# MODELLING BASQUE INTONATION USING FUJISAKI'S MODEL AND CARTS.

Eva Navas, Inma Hernáez, Ana Armenta, Borja Etxebarria, Jasone Salaberria

(eva, inma, ana, borja@bips.bi.ehu.es)

## Abstract

In this paper we present an analysis of the intonation for the Basque language. Based on Fujisaki's model, parameters for a set of about 300 sentences have been manually and automatically calculated and used to train a set of regression trees. Results show that the model is valid for Basque.

## 1 Introduction

Our previous Text to Speech Converter for Basque Intonation used a simple intonation model based on peaks assigned to the accented syllables and valleys between them, with very simple rules to assign declination and to characterise the different intonation structures. Now, we have chosen Fujisaki's model [1] to synthesise Basque intonation, mainly because it has proven to be a valid model for many languages [2][3][4].

## 2 Speech material

To prove the validity of the model for Basque, a first corpus including about 300 declarative and interrogative simple sentences was used. All of them were uttered by a female speaker in her native variety (see [2] for a description of this variety). A database has been created for the purpose of intonation studies and specifically for the analysis of the influence of the position of the focus and the lexically stressed words in the overall intonation curve [5]. It was recorded on a minidisk in a silent environment at the speaker's home, and digitised at 16KHz with 16 bits per sample.

It includes 121 short declarative sentences, 129 declarative sentences including lexically accented words and 98 question sentences, composed by 2, 3 or 4 short phrases formed with words of 2,3 or 4 syllables. Focus position and the position of lexically accented and unaccented words were combined in different ways.

All the sentences have been manually segmented into phonemes, words and phrases and their F0 curves have been calculated using a method based in [6].

## 3 Database labelling

The database parameterisation has been done both by hand and automatically. The goal of the manual task was to deduce linguistic constraints for the model that would help the automatic program. Having these manual parameters, they have also been used to validate the automatic process and a comparison between the two parameterisations has been made.

### *3.1 Manual process tool*

Manual parameterisation has been performed with the help of a graphic tool called AhoFuj, specifically designed to perform this task. It dynamically represents the F0 curve calculated from the Fujisaki's model parameters. It also allows the acoustical evaluation of the results.

Affiliation: *Dept. of Electronics and Telecommunication.*
*University of the Basque Country*

The impulses and pulses of the model can be inserted and modified graphically using the keyboard and saved to a text file (from where they can be loaded later). The effects of the modifications of the parameter values on the final synthetic pitch curve are seen on the fly, which permits a very fast manual adjustment of the model to the real pitch curve.

A vocoder has been incorporated to the program, so that original and synthetic pitch can be exchanged in the coded signal and both coded signals can be played, which permits an immediate perceptual evaluation of the parameters being edited.

Labels and time marks, taken from a text file, can be displayed to help the user to fulfil linguistic constraints if convenient. The program is configured (status and display options, file location, Fujisaki's model constants, frame rate of F0 synthetic curve and so on) trough a text file interface.

Figure 1 shows a typical AhoFuj session. In the upper window synthetic (dark line) and natural (light line) pitch curves are displayed. In the lower window, from top to bottom we can see the pulses and impulses of Fujisaki's model, accent group labels, position of the cursor in time axis and F0 axis (plus selection of natural or synthetic curve and F0min value) and selected pulse or impulse information. Finally the last line is to interact with the user.

