An innovative F0 modeling approach for emphatic affirmative speech, applied to the Greek language

Georgios P. Giannopoulos & Aimilios E. Chalamandaris

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, Marousi, 151-25, Athens, Greece
ggia@freemail.gr, achalam@ilsp.gr

Abstract

Prosody generation engine which is is responsible for the naturalness of the synthetic speech, remains one of the most important component of a Text-to-Speech synthesis system. In this paper we present an innovative algorithm for modelling the fundamental frequency F_0 for the Greek language, for sentences containing emphatic segments. The main idea of our approach is the definition of a specific set of intonation word models, derived from a spoken corpus, the use of which is sufficient in modeling the pitch contour of arbitrary long sentences similarly structured. Our method is based on a prosodic unit selection approach. This is tested to ILSP's TtS system for the Greek language Ekfonitis+ [1], which is customized to utter weather reports with virtually natural synthetic voice. The system was designed and trained on a spoken corpus of 120 naturally uttered sentences of weather forecasts, containing emphasis segments and has proved to be very efficient in coping with similarly structured sentences. In the first section of the paper we present a brief review of the existing literature on this field, in addition with analogous approaches for other languages. In the second section we present our method and the design procedure. The last two sections contain the preliminary results acquired from our experiments as well as conclusions and refer to future work that needs to be carried out.

1. Introduction and background

The modeling of the pitch contour in Text-to-Speech synthesis systems remains one of the most challenging areas of research. Prosody and in particular intonation plays a key role in the perceived naturalness of synthetic speech. Many different approaches are suggested for coping with this topic.

Some algorithms are trying to simulate the human production mechanism using the filter-based model of Fujisaki [2] (Japanese TtS). This pioneer work has been applied to many languages, including German [3], English, Estonian, Greek etc. [2]. Other algorithms follow the ideas of Pierrehumbert's thesis for American English [4] in which we can describe (prosodically label) an intonation contour as a series of high and low tones using ToBI (Tones and Break Indices) transcription. Using Pierrehumbert theory many researchers developed ToBI transcriptions for several languages including one for the Greek language (Gr-ToBI) [5]. In this approach, F_0 generation module uses a target interpolation scheme with accent and boundary markers that are ToBI labels. Targets are placed with reference to syllable structure, within a pitch range specified by top and base lines [6]. A recent work [7] highlights several flaws of ToBI system that have limited its acceptance and application. Many algorithms can also be classified as data-driven methods. In these methods we have a phonetically and prosodically segmented and labeled spoken corpus and in a TtS system, we predict the F_0 using parameters and data extracted from the labeled corpus. Many TtS systems for several languages use data-driven modeling (aka corpus-based synthesis). For example the Tilt model for English [6, 8] uses accents that are marked in the speech database and are parametrized using a parabolic approximation. As far as Spanish is concerned, another method has been suggested applying statistical modelling of F_0 contours using Bezier functions [9], while for French we have two data-driven approaches for synthetic F_0, the SFC trainable prosodic model [10] and the INTSINT prosody annotation scheme together with the perception stylized modelling of F_0 through MOMEL algorithm [11].

For the Greek language there have been several different approaches. In [12, 13] F_0 generation is corpus-based using syllable's features, associated turning points (F_0 maxima and minima) and four declination lines. In a recent corpus-base approach [14], the prosody module is using high-level linguistically annotated corpora, ToBI and CART. Greek texts preserve three types of expression: the affirmative, the exclamative and the interrogative. 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. In the Greek language the position of the segmental stress within a word can only be attributed to the vowel of one of the 3 final syllables of the word, and it is annotated in the orthographic representation. In our case, we deal only with affirmative speech sentences.

2. Our approach

Our approach is based on the observation that intonation during Text-to-Speech sessions for the Greek language can be rendered with a set of intonation models containing only the perceptually significant variability encountered in natural speech [15, 16], along with the respective carrier declination trend [17]. In this paper we extend this method to the modelling of emphatic speech. The main idea of our approach is the prosodic unit selection approach, according to which the pitch contour of any affirmative sentence can be modeled by means of a set of predefined curves, corresponding to different intonation words, by simple concatenation and interpolation of the appropriate models. The selection of the optimal prosodic segments is performed via a cost function, which considers segmental and supra-segmental parameters such as the length of the IW, context, punctuation marks, position in the sentence, part of speech etc. The general structure and function of our system becomes clearer in the following paragraphs.

