

mCART, a multiple intonation model training package.

Pablo Daniel Agüero, Jordi Adell, Javier Perez and Antonio Bonafonte

May 11, 2007

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Intonation modelling</b>	<b>4</b>
2.1	Corpus based training of intonation models . . . . .	5
2.1.1	Parameterization . . . . .	6
2.1.2	Model training . . . . .	7
<b>3</b>	<b>Joint Extraction and Modelling Approach</b>	<b>9</b>
3.1	JEMA methodology . . . . .	9
3.2	Bézier intonation model . . . . .	11
3.2.1	Superpositional Bézier intonation model using JEMA . . . . .	11
3.3	Fujisaki’s intonational model . . . . .	13
3.3.1	Closed-form determination of amplitude parameters . . . . .	14
<b>4</b>	<b>mCART usage</b>	<b>16</b>
4.1	Command-line parameters . . . . .	16

# 1 Introduction

mCART is a complete intonation model training package. This software eases the generation of intonation models for TTS with many command-line configuration parameters. Three different mathematical formulations are implemented: Bézier, Fujisaki and Tilt. Each formulation can be trained by means of the two available procedures: SbS and JEMA. Several training modes are available too: train&test, n-FOLD cross-validation and full-training. Some of these modes can be used for research purposes to study the performance of each training procedure (and look for improvements by refining the techniques).

In this document you can find three different parts. Some of them may be skipped depending on the needs of the reader.

Section 2 is a brief introduction to intonation modelling with a critic view of the field.

Section 3 is a complete review of our Joint Extraction and Modelling Approach proposed in several papers [Agü04a, Agü04b, Roj05]. This technique is compared with state-of-the-art approaches (sentence-by-sentence parameter extraction: SbS). Both approaches (JEMA and SbS) are implemented in mCART.

In Section 4 is explained "how-to" use the software. Formats of the input files are defined and command-line parameters are described.

For any further information you may contact the author of the software: Pablo Daniel Agüero (pdaguero@gps.tsc.upc.edu or pdaguero@fi.mdp.edu.ar).

## 2 Intonation modelling

The intonation model is an important component of the Prosody Module of text-to-speech systems. It generates a fundamental frequency contour that is suitable for the input text that is synthesized. The generation process uses information provided by upstream components, such as syllabification, stress, phonetic transcription, part-of-speech tagging, syntactic analysis, prosodic boundaries. In the following example we show the input text with additional information about prosodic boundaries (\$), pauses (\$\$), phonetic transcription, stress (') and syllabification (-):

”Mr President, \$\$ on behalf of the European Liberal Democrat \$ and Reform Group \$\$ I congratulate the president of the commission \$\$ on assembling a talented team \$ of new commissioners \$\$ from the ten new member states, \$\$ France \$\$ and Spain.”

m,I,s,-,t,'@,|,p,r,'e,-,z,@,-,d,@,n,t,|,\_,|,Q,n,|,b,'I,-,h,A:,f,|,Q,v,|,D,'@,|,j,U@,-,r,@,-,p,'I@,n,|,l,'I,-,b,@,-,r,@,l,|,d,e,-,m,@,-,k,r,'{,t,|,\_,|,'{,n,d,|,r,I,-,f,'0:,m,|,g,r,'u:,p,|,\_,|,'aI,|,k,@,n,-,g,r,'{,-,tS,@,-,l,eI,t,|,D,'@,|,p,r,'e,-,z,@,-,d,@,n,t,|,Q,v,|,D,'@,|,k,@,-,m,I,-,S,'@,n,|,\_,|,Q,n,|,@,-,s,'e,m,-,b,l,-,@,N,|,'@,|,t,'{,-,l,@,n,-,t,@,d,|,t,'i:,m,|,Q,v,|,n,j,'u:,|,k,'@,-,m,I,-,S,@,-,n,@,z,|,\_,|,f,r,'@,m,|,D,'@,|,t,'e,n,|,n,j,'u:,|,m,'e,m,-,b,@,|,s,t,'eI,t,s,|,\_,|,f,r,'A:,n,s,|,\_,|,'{,n,d,|,s,p,'eI,n

Therefore, the available information for the intonation model consists of words, punctuation, syllables, phonemes, prosodic boundaries, accents, etc. In some situations additional labels may be available: gender, speaker, emphasis, attitude, emotion, dialog act, etc [Shr98, Cam04]. Using this information, the intonation model generates a suitable fundamental frequency contour for the text. This contour is used by the following modules to produce the synthesized speech.

We can consider the intonation model as a function that maps linguistic and paralinguistic input features onto a fundamental frequency contour at the output:

$$G(F) = \hat{F}_0$$

where  $F$  are the set of input linguistic and paralinguistic features,  $G()$  is the mapping function (intonation model), and  $\hat{F}_0$  is the generated fundamental frequency contour (prediction of the model).

The resulting fundamental frequency contour may be different to the fundamental frequency contour produced by a human given the input text (reference contour). The generated contour has an error ( $\epsilon$ ) compared to the reference contour:

$$G(F) = F_0 + \epsilon$$

The intonation model is built minimizing the error  $\epsilon$ . The predicted contour must be as close as possible to the fundamental frequency contour produced by a human.

The intonation model can be built using two approaches: manual and automatic methods.

Manual methods rely on human expertise to produce a set of rules that explain the fundamental frequency contour behaviour given the linguistic and paralinguistic features. These rules are used in speech synthesis to generate the fundamental frequency contour. They are hard to write because it is difficult to cover all possible situations that may arise. Additionally, migration to a new domain requires new analysis by the expert and it is a time-consuming process.

