Modeling Prosodic Structures in Linguistically Enriched Environments

Gerasimos Xydas, Dimitris Spiliotopoulos, and Georgios Kouroupetroglou

University of Athens
Department of Informatics and Telecommunications, Greece
Email: gxydas@di.uoa.gr, dspiliot@di.uoa.gr, koupe@di.uoa.gr

Abstract. A significant challenge in Text-to-Speech (TtS) synthesis is the formulation of the prosodic structures (phrase breaks, pitch accents, phrase accents and boundary tones) of utterances. The prediction of these elements robustly relies on the accuracy and the quality of error-prone linguistic procedures, such as the identification of the part-of-speech and the syntactic tree. Additional linguistic factors, such as rhetorical relations, improve the naturalness of the prosody, but are hard to extract from plain texts. In this work, we are proposing a method to generate enhanced prosodic events for TtS by utilizing accurate, error-free and high-level linguistic information. We are also presenting an appropriate XML annotation scheme to encode syntax, grammar, new or given information, phrase subject/object information, as well as rhetorical elements. These linguistically enriched has have been utilized to build realistic machine learning models for the prediction of the prosodic structures in terms of segmental information and ToBI marks. The methodology has been applied by exploiting a Natural Language Generator (NLG) system. The trained models have been built using classification via regression trees and the results strongly indicate the realistic effect on the generated prosody. The evaluation of this approach has been made by comparing the models produced by the enriched documents to those produced by plain text of the same domain. The results show an improved accuracy of up to 23%.

1 Introduction

One of the most important tasks in Text-to-Speech (TtS) synthesis is the prediction of the prosodic structure of the utterance prior to prosody rendering. For example, phrase break prediction is fundamental for F0 contour generation, duration models and pause insertions [1]. We define the prosodic structure as a set of features related to the position and the type of (a) prosodic phrase breaks, (b) pitch accents, (c) phrase accents and (d) boundary tones. The rule-driven approaches for their prediction fail to capture the richness of human speech, are generally difficult to write, to adapt to new domains and new set of features, and usually provide the prosody generation module with poor input. On the other hand, machine learning planning can yield more reasonable results provided that the size of the sample data increases along with the size of the selected features and their variability.

Prosody construction is a complex process that involves the analysis of several linguistic phenomena, which is usually prone to errors. For instance, part-of-speech (POS) identification fails in 5% of the cases for Greek using statistical taggers [2], while syntax and metric trees are hard to construct. The generation of tones and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [3]. Former works show that certain

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2004, LNAI 2448, pp. 1–8, 2004.

relations can affect pitch assignment and placement, such as discourse structure [4], already given or new information [5] and contrast [6].

However, enriched information like focus prominence and rhetorical relations is difficult to be extracted from plain texts. Concept-to-Speech (CtS) systems (i.e. an NLG coupled with a TtS [7]) can provide linguistic information which can be used in prosody modeling [8,15].

In order to study the effects of the introduction of linguistic meta-information in documents, we compare prosodic models made by linguistically poor to enriched information. Due to the lack of a sophisticated appropriate linguistic analyzer, we have used a Natural Language Generation (NLG) system that can generate texts annotated with high level error-free linguistic factors in contrast to plain texts [9]. As NLG systems deal with written text and fail to represent spoken language, we have extended an XML markup scheme (SOLE [10]) to provide more evidence of stress and intonational focus information in documents. Using this meta-information, we built 3 CART models [11] for prosodic phrase breaks, pitch accents and endtones using a linguistically enriched annotated voice corpus. The results show improved classification of the selected features in the case of the annotated documents, as presented at the end of this paper.

2 Towards intonational focus prominence

One of the many factors that affect speech prosody is *intonational focus* prominence. This is a property that is well hidden in language and manifests itself in utterances. Strong leads towards identification of the intonational focus (phonological stress) points in each phrase can be revealed by analyzing the linguistic information [12]. Intonational focus points are prosodic instances where (mainly) the pitch is used to denote the center of meaning for a phrase. However the above information, although valuable, is not enough for all occasions. 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 semantics and pragmatic factors [13]. So, even for the limited number of sentence structures generated for this domain several more useful features exist inside the language generation stages that can be of value to the speech synthesis.

