ON THE USE OF CART-TREE FOR PROSODIC PREDICTIONS IN THE ITALIAN FESTIVAL TTS

Piero Cosi, Cinzia Avesani

ISTC-SFD - (ex IFD) CNR
Istituto di Scienze e Tecnologie della Cognizione
Sezione di Fonetica e Dialettologia
(ex Istituto di Fonetica e Dialettologia)
Consiglio Nazionale delle Ricerche
e-mail: {cosi, avesani}@csrf.pd.cnr.it
www: http://nts.csrf.pd.cnr.it/
www: http://www.csrf.pd.cnr.it/

Fabio Tesser, Roberto Gretter, Fabio Pianesi

ITC-IRST
Istituto Trentino di Cultura
Istituto per la Ricerca Scientifica e Tecnologica
e-mail: {gretter, tesser, pianesi}@irst.itc.it
www: http://www.itc.it/IRST/index.htm

1. ABSTRACT

A new data-driven prosodic module for the Italian FESTIVAL Text-To-Speech (TTS) synthesizer is described. This module, based on the statistically motivated "Classification and Regression Trees" (CART) theory, is described and compared with a previously developed rule-based module[1].

2. INTRODUCTION

Prosody is "the organizational structure of speech" [1]. It refers both to the phonological organization of segments in higher level constituents and to the pattern of relative prominences within these constituents and to the phonetic reflexes of this organization in the patterns of F0, duration, amplitude and segment quality in an utterance.

Three main aspects are involved in this quite complex phenomenon: ***phrasing***, the 'chunking' of an utterance in a hierarchy of prosodic constituents such as major and minor intonational phrases, prosodic words etc., which correspond to meaningful units of information; ***accentuation***, that refers to the location and type of pitch accents in the text; and ***tune***, the melodic composition of an utterance.

With good controlling of suprasegmental features, different syntactic patterns and even emotions can be well modeled in speech: however, almost everything seems to have effect on prosodic features of natural speech, which makes accurate modeling very difficult.

In the past decade there has been a growing appreciation of the importance for a TTS and Concept-To-Speech (CTS) system to generate appropriate linguistic representation of prosody, from which appropriate acoustic patterns can be obtained, which will be manifested in the output speech waveform. The output of a TTS system, with such a prosodic component is still a sequence of phones, each of which has an energy, a duration and an F0 (pitch) value. Natural prosody heavily relies on syntax, semantics and pragmatics. Since very few data is currently available on the last two linguistic aspects, TTS systems merely concentrate, if they do, on syntax. A high quality Text-To-Speech

[1] Part of this work has been executed in the framework of the research project named *MPIRO* (*Multilingual Personalized Information Objects*. European Project IST-1999-10982 - WWW page: http://www.ltg.ed.ac.uk/mpiro/), founded by the *European Commission* and *TICCA* (*Tecnologie Cognitive per l'interazione e la cooperazione con agenti artificiali*), co-founded by *CNR* and *Provincia Autonoma Trentina*.

system should at least comprise a morpho-syntactic analyzer able to reduce a given sentence into a sequence of parts-of-speech, and to further describe it in the form of a syntax tree to unveil its internal structure. Accurate phonetic transcription can only be achieved provided the part of speech category of some words is available, as well as if the dependency relationship between successive words is known, and the internal structure of a sentence drives strongly its pitch pattern. At the present time, without a complex syntactic analyzer, the module that gives a Text-To-Speech synthesizer an acoustic description of prosody, in the form of a sequence of phoneme energy, durations and F0 targets, is surely the weakest part of the system.

