

Analysis and Modeling of F_0 Contours of Portuguese Utterances Based on the Command-Response Model

Hiroya Fujisaki¹, Shuichi Narusawa², Sumio Ohno³ and Diamantino Freitas⁴

¹Professor Emeritus, University of Tokyo, Japan

²Graduate School of Information Science and Technology, University of Tokyo, Japan

³Faculty of Engineering, Tokyo University of Technology, Japan

⁴Faculty of Engineering, University of Porto, Portugal

fujisaki@alum.mit.edu narusawa@gavo.t.u-tokyo.ac.jp

ohno@cc.teu.ac.jp dfreitas@fe.up.pt

Abstract

This paper describes the results of a joint study on the applicability of the command-response model to F_0 contours of European Portuguese, with an aim to use it in a TTS system. Analysis-by-Synthesis of observed F_0 contours of a number of utterances by five native speakers indicated that the model with provisions for both positive and negative accent commands applies quite well to all the utterances tested. The estimated commands are found to be closely related to the linguistic contents of the utterances. One of the features of European Portuguese found in utterances by the majority of speakers is the occurrence of a negative accent command at certain phrase-initial positions, and its perceptual significance is examined by an informal listening test, using stimuli synthesized both with and without negative accent commands.

1. Introduction

Portuguese belongs to the family of Romance languages and is among the top ten languages most widely spoken in the world, with nearly 200 million speakers living in Portugal, Brazil and other countries. It is thus quite important from the viewpoint not only of basic speech science but also of information technology to elucidate the acoustic-phonetic characteristics of Portuguese, and to utilize the findings to establish effective means for synthesis and recognition of speech of Portuguese. Relatively little has been published, however, on the quantitative characteristics of its prosody, especially of its intonation, of which the primary acoustic correlate is the contour of the fundamental frequency of voice (henceforth the F_0 contour).

It has been shown by Fujisaki and his coworkers that the command-response model, originally developed for the process of generation of F_0 contours of Common Japanese, applies quite well, after certain language-specific modifications, to F_0 contours of many other languages including Chinese, English, German, Greek, Korean, Spanish, Swedish[1] and Thai[2]. The model has also been shown to apply to F_0 contours of Basque[3], French[4], and Italian[5].

The present paper describes the results of a joint study conducted by the authors to test the applicability of the command-response model to F_0 contours of utterances of European Portuguese, as well as to find specific features that will have to be added to the original model, with an aim to utilize it in a TTS system of European Portuguese (henceforth 'Portuguese' for short)[6, 7].

2. A model for the generation process of F_0 contours of Portuguese utterances

Careful observation of F_0 contours of Portuguese utterances suggests that the mechanism of laryngeal control for Portuguese intonation is essentially the same, at least qualitatively, as that for other languages investigated so far, but requires accent commands of both positive and negative polarities in order to be able to generate the fall-rise patterns found in F_0 contours of many utterance samples. Figure 1 shows the model for the process of generation of F_0 contours of Portuguese utterances we propose on the basis of these considerations. The phrase commands are impulses while the accent commands are pedestal functions. These commands are applied to the respective control mechanisms which are assumed to be critically-damped and produce phrase and accent components. These components are then added to a constant component $\ln F_b$ to produce the final $\ln F_0(t)$. For the rest of the paper, we shall use the word ' F_0 -contour' to indicate $\ln F_0(t)$. Physiological and physical evidences supporting the model were presented elsewhere[1].

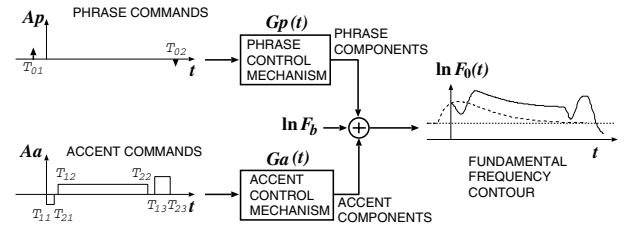


Figure 1: The command-response model for the process of generating F_0 contours of Portuguese utterances.

