Language analysis and prosodic feature annotation for high quality speech synthesis

Dimitris Spiliotopoulos[*]

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
dspiliot@di.uoa.gr

Abstract. This thesis addresses the problem of the identification, extraction, annotation and modelling of prosodic features aimed for high quality speech synthesis. For document complex visual structures, a method for the acoustic specification modelling of simple and complex data tables, derived from the human paradigm, based on a series of psychoacoustic experiments that were set-up for providing speech properties obtained from prosodic analysis of natural spoken descriptions of data tables according to their properties is analysed. In the case of plain and enriched texts, this work presents a language independent framework for language analysis and semantic annotation for use in plain and enriched texts, as well as the approach auditory universal accessibility acoustic rendition of data tables using naturally derived prosody specification for Document-to-Audio systems. The framework basis, linguistic resources and language analysis procedures (word/sentence identification, part-of-speech, prosodic feature annotation) for text annotation/processing for plain or enriched text corpora aims to produce an automated XML-annotated enriched prosodic markup for English and Greek texts, for improved synthetic speech. The prosodic features are classified according to the analysis of the speech-specific characteristics for their role in prosody modelling and passed through to the synthesizer via an extended SOLE-ML description. Evaluation results show that using selectable hybrid methods for part-of-speech tagging high accuracy is achieved. Annotation of a large generated text corpus containing 50% enriched text and 50% canned plain text produces a fully annotated uniform SOLE-ML output containing all prosodic features found in the initial enriched source. Furthermore, additional automatically-derived prosodic feature annotation and speech synthesis related values are assigned, such as word-placement in sentences and phrases, previous and next word entity relations, emphatic phrases containing proper nouns, and more.

1. Introduction

Text annotation is a procedure where certain meta-information gets identified and associated with the entities in a text corpus. Such information is commonly used in computational linguistics for language analysis, speech processing, natural language processing, speech synthesis, and other areas. The type of information that is analyzed and associated to text units may span the linguistic analysis tree (grammatical,

[*] Dissertation Advisor: Georgios Kouroupetroglou, Assistant Professor

syntactic, morphological, semantic, pragmatic, phonological, phonetic), as well as include any other description that may be of use.

General purpose Text-to-Speech (TtS) systems use certain language processing subsystems, such as sentence segmentation and part-of-speech tagging, for the analysis of the written text input. Depending on the actual system, such analysis may suffer from inherent statistical error accuracy that may be due to the design and implementation of the respective modules or language ambiguity. However, TtS systems may employ language analysis modules that are designed for high accuracy in specific thematic domains for which they seem to perform adequately. The respective accuracy when used for generic or other thematic domains may fall under unacceptable levels. Additionally, the language processing modules embedded in TtS systems are not usually designed to identify and extract higher-level linguistic information, such as semantic or pragmatic factors, that may be used to aid speech synthesis.

Previous works that have explored natural language generated texts show that linguistically enriched annotated text input to a speech synthesizer can lead to improved naturalness of speech output [1], [2]. Generation of tones and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [3]. When such input can be provided, the language processing from the TtS system can be superseded.

In the case of visual document structures, the aural rendition of data tables constitutes a hard task because of the difficulty in accessing the semantic information under the visual structure. The complex visual structures bear a distinct association between the physical layout and the underlying logical structure [4]. The columns and rows of a table represent the logical connections [5]. Hurst presented a table theory that "views the table as a presentation of a set of relations holding between organized hierarchical concepts" [6]. Previous works also show that information about the semantic structure of HTML tables can be used to aid navigation and browsing of such visual components [7]. 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 [8]. 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 [10]. Other suggestions include automated approaches for retrieval of hierarchical data from HTML tables [11]. Smart browsers are used to access critical information for use in reading tables as well as linearization techniques are employed for transforming tables into a more easily readable form by screen readers [12]. Table browsing techniques include the use of Conceptual Graphs for the classification of header and data cells using Hidden Markov Models for identification [13] as well as systems that decompile 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 [15]. 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 similarly complex visual structures, such as frames – either automatically or manually [16]. However, since the problem is addressed on the visual level, the transfer of the linearized visual stucture to the actual spoken form remains problematic.