3. ITALIAN FESTIVAL TTS

Festival is a general multi-lingual speech synthesis system developed at the CSTR (Center for Speech Technology Research, Edinburgh, Scotland, UK) [2-3] offering a full text to speech system with various APIs (Application Program Interfaces), and an environment for development and research of speech synthesis techniques. It is written in C++ with a SCHEME-based command interpreter [4] for general control. Festival is a general-purpose concatenative text-to-speech (TTS) system that uses the residual-LPC [5], the OGI-residual-LPC [6] and MBROLA [7] synthesis techniques, and is able to transcribe unrestricted text to speech. Festival contains standard tools for intonation and duration prediction. Generally, intonation is generated in two steps: prediction of pitch accents and prediction of F0 values. In the simplest case intonation parameters are just set to constant values at the start and at the end of the utterance (130, 110 Hz); more complex modules use the so called Classification and Regression Trees (CART) [8]. The Italian Festival modules [9-11] are described in Figure 1.

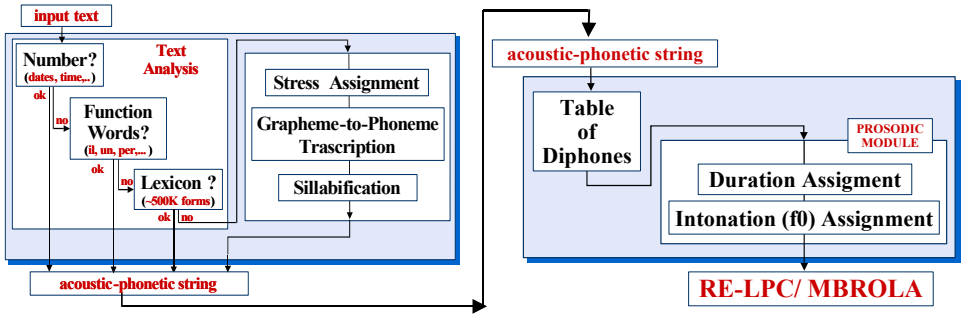


Figure 1. Text/Linguistic/Prosodic modules for Italian TTS followed by the synthesis engine module (RE_LPC and MBROLA).

A first module implements a simple grammar to convert strings of numbers into word sequences. Numbers are expanded at the word and phoneme level distinguishing among time, dates, telephone numbers, etc.. Non-numerical data are divided into “function-words” and “content-words”. In the prosodic modules, “function words” are treated differently from lexical words. All the other words are phonemically transcribed in a wide lexicon, compiled in Festival format to speed up search procedures. If they are not present in the lexicon, they are phonemically transcribed via explicit stress-assignment, letter-to-sound and syllabification rules. The Lexicon, compiled in Festival format, comprises 500k stressed forms, phonemically transcribed, divided in syllables and labeled after their

grammatical class or part-of-speech (POS). The correct diphones are then extracted from the acoustic database, where for each unit, specific information relative to its mean duration and pitch is included, and, finally, duration and F0 assignment is applied before activating the speech synthesis engine (residual LPC, OGI residual LPC, MBROLA) to generate the speech waveform.

4. PROSODIC MODULE

The aim of this module is to modify, by explicit rules, average duration and F0 values assigned by default to each phoneme by a factor ranging between a maximum and minimum threshold, in order to possibly capture all the phenomena that are known to influence prosody.

Segment's intrinsic duration can vary according to a number of segmental and suprasegmental factors, among which we can mention tempo, speaking style (read *vs*. spontaneous; accurate *vs*. sloppy, natural *vs*. reiterant etc.), phonetic context, stress, accent, focus, position in prosodic phrase.

Segments display different F0 values according to their intonational status, i.e. to whether they belong to pitch accented syllables, to the type of pitch accents they are associated with, to the role pitch accents play in the pitch contour, to their flanking intonational phrase boundaries etc. Moreover, variations in overall pitch contour may be due to gender, physical and emotional state, syntactic mood, speaker attitude, speaker beliefs, position in discourse structure etc. At present, only a very small set of these factors are taken into consideration.

