High Quality Intonation Algorithm for the Greek TTS - System

George Epitropakis, Nikos Yiourgalis, George Kokkinakis
Wire Communications Laboratory,
University of Patras, Greece

ABSTRACT

This article presents a complete algorithm for the generation of intonation (F_0 contours) for the Greek Text-To-Speech system, based on a multi-layer label structure that is constructed for the phonemes representing the input text. This structure consists of the phoneme's distinctive features, the position of the syllable that the phoneme belongs to, the prosodic label of the word that the phoneme belongs to, and the phoneme's prosodic context in the sentence. According to the contents of that structure, the algorithm assigns to each phoneme of the input sentence a target pitch level to be reached either at the beginning, or the middle, or the end of the phoneme. When all the phonemes have been assigned the appropriate F_0 level, the overall pitch contour is constructed by linear interpolation between the successive F_0 levels. Although the method proposed seems to be a rather abstract approach, it takes into consideration linguistic, phonotactics and metrical constraints of the input and not linguistic constraints alone. In addition, the method is especially suited for languages, such as Greek, which are inflectionally rich and have great freedom of word-order.

PURPOSE AND SCOPE

To improve the naturalness of the synthetic speech, numerous methods have been proposed, all of which differ in both the phonetic and the phonological representation of intonation (Hirst, 1992). However, from the most radically concrete positions of Pierrehumbert (1980) to the abstract representations of 't Hart, Collier, and Cohen (1991), and to the even more abstract approaches of Carlson and Granstrom (1973), it is evident that the choice of a particular prosodic model strictly depends on its purpose. The aim of our model is to produce high quality intonation contours for the Greek TTS-system, developed in Wire Communications Laboratory for the practical application of one-way information broadcasting i.e. reading newspapers and books.

To this end, the model is based on the results obtained from an extensive study of Greek intonation that was carried out on text material recorded from just one speaker, who was strictly instructed to speak in a "neutral" reading style. The pitch contours extracted from natural speech were "stylized" by following the guidelines of the "perceptual approach", developed at the Institute for Perception Research of Eindhoven (IPO), but no standardized pitch movements and pitch configurations were extracted. On the contrary, our methodology is based on the description of all the turning points produced in the "stylized" pitch contours in terms of discrete textual phenomena.

To extract and formulate the rules that establish the correspondence between the linguistic constraints and prosodic properties of any input sentence, we first drew up a list of 200 test sentences that covered the greatest possible syntactic structures of the language. The material to be recorded was created by combining simple types of Noun Phrases and Verb Phrases, and by increasing successively their complexity, always producing meaningful sentences. In addition, sentences with identical syntactic structures, but with different numbers of syllables were created. These sentences were recorded by 1 male speaker who was asked to speak in a neutral manner.

Through the steps of "perceptual equivalence" methodology ('t Hart, Collier and Cohen, 1991), a resynthesized version of each recorded sentence has been created where the original F_0 contour is approximated by the smallest possible number of straight-line segments, so that only the fluctuations that do essentially contribute to the perception of prosody are to be accounted for. The final step, was to determine all the turning points appearing in the resynthesized F_0 versions, and to describe them in terms of discrete textual events.

ACCENTS, PHRASING, AND PHONOLOGICAL REPRESENTATION

Accentuation and phrasing are two of the weak points in TTS-systems (Traber 1992), partly because syntactic information is not enough to derive a reasonable stress pattern and phrasing, and partly because the accentuation and phrasing rules are not elaborate enough. As far as the accentuation of the Greek language is concerned, there is no problem, because word accents are included in the orthographic representation of the input. However, the phrasing is a more complicated problem for the following reasons:
a. Greek is an inflectionally rich language and the problems start from the correct and unambiguous text labelling achieved in real-time.
b. It is a language where different syntactic structures can be used to express the same meaning. In addition, the position of constituents in a sentence is almost free and the constituents can be moved to reflect pragmatic factors.

As a consequence, it was our deliberate choice to lay greater emphasis on phonotactics and metrical data in the development of the model rather than upon syntactic phenomena. However, a syntactic-prosodic parsing of the input is carried out (Epitropakis, 1993), in order to determine the phrasal units of the sentence that are necessary to determine some of the pitch movements that are not signalled by phonotactics and metrical events.