This work examines the problem of semantic feature annotation text-to-speech systems and the acoustic representation of complex visual structures for document-to-audio systems around two main issues:

1. The identification, extraction, analysis, annotation of prosodic features from plain or enriched texts and the automatic generation of a uniform enriched text description, and the design, implementation and evaluation of a methodology for automatic annotation of large domain-dependent Greek text corpora.
2. The analysis of the visual and semantic characteristics of data tables the utilisation of prosodic attributes based on a series of psychoacoustic experiments on the spoken format of data tables.

The work in this thesis produced the following results:

1. A language-independent framework for language analysis and prosodic feature annotation.
2. A methodology for modelling emphatic events in plain and enriched texts as well as prosodically enriched input for speech synthesis.
3. A methodology for universal prosody specification of data tables in documents for acoustic rendition via the document-to-audio approach

2 Semantics, complexity and navigation of data tables

Tables are visual structures spanning two dimensions, used as a means of grouping information.

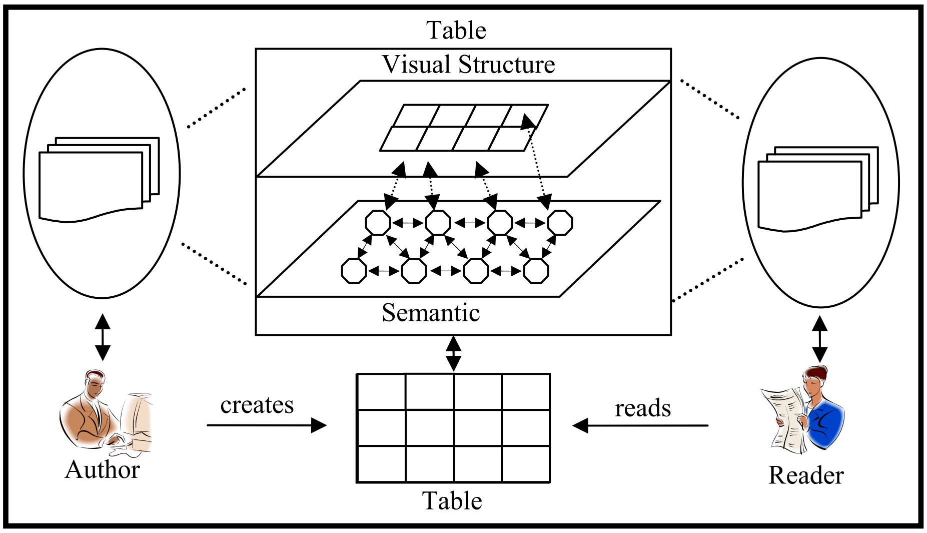

Fig. 1. Table Structural Information – Creating and Visually Reading a Data Table

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).

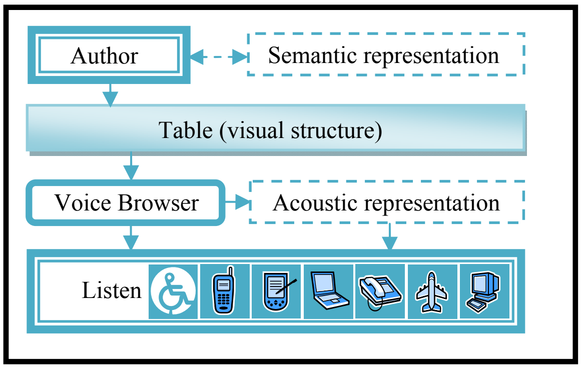

Fig. 2. Spoken Format Rendering of Tables

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. 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 [17].

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 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. 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.

Complex table can be thought of as a three-dimensional structures, compared to the two-dimensional simple data tables. Fig. 3 shows a 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.

