Diction Based Prosody Modeling in Table-to-Speech Synthesis

Dimitris Spiliotopoulos, Gerasimos Xydas and Georgios Kouroupetroglou

University of Athens
Department of Informatics and Telecommunications
{dspiliot, gxydas, koupe}@di.uoa.gr

Abstract. Transferring a structure from the visual modality to the aural one presents a difficult challenge. In this work we are experimenting with prosody modeling for the synthesized speech representation of tabulated structures. This is achieved by analyzing naturally spoken descriptions of data tables and a following feedback by blind and sighted users. The derived prosodic phrase accent and pause break placement and values are examined in terms of successfully conveying semantically important visual information through prosody control in Table-to-Speech synthesis. Finally, the quality of the information provision of synthesized tables when utilizing the proposed prosody specification is studied against plain synthesis.

1. Introduction

Text material is primarily optimized for visual presentation by embedding several visual components. These range from simple "bold", "italic", or coloured letters directives to more complex ones such as those that define a spatial layout (tables, forms, etc.). Transferring a structure from the visual modality to aural is by no means an easy task. For example, tables are characterized by many qualitative and quantitative aspects that should be taken into consideration since successful vocalization is greatly affected by them. Most common approaches tend to linearize two-dimensional elements prior to their acoustic presentation. However, most of the semantic meaning of their enclosed text is implicit to the visual structure. This work is concerned with the vocalization of *data tables*, the most widely used two-dimensional structure in documents.

Data tables are categorized into *simple* and *complex*. Simple tables have up to one row and one column of header cells, while complex ones contain more than one level of logical row or column headers. This means that header and data cells can be expanded to encompass more than one row or column forming nested tables. Hence, complex tables can be thought of as three-dimensional structures [1], compared to the two-dimensional simple data tables. The third dimension of the semantic structure is embedded inside the two dimensional visual structure.

Complex visual structures bear a distinct association between the physical layout and the underlying logical structure [2]. Previous works show that appropriate mark-up can be used to assign logical structure to table cells [3] and suggest additional

mark-up annotation to existing tables for adding context in order to improve navigation [4]. Other suggestions include automated approaches for retrieval of hierarchical data from HTML tables [5] or transforming tables into a linear, easily readable form by screen readers [6]. The semantic structure of HTML tables can be used to aid their navigation and browsing [7]. However, since the problem is addressed on the visual level, the major handicap of the linearized transformation approach to the actual spoken form remains.

Previous studies focusing on the speech representation show that one-dimensional elements such as bold and italic letters, gain their acoustic representation by the use of prosody control [8][9], while others deal with the acoustic representation of linear visual components using synthesized speech [10].

In [11], a script-based open platform for rendering meta-information to speech using a combination of prosody modifications and non-speech audio sounds has been presented. However, the exploitation of synthetic speech prosody parameterization necessitates the utilization of the human natural spoken rendition for tables. Our motivation is to examine and model the natural speech specification of table meta-information by analyzing spoken paradigms from human readers in order to aid speech synthesis.

Using the acquired analyzed speech data from the most preferred human spoken renditions [12], we derived a prosody model concerning phrase accents and pause breaks. In the following sections we present the resulted prosody specification as well as the psychoacoustic experiments on the corresponding speech-synthesized data tables.

2. Requirements for Human Spoken Rendition of Tables

Recent research shows that advanced *browsing* techniques may be used to create table linearization that analyses the implicit structural information contained in tables so that it is conveyed to text (and consequently to speech) by navigation of the data cells [1]. It is obvious that complex data tables are much more complicated to browse since they may have multiple levels of structures in a hierarchical manner. Pure linear as well as intelligent navigation for the tables are accounted for in this work. Intelligent navigation is a process that takes place before the actual synthesis and, for the case of simple tables, results in header-data cell pair linearization while, for the case of complex tables, a decomposition revealing the respective sub-tables takes place.