We finally came out to a set of linguistic information (table 1) according to which each phoneme is labelled with a target pitch level, so that all the pitch configurations presented in the results obtained from the resynthesized F_0 contours data-base analysis to be synthesized. Generally, each phoneme of the input is labelled according to:

- certain distinctive features (vowel/consonant, stressed/unstressed, etc.),
- the lexical position of the syllable that the phoneme belongs to (antepenultima, penultima, ultima), and
- the phoneme's relative position to prosodically significant events, where the boundaries between sentences, sentence and relative phrase, verb phrase (VP) and noun phrase (NP), VP and adverbial phrase (AP), VP and prepositional phrase (PP) are taken into account.

1	First phoneme of the phrase	16	Boundary (VP-AP) vowel
2	Last phoneme before the first accent of the phrase	17	Voiced consonant before the end of the sentence
3	Accented phoneme	18	First vowel after the boundary (VP-AP)
4	Phoneme between two accents	19	Boundary (VP-PP) vowel
5	First pre-boundary (NP-VP) phoneme	20	Pre-boundary (VP-PP) vowel
6	Second pre-boundary (NP-VP) phoneme	21	First vowel after the boundary VP-PP
7	Penultima	22	Voiced consonant before the boundary VP-PP
8	Boundary phoneme (NP-VP)	23	Final phoneme of the phrase
9	Ultima vowel after an accented penultima	24	Pre-boundary (VP-AP) vowel
10	First phoneme after the boundary NP-VP	25	Boundary (,) vowel
11	Voiced consonant before an accent	26	First pre-boundary (,) vowel
12	First phoneme just after the accent	27	Second pre-boundary (,) vowel
13	Pre-final phoneme of the phrase	28	First vowel just after the boundary (,)
14	Final vowel of the phrase	29	Voiced consonant before the boundary (,)
15	Voiced consonant before the NP-VP boundary	30	Voiced consonant before the boundary VP-AP

Table 1. *Linguistic properties that are used for labelling the input phonemes.*

A set of rules that could constitute the "grammar of the intonation" of the Greek language for neutral speaking style has been extracted. These rules are of type:

a, b, c, .. --> F_0 level,

where *a,b,c,..* is the above presented linguistic information labelling the phonemes of the input. *F_0 level,* is a label used to assign at each phoneme one of the following tone levels: *Base, mid and top.* It is not an absolute F_0 value (in Hz), but a dynamic estimation of the corresponding phoneme's pitch value in one of the three declined lines used for the Greek language (baseline, midline and topline). Both the starting point (Hz) and the declination slope in semitones/sec of these lines have been experimentally extracted by statistical analysis of the original pitch contours of the speech data-base. Equations 1 and 2 give the starting point B_{start} and the slope B_{slope} of the baseline respectively according to the sentence's duration *t*. Additional information for the determination and the implementation of the necessary reset points has also been extracted.

$$B_{start} = 120 * e^{(-0.061/t)} \quad (1) \qquad B_{slope} = 3.05 \ / \ (t+0.505) \quad (2)$$

INTONATION ALGORITHM

The intonation algorithm assigns to each phoneme of the input sentence a target pitch *level* to be reached at the middle or the end or the start of the phoneme, according to the rules of the intonational grammar.

For example, given the input *"O aEt'Os tu f'ilu mu tu t'aKi pu 'idamE xTEs, pEt'ai psil'a"* (=the eagle of my friend Takis, which we saw yesterday, flies high) after the linguistic analysis (Epitropakis, 1993), which obtains data such as those shown in table 1, the algorithm generates the overall pitch contour in the following steps:

1. Firstly the declined lines and the appropriate reset points are determined.
2. Input in the algorithm is the phonetic representation of the input sentence plus the corresponding labels describing the phonemes properties. The appropriate F_0-levels are assigned to the corresponding phonemes according to the rules of the grammar (table 2 gives the rules used in this example). A total of 67 such rules constitute the grammar for Greek intonation. The phonological representation for the example given, is as follows: *O[1] aE[2,7]t'O[3]s tu[4,12] f'i[3,7]lu[9,12] mu tu[26] t'a[3,7,27]Ki[8,9,12,25] pu[28] 'ida[5,7,12,26]mE[6,27] xTE[8,25]s, pE[10,28]t'a[3,7]i[9,12] psil[11,17]'a[3,14,23,30]*

