MODULE FOR GENERATING THE F0 CONTOUR USING AS INPUT A TEXT STRUCTURED BY PROSODIC INFORMATION

Vasile APOPEI, Doina JITCĂ

Institute for Computer Science, Romanian Academy – Iaşi Branch

Corresponding author: Vasile APOPEI,
vapopei@iit.tuiasi.ro

The paper presents a prosodic implementation for Romanian speech synthesis by applying a four-level intonational hierarchy model [7]. Using this hierarchy it follows good descriptions of intonational contour meaning. The input in this application consists in a text structured by an intonational unit tree, based on this hierarchy. The prosodic implementation involved the design of a rule-based submodule for F0 contour generation. The synthesised F0 contour and the phoneme sequence are directed to the input of a Klatt type synthesiser.In this paper, the intonational unit hierarchy is described by defining each intonational unit level. After a presentation of some F0 contour patterns corresponding to different types of phrasal events by which the phrasal units are delimited, a case study is analyzed in order to exemplify the results in the F0 frequency contour synthesis based on our intonation modelling. This application represents an improved implementation of intonation compared to that presented in [5], which is based on a two-level prosodic hierarchy only.

Key words: Intonational hierarchy, utterance tree, phrasal event, F0 Contour, speech synthesis.

1. INTRODUCTION

The main modules involved in the implementation of prosody in speech synthesis are the following: the linguistic processing module, the phonetic module, and the speech synthesizer module. They are illustrated by internal components in the block diagram of figure 1.

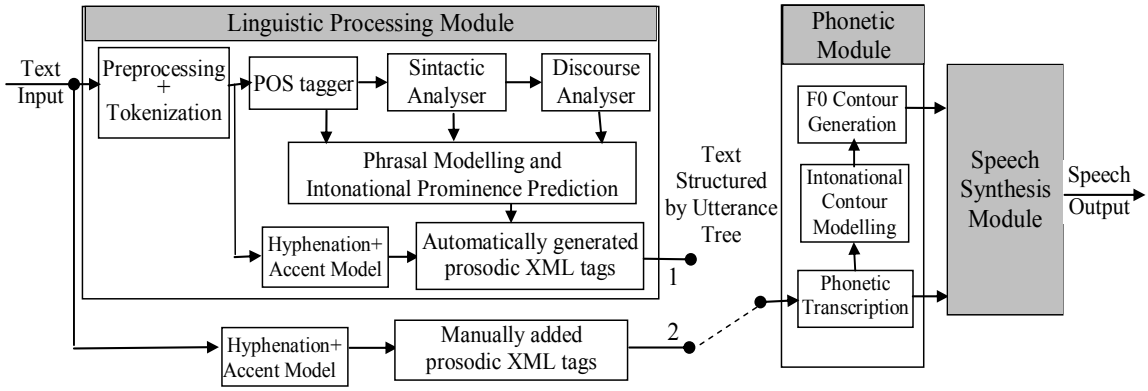

Figure 1. A block diagram of prosodic speech synthesis.

The output of the first module consists in a hyphenated text, with the syllables grouped in a tree structure of intonational units. The structuring must be based on a defined multi-level hierarchy of intonational units and an implementation of a phrasing submodule. The intonational phrasing submodule uses as input the text annotated and structured by a POS tagger, a syntactic analyser and a discourse parser. We agree with the approach in [1] that phrasing modelling and prominence prediction are related and they can be thought of as a single submodule responsible for an automatic generation of a prosodic description

corresponding to the input text. Based on the output of this submodule, the syllable sequence generated by the hyphenation submodule is structured according to the utterance tree and it represents the input of the phonetic module.

In paper [3], the linguistic module from figure 1 is divided into a *linguistic front-end* responsible for parsing the input text into linguistic structures, and a *prosodic front-end* responsible for generating the linguistic prosodic structure.

