Joint extraction and prediction of Fujisaki's intonation model parameters

Pablo Daniel Agüero, Klaus Wimmer, Antonio Bonafonte

Speech Processing Group
Department of Signal Theory and Communications
Polytechnical University of Catalonia, Barcelona, Spain
pdaguero@gps.tsc.upc.es

Abstract

This paper presents a joint extraction and prediction framework for intonation modeling applied to Fujisaki's intonation model for text-to-speech conversion. Previous methods in the area extract the parameters of accent and phrase commands for each sentence. Then, these parameters are related to linguistic features for prediction. In our approach commands that share the same linguistic features are globally estimated. This approach intends to overcome some consistency problems of the extracted model parameters. The global nature of the parameter optimization avoids the interpolation step, which sometimes can produce a bias in the extracted parameters. Experimental results show that the higher consistency of the parameters result in a higher accuracy when the fundamental frequency contours are predicted.

1. Introduction

The intonation model is a key component in text-to-speech conversion, because it provides information that increases the intelligibility and naturalness.

Several intonation models use machine learning techniques to predict fundamental frequency contours from linguistic information. This approach has the advantage of a shorter development and adaptation time to a new domain than rule-based approaches.

In order to provide a compact parametric representation of the fundamental frequency contour suitable for machine learning techniques, some methods [2, 3, 7, 11] rely on parametric models that provide the neccesary flexibility to approximate the fundamental frequency contours. A previous step to parameterization is filtering (microprosody, boundary effects and noise removal) and interpolation of unvoiced segments.

After the parameterization step, machine learning techniques are used to map linguistic features onto sets of parameters.

Prediction accuracy depends on several factors:

- **Consistency of the parameterization**. In some intonation models multiple possible sets of parameters can provide good approximations to the fundamental frequency contour. This makes the prediction task more difficult.
- **Filtering and interpolation**. The steps of filtering and interpolation are critical. In some cases they generate contours that are not suitable for parameterization. As a consequence, the extracted parameters are biased.
- **Relationship between parameterization and linguistic information**. As pointed out in [1, 9], it is useful to relate the parameterization to linguistic information to improve the prediction accuracy.

This paper proposes an algorithm to overcome the previously mentioned problems. This algorithm is applied to Fujisaki's intonation model.

Fujisaki's intonation model has a physiological basis [5]. The logarithm of the fundamental frequency contour is modeled superposing the output of two second order critically damped filters and a constant base frequency. One filter is excited with deltas (phrase commands), and the other with pulses (accent commands). The phrase commands are related to the slow varying component of intonation, and accent commands are related to fast changes.

Several parameter extraction algorithms have been proposed to obtain Fujisaki's intonation model parameters from a given fundamental frequency contour. Some methods [4, 8] rely on an initial stylization of the fundamental contour. The number of commands and their values are highly influenced by the stylization. As a consequence, the parameters are not necessarely related to physiological phenomena. Furthermore, the commands may not easily be related to the linguistic content of the utterance. Some methods tackle this problem imposing linguistic constraints during the parameter extraction step [1, 6, 9, 10].

In a previous paper [1], we presented the basic features of a two-stage approach. First, command parameters were extracted using strong linguistic constraints. Next, classification trees were trained to predict these parameters from linguistic features derived from text. In section 2 we present some modifications of this algorithm that improved the results of previous experiments.

In section 3 we propose a joint parameter extraction and prediction of Fujisaki's intonation model parameters using global optimization of parameters. The aim is overcoming the problems of inconsistency due to initial stylization and sentence-by-sentence extraction of two-stage approaches.

Section 4 shows experimental results for both approaches, using objective measures and perceptual tests. Section 5 presents the conclusions.

2. Two-stage method

Previously, we used the classical two-stage approach for intonation modeling [1]. In the first stage, command parameters are extracted sentence by sentence yielding optimal parameters for each contour of the training set. Next, the resulting parameters are used to train classification trees based on vector clustering.

The main characteristic of the extraction procedure is the application of strong linguistic constraints:

- Each minor phrase is modeled by one phrase command. The phrase command can only appear within a window

of 200 ms centered at the beginning of the minor phrase.

- The number of accent commands inside an accent group is limited to one.