4.1 *Rule- based prosodic module*

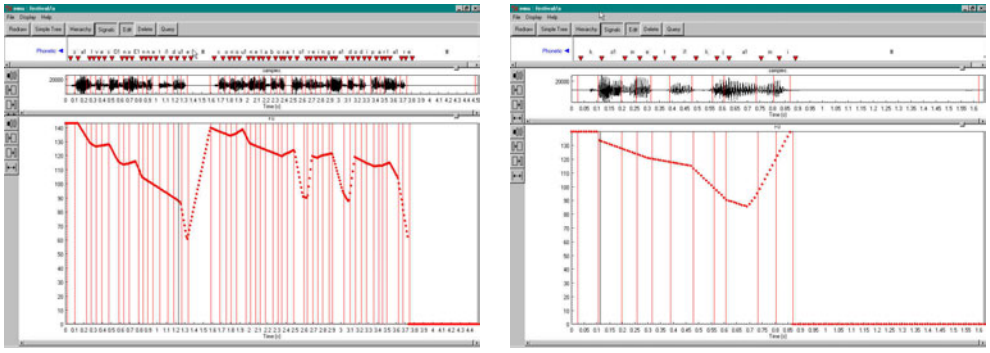
The rule-based prosodic module actually present in the Italian version of Festival is quite simple and relies essentially on punctuation marks and function words.

Each phoneme is assigned a mean duration, which was statistically computed by analyzing a wide corpus of Italian sentences produced by various RAI Italian television announcers [12]. The duration of stressed vowels is augmented by 20% relative to the average vowel duration. Pauses between words are divided in 2 categories: short pauses of 250 ms, associated with punctuation marks such as [\ , ;] and long pauses of 750 ms associated with main conclusive punctuation marks such as [? . : !].

As for intonation, declarative sentences are segmented in intonational phrases based on punctuation marks and presence of function words; each phrase is assigned a baseline starting at 140Hz and ending at 60Hz (for a typical male voice). For any stressed syllable, the F0 contour is raised by approximately 10Hz over the baseline, while the last syllable has a steeper inclination relative to the baseline. A resetting of the baseline is triggered by the presence of any punctuation marks and by any function word. Interrogative sentences are also segmented in intonation phrases and assigned baseline values. Relative to declaratives, these values are modified in phrase final position to produce a falling-raising pattern associated with the last word in the sentence.

Specifically, for an interrogative sentence a "Target Point" (TP) is assigned to the last stressed vowel, and is aligned at 3/4 of its duration: at that point the F0 curve reaches a value corresponding to 80% of the baseline, falling from a value equal to the baseline assigned to the end of the preceding vowel. Starting from TP, F0 raises up to F0max with an inclination that spans over the post-tonic unstressed syllables. The last syllable is

assigned a faster inclination. An example of a declarative and interrogative sentence synthesized by FESTIVAL is given in Figure 2.



(a) **(b)**

Figure 2. F0 contour for two declarative sentences (a): "Salve, sono NT2. Sono un elaboratore in grado di parlare" (Hello, I am NT2. I am a computer able to speak.); and for one interrogative sentence (b): "Come ti chiami?" (What's your name?), synthesized with a male voice with the rule-based prosodic module.

4.2 *Statistically- based "CART" prosodic module*

A CART [8] is a statistical method for predicting data from a set of feature vectors. In particular, a CART is a binary branching tree with questions about the influencing factors at the nodes and best predicted values at the leaves. The tree contains yes/no questions about the features and provides either the probability distribution or a mean and standard deviation. Decision trees are obtained by finding the question that splits the data minimizing the mean "impurity" of the partition; while the impurity is small when the items are similar.

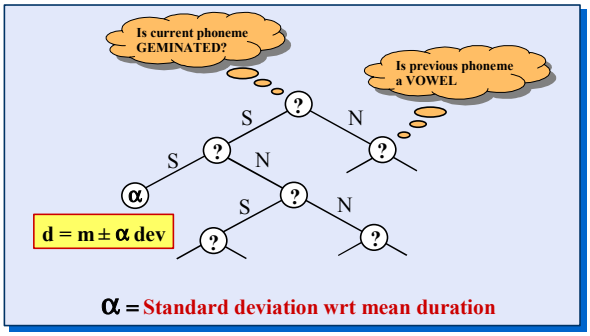


Figure 3. Graphical representation of the statistical method named CART.