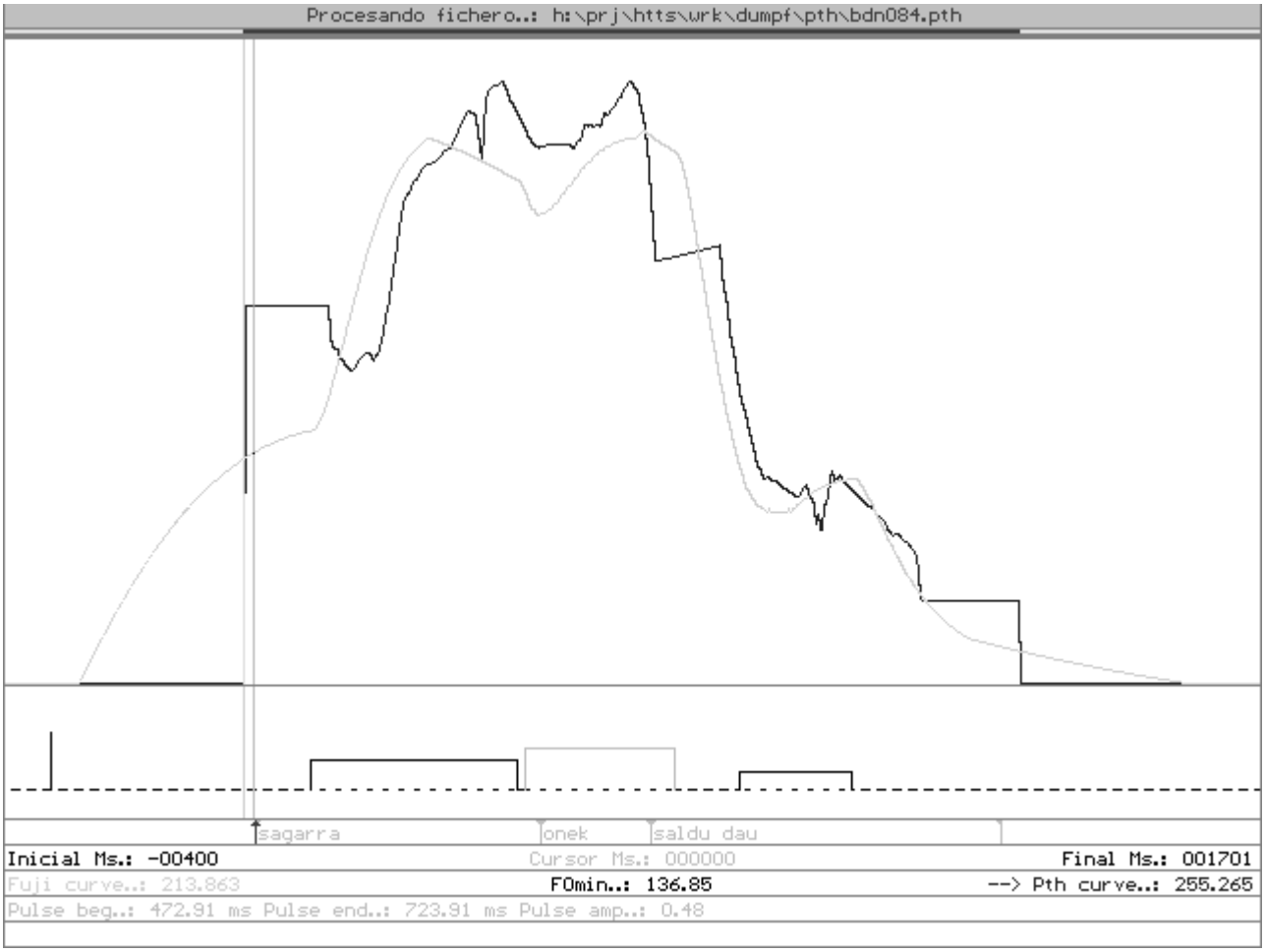


Figure 1: Manual labelling of pitch curves.

The AhoFuj program is also being used to check the fitting of the automatic parameterisation. The automatic parameters are loaded from a text file and the match between synthetic and natural pitch curves is visually and acoustically evaluated.

The program presently runs only under MSDOS and it is being ported to Windows9x and Linux.

