Memoirs of the Scientific Sections of the Romanian Academy
Tome XXXV, 2012

COMPUTER SCIENCE

A ROMANIAN PROSODY PREDICTION MODULE BASED ON A FUNCTIONAL INTONATIONAL MODEL

DOINA JITCĂ, VASILE APOPEI and OTILIA PĂDURARU

Institute of Computer Science of Romanian Academy
Romanian Academy, Iaşi Branch, Romania
Corresponding author: doina.jitca@iit.academiaromana-is.ro

This paper presents a prosodic prediction module used by the Romanian Text-to-Speech (TtS) system in intonation synthesis. The prosody prediction refers to the surface generation of the F0 contour, based on the F0 patterns assigned to the functional categories of the prosodic units. Prior to the prediction module presentation, the paper includes a summary of these functional categories and the partial melodic contour descriptions based on functional labels. The block diagram of the prediction module outlines two main processing steps: the phrasing prediction for building the utterance tree and the selection of the melodic contours of its groups. Both processing steps are exemplified within a case study of Romanian text speech synthesis. The prosody prediction results are discussed and compared with natural F0 contours of different speakers.

Key words: functional categories, prosodic unit, phrasing prediction, melodic contour selection.

1. INTRODUCTION

A Prosody Prediction Module (PPM) was previously introduced into a Romanian Text-to-Speech system [1], [5] to improve the quality of the speech output. The main components of the TtS system, illustrated in the block diagram of Figure 1, are the linguistic module, the phonetic module and the speech synthesizer. After analyzing the input text, the PPM extracts relevant information at the prosodic level, then outputs a text structured by an utterance tree and annotated by prosodic attributes. The output data structure of the PPM is the input to the Prosody Control Module (PCM), which generates the F0 contour [5]. The Phonetic Module produces a phoneme sequence and an F0 contour for the speech synthesizer by converting prosodic information into events in the tonal space of the synthesized F0 contour [6]. The PPM and the PCM modules are based on a functional intonational model that applies a function at the communicative act level [4] to each prosodic unit. This function is relative to the parent group that includes the prosodic unit.

The prediction module requires two main processing steps: the identification of the prosodic groups at different levels of a prosodic hierarchy and the selection of the appropriate melodic contours for the intonational phrases related to the text groups. Both of them are based on prosodic indications derived from text analysis. The first step is implemented by a Phrasing Prediction submodule that generates the utterance tree of the synthesized utterance. The Melodic Contour Selection (MCS) submodule that encodes relevant information for selecting an appropriate melodic contour from a Melodic Contour Dictionary (MCD) implements the second step. The output of the PPM is a prosodic annotated text, represented by the prosodic structure of the input text and the melodic contour descriptions at each prosodic unit.

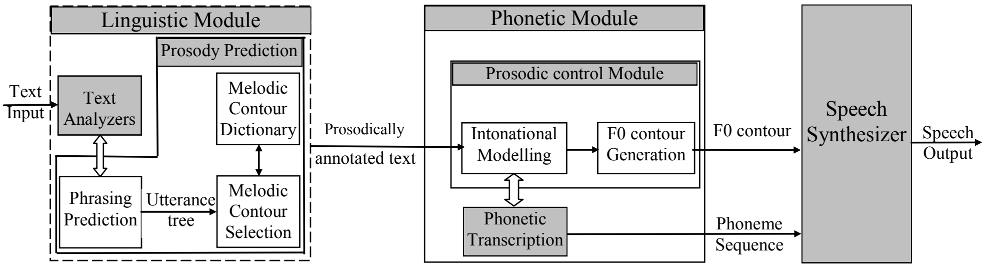


Fig. 1. The block diagram of a Text-to-Speech system.

The PPM generates an utterance tree by using an input text annotated only at the morphological level, and not syntactically structured.

The predictor of the English intonation presented in [2] has used both morphologic and syntactic annotations of the input text. Consequently, the prediction algorithm has used a syntactic hierarchy at its input and aimed to transform it into a prosodic hierarchy. In the third step, the Black-Taylor algorithm has performed a flatness of the syntactic tree by merging the syntactic units that are not separated by breaks at the prosodic level. Our PPM has used a sequence of POS (part of speech) units at the input and built a prosodic unit hierarchy.

From this point of view, our prediction module is closer to the one of the commercial text-to-speech system ACTOR, that uses only a POS annotation based on a lexicon of functional words (small variant analysis) or a word classifier (large variant analysis), for intonation prediction. Based on the POS annotation, a prediction module predicts pitch accents and simple word groups (noun + verb, pronoun+ auxiliary + verb) in order to transform them into prosodic units [9]. The way prosodic indications are extracted from a text does not generate essential conceptual differences between various implementations. The differences consist in the intonational model and in the prosodic description synthesis from the amount of prosodic indications.

Currently, our TtS system does not contain an internal morphologic analyzer. The POS unit sequences are generated by an external POS tagger module that structures and annotates the text with POS information.

The data output of the PPM is a hierarchically structured text of prosodic units placed in the nodes of the utterance tree (Fig. 2). The prosodic information generated by the PPM module contains descriptions of the melodic contours corresponding to the elementary prosodic units (Accentual Units – AUs) and to the non-elementary units (Accentual Unit Groups – AUGs).

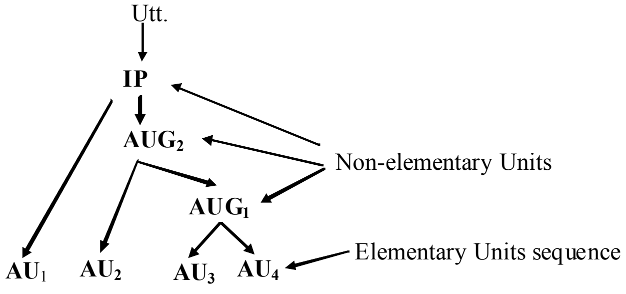


Fig. 2. An example of an intonational tree.

