A questionnaire can be created and given to readers to annotate the logic of the sections in books.

4 Delivering Books into Acoustic Modality

When conveying a book in auditory modality, the functionality is focused on representing plain text stripping all visual, structural and layout information included. So the acoustic presentation lacks of the additional information that is provided. In DTBs for navigation purposes the structural elements are spoken out separately from text.

While the reader uses the navigation features provided in a DTB, it is required these information – metadata conveyed in acoustic modality. The goal is not to separate or remove the metadata from the text itself. Trying to accomplish this goal arises the need of defining which auditory elements these features affect. For example, the title should be spoken out in a different manner than the body of text, changing the appropriate prosodic elements of the synthetic speech or using specific sounds like earcons, so the user to understand which is the attribute of the text that was spoken out without to be defined separately.

In this section we will describe the acoustic mapping of semantic meta-data in books. As we described above, the book is formatted in LogicML. Following previous works on Document-to-Audio conversion [30], the semantics meta-data can be acoustically represented by specific auditory elements like (a) alternative text insertion in the document's text stream, (b) modifications in the prosody, (c) switching voices and (d) inserting other sounds like earcons and auditory icons in the waveform stream, according to the class of meta-data provided in the e-book. The user can be trained to recognize and to combine speech and sounds with specific commands and events.

The book (in LogicML) through the Dialog and Navigation Manager is passed to the Document-to-Audio platform (Fig. 3) [29][30]. It will be mapped in specific acoustic elements, as mentioned above, producing a new annotated document and auditory synthesizer will implement the mapping (the output files can be e.g. Mpeg4, SMIL and audio files).

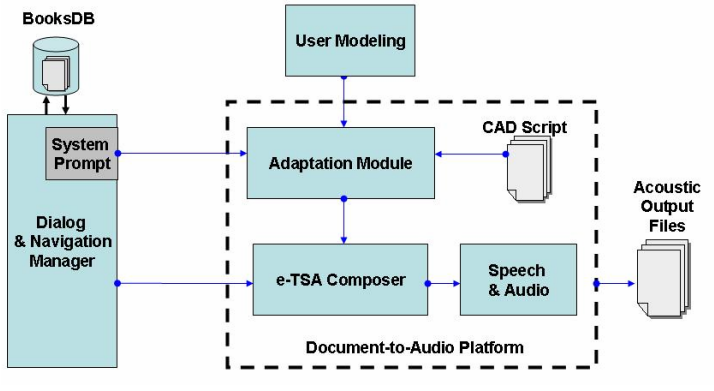

Fig. 3. The Acoustic Mapping module

The mapping will be done according the information provided to DtA platform. The rules that describe the semantic relation of metadata and acoustic representation are defined with the use of Cluster Auditory Definition (CAD) scripts as it is proposed in e-TSA composer platform. Through CAD scripts it is possible to be defined the prosodic behavior and audio insertions, like earcons, in each pair of metadata along with its content text. The DtA platform will give greater priority to user preferences than the default CAD rules. For example, the user might need to hear faster the content of the book but there are elements that should be read slower (defined by CAD). Then the Adaptation Module will give higher priority to the rules provided by User Modeling. Another functionality of Adaptation Module is how to handle System Prompts. The system prompts can be mapped to different elements than these of the content, so to be distinct to the listener.

The rules of mapping can derive either by a "test and error" process or from a series of experiments, like in [8] [26].

5 Conclusions

We have presented an augmented approach for both the production and the reading processes of DAISY/NISO compliant DTBs. Our methodology includes a system architecture for multimodal accessibility of books. In this paper we have emphasised the acoustic modality. Our aim was to propose a methodology for presenting the independency of the logic, hidden in the elements of a book for multimodal accessibility.

Acknowledgements. The work described in this paper has been funded by the European Social Fund and National Resources under the HOMER project of the Competitiveness Programme: PENED, Greek General Secretariat of Research and Technology.

References

1. Küpper, N.: Recording of Visual Reading Activity: Research into Newspaper Reading Behaviour (1989) Available as pdf from http://calendardesign.de/leseforschung/eyetrackstudy.pdf
2. Holmberg, N.: Eye movement patterns and newspaper design factors. An experimental approach. Master Thesis, Lund University Cognitive Science (2004)
3. Holmqvist, K., Holsanova, J., Barthelson, M., Lundqvist, D.: Reading or scanning? A study of newspaper and net paper reading. In: Hyönä, J.R., Deubel, H. (eds.) The mind's eye: cognitive and applied aspects of eye movement research, pp. 657–670. Elsevier, Amsterdam (2003)
4. Holmqvist, K., Wartenberg, C.: The role of local design factors for newspaper reading behaviour – an eye-tracking repspective. Daily newspaper layout-designer's prediction of reader's visual behaviour- a case study Lund University Cognitive Studies, 127. Lund: LUCS (2005)