### *3.2 Linguistic constraints*

After having parameterised a significant number of sentences of each type (declarative and question), we found that accent group marks, accent positions and sentence type are especially meaningful to fix the variation range of each parameter. Having this in mind, several restrictions to model parameters were deduced:

- Each sentence will have only one delta parameter that will give the global slope of F0 curve. The position of this parameter will be fixed 320 ms before the sentence starts. The amplitude only depends on the sentence type.
- Each accent group will be modelled by only one pulse command. The position of this command will depend on the position of the accent into the accent group:

  - If the accent goes in the first syllable of the accent group, the pulse may start before the start of the accent group and ends before the end of the accent group.

  - If the accent goes in the second syllable of the accent group, the pulse starts and ends inside the accent group.
- In the case of question sentences an extra final pulse will be added to model the final positive slope of the F0 curve. This pulse starts after 60% of the last accent group duration has passed.
- A minimal distance between pulses and a minimal pulse duration has been observed too. We have set those values to 10-20 ms and 80-100 ms respectively.
- Small variations of the pulse positions, lengths and amplitudes are not perceived.

### *3.3 Automatic process*

To obtain a good model of intonation for Basque, i.e. one that is representative enough, many of sentences must be analysed and labelled. So it was necessary to speeding the process of pitch-curve approximation and an automatic program called AutoFuji has been developed to perform this task.

Automatic extraction of the parameters is done using Analysis by Synthesis. This automatic process tries to obtain F0min and the parameters which define the phrase and accent commands of Fujisaki's model that generate the best synthetic pitch curve, i.e. the one that has the minimum square error with the natural curve. The information needed by this process is the real pitch curve and some linguistic information about the sentence.

To obtain the best approximation and find the minimum MSE solution an exhaustive search is made. The process implies testing many values for each parameter and the computational time becomes prohibitive if no linguistic constraints are applied. So the search is limited to the set of parameter combinations allowed by those linguistic restrictions.

Beside those linguistic constraints that may set the initial and/or the final value for that parameter, we also limit the search domain for each parameter by doing quantization of the parameters.

A configuration program uses linguistic information to determine the initial values and ranges of variations of the parameters is used together with the natural pitch curve by the AutoFuji program which returns the best parameter combination among the allowed ones. Figure 2 shows the structure of automatic parameter extraction process.

Fujisaki's model produces a continuous curve, without making any difference between the voiced and unvoiced frames. Our automatic program allows the user to consider the error in the unvoiced frames to calculate the MSE solution.

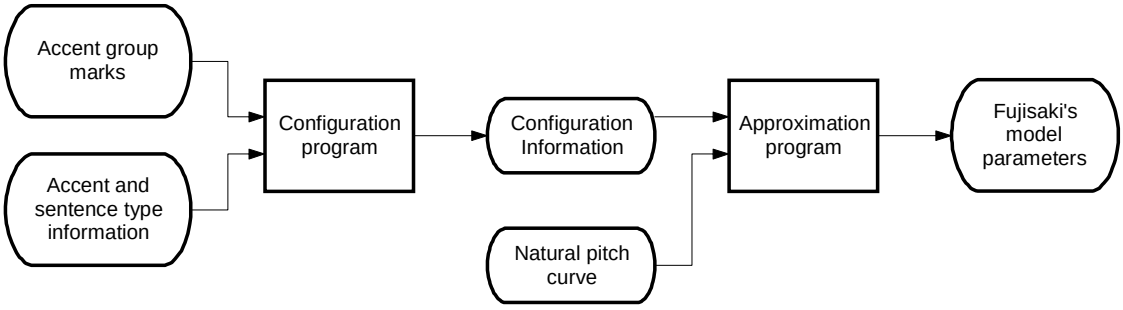


Figure 2: Automatic labelling of pitch curves.

The Figure 3 depicts the result of the automatic parameterisation of the same sentence showed in Figure 1.