Up to this stage, we did not implement a linguistic processing module in our speech synthesizer and, in this case, the input to the phonetic module is generated manually. This module consists in text structured by an intonational unit tree characterised by intonational prominence information (pitch accent type, duration, energy). We are interested in using the results of Romanian natural language parsing for an automatically phrasing and prominence prediction from the Romanian text.

The *Phonetic Transcription* submodule performs the syllable sequences and generates the utterance phoneme sequence. Each phonetic element of the output sequence has a duration attribute and the resulted time series is directed to the speech synthesizer module. To implement the intonation, a parallel time series have to be generated, containing the corresponding tone values of the F0 contour.

The submodule *Intonational Contour Modelling* translates the intonational unit structure, from the XML input file, into a sequence of F0 contour events. The submodule *F0 Contour Generation* selects a corresponding F0 pattern for each event from the sequence of the previous step accordingly to the tonal indications from the XML input file and the phonetic context. The selected patterns are used to generate the most significant segments of the synthesized F0 contour. The segments resulted are linked by linear interpolation.

The phoneme sequence synchronised with the synthesised F0 contour is the input to the speech synthesiser module.

The paper presents a prosodic implementation for Romanian speech synthesis corresponding to the second variant illustrated in figure 1. This application provides an improved implementation of some intonation elements, compared to that presented in [5]. The previous version was based on a two-level prosodic hierarchy, consisting of the *intonational phrase* level and a sublevel called *intermediate phrase*.

Our intention for creating a Romanian text corpus with prosodic annotation led us to the conclusion that a two-level prosodic hierarchy is not adequate enough to describe, at phrasal level, the variety of F0 contours extracted from natural utterances. For this purpose, we defined a *four-level intonational unit hierarchy* by dividing the intermediate phrase into one or multi-rhythmical units, each of them composed by one or multi-accentual units. Based on this hierarchy, an XML schema was defined and used to structure the input text in our speech synthesis [6], [7].

The present paper expresses the results of speech synthesis by applying the new prosodic hierarchy in intonation modelling. The proposed four-level hierarchy is shown to describe better the meaning of intonational contours.

In the second section of the paper, we present our intonational unit hierarchy defining each intonational unit level. In the third section we present some considerations about the F0 contour generation. The fourth section encloses a case study to exemplify the results in the F0 frequency contour synthesis based on this intonation modelling.

2. INTONATIONAL UNIT HIERARCHY

The perspective of our intonational model is generated by the autosegmental-metrical theory of intonation and it is based on a prosodic hierarchy in [7], presented in figure 2.

As it can be seen in figure 2, the intonational units marked in our prosody annotation are the following: Intonation Phrases (IPs), intermediate phrases (ip_s), Rhythmical Units (RUs), and Accentual Units (AUs). The words and syllables that compose this phrasal unit are used for setting the positions of the prominence attributes. The syllable is the smallest unit within this hierarchy, to which we can assign target tones in the F0 contour.

The smallest intonational unit with a melodic contour is the accentual unit (AU). The AU can contain one stressed or accented word and eventually other clitic words. In certain cases, an AU can contain two nonclitic words, when one of them looses the stress in the neighbourhood of the other.

Within an AU, the F0 contour pattern has a tonal curve during the unaccented syllables that precede the accented syllable, a pattern curve for the stress or accented syllable, and a tonal curve during the unaccented

syllables that follows the accented syllable. The tonal variation along the three curves generates the melodic contour of an AU, and a significant contrast between accented and non-accented syllables.

The difference between this account and that in paper [5] consists essentially in adding two new intonational unit levels under the intermediate phrase level.

The unit of intermediate phrase type corresponds to the Pierrehumbert's [8] definition related to the down-stepping trends, consisting of decreasing the target tones of accent pitch events in the F0 contour. If the reset of this trend occurs and the final of intonational phrase is not reached, then an intermediate phrase end is assigned.

