Prediction of Accent Commands for the Fujisaki Intonation Model

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 a model to predict the accent commands (henceforth ACs) of the Fujisaki Model for the F0 contour, being known the phrase commands (henceforth FCs). Accent commands are associated with syllables. For each syllable, an artificial neural network (ANN) decides, with an accuracy of 89.4% whether there will be an associated AC or not. For syllables with associated AC, the amplitude, Aa, the onset time anticipation, T1a, and the offset time anticipation, T2a, are predicted by additional ANNs, with resulting linear correlation coefficient of 0.602, 0.743 and 0.650, respectively. The features used for each ANN are presented and discussed. Finally a comparison between target and predicted F0 contour is presented.

1. Introduction

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

F0 is the most perceptually relevant component in prosody. The Fujisaki model for F0 contour has been proven to be well adapted, with very high naturalness, to several languages [2] like Japanese, Korean, Spanish, Polish, Greek, Swedish, English [3], German [4], Basque [5] and now also Portuguese [1][6]. The model [2] consists of the logarithmic addition of baseline fundamental frequency, F_b, phrase components and accent components. The baseline fundamental frequency is constant in each utterance. Phrase commands are represented as a set of impulses and the accent commands as a set of pedestal functions. The corresponding components are represented as responses of respective control mechanisms to these commands.

The model to predict FCs was described in [1]. It works in two steps. In the first step FCs associated with the beginning of accent groups are inserted, based on orthographic marks and weighted candidates. In the second step the exact position T0 and the magnitude Ap of each FC are predicted by means of two ANNs.

To complete the automatic operation of the model for European Portuguese, this paper is dedicated to describe the prediction of the remaining components, the ACs, assuming that the FCs are already known. The denominated beta, the natural angular frequency parameter of the AC, is here considered to be constant for all ACs.

2. Text and Speech Corpus and F0 Labeling

The data used for training and testing were extracted from the FEUP-IPB database [7]. This database consists of several labeled speech tracks of read speech. The speech waveforms were manually labelled in three levels: the phonetic, word and phrase levels. Seven tracks of the database were used, in a total of 101 utterances relative to paragraphs of high variety of lengths, from 1 to 100 words. 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 syllables and the test set consisting of the remaining 15%.

The Fujisaki parameters for the database were extracted using a specifically developed tool. The process of labeling consisted in a manual optimization of the FC and introduction of no more than one AC for each syllable that has its own F0 movement. The optimization was oriented towards the best fit of the model predicted contour relative to the original F0 in voiced parts. The manual optimization allowed the improvement of the root-mean-squared error (rmse), resulting in a value of 3.97Hz, calculated along the whole corpus, between the estimated and the original F0 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.

3. Accent Commands

The connection between ACs and syllables were followed. The present approximation is different from the approximations used by Mixdorff [4] or even by Navas [5] that consider one AC by accent group to be enough.

A refinement of the parameter estimation process was needed, in order to guarantee that not more than one AC is associated to each syllable. One AC is here considered associated to each syllable when the area 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 model 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 relatively to the syllable boundaries. Namely, T1 is determined as the beginning of the voiced segments of syllable minus an initial anticipation (1), and T2 is the end of the voiced segments of the syllable minus a final anticipation (2). These anticipations, from now on T1a and T2a, are the timing parameters to be predicted, once the beginning and end of the voiced part of the syllable are known:

$$T1 = Bvs - T1a \quad (1)$$

$$T2 = Evs - T2a \quad (2)$$

where: Bvs – beginning time of voiced segments; Evs – end time of voiced segments; T1a – T1 anticipation; and T2a – T2 anticipation.

The additional parameters to be predicted are:

- Ca – logical, indicating the presence or absence of an AC associated with the syllable
- Aa – the amplitude of the AC

Feed-forward ANNs trained with back-propagation algorithms were used to predict these four parameters.

