Evaluation of a System for F_0 Contour Prediction for European Portuguese

João Paulo Teixeira, Diamantino Freitas** and Hiroya Fujisaki****

{*Polytechnic Institute of Bragança, **Faculty of Engineering of University of Porto}, Portugal
***University of Tokyo, Japan
joaopt@ipb.pt, dfreitas@fe.up.pt, fujisaki@alum.mit.edu

Abstract

This paper presents the evaluation of a system for speech F_0 contour prediction for European Portuguese using the Fujisaki model. It is composed of two command-generating sub-systems, the phrase command sub-system and the accent command sub-system. The parameters for evaluating the ability of each sub-system are described. A comparison is made between original and predicted F_0 contours. Finally, the results of a perceptual test are discussed.

1. Introduction

This paper reports the evaluation of the work related to a part of the prosody model for TTS that was developed for the European Portuguese (EP) language [1].

F_0 is the most perceptually relevant component in prosody. The Fujisaki model for F_0 contour has been reported as giving very high naturalness for several languages [2] including Japanese, Korean, Spanish, Polish, Greek, Swedish, English [3], German [4], Basque [5] and now also Portuguese [1]. The Fujisaki approach [2] consists of logarithmic addition of baseline fundamental frequency, F_b, phrase components and accent components. The baseline fundamental frequency is assumed to be constant within each utterance. Phrase commands (PCs) are represented as a set of impulses and the Accent Commands (ACs) as a set of pedestal functions. The corresponding components are represented as responses of the respective control mechanisms to these commands.

Two processes are involved in the Command Generating System (CGS). The first one, as depicted in Fig.1, concerns with the development of the system. The second one, depicted in Fig. 2, is the production of the synthetic F_0 contour, using the CGS.

The first process is a bottom-up process because it departs from the bottom level (the F_0 contour) and expands to a higher level (the estimated commands) and is named as the estimation process. The accent and phrase commands are estimated regarding the Fujisaki model, the F_0 contour and eventually the syntax and discourse structures (dotted arrow in Fig. 1). The estimated commands and the output of the text analysis are the information used in the interpretation process. In this system, Artificial Neural Networks (ANN) were the essential tools to interpret the process and basically consist in the CGS. More details are presented in the next section.

The second process is a top-down process because it departs from the higher level information (the text), and goes down to the lower level (the F_0 contour). It is named as the prediction process. In this process the developed system uses the syntax and discourse structure to predict the accent and phrase commands. Finally, the predicted commands are used with the Fujisaki equations to produce the synthetic F_0 contour.

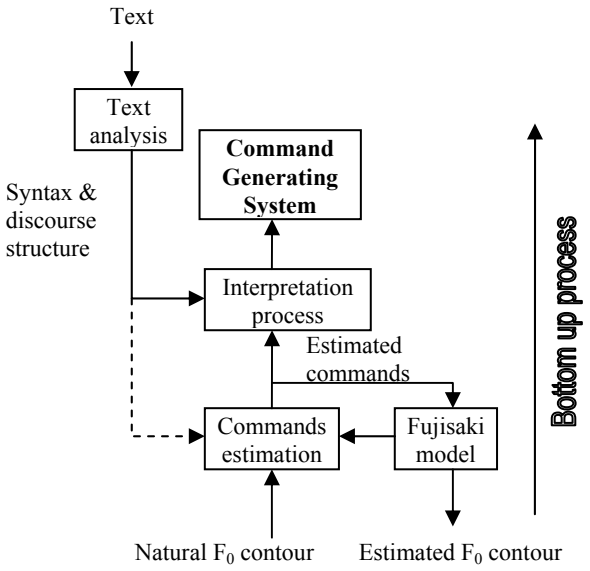

Figure 1: Estimation of commands and development of the system.

