A description language at the accentual unit level for Romanian intonation

Doina Jitcă [1], Vasile Apopei[1], Magdalena Jitcă [2]

[1] Institute for Computer Science, Romanian Academy, Iasi Branch, Romania
[2] University "Alexandru Ioan Cuza", Iasi, Romania
jdoina@iit.tuiasi.ro, vapopei@iit.tuiasi.ro, magdalena.jitca@infoiasi.ro

Abstract

The paper presents a classification of accentual unit patterns (AU patterns) and a corresponding label set used for generating an AU label based description language of intonation. Our AU patterns classification performed over a Romanian speech corpus, is based on the consideration that each intonational phrase corresponds to a basic discourse unit (BDU) or a subunit of BDUs. Therefore, we assign to each AU category a function in the spoken discourse. The description of the F0 contour by AU labels is suited for a text-to-speech system to create a description language of intonation for building the output of the linguistic module. It structures the input text into intonational units including the F0 contour characterizations as attributes. The structured text will be used as input for the phonetic module that generates the F0 contour for the synthesizer.
Index Terms: F0 contour, AU patterns, discourse events, speech synthesis

1. Introduction

In paper [1] was presented a solution for an F0 contour generating module of a Romanian TtS. The module has used an XML input text file, manually generated, with the text structured by intonational information using ToBI annotation labels. The intonational description is loosely related to the F0 contour patterns and it generates the following difficulties: at the linguistic module, in automatically generating correlations between the prosodic information and the syntactic structures, and at the phonetic modules, in translating the prosodic description into pattern segments within the F0 scale.
The present paper proposes an intonational description language based on labels assigned to accentual units (stress units). In order to achieve this goal the accentual unit was considered to be the basic pattern unit (BPU) within F0 contour and we have grouped the AU patterns over a Romanian speech corpus taking into consideration their different functions in spoken discourse. A solution for the F0 contour synthesis by concatenating AU natural patterns is presented in [2]. The model defines three functional types of AUs in different acoustic and phonetic contexts. In our model we increase the number of functional types of AU and the final result is a language for describing the melodic contours.
The intonational hierarchy from figure 1 illustrates our perspective over the melodic contours. An intonational phrase/intermediate phrase (IP/ip) consists of several AUs, depending on the number of stressed words. The AU sequence is nonlinear, as AUs alternate with AU groups on the same level of hierarchy. There are objective reasons for intermediate grouping between AUs and IPs/ips levels [3]. We called these groups accentual unit groups (AUGs). On the lower level of this hierarchy, an AU sequence results by splitting the AUGs.

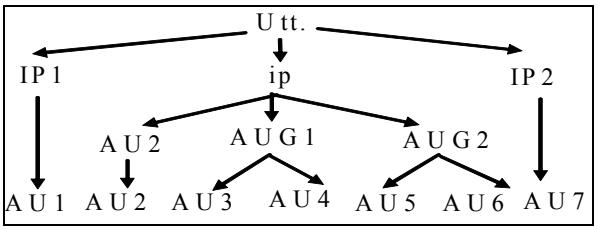

Figure 1: *The intonational hierarchy*

For classifying the BPUs we have used the F0 contour (AU pattern types) as a symbolic correlate between the spoken discourse events and the acoustic events [4]. As, referring to this idea we found that, generally, an IP should contain an AU that contributes to bring into attention the new basic discourse unit (BDU) and also, another AU to end it and eventually to prepare the audience for the next BDU. We called them AUs of PUSH/POP type. We found that the AUs having these functions usually manifest large F0 frequency variations during their accented and eventually, the following unaccented syllables. Within the first ones the F0 contour reaches highest tonal level (Top level) and within the last ones it reaches the lowest tonal level (Bottom level) of the respective IP.
There are other types of AUs that have a role in word focusing by highlighting certain tonal levels within an IP/ip. The highlighted tonal levels can be implied in generating the semantic focus by pitch accents widen around them or by contrasts with other tonal levels/targets from the adjacent AUs. We distinguished non-neutral tonal focus patterns, that manifest significant pitch accents, and weak tonal focus patterns generated by small pitch movements around a highlighted tonal level. The AU patterns having PUSH/POP function within an IP can manifest a tonal focus too.
The IP structure can contain AUs that have the F0 contour widen along the interpolate line between two tonal target/levels highlighted by AUs of previously mentioned types.
The AUG expresses either a semantic relation between the corresponding words, or just a rhythmical one. Generally, the AUG structure can be viewed as an IP structure at a small frequency scale that have a local top and bottom coordinate. Therefore, an AUG consists of two or three AUs equivalent to the PUSH/POP ones from IPs, or to the tonal focusing AUs.
We build a set of annotation labels consisting in mnemonics that suggest the function of AU in spoken discourse. Each label has a set of attributes, to indicate a particular F0 contour pattern for the corresponding AU. For example, the pitch accent type on accented syllables (as in ToBI annotation system) is explicitly descripted as an attribute of the AU label, only in case a certain pattern has no default pitch type defined for the respective label.
A Romanian speech corpus analysis has led us to define F0 contour patterns as prototypes for each category. The variability within the defined functional categories is