The descriptions of the elementary units (containing a single accentuation event), include information referring to the pitch accent type. The RO-ToBI annotation system [7] has been used to mark the pitch events of the elementary patterns.

The descriptions of the non-elementary units by a set of functional categories are materialized in sequences of functional labels which partition the prosodic groups (Intonational Phrases – IPs, intermediate phrases – ips and phonologic groups) into units with a specific role in the communication act.

Each prosodic group of the utterance tree has been assigned a sequence of functional labels. In addition, by joining the two description levels mentioned above to annotate a F0 contour, each AU (prosodic word) has been assigned both a functional label and a label from the Ro-ToBI set.

The functional categories of the accentual units are presented in the next section, while the data processing flow of the PPM is discussed in section 3. The block diagram of the melodic contour selection submodule is analyzed in section 4. Section 5 includes a case study which illustrates the data processing flow within the PPM.

2. THE FUNCTIONAL CATEGORIES OF THE ACCENTUAL UNITS

At the prosodic unit level, the description of the melodic contours by sequences of functional units, leads to group partitioning and marks the position of the units carrying a focus. The set of functional unit categories used in this paper is presented in Table 1, by labels associated with functional descriptions at the communication level. This table also contains descriptions of the F0 contour patterns by RO-ToBI labels, when the communicative functions apply/refer to elementary units (prosodic words). These descriptions provide variants of pitch accent realizations within the F0 patterns.

Table 1

The functional categories at the communication level

Functional label	**Pitch Accent**	**Communicative Function**
PH	H* L+>H*	PUSH – "pushes" forward the communication act by rising pitch movements that reach high tones in a group. These patterns are prominent at the perception level, without reaching the highest level on the accented syllable.
PO	H* L* H+L*	POP – relaxes the communication act either at the end of a group (e.g. in statements), or at its beginning, before a prominent tonal rise (e.g. in yes-no questions).
L	L* !H*	LINK – establishes a functional link between two subunits with contrasting target tones, (e.g. PH/PO), by targeting medium tonal levels, without generating a focus.
L+F	~L+H* ~L*	Establishes a functional link between two subunits with contrasting target tones and generates the focus event of a group by positive or negative emphases.
PO+F	H+L* L*+>H	Gives rise to a POP event and to a pronounced emphasis by target tones reaching minimal levels on the accented syllable.
PH+f	~H*	Gives rise to a PUSH event and to a small pitch variation reaching the maximal tonal level.
PU+F	~L*+H H%	Gives rise to a POP-UP event and to an emphasis by target tones reaching the bottom level of the tonal space, followed by a rising movement on the accented syllable to a high tone.
PD	H* L% L+H*L%	PUSH-DOWN – gives rise to a PUSH event, then brings the tone to a low level, without generating a focus (e.g. in yes-no questions with the sentence emphasis in a non-final position).
PU	L*+HH%	POP-UP - gives rise to a POP event, without generating a focus, by maintaining the tone at a low level on the accented syllable, then rises the tone on the next unaccented syllable.
L+f	~H*	Establishes a functional link between two subunits with contrasting target tones and maintains a small pitch variation around a high or medium tonal level.
PH+F	H* H*+>L L+H*	Gives rise to a PUSH event and to an emphasis by target tones reaching the top level of the tonal space, without a falling movement on the accented syllable.
PD+F	H*+LL% ~H*+LL%	Gives rise to a PUSH event, having a rising pitch movement followed by a fall on the last part of the accented syllable and, sometimes, on the next unaccented syllable.

Therefore, the functional analysis of an intonational contour divides each AUG of the utterance tree into sequences of functional units described by sequences of functional labels (e.g. PH / PO, PH / L+F / PO, PH + F / PO, PH / PO+F, etc.). The functional perspective of our model on the Romanian intonational contours leads to a more informational description than the "weak-strong" partitions ("nonfocused-focused" partitions) of Ladd's hierarchical structure of the focus [8].

3. THE PROSODY PREDICTION MODULE

The aim of the PPM consists in building the utterance tree with all prosodic information related to the generation of the F0 contour for input text synthesis. The processing flow of the PPM in Figure 3 presents all steps in achieving this data output, used by the current implementation. The input data consists in a text annotated by a POS tagger. In the editing window, the user can add some macroprosodic annotations.

During the first processing step of the PPM, the input text is stored as a list of words with their POS labels and possibly macroprosodic information (markers for word grouping and functional labels at the word group level).

During the next step, a prediction of the prosodic words (AUs) is made. The output is represented by the list of elementary prosodic unit (Accentual Unit list). The prediction is based on a set of rules extracted from a Romanian corpus by taking into account different contexts of accented and clitic words (in subsection 5.1). The list of accentual units partially keeps the POS information due to a reduction of the POS categories. The AU list underlies the utterance tree building and can be thought as a projection of the tree on the time axis. Our method includes a bottom-up building of the utterance tree. To this moment, the AU instances contain the text divided into syllables. The stress position is also specified.

Before the AU grouping steps, the PPM infers a set of prosodic indications related to the following aspects:

- Marks the AUs that cannot be separated (prosodic related AUs- PRAUs) and must to be taken together in a group or in an intonational phrase. This is the case of words belonging to a syntagm or of POS related words (e.g. a noun and its determinants). These marks are taken into account by the Phrasing Submodule. The PRAU tagging is performed by a set of rules that search certain POS sequences in the AU list generated in the previous step.
- Marks the words that must be focused in the utterance. This is the case of deictic words that refer to an action, to an action characteristic or to the spatial-temporal coordinates of actions.
- Marks the words that are related to prominent tones in the utterance. The prominent tones are those reaching maximal levels within phrases/IPs having descending F0 contours (affirmative sentences) or minimal levels within phrases/ IPs having ascending F0 contours (yes-no questions). For example, the negation word "*nu*" ("no") has an emphasis and reaches the top level of the phrase, when this phrase is a statement. An example of ascending F0 contour is the contour of a yes-no question reaching the minimal tonal level on the verb in the initial position (when this verb carries an emphasis) or on the last word, when the emphasis is placed in the final position.

