ANALYSIS AND MODELING OF FUNDAMENTAL FREQUENCY CONTOURS OF GREEK UTTERANCES

Hiroya Fujisaki, Sumio Ohno and Takashi Yagi

Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

ABSTRACT

A quantitative model for the process of F_0 contour generation, originally developed for Japanese by Fujisaki and his co-workers, has already been shown to be valid for several other languages. The present study aims at testing its applicability to F_0 contours of Greek utterances. Analysis of F_0 contours of 200 utterances by two native speakers of Greek, produced by reading texts of narrations and conversations, has shown that the model is essentially valid, and suggests the model's usefulness for Text-to-Speech synthesis of Greek.

1. INTRODUCTION

In many languages of the world, the contour of the fundamental frequency of voice (henceforth the F_0 contour), along with the intensity and the duration of various speech units, plays an important role in conveying linguistic information concerning the lexical tone/accent, syntax, and focus of the message, as well as para- and non-linguistic information concerning the intention, attitude, speaking style, gender, physical and emotional states of the speaker [1]. While linguistic information is primarily symbolic and is thus discrete in nature, para- and non-linguistic information can be both discrete and continuous. For example, the distinction between statement and question is discrete, but the degree of doubt, generally expressed by an interrogative intonation, is continuous. On the other hand, the F_0 contour is essentially a continuous function both in amplitude and in time.

Because of this intrinsic difference in the nature of prosodic information and the corresponding F_0 contour, it has always been a difficult and challenging problem to find the exact relationship between the two. Prosodic labeling methods use finite sets of symbols, and thus can describe only the discrete aspects of prosodic information. On the other hand, methods based on stylization or modeling generally attempt to characterize an F_0 contour by a finite number of points associated with continuous parameter values, and thus have the possibly of representing both discrete and continuous aspects of prosody. If, however, the parameter values are limited to a finite set, these methods also fail to capture the continuous aspects of prosodic information.

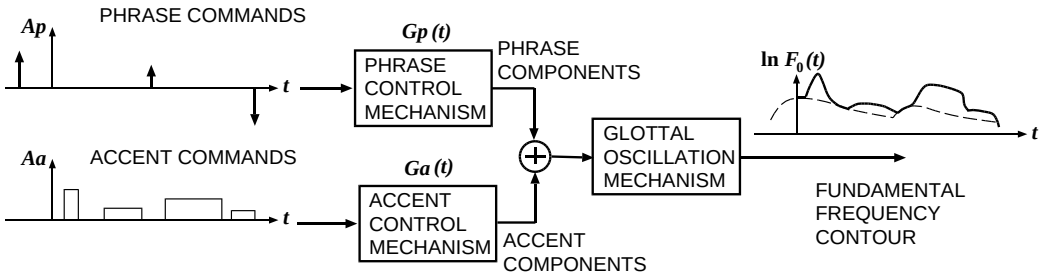
If we have sufficient knowledge on the mechanisms and processes responsible for generating the F_0 contour from the underlying information, we can construct a quantitative model, which will allow us to synthesize an F_0 contour from the underlying information. Once such a model is constructed it can also be used to solve the inverse problem, i.e., to find the underlying information from an observed F_0 contour.

On the basis of these considerations, Fujisaki and his coworkers have constructed a quantitative model for the mechanisms and processes involved in the generation of the F_0 contour [2,3]. The validity and utility of the model have been demonstrated at first for the analysis and synthesis of F_0 contours of words and sentences of the common Japanese, and have since been extended to the analysis and synthesis of F_0 contours of other dialects of Japanese as well as of several other languages including Chinese, English, German, Korean, Spanish, and Swedish [4-9]. These studies have revealed that the major features of the original model for the F_0 contours of the common Japanese are valid for those of other dialects and languages, but certain modifications and elaborations need to be introduced to represent those features that are specific to certain dialects and languages.

The present paper describes the method and results of a preliminary study on the analysis and modeling of the F_0 contours of Greek utterances (text reading), based on the approach that has been proved to be successful for the F_0 contours of Japanese and other languages.

2. MODELING OF F_0 CONTOURS OF GREEK UTTERANCES

It is widely accepted that F_0 contours of many languages are characterized by relatively slow undulations (henceforth phrase components) which roughly correspond to larger phrases, clauses, and sentences, and by relatively fast rise/fall patterns (henceforth accent components) which correspond to either lexical tones of syllables or lexical accent of words. It is also accepted that these two kinds of components can be considered to be additive if one adopts the logarithmic scale (or equivalently, the semitone scale) for F_0 as a function of time. For example, in the stylized model for F_0 contours of Greek utterances used for Text-to-Speech Synthesis [10,11], the former



F_0 contour generation.

is represented by a declining straight baseline in the semitone scale of fundamental frequency, while the latter is represented by positive deviations from this baseline. Our previous studies on F_0 contours of various languages indicate, however, that the F_0 contour, when plotted on the logarithmic frequency scale as a function of time, can be very closely approximated as the sum of phrase components which initially rise and then gradually decay to approach zero, and positive accent components which occur at accented syllables or morae, both added to a constant baseline.

