ISCA Archive
http://www.isca-speech.org/archive

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

TOWARDS AN INTONATION MODULE FOR A PORTUGUESE TTS SYSTEM

Diamantino Freitas and Daniela Braga

Faculty of Engineering of the University of Porto and
Polytechnic Institute of Porto, Portugal
dfreitas@fe.up.pt
dbraga@fe.up.pt

ABSTRACT

In this paper, a correlation between the linguistic structure of the written text and the real intonation behavior of the read speech in European Portuguese language (EP) is presented. It is our belief that intonation behavior in EP can be strongly predicted from two main coordinates: the syntactic structure of the sentence and its pragmatic communicative function, in one way, combined with the phonological and syntactic nature of the words, in the other way.

The purpose of our work is to identify in real speech the main intonation elements, which are relevant to speech naturalness as well as to analyze the factors that determine them. This work addresses the cases of declarative/imperative, interrogative and enumerative phrases. Basic categorizations of the intonation elements, in correlation with the underlying factors are presented. General regularities and correlations as well as the resulting rules, that may be a starting point for practical implementation of an intonation module, are presented and demonstrated, under a Fujisaki's phonetic/physiological approach.

The methodology was based on the observation and modeling of a significant prosodic corpus where different intonation patterns occur in a diversity of text structures. It is our goal to contribute with practical techniques and experience in order to perform a more accurate intonation modeling of Text-to-Speech (TTS) applications, using a rule-based approach.

1. INTRODUCTION

In the past a considerable number of studies of EP prosody has been produced [1] [2] [3] [4] with outstanding theoretical results. However, aiming at the use of TTS systems, a big gap has remained to be filled, specifically, in what concerns practical rules or methods that allow building a prosodic control module. As far as the authors know there are no previously published studies in EP that bring these methodologies to the engineering or programming areas for implementation.

Taking in consideration the theoretical developments, a more practical study was started a while ago with the purposes of, firstly, to control the intonation parameters of the synthetic speech waveform and secondly, to find the appropriate rules and methodologies for implementation of this control in a natural sounding way.

For the first aspect, the Fujisaki's approach was selected amongst others because of its parametric nature, good accuracy and physiological base. For the second aspect, a rule-based approach for generation of intonation scheme was chosen, taking in consideration the more or less deterministic way in which intonation is produced in read speech and its quite clear correlation with text structure, including syntax, semantics and pragmatics. In our work, only syntactic aspects and communicative intention are used for the moment.

2. LINGUISTIC TREATMENT OF THE TEXT

As shown elsewhere, we have developed a linguistic rule-based platform in order to generate intonation in a predictable way. In a TTS system, the step after the text pre-processing (numerals, acronyms, grapheme-phoneme conversion) is the linguistic analysis of the text. There are basically three types of linguistic information provided by the written text that we consider decisive for f0 manipulation: the pragmatic and communicative objectives of the utterances, the phonological structure of the words, and the syntactic analysis of the text.

The pragmatic orientation of the utterance, whether it is a question or a statement, is traditionally extracted through punctuation marks. Although these conventional signs are just rough prosodic transcribers, when observing a small sample of intonation possibilities, they can indeed be extremely useful in correlation with other linguistic information, such as the syntactic content and function of the words. Anyhow, punctuation marks happen to be the first prosodic parsers in a text, since they immediately mark-up important phrase boundaries where intonation patterns will be anchored.

The phonological information that can be useful for f0 control is the tonic/ non-tonic syllable identification and the word length (number of syllables). As TTS systems start from written texts, both phonological parameters are conventionally treated following the syllabic division prescribed by orthographic standard rules for EP. Nevertheless, some co-articulation events and phonetic reductions are obviously considered in our experiments.

