PAPER Special Section on Corpus-Based Speech Technologies

Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora

Gerasimos XYDAS[†], *Student Member*, Dimitris SPILIOTOPOULOS[†], *and* Georgios KOUROUPETROGLOU[†a)], *Nonmembers*

SUMMARY Synthetic speech usually suffers from bad F0 contour surface. The prediction of the underlying pitch targets robustly relies on the quality of the predicted prosodic structures, i.e. the corresponding sequences of tones and breaks. In the present work, we have utilized a linguistically enriched annotated corpus to build data-driven models for predicting prosodic structures with increased accuracy. We have then used a linear regression approach for the F0 modeling. An appropriate XML annotation scheme has been introduced to encode syntax, grammar, new or already given information, phrase subject/object information, as well as rhetorical elements in the corpus, by exploiting a Natural Language Generator (NLG) system. To prove the benefits from the introduction of the enriched input meta-information, we first show that while tone and break CART predictors have high accuracy when standing alone (92.35% for breaks, 87.76% for accents and 99.03% for endtones), their application in the TtS chain degrades the Linear Regression pitch target model. On the other hand, the enriched linguistic meta-information minimizes errors of models leading to a more natural F0 surface. Both objective and subjective evaluation were adopted for the intonation contours by taking into account the propagated errors introduced by each model in the synthesis chain.

key words: *prosody modeling, text-to-speech, linguistic meta-information, synthetic prosody evaluation*

1. Introduction

One of the most important tasks in Text-to-Speech (TtS) synthesis is the prosody generation for a given utterance. Prosody construction is a complex process that involves the analysis of several linguistic and acoustic phenomena. Traditionally ([1], [2]), this involves the following modules:

part-of-speech → syntactic tree → breaks → pitch accents
→ boundary tones → F0 pitch targets

Each of these modules either generates or predicts a set of features to be used by their successors in the TtS chain [1]. The rule-driven modeling approaches are generally difficult to write, to adapt to new domains and new feature sets, fail to capture the richness of human speech in cases of acoustic elements such as tones (pitch accents and boundary tones) and usually provide the F0 target module with poor input. On the other hand, machine learning planning can yield more realistic results provided that the size of the learning data increases along with the size of the selected features and their variability [3].

All the above mentioned modules are usually prone to errors. For instance, part-of-speech (POS) identification scores 95% in most European languages [4], while syntax and metric trees are hard to construct. In contrast, the generation of tones and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [5]. Former works have shown that certain relations can affect pitch assignment and placement, such as discourse structure [6], already given or new information [7] and contrast [8].

However, linguistically enriched information like focus prominence and rhetorical relations is difficult to extract from plain texts. Concept-to-Speech (CtS) systems (i.e. a Natural Language Generation — NLG — system coupled with a TtS system [9]) can provide linguistic information which can be used in prosody modeling [10], [11].

The key idea of the present study is to model prosody structuring and F0 contour generation by utilizing advanced linguistic factors. Corpus-based techniques were adopted to build machine learning models based on the commonly used classification and auto-regression trees (CART) [12] and linear regression [13] algorithms. We chose to focus on a limited domain to ensure control over the enriched generated phenomena. Thus, speech data were collected from a museum guide tour and set up appropriately to include large amounts of emphatic events that can lead to focus prominence determination. These reflect the presence of features like "new or already given information" to the listener or how many times a subject was mentioned before.

In order to study the effects of the introduction of linguistic meta-information in documents, we compare the prosodic structure models (breaks, tones and boundaries) built by linguistically poor information against those built by counting enriched information to find out how well the models classify in either case. Furthermore, we have evaluated the performance of the actual perceptual tonal output of the system, i.e. the F0 curve, by comparing the F0 contour generation models in cases of plain and enriched environments.

Due to the lack of an appropriately sophisticated linguistic analyzer to extract the required features, we have used an NLG system that can generate texts annotated with high-level error-free linguistic factors in contrast to plain texts [14]. As NLG systems deal with written text and fail to represent spoken language, we have extended an XML markup scheme (SOLE-ML [15]) to provide more evidence

Manuscript received July 1, 2004.
Manuscript revised October 14, 2004.

[†] The authors are with the Department of Informatics and Telecommunications, University of Athens, Panepistimiopolis, Ilisia, GR-15784, Athens, Greece.
a) E-mail: koupe@di.uoa.gr

DOI: 10.1093/ietisy/e88–d.3.510

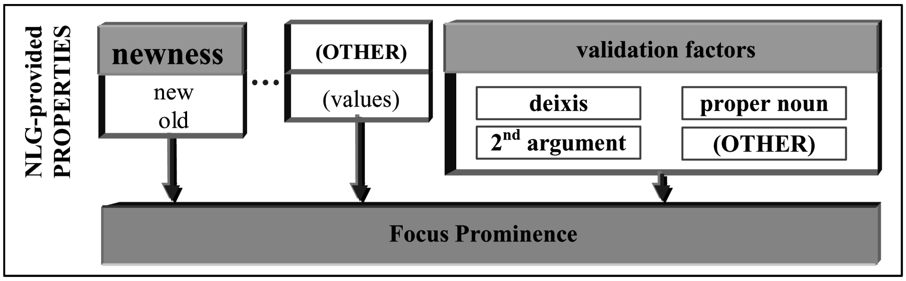


Fig. 1 Noun-phrase focus prominence elements.

of stress and intonational focus information in documents. This enriched output is then used as an input to the TtS synthesizer [16].