On the other hand, automatic methods make use of machine learning techniques to produce an intonation model given input data. Input data is analyzed looking for regularities. These regularities are expressed as rules that explain the behaviour of the data (training phase).

Manual methods are useful because they provide a deeper insight about the task. However, it is a time-consuming process. Therefore, it is advisable to use these manual methods to get knowledge that can be applied to automatic methods. Machine learning techniques may take advantage of these features observed in the manual analysis to build better intonation models.

We must point out that the achievable quality of the intonation model has an upper-bound limit because of several reasons that are far away from being solved using nowadays techniques:

- **The fundamental frequency contour depends on the choice of the speaker.** There is a mapping of one-to-many from a sentence to the possible fundamental frequency contour space. The speaker may choose many different realizations of the fundamental frequency contour and the choice is completely random without any possible explanation for it. It is difficult to measure the accuracy of a system and the modelling task because it is not possible to have all possible references to compare with. The reference is dependant on the speaker's choice and introduces difficulties to objectively measure the accuracy of the system. Objective measures are in general pessimistic for prosody analysis.
- **The information in a text is not enough to perform natural intonation.** The fundamental frequency contour is closely related to the interpretation of the meaning of the sentence by the speaker. The speaker uses information that is not contained in the input text to produce an adequate intonation. This information is the knowledge that we acquire during our life and is used to understand the text. The strong limitations in the understanding capabilities of the machines is an upper-bound to the achievable quality of the intonation. A computer can not transmit an intention nor emotions analyzing the input text because a computer can not understand a text.

For example, in the sentence "The police found two grams of cocaine in the bag of the student" the word "cocaine" may be emphasized because the speaker may want to highlight the drug that was found. On the other hand, in the sentence "The police found half kilogram of cocaine in the bag of the student" the words "half kilogram" and "cocaine" may be highlighted because it is also an important thing the amount of drug that was found.

How can a computer infer the words that must be highlighted in a text? It is a complex task that needs of world knowledge that is far from being part of the information a computer can analyze. Hence, it is not possible to produce high quality intonation models because of the lack of knowledge of computers.

- **Fundamental frequency measurement errors.** The fundamental frequency contour extracted from an utterance with training purposes has measurement errors and microprosody. Common pitch extraction errors are:
  - Pitch halving. The detected pitch is half the real pitch.
  - Pitch doubling. The detected pitch is double the real pitch,
  - Voiced to unvoiced transitions. It is difficult to measure the pitch in voiced to unvoiced transitions because of coarticulation effects.
  - Microprosody. The articulation of phonemes introduces perturbations in the pitch (microprosody).

Filtering techniques are applied to remove these effects, but some important information can also be lost or some bias may be introduced.

## 2.1 Corpus based training of intonation models

Nowadays intonation models are generated using corpus-based approaches. Machine learning techniques are applied to the corpus to perform data mining and extract regularities in the behaviour of fundamental frequency contour related to linguistic and paralinguistic features (intonation model training).

Machine learning techniques are widely used in intonation modeling because of several advantages:

- **Automatic training.** Hand-written rules require highly skilled persons and a long development time. It is preferable to use automatic techniques that may find regularities in the training corpora.
- **Fast adaptation to new domains using corpora.** The adaptation time to new domains is important to avoid developing from scratch a system. Machine learning techniques may extrapolate knowledge from one domain to the other. It is also possible to train an intonation model from scratch with little effort.
- **Management of continuous and discrete features.** Linguistic features used in text-to-speech systems include continuous (e.g.: number of syllables) and discrete features (e.g.: part-of-speech tags). Machine learning techniques may provide new knowledge about the task analyzing the rules obtained after training.
- **Unseen cases.** Machine learning techniques can find structure in the data and extrapolate to unseen cases. However, limitations appear because of missing input features. Many features can not be obtained from text because of limitations in the understanding capabilities of computers.

- **Objective scoring.** Machine learning techniques use scores to measure the goodness in the process of training. These goodness measures are objective measures that in some cases are not closely related to the psychoacoustics. In general, they measure the correlation or root mean squared error between the original and predicted contours. These global measures do not focus on certain local effects that produce a lower mean opinion score (MOS), for example: a time shift in stress.

According to the training algorithms proposed in the literature, the training process using machine learning techniques can be divided in two steps: parameterization and model training.

### 2.1.1 Parameterization

Parameterization is an important step because the pitch contour can be described using a fixed set of parameters. A parameterization eases the treatment of pitch contours with different durations and pitch events located in different positions. These facts will be explained more clearly with the Tilt parameterization.

Taylor [Tay00] proposed an intonation model characterised by a sequence of phonetic intonational events: pitch accents and boundary tones.

