Building Prosodic Structures in a Concept-to-Speech System

Gerasimos Xydas, Dimitris Spiliotopoulos and Georgios Kouroupetroglou

Speech Group
University of Athens
Department of Informatics and Telecommunications
Division of Communication and Signal Processing
Panepistimiopolis, Ilisia, 15784 Athens, Greece
{gxydas, dspiliot, koupe}@di.uoa.gr

Abstract: The prosodic structure of utterances in terms of breaks and tones is a significant problem in speech synthesis. In this work we present the results from models used to predict accurate and realistic prosodic structures within the context of a Concept-to-Speech system for a virtual museum guide. We have used a Natural Language Generator system for providing error-free enriched linguistic information, such as syntax and Part-of-Speech, to a Speech Synthesizer. An XML annotation has been used as a means for this transfer of linguistic data. The annotated data was used to build classification trees for the prediction of prosodic phrase breaks, pitch accents and endtones (phrase accents and boundary tones). The annotation of utterances included segmental information, ToBI marks, syntax, grammar and some domain specific features such as new/given and phrase subject/object information. The linguistic nature of the domain allowed us to carefully select the set of features and the training conditions and also to utilize speech-oriented information from the written language produced by the Natural Language Generator component, such as evidence of stress and intonational focus. A speech corpus of 516 utterances has been used for training and evaluation purposes. To optimize the generated models, we used exhaustive training upon the domain data, achieving a correlation between the observed and the predicted elements of 97.286% for phrase breaks, 99.349% for pitch accents and 99.992% for endtones.

Keywords: prosody prediction, Concept-to-Speech, SOLE

1. Introduction

One of the most important tasks in Text-to-Speech (TtS) synthesis is the prediction of the prosodic structure of the utterance to be spoken. This forms the basis of the rendering of the segmental durations and the fundamental frequency's contour or the selection of the appropriate units in corpus based synthesis. The description of the prosodic structure is usually defined by the position and the type of (a) prosodic phrase breaks, (b) pitch accents, (c) phrase accents and (d) boundary tones. The last two are usually grouped together (they do not co-occur in the tone tier) and referred as endtones. Two approaches are usually followed for the identification of the above elements in an utterance: rule-driven and machine learning. The former fails to capture all the richness of human speech, is generally difficult to write and to adapt to new domains and usually provides poor input to the prosody generation module, while the latter can yield reasonable results as long as the size of the sample data increases with the size of the domain of the application.

Prosody generation is a complex process that involves the analysis of several linguistic phenomena. Dynamic approaches are usually prone to errors. For instance, part-of-speech (POS) identification fails in 5% of the cases for Greek using statistical taggers [Petasis et al., 1999], while syntax and metric trees are hard to construct. A solution that overcomes that is offered by (a) limiting the domain to which the TtS applies to and thus limiting the linguistic phenomena, and (b) using a Concept-to-Speech (CtS) system [Theune et al., 2001]. The advantage of the latter is that the generated texts are annotated with high level linguistic factors in contrast to plain texts [Reiter and Dale, 1997]. Annotation varies in terms of

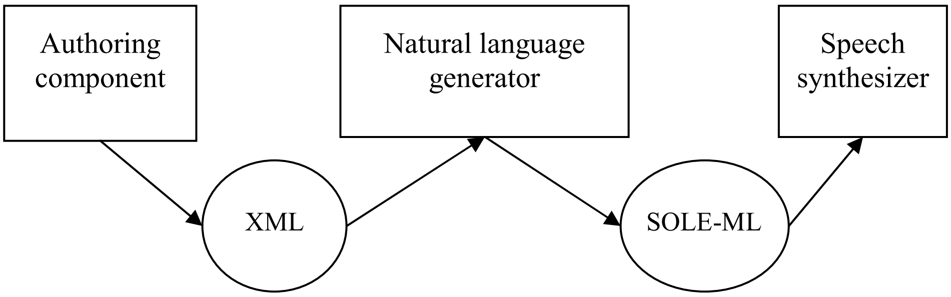

Figure 1 The M-PIRO Concept-to-Speech system

available words, grammatical rules, syntactic rules, concatenation and comparison principles, language specific rules, domain type and size, general notions and concept, and so on. While, Natural Language Generation systems usually deal with written text and fail to represent spoken language, although they are able to produce linguistically enriched output (e.g. syntactic structure, rhetorical relations etc) [Somers et al., 1997], the generation of tones and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [Black & Taylor, 1994].