Figure 1 shows the configuration of the model. The phrase commands are assumed to be impulses applied to the phrase control mechanism to generate the phrase components, while the accent commands are assumed to be positive stepwise functions applied to the accent control mechanism to generate the accent components. Both mechanisms are assumed to be critically damped second-order linear systems, and the sum of their outputs, i.e., the phrase components and accent components, is superposed on a baseline value (ln Fb) to form an F_0 contour, as given by the following equation:

$$\begin{aligned}\ln F_0(t) = \ln Fb &+ \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) \\ &+ \sum_{j=1}^{J} Aa_j [Ga(t - T_{1j}) - Ga(t - T_{2j})],\end{aligned} \tag{1}$$

$$Gp(t) \begin{cases} = \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \tag{2}$$

$$Ga(t) \begin{cases} = \min[1 - (1 + \beta t)\exp(-\beta t),\ \gamma], & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \tag{3}$$

where $Gp(t)$ represents the impulse response function of the phrase control mechanism and $Ga(t)$ represents the step response function of the accent control mechanism.

The symbols in these equations indicate

Fb : asymptotic value of fundamental frequency,
I : number of phrase commands,
J : number of accent commands,
Ap_i : magnitude of the ith phrase command,
Aa_j : amplitude of the jth accent command,
T_{0i} : timing of the ith phrase command,
T_{1j} : onset of the jth accent command,
T_{2j} : end of the jth accent command,
α : natural frequency of the phrase control mechanism,
β : natural frequency of the accent control mechanism,
γ : relative ceiling level of accent components.

The parameters α and β are assumed to be constant at least within an utterance, while the parameter γ is set equal to 0.9. It has been shown that the ability of the model to produce vary accurate approximations to observed F_0 contours has its basis in the physiological and physical mechanisms of the larynx [12].

3. ANALYSIS-BY-SYNTHESIS OF F_0 CONTOURS OF GREEK UTTERANCES

3.1. Speech Material

The speech material consists of recordings of narrations and conversations from a textbook of contemporary Greek, read or acted by two native speakers (a male and a female), each reading about half of the total material. Each sentence is read at two speech rates : slow and normal. The recordings, in two cassette tapes, were supplied by the publisher. The speech signal was digitized at 10 kHz with 16 bits. Fundamental frequencies were extracted at every 10 ms by a modified autocorrelation analysis of the LPC residual.

3.2. Analysis Procedure

The validity of the proposed model can be tested by Analysis-by-Synthesis, i.e., by constructing the best approximation to an observed F_0 contour, and by examining the closeness of the approximation. The optimization is carried out by minimizing the mean squared error in the ln $F_0(t)$ domain through a hill-climbing search in the space of model parameters. This allows one to decompose a given F_0 contour into its constituent components, and to estimate their underlying commands by deconvolution.

3.3. Experimental Results and Discussion

Figure 2 shows the results of F_0 contour analysis of one utterance each of the following four sentences:

(a) “Σήμερα πολλοί άνθρωποι απ’όλον τον κόσμο θέλουν να πάνε στην Ακρόπολη.” (Nowadays many people from all over the world want to go the Acropolis.),

