Modeling of Intonation Bearing Emphasis for TTS-Synthesis of Greek Dialogues

D. Galanis, V. Darsinos & G. Kokkinakis

Wire Communications Laboratory, University of Patras, Greece

ABSTRACT

TTS-synthesis of neutral style Greek with good intelligibility and quality has been achieved some time ago. As a further step towards expanding the applications domain of the TTS-system developed in our laboratory, the incorporation of emphasis into speech used in man-machine dialogues according to their context has been studied recently. In this paper the method applied for the analysis of intonation patterns, the results of this analysis, the algorithm established for the creation of desired intonation patterns and its implementation in the existing TTS-system for Greek is reported.

1. INTRODUCTION

The need for the exploitation of currently developed speech synthesis systems in expanded application areas [1][2] which do not necessarily address to the problem of the efficient one way information transfer [3], involves the capability of synthesizing various types of speech. Intonation functions over several levels of spoken language, e.g. context and semantics [4], and thereby its variation in natural speech conveys information about the interaction and provides useful cues to the human participants as well. Therefore the human-machine interaction in a dialogue system could be efficiently improved if the synthetic speech could convey information in a more natural manner by including emphasis according to the context and thus interactionally appropriate intonation patterns.

As the first step for achieving this goal, in this paper we present the analysis and the modeling results of context dependent emphatic intonation patterns. The investigation of the F0 contours of the collected speech material revealed the mechanism for realization of emphasis in Greek in terms of extent and duration of F0 rise and fall in each specific case. In addition, we present an algorithm for the automatic generation of such patterns incorporated in the existing Greek TTS-system.

2. SPEECH MATERIAL

Since we are interested in the intonational marking and realization of intentional focus and furthermore the possible use of the results in an interactive dialogue system, a set of 100 sentences were designed and embedded into several information exchange dialogues from the public transport domain. The dialogues were performed by two (2) male non-professional speakers (informer and inquirer), each time in a different pre-defined focusing context. The speech samples were recorded in a sound treated room and digitized at 8 kHz.

The dialogues consisted of sequences of simple statements and questions (including the 100 sample phrases) uttered either by the inquirer or the informer. Since each one of the sample phrases could be included in different dialogues, we performed a listening test to verify and judge the different intonational realizations as far as their context dependency is concerned. Two subjects were involved in this verification task which consisted of the following steps: the dialogues were presented to the listeners in a whole while the test phrases were clearly pointed out for each case. Then the listeners were instructed on the one hand to determine whether the intonational realization of the indicated phrases was or was not appropriate according to the dialogue structure and on the other hand to specify for each one of the well judged cases the lexical elements (words or phrases) characterized by emphasis. Finally, the fundamental frequency contours for the phrases extracted from the listening test was automatically computed.

3. ANALYSIS AND RESULTS

3.1. Analysis procedure

Prior to the work reported in this paper the target of the prosodic analysis was the determination of a neutral speech output style [5], that is appropriate for one way applications. The basic idea that was followed for the expansion of the existing intonation model for emphatic speech and furthermore for dialogue purposes was to determine and model the contrastive differences of:

- the intonation patterns observed in the emphatic sample phrases and the corresponding patterns as they are determined from the neutral intonation model and
- the emphatic intonation patterns of the same sample phrase as observed in the different dialogues with respect to the location of the lexical items bearing emphasis according to the dialogue structure.

As far as emphatic intonation patterns are concerned the analysis and modeling is interpreted in terms of pitch range and timing of the F0 movements. From the investigation of the intonation material in our dialogues we observed that the realization of emphasis in Greek is straight-forward with respect to these parameters. It is realized with a major F0 rise (thus expanding the F0 range) aligned with the stressed syllable of the lexical item bearing emphasis and a F0 fall starting immediately after the end of it. The duration of the fall varies between one and three syllables and one of the most striking properties is the accentless intonation to the end of the phrase. In cases where the item bearing emphasis is placed far from the beginning of the utterance we observe F0 rises and falls of moderate size synchronized with stressed syllables and phrase boundaries up to the position of the emphatic lexical item.

The observations reported above are clearly shown in Figure 1 and furthermore indicate a three level segmentation to be adopted through the analysis and the modeling procedure as well. The sample phrases were segmented into three parts (pre-focal, focal

and post-focal) according to the position of the emphatic item with respect to the results of the listening test.