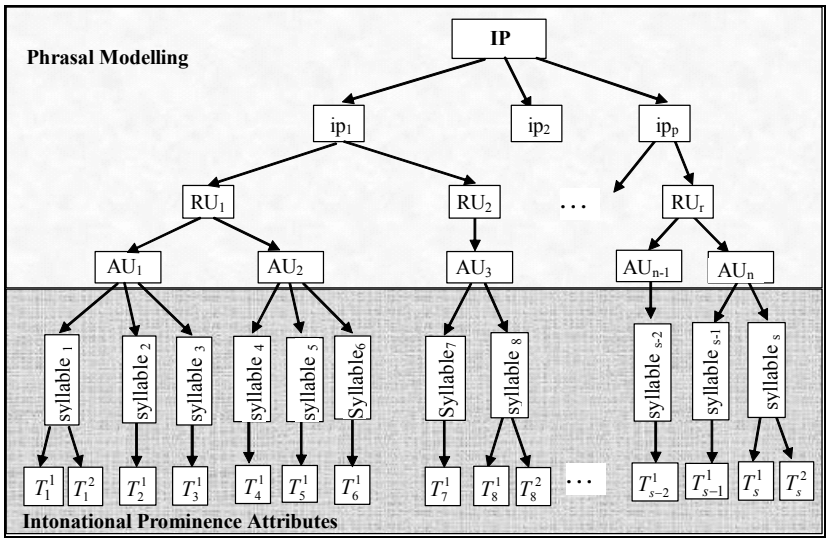

Figure 2. Intonational Unit Hierarchy.

We noticed downstepping and upstepping trends in natural F0 contours. A short break and an accent phrase tone were introduced in ToBI annotation at the end of an intermediate phrase (Annex). The intonational phrases are ending with boundary tones and they are followed by long breaks (Annex).

The changes of these tendencies in F0 contour are called *F0 resets* and several pitch event patterns can be assigned to them. In some cases, the F0 reset are not visible by prominent tonal events, but the change in slope of downstepping/upstepping tendency is present.

The model considers that an intonational unit consists of at most one intermediate phrase. In certain cases, within an intermediate phrase containing several words, one observes a grouping of them implied by phrasing or prominence reasons. In respect to the intermediate phrase definition, if a change in downsteping/upstepping tendency doesn't appear, we have to introduce a sublevel of the intermediate phrase. We called it the *level of rhythmical unit* [7]. Mapping rhythmical units into an F0 contour pattern space, we found some tonal event type combinations to describe them. A rhythmical formula often met in reading aloud of a text, in affirmative sentences, can be described using ToBI labels by the tone sequences in the relation (1)

The significance of the labels used in this paper to describe the intonational contour is presented in the Appendix.

$$[H^* \ L+!H^*] \ , \ [H^* \ L^*+!H] \tag{1}$$

These sequences are rhythmical units composed by two Accentual Units (AUs). Within the first AU, a monotonal pitch accent of high H* type occurs and, in the second, a bitonal pitch accent occurs with the target High tone at lower level. Within rhythmical units, significant tonal contrasts occur as a combination of differences in the mean tonal level of accentual units by which they a composed, differences between principal accent prominence, and/or differences between principal accent types. For the case of previous sequences, the contrast is generated by the tonal difference between the target of accents H* and !H*, weighted by the difference between a monotonal and a bitonal pich accent.

In paper [5], a case study was presented using the natural utterance of the text "*Avea sentimetul că mai fusese prin cartierul respectiv odată ...*" (He/She had the feeling that he/she already visited once the neighbourhood …) which is illustrated by the F0 contour in figure 3.

Using the perspective of two-level hierarchies, the F0 curve was interpreted as it is expressed in the variant (2). The novel perspective of the four-level hierarchy generates a new tree for this utterance described by the variant (3). In the text annotation of the variants (2) and (3), the slash separator was used to delimit the *AUs*, the round parenthesis to enclose the *RUs*, the square brackets to enclose the *ip,* and the brackets to enclose the *IP*.