An optimization of the ANN architectures, numbers of layers, numbers of nodes per layer, activating functions, training algorithms and sets of features was done. The best performances for the parameters were achieved with different ANNs. This may be explained by the low correlation that exists between parameters, that is: r(Aa,T1a) = 0.33; r(Aa,T2a) = 0.43; r(Aa,Ca) = 0.61; r(T1a,T2a) = 0.34; r(T1a,Ca) = 0.29; r(T2a,Ca) = 0.49.

Table 1 summarizes the architectures used for each ANN to predict required parameters. Numbers of nodes in layers are presented in the first column, considering the input node as the first node. The ANNs that have 25 nodes in input layer just use the first 25 features of Table 2.

Table 1: *Architectures and performances of the ANNs selected to predict the output parameters.*

	Nodes/layer	Activating functions	r
Ca	27-10-1	Hyp. Log. - Linear	0.654
Aa	27-6-1	Hyp. Log. - Linear	0.602
T1a	25-10-1	Hyp. Log. - Linear	0.743
T2a	25-7-5-1	Hyp. Tang. -Hyp. Log. - Linear	0.650

In the training set, the target values of ANN for Ca are 1 or 0 depending on whether an AC is associated with a syllable or not, respectively. The predicted value is set to 1 or 0 as the output of the ANN is higher or lower than the threshold L = 0.5.

The output values of Aa's, T1a's and T2a's ANNs are 85% of the parameter's value divided by its maximum and normalized to have a null average and standard deviation equal to 1.

Training was done over the training set and using the test set for cross-validation in order to avoid over-fitting. The test vector was used to stop training early if further training on the training set will hurt generalization to the test set. The mean squared error between output and target values was used as the cost function. The training algorithm was the back-propagation Levenberg-Marquardt.

For the output parameters Aa, T1a and T2a, training sessions using just the syllables with associated AC were experimented. This sub-set for training gave a better performance just for the T2a's ANN.

3.1. Features

The sets of features were built taking into account the known and foreseeable dependencies as well as local contextual information. An optimization followed in the composition of the sets and the ways of coding. Although some features present a very low correlation with the output parameter, as can be seen in Table 2, their ensemble use in the whole set of features improves the final performance.

Any listed feature is coded in one node of the input layer. A short comment concerning each feature follows: **F1**: Strongly correlated with the presence of AC, its amplitude, Aa, T1a and T2a. **F2**: It is even more strongly correlated with the presence of AC, because syllables without voiced segments don't have an associated AC. It is also strongly correlated with Aa and T2a, but negatively correlated with T1a. What mean that the longer the voiced part of syllables is, the later is the onset time of AC.

Table 2: *List of features and their correlation, r, with Ca, Aa, T1a, and T2a.*

F #	Feature description	r(Ca)	r(Aa)	r(T1a)	r(T2a)
1	Syllable duration	0.27	0.17	0.11	0.43
2	Dur. of syll. voiced part	0.40	0.17	-0.21	0.43
3	Vowel duration	0.41	0.22	0.04	0.42
4	Type of syllable	0.51	0.39	0.23	0.38
5	Tonic syllable	0.26	0.23	0.08	0.24
6	Type of vowel in syll.	0.41	0.31	0.12	0.39
7	Dist. (s) to end of sent.	0.05	0.07	0.01	-0.03
8	Dist. (s) to beg. of phrase	-0.04	-0.05	-0.05	0.00
9	# of AC from beg. of phrase	-0.02	-0.04	-0.07	0.00
10	Dist. (s) to beg. of FC	0.02	0.09	0.01	0.04
11	# of AC from beg. of FC	0.02	0.06	-0.03	0.02
12	Dist. (s) to next FC	-0.02	-0.00	0.01	-0.07
13	Last word of paragraph	-0.06	-0.06	-0.02	0.08
14	Last syll. Of paragraph	-0.10	-0.10	-0.04	0.04
15	Last word of sentence	-0.06	-0.08	-0.02	0.10
16	Last syll. of sentence	-0.10	-0.13	-0.05	0.09
17	Syll. number in the word	-0.08	-0.13	-0.11	0.00
18	# of syll. to end of word	0.01	0.12	0.07	-0.05
19	# of syllables in the word	-0.06	-0.01	-0.03	-0.04
20	Dur. (s) of the word	0.02	0.02	0.01	0.10
21	Aa of previous AC	-0.03	0.07	0.01	0.03
22	Dur. (s) of previous AC	0.01	-0.02	-0.05	-0.01
23	Dist. (s) to previous T2	0.00	0.01	0.22	0.04
24	Dist. (s) to previous pause	-0.05	-0.05	-0.08	-0.03
25	Dist. (s) to next pause	0.01	0.06	0.04	-0.03
26	Last tonic syll. of ISWIW*	0.02	0.02	-0.03	-0.01
27	ISWIW*	-0.07	-0.04	-0.02	-0.03

