Automatic Analysis and Synthesis of Fujisaki's Intonation Model for TTS

Pablo Daniel Agüero, Klaus Wimmer and Antonio Bonafonte

Department of Signal Theory and Communications
TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona - Spain
{pdaguero; alwk; antonio}@gps.tsc.upc.es

Abstract

This paper deals with the automatic analysis and synthesis of intonation using Fujisaki's model. We propose an analysis method which imposes strong linguistic constraints. This method produces good representations of the F0 contour when compared to other current methods which do not impose such constrains. Furthermore, this option limits the variability and is more predictable so it is the best option for prediction (at least when accent commands are related to accent groups). Several prediction algorithms are evaluated. The results show that VCART (an extension of CART to predict vector values) gives the best performance when compared with standard CART or with neural networks. The paper also analyzes which features are more relevant to predict the parameters of Fujisaki's model.

1. Introduction

Fujisaki's model [1] is a well known representation of intonation. In this model, the lnF_0 is expressed by the superposition of the baseline value of the fundamental frequency (F_b) with the outputs of two critically-damped second-order filters. The first filter is excited by deltas and accounts for the slow-varying phrase component. The second filter is excited by pulses and accounts for the fast-varying accent component.

Fujisaki's model is very compact: there is no redundancy in phrase and accent commands. It is able to cover many intonation phenomena (although the model needs to be validated on different speech styles). And the commands can be linguistically interpreted as phrase and accent movements. However, the representation of the F_0 contour is not unique. In fact, the F_0 contour can be approximated by the output of the model with arbitrary accuracy, if a large number of commands is used. Therefore, there is always a trade off between minimizing the approximation error and obtaining a set of linguistically meaningful commands.

With respect to automatic analysis, recently some methods were proposed for extracting the parameters of Fujisaki's model. Mixdorff [2] and Narusawa et al. [3] extracted the parameters without using any linguistic information. The algorithms produce contours that match closely the original contour. The main problem of this approach is that it may be difficult to relate the extracted parameters to linguistic information. To increase the linguistic interpretability, Möbius constrained Fujisaki's parameters by intonational entities directly related to linguistic features [4]. In this paper we analyze if these constrains degrade the aproximation to the original contour.

This work has been sponsored by the Spanish Government under grant TIC2002-04447-C02.

Concerning to synthesis, it is necessary to predict the position and amplitude of phrase and accent commands from linguistic features. Navas et al. [5] proposed to predict Fujisaki's intonation model parameters using CART which is well suited to classify and predict values based on unordered discrete values. In the standard implementation, if several variables are predicted, they are handled independently, building a specific tree for each one. This can produce ill effects if the values are correlated. Neural networks are another prediction method which is frequently used in text-to-speech synthesis. Although they are not very well suited to deal with unordered discrete features, they give good results in many problems. Mixdorff et al. [6] predicted the parameters of each syllable using a feed-forward neural network. It jointly predicted the parameters of Fujisaki's intonation model and other prosodic patterns. In this paper we compare these two methods with an extension to the CART algorithm: one unique regression tree is built to jointly predict correlated parameters.

This paper is organized as follows. In section 2, we analyze the extraction methods proposed by Mixdorff [2] and Narusawa et al. [3]. Some modifications are proposed and evaluated. It will be seen that the predicted commands are not easily related to linguistic features. Therefore, we propose a method that extracts the commands imposing severe linguistic constraints. The performance of this method is comparable to the other ones, validating the imposed constraints. Section 3 compares the three prediction algorithms mentioned previously. This section also analyzes which features are more useful to predict each command. Finally, the main conclusions of this work are drawn.

2. Analysis: extraction of the parameters

In Fujisaki's intonation model, the lnF_0 satisfies the following equation:

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

$G_p(t)$ represents the impulse response function of the phrase control mechanism and $G_a(t)$ represents the step response function of the accent control mechanism. These functions are fixed for a wide range of speakers and languages.