{[Avea sentimetul]$_{ip}$ [că mai fusese] $_{ip}$ [prin cartierul] $_{ip}$ [respectiv] $_{ip}$ [odată] $_{ip}$} $_{IP}$ (2)

{[(Avea/sentimetul) $_{RU}$]$_{ip}$ [(că mai fusese/ prin cartierul)$_{RU}$ (respectiv/ odată) $_{RU}$] $_{ip}$}$_{IP}$ (3)

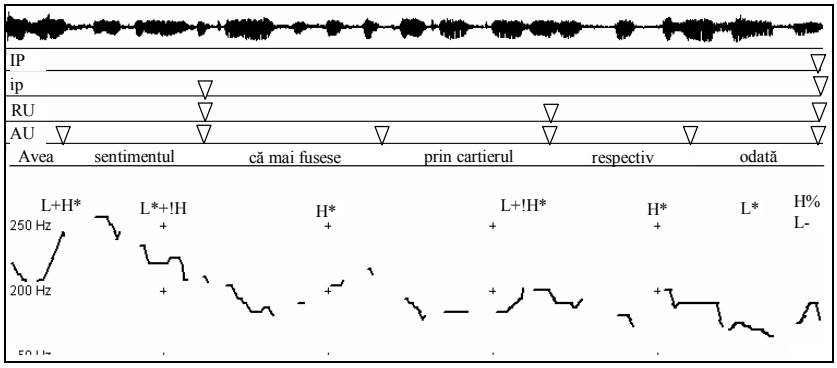

Figure 3. The F0 contour of the text utterance
"*Avea sentimentul că mai fusese prin cartierul respectiv odată …*"
(He/She had the feeling that he/she already visited once the neighbourhood …)

In figure 3, the F0 contour is annotated in the four-level intonational hierarchy perspective described by variant (2). In this interpretation, grouping the last four AUs in two RUs gives the possibility to relate the two pitch accents within each RU and to define some tonal patterns, characterised by small differences between target tones, and certain combinations of pitch accent type. The three levels above the accentual unit give the possibility to build the intonation for spoken structured phrases.

All intonational models have at most two levels: intonational and intermediate phrase levels. We consider the accentual unit level to be an implicit one, because all nonclitic words, alone or, eventually, together with some clitics, must be accented or, simply, stressed. This creates in the F0 contour, a low level F0 contour pattern unit. The target points in AU pattern are mapped into the frequency range, taking into account its position in the upper level intonational units. Introducing the rhythmical unit level, we can substructure an intermediate phrase that keeps unchanged the linebase trend into smaller units, with certain pattern characteristics for which linguistic equivalences can be found. These reasons are important in generating the F0 contour for speech synthesis. A multilevel intonational hierarchy gives us the possibility to describe an intonational contour in terms of pattern units for low level events grouped into pattern units for higher level events.

3. CONSIDERATIONS ON THE F0 CONTOUR GENERATION

We split the implementation of F0 contour generation into two submodules that compose the *Phonetic module: Intonational Contour Modelling* and *F0 Contour Generation* (figure 1). Their design was based on a hand-crafted rule set and an F0 contour pattern set resulted from a text corpus annotated at prosodic information. The submodule *Intonational Contour Modelling* translates the intonational unit structure, from the XML input file, into a sequence of events that have assigned a "point" in the F0 contour. The "points" are characterized by a position within the F0 contour trends decided by rules for each IP/ip units from the input utterance tree. The establishment of these trends represents the first step in mapping the F0 contour into F0 frequency scale. The "points" are in fact the segments of F0 contour for which the submodule *Intonational Contour Modelling* assigns a proper frequency range, within the overall frequency range of the synthesized utterance, and a length on time scale, in relation with the position in intonational structure of the unit containing that event. The events are related to the corresponding Accentual Units within which they spread around the accented syllable, involving the previous and the next syllable too.