ISWIW* – interrogative sentence without interrogative word.

F3: Duration of vowel or diphthong of the syllable. Is zero in cases of syllables where the vowel was suppressed. It is also strongly correlated with the presence of AC, its amplitude, Aa and T2a. In fact, these first three features are quite highly correlated with each other, but do not carry exactly the same information. **F4**: This feature codes the type of syllable according to vowel (V) and consonant (C) sequences. Codification is the following, from lower to higher correlation: 1-C; 2-CC; 3-V; 4-VC; 5-VCC; 6-CCVC; 7-CCV; 8-CVC; 9-CV. This feature is the most correlated with the presence of AC, its amplitude, Aa and T1a. It is also strongly correlated with T2a. **F5**: Signalizes if the syllable is tonic or not. Tonic syllables have a strong correlation with the presence of AC, its amplitude, Aa and T2a. **F6**: Vowels were divided in five groups according to the average length and category. Codification is the following, from lower to higher correlation (phonemes are presented in SAMPA code): 1-short vowels (u and @); 2-median vowels (i and 6); 3-

diphtongs; 4-nasal vowels; 5-long vowels (a, E, e, o and O). It is strongly correlated with all output parameters. **F7**: Distance, in seconds, from the beginning of the syllable to the end of the sentence. It is slightly correlated with the presence of AC and its amplitude, Aa. **F8**: Slightly negatively correlated with the presence of AC, its amplitude, Aa and T1a. **F9**: Slightly negatively correlated with T1a. **F10**: Distance in seconds from the beginning of FC to the beginning of the syllable. Is slightly correlated with Aa. **F11**: Slightly correlated with Aa. **F12**: Slightly negatively correlated with T2a. **F13**: Signalizes if the present syllable belongs to the last word of the paragraph. It is coded as yes/no (1/0). It is slightly negatively correlated with the presence of AC, and slightly correlated with a longer anticipation of the off set time. **F14**: Signalizes if the present syllable is the last one of the paragraph. It is coded as yes/no (1/0). It is negatively correlated with the presence of AC, and its amplitude, Aa. **F15**: Signalizes if the present syllable belongs to the last word of the sentence. It is coded as yes/no (1/0). It is slightly negatively correlated with the presence of AC, its amplitude, Aa and is correlated with a longer anticipation of offset time. **F16**: Signalizes if the present syllable is the last one of the sentence. It is coded as yes/no (1/0). It is negatively correlated with the presence of AC and its amplitude, Aa, and is slightly correlated with longer anticipation of the offset time. **F17**: Position in word - codes the number of syllables existing towards the beginning of word. It is negatively correlated with the presence of AC, its amplitude, Aa and T1a. **F18**: Position in word - codes the number of syllables towards the end of word. It is correlated with Aa and T1a. **F19**: Word length is the total number of syllables in the word. **F20**: Word duration. It is correlated with T2a. **F21**: Slightly correlated with Aa. **F22**: Length of previous AC. **F23**: Distance in seconds to offset time of previous AC. The farther is the offset time of the previous AC, the longer is the anticipation of the onset time of present AC. **F24**: Slightly negatively correlated with all parameters. F25: This is slightly correlated with Aa. **F26**: This feature codes the condition of the present syllable being the last tonic syllable of an interrogative sentence of type without interrogative word. It is coded as yes/no (1/0). **F27**: This feature codes the condition of the syllable belonging to an interrogative sentence of type without interrogative word. It is coded as yes/no (1/0). This and the previous features intend to code the situation of the last tonic syllable in an interrogative sentence of type without interrogative word, which is known to have a rise-fall F0 contour.

