

Auditory universal accessibility of data tables using naturally derived prosody specification

Dimitris Spiliotopoulos · Gerasimos Xydas ·
Georgios Kouroupetroglou · Vasilios Argyropoulos ·
Kalliopi Ikospentaki

© Springer-Verlag 2009

Abstract Text documents usually embody visually oriented meta-information in the form of complex visual structures, such as tables. The semantics involved in such objects result in poor and ambiguous text-to-speech synthesis. Although most speech synthesis frameworks allow the consistent control of an abundance of parameters, such as prosodic cues, through appropriate markup, there is no actual prosodic specification to speech-enable visual elements. This paper presents a method for the acoustic specification modelling of simple and complex data tables, derived from the human paradigm. A series of psychoacoustic experiments were set up for providing speech properties obtained from prosodic analysis of natural spoken descriptions of data tables. Thirty blind and 30 sighted listeners selected the most prominent natural rendition. The derived prosodic phrase accent and pause break placement vectors were modelled using the ToBI semiotic system to successfully convey semantically important visual

information through prosody control. The quality of the information provision of speech-synthesized tables when utilizing the proposed prosody specification was evaluated by first-time listeners. The results show a significant increase (from 14 to 20% depending on the table type) of the user subjective understanding (overall impression, listening effort and acceptance) of the table data semantic structure compared to the traditional linearized speech synthesis of tables. Furthermore, it is proven that successful prosody manipulation can be applied to data tables using generic specification sets for certain table types and browsing techniques, resulting in improved data comprehension.

Keywords Data tables · Universal accessibility · Acoustic rendition · Auditory interfaces · Text-to-speech

1 Introduction

Text material is primarily optimized for visual presentation by embedding various visual components. These range from simple formatting elements (e.g., “bold”, “italic”, or coloured letters) to more complex ones, such as those that define a spatial layout (e.g., tables, forms, etc.). Transferring a structure from the visual modality to the aural one and retaining the associated semantic relations between the enclosed data involve specific assumptions. These assumptions include the structure characteristics (visualization, size, type, etc.) and the data relations.

Linear textual information carried by plain text documents can be straightforwardly rendered to speech by text-to-speech systems or reading applications (such as screen readers or other types of web page and document readers). On the other hand, complex visual document structures that

D. Spiliotopoulos (✉) · G. Xydas · G. Kouroupetroglou
Department of Informatics and Telecommunications, University
of Athens, Panepistimiopolis, Ilisia, 15784 Athens, Greece
e-mail: dspiliot@di.uoa.gr

G. Xydas
e-mail: gxydas@di.uoa.gr

G. Kouroupetroglou
e-mail: koupe@di.uoa.gr

V. Argyropoulos
Department of Special Education, University of Thessaly,
Argonafton & Filellinon St., 38221 Volos, Greece
e-mail: vassargi@sed.uth.gr

K. Ikospentaki
Cognitive Science Laboratory, University of Athens,
Panepistimiopolis, Ilisia, 15784 Athens, Greece
e-mail: kikospe@phs.uoa.gr

yield non-linear information are not perceptually transferred in speech-oriented user interfaces unless the structural content is transferred along. Tables, the most widely used multi-dimensional visual structure in documents, feature many qualitative (such as type, design, complexity) and quantitative (such as size, amount of cell data) aspects that should be taken into consideration, since they greatly affect successful vocalization. Most common approaches tend to linearize these two-dimensional elements prior to their acoustic presentation. However, most of the semantic meaning of their enclosed text is implicit to the visual structure.

Table is a generally used term to denote certain structural layout. The World Wide Web Consortium (W3C) provides a wide range of recommendations and guidelines to make the content that is written for the web “accessible to a wider range of people with disabilities, including blindness and low vision, deafness and hearing loss, learning disabilities, cognitive limitations, limited movement, speech difficulties, photosensitivity and combinations of these” [3]. The guidelines identify two distinct types of table structures, the ones that are used to organize data (data tables), and the ones that are used to create a visual layout of the page (layout tables). According to the W3C Web Content Accessibility Guidelines 1.0, tables should only be used to mark up truly tabular information [4].

Data tables can be classified 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 [19], compared to the two-dimensional simple data tables. The third dimension of the semantic structure is embedded inside the two-dimensional visual structure. The term *genuine tables* is sometimes used as a definition for tables where the two-dimensional grid is semantically significant [15].

The aural rendition of tables constitutes a hard task because of the difficulty in accessing the semantic information implied in the visual structure of tables. Complex visual structures bear a distinct association between the physical layout and the underlying logical structure [23]. The columns and rows of a table represent the logical connections [24]. Hurst [8] presented a table theory that “views the table as a presentation of a set of relations holding between organized hierarchical concepts”. Previous works also show that information about the semantic structure of HyperText Markup Language (HTML) tables can be used to aid navigation and browsing of such visual components [18]. Earlier 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 [16]. Appropriate markup can be used to assign logical structure arrangement to table cells [9], while navigation can be improved by additional markup annotation to add context to existing tables [7]. Other suggestions include automated approaches for retrieval of hierarchical data from HTML tables [13]. Smart browsers are used to access critical information for use in reading tables, whereas linearization techniques are employed for transforming tables into a more easily readable form by screen readers [35]. Table browsing techniques include the use of Conceptual Graphs for the classification of header and data cells using Hidden Markov Models for identification [10] as well as systems that de-compile tables into discrete HTML documents using an HTML index for navigation [14].

Tables can be processed by identifying their dimension, which is directly proportional to the complexity, and therefore deriving the logical grid [6]. 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—or similar complex visual structures, such as frames—either automatically or manually [17]. However, since the problem is addressed at a visual level, the transfer of the linearized visual structure to the actual spoken form remains problematic.

The W3C actively works towards providing resources for defining several markup languages for applications supporting speech input and output, collectively mentioned as the W3C Speech Interface Framework [11]. Recommendations for acoustic rendition of web documents by synthetic speech include the Speech Synthesis Markup Language (SSML) [2], Aural Cascaded Style Sheets (ACSS) [12], and Cascading Style Sheets Level 3 (CSS) [21]. These recommendations, nevertheless, only apply to the creation of documents where the author is expected to embrace the acoustic representation of the document, providing no means for managing single or multi-dimensional visual structures in synthetic speech. Studies focusing on the speech representation show that one-dimensional visual elements, such as bold and italic in letters or words, gain their acoustic representation by the use of prosody control [33], while others deal with the acoustic representation of linear visual components (such as typesetting for mathematical representations in LaTeX) using synthesized speech [22]. However, the exploitation of synthetic speech prosody parameterization necessitates the utilization of the human natural spoken rendition for tables.

The work presented in this paper addresses the universal accessibility of complex document visual structures for

both the visually capable and the visually impaired using speech-only user interfaces, as part of design-for-all approach [29] for the auditory rendition of visual documents. Electronic documents that contain complex visual structures should be fully accessible. The accessibility of data contained in such structures necessitates a sophisticated generic approach that should not be specific to a particular (screen) reading application. The resulting spoken format of data tables should be accessible for everyone, including the visually impaired, as well as from anywhere (e.g., content access from a remote location through telephone, mobile computer, or during travel).

To achieve the vocalization of data tables, this work focuses on the representation of the corresponding table meta-information through prosody control derived from the human paradigm. Natural speech samples from human readers have been examined and modelled in order to aid speech synthesis. The aim is to derive a generalized specification using spatial directives for application to a design-for-all [28] document-to-audio approach.