Each event has a rise component, a fall component, or both. The parameterization is shown in Figure 1. The amplitude and duration parameters describe the trajectory of the pitch curve. However, these parameters are relative to the other two parameters  $F0_{peak}$  and  $t_{peak}$ .  $F0_{peak}$  and  $t_{peak}$  can be absolute or relative parameters.  $F0_{peak}$  may depend on the pitch range of the speaker while  $t_{peak}$  is relative to the time when the nucleus begins,

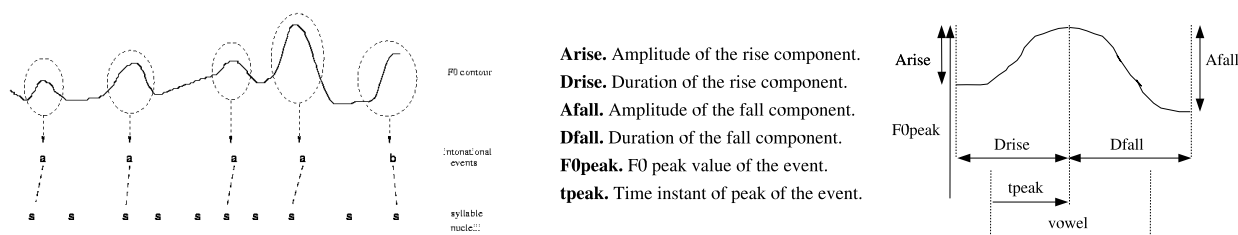


Table 1: Tilt parameters.

Therefore, using the Tilt parameterization we may explain pitch curves with different pitch ranges and event time instants with the same parameters. This homogeneity property is useful to find rules that explain the different set of values of the parameters given the linguistic and paralinguistic input features. These rules will have generalization properties that may help to predict pitch contours for unseen cases in the training data.

The parameterization is performed making some assumptions which may introduce noise in the extracted parameters:

- **Removal of noise and microprosody.** The fundamental frequency extraction of a speech signal is a task prone to errors: pitch halving, pitch doubling, microprosody and measurement errors in voiced-unvoiced boundaries are sources of noise. Smoothing techniques are applied to remove such effects. However, this smoothing process may introduce new noise. Contours that should have the same shape after filtering are not equal because filtering was driven to provide dissimilar results due to differences in pitch extraction errors.
- **Continuity of the fundamental frequency contour.** Some intonation models need continuous fundamental frequency contours to perform parameterization. Interpolation techniques are used to fill the unvoiced regions of speech. The main drawback is that the resulting contours may bias the parameter extraction.
- **Consistency of the parameterization.** The parameterization of a fundamental frequency contour is considered to be unique. However, some intonation models can not ensure this (e.g.: Fujisaki). In some intonation models multiple possible sets of parameters can provide good approximations to the fundamental frequency contour. This makes the prediction task more difficult, because similar contours can have different parameterizations, increasing the dispersion of the parameters. This effect occurs when the extraction is individually performed for each sentence.

### 2.1.2 Model training

Model training is the process of generation of a intonation model given a corpus of training data. A set of parameters are extracted for each prosodic unit and the linguistic and paralinguistic features are generated.

The training data format is shown in Table 2. Each row correspond to a prosodic unit in the training data. Columns  $P_1$ ,  $P_2$ , etc, are the parameters extracted from the pitch contour. Columns  $F_1$ ,  $F_2$ , etc are the linguistic and paralinguistic features that will be used to predict the parameters of the first columns.

$P_1$	$P_2$	...	$F_1$	$F_2$	...
$p_{1_1}$	$p_{2_1}$	...	$f_{1_1}$	$f_{2_1}$	...
$p_{1_2}$	$p_{2_2}$	...	$f_{1_2}$	$f_{2_2}$	...
$p_{1_3}$	$p_{2_3}$	...	$f_{1_3}$	$f_{2_3}$	...
$p_{1_4}$	$p_{2_4}$	...	$f_{1_4}$	$f_{2_4}$	...
...	...	...	...	...	...

Table 2: Training data.

In the literature several machine learning techniques are proposed to predict a set of parameters given a set of input features: classification and regression trees (CART), neural networks (NN), instance based learning (IBL), support vector machines (SVM), etc.

In this package CART will be used. A CART consists of a set of questions that splits the input data into subsets. The questions are about discrete and continuous features of the parameter or parameters that are being predicted.

For example, given the training data of Table 3 two possible trees are possible depending on the variable that is predicted: A (discrete variable-classification problem) or B (continuous variable-regression problem). In this example  $F_1$  and  $F_2$  are the features.  $F_1$  is a discrete feature and  $F_2$  is a continuous feature.

$A$	$B$	$F_1$	$F_2$
a	2	Y	10
b	3	X	9
c	10	X	2
a	3	Y	11
b	12	X	4

Table 3: Training data for CART.

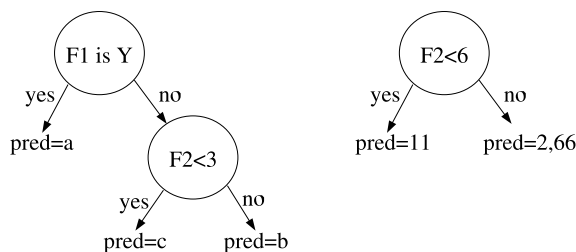


Figure 1:

Classification and regression trees have some advantages over other machine learning techniques:

- **Simplicity of results.** In most cases, the interpretation of results in a tree is very simple. The simplicity is not only useful for computational reasons (it is much more easy to evaluate a condition than calculating more complex mathematical formulations), but can also yield a much simpler model for explaining why the predictions are performed in a particular manner.
- **Nonparametric and nonlinear.** There is no implicit assumption about the relationships between the variable being predicted and the features. Thus, trees are particularly well suited for data mining tasks

where there is little a priori knowledge. Trees can often reveal simple relationships between just a few variables that could have gone unnoticed using other analytic techniques.

### 3 Joint Extraction and Modelling Approach

In the literature intonation models are trained using the two step approach explained in previous sections: parameterization and model training. It is also common to perform a smoothing of the original contours to remove noise and microprosody and interpolate unvoiced regions. The training scheme is shown in Figure 2.



