Analysis and Modelling of the Carrier Declination for the Greek Language

Georgios P. Giannopoulos, **Stavroula-Evita F. Fotinea**, **Aimilios E. Chalamandaris**, **Theologos D. Athanaselis**, and **George V. Carayannis**

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, Marousi, 151 25, Athens, Greece
ggia@freemail.gr, {evita,achalam,tathana,gcara}@ilsp.gr

Abstract

This paper presents a methodology for analysing and modelling the carrier declination of the affirmative F_0 contours for arbitrarily long texts. During synthesis, formulation of the pitch patterns of arbitrarily long and complex sentences of the Greek language can be rendered with a finite set of intonation models for each intonation word, along with the respective carrier declination trend. This paper provides a non-linear modelling of the declination trend of the Greek language which is compatible with these intonation models. The effect of pause and word prominence into the carrier trend is also investigated. In such cases, the speakers tend to reset their register, and carrier trend locally increases.

1 Introduction

Intonation during Text to Speech Synthesis Sessions for the Greek language can be rendered with a set of intonation models containing only the perceptually significant variability encountered in natural speech [1, 3]. Greek texts preserve three types of expression. The affirmative which is indicated by the full stop (.), the exclamative which is indicated by the exclamation mark (!) and the interrogative which is indicated by the question mark (;). The intonation patterns observed for any of the above cases entail complexity that is mainly attributed to the various syntactic phenomena implemented, and the freedom of the position of the segmental stress within a word.

The example of the Greek sentence: "Το παιδί διαβάζει στο δάσκαλο το γράμμα." (The child reads to the teacher the letter) contains the following 4 IWs (=Intonation Words): "topeð'i ðjav'azi stoð'askalo toɣr'ama". The synthetic F_0 contour can be generated with concatenation of 3 basic IW models [3]: the Introductory (In), the Middle (Mw) and the Conclusive (Cv) model in the order: $In + Mw + Mw + Cv$. This basic rule applies in the case of affirmative sentences of the Greek language for a vast variety of syntactic phenomena. However, in the cases of double stress occurences, pause or triggering of the intrinsic emphasis mechanism, this basic rule is accordingly modified with appropriate inclusion of Pausive (Ps), Emphasis (Em), Double Stress Introductory (Db-I) and Double Stress Conclusive (Db-C) models and subsequent adjustment of the carrier [1, 2, 3]. These intonation patterns can be then applied onto a carrier, the development and trend of which seem to be also very important for the correct acoustic registration of the type of expression and the synthetic speech output naturalness. It has been noted in many languages as well as in Greek [5, 6], that the carrier onto which F_0 curves develop, tends to decrease with time in the affirmative type of expression.

As also proposed in [7], the declination is not modelled by decreasing top and baselines but by a downstep from peak to peak. Hence, the conventional approach, where the carrier is a sloped line and the declination is modelled by changing the slope, is no longer used, since it deprives naturalness. According to the proposed approach the carrier declination is not modelled as a monotonically decreasing line since this is not the case in the natural utterances; instead we attemp to "capture" the natural variability of the carrier declination curve evolution for every IW in the sentence in a non-linear way. During the TTS sessions these data are used to simulate a carrier trend that resembles the natural.

2 Analysis of the F_0 Carrier Trend

In order to "capture" the natural variability of carrier declination curve we have analized the F_0 trend from natural utterances. We formed a corpus of affirmative sentences consisting of one through nine IWs and a corpus of affirmative sentences with one IW with prominence (emphasis) and pause. These corpora were utterated by native Greek speakers and then these data were manually segmented, labelling all the IWs bound-

ISBN 1-876346-48-5

aries. From these labels, the F_0 range for each IW was derived automatically.

2.1 Spoken corpus creation

During the corpus generation we formed affirmative sentences with multivowel words. Each word consisted of at least two leading and two trailing vowels thus, allowing for the complete development of word-level F_0 contours.

Corpus of affirmative sentences comprising 1-9 IWs: the main sentence level parameter that was taken into account in order to investigate the carrier trend was the number of IWs. We formed a corpus of 27 sentences with 1-9 IWs (3 sentences for each case). Then four Greek native speakers with no hearing or speaking disability whatsoever (2 male and 2 female) uttered twice this corpus at their natural speaking tempo; hence we gathered $27 * 4 * 2 = 216$ utterances. These data were used to extract the carrier declination patterns for simple affirmative sentences (affirmative sentences without IWs bearing emphasis or pause).