Decision trees have the interesting property of providing a straightforward mechanism for combining the virtues of probabilities with knowledge from rule-based approaches. Because the structure of the knowledge from the rules is preserved, these methods can be used not only to model the phenomena of interest but also to test hypotheses and to gain knowledge about which information sources are providing the most gain in the models [13]. The practical advantages of CARTs are that standard tools for their generation are widely

available, and that the computed regression tree is interpretable. The disadvantage lies in the fact that it needs a large amount of training data.

As for Italian Festival, two CARTs were trained in order to identify correlations between linguistic information and duration and intonation contours from two set of training data. The *duration* of each phoneme and the *F0 values* of each syllable (*start/end of the syllable and mid of the vowel*) were independently predicted by two CARTs applied to two corpora of different type of natural speech, spoken in different speaking styles: a broadcasted news corpus, spoken by a national TV announcer (RAI-news) [11] for duration and a corpus of read child stories, spoken by a professional speaker (RAI-fiabe) [14] for F0 intonation. No ToBI-based [15] or Tilt-based [16] intonation transcription is still available for these corpora, so direct F0 contours become the input/output for the corresponding CART tree. At the moment, only text-type segmental, lexical and syntactic information, indicated in Table 1 for duration and in Table 2 for intonation (F0), are used while building the classification trees.

- **REALISED SEGMENT**
  - *segment identity*: SAMPA phone name
  - *segment type*: vowel, consonant, glide
  - *vowel length*: short, long, diphthong, schwa
  - *vowel height*: high, mid, low
  - *vowel frontness*: front, mid, back
  - *lip rounding*: yes, no
  - *consonant type*: stop, fricative, affricate, nasal, liquid
  - *place of articulation*: labial, alveolar, palatal, labio-dental, dental, velar
  - *consonant voicing*: yes, no
  - *geminate*: yes, no
- **POSITION OF SEGMENT IN SYLLABLE**
  - *position in syllable initia*l: integer
  - *syllable par*t: onset, coda
- **SYLLABLE**
  - *lexical stres*s: stressed, unstressed
  - *break level after this syllable*: integer
- **POSITION OF SYLLABLE IN WORD**
  - *position type*: single, initial, final, mid
- **WORD**
  - *function word*: yes, no
  - *function word type*: conjunction, pronoun, etc…
  - *first function word*: yes, no
  - *part-of-speec*h: noun, adjective, etc..
  - *length in syllable*s: integer
  - *break level after this word*: integer
- **REALISED PREVIOUS AND FOLLOWING SEGMENTS OR SYLLABLES OR WORDS:**
  - *the features above are also computed for the previous and following segments, syllables and words in order to consider coarticulation effects.*

Table 1. Factors used for the duration prediction CART module. For each realized sound segment the following factors were extracted from the database. Domains are presented in small capitals, factors in italics, and the possible values in standard font.

- **REALISED SYLLABLE**
  length in segment: integer
  lexical stress: stressed, unstressed
  break level after this syllable: integer
  syllable duration: float
  syllable vowel: SAMPA vowel name
  syllable first segment type: vowel, consonant (stop, fricative, affricate, nasal, liquid) glide, geminate
- **POSITION OF SYLLABLE IN WORD**
  position type: single, initial, final, mid
  number of syllables since last phrase break: integer
  number of syllables to next phrase break: integer
  number of syllables to next stressed syllable: integer
- **WORD**
  function word: yes, no
  function word type: conjunction, pronoun, etc…
  first function word: yes, no
  part-of-speech: noun, adjective, etc..
  length in syllables: integer
  break level after this word: integer
  word duration: float
- **POSITION OF WORD IN PHRASE**
  position of this word: integer
  number of words to end of this phrase: integer
  number of non-major phrase breaks since last major phrase break: integer
- **PHRASE**
  phrase type: dichiarativa, interrogativa
- **REALISED PREVIOUS AND FOLLOWING SEGMENTS OR SYLLABLES OR WORDS:**
  the features above are also computed for the previous and following segments, syllables and words in order to consider coarticulation effects.