It is affected by specific linguistic information factors, alone or in combination, such as syntax, rhetorical relations, discourse structure, contrast, already given or new information, and more. These properties require sophisticated linguistic analysis during TtS synthesis in order to be extracted. This information is not straightforwardly present in plain texts since the written form is stripped from it. However, NLG systems can generate it and provide it to the TtS in the form of annotated text.

In this work, useful information in the form of specific properties for lexical items is utilized to aid intonational focus (Fig. 1).

By examining the above properties the chances of having intonational focus in a syllable within a particular phrase is computed. Focus prominence is assigned to lexical items that are parts of Noun Phrases (NPs) in varying degrees as shown below:

Strong focus prominence:	newness=new	AND	validation=passed
Normal focus prominence:	newness=old	AND	validation=passed
Weak focus prominence:	newness=new	AND	validation=failed
No focus prominence:	newness=old	AND	validation=failed

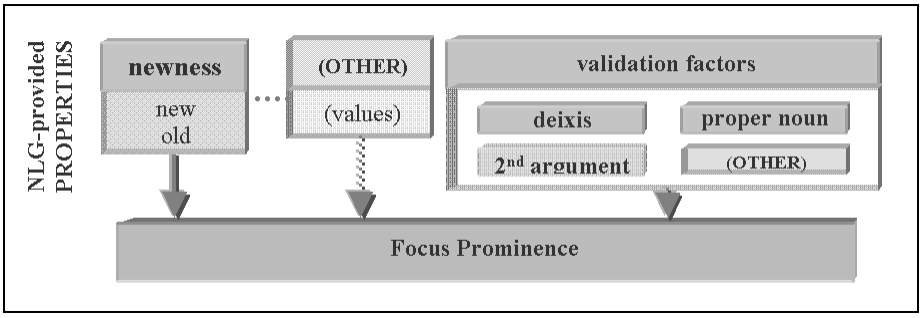

Fig. 1. Noun-phrase focus prominence elements

In our case, an implementation of the ILEX [14] NLG has been used. The SOLE markup output of the NLG provides enumerated word lists and syntactic tree structures to the TtS (DEMOSTHeNES) [16]. As shown in Figure 2, on the syntactic tree, error-free information exists at the phrase level about the phrase type (sentence, noun phrase, prepositional phrase, relative clause, etc) as well as at word level about the part-of-speech (determiner, noun, verb, preposition, etc.). The annotated text of the chosen domain (museum exhibits [20]) contains sentences of a fairly straightforward (SVO) structure. However, enough variation is provided in the domain for the range of phrase types and lexical categories mentioned above to occur in sentences. The particular generator can produce such detailed meta-information.Since the SOLE specification was not speech aware, it was extended in order to accommodate those elements that were used towards identification {ID} and validation {VAL} of intonational focus. These properties are attached to NPs:

{ID} New or already given information: newness [new/old]
{VAL} Whether NP is second argument to the verb: arg2 [true/false]
{VAL} Whether there is deixis: genitive-deixis, accusative-deixis [true/false]
{VAL} Whether there is a proper noun in the noun phrase: proper-group [true/false]

3 The corpus setup

The *FULL* corpus was constituted of 516 utterances (5380 words and 13214 syllables) of descriptions of museum exhibits. However, the 48.03% of the words was delivered without any linguistic information from the NLG component. These were marked as “canned” phrases (2719 words and 6700 syllables) and constituted the *CANNED* corpus subset. We also filtered out a pure *ENRICHED* subset of 192 enriched utterances (1534 words and 3794 syllables). A comparison of the aforementioned sets follows to show the improvements achieved in the *ENRICHED* subset case.