The last important task was to accommodate and evaluate error propagation throughout the modules in the aforementioned TtS chain and see how this can be minimized by the use of highly accurate predictors. While other works evaluate the F0 predictors using human-labeled (or manually-corrected) data, e.g. original ToBI marks, we have also looked into the effects caused in F0 contour by the errors of the ToBI predictors since breaks, accents and endtones are included in the feature set of the F0 target models.

2. The Enriched Linguistic Environment

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 [17]. 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 [18]. 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.

Synthetic prosody 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

In our case, an implementation of the ILEX [19] NLG was used. The SOLE markup output of the NLG provides enumerated word lists and syntactic tree structures to the TtS (DEMOSTHeNES) [16]. As shown in Fig. 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 Subject-Verb-Object (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 "shallow" syntactic representation of the natural language generator provides annotation of phrase-level word groupings of a two-level depth (Fig. 3). Noun phrase and prepositional phrase (PP) exist under the top level "sentence" grouping. Also NPs can be children of other NPs or PPs. The rather restricted form of generated text does not allow for too many phrases within the sentence, thus only one level of children phrases can exist under any top-level phrase on the syntactic tree. This ensures that the syntax information used to generate prosody is not too broad. The resulting corpus is well-balanced, that is with adequate yet not too complicated syntactic description making it an ideal basis to achieve successful results in building prosody.

The lexical item description includes the following categories: determiner (DT), common noun (N), proper noun (PN), verb (V), preposition (IN), adverb (ADV), adjective (ADJ), personal pronoun (PRP), indicative pronoun (IP), conjunction (CC), and cardinal (CD). Again, the domain de-

```
<utterance>
<relation name="Word" structure-type="list">
<wordlist>
<w id="w1">αυτό</w>
<w id="w2">το</w>
<w id="w3">έκθεμα</w>
<w id="w4">είναι</w>
<w id="w5">ένα</w>
<w id="w6" punct=",">ειδώλιο</w>
<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>
<relation name="Grouping" structure-type="list"/>
<relation name="Syntax" structure-type="tree">
<elem phrase-type="S">
<elem phrase-type="NP" newness="new">
<elem lex-cat="IP" href="words.xml#id(w1)"/>
<elem lex-cat="DT" href="words.xml#id(w2)"/>
<elem lex-cat="N" href="words.xml#id(w3)"/>
</elem>
<elem lex-cat="V" href="words.xml#id(w4)"/>
<elem phrase-type="NP" newness="new"
arg2="true">
<elem lex-cat="DT" href="words.xml#id(w5)"/>
<elem lex-cat="N" href="words.xml#id(w6)"/>
</elem>
<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)"/>
<elem lex-cat="DT" href="words.xml#id(w10)"/>
<elem lex-cat="N" href="words.xml#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="ADJ" href="words.xml#id(w13)"/>
<elem lex-cat="N" href="words.xml#id(w14)"/>
</elem>
</elem>
</elem>
</relation>
</utterance>
```

Fig. 2 A SOLE-ML example.

scription ensures that items of all categories exist in abundance.

The particular generator can produce such detailed meta-information. Since the SOLE-ML 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 Speech Corpus Setup

The corpus data were taken from the description of museum exhibits in the Greek language. It consists of 482 utterances (5484 words and 13467 syllables). Since the NLG component was not able to provide the complete corpus with annotations (40.87% of the sentences were delivered as plain text – "canned"), in order to facilitate our experiments, we formulated two subsets from the corpus data: (a) the *ENRICHED* set (285 utts., 2533 wrds and 6284 syls.) and (b) the *CANNED* set (197 utts., 2951 wrds and 7183 syls.). The utterances in the *CANNED* subset are delivered in a plain form. In the *ENRICHED* subset case, they are accommodated with the enriched meta-information. A single exhibit description could contain both *CANNED* and *ENRICHED* ones. At a first sight, we can see how different the *ENRICHED* and the *CANNED* sets are: 8.9 wrds/utts in *ENRICHED* vs. 15 wrds/utts. This is justified by the constraints introduced by the NLG component.

