Modeling arbitrarily long sentence-spanning F0 contours by parametric concatenation of word-spanning patterns.

Evita F.Fotinea, Michael A.Vlahakis† and George V.Carayannis†
National Technical University of Athens, Electrical and Computer Engineering Dpt, Division of Computer Science,
Digital Signal Processing Lab., 9, Heroon Polytechniou St. Zographou 157 73, Athens, Greece.
e-mail: evita@ilsp.gr or efotin@image.ntua.gr
†Institute for Language and Speech Processing, 22, Margari St. Athens 11525, Greece
Tel : +301 6712250 Fax : +301 6741262, E-mail: gcara@ilsp.gr.

ABSTRACT

Modeling F0 contours of arbitrarily long and complex sentences of the Greek language may prove to be a difficult task if one considers the various parameters involved, namely focus, position of the prominent vowel within words, syntactic structure and type of expression. None the less, this complexity may be significantly reduced if the expressive requirements of the application area in mind are taken into account. Study of the expressive requirements of the information broadcasting applications revealed that the affirmative type of expression is heavily used and regardless of size and complexity, each sentence-spanning contour may be composed of only four word-spanning patterns. This result not only leads to significant savings in the resources required for a natural sounding speech output but also indicates a highly structured intonative component.

1. INTRODUCTION

Single word utterances are extensively used in everyday discourse and provide the simplest form of an oral message. Earlier work [1], reported on the modeling of the F0 pattern by which these utterances convey affirmative statements. The shape of the pattern depends on the position of the prominent vowel as well as the number of vowels preceding (which are called leading) and the number of vowels that follow the prominent one (which are called trailing). Study of the oral inflection of Greek verbs, a slightly more complicated utterance, revealed also that the patterns reported in [1] may be successfully used to model the inflectional environment. The fact that the structure of each entity mentioned above is merely an apposition (of similar objects-verbs), a substructure often found in the language, suggested that large sentences should also be examined. Are the above mentioned patterns encountered in other speech structures as well or do sentence-level F0 patterns develop in a different way due to sentence-level factors ? In order to answer this question an extensive experiment was formulated by taking into account the following prosodic attributes and parameters.

2. SENTENCE LEVEL PROSODIC ATTRIBUTES

Word Prominence: Even though the "Emphasis" or "Word Prominence" mechanism is examined by other researchers [2] and is well understood, it is reconsidered here mainly for the following reason. In our experiments it was found that even though the sentences used, were uttered with no emphasis placed whatsoever on any of the words used, there do exist certain syntactic structures realized only through Word Prominence. A distinction should be therefore drawn between Word Prominence as a mechanism for intentionally altering the focus within a sentence and as an intrinsic property of certain syntactic structures. In this paper only the second case is considered.

Type of expression: It is known [3], that some of the most important applications of Speech Synthesis lie in the information broadcasting area. In this respect, only affirmative statements were considered in this experiment because the results obtained could be easily integrated in our current TTS system and possible applications could immediately be addressed.

3. SENTENCE LEVEL PARAMETERS

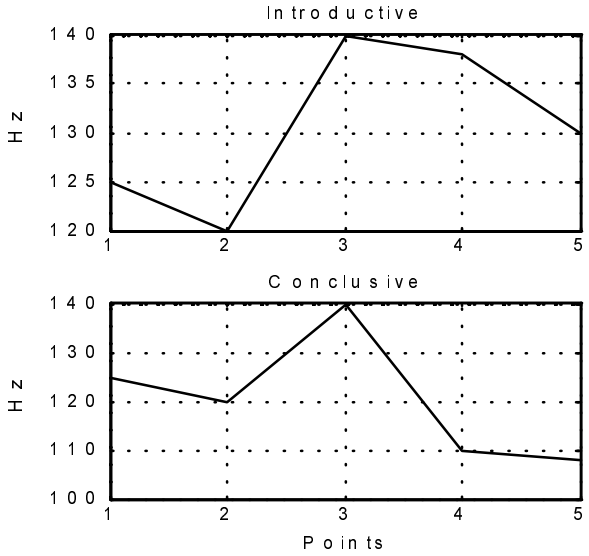
Syntactic structure: It is the main sentence level parameter that was taken into account in order to investigate the dependence of the F0 pattern structure on the syntactic structure of the utterance. The list of the syntactic structures used is exhaustive and covers a vast variety of texts that may be encountered in the application areas of interest. The presentation develops in three levels and for each level, the syntactic phenomena investigated are reported in short below.

Phrase level.