The underlying phrase and accent commands of an utterance ($A_{p_i}, T_{0_i}, A_{a_j}, T_{1_j}$ and T_{2_j}) cannot be directly inferred,

because there is no analytical solution to the inversion of Fujisaki's model. Usually, approximation algorithms are used to infer the underlying commands. A good initial estimation of the number and position of phrase and accent commands is crucial to obtain linguistically meaningful commands. This section compares the methods of Mixdorff [2] and Fujisaki et al. [7, 3, 8]. Some modifications are proposed to the previous methods. Finally, we introduce a new method that imposes linguistic constraints.

2.1. Algorithms based on Mixdorff's method

Mixdorff proposed a multi-stage approach performing a spectral decomposition in low frequency (LFC) and high frequency (HFC) components of fundamental frequency contour [2]. The first step of the algorithm consists of a quadratic spline stylization of the curve. Then, the smoothed fundamental frequency contour is decomposed in two components, using low-pass and high-pass filters. The output of the high-pass accounts for the faster movements in the F0 contour. Therefore accent command onsets and offsets are detected in the high-pass filtered contour, corresponding to two subsequent minimum points. Since the low-pass filtered contour contains the slower variations of the F0 movements, the onset of a phrase command is characterized by a local minimum of the low frequency contour. This initial command parameter sequence is refined by a three-step hill-climbing search, which minimizes the overall mean-square-error. In our study we use the freely available implementation of Mixdorff, herein after called *M1*.

However some problems were observed: the extraction procedure tends to miss the first accent command, if the high-pass filtered contour starts with an absolute minimum. Therefore we modified the command initialization procedure and placed the first accent command already in the beginning of the phrase, if it is indicated by the absolute minimum. This applies analogously to the last accent command. Furthermore algorithm *M1* is optimized for German and misses many commands in our Spanish speech corpus. For this reason the parameters of the initial smoothing procedure are adapted. The modified algorithm is denoted hreafter as *M2*.

2.2. Algorithms based on Fujisaki's method

Fujisaki et al. [7, 3, 8] suggested to extract parameters applying a preprocessing procedure that results in the continuous third-order polynomial stylization. Commands are searched in the derivate of this function. First preprocessing is performed to eliminate gross errors and border effects (e.g. micro-prosodic disturbances). Then the contour is smoothed using piecewise cubic interpolation, which results in a curve, that is differentiable everywhere. A sequence of maximum and minimum of the first derivative of the smoothed F0 contour corresponds to the onset and offset of an accent command. The effect of accent commands is subtracted from the smoothed F0 contour. The phrase commands should approximate this residual F0 contour.

Since no detailed description of the phrase command estimation procedure was given, we applied the following method: maximum points in this residual contour correspond to phrase command onsets, with a delay equal to the damping time of the phrase control filter. The minimum distance of two subsequent phrase commands is limited to 500 msec. The integral of the residual contour serves for amplitude estimation of the phrase command. This method is named *F1*.

The above method relies on the derivative of the smoothed contour, which also contains minor effects of the phrase command response. This may sometimes confuse the accent command initialization. Therefore we also applied accent and phrase command estimation in reverse order. Extracting phrase commands first and subtracting them from the smoothed contour yields a residual curve that should only contain the effects of accent command responses. This should facilitate the estimation of the onsets and offsets of accent commands. On the other hand the estimation of phrase commands needs to be more precise. We refer to this as version *F2*.

2.3. Linguistically Motivated Method

The previous methods do not impose any constrain in the parameters. This flexibility allows the algorithms to produce contours that closely match the original contours. However, it may be difficult to interpret linguistically and to predict the commands from the text. Our proposal is a parameter extraction algorithm that uses information about accent groups and prosodic phrases, which is available both in the TTS system and in the training corpus. Accent groups are defined as one content word and all the preceding function words. This intonation unit has being used frequently to describe Spanish intonation and has being used by recent proposals (see for instance [9]), obtaining good representations. The algorithm applies the following constraints:

- Each prosodic phrase is modeled by one phrase command, that can only appear within a window of 200 msec centered at the beginning of the prosodic phrase.
- The number of accent commands inside each accent group is limited to one.

