Workshop on Text Processing for Modern Greek: from Symbolic to Statistical Approaches, 6th International Conference of Greek Linguistics , Rethymno, Crete, Greece, September 18-21, 2003

Prosody Prediction from Linguistically Enriched Documents Based on a Machine Learning Approach

Gerasimos Xydas, Dimitris Spiliotopoulos and Georgios Kouroupetroglou
University of Athens
Department of Informatics and Telecommunications

{gxydas, dspiliot, koupe}@di.uoa.gr

Abstract: *One of the main aspects in text-to-speech synthesis is the successful prediction of prosodic events. In this work we deal with the prediction of prosodic phrase breaks, accent tones and boundary tones from a linguistically XML-based enriched input (SOLE-ML) produced by a Natural Language Generator (NLG) system. We first extended the original specification of SOLE-ML in order for the NLG to produce a more spoken aware output providing evidence of stress and intonational focus. We then used a machine learning approach (CART) to statistically analyze documents as sequences of Part-of-Speech (POS), already given or new information, object-subject information and other domain features, in order to predict prosodic phrase breaks, accent tones and boundary tones. We applied this approach on a specific domain of Greek descriptions of museum exhibits. An important task of this work was the optimization of the set of features used for training, after which the correlation between the observed and the predicted aforementioned prosodic elements became 97,72%, 96,77% and 100,00% respectively. The large amount (48.03%) of untagged text in the above corpus shows that the produced trained models can be applied to plain text of the same domain as well with success.*

Keywords: prosody, machine learning, CART, SOLE, text-to-speech

1. Introduction

During the last years, while the segmental quality of speech synthesis has been highly improved [Dutoit et al., 1996] [Black & Lenzo, 2000], the production of realistic and natural prosody is still a difficult problem. The improvements of limited-domain synthesis have led to natural sounding domain synthesizers, while free tools produce quite natural results but still for small domains.

During the generation of prosody in Text-to-Speech (TtS) systems, the rendering of segmental durations and the fundamental frequency's contour usually require a set of prosodic events to apply on. Such events are the position and the type of (a) prosodic phrase breaks, (b) accent tones and (c) boundary tones. The prediction of these elements targets to a realistic and natural placement of the appropriate type of each of them within a synthetic speech utterance. Two approaches are usually followed: 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 linguistic phenomena. In the aforementioned cases the achievement of highly accurate prediction is failing due to errors that are produced and propagating by such analysis, closely related to the generation of prosody. For example, 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. Moreover, prosody generation in generic TtS systems seems to be moderate as for example intonation events are often grouped to 4-5 categories in order to reduce the probabilities of getting artificial and unnatural F0 curves: e.g. ToBI annotation offers a variety of marks for pitch accents and usually these are quantized to less (e.g. only 5

are being used in standard English voices in Festival speech synthesis system [www.festvox.org]), thus failing to represent unrestricted prosodic phenomena.

A solution that overcomes both the above mentioned problems is offered by (a) limiting the domain to which the TtS applies to and thus limiting the linguistic phenomena, offering a more concrete set of analysis data to predict prosody and (b) coupling the TtS system with a Natural Language Generation (NLG) system forming a Data-to-Speech or Concept-to-Speech (CtS) system [Theune et al., 2001]. From the speech synthesis point of view, input texts can be categorized as plain (simple unlabelled) or linguistically enriched. The latter is used to denote the written texts that are annotated with high level linguistic factors as they are generated by the generation system while the former is stripped of such information [Reiter and Dale, 1997]. NLG systems can produce many sorts of texts depending on the information used as input. That information varies in terms of available words, grammatical rules, syntactic rules, concatenation and comparison principles, language specific rules, domain type and size, general notions and concept, and so on. However, NLG systems usually deal with written text and fail to represent spoken language, although they are able to produce linguistically enriched outputs (e.g. syntactic structure, rhetorical relations etc) [Somers et al., 1997].

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.

The generation of intonation events and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [Black & Taylor, 1994]. Trained models can produce rich forms of prosodic variations as long as the set of features used to apply on is carefully selected. Here, 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, accent tones and boundary tones CART trees for the above domain.

2. The SOLE-ML pipeline

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

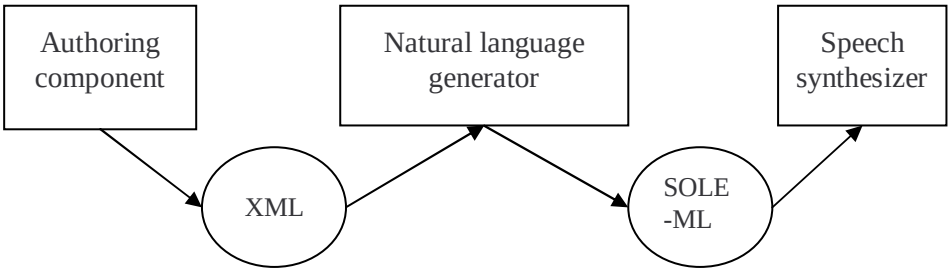

Figure 1 *The M-PIRO CtS system*

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). Those data are exported from the authoring component according to a proprietary XML format that represents the input for the Natural Language Generator (NLG). The NLG produces valid texts accompanied by linguistic information in the XML based SOLE markup language through the SOLE component.