generated by the position of the accented syllable, the type of pitch accent, the prominence of the AU, the number of syllables etc. The labels give information about the position of the AUs in the F0 frequency scale, about types of pitch movements within it, about the amplitude of F0 frequency variations.
In conclusion, the perspective of this model gives a meaning to the melodic contour of an IP, close related to the F0 contour pattern for AUs. The description based on the corresponding labels will be more easily converted by a phonetic module of a TtS system into a sequence of patterns for building the F0 contour.

2. The label set for intonation description

The analysis of Romanian intonation over a speech corpus containing neutral text reading led us to identify some AU pattern types and to develop a set of corresponding labels for intonation annotation, divided into five categories: labels for AUs of PUSH/POP type at IP/ip level, labels of push/pop type at AUGs level, labels for tonal focusing, derived labels for AUs that have both preceding functionality and a category with AUs that link two tonal levels (tonal link labels). The labels are presented in the following paragraphs using a symbolic description of intonation consisting in AU labels separated by slash "/" and grouped by round parenthesis "()" into AUGs and by squared parenthesis into IPs/ips. In order to use them for an XML description, the labels have been transformed into values of the functional attribute of the AU tag that marks the text of an accentual unit.

2.1 The PUSH/POP labels

Introducing into attention a new BDU is performed usually by an AU that manifests large F0 variation during its accented syllable and/or the next unaccented one, up to the top level of the IP. We labeled these AU by „PH" label (PUSH). In the corresponding manner, the IP contains an AU that marks the end of the BDUs and we annotate it by „PO%" label (POP) and „PO" label in the IP and ip ending, respectively (the *ip* having "L-" type accent phrase).
In neutral case, an AU of "PH" type has a pitch accent of H* type and corresponds to the so-called "accent without focus" defined in [5] or to an initial accent (AI) defined in [6].
The AUs of PUSH type usually have an initial position within the IP (except yes-no question cases), but in case the first word/words must be focused at low tonal level, in neutral manner („f" labeled AU), a delay occurs to the PUSH event with the corresponding length of the focused word/word group.
The end of an IP is generated within an AU labeled with „PO%", in case an end point is present in the corresponding text. The „PO%" pattern is characterized by a decreasing F0 variation until the bottom level of IP is reached.
Another kind of AU pattern for BDU ending must be defined when both the end of the current BDU and the beginning of the next BDU are marked in spoken discourse. In this case, after the F0 contour reaches the lowest level during the last accented syllable, a rising F0 contour segment corresponding to a high boundary tone begins. We labeled this type of AU by „PU%" (POP-UP) and „PU" label in the IP and ip ending, respectively (the *ip* having "H-" type accent phrase).
Figure 3 illustrates the F0 contour of the utterance of the Romanian text *Avem de discutat lucruri serioase*. The "PH" label is assigned to the word *Avem* and the POP event corresponds to the word *serioase* labeled by a derived label "PO%+F" that suggest the occurrence of both POP event and focus event.

2.2 Tonal focusing labels

In an IP, between an AU of type "PH" and an AU of type "POP/POP-UP" there are one or several AUs within which certain tonal levels are highlighted. In figure 3, two types of tonal focus are present, corresponding to the words *lucruri* and *discutat.*

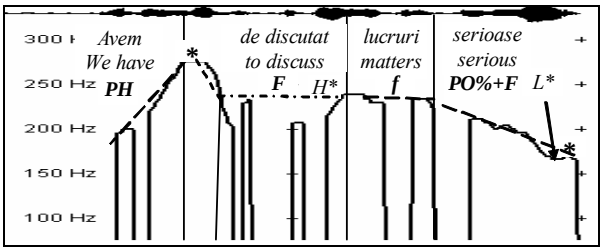

Figure 3: *The natural and the stylized F0 contour of the utterance of the text „Avem de discutat lucruri serioase" ("We have to discuss about serious matters")*

