# PARAMETER ESTIMATION AND PREDICTION FROM TEXT FOR A SUPERPOSITIONAL INTONATION MODEL

*Humberto M. Torres, Jorge A. Gurlekian*

*Laboratorio de Investigaciones Sensoriales,*
*Instituto de Neurociencias Aplicadas,*
*Hospital de Clínicas, Buenos Aires, Argentina*
*jag@fmed.uba.ar*

**Abstract:** This work presents an approach for parameter estimation and prediction of the Fujisaki model for Argentine Spanish. Language hypotheses were proposed for estimation and tested by means of genetic algorithms. These hypotheses were validated by comparison of the estimation performance relative to the standard method. Prediction was then calculated based on input text information and performed by CARTs reached a performance comparable to approximations presented for Japanese and English. Objective measures showed that the predicted F0 contours are not close copies of the originals, nevertheless perceptual judgments revealed an impression of good quality for the intonation which makes the method a valuable tool for the development of TTS systems.

## 1 Introduction

The fundamental frequency contour is the key parameter to achieve high quality synthesis in Text-to-Speech (TTS) systems. There are marked differences between intonation contours of different Spanish variants. The Argentine Spanish intonation contour in particular shows the strong influence of various romance languages: e.g. Neapolitan Italian which makes it different from Madrid or Central American Spanish [2]. An intonational model that has been tested for different languages was proposed by [4]. This model -called superpositional- is hierarchical, additive, parametric and continuous in time. It allows the calculation of a reduced parameter set that represents real intonation contours in a compact and automatic way. We choose this model for our proposal based on the validity of the model for Argentine Spanish [5]. This model analytically describes the F0 contour in a log scale, as the superposition of three components: a base frequency ($F_{min}$), tonal accents and phrase accents. Phrase accents are calculated as the response to a second order lineal filter critically excited with a delta function called phrase command. Tonal accents resulted from the response to the same filter, excited with a step function called accent command.

Parameters $\alpha$ and $\beta$ in Fujisaki equation, characterize the dynamic properties of the laryngeal mechanisms of phrase and accent control. Together with $\gamma$ they can be considered practically constant for all speakers. $F_{min}$ must be estimated for each emission. Finally, the parameters to be calculated are the existence or not of phrase commands, amplitude and position values of the phrase accents ($Af$ and $T0$), amplitude and position values of tonal accents ($Aa$, $T1$ and $T2$). The complex formulation of the model requires an automatic parameter extraction approach from F0 measurements. One of the standard methods is analysis-by-synthesis [9]. This requires a complete search by quantified steps, within a reasonable range for each parameter. The iterative process continues until the best value combination fits the measured contour.

Our database consists of 741 declarative sentences extracted from Buenos Aires Argentine newspapers. The sentences contain 97% of all Spanish syllables, in both stress conditions and all possible syllabic positions within the word. A native female speaker read the sentences in a sound proof chamber. Recordings were made with an AKG dynamic microphone and 16 Khz/16bit conversion. The speaker was instructed to read the sentences with natural tonal variations. Each sound file was manually labeled twice. The files were labeled in different layers: phonetic, orthographic, pause levels between words, and tonal marks according to an extended ToBI method for Argentine Spanish [7]. Parts of speech and syntactic layers were also indicated.

The structure of this work consider that the way to go from the input text to the corresponding F0 contour for a sentence requires estimation defined as the process by which the model parameter based on the required F0 contour is calculated, and prediction defined as the process for the model parameter calculation based on the input text. In Section 2 we present the model estimation method, in section 3 the model's prediction procedures. Perceptual tests are evaluated in section 4 and finally, in section 5 we present our conclusions.

## 2 Model estimation

A new method based on Genetic Algorithms (GAs) and linguisitc restrictions for the superpositional F0 model parameter estimation is presented. The idea is to reduce the speaker dependent parameters as much as possible and then estimate the remaining parameters, which are supposed to be associated with the text structure. Those can be fixed in advance or limited in range according to our linguistic hypothesis. Others will depend on upper level information, such as phrase type, intentionality, speaker mood, etc. Since this information is not available in conventional TTS systems, we will suppose that the values are only influenced by the text. In summary, our hypothesis is that model parameters will only depend on the text and that the speaker characteristics will remain invariable.