Thus the F_0 contour as a function of time can be expressed by the following equations:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^I A p_i G p(t - T_{0i}) + \sum_{j=1}^J A a_j \{G a(t - T_{1j}) - G a(t - T_{2j})\}, \quad (1)$$

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

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

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

- F_b : baseline value of fundamental frequency,
- I : number of phrase commands,
- J : number of accent commands,
- Ap_i : magnitude of the i th phrase command,
- Aa_j : amplitude of the j th accent command,
- T_{0i} : timing of the i th phrase command,
- T_{1j} : onset of the j th accent command,
- T_{2j} : offset of the j th accent command,
- α : natural angular frequency of the phrase control mechanism,
- β : natural angular frequency of the accent control mechanism,
- γ : relative ceiling level of accent components.

Parameters α and β are known to be almost constant within an utterance as well as across utterances of a particular speaker. Although certain individual differences exist across speakers, it has been shown that $\alpha = 3.0$ and $\beta = 20.0$ can be used as default values. Parameter γ may be variable across utterances and speakers, but it has also been shown that $\gamma = 0.9$ can be used as a default value.

3. Speech material and method of analysis

3.1. Speech material

The speech material for the present study was recorded at the Faculty of Engineering of the University of Porto. It consists of recordings of three texts read by five native speakers of European Portuguese. Table 1 lists the texts and their English translations. Text A consists of two declarative sentences. Text B is a long sentence consisting of three imperative clauses followed by three wh-question clauses. Text C is a yes/no question sentence.

Each text was uttered five times at a normal speech rate for each individual speaker. The speakers are three male (DF, JT, and FM) and two female (mb and db) native speakers of European Portuguese. The average speech rate for the five speakers are 6.4, 7.3, 7.2, 7.3, and 7.1 syllables per second, respectively.

3.2. Analysis procedure

The speech signal was digitized at 10kHz with 16bit precision. The fundamental frequency was extracted at 10ms intervals by the modified autocorrelation analysis of the LPC residual. The measured F_0 contour was aligned with the speech waveform by visual inspection of the waveform whenever possible.

Using the model’s formulation described in Section 2, it is possible to extract the model parameters from an observed F_0 contour by Analysis-by-Synthesis. This is performed in the following two steps:

1. Obtain a fairly good approximation by using an algorithm for automatic extraction of model parameters[8].
2. Correct, if necessary, any errors in the results of automatic extraction, and refine the results by automatic Analysis-by-Synthesis, viz., automatic optimization of the modified set of parameters by successive approximation.

The second step is introduced since the purpose of the current study is not to assess the performance of the algorithm itself, but to obtain most accurate results that are compatible with the linguistic units and structures of the text of the utterances.

4. Experimental results

The panels (a) and (b) in Figure 2 show the results of analysis of one sample each of the five utterances of Text A by one male (DF) and one female (mb) speaker. Each panel shows, from top to bottom, the speech waveform, measured F_0 values (+ symbols), the model-generated best approximation (solid line), the baseline frequency (dotted line), the phrase commands (impulses), and the accent commands (square pulses). The dashed lines indicate the contributions of phrase components, and the differences between the F_0 contour and the phrase components correspond to the accent components. The results in these panels indicate that negative accent commands are necessary to account for the fall-rise patterns near the onset of some but not all of the phrases, and the use of this negative command is dependent on the speaker as well as on the way the speaker produces prosodic (not syntactic) phrases.

For instance, there is essentially no fall-rise at the onset of the first sentence in both speakers. On the other hand, a fall-rise of appreciable magnitude occurs near the onset of the second sentence in both speakers, while another fall-rise occurs only in the utterance of (a) near the onset of the next prosodic phrase starting at “se” but not in the utterance of (b) in which the next prosodic phrase starts at “na”.

Figure 3 shows the result of analysis of one of the five utterances of Text B by another male speaker (JT). It shows that there is no phrase-initial fall-rise in the F_0 contour, so that it can be approximated and hence can be reproduced quite accurately without using negative accent commands. The result also shows clearly the use of positive accent commands for the continuations rises (indicating non-finality) toward the end of clauses except for the end of the third and the sixth clauses, the

Table 1: Three Portuguese texts and their English translations for the speech material.