The utterances were built based on the Heterogeneous Relation Graph (HRG) [21] model and so all feature names used hereafter conform to this. Texts were exported in a properly visualized and readable RTF format (Fig. 4). A professional speaker captured the spoken expressions of a museum guided tour, and, by following the annotation direc-

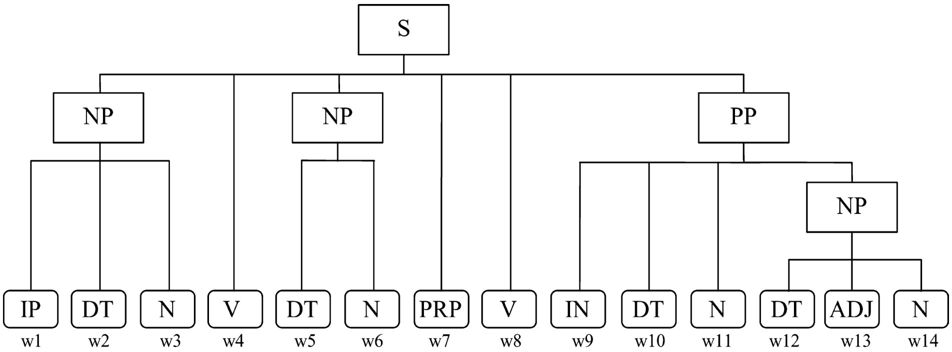
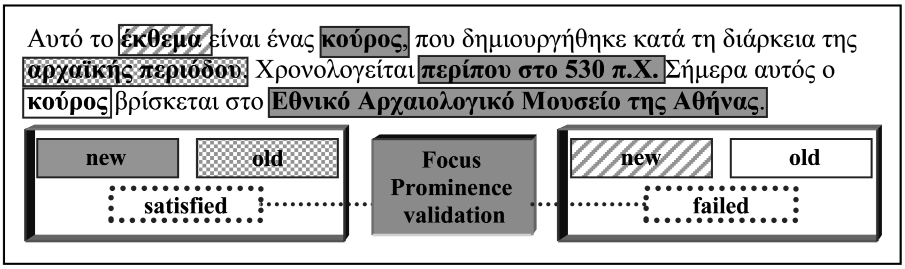


Fig. 3 Generated syntactic representation example.



Αυτό το **έκθεμα** είναι ένας **κούρος**, που δημιουργήθηκε κατά τη διάρκεια της **αρχαϊκής περιόδου**. Χρονολογείται **περίπου στο 530 π.Χ.** Σήμερα αυτός ο **κούρος** βρίσκεται στο **Εθνικό Αρχαιολογικό Μουσείο της Αθήνας**.

Fig. 4 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 1 Occurrences of break indices.

	CANNED		*ENRICHED*	
Break	#	%	#	%
0	1009	34.19	749	29.57
1	1359	46.05	1226	48.40
2	358	12.13	249	9.83
3	225	7.63	309	12.20
Total	2951	100.00	2533	100.00

Table 2 Accent groups (pitch accents).

	CANNED		*ENRICHED*	
Accent	#	%	#	%
L+H*	578	27.72	453	25.93
L*+H	604	31.05	565	32.34
H*+L	285	14.65	404	23.13
H*	233	11.98	218	12.48
L*	245	12.60	107	6.12
Total	1945	100.00	1747	100.00

Table 3 Endtone groups (phrase accents, boundary tones).

	CANNED		*ENRICHED*	
Endtone	#	%	#	%
L-	39	7.44	16	3.11
H-	288	54.96	213	41.44
L-L%	185	35.30	282	54.86
H-H%	6	1.15	2	0.39
L-H%	6	1.15	1	0.20
Total	524	100.00	514	100.00

Table 4 F0 mean and standard deviation of the original speech, in the cases of *CANNED* spoken utterances and *ENRICHED* ones.

	CANNED		*ENRICHED*	
	Mean (Hz)	σ	Mean (Hz)	σ
start	145.0	28.36	151.8	30.70
mid-v	148.5	30.10	155.2	33.47
end	145.0	29.51	151.0	32.76

tions, rendered the different levels of focus according to the properties attached to lexical items provided by the NLG. The produced speech corpus was further automatically segmented and hand annotated using the GR-ToBI marks [22] providing description of the tonal events.

3.1 Original Prosodic Structure Characteristics

As the frequency of some ToBI marks is low in the corpus, we grouped them, while they can be useful when more data is available. Break indices mark boundaries (0 to 3) that are represented by a subjective notion of disjunction between words (Table 1). Additional tonal events usually marked on the Break Index tier using special ToBI diacritics — Sandhi (s), mismatch (m), pause (p), and uncertainty (?) — were not accounted for since their occurrence was sparse (< 0.1%) and thus usefulness was negligible in this work.

Pitch accents are represented by 5 binary features (Table 2) and endtones (ToBI phrase accents and boundary tones grouped together since GR-ToBI does not allow them to co-occur) by 4 features (Table 3).

3.2 Original F0 Contour Characteristics

We have chosen to model F0 targets by following a commonly used strategy [13] in order to produce results comparable to other works: for each syllable we model the target at the start point of the syllable (start), at the middle of the vowel (mid-v) and at the end of the syllable (end). Table 4 shows the mean values and the standard deviations of the corresponding F0 targets produced by the professional reader in the parts of the *CANNED* and the *ENRICHED* utterances. The high standard deviation shown in Table 4 (> 30 Hz in the *ENRICHED* case) reflects the variability of the pitch targets in the original speech due to the presence of rich emphatic events. The pitch accent lies somewhere inside the vowel. However, the 3-point approach followed here, as well as in other F0 modeling studies, models the center point of the vowel in a syllable. In this work, we have not looked into the effects caused by accent alignment accuracy.