To simplify the command estimation procedure we make use of the following observation: phrase and accent commands only influence future commands. Since the distances of the commands are normally rather large, they often do not interact at all. Therefore we can obtain the initial guess of each command independently of the others, using a left-to-right procedure. First, the $\log F_0$ contour is interpolated in the voiceless segments and the base frequency is removed. The result is filtered using a median filter. The onset of the phrase command is limited to the beginning of the phrase ± 100 msec. The phrase command amplitude is computed so that the phrase command does not cut the original contour. In this way, the accent command responses can be superposed later on. Each phrase command is removed from the original contour, resulting in a residual contour. Accent commands are searched in the residual contour, computing the optimal magnitude for each combination of onset and offset. The command that gives the least mean square error is chosen. The minimal duration of an accent command is fixed to 50 msec.

This initial approximation is improved by a hill-climbing algorithm based on gradient descent, that minimizes the mean-square-error in the log-domain between the original F0 contour and the synthetic one, optimizing all commands jointly (the whole method is called *L1* in the sequel).

We also propose an alternative, which forces the accent commands to be closely related not only to accent groups, but also to the stressed syllables. The search of initial accent command positions is restricted to the stressed syllable inside the accent group, the preceding and the following syllable. The error function of the hill-climbing procedure is weighted to further favor the placement of accent commands near stressed syllables. The developed algorithm is referred to as *L2*.

2.4. Experimental results

The employed speech corpus consists of 500 declarative sentences, that were read by a female speaker. The fundamental frequency contour was derived from the laryngograph signal. Altogether it contains 2893 accent groups and 937 prosodic phrases. The value of the baseline frequency F_b for the Fujisaki model was set to 107 Hz.

Table 1 shows the results of the algorithms described in this section. For each algorithm the mean square error (MSE) and the correlation (ρ), is computed comparing the contour produced by the extracted commands and the original $\log F_0$ contour (after interpolation). The table also shows the number of phrase commands and the accent commands normalized by the number of prosodic phrases and accent groups respectively. The results reveal the correlation between the error and the number of commands. In general, the algorithms *M1* and *M2* tend to place fewer commands than the other methods. Consequently, the MSE of methods *M1* and *M2* is relatively high compared to the other approaches. As already mentioned, the method *M1* seems to ignore minor accent commands and also miss some phrase commands. The modified method *M2* found more accent commands than the original method *M1* and consequently gets slightly lower error. On the other hand, the methods *F1* and *F2* have the lowest errors, but they tend to use more commands than the other methods. The method *F2* gives better results than *F1*. This indicates that primary estimation of phrase commands was successfully applied. Finally, in spite of the strong restrictions applied, the algorithms *L1* and even *L2* produce good approximations to the original F0 contour.

Table 1: *Results of command extraction: #pc: number of phrase commands / number of phrases; #ac: number of accent commands / number of accent groups; MSE: mean square error; ρ: correlation coefficient.*

Algorithm	#pc	#ac	MSE	ρ
M1	0.57	0.56	$21.6 * 10^{-3}$	0.79
M2	0.59	0.77	$19.0 * 10^{-3}$	0.81
F1	1.47	1.53	$2.0 * 10^{-3}$	0.96
F2	1.60	1.60	$1.1 * 10^{-3}$	0.98
L1	1.00	1.00	$5.5 * 10^{-3}$	0.87
L2	1.00	1.00	$5.8 * 10^{-3}$	0.87

2.5. Experimental results with limited commands

We propose to relate accent commands to the accent group. The algorithms *M1*, *M2*, *F1* and *F2* may detect several accent commands in each accent group. Thus, the number of accent commands has to be predicted. Several experiments were done to predict this number from the features available in the TTS intonation module but the classification performance was very poor.

Alternatively, the commands extracted were filtered to select for each accent group only the accent command with the largest area. However, the deletion of the small accent commands degrades dramatically the results, specially for the *F1* and *F2* algorithms. For instance, for the *F1* method the MSE increases to $16 * 10^{-3}$.