The text corpus was first interpreted by the Heterogeneous Relation Graph (HRG) [17] component of the TtS and then it was exported in a properly visualized and readable RTF format (Fig. 3). A professional speaker captured the spoken expressions of a museum guided tour, and, by following the annotation directions, rendered the different levels of focus according to the properties attached to lexical items provided by the NLG. The produced

```
<utterance>
<relation name="Word" structure-type="list">
<wordlist>
…
<w id="w7">που</w>
<w id="w8">δημιουργήθηκε</w>
<w id="w9">κατά</w>
<w id="w10">τη</w>
<w id="w11">διάρκεια</w>
<w id="w12">της</w>
<w id="w13">αρχαϊκής</w>
<w id="w14" punct=".">περιόδου</w>
…
</wordlist>
</relation>
…
<elem phrase-type="S">
<elem lex-cat="PRP" href="words.xml#id(w7)"/>
<elem lex-cat="V" href="words.xml#id(w8)"/>
<elem phrase-type="PP">
<elem lex-cat="IN" href="words.xml#id(w9)..id(w11)"/>
<elem phrase-type="NP" newness="new" arg2="true" proper-group="true"
genitive-deixis="true">
<elem lex-cat="DT" href="words.xml#id(w12)"/>
<elem lex-cat="N" href="words.xml#id(w13)..id(w14)"/>
</elem>
</elem>
</elem>
…
</relation>
</utterance>
```

Fig. 2. A SOLE-ML example

voice corpus was further automatically segmented and hand annotated using the GR-ToBI marks [18] providing description of tonal events. As the frequency of some marks is low in the corpus, we grouped them, while they can be useful when more data is available. Thus, pitch accents are represented by 5 binary features (Table 1) and endtones (ToBI phrase accents and boundary tones grouped together since GR-ToBI does not allow them to co-occur) by 4 features (Table 2). Break indices mark boundaries (0 to 3) that are represented by a subjective notion of disjunction between words. The additional tonal events - Sandhi (s), mismatch (m), pause (p), and uncertainty (?) - diacritics were eliminated (Table 3).

Table 1. Accent groups (pitch accents) in the ENRICHED subset.

Accent	L*	H*	L*+H	L+H*	H*+L	Total
# occurrences	48	138	321	285	263	1055
Occurrences **%**	4.55	13.08	30.43	27.01	24.93	100.00

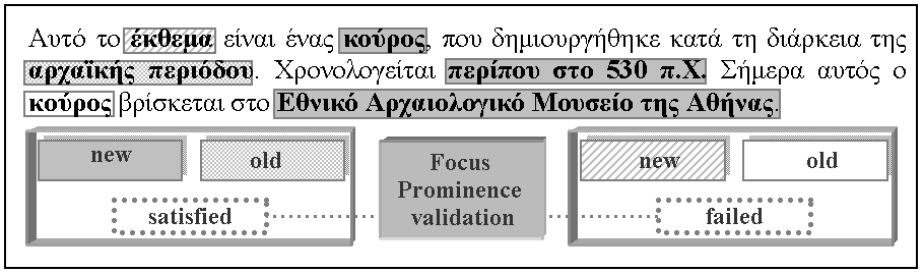

Fig. 3. RTF format document sample (This exhibit is a kouros, created during the archaic period. It dates from circa 530 B.C. Currently this kouros is in the National Archaeological Museum of Athens.)

Table 2. Endtone groups (phrase accents, boundary tones) in the ENRICHED subset.

Endtone	L-	H-	L-L%	H-H%	Total
# occurrences	9	114	160	2	285
Occurrences **%**	3.16	40.00	56.14	0.70	100.00

4 Building the models

For the prediction of the GRToBI marks we used the wagon [19] program to build classification trees. The features selected for the training were: part of speech, number of stressed/unstressed syllables/words since last major break and until next major break, stress, punctuation, phrasing information (*ENRICHED*), phrase type (*ENRICHED*) and focus (*ENRICHED*). The focus feature is computed from the "newness", "arg2", "deixis" and "proper noun" features (Fig. 1). A window of 5 items (current, 2 before and 2 after) has been used in all cases, leading to a set of 30 attributes for breaks, 35 for accents and 40 for endtones. In the *CANNED* subset only POS was used as part of the linguistic analysis. The tables below show the classification matrix for each model using the 10-fold cross validation method.

Table 7 illustrates the significant enhancements in the correlation between the observed and the train data in the case of the *ENRICHED* subset. The *CANNED* subset can be seen as untagged, plain text. The *FULL* set is a mix of tagged and untagged information, while the *ENRICHED* subset contains enriched meta-information about the text.