The prosodic indications are stored in AU instances. The phrasing indications are taken into account in the grouping steps.

The aim of the first grouping step is to predict the breathiness groups by analyzing the phrasing indications, the punctuation marks and the length of the words to generate a reasonable length for these groups. In this step, the PPM creates AUG instances for all predicted groups. The AUG instances contain a link to their AU children from the AU list.

After the first grouping step, additional grouping conditions are tested. They refer to the existence of IP beginning/ending marks. An appropriate number of IP instances are created, with links to their AUG child instances of the output list from the previous step.

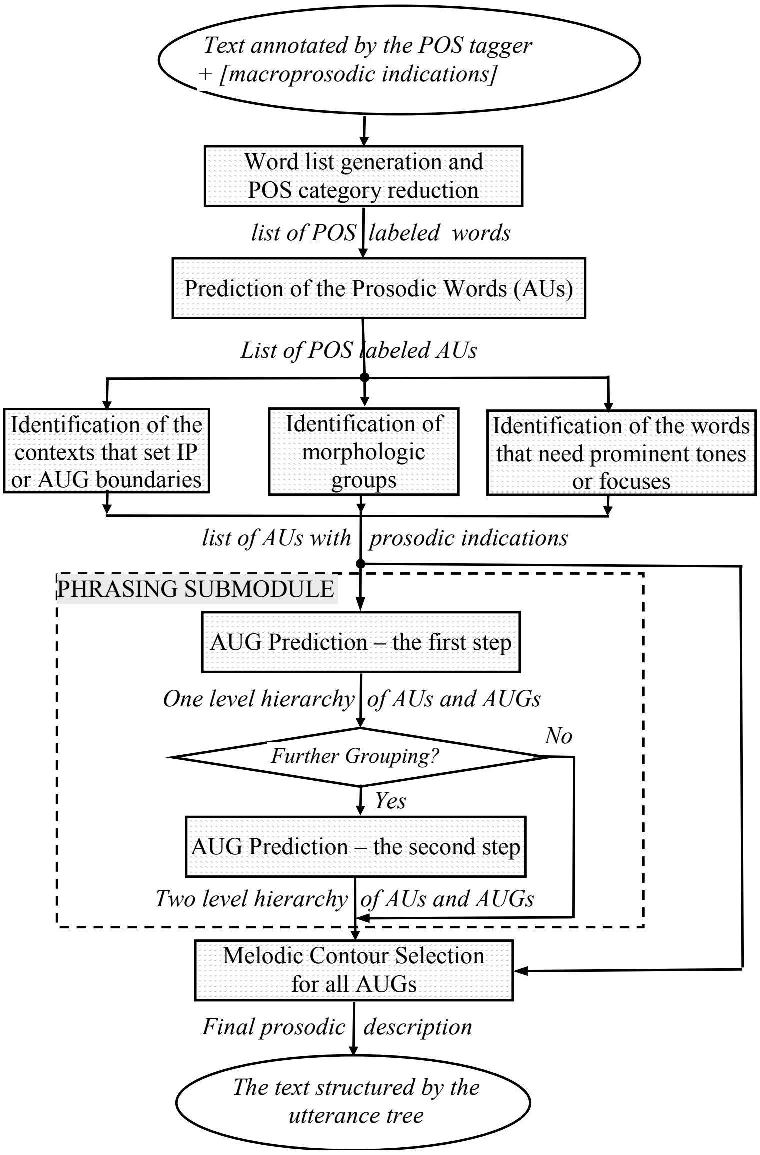

Text annotated by the POS tagger
+ [macroprosodic indications]
Word list generation and POS category reduction
list of POS labeled words
Prediction of the Prosodic Words (AUs)
List of POS labeled AUs
Identification of the contexts that set IP or AUG boundaries
Identification of morphologic groups
Identification of the words that need prominent tones or focuses
list of AUs with prosodic indications
PHRASING SUBMODULE
AUG Prediction – the first step
One level hierarchy of AUs and AUGs
Further Grouping?
No
Yes
AUG Prediction – the second step
Two level hierarchy of AUs and AUGs
Melodic Contour Selection for all AUGs
Final prosodic description
The text structured by the utterance tree

Fig. 3. The PPM data processing flow.

After the phrasing steps are performed and the top part of the utterance tree is built, the PPM assigns melodic contour descriptions to all its AUGs. These descriptions consist in functional label sequences corresponding to their child prosodic unit sequences. The melodic contours are selected from a dictionary by taking into account the prosodic indications deduced from the text analysis.

4. THE MELODIC CONTOUR SELECTION MODULE

The MCS module, depicted in Figure 4, processes the utterance tree generated by the Phrasing Module in a top-down manner and converts it into a sequence of groups corresponding to the nodes of the tree. For each group, the Key Generator submodule calculates a key of variable length, equal to the number of AUs. The Key Generator submodule encodes the specifications referring to:

- the position of the AU that carries the focus of a group;
- the position of the AU that carries the PUSH event of a group;
- subgroups of the group derived from the syntagmatic relation of the corresponding words or of the morphologic subgroups;
- the presence of a punctuation mark at the end of the corresponding text.

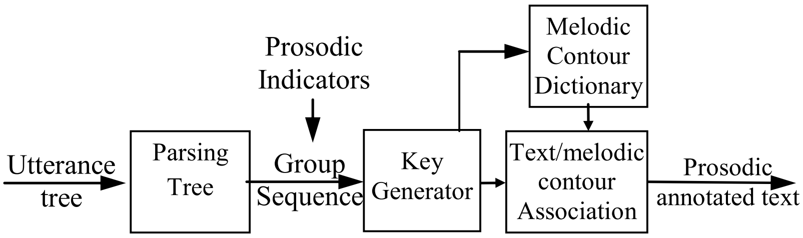


Fig. 4. The block diagram of the Melodic Contour Selection Module.