We are mostly interested in exploring rhetorical relations information effect on prosody for Greek speech synthesis. CtS systems can provide such relations which can be used in prosody modeling [McKeown & Pan, 2000]. Former works show that they can also affect pitch assignment and placement, such as discourse structure [Grosz & Hirschberg, 1992], old (already given) or new information [Hirschberg, 1993], contrast [Prevost, 1995], etc. This work involves a selective use – encoding in a markup language, evaluating and selecting for input to the TtS system - of rhetorical relations for improved prosody, pitch prediction and assignment.

We have tried to optimize a set of features that includes the traditional POS and syntax and is extended by elements of focus. We exploit the constraints of a museum exhibits domain in order to deliver more natural synthetic speech by accentuating focus information. We used the M-PIRO CtS system [Androutsopoulos et al., 2001] in order to set up an enriched pipeline between the NLG (Exprimo) [O'Donnel et al, 2001] and the TtS (DEMOSTHeNES) [Xydas and Kouroupetroglou, 2001a & b] subsystems. This pipeline is based on the SOLE markup notation [Hitzeman et al. 1999] and has been extended as to provide more evidence of stress and intonational focus information in documents. Using this meta-information, we optimized 3 sets of features for training prosodic phrase breaks, pitch accents and endtones CART trees for the above domain.

2. The M-PIRO Concept-to-Speech system

The M-PIRO CtS system [Androutsopoulos et al. 2001] involves 3 distinct subsystems: The authoring component, the natural language generator and the speech synthesizer (Fig. 1)

In our case, the Exprimo natural language generator has been used, which is an implementation of the ILEX generator [O'Donnel et al, 2001]. The information used by the generation subsystem is entered initially and updated through the authoring component [Androutsopoulos et al. 2002]. The entered data represent the whole domain (words, grammar, syntax, user modeling, personalization, concepts, etc).

The SOLE markup, which lies between the NLG and the TtS, provides enumerated word lists and syntactic tree structures. On the syntactic tree, 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.). This structure carries error-free phrasing and POS information (Fig. 2).

```
<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>
```

Figure 2 A SOLE example

The annotated text contains sentences of a fairly limited structure (Subject/Verb/Object - SVO). However enough variation is provided in the domain for the range of phrase types and lexical categories mentioned above to occur in sentences.

The analysis of the above linguistic information strongly leads towards identification of the intonational focus (phonological stress) points in each phrase [Cruttenden, 1986]. Intonational focus points are prosodical instances where the pitch (mostly, but duration and loudness can vary, as well) 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 [Bolinger, 1989]. 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. However those were not supported by the initial SOLE description, thus an extension was needed.

Most of the sentences generated by Exprimo in M-PIRO can be annotated with such detailed meta-information. Pieces of canned text integrated in the presentation are marked as “CANNED-TEXT” without any POS or phrasing information. This missing information is retrieved by the standard NLP modules of the TtS.

Taking into account the capabilities of Exprimo, we extended the SOLE specification to accommodate elements that could directly or indirectly imply emphasis for the specific domain. These elements stand for noun phrases and are:

- Newness or given information: newness [new/old]
- Number of times mentioned before: mentioned-count [integer]
- Whether they are a second argument to the verb: arg2 [true/false]
- Whether there is deixis: genitive-deixis, accusative-deixis [true/false]
- Whether there is a proper noun in the noun phrase: proper-group [true/false]

By examining and combining the above we can compute on the chances of having the intonational focus in a syllable within a particular phrase. By this approach, focus priority is assigned to nouns that are parts of NPs where the following stand:

Strong focus: [newness=new] AND [arg2=true] AND [proper-group=true] AND
[(genitive-deixis) OR (accusative-deixis)]
Normal focus: [newness=old] AND [arg2=true] AND [proper-group=true] AND
[(genitive-deixis) OR (accusative-deixis)]
Weak focus: [newness=old]

3. The corpus

The corpus is constituted of 516 utterances, 5380 words and 13214 syllables. In order to achieve concrete results we grouped together some low-frequency features. A professional speaker was used in order to capture the spoken expressions of a museum guided tour. The speaker was instructed to render the 3 levels of focus presented above.