The first pattern, illustrated by the word *lucruri*, refers to the case of small variations around a tonal level, after having reached the focusing tonal level. We annotate this kind of AU pattern by „f" label, considering it a weak focus, usually occurring on the topic word.
The second pattern illustrated by the word *discutat* refers to the strong focus generated by a pitch accent of H* type. For this type of AU the tonal level on the last syllable equals the one from the beginning of the stressed word. Keeping the beginning and ending tones within a word on the same level, in the presence of a major pitch accent (H* or L*), one generates a strong focus on a certain word. We annotate these AU patterns by „F" label. In case the accented syllable is in the initial position of a focused word, the well-known shape of "peak" is generated within the corresponding AU and then the tone on the last unaccented syllables falls until the initial level is reached.
The AU of the last word *serioase* performs BDU ending and highlights the final target tone by a pitch accent of L* type generated by a slope during last accented syllable. Therefore, the AU was annotated by "PO%+F" derived label. The description of the melodic contour is the following:
PH / F / f / PO%+F
An AU having a F0 pattern of type "f", positioned at low level before or after an AU with high target tone, carries a semantic focus generated by the tonal contrast between the target tones of two adjacent AUs (figure 4). The semantic focus based on tonal contrast (corresponding to the metrical view of sentence stress defined in [5]) is frequently used in Romanian neutral rendering. We have no special label to annotate the semantic focus generated by the tonal contrast between the target tones of two adjacent AUs, because a new label can't add other information about the F0 pattern. The attribute for tonal levels ("-l") should be used with the labels of the implied AUs in order to change the size of tonal contrast.
Figure 4 illustrates the F0 contour of the utterance of the Romanian text *Era genul de om… (He was the type of man…).* The verb *era* is rendered at a low level with a "f" pattern and becomes focused by the following high target tone reached during the next AU of PUSH type.
The label sequence f / PH / PU% -l:m describing the F0 contour from figure 4 specifies a medium level for the boundary target tone of the "PU%" label, using the attributes "-l:", specifying that the last rising segment hasn't got a large amplitude.

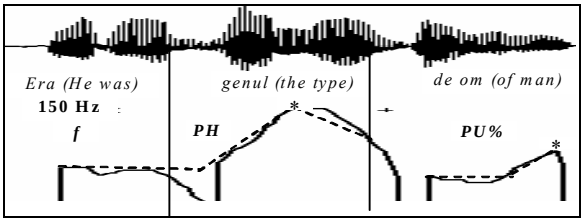

Figure 4*: The natural and the stylized F0 contour of the utterance of the text „Era genul de om…" ("He was the type of man…")*

The information about contrast focus is implicitly presented in the description. Before an AU of "PH" type, the "f" pattern corresponding to the first AU has the level attribute, "-l:", at the implicit low value.

2.3 Derived labels

The beginning of a BDU can be conveyed by an AU that reaches a high tonal level by stepping upward before its first syllable. During the AU the F0 frequency keeps this high value manifesting small variations in amplitude. We annotated such an AU pattern by „PH+f" label.

Figure 5 illustrates the F0 contour of the utterance of the Romanian text *Bine, atunci luați loc şi să stăm de vorbă mai comod (OK, then sit down and let's talk more comfortable).* The „PH+f" label corresponds to a beginning IP at a high focused tonal level.

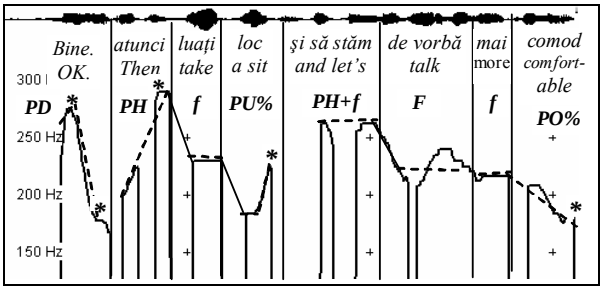

Figure 5: *The natural and the stylized F0 contour of the utterance of the text „Bine, atunci luați loc şi să stăm de vorbă mai comod" ("OK, then sit down and let's talk more comfortable")*

The utterance contains three IPs, the third one begins at high level and its first AU (corresponding to the verb *să stăm*) keeps this high level with small variations.