Corpus of affirmative sentences containing IWs with prominence and pause: the main sentence level parameters that was taken into account was first the position of the word prominence and pause into the utterance and secondly the sentence length. We formed a corpus of 23 sentences (13 sentences containing pause and 10 sentences containing the word prominence phenomenon). Again the same four Greek native speakers as reported above, uttered twice the corpus at their natural speaking tempo and we received $23 * 4 * 2 = 184$ utterances. These data were used to investigate the carrier trend behaviour in word prominence and pause.

2.2 Measurements

The data of all the utterances were manually segmented, labelling all the IW boundaries. With a script in Praat [8] we used theses labels, isolating the waveform of each IW and extracted the pitch contour for the F_0 range ($minF_0$ and $maxF_0$ values) of each IW.

3 Reporting the Results

For each template sentence used, we calculated an averaged carrier trend which was afterwards normalized in the interval $[0, 1]$. Each averaged carrier trend (defined by an upper and a lower limit curve) represents the carrier trend pattern of the corresponding sentence template. During Text to Speech Synthesis Sessions we used these patterns for the generation of the carrier evolution of the synthetic F_0 contour.

In Figures 1, 2, 3, 4 and 5, the Y axis represents the normalised F_0 range and the X axis depicts the locations of the IWs inside the sentence. These patterns describe for each IW the F_0 range that should be used in synthetic F_0 contour generation.

3.1 Simple affirmative senteces

We extracted 9 patterns of the carrier trend for each template sentence spanning from 1 to 9 IWs.

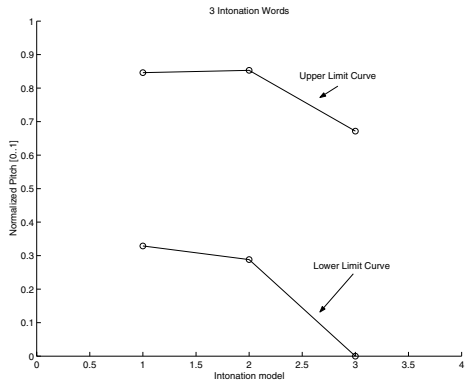

Figure 1: Carrier trend: for small sentences (3 IWs)

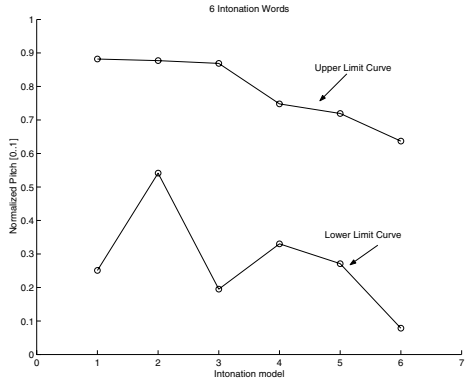

Figure 2: Carrier trend: for medium sentences (6 IWs)

In Figures 1, 2 and 3, one can see the patterns of the carrier trend for small, medium and large simple affirmative sentences. As we observe in figures, the carrier trend has a declination which is not linear and varies between different sentence templates. From the experimentation, we confirm, that the pattern found for the 3 IWs in Figure 1 is similar to the part of the pattern for 6 IWs in Figure 2, comprising the first three words only. Similar observations can be made for 6 and 9 IW patterns. A slight deviation is observed in each sentence final IW, which is quite normal, since the final (conclusive) IW in affirmations is always applied at a lower level (namely, the carrier tends to decrease more in word final positions). This variation of the carrier declination which during Text to Speech Synthesis Sessions is very important for the correct acoustic registration of the type of expression and mostly for the synthetic speech output naturalness.

ISBN 1-876346-48-5

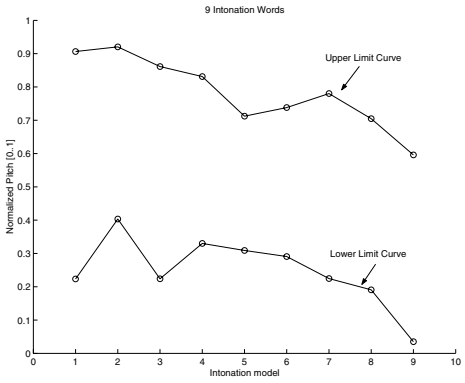

Figure 3: Carrier trend: for large sentences (9 IWs)

3.2 Affirmative sentences containing pause or prominence

In Figures 4 and 5, we see the carrier trend pattern for medium sentences (6 IWs) containing word prominence in the 4th IW and large sentences (10 IWs) containing pause between the fifth and sixth IW.