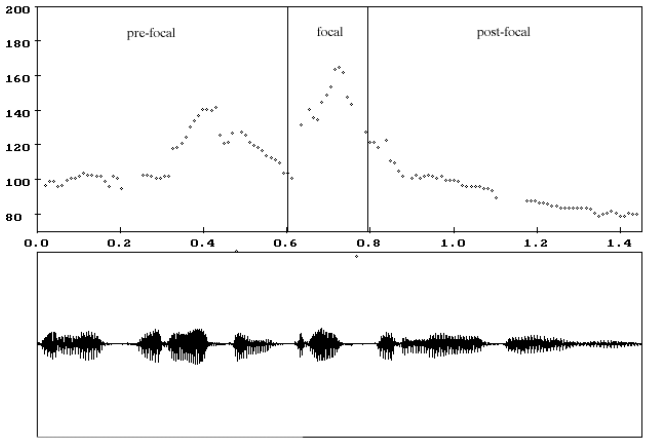


Figure 1 *F0 contour of the sentence "The three o'clock bus stops at Egio" with emphasis on "three o'clock". The pre-focal, focal and post-focal parts are indicated.*

3.2. F0 modeling

In order to capture the basic characteristics of all the different emphatic intonation patterns which are not determined from the neutral model and express them in terms of extent and timing of F0 variations, we created an alternative version for each one of the contours based on its approximation with the smallest possible number of straight line segments. The resulting stylized pitch contours consist of sequences of pitch movements which vary between four declined lines (BASE, MID, TOP and FOCUS) according to the sentence mode. Consequently, all the turning points of the F0 contour (maxima and minima) belong in one of these four declined lines. Through statistical analysis of the F0 measurements of every turning point and consequently each declined line and with respect to the segmentation of the sample sentences in the pre-focal, focal and post-focal parts we concluded the following:

There are no pitch movements in the pre-focal part which exceed in size the distance between the lower (BASE) and the higher (TOP) declined lines as it is quantitatively determined by the neutral intonation model. According to this model, which was obtained through statistical analysis for a neutral speaking style, the quantitative attributes of the lower declined line, that is its slope B_{SLOPE} and starting F0 value B_{START}, are given from Eq.1 and Eq.2 respectively as a function of the sentence's total duration t. The MID and TOP lines are defined as having a standard distance of 3.5 and 6.8 semitones (ST) above BASE line. By measuring the BASE line slope and starting frequency for each one of the emphatic sample sentences and comparing them with the corresponding neutral values obtained from Eq.1 and Eq.2 we came up with a standard deviation of 0.25 ST/sec for B_{SLOPE} and 1.03 Hz for B_{START}. These results imply that the size of the F0 variations observed in the pre-focal part and the slope and starting frequency values of BASE line can be modeled accurately enough by the neutral intonation model.

$$B_{SLOPE} = \frac{3.05}{t + 0.505} \text{ (ST/sec)} \qquad \text{(Eq.1)}$$

$$B_{START} = 100e^{-0.061/t} \text{ (Hz)} \qquad \text{(Eq.2)}$$

F0 maxima with values greater than those resulting from the definition of the TOP declined line are observed only during the focal part and they are aligned with the stressed syllables of the items characterized by emphasis. In the cases where the emphatic stressed syllables are more than one, the last one seems to be the most prominent. This seems to be the case when prominence is expressed in terms of absolute F0 values or in terms of F0 overshoot above the TOP line. In the first case the mean value for the last F0 peak is 179.3 Hz while the corresponding value for all the preceding peaks is 172.7 Hz. In the second case the overshoot is 19.5 and 6.2 Hz respectively. Furthermore, these findings are similar to those of the cases where there is only one emphatic stressed syllable (mean F0 peak value 180.4 Hz, mean overshoot 20.1 Hz).

From the results reported above we concluded that the distance of the first (in the cases of more than one) emphatic F0 peaks from the TOP line is not significant enough to be modeled. On the other hand, the values of the last (or the only) emphatic F0 maximum should be used for the determination of the FOCUS line. As it can be seen in Table 1 the mean values of these maxima and their distance from the TOP line as well are proportionally related to the total duration of the sentence.