Figure 2: Two step approach scheme.

This approach has some limitations explained in previous sections:

- **Interpolation of unvoiced regions is mandatory.** This process may introduce a bias in the extracted parameters.
- **Sentence-by-sentence parameter extraction.** The parameters are extracted sentence-by-sentence. This may introduce inconsistencies in the parameterization. Contours with similar shape may have a different set of extracted parameters. Some intonation models have mathematical formulations that are prone to it. Many sets of parameters can approximate the original contour.

#### 3.1 JEMA methodology

In our work we propose an intonation model training framework to avoid the problems of inconsistency and parameter extraction bias.

The framework consists of a combination of the parameter extraction and model training into a loop. In this way successive iterations are performed to train the intonation model.

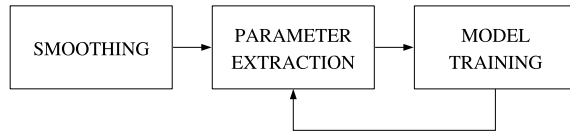


Figure 3: Joint extraction and mapping approach scheme.

In this joint approach, trees are used because of the previously mentioned advantages.

In figure 4 a corpus of two sentences is shown. First sentence has three prosodic units (1,2 and 3) and second sentence has two prosodic units (4 and 5). These prosodic units can be minor phrases, accent groups, syllables, etc. In this example we will consider accent group as the prosodic unit.

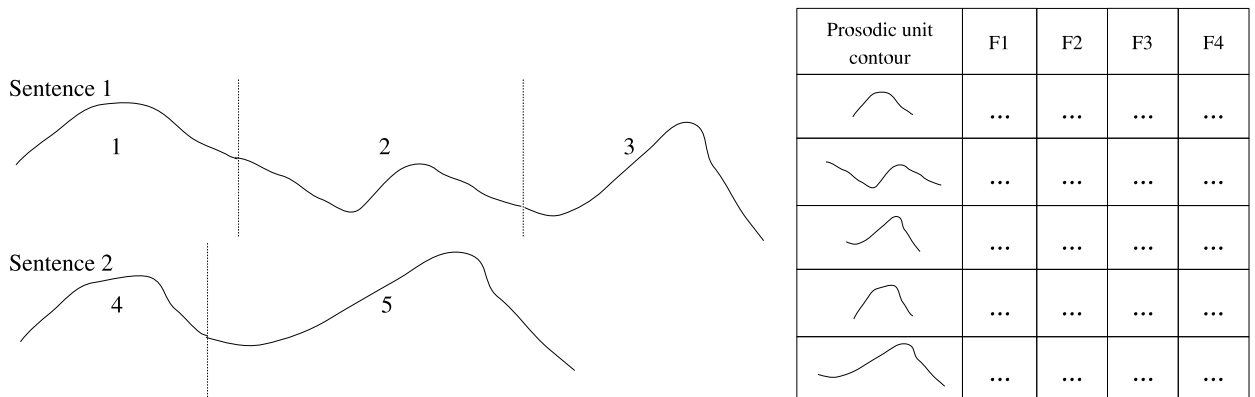


Table 4: Example of a corpus with two sentences. Prosodic units are numbered from 1 to 5.

The training database consists of all prosodic units and their corresponding linguistic and paralinguistic features. Figure 4 shows the training database. Each row correspond to a prosodic unit (in this example five

prosodic units). The first column is the original contour of the prosodic unit while the following columns are the features of the prosodic unit.

At the beginning, all prosodic units in the training database are considered to belong to the same class (class 0). All prosodic units are approximated by the same class contour (class 0). Class 0 contour is the contour that minimizes the approximation error over all contours of the prosodic units in the training data. The parameters of class 0 contour are globally optimized. Figure 4 shows the approximation (in red) for the corpus with two sentences.

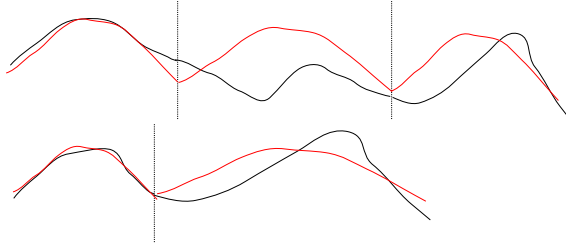


Figure 4: Approximation with class 0 contour.

The splitting of the training data into subclasses performing questions about linguistic and paralinguistic features reduces the approximation error. However, the increasing number of classes can also overfit training data and the generalization of the model will be poor. Stopping conditions are necessary to avoid this effect.

In Figure 5 is shown the best split after trying all possible questions over the features. Two new classes are created after splitting class 0. This splitting approximates well 4 prosodic units in the database. Prosodic unit 2 is not properly modelled yet. The approximation with two classes is shown in Figure 6.

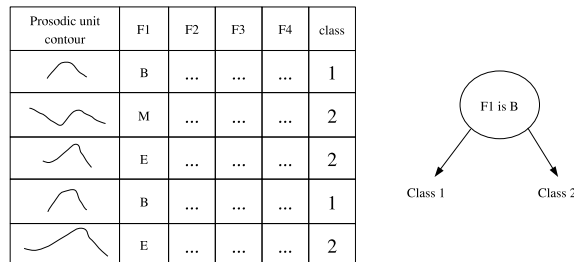


Figure 5: Approximation with two classes.