4. Results

To evaluate the performance of the Ca's ANN four parameters were used: linear correlation coefficient (r), accuracy (A – given by (3)), recall rate (R – given by (4)) and precision rate (P – given by (5)):

$$A(\%) = \frac{\text{number of correct decisions}}{\text{number of syllables}} \times 100\% \quad (3)$$

where the number of correct decisions is the number of times that the output matches the target as to having an associated AC or not;

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

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

where C is the number of correctly inserted ACs, D is the number of deleted (i.e., not inserted) ACs, and I is the number of inserted errors (i.e., incorrectly inserted ACs).

Table 3: *Performance values for the best Ca's ANN.*

A(%)	r	P(%)	R(%)
89,3	0,654	97,3	91,5

Table 3 reports the performance for Ca's ANN in test set.

For Aa's, T1a's and T2a's ANNs the performance was measured by the linear correlation coefficient and is presented in Table 1 above.

However, the statistics presented for performance of individual parameters of the AC prediction model is not enough to efficiently evaluate the prediction of Accent Commands.

Figures 1 and 2 display one application example each of the model to a part of a paragraph from the test set. In Figure 1 the AC-prediction model is applied over the set of labeled FCs. In Figure 2 both FCs and ACs are predicted by the developed methods presented in [1] and here, respectively. The English translation of the text presented in the Figures is: "… and are certainly important for all, particularly for those who have responsibilities …".

In these figures, the following data is presented from top to bottom: sound waveform; + marks - measured F0; thin line - estimated F0 contour and phrase component added to base line frequency; thick line – predicted F0 contour using AC model; phrase commands in Figure 1, and estimated and predicted phrase commands with thick and dashed lines, respectively, in Figure 2; estimated and predicted ACs with dashed and solid lines, respectively; 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.

5. Conclusions

An ANN-based model to predict ACs was presented in this paper. It is the last part of a prosody system that has been developed for text-to-speech synthesis of EP. This system consists of a specific model for prediction of the segmental durations [8] and two other models for prediction of F0 contours based on Fujisaki's FC and AC. The complete prosody system produces contours that modulate the speech that is to be produced from the given text. The model described here allows one AC to be associated with each syllable. The correlations between predicted and target values of individual parameters are quite good compared with the ones produced in similar works [4][5].

The model considers γ and β as constants with values 0.9 and 20 /s, respectively. Although β=30 /s gives a better fitting between predicted and original F0 contours, the re-synthesized speech does not sounds quite natural. With β=20 this problem seems to be reduced.

The produced F0 contour with the predicted parameters approximately follows the measured F0. The major differences are coming from the difficulty in emphasizing the "focus" word due to the absence of this information in the