4. Building the Models

The features selected for the training were most of the commonest in the literature [13], [23], applied in a 5-instance window in cases of categorical features and in a 3-instance window in cases of continuous features. The following will be referred hereafter as the "common" feature set:

- stress: indicates whether a syllable carries a lexical stress or not. Values: 0, 1.
- syl_in, syl_out: the number of syllables since and until major breaks. Values: integer.
- word_in, word_out: the number of words since and until major breaks. Values: integer.
- ssyl_in, ssyl_out: the number of stressed syllables since and until major breaks. Values: integer.
- last_ssyl: indicates whether a syllable is the last stressed in a phrase. Values: 0, 1.
- last_asyl: indicates whether a syllable is the last accented in a phrase. Values: 0, 1.
- syl_position: the syllable position in a word. Values: initial, mid, final and single.
- onset_size: the number of the phonemes before the vowel in a syllable [24]. Values: integer.
- coda_size: the number of the phonemes after the vowel in a syllable [24]. Values: integer.
- stress_structure: indicates in which syllable of the word the lexical stress is. The values for the Greek language are: final, penultimate, antepenultimate and none.
- gpos: the part-of-speech of the word. Values: verb, noun, proper noun, indicative pronoun, preposition, determiner, personal pronoun, adverb, adjective, conjunction, pronoun, cardinal.
- punctuation: based on the punctuation marks, this feature indicates minor punctuation breaks (commas), major ones (full stops, exclamation marks, question marks) or none.

Additionally (for the *ENRICHED* subset), we have introduced the following "enriched" feature set:

- phrase_boundaries: explicitly marked start and end syllables of phrases and sentences. Values: pstart, sstart, pend, send and none.
- phrase_type: the type of the corresponding phrase. Values: sentence, noun phrase, and prepositional phrase.
- syntax_tree_depth_level: phrases and/or lexical items contained by other phrases are explicitly marked so (see Fig. 3). Values: integer.
- focus: The intonational focus feature that is computed from the "newness", "arg2", "deixis" and "proper noun" tags (see Fig. 1). Values: strong, normal, weak and none.

Tables 1, 2 and 3 in Sect. 3 presented the occurrences of breaks, pitch accents and endtones. Breaks always occur at the end of a word and all words have a break. On the other hand, it is not clear where pitch accents occur in words and also not all words get an accent. Thus, since the learning set of the accent models consists of syllable instances, the model should also take into account the *UNACCENTED* class and predict whether a syllable deserves an accent and, if so, which one. Table 5 presents the occurrences of accented and unaccented syllables in the corpus.

Table 5 Accent and unaccented syllables.

	CANNED		*ENRICHED*	
	#	%	#	%
accented	1945	27.08	1747	27.80
unaccented	5238	72.92	4537	72.20
Total	7183	100.00	6284	100.00

Table 6 Tonal-boundary and non-tonal-boundary words.

	CANNED		*ENRICHED*	
	#	%	#	%
endtones	524	17.76	514	20.29
none	2427	82.24	2019	79.71
Total	2951	100.00	2533	100.00

Table 7 The several configurations used to evaluate the performance of the models.

configuration	utterances	feature set
C1	CANNED	common
C2	ENRICHED	common
C3	ENRICHED	common + enriched

Similarly, endtones always occur at the end of a word, defining a tonal boundary. However, not all words have endtones. Thus, the class *NONE* should be also included in the possible values of the endtone prediction. Table 6 presents the occurrences of tonal-boundary and non-tonal-boundary words in both subsets.

In order to inspect on performance enhancements caused by the introduction of the enriched data, we created three (3) different configurations (Table 7). C1 consists of utterances from the *CANNED* subset. The prosodic structure and F0 models in this case were built by utilizing the *common* feature set (Sect. 4). Configuration C2 uses the same feature set, but the utterances come from the *ENRICHED* subset. This way we inspect how the restricted grammar introduced in the *ENRICHED* case affects the models, though both subsets originate from the same domain. Finally, the last configuration exploits the additional *enriched* features upon data from the *ENRICHED* subset. This is to evaluate the effect of the introduction of the enriched linguistic meta-information.

5. Evaluation

The evaluation is divided in two parts: the first one deals with the performance of the prosodic structure predictors (ToBI marks), while the second part copes with the generated F0 contours.

5.1 Prediction of ToBI Marks

For the prediction of the ToBI marks we used the wagon CART building program [25] to build classification trees. Wagon uses a greedy algorithm that incrementally finds the best single feature to improve the prediction. For each of the above configurations (Table 7) we built 3 models: break, accent and endtone. Our validation approach is based on the 10-fold cross validation method.