The text corpus was first annotated by DEMOSTHeNES and was then produced in a properly visualized and readable RTF format for the speakers to read them out loudly following the annotation directions. This annotation was achieved through the XML export component of DEMOSTHeNES that enables the presentation of any information available in the Heterogeneous Relation Graph (HRG) [Taylor et al., 2001] component. In our case we represented the above assigned focus information in a readable form. The produced voice corpora were further automatically segmented and hand annotated using the GR-ToBI marks [Arvaniti and Baltazani, 2000]. 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, accent tones (i.e. ToBI pitch accents) are represented by 5 binary features (Table 1) and endtones (i.e. ToBI phrase accents and boundary tones grouped together as the grammar of GR-ToBI does not allow them to co-occur) by 8 features (Table 2).

Feature		accent 1	accent 2	accent 3	accent 4	accent 5	**Total accented**
Main accent		L*	H*	L*+H	L+H*	H*+L	
diacritics	downstep		!H*	L*+!H	L+!H*	!H*+L	
	weak			wL*+H			
	early			>L*+H			
	late			<L*+H			
	low point	wL*					
# occurences		**332**	**439**	**1175**	**976**	**676**	**3598**
Occurrences %		**9.23**	**12.20**	**32.66**	**27.12**	**18.79**	**100**

Table 1: *Accent groups in the processed corpus.*

Feature	endtone 1	endtone 2	endtone 3	endtone 4	endtone 5	endtone 6	endtone 7	endtone 8
Main tone	L-	H-	L%	H%	L-L%	L-H%	H-L%	H-H%
Downstep diacritics		!H-		!H%		L-!H%	!H-L%	!H-H%
								H-!H%
								!H-!H%
# occurences	**45**	**449**	**0**	**0**	**417**	**4**	**0**	**8**
Occurrences %	**4.88**	**48.7**	**0**	**0**	**45.19**	**0.47**	**0**	**0.88**

Table 2: *Endtone groups in the processed corpus.*

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.

Break index	Occurrences #	Occurrences (%)
0	1727	32.1
1	2541	47.23
2	596	11.08
3	516	9.59
Total	**5380**	100

Table 3: *Occurrences of break indices in the corpus*

4. Predicting the prosodic structure

To predict the GRToBI marks we used the linguistic factors presented in the SOLE documents as features to produce the trained prosodic models, using classification trees [Breiman et al. 1984]. We used the `wagon` [Taylor et al., 1998b] program for this purpose and we built three models:

- Prosodic phrase break model, where break indices were assigned to words.
- Accent model, where pitch accents were assigned to stressed syllables.
- Endtone model, where phrase accents and boundary tones were assigned to syllables at phrase boundaries.

For each case, we tried an exhaustive classification, optimized for the specific domain. The initial features we used were too many and were eliminated after a lot of trials to the following generic ones (syllable relation - Utterance structure of HRG):

- `R:SylStructure.parent.gpos`: the Part-of-Speech of the corresponding word.
- `stress`: an binary indication of lexical stress.
- `syl_in, word_in`: number of syllables/words since last phrase break.
- `syl_out, word_out`: number of syllables/words until next phrase break.
- `ssyl_in, sword_in`: number of stressed syllables/words since last phrase break.
- `ssyl_out, sword_out`: number of stressed syllables/words until next phrase break.
- `R:SylStructure.parent.R:Phrase.parent.punc`: the punctuation of a phrase.

and M-PIRO/SOLE specific ones:

- `R:SylStructure.parent.R:Phrase.parent.newness`: new or given information provided by the text generator.
- `R:SylStructure.parent.R:Phrase.parent.arg2`: arg2 information provided by the text generator.
- `R:SylStructure.parent.R:Phrase.parent.deixis`: an indication of deixis (accusative/genitive/none) information provided by the text generator.

For all the cases we assumed the above features for a context of two items before (p – previous and pp – previous, previous) and two items after (n – next and nn – next, next) (five in total) the current item, in Syllable, Word and Phrase relation, leading to a set of 40 parameters for each vector. For the part-of-speech feature (gpos) we used the bellow tagset:

Vb	VerB
Aj	AdJective
No	Noun
At	ArTicle
Cj	ConJuction
Pn	ProNoun
Pp	PrePosition
Ad	Adverb
Pt	Particle

Table 4: *The POS tagset used for training.*

We have not made any attempt to optimize the tagset and experiment with smaller (e.g. function/content words only) or bigger (e.g. including declensions, gender) ones.