Using the key computed for each group, a melodic contour meeting all the requirements derived from the text is extracted from the MCD. Then, the "Text-melodic contour Association" submodule assigns functional labels from the MCD output sequence to the text units of each group. This submodule provides the PPM output consisting in a text structured by the utterance tree and labeled by Ro-ToBI and functional labels.

Several examples of melodic contours for groups containing two, three or four elementary units are presented in Table 2. Within a sequence, the functional labels are separated by "/". In contour descriptions, the functional labels corresponding to the prosodic words are accompanied or not by an explicit specification of the pitch accent type. For the former case, the RO-ToBI label is separated from the functional one by the "–" sign. For example, for the PO label, the low pitch accent (L*) can be considered as implicit. Instead, a high type accent can be explicitly specified, sometimes accompanied by a boundary tone (e.g. H*L-).

When a melodic contour contains an inner group, the labels are structured by rounded brackets for a single level of subgrouping or by rounded and square brackets when two lower levels are present. The functional label assigned to a group is attached as an index and indicates the role of the prosodic group at the communication act level, and, if necessary, the presence of a focus event.

Table 2
Examples of partial melodic contours

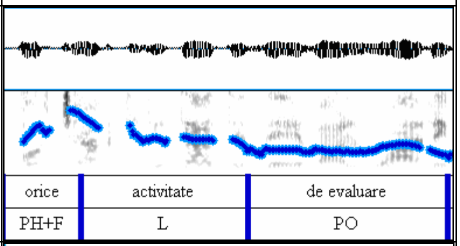
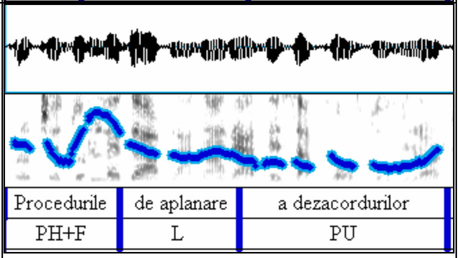
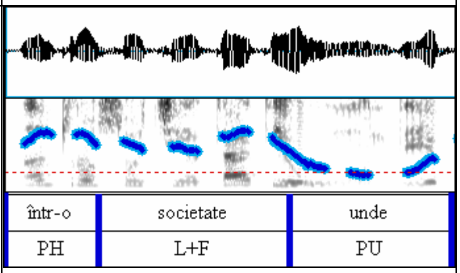
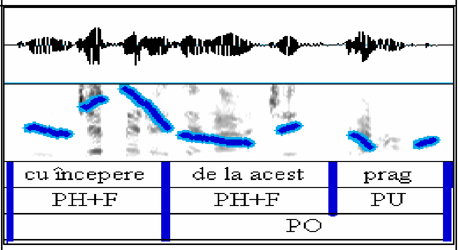
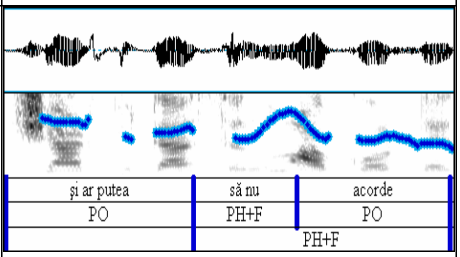
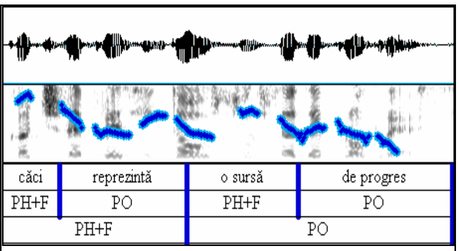
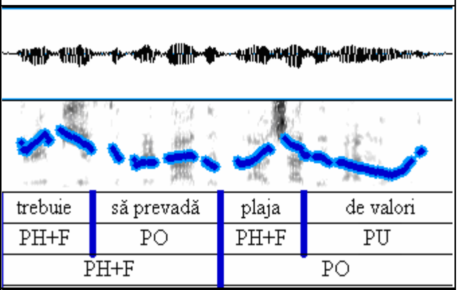
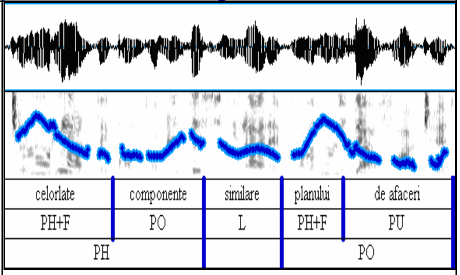
Label sequence	**Descriptions**	**F0 pattern**
PH+F /L-H*/ PO	is a F0 contour corresponding to a three AU group with a downstepping tendency and having the initial position within the parent phrase. It applies the tonal maximum to the accented syllable of the AU in the initial position that generates both a PUSH and a focus event. The melodic contour of the group ends in a POP event with a !H* pitch accent.	
PH+F /L-H*/ PU	is a F0 contour corresponding to a three AU group with a downstepping tendency and having the initial position within the parent phrase. It applies the tonal maximum to the accented syllable of the AU in the initial position that generates both a PUSH and a focus event. The melodic contour of the group ends in a POP-UP event that suggests the parent phrase continuation.	
PH /L+F/ PU	is a F0 contour corresponding to a three AU group in a non-final position of the parent phrase. It applies the tonal maximum to the accented syllable of the AU in the medial position, and generates the focus of the phrase here. The AU in the initial position generates a PUSH event. The melodic contour of the group ends in a POP-UP event that suggests the parent phrase continuation.	
PH+F/ (PH+F/PU)$_{PO}$	is a F0 contour corresponding to a three AU group that marks the morphologic subgroup of the last two AUs at the prosodic level. The contour spreading over the last two AUs appears in the low part of the tonal space of the group, leaving the first AU in the top part of the tonal space (generating a PH+F event). A PH+F event at the subgroup level is produced by the second AU. The melodic contour of the group ends in a POP-UP event that suggests the parent phrase continuation.	
PO/ (PH+F/PO)$_{PH+F}$	is a F0 contour corresponding to a three AU group that marks the morphological subgroup of the last two AUs at the prosodic level. The contour of the AU in the medial position reaches the maximal tonal level and focuses the corresponding word. The first AU raises the F0 contour during its accented syllable to an intermediate high level, leading to a POP event. The melodic contour of the group ends in a POP-UP event that suggests the parent phrase continuation.	