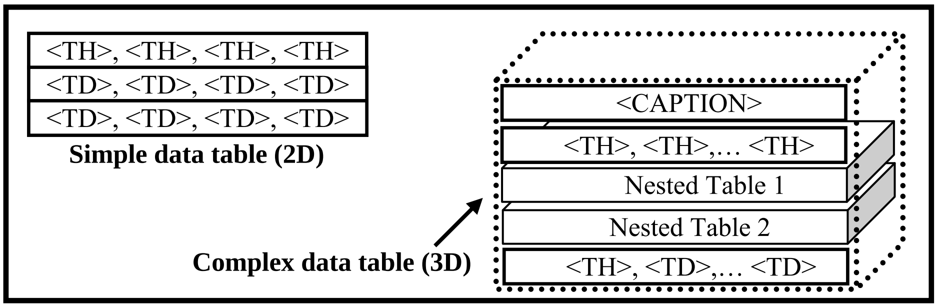

Fig. 3. Simple and Complex Data Table Dimensional Comparison

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. Additionally both header and data cells in nested tables are governed by the top-level headers depicted in Fig. 3, 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 which the header is associated with. Moreover, using the scope attribute on data cells forces them to behave like header cells.

3. Enriched Text Annotation

TtS systems generally accept plain (or “raw”) text as input, using specialized algorithms to internally generate the needed natural language data prior to synthesis. However, the algorithms that are usually implemented for such tasks are not powerful enough to broadly identify additional information about several linguistic phenomena from the plain text form, thus limiting the depth of text analysis and the derived description. A valuable alternative is to use pre-processed annotated text as input to the speech synthesizer. Enriched text of that kind exhibits major advantage over plain text as it retains structural and discourse level information. Each of the above types of linguistic information is described by sets of features that can be used to generate improved prosody in speech synthesis. Depending on the domain as well as the type of text, different sets of features may be used for maximum improvement.

As an alternative to generated text, existing plain text can be adequately processed to derive annotated NLG-similar output, essentially gaining advantage for the prosody modelling stage in speech synthesis. In order to do that efficiently, automated analysis and annotation should be made available for the most language analysis stages. A breakdown of the identifiable distinct processes is:

- Word/Sentence identification and segmentation.
- Morphological analysis (part-of-speech tagging and noun-phrase identification).
- Calculation/annotation of prosodic features.
- Creation/export to appropriate XML format description for speech synthesis.

As described in the following paragraphs, fully automated analysis can be achieved for all processes. The enriched linguistic annotation needs to be exported to a well-tested and reliable standard markup, such as XML. All the above processes have been implemented through the utilisation of the Ellogon Language Engineering Platform [18] platform and implemented the speech-oriented natural language analysis and annotation components [19].

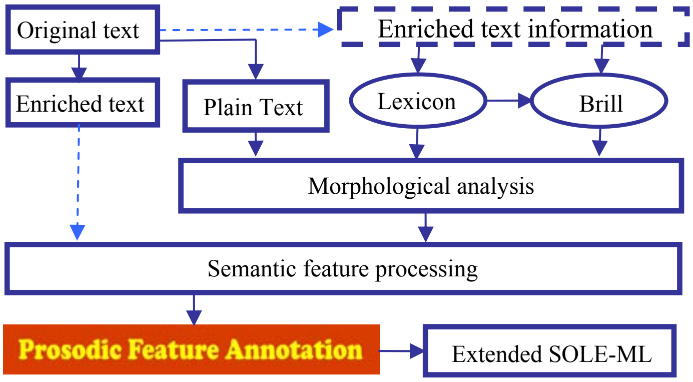

Fig. 4. The annotation workflow

As shown in Figure 4, the input may be either fully or partially annotated text (e.g. from a Natural Language Generator) or plain unformatted text. Information from the enriched input is extracted and used for the annotation of the plain text. The prosodic feature annotation assigns prosodically important values for the calculation of intonation focus for higher quality speech synthesis.

4. Morphological Analysis

Pre-processing mainly includes word and sentence identification, as well as part-of-speech (POS) tagging. For English texts, a POS tagger based on machine learning is used, while for Greek texts a combination of lexicon-based and machine learning analysis is preferred. Word and sentence identification are performed by a rule-based component that presents an accuracy that approaches 100% for both languages. For part-of-speech tagging, the implementation is based on Transformation-based Error-driven learning [20] and provides models for English with accuracy that approaches 97%, measured as average of several accuracy measurements performed on various thematic domains.