This paper first presents an analysis of the visual and semantic characteristics of data table structures to determine the types of attributes that should be taken into consideration for the aural rendering of data tables. Then, it describes a series of psychoacoustic experiments on spoken format of data tables utilizing prosodic attributes. Blind and sighted listeners were asked to reconstruct simple and complex data tables from naturally spoken descriptions by expert readers. From the listener feedback it was deduced that consistent prosodic rendering can model the underlying semantic structure of tables. Using the acquired analysed speech data from the most preferred human spoken renditions, a corresponding prosody specification is introduced, including phrase accent and pause break information. Finally, this specification is evaluated by human subjects using aural renditions of several seen and unseen data tables and compared against the traditional linear vocalization approaches.

2 Table specification

Tables are used in abundance in electronic documents today. There are many document type specifications that contain table descriptions allowing implementation of tables. One of the most common document types, used by the on-line community on an everyday basis, is HTML. HTML provides a set of meta-elements that specifies the visual presentation of text. According to the W3C, “The HTML table model allows authors to arrange data into rows and columns of cells” [20].

This paper focuses on the vocalization of table structures in auditory user interfaces. Transferring a structure from

the visual modality to aural is a complex task. Tables are characterized by many qualitative and quantitative aspects that should be taken into consideration to obtain successful vocalization. For table construction, the following factors are taken into account:

- *Size of table* can range from just a couple of cells (e.g., 1×2 table size) to any number of rows and columns.
- *Design of table* refers to the implementation of the spatial layout by the designer. There is more than one way of successfully presenting the same information using different tables. Those tables in effect should convey the data in the way intended by the designer. The designer may use different header cells or grouping of data. The differences in design may be in the use or non-use of certain attributes. All sorts of data can be supported inside table cells.
- *Amount of data in each cell*, e.g., a whole sentence against one letter.
- *Type of data in each cell* may vary. The data in HTML table cells, for instance, may be text, preformatted text, images, links, forms, form fields, other tables, etc.
- *Visual enhancements* (such as colour or bold letters) are commonly used by designers in order to emphasize certain parts (cells, rows, etc.) of a table.
- *Complexity*, e.g., nested tables, merged cells.
- *Inherent underlying semantic structure*, the most difficult parameter to utilize.
- *Browsing (linearization)* is not a table characteristic but rather a manner of navigation and/or linearization of a table, discussed in detail later.

This work describes experiments on the rendition of visual table structures to voice. To achieve that, the W3C recommendations were followed in order to ensure that the data tables used in the experiment were designed to be processed aurally as well as to conform to certain guidelines for wellformedness. In respect to that, certain considerations pertaining to the above discussion were made, and appropriate example tables were selected based on table accessibility guidelines, complexity, and browsing considerations as discussed below.

2.1 Table accessibility guidelines

Special recommendations to promote accessibility, containing guidelines on how to make the web content accessible to people with disabilities, are provided by the W3C [5]. According to these, the use of <TH> (for headers) and <TD> (for data cells) is mandatory. The use of <THEAD>, <TFOOT>, <TBODY> and <TR> to group rows and <COL> and <COLGROUP> to group columns is also required to associate data and header cells. Moreover, W3C table description states that HTML data

tables should not be used for visual layout purposes, that being the task of style sheets.

This work is concerned with model data tables which by design are compliant to high priority Web Accessibility Initiative (WAI) guidelines recommendation. The term data tables, from this point onward, will refer to HTML TABLE containers 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. Such tables use HTML markup reserved for data tables, such as <TH>, <TD>, <CAPTION>, etc. These elements are used by agents in order to identify and manipulate the data before vocalization.

2.2 Underlying semantics and browsing considerations

Tables are visual structures spanning two dimensions, used as a means of grouping information. The designer of a table is mostly concerned with the visual representation of the data. The resulting table is a visual structure formed to accommodate the semantic relationship of the data. In fact, the semantic relationship is the cause for choosing a table as the most appropriate visual structure for modelling the information. The table as a structure models semantic information visually. Such semantic information is then inherited inside the structure in the visual modality and ideally can be retrieved by the visual reader (Fig. 1). This is possible because the reader can see the whole structure and infer the semantic relationships between the header and data cells by deciphering the visual representation. There is a direct connection between the original data (natural language), the structural information (written language), and the data reconstruction (natural language).

Accessing table structures by visually impaired or blind people is a far more difficult task, since the visual structure has to be processed in order to render it in the aural modality. Auditory user interfaces access the tables and try

to present the information to the listeners. There are two major considerations in this case. The first is non-visual browsing of tables, which involves the manner of linearization and presentation of the embodied data for aural rendition. The second is the quality of the acoustic representation of the linearized table that should allow the conversion of the visual structure into understandable speech faithful to the intended semantic structure (Fig. 2).

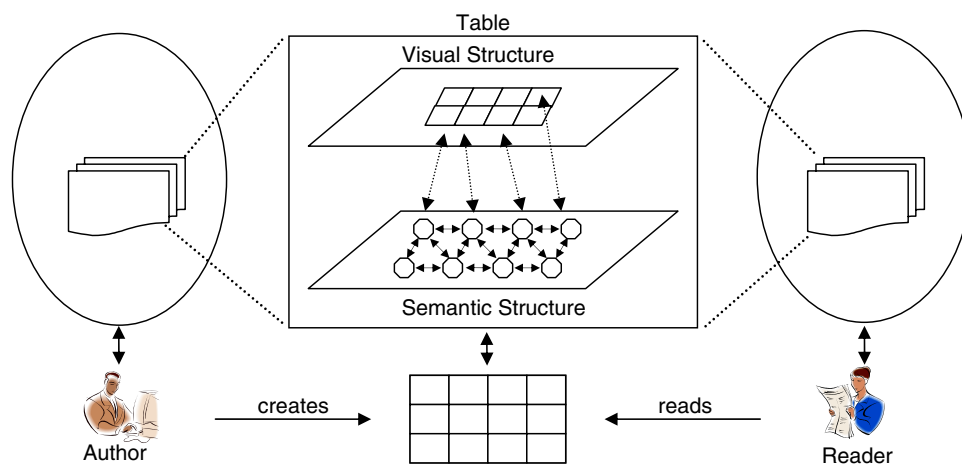
Tables can be rendered to speech in several ways, depending on the approach selected for linearization. A common screen reader would linearize a table row by row, resulting in a loss of semantic information between the cell data. Figure 3 shows how linear reading of the example table retains no structural meaning, rendering the output almost useless to the listener.

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 into text (and consequently to the listeners) by navigation of the data cells [19]. Figure 4 depicts examples of the same table browsed in two different ways following browsing specifications leading to either header–data pairs or header–data–data matches. The bottom linear reading form effectively renders the columns while the top form renders the rows.

2.3 Table complexity

Completing the previous considerations, emphasis is given to the use of well-formed data tables taking into account the variation in complexity that these structures may carry. As mentioned previously, tables are structures that are used to arrange content into a two-dimensional yet semantically coherent assembly. Cells are the basic blocks that contain data. The type of data may be “header” information, which is used to describe the data in other cells, or “data” information. Cells are arranged into rows and columns that

Fig. 1 Table structural information—creating and visually reading a data table



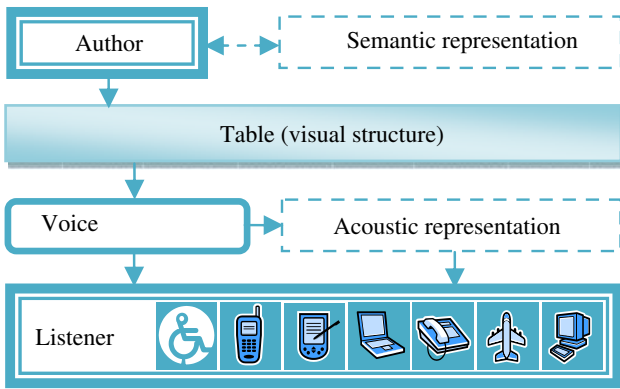


Fig. 2 Spoken format rendering of tables

are perceived as groups of data labelled by the header information, thus forming the table.