We found a systematic presentation of different melodic contours for Romanian intonation in paper [2], grouped into various intonational categories and subcategories.

The submodule *F0 Contour Generation* selects a corresponding F0 pattern for each event from the sequence of the previous step accordingly to the tonal indications from the XML input file and the phonetic context. The selected patterns are then mapped into the F0 frequency ranges deduced by submodule *Intonational Contour Modelling* and the most significant segments of the synthesized F0 contour are generated. During the next step in F0 contour generation, the segments resulted in the previous step are linked by linear interpolation.

The design of this submodule involves building an F0 contour event inventory (described by one or more tone labels) and, for each of these events, to assign several patterns corresponding to different phonetic contexts and accented syllable positions [4].

Our F0 pattern unit inventory contains two important categories assigned to the following events: events that appear in the beginning of phrasal units (IP, ip, RU), and events that appear in the end of phrasal units. We denoted these events as *phrasal events*. These events occur within an AU unit and they are speeded during the accented syllable and the previous or the next syllable can be involved too. In relation with the accented syllable position there exist different F0 patterns for events which occur on the first syllables of an AU and events which occur on the last syllables of an AU. The examples used in this paper are selected from the second category of F0 patterns.

A *phrasal event* can be described by a sequence of ToBI labels which include a pitch accent and other tones from the following category: phrase accent (L-, H-), boundary tones (%L, %H, L%, H%), and significant tones in previous or next position of accented syllables (H+, L+).

The phrasal event prominence is not identical to the prominence of the pitch accent that is included. The pitch accent prominence is related to presence/absence of the syntactic accent. The phrasal event prominence signifies the boundary prominence of phrasal units.

A phrasal unit must have at least one prominent event (in the beginning or in the end) but it may have prominent the both events.

We modelled a phrasal event, in the beginning position, which contains a pitch accent of "high" type, by four pitch segments (figure 4.a):

- a pitch variation segment during the previous syllable of the accented syllable (drawn with circled line) - *I segment;*
- an increasing step segment during a part of the accented syllable corresponding to unvoiced consonant if they exist within the accented syllable (drawn with dotted line)- *II segment;*
- a continuous pitch variation segment during the accented syllable (drawn with solid line)- *III segment;*
- a pitch variation segment or a pitch step segment during the next syllable of the accented syllable (drawn with circled line)- *IV segment*.

A particular pattern depends on the phonetic context and the prominence level of a phrasal event. Both determinants led to some ratios between the projection of these segments on time and frequency axis.

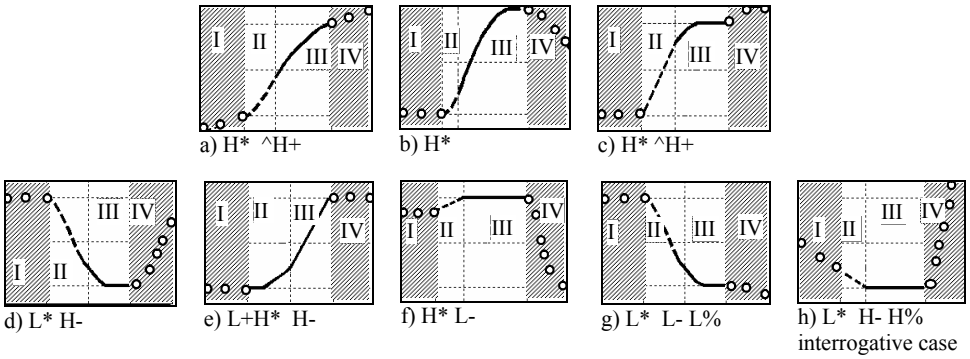

Figure 4. F0 contour patterns in the beginning and in the ending of a phrasal units.

A syntactic accent of H* type within a phrasal beginning event can be generated by increasing the frequency range of the *III segment* and decreasing the frequency range of the *II segment*. Furthermore, we change the *IV segment*, from an increasing one to a decreasing one since the maximum high tone of phrasal unit is reached during pitch accent (accented syllable) – figure 4.b.