Text	Portuguese	English
A	Conhece a situação na pele.	He knows the situation in the skin (i.e., very well).
	Aprendeu-a na idade em que se aprende e se não esquece.	He learned it at an age when one learns and does not forget.
B	Fala com eles, conhece os seus sentimentos,	Talk to them, learn their feelings,
	faz com eles as mesmas contas que um dia fez: quanto custa uma casa lá na terra deles, quanto se ganha, quanto é necessário para se viver.	do with them the same counts as you once did: how much does a house cost in their land, how much does one earn, how much is necessary to live.
C	Deseja escolher mais algum produto desta lista?	Do you want to choose one more product from this list?

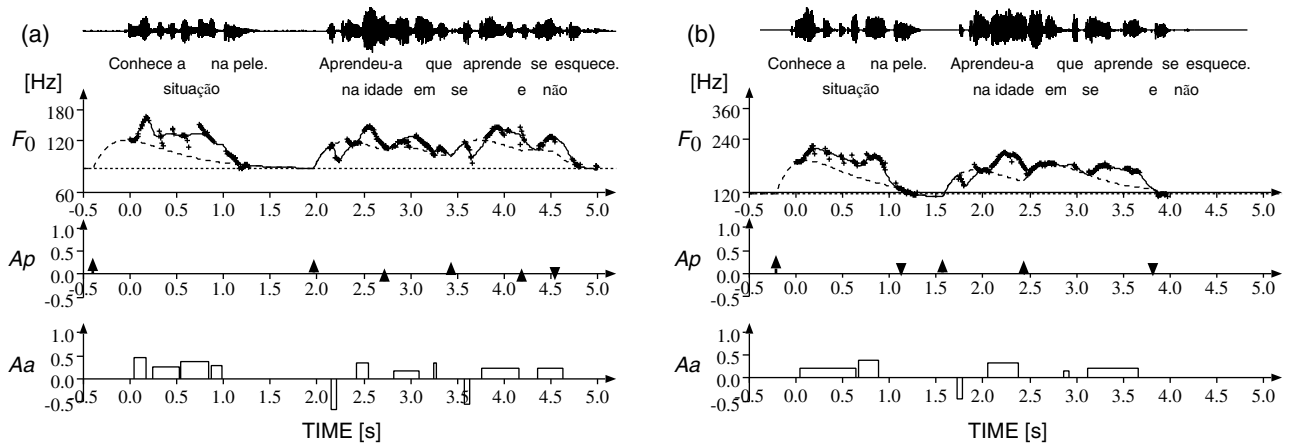


Figure 2: An example each of the F_0 contour analysis results for the utterances of Text A by one male and one female speaker of European Portuguese. (a) male speaker DF, (b) female speaker mb.

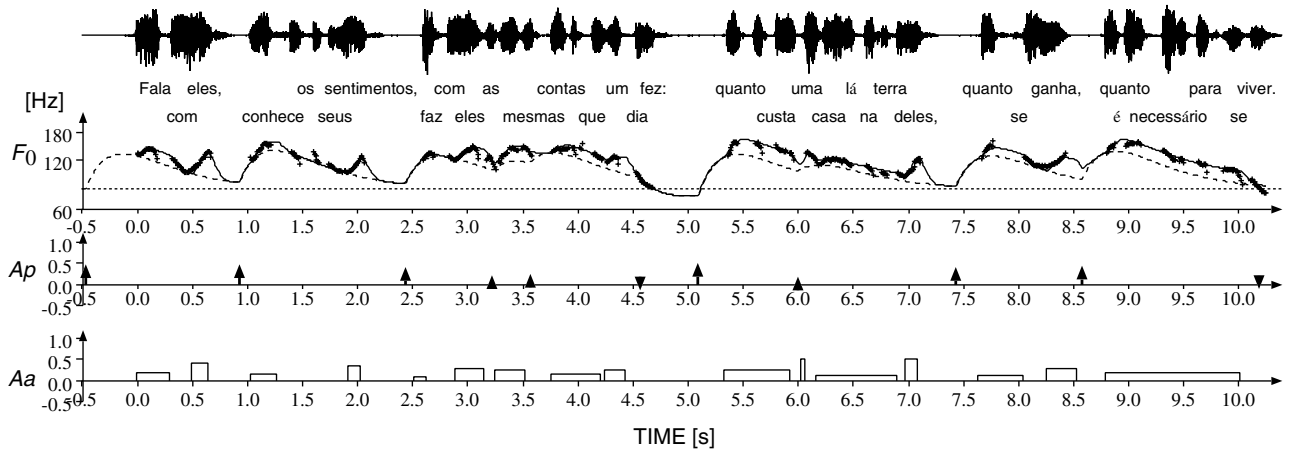


Figure 3: An example of the F_0 contour analysis results for the utterances of Text B by another male speaker (JT) of European Portuguese.