In terms of complexity, simple data tables have up to one row and up to one column of header cells. Figure 5 shows an example of a simple table used in the literature. It is easy to observe that the first row is the only row of header cells while there is no respective column of headers.

Complex data tables contain more than one level of logical row or column headers. This means that header cells can be expanded to encompass more than one row or column. Moreover, the same can be true for data cells. This can lead to complex tables made up of nested tables. From the example of complex table shown in Fig. 6, it is obvious that complex data tables are more difficult to browse, since they may contain multiple levels of structures in a hierarchical manner. Not depicted in Fig. 6 is the `<DATA>` cell information contained in the nested simple tables.

The example complex table can be thought of as a three-dimensional structure [19], compared to the two-dimensional simple data table of Fig. 5. Figure 7 shows a

Name	Phone Number	Age	Weight
Steve Nelson	425/555-2186	54	130 lbs
Maria Sanchez	425/555-8741	43	120 lbs

Fig. 5 A simple data table example

semantic structure dimensional comparison view of the two tables. The third dimension of the semantic structure of the complex table is embedded in the two-dimensional visual structure.

For complex tables, intelligent browsing may be realized in HTML by the use of *headers* and *id* attributes or *scope* in order to accommodate and handle more than one logical level in a table, and when it is necessary to link more than two headers with a data cell. This means that data cells inside the nested tables are semantically related to their respective header cells. In addition, both header and data cells in nested tables are governed by the top-level headers depicted in Fig. 7, complex table, second row. Similarly, the bottom row summarizes from data contained in both the nested tables, also semantically related to the same header cells (second row). In order to browse such complex HTML table, the scope attribute is used with header information in order to provide the renderer with the data cells to which the header is associated with. Moreover, using the scope attribute on data cells forces them to behave like header cells.

3 Natural rendition experiment

From the above considerations, it emerges that both simple and complex data tables have to be considered in order to observe data semantic coherence to gain enough knowledge for representing such visual structures acoustically.

Fig. 3 Common linear reading of table data

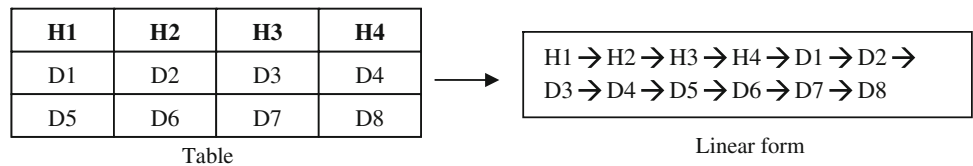


Fig. 4 Intelligent browsing of table data

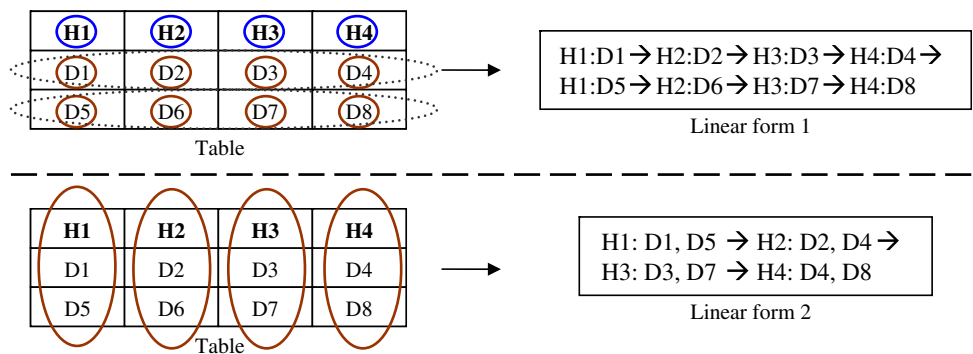


Fig. 6 A complex data table example

Travel Expense Report				
	Meals	Hotels	Transport	subtotals
San Jose				
25-Aug-97	37.74	112.00	45.00	
26-Aug-97	27.28	112.00	45.00	
subtotals	65.02	224.00	90.00	379.02
Seattle				
27-Aug-97	96.25	109.00	36.00	
28-Aug-97	35.00	109.00	36.00	
subtotals	131.25	218.00	72.00	421.25
Totals	196.27	442.00	162.00	800.27

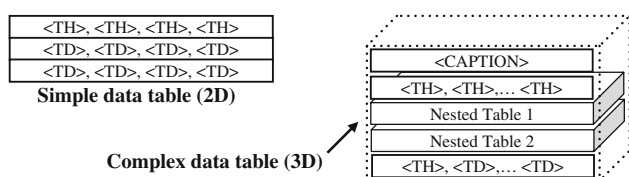


Fig. 7 Simple and complex data table dimensional comparison

The three-dimensional semantic correlation for complex table data along with the straightforward relations of simple table data needs to be modelled for the same acoustic characteristics and described by the resulting prosody model specification for table-to-speech synthesis. The semantic representation must be consistent and straightforward for the readers and the listeners to accurately fathom. The examined tables have to be well-formed data tables conforming to W3C WAI guidelines.

The aim of this psychoacoustic experiment was the examination of speech prosody characteristics that describe the human aural rendition of data tables. Natural spoken rendition required human subjects as readers of data tables. The tables selected for this experiment were similar to the W3C reference simple and complex tables [4]. The data in both tables were in Greek, the native language of all readers and listeners. Tables 1 and 2 show the exact simple and complex data tables used in the experiment.

Table 1 The simple data table used in the experiment (customer information)

Name	Telephone number	Age	Residence
Thomas Mizas	4850918	54	Ilisia
Maria Ripi	7776755	29	Marousi
Elena Riga	5640089	43	Pagrati
Nikos Mistas	3867244	32	Thisio

Table 2 The complex data table used in the experiment (travel expense report)

	Meals	Hotels	Transport	Total
Paris				
25 August 1997	37.00	112.00	31.00	
26 August 1997	28.00	112.00	31.00	
Subtotal	65.00	224.00	62.00	351.00
London				
27 August 1997	96.00	109.00	16.00	
28 August 1997	35.00	109.00	16.00	
Subtotal	131.00	218.00	32.00	381.00
Grand total	196.00	442.00	94.00	732.00

3.1 Participants

Thirty sighted and thirty blind listeners participated in this study. The sighted participants were students of the University of Athens and ranged in age from 20 to 25 years (mean age 23 years). Some of the blind participants were also students of the University of Athens and some others were members of the Greek Association of the Blind and ranged in age from 20 to 28 years (mean age 24 years). None had any history of hearing deficit, while all were familiar with the synthetic voice MBROLA ‘gr2’ used in the evaluation process [31]. The evaluation sessions were performed under the same conditions with a different group of 30 sighted experienced listeners of synthetic speech, aged from 20 to 23 years (mean age 22 years).