This level of analysis, in particular the syllabic quantity and division of the word, is of extreme importance for artificial f0 generation using the Fujisaki model, since the tonic syllable is where the accent commands are moored. Normally, each tonic syllable corresponds to an accent command. However, experiments reported below demonstrate how relevant non-tonic syllables can be in certain positions in the utterance, especially regarding the initial, intermediate and final rise/fall movements. The length of the first and last words in number of syllables can, by its turn, determine the length of the initial and final fall/rise movements, because the former ends at the onset

of the first tonic syllable and the latter starts at the onset of the last tonic syllable of the utterance. We therefore can see that the length and the tonic syllable position play a joint role in those movements.

The purpose of the syntactic analysis of the text is to recognize hierarchically disposed phonological units with semantic meaning that provide the boundaries from where intonation patterns will be obtained. With this linguistic-based process correlation regularities are detected. Its functioning begins with a morpho-syntactic labeling of the words. This aspect is important to the first intonation factor: the consideration of a 'single word' or a 'multiple word', that has to do with the nature of the word – weather there is a content word (nouns, verbs, adjectives, adverbs) or a function word (prepositions, articles, some pronouns, some connectors). This has obviously repercussions on the f0 trends, as shown below. For example, the Figure 2, below, shows that the article "o" (the) is treated as non-tonic because it is a function word.

As we proceed in the text analysis, a set of rules based on the articulation of the Valences grammar and the Generative grammar interpret and combine words into syntagmas, in a first stage, and from syntagmas into phrasal groups. A typology of categories for each linguistic tier was also developed. The marks that result from the analyzed text support the prosodic patterns.

The TTS system is prepared to receive and to store the information obtained from the morpho-syntactic labelling of the text. This information is hierarchically organized, from the morphological value of the word up to the highest level, the Phrasal Group level. Prosodic labelling is also obtained from the linguistic mapping of the sentences. Briefly, we have considered 4 levels of linguistic analysis, the word level (WL), containing the syntactic information of the word, the syntactic level (SL), consisting in a unit that may receive a syntactic function, the phrasal group level (PGL), which the highest unit of meaning where a breathing stop is allowed and the prosodic level (PL), totally anchored on the phrasal group labels.

From each of the Phrasal Groups derives one prosodic pattern. For each of these linguistic levels we have proposed a typology [5] of labels that are presently being implemented in the PROLOG analyzer.

3. PROSODIC ELEMENTS AND FACTORS

We followed Fujisaki's model for intonation pattern generation, which is based on a small set of elements such as phrase and accent contours, whose features (parameters) are time positions and durations, intensities, "time constants" and limit amplitude [6] for which we have demonstrated that a combination of those contours is capable of describing a smoothed intonation pattern of a random phrase in EP [5].

The next step is to analyze real speech f0 contours and detect the presence and features of so-called intonation elements that emerge in coincidence with, or only in the proximity of, the candidate position marks, of syntactic and lexical origin, merged with the physiological positions recommended by Fujisaki [6].

3.1. Intonation elements

The chosen elements have some similarity but don't coincide with the intonation elements recommended in other intonation analysis approaches like TOBI and INTSINT. The method used in selecting the present elements was the observation of the prevalence of relevant f0 movements in the phrase contours, amongst a significant set of phrase materials and verification of their subjective importance for the naturalness of the phrase f0 contour. For this purpose manipulation of the f0 contours through manual linearization was done in the PRAAT [7] environment.

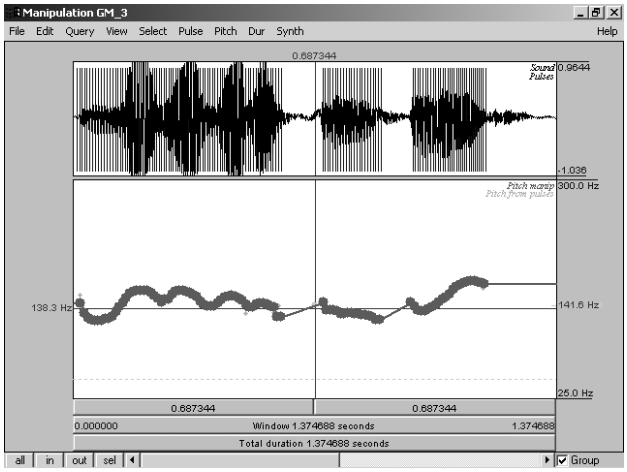

Figure 1: Yes / no question: "Deseja ver as promoções?".