3. Analysis of Spoken Corpus

The spoken corpus we used for training is consisted of 110 sentences mainly revolving around weather forecast topics and news, uttered in a natural tempo by a professional female speaker. The weather forecasts were chosen intentionally because of the emphatic segments they include.

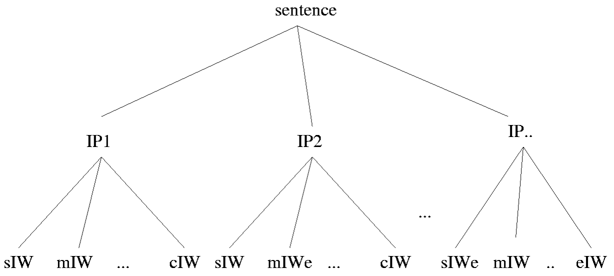

Figure 1: *Decomposition of a sentence into IPs and IWs*

The phonetic and prosodic annotation of the corpus was performed semi-automatically, in order to ensure that there were no mistakes in the segmentation process, which would affect our models. The boundaries of the IPs (=Intonation Phrases or breath groups [18]) and the boundaries of the IWs (=Intonation Words or stress groups [18]) as it is depicted in Figure 1 and Table 1 were also manually defined, since we have not reached yet a specific set of objective rules that would allow us to automatically annotate the IWs. Nevertheless, this is a task we aim to fulfill in the immediate future, since such tool would allow us to investigate fast large corpora and different contexts. The program used for performing the segmentation task was Praat [19]. In most of the times emphasis appears with pitch rising, which affects the neighboring IWs pitch contours, with vowels durations longer than average. There were cases where emphasis was followed by a short pause, suggesting a new intonation phrase boundary. In general we observed that there were quite a few non-linear phenomena for attributing emphasis to a word, and these were the most difficult parts to define and foresee in our application. In Table 1 one can observe the final set of intonation word models we derived.

Table 1: *Intonation Words (IWs) Annotation formulations*

model	observations
sIW	Starting IW – beginning of IP
mIW	Middle IW – inside of IP
cIW	Continuation IW – end of IP, follows continuation of information
eIW	End IW – end of IP, sentence closure
$sIWe$	Starting IW Emphasis – beginning of IP
$mIWe$	Middle IW Emphasis – inside of IP
$cIWe$	Continuation IW Emphasis – end of IP, follows continuation of information
$eIWe$	End IW Emphasis – end of IP, sentence closure

These results derived from manual investigation of the F_0 evolution within the used spoken corpus and are in accordance to previous literature [20]. The Greek sentence: /Το παιδί διαβάζει στο δάσκαλο το γράμμα/ (The child reads the letter to the teacher) contains the following 4 IWs (=Intonation Words or stress groups [18]) in IPA (International Phonetics Alphabet): "/topeðí ðjavázı stoðáskalo toɣráma/. The synthetic F_0 contour can be generated with concatenation of 3 basic IW models [16]: the introductory-Start (sIW), the Middle (mIW) and the conclusive-End (eIW) model in the following order: $sIW + mIW + mIW + eIW$ (Figure 2).

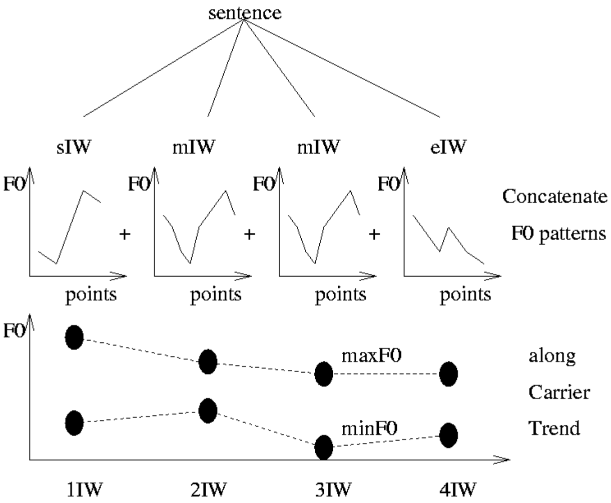

Figure 2: F_0 - *carrier: /topeðí ðjavázı stoðáskalo toɣráma/*

This rule can be applied in most cases of affirmative sentences in Greek language, for a vast variety of syntactic phenomena. Each IW model is a vector of F_0 values for the phonemes inside the IW. The F_0 values of the IW models are extracted from the annotated spoken corpus. We can generate the synthetic F_0 for this sentence by simply concatenating and interpolating these four pitch patterns and appling a carrier trend over them [17]. The carrier trend pattern for this sentence of the 4 IWs, consists of four points defining the pitch contour bandwidth for each IW F_0 model. In the above example the sentence, which is also one single intonation phrase (IP), is decomposed directly to four intonation words (IWs).