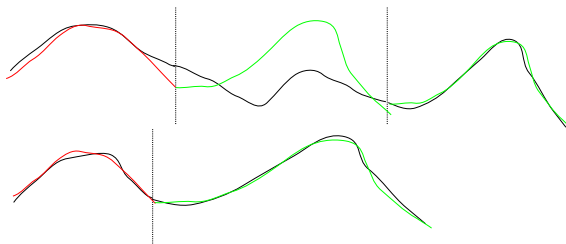


Figure 6: Approximation with two classes.

This splitting process continues until a stopping condition is reached.

The process for intonation model training using JEMA can be summarized into these steps:

- **Initialization.** Initially only one class exists, because the tree has only the root node. In this way, all prosodic units (accent groups, minor phrases) will be represented by the same set of parameters. These parameters are obtained using a global optimization algorithm over all training data.
- **Splitting.** Linguistic and paralinguistic features are used to do questions in the tree to split training data. After a new question is done, the training data will have two new classes obtained from the splitting of the previous class.

- **Optimization.** When the new classes are obtained, a global optimization algorithm is used to find the new optimal parameters. Depending on the parameterization, this optimization step can be time consuming if the optimal solution has not closed-form (e.g.: Fujisaki’s intonation model). In such cases hill-climbing algorithms are used to find the optimal solution.
- **Scoring of the splitting.** The new parameterization is used to measure the improvement of the goodness measure compared to its value previous to the splitting.
- **Selection of the highest improvement.** After all possible splittings were tried, the splitting with the highest improvement is chosen as the best split and the tree is updated for the next iteration.
- **Stopping condition.** The decision of another iteration for an additional splitting is performed taking into account a minimum number of elements on each leaf and a minimum improvement of the goodness score.

This approach can be applied to several parametric intonation models, because is a general technique to train intonation models. In some cases, such as in Fujisaki’s intonation model, is more time consuming, because the parameter optimization requires high-climbing optimization. There is not closed-form solution for this intonation model. On the other hand, Bézier intonation model has a closed-form solution, and the flexibility of the mathematical formulation allows a more complex modelling.

In this approach the parameters are extracted using a global optimization algorithm. As a consequence, the interpolation is not needed and the parameters are more consistent and not biased by interpolation. This approach provides improvements in models that have the intrinsic problem that several sets of parameters optimally approximate a given contour. This is the case of Fujisaki’s intonation model.

In this work we show how we can apply this training framework to two intonation models: Bézier and Fujisaki’s intonation models.

## 3.2 Bézier intonation model

Escudero et al [Esc02] proposed to model accent groups in Spanish using Bézier curves.

Bézier curves are based on a polynomial function. The coefficients allow a representation that is more meaningful than the resulting polynomial coefficients in expanded form.

The polynomial formulation is shown in equation 1 and the shape of the base polynomials for a fourth order curve are shown in Figure 7.

$$P(t) = \sum_n^N \alpha_n \binom{N}{n} t^n (1-t)^{(N-n)} \quad (1)$$

Figure 8 shows an approximation of a fundamental frequency contour using Bézier curves for accent groups, with continuity constraints up to the first derivative.

In his thesis, Escudero analyzes various ways to classify accent groups based on the proposals of Lopez [Lóp93], Garrido [Gar96], Vallejo [Val98] and Alcoba. In this way, a fundamental frequency contour can be predicted based on the linguistic features of the accent group. Here we propose to extend his work by training intonation models based on Bézier curves using JEMA. This technique will be applied to a superpositional Bézier intonation model (minor phrase and accent group components) and a non-superpositional Bézier intonation model (accent group component).

### 3.2.1 Superpositional Bézier intonation model using JEMA

The Bézier intonation model can be trained using the JEMA approach. The joint optimization framework imposes that the formulation to extract the optimal polynomial coefficients is modified. The optimization is performed minimizing the mean squared error, but taking into account that:

- The error that is minimized is the global mean squared error.
- Two components are combined using Bézier curves: minor phrase and accent group.
- The group of coefficients corresponding to a Bézier curve depend on a vector that maps minor phrase or accent group classes with positive integers (class number).

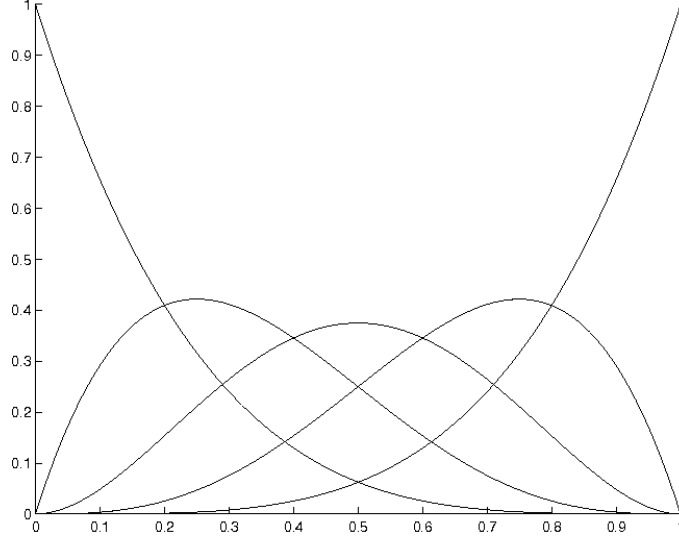


Figure 7: Bézier polynomials

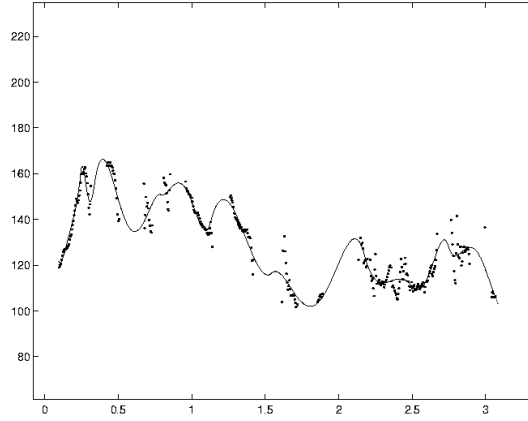


Figure 8: Fundamental frequency contour approximated using Bézier curves with five coefficients