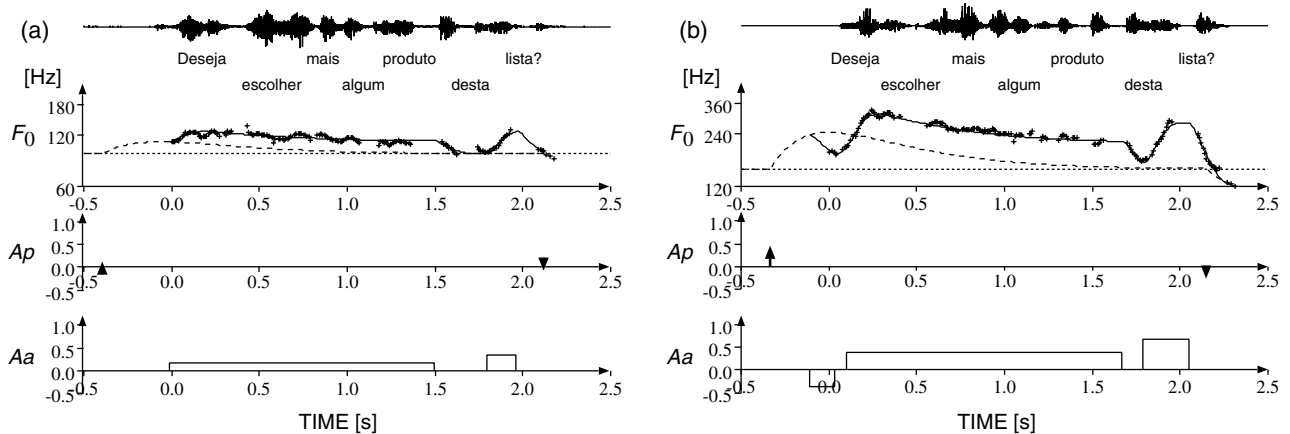


Figure 4: An example each of the F_0 contour analysis results for the utterances of Text C by still another male (FM) and another female (db) speaker of European Portuguese.

third being the last of the three imperative clauses and the sixth being the last of the three wh-question clauses.

The panels (a) and (b) in Figure 4 show the results of analysis of one sample each of the five utterances of Text C by still another male (FM) and another female (db) speaker. These pan-

els indicate that F_0 contours of yes/no questions are generally characterized by rather monotone intonation except for the final rise, which, unlike the continuation rise, occurs only at the main syllable of the utterance-final word (the penultimate syllable in this case) so that the rise is accompanied by a subsequent

fall. Comparison of panels (a) and (b) also indicates that there is no fall-rise of appreciable size in the utterance of the speaker (FM) in (a) but the fall-rise is quite large in the utterance of the speaker (db) in (b).

Analysis of all the 75 speech samples indicates that the presence/absence of negative accent commands at certain phrase-initial positions are quite consistent within one speaker. Although we need to analyze much larger amount of data to come to a reliable conclusion, the results so far obtained from the current speech material show that the use of fall-rises is definitely dependent on the speaker, and suggest that it may be influenced by regional, dialectal or stylistic factors.

Although our experiences on other languages indicate that the parameters α , β and F_b have very little variations within a speaker and can be regarded to be constants, and that inter-speaker variations of α and β are also rather small so that they can be regarded as speaker-independent constants for the purpose of analysis and synthesis, the current material allows us to compare inter-speaker variability of these parameters. Table 2 lists the averaged values for the three parameters of the model obtained from the analyses of 15 utterances each from each speaker.

Table 2: Averaged values of α , β and F_b for each speaker.

Speaker	DF	JT	FM	mb	db
α [s ⁻¹]	2.7	2.9	3.0	2.8	3.0
β [s ⁻¹]	20.2	22.2	20.0	20.5	20.0
F_b [Hz]	82.5	79.1	93.2	120.5	137.3

5. Preliminary perceptual experiment

The wide variations in the rate of occurrence, location as well as the speaker-dependency of the fall-rise patterns observed at certain phrase-initial positions motivated us to examine their perceptual significance by the following preliminary listening test.