5.1.1 *N*-Fold Cross Validation

Cross validation is a computationally demanding method for validating a procedure for model building, which avoids the requirement for a new or independent validation dataset. In *N*-fold cross validation, the learning dataset is randomly split into *N* sections, stratified by the outcome variable of interest. This ensures that a similar distribution of outcomes is present in each of the *N* subsets of data. One of these subsets of data is reserved for use as an independent test dataset, while the other $N - 1$ subsets are combined for use as learning datasets in the model-building procedure. The entire model-building procedure is repeated *N* times, with a different subset of the data reserved for use as the test dataset each time. Cross validation is based on the fact that the average performance of these *N* models is an excellent estimate of the performance of the original model (produced using the entire learning dataset) on a future independent set of measurements.

5.1.2 Evaluation Metrics

The performance was estimated by using the precision and recall metrics per break, pitch accent and endtone class, as they have been explained in Sect. 3.

Per class precision (P_{class}) is defined as the number of correctly identified instances of a class (tp), divided by the number of correctly identified instances, plus the number of wrongly selected cases (fp) for that class:

$$P_{class} = \frac{tp}{tp + fp} \tag{1}$$

Per class recall (R_{class}) is estimated as the number of correctly identified instances of a class (tp), divided by the number of correctly identified instances plus the number of cases the system failed to classify for that class (fn):

$$R_{class} = \frac{tp}{tp + fn} \tag{2}$$

5.1.3 Results

Tables 8, 9 and 10 present the results for the break, the accent and the endtone model respectively for the three configurations C1, C2 and C3. In general, there is an improvement on the performance of the models in C2 case (compared to C1). This is mainly caused by (a) the restricted grammar used in the *ENRICHED* utterances and (b) the shorter average length of the *ENRICHED* utterances compared to the *CANNED* ones as explained in Sect. 3.

Table 8 Results from the 10-fold cross validation of the prosodic phrase break models. (Conf. = configuration, Corr. = correctly classified instances, CXp = CX precision, CXr = CX recall).

Conf.	Corr. (%)	Break			
		0	1	2	3
C1p	83.27	0.89	0.81	0.76	0.98
C1r		0.76	0.97	0.38	0.83
C2p	87.65	0.81	0.88	0.81	0.98
C2r		0.85	0.94	0.53	0.97
C3p	92.35	0.86	0.95	0.85	0.97
C3r		0.93	0.94	0.79	0.97

Table 9 Results from the 10-fold cross validation of the accent models. (Conf. = configuration, Corr. = correctly classified instances, CXp = CX precision, CXr = CX recall, UN = unaccented).

Conf.	Corr. (%)	Accent					
		UN	LH*	L*H	H*L	H*	L*
C1p	71.67	0.91	0.39	0.32	0.32	0.11	0.25
C1r		0.85	0.44	0.56	0.28	0.08	0.21
C2p	81.07	0.94	0.50	0.39	0.63	0.41	0.33
C2r		0.95	0.47	0.70	0.31	0.08	0.13
C3p	87.76	0.95	0.70	0.64	0.73	0.50	0.40
C3r		0.98	0.63	0.75	0.73	0.27	0.16

Table 10 Results from the 10-fold cross validation of the endtone models. (Conf. = configuration, Corr. = correctly classified instances, CXp = CX precision, CXr = CX recall, N = NONE).

Conf.	Corr. (%)	Endtone					
		N	LL%	LH%	HH%	H-	L-
C1p	96.59	0.98	0.88	0	0	0.65	0
C1r		0.99	0.90	0	0	0.61	0
C2p	98.69	0.99	0.95	0	0	0.88	0
C2r		0.99	0.95	0	0	0.82	0
C3p	99.03	1	0.92	0	0	0.92	0.82
C3r		0.99	0.97	0	0	0.96	0.93

Though the accuracy of accent prediction was increased in configuration C2, the performance of the accented classes (i.e. all classes apart from *UNACCENTED*) was slightly improved. This was due to the fact that the CART models produce more accurate results in cases where enough data are provided. The above model was trained on syllable instances and the unaccented syllables in the *ENRICHED* utterances constitute the 72.2% of the total syllables (Table 5). Consequently, there is an improvement in the CART tree on predicting the *UNACCENTED* class and that greatly raises the total accuracy of the model. This however does not affect the accuracy of the accented classes, as shown by their recall and precision metrics. Concerning configuration C3, we did not expect high scores from the CART in the cases of L* (6.12% — Table 2) and H* (12.48%) as there are less instances of them related to the other classes. However, the introduction of the enriched features provides better prediction of accents.

Concerning the endtone prediction, the CART framework did not achieve good results in the cases of low-frequency occurrences, as expected. In configuration C1, L-H% constitutes the 0.2%, H-H% also the 0.2% and L- the 1.3% (see Tables 3 and 6). In the remaining two configurations, distributions are even lower: L-H% is 0.04%, H-H% is 0.08% and L- is 0.6%. However, the introduction of the enriched feature set provided a good input to the model in the L- case. Other machine learning approaches (e.g. Bayesian Networks) show improved classification for low-frequency ToBI classes but a worse one for high-frequency cases [26]. However, such optimization was out of the scope of this study.