### 2.1 Statistical analysis of parameter values

From statistical analysis of the results obtained before [7], we can fix $\alpha = 2$, $\beta = 20$ and $\gamma = 0.9$.

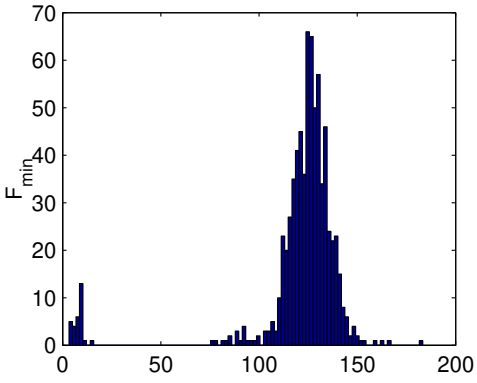


**Figure 1** - Base frequency histogram, estimated by the analysis-by-synthesis method [7].

Base frequency $F_{min}$ is considered approximately fixed for a particular speaker, since it models the vocal folds in static conditions. Fig. 1 shows a histogram of $F_{min}$ values obtained before [7], and derived from that data we will fix $F_{min}$ = 130 Hz. This value was validated empirically for a data base sentence set.

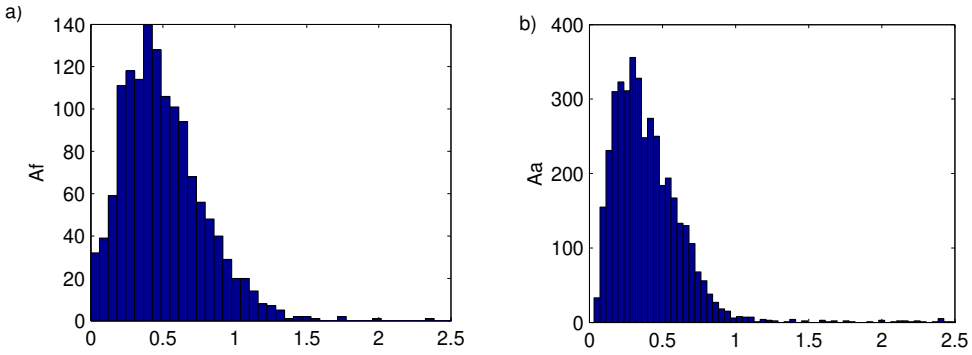


**Figure 2** - Model parameters histogram, estimated by the analysis-by-synthesis method [7]. a) Phrase commands amplitude $Af$, and b) accent commands amplitude $Aa$.

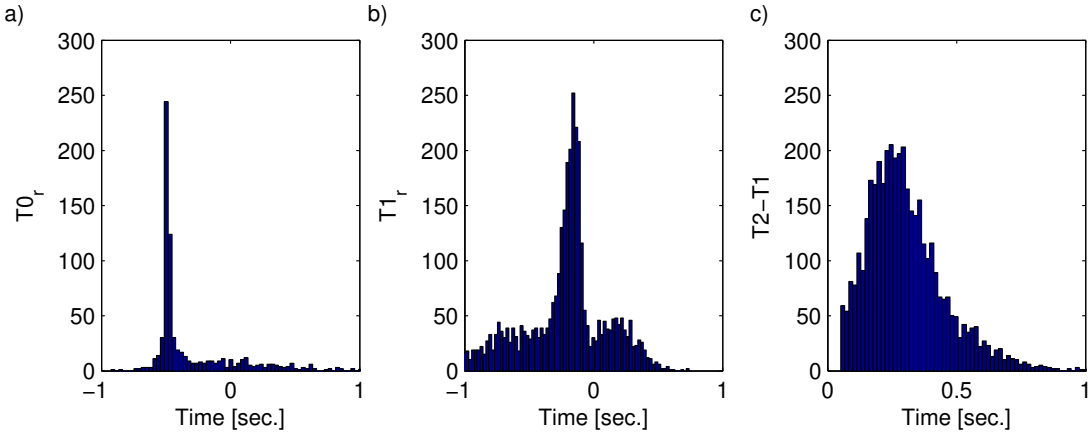


**Figure 3** - Model parameters histogram, estimated by the analysis-by-synthesis method [7]. a) Distance between phrase commands location and the beginning of the intonational phrase, b) distance between accent commands location and next stressed vowel midpoint, and c) accent command durations.