Table 2. Factors used for the intonation prediction CART module. For each realised sound segment the following factors were extracted from the database. Domains are presented in small capitals, factors in italics, and the possible values in standard font.

Automatic segmentation, obtained through the use of quite an efficient phonetic recognition system [17], pitch marks, F0 and feature extraction computed by Praat [18] (for LPC residual synthesis), have been executed before the CART training, using various Tcl/Tk modules [19]. Moreover, for an homogeneous treatment of the data - that is to factor out the influence of the intrinsic duration and intonation - the absolute values were first converted to z-scores, and the mean and the standard deviation of each sound were stored in a separate file. The CART was then trained on the training corpora with the program "WAGON" a tool, from the Edinburgh Speech Tools Library [20], available with the FESTIVAL Speech Synthesis system.

5. OBSERVATIONS AND CONCLUDING REMARKS

Visual inspection and preliminary listening tests of various examples, such those illustrated in Figure 4 for a simple sentence, revealed that the prosodic results obtained by CART seem good and natural. As for duration, the overall quality of the system looks more accurate and natural when compared with that obtained with the rule-based prosodic module. It is clear that much more accurate perceptual experiments shall be designed in the future to support these preliminary observations.

As far as intonation is concerned, we can compare the four plots in Figure 4. They represent: (a) the original F0 pattern of the Italian sentence: “Quando Stefano Roi compí i dodici anni, chiese in regalo a suo padre, capitano di mare e padrone di un bel veliero, che lo portasse con sé a bordo” ("When Stefano Roi was twelve, asked to his father, who was a captain and owner of a beautiful sailingship, to be taken on board with him"), extracted from a short novel written by the Italian writer Dino Buzzati, pronounced by an Italian professional speaker; its corresponding synthesized versions for the three different prosodic modules: (b) explicit hand-rules, (c) CART trained on the RAI-news corpus, and (d) CART trained on RAI-fiabe corpus.

It can be observed that, when the explicit rules are applied for F0 pattern prediction, the pitch starts at a high level and decreases slowly along the sentence, a reset of F0 is activated at each functional group, and a significant fall in F0 values is present at the end of sentence. As for CART modules, it seems that, with this statistically motivated technique, the system “learns” quite effectively the speaker style characteristics. Looking at plot (c), which refers to the CART trained on RAI-news, we can see that F0 starts again at a very high level at the beginning of the sentence and that underlines, in fact, the common intention of typical TV news announcers of “capturing attention”. Finally, comparing plots (c) and (d), a part from the different pitch range, the F0 pattern predicted in (d) is more similar to the original one than those obtained in (c). This result could have been predicted, because the prosodic style of the original sentence, the reading of a short novel, is much more similar to the prosodic child-story reading style of the RAI-fiabe corpus than to the news-reading style of the RAI-news corpus .

Furthermore, we should remind that the speech material used to train the CARTs was not “prosodically” transcribed, i.e. labeled with ToBi or Tilt annotation schemes. Therefore, we could be confident that better results will be obtained when these factors will be included in our system in the near future.

6. FUTURE TRENDS

Various corpora will be ToBi-labelled and annotated with linguistic and discourse information and new CARTs will be trained using also those new informations for phrasing, accent placement and F0 prediction.

Another important research activity will be focused on emotions. The speaker's feelings and emotional state influence speech in many ways and the proper implementation of these features in synthesized speech may considerably increase its quality [21-23]. Within text-to-speech systems this is a rather difficult task, because written text usually contains no information of these features. However, this kind of information may be provided to a synthesizer with some specific control characters or character strings.