Each command is represented as one vector of parameters: $[A_p,T_0]$ for phrase comands and $[A_a,T_1,T_2]$ for accent commands. In the command prediction stage, a clustering of command parameter vectors is performed using regression trees. One of the trees is related to accent groups (accent commands) and the other to minor phrases (phrase commands). The questions of these trees are related to the linguistic features of accent groups and minor phrases. The centroids of the clusters are the parameter vectors that minimize the mean distance to the other vectors of the cluster. In this way, possible deformations due to individual prediction of each command parameter is avoided.

Some observations led to improvements of the extraction procedure presented in [1]:

- The default value $\alpha = 3.0$ for the phrase control mechanism was not optimal. In several experiments the global optimum value was found to be $\alpha = 1.8$ for the whole corpus.
- The window for accent command timing extends 50 ms the accent group boundaries. This extension overcomes the problems of F0 peaks that extend their influence outside the accent group.

Concerning command prediction, we found it is advantageous to predict accent command onset and offset, rather than to predict onset and duration. Compared to the results presented in [1], these modifications resulted in better fitting accuracy as well as in better prediction performance. The synthesis capabilities of this method will serve as baseline for the evaluation of the new combined extraction and prediction algorithm.

3. Joint parameter extraction and prediction algorithm

In this section we propose a novel algorithm that performs a joint parameter extraction and prediction of Fujisaki's intonation model parameters.

As in the previous method, two regression trees are grown using linguistic features as questions in the nodes. One tree is related to minor phrases, and the other to accent groups. As before, we assume that each minor phrase is modeled by one phrase command, and each accent group is modeled by one accent command.

Thus, each leaf of the tree collects a set of fundamental frequency contours from the training corpus that must be approximated with command responses. A hill-climbing procedure is used to find the parameters that provide a global optimal approximation to all the fundamental frequency contours.

Due to the superpositional nature of the intonation model, each partition of one tree affects the optimal solutions of the parameters of the other tree. Therefore, the optimization must be jointly performed for phrase and accent commands.

The steps of the algorithm are:

1. Each tree (accent group tree and minor phrase tree) has an initial root node, which groups all the contours. An initial optimal solution is found that approximates all contours with the same phrase command for each minor phrase, and with the same accent command for each accent group.
2. All possible questions are examined in the leafs. For each question, the optimal parameters for phrase and accent commands are determined, and the approximation error is obtained. The optimization is performed using a hill-climbing algorithm.
3. The splitting questions for phrase and accent command trees are chosen. The selection criterion of the optimal node question is the minimization of the approximation error.
4. Then, the global optimal values for α, β and F_b are searched using a grid of values.
5. The process is iterated from the second step, until a minimum number of elements in the leafs is reached or the differential gain on accuracy is lower than a threshold.

The global optimization avoids the interpolation step of the stylization process, which can cause a bias in the parameter extraction. Another advantage of global optimization is the consistency of the parameters. Non-consistent parameters increase the dispersion, and limit the prediction capabilities of machine learning techniques.

Classification and regression trees are chosen because they allow the use of discrete and continuous features. Furthermore, the representation provides useful information to increase the knowledge about the task. This information can be used for future improvements of the system.

4. Experimental results

The experiments were performed for a female voice of a Spanish corpus of 500 sentences. The utterances were manually segmented in demiphones, and the fundamental frequency contour was obtained from the laryngograph channel. The train set for the experiments was 70% of the corpus and the test set was 30% of the corpus.

4.1. Previous intonation models

In addition to the two intonation models presented in this work, the UPC text-to-speech system has another two intonation models:

- **Piece-wise linear intonation model**: This intonation model is rule-based. The fundamental frequency contour is represented by a set of straight lines, and accents are modeled with peaks.
- **Intonation model using superposition of Bézier curves**: The joint parameter extraction and prediction approach has been applied to model fundamental frequency contours using Bézier curves. Accent groups and minor phrases are modeled using a superposition of two components. Each component is modeled using Bézier curves. The global optimization of the parameters enables the separation of the effects of each component.

4.2. Objective evaluation

The global results for all approaches are shown in Table 1. The joint parameter extraction and prediction algorithm has a better performance with respect to the RMSE and correlation measures than the two-stage approach. The higher consistency is the result of the global optimization procedure. The intonation model based on Bézier curves has the best performance. This is due to the flexibility of the Bézier components to model fundamental frequency contours more accurately.