Table 2 continous

(PH+F/PO)$_{PH+F}$ (PH+F/PO)$_{PO}$	is a F0 contour corresponding to a four AU group that marks, at the prosodic level, the morphological subgroups of the first two AUs (conjunction+verb) and of the last two AUs (noun+noun). The reset F0 contour between the second and the third AU marks the beginning of the second subgroup. The contour generates a local focus on the third AU and the global focus on the first AU. This F0 contour corresponds to the final group within the parent phrase (prior to the dot).	
(PH+F/PO)$_{PH+F}$ (PH+F/PU)$_{PO}$	is a F0 contour corresponding to a four AU group that marks, at the prosodic level, the morphological subgroups of the first two AUs (verb+verb) and of the last two AUs (noun+noun). The reset F0 contour between the second and the third AU marks the beginning of the second subgroup. The contour generates a local focus on the third AU and the global focus on the first AU. The melodic contour of the group ends in a POP-UP event that suggests the parent phrase continuation	
(PH+F/PO)$_{PH+F}$ /L/ (PH+F/PU)$_{PO}$	is a F0 contour corresponding to a five AU group that marks, at the prosodic level, the morphological subgroups of the first two AUs (adjective+noun) and of the last two AUs (noun+noun). The reset F0 contour at the beginning of the fourth AU marks the beginning of the second subgroup. The F0 contour generates a local focus on the fourth AU and the global focus on the first AU. The third AU is a link between two subgroups. The melodic contour of the group ends in a POP-UP event that suggests the parent phrase continuation.	

The Phonetic Module of the TtS system will convert these descriptions into sequences of coordinates and F0 patterns in the tonal space of the synthesized utterance, generating the F0 contour.

The data processing in each block of the diagram is illustrated by the case study presented in the next section.

5. APPLAYING PROSODY. A CASE STUDY

This section illustrates the data processing flow within the PPM, by using the Romanian text: ”*Dacă lucrurile ar fi normale, atunci o comisie de evaluare colectivă, ar putea să depisteze un plagiat, şi ar putea să nu acorde dreptul solicitat vinovatului.*” (*If things were normal, then a commission of collective evaluation, could detect plagiarism, and might not grant the requested right to the guilty* *person*) as an input. This text was first edited in the POS tagger input window, available at [10]. The POS tagger generated the output in Figure 5, where the POS labels are marked by characters in bold.

```
Dacă|dacă|C|Csssp lucrurile|lucru|NPRY|Ncfpry ar|avea|VA3|Va--3 fi|fi|VA|Vanp
normale|normal|APN|Afpfp-n ,|,|COMMA|COMMA atunci|atunci|R|Rgp o|un|TSR|Tifsr
comisie|comisie|NSRN|Ncfsrn de|de|S|Spsa evaluare|evaluare|NSRN|Ncfsrn
colectivă|colectiv|ASN|Afpfsrn ,|,|COMMA|COMMA ar|avea|VA3|Va--3
putea|putea|VN|Vmnp să|să|QS|Qs depisteze|depista|V3|Vmsp3 un|un|TSR|Timsr
plagiat|plagiat|NSN|Ncms-n ,|,|COMMA|COMMA şi|sine|PXD|Px3--d--------w
ar|avea|VA3|Va--3 putea|putea|VN|Vmnp să|să|QS|Qs nu|nu|QZ|Qz
acorde|acorda|V3|Vmsp3 dreptul|drept|NSRY|Ncmsry solicitat|solicitat|ASN|Afpms-n
vinovatului|vinovată|NSOY|Ncmsoy .|.|PERIOD|PERIOD
```

Fig. 5. A Romanian phrase annotated by the POS tagger [10].

The accented syllables have been marked manually in the input window of the TtS system. The first step of the PPM data processing flow is a pre-processing step that consists in word extraction from the input window and in building the word list. The data structure assigned to each word contains attributes related to the POS category label and, possibly, to macroprosodic indications (beginning group, ending group, functional label group).

The first step of the prosodic prediction consists in converting the word list into a prosodic word (AU) list that will be presented in the next subsection.

5.1. THE PROSODIC WORD PREDICTION

The accentual units are elementary pitch contour segments that contain a single pitch event of pitch accent type. They generally correspond to one accented word. Structuring the text into accentual units implies assigning each clitic word (e.g. prepositions, conjunctions, articles, etc.) to one accented word that generates one accentual unit.

The accentual unit prediction is based on a consistent set of rules of the following types:

{left_POS_context, POS_sequence, right_POS_context, POS_label_Out} (1)

A rule is verified by a POS context, then an accentual unit object is created, which stores the text corresponding to the POS_sequence of the rule. The text of an AU is divided into syllables. At this data processing stage, a reduction of the POS label number is performed. The POS label from the *POS_label_Out* field of the rules is applied to the new AU. The POS labels at the AU level are taken into account by the Prosodic Indication Extraction module (Fig. 4).

An example of an AU prediction rule is presented in (2). The rule identifies the *reflexive pronoun/ auxiliary verb/ past participle verb* POS sequence and generates an AU with a "V" predicative verb label.

```
{Anything, _T("PxyVaVmp"), Anything, _T("V")} }    (2)
```

After processing the word list corresponding to the text in Figure 5, the PPM generates a list of 18 AUs, presented in Table 3.