Table 2: *1693 IWs F_0 patterns in 8 model classes extracted*

model class	F_0 points vectors	model emphasis class	F_0 points vectors
sIW	432	$sIWe$	52
mIW	549	$mIWe$	183
cIW	235	$cIWe$	57
eIW	161	$eIWe$	24

3.1. Extracting F_0 and carrier declination patterns

Our annotated speech corpus contains 110 sentences which consist of 498 IPs, 1693 IWs, 13931 phonemes and 360 pauses. Each sentence contains on the average 4.5 IPs, 15 IWs, 130 phonemes and 3 pauses. Using a Praat Script [19] we extracted automatically from the annotated spoken corpus all F_0 model patterns for each IW model of the Table 2 and all carrier declination patterns for each IP respectively. For the actual pitch values we made use of the implemented pitch extraction algorithm in Praat [19, 21] which provides satisfactory results with speech signals.

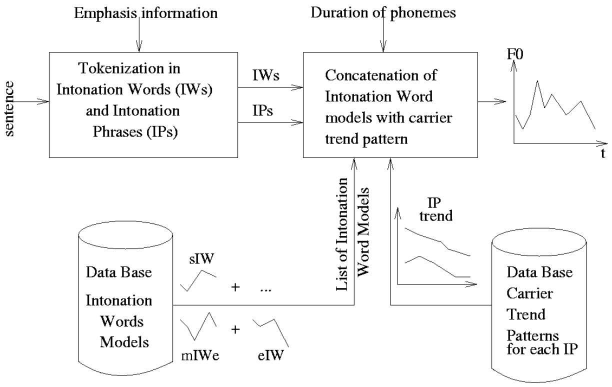
Emphasis information
Duration of phonemes
sentence
Tokenization in Intonation Words (IWs) and Intonation Phrases (IPs)
IWs
IPs
Concatenation of Intonation Word models with carrier trend pattern
F0
t
List of Intonation Word Models
IP trend
Data Base Intonation Words Models
sIW
+ ...
mIWe
+
eIW
Data Base Carrier Trend Patterns for each IP

Figure 3: F_0 *generator: intonation models + carrier trend*

4. Designing the F_0 component of Prosody

The general structure of the F_0 modeling engine is illustrated in Figure 3. The input sentence is tokenized in intonation phrases (IPs) and then each IP is additionally tokenized to a lower level of intonation words. For each IW we select the appropriate model pattern from the database acquired during the training stage, along with the corresponding carrier trend, which is, in other words, the bandwidth of the pitch contour within every intonation word. For every intonation word in the data-base we store not only the pitch contour samples but store phonetic information, contextual parameters, such as stress position, length of the word, similarity of phoneme types according to their manner and place of formation in the vocal tract, as well [22]. This information is important for the algorithm among the intonation model patterns which is performed by means of the optimal path of a cost function. The fine tuning and the parameters used in the cost function need to be additionally investigated. As it is currently implemented, in case the same intonation word is found in the database, the prosodic unit selection algorithm is biased to choose entire IW models, otherwise, the sub-optimal intonation model selected is the one which best fits its characteristics. It is obvious that the larger the training corpus is, the better performance we achieve, taking into account that the larger corpus provides not only better coverage of the intonation models but more variety and naturalness as well.

5. Application of F_0 in a TTS-system

In order to investigate the performance of our algorithm we incorporated our F_0 generator component to the TtS application that has been developed at ILSP. After the training session of the system, during which only 110 sentences have been used, we tried to re-synthesize 10 sentences that were not included in the training data-set. The sentences were chosen randomly. Most of the results were virtually as natural as the original utterances. One can observe the original and the synthetic F_0 contour for the sentence in Figure 5. In Figure 4, we see an example of resynthesis of a sentence from the training corpus. The overall results, although preliminary, were very promising and exceedingly satisfactory. However there were cases where the selected intonation word models did not fit optimally the target prosody. This was mainly due to the fact that there was not enough coverage of all possible intonation words in every category in addition due to the fact that the search algorithm still needs refinement, because of its lack of robustness. Special attention needs also to be given to non-linear phenomena, as already mentioned above, where for example emphasis is depicted with arbitrary pauses by the speaker. These could possibly be dealt by other search algorithms, such as CART.