Sometimes, after a prominent H* pitch accent during a PUSH event, a tonal return to the value from the beginning of the word occurs in the end of the AU. Therefore, a focus is being generated and we label the corresponding AU with „PH+F". A focus can also be generated by a decreasing L* pitch accent after the high tonal level is reached and, in this case, we use the „PH+F" label and the attribute "–p:" (pitch) appears explicitly in the description with the "l" value.

In a similar manner, we can generate derived label for POP/ POP-UP events in case a prominent pitch accents occurs by a tonal decrease on the accented syllable, down to a minimum value. We label these Aus with "PO%+F"/ "PU%+F".

Another type of L* pitch accent within a POP-UP event is generated by just mantaining the low level during the accented syllable, reached before, and then by the rise of the F0 contour, in order to generate the boundary tones. This pitch accent characterization corresponds to the pattern of underived labels "PU%", illustrated in figure 5 by the AU corresponding to the word *loc*. The intonational description of the three IPs from figure 5 is the following:

[PD] [PH / f / PU%] [PH+f / F / f / PO%]

The label PD corresponds to a PUSH-DOWN event. We consider that in yes-no questions the PUSH event occurs in the end of the BDU in the case of nonfinal emphasis. The final rising on accented syllable by a H* pitch accent generates an PUSH event (oxitone final word) or a PUSH-DOWN event (nonoxitone final word). For yes-no questions an attribute for F0 range ("-r:") is used with large "l" value in order to characterize their final rising up to high levels.

An example of using "PD" labels in affirmative intonation is illustrated by the word *bine* in figure 5 where two target tones can be distinguished within the AU, both being significantly highlighted. The first one corresponds to a high target tone of the pitch accent and the second to a lower boundary tone. The F0 contour rises on the first syllable and falls down on the last syllable.

Another type of PUSH-DOWN event in affirmative intonations is generated within an AU during the initial unaccented syllables, followed by the fall of the F0 contour to the level where a focus occurs, starting from the accented syllable. We labeled this type of AU with "PD+f", in case of neutral focus and with "PD+F" in case of strong focus (L* pitch accent).

2.4 The push/pop labels

The annotation of AUGs must consist of a label sequence corresponding to its AU components and it may be assigned a label that characterizes its function within the IP/ip.

To annotate the AU component we introduce a set of mnemonics equivalent to those used at IP/ip level, based on their functional resemblance, as follows: „ph" labels for the first AU, "po/pu" labels for the last AU and F/f labels for the focused AUs. The AUs that have functions at IP/ip level too, keep the label at this level. The AU patterns of „ph" or „po" type contain a H* pitch accent and the patterns of „pu" type contain a L* pitch accent. In figure 6 the arrows mark the target tones of all pitch accents of H* type within the four AUGs. The F0 patterns at AUG level are also characterized by the difference between the AU target tone levels. In figure 6, it has an implicit positive value for the first three AUGs and an implicit zero value for the last one. The last AUG having the second target tone at a level equal to its first target tone, generates the *ip* ending with an accent phrase of "H-" type(the AU of "PU" type).

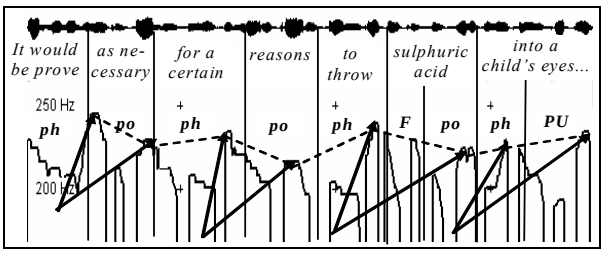

Figure 6: *The natural and the stylized F0 contour of the utterance of the text „...s-ar dovedi necesar dintr-un motiv oarecare să aruncați acid sulfuric în ochii unui copil..."*

The description of the F0 contour from figure 6 at AU level, corresponding to the middle and the end of an intermediate phrase, is the following (the PUSH AU is not presented in figure 6):

...(ph /po) /(ph / po) / (ph / F /po) / (ph / PU)

The labels for AUGs are the same as those used for ungrouped AUs at IP/ip level annotation, conveying they have equivalent functions. For example, an AUG in the beginning of the phrase described by (PH/po) is labeled by "PH", or an AUG in the ending of the phrase described by (ph/PO%) is labeled by PO% and a focused AUG (ph/po) can be described by the "F" label.