Table 3
The AU list corresponding to the input text in Figure 5

AU	Text	AU	Text
AU1	dácă	AU10	să depistéze
AU2	lúcrurile	AU11	ún
AU3	ar fí	AU12	plagiát,
AU4	normále,	AU13	şi ar puteá
AU5	atúnci	AU14	să nú
AU6	o comísie	AU15	acórde
AU7	de evaluáre	AU16	dréptul
AU8	colectívă,	AU17	Solicitát
AU9	ar puteá	AU18	vinovátului.

The AUs are assigned to the leaf nodes of the utterance tree. These nodes are generated by the Phrasing Prediction processing block (Fig. 4) that will be presented in the next section.

5.2. THE PHRASING PREDICTION

The first phrasing step aims to generate reasonable lengths for the breathiness periods. The punctuation marks and a set of words representing markers for the IP beginning are taken into account for AU grouping. A non-final list of AUG instances is generated at the output. In a general case, the AUG list may contain ungrouped AUs. Each AUG instance keeps a link to the list of the AU children.

In the case of the text in Figure 5, the first phrasing prediction step detects two IP beginning markers, corresponding to the words *IF and THEN.* The second marker is used to force the AUG2 beginning. The second comma followed by the verb (*ar putea* – "could") represents the condition for the AUG3 beginning. The third comma followed by the conjunction "şi" ("and") marks the AUG4 beginning. The last six AUs are included in the last AUG. The output list after the first phrasing step contains four AUG instances (Fig. 6).

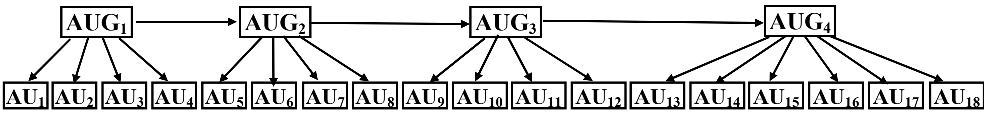


Fig. 6. The utterance tree after the first phrasing step.

If there are conditions for generating more than one IP for the synthesized utterance, then a second phrasing step processes the previous list of AUGs, to group them into several intonational phrases (IPs). For our example, three conditions for IP beginning are detected. Consequently, the phrasing submodule generates a list of four intonational phrases: IP1, IP2, IP3 and IP4.

The IP beginning markers refer to the following rules of the PPM:

- The clauses introduced by the markers *IF* and *THEN* are uttered in different IPs when the IF clause is longer than three words.
- An active verb preceded by a comma is a marker of an IP beginning.
- The conjunction "*şi*" preceded by a comma is a marker of an IP beginning.

Consequently, the AUG1-AUG4 become IP units (IP1-IP4). At this moment, the top part of the utterance tree is built (Fig. 7).

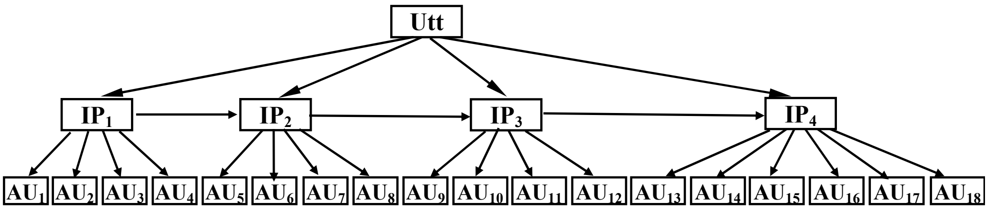


Fig. 7. The utterance tree after the second phrasing step.

When no condition for the IP beginning is detected, the phrasing submodule generates an implicit solution for IP grouping.

As Figure 7 illustrates, the IP1-IP4 are large groups of four or six AUs. The Melodic Contour Selection submodule can assign them melodic contours having one or two subgrouping levels, according to prosodic or macroprosodic indications.

5.3. THE MELODIC CONTOUR SELECTION

The MCS submodule performs a top-down processing of the utterance tree and assigns a melodic contour to each IP/AUG. The melodic contours are described by functional label sequences. The length of a sequence equals the length of the child list. The MCS selects the melodic contours from a dictionary (MCD), by calculating a search key based on prosodic indications. The key contains two characters (separated by "/") for each AU, where the melodic and phrasing constrains are coded. The phrasing constrains at the key level can introduce one or two phrasing levels.

The keys and the melodic contour descriptions extracted from the MCD and applied to IP1-IP4, are presented in Table 4.

Table 4

The keys and the asociated melodic contours of IP1-IP4

AUG ID	Key for the MCD	Melodic Contour Description
IP1	K1=MIN_ /_1/__/_21	M1=PO/(PH+F/L/PU)$_{PH}$
IP2	K2=MAX_ /_1/_1/_21	M2= PH+F/(PH+F/(PH/PU)$_{PO}$)$_{PO}$
IP3	K3=MAX_/_2/_1/_21	M3=(PH+F/PO)$_{PH}$/(PH/PU)$_{PO}$
IP4	K4=_1/MAX_1/_3/_1/__/_20	M4=(PO/(PH+F/PO)$_{PH+F}$)$_{PH+F}$/(PH+F/L/PO)$_{PO}$

The detection of the IF-THEN clauses from the text analysis leads to constrains to the IP2 and IP2 melodic contours. The implicit tones for the IP1/IP2 beginning (corresponding to the words "DACĂ" ("IF") and "ATUNCI" ("THEN"), respectively) are low for IP1 and high for IP2. These tonal constrains correspond to the MIN and MAX tonal information at the K1 and K2 key level (Table 4). The first phrasing level within M1 and M2 separates the key words "DACĂ" and "ATUNCI" from the rest of the text of the clauses introduced by "DACĂ" and "ATUNCI". The second phrasing level in M2 is required by the prosodic indications deduced from the text analysis, which mark the morphologic group noun+adjective ("*evaluare colectivă*"). The first prosodic group of two AUs in M3 corresponds to the *modal verb+ predicative verb* morphologic group. The second group results from rhythmic constrains, which accentuate the "un" article and the "plagiat" noun within two separate AUs. The high level prosodic group in M4 corresponds to a verbal group and a nominal group, respectively. The low level prosodic group results from the link between the negation word "nu" and the predicative verb.