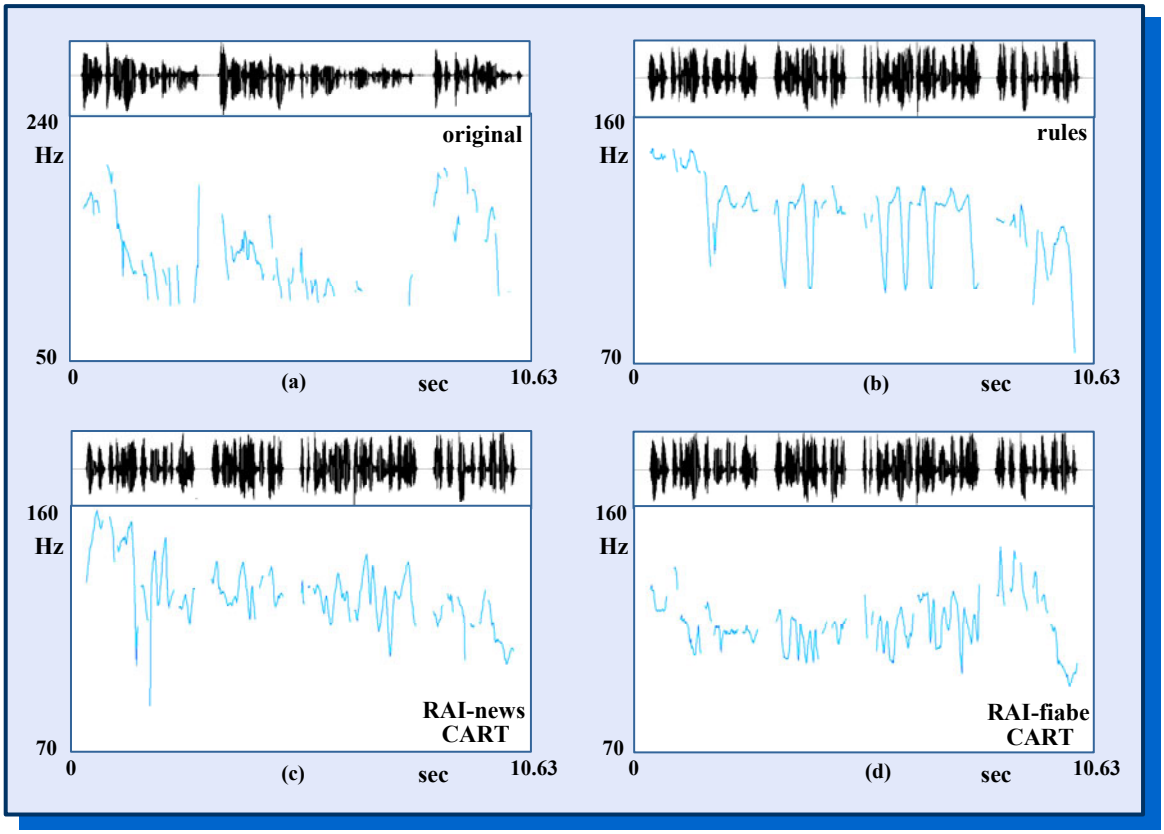


Figure 4. Waveform and F0 contour for the Italian sentence: “Quando Stefano Roi compí i dodici anni, chiese in regalo a suo padre, capitano di mare e padrone di un bel veliero, che lo portasse con sé a bordo” ("When Stefano Roi was twelve, asked to his father, who was a captain and owner of a beautiful sailingship, to be taken on board with him") pronounced by an Italian professional speaker (a) and synthesised by rules (b), by CART trained on RAI-news corpus (c) and by CART trained on RAI-fiabe corpus (d).

REFERENCES

[1] Beckman M., The parsing of prosody, *Language and Cognitive Processes*, Vol. 11, 1/2, 1996, pp. 17-67.