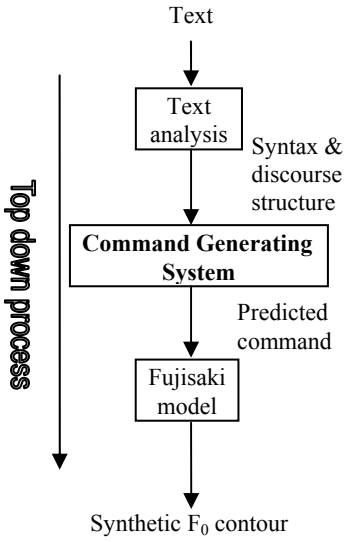

Figure 2: Command prediction process.

2. Implementation of the System

Since the Fujisaki model uses AC and PC where the respective components are added (in the logarithmic domain), as depicted in Fig. 3, two sub-systems were developed to deal with each type of command.

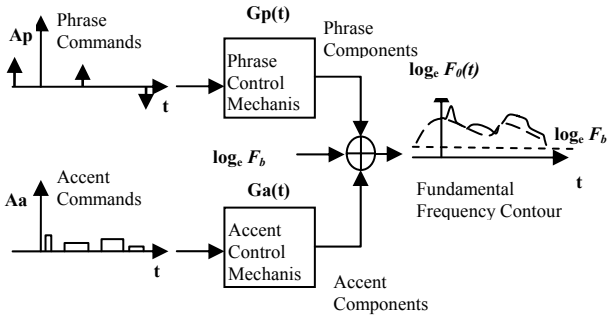

Figure 3: Fujisaki model for the process of generating F_0 contours [2].

F_0 is given by Eq. (1) and the phrase and accent control mechanisms are governed by Eq. (2) and Eq. (3), respectively.

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \left\{ G_a(t - T_{1j}) - G_a(t - T_{2j}) \right\} \quad (1)$$

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

$$G_a(t) = \begin{cases} \min\left[1 - (1 + \beta t)\exp(-\beta t), \gamma\right], & t \geq 0, \\ 0, & t < 0 \end{cases} \quad (3)$$

where the parameters are: Fb – baseline fundamental frequency; I – number of PC; J number of AC; Ap_i – magnitude of the ith PC; Aa_j – amplitude of the jth AC; T0i – timing of the ith PC; $T1_j$ – onset time of the jth AC; $T2_j$ – offset time of the jth AC; α – natural angular frequency of the phrase control mechanism; β - natural angular frequency of the accent control mechanism; γ – relative ceiling level of the accent components.

2.1. Command estimation process

The data used for training and testing was extracted from the FEUP-IPB database [6]. Mainly declarative and interrogative types of sentences were selected, in a total of 21 minutes of speech and 7,500 syllables. The corpus was divided into two sets, a training set consisting of 85% of the used database and the test set consisting of the remaining 15%.

It was experimentally determined that the baseline frequency is 75 Hz, the parameters α and β are respectively, 2.0 s^{-1} and 20 s^{-1}. These values were kept constant in the model. γ was kept constant and equal to 0.9.

The stream of Fujisaki parameters for the database were extracted in two phases using a specifically developed tool. In the first phase, the optimization was oriented by placement of PC and AC towards the best fit of the model-predicted contour relative to the original F_0 in voiced parts, without regarding the syntax and discourse structure. A strong correlation between AC and syllables was noticed. Therefore, the second phase of labeling took place considering that no more than one AC for each syllable that has its own F_0 movement should be introduced. The manual optimization allowed the improvement of the root-mean-squared error (rmse), resulting in a value of about 4 Hz, calculated along the whole corpus, between the estimated and the original F_0 contours in voiced parts. Moreover, the naturalness of the re-synthesized speech with the estimated parameters was found to be practically indistinguishable from the original.

2.2. Phrase command generating sub-system

The sub-system for predicting PC, described in detail in [1] and [7], performs in two steps. In the first step it inserts PC associated with the beginning of accent groups, based on orthographic marks and weighted candidates. The second step predicts the exact position of the PC namely T0a and PC magnitude Ap by means of two ANNs.

The location of inserted PC are found to be consistent with text and with labeled PC. Table 1 presents the number of correctly inserted PC (C), the deletion errors (D), the inserted errors (I), the recall rate (R), given by Eq. (4), and the precision rate (P), given by Eq. (5).