Fig. 2 shows phrase and accent command amplitude histograms. As we can see, their values are clearly limited to a small range, with an average close to 0.5. As the base frequency is fixed, we can suppose that $Af$, and in a lesser extent $Aa$, could be slightly greater, in order to compensate for cases where the value of $F_{min}$ is lower than the desirable one. Thus, we can suppose that the range of possible command amplitude values could extend beyond value 1 (see Fig. 2). Amplitude range was set at [0; 2.55]. These range of values were also validated empirically for the data base sentences set.

In order to define phrase commands completely, possible $T0$ need to be defined. First, we can assume that only one phrase command exists by intonative phrase, and that in addition its location is close to the beginning of the phrase. In Fig. 4.a) the distance histogram between an intonative group beginning and a phrase command location - denoted as $T0_r$- is shown.

From Fig. 3.a) we can assume that the value of $T0_r$ is, in most of the cases, approximately -0.5 sec. As mentioned before, the effect of the phrase command is a global F0 movement, which is added to $F_{min}$, that will remain constant. Thus, the histogram in the Fig. 3.a) loses force, since it was calculated when $F_{min}$ was variable. Preliminary experiments showed that if we let $T0_r$ vary in [-0.512; 0] sec., excellent results are obtained.

Accents commands will produce F0 local movements that are associated with tonal accents. For most languages included Spanish, tonal accents only occur at the stressed syllable of content

words. Then it is possible in turn to associate accent commands with the content word stressed syllable.

Even though not all stressed syllables produce tonal accents, we will initially assume that all stressed syllables in content words have an associated accent command. Accent location ($T1$), is measured relative to a reference time located at the center of the closest stressed vowel, and it is indicated as $T1_r$. In Fig. 3.b) the $T1_r$ histogram is shown. In average, $T1_r$ is approximately -0.1 sec. The figure confirms our hypothesis that accent command location and stressed syllables location can be associated. In this work $T1_r$ is allowed to vary between $\pm$ 0.512 sec. In order to complete the model, accent command duration ($T2-T1$) must be set. In Fig. 3.c) the $T2-T1$ histogram can be observed, and from this, the range of its possible values is limited to [0; 0.512].

In Table 1 we have summarized the possible model parameter values fixed in this paper.

**Table 1** - Values and ranges of possible values of model parameters. Time values are expressed in seconds. $T0_r$ is measured relative to intonative phrase beginning, and $T1_r$ is measured relative vowel center.

| $F_{min}$ | $\alpha$ | $\beta$ | $\gamma$ | $Af$ | $Aa$ | $T0_r$ | $T1_r$ | $T2-T1$ |
|---|---|---|---|---|---|---|---|---|
| 130 | 2 | 20 | 0.9 | [0 ; 2.55] | [0 ; 2.55] | [-0.512 ; 0] | [-0.511 ; 0.512] | [0 ; 0.512] |

## 2.2 Genetic algorithms

Genetic Algorithms are a procedure set for optimization and problem search resolution. They are based on biological evolution precepts: population based selection, sexual reproduction and mutation [6]. Through selection we establish which member of a population will survive to reproduce and be parents in the next generation; and through reproduction we get the mixture and recombination of the descendants. This gene mixture allows the species to evolve more quickly than they would if they only had the copy of the genetic material of one of their parents. The principles of GAs are based in scheme theory [6], in which the fundamental theorem of the GA ensures its convergence.

## 2.3 Experiments and Results

When implementing a GA, we have to define a set of parameters: codification type, selection method, fitness function, cross and mutation methods, cross and mutation rates, number of individuals, and stop condition. In this work we used: binary codification, coding the variables corresponding to amplitudes ($A_f$ and $A_a$) with a resolution of 0.01, and times ($T0$, $T1$ and $T2$) with a resolution of 0.001; roulette method, with elitism for the selection; the evaluation was made with fitness function define in the Eq. 1:

$$f_i(n) = \frac{1}{1+MSE_i(n)} \tag{1}$$

where $f_i(n)$ is the individual's fitness $i$ in the $n$ generation; the reproduction was made with two points cross, with a probability of 0.5, and a uniform mutation with a probability of 0.1; the number of individuals was fixed to 80; the maximum fitness of 0.1 or 50000 generations, as stop criterion.