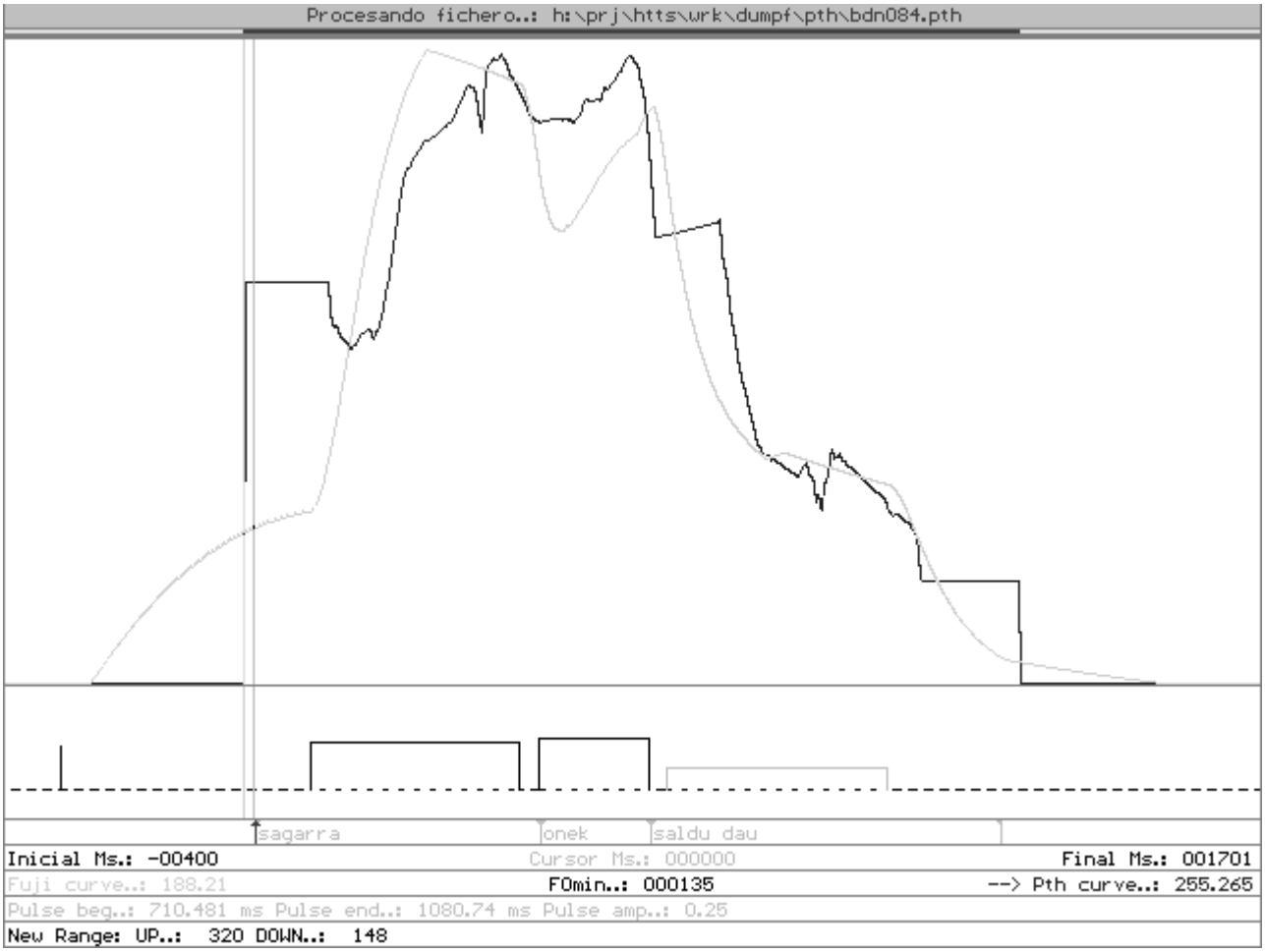


Figure 3: Result of automatic parameterisation.

