Machine Learning Models and Selection Methods for the Auditory Representation of Documents via Synthetic Speech with Enriched Prosody

Gerasimos Xydas*

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

Abstract. This thesis addresses the problem of the quality of the synthetic speech and the auditory representation of documents. We introduce the language platform DEMOSTHeNES, which we exploit to design a complete Text-to-Speech (TtS) system for the Greek language featuring novel properties, like the non-standard word normalization sub-system for inflecting languages, the support of polyglot texts and different pronunciation idioms and more. In addition, starting from the text processing, we deal with the lack of text meta-data auditory rendering provision from TtS systems. The proposed script-based Document-to-Audio architecture manages to aurally represent meta-data along with the text by using synthetic speech with controlled prosodic properties, voice switching and audio insertions. Moving to the problem of synthetic prosody, we present a modeling method that utilizes enriched linguistic features during the prediction of the intontational structure and the F_0 targets in CART and linear regression models. Our approach clearly shows the improvements in F_0 surface, as the enriched features increase the correlation of the original and the synthetic curve by 19.5%. Furthermore, we introduce an F_0 modeling approach based on selecting abstracted tone groups. Groups are formed according to the tonal distinction of significant and functional syllables. This distinction is experimentally proved as valid by both objective and perceptual measurements. The proposed model is judged to be more natural in 75% of the cases, when compared against a typical linear regression one. Finally, a series of psycho-acoustic evaluative experiments shows that the model achieves greater than 70.8% accuracy in rendering three levels of emphasis during speech synthesis.

1. Introduction

Speech synthesis has become a major interface in human-machine interaction. During the last years, rule-based robotic-sounding speech has given place to corpus-based natural-sounding speech in our everyday life interactions with information systems, from embedded systems and mobile devices to large-scale telecommunication services. However, communication with computer-generated speech still lacks the

* Dissertation Advisor: Georgios Kouroupetroglou, Assis. Professor

transfer of content or discourse related emotions and this result to the neutral expressions or monotonous speech that most Text-to-Speech (TtS) systems produce.

This dissertation examines the problem of the quality of synthetic speech and the auditory representation of documents around two main issues:

1. the design of an open and flexible platform to convert documents to synthetic speech and audio as a base for the deeper understanding of the synthesis related procedures and the problems in both experimental and production environments
2. the study and the invention of new algorithms for dealing with the above problems in each step of the synthesis chain.

We first introduce DEMOSTHeNES language platform and then a deeper study in synthetic prosody follows.

2. The DEMOSTHeNES Language Platform

Most TtS systems have as a common ground the distinction and the autonomy of the participating linguistic procedures. They mainly differ in two ways: (a) in the knowledge representation approach and (b) in the fundamental language components and tools they offer for linguistic and phonologic development. FESTIVAL [1] and its predecessor CHATR [2] are the most famous open source TtS systems. FESTIVAL takes advantage of the Heterogeneous Relation Graphs (HRG) [3], which is a very flexible formalism for knowledge representation. It is also accommodated by a powerful toolkit, the Edinburgh Speech Tools, which support the development of new languages and voices. FLITE [4] is a small-footprint version of FESTIVAL; it is portable and has been written from scratch in C language. It targets to embedded systems and mobile devices. FreeTTS [5] is a Java version of FLITE. ProSynth [6] uses an XML scheme, namely ProXML, for data manipulation through the several procedures, while EULER [7] implements a set of list layers, aligned over a common time axis, the Multilayer Data Structures. EULER utilizes the MBROLA diphone synthesizer [8], which is also supported in all the above systems.

However, documents do not consist of normalized text, i.e. strings that their pronunciation can be looked up in a lexicon. Non-Standard Words [9] and other text meta-data carry additional expressive information that the typical TtS paradigm strips out during synthesis. Several works have been experimenting with the auditory representation of meta-information, like mathematic formulas [10] and visual components [11][12][13][14]. All studies agree that meta-information is not efficiently represented by simply replacing them with text descriptions, but more aural-oriented elements, such as prosody control and audio sounds, should be embedded in the audio format of a document. During this thesis we found out that such approaches should be also applied during the normalization procedure of plain texts. Dates, hard to remember telephone numbers, parenthetical texts, punctuation marks (especially when used repetitive) etc, should raise different speech expressions than normal text.