3.2 Procedure

Three adept readers were asked to record the natural renditions (referred to as SR-1, SR-2, SR-3) in the recording studio of the Speech Lab., University of Athens. The readers were fluent individuals bearing adequate familiarization

with table structures and speech prosody, and they were experts in the fields of linguistics, speech synthesis, and musicology/acoustics. They were asked to provide two spoken renditions for every table type (simple and complex). Readings were made using the plain linear browsing and the guided navigation matching the intelligent browsing. The readers were encouraged to provide their own prosodic rendition of the tables in accordance with their subjective judgment and expertise.

The listening sessions took place in the University of Athens sound-proof Speech Lab. All sessions lasted approximately 60 min for each participant. The evaluation sessions lasted 30 min. The participants were asked to listen to all spoken representations of the data tables in random order, and reconstruct them back to written form on paper (sighted subjects) or on Braille (blind subjects). An initial training session took place where the participants were shown sample data tables (printed matter for the sighted, Braille for the blind) of simple and complex tables and were given ample time to familiarize themselves with the table structures.

The objective task was the reconstruction of the table information from the spoken rendition back to written form so that the implicit structure should be retained. The level of table reconstruction by the listeners would lead to the evaluation and selection of the human spoken rendition that provides the best transfer of all tables to the acoustic modality. For each table rendition style (linear and intelligent for both simple and complex tables), the following marking was followed (the objective or subjective quality of each input is shown in the parentheses):

- Objective evaluation score for the reconstructed data table distributed according to the 2D pattern reconstruction, comprehension, and data inclusion: (*obj*);
- confidence factor, 1 (lowest) to 5 (highest): (*subj-1*);
- acceptance factor, subjective evaluation score, 1 (lowest) to 5 (highest): (*subj-2*);
- brief justification and/or problem indication for subjective evaluation score: (*subj-3*);
- subjective preference of reader rendition (choice of acoustic rendition): (*subj-4*).

The *subj-3* requirements were adjusted according to the feedback received from a smaller pilot experiment, so that

Table 3 Subjective listener input (*subj-3*)

Features	Comments
Linearization	Unacceptable, not good, good enough, v. good, ideal
Pause breaks	Too short, slightly short, slightly fast, too fast, uneven, correct
Accents	Unacceptable, correct
Other	(additional remarks)

specific input was required from the listeners [26]. The brief justification was a mandatory requirement that included specific marks about the quality of the aural rendition style and prosodic features that prompt the subjective acceptance scores as shown in Table 3.

3.3 Analysis of the results

The experimental results were evaluated in order to classify table reconstructions from the subjects. The confidence score from all listeners was 4 or above from scale 1 (lowest) to 5 (highest) which was enough to validate all scoring. The table reconstruction was examined and marked by an examiner evaluation score with a range of 1 (lowest) to 5 (highest) that was distributed according to the following considerations:

1. non-reconstructive 2D pattern, non-comprehensible, vast lack of data;
2. existence of a vague reconstructive 2D pattern, non-comprehensible, a great amount of data is missing;
3. existence of a clear reconstructive 2D pattern, quite comprehensible but still an amount of data is missing;
4. existence of a clear 2D reconstructive pattern, comprehensible with coherence, few data is missing;
5. perfect representation of a 2D pattern, comprehensible with coherence, all data are included in the representational form.

Table 4 presents the listener subjective evaluation score (LSSES) and examiner objective evaluation score (EOES) averages from blind and sighted subjects for all four data table natural speech renditions from all readers. LSSES refers to the *subj-2* quality while EOES to the *obj* as described above.

Initial examination of the results shows that common linearization of the complex data table resulted in poor comprehension from all listeners with the average score of 2.3. The *t* test statistical criterion was also conducted on these results and showed no statistical significance.

On the other hand, the average score for the respective rendition of the complex data table following the intelligent browsing of such structures was 3.9, confirming the previous related research that intelligent browsing is crucial to acoustic rendering. For simple tables, the scores for linear and intelligent browsing were almost identical, showing that either linearization is acceptable. Further observation shows that the sighted subjects performed better, something that was actually expected. Blind subjects scored half a mark (0.5) lower than sighted subjects in the simple table cases, while about 0.9 lower for complex tables. That was a very consistent behaviour, and a possible indication of a pattern that may be useful in further similar experiments.

Table 4 Partial and overall scores for simple and complex structure reconstruction and user feedback (*obj*)

	Simple (linear)		Simple (intelligent)		Complex (linear)		Complex (intelligent)	
	LSES	EOES	LSES	EOES	LSES	EOES	LSES	EOES
Blind	3.8	4.1	3.7	4.0	2.2	1.6	3.4	3.7
Sighted	3.8	4.6	4.2	4.9	2.8	2.7	3.7	4.6
All	3.8	4.35	3.95	4.45	2.5	2.15	3.55	4.15
Overall	4.1		4.2		2.3		3.9	

Table 5 *subj-2* and *obj* scores for simple and complex table renditions (SR: spoken rendition)

	Simple (linear)			Simple (intelligent)			Complex (linear)			Complex (intelligent)		
	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3
Blind (<i>subj-2</i>)	4.13	3.53	3.70	4.20	3.40	3.40	2.13	2.86	1.70	3.30	3.70	3.20
Blind (<i>obj</i>)	4.30	4.00	4.10	4.20	4.00	3.80	1.50	1.86	1.46	3.70	4.00	3.50
Sighted (<i>subj-2</i>)	4.56	3.20	3.70	4.43	4.33	3.83	2.67	3.37	2.43	4.17	4.20	2.77
Sighted (<i>obj</i>)	5.00	4.07	4.80	5.00	4.93	4.80	2.77	3.37	2.10	4.77	4.87	4.23
Overall	4.50	3.70	4.08	4.41	4.17	3.96	2.27	2.87	1.92	3.99	4.19	3.42
Preference	SR-1			SR-1			SR-2			SR-2		

Amongst the three spoken renditions (for each of the four table/rendering style combinations), all listeners (and the respective table reconstruction evaluation) seemed to agree on specific renditions for the simple and complex tables. Table 5 shows the respective scoring for the subjective and objective analysis from the subjects, while Table 6 shows the spoken rendition preference analysis. Consistency of the results was preserved between reading styles.

From this point onwards, “SR-X” will refer to the spoken rendition from a reader for a particular table-style combination, where “X” is the reader number.

The spoken renditions made by reader 3 (SR-3) scored poorly and were not favoured for any of the tables and reading styles, although they were highly valued amongst many listeners. The reason is that the other two renditions offered a combination of characteristics that made them perform better. Moreover, SR-3s were judged to be too fast in terms of speech rate as well as too long/short in pause breaks. Only in the case of the simple table linear rendition, SR-3 was a second favorite, due to the very short pauses of SR-2, which the listeners could not keep up with. As a note for the future, from the respective subjective comments, the SR-3s could be thought of as a lower threshold in speech rate and pause breaks for reading tables or similar structures or perhaps, as suggested by some listeners, could form a basis for parameterization for use when repeating a structure. Hence, the SR-3 analysis results are not directly comparable and, therefore, should not be projected against the ones from the other renditions. Tables 7 and 8 depict

the subjects’ feedback on the justification and problem indication of the simple and complex table renditions that shows how each of the three shaping factors (linearization, breaks, tones) affected the level of understanding of the table data.