Natural spoken rendition required human subjects as readers of data tables. Selecting appropriate sample tables for rendering to natural speech required several factors to be taken into consideration. Table wellformedness is ensured through certain compliance to W3C table specification [13] and W3C WAI recommendations [14]. In this work, only pure *data tables* are considered, that is tables used solely to convey information comprising of pure data of certain relationship, not used for page layout and without any visual enhancements or styles applied. Moreover, for human spoken rendition the so-called *genuine tables* [15], that is tables where the two dimensional grid is semantically significant are considered.

3. The Prosody Model Specification

The human spoken table rendition feedback from the listeners has led to the design of an initial specification for prosodic modeling of simple and complex table structures for synthetic speech. For prosody markup, the ToBI annotation model [16] conventions were used as a means of qualitative analysis. Rules pertaining to phrase accent and boundary tone assignment (L- describing the fall in pitch at the end of spoken data cells, H- describing the rise in pitch) were constructed according to the experimental data. Moreover, pause break parameters were set up according to the preferred natural language rendition adjusted by the listeners' proposed modifications.

Linear browsing was modeled as shown in Table 1, the phrase accent model specification describing position and type of phrase accent tone according to conditions that apply for either simple or complex table. Table 2 shows the respective specification derived for intelligent browsing of simple and complex tables. An obvious observation is the simplicity of the model for intelligent browsing as a result of semantic resolution prior to vocalization.

Table 1. Simple and complex table (linear browsing) phrase accent specification

position	tone	conditions (simple table)	conditions (complex table)
Header cell	L-	row-final	row-final
Header cell	H-	not-row-final	not-row-final
Data cell	H-	row-penultimate	row-initial AND row-is-nested-table-final
Data cell	L-	not-row-penultimate	(row-initial AND row-is-not-nested-table-final) OR row-final

Table 2. Simple and complex table (intelligent browsing) phrase accent specification

position	tone	conditions
Header cell	L-	not-part-of-pair
Header cell	H-	row-final
Data cell	L-	(none)

Pause breaks have been assigned at the end of cells and rows as absolute values in milliseconds, calculated as multiples of the shortest pause selected according to the experimental data analysis.

Table 3. Simple table (linear) pause breaks

position	ms	multiplier
Cell	600	x1.00
Row	900	x1.50

Table 4. Simple table (intel.) pause breaks

position	ms	multiplier
header cell	200	x1.00
data cell	500	x2.50
Row	750	x3.75

Tables 3 and 4 show the actual values and multiplier factors for linear and intelligent browsing for simple tables, while tables 5 and 6 the respective data for complex tables.

Table 5. Complex table (linear) pause breaks

position	ms	mult.
header row	750	x2.50
data row	750	x2.50
Table header cell	300	x1.00
Nested table header cell	600	x2.00
Data cell	525	x1.75

Table 6. Complex table (intel.) pause breaks

position	ms	mult.
Nested table header cell	750	x3.75
header cell	200	x1.00
data cell	750	x3.75
row	1250	x6.25

4. Experiments

We carried out a set of psychoacoustic experiments using a group of 10 experienced listeners, 21-24 years old. They were asked to take part in a formal listening to known and unseen synthesized spoken tables. We used the DEMOSTHeNES Document-to-Audio platform [11] to host the derived prosody specification by the means of two auditory scripts for the simple and the complex table respectively. Table de-compilation to logical layer was followed by the application of the above mentioned prosodic phrase accent and pause break parameters. The selected tables were rendered using both plain speech synthesis (without parameterization) and enhanced one (with prosody model parameterization) by the newly acquired parameterization in order to experiment with the proposed Table-to-Speech approach.

The aim of the first experiment was a comparative subjective analysis of plain and enhanced speech synthesis renditions of already known example tables (as introduced in [12]) in order to measure the impact of the new prosodic adjustment model. The second experiment involved unseen test tables that were used to determine the competence of the prosodic model measured by the resulting understanding of table data, as well as subjective listener input for each rendition described by the model. The tables that were selected for this experiment were similar to the reference simple and complex tables described in the literature. The text data in both tables were provided in Greek, which is the native language of all listeners.