By applying the utterance tree in Figure 7 at the MCS input, its depth increases by two levels since the melodic contours of the groups IP2 and IP4 introduce two lower levels. The final variant of the utterance tree is illustrated in Figure 8.

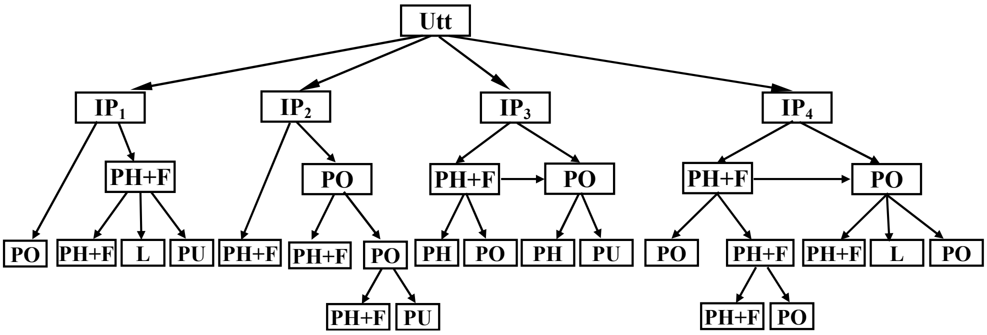


Fig. 8. The utterance tree after applying melodic contours at the group level.

The resulted utterance tree and the description of the melodic contour for each group assigned to a node represent the prosodic information at the output of the PPM module. This module structures the input text into prosodic units and annotates them with descriptions of their melodic contours. This information is applied to the Phonetic Module (Fig.1) that prepares the input (the phoneme sequence and the F0 contour) for the speech synthesizer.

5.4. DISCUSSION OF THE SPEECH SYNTHESIS RESULTS

In this section, the results of our speech synthesis, based on a PPM module, will be compared with natural F0 contours produced by different speakers uttering the same text. The F0 contours in Figure 9 are extracted from two utterances of the text used in the analyzed case study: "*Dacă lucrurile ar fi normale, atunci o comisie de evaluare colectivă, ar putea să depisteze un plagiat, şi ar putea să nu acorde dreptul solicitat vinovatului*." (T1).

The two F0 contours are divided into four segments that correspond to the four Intonatinal Phrases (IP1-IP4).

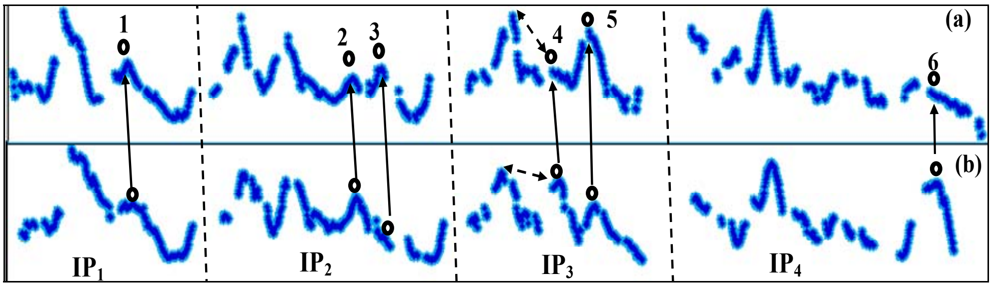


Fig. 9. The F0 contours extracted from two natural utterances of the T1 text.

The melodic descriptions of the IP1-IP4 contours in Figure 9 are presented in Table 5. The differences between them are underlined in the "Comments" column.

Table 5

The functional descriptions of the contours in Figure 9

Functional description	Comments
IP1-(a): PO/(PH+F/(PH+F/PU)$_{\mathrm{PO}}$)$_{\mathrm{PH+F}}$ IP1-(b): PO/(PH+F/L+f/PU) $_{\mathrm{PH+F}}$	The difference consists in the second small peak, present only in the first contour and marking the beginning of the last subgroup (Fig. 9 - point mark no.1)
IP2-(a): PH+F/(PH+F/(PH+F/PU)$_{\mathrm{PO}}$) $_{\mathrm{PO}}$ IP2-(b): PH+F/(PH+F/PO)$_{\mathrm{PH+F}}$ /PU)$_{\mathrm{PO}}$	The prosodic subgroups related to the nominal group are different. PH+F/PU groups the last two words (*"evaluare colectivă*") in (a). PH+F/PO groups the first two words ("*o comisie de evaluare*") in (b). (Fig. 9 - point marks no. 2, 3).
IP3-(a): (PH+F/PO-L*)$_{\mathrm{PH+F}}$/(PH+F/PU)$_{\mathrm{PO}}$ IP3-(b): (PH+F/PO-H*)$_{\mathrm{PH+F}}$/(PH+F/PU)$_{\mathrm{PO}}$	The groups in (a) and (b) display a significant tonal difference and a small tonal difference, respectively. In Fig. 9, this difference is marked by a dashed line between the target tones (Fig. 9 - point marks no.4, 5).

IP4-(a):
(PO/(PH+F/PO) $_{\mathrm{PH+F)}}$ $_{\mathrm{PH+F}}$ /(PH+F/L/PO)$_{\mathrm{PO}}$
IP4-(b):
(PO/(PH+F/PO) $_{\mathrm{PH+F)}}$ $_{\mathrm{PH+F}}$ /(PH+F/PO)$_{\mathrm{PO}}$/PD
	Here, the difference is generated by the AU represented by the "*vinovatului*" word, which is subordinated within the last prosodic group in (a) and within the first prosodic group related to a verbal group in (b).