The mathematical formulation is shown in equation 2.

$$F_0^k(t) = \sum_i^{N_{MP}^k} P_{MP_i}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG_j}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \quad (2)$$

where:

$N_{MP}^k$  is the number of minor phrases of the  $k$ th sentence.

$N_{AG}^k$  is the number of accent groups of the  $k$ th sentence.

$t_{MP_i}^k(t)$  is the temporal axis of the  $i$ th minor phrase of the  $k$ th sentence.

$t_{AG_j}^k(t)$  is the temporal axis of the  $j$ th accent group of the  $k$ th sentence.

$C_{MP_i}^k$  is the number of the minor phrase class assigned to the  $i$ th minor phrase of the  $k$ th sentence.

$C_{AG_j}^k$  is the number of the accent group class assigned to the  $j$ th accent group of the  $k$ th sentence.

In this function,  $P_{MP}$  and  $P_{AG}$  are the Bézier curves of the minor phrase and accent group components, respectively. Each curve has its own associated time axis,  $t_{MP}(t)$  and  $t_{AG}(t)$ . The time axis range is zero to one. These curves are zero elsewhere.

The joint cost function is shown in equation 3. The goal is to minimize the mean squared error. This equation has a unique analytical minimum that is found using a set of linear equations.

$$e = \sum_k^{N_s} \left( \sum_t^{T_k} \left( f_0^k(t) - \left( \sum_i^{N_{MP}^k} P_{MP_i}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG_j}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \right) \right) \right)^2 \quad (3)$$

where:

$N_s$  is the number of sentences.

$T_k$  is the duration of the sentence.

This polynomial representation of the fundamental frequency contour provides a high flexibility in the modelling. In order to avoid non closed-form solution, the knots must be in fixed positions. In this way the model flexibility is reduced but the calculus of the optimal coefficients is simplified.

Another drawback of the Bézier superpositional approach is that it is not possible to set continuity constraints for all contours. In this way, discontinuities appear in the joining point of the components. Depending on the choice of the prosodic units, this limitation can cause effects that are not desired. This problem can be overcome using smoothing in the joining points for each component.

### 3.3 Fujisaki's intonational model

The Fujisaki's intonational model (Fujisaki et al. [Fuj84]) is based on a physical model of the fundamental frequency production system. It is represented by two second-order filters. One filter is excited with pulses, and the other with deltas. The latter are related to phrase commands (time response function is shown in equation 5), and pulses are related to accent commands (time response function is shown in equation 6). A DC value ( $F_b$ ) is added to the output of these filters (see equation 4).

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^I A_{p_i} G_p(t - T_{0_i}) + \sum_{j=1}^J A_{a_j} G_a(t - T_{1_j}) - G_a(t - T_{2_j}) \quad (4)$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (5)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t) e^{-\beta t}, \gamma] & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (6)$$

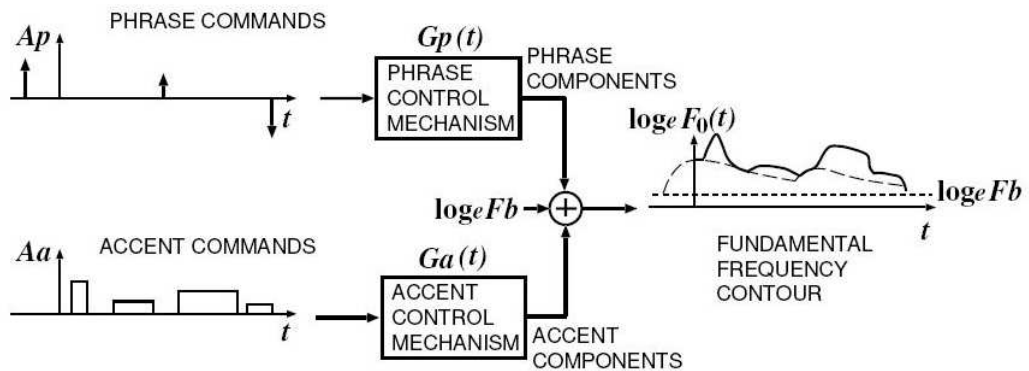


Figure 9: Fujisaki's model scheme.

This model has the support of a physiological basis for the mathematical formulation [Fuj00a]. However, its main drawback is that it is not possible to obtain a closed-form solution. In this way, hill-climbing techniques must be used to obtain an optimal solution.

