PROSODY ASSIGNMENT TO TTS-SYSTEMS BASED ON LINGUISTIC ANALYSIS

G. Epitropakis, N. Yiourgalis, G. Kokkinakis

Wire Communications Laboratory
University of Patras

ABSTRACT - In this paper we present a)a complete method for formulating the rules needed for assigning prosody to a Text-To-Speech system on the basis of linguistic knowledge extracted from text and b)the implementation of the method in the Greek TTS-system developed at our laboratory (Yiourgalis & Kokkinakis, 1991).

INTRODUCTION

Quality problems are focused in modern TTS systems on the lack of naturalness. High quality synthetic speech can only be obtained by the correct assignment of intonational contours and the proper location of appropriate in duration pauses. Thus, knowledge of how the system could meaningfully manipulate these prosodic features (pitch and pauses) for any input text, must be supplemented to the system. To this end, we retrieved, from the abundance of possible physiological measurements, the discrete events that do contribute essentially to the perception of the speech prosody. Rules between these discrete events, and the F_0 movements and the speech pauses appearance were finally established.

Our approach, whose main characteristic is the continuous confrontation of the experimental measurements of speech with the way in which the produced prosody is perceived, is based on the following basic assumption (Cohen and 't Hart, 1967):

> "The pitch movements that are interpreted as relevant by the listener are related to corresponding activities on the part of the speaker. These are assumed to be characterized by discrete commands to the vocal cords and should be recoverable in the resulting pitch contour, which may present themselves at first sight as continuous variations in time."

The method is tedious and time-consuming but it greatly improves the quality of the synthetic speech. Furthermore it is general and can be applied to other languages. Nevertheless, language specific resources and researcher skills are necessary in each case.

THE METHOD

Data Collection

Given the purpose of our study, viz. to extract and formulate the rules that establish the correspondence between the syntactic constraints and prosodic properties of any input sentence, we first drew up a list of 200 test sentences that cover the greatest possible syntactic structures of the language. These possibly appearing structures have been extracted from an exhaustive analysis of a 120,000 words text corpus, previously labelled by analysts. The material to be recorded has been achieved by primarily combining simple types of Noun Phrases (NP) and Verb Phrases (VP), and by increasing successively their complexity, always producing meaningful sentences. In addition, sentences with identical syntactic structures, but implemented with different in size (number of syllables) words have been constructed in order to achieve the necessary variety in the number of syllables that precede the prosodically important point (syntactic boundary or stressed syllable).

These sentences were recorded by 3 male trained speakers, sampled at 10KHz and stored on disk. It is well known that the same text can be read in different ways that reflect, in some cases, to different meaning. Thus, it has been asked from the speakers to read the sentences in such a way that the same meaning is perceived since we are aiming at rules that provide a neutral reading style.

Pitch Extraction

Acoustic analysis with pitch extraction and pause duration measurement included, were performed directly on the digitized speech waveform, for every sentence in the speech database.

The pitch extraction method is based on the SIFT algorithm (Markel & Gray, 1976). An end-point detection algorithm based on energy calculation is previously performed in order to avoid the influence of the recording noise. Finally, a differentiation module is added in order to reject the wrong values extracted from the SIFT algorithm. A special graphical environment describes the F_0 curve of the utterance time-aligned with the graphical representation of the speech waveform.

Manual Speech Synthesis

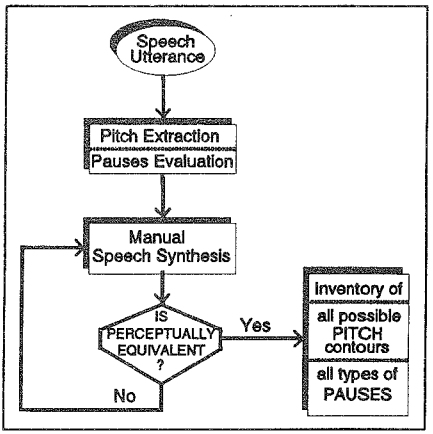


Figure 1. The main steps of the method.

Using a specially designed tool developed in the context of our TTS system each test sentence can be re-synthesized by manually assigning any desired F_0 contour. The purpose of this step, is to obtain a new F_0 curve for each test utterance that causes the resynthesized version to be acoustically indistinguishable from the natural one. The ultimate aim of this procedure is to gather a manageable set (Fig. 2: inventory) of naturally sounded styles of F_0 contours, called "stylized contours", that cover all the possible syntactic structures of the language. These contours will be further analyzed in order to form the set of precepts needed for constructing intonation during speech synthesis.

The stylised contours were decided upon two additional criteria besides the perceptual equivalence:

1) the original F_0 contour of the sentence is approximated by the smallest possible number

of straight-line segments and,

2) the changes in the resynthesized version must be marked by corresponding linguistic properties of the sentence (syntactic boundary, word accent, etc.).