For Greek, the common approach for most embedded systems is the use of Lexicon-based POS taggers. This approach is used by most speech synthesisers and yields accuracy between 75-85% depending on the domain of the text corpora. This low accuracy in most cases hinders poor final prosody prediction. This is due to Greek being an inflectional language with vast vocabulary that cannot be covered by lexicons. In order to increase the accuracy of POS tagging when processing documents in the Greek language, we used a hybrid approach, a combination of a lexicon-based POS tagger and a rule-based (Brill) POS tagging component. Two morphological lexicons for the Greek language have been combined in order to build a lexicon-based POS tagger with the highest possible coverage. The first lexicon is a large-scale morphological lexicon for the Greek language, developed exclusively for the system [21]. The lexicon consists of ~60,000 lemmas that correspond to ~710,000 different word forms (Greek is an inflexional language). The second lexicon is property of the Speech Group, University of Athens, used in the DEMOSTHeNES speech composer [22] and contains ~60,000 lemmas, which correspond to ~650,000 word forms. Both lexicons yield a word form identification span of ~880,000.

The hybrid approach was applied to the full generated corpus using two different ways in order to examine and evaluate the best procedure:

In the first approach, the built-in POS tagger and the lexicon-based POS tagger are both applied independently. Depending on the actual corpus and relative precision of the lexicon and HBrill modules, a word can be set to be assigned a value by either tagger (or both). The default state is that if a word contained in any of the two lexicons and thus is assigned a POS category by the lexicon-based tagger, this categorization becomes the final POS of the word, ignoring any categorization performed by the machine learning POS tagger. On the other hand, if a word is not found in any of the two lexicons, the categorization presented by the built-in POS tagger is assigned. The machine learning based POS tagger uses an extension of the Penn Tree Bank tagset, which contains additional information regarding number and gender of words [23]. This approach achieved an accuracy of 95%.

The second approach saw that the lexicon-based component was always followed be the machine learning POS tagger. Initial values were extrapolated from the lexicons and fed as initial states for the machine learning algorithm which produced the final value. In the case of partially annotated texts, the values of the pre-annotated word tokens were used for initial values in similar word forms since they were 100% correct coming from the natural language generator. This approach yielded total accuracy >97% for plain and >98% for partially annotated Greek texts and was the preferred choice.

5. Prosodic Feature Annotation

Previous research shows that higher-level linguistic information such as semantic features can be used to improve prosody modelling for speech synthesis [24]. This is because part-of-speech and phrase type information alone cannot always infer certain intonational focus points since those are not only affected by syntax but also by semantic and pragmatic factors [25]. For prosody modelling in speech synthesis, these factors can be used for calculation, deduction and verification of focus prominence and are accounted for by enriching the text corpus accordingly.

In our corpus, the plain text was annotated using the hybrid part-of-speech technique. Then, the results were validated and updated using the part-of-speech information from the enriched corpus. The benefit is twofold, the values are checked with the correct ones from the enriched text (if such is available for a lexical item) and key items are assigned specific values where appropriate. After that, certain semantic factors are calculated and added to the meta-information pool

Figure 5 shows an example of how semantic factors such as *newness (new or old infromation)*, *contrast*, *explicit emphasis*, *first or second argument to verb* may be used for determining intonational focus prominence.

This exhibit is an amphora, created during the archaic period. It dates from the early fifth century before Christ. It was found in Beotea but it was made in Athens. It depicts a warrior performing splachnoscopy before leaving for battle. Splachnoscopy is the study of animal entrails, through which people tried to predict the future. It was one of the most common divination methods used in the archaic period. This amphora was painted by the painter of Kleofrades and was decorated with the red figure technique.