To get a quantitative idea of the difference between manual and automatic parameterisation, a comparison has been made, using most of the sentences of the database. Table 1 shows the medium error (calculated by summing the absolute value of the differences between each manual and automatic parameter and dividing the result by the number of cases studied) and the medium relative error (calculated in the same way, but dividing each difference by the absolute value of the manual parameter).

| | **Pulse amplitude** | **Pulse position** | **Pulse duration** |
|---|---|---|---|
| Medium error | 0.127 | 120.888 ms | 88.828 ms |
| Medium relative error | 0.307 | 1.198 | 0.357 |

Table 1: Errors between manual and automatic parameterisation.

An error was expected because in the manual case, the delta parameters were variable, whilst in the automatic case they were fixed to a medium value, so pulse parameters had to compensate for this effect. We should not forget the strong quantization introduced in the automatic process.

# 4 Parameter estimation from text

Statistical analysis of the parameters has been made using commercial software to train binary regression trees [7]. For the pulse commands one tree has been built for each parameter, i.e. one for the amplitude, one for the position and another one for the duration of the pulse. In our data set, delta parameters were very uniform, so no tree has been used to predict them. Linguistic constraints are used here as well. Each parameter corresponding to a pulse is directly related to the characteristics of the accent group and sentence that correspond to that particular pulse. To train a pulse we just consider the acoustical properties of the accent group and sentence to which it belongs. The features extracted from text are listed and described in Table 2.

| Var name | Description | Var name | Description |
|---|---|---|---|
| msi<br>msi_rsn | AG* start time, idem normalised to sentence duration | pal<br>pal_rnpalsn | position of current AG, idem normalised to nº of AG in sentence |
| msf<br>msf_rsn | AG ending time, idem normalised to sentence duration | msi_acc<br>msi_acc_rp<br>msi:acc_rsn | distance from AG's start to first accent of AG, idem normalised to AG duration and to sentence duration |
| dpal<br>dpal_rsn | AG duration, idem normalised to sentence duration | dgl<br>dgl_rsn | distance between the focus and the current AG, same normalised to sentence duration |
| tsn | type of sentence | palsn | position of AG in the sentence |
| tpul | type of pulse | gl | index of focused AG |
| tacc | accent type: lexically marked or unmarked | gl-pal | nº of AG between the focus and the current AG |
| df | utterance duration | acc | position of the first accent in the AG |
| ipul | pulse index | | |

Table 2: Predictor variables used to train the trees (AG* = Accent Group).

For the sentences that had not lexically accented words, i.e. accent position was the same for all the accent groups in the database, a first experiment was made without using accent information. Two trees were built, one with the parameters of manual parameterisation and another one with the automatic extracted parameters. Curves predicted by those trees can be seen in Figure 4, which also shows the curves corresponding to parameters obtained in the manual and automatic processes for the same test sentence.