DEMOSTHeNES is an open and flexible language platform that allows both the interconnection with several research tools (e.g. FESTIVAL, PRAAT, WEKA), but also introduces novel features such as the Document-to-Audio framework for

vocalizing meta-data, the text normalization component for inflected languages [25], the heterogeneous object lexicon, the XSLT authoring tool for modules and the extensive support by several auxiliary procedures for experiments.

From Document to Audio

In order to cover the need for a more precise auditory representation of documents, we introduce the Document-to-Audio (DtA) framework. Figure 1 presents the general architecture of DtA.

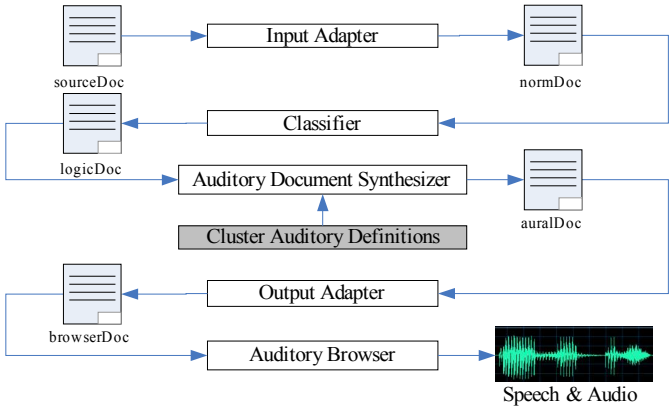

Fig. 1. The architecture of the DtA framework.

The auditory image of each cluster of information in the source document is defined by a Cluster Auditory Definition (CAD) XSLT-based scripts. Thus, any class of meta-information can be transferred in the audio modality by any of the below properties:

- prosody modification
- sound insertion
- voice switch
- pronounceable description

or any other property of the underlying TtS. The integration details of the DtA framework with DEMOSTHeNES can be found in [15][16]. A series of pilot psychoacoustic experiments concerning the representation of visual documents and simple/complex tables using the DtA framework, showed the perceptual enhancement of structure understanding by both visual capable and visual impaired people [20][21] [22][23][24].

3. Prosody Modeling Using Linguistically Enriched Features

To improve the naturalness and the realism of synthetic prosody, we study the effects of the introduction of high-level linguistic information in the prosodic structure models of phrase breaks, pitch accents and boundary tones. Due to the lack of a sophisticated appropriate linguistic analyzer to extract the required features, we used an extended SOLE-ML markup scheme to provide the TtS with more evidence of stress and intonational focus information in documents.

The speech corpus used from modeling was derived from a museum exhibits descriptions corpus [17]. 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 enriched linguistic meta-information.

Prediction of ToBI marks

For the prediction of the ToBI marks we followed the common in other languages approach of CART classification trees. Tables 1, 2 and 3 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.

Table 1. 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 2. 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

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 were 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. 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%) 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.

Table 3. 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

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%. In the rest 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.

F_0 curve generation

To build the F0 model we also chose the commonly adopted Linear Regression [18] method (F0-LR). Three models were built to predict the F_0 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. In the first group, we evaluate the F_0 models against the original supplied ToBI values (i.e. from the hand-labeled annotations). In the second group, we use the predicted ToBI marks from the TtS chain to evaluate the actual synthetic F_0 contour.

Table 4 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 4. 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 5 tabulates the performance of the F_0 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 F_0. Interesting points can be deduced from this table. First of all, the accuracy of the ToBI accent models presented in Table 2 is not depicted in the correlation of F_0 in the cases of C1 and C2, where we have a mean decrease of 17.6% and 13.0% respectively compared to Table 4, 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 F_0 targets by 19.6%.

Table 5. 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

4. The Tone-Group Selection Model

Taken into account the naturalness that speech unit selection generates, but also the complexity in achieving adequate coverage from corpus-based models, we propose a data-driven method for F_0 modeling based on selecting Tone-Groups (TG) [19]. We suggest that if a TG with a specific prosodic structure is spoken in one way, then another TG with a similar structure should be spoken in a similar way in terms of speech synthesis. We will present the design, implementation and evaluation of this approach.