Thus, through trial-and-error manipulation, an output is achieved which is acoustically as close as possible to the natural speech. The basic idea here, is to distinct the changes in the F_0 contour that do not contribute to the perception of the sentence prosody. These changes are due, among others, to micro-intonation phenomena, such as the vowel intrinsic pitch and the influence of consonants to the adjacent vowels (Reinholt, 1986).

In Fig. 3 we show an example of the stylised F_0 contour assignment. All the pitch movements are restricted between three declined lines (Pierrehumbert, 1976) that have experimentally been evaluated. Generally, it can be reported that differences of about 15Hz are not acoustically perceived.

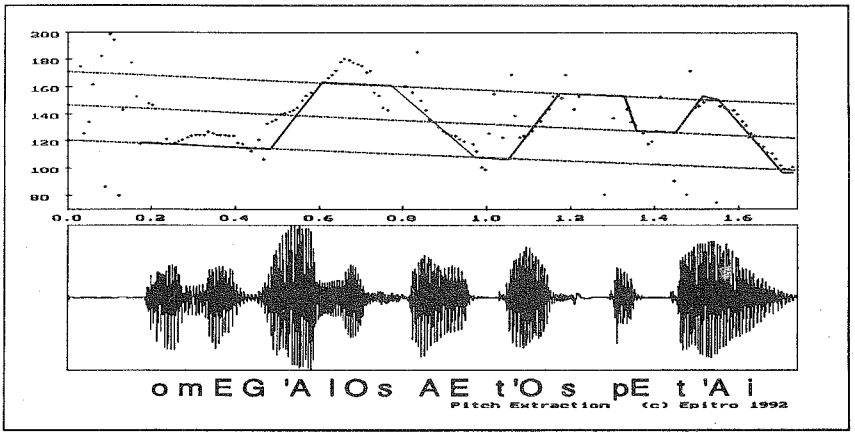


Figure 2. An example of the pitch contour stylization procedure.

Pause Measurement

The pauses are measured by cutting and pasting the interesting interval via another special graphical environment . The results are stored in a database file, whose fields reflect the way the whole speech material is organized, i.e. there is information about the possible syntactic structures of the sentences, the number of syllables preceding the boundary and the location of the word accent. After the analysis has been performed, one can retrieve the appropriate results by simply prompting the appropriate questions to this data base.

Intonational Rule Formulation

A set of rules describing the pitch contours is determined by examining the stylized pitch contours in connection to the following linguistic information of a sentence:

i) The word stress points.

ii) The syntactic/intonational boundaries of the sentence.

iii) The length (in syllables) of the pre-boundary phrase.
iv) The punctuation marks.

Each rule describes:

i. The direction of the F_0 change (Rise / Fall).
ii. The on-set time of the F_0 change (in the start of the vowel / after 5ms).
iii. The slope of the F_0 change (when the maximum/minimum value must be achieved: start / end of the vowel).
iv. The size of the change (1: when there is a movement between the upper or lower declined line and the Midline / 2: when there is a movement between Topline and Bottom).

A complete rule describing how an accented syllable to be synthesized is treated for the Greek language, is given below:

Rule 1.

All the stressed syllables in the word are marked by a pitch rise exercised at the beginning of the vowel. It starts at the F_0 value of the preceding syllable and at the end of the vowel, it reaches the Topline. The unaccented syllables located between two accents are marked by a fall movement down to the Bottom, which is reached at the end of the last unstressed syllable.

SPECIAL CASES:

i. If the accent is located on the penultimate syllable of the word, and the ultimate syllable is at least two syllables apart from the following intonational boundary, then the ultimate syllable keeps on.

EXCEPTIONS:

ii. If the preceding syllable has reached the Topline, the new (stressed) syllable keeps on.
iii. The one-syllable words are not stressed.
iv. If the stress is located two or less syllables before a syntactic boundary, then is a pitch fall instead. The end of the fall coincides with the start of the pre-boundary syllable, which is marked by a pitch rise.
v. The first stressed syllable of the second NP after the conjunction "και"(=and) is marked by a pitch rise up to the Midline.

1. The rule describing how the intonation contour for an accented syllable to be synthesized, is constructed.

Additional rules have been extracted in order to evaluate the pauses of a sentence to be synthesized. These rules determine the location of the silence interval plus its duration. The later depends on the type of the existing syntactic boundary and the number of preceding syllables as well (see Results)..

THE SYSTEM

Implementation

The intonation contour fed to the synthesizer, is constructed by an algorithm whose resources are the linguistic information of the sentence and the intonational rules extracted and formulated with the method presented above. The linguistic information is retrieved during TTS-synthesis by special modules (Syntactic Parser and Text-preprocessor). Additional information needed for the location of the accented syllable (antepenultimate, penultimate or ultimate syllable) is extracted by the Segmentation module of the system.