The subjects listened to the synthesized tables in random order and were asked to assert their understanding of the data in the range of 1 (lowest) to 5 (highest). The results have shown that the parameterization has led to significant increase in their understanding of the table data semantic structure (Fig. 1).

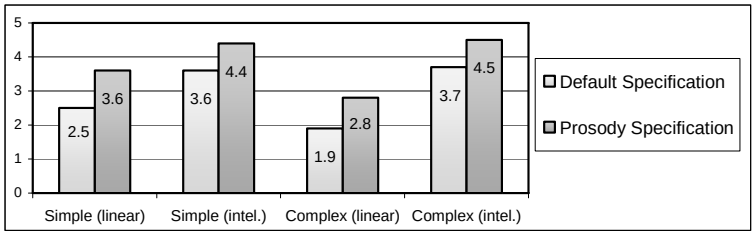

Fig. 1. Side-by-side comparison of prosodic model against default specification

The second part of this included synthesized spoken formats of unseen Tables 7 and 8 much larger than the initial experiment ones. The simple table linear spoken format included repeats of the header row, a usual practice for larger tables that contain several rows of data. The translation of the content to English is provided by italicized text in square brackets.

Table 7. The larger simple data table contains one header row and eight data rows.

Πρόγραμμα μεταδόσεων αθλημάτων από το ραδιόφωνο.
[*Radio-transmitted sports schedule*].

Ημέρα [***Day***]	**Άθλημα** [***Sport***]	**Έναρξη** [***Start time***]	**Λήξη** [***End time***]
Δευτέρα [*Monday*]	Στίβος [*Athletics*]	11.00	18.00
Τρίτη [*Tuesday*]	Τένις [*Tennis*]	20.00	23.00
Τετάρτη [*Wednesday*]	Στίβος [*Athletics*]	09.00	18.00
Πέμπτη [*Thursday*]	Γυμναστική [*Gymnastics*]	16.00	21.00
Παρασκευή [*Friday*]	Πόλο [*Water polo*]	12.00	15.00
Σάββατο [*Saturday*]	Γυμναστική [*Gymnastics*]	16.00	18.00
Σάββατο [*Saturday*]	Ποδόσφαιρο [*Football*]	21.00	23.00
Κυριακή [*Sunday*]	Στίβος [*Athletics*]	09.00	12.00

During the second experiment, the subjects were asked to listen to each synthesized rendition and answer carefully selected key questions (asked beforehand and chosen in random order) designed to retrieve data from the tables. The listeners were asked to look for specific information and expected to recognize nested tables, data spanning several rows or columns, etc, in order to answer. Moreover at the end of each session they were asked to provide subjective opinion for overall impression on the quality of rendition, the listening effort required to understand each table, and their acceptance.

Table 8. The complex data table contains three nested sub-tables

Πόλεις και ο καιρός τους τις επόμενες μέρες.
[*Cities and their weather for the following days*]

	Δευτέρα [*Monday*]	**Τρίτη [*Tuesday*]**	**Τετάρτη [*Wednesday*]**	**Πέμπτη [*Thursday*]**	**Παρασκευή [*Friday*]**
Αθήνα [*Athens*]					
Θερμοκρασία [*Temperature*]	23	24	26	22	18
Άνεμος [*Wind*]	Βορειοδυτικός [*Northwest*]	Δυτικός [*West*]	Νοτιοδυτικός [*Southwest*]	Βορειοδυτικός [*Northwest*]	Βόρειος [*North*]
Θεσσαλονίκη [*Salonika*]					
Θερμοκρασία [*Temperature*]	16	17	20	16	13
Άνεμος [*Wind*]	Βόρειος [*North*]	Βόρειος [*North*]	Δυτικός [*West*]	Βόρειος [*North*]	Βορειοδυτικός [*Northwest*]
Πάτρα [*Patra*]					
Θερμοκρασία [*Temperature*]	19	22	23	20	19
Άνεμος [*Wind*]	Βορειοδυτικός [*Northwest*]	Δυτικός [*West*]	Νότιος [*South*]	Νοτιοδυτικός [*Southwest*]	Νοτιοδυτικός [*Southwest*]

