NEW RULE-BASED AND DATA-DRIVEN STRATEGY TO INCORPORATE FUJISAKI'S F0 MODEL TO A TEXT-TO-SPEECH SYSTEM IN CASTILLIAN SPANISH

J.M. Gutiérrez-Arriola, J.M. Montero, D. Saiz, J.M. Pardo*

Grupo de Tecnología del Habla, Dpto. de Ingeniería Electrónica, ETSIT, Universidad Politécnica de Madrid.

*Also in Dpto. de Ingeniería de Circuitos y Sistemas, EUITT, Universidad Politécnica de Madrid.

Contact: juana@die.upm.es, http://www-gth.die.upm.es

ABSTRACT

We will present the analysis of a Spanish prosody database by estimating the parameters of Fujisaki's model for F0 contours. These parameters are classified attending to linguistic features and they form the analysis database. When synthesizing F0 contours we extract the linguistic features from the text and perform a k-Nearest Neighbour search. Linguistic feature comparison distance is trained using data from the prosody database. To avoid artifacts we perform a rule-base filtering on synthesis parameters.

The results of our evaluation test show that the proposed system is significantly better than the previous neural network approach. This evaluation confirms the ability of Fujisaki's model to represent prosody information based on linguistic features.

1.INTRODUCTION

The first research to generate F0 contours in the Speech Technology Group was aimed at a rule-based model [1]. It was judged to be monotonous and rather unnatural. Later research led us to use a neural network to model the F0 contour [2]. We recorded a prosody database and used it to train a three-layered Perceptron (one hidden layer and 55 binary coded input parameters) with one output corresponding to the syllable pitch.

These two models (rule-based and NN-based) were compared in a formal text and the neural network prosody was judged as significantly better [2]. Nevertheless questions and exclamations were not modeled very well due to the lack of examples to train the neural network.

The problem with the neural network is that there is no direct relationship between linguistic features and generated F0. This fact makes this model of little use when trying to transform styles or voices.

Fujisaki's model has been successfully applied as a F0 contour model in several languages, also to Argentinean Spanish [4]. Recently new data-driven strategies are being investigated to compute command parameters from text [6].

In this paper we introduce a new kNN-based and rule-based strategy for synthesizing Fujisaki's parameters.

2.BRIEF DESCRIPTION OF THE MODEL

Fujisaki's model expresses the intonation as the sum of three prosodic levels: minimum F0, phrase level and word level [3], [5].

The minimum F0 is defined as a constant. The phrase level is defined by three parameters: T0 is the place where the component starts, Af is the amplitude of the phrase command and α is the damping factor of the phrase control system (figure 1a and 1b).

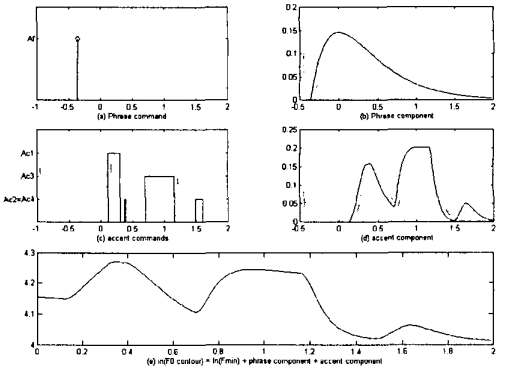

Figure 1. Superpositional representation of Fujisaki's model

The accent level defines one accent command for each stressed vowel in the phrase. The accent component is defined by the beginning and end points of the accent command (T1 and T2), its amplitude (Ac) and β, the damping factor of the accent control system (see figure 1c, 1d).

The linguistic interpretation of this superpositional model is that the phrase component establishes the baseline of the F0 contour, while accent components account for local variations [5].

0-7803-7041-4/01/$10.00 ©2001 IEEE

3.ANALISYS OF THE DATABASE

A male speaker read several texts containing a variety of styles such as formal speech, interviews, dialogues, short exclamative sentences. The database consists of about 750 intonational phrases between pauses. The recordings are made in five sessions, in a silent environment with a sampling frequency of 16KHz and 16 bits per sample.

Each phrase is marked with its final punctuation mark or with the reason for the speaker to make the pause, attending to one of the following categories: exclamation mark, open bracket, close bracket, comma, full stop or semicolon, colon, question mark, preposition, coordinate clause, attributive verb, subordinate clause, verb, adverb and quotation marks or dash or dots.