As conclusion, the methods *L1* and *L2* are selected as the analysis method. It seems that the best option to relate the commands to the accent group is to impose linguistic constraints in the extraction algorithm since the initialization.

3. Synthesis: prediction of the parameters

3.1. Prediction methodology

As stated in previous section, we propose to relate the accent command to the accent group and the phrase command to the prosodic phrase. The selected extraction algorithm are the ones named (*L1* and *L2*). For each accent group, the parameters of the unique command (A_a, T_1, and $T_2 - T_1$) are predicted. All the times are referred to the beginning of the accent group. This procedure is iterated for each accent group from left to right. Furthermore, the algorithm predicts the parameters of the phrase command (A_p and T_0) for each prosodic phrase.

3.2. Features for prediction

In text-to-speech synthesis, the F_0 contour needs to be predicted from linguistic features derived from input text. Furthermore, in many systems, phrase detection and duration assignment is already done before predicting the F_0 contour. We consider all the available features which may be related to the F_0 contour. Many of them were proposed by Navas et al. [5]. These features include information as duration, position, type, etc. from several prosodic units as the stressed syllable, accent group, and prosodic phrase. Some additional features were added. For instance, to predict the accent command, we considered the duration of the stressed syllable, the POS of the first and last word of the accent group and the amplitude of the last phrase command.

3.3. Prediction algorithms

Three prediction algorithms are evaluated. The CART algorithm independently predicts each one of the five values. CART algorithm has been successfully used in many problems, but in the standard implementation does not exploit correlation between the values to be predicted. The second prediction algorithm is based on neural networks. Two neural networks are trained, one for the accent command parameters and other for the phrase command parameters. Neural networks predict jointly the parameters but they are not very well suited to deal with unordered discrete values. Finally, the CART algorithm is extended to predict vector values (this extension will be referred as VCART). At each node of the tree, the chosen question is the one that minimizes the mean Mahalanobis distance to the centroid. This algorithm can be seen as a vector clustering based on features. As in the neural networks case, two VCART are built, one to predict accent commands and the other for phrase commands. The advantage of this method with respect to standard CART is that it avoids some ill effects that occur when correlated parameters are predicted independently.

3.4. Experimental results

Several experiments are performed to evaluate the performance of the different methods. The classifiers are trained using the commands extracted from 400 declarative sentences. For the 100 test sentences, the commands are predicted and the $\log F_0$ contour is generated. Then, the mean square error (MSE) and the correlation (ρ) are computed as defined in section 2.4.

Table 2 compares the performance of each prediction method (CART, neural networks and VCART) for the two methods selected to extract the parameters of Fujisaki's model from the training corpus. For the extraction method *L1*, neural networks and VCART perform significantly better than CART. This reveals the importance of exploiting the correlation between the parameters of the commands. VCART is in this case

Table 2: *Results of contour prediction for the selected extraction methods* L1 *and* L2*.*

Pred. algorithm	CART		Neural Networks		VCART	
Extrac. algorithm	MSE	ρ	MSE	ρ	MSE	ρ
L1	$34.6 * 10^{-3}$	0.44	$23.4 * 10^{-3}$	0.47	$20.6 * 10^{-3}$	0.50
L2	$22.5 * 10^{-3}$	0.51	$22.0 * 10^{-3}$	0.49	$22.6 * 10^{-3}$	0.50

the best solution. For the *L2* algorithm, the prediction method does not affect significantly the results. We believe that as the extraction method impose so strong constraints in the positions, these values are almost fixed and therefore the parameters are less correlated. The use of *L1* with VCART gives the best results (in MSE) but *L2* gets similar results (even better in correlation). We conclude that imposing strong restriction on the position of Fujisaki's commands does not degrade the performance and makes more easy to interpret and to predict the commands.