Figure 2 shows overall impression (5 = excellent, 1 = bad) of synthesized speech rendering of tables as well as listening effort needed by the listeners in order to answer the key questions (5 = no meaning understood with any feasible effort, 1 = no effort required). It is worth mentioning that half of the listeners were unhappy with linear rendition of simple table while 8 out of 10 were unable to understand the linear rendition of the complex table. This shows that, for linear navigation, prosody control fails to replace semantic structure when that is completely lost, less so for simpler tables where some of it may be retained.

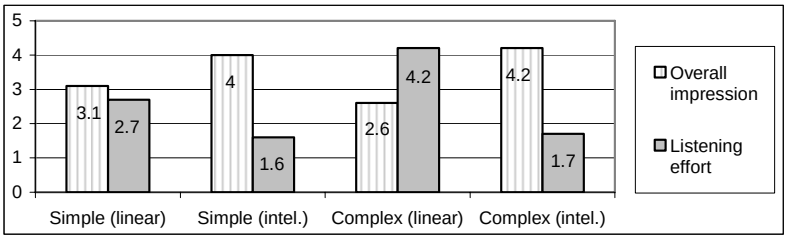

Fig. 2. Overall impression (higher=better) and listening effort (higher=worse)

It is obvious that linear reading of complex tables really failed to render the semantic relationship of the data understandable, which was the case for the natural

speech rendition during the initial experiments as well. However, the prosody model worked successfully for the other cases, the result being improvement in acoustic representation as well as reduced effort.

As an overall assessment of the results from these experiments, it can be deducted that the prosodic model provided a promising approach to modeling visual structures and to identify potential implementation issues in Table-to-Speech synthesis from real speech derived data. Navigation manner makes a strong impact on the final result and for that it should be pursued. Furthermore, it is concluded that by careful prosody modeling, a degree of semantic structure essence is retained in the resulting synthesized tables, thus making the content easier for the listener to comprehend. Finally, there is strong indication of several structural elements (e.g. rows, header-data cell pairs) that contain semantic importance for data understanding, and can be used by the synthesis system.

5. Conclusions

We presented an experimental study of vocalizing data tables. A set of prosodic parameters was derived by natural speech data that were analyzed in terms of phrase accent tones and pauses, clearly illustrating consistency against cell content and visual structure. The deducted specification formed the basis for a prosody model presented in this work that was used for automated rendering through synthetic speech. Experienced listeners through a formal acoustical assessment examined the generated utterances in cases of simple and complex tables. It was shown that such prosody modeling approach can successfully lead to improved understanding of synthesized speech rendition of tables, eventually conveying semantically important visual information to speech by prosody control.

Direct comparison of the prosody model aided synthetic speech against the default parameters used by a TtS system revealed the fact that certain semantic information can be carried from the visual structure to the spoken output through the use of phrase accent and pause break parameters.

It is concluded that further investigation should be granted to this area, especially in terms of determining which prosodic features have the most significant role in conveying semantic content. Furthermore the positive results encourage further analysis on the real speech data including additional parameters such as speech rate and duration.

Acknowledgments

The work described in this paper has been partially supported by the HERACLITUS project of the Operational Programme for Education and Initial Vocational Training (EPEAEK) of the Greek Ministry of Education under the 3rd European Community Support Framework for Greece. Special thanks to M. Platakis, O. Kostakis, P. Kollias, G. Marinakis, P. Karra, E. Kouvarakis, and L. Perellis for their participation in the psychoacoustic experiments described in this work.