Method	**RMSE** [Hz]	ρ
Piece-wise linear	20.46	0.5874
Superpositional Bézier	18.08	0.7509
Two-stages approach	21.79	0.6820
Joint extraction and prediction	18.67	0.7315

Table 1: *Global results.*

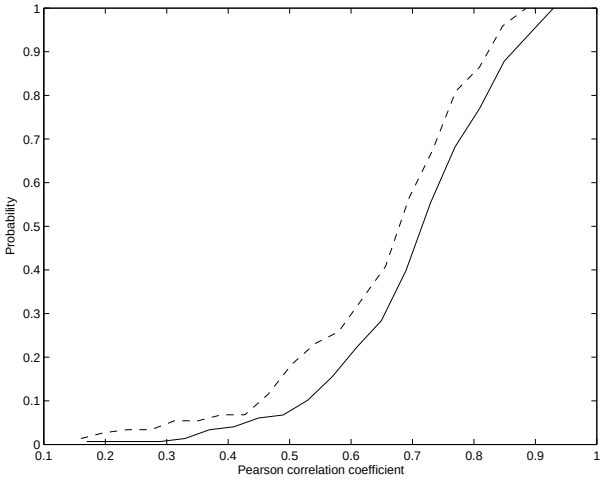

Figure 1: *Correlation coefficient cumulative density function (dotted line: two-stage aproach, solid line: joint parameter extraction and prediction approach).*

Figures 1 and 2 show the cumulative probability density function of the root mean squared error (RMSE) and Pearson correlation coefficient (ρ). With these graphics it is possible to measure the quality of the prediction in terms of dispersion. The joint extraction and prediction approach shows a higher probability of having higher correlation and lower RMSE than the two-stage approach. For instance, the probability of a F0 contour with $RMSE < 20Hz$ is 0.75 for the joint approach, which is significantly higher than the probability of 0.5 of the two-stage approach.

4.3. Subjective evaluation

In order to obtain a perceptual measure of the quality of the intonation for each model, we performed a listening test. The number of evaluators was 19. They were asked to judge the naturalness of several sentences using a five point scale (1:unnatural, 5:natural).

Figure 3 shows the results of perceptual evaluation of naturalness for all methods. A natural intonation is included in the test to measure the quality of the evaluators.

The results do not show a clear preference of the proposed method over our previous two-stage approach. The two-stage approach has a higher dispersion in the scores than the joint extraction and prediction methods.

This evaluation shows that the accuracy gain in the RMSE and correlation coefficient measures is not high enough to show a subjective improvement using joint approaches. In addition, these measures are not enough to measure the quality of a fundamental frequency contour, because some perceptual local effects are not taken into account.

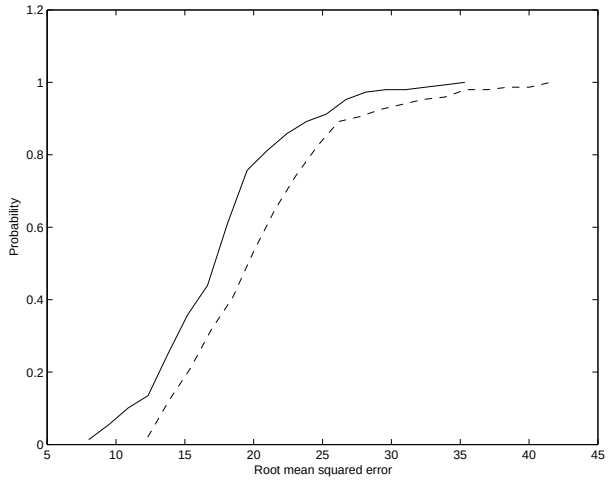

Figure 2: *RMSE cumulative density function (dotted line: two-stage aproach, solid line: joint parameter extraction and prediction approach).*