The SOLE component lies between the NLG and the TtS. The SOLE markup language itself 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*

Analyzing the above linguistic information strongly leads towards identification of the intonational focus (phonological stress) points in each phrase can be revealed [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 in 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 as shown later.

3. The corpus

The train corpus is constituted of 516 utterances, 5380 words and 13214 syllables, of which 1509 were used as test data. For the evaluation of the models we used a smaller set of unseen data of 615 syllables and 22 utterances. The test data was carefully selected in order to include a reasonable distribution of the features of interest. In order to achieve concrete results we also tried to group some low-frequency features. Two professional speakers (male and female) were used in order to capture the spoken expressions of a museum guided tour. The results presented here have been extracted from the male database. The speakers were instructed to render 3 levels of focus using the above information. Focus priority has been assigned to nouns that were parts of NPs where the following standed:

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]

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 6 features (Table 2).

Feature		accent 1	accent 2	accent 3	accent 4	accent 5
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*				
Occurrences %		**9.23**	**12.20**	**32.65**	**27.13**	**18.79**

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%
Occurrences %	**4.89**	**48.8**	**0**	**0**	**45.33**	**0.43**	**0**	**0.54**

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 (%)
0	32.1
1	47.33
2	10.97
3	9.6

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

Also, the above mentioned post processing (grouping) eliminated almost every human annotation evaluation differences between the 3 linguists, thus allowing the use of the full corpus. Most importantly, it allowed the use of less annotation marks resulting to more robust results after training.

4. Training the prosodic models

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 and Regression Trees [Breiman et al. 1984]. We used the `wagon` [Taylor et al., 1998] program for this purpose and we built three models:

- Prosodic phrase break model, where break indices were assigned to syllables at word boundaries.
- Accent model, where accent tones were assigned to stressed syllables.
- Endtone model, where ending tones were given to syllables at phrase boundaries.

For each case, we tried an exhaustive classification, suitable 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`: number of syllables since last phrase break.
- `syl_out`: number of syllables until next phrase break.
- `ssyl_in`: number of stressed syllables since last phrase break.
- `ssyl_out`: number of stressed syllables 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 and pp) and two items after (n and nn) (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 break model

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, 1996]. Using all the above mentioned features we achieved a correlation of 97,72% between the observed and the predicted values. Bellow is the output of `wagon`:

train test	0	1	2	3	Score	Cor.
0	104	1	0	0	104/105	99,048
1	7	341	0	0	341/348	97,989
2	1	3	92	0	92/96	95,833
3	1	1	0	64	64/66	96,970

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

After a lot of trials, the set of the selected features includes the `gpos, syl_in, syl_out, newness, deixis, stress` and `punc`.

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. Bellow is the output of `wagon`:

train test	NONE	L+H*	L*+H	H*+L	H*	L*	Score	Cor.
NONE	460	0	0	0	0	0	460/460	100,000
L+H*	0	48	0	0	0	0	48/48	100,000
L*+H	0	1	56	0	0	0	56/57	98,246
H*+L	0	0	0	26	0	0	26/26	100,000
H*	1	1	0	0	14	0	14/16	87,500
L*	1	0	0	1	0	6	6/8	75,000

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

The overall achieved correlation was initially 99,187%. However, NONE accent is always predicted 100% in the right place (i.e. on top of unstressed syllables). The 2 exceptions shown in Table 6 (H* and L*) concerning the NONE value were caused because of faulty break indices (a value of 0 allows only one accent per prosodic word; in that cases we had two accents in prosodic words). By removing the NONE from the possible values we get a more valuable result of 96,77%. The set of the selected features is: `syl_in, syl_out, gpos, bi, newness, arg2` and `deixis.`

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	570	0	0	0	0	0	570/570	100,000
L-L%	0	20	0	0	0	0	20/20	100,000
L-H%	0	0	1	0	0	0	1/1	100,000
H-H%	0	0	0	10	0	0	1/1	100,000
H-	0	0	0	0	23	0	23/23	100,000
L-	0	0	0	0	0	2	2/2	100,000

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

This absolute percentage 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. In this case, `syl_in`, `bi` and `punc` were the most important features.

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 in an extended SOLE-ML format that improved the evidence of stress and intonational focus in phrases. This meta-information, along with the meta-information about the POS and the syntactic structure, was used to optimize CART-based predictors for prosodic phrase breaks, pitch accents and boundary tones. The application of these models to a constrained domain of museum exhibits in the Greek language resulted in highly accurate prediction of these elements.

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 Lenzo, K. (2000) “*Limited domain synthesis*”. In Proceedings of ICSLP, Beijing, China, 2000

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.

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

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Van der Vrecken, O. (1996) "*The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for NonCommercial Purpose*", proc ICSLP'96,vol.3, p1393-1396.

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, pages 269 - 276.

Taylor, P., Caley, R., and Black, A. (1998) "*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, 2001

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):47-86, 2001.

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.