#	Lexical item	Focus	Prosodic features
1	amphora	3	[newness_TRUE, arg2]
2	archaic period	3	[newness_TRUE, arg2]
3	Christ	2	[newness_FALSE, proper-noun]
4	It was … in Athens	-	[contrast]
5	Beotea	3	[newness_TRUE, arg2, ***proper-noun***]
6	Athens	3	[newness_TRUE, arg2, ***proper-noun***]
7	splachnoscopy	3	[newness_TRUE, arg2, ***emphasis***]
8	Splachnoscopy	3	[newness_FALSE, arg1, ***emphasis***]
9	archaic period	1	[newness_FALSE, arg2]
10	amphora	1	[newness_FALSE, arg2]
11	the painter of Kleofrades	3	[newness_TRUE, arg2, ***proper-noun***]
12	red figure	3	[newness_TRUE, arg2, ***proper-noun***]

Fig. 5. Focus prominence identification from semantic factors

The intonational focus is assigned in a scale of three, strong focus '3', normal focus '2', and weak focus '1'. The features in bold are the ones computed from the information provided by the enriched portion of the text. Although *newness* is a key factor for strong intonational focus, certain validation checks in the algorithm make sure that only the proper lexical items are assigned. Validation factors are proper-noun and second-argument-to-verb (arg2) as well as explicit factors such as *emphasis* and *contrast*. As a result, strong focus ‘3’ is assigned when validation factors arg2 and/or proper-noun exist for a new information (e.g., #1-2) while old information (e.g. #8-9) gets weak focus, as shown below:

Strong focus prominence:	newness_TRUE (validation=passed)
Normal focus prominence:	newness_FALSE (validation=passed)
Weak focus prominence:	newness_TRUE (validation=failed)
No focus prominence:	newness_FALSE (validation=failed)

However, it can be seen that explicit factors elevate the focus prominence, clearly providing explicit emphatic events as in the case of *splachnoscopy* (# 8) where if it were not for the *emphasis* factor it would have been assigned weak focus since it is an already given piece of information.

Contrast is a rather generalized annotation that was implemented as a rule in the process and was initiated due to the fact that domain contained several instances of similarly NLG-derived phrases. The rule applies to both Greek and English text and elevates the main verb(s) and the conjunction to a mid-level emphasis, thus assigning explicitly a normal focus prominence marker (not showing in Figure 5).

From the above, it is obvious that the precision of part-of-speech identification is quite important since certain lexical items are validated for their assigned focus prominence using the part-of-speech information against the identified prosodic features.

6. SOLE-ML description

The enriched text meta-information is encoded using an open XML schema. It is an extension (to cater for the semantic/prosodic description) of the SOLE-ML description [26], and was originally built as an annotation scheme for CtS synthesis, used as markup for the enriched text output of the ILEX generator.

```
<utterance>
<relation name="Word" structure-type="list">
<wordlist>
<w id="w20">It</w>
<w id="w21">was</w>
<w id="w22">found</w>
<w id="w23">in</w>
<w id="w24">Beotea</w>
<w id="w25">but</w>
<w id="w26">it</w>
<w id="w27">was</w>
<w id="w28">made</w>
<w id="w29">in</w>
<w id="w30" punct=".">Athens</w>
</wordlist>
</relation>
<relation name="Group" structure-type="list">
</relation>
<relation name="Syntax" structure-type="tree">
<elem phrase-type="S">
<elem phrase-type="prosody" contrast>
<elem lex-cat="PRONOUN" href="w#id(w20)"/elem>
<elem phrase-type="prosody" mid-emphasis-verb>
<elem lex-cat="VERB" href="w#id(w21)"/elem>
<elem lex-cat="VERB" href="w#id(w22)"/elem>
</elem>
<elem lex-cat="PREPOS" href="w#id(w23)"/elem>
<elem phrase-type="prosody" newness="true", arg2, proper-noun>
<elem lex-cat="NOUN" href="w#id(w24)"/elem>
</elem>
<elem phrase-type="prosody" mid-emphasis-conj>
<elem lex-cat="CONJNCT" href="w#id(w25)"/elem>
</elem>
<elem lex-cat="PRONOUN" href="w#id(w26)"/elem>
<elem phrase-type="prosody" mid-emphasis-verb>
<elem lex-cat="VERB" href="w#id(w27)"/elem>
<elem lex-cat="VERB" href="wl#id(w28)"/elem>
</elem>
<elem lex-cat="PREPOS" href="w#id(w29)"/elem>
<elem phrase-type="prosody" newness="true", arg2, proper-noun >
<elem lex-cat="NOUN" href="w#id(w30)"/elem>
</elem>
</elem>
</elem>
</relation>
</utterance>
```