Pitch is automatically extracted and manually revised for each phrase. To smooth the F0 contour we apply a media filter to the pitch curve, before trying to extract Fujisaki's parameters from the training database.

According to [5] and to our preliminary experiments, we impose some restrictions to the analysis, for an easier parameter calculation:

- There is only one phrase command per intonational phrase.
- The phrase component cannot exceed the F0 contour at any point of the curve.
- We set $T0=-1/\alpha$ to get the maximum of the phrase component at the beginning of the phrase, as can be observed in figure 1b. This restriction makes impossible for the search algorithm to try to model the first stressed vowel by means of the phrase command.
- We consider an accent command for each stressed vowel in the phrase.
- Accent commands cannot overlap.

For each parameter we define a search space based in a quantization step as shown in table I. The approximation tries to minimize the mean square error in the logF0 domain. The analysis algorithm is as follows:

1. Optimization of Fmin, α and Af.
2. The new objective is calculated as F0 contour minus the optimized phrase component.
3. From left to right we assign an accent command to each stressed vowel in the intonation phrase and we optimize the F0 contour around this vowel.

Figure 2 shows the fitting procedure for a phrase in Castillian Spanish.

After automatic extraction, a manual revision is done to avoid bad analyzed data.

The main errors are due to the lack of pitch marks and to misalignments between the expected intonation and what the speaker said. In some cases he does a rising F0 contour in absence of stress, and sometimes there is no stressed F0 pattern around the stressed vowel.

Parameter	Min	Max	Step
Fmin	30	70	1 Hz
α	-3	3	0.1
β	8	32	4
Af	0	2	0.1
Aa	0	2	0.1
T1j	Max(T2j-1, Tini-λDur)	Tini+λDur	0.015 sg
T2j	T1j	Tfin+λDur	0.015 sg

Table I. Tini and Tfin are the beginning and end point of the stressed vowel; Dur is the vowel duration.

We find another special case when addressing yes-no questions. In Spanish this intonation is characterized by a final rising contour. This possibility is not considered in the approximation procedure. In the revision phase an accent command is added to model this final effect. We have made experiments for modeling this final rise with a phrase [5] or accent command. The use of an accent command results in better modeling of the final rise.

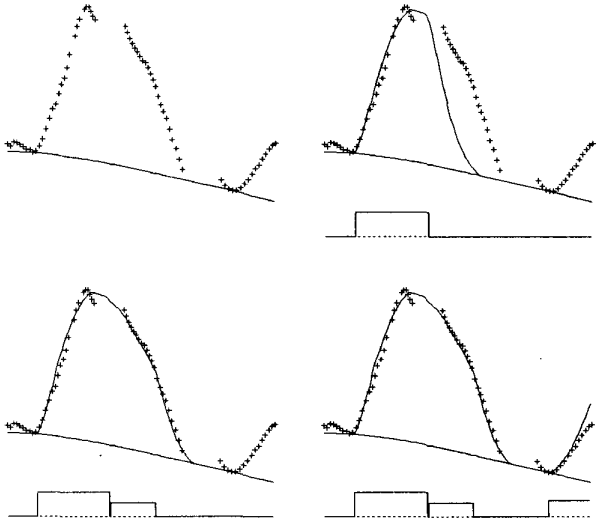

Figure 2. Automatic extraction of model parameters for the phrase 'Al leer la palabra plan'.

4.SYNTHESIS OF F0 CONTOURS

To include the information obtained in the analysis phase in a text-to-speech system, we build a database with one entry per accent command; each entry has ten input linguistic parameters and seven output Fujisaki's parameters. The inputs are:

1. Type of pause at the end of the intonational phrase.
2. Total number of stressed vowels in the prosodic unit.
3. Duration of the stressed vowel.
4. Position of the stressed vowel relative to the number of stressed vowels (1 for the first stressed vowel and N for the N-Th one).

5. Position of the stressed syllable in the phrase, relative to the total number of syllables.
6. Distance to previous stressed syllable.
7. Distance to next stressed syllable.
8. ¿Is it the last stressed vowel?
9. Type of question: yes-no question, wh-question or or-question.
10. Type of sentence: declarative, question or exclamation.