- Agreement between article and noun
- Agreement between adjective and noun
- Noun phrase. Noun + modifier
- Noun phrase. Noun + complement
- Coordination Depending on the kind of conjunction, the "and", "or", "neither nor" and "either or" types of coordination were investigated.

Sentence level.

- Tense.
- Arity: The following cases were investigated
  Subject + verb. (SV).
  Subject + verb + object. (SVO).
  Subject + verb + object + direct object (SVOO).
- Negation. Δεν (do/does not) and Μη/Μην (do not).
- Quantifiers ("some").
- Adverbials.
- Passive voice.
- Clitics. Both enclitics and proclitics were examined.



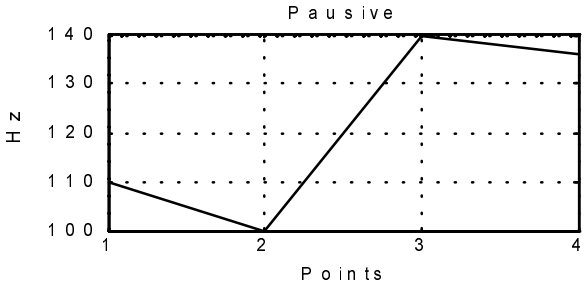


Figure 1

- Reflexives.
- Comparison.: Relative, Absolute and Coincidence.
- Extraposition: the terms of the SVO structure are reordered without any modification of the meaning.
- Extensions of the basic SVO structure, including sentences constructed by adding a verb or even more terms to all of the above phrases, as well as sentences containing prepositional phrases.

Connected sentences level.

- Coordination: Depending on the kind of conjunction used, several cases were investigated
- Subordination: Depending on the kind of conjunction several cases were also investigated
- Subordination.

Speaking rate: This parameter affects the shape of the patterns. Especially during fast rated speech, F0 patterns do not fully develop due to poor articulation. In our experiment the material was uttered in normal rate, the normal rate being considered as the target speech rate of our TTS system.

Pauses: Pauses are inserted either at the position of a comma or a fullstop or intentionally at any other position within an utterance in order to render the speech message clear. In this experiment, the speakers were left free to pause at any time it seemed convenient in order to produce as clearer speech material as possible.

4. WORD LEVEL PARAMETERS

The only parameter that was taken into account is the position of the prominent vowel of a word pronounced within a sentence. In Greek, only one of any of the last three vowels may be prominent within a word. The range and the shape of the F0 patterns spanning a single word, are affected by both the position of the prominent vowel, and the position of the prominent vowel of the preceding and the following word.

5. FORMULATION OF EXPERIMENTS

For the actual experiment, a set of prototype template sentences - each corresponding to a specific syntactic structure - was formed by using multivowel words. Each word consisted of at least two leading and two trailing vowels thus, allowing for the complete development of the word-level F0 patterns. Our investigation took into account a great number of possible variations [2] of the template sentence. Moreover, more complicated structures were uttered and examined by concatenating a great variety of prototype template ones.

The sentences were uttered twice, in a relaxed manner by a male as well as a female native Greek speaker. In this way, a total of 914 utterances was obtained. The structure of each sentence as well as the expression conveyed (non emphatic affirmations) was informally confirmed by Greek students unfamiliar whatsoever with the speakers. For each sentence, the F0 curve was extracted, observed, modeled and the resulting model quality was informally evaluated.

6. PRESENTATION OF THE RESULTS

6.1. Modeling of the word-spanning patterns

According to the thorough examination of the utterances used for this experiment the following patterns were extracted and modeled using piecewise linear approximation.

Introductive [In]: The pattern is used for the introduction of a main or a subordinate clause and is used in every non final and non pre-pausal position. It conveys the message that more information follows and the sentence has not terminated. In figure 1 the piecewise linear approximation to this pattern is shown. Point 1 coincides with the beginning of the word, point 2 with the beginning of the syllable that contains the prominent vowel, point 3 with the beginning of the ultimate, point 4 with the middle with the ultimate vowel while 5 with the end of the word.

Emphasizing [Em]: The pattern is encountered at a sentence non final position and its mission is the intentional alteration of the focus. Even though it is similar to the Introductive pattern, its alignment is significantly different. Point 1 coincides with the beginning of the word, point 2 with the beginning of the

prominent vowel, point 3 with the end of the ultimate vowel, point 4 with the beginning of the ultimate while point 5 with the end of the word.

Pausive [Ps]: This pattern is encountered at a pre-final, pre-pausal position and in figure 1 its piecewise linear approximation is shown. Point 1 coincides with the beginning of the word, point 2 with the beginning of the ultimate and point 3 with the end of the word. In cases where the duration of the ultimate vowel is severely lengthened then the part 3-4 is also developed, spanning the second half of the ultimate vowel.