The linear correlation coefficient of the predicted Ap and T0a are 0.772 and 0.649 respectively.

Table 1: Results of the inserted PC

C	D	I	R	P
435	211	208	67.3%	67.7%

$$R(\%) = \frac{C}{C+D} \times 100\% \quad (4)$$

$$P(\%) = \frac{C}{C+I} \times 100\% \quad (5)$$

2.3. Accent command generating sub-system

The connection between AC and syllables were followed as described in [1] and [8].

The parameter estimation process was refined, in order to guarantee that not more than one AC is associated with each syllable. One AC is here considered associated with each syllable when the amplitude of the components within the voiced part of the syllable is higher than 35% of its maximum amplitude.

So, each AC is associated with just one syllable and each syllable can have one or no associated AC. Therefore the system has to decide for each syllable with voiced segments if it has an associated AC or not, and in case the decision is positive, to calculate the parameters of the respective AC.
For each AC three parameters have to be predicted: amplitude - Aa, onset time – T1 and offset time – T2. T1 and T2 are determined relative to the syllable boundaries by prediction of its anticipation T1a and T2a.

The parameters to be predicted are: Ca – logical parameter, indicating the presence or absence of an AC associated with the syllable; Aa – amplitude of AC, T1a anticipation of onset time; T2a anticipation of offset time.

Feed-forward ANNs, trained by the back-propagation algorithm, were used to predict these four parameters.

Table 2 presents the objective results of the prediction of each parameter. In case of Ca, the accuracy given by Eq. (6) and the correlation coefficient (r) are presented, meanwhile for the other parameter only r is presented.

Table 2: Correlation (r) and accuracy (A) between estimated and predicted parameters

Perform.	Ca	Aa	T1a	T2a
r	0.654	0.602	0.743	0.650
A(%)	89.3	-	-	-

$$A(\%) = \frac{\text{number of correct decisions}}{\text{number of syllables}} \times 100\% \qquad (6)$$

3. Evaluation

A final perceptual test was performed using 5 paragraphs of the test set, in a top-down process, as depicted in Fig. 2.

The standard category-judgment test [9] was conducted. This test starts with presentation of references of scale, namely, the references of excellent and unsatisfactory. The reference of excellence was the original recorded sound. The unsatisfactory reference was decided to be a flat F_0 contour with the average F_0 value (103 Hz), named as No model. The following stimuli were used in the perceptual test: No model; Original; Estimated – re-synthesis with F_0 obtained from estimated commands; Predicted AC – re-synthesis with F_0 obtained from estimated PC and predicted AC; F_0 model - re-synthesis with F_0 obtained from predicted PC and predicted AC.

Nineteen subjects judged the 5 paragraphs of each stimuli in a scale from 1 to 5 (1- **Unsatisfactory**, 2- **Poor**, 3- **Fair**, 4- **Good**, 5- **Excellent**). Table 3 presents the average root-mean-squared-error (rmse) and correlation coefficient (r) between the stimuli and original in the 5 paragraphs, as well as the Mean-Opinion-Score (MOS) of the perceptual test.

Table 3: Performance in the perceptual test set.

	No mod.	Original	Estim.	Pred. AC	F_0 mod.
rmse (Hz)	19.1	-	4.2	16.3	16.1
r	-	-	0.97	0.55	0.49
MOS	1.2	4.6	4.4	3.1	3.1

Figure 4 displays one application example of the model to a part of a paragraph from the test set. The English translation of the text is: "… and are certainly important for all, particularly for those who have responsibilities …".

In this figure, the following data is presented from top to bottom: sound waveform; + marks - measured F_0; thin line - estimated F_0 contour and phrase component added to base line frequency; thick line – predicted F_0 contour and predicted phrase component added to base line frequency; dashed lines – estimated commands; thick lines – predicted commands; solid line indicating syllables (thick lines represent tonic syllables) – descending sequence of syllables form one accent group; orthographic marks in text; words; phoneme labels. Vertical lines denote word boundaries.