Table 3. Occurrences of break indices in the ENRICHED subset.

Break	0	1	2	3	Total
# occurrences	468	730	144	192	1534
Occurences %	30.50	47.59	9.39	12.52	100.00

Table 4. Confusion matrix and accuracy by class for the phrase break model (cor. 89.12%).

Break	0	1	2	3	Precision	Recall
0	420	46	2	0	0.857	0.897
1	68	644	18	0	0.893	0.882
2	2	29	113	0	0.85	0.785
3	0	1	0	191	1	0.995

Table 5. Confusion matrix and accuracy by class for the pitch accent model (cor. 85.56%).

Accent	NONE	L+H*	L*+H	H*+L	H*	L*	Precision	Recall
NONE	2626	27	33	43	7	3	0.919	0.959
L+H*	58	149	47	16	15	0	0.696	0.523
L*+H	68	14	221	4	13	1	0.644	0.688
H*+L	42	9	4	203	3	2	0.712	0.772
H*	41	12	28	9	47	1	0.534	0.341
L*	22	3	10	10	3	0	0	0

The example below exploits the produced models and illustrates the well placed pitch accents, their realistic variation and the natural sounding choice of break index 0 in the second phrase ("*μία υδρία*" – *a hydria*) and in the third phrase ("*κατά τη διάρκεια*" – *during the*). The latter leads to the correct placement of the intonational focus to the nouns "*υδρία*" and "*διάρκεια*". The words are enclosed in brackets in the form of [syl1 syl2$^{<\text{pitch_accent}>}$ … sylN]$_{<\text{break\_index}>}$ /<endtone>

"*Αυτό το έκθεμα είναι μία υδρία που δημιουργήθηκε κατά τη διάρκεια της κλασσικής περιόδου*" (*This exhibit is a hydria, created during the classical period.*)

[a fto^{L+H*}]$_1$ [to]$_0$ [e^{H*+L} kTe ma]$_2$ /H-

[i^{L*+H} ne]$_1$ [mi^{L*+H} a]$_1$ [i Dri^{H*+L} a]$_2$ /H-

[pu]$_0$[Di mi u rji^{L*+H} Ti ce]$_1$ [ka ta]$_0$ [ti]$_0$ [Dja^{H*} rci a]$_1$ [tis]$_0$ [kla si cis^{L*+H}]$_1$ [pe ri o^{H*+L} Du]$_3$ /L-L%

Table 6. Confusion matrix and accuracy by class for the endtone model (cor. 99.10%).

Endtone	NONE	L-L%	H-H%	H-	L-	Precision	Recall
NONE	3487	14	0	8	0	0.999	0.994
L-L%	4	156	0	0	0	0.907	0.975
H-H%	0	2	0	0	0	0	0
H-	1	0	0	113	0	0.897	0.991
L-	0	0	0	5	4	1	0.444

Table 7. Correlation of observed and test data in the three sets of the corpus.

Set →	***CANNED***	***FULL***	***ENRICHED***
Breaks	66.01%	72.69%	89.12%
Accents	71.67%	76.45%	85.56%
Endtones	97.59%	97.93%	99.10%
# syllables	6700	13214	3794

5 Discussion and conclusions

Carefully selected and properly structured linguistic meta-information has been used to improve the prediction of phrases and intonational events. An extended SOLE-ML specification has been formulated to accommodate the required factors that can imply focus prominence. Thus, using an NLG system we provided the speech synthesizer with evidence of stress and intonational focus. The improvement in the delivery of prosody in cases where linguistically enriched information was available was shown. However, the CART predictors have been only slightly (1–2%) improved by the introduction of features that were expected to have a strong influence on focus identification. The main reasons were: a) the restricted nature of the syntactic structure of the specific domain, which could imply these features by the combination of other linguistic factors (such as POS, syllabic distances and break indices), and b) the limited capabilities of the NLG component to provide the TtS with more speech oriented information. In overall, we have achieved a moderate classification concerning the pitch accents, as the high score is mainly caused by the well classified NONE accents. On the other hand, this illustrates a good accented/unaccented classification. Also, prosodic phrase breaks and endtones are very well classified, as shown by the 10-fold cross validation. The application of the trained models to a linguistically enriched restricted domain of museum exhibits in the Greek language resulted in a highly accurate prediction of realistic prosodic structures. This accuracy amounts to 23% compared to non-enriched cases as shown in Table 7.