5.2 F0 Model

To build the F0 model we chose the commonly adopted Linear Regression [13] method (F0-LR). The training was carried out by the ols (Linear Regression by ordinary least squares) [25] program. Following the F0 framework described in Sect. 3.2, three models are built to predict the F0 targets in the start, mid (vowel) and end point of a syllable. In all cases, the validation of the models was performed by holding out a balanced 10% of the learning data set that formed the test set.

5.2.1 Objective Evaluation

The evaluation process was focused on the accuracy of the pitch target prediction and the F0 contour on top of the synthetic utterances. For the objective evaluation of the models' results we computed the root mean square error (RMSE) and the correlation coefficient (*r*), which have been commonly used in other works. In order to inspect on performance enhancements caused by the introduction of the enriched data, we created two groups of experimental setups in order to look into the effects caused in F0 by (a) different kinds of input and (b) models' error propagation in the TtS chain. All the configurations of Table 7 are contributing in these two groups.

In the first group, we evaluate the F0 models against the original supplied ToBI values (i.e. from the hand-labeled annotations). These cases do not encapsulate any TtS-related effects, so no ToBI prediction is taking place. In the other group, we use the predicted ToBI marks from the TtS chain to evaluate the actual synthetic F0 contour.

Table 11 actually presents the optimum target RMSE and correlation. Looking at the columns of the configurations C1 and C2, it is clear that we achieve slightly better performance in cases of syntactically restricted input text, as in the case of C2. Also, the shorter average length in the *ENRICHED* utterances seems to provide better classification in the models. By introducing the enriched features (C3) along with input data identical to C2, we get an actual improvement of ~9.5% in the correlation of the predicted F0 curves against the original ones.

Table 11 Performance of the F0-LR models in the C1, C2 and C3 configurations using the **original** ToBI marks. (s = start, m = mid_v and e = end).

	C1		**C2**		**C3**	
	RMSE	*r*	RMSE	*r*	RMSE	*r*
S	20.6	0.71	17.3	0.75	16.3	0.82
M	21.2	0.72	18.3	0.74	18.6	0.84
E	20.7	0.71	18.1	0.74	15.9	0.82

Table 12 Performance of the F0-LR models in the C1, C2 and C3 configurations using the **predicted** ToBI marks. (s = start, m = mid_v and e = end).

	C1		**C2**		**C3**	
	RMSE	*r*	RMSE	*r*	RMSE	*r*
s	23.2	0.60	22.1	0.65	20.1	0.78
m	24.8	0.58	24.2	0.64	21.3	0.77
e	25.4	0.58	23.2	0.65	20.8	0.79

Table 12 tabulates the performance of the F0 models through the TtS chain. In these setups the ToBI marks are predicted using the CART models presented before. The high values in RMSE are explained by the also high standard deviation of the original F0. Interesting points can be deduced from this table. First of all, the accuracy of the ToBI accent models presented in Table 9 is not depicted in the correlation of F0 in the cases of C1 and C2, where we have a mean decrease of 17.6% and 13.0% respectively compared to Table 11, while in the C3 case the mean decrease is just 5.7%. This confirms the fact that the low performance of the accented classes of the CART based ToBI predictors is hidden by their apparently high accuracy. Furthermore, the introduction of the enriched feature set has increased the correlation in the F0 targets by 19.6%.

Figure 5 illustrates an example utterance, both original and synthetic (using predicted ToBI values): in the enriched input (configuration C3) the curve goes up and down almost synchronized with the original curve. In the plain input case (configuration C2), some events have been missed due to errors in ToBI pitch accent and break indices in previous modules that have propagated to the F0 model.

5.2.2 Subjective Evaluation

Further to the objective evaluation of the models, the nature of the domain (enriched and canned text) and the involvement of properties such as "newness" that are employed in a description that spans more than a single sentence require an additional subjective evaluation that can add a more qualitative aspect to the results presented above pictures the perceptual effects of the linguistically enriched information to the listeners.

A group of 10 trained and untrained listeners were asked to take part in an intonation evaluation assessment.