Linguistic attributes	F_0-level and timing
1	Base at the start of the phoneme
2,7	Base at the end of the phoneme
3	Top at the end of the phoneme
3,7	Mid at the start of the phoneme
3,7,27	Base at the end of the phoneme
3,14,23,30	Top at the middle and base at the end of the phoneme
4,12	Base at the end of the phoneme
5,7,12,26	Top at the start of the phoneme
6,27	Base at the end of the phoneme
8,25	Top at the end of the phoneme
8,9,12,25	Top at the end of the phoneme
9,12	Top at the end of the phoneme
10,28	Base at the end of the phoneme
11,17	Mid at the middle of the phoneme
26	Base at the end of the phoneme
28	Base at the end of the phoneme

Table 2. *The rules of the intonational grammar used for the given example.*

3. When all the phonemes have been assigned a F_0-level, the overall pitch contour is constructed by linear interpolation between the successive levels. Figure 1 gives the overall pitch contour (solid line) constructed for the example sentence according to the F_0-levels (solid dots) that have been determined. The original pitch contour is also given (dotted line).

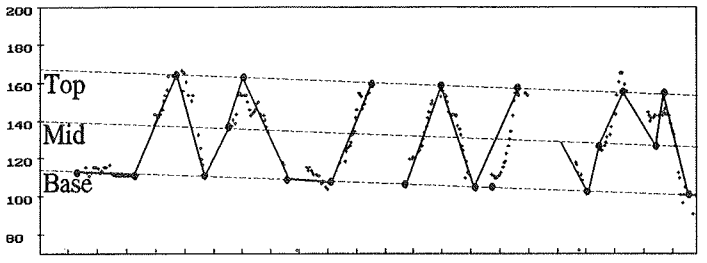

Figure 1. *Synthesized pitch contour for the example mentioned above.*

EVALUATION RESULTS - CONCLUSIONS

A complete methodology for constructing intonation contours in the context of TTS-systems has been presented. It has been shown that a successful intonation model can be created for a language such as Greek, which has peculiar difficulties, by combining syntactic, phonotactics and metrical data. The particular combination of data exploited in this model permits to achieve great accuracy and to be computationally tractable in real-time. To evaluate this methodology, two different tests have been carried out:

The first test estimates the performance of the algorithm in correctly determining the prosodically important turning points in the pitch contours. For this purpose, a set of sentences have been recorded and the appropriate turning points have been manually determined. Then the proposed algorithm has been used for the same set. Analysis showed that 71% of these points were correctly determined. A percentage of 18% were not been determined, but the resulting overall pitch contour was almost equivalent to the expected contour. Finally, 11% of the turning points have been determined incorrectly or have been totally missed. This omission led to unacceptable pitch contours. Generally, the method proved to be robust, but for the extraction of the grammar rules only the rules with high statistical correlation (above 80%) in the speech data-base, are included. The inclusion of heuristic rules with lower occurrence must be taken into account in the future.

The second test concerns the evaluation of the speech output quality. For this reason, preliminary tests with 4 listeners were carried out. Two different text subjects were synthesized by the Greek TTS-system consisting of 39 sentences (650 words). The results show that the naturalness of the resulting speech is very high. Refinement of the rules and Spline function interpolation are in progress in order to further improve the speech quality. In addition, a novel phrasing model (Michos et al.) is under consideration.

REFERENCES

R. Carlson and B. Granstrom (1973), "Word accent, emphatic stress and syntax in a synthesis by rule scheme for Swedish", *OSPR-STL*, Vol. 2-3, pp.31-36, KTH.

G. Epitropakis, N. Yiourgalis, and G. Kokkinakis, "Prosody control of TTS-Systems based on linguistic analysis", *Eurospeech '93*, (to be presented).

D. Hirst (1992), "Prediction of prosody: An overview", *Talking Machines: Theories, Models, and Designs*, Baily, Benoit, and Sawallis (eds), pp.199-204, Elsevier Science Publishers B.V.

J. t'Hart, R. Collier, and A. Cohen, "A perceptual study of intonation", Cambridge University Press.

S. Michos, G. Epitropakis, N. Fakotakis, and G. Kokkinakis, "A novel phrasing method for high quality prosody in TTS-systems", to be submitted to ICASSP '94.

J. Pierrehumbert (1980), "The Phonology and Phonetics of English Intonation", *MIT Ph.D. dissertation.*

C. Traber (1992), "F_0 generation with a database of natural F_0 patterns and with a neural network", *Talking Machines: Theories, Models, and Designs*, Baily, Benoit, and Sawallis (eds), pp.199-204.