Conclusive [Cv]: This pattern is always used at the sentence final position and conveys the message that the utterance has terminated. In figure 1 its piecewise linear approximation is shown. Point 1 coincides with the beginning of the word, point 2 with the beginning of the prominent vowel, point 3 with the end of the prominent vowel, point 4 with the middle of the syllable following the prominent one while point 5 with the end of the word. It is worthwhile noting that in the case of an affirmative sentence composed of just a single word, the sentence spanning contour is reduced to this word spanning pattern.

Parameters affecting the shape of the patterns.

The shape of the patterns depend on the length of each word as well as on the position of the prominent vowel. Moreover, they depend on the length as well as on the position of the prominent vowel of the word that follows. In case that the prominent vowel is found at the ultimate and the prominent vowel of the word that follows is found at the beginning, a pattern linkage phenomenon is observed. In some other instances the patterns are not fully developed and only the pieces of the patterns in the vicinity of the prominent vowel seem to be generated.

6.2. Modeling of the sentence spanning contour

It was found that as far as F0 is concerned, the template sentences may be classified into two categories. For the first, comprising a total of 10, the trend of the F0 contour is flat. For the second, a falling F0 trend starting from the position of a special word (usually a particle or conjunction "δεν/not", "μην/do not", ούτε/neither", etc.) up to the end of the utterance is observed. This is the way that a negation, comparison or disjunction is clearly conveyed. It seems that the emphatic manner in which these words are uttered, is intrinsically unavoidable even though the sentences examined were uttered with no emphasis placed on any word used. The suggestion is further supported by the fact that the utterance of other sentences in which the focus was intentionally placed on some words demonstrated the same pattern. The syntactic structures that fall into this category are presented below.

(a) Negation.

In negative sentences the peak of the pitch contour coincides with the position of the negative adverb "δεν/not" or "μην/do not". From that point on, the F0 decreases until the end of the sentence. On the other hand, in double negation, negative sentences, the F0 decreases just after the second word used for the double negation. For example in [Ο διαπραγματευτής δεν δίνει καμμία λύση/The negotiator does not give any solution] the F0 is decreased just after the word "καμμία" (any).

(b) Relative comparison.

The observation made for the case (a) holds also here A peak is observed at the position of the comparative "πιό / more" or the prominent vowel of the comparative adjective "μεγαλύτερος/older" and from that point on the F0 decreases.

(c) Coordination.

The observations made for cases (a) and (b) also hold here for the conjunction words "είτε/either-or, ή/or, ούτε/neither-nor" etc.

(d) Subordination.

In both cases of conjunction words and relative pronouns the values of F0 of the intonation word that precedes the comma (pause) increase only at the last syllable, regardless of the position of the prominent vowel.

(e) Clitics.

In the case of enclitics where the enclitic word resulted in the appearance of a second stress within the same grammatical word, the resulting pattern may be broken down into two distinct patterns and modeled like in the case of two consecutive intonation words. In this case the shape of the resulting patterns depend on the positions of the prominent vowels. For example "Το μάθημά μου τελείωσε" (My lesson is over) may be broken down into the following intonation words and modeled accordingly "Το μάθη - μάμου τελείωσε".

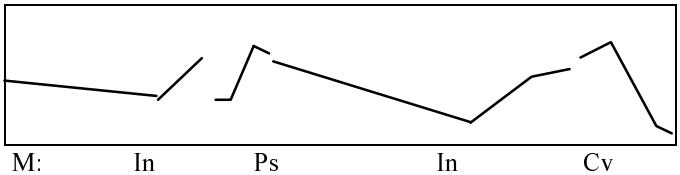
In the case where no second stress appears, the clitic word is simply incorporated into the intonation word to which it refers.

6.3. Implementation

Below, some qualitative examples on the implementation of the sentence spanning contours are given.. Grammatical words (G), Translation (T), Phonetic description (P), Intonation words (I), F0 and the Model sequence (M) are depicted in all examples.

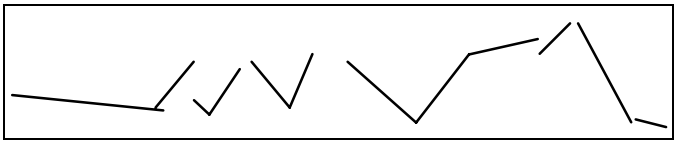
Example 1.