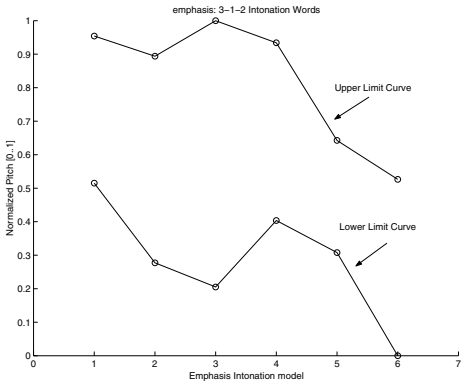

Figure 4: Carrier trend: Sentence with 6 IWs with word prominence (emphasis) in the 4th IW

For small sentences the carrier trend pattern was similar with the pattern of the medium sized simple affirmative sentences (Figure 2). For large sentences (Figure 3) we observe a carrier declination till the intonation word before pause or word prominence; then the speaker resets his register and the carrier trend temporary increases before its declination. The register reset signals the subphrasing of the sentences on the boundary of the pause. The second part of the sentence, after the reset, seems then to be starting again at a pitch level lower than that of the beginning of the sentence. This is mentioned in the rules proposed in [2] and also confirmed by this set of experiments. Note, that the patterns of Figure 5 after pause, namely from sixth to tenth IW, is quite similar with the one depicted in Figure 2 that applies to medium size sentences (the latter found for 6 IWs).

Similar observation could be made in the case of word prominence, namely in (intrinsic) emphasis triggered

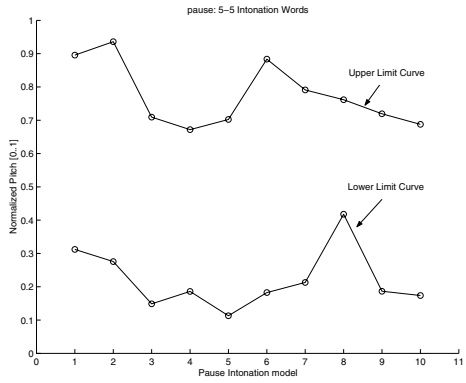

Figure 5: Carrier trend: Sentence with 10 IWs with pause between 5-6 IW

by some special words. Special words [3] are usually particles or conjunctions through which negation, relative comparison or some co-ordination cases are realised. Special words are also considered the words where intentional alteration of the focus, ie. intentionally placed emphasis, exists. It is worth mentioning here that, as also proved in [2] when the emphasis mechanism is triggered, then the rest of the sentence is uttered while the slope declines rather quickly, since the significant part of the sentence has been already conveyed. Yet, the emphasis mechanism is a complicated phenomenon and needs special treatment.

The observations for the carrier reset and subsequent re-initialisation are very important for the correct acoustic registration of the specific phenomenon (pause or emphasis) during TTS sessions.

4 Evaluating the F_0 Carrier Trend in a TTS system

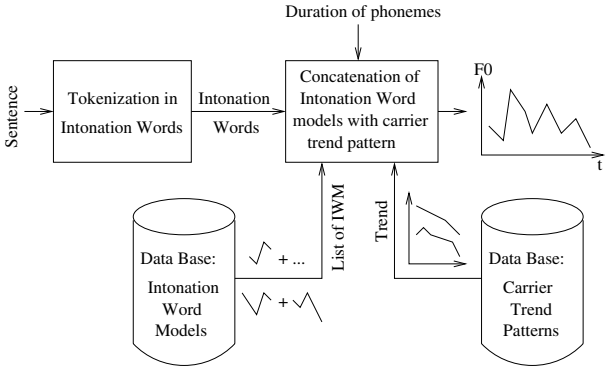

Figure 6: F_0 generator: intonation models + carrier trend

In Figure 6 we present the structure of the F_0 generator where the results of the previous experiments have been evaluated.

The input sentence is tokenised in intonation words.