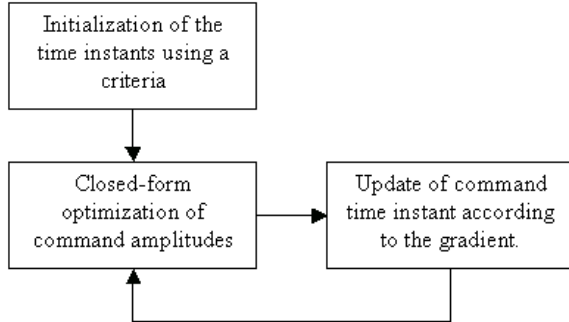


Figure 10: *Update loop.*

$f_b$	Scalar that represents the base frequency.
$u$	Vector of ones.
$G_p$	Matrix of the phrase command component.
$A_p$	Vector of the amplitude of the phrase commands.
$G_a$	Matrix of the accent command component.
$A_a$	Vector of the amplitude of the accent commands.

Table 5: Variable meaning

However, it is possible to obtain an optimal solution for the amplitude commands assuming that the time instants are known ( $T_0$ ,  $T_1$  and  $T_2$ ).

Several papers have proposed different ways to obtain Fujisaki’s parameters from the fundamental frequency contour.

Fujisaki et al [Fuj00b] suggest to extract parameters applying a preprocessing procedure which results in a continuous third-order polynomial stylization. Then, accent commands are searched in the derivative of this function. The residual after the subtraction of the detected commands from the original contour is used to detect phrase commands.

Mixdorff [Mix00] proposed a decomposition approach performing a spectral decomposition in low frequency (LFC) and high frequency (HFC) components of fundamental frequency contour.

Navas et al. [Nav02] proposed parameter extraction using grid search applied to the Basque language.

### 3.3.1 Closed-form determination of amplitude parameters

In this closed-form formulation it is assumed that the time instants are obtained using another approach. The optimal values of the time instants can be found using grid search or gradient descent techniques. In our case we used gradient descent, because it provides a more accurate solution.

The optimization loop is shown in figure 10. It is a combination of closed-form optimization of amplitude values, and the update of the values of the time instants according to the gradient. After some iterations, the optimal solution is found. This process is faster than the optimization of all the parameters using gradient descent.

The Fujisaki’s intonation model formulation can be expressed as shown in equation 7.

$$\hat{f}_0 = (\ln f_b)u + G_p A_p + G_a A_a \quad (7)$$

The meaning of the variables is shown in Table 5. In equation 7, the base frequency ( $\ln f_b$ ), the phrase command amplitude ( $A_p$ ) and the accent command amplitude ( $A_a$ ) are jointly optimized.

The  $f_0$  vector is a concatenation of all the contours of the training set. The  $G_a$  and  $G_p$  matrices have the corresponding components of the phrase and accent commands, respectively. The contours are normalized to one, and can be obtained because it is assumed that the time instants are known.

It is possible to have different classes of phrase and accent commands (with amplitudes given by  $A_p$  and  $A_a$ ). Each class approximates different segments of the fundamental frequency contour. The assignment is performed by the algorithm explained in section 3.1.

In order to find the optimal parameters of the model ( $lnf_b$ ,  $A_p$  and  $A_a$ ) it is necessary to formulate an optimization function. In this case, we will minimize the mean squared error of the approximation, as shown in equation 8.

$$e^2 = (f_0 - \hat{f}_0)^T (f_0 - \hat{f}_0) \quad (8)$$

Taking the derivatives with respect to  $lnf_b$ ,  $A_p$  and  $A_a$ , we obtain a system of equations that can be expressed in matrix notation as in equation 9.

$$\begin{bmatrix} u^T f_0 \\ G_p^T f_0 \\ G_a^T f_0 \end{bmatrix} = \begin{bmatrix} u^T u & u^T G_p & u^T G_a \\ G_p^T u & G_p^T G_p & G_p^T G_a \\ G_a^T u & G_a^T G_p & G_a^T G_a \end{bmatrix} \begin{bmatrix} lnf_b \\ G_p \\ G_a \end{bmatrix} \quad (9)$$

This set of equations allow a fast and accurate solution of the command amplitudes. In this way the training speed is improved for the joint extraction and prediction approach.

## 4 mCART usage

### 4.1 Command-line parameters

The command-line parameters of mCART are:

- **-agfields -f**

Name of the field description file. Two types of fields are admitted: enum (discrete) and float (continuous).

**Default value:** ag.desc

**Example:**

```
field1 enum
field2 enum
field3 float
field4 enum
```

- **-agboundary -b**

Boundaries of each accent group in the data.

**Default value:** ag.dat

**Example:**

```
3 0 1.174 1.578 1.578 2.062 2.062 3.098
2 1 1.19 1.75 1.75 2.526
```

**Description.** In this file we describe that the corpus has two sentences (two lines). The first sentence has three accent groups with the following time boundaries: [1.174-1.578], [1.578-2.062] and [2.062-3.098]. The second sentence has two accent groups with the following time boundaries: [1.19-1.75] and [1.75-2.526].

- **-agdata -d**

Values for each field of the accent groups in the data.

**Default value:** ag.data

**Example:**

```
A M 2 XX
B O 2 XX
C O 2 ZZ
B P 2 ZZ
```

B P 2 XX

**Description.** In `-fields` we defined that we have with four fields: three discrete (`field1`, `field2` and `field4`) and one continuous (`field3`). In `-agboundary` we defined that we have two sentences with three and two accent groups, respectively. In this file we describe the values of the fields for each accent group, sequentially.

- **`-agtree -t`**

Output tree for accent group after stopping condition is reached.

**Default value:** `ag.tree`

- **`-mpfields -g`**

Idem to `-agfields`, but it refers to minor phrase data.