And the outputs are:

1. Fmin.
2. Ac, T1, T2 and β.
3. Af and α corresponding to the phrase command that includes the accent command.

In the F0 generation module, the ten input parameters of the database are obtained directly from the text to synthesize. With these parameters we search the database to find the nearest neighbour entry and we select the corresponding accent command.

When finding the phrase command, we reduce the search space to the phrase commands associated to the selected accent commands. The criteria is to select the phrase command with an average F0 that is the closest to the previously selected phrase command. We adopt this decision after some experiments trying to obtain the phrase command only from linguistic parameters. Due to the variety of intonation patterns in the database, the average F0 of two independently selected F0 contours could be quite different, making the utterance to sound unnatural.

5.WEIGHTED DISTANCE OPTIMIZATION

As mentioned above we compare the linguistic parameters of synthesis text to the entries of the analysis table and find the nearest one. To do this we apply a weighted distance as follows:

$$\delta_{jmin} = min(\delta_{j1}, \delta_{j2}, ..., \delta_{jL-1}, \delta_{jL})$$

$$\delta_{jl} = \sum_{n=1}^{N} P_n |C_{nj} - K_{nl}|$$

Where: L is the number of entries in the table.
N is the number of parameters to consider.
C_{nj} is the n-th parameter extracted from the text.
K_{nl} is the n-th parameter in the table entry.
P_n is the weight for the n-th parameter.

For the type of sentence we use a binary distance in the search; for example, if the phrase is part of a question, we only examine entries marked as taken from questions (the same applies to declaratives and exclamations). This is the reason for not including the type of sentence in the distance above.

The weights Pn are obtained automatically from the regeneration of eight F0 contours from the database. The range of the weights was between 1 and 3. We tried all the possible combinations of the nine weights to select the parameters for the eight phrases.

The more important parameters are the total number of stressed vowels, the relative position of the stressed vowel, whether the vowel is the last stressed one or not, and the type of question.

6.RULE-BASED POST-PROCESSING OF SELECTED PARAMETERS

After listening to the first synthesized results we discover some problems with the system.

We cannot synthesized an F0 under 60Hz or over 140Hz because the voice is severely distorted by the prosody modification algorithm. To avoid these problems we adopt two solutions: if the F0 contour is less than 60Hz at any point of the prosodic unit, we raise the value of Fmin and, if the F0 contour is greater than 140Hz, we reduce Af.

When the pause is produced by a full stop or the sentence is a wh-question, we find that the final declination is not emphasized enough. So at the end of those prosodic units we add a negative phrase command [5].

When we add phrase and accent components proceeding from different phrases, we can find some unnatural rising or falling edges. To avoid this, we limit the relationship between the maximum of the accent component and the value of the phrase component at the same point.

When synthesizing an exclamation we increase the amplitude of the accent commands to make the exclamation more emphatic. This increase is empirically set to 0.3.

We also eliminate the phrases with negative α from the database, because they give bad results when synthesizing.

7.EVALUATION

Two evaluations are carried out: first we re-synthesize the texts from the database and obtaine the number of right elections of the commands and the error between the original and synthesized F0 contour. We also make a subjective evaluation to compare the F0 contours synthesized by Fujisaki's model to the contours predicted by the neural network.

We can not test all the type of pauses, because the synthesizer we used do not have a pause generator, so we only re-synthesize the prosodic units related to punctuation marks.

The results of the evaluation are shown in table II and figures 3 and 4.

Sentence type	Number of accent commands	Right	Wrong	Percentage
Declarative	459	375	84	81.7%
Exclamation	133	121	12	91.0%
y-n question	217	199	18	91.7%
wh-question	44	42	2	95.5%
or-question	5	2	3	40.0%
Total	858	739	119	86.1%

Table II. Objective evaluation results

The errors are basically due to the shortest prosodic units, because of the similarity of their parameters (the search algorithm will assign the first entry with the minimum error).

In figure 3 we can see the distribution of the error when we do not post-process the F0 contour. As it was expected the error is smaller when we do not make any correction. This is explained because the modifications are not imposed by the database but by the synthesis process, so although it may sound better, it also differs more from the original data. Figure 4 shows the distribution when applying the modifications.