The analysis of speech signals of the spoken renditions themselves revealed several qualitative differences. It involved an analysis of the prosody of the signals in terms of phrase accent tones, pause breaks, as well as overall speech rate. For prosody markup, the ToBI annotation model [25] was used as a means of qualitative analysis. In this experiment, a key factor is the realization that the structural layout of the tables governs the speech rendition. Thus, the analysis was focused specifically on boundary tones (L- describing the fall in pitch at end of spoken data cell; H- describing the boundary rise) and respective pause breaks between cells and rows, for each of the four spoken renditions. Table 9 depicts the placement of boundary tones for SR-1 and SR-2 for the linear simple table renditions. In the initial experiment, the five blind listeners marginally preferred SR-2 and enjoyed the change in pitch on the penultimate cell of the <TD> rows, as a clear indication of the end of row. In the main experiment, the blind listeners argued that both accent assignments were correct, and although they were marginally happier (59%) with the SR-2 accents, the overall rendition and mostly pause breaks of SR-1 were the final preference. In fact, due to the simplicity of the tables, pause breaks were more crucial. The sighted subjects assigned very little importance to boundary tone differences between SR1 and SR-2.

Table 6 Natural speech rendition preference (*subj-4*) analysis (SR: spoken rendition)

	Simple (linear)			Simple (intelligent)			Complex (linear)			Complex (intelligent)		
	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3
Blind	14 (47%)	5 (16%)	11 (37%)	21 (70%)	5 (16%)	4 (14%)	4 (14%)	21 (70%)	5 (16%)	12 (40%)	14 (46%)	4 (14%)
Sighted	25 (84%)	1 (3%)	4 (13%)	16 (53%)	11 (37%)	3 (10%)	5 (16%)	19 (64%)	6 (20%)	11 (36%)	19 (64%)	0 (0%)
Overall	39 (65%)	6 (10%)	15 (25%)	37 (62%)	16 (27%)	7 (11%)	9 (15%)	40 (67%)	11 (18%)	23 (38%)	33 (55%)	4 (7%)
Preference	SR-1			SR-1			SR-2			SR-2		

Table 7 Justification and problem indication (*subj-3*) feedback for simple table renditions

	Linear browsing			Intelligent browsing		
	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3
Linearization	Good	Good	Good	Ideal/v. good	Ideal/v. good	Ideal/v. good
Pause breaks	Correct	Too short	Slightly short	Slightly long	Slightly long	Too long
Boundary tones	Correct	Correct	Incorrect	Correct	Correct	Correct

Table 8 Justification and problem indication (*subj-3*) feedback for complex table renditions

	Linear browsing			Intelligent browsing		
	SR-1	SR-2	SR-3	SR-1	SR-2	SR-3
Linearization	Unacceptable	Unacceptable	Unacceptable	Ideal/v. good	Ideal/v. good	Ideal/v. good
Pause breaks	Correct	Correct	Too short	Correct	Correct	Too short
Boundary tones	Correct	Correct	Correct	Correct	Correct	Correct

Table 10 shows that both intelligent spoken renditions for the simple table shared the same characteristics, fully approved by the listeners. All <TH> cells had high phrase accent and all <TD> cells low. Similar results were obtained for complex table SRs. As shown in Table 11, a point of high impact was observed for the linear browsing renditions. The change of pitch when reading the 5th and 9th rows (the last row of each nested data table—worked out as indication of end of block of data) clearly resulted in better reconstruction. Both readers tried to acoustically represent the existence of nested tables, clearly rendering the headers of the nested tables as L (first cell data in 3rd–4th and 7th–8th rows) as well as the end of those tables (SR-2).

The pitch analysis for the intelligent SRs for the complex table showed identical results and the same pattern as for the simple table (see Table 6).

Pause breaks were expected to play a significant role in the spoken format of tables, since pausing could prove to be very helpful in distinguishing between rows. That is true for data tables because rows are used by the designer to represent semantic correlation between the data they encompass. Besides, pause breaks would prove to be most useful in cases where the pitch remains unchanged. Such example is shown in Tables 12 and 13.

For simple tables (linear browsing), the pause at the end of each cell data and end of row of cell data proved more significant than change in pitch for the sighted listeners. The blind listeners were more sensitive to pitch and provided more feedback, as also shown by previous works [32]. In the initial experiment, the five blind subjects preferred the shorter pause breaks of SR-2 and stated that the simple tables are easy enough to identify when an end of row occurs. The result was also reinforced by the fact that

Table 9 Simple table (linear) boundary tones

	SR-1	SR-2
<CAPTION>	L-	L-
<TH> rows	H- H- H- L-	H- H- H- L-
<TD> rows	L- L- L- L-	L- L- H- L-

Table 10 Simple and complex table (intelligent) boundary tones

	SR-1	SR-2
<CAPTION>	L-	L-
<TH>	H-	H-
<TD>	L-	L-

Table 11 Complex table (linear) boundary tones

	SR-1	SR-2
<CAPTION>	L-	L-
1st row	H- H- H- L-	H- H- H- L-
2nd row	L-	L-
3rd row	L- H- H- L-	L- H- H- L-
4th row	L- H- H- L-	L- H- H- L-
5th row	L- H- H- H- L-	H- H- H- H- L-
6th row	L-	L-
7th row	L- H- H- L-	L- H- H- L-
8th row	L- H- H- L-	L- H- H- L-
9th row	L- H- H- H- L-	H- H- H- H- L-
10th row	L- H- H- H- L-	L- H- H- H- L-

Table 12 Simple table (linear) breaks (in seconds)

	SR-1	SR-2
End of cell	2.31	1.09
End of row	2.91	1.75

Table 13 Simple table (intelligent) breaks (in s)

	SR-1	SR-2
End of header cell	0.60	0.82
End of data cell	2.41	2.10

Table 14 Complex table (linear) breaks (in s)

	SR-1	SR-2
End of header-row	2.41	2.55
End of data row	2.80	2.66
Top row header cell	1.42	0.92
Other header cell	2.34	2.00
Data cell	1.69	1.74

the blind subjects were experts in synthetic speech. Here, the blind subjects achieved better results from SR-1 and showed a clear preference for the much slower rendition.

The intelligent renditions were also affected. The difference in pause breaks made the header–data pairs more distinct for the listeners. Shorter break between the two pieces of data combined with much longer pause between pairs resulted in better scoring. In this case, however, the blind subjects clearly showed preference for SR-1, justified by the fact that the notion of pairs was granted by a shorter pause after header cell data, while providing enough length for clear identification of end of row. All subjects reported

Table 15 Complex table (intelligent) breaks (in seconds)

	SR-1	SR-2
Nested table header	1.84	2.79
End of row	2.32	3.81
Header cell	1.10	0.68
Data cell	1.94	2.83

that the pause break differences were ideal; however, over 80% of them found the actual length to be a bit longer than preferable and urged for slightly shorter breaks, thus faster overall table rendition.

For complex tables, pitch variation had a greater impact than pause length. In the first case (linear browsing), there was little difference in pause breaks, as shown in Table 14. The pitch variation made all the difference, although the linear rendition of the structure left only little to be salvaged. The situation was different for the intelligent renditions (Table 15). This time, semantics were carried over to the listeners and allowed prosody—in this case, the pause length—to outline the structure. Longer pauses in SR-2 (ranging from 49 to 64% longer than SR-1) and enough variation between header–data cell pairs and “end of row” marking provided the most successful vocalization.

Speech rate in this experiment had minimal effect, since there was no connected speech involved. In fact, the strong presence of long pause breaks renders any speculation worthless.

4 The prosody model specification

The human spoken table rendition feedback from the listeners led to the design of a formal specification for prosodic modelling of simple and complex table structures to be used by speech synthesis systems. The ToBI annotation model conventions were used for the phrase accent descriptions, while relative values have been used for pause breaks. 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.