The following compact description results for the F0 contour by using only AUG labels:

...(F) / (F)/ (F) / (PU)

The compact description is useful for searching through a lexicon of stylized melodic contours at IP/ip level.

2.5 Tonal linking labels

The labels of type "L" are used for annotating the AUs that link two tonal levels. Their F0 contour patterns manifest a downsteping or upsteping trend. If a pitch accent occurs during the accented syllable of an AU of this type, then its label becomes "L+f". Using the attribute "t" (trend) and one of the values "u" or "d" helps to specify the upstep direction ("u" value) or downstep direction ("d" value). Figure 7 illustrates the F0 contour of the utterance of the Romanian text *Sunt destule scaune?(Are there enough chairs?*). The linebase has a decreasing tendency until the lowest tone is reached, before the final rising of the yes-no question.

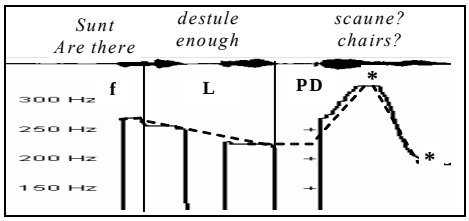

Figure 7: *The natural and the stylized F0 contour of the utterance of the text „Sunt destule scaune" ("Are there enough chairs?")*

The AU pattern of the word *destule* fits the implicit downstep trend beginning with the medium level at which the first AU of auxiliar verb "*sunt*" performs a weak tonal focus. The description of the melodic contour from figure 7 is the following:

f -l : m/ L / PD

3. The intonational melodic contour description using AU labels

In the perspective of this intonational model, the melodic contours of IPs/ ips can be stylized to a sequence of tonal floors and target tones. The target tones are reached during PH / PO/ ph / po events. Along each tonal floor the variations of an "f or F" AU pattern are spread. The "L" AU patterns link two consecutive tonal floors/targets. The size of the difference between adjacent tonal coordinates of the stylized contour can be used to control the semantic focuses prominence based on tonal contrast. The prominence of PUSH or POP/POP-UP event influences the whole F0 frequency range of the IP, respective top and bottom level. An IP has a number of tonal floors at intermediate levels. We view the IP as a temporal sequence of tonal floors and target points that generate a tonal skeleton to which the contained AU patterns are anchored. The F0 stylized contours from figures 3-7 illustrate the tonal skeletons of the corresponding natural curves.

Based on the speech corpus analysis we built two lexicons, one containing tonal skeleton prototypes described by different label sequences and another containing the AU patterns corresponding to each label in different lexical contexts. By using the two lexicons a phonetic module can be designed in order to generate the F0 contour in Romanian speech synthesis.

4. Conclusions

We consider that this description language for Romanian intonation can be understood both by NLP researchers and speech technology researchers that are interested in spoken language. The description using AU labels characterizes better what happens within F0 contour between pitch accents.

Our future task consists in building a Romanian speech corpus annotated at intonational level using the AU label set and at morfo-syntactic level, and then training a module to predict an intonational structure for an input text in Romanian text-to-speech systems. We think the syntactic units can be more easily assigned to a structured sequence of AU labels. Furthermore, the F0 contour generating module must be modified in order to use the pattern lexicons at AU level and IP/ip level.

The results will be useful in connecting a TtS system to the eDTLR (electronic Romanian dictionary), performing read aloud sense definitions of the words contained in the dictionary, in order to be used by the persons with sight disabilities.

5. Acknowledgements

This research was performed in the Romanian Academy and is partially supported by the grant PNCDI2 nr.910013/18.09.2007 entitled "eDTLR – Dicționarul Tezaur al Limbii Române în format electronic".

6. References

[1] Apopei V. and Jitcă D., "Module for generating the F0 Contour using as input a Text structured by prosodic information", Advances in Spoken Language Technology, Romanian Academy, 119-126, Bucharest, 2007.

[2] Heggtveit, P. O. and Natvig, J. E., "Intonation Modelling with a Lexicon Natural F0 Contours", Eurospeech, 2001.

[3] Sun-Ah Jun, "Intonational phonology of Seul Korean" Revisted, Japanese /Korean Linguistics Conference, Tucson, Arizona, nov.5-7, 2004

[4] Morton, K., Tatham, M. and Lewis, E., "A new Intonation model for text-to-speech Synthesis"

[5] Ladd, D. R. , "Intonational Phonology", Cambridge University Press, 1996

[6] Mertens, P., "Synthesizing elaborate Intonation Contour in Text-to-speech For French"