Sentence total duration (sec)	Mean value of emphatic F0 maxima (Hz)	Mean overshoot of emphatic F0 maxima above the TOP line (Hz)
0 - 1.5	178.2	18.2
1.5 - 2	178.6	18.7
2 - 2.5	179.6	19.7
2.5 - 3	180.2	21.0
3 - 3.5	181.6	21.4

Table 1 *Values of emphatic F0 maxima and their dependency on the sentence duration.*

The FOCUS line was finally quantitatively determined through statistical analysis of the original pitch contours (Fig.2) and it is expressed in terms of the F0 overshoot D in semitones above the TOP line according to the sentence's total duration t_{tot} and the position in time of the emphatic stressed syllable t_F (Eq.3).

$$D = \frac{15.42}{0.57 + (t_F / t_{tot}^2)} \text{ (ST)} \qquad \text{(Eq.3)}$$

As far as the post-focal part is concerned no major pitch variations were observed and the F0 contour followed the BASE line to the end of the sentence as described from the neutral intonation model.

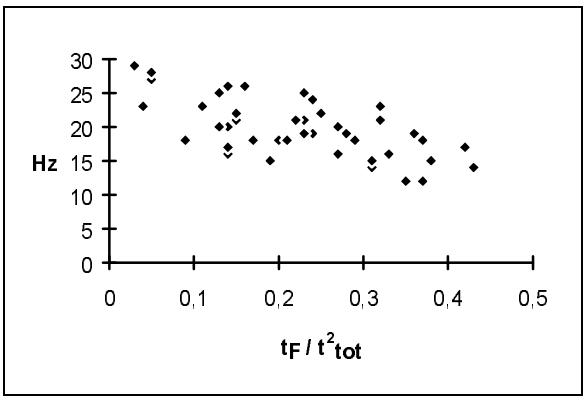


Figure 2 *Overshoot of emphatic F0 maxima above the TOP line as a function of their position in time.*

4. F0 GENERATION

For the purpose of the accurate regeneration of the intonation patterns described in the previous section the basic idea was to capture the F0 contour's characteristics by the determination of all its turning points (maxima and minima) in association with discrete textual phenomena along with information about the location of emphasis. For this reason the syllables of the input text were labeled in terms of a set of discrete features (Table 2) and a set of rules which assigns a target F0 level (BASE, MID, TOP or FOCUS) for every syllable was extracted. The kind of textual information used for the syllable's labeling was selected on the basis of its unambiguous extraction directly from the input text except for the information concerning the location of emphasis which is manually provided. The intonation rules have the form:

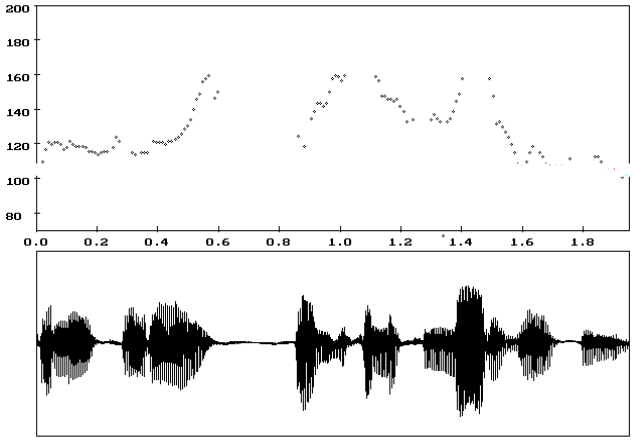
a,b,c, … = F0 level

The rules do not produce absolute F0 values for every syllable but rather the syllable's corresponding pitch value according to the calculation of the four declined lines with respect to the sentence's duration and the location in time of the emphatic items.

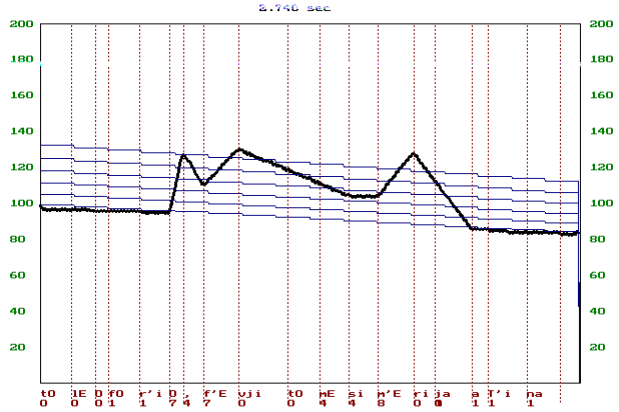
Syllable's features
stressed/unstressed syllable
ultimate/penultimate/antepenultimate syllable
distance in syllables from the previous stressed syllable
distance in syllables from the next stressed syllable
distance in syllables from the phrase boundary
emphatic/non emphatic syllable according to the segmentation of the sentence in the pre-focal, focal and post-focal parts