The prosodic basis of the table-to-speech synthesis was a well-established CART-based ToBI model, rendered by a Linear Regression (LR) F0 generation model. Data have been gathered from a 600-utterance voice corpus. Table 16 provides summary information about this model, in terms of ToBI elements prediction using CART trees for breaks,

Table 16 Basic prosodic model performance

	Break				Accent					Endtone	
	0	1	2	3	LH*	L*H	H*L	H*	L*	L-	H-
Precision	0.89	0.81	0.76	0.98	0.39	0.32	0.32	0.11	0.25	0.88	0.65
Recall	0.76	0.97	0.38	0.83	0.44	0.56	0.28	0.08	0.21	0.90	0.61
Classification rate (%)	83.27				71.67					96.59	

Table 17 Linear regression model performance

	RMSE (Hz)	<i>r</i>
Start	20.6	0.71
Mid-nucleus	21.2	0.72
End	20.7	0.71

accents and endtones. This model has been validated with the *N*-fold cross method.

The F0 generation model is based on a LR approach [1], using the above-mentioned predicted ToBI values. Three LR models were used: one for the F0 target at the start of a syllable, one at the mid of the nucleus and one at the end of the syllable. Table 17 presents the performance of the LR models in terms of root mean square error (RMSE) and correlation (*r*).

The optimized specifications were constructed based on the user-selected most prominent natural renditions (SR-1 for simple tables, SR-2 for complex tables), to alter the above-mentioned predicted values. The optimizations were derived by the listener feedback for justification and problem indication (*subj-3*). The boundary tone model specification describing position and type of phrase accent tone according to conditions that apply for either simple or complex table and linear/intelligent browsing is shown in Table 18. An obvious observation is the simplicity of the model for intelligent browsing, as a result of semantic resolution prior to vocalization.

Table 18 Boundary tone specification

Table (browsing)	Tone	Header cell	Data cell
Simple (linear)	L-	Final-row	Not-row-penultimate
	H-	Not-final-row	Row-penultimate
Simple (intel.)	L-	Not-header-row	(all)
	H-	Header-row	NA
Complex (linear)	L-	Final-row	Row-final Row-initial AND not-nested-table-final-row
	H-	Not-final-row	Row-initial AND nested-table-final-row
Complex (intel.)	L-	Not-header-row	(all)
	H-	Header-row	NA

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

Table 19 shows the actual values and multiplier factors for linear and intelligent browsing for simple and complex tables.

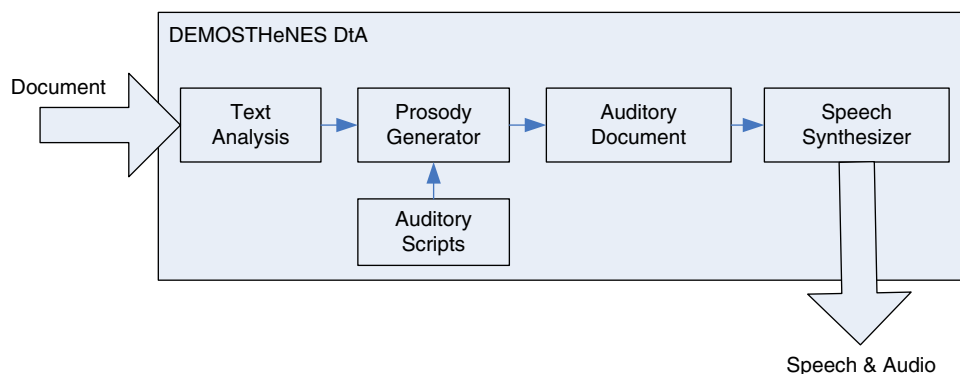
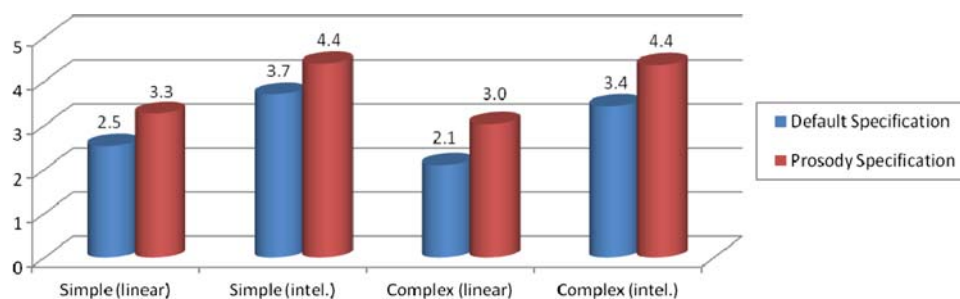
The specification is applicable to all data tables, since it references generic cell information without assuming that specific headers or id attributes are used.

5 Evaluation

The subjects were asked to take part in a formal listening experiment to evaluate synthesized spoken tables. The DEMOSTHeNES document-to-audio platform featuring the MBROLA gr2 diphone database [31] was used to synthesize speech carrying the derived prosody specification (Fig. 8). This was achieved by means of two auditory scripts for the simple and the complex tables. 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 (using the DEMOSTHeNES built-in generic prosody assignments) and enhanced spoken rendition by the newly acquired specification in order to experiment with the proposed table-to-speech approach.

Table 19 Pause break specification

Duration	Table (browsing)	Header cell	Data cell	Header-row	Data row	%%Nested table header cell
ms (mult.)	Simple (linear)	700 ($\times 1.00$)	700 ($\times 1.00$)	1000 ($\times 1.40$)	1000 ($\times 1.40$)	NA
	Simple (intel.)	200 ($\times 1.00$)	500 ($\times 2.50$)	750 ($\times 3.75$)	750 ($\times 3.75$)	NA
	Complex (linear)	300 ($\times 1.00$)	525 ($\times 1.75$)	750 ($\times 2.50$)	750 ($\times 2.50$)	600 ($\times 2.00$)
	Complex (intel.)	200 ($\times 1.00$)	750 ($\times 3.75$)	1250 ($\times 1.75$)	1250 ($\times 1.75$)	750 ($\times 3.75$)

Fig. 8 The document-to-audio platform**Fig. 9** Side-by-side comparison of prosodic model against default specification

The aim of the first evaluation was a comparative subjective analysis of plain and enhanced speech synthesis renditions of simple and complex example tables as introduced in the initial experiment [27] in order to measure the impact of the new prosodic adjustment model. 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 (Fig. 9) have shown that the parameterization has led to significant increase (14–16% for simple tables, 18–20% for complex tables) in their subjective understanding of the table data semantic structure.

The second evaluation involved listening to unseen and unknown data tables. The results 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 synthesized spoken formats of the unseen and unknown simple and complex data tables used in the evaluation were large enough to contain several bits of

information (Tables 20, 21). 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 shown by italicized text in square brackets.

The subjects were asked to listen to each synthesized rendition and carefully answer 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 the questions. Moreover, at the end of each session, they were asked to provide their subjective opinions on the overall quality of rendition, the listening effort required to understand each table, and their acceptance. For both evaluation parts, feedback collection was achieved using a modified version of the mean opinion score questionnaire [30].

