INTONATION MODELING FOR TTS USING A JOINT EXTRACTION AND PREDICTION APPROACH

Pablo Daniel Agüero and Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Campus Nord D-5, C/Jordi Girona 1-3, 08034 Barcelona - Spain
pdaguero@gps.tsc.upc.es, antonio.bonafonte@upc.es

ABSTRACT

The intonation model is a key component in text-to-speech synthesis to improve 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 new domains. In order to provide a compact parametric representation of the fundamental frequency contour suitable for machine learning techniques, some methods rely on parametric models that provide the neccesary flexibility to approximate the fundamental frequency contours. After the parameterization step, machine learning techniques are used to map linguistic features onto sets of linguistic parameters. Prediction accuracy depends on several factors:

- In some intonation models multiple possible sets of parameters can provide good approximations to the fundamental frequency contour (inconsistency).
- In some cases filtering and interpolation of unvoiced regions generate contours that are not suitable for parameterization and the extracted parameters are biased.

This paper proposes an approach to overcome the previously mentioned problems. This approach performs a joint parameter extraction and prediction of intonation contours with global optimization of parameters. It has been successfully applied to Bèzier and Fujisaki parameterisations.

1. INTRODUCTION

Intonation modelling is an important task in a text-to-speech system, affecting intelligibity and naturalness.

Several models are proposed in the literature, such as ToBI [1], Fujisaki [2], IPO [3], Tilt [4], PaIntE [5], INTSINT [6] and Bézier [7]. They try to overcome with the difficulties of intonation modeling:

- The fundamental frequency contour depends on the choice of speaker. As a consequence, there is a mapping of one-to-many from a sentence to the possible fundamental frequency contour space. Furthermore, it is difficult to measure the accuracy of a system and the modeling task.
- The information in a text is not enough to perform natural intonation. Other sources of information should be used (semantic and pragmatic information). Valuable information can be provided by part-of-speech taggers and syntactic analyzers.
- The fundamental frequency contour extracted from an utterance has measurement errors and microprosody. Filters are applied to remove these effects, but some important information can also be lost.

In general, intonation models trained automatically perform a two step process.

The first step consists of a smoothing process, with interpolation of unvoiced regions. The smoothing process removes measurement errors and microprosody. The interpolation is performed because some models need a continuous fundamental frequency contour to extract parameters. These parameters represent fundamental frequency contour in a compact way, suitable for prediction.

In the second step machine learning techniques are used to relate the parameters to linguistic information. The output of this step is a model that predicts fundamental frequency contours from linguistic information.

This two step process can have unexpected consequences in the final prediction accuracy, because there is a strong interaction between the two steps.

The Fujisaki intonation model [2] represents the fundamental frequency contour as the linear combination of two

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

components, that are associated with minor phrase and accent group. The initial extraction of the parameters (amplitudes and time instants of phrase and accent commands) is a complex task. There are several sets of parameters that can represent the same fundamental frequency contour. Furthermore, the shape of the stylized contour depends on the smoothing and interpolation parameters. As a result, the second step can be provided with a non-consistent set of parameters which are different for the same linguistic information. The final result is poorer prediction accuracy. Some researchers avoid this effect by applying additional knowledge in the extraction process [8] [9].

Although previously proposed intonation models performed successfully with the two-step process, a joint framework could improve the results of these methods.

Our proposal is a joint parameter extraction and prediction framework to train an intonation model.

The idea is to extract a set of parameters that closely represent the fundamental frequency contour taking into account linguistic features. This approach allows a consistent representation to increase prediction capabilities. A global optimization algorithm which takes into account all training data must be used to obtain the optimal parameters.

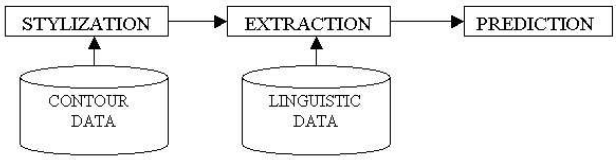

Fig. 1. Classical scheme where extraction and prediction are two separate steps.

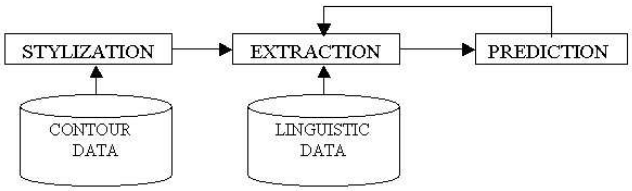

Fig. 2. Proposed scheme where extraction and prediction steps interact.

In this paper this joint process will be explored using a superpositional approach. There is evidence of the existence of differents effects that are linearly combined to produce the final fundamental frequency contour in the literature [2] [10]. In this work the components that are linearly combined are related to minor phrase and accent group.

The fundamental frequency contour is represented using two intonation models: Bézier curves (Bézier intonation model) [11], and accent and phrase commands (Fujisaki's intonation model) [12]. The polynomial representation has been used to model intonation in Spanish [7]. Few coefficients allow a very accurate representation of each component of the fundamental frequency contour.