Table 2 *Discrete syllable features used for the association of the turning points with the input text.*

For the generation of the appropriate F0 contour the input to the intonation algorithm is the text string enriched with emphatic markings which reflect speaker's intentional focus. First, the declined lines are determined according to the sentence duration and location of emphasis. Then, the input text is processed and each syllable is assigned a unique vector representing its attributes according to Table 2. Finally, every syllable is assigned a F0 level according to the rules and the final contour is constructed by linear interpolation between the successive levels. The resulting pitch contour is a fairly accurate reproduction of the original one as can be seen in Figure 3, as far as the patterns used for the analysis are concerned.



(a)



(b)

Figure 3 *Pitch contours of the sentence "The bus is leaving at noon for Athens" with emphasis "at noon": (a) Original contour (b) Automatically synthesized contour.*

5. EVALUATION

For the evaluation of the algorithm's performance a listening test with five listeners was carried out. The purpose of the test was to verify the capability of the algorithm to produce: a) unambiguous emphatic F0 patterns as far as the location of emphasis is

concerned and b) emphatic patterns adequate to dialogues different from those employed in the analysis.

A set of 60 test sentences included in several dialogues and characterized by emphasis in various contexts was collected. After manual determination of the emphatic items, the synthesized versions of the test phrases were presented to the listeners in two ways. The first time the isolated sentences were randomly used and the listeners had to indicate the items bearing emphasis. In the second test three synthetic versions for every test phrase were created, each one expressing emphasis on a different lexical component including the case realized in the original sentence. Then these versions were embedded in the corresponding dialogues and presented to the listeners in a whole. The listeners were instructed to point out which one of the intonational realizations of the test sentence was appropriate according to the dialogue structure by rating them in a two level scale (acceptable or unacceptable).

From the evaluation results for both tests shown in Table 3 we can see that the listeners were able to perceive and determine emphasis with great accuracy even if they were not familiar with synthetic speech. The rather high success rates of the second test for both groups are due to the fact that the subjects were instructed to label only one out of the three sentences as being appropriate for the dialogue according to their opinion.

	1st test	2nd test	Total
Group 1	98.3	92.7	95.5
Group 2	96.2	90.4	93.3
Total	97.46	91.78	

Table 3 *Success rates of the evaluation test. Listeners familiar with synthetic speech are in Group 1; those unfamiliar are in Group 2.*

In both tests the listeners were provided with a writen version of the test phrase in order to prevent voice quality parameters from being involved in the evaluation. However, the evaluation results indicate that this was not completely avoided, especially in the case of listeners of Group 2.

6. REFERENCES

1. Jean-Yves Magadur, Frederic Gavignet, Francois Andry, Francis Charpentier (1993), "A French Oral Dialogue System for Flight Reservations Over the Telephone", *Proc. Eurospeech 93, Germany, pp. 1789-1792.*
2. Compe, A. Kieβling, T. Kuhn, M. Mast, H. Niemann, E. Noth, K.Ott, A.Batliner (1993), "Prosody takes over: A prosodically guided dialog system", *Proc. Eurospeech 93, Germany, pp. 2003-2006*.
3. King R.W. (1989), "Layout, Processing, User Control and Prosody Insertion in an on-line Synthetic Speech System", *Proc. Eurospeech 89, Paris, pp. 121-124.*
4. Hirschberg J., "Using discourse context to guide pitch accent decisions in synthetic speech", *Talking Machines: Theories, models and Designs, Bailly G., Benoit C. and Sawallis T.R., France, 1992, Elsevier Science Publishers.*
5. Epitropakis, N. Yiourgalis & G. Kokkinakis (1993), "High quality intonation algorithm for the Greek TTS-system", *ESCA Workshop on Prosody, 27-29 September 1993, Lund, Sweden, pp. 70-73.*