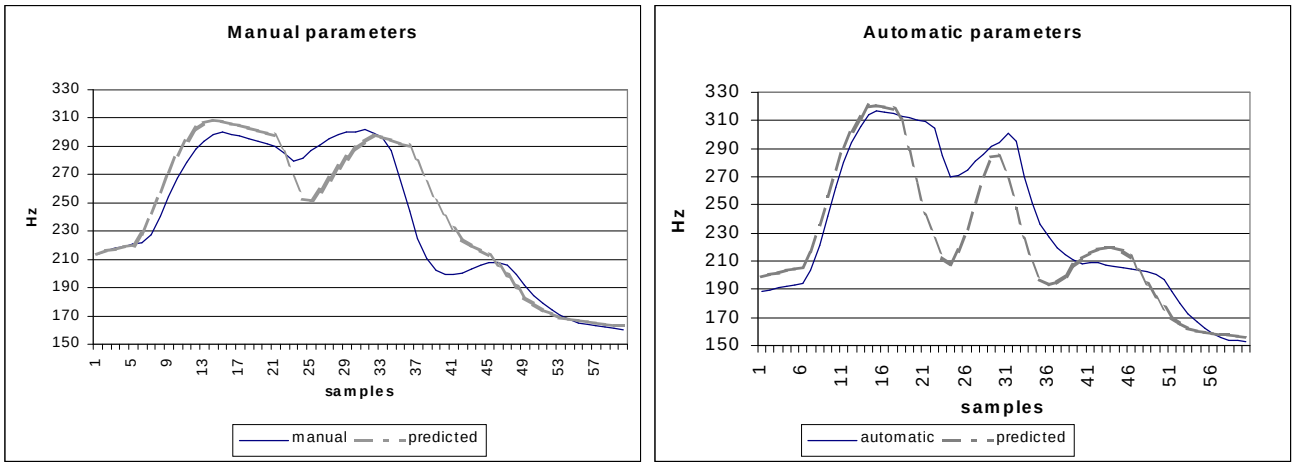


Figure 4: Comparison between pitch curves generated with predicted and "real" parameters for a sentence with three accent groups without lexically marked words.

For the sentences having lexically accented words, a first set of trees was built without accent information. In a second set of trees accent information was introduced. Dealing with this type of sentences, the accent information has proven very useful as shown in Table 3 which summarises the variables that have been used to predict each pulse parameter. Particularly for the pulse position it is essential because this parameter is determined only by accent position. The second meaningful variable is focus position, but it is much less important. Amplitude trees show that amplitude depends mainly on the focus position, then on the ending time of the accent group and finally on the accent position. Pulse duration

depends on accent group starting and ending times, then on the position of the accent group in the sentence and finally on the duration of the accent group.

We justify the dependency on ending time to the special characteristics of our database, and we would expect this variable to be ignored in a heterogeneous database.

| Position trees | | Amplitude trees | | Duration trees | |
|---|---|---|---|---|---|
| **variable** | **importance** | **variable** | **importance** | **variable** | **importance** |
| msi_acc | 100.00 | dgl_rsn | 100.00 | msf_rsn | 100.00 |
| msi_acc_rp | 79.18 | dgl | 69.89 | pal_rnpa | 94.64 |
| msi_acc_rsn | 76.71 | gl_pal | 67.61 | msi_rsn | 64.31 |
| dgl_rsn | 18.99 | msf_rsn | 58.67 | tpul | 19.86 |
| dgl | 17.82 | ipul | 44.66 | dpal | 16.91 |
| gl_pal | 17.36 | acc | 24.94 | dpal_rsn | 11.44 |

Table 3: Variable importance in pulse parameter prediction.

# 5 Conclusions and future work

Here we have presented the results corresponding to sets of trees trained separately for each sentence type, but global training has also been done. All this work is the preliminary phase to a more general analysis that will consider more complex sentences including pauses and more varied syntactic structures. A new database is being created, standard Basque has been used and data are expected to be much more heterogeneous.

# 6 Bibliography


[1] H. Fujisaki, K. Hirose
*Analysis of voice fundamental frequency contours for declarative sentences of Japanese*
Journal of Acoustic Society. Jpn. (E) 5, 4. 1984

[2] H. Fujisaki, S. Ohno
*Analysis and modelling of fundamental frequency contours of English utterances*
Eurospeech'95, pp 985-988. Madrid, Sep. 1995

[3] H. Fujisaki, S. Ohno, T. Yagi
*Analysis and modelling of fundamental frequency contours of Greek utterances*
Eurospeech'97, pp 465-468. Rhodes, Sep. 1997

[4] C. Wang, H. Fujisaki, S. Ohno, T. Kodama
*Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model*
Eurospeech'99, pp 1655-1658. Budapest, Sep. 1999

[5] G. Elordieta, I. Gaminde, I. Hernáez, J. Salaberria, I. Matín de Vidales
*Another step in the modelling of Basque intonation: Bermeo*
Lecture Notes in Computer Science; Vol 1692: Lecture Notes in Artificial Intelligence pp 361-364, 1999

[6] D. Griffin, J. S. Lim
*Multiband excitation vocoder*
IEEE Trans. ASSP. Vol 36, N 8. August 1988

[7] L. Breiman, J.H. Friedman, R.A. Olsen, C. J. Stone
*Classification and Regression Trees*
Chapman&Hall, 1984