References

1. Pontelli, E., Xiong, W., Gupta, G., & Karshmer, A. (2000). A Domain Specific Language Framework for Non-visual Browsing of Complex HTML Structures. *Proc. ACM Conf. Assistive Technologies - ASSETS 2000*, 180-187.
2. Ramel, J-Y., Crucianou M., Vincent, N., & Faure, C. (2003). Detection, Extraction and Representation of Tables. Proc. *7th Int. Conf. Document Analysis and Recognition - ICDAR 2003.* 374-378.
3. Hurst, M., & Douglas, S. (1997). Layout & Language: Preliminary Experiments in Assigning Logical Structure to Table Cells. *Proc. 4th Int. Conf. Document Analysis and Recognition – ICDAR 2003*, 1043-1047.
4. Filepp, R., Challenger, J., & Rosu, D. (2002). Improving the Accessibility of Aurally Rendered HTML Tables. *Proc. ACM Conf. on Assistive Technologies - ASSETS 2002*, 9-16.
5. Lim, S., & Ng, Y. (1999). An Automated Approach for Retrieving Hierarchical Data from HTML Tables. *Proc. 8th ACM Int. Conf. Information and Knowledge Management - CIKM 1999*, 466-474.
6. Yesilada, Y., Stevens, R., Goble, C., & Hussein, S. (2004). Rendering Tables in Audio: The Interaction of Structure and Reading Styles. *Proc. ACM Conf. Assistive Technologies - ASSETS 2004*, 16-23.
7. Pontelli, E., Gillan, D., Xiong, W., Saad, E., Gupta, G., & Karshmer, A. (2002). Navigation of HTML Tables, Frames, and XML Fragments. *Proc. ACM Conf. on Assistive Technologies - ASSETS 2002*, 25-32.
8. Xydas, G., Argyropoulos, V., Karakosta, T., & Kouroupetroglou, G. (2005). An Experimental Approach in Recognizing Synthesized Auditory Components in a Non-Visual Interaction with Documents. *Proc. Human-Computer Interaction - HCII 2005*, to appear.
9. Xydas, G., Spiliotopoulos D., & Kouroupetroglou, G. (2003): Modeling Emphatic Events from Non-Speech Aware Documents in Speech Based User Interfaces. *Proc. Human-Computer Interaction - HCII 2003*, *Theory and Practice*, 2, 806-810.
10. Raman, T. (1992). An Audio View of (LA)TEX Documents, *TUGboat, Proc. 1992 Annual Meeting*, 13, 3, 372-379.
11. Xydas, G., & Kouroupetrolgou, G. (2001). Text-to-Speech Scripting Interface for Appropriate Vocalisation of E-Texts. *Proc. 7th European Conf. Speech Communication and Technology - EUROSPEECH 2001*, 2247-2250.
12. Spiliotopoulos D., Xydas, G., Kouroupetroglou, G., and Argyropoulos, V., (2005), "Experimentation on Spoken Format of Tables in Auditory User Interfaces". Universal Access in HCI, *Proc. HCI International 2005: The 11th International Conference on Human-Computer Interaction (HCII-2005)*, 22-27 July, 2005, Las Vegas, USA, to appear.
13. Raggett, D., Le Hors, A., & Jacobs, I. (1999). Tables, HTML 4.01 Specification, W3C Recommendation, *http://www.w3.org/TR/REC-html40*
14. Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). Web Content Accessibility Guidelines 1.0, W3C Recommendation, 5 May 1999, *http://www.w3.org/TR/WAI-WEBCONTENT/*
15. Penn, G., Hu, J., Luo, H., & McDonald, R. (2001) Flexible Web Document Analysis for Delivery to Narrow-Bandwidth Devices, *Proc. 6th Int. Conf. on Document Analysis and Recognition - ICDAR 2001*, 1074-1078.
16. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A Standard for Labeling English Prosody. *Proc. Int. Conf. Spoken Language Processing - ICSLP-92*, 2, 867-870.