The selected intonation elements, sorted roughly by decreasing importance for the communicative intention, were:

- Final rise/fall
- Phrase declination
- Initial fall/rise
- Intermediate fall/rise and rise/fall.

Figures 1 and 2 show examples of f0 contours where the above elements can be clearly observed. In figure 1the first, second and third elements can be observed.

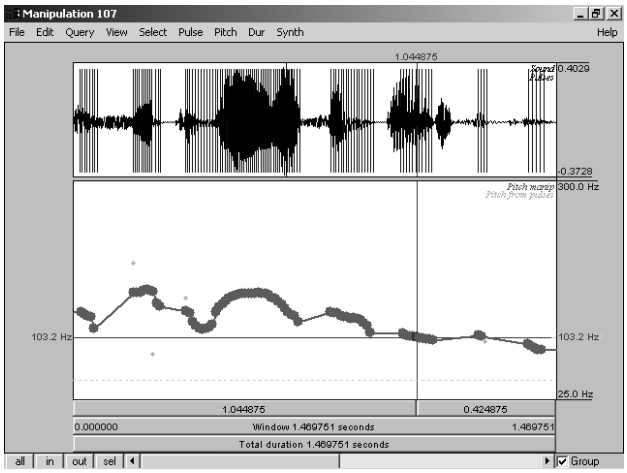

Figure 2: Declarative: "O futebol é fantástico."

In figure 2, at the cursor's position a small but observable f0 decrease below the declination can be observed, as a "negative" intermediate accent.

As far as the authors know, the initial rise/fall element has not been mentioned before in the literature as a relevant element in EP intonation patterns, but we can now prove that its relevance is quite large in the perspective of naturalness. Technically, the above elements are easily realizable by means of Fujisaki's contours what makes them even more interesting.

It is important to state at this stage that, especially in long phrases, a number of intonation movements of lexical origin are normally present giving the f0-smoothed contour a quite complex shape. Although it is possible to model this complexity with Fujisaki's elementary contours we have soon concluded that most of the accents are not relevant in a first approach to naturalness. It was in the definition of which were the relevant elements that we decided to keep the fourth category above.

3.2. Intonation categories considered

Due to practical reasons we decided to limit this study to a set of most important categories of communicative intention in read speech. We selected from the above list the following prosodic categories:

- Declarative, imperative
- Interrogatives, wh- and yes/no questions
- Enumerative.

For these categories a number of phrases was studied. The phrases were selected according to the dissemination in the space spanned by the most common types of phrases encountered in non-sophisticated applications such as e-commerce, news and information systems. The metrics of these sentences was carefully prepared taking in consideration a number of syntactic and lexical factors and their combinations in order to highlight the existing invariant elements amongst similar phrases as well as the specific elements of the phrases categories (minimal pair approach).

3.3. Intonation factors

Investigation of occurrence and features of the above intonation elements was conducted along the following coordinates or intonation factors presenting a clear or potential influence:

- Single word / multiple words phrase: it is evident that a single word utterance has a simpler intonation with fewer elements than one with more words
- First tonic syllable position: the duration of the pre-tonic segment can be significant in longer words, how does intonation start?
- Morpho / syntactic type of first word: syntactic emphasis can affect intonation of the first word
- Number of syllables in first word: a factor that potentially connects with the previous three
- Phrasal group (see above) delimitation: clear intonation movements exist in general at phrasal groups boundaries
- Number of syllables of last word: the final f0 movement can potentially be influenced by this factor
- Position of tonic syllable in last word: similar to the previous argument but stronger.
- Position of the verb and or intensity adverbs in the phrase: a strong influence may be observed specially when placed near the beginning or the end.