The number of individuals, as well as cross and mutation rates, were fixed in an empirical way after a series of tests with a corpus sentences set. With this set up, five runs were made over each of the 741 sentences. Results are shown in Table 2. Root Mean Square Error (RMSE)

and average Correlation Coefficient $R^2$ values are similar to those obtained with the analysis by synthesis method [7, 9] where RMSE was 16 HZ and $R^2$ was 0.93 employing the same data base.

In Fig. 4 an example of F0 modeling with the superpositional model is shown, where the parameters were estimated with GAs. In Fig. 4.a) the real (dotted line) and estimated (solid line) F0 contours are shown; and in Fig. 4.b) the phrase commands ($Af$) and the accent commands ($Aa$) are shown. The dotted lines correspond to SAMPA labeling. Estimated F0 values follow well the movement of real values. For this sentence, an RMSE of 32.7537 Hz. and a $R^2$ of 0.8985 were obtained. In Fig. 5 and Fig. 6 histograms of model parameters obtained with GAs are presented. Amplitude parameter shapes a) and b) are similar to those presented before. Main diferences are noted in time parameters.

**Table 2** - RMSE and $R^2$ obtained for 741 sentences. The values are expressed in Hz.

| | **Mean** | **Standard deviation** | **Minimum value** | **Maximum value** |
|---|---|---|---|---|
| RMSE | 16.27 | 6.94 | 7.34 | 53.01 |
| $R^2$ | 0.91 | 0.08 | 0.033 | 0.99 |

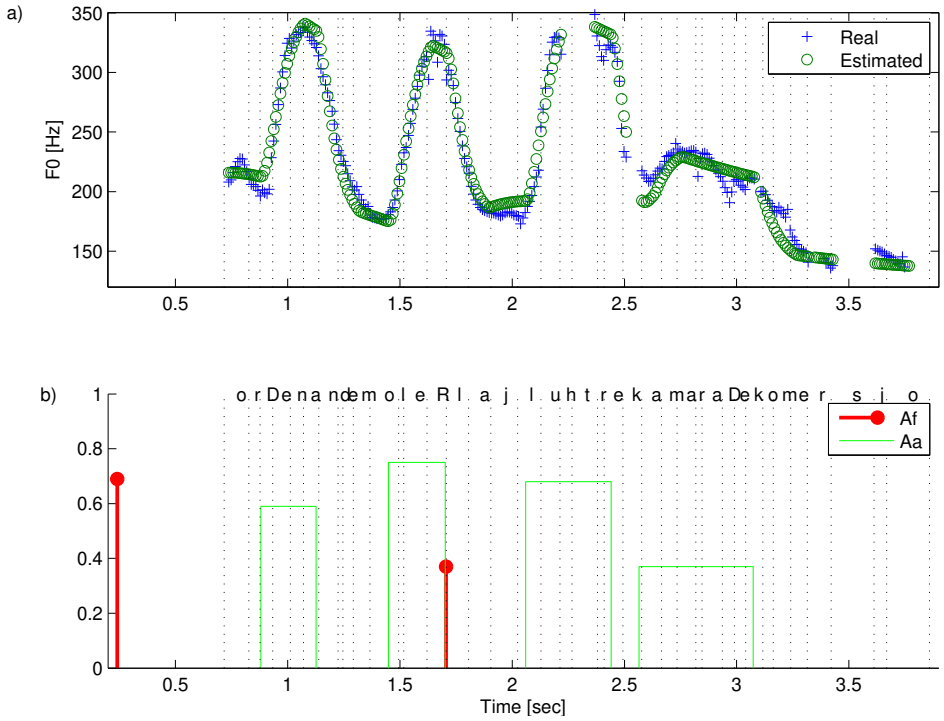


**Figure 4** - Result of the superpositional model for the phrase *Ordenan demoler la ilustre cámara de comercio* (They ordered to demolish the distinguished chamber of commerce), where the model parameters were estimated by a GAs. a) Real F0 (dotted line) and estimated (solid line) contours, and b) Phrase commands ($Af$) and accent commands ($Aa$). The dotted lines correspond to SAMPA labeling.