Two sets of synthetic stimuli were prepared starting from speech samples whose observed F_0 contours clearly contained fall-rise patterns. One was a set of resynthesized speech with F_0 contours generated from the parameters of the model with accent commands of both positive and negative polarities. The other was a set of resynthesized speech with F_0 contours generated from the parameters of the model with accent commands of only positive polarity. The parameters of the latter model were extracted from observed F_0 contours after removing the local fall-rise patterns as though they were micro-prosodic disturbances, whose perceptual effects are known to be quite small, if not at all. Analysis-resynthesis was conducted by the Psola method, which introduced negligibly small degradations in speech quality.

The test procedure was based on paired comparison. The subject's task was to tell whether the pair of stimuli were judged to be prosodically the same or different. Since native speakers were not available, the subjects were two non-native speakers of European Portuguese. The results indicated, however, that the presence/absence of initial fall-rise is hardly perceived except in the case of yes/no questions. Although we need to confirm this by a formal listening test involving native speakers and greater variety of stimuli, the result of this preliminary experiment at least suggests that the role of the fall-rise may not necessarily be linguistic but may be para- and non-linguistic. Further study is certainly necessary to elucidate its nature and possible functions.

6. Summary and conclusion

This paper has described preliminary results of a joint work on the use of the command-response model for the analysis and synthesis of F_0 contours of European Portuguese. It was shown that the observed F_0 contours of European Portuguese can be approximated quite well by a model with provisions for both positive and negative accent commands, indicating the model's usefulness for TTS systems for European Portuguese. Analysis of a number speech samples by five native speakers, containing declarative, imperative, wh-question and yes/no question sentences/clauses indicated that negative accent commands are occasionally used in phrase-initial positions but their occurrence is not deterministic but is influenced by various factors. Further work is obviously necessary to elucidate its role as well as to establish a reliable method for automatic estimation of model parameters.

7. Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research on Priority Areas (B) No.12132102 "Analysis, Formulation, and Modeling of Prosody" (Principal Investigator: Hiroya Fujisaki) from the Ministry of Education, Culture, Science and Technology of Japan. The authors wish to acknowledge the colleagues at the University of Porto and the Polytechnic Institute of Bragança for their cooperation and assistance.

8. References

- [1] Fujisaki, H., "The fundamental frequency contour of speech — Its modeling, underlying mechanisms, and application to multilingual speech synthesis," Proceedings of ICSP '99, vol. 1, pp. 19–26, 1999.
- [2] Fujisaki, H., Ohno, S. and Luksaneeyanawin, S., "Analysis and synthesis of F_0 contours of Thai utterances based on the command-response model," To appear in Proceedings of ICPHS 2003, Barcelona, 2003.
- [3] Navas, E., Hernandez, I. and Sanchez, J. M., "Basque intonation modeling for text to speech conversion," Proceedings of ICSLP 2002, vol. 4, pp. 2409–2412, 2002.
- [4] Bailly, G., Contribution a la determination automatique de la prosodie du Francais parle a partir d'une analyse syntaxique. Etablissement d'un modele de generation, PhD thesis, Institut National Polytechnique, Grenoble, 1983.
- [5] Rossi, P. S., Palmieri, F. and Cutugno, F., "A method for automatic extraction of Fujisaki-model parameters," Proceedings of Speech Prosody 2002, pp. 615–618, 2002.
- [6] Freitas, D., Braga, D., Barros, M. J., Latsch, V. and Teixeira, J. P., "Correlation between phonetic factors and linguistic events regarding a prosodic pattern of European Portuguese: a practical proposal," Proceedings of ISCP 2001, vol. 2, pp. 975–980, 2001.
- [7] Freitas, D. and Braga, D., "Towards an intonation module for a Portuguese TTS system," Proceedings of ICSLP 2002, vol. 1, pp. 161–164, 2001.
- [8] Narusawa, S., Minematsu, N., Hirose, K. and Fujisaki, H., "Automatic extraction of model parameters from fundamental frequency contours of English utterances," Proceedings of ICSLP 2002, vol. 3, pp. 1725–1728, 2002.