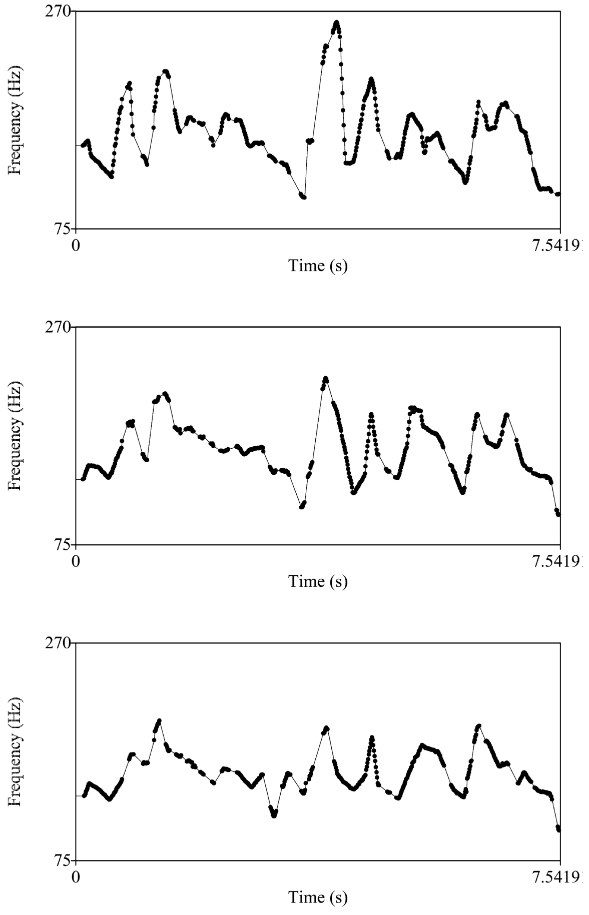


Fig. 5 The F0 contour of the original (top), enriched input — C3 (middle) and plain input — C2 (bottom).

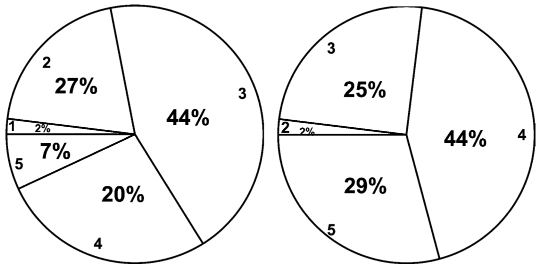


Fig. 6 Scoring distribution for plain input text (C2 — left) and enriched input text (C3 — right).

For the total of 8 sentences, listeners were asked to listen to 2 pairs of synthesized speech for each of the sentences. The first pair consisted of a carrier of the original pitch curve against the speech output from the enriched text input (Table 12, configuration C3). The second pair consisted of the original pitch as before and the speech output from the plain text input (Table 12, configuration C2). The listeners were asked to evaluate the similarity in intonational focus, tone and break (prosody prediction successfulness) for the two models against the original, in a scale from 5 (identical) to 1 (totally different). The results (Fig. 6) show that the enriched text sentences had a 44% scoring of 4 (29% of 5, 25% of 3) while the plain text sentences had a 44% scoring of 3 (27% of 2, 20% of 4). A general listener opinion was that the breaks between the words in plain text model suffered the most (which reflects the low score of the break index model in the *CANNED* case — Table 8), however failing to predict successful breaks leads inevitably to misplacement and/or wrong tone assignment.

6. Conclusions

Our aim was to study the effects of several linguistic features in prosody generation. Using a CtS system, we utilized a linguistically annotated corpus. The provided properly structured linguistic meta-information has been used to improve the prediction of tones, breaks and pitch targets. An extended SOLE-ML specification has been formulated to accommodate the required factors that can imply focus prominence. The improvement in the delivery of prosody in cases where linguistically enriched information was available was shown by the CART prosodic structure models. We then compared the performance of the linear regression F0 model in cases of enriched XML input against plain text input using the original ToBI marks, as well as the predicted ones, like in real applications, to evaluate the performance of the whole prosody generation component, accommodating error-propagation from module to module. We concluded that the generated F0 curve correlates 19.5% better upon the introduction of the enriched input.

Acknowledgments

The work of this paper has been partially supported by the GR-PROSODY project of the KAPODISTRIAS Program, Special Account for Research Grants, National and Kapodistrian University of Athens and by the HERACLITUS project of the Operational Programme for Education and Initial Vocational Training (EPEAEK), Greek Ministry of Education, under the 3rd European Community Support Framework for Greece.

We would like to thank the reviewers for their constructive and useful comments.

References

[1] P. Taylor, A. Black, and R. Caley, “The architecture of the festival speech synthesis system,” Proc. 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, pp.147–151, 1998.

[2] T. Dutoit, An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Dordrecht, 1997.

[3] P. Taylor and A.W. Black, “Assigning phrase breaks from part-of-speech sequences,” Computer Speech Lang., vol.12, no.2, pp.99–117, 1998.

[4] G. Petasis, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, and I. Androutsopoulos, “Using machine learning techniques for part-of-speech tagging in the Greek language,” Advances in Informatics: Proc. Panhellenic Conf. on Informatics, ed. D.I. Fotiadis and S.D. Nikolopoulos, pp.273–281, World Scientific, 2000.

[5] A. Black and P. Taylor, “Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input,” Proc. 3rd International Conf. on Spoken Language Processing,

pp.715–718, Yokohama, Japan, 1994.

[6] B. Grosz and J. Hirschberg, "Some intonational characteristics of discourse structure," Proc. 2nd International Conf. on Spoken Language Processing, vol.1, pp.429–432, Banff, Canada, 1992.

[7] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," Artif. Intell., vol.63, pp.305–340, 1993.

[8] S. Prevost, A semantics of contrast and information structure for specifying intonation in spoken language generation, Ph.D. Thesis, University of Pennsylvania, 1995.