Fig. 6. The XML description

SOLE-ML has been successfully used in earlier works, a well-tested means of representing enriched linguistic information, and is now standard input of the

DEMOSTHeNES speech composer. The automatic extraction to the extended XML description based on SOLE-ML encodes all prosodic features. Figure 6 shows the XML output for the sentence *"It was found in Beotea but it was made in Athens."* from the text paragraph shown in Figure 5.A wordlist of all tokens (words) and punctuation values takes up the first part (<*wordlist*>), followed by the syntax tree, prosodic features, and other high-level information (<relation>). This is the input for the speech synthesizer.

7. Results and conclusion

The proposed framework utilises the meta-information contained in enriched automatically generated texts in order to compute and annotate both the enriched and the plain text with prosodic features that aid focus prominence in synthetic speech. The uniformly annotated target text contains enough elements to aid focus prominence using the modified speech synthesizer for Greek or an equivalent for English. An evaluation of the performance of the hybrid morphological analysis methods was performed for both Greek and English texts, shown in Table 1.

Table 1. Plain text part-of-speech annotation

Corpus (plain text)		Lexicon	Brill	Hybrid
English	precision	0.90	0.97	0.98
	recall	0.77	0.92	0.98
Greek	precision	0.88	0.94	0.98
	recall	0.75	0.84	0.92

These results include the prime importance validation factor in our approach *proper-noun,* while exclude all other features that are calculated later in the process.

Enriched text annotation using naturally generated meta-information for a specific domain greatly enhances the intonational focus prominence predictors of a speech synthesizer. A strong indication of focus based on the new or already given information validated by the type of the lexical item works exceptionally well for domain-dependent corpora where the prosodic features can be more easily calculated automatically. This leads to enhanced input for speech synthesis, while bypassing all internal language analysis modules of the synthesizer, results on improved prosody prediction.

This work main contribution is a framework and methodology for prosodic feature annotation for improved synthetic speech quality, both for plain and automatically generated texts. Using large generated corpora and the automatic linguistic analysis and annotation processes, prosodic models were produced and tested as input for a speech synthesiser resulting in high quality speech synthesis. Moreover, for documents that contain visual structures, prosody specifications for simple and complex data tables were calculated using the results from expert human renditions and respective psychoacoustic experimentation using both visually capable and blind subjects.

References

1 Pan, S., McKeown, K., and Hirschberg, J., “Exploring features from natural language generation for prosody modeling“ Computer Speech and Language, 16:457-490, 2002

2 Xydas, G., Spiliotopoulos, D., and Kouroupetroglou, G., "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora". IEICE Trans. of Inf. and Syst., Special Section on "Corpus-Based Speech Technologies", vol. E88-D, no 3, March 2005, pp. 510-518.

3 Black, A., and Taylor, P., “Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input” Proc. 3rd Int. Conf. on Spoken Language Processing, pp.715–718, Yokohama, Japan, 1994.

4. Ramel, J-Y., Crucianou M., Vincent, N., Faure, C.: Detection, Extraction and Representation of Tables. Proc. 7th Int. Conf. Document Analysis and Recognition (ICDAR), pp.374-378 (2003)

5. Silva, A.C., Jorge, A.M., Torgo, L.: Design of an end-to-end method to extract information from tables, International Journal of Document Analysis and Recognition 8(2), Special issue on detection and understanding of tables and forms for document processing applications, pp. 144-171 (2006)