[2] Black A.W., Taylor P.A., Caley R., The Architecture of the Festival Speech Synthesis System”, in *The Third ESCA Workshop in Speech Synthesis*, Jenolan Caves Mountain House, Blue Mountains, Australia, November 26-29, 1998, pp. 147-151.

[3] Black A.W., Taylor P.A., Caley R., *The Festival speech synthesis system 1.4.2, www:* http://www.cstr.ed.ac.uk/projects/festival/manual/festival-1.4.0.ps.gz.

[4] *SCHEME, Computer Programming Language,* www: http://www-swiss.ai.mit.edu/~jaffer/Scheme.html.

[5] Dutoit T., *An Introduction to Text-To-Speech Synthesis*, Kluwer Academic Publishers, pp. 1996, 326.

[6] Macon M., Cronk A., Wouters J., Kain A., *OGIresLPC: Diphone synthesiser using residual-excited linear prediction*, num. CSE-97-007, Department of Computer Science, OGI, Portland, OR, Sep, 1997.
[7] Dutoit T., Leich H., MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database, *Speech Communication*, Elsevier Publisher, December 1993, vol. 13, n° 3-4, pp. 435-440.
[8] Breiman L., Friedman J., Stone C.J., Olshen R.A., "Classification and Regression Trees". Chapman & Hall/CRC, 1984.
[9] Cosi P., Tesser F., Gretter R., C. Avesani C., Festival Speaks Italian!, in *Proceedings of EUROSPEECH 2001*, Aalborg, Denmark, Sep 3-7 2001, pp. 509-512.
[10] Cosi P., Gretter R., Tesser F., FESTIVAL parla italiano!, in *Atti delle XI Giornate di Studio del G.F.S.*, Padova, Italy, November 29-30, Dicember 1 2000, pp. .
[11] Cosi P., Gretter R., Tesser F., Recenti sviluppi di FESTIVAL per l'italiano, in *Atti delle XII Giornate di Studio del G.F.S.*, Macerata, Italy, Dicembre 13-15, 2001, (in press).
[12] Federico M., Giordani D., Coletti P., Development and evaluation of an Italian broadcast news corpus, in *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2000
[13] Price P., Hirschberg J., Session 13: Prosody, Proceedings of the Speech and Natural Language Workshop, DARPA, 1992, pp. 416-417.
[14] www: http://www.rainet.it/bambini
[15] Beckman M., Ayers E.G., *Guidelines for ToBI Labelling*, Ohio State University. http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html
[16] Taylor P., The Tilt Intonation Model, in in *Proceedings of International Conference on Spoken Language Proessing. (ICSLP-1998)*, Sydney Australia, 30th November - 4th December 1998, Paper 827, Vol. IV, pp. 1383-1386.
[17] Cosi P., Hosom J.P., High Performance "General Purpose" Phonetic Recognition for Italian, in *Proceedings of International Conference on Spoken Language Processing. (ICSLP-2000)*, Beijing, Cina, 16-20 October, 2000, Vol. II, pp. 527-530.
[18] Boersma P., Praat, a system for doing phonetics by computer, *Glot International* 5(9/10), pp. 341-345.
[19] Tcl/Tk: K. Ousterhout - ouster@sprite.berkeley.edu.
www: http://sol.brunel.ac.uk/tcl/Tcl.html.
[20] *The Edinburgh Speech Tools Library*. http://www.cstr.ed.ac.uk/projects/speech_tools/.
[21] Murray I.R. and Arnott J.L., Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion, Journal of Acoustical Socity of America, J.A.S. A., 93, 1993, pp. 1097-1107
[22] Schroeder M., Emotional Speech Synthesis: A Review, *in Proceedings of EUROSPEECH 2001*, (Aalborg, Denmark, 3-7 September), vol. 1, pp. 561-564.
[23] Schroeder M., Cowie R., Douglas-Cowie E., Westerdijik M. and Gielen S., Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis, Proceedings of the EUROSPEECH 2001, (Aalborg, Denmark, 3-7 September), vol.1, pp. 87-90.