6 Acknowledgments

The work described in this paper has been partially supported by the HERACLITUS project of the Operational Programme for Education and Initial Vocational Training (EPEAEK) of the Greek Ministry of Education under the 3rd European Community Support Framework for Greece.

References

1. Taylor, P. and Black, A. W.: Assigning Phrase Breaks from Part-of-Speech Sequences. Computer Speech and Language, 12(2), (1998) 99–117.
2. Petasis, G., Karkaletsis, V., Farmakiotou, D., Samaritakis, G., Androutsopoulos, I. and Spyropoulos, C.: A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker. Proceedings of the 8th Panhellenic Conference on Informatics, (2001) 80–89.

3. Black, A. and Taylor, P.: Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. ICSLP '94, (1994) 715–718.
4. Grosz, B., & Hirschberg, J.: Some intonational characteristics of discourse structure. Proceedings of 2nd of International Conference on Spoken Language Processing, 1, (1992) 429–432.
5. Hirschberg, J.: Pitch accent in context: predicting intonational prominence from text. Artificial Intelligence 63, (1993) 305-340.
6. Prevost, S.: A semantics of contrast and information structure for specifying intonation in spoken language generation. Ph.D. Thesis, University of Pennsylvania, (1995).
7. Theune, M., Klabbers, E., Odijk, J., De Pijper, J.R., and Krahmer, E.: From Data to Speech: A General Approach. Natural Language Engineering, 7(1), (2001) 47–86.
8. McKeown, K., and Pan, S.: Prosody modelling in concept-to-speech generation: methodological issues. Philosophical Transactions of the Royal Society, 358(1769), (2000) 1419–1431.
9. Reiter, E., and Dale, R.: Building Applied Natural Generation Systems. Natural Language Engineering, 3 (1997) 57–87.
10. Hitzeman, J., Black, A., Mellish, C., Oberlander, J., Poesio, M., and Taylor, P.: An annotation scheme for Concept-to-Speech synthesis. Proceedings of the European Workshop on Natural Language Generation, Toulouse France, (1999) 59–66.
11. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York, (1984).
12. Cruttenden. A.: Intonation. Cambridge University Press, Cambridge, UK, (1986).
13. Bolinger, D.: Intonation and its Uses: Melody in grammar and discourse. Edward Arnold, London, (1989).
14. O'Donnel, M., Mellish, C., Oberlander, J., & Knott, A.: ILEX: An architecture for a dynamic hypertext generation system. Natural Language Engineering, 7(3), (2001) 225–250.
15. Xydas G. and Kouroupetroglou G.: Augmented Auditory Representation of e-Texts for Text-to-Speech Systems. Lecture Notes in Artificial Intelligence, 2166, (2001) 134–141.
16. Xydas G. and Kouroupetroglou G: The DEMOSTHeNES Speech Composer. Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, (2001) 167–172.
17. Taylor, P., Black, A., and Caley, R.: Heterogeneous Relation Graphs as a Mechanism for Representing Linguistic Information. Speech Communications 33, (2001) 153–174.
18. Arvaniti, A., and Baltazani, M.: Greek ToBI: A System For The Annotation Of Greek Speech Corpora. Proceedings of Second International Conference on Language Resources and Evaluation, 2, (2000) 555–562.
19. Taylor, P., Caley, R., and Black, A.: The Edinburgh Speech Tools Library. The Centre for Speech Technology Research, University of Edinburgh, 1.0.1 edition, (1998). `http://www.cstr.ed.ac.uk/projects/speechtools.html`.
20. Androutsopoulos, I., Kokkinaki, V., Dimitromanolaki, A., Calder, J., Oberlander, J., and Not, E.: Generating Multilingual Personalized Descriptions of Museum Exhibits – The M-PIRO Project. Proceedings of the 29th Conference on Computer Applications and Quantitative Methods in Archaeology, (2001).