Otherwise, for generating a beginning phrasal event with a less prominent pitch accent (no syntactic accent) we increase the frequency range of the *II segment* (if it exists), decreasing the frequency range of the

III segment, and allocate a range frequency to the *IV segment* in the increasing variant (figure 4.c). In these conditions, the phrasal event keeps its prominence as in the previous case but eliminate de syntactic accent.

The events in the end of a phrasal unit can be classified into events with high boundary tone and events with low boundary tone. The events in the first case can have a pitch accent of H* type (patterns are the same as those in the beginning of phrasal unit), or a pitch accent of L+H* type (figure 4.e), or a pitch accent of L* type (figure 4.d, h). In the second case, the patterns can include a pitch accent of high type (figure 4.f) before the final fall, or a pitch accent of L* type (figure 4.g).

In F0 pattern inventory we added to phrasal event patterns some other patterns corresponding to AUs that don't contain on accented syllable a pitch accent.

In the following section we illustrate by an example the results of F0 contour generation module of our implementation.

4. A CASE STUDY

We exemplify in our Romanian speech synthesis, the manner in which a natural intonational contour is related to a synthesised one corresponding to the sentence "*Winston îşi duse paharul la buze cu oarecare nerăbdare.*" ("Winston brings the glass to lips with some impatience."), chosen from the text corpus annotated with prosodic information [6], [7]. The F0 contour extracted from this natural utterance is illustrated in figure 5 and it is annotated using ToBI labels [9].

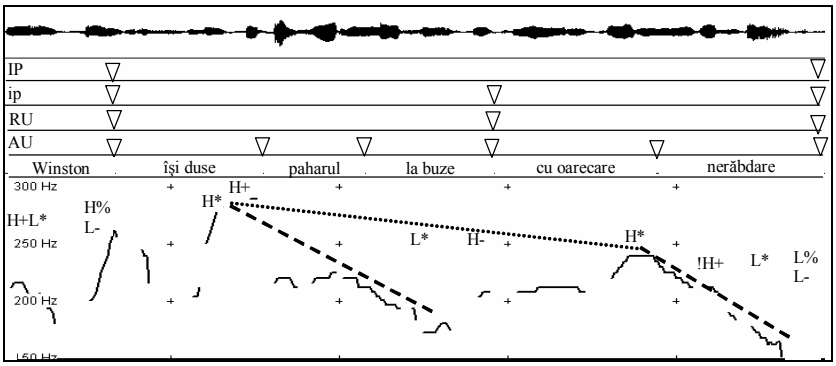

Figure 5. The F0 contour extracted from the natural utterance of the text "*Winston îşi duse paharul la buze cu oarecare nerăbdare.*" ("*Winston brings the glass to lips with some impatience.*").

One intonational variant of the text phrasing is described as follows:

$\{[\text{Winston}]_{ip}\}_{IP}\ \{[(\text{îşi duse/ paharul / la buze})_{RU}]_{ip}\ [(\text{ cu oarecare/ nerăbdare})_{RU}]_{ip}\ \}_{IP}$ (4)

The subject *"Winston"* is uttered within an intonational phrase finished with an H% boundary tone. This is pointed out by an H+L* accent type. Until the sentence ends, the words are accented within the following five accentual units: *îşi duse/ paharul/ la buze/ cu oarecare/ nerăbdare*. These AUs are grouped into two intermediate phrases separated by the rising tone (H-) on the last syllable of the word *"buze"*. Concerning the position of the two intermediate phrases within the intonational phrase (final/nonfinal), a different F0 contour pattern occurs: L* H- in the first case, L* L- in the second case. The F0 contour patterns of the accentual units in the beginning of the two intermediate phrases (/*îşi duse*/ and /*cu oarecare*/) are adequate to their specific phonetic context, but they both fix the higher target tone from which the downstepping trends begin. The downstepping trends are marked in figure 5 by dashed lines. These target tones decrease along downstepping trends at the intonational phrase level, marked in figure 5 by a doted line.