4. Discussion

Original stimuli were very well classified within the level of **Excellent**. The unsatisfactory reference, No model, was classified with the level of **Unsatisfactory**. The F_0 of this stimulus was no correlation with original because it is a constant value.

MOS of estimated stimuli (4.4) is at the level of a **Good** acceptability of naturalness. The distance to the natural stimuli were very low (0.23), proving the closeness to the original. If there is a strong correlation between rmse and MOS, as it was proved in [1], this result can be extended to the complete database, once the mean rmse of paragraphs in the perceptual test (4.2 Hz) is at same level as the rmse in the complete database (3.97 Hz).

The predicted AC stimuli and F_0 model stimuli were classified at the same level in the rmse, r and MOS evaluation parameters, but with a significant reduction denoting some degradation introduced by the AC sub-system. Anyhow, the complete F_0 model received the final classification level of **Fair**.

The difference in MOS between estimated and predicted AC stimuli can be ascribed to the degradation due to the AC sub-system, because only the AC became different. Following the same idea, the difference between predicted AC stimuli and F_0 model stimuli could eventually be interpreted as the degradation due to the PC sub-model, because it is the new component added to the F_0 model. Anyhow, it is not exactly this way because the error components are not additive. The measured error (rmse, r or MOS) is the addition of non-orthogonal error components. These components are the accent and phrase components errors, which were considered as non-orthogonal in the generating system because the AC sub-system uses features dependent on the PC.

An analysis of the error introduced by the components of AC and PC in two situations, SA and SB follows. The situation SA corresponds to the produced stimulus of predicted AC of the perceptual tests, where the F_0 contour was determined by the prediction of only the AC using the estimated PC. The situation SB corresponds to the produced stimulus of F_0 model, where the F_0 contour was determined by the prediction of PCs and ACs. It is considered that no AC error is introduced between situations SA and SB. The error considered in the following analysis can be any of the parameters rmse, r, or MOS.

Figure 5 represents the error vector eSA in situation SA, and eSB in situation SB,, considering that the PC and AC error axis (PCe and ACe) are non-orthogonal. The eSA has only AC error component (ACe), with the value eAC, and no PC error component (PCe), once the estimated PCs were used and theoretically, it has no error. The situation B, SB, now with phrase component error δePC, has the same AC component eAC, but a different absolute error eSB. This example shows that even a decrease in the absolute error from SA to SB, δe, can correspond to a significant increase in the phrase command component error, δePC.

A group of experiments and measurements can be considered with the objective of measuring the angle between the PC and AC components.

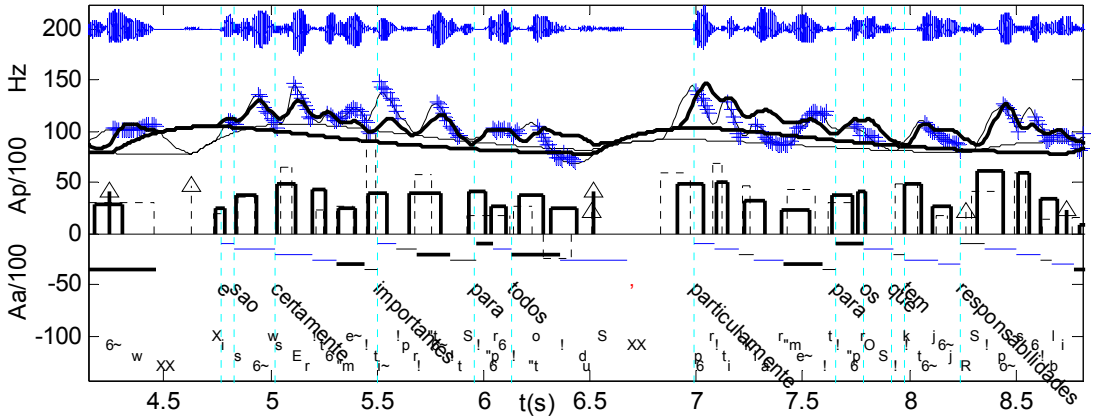

Figure 4: F0 contour predicted with F0 contour prediction system.