[9] M. Theune, E. Klabbers, J. Odijk, J.R. De Pijper, and E. Krahmer, "From data to speech: A general approach," Natural Language Engineering, vol.7, no.1, pp.47–86, 2001.

[10] K. McKeown and S. Pan, "Prosody modelling in concept-to-speech generation: Methodological issues," Philosophical Transactions of the Royal Society, vol.358, no.1769, pp.1419–1431, 2000.

[11] G. Xydas and G. Kouroupetroglou, "Augmented auditory representation of e-texts for text-to-speech systems," Lecture Notes in Artificial Intelligence, vol.2166, pp.134–141, 2001.

[12] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.

[13] A. Black and A. Hunt, "Generating F0 contours from the ToBI labels using linear regression," Proc. 4th International Conf. on Spoken Language Processing, vol.3, pp.1385–1388, Philadelphia, USA, 1996.

[14] E. Reiter and R. Dale, "Building applied natural generation systems," Natural Language Engineering, vol.3, pp.57–87, 1997.

[15] J. Hitzeman, A. Black, C. Mellish, J. Oberlander, M. Poesio, and P. Taylor, "An annotation scheme for concept-to-speech synthesis," Proc. 7th European Workshop on Natural Language Generation, pp.59–66, Toulouse France, 1999.

[16] G. Xydas and G. Kouroupetroglou, "The DEMOSTHeNES speech composer," Proc. 4th ISCA Workshop on Speech Synthesis, pp.167–172, Perthshire, Scotland, 2001.

[17] A. Cruttenden, Intonation, Cambridge University Press, Cambridge, UK, 1986.

[18] D. Bolinger, Intonation and its uses: Melody in grammar and discourse, Edward Arnold, London, 1989.

[19] M. O'Donnel, C. Mellish, J. Oberlander, and A. Knott, "ILEX: An architecture for a dynamic hypertext generation system," Natural Language Engineering, vol.7, no.3, pp.225–250, 2001.

[20] I. Androutsopoulos, V. Kokkinaki, A. Dimitromanolaki, J. Calder, J. Oberlander, and E. Not, "Generating multilingual personalized descriptions of museum exhibits – The M-PIRO project," Proc. 29th Conf. on Computer Applications and Quantitative Methods in Archaeology, Gotland, Sweden, 2001.

[21] P. Taylor, A. Black, and R. Caley, "Heterogeneous relation graphs as a mechanism for representing linguistic information," Speech Commun., vol.33, pp.153–174, 2001.

[22] A. Arvaniti and M. Baltazani, "Greek ToBI: A system for the annotation df Greek speech corpora," Proc. International Conf. on Language Resources and Evaluation, vol.2, pp.555–562, Athens, Greece, 2000.

[23] X. Sun, "Predicting underlying pitch targets for intonation modeling," Proc. 4th ISCA Workshop on Speech Synthesis, pp.143–148, Perthshire, Scotland, 2000.

[24] J.P.H. van Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours," Proc. 3rd International Conf. on Spoken Language Processing, vol.2, pp.719–722, Yokohama, Japan, 1994.

[25] P. Taylor, R. Caley, and A. Black, "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.

[26] P. Zervas, N. Fakotakis, and G. Kokkinakis, "Pitch accent prediction from ToBI annotated corpora based on Bayesian learning," Lecture Notes in Artificial Intelligence, vol.3206, pp.545–552, 2004.

Gerasimos Xydas received the B.Sc. degree in Informatics from the Department of Informatics and Telecommunications, University of Athens in 1998 and the M.Sc. degree in Distributed Systems from the University of KENT at Canterbury in 1999. During 1999, he stayed in British Telecom's Adastral Park, in the Middleware Project. He has received the "Ericsson Award of Excellence in Telecommunications" in 1999 and since then he is researching in the field of speech synthesis, providing a set of public domain Greek speech resources. He is now in the Speech Communication Group of the University of Athens, studying towards a PhD degree. Contact: gxydas@di.uoa.gr

Dimitris Spiliotopoulos received the B.Sc. and MPhil degrees in Computation from the University of Manchester Institute of Science and Technology in 1995 and 1999, respectively, and the M.A. degree in Linguistics from the University of Manchester in 1998. During 2000–2002, he worked for the Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Athens, Greece, as a research fellow. Now studying towards a PhD degree in the Department of Informatics and Telecommunications, University of Athens, Athens, Greece. Contact: dspiliot@di.uoa.gr

Georgios Kouroupetroglou received the B.Sc. degree in Physics and the Ph.D. degree in Communications and Signal Processing from the University of Athens in 1979 and 1982, respectively. He is now an Assistant Professor, Division of Communication and Signal Processing, Department of Informatics and Telecommunications, University of Athens and head of the Speech Communication Group. His research interests include Speech Synthesis, Spoken Dialogue Systems, Computer-based Augmentative and Alternative Communication Aids and Information Systems for Disabled People. Contact: koupe@di.uoa.gr