## 2.4 Discussion

We have presented a novel method to estimate model parameters from speech waveforms and their corresponding texts. It is based on a set of empirical rules that allows the association of

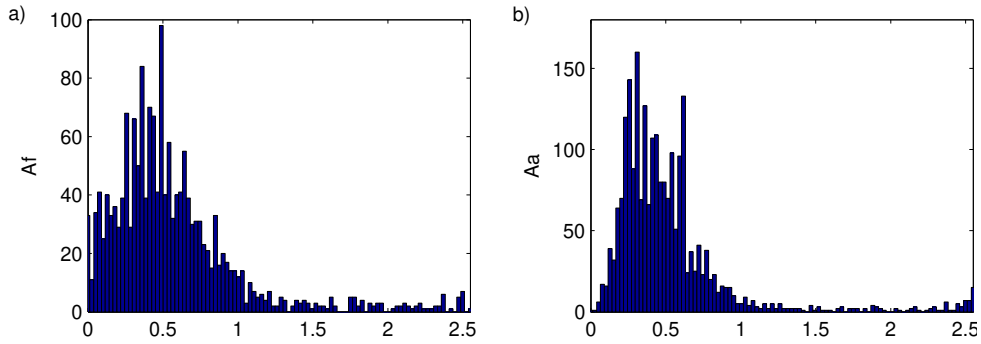


**Figure 5** - Amplitude command histograms estimated by GAs and linguistic restrictions. a) Phrase command amplitude $Af$, and b) accent command amplitude $Aa$.

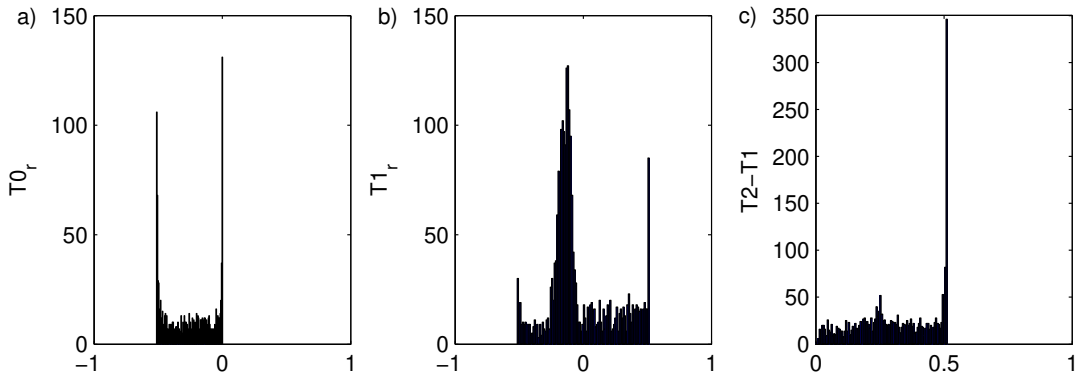


**Figure 6** - Location and duration command histograms estimated by GAs and linguistic restrictions. a) Distance between the phrase command location and the beginning of the intonational phrase, b) distance between the accent command location and the midpoint of the nearest stressed vowel, and c) accent command duration.

syntactic structures to model components. Model parameters are delimited prior to their value estimation by GAs. Results show that the method performance is similar to other algorithms previously used and principally it helps to verify existing relations between syntactic structure and model parameter values. The main disadvantage is a high computational cost, but it occur offline in a complete TTS system.

In the following section, estimation results are used as a base to propose a parameter model prediction method from text.

## 3 Model prediction from the text

Classification trees are one of the more used non-parametric supervised inductive learning methods in the areas of natural language processing and speech processing. A classification tree is a way of representing the knowledge obtained in an inductive learning process. The algorithm proposed by [1], well-known as Classification and Regression Tree (CART) will be implemented. In this section we will focus on describing the experiments performed to predict the superpositional model parameter values using CARTs. Different input types were used to feed the trees: location and length of the intonative phrase; identity, location and length of the stressed vowel; identity of the context phonemes, in a five window length; Part-Of-Speech (POS), in a context of two words before and two later; distance to the previous and following

stressed vowel that has an accent command associated; parameters values of the previous accent and/or phrase command.