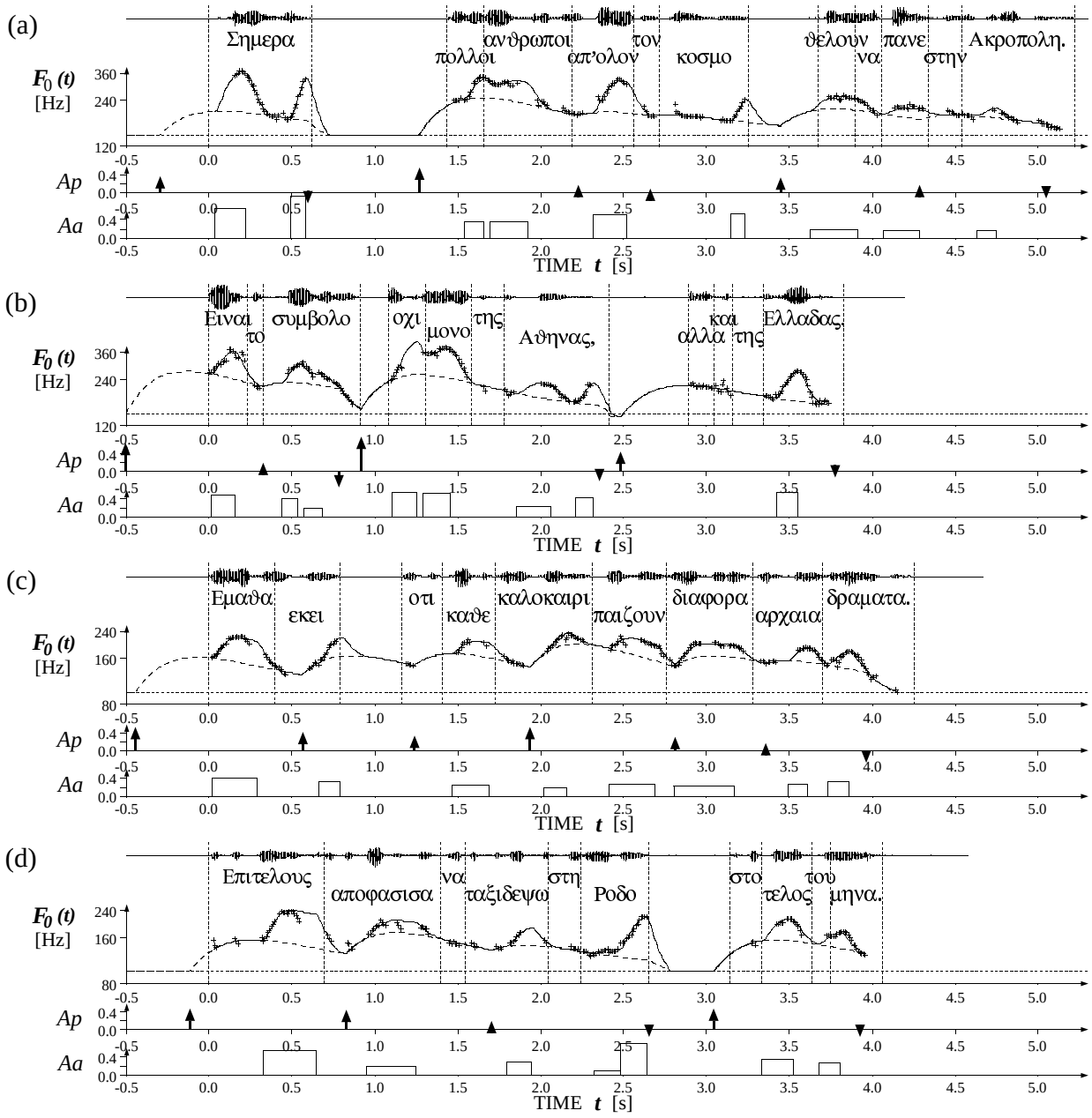
(b) “Είναι το σύμβολο όχι μόνο της Αθήνας, αλλά και της Ελλάδας.” (It is the symbol not only of Athens, but also of Greece.)

(c) “Έμαθα εκεί ότι κάθε καλοκαίρι παίζουν διάφορα αρχαία δράματα.” (I learned there that each summer various ancient dramas are played.)

(d) “Επιτέλους αποφάσισα να ταξιδέψω στη Ρόδο στο τέλος του μήνα.” (Finally I decided to travel to Rhodes at the end of the month.)

Utterances (a) and (b) were produced by the female speaker while (c) and (d) were produced by the male speaker, all at the normal speech rate. Each panel displays, from top to bottom, the speech waveform, approximate positions of word boundaries (vertical dotted lines), measured F_0 values (+ symbols), the best approximation by the model (solid line), the baseline frequency (horizontal dotted line), the phrase commands (impulses), and the accent commands (stepwise functions). The broken lines indicate the phrase components wherever they differ from the model’s approximation, and the differences between the solid line and the broken lines indicate the accent components.

As shown by all the four panels, the model is capable of producing very close approximations to the measured F_0 contours, and the commands and resulting components clearly indicate the prosodic structure of the utterances,



F_0 contours of four Greek utterances by two native speakers. Fundamental frequencies are plotted on a logarithmic scale.

which reflect the syntactic and lexical information of the message in most cases.