Figure 10 shows overall impression (5 = excellent, 1 = bad) of synthesized speech rendering of tables as well

Table 20 A larger unknown simple data table that contains one header-row and eight data rows (radio-transmitted sports schedule)

Day	Sport	Start time	End time
Monday	Athletics	11.00	18.00
Tuesday	Tennis	20.00	23.00
Wednesday	Athletics	09.00	18.00
Thursday	Gymnastics	16.00	21.00
Friday	Water polo	12.00	15.00
Saturday	Gymnastics	16.00	18.00
Saturday	Football	21.00	23.00
Sunday	Athletics	09.00	12.00

Table 21 An unknown complex data table that contains three nested sub-tables (weather report for the coming week)

	Monday	Tuesday	Wednesday	Thursday	Friday
Athens					
Temperature	23	24	26	22	18
Wind	Northwest	West	Southwest	Northwest	North
Salonika					
Temperature	16	17	20	16	13
Wind	North	North	West	North	Northwest
Patra					
Temperature	19	22	23	20	19
Wind	Northwest	West	South	Southwest	Southwest

as listening effort needed by the listeners in order to answer the key questions (5 = easy to understand, no effort required, 1 = no meaning understood despite any feasible effort). The acceptance of the overall rendition could only be marked as “yes” or “no” and the resulting percentage is therefore shown below in scale. It is worth mentioning that half of the listeners were unhappy with the linear rendition of the simple table, while 8 out of 10 were unable to understand the linear rendition of the complex table, marking them as unacceptable. 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.

It is obvious that the linear rendition of complex tables failed to make the semantic relationship of the data understandable, which was the case for the natural speech rendition during the initial experiments as well. In this case, the improvement from the prosodic enhancement and intelligent browsing was significant (26% improvement), successfully rendering certain key semantic data to speech. However, since prosody enhancement cannot be well supported by linear browsing (low acceptance), the listeners clearly prefer intelligent browsing with the prosody specification. The prosody model worked successfully for all other cases, the result being an improvement in acoustic representation as well as reduced effort.

Participants’ responses were marked according to their overall impression, their listening effort and the acceptance. The results show that the proposed specification exhibits significant improvement in both cases of simple table renditions. Moreover, from the table (complex vs. simple) × rendition (linear vs. intelligent) conducted on these data showed main effects for tables [$F(1,60) = 4.036, P < 0.01$] in favour of the complex tables and for rendition [$F(1,60) = 3.033, P < 0.01$] in favour of the intelligent rendition. This confirms the listener feedback (as illustrated in Fig. 10) that the specification performs extremely well for acoustic rendering of complex tables using intelligent browsing.

Precision and recall metrics were calculated from the correct and non-correct answers to the evaluation questions of the unseen and unknown synthesized table renditions. As shown in Table 22, in all cases, the prosodic specification was a great improvement over the default prosody control. In the cases where the prosody specification was used, the listeners had a very high hit rate with very few non-correct answers and only a few null answers. For the default specification, precision was marginally lower (in all but one case) while recall was substantially lower. This reflects the fact that the subjects had a lot of null answers, which means that the listeners failed to find correct answers yet still retained a low error rate. Ultimately, this means that the subjects either find the correct answer or they do not. The comparison of the recall figures shows the great

Fig. 10 Overall impression (higher = better), listening effort (higher = easier) and acceptance (higher = better)

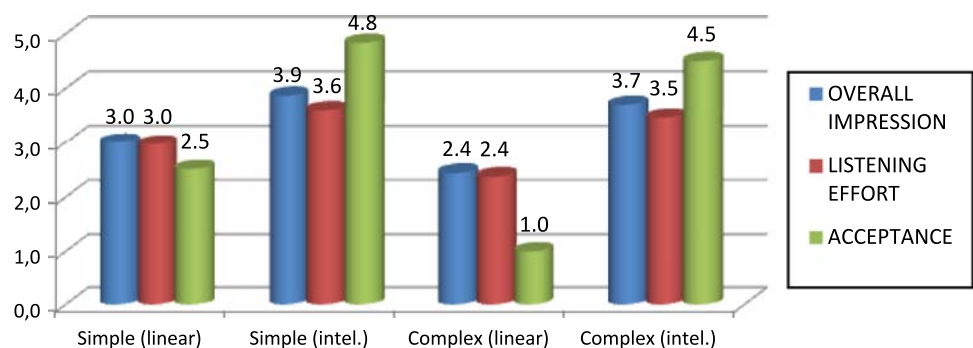


Table 22 Evaluation of unseen tables with and without prosody specification

		Prosody	Non-prosody
Simple (linear)	Precision	0.99	0.95
	Recall	0.96	0.68
	f-measure	0.98	0.79
Simple (intelligent)	Precision	0.99	0.98
	Recall	0.97	0.79
	f-measure	0.98	0.87
Complex (linear)	Precision	0.95	0.79
	Recall	0.59	0.26
	f-measure	0.73	0.40
Complex (intelligent)	Precision	0.99	0.96
	Recall	0.86	0.74
	f-measure	0.92	0.83

improvement of prosody specification in the identification of the correct information in the spoken rendition of visual structures. This is shown vividly by the f-measure, the weighted harmonic mean of precision and recall.

6 Conclusions

This paper addresses the complex problem of table vocalization. Simple linear and intelligent browsing techniques were used to convert data tables into text and thus conventional text-to-speech synthesis technology could be used for table-to-speech synthesis. In order to access “the semantic information under the visual structure” through synthesized speech, the authors propose diction-based prosody modelling for synthesis of simple and complex data tables.

The prosody models were constructed based on the analysis of an experimental study of spoken presentation for complex visual document structures such as data tables. Human readers were employed to render both simple and complex visual structures which were then evaluated by blind and sighted listeners. Sets of prosodic parameters were analysed in terms of boundary tones and pauses, clearly illustrating consistency against cell content and visual structure. The deduced specification formed the basis for the auditory scripting modelling of tables to aid automatic rendering using synthetic speech. Through a formal acoustical assessment, the direct comparison of the prosody model-aided synthetic speech against the default parameters used by a TtS system revealed 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. The results also show that both the visually impaired and the visually capable

show equally increased comprehension of semantic relations between data from speech renditions of tables.

As an overall assessment of the results from these experiments, it can be deduced that the prosodic model provided a promising approach to modelling visual structures and to identify potential implementation issues in table-to-speech synthesis from natural speech derived data. The navigation style makes a strong impact on the final result, and therefore it should be pursued and modelled accordingly. Furthermore, it was deduced that by careful prosody modelling, 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.

Though this work has reached its initial target of the implementation of a prosodic specification based on the natural human rendition that could deliver table information, there are two aspects of prosody that have not been accommodated in the current prosodic model and have been planned for further work. The first one is duration modelling. This is the rhythmic aspect of prosody, which is currently not always aligned with the intonational events, but follows a grammatical driven model for assigning segmental durations. This is the most common approach for text-to-speech synthesizers, as a larger data corpus is required to extract adequate duration-pitch correlated information. The second aspect is the handling of micro-prosodic phenomena, i.e., the role of segmental prosody to the F0 surface, such as the F0 difference between open and close vowels. This speech characteristic was not captured in the basic prosodic model, because, as has been pointed in [34], larger corpora are also required in order to collect adequate data for such analysis and modelling.

It is concluded that prosody control can successfully lead to improved understanding of synthesized speech rendition of data tables, eventually conveying semantically important visual information to speech by prosody control alone. This reveals the potential benefits of the use of prosody modelling in conjunction with other means of speech and audio modifications such as voice alternation, auditory icons, and earcons, and certainly warrants further research toward such approach.

Acknowledgments The work described in this paper has been partially supported by the European Social Fund and Hellenic National Resources under the HERACLITUS project of the EPEAEK II programme, Greek Ministry of Education. We would like to thank Manolis Platakis, Dimitris Sifakis, and the students of the University of Athens, Department of Informatics and Telecommunications, as well as the members of the Panhellenic Association of the Blind for their participation in the experiments described in this work.