5. Laarni, J.: Effects of color, font type and font style on user preferences. In: Stephanidis, C. (ed.) Adjunct Proceedings of HCI International 2003, pp. 31–32. Crete University Press, Heraklion (2003)
6. House, D., Bell, L., Gustafson, K., Johansson, L.: Child-directed speech synthesis: Evaluation of prosodic variation for an educational computer program. In: Proc. Eurospeech 1999, Budapest, Hungary (1999)
7. Johnson, W.L., Narayanan, S., Whitney, R., Das, R., Bulut, M., LaBore, C.: Limited domain synthesis of expressive military speech for animated characters. In: IEEE Speech Synthesis Workshop, Santa Monica, USA (2002)
8. Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G., Argyropoulos, V.: Experimentation on Spoken Format of Tables in Auditory User Interfaces. Universal Access in HCI. In: Proceedings of the 3d International Conference on Human-Computer Interaction (HCII-2005), Las Vegas, USA, vol. 8 (July 22-27, 2005)
9. Schröder, M.: Expressing degree of activation in synthetic speech. IEEE Transactions on Audio, Speech and Language Processing 14(4), 1128–1136 (2006)
10. Kramer, G. (ed.): Auditory Display: Sonification, Audification, and Auditory Interfaces. In: Proceedings Santa Fe Institute Studies in the Sci, Addison – Wesley, vol. 18 (1994)
11. Gorny, P.: Typographic semantics of Webpages Accessible for Visual Impaired Users, Mapping Layout and Interaction Objects to an Auditory Interaction Space. In: International Conference on Computer Helping with Special Needs, pp. 17–21 (2000)
12. Brewster, S.A., Raty, V.-P., Kortekangas, A.: Earcons as a Method of Providing Navigational Cues in a Menu Hierarchy. In: Proceedings of HCI'96, Imperial College, London, UK, pp. 167–183. Springer, Heidelberg (1996)
13. Mynatt, E.D.: Designing with auditory icons: how well do we identify auditory cues? In: Plaisant, C. (ed.) Conference Companion on Human Factors in Computing Systems, Boston, Massachusetts, United States, April 24–28, 1994, pp. 269–270. ACM Press, New York (1994)
14. Djennane, S.: 3D-Audio News Presentation Modeling. In: User Interfaces for All, pp. 280–286 (2002)
15. Duarte, C., Carriço., L.: Identifying adaptation dimensions in digital talking books. In: Proceedings of the 9th international conference on Intelligent user interface IUI '04 (2004)
16. Daisy Consortium http://www.daisy.org
17. ANSI/NISO.: Specifications for the digital talking book (2002) http://www.niso.org/standards/resources/Z39-86-2002.html
18. OASIS, Organization for the Advancement of Structured Information Standards http://www.oasis-open.org/home/index.php
19. Web Content Accessibility Guidelines, W3C http://www.w3.org/TR/WAI-WEBCONTENT/
20. Raman, T.V.: An Audio view of (LA)TEX Documents. In: Proceedings of the Annual Meeting. TUGboat, vol 13(3), pp. 65–70 (1992)
21. Hakulinen, J., Turunen, M., Rih, K-J.: 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, pp. 65–70 (1999)
22. Xydas, G., Argyropoulos, V., Karakosta, T., Kouroupetroglou, G.: An Open Platform for Conducting Psycho-Acoustic Experiments in the Auditory Representation of Web Documents. In: Proc. Conference ACOUSTICS (2004)
23. Stephanidis, C., Akoumianakis, D., Sfyrakis, M., Paramythis, A.: Universal accessibility in HCI: Process-oriented design guidelines and tool requirements. In: Proceedings of the 4th ERCIM Workshop on User Interfaces for All, Stockholm, Sweden (1998)

24. EDEaN.: Design for All Network of Excellence (IST-2001-38833) European Design for All e-accessibility Network (EDEaN) (2001) http://www.d4allnet.gr/
25. Europe's Information Society, Design for All http://europa.eu.int//information_society/policy/accessibility/dfa/index_en.htm
26. Xydas, G., Argyropoulos, V., Karakosta, T., Kouroupetroglou, G.: An Experimental Approach in Recognizing Synthesized Auditory Components in a Non-Visual Interaction with Documents. In: Proceedings of the 11th International Conference on Human-Computer Interaction (HCII2005), Las Vegas, Nevada SA, 22-27 July 2005, vol. 3, pp. 411–420 (2005)
27. Kilov, H.: Semantics in open hypermedia systems: Information modeling for document understanding. In: ECHT'94 Workshop on Open Hypermedia Systems Edinburgh, Scotland (1994)
28. Kerscher G.: Theory Behind the DTBook DTD. Daisy Consortium (2001) http://www.daisy.org/ publications/docs/theory_dtbook/theory_dtbook.html
29. Xydas, G., Kouroupetroglou, G.: 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 pp. 2247–2250 (2001)
30. Xydas, G., Kouroupetroglou, G.: Augmented Auditory Representation of e-Texts for Text-to-Speech Systems. In: Matoušek, V., Mautner, P., Mouček, R., Tauser, K. (eds.) TSD 2001. LNCS (LNAI), vol. 2166, pp. 134–141. Springer, Heidelberg (2001)
31. Fink, J., Kobsa, A., Nill, A.: Adaptable and Adaptive Information Provision for All Users, Including Disabled and Elderly People (1999) Available from http://citeseer.ist.psu.edu/fink99adaptable.html
32. Hjalmarsson, A.: Adaptive Spoken Dialog Systems. In: GSLT, Speech Technology 1 Closing Seminar (2005)
33. McTear, M.F.: Spoken dialogue technology: enabling the conversational user interface. ACM Comput. Surv. 34(1), 90–169 (2002)
34. Turunen, M., Hakulinen, J., Räihä, K., Salonen, E., Kainulainen, A., Prusi, P.: An architecture and applications for speech-based accessibility systems. IBM Syst. J. 44(3), 485–504 (2005)