Namely, prosodic phrasing is in broad agreement with the syntactic structure. The only discrepancy is found in the utterance (c), where a phrase command occurs between the word 'κάθε (each)' and 'καλοκαίρι (summer).' The discrepancy may well have been caused by a careless mistake on the part of the speaker, since it is not found in the utterance of the same sentence produced at a slow speech rate. Likewise, the occurrence of the accent command agrees almost perfectly with the position of the lexical accent, though the amplitude of the accent command varies widely depending on the degree of prominence that is given to a particular word. The only exception is the word 'κόσμο (world)' in the utterance (a), where the accent command does not occur at the syllable of the lexical accent, but is shifted to the final syllable, expressing a non-terminal intonation. The expression of non-finality (non-terminal intonation) in statements is seen to depend on whether the final syllable of the phrase in question has an intrinsic lexical accent or not.

The current method of F_0 contour analysis is capable of extracting both the discrete and the continuous aspects of information contained in the F_0 contour. It is needless to say that symbolic annotations are not sufficient to reproduce the entire prosodic features of an utterance. Quantitative features, represented by the actual timing and the magnitude of the phrase and accent commands, are indispensable for reproducing the original F_0 contour, or for generating an F_0 contour which is capable of giving a natural intonation.

Although the above discussion referred only to the results shown in Fig. 2, analysis has already been made on all the 200 utterances, and the results indicate that the model can always generate very close approximations to all the observed F_0 contours from a small number of linguistically meaningful commands. These results suggest that the current model is useful for the analysis, modeling, annotation, and synthesis of F_0 contours of Greek utterances. Further work is under way to investigate the language-specific features of Greek intonation as compared with features of other languages thus far analyzed by the current method, as well as the variations of F_0 contour parameters due to differences in speech rate, speaking style (narration vs. conversation) and speakers in quantitative terms.

4. SUMMARY

The applicability of a model-based analysis method, initially developed for F_0 contours of Japanese and later extended to a number of other languages, has been tested for F_0 contours of Greek utterances. Although the speech material for the current study is limited both in the number of speakers and in the number of utterances, the model was shown to be capable of generating very close approximations to all the observed F_0 contours from a limited number of linguistically meaningful commands. Thus it opens a way to a highly accurate quantification of F_0 contours of Greek utterances which is useful for high-quality speech synthesis in Text-to-Speech systems, as well as for prosodic annotation/labeling. This work is supported by a Grant-in-Aid for Scientific Research (No. 08458090) from the Ministry of Education, Science and Culture of Japan. The authors also would like to thank Prof. George Kokkinakis for kindly supplying some of the speech material for the current study.

REFERENCES

[1] H. Fujisaki, "From information to intonation," *Proceedings of the 1993 International Symposium on Spoken Dialogue*, pp. 7–18, 1993.

[2] H. Fujisaki and S. Nagashima, "A model for the synthesis of pitch contours of connected speech," *Annual Report of the Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.

[3] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn. (E)*, vol. 5, pp. 233–242, 1984.

[4] H. Fujisaki, K. Hirose, P. Hallé and H. Lei, "Analysis and modeling of tonal features in polysyllable words and sentences of the Standard Chinese," *Proceedings of the 1990 International Conference on Spoken Language Processing*, vol. 2, pp. 841–844, 1990.

[5] H. Fujisaki and S. Ohno, "Analysis and modeling of fundamental frequency contours of English utterances," *Proceedings of the 4th European Conference in Speech Communication and Technology*, vol. 2, pp. 985–988, 1995.

[6] H. Fujisaki and H. Mixdorff, "Analysis of voice fundamental frequency contours of German utterances using a quantitative model," *Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 4, pp. 2231–2234, 1994.

[7] H. Fujisaki, "Analysis and modeling of fundamental frequency contours of Korean utterances," In *Phonetics and linguistics* – In Honour of Prof, H. B. Lee, (M. Yn, ed.) PP. 640–657, 1996.

[8] H. Fujisaki, S. Ohno, K. Nakamura, M. Guirao and J. Gurlekian, "Analysis of accent and intonation in Spanish based on a quantitative model." *Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 1, pp. 355–358, 1994.

[9] H. Fujisaki, M. Ljungqvist and H. Murata, "Analysis and modeling of word accent and sentence intonation in Swedish." *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 211–214, 1993.

[10] G. Epitcopakis, N. Yiourgalis, and G. Kokkinakis, "High quality intonation algorithm for the Greek TTS-system," *Working Papers 41, Dept. of Linguistics and Phonetics, Lund University*, pp. 70–73, 1993.

[11] D. Galanis, V. Darsinos and G. Kokkinakis, "Modeling of intonation bearing emphasis for TTS-synthesis of Greek dialogues," *Proceedings of 1996 International Conference on Spoken Language Processing*, vol. 3, pp. 1357–1360, 1996.

[12] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," In *Vocal Physiology: Voice Production, Mechanisms and Functions* (O. Fujimura, Ed.), pp. 347–355, Raven Press, New York, 1988.