We propose a TG patterning scheme where the encoding of each unit is based only on the intonational structure of the perceptually significant syllables, while the rest

(null) syllables F_0 contours are being approximated during runtime. Thus we achieve to:

- minimize the database size
- maximize the coverage of each unit in the inventory, by abstracting the perceptually surplus information
- efficiently encode TGs content in small-footprint cases

Figure 2 shows the TG structure of an utterance.

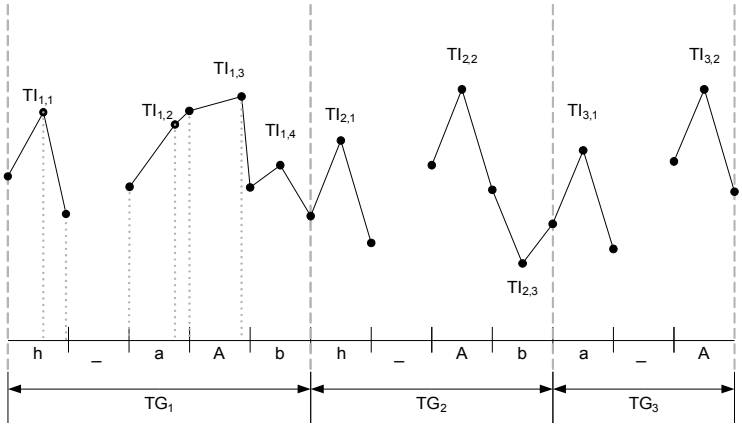

Fig. 2. Tone-Groups, *haAb* (significant) patterns and Tone-Items. The horizontal grid represents the syllables, whereas the dotted lines show the positioning details of each target.

During the candidate grid construction and the calculation of the best path, we use the F_0 difference measured in cents between the two adjacent Tone-Items of the TGs to be connected and the F_0 slope difference as a binary criterion to accept or reject a connection, as steep slopes in connections usually result in trembling voice. Other weighted measures such as *start_f0*, *end_f0*, *max_f0*, *mean_f0* and *stddev_f0* are also calculated. Thus, we achieve to minimize the total distortion of the curve caused by F_0 discontinuities.

Objective evaluation

The re-synthesis of the original pitch curve based on the TG model, achieved a 0.94 correlation and 15.23 RMSE between the original pitch and the one produced by the approximation of the null syllables F_0 with straight lines. These values correspond to what we miss when we encode pitch contours following the significant syllable scheme. As expected, the simplistic straight-line interpolation strongly affects the performance. On the other hand, since these parts do not represent any nuclear accent nor significant pitch movements, we assumed that a more phonologically aware rendering would enhanced the performance. The approach we followed here is based on a scalable progression (rising or falling) of the pitch in between the boundaries of the *null* syllables, using data from the pitch baseline, the adjacent *significant* syllables

and the word morphology. This plain approach shows great improvement in the performance of the model (correlation=0.96, RMSE=9.20) and partially supports our assumption that we have managed to represent the pitch curve by modeling only the important syllables, whereas some reasonable pitch scales can successfully replace the rest portions.

Listening tests

We set up an experimental environment to subjectively evaluate the TGS model. A professional speaker uttered 12 sentences with no major pitch and intensity alterations in his voice, i.e. avoiding emphasizing or de-emphasizing any part of the utterances. These utterances were further modeled using the TGS model. We used two configurations, the first one using a natural voice carrier and the second one using the MBROLA diphone database gr2[1].

The first listening test targeted the evaluation of the capability of the TGS approach to sufficiently represent the F_0 surface from a perceptual view. This was carried out to answer, How natural does TG selection sound? . For that reason, we compared against a well-established Linear Regression (LR) model [17]. Figure 3 shows 31 listeners preferences in each stimulus when using a natural voice carrier (left) and diphone concatenation (right). In total, the TGS model was preferred in 75% of the cases, while the LR in 25%.