The text structured by prosodic tags in the XML file becomes input to the phonetic module. In this example, the content of the obtained XML file is listed in figure 6.

In this annotation, the tags <IP>, <ip>, <RU>, <AU> and <Syl> have been used to mark the corresponding intonational units: intonational phrase, intermediate phrase, rhythmical unit, accentual unit, and syllable. The indications for pitch accents, phrasal tones, and boundary tones are given using attributes in the corresponding tags [7].

```
<IP BeginToneLabel="%L" BoundaryToneLabel="H%">
  <ip PhraseToneLabel="L-">
    <RU><AU><Syl Accent="H+L*">win</Syl><Syl >ston </Syl></AU></RU>
  </ip>
 </IP>
 <IP BeginToneLabel="%L" BoundaryToneLabel="L%">
  <ip PhraseToneLabel="H-">
    <RU><AU><Syl >îşi</Syl ><Syl Accent="H*">du</Syl> <Syl TargetTone=" H+"> se</Syl></AU>
          <AU><Syl>pa</Syl><Syl Accent="Stress">ha</Syl><Syl>rul</Syl></AU>
          <AU><Syl>la</Syl><Syl Accent ="L*">bu</Syl><Syl>ze</Syl></AU>
    </RU>
  </ip>
  <ip PhraseToneLabel="L-">
     <RU><AU><Syl>cu</ Syl ><Syl>oa</ Syl ><Syl>re</ Syl >< Syl Accent="^H*"> ca</Syl>
                <Syl>re</Syl></AU>
          <AU ><Syl>ne</Syl><Syl TargetTone="!H+">răb</Syl><Syl Accent="L*">da</Syl>
                <Syl>re</Syl></AU>
     </RU>
  </ip>
 </IP>
```

Figure 6. The content of the prosodic XML file for the text "*Winston îşi duse paharul la buze cu oarecare nerăbdare.*" ("*Winston brings the glass to lips with some impatience.*").

Based on this XML prosodic file, the phonetic module generates the F0 contour illustrated in figure 7.

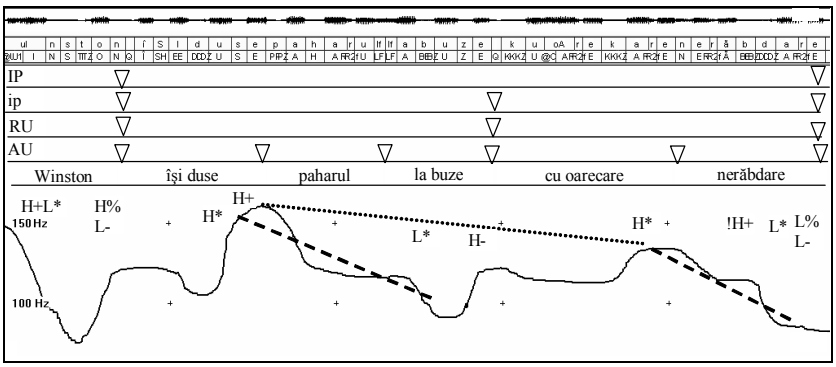

Figure 7. The F0 contour extracted from the synthesised utterance of the text "*Winston îşi duse paharul la buze cu oarecare nerăbdare*"

One can observe that all significant tones in the F0 contour in figure 5 are present in the synthesised F0 contour, too. The F0 frequency range in this type of synthesiser is situated between 70 and 150 Hz, corresponding to a male voice, distinct from the natural utterance that corresponds to a female voice. The F0 contour from figure 7 gives a good quality at the perception level to the synthesised wave.