For inputs corresponding to location, distance and length, different measurement units were used: time and number of phonemes, syllables, words and intonative phrases. These measurements were introduced as absolute and relative values. A tree was used for each of the parameters to be estimated: phrase command location, relative to the beginning of intonative phrase; phrase command amplitude; accent command location relative to stressed vowels of content words, except the last one which does not have any associated command; accent command amplitude; accent command duration.

In order to avoid the over-training effect crossed validation was employed in the following way: the total of the data was divided at random in a training set (80% of the available data), used to estimate the CART, and a test set to estimate the performance (20% of the available data). The percentage designated for training is a critical factor, because CART's performance is very sensitive to the quality and quantity of the data used to estimate the model. If the percentage to test is low, a poor generalization will be obtained, and if it is high there will not be enough data to train the models. For example, going from a 70% to 90% for training, the classification performance improves between 4 and 5%. The experiments were made over five possible partitions of the available data set and the results were averaged.

### 3.1 Results

For different data sets, a series of experiments were performed to find CART parameters that minimize the classification error. During the experiments, we tried to minimize the classification error of the test data by changing the depth of the tree, but at the same time avoiding that the error was smaller than that obtained with the training data. In all cases, this depth was three. By increasing the depth, errors diminished for the training set, but increased for the test set.

In Table 3 the results obtained for the five possible test sets are shown. During CART's construction, the method automatically selects the inputs to be used.

**Table 3** - RMSE percentages obtained by predicting the superpositional model parameter values using CARTs for five different test set.

| **Set #** | $Af$ | $Aa$ | $T0_r$ | $T1_r$ | $T2-T1$ |
|---|---|---|---|---|---|
| I | 7 | 17 | 16 | 24 | 29 |
| II | 8 | 20 | 16 | 25 | 27 |
| III | 9 | 16 | 15 | 25 | 29 |
| IV | 9 | 18 | 15 | 25 | 29 |
| V | 8 | 19 | 16 | 26 | 27 |

During CART's construction, the method automatically selects the inputs to be used. The different features that were chosen as inputs for each of the predicted parameters are listed below:

- $Af$: previous and next phoneme identity; end of phrase distance, in seconds; previous accent command duration, in seconds.
- $T0_r$: stressed vowel location of the first content word, in seconds. phrase length, in seconds. previous accent command duration, in seconds. previous phrase command amplitude

- *Aa*: previous and next phoneme identity; next phoneme sound; content word POS and previous, next and next of the next word; number of intonative phrases in the sentence; phrase length, in seconds.
- $T1_r$: previous and next phoneme identity; distance to intonation phrase end, in seconds.
- $T2-T1$: next phoneme identity; phrase length, in seconds; distance to intonation phrase beginning, in seconds; distance to intonation phrase end, in seconds; previous accent command duration, in seconds.

In Table 4, error results of objective evaluation in different scales can be observed. RMSE is measured as the difference between the original and the predicted F0, considering only voicing segments without interpolation of unvoiced segments. MSE values were also calculated with F0 in the logarithmic domain to make it comparable with [10] results.

**Table 4** - Errors in different F0 scales, obtained by CARTs, for the different test sets. ST: Semitones; Ln: natural logarithm scale.

| Set # | RMSE Hz | RMSE ST | RMSE Ln | RMSE ERBs | MSE Ln |
|---|---|---|---|---|---|
| I | 50 | 3.7 | 0.21 | 0.89 | 0.078 |
| II | 55 | 3.8 | 0.22 | 0.93 | 0.085 |
| III | 45 | 3.6 | 0.21 | 0.84 | 0.070 |
| IV | 52 | 3.9 | 0.22 | 0.92 | 0.083 |
| V | 51 | 3.7 | 0.21 | 0.89 | 0.076 |
| Average | 51 | 3.7 | 0.21 | 0.90 | 0.078 |

In Fig. 7 we present the original and the predicted F0. For this example, the predicted F0 follows the tonal movements present in the original, but there is a little delay and the amplitudes do not coincide exactly.