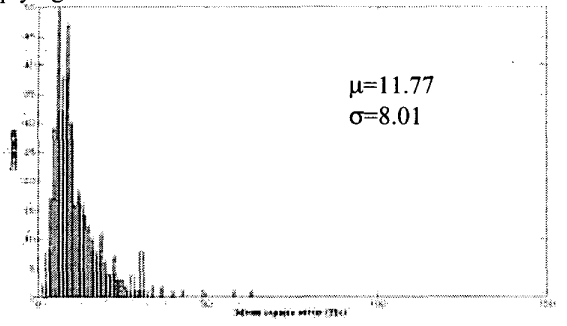

Figure 3. Square error distribution without post-processing.

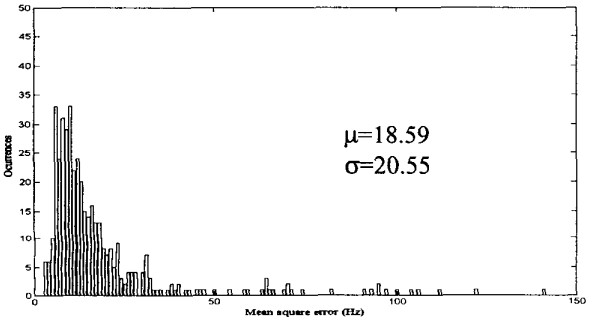

Figure 4. Square error distribution with post-processing.

In addition to this objective test, we make a subjective test as proposed in [2]. 50 listeners are asked to listen to 30 sentences and to judge its intonation in a range from 1 (=very bad) to 5 (=very good). The 30 sentences are 2 repetitions of the same 15 sentences, one synthesize with the neural network prosody and the other one using Fujisaki's model. The utterances are presented randomly.

The results are shown in figure 5. The significance of the result is verified with a T-Student test for related small experiments. For a 95% confidence level we calculate a t=5.29, which proves that Fujisaki's model is significantly better.

It is also interesting to compare the results of the subjective test by types of sentence. The improvement is judged to be greater in exclamations and questions, which were the weak point of the neural network as we pointed out previously. Results of this test are shown in table III.

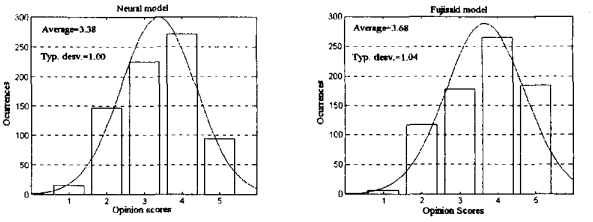

Figure 5. Results of the subjective test.

Type	Neural network	Fujisaki
Modal	3.58	3.49
Question	3.10	3.80
Exclamation	3.37	3.89
All	3.38	3.68

Table III. Results of the subjective test by type of sentence.

8.CONCLUSIONS AND FUTURE RESEARCH

We have applied Fujisaki's F0 model to a text to speech system in Spanish significantly improving the results of the previous NN-based system.

Fujisaki's model has proved to be good at relating linguistic to prosodic parameters in Spanish. One important advantage of using this model is the possibility of adapting the parameters when trying to synthesize different speaking styles or even different speakers.

We should improve the analysis system to make it fully automatic, and we should try other data-driven strategies such as applying a neural network approach to generate Fujisaki's parameters (not to generate F0 values).

7.REFERENCES

[1] *"Improving naturalness in a text to speech system with a new fundamental frequency algorithm"*, P.J. Moreno et al. Eurospeech 89, pp 360-363, 1989.

[2] *"Mejora de la frecuencia fundamental en la conversión de texto a voz"*, Jose Angel Vallejo, PhD Thesis, ETSI Telecomunicación, UPM, 1998.

[3] *"Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations"*, H. Fujisaki. STL-QPSR, Jan. 1981.

[4] *"Analysis of accent and intonation in Spanish based on a quantitative model"*, H Fujisaki, S. Ohno, K. Nakamura, M. Guirao and J. Gurlekian. Proc. ICSLP94, 1, pp. 355-358, 1994.

[5] *"Analysis and synthesis of F_0 contours by means of Fujisaki's model"*, B. Möbius, M. Pätzold and W.Hess., Speech Communication 13, pp. 53-61, 1993.

[6] *"Data driven intonation modeling using a neural network and a command response model"* Atsuhiro Sakurai, Nobuaki Minematsu, Keikichi Hirose. ICSLP 2000.