5. Conclusions

A scheme for speech F_0 contour generation based on the CGS for the Fujisaki model was evaluated. The CGS consists of the AC and PC generating sub-systems. The development of the AC and PC generating sub-systems in a bottom-up process was presented. The PC generating sub-system inserts PCs with a recall rate and precision rate both at 67%. There is a room for improvement using a semantic knowledge in the PC generating sub-system. The magnitude and anticipation of the PC are predicted with an r of 0.77 and 0.65, respectively. These results represent an improvement compared with a similar work on German [4], but there is still a room for improvement. A strong consistency in the phrasing process is fundamental. The association of AC with syllables in the AC generating sub-system produces F_0 movements very similar to the original ones, as depicted in Fig. 4. Again, the correlation and accuracy results present some improvement, but still far from excellence. The lack of focus information does not allow the system to produce strong accents at focus positions.

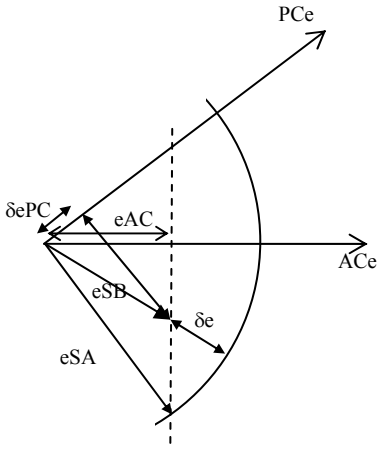

Figure 5: PC and AC error components in two situations, considering non-orthogonal axis.

Therefore, focus information is crucial. Experiments suggest that focus depends on phonological, linguistic, paralinguistic (stylistic) and even nonlinguistic (emotional) factors.

Perceptual tests confirmed that the sub-systems produce fair prosody, but still need improvements to be accepted as a good prosody model.

6. References

[1] Teixeira, J. P. 2004. "A Prosody Model to TTS Systems",*Doctoral thesis in Electrotechnical and Computer Engi.* in Faculdade de Engenharia da Universidade do Porto. http://www.ipb.pt/~joaopt/publicacoes/publicacoes.html

[2] Fujisaki, H., 2002. Modeling in the Study of Tonal Features of Speech with Application to Multilingual Speech Synthesis. *Proceedings of Joint International Conference of SNLP and Oriental COCOSDA*. Hua-Hin, Thailand, D1-D10.

[3] Fujisaki, H., Ohno, S., 1995. Analysis and Modeling of Fundamental Frequency Contours of English Utterances. *Proceedings of Eurospeech'95,* Madrid, 985-988.

[4] Mixdorff, H., 2002. An Integrated Approach to Modeling German Prosody. *Thesis for Dr.-Ing. Habil.*, Technical University of Dresden.

[5] Navas, E., Hernáez, I., Sánchez, J. M., 2002. Basque Intonation Modelling for Text to Speech Conversion. *Proceedings of ICSLP'02*, Denver, 2409-2412.

[6] Teixeira, J. P., Freitas, D., Braga, D., Barros, J., Latsch, 2001. Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB. *Proceedings* of *Eurospeech'01*, Aalborg, 1707-1710.

[7] Teixeira, J. P., Freitas, D., Fujisaki, H., 2003. Prediction of Fujisaki Model's Phrase Commands. *Proceedings of Eurospeech 2003*. Geneva, 397-400.

[8] Teixeira, J. P.; Freitas, D. and Fujisaki, H.. (2004). Prediction of Accent Commands for the Fujisaki Intonation Model. *Proceedings of Speech Prosody 2004,* Nara - Japan. Pages 451-455.

[9] Standard Publication No. 297, IEEE, (1969). IEEE Recommended Pratice for Speech Quality Measurements. *IEEE Transations on Audio and Electroacoustics*. Vol. AU-17, no.3. 1969.