Some additional experiments are done to evaluate the importance of each feature to predict the commands. The VCART prediction method and the *L1* extraction method are selected to perform this analysis. Starting from the complete classifier, the feature that less influences the prediction results is deleted from the set of available features. The degradation of the classifiers indicates the importance of that feature. This procedure is iterated until only one feature remains.

With respect to the parameters of the command phrase, the T_0 value is not very critical because the extraction method forces the command to be near the beginning of the phrase. On the other hand, the amplitude A_p seems to be strongly correlated with the duration of the phrase and with the position of the phrase in the sentence: long phrases need larger amplitudes and the amplitudes decrease along the sentence. With respect to the accent command, most of the prediction power comes from the duration of the accent group, the number of remaining accent groups in the phrase and the amplitude of the last phrase command: we observe that the amplitudes of the accent commands decrease along the phrase; furthermore, if the phrase command amplitude gets a large value, then the accent commands require small amplitudes to produce natural contours.

4. Conclusions

In this work the parameters of the Fujisaki's intonation model are derived automatically from data. Accent commands are related to the accent group. This intonation unit has offered good results in several works on Spanish intonation. Some preliminary experiments show that the F_0 contours are better predicted if the number of accent commands in each accent group is limited to one.

A new method is proposed to extract the parameters of the Fujisaki intonation model. The method makes an initial guess assuming that there is only one accent command for each accent group and one phrase command at the beginning of each prosodic phrase. Afterwards, the locally optimal parameters are found using the gradient method. This method is compared with several methods derived from recent proposals [2, 7, 3]. The results show that there is a trade-off between number of commands and accuracy of the approximation and the new method gives a very good compromise. Furthermore, we tried to use the other methods in our prediction strategy (with the accent group as intonation unit) but we did not get good results. making the new method the only option.

Several prediction algorithms were evaluated, including VCART (vector clustering using classification trees). This technique avoids some ill effects of independent prediction of each model parameters. The results of the experiments show that this method is better than building one regression tree for each parameter or using neural networks.

Many features were considered to predict the parameters of Fujisaki's intonation model. Afterwards, the influence of each feature was evaluated. We conclude that a small number of features gives most of the prediction power. The result of the analysis is in accordance to the authors experience with the model.

The predicted contours offer very good quality in our TTS system. However, it should be noted that all the experiments were done using a read-style corpus of declarative sentences. Further work needs to be done to validate the use of only one accent command for each accent group in other types of sentences and other speech styles.

5. References

[1] H. Fujisaki and K. Hirose, 1984, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," in *Journal of the Acoustical Society of Japan(E)*, vol. 5, 233–242.

[2] H. Mixdorff, 2000, "A novel approach to the fully automatic extraction of fujisaki model parameters," in *Proceedings of ICASSP*, vol. 3, Istanbul, Turkey, 1281–1284.

[3] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, 2002, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proceedings of ICASSP*, Orlando, Florida, 509–512.

[4] B. Möbius, M. Pätzold, and W. Hess, October 1993, "Analysis and synthesis of german f0 contours by means of fujisaki's model," *Speech Communication*, vol. 13, no. 1–2, pp. 53–62.

[5] E. Navas, I. Hernaez, and J.M. Sanchez, 2002, "Basque intonation modelling for text to speech conversion," in *Proceedings of ICSLP*, Denver, Colorado, USA.

[6] H. Mixdorff and O. Jokisch, 2001, "Implementing and evaluating an integrated approach to modeling german prosody," in *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland, 211–216.

[7] H. Fujisaki, S. Narusawa, and M. Maruno, 2000, "Preprocessing of fundamental frequency contours of speech for automatic parameter extraction," in *Proceedings of ICSP*, Beijing, China, 722–725.

[8] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, 2002, "Automatic extraction of model parameters from fundamental frequency contours of english utterances," in *Proceedings of ICSLP*, Denver, Colorado, USA, 1725–1728.

[9] David Escudero, Valentín Cardeñoso, and Antonio Bonafonte, 2002, "Corpus based extraction of quantitative prosodic parameters of stress group in spanish," in *Proceedings of ICASSP*, Orlando, Florida, USA, 481–484.