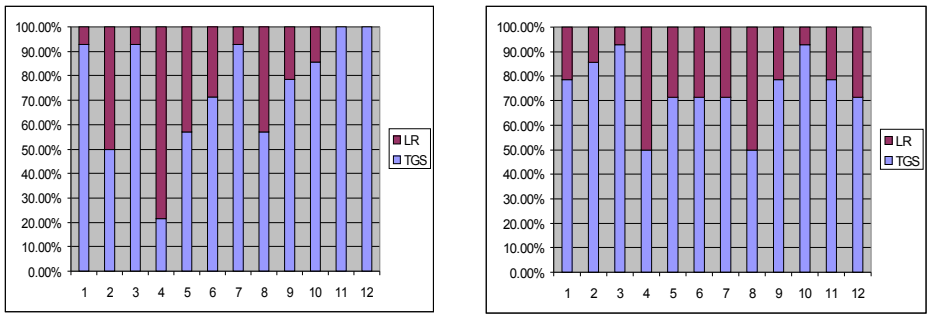

Fig. 3. Listeners choices between LR and TGS (left = natural voice carrier, right = diphone voice). Vertical axis shows the stimulus index.

Both cases clearly show that the TGS model sounded more natural to the users, though it models less syllables (only 4 at maximum per TG) than the LR model. This supports our assumption that we can encode only the intonationally most important syllables without losing in F_0 definition.

The second listening test concerned the provision of emphasis in utterances. Emphasis feature was applied in three levels, minor, major and none. Each sentence featured at least one major focus and zero or more other prominent points of any level. For each sentence, listeners were asked to mark the level of emphasis they

[1] .All the speech waves used for the tests can be found in http://www.di.uoa.gr/~gxydas/en/tone_group_selection.shtml

perceived at the pre-defined points *A* and *B* as follows: 0 for null, 1 for minor and 2 for major.

We tested the null hypothesis that the listeners perceived the target emphasis in more than 1 level of difference than the intended level. Thus, we calculated the probability *p* that the mean absolute difference *M* would be as different or more different from 1. We also made the assumption that our data are normally distributed. The results showed that we can reject the null hypothesis in the majority of the cases ($p<0.05$), while the mean difference between the target emphasis and listeners perception was 0.27 in the case of the natural voice carrier and 0.23 in the case of the MBROLA voice. It can be derived that listeners perceive 15% clearer the focus information in diphone synthesis (not diphone-selection), where the intensity is constant throughout the whole stimuli and the signal features flat dynamics.

References

1. Black, W.A., Taylor, P. and Caley, R.: The FESTIVAL Speech Synthesis System, (1998) http://www.festvox.org
2. Black, W.A., and Taylor, P.: CHATR: a generic speech synthesis system, Proc. the 15th Int l Conf. Computational Linguistics (COLING94), Kyoto, Japan., vol.2, (1994) 983-986.
3. Taylor, P., Black, W.A. and Caley, R.: Heterogeneous Relation Graphs as a Mechanism for Representing Linguistic Information, Speech Communication, vol. 33, (2001) 153-174.
4. Black, W.A. and Lenzo, K.A.: FLITE: a small fast run-time synthesis engine, Proc. 4th ISCA Workshop on Speech Synthesis (SSW4), Perthshire, Scotland, (2001) 204-207.
5. FreeTTS; http://freetts.sourceforge.net
6. Huckvale, M.: Representation and Processing of Linguistic Structures for an All-Prosodic Synthesis System Using XML, Proc. 6th Eu. Conf. Speech Communication and Technology (EUROSPEECH 99), Budapest, Hungary, (1999) 1847-1850.
7. Dutoit, T., Bagein, M., Malfrere, F., Pagel, V., Ruelle, A., Tounsi, N. and Wynsberghe, D.: EULER: an Open, Generic, Multi-lingual and Multi-Platform Text-To-Speech System, Proc. 2nd Int l Conf. Language Resources and Evaluation (LREC 2000), Athens, Greece, (2000) 563-566.
8. Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vreken, O.: The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, Proc. 4th Int l Conf. Spoken Language Processing (ICSLP 96), Philadelphia, PA, USA, vol. 3, (1996) 1393-1396.
9. Sproat, R., Black, W.A., Chen, S., Kumar, S., Ostendorf, M. and Richards, C.: Normalization of non-standard words, Computer Speech and Language, vol. 15, no. 3, (2001) 287-333.
10 Raman, V.T.: An Audio View of (LA)TEX Documents, Proc. TexUsers Group, vol. 13, no. 3, (1992) 372-379.
11. Shriver, S., Black, W.A. and Rosenfeld, R.: Audio Signals in Speech Interfaces, Proc. Int l Conf. Speech and Language Processing (ICSLP 2000), Beijing, China, vol. 1, (2000) 142-145.
12. Blattner, M.M., Sumikawa, D.A. and Greenberg, R.M.: Earcons and Icons: Their Structure and Common Design Principles, Human Computer Interaction, vol. 4, (1989) 11-14.
13. Gorny, P.: Typographic semantics of Webpages Accessible for Visual Impaired Users, Mapping Layout and Interaction Objects to an Auditory Interaction Space, Proc. Int l Conf. on Computer Helping with Special Needs, (2000) 17-21.