6. Hurst, M.: Towards a theory of tables, International Journal of Document Analysis, 8(2-3), pp.123-131 (2006)

7. Pontelli, E., Gillan, D., Xiong, W., Saad, E., Gupta, G., Karshmer, A.: Navigation of HTML Tables, Frames, and XML Fragments. Proc. ACM Conf. on Assistive Technologies (ASSETS), pp.25-32 (2002)

8. Pitt, I., Edwards, A.: An Improved Auditory Interface for the Exploration of Lists. ACM Multimedia 1997, pp. 51-61 (1997)

9. Hurst, M., Douglas, S.: Layout & Language: Preliminary Experiments in Assigning Logical Structure to Table Cells. Proc. 4th Int. Conf. Document Analysis and Recognition (ICDAR), pp.1043-1047 (1997)

10. Filepp, R., Challenger, J., Rosu, D.: Improving the Accessibility of Aurally Rendered HTML Tables. Proc. ACM Conf. on Assistive Technologies (ASSETS), pp.9-16 (2002)

11. Lim, S., Ng, Y.: An Automated Approach for Retrieving Hierarchical Data from HTML Tables. Proc. 8th ACM Int. Conf. Information and Knowledge Management (CIKM), pp.466-474 (1999)

12. Yesilada, Y., Stevens, R., Goble, C., Hussein, S.: Rendering Tables in Audio: The Interaction of Structure and Reading Styles. Proc. ACM Conf. Assistive Technologies (ASSETS), pp.16-23 (2004)

13. 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, Proc. of the ACM Conf. on Universal Usability, Vancouver, Canada, pp. 134-141 (2003)

14. Oogane, T., Asakawa, C.: An Interactive Method for Accessing Tables in HTML, Proc. of Intl. ACM Conf. on Assistive Technologies, 1998, pp. 126-128 (1998)

15. Embley, D.W., Hurst, M., Lopresti, D.P., Nagy, G.: Table-processing paradigms: a research survey. International Journal of Document Analysis, 8(2-3), pp.66-86 (2006)

16. Pontelli, E., Gillan, D.J., Gupta, G., Karshmer, A.I., Saad, E., Xiong, W.: Intelligent non-visual navigation of complex HTML structures. Universal Access in the Information Society 2(1), pp. 56-69 (2002)

17. Pontelli, E., Xiong, W., Gupta, G., Karshmer, A.: A Domain Specific Language Framework for Non-visual Browsing of Complex HTML Structures. Proc. ACM Conf. Assistive Technologies (ASSETS), pp.180-187 (2000)

18 Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., and Spyropoulos, C.D., “Ellogon: A New Text Engineering Platform”. Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), pp. 72-78, Las Palmas, Canary Islands, Spain, May 2002.

19 Ellogon Language Engineering Platform, Speech tools add-ons, *www.ellogon.org/speech/*

20 Brill, E., “Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging”. Computational Linguistics, 21:543-565, 1995.

21 Petasis, G., Karkaletsis, V., Farmakiotou, D., Androutsopoulos, I., and Spyropoulos, C.D., “A Greek Morphological Lexicon and its Exploitation by Natural Language Processing Applications”. Lecture Notes on Computer Science, vol.2563, Springer Verlag, 2003.

22 Xydas, G. and Kouroupetroglou, G., “The DEMOSTHeNES Speech Composer”, Proc. 4th ISCA Workshop on Speech Synthesis, Perthshire, Scotland, pp.167-172, 2001.

23 Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., and Androutsopoulos, I., “Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques”. Fakotakis et al. (Eds.), Machine Learning in Human Language Technology, pp. 29-34, 1999.

24.O’Donnel, M., Mellish, C., Oberlander, J., and Knott, A., “ILEX: An architecture for a dynamic hypertext generation system”, Natural Language Engineering, vol.7, no.3, pp. 225-250, 2001

25 Bolinger, D., Intonation and its Uses: Melody in grammar and discourse, Edward Arnold, London, 1989.

26 Hitzeman, J., Black, A., Mellish, C., Oberlander, J., Poesio, M., and Taylor, P., “An annotation scheme for Concept-to-Speech synthesis”, Proc. 7th European Workshop on Natural Language Generation, Toulouse France, pp. 59-66, 1999.