References

1. Black, A., Hunt, A.: Generating F0 contours from the ToBI labels using linear regression. In: Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, USA, vol. 3, pp. 1385–1388 (1996)
2. Burnett, C.D., Walker, R.M., Hunt, A. (eds.): Speech Synthesis Markup Language (SSML) Version 1.0, W3C Recommendation. <http://www.w3.org/TR/speech-synthesis/> (2004). September 2004
3. Caldwell, B., Cooper, M., Guarino Reid, L., Vanderheiden, G. (eds.): Web Content Accessibility Guidelines 2.0, W3C Candidate Recommendation. <http://www.w3.org/TR/WCAG20/> (2008). 30 April 2008
4. Chisholm, W., Vanderheiden, G., Jacobs, I. (eds.): HTML Techniques for Web Content Accessibility Guidelines 1.0, W3C Note. <http://www.w3.org/TR/WCAG10-HTML-TECHS/> (2000). 6 November 2000
5. Chisholm, W., Vanderheiden, G., Jacobs, I. (eds.): Web Content Accessibility Guidelines 1.0, W3C Recommendation. <http://www.w3.org/TR/WAI-WEBCONTENT/> (1999). 5 May 1999
6. Embley, D.W., Hurst, M., Lopresti, D.P., Nagy, G.: Table-processing paradigms: a research survey. *Int. J. Document Anal.* **8**(2–3), 66–86 (2006)
7. Filepp, R., Challenger, J., Rosu, D.: Improving the accessibility of aurally rendered HTML tables. In: Proceedings of ACM Conference on Assistive Technologies (ASSETS), pp. 9–16 (2002)
8. Hurst, M.: Towards a theory of tables. *Int. J. Document Anal.* **8**(2–3), 123–131 (2006)
9. Hurst, M., Douglas, S.: Layout & language: preliminary experiments in assigning logical structure to table cells. In: Proceedings of 4th International Conference on Document Analysis and Recognition (ICDAR), pp. 1043–1047 (1997)
10. 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: Proceedings of the ACM Conference on Universal Usability, Vancouver, Canada, pp. 134–141 (2003)
11. Larson, J.A. (ed.): Introduction and Overview of W3C Speech Interface Framework, W3C Working Draft. <http://www.w3.org/TR/voice-intro> (2000). 4 December 2000
12. Lilley, C., Raman, T.V.: Aural Cascading Style Sheets (ACSS), W3C Working Draft. <http://www.w3.org/TR/WD-acss> (1999). 2 September 1999
13. Lim, S., Ng, Y.: An automated approach for retrieving hierarchical data from HTML tables. In: Proceedings of 8th ACM International Conference on Information and Knowledge Management (CIKM), pp.466–474 (1999)
14. Oogane, T., Asakawa, C.: An interactive method for accessing tables in HTML. In: Proceedings of International ACM Conference on Assistive Technologies, pp. 126–128 (1998)
15. Penn, G., Hu, J., Luo, H., McDonald, R.: Flexible web document analysis for delivery to narrow-bandwidth devices. In: Proceedings of 6th International Conference on Document Analysis and Recognition (ICDAR), pp. 1074–1078 (2001)
16. Pitt, I., Edwards, A.: An improved auditory interface for the exploration of lists. In: ACM Multimedia, pp. 51–61 (1997)
17. Pontelli, E., Gillan, D.J., Gupta, G., Karshmer, A.I., Saad, E., Xiong, W.: Intelligent non-visual navigation of complex HTML structures. *Univ. Access Inf. Soc.* **2**(1), 56–69 (2002)
18. Pontelli, E., Gillan, D., Xiong, W., Saad, E., Gupta, G., Karshmer, A.: Navigation of HTML tables, frames, and XML fragments. In: Proceedings of ACM Conference on Assistive Technologies (ASSETS), pp. 25–32 (2002)
19. Pontelli, E., Xiong, W., Gupta, G., Karshmer, A.: A domain specific language framework for non-visual browsing of complex HTML structures. In: Proceedings of ACM Conference on Assistive Technologies (ASSETS), pp. 180–187 (2000)
20. Raggett, D., Le Hors, A., Jacobs, I. (eds.): Tables, HTML 4.01 Specification, W3C Recommendation. <http://www.w3.org/TR/REC-html40> (1999)
21. Raggett, D., Glazman, D., Santambrogio, C. (eds.): CSS3 Speech Module, W3C Working Draft. <http://www.w3.org/TR/css3-speech> (2004). December 2004
22. Raman, T.: An audio view of (LA)TEX documents, TUGboat. In: Proceedings of 1992 Annual Meeting, vol. 13, no. 3, pp. 372–379 (1992)
23. Ramel, J.-Y., Crucianou, M., Vincent, N., Faure, C.: Detection, extraction and representation of tables. In: Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR), pp. 374–378 (2003)
24. Silva, A.C., Jorge, A.M., Torgo, L.: Design of an end-to-end method to extract information from tables. *Int. J. Document Anal. Recogn.* **8**(2), 144–171 (2006) (special issue on detection and understanding of tables and forms for document processing applications)
25. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: a standard for labeling english prosody. In: Proceedings of International Conference on Spoken Language Processing (ICSLP), vol. 2, pp. 867–870 (1992)
26. Spiliotopoulos, D., Xydias, G., Kouroupetroglou, G., Argyropoulos, V.: Experimentation on spoken format of tables in auditory user interfaces. Universal access in HCI. In: Proceedings of 11th International Conference on Human–Computer Interaction (HCI-2005), Las Vegas, USA, 22–27 July (2005)
27. Spiliotopoulos, D., Xydias, G., Kouroupetroglou, G.: Diction based prosody modeling in table-to-speech synthesis. *Lecture Notes in Artificial Intelligence*, vol. 3658, pp. 294–301. Springer, Berlin (2005)
28. Stephanidis, C.: Designing for all in the information society: challenges towards universal access in the information age. ERCIM ICST Research Report, ICS-Forth, Heraklion, Crete, pp. 21–24 (1999)
29. Stephanidis, C.: User interfaces for all: new perspectives into human–computer interaction. In: Stephanidis, C. (ed.) *User interfaces for all*, pp. 3–17. Lawrence Erlbaum, Mahwah, NJ (2001)
30. Viswanathan, M., Viswanathan, M.: Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Comput Speech Lang* **19**(1), 55–83 (2005)
31. Xydias, G., Kouroupetroglou, G.: Text-to-speech scripting interface for appropriate vocalisation of E-texts. In: Proceedings of 7th European Conference on Speech Communication and Technology, pp. 2247–2250 (2001)
32. Xydias, 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 Human–Computer Interaction (2005)
33. Xydias, G., Spiliotopoulos, D., Kouroupetroglou, G.: Modeling emphatic events from non-speech aware documents in speech based user interfaces. In: Proceedings of International Conference on Human–Computer Interaction (HCI), Theory and Practice, vol. 2, pp. 806–810 (2003)
34. Xydias, G., Kouroupetroglou, G.: Tone-Group F0 selection for modeling focus prominence in small-footprint speech synthesis. *Speech Commun.* **48**(9), 1057–1078 (2006)
35. Yesilada, Y., Stevens, R., Goble, C., Hussein, S.: Rendering tables in audio: the interaction of structure and reading styles. In: Proceedings of ACM Conference on Assistive Technologies (ASSETS), pp. 16–23 (2004)