In 5.1–5.3 subsections, we have presented the prosodic aspects deduced by the PPM module when performing a morpho-syntactic analysis of the input text, including the resulted utterance tree structure and the melodic contour description at each prosodic group level. Based on the output of the PPM module, the Phonetic Module generates the F0 contour for the speech synthesizer. In Figure 10, the F0 contour of the synthesized utterance of the T1 text (b) is compared with its corresponding natural F0 contour (a).

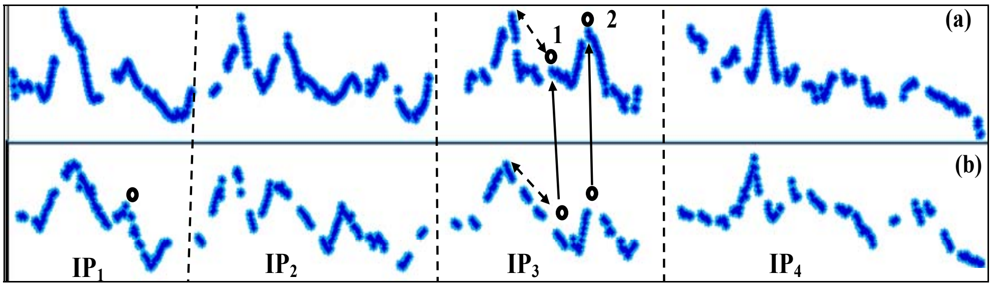


Fig. 10. The F0 contours extracted from the natural and synthesized utterances of the T1 text (a-natural contour; b-synthesized contour).

The comparison between the two F0 contours, based on the prosodic events used in [3], leads to the conclusion that they have the same IP breaks, pitch accents and boundary tones. The perception test also sustains this visual observation. Some dissimilarities can still be perceived. They are related to the wideness of the overall pith range, the tonal differences between the targets of a group, the lengths of the accented vowels and the F0 patterns at the accentual unit level.

In Figure 10, making the synthesized F0 contour closer to the natural one, would require:

- to change the prosodic grouping rules corresponding to a nominal group of noun+ noun+ adjective type, from noun+ (noun+ adjective) to (noun+ noun)+ adjective;
- to introduce a rule to subordinate the nominal group of the direct complement to the verbal group;
- to introduce a semantic submodule in the PPM, to establish the subordination of the "*vinovatului*" word to the verbal group.

The PPM implements a set of rules leading to an implicit prosodic variant. In the future, a set of macroprosodic markers should be developed for helping the prosodic prediction module to generate other intonational variants.

6. CONCLUSIONS

Our intonation model can be a good starting point for prosodic prediction. It offers a perspective on complex F0 contours by converting each pitch contour into a hierarchy of partial melodic contours. The complex problem of the F0 contour prediction is decomposed into two subproblems: the building of the utterance tree and the partial melodic contour selection at each group level. Both prediction subproblems are solved by using prosodic indications deduced by rules that evaluate different lexical and morphological contexts. The prediction results depend on the set of rules at each submodule level.

When no prosodic indication is detected, the predictor outputs implicit solutions for phrasing and for melodic contour selection. Our implementation of intonation prediction creates a framework for integrating a melodic contour dictionary and decision modules for phrasing generation and for the melodic contour selection at the group level. The implicit neutral intonational contours provided by the prediction module have generated good results in perception tests at level of the IP breaks and boundary tones. The set of rules for each submodule can be further increased, leading to an improved prediction.

Acknowledgments: The research presented in this paper was developed within the Institute of Computer Science of Romanian Academy research program

Authors' contributions: D. Jitcă and V. Apopei designed the intonation model and implemented the PPM in the TtS system; D. Jitcă and O. Paduraru built the speech database and wrote the paper.

REFERENCES

1. APOPEI V., JITCĂ D., *Module for generating the F0 Contour using as input a Text structured by prosodic information*, in *Advances in Spoken Language Technology* (C. Burileanu and H. N. Teodorescu, Eds.), Publishing House of the Romanian Academy, Bucharest, 2007, 119-126.
2. BLACK A., TAYLOR P., *Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input*, Proceedings of the 3rd Int. Conf. on Spoken Language Processing, Yokohama, Japan,1994., 715–718.
3. BULYKO I., OSTENDORF M., *A bootstrapping approach to automating prosodic annotation for limited-domain synthesis,* Proceedings of the IEEE Workshop on Speech Synthesis, Santa Monica, 2002,115 - 118.
4. JITCĂ D., APOPEI V., JITCĂ M., *The F0 Contour Modeling as Functional Accentual Unit Sequences*, International Journal of Speech Technology, 2009, **12**, *2*, 75-82.
5. JITCĂ D., APOPEI V., JITCĂ M., *How can a functional perspective be used in intonation modeling*, International Conference Speech Prosody, Chicago, SUA, May 11-14, 2010.
6. JITCĂ D., APOPEI V., *An intonation prediction module for Romanian TTS system, as a prosodic tree generato*r, Proceedings of the Speech Technology and Human-Computer Dialogue (SpeD), May 18-21, 2011.
7. JITCĂ D., APOPEI V., PADURARU O., *Transcription of Romanian Intonation-RoToBI*, www.etc.tuiasi.ro/sibm/romanian_spoken_language/RoToBi/RoToBi_System.htm
8. LADD D. R., *Intonational phonology*, Cambridge University Press, 1996.
9. QUAZZA S., DONETTI L., MOISA L., SALZA P. L., *Actor: a multilingual unit-selection speech synthesis system,* Proceedings of the 4th ISCA, 2001.
10. ⁂ Research Institute for Artificial Intelligence of Romanian Academy, *Text Processing by Xml Web Services*, www.racai.ro/webservices/TextProcessing.aspx

Received April 4, 2012