The next two sections give a detailed explanation of our approach, with a brief explanation of the mathematical models and training process. Section 4 shows experimental results of this approach compared with a two-step approach. Section 5 gives the conclusions about this work.

2. BÉZIER INTONATION MODEL

2.1. Mathematical model

The joint optimization framework imposes that the formulation to extract the optimal polynomial coefficients is modified. The optimization is performed minimizing the mean squared error, but taking into account that:

- The error that is minimized is the global mean squared error.
- Two components are combined using Bézier curves.
- The group of coefficients corresponding to a Bézier curve depend on a vector which maps minor phrase or accent group classes with positive integers (class number).

The mathematical formulation is shown in equation 1.

$$F_0^k(t) = \sum_i^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \quad (1)$$

where:

N_{MP}^k is the number of minor phrases of the kth sentence.

N_{AG}^k is the number of accent groups of the kth sentence.

$t_{MP_i}^k(t)$ is the temporal axis of the ith minor phrase of the kth sentence.

$t_{AG_j}^k(t)$ is the temporal axis of the jth accent group of the kth sentence.

$C_{MP_i}^k$ is the number of the minor phrase class assigned to the ith minor phrase of the kth sentence.

$C_{AG_j}^k$ is the number of the accent group class assigned to the jth accent group of the kth sentence.

In this function, P_{MP} and P_{AG} are the Bézier curves of the minor phrase and accent group components, respectively. Each curve has its own associated time axis, $t_{MP}(t)$ and $t_{AG}(t)$. The time axis range is zero to one. These curves are zero elsewhere.

The joint cost function is shown in equation 2. The goal is to minimize the mean squared error. This equation has a unique analytical minimum that is found using a set of linear equations.

$$e = \sum_{k}^{N_s} \left(\sum_{t}^{T_k} \left(f0^k(t) - \left(\sum_{i}^{N^k_{MP}} P_{MP}^{C^k_{MP_i}}(t^k_{MP_i}(t)) + \sum_{j}^{N^k_{AG}} P_{AG}^{C^k_{AG_j}}(t^k_{AG_j}(t)) \right) \right)^2 \right) \quad (2)$$

where:

N_s is the number of sentences.
T_k is the duration of the sentence.

2.2. Model training process

The idea behind the training process is to find a set of minor phrase and accent group clusters (obtained using linguistic information) that are optimal in the sense of mean squared error and Pearson correlation coefficient.

Mean squared error and Pearson correlation coefficient are chosen as the optimization indexes because there is a common consensus on intonation modelling about using them to measure the prediction accuracy.

There are many ways to perform a clustering based on a set of parameters. Classification and regression trees are chosen, because of the capabilities to classify using continuous and discrete features. The information provided by the final tree can be valuable for future improvements or to get an insight of the main features related to the problem.

Because of the superpositional approach, two independent trees are trained (accent group component tree and minor phrase component tree), with a joint optimization cost (Pearson correlation coefficient).

Initially, each tree has a unique root node. As a consequence, there is only one minor phrase and accent group class.

The steps performed to grow the trees are:

- Consider each possible splitting for each tree, according to linguistic parameters extracted from text.
- Find the optimal polynomial coefficents (α's and β's associated to minor phrases and accent groups) for each splitting.
- Select the split which maximizes the Pearson correlation coefficient.

The trees are grown until the Pearson correlation coefficient gain is less than a predefined threshold. The number of elements in each leaf is bounded to be superior than a predefined threshold (in our experiments, this threshold is 40), in order to prevent a weak modeling of cluster due to small data size.

3. FUJISAKI'S INTONATION MODEL

The Fujisaki's intonation model has been widely described in the literature (e.g.: Fujisaki et al [2]). We will not dedicate time to the description of this model.

The proposed framework is applied to Fujisaki's intonation model replacing the polynomial representation of Bézier intonation model by accent and phrase commands. Phrase commands are related to minor phrases, and accent commands are related to accent groups. There is only one command assigned to minor phrases and accent groups.

The main problem of Fujisaki's intonation model is that it does not have a closed-form solution. As a consequence, it is neccesary to perform iterations in order to obtain an optimal solution. The Fujisaki's intonation model is more time consuming in the training process.

4. EXPERIMENTS

4.1. Experimental conditions

The corpus used in this work are CMU-ARCTIC (recently delivered by Carnegie Mellon University, and can be downloaded from the FestVox site [13]) and ESMA (Universitat Politècnica de Catalunya corpus).

The experiments consisted in intonation modeling of:

- Two speakers of ARCTIC database (BDL and STL) (American English).
- Female speaker of database ESMA (Spanish).

Data is divided in training (70%) and testing sets (30%).

The results are analyzed using mean squared error and Pearson correlation coefficient measures. Mean squared error measures the distance, and Pearson correlation coefficient measures the shape similarity.

4.2. Experimental results

Tables 1 and 2 show the global RMSE and Pearson correlