Acoustic Modeling of Dialogue Elements for Document Accessibility

Pepi Stavropoulou[1,2], Dimitris Spiliotopoulos[1],
and Georgios Kouroupetroglou[1]

[1] Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
[2] Department of Linguistics, University of Ioannina,
GR-45110, Ioannina, Greece
{pepis,dspiliot,koupe}@di.uoa.gr

Abstract. Document-to-Audio accessibility assumes that all meaningful presentaion elements in the document, such as bold, italics, tables or bullets, should be properly processed and acoustically modeled, in order to convey the intended meaning to the listeners in a complete and adequate manner. Similarly, several types of documents may contain reported speech and dialogue content signaled through punctuation and other visual elements that require further processing before being rendered to speech. This paper explores such dialogue elements in documents, examines their actual indicators and their use, and investigates the most prominent methods for their acoustic modeling, namely the use of prosody manipulation and voice alternation. It further reports on a pilot experiment on the appropriateness of voice alternation as means for the effective spoken rendition of dialogue elements in documents. Results demonstrate a clear listener preference for the "multiple voice" renditions over the ones using a single voice.

Keywords: Acoustic modeling, document accessibility, dialogue, reported speech, Text to Speech synthesis, voice alternation, Document-to-Audio.

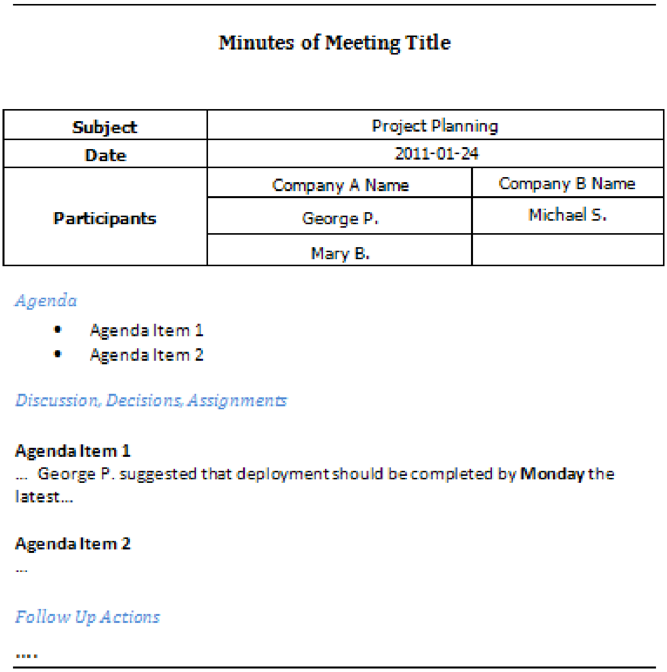
1 Introduction

Work on universal access to documents aims towards making document content accessible to the widest possible range of end users including people with disabilities. Proper adjustments to document layout and text formatting as well as utilization of different modalities are key means to accommodating users "with different abilities, requirements and preferences in a variety of contexts of use" [26]. Accordingly, Text to Speech (TtS) systems transfer document content to the acoustic modality making it accessible to the visually impaired, in eyes-busy situations, in spoken dialogue applications and so forth. For the transfer to be effective document metadata must be utilized and visual elements must be properly rendered to speech as part of a complete Document-to-Audio (DtA) process.

C. Stephanidis (Ed.): Universal Access in HCI, Part IV, HCII 2011, LNCS 6768, pp. 175–184, 2011.

2 DtA - Acoustic Modeling of Visual Elements in Documents

In transferring documents from the visual to the aural modality, elements optimized for vision need to be correctly identified, appropriately processed and subsequently vocalized in a manner that improves naturalness, aids comprehension, and minimizes listening effort. Visual presentation elements such as tables, paragraphs, headings and bullets convey semantic and pragmatic information critical for understanding the intended meaning of the text, thus necessitating the use of appropriate acoustic modeling of the underlying logical association between visual structure and the meaning itself. How does one read tables? What is an appropriate prosodic specification for bold or italics? How can bullets or quotation marks be acoustically perceived? Take the minutes of a meeting for instance. A brief inspection of a sample "meeting of minutes" document (Figure 1) reveals several visual elements that on the one hand pose certain challenges for the effective acoustic rendition, while on the other hand provide important information for accessing, interpreting and subsequently communicating the underlying semantic and pragmatic content to the listener.



Minutes of Meeting Title

Subject	Project Planning	
Date	2011-01-24	
Participants	Company A Name	Company B Name
	George P.	Michael S.
	Mary B.	

Agenda

- Agenda Item 1
- Agenda Item 2

Discussion, Decisions, Assignments

Agenda Item 1

... George P. suggested that deployment should be completed by **Monday** the latest...

Agenda Item 2

...

Follow Up Actions

....

Fig. 1. Sample "minutes of meeting" document

More specifically, these elements are:

- 1. Headings and subheadings indicating basic discourse segmentation and topic hierarchy. Discourse structure in general is marked through variation in prosodic parameters such as pitch range, intensity and preboundary lengthening, as suggested in a number of previous studies [7, 14, 27]. However, most synthesizers today disregard structure beyond the utterance level, making no use of meta-information on paragraph breaks or document sectioning.

- 2. Table structures. A significant amount of work has been devoted to the extraction and meaningful reconstruction of the logical relations that are implicit in the tabular layout [2, 3, 4, 11, 15, 18, 20, 28]. Once the underlying logical relations are extracted and reconstructed in a way suitable for non-visual modalities, the resulting semantic representation may be utilized through proper parameterization of Text to Speech synthesis [22, 23]. Previous work has demonstrated that appropriate prosody control and the use of earcons and spearcons improve naturalness, acceptance and listening effort especially in the case of complex table structures in which simple linearization techniques are proven inadequate [24].
- 3. Types of list structures such as bulletin. In a study of spoken lists [19] showed that appropriate use of rhythm and pitch decreases cognitive effort and aids recall. Xydas et al. [29] demonstrated listeners' preference for a combination of modifications in volume, pitch and use of earcons for the discrimination among bold letters, italics and bulletin respectively. In general, use of non speech sounds such as tones or beeps is considered a particularly effective way for introducing list items and vocalizing bullets [5].
- 4. Bold formatting indicating emphasis. In a way similar to the aural rendition of other metadata and visual emphatic events, pitch modification on the bold word and increase in volume have been utilized in cueing bold letters [26, 29].
- 5. Reported speech elements. While our example involves an instance of indirect reported speech ("George P. suggested that…") that is not typographically set off, there are many types of documents such as interviews, transcripts, proceedings and narratives that include an abundance of visual elements cueing the presence of direct speech and dialogue. Such elements and their subsequent effective vocalization constitute the focus of this paper.

In particular, this paper examines the appropriateness of synthetic voice alternation as a means for transferring reported direct speech and dialogue content in written documents to the acoustic modality. In the following sections, we first briefly analyze the basic types of reported speech along with the visual components and other cues used in documents for denoting them. Next, the most appropriate means for successfully rendering them to the aural modality are presented, namely prosody parameterization and voice alternation. A pilot psychoacoustic experiment comparing listeners' perception of a single voice rendition to their perception of the one using voice alternation is presented in sections 3 and 4. Major findings and directions for future work are discussed last.

3 Speech and Dialogue Elements in Text

Following Sinclair [21] there are two main ways for reporting one's words when writing, namely *quote structures*, also referred to as *direct speech*, and *report structures*, also referred to as *indirect speech*. In the first case, the speaker's exact words are reported, while in the second case there is no exact reproduction involved, rather certain changes apply on the original utterance’s grammar structure and content. Intuitively, the action of reporting another utterance constitutes a universal communicative and linguistic phenomenon, while at the same time a clear cut distinction between direct and indirect speech is found encoded in a good number of languages around the world [8].

Though indirect speech is indicated in written text solely on a lexico-syntactic basis through the use of reporting phrases such as “John said” along with changes in personal, temporal or locative references, direct speech is, in addition, cued through certain visual components applied directly onto the text or in the form of meta-information embedded in the source document, These components provide visual cues to the existence of direct speech and other dialogue elements such as dialogue turns. Following are the most common and widely used visual indications to speech and dialogue elements within written discourse:

Quotation mark pairs are most often used to delimit the beginning and end of direct speech. Depending on specific language conventions quotation marks may come in different forms such as single (‘’) or double (“”) inverted commas, double angle quotes («») or corner brackets. Speakers may also alternate between different forms, in order to denote the presence of a nested quote (a quote within a quote). The accompanying reporting phrase is placed outside the quotation marks and can be positioned at the beginning, at the end or within the quote structure; when the reported “voice” has already been established in context, the reporting phrase is often omitted.

When dialogue is reported, turn taking is indicated through the use of line or paragraph breaks. In dialogue inverted commas are often replaced by a quotation dash corresponding to a single dialogue turn. Furthermore, the name of the speaker may be used followed by colons (:) or a quotation dash (.-). This explicit mentioning of speakers’ names greatly simplifies the task of matching each speaker to the correct turn. Finally, in interviews dialogue is sometimes indicated through the use of bold letters for representing the interviewer’s utterance. Again turn taking is cued through line or paragraph breaks. Table 1 summarizes quoting styles and visual cues to reported speech providing examples for each one.

Table 1. Reported speech indicators

	Indirect Speech	**Direct Speech**
Deictic references and syntactic dependence	He complained ***that*** nobody came to ***his*** party	–
Reporting phrases	***He complained*** that nobody came	“Nobody came”, ***he complained***
Quotation marks	–	“..” / «…» / ‘…’ etc
Quotation dash	–	**-** Nobody came, he said.
Colons	–	**JUDGE:** Can the defendant please rise?
Line breaks	–	Line or paragraph breaks indicating speakers’ turns

At this point it should be noted that, while there is an abundance of visual cues that are more or less consistently used, the functional load of each element (e.g. the use of colons before enumeration or the use of inverted commas to denote irony and so forth) as well as the complexity of certain dialogues may pose several challenges to the identification of dialogue within text, including the identification of dialogue participants and the correct assignment of each turn to the respective participant. Addressing these issues, however, is beyond the scope of this paper.

Punctuation and other visual cues within written text often serve as a substitute of prosody in speech [9]. Subsequently, visual markers of reported speech such as quotes or line breaks should *in a broad sense* correspond to prosodic markers of reported speech in spoken discourse. Accordingly, in a study of informal conversations examining the relationship between reported speech and prosody in English, Klewitz and Cooper [13] found instances of reported speech to correlate with shifts in pitch range, intensity, speech rate and perceptually isochronous rhythmic patterns, as well as paralinguistic expressive qualities such as breathy or nasal voice. Their data further suggests that in dialogue distinct prosodic marking may be assigned to the different "voices" (i.e. interlocutors) reported, facilitating the hearer's task of keeping track of *who is speaking now*. In addition, Jansen [12] reported on a statistically significant expansion of overall pitch range of direct speech compared to both surrounding narrative segments as well as instances of indirect speech. Furthermore, direct speech was found to be more often preceded by stronger intonational breaks. In contrast, no statistically significant differences were reported between indirect speech and surrounding narration with regards to the prosodic parameters attested. Similar effect of pitch range has also been demonstrated for direct speech in Brazilian Portuguese [17]. In short, taking into account that prosody has been shown to function as a significant marker of discourse structure in general [1, 7, 10, 14, 16] (among others), and reported speech instances constitute more or less clearly demarcated discourse segments with particular discourse functions, the appropriate manipulation of prosodic parameters such as pitch range, intensity, lengthening and pausing is expected to play a key role in successfully transferring reported speech back to the aural modality.

In addition to prosody manipulation, in the process of rendering documents to speech, TtS systems can provide further means for signaling reported speech and dialogue. Intuitively, voice alternation – switching between different synthetic voices over the course of the interaction – is expected to be the most appropriate other medium for signaling multiple voices in written dialogue. In a similar vein, voice alternation has been used for landmarking and context setting in automated spoken dialogue systems. Association of different synthetic voices to specific dialogue states may improve system navigation, and increase user confidence and awareness of dialog progress.

The study presented in this paper focuses on the use of voice alternation for the effective acoustic rendition of direct speech and dialogue in particular. Direct speech was preferred over indirect speech as it is more explicitly related to the existence of a different – other than the narrator's – voice, bringing forth a distinct speech situation and reproducing the perspective of the original utterance. As such, it is better suited for switching to a different voice. In contrast, indirect speech is more tightly integrated within the embedding context, maintaining the perspective of the narrator and lacking syntactic independence and expressive properties (e.g. use of exclamation marks). Accordingly, it is not distinguished typographically and speakers do not tend to prosodically mark it [12]. Furthermore, dialogue was suited for testing more than two voices as well as assessing the effect of voice alternation on facilitating comprehension. Written dialogue often lacks explicit “tagging” of interlocutors through the use of reporting phrases (e.g. “he said”). In effect line breaks are often the only means left signaling turn taking, rendering their effective transfer to the acoustic modality crucial for determining *speakership*, i.e. understanding “who is speaking now”.

4 Experimental Setup

In accordance with the above a pilot experiment was carried out comparing two versions of synthetic speech renditions, one using voice alternation and one in which only a single synthetic voice is used. The two renditions were compared in terms of both objective and subjective criteria. As part of the objective evaluation, participants were asked to answer a set of questions aiming to assess their degree of comprehension and appropriate alignment of turns to interlocutors. For the subjective evaluation participants were asked to assign a score on a Likert scale ranging from 1 to 5 evaluating overall impression, acceptance, ease of comprehension and naturalness of each rendition.

Materials consisted of three texts rendered in both – single voice and multiple voices – conditions. The first text was a narrative passage with two characters, the narrator being one of the characters. Dialogue turns were marked by quotation dashes, line breaks and reporting phrases. Reporting phrases were either omitted or placed right before the quote. The second text was also a narrative passage with two characters; only this time the narrator was not a character in the story, and thus there were three voices involved in total. Dialogue turns were marked through double inverted commas, line breaks and reporting phrases. The reporting phrases were either omitted, placed in the beginning, at the end or in the middle of the quote. In the later case, the reported speaker’s dialogue turn was interrupted by the narrator’s voice. The third text was a “transcripts of trial” document with three voices and no narrator. Dialogue turns and speakers’ identity were marked explicitly through the use of the speaker’s name followed by colons at the beginning of each turn. Thus no ambiguity regarding speakership arose. For the multiple voice rendition of the third text, each speaker’s name was mentioned only once when it first occurred and was afterwards omitted. In other words, once the mapping between speaker and synthetic voice had been established, the identification function performed by the mentioning of the speaker’s name at the beginning of each dialogue turn was now fulfilled through switching to a different voice, in an attempt to achieve faster and still intelligible interaction.

The materials were presented to two groups of four first-time listeners. Each group was presented with the materials in reverse order to ensure that order or memory effects were factored out and did not bias the results. For the narrative passages the following procedure was followed: After having listened to one condition, subjects answered a set of comprehension questions regarding text’s content. For each answer they further provided a degree of confidence ranging from 1 to 5, with 5 being “absolutely certain that my answer is correct” and 1 being “completely unsure”. In addition, they again assigned a score on a scale from 1 to 5 grading acceptance, naturalness and listening effort of the rendition as a whole. Next, they listened to the second condition. At the end they declared their preference between the two renditions, ranked each rendition for overall impression and made any other comments that they considered helpful. For the trial transcripts no comprehension questions were asked, as the structure of the document unambiguously determined speakership.

5 Results

Overall, subjects preferred the multiple voice renditions to the single voice ones. They considered the former to be more appropriate and easier to understand. More specifically, the multiple voice renditions were preferred in 21 out of 24 cases in total. One particular listener declared a preference on the use of a combination of features found in each rendition separately, namely the explicit use of the interlocutor's name before each turn along with the switch to a different voice.