4.1. Prosodic phrase breaks

The identification of prosodic break prediction is the base for the remaining processes and it is an important problem in text-to-speech synthesis. Break prediction is fundamental for F0 contour generation, duration models and pause insertions [Taylor et al, 1998a]. Using all the above mentioned features we achieved a correlation of 97.286% between the observed and the predicted values (96.580% when excluding the M-PIRO specific features). Bellow is the classification matrix:

train test	0	1	2	3	Score	Cor.
0	1715	11	0	1	1715/1727	99.305
1	71	2461	8	1	2461/2541	96.852
2	11	26	555	4	555/596	93.121
3	2	3	8	503	503/516	97.481

Table 5: *Observed and predicted break indices for the prosodic phrase break model.*

4.2. Accent and Endtone models

For the prediction of accents and endtones, we used the same features, plus the `R:SylStructure.parent.bi` (break index) for the accent and the endtone model and the `accent` for the endtone one. Table 6 presents the classification matrix. The overall achieved correlation here is 99.349%. Without the M-PIRO specific features this decreases to 99.152%.

train test	NONE	L+H*	L*+H	H*+L	H*	L*	Score	Cor.
NONE	9612	2	1	1	0	0	9612/9616	99.958
L+H*	6	964	2	1	3	0	964/976	98.770
L*+H	8	13	1151	0	3	0	1151/1175	97.957
H*+L	4	4	0	667	1	0	667/676	98.669
H*	8	4	11	2	414	0	414/439	94.305
L*	6	2	0	3	1	320	320/332	96.386

Table 6: *Observed and predicted tones for the accent model.*

For the endtone model, we expected to have reasonably good results, as the distribution of the endtones (Table 2) showed that for the specific domain almost only two values were observed (H- and L-L% = 94,13%).

train test	NONE	L-L%	L-H%	H-H%	H-	L-	Score	Cor.
NONE	12293	0	0	0	1	0	12293/12294	99.992
L-L%	0	417	0	0	0	0	417/417	100.000
L-H%	0	0	4	0	0	0	4/4	100.000
H-H%	0	0	0	5	0	0	5/5	100.000
H-	0	0	0	0	449	0	449/449	100.000
L-	0	0	0	0	0	45	45/45	100.000

Table 7: *Observed and predicted tones for the endtone model.*

This absolute high score in all endtones was produced because L-L% always occurred after a full stop and can be accurately predicted and the extremely low-frequency H-L% and H-H% were caused in cases of questions. All the H- and L- occurrences were successfully predicted as well. An interesting thing to mention is that, when we used word-level training the result was slightly lower (99.870%), while M-PIRO specific features did not affect the results (99.985%).

All the experiments were done on seen data due to the restricted nature of the specific domain, allowing us to achieve the highest performance. However, doing experiments on unseen data, the aforementioned scores decrease to 97.230% for the breaks, 73.747% for the accents and 94.834% for the endtones. These scores, along with the fact that 48.03% of the text produced by Exprimo was (a) free of any linguistic information and (b) more delicate (CANNED text), show that the models can be used with success in generic domains, not necessarily with annotated texts. Such an evaluation is a task to be done.

4.3. Example

Bellow is the predicted values for the sentence

"Αυτό το έκθεμα είναι ένας στατήρας που δημιουργήθηκε κατά την διάρκεια της ελληνιστικής περιόδου."

using the above models. Break indices are indicated by sub-script numbers, accents by super-script marks and endtones are at the end of each phrase.

[a - fto^{L+H*}]$_{1}$ – [to]$_{0}$ – [e^{H*+L} - kTe – ma]$_{2}$ - [H-]

[i^{L*+H} – ne]$_{1}$ – [e – nas]$_{0}$ – [sta - ti^{H*+L} - ras]$_{2}$ - [H-]

[pu]$_{0}$ – [Di - mi - u - rji^{L*+H} - Ti – ce]$_{1}$ – [ka – ta]$_{0}$ – [ti]$_{0}$ – [Dja^{H*} - rci – a]$_{2}$ – [H-]

[tis]$_{0}$ – [e - li - ni - sti - cis^{L*+H}]$_{1}$ – [pe - ri - o^{H*+L} – Du]$_{3}$ - [L-L%]

Looking at the above example we can see (a) the well-placed accents, (b) their realistic variation and (c) the natural sounding choice of break index 0 in phrase 2 ("ένας στατήρας") and phrase 3 ("κατά τη διάρκεια"), which leads to the correct placement of focus to the nouns "στατήρας" and "διάρκεια".