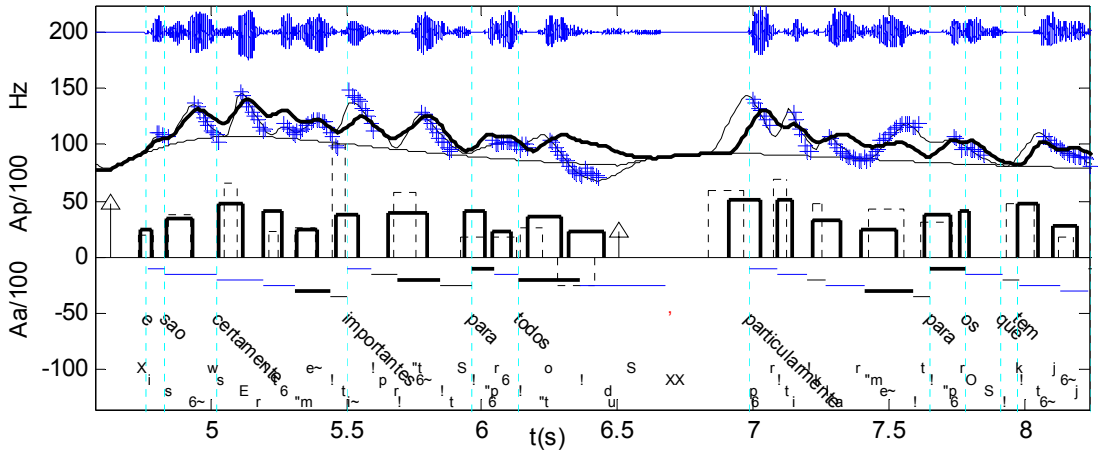

Figure 1: *F0 contour predicted with AC model over labeled FCs.*

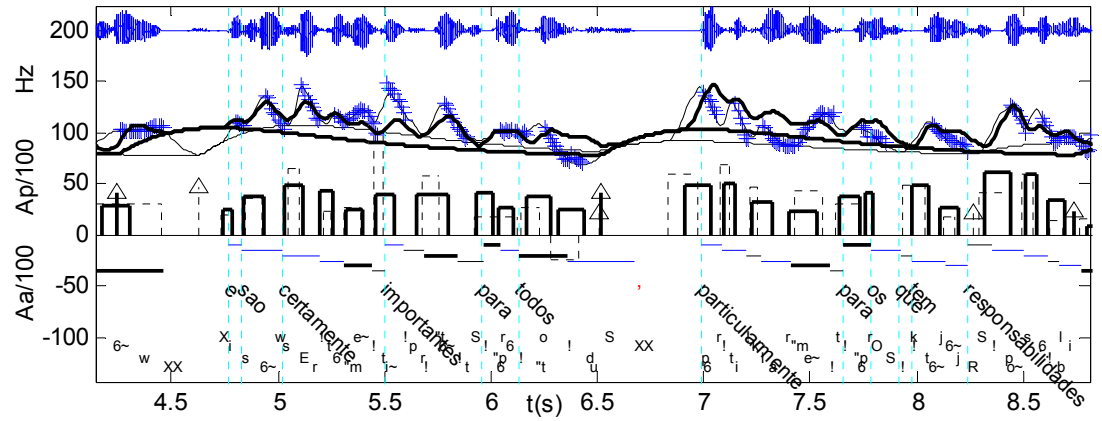

Figure 2*: F0 contour predicted with AC model over predicted FCs.*

training phase of the model. The final speech signal, produced by re-synthesis with the predicted F0 contour is not completely natural yet, but is considered as acceptable.

It is fair to mention that the model uses just some of the available linguistic information. For instance, syntax information has not been used. Moreover, paralinguistic information is not extracted by the model and several times the speaker produces a higher F0 movement, which can be explained by this kind of information that is not followed.

A perceptual test using a MOS scale is under way using as stimuli speech waveforms produced by F0 modification with a PSOLA algorithm. Preliminary work with 15 subjects gave an average score of 3.2 and 3.1 for the ACs predicted with labeled and predicted FCs, respectively, against 4.6 for the original stimulus. The ensemble usage of the whole prosody system, for durations [8] and F0, achieves a score of 3.0. The general scores achieved of 3 are at the "fair" level in a MOS scale.

6. References

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

[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*. May 2002, 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] Fujisaki, H., Narusawa, S., Ohno, S., Freitas, D., 2003. Analysis and Modeling of F0 Contours of Portuguese Utterances Based on the Command-Response Model. *Proceedings of Eurospeech 2003*. Geneva, 2317-2320.

[7] 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. *Proc.* of *Eurospeech'01*, Aalborg, 1707-1710.

[8] Teixeira, J. P., Freitas, D., Segmental Durations Predicted with a Neural Network. *Proceedings of Eurospeech 2003*. Geneva, 169-172.