Figure 2 illustrates listeners' ranking of each rendition as far as overall impression, listening effort and naturalness are concerned. Ranking scale ranged from 1 to 5, 5 being the optimal rank. As can be seen from figure 2, the multiple voice renditions scored higher, achieving an average 0.8 points improvement on all criteria examined. Furthermore, subjects proclaimed a higher degree of confidence with regards to their feedback in the comprehension task.

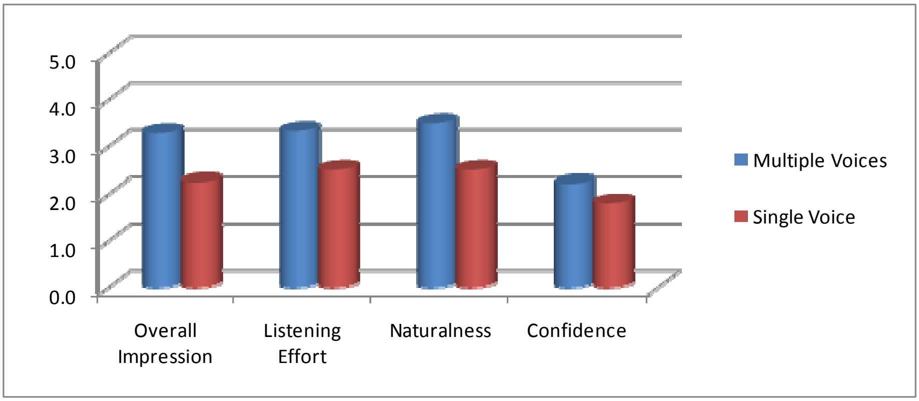


Fig. 2. Subjective evaluation criteria

Subjective evaluation results were indeed corroborated by subjects' performance on the comprehension task. The subjects' answers were 94% accurate in the case of the multiple voice renditions contrary to the single voice renditions for which accuracy was limited to 75%. Response times were also 10-20% faster for the multiple voice renditions. Still, even the multiple voice renditions were considered inadequate in some instances: more specifically, 25% of the total answers to the question "Do you consider this type of rendition acceptable?" were *NO*. For the single voice renditions, there was only one *YES* answer.

6 Discussion – Future Work

The results of the experiment presented here confirm the main hypothesis that voice alternation can effectively be used to model dialogue elements in documents, minimizing listening effort and facilitating comprehension. The utilization of different synthetic voices is particularly useful for determining speakership, that is the alignment of interlocutors to particular turns. Speakership itself is a particular instance of footing [6] that places a significant load on cognition.

Nevertheless, voice alternation alone sometimes fails to meet user expectations and reach level of acceptance. The latter necessitates the use of other means such as prosody control for improving the quality and legibility of the rendition. As noted, human speakers use pausing, alter pitch or employ paralinguistic devices – among others – to signal reported speech. Thus, further experimentation is required for determining the most appropriate combination of prosodic parameterization and voice alternation.

On a final note, modeling of reported speech may also prove to be a promising area for the study of emotional speech as part of affective computing. In particular, several reporting phrases used in quote structures may indicate the manner of speaking as well as the speaker's emotional state. Verbs and phrases such as "scream", "yell", "whisper", "plead", "reply angrily" or "cry" can serve as cues for detecting and simulating emotions in an effort for more natural and expressive interaction. In fact, the human paradigm suggests that human narrators employ certain paralinguistic devices in order to convey the reported speaker's emotional state and attitude [13].

7 Conclusion

The empirical evidence presented in this paper favors the use of voice alternation for the acoustic modeling of dialogue elements in written documents, in an attempt to make document content universally accessible. When speech reported within written text is rendered back to the aural modality, there should be certain acoustic cues to the beginning and end of the reported speech stretch as well as to each dialogue turn, in order for the listener to correctly identify and comprehend the intended dialogue structure. Voice alternation is one of them. Appropriate prosody modeling is another medium that certainly calls for further research. In any case, prior to any acoustic modeling, the correct identification of reported speech stretches within text as well as the appropriate assignment of "speakership" require extensive document preprocessing that constitutes a demanding and important task in its own right.