5. Conclusions

Using a Concept-to-Speech system we managed to provide the speech synthesis component with carefully selected and properly structured enriched linguistic meta-information that improved the prediction of phrases and intonation events. The variation of the features that were used offered to us the luxury to try out several combinations for improvement in prediction. We showed that the introduction of specific features that were expected to have a strong influence in focus identification, failed to do so. The main reason was the restricted nature of the syntactic structure of the texts, which could imply these features by other standard linguistic factors (such as POS, syl_in and syl_out). The restricted texts have not allowed for the exhaustive tries and there is a belief that certain features could be useful in others texts. The application of the trained models to a constrained domain of museum exhibits in the Greek language resulted in highly accurate prediction of the prosodic structures. Moreover, the large amount (48.03%) of untagged text in the training data shows that the produced trained models can be applied to plain text of the same domain as well with success.

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.

7. References

Androutsopoulos, I., Kokkinaki, V., Dimitromanolaki, A., Calder, J., Oberlander, J., and Not, E. (2001) "*Generating Multilingual Personalized Descriptions of Museum Exhibits – The M-PIRO Project*". Proc. 29th Conference on Computer Applications and Quantitative Methods in Archaeology, Gotland, Sweden, 2001.

Androutsopoulos, I., Spiliotopoulos, D., Stamatakis, K., Dimitromanolaki, A., Karkaletsis, V., and Spyropoulos, C., (2002) *"Symbolic Authoring for Multilingual Natural Language Generation"*. Lecture Notes in Artificial Intelligence (LNAI), Vol. 2308, 2002, pp 131-142.

Arvaniti, A., and Baltazani, M. (2000) "*Greek ToBI: A System For The Annotation Of Greek Speech Corpora*". Proceedings of Second International Conference on Language Resources and Evaluation (LREC2000), vol 2: 555-562.

Black, A. and Taylor, P. (1994) *Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input*, ICSLP94, Yokohama, Japan.

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

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J, (1984). *"Classification and Regression Trees"*. Chapman & Hall, New York, 1984.

Cruttenden. A. (1986) *Intonation*. Cambridge University Press, Cambridge, UK.

Grosz, B., & Hirschberg, J., (1992) "*Some intonational characteristics of discourse structure*". In Proceedings of 2nd of International Conference on Spoken Language Processing, 1992, Vol 1, pp. 429-432.

Hirschberg, J., (1993) "*Pitch accent in context: predicting intonational prominence from text*". Artificial Intelligence 63, pp. 305-340.

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

McKeown, K., and Pan, S., (2000) *Prosody modelling in concept-to-speech generation: methodological issues*. Philosophical Transactions of the Royal Society, 358(1769):1419-1431, 2000.

O'Donnel, M., Mellish, C., Oberlander, J., & Knott, A., (2001) "*ILEX: An architecture for a dynamic hypertext generation system*". In Natural Language Engineering, 7(3): 225-250, 2001.

Petasis, G., Karkaletsis, V., Farmakiotou, D., Samaritakis, G., Androutsopoulos, I. and Spyropoulos, C. (2001) "*A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker*". In Proceedings of the 8th Panhellenic Conference on Informatics, 8 - 10 November 2001, Nicosia, Cyprus.

Prevost, S., (1995) "*A semantics of contrast and information structure for specifying intonation in spoken language generation*". Ph.D. Thesis, University of Pennsylvania, 1995.

Reiter, E., and Dale, R., (1997) "*Building Applied Natural Generation Systems*". In Natural Language Engineering, 3:57--87, 1997.

Somers, H., Black, B., Nivre, J., Lager, T., Multari, A., Gilardoni, L., Ellman, J., and Rogers, A. (1997) "*Multilingual generation and summarization of job adverts: the TREE project*". In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 269 - 276.

Taylor, P. and Black, A. W. (1998a) *"Assigning Phrase Breaks from Part-of-Speech Sequences"*. Computer Speech and Language, 12(2), pp. 99-117.

Taylor, P., Caley, R., and Black, A. (1998b) "*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.

Taylor, P., Black, A., and Caley, R. (2001) *"Heterogeneous Relation Graphs as a Mechanism for Representing Linguistic Information"*, Speech Communications 33, pp 153-174.

Theune, M., Klabbers, E., Odijk, J., De Pijper, J.R., and Krahmer, E. (2001) "*From Data to Speech: A General Approach*". Natural Language Engineering, 7(1), pp. 47-86.

Xydas G. and Kouroupetroglou G. (2001a) *"Augmented Auditory Representation of e-Texts for Text-to-Speech Systems"*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 2166, 2001, pp. 134-141.

Xydas G. and Kouroupetroglou G. (2001b) *"The DEMOSTHeNES Speech Composer"*, Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, August 29th - September 1st, 2001, pp 167-172.