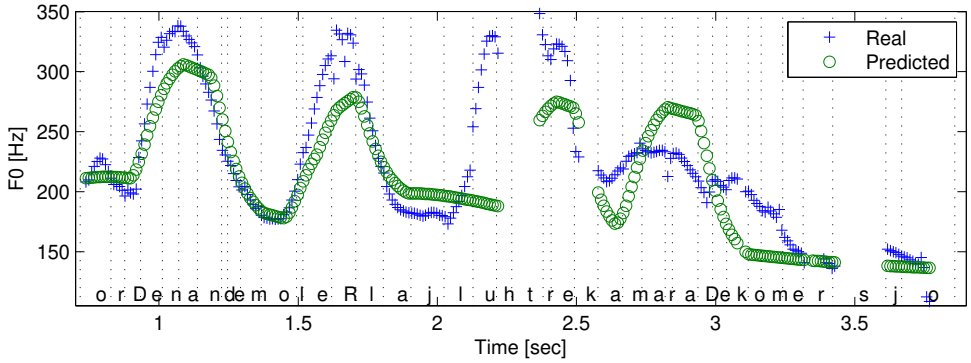


**Figure 7** - Original F0 (+ line) and predicted (o line) for the phrase *Ordenan demoler la ilustre cámara de comercio* (They ordered to demolish the distinguished chamber of commerce).The dotted lines correspond to SAMPA labeling.

### 3.1.1 Discussion

Objective measures showed a higher error than those obtained by estimation. Besides, it exceeds two and a half times the perceptible error that is approximately 1.5 semitones. Nevertheless, the

error values obtained are similar to the results obtained for Japanese [10]. In Table 3, parameter values related to the phrase commands show less RMSE in the prediction. In general, the errors are due to an underestimation of the values, principally in the case of the phrase command location. Besides, the system only caters for four possible values. This last point reflects one of the greatest weaknesses of CARTs: when its output is a numerical value belonging to the real ones, it does not have the capacity to interpolate and/or approximate values not seen in the construction stage. At the moment, the system is less efficient at predicting the duration values of the accent command.

## 4 Perceptual test

A perceptual test of the F0 contours predicted by CARTs was prepared. Twenty sentences were re-synthesized from the test set using the PSOLA method of the program Praat. Mean Opinion Score (MOS) and Degradation MOS (DMOS) scales [11] were used. The test was presented to ten listeners, who perceived natural and modified sentences through a headphone [3] in an acoustic chamber. They were instructed to evaluate the intonation quality of the re-synthesized sentences according to the proposed scales. On average, MOS gave a result of 3.75, and DMOS of 3.88, calculated over all the listeners and all the test sentences. These results correspond to a good intonation quality, with a certain degree of distortion that did not bother. Present results are similar to others as reported in [10] for Japanese and in [8] for Spanish.

In Fig. 8.a) boxplots are presented for MOS and in Fig. 8.b) for DMOS. These figures allow us to observe that in general the tests do not have a great variance, and that sentences #18 and #19 present the worst results in the test. On average, these two sentences were evaluated with a regular quality intonation, and present tonal accents shifted one syllable to the right. This maybe due to the proximity of a boundary intonation phrase without silence.