G: διαπραγματευτής φεύγει, γιατί η διαπραγμάτευση λήγει.
T: the negotiator leaves because the negotiation ends.
P: o djapraγmateft'is f'evγi, jat'i 'i djapraγm'atefsi l'iγi.
I: οδιαπραγματευτής φεύγει, γιατίηδιαπραγμάτευση λήγει.



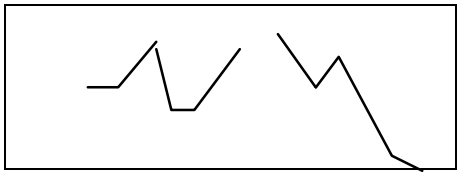
Example 2.

G: ο διαπραγματευτής ξέρει, γιατί η διαπραγμάτευση λήγει.
T: the negotiator knows why the negotiation ends.
P: o djapraγmateft'is ks'eri, jat'i 'i djapraγm'atefsi l'iγi.
I: οδιαπραγματευτής ξέρει γιατί ηδιαπραγμάτευση λήγει.



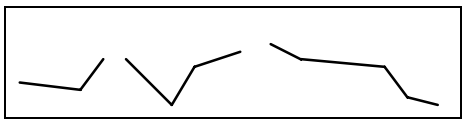
M: In In In In Cv

Example 3.
G: το μάθημά μου τελείωσε.
T: my lesson is finished.
P: to m'aθim'a mu tel'iose.
I: τομάθημάμου τελείωσε.



M: In In Cv

Example 4.
G: το παιδί διαβάζει το μάθημά του.
T: the child studies its lesson.
P: to peδ'i δjav'azi to m'aθim'a tu.
I: τοπαιδί διαβάζει τομάθημάτου.



M In In Cv

For the actual implementation the following resources are required :

- A dictionary containing the indeclinable words.
- The piecewise linear approximation models of the word-spanning patterns.
- The rules for implementing the sentence-level trends.

The algorithm proceeds as follows on a per sentence basis :

- The orthographic text is converted to phonetics.
- The commas and prominent vowels are located.
- The indeclinable words are located and identified.
- The intonation words are formed.
- The sentence level trends are implemented.
- The Conclusive and Pausive patterns are applied.
- The Introductive patterns are applied.

The quality of the resulting models was informally evaluated using the methods described below. At first, each sentence was LPC coded and resynthesized using a lattice synthesizer, the natural F0 being replaced with the one represented by the model introduced herein. Then the TTS system was used. In all cases, both the expression conveyed as well as the proper acoustic registration of the position of the special words or the position of the prominent vowels were evaluated. The speech generated was not monotonous, fluent and therefore easy and pleasant to listen to.

7. CONCLUSION

In this paper it is proposed that arbitrarily long F0 contours expressing affirmative statements, can be modeled by using only four word-spanning patterns, a small size dictionary and an emphasis rule. The result, although simple, proposes an interesting principle. During a discourse, new information is introduced by using the "Introductive" pattern and prompting the listener that more information follows until the speaker decides either to pause or complete. In either case the corresponding pattern is used. In cases where the focus needs to be altered, a peak of the F0 curve renders a certain word prominent. From that point on up to the end of the sentence, the F0 deviation is severely decreased thus contributing more to the prominence of the word. The acoustic effect perceived, is that further prosodic elaboration is not required because the significant part of the meaning is already transmitted.

The approach does not require any extensive syntactic and/or morphological analysis [6], may still cope with a great variety of arbitrarily long texts and is suitable for real time implementation. Even though elaborate expressive phenomena may not be produced the speech obtained is natural, pleasant not causing any fatigue whatsoever.

REFERENCES

[1] M.Vlahakis, E.Fotinea, G.Carayannis (1994)."Word level stress as an intrinsic property of word level prosody and pitch modeling through a pencil of functions", EURASIP/EUSIPCO-94, 13-16 September 1994, Edinburgh, Scotland, Signal Processing VII "Theories and Applications", Vol.I, pp.1-3.

[2] Botinis (1989). Stress and prosodic structure in Greek. Lund university press.

[3] Lawrence R. Rabiner, Fellow IEEE, (1994). Applications of Voice Processing in Telecommunications. Proceedings of the IEEE, Vol 82, No 2, February 1994.

[4] Malavakis (1985). Syntactic and intonative phenomena. Proc of the Vith Congress of Greek linguistics, Thessaloniki, 41-45.

[5] Malavakis (1987). Intonation patterns in Greek. 11th ICPhs-1987 Tallinn USSR.

[6] George Epitropakis, Nickolas Yiourgalis, George Kokkinakis (1993). Prosody control of TTS-Systems based on linguistic analysis. Eurospeech 93.