3.4. Illustration of some typical cases

3.4.1. Final rise/ fall:

The final fall/rise movement may exist, frequently as a single rise movement, in yes / no questions. See the example linearized f0 contour in figure 3 below.

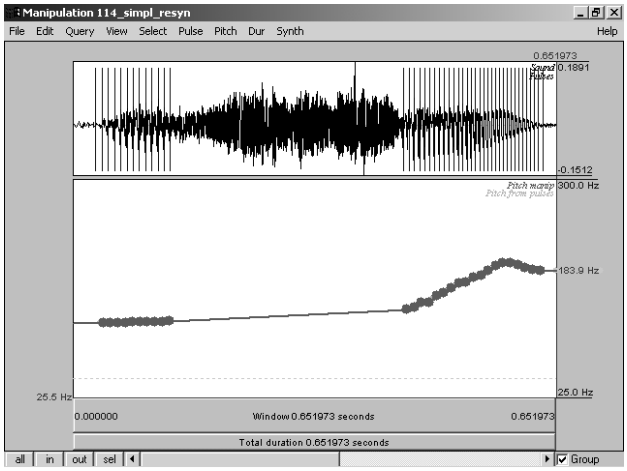

Figure 3: Yes / No -Interrogative: "Disse sim?"

In wh-questions the final pattern is similar to the declarative sentences, that is with a smooth ending of declination added with an optional tonic vowel accent. See the example in figure 4 below.

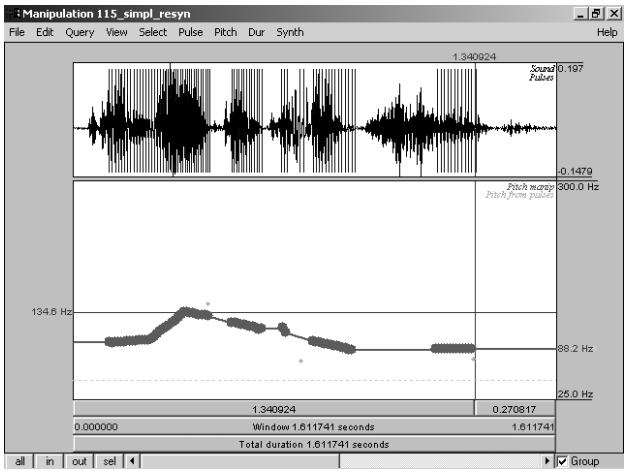

Figure 4: Wh- interrogative: "Qual é a quantidade desejada?"

3.4.2. Phrase declination:

Phrase f0 contour declination happens in every phrase. For instance in the example of figure 5 below the linearized f0

contour is rather typical of an utterance with two phrasal groups but a single global descent

In single word phrases the declination effect also happens as can be seen in figure 6.

3.4.3. Initial fall/rise:

This movement is presently one of the most interesting because of the increased naturalness that occurs when it is considered in real text-to-speech systems, in comparison with a situation with its absence. The initial fall/rise is generally steeper than the declination itself, starts from a small value of f0 but doesn't go down below the base value. The duration of this element can be pronounced, like in the phrase of figure 5, short like in the phrase of figure 2 or even non-existent like in the single word utterance of figure 6. The normal extent of this element is generally from the start of the utterance to the onset of the first tonic syllable.

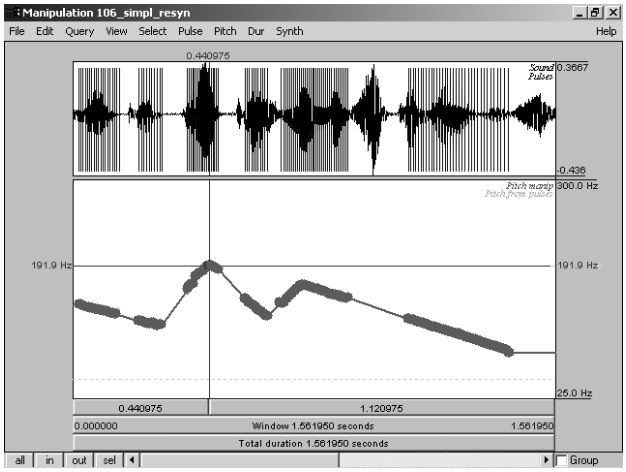

Figure 5: Imperative: "Multiplique os números escolhidos!".