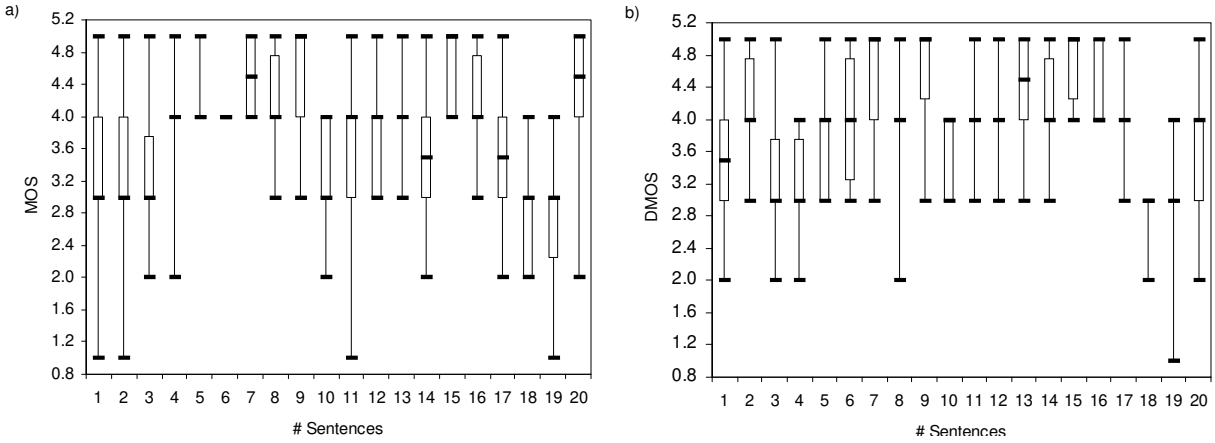


**Figure 8** - a) Box diagram of MOS and b) DMOS for the model of the predicted F0 using CARTs, for the 20 test sentences.

## 5 Conclusions

In this paper a hypothesis that correlates the intonation model parameters with the input text of a TTS system was outlined. The model parameter values are fixed or range limited. In order to confirm the validity of the hypothesis, we made a model estimation using GAs that achieved a similar performance to the obtained with a standard estimation method.

CARTs achieved a good performance, comparable to other approximations used for different variations of Spanish. Although the error is a little bit high, the perceptual tests qualified the obtained intonation as good, which encourages the use of the proposed method as part of a text-to-speech conversion system.

## 6 Acknowledgements

This research has been carried out with the support of the Ministerio de Ciencia y Tecnología and the Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

## References

[1] BREIMAN, L., J. FRIEDMAN and C. STONE: *Clasification and regresison tree*. Chapman & Hall, New York, 1984.

[2] COLANTONI, L. and J. GURLEKIAN: *Convergence and intonation: historical evidence from Buenos Aires Spanish*. Bilingualism: Language and Cognition, 7(2):107–118, August 2004.

[3] DIMOLITAS, S., F. CORCORAN and C. RAVISHANKAR: *Dependence of opinion scores on listening sets used in degradation category rating assesments*. IEEE Transactions on Speech and Audio Porocessing, 3(5):421–424, 1995.

[4] FUJISAKI, H. and K. HIROSE: *Analysis of voice fundamental frequency contours for declarative sentences of Japanese*. Journal of Acoustic Society, 5(4):233–242, 1984.

[5] FUJISAKI, H., S. OHONO, K. ICHI NAKAMURA, M. GUIRAO and J. GURLEKIAN: *Analysis of accent and intonation in Spanish based on a quantitative Model*. In *Proc. of ICSLP 94*, pp. 355–358, Yokohama, September 1994.

[6] GOLDBERG, D.: *Genetic Algorithms in search, optimization and Machine Learning*. Addison-Wesley, 1989.

[7] GURLEKIAN, J. A., H. M. TORRES and L. COLANTONI: *Evaluación de las descripciones analítica y perceptual de la entonación de una base de datos de oraciones declarativas de foco amplio para el español hablado en Buenos Aires*. Estudios de Fonética Experimental, XIII:275–302, 2004.

[8] GUTIERREZ-ARRIOLA, J., J. MONTERO, D. SAIZ and J. PARDO: *New Rule-Based and Data-Driven Strategy to Incorporate Fujisaki's F0 Model to a Text to Speech System in Castillian Spanish*. In *Proc. of ICASSP 2001*, vol. 2, pp. 821–824, Salt Lake City, UT, USA, May 2001.

[9] MIXDORFF, H.: *A novel approach to the fully automatic extraction of Fujisaki model parameters*. In *Proc. of ICASSP 2000*, vol. 3, pp. 1281–1284, Istanbul, June 2000.

[10] SAKURAI, A., H. K and N. MINEMATSU: *Data-driven generation of F0 contours using a superpositonal model*. Speech Comunication, 40(4):535–549, June 2003.

[11] THORPE, L. and B. SHELTON: *Subjetive Test Methodology: MOS versus DMOS in evaluation of Speech Coding algoritms*. In *Proc. of IEEE Workshop Speech Coding Telecomun.*, pp. 73–74, Sainte-Adèle, Canada, October 1993.