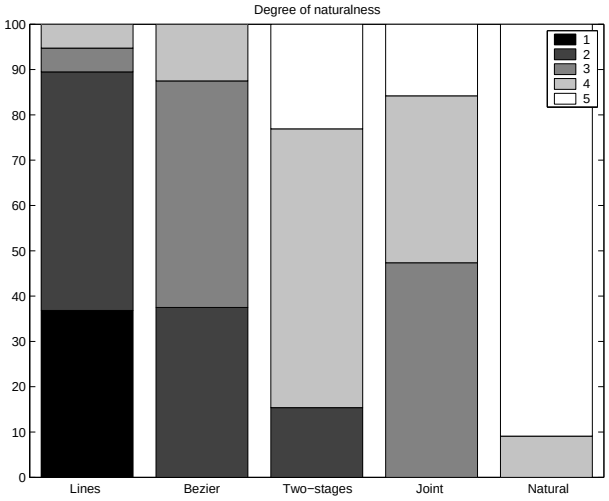

Figure 3: *Perceptual evaluation. (1:Unnatural, 5:Natural)*

5. Conclusions

This paper presents a novel approach for parameter extraction and prediction of intonation models, using a joint optimization framework. This approach is successfully applied to Fujisaki's intonation model.

The higher consistency of the parameters is due to the global optimization of the parameters. In this way, manipulations of the fundamental frequency contour for interpolation of unvoiced segments is avoided. The consistency of the extracted parameters decrease the dispersion of their values. This is important to increase the performance of machine learning techniques.

The experiments support the theoretical advantages of this approach, outperforming our previous approach for intonation modeling using Fujisaki's model. The system has a higher prediction accuracy and an inferior dispersion of RMSE and Pearson correlation coefficient in the cumulative density functions.

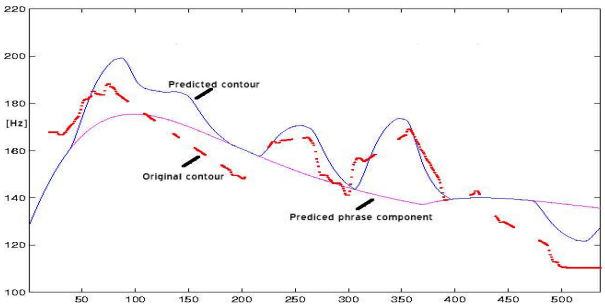

Figure 4: *Predicted fundamental frequency contour.*

However, perceptual experiments do not show a preference for the joint parameter extraction and prediction method over the other methods. This is due to two reasons: the RMSE is high and the correlation is low to obtain natural contours with any evaluated method of this paper, and the objective similarity measures lack from a psychoacustic basis. Further work will be dedicated to overcome these two problems.

6. Acknowledgements

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado).

7. References

[1] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Automatic analysis and synthesis of Fujisaki's intonation model for TTS," *Speech Prosody*, pp. 427–430, 2004.

[2] D. Escudero, V. Cardeñoso, and A. Bonafonte, "Experimental evaluation of the relevance of prosodic features in Spanish using machine learning techniques," *Proceedings of Eurospeech*, pp. 2309–2312, 2003.

[3] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of Acoustics Society of Japan*, pp. 233–242, 1984.

[4] H. Fujisaki, S. Narusawa, and M. Maruno, "Preprocessing of fundamental frequency contours of speech for automatic parameter extraction," *Proceedings of ICSLP*, pp. 722–725, 2000.

[5] H. Fujisaki, S. Ohno, and S. Narusawa, "Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common Japanese and the standard Chinese," *Proceedings of the 5th Seminar on Speech Production*, pp. 145–148, 2000.

[6] K. Hirose, Y. Furuyama, S. Narusawa, and N. Minematsu, "Use of linguistic information for automatic extraction of F0 contour generation process model parameters."

[7] D. Hirst, N. Ide, and J. Veronis, "Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project," *Proceedings of 2nd ESCA/IEEE Workshop on Intonation*, 1994.

[8] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proceedings of ICASSP*, pp. 1281–1284, 2000.

[9] B. Möebius, "Components of a quantitative model of German intonation," *Proceedings of ICPhS 95*, pp. 108–115, 1995.

[10] E. Navas, I. Hernaez, and J. Sanchez, "Basque intonation modelling for text to speech conversion," *Proceedings of ICSLP*, pp. 2409–2412, 2002.

[11] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of Acoustical Society of America*, pp. 1697–1714, 2000.