All pitch movements (F_0 changes) are arranged on a basic contour including three declined lines (Bottom, Midline and Topline). The declination rate of these lines depends on the total duration of the synthesized sentence.

As an example, consider the development of a pitch contour through the application of the appropriate rules (Fig. 3):

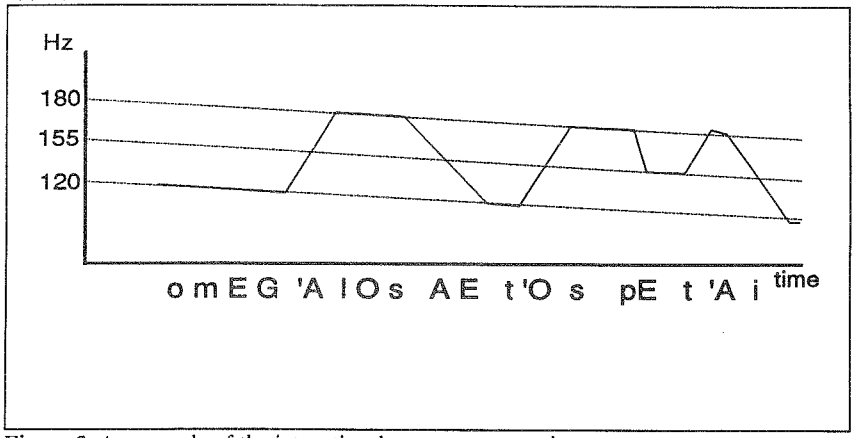


Figure 3. An example of the intonational contour construction.

The pitch remains constant to the Bottom line until it reaches the first accented syllable (G'A), where there is a rise movement up to the Topline (rule 1.). It remains there for one additional syllable (the ultima of the word "mEG'AlOs"), since the accented syllable is the penultima of the word and there isn't a syntactic boundary near by (rule 1.i). After the ultima (lOs), there is a fall movement of size 2. It reaches the Bottom just before the next accented syllable (t'Os), where it begins raising again up to the Topline (rule 1.). The fall starts at the end of the stressed syllable and it reaches its minimum value (Midline) at the end of the last unstressed syllable (pE) before the next accent (rule 1.i). The last rise coincides with the last accented syllable (t'A) and it immediately falls up to the Bottom line because that is the end of a declarative sentence.

RESULTS

Analysis of the stylised contours has lead to the following decisions:

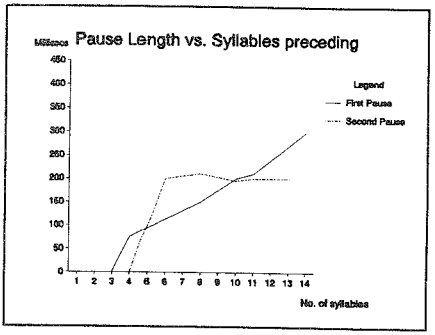
- There are 8 types of syntactic boundaries that completely characterize the intonational

cues of the Greek language. Namely:

1. Sentence - Sentence.
2. Sentence - Subordinate clause.
3. NP - VP.
4. NP - NP or VP - VP that are connected by a conjunction.
5. VP - Prepositional Phrase (PrepP), when the latter contains more than two words.
6. In commas, parenthesis.
7. NP - Subordinate clause, when they are separated by a comma. The difference of this type from type 2 above, is that here there is a reset in the declined lines.
8. NP - VP, when both are larger than 9 syllables. (There is also a declination reset)

- Up to 35 (the exceptions included) rules were enough for the system to define the appropriate synthetic pitch contour, for any input sentence.

- Two major categories of pauses are present in spoken Greek. The first is almost independent from the number of syllables preceding the boundary while the second depends highly. The first one appears at major syntactic boundaries (types 1,2,6,7) and the second at the other types.



CONCLUSIONS

A syntax-to-prosody method has been implemented into the Greek parametric formant speech synthesizer, that has greatly influenced its quality. Improvements were justified in both the rhythm of the synthetic speech, by insertion of pauses in the proper locations in the sentence to be spoken, and its naturalness, by the application of a more "natural" resembling and well documented intonational contour.

REFERENCES

Markel J.D. and Gray A.H. (1976) *Linear Prediction of Speech*, New York:Springer Verlag, 1976.

Pierrehumbert J. (1979) *The perception of fundamental frequency declination*, Journal of the Acoustical Society of America 66: 363-369.

Reinholt Petersen N. (1986) *Perceptual compensation for segmentally conditioned fundamental frequency perturbation*, Phonetica 43: 31-42.

t'Hart J., Collier R. and Cohen A. (1990) *A perceptual study of intonation*, Cambridge University Press, New York, 1990.

Yiourgalis N., Kokkinakis G. (1991) *A TTS system for Greek*, ICCASP '91, pp.525-529, Toronto, Canada.