**Default value:** `mp.desc`

- **`-mpboundary -c`**

Idem to `-agboundary`, but it refers to minor phrase data.

**Default value:** `mp.dat`

- **`-mpdata -e`**

Idem to `-agdata`, but it refers to minor phrase data.

**Default value:** `mp.data`

- **`-mptree -u`**

Idem to `-agtree`, but it refers to minor phrase output tree.

**Default value:** `mp.tree`

- **`-f0 -p`**

F0 values for each sentence in the data.

**Default value:** `f0.dat`

**Example:**

1506 1.19 240.68 1.195 225.265 1.2 209.009 1.205 192.147 1.21 182.738 1.215 180.168 1.22 178.66 1.225  
176.553 1.23 176.116 1.235 183.211 1.24 188.702 1.245 194.84 1.25 195.047 1.255 206.759 1.26 211.942 1.265  
212.795 1.27 213.703 1.275215.748 1.28 217.735 1.285 219.36 1.29 220.495 1.295 220.809 1.3

**Description.** Each line of the file has the number of points (1506 in our example), and the time instant and fundamental frequency for each point.

- **-nucleus -n**

Time instant of the syllable nuclei for each accent group.

**Default value:** `syl.dat`

**Example:**

```
3 0 1.400 2.010 2.912
2 1 1.491 1.973
```

**Description.** Each line of the file has the time instants of the accent groups of the sentences.

- **-folds -F**

File that contains the fold number for each sentence.

**Default value:** `folds.dat`

- **-stop -s**

Stop condition relative to the RMSE reduction in an iteration.

**Default value:** `0.01`

- **-minsize -i**

Stop condition relative to the minimum number of elements in each leaf. If it is not possible to perform a splitting due to this, the program stops.

**Default value:** `20`

- **-resolution -r**

Number of values that are explored for continuous fields.

**Default value:** `10`

- **-balance -a**

Minimum percentage of elements in each leaf after a splitting. This allows to preserve the balance in the tree.

**Default value: 10**

- **-nodes -o**

Number of coefficients for Bézier intonation model.

**Default value: 5**

- **-model -l**

Intonation model that is used for training:

- Model 0. Superpositional Bezier using JEMA.
- Model 1. Bezier that only uses accent groups.
- Model 2. Fujisaki using JEMA.
- Model 3. Fujisaki.
- Model 4. Tilt using JEMA.
- Model 5. Tilt.
- Model 6. Bezier using JEMA (only accent groups).

**Default value: 0**

## References

- [Agü04a] Pablo Daniel Agüero, and Antonio Bonafonte, “Intonation modeling for TTS using a joint extraction and prediction approach”, *Proceedings of the International Workshop on Speech Synthesis*, 2004.
- [Agü04b] Pablo Daniel Agüero, Klaus Wimmer, and Antonio Bonafonte, “Joint extraction and prediction of Fujisaki’s intonation model parameters”, *Proceedings of International Conference on Spoken Language Processing*, 2004.
- [Cam04] N. Campbell, “Speech & expression: the value of a longitudinal corpus”, *LREC 2004*, 2004.
- [Esc02] D. Escudero, and V. Cardenoso, “Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish”, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pags. 481–484, 2002.
- [Fuj84] H. Fujisaki, and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, *Journal of the Acoustical Society of Japan*, Vol. 5, pags. 233–242, 1984.
- [Fuj00a] H. Fujisaki, S. Ohno, and S. Narusawa, “Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common Japanese and the standard Chinese”, *Proceedings of the 5th Seminar on Speech Production*, pags. 145–148, 2000, bavaria, Germany.
- [Fuj00b] Hiroya Fujisaki, Shuichi Narusawa, and Masako Maruno, “Pre-processing of fundamental frequency contours of speech for automatic parameter extraction”, *Proceedings of the International Conference on Signal Processing*, pags. 722–725, 2000.
- [Gar96] J.M. Garrido, “Modelling spanish intonation for text-to-speech applications”, *PhD Thesis, Universidad Autónoma de Barcelona*, 1996.
- [Lóp93] E. López, “Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto voz en Español basados en concatenación de unidades”, *PhD Thesis, E.T.S. de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid*, 1993.
- [Mix00] H. Mixdorff, “A novel approach to the fully automatic extraction of Fujisaki model parameters”, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pags. 1281–1284, 2000.
- [Nav02] E. Navas, I. Hernaez, and N. Ezeiza, “Assigning phrase breaks using CART’s in Basque TTS”, *Proceedings of the International Conference on Speech Prosody*, pags. 527–531, 2002.
- [Roj05] Matej Rojc, Pablo Daniel Agüero, Antonio Bonafonte, and Zdravko Kacic, “Training the Tilt intonation model using the JEMA methodology”, *Proceedings of Eurospeech*, 2005.
- [Shr98] E. Shriberg, R. Bates, A. Stolleke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, “Can prosody aid the automatic classification of dialog acts in conversational speech?”, *Language and Speech*, , nº 41, pags. 439–487, 1998.
- [Tay00] P. Taylor, “Analysis and synthesis of intonation using the Tilt model”, *Journal of the Acoustical Society of America*, Vol. 107, nº 3, pags. 1697–1714, 2000.
- [Val98] J.A. Vallejo, “Mejora de la frecuencia fundamental en la conversión de texto a voz”, *PhD Thesis, E.T.S.I de Telecomunicaciones, Universidad Politécnica de Madrid*, 1998.