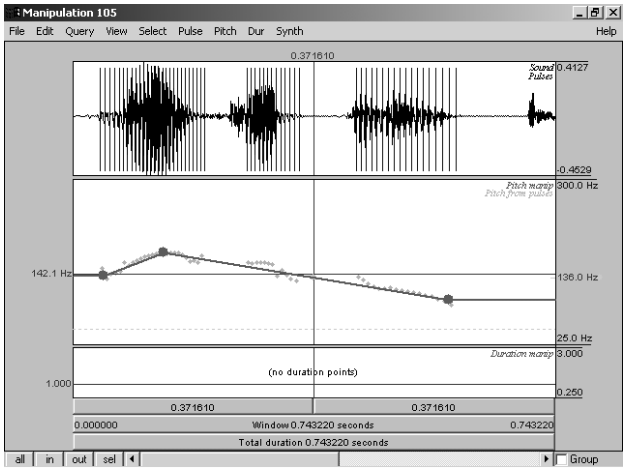

Figure 6: Imperative: "Multiplique!".

3.4.4. Intermediate fall/rise and rise/fall:

This element is most of the times an optional one, meaning that its occurrence is not essential for a natural intonation. Its presence brings normally more style to the utterance. It is basically defined as an alternation or rhyme in f0 placed in adjacent syllables, in general. This element can be well observed in the utterance illustrated in figure 2.

4. CONCLUSIONS

We presented a study of basic intonation aspects in European Portuguese consisting of identification of intonation elements and their relationship to intonation factors in the phrase and to the type of communicative intention. The study extended only to the most frequent types of sentences, and addressed the most important intonation elements. The present results, including the characterization of the initial fall/rise that has not been addressed in the past, permit the realization of synthetic intonation with a very good degree of naturalness using the Fujisaki model approach with parameters that can be easily extracted from the elements characterization in the space of the intonation factors of the sentence.

In the near future, a more detailed description of the present rules is foreseen, allowing to synthesize an even more natural intonation pattern, that together with a specially developed durations model, described elsewhere [8] will permit to achieve a globally more perfect and practically realizable prosody.

5. ACNOWLEDGEMENTS

The authors wish to acknowledge AdI –Agência de Inovação, NEOSIS, FEUP, IPP and COST 258 for the support given to this work.

6. REFERENCES

[1] Frota, Sónia, Prosody and Focus in European Portuguese, New York, Garland Publishing, Inc., 2000.

[2] Cruz-Ferreira, Madalena, "Intonation in European Portuguese", in Hirst, D.; Di-Cristo, A.; Intonational Systems, Cambrige University Press, 1998.

[3] Vigário, M., The Prosodic Word in European Portuguese, PhD dissertation, University of Lisbon, 2001.

[4] Martin, Philippe, "Modeling F0 in various romance languages", in Improvements in Speech Synthesis, Keller, E. et al., editors, Wiley: 104-119, 2002.

[5] Freitas, D. et al., "Correlation between phonetic factors and linguistic events regarding a prosodic pattern of European Portuguese: a practical proposal", *Proceedings of ICSP2001*, Taejeon, Korea.

[6] Sagisaka, Y. Et al; *Computing Prosody*, Spring New York, USA, ISBN 0-387-94804-X, ch 3, pp. 27-40, 1997.

[7] PRAAT, Copyright 1992-2001, by Paul Boersman and David Weenink, www.praat.org.

[8] Teixeira, J., "A segmental durations model for Portuguese Text-to-speech", submitted to ICSLP 2002.