5. CONCLUSIONS

Our perspective in introducing the intonation in speech synthesis can be expressed by the superposition of two components: a background one generated by the phrasal units structure, and a foreground one generated by the syntactic accent prominence and position. We can talk about a "neutral" F0 contour resulted from the phrasal event structure by respecting some rules deduced for phrasal event patterns and for their relative position into the F0 frequency scale. We can add prominent syntactic accents by modifying the phrasal event pattern and event phrasal unit structure.

In this stage of implementation of intonation in Romanian speech synthesis we focused our attention on building a set of rules to achieve the following tasks: to translate the intonational unit structure, from the

XML input file, into a sequence of events on the F0 contour trends corresponding to IP/ip unit structure, and to create a large event pattern inventory.

The experiments with our prosodic implementation module using four-level hierarchy intonation led us to the conclusion that it generates an improved intonation in relationship with the interpretation of the text from the XML input file, because the synthesised F0 contour follows better the events in the utterance.

We are interested in using the results of Romanian natural language parsing for an automatically phrasing and prominence prediction from the Romanian text, relating the intonational structures to the corresponding syntactic and semantic ones, in order to generate an adequate prosodic variant of the input text.

REFERENCES

1. Di CRISTO A., *La Prosodie au carrefour de la phonétique, de la Phonologie et l'articulation formes-functions,* Travaux Interdisciplinaires du Laboratoire Parole et Langage, **Vol.23**, pp. 63-211, 2004.
2. DASCĂLU-JINGA L., *Melodia vorbirii în limba română*, Editura Univers enciclopedic, Bucureşti, 2001.
3. VANIO M., SUNI A., SIRJOLA P., *Developing a Finnish Concept to Speech System*, Proceedings HLT 2005, pp. 201-206, 2005.
4. JILKA M., MÖHLER G., DOGIL G., *Rules for the Generation of ToBI-based American English Intonation*, Speech Communication 28, pp. 83 – 108, 1999.
5. APOPEI V., JITCA D., TEODORESCU H.N., *Implementation of Stress and Emotion Rendering Rules in Synthesized Speech,* Proceedings of the 3rd Conference Speech Technology and Human-Computer-Dialogue "SpeD 2005", pp.67-72, 2005.
6. APOPEI V., JITCA D., *A Set of Intonational Category for Romanian Speech and Text Annotation,* European Conference on Intelligent Systems and Technologies, Proceedings ECIT2006 , pp. 117-124, 2006.
7. APOPEI V., JITCA D., *Schema XML de adnotare a intonației în cadrul corpusurilor de text,* "Lucrările atelierului „Resurse lingvistice şi instrumente pentru prelucrarea limbii române", pp 9-14, 2006.
8. BECKMAN M., PIERREHUMBERT J., *Intonational Structure in Japanese and English*, Phonology Yearbook III, pp. 15-70, 1986.
9. BECKMAN M., AYERS G., *Guidelines for ToBI Labelling*, ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf, 1997.

APPENDIX

Our annotation of intonation follows the conventions of ToBI system [9]. We used the following labels grouped into four categories:

- Pitch accent labels corresponds to certain accentuated syllables:
  - L* - a monotonal pitch accent, a movement to a low target tone;
  - H* - a monotonal pitch accent, a movement to a high target tone;
  - L+H* - a bitonal pitch accent, a movement from a low tone in the beginning of the accented syllable, to a high target tone during its voiced part;
  - L*+H - a bitonal pitch accent, a movement to a low tone during the first half of the accented syllable, followed by a movement to a high target tone during the last voiced part.
- Phrase accent labels for marking the final tones of intermediate phrases (ip): L-, H-.
- Boundary phrase labels for marking the beginning and the final tones of intonational phrases (IP): %L, %H, L%, H%.
- Labels for marking the significant target tones in previous or next position of accented syllables: L+, H+.

The high tone labels can be altered by the signs " ! " and " ^ " to mark the present tone as being lower and, respectively, higher than the preceding high tone.