References

1. Den Ouden, H., Noordman, L., Terken, J.: The prosodic realization of organizational features of texts. In: Proc. Speech Prosody 2002, pp. 543–546 (2002)
2. Chen, H.-H., Tsai, S.-C., Tsai, J.-H.: Mining tables from large scale html texts. In: Proceedings of the 18th International Conference on Computational Linguistics, Saarbrucken, Germany (2000)
3. 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)

4. 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)
5. Fröhlich, P.: Increasing Interaction Robustness of Speech-enabled Mobile Applications by Enhancing Speech Output with Non-speech Sound. In: Proc. ROBUST 2004, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction, Norwich, England (August 2004)
6. Goffman, E.: Forms of talk. Basil Blackwell, Oxford (1981)
7. Grosz, B., Hirschberg, J.: Some intonational characteristics of discourse structure. In: Proceedings of the 2nd International Conference on Spoken Language Processing, Banff, Canada, pp. 429–432 (1992)
8. Haberland, H.: Reported Speech in Danish. In: Coulmas, F. (ed.) Direct and Indirect Speech. Trends in Linguistics, Studies and Monographs, vol. 31. Mouton de Gruyter, Berlin (1986)
9. Halliday, M.A.K.: Spoken and written language. Deakin University Press, Geelong (1985)
10. Herman, R.: Intonation and discourse structure in English: Phonological and phonetic markers of local and global discourse structure. PhD Thesis (1998)
11. 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)
12. Jansen, W., Gregory, M.L., Brenier, J.M.: Prosodic correlates of directly reported speech: Evidence from conversational speech. In: Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Banks, NJ, pp. 77–80 (2001)
13. Klewitz, G., Couper-Kuhlen, E.: Quote-unquote? The role of prosody in the contextualization of reported speech sequences. Pragmatics 9(4), 459–485 (1999)
14. Lehiste, I.: Some Phonetic Characteristics of Discourse. Studia Linguistica 36(2), 117–130 (1982)
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. Nakatani, C., Hirschberg, J., Grosz, B.: Discourse Structure in Spoken Language. Studies on Speech Corpora (1995)
17. Oliveira, M., Cunha, D.A.C.: Prosody as Marker of Direct Reported Speech Boundary. In: Speech Prosody 2004, Nara, Japan (March 23-26, 2004)
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. Pitt, I., Edwards, A.: An Improved Auditory Interface for the Exploration of Lists. ACM Multimedia 1997, 51–61 (1997)
20. 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)
21. Sinclair, J.: Collins Cobuild English Grammar. Harper Collins, London (2002)
22. 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)
23. Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G., Argyropoulos, V., Ikospentaki, K.: Auditory Universal Accessibility of Data Tables using Naturally Derived Prosody Specification. Univ. Access Inf. Soc. 9(2), 169–183 (2010)

24. Spiliotopoulos, D., Stavropoulou, P., Kouroupetroglou, G.: Acoustic Rendering of Data Tables using Earcons and Prosody for Document Accessibility. In: Stephanidis, C. (ed.) UAHCI 2009. LNCS, vol. 5616, pp. 587–596. Springer, Heidelberg (2009)
25. Stephanidis, C., Akoumianakis, D., Sfyrakis, M., Paramythis, A.: Universal accessibility in HCI: Process-oriented design guidelines and tool requirements. In: Stephanidis, C., Waern, A. (eds.) Proceedings of the 4th ERCIM Workshop on User Interfaces for All, Stockholm, Sweden, October 19-21 (1998)
26. Truillet, P., Oriola, B., Nespoulous, J.L., Vigoroux, N.: Effect of Sound Fonts in an Aural Presentation. In: 6th ERCIM Workshop, UI4ALL, pp. 135–144 (2000)
27. Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., Price, P.: Segmental durations in the vicinity of prosodic phrase boundaries. Journal of the Acoustical Society of America 91(3), 1707–1717 (1992)
28. 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)
29. Xydas, G., Argyropoulos, V., Karakosta, T., Kouroupetroglou, G.: An Experimental Approach in Recognizing Synthesized Auditory Components in a Non-Visual Interaction with Documents. In: Proc. Human-Computer Interaction - HCII 2005 (2005)