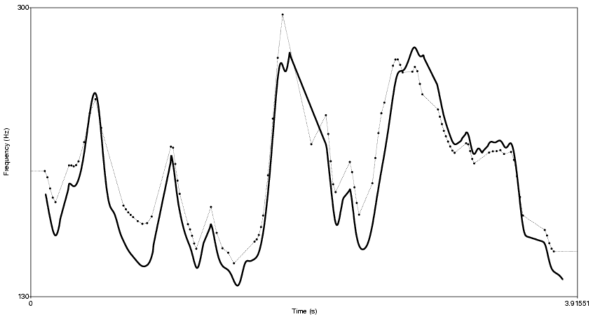

Figure 4: *Example of* F_0 *resynthesis. The line with solid black line is the natural* F_0 *and the dotted line is the synthetic. In this sentence we have IP1: /*apoávrıo/, IP2: /cemeθávrıo/ and IP3: /tafenómena θaıpoχorísun/. The concatenated IWs models are: sIW, sIW and $sIW + eIW$.

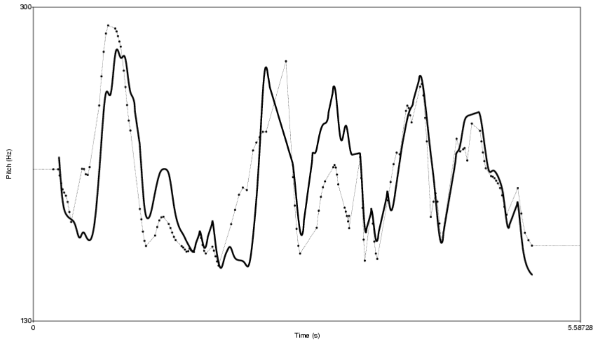

Figure 5: *Example of* F_0 *synthesis. The line with solid black line is the natural* F_0 *and the dotted line is the synthetic. In this sentence we have IP1: /taedonótera fenómena apotovráðı/ and IP2: /θastréfode stısıpólıpes perıoçés/. The concatenated IWs models are:* $sIWe + mIW + cIW$ *and* $sIW + mIW + mIW + eIW$.

6. Discussion

The aforementioned method proved to be very efficient in the framework of our TtS system for Greek language. To this end contributed the fact that Greek language is not a tonal language, but instead a straightforward one, as its orthographic representation depicts the word stress. It has been shown that the definition of only eight intonation word models in our case was efficient and complete enough to model the majority of the uttered sentences. However, we have not yet ensured that this approach, which was successfully applied to the speaker we used, will be successfully applied to other speakers as well. A larger scale investigation must be carried out in order to make sure that our method applies with the least possible exceptions for most Greek native speakers. It is also worth mentioning that the aforementioned research addresses only affirmative speech and not other types, such as exclamative or interrogative. These types need further investigation and possibly more complex models to be described.

7. Conclusions - Future Work

In this paper we presented an innovative method for modeling the F_0 contour in emphatic sentences for a Greek TtS system. It consists of a data-driven method, that aims to define a finite set of intonation models, which can be combined via a cost function and efficiently simulate the F_0 contour for arbitrary long sentences of similar structure. The algorithm was designed and applied to the Greek TtS system developed at ILSP, and has proved to be very efficient in the specific framework. Preliminary experiments have shown that this algorithm performs exceedingly well in collaboration with a TtS, which incorporates unit selection in its function. Nevertheless, more work needs to be performed in order to include more cases of prosodic patterns as well as to optimize the pattern selection algorithm, which is responsible for attributing the optimal models to each case. Our immediate future plans are to investigate non-linear methods for optimal path search, such as regression trees CART etc. Last but not least, a field worth investigating is the automatic prosodic annotation of a spoken corpus, providing automatically prosodic cues, according to our method. A preliminary research has been carried out, defining some first objective metrics for automatically discriminating prosodic cues, however further refinement and research is necessary.

8. Acknowledgements

This work has been carried out in the framework of the speech synthesis project in the premises of the Institute for Language and Speech Processing. We would like to thank all members of ILSP's speech synthesis team. In particular, we would like to thank Dr. S. Raptis, P. Tsiakoulis, S. Karabetsos and C. Kalimeris for their valuable help and advice.