ISBN 1-876346-48-5

Then for each IW the F_0 contour is generated using the corresponing IW model, the boundaries of F_0 from carrier trend pattern and the duration of each phoneme. There is a database which contains the set of the different IW models and a database with the carrier trend patterns for each sentence template. The F_0 generator parses the input sentence, extracts the corresponding carrier trend pattern and the list of IW models that should be concatenated in order to produce the synthetic F_0 contour. If the sentence to be synthesized is longer than 9 IW appropriate interpolation for the carrier trend is performed. The duration of the phonemes is given from a different module which parses the input sentence using a database of duration rules.

4.1 Experiments and results – mean opinion score (MOS)

In order to examine the consistency and quality of our observations, we carried out a series of experiments on the ILSP's TTS engine [4], using the prosody generating module based on the above described experimentation and results. A group of native greek speaking listeners was used in order to rate the naturalness of the synthetic speech, as far as the prosody was concerned. They were given a small corpus of 9 affirmative sentences, one utterance with a monotonically decreasing carrier trend and another with our non-linear carrier trend patterns, depending on the length of the sentence. The sound samples were mixed and the listeners had to provide a score for each sentence, rating with five the virtually naturally uttered sentence, and with 1 the non-natural and eventually annoyingly spoken sentence.

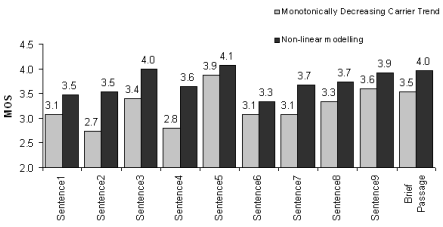

Figure 7: Mean Opinion Score

Finally, a brief meaningful passage was given to the listeners, again in two different versions, one with the modelling and another one with the monotonically decreasing carrier trend and they were asked to rate them again as far as the naturalness was concerned. The mean opinion scores are shown in Figure 7 where one can note the listeners preference for the proposed non-linear carrier declination model.

5 Conclusion

We have presented a non-linear modelling of the declination trend for the Greek Language which consists of concatenating the appropriate F0 models for each intonation word, taken from a set of stored patterns, and of applying at the same time the non-linear model of carrier trend taken again from a set of predefined patterns. These patterns for the intonation words and for the carrier trend were extracted automatically from a spoken corpus by four native Greek speakers. The experiments we carried out, depicted the audience preference for this method over a monotonically decreasing carrier trend on a TTS system developed by ILSP. Further research will be carried out for the cases of interogative and exclamative sentences.

References

[1] Stavroula-Evita F. Fotinea, Michael A. Vlahakis and George V. Carayannis. "Modeling arbitrarily long sentence-spanning F0 contours by parametric concatenation of word-spanning patterns", Rhodes, Greece: ESCA Eurospeech97, Sep 1997, vol 2, pp. 315-318.

[2] Stavroula-Evita F. Fotinea, "Sentence-level Prosodic Modeling of the Greek language with Applications to Text-To-Speech synthesis", PhD Thesis (in Greek), National Technical University of Athens, University Press, 1999.

[3] Stavroula-Evita F. Fotinea, Michael A. Vlahakis and George V. Carayannis. "On the improvement of acoustic registration of tempo and intonation over large sentences for text to speech synthesis in the Greek language", Patra, Greece: Euronoise2001, Jan 2001, pp. 597-607.

[4] Institute for Speech and Language Processing, Text-To-Speech Synthesizer, commercially available, more information on http://www.ilsp.gr/ekfonitis_plus_eng.html, 2003.

[5] Antonis Botinis, "Stress and Prosodic structure in Greek: A Phonological, Acoustic, Physiological and Perceptual Study", Lund: Lund University Press, 1989.

[6] Fujisaki, H., "Prosody, Models and Spontaneous Speech", in Computing Prosody, Springer-Verlag, New York 1997.

[7] Barbara Heuft, Thomas Portele, "Synthesizing prosody: a prominence-based approach", Proc. ICSLP 96, vol. 3, pp. 1361-1364, Philadelphia, PA, USA, October 1996.

[8] Paul Boersma, David Weenink, Praat 4.0 – a system for doing phonetic by computer, http://www.praat.org, 1992-2001.

[9] Christos S. Malliopoulos, "Methods for speech analysis and synthesis with rules for the Greek language", PhD Thesis (in Greek), National Technical University of Athens, University Press, 1999.