14.Truillet, P., Oriola, B., Nespoulous, J.L. and Vigoroux, N.: Effect of Sound Fonts in an Aural Presentation, Proc. of 6th ERCIM Workshop, UI4ALL, (2000) 135-144.
15.Xydas G. and Kouroupetroglou G.: Augmented Auditory Representation of e-Texts for Text-to-Speech Systems. Lecture Notes in Artificial Intelligence, Vol. 2166, (2001) 134-141.
16.Xydas G. and Kouroupetroglou G: The DEMOSTHeNES Speech Composer. Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, (2001) 167-172.
17.Xydas, G., Spiliotopoulos, D. and Kouroupetroglou, G.: Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora, IEICE Trans. of Information and Systems, Special Section on "Corpus-Based Speech Technologies", vol. E88-D, no 3, (2005) 510-518.
18.Black, W.A. and Hunt, A.: Generating F0 contours from the ToBI labels using linear regression, Proc. 4th Int l Conf. Spoken Language Processing (ICLSP 96), Philadelphia, USA, vol.3, (1996) 1385-1388.
19.Xydas, G. and Kouroupetroglou, G.: Tone-Group F0 selection for modeling focus prominence in small-footprint speech synthesis, Speech Communication, vol. 48, issue 9, (2006) 1057-1078.
20.Spiliotopoulos, D., Xydas, G. and Kouroupetroglou, G.: Diction Based Prosody Modeling in Table-to-Speech Synthesis. Lecture Notes in Computer Science, LNCS 3658, Springer-Verlag Berlin Heidelberg, (2005) 294-301.
21.Spiliotopoulos, D., Xydas, G., Kouroupetrolgou, G. and Argyropoulos, V.: Experimentation on spoken format of tables in auditory user interfaces, Proc. 11th Int l Conf. Human-Computer Interaction (HCII2005), Las Vegas, Nevada SA, (2005) 361-370.
22.Xydas, G., Argyropoulos, V., Karakosta, T. and Kouroupetroglou, G.: An Experimental Approach in Recognizing Synthesized Auditory Components in a Non-Visual Interaction with Documents, Proc. 11th Int l Conf. Human-Computer Interaction (HCII 2005), vol. 3, (2005) 411-420.
23.Xydas, G., Argyropoulos, V., Karakosta, T. and Kouroupetroglou, G.: An Open Platform for Conducting Psycho-Acoustic Experiments in the Auditory Representation of Web Documents, Proc. National Conf. ACOUSTICS 2004, (2004) 157-164.
24.Xydas, G., Spiliotopoulos, D. and Kouroupetroglou, G.: Modelling Emphatic Events from Non-Speech Aware Documents in Speech Based User Interfaces, Proc. 10th Int l Conf. on Human - Computer Interaction (HCII2003), Crete, Greece, (2003) 806-810.
25.Xydas, G., Karberis, G. and Kouroupetroglou, G.: Text Normalization for the Pronunciation of Non-Standard Words in an Inflected Language, Methods and Applications of Artificial Intelligence, LNAI 3025, Springer-Verlag Berlin Heidelberg, (2004) 390-399.