9. References

[1] Institute for Speech and Language Processing, Text-To-Speech Synthesiser, commercially availabe, more information on *http://www.ilsp.gr/ekfonitis_plus_eng.html*, 2003.

[2] Fujisaki Hiroya, "Information, Prosody, and Modeling – with Emphasis on Tonal Features of Speech", In *Proceedings of Speech Prosody 2004*, Nara, Japan, 1-10.

[3] Mixdorff Hansjorg, "An Integrated Approach to Modeling German Prosody", Post-Doc Thesis, Technical University of Dresden, 2002

[4] Pierrehumbert, J. B., "The Phonology and Phonetics of English Intonation", PhD Thesis, MIT, Published by Indiana University Linguistics Club, 1980.

[5] Arvaniti A., M. Baltazani, "Greek ToBI: A System For The Annotation Of Greek Speech Corpora". In *Proceedings of Second International Conference on Language Resources and Evaluation*, LREC2000, vol 2: 555-562.

[6] Ann K. Syrdal, Gregor Mohler, Kurt Dusterhoff, Alistair Conkie, and Alan W. Black., "Three methods of intonation modeling". In *Third International Workshop on Speech Synthesis, Jenolan Caves*, Australia, 1998.

[7] Wightman W. Colin, "ToBI Or Not ToBI?", In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France.

[8] Taylor P., "Analysis and synthesis of intonation using the tilt model". In *Journal of the Acoustical Society of America*, 2000, 107(3):1697-1714.

[9] Cardenoso V., Escudero D., "Statistical Modelling of Stress Groups in Spanish", In *Proceeding of Speech Prosody 2002*, Aix-en-Provence, France.

[10] Bailly Gerard, Holm Bleicke, "SFC: A trainable prosodic model", In *Speech Communication 46*, 2005, pp. 348-364.

[11] Hirst Daniel, Di Cristo Albert, Espesser Robert, "Levels of representation and levels of analysis for intonation" In *M. Horne (ed) Prosody : Theory and Experiment* Kluwer, Dordrecht, 2000.

[12] Epitropakis G., Yiourgalis N., and Kokkinakis G., "High Quality Intonation Algorithm for the Greek TTS-System", In *ESCA Workshop on Prosody Lund*, Sweden, Sept. 1993, 27-29.

[13] Galanis D., Darsinos V., Kokkinakis G., "Modeling of Intonation Bearing Emphasis for TTS-Synthesis of Greek Dialogues", In *Proceedings of The 4th International Conference on Spoken Language Processing*, 1996, vol. 3.

[14] Xydas G., Spiliotopoulos D. and Kouroupetroglou G. "Modelling Improved Prosody Generation from High-Level Linguistically Annotated Corpora", In *IEICE Trans. Inf. & Syst., Special Issue on Corpus-Based Speech Technologies*, Vol.E88D, No.3 March 2005, pp 510-518.

[15] 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", In *ESCA Eurospeech97*, Sep 1997, Rhodes, Greece, vol 2, p. 315-318.

[16] 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", In *Euronoise2001*, Jan 2001, Patra, Greece, p. 597-607.

[17] Giannopoulos P. Georgios, Stavroula-Evita F. Fotinea, Chalamandaris E. Aimilios, Theologos D. Athanaselis and George V. Carayannis, "Analysis and modelling of the Carrier Declination for the Greek language", In *Proc. of the 15th International Congress of Phonetic Sciences - ICPhS03*, 3-9 August 2003, Barcelona, p. 555-558.

[18] Botinis Antonis, Granstrom Bjorn, Mobius Bernd, "Developments and paradigms in intonation research", In *Speech Communication 33*, 2001, p. 263-296.

[19] Boersma, Paul & Weenink, David (2005). Praat: doing phonetics by computer (Version 4.4) Computer program. Retrieved December 19, 2005, from http://www.praat.org/

[20] Dutoit Thierry, 1997. An introduction to Text-to-Speech synthesis, Kluwer Academic Publishers, ISBN: 0792344987.

[21] Boersma, Paul, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ration of a sampled sound.", In *Proceedings of the Institute of Phonetic Sciences*, 1993, University of Amsterdam, 17:97-110

[22] Founda M., Tambouratzis G., Chalamandaris A., and Carayannis G., "Reducing spectral mismatches in concatenative speech synthesis via systematic database enrichment", In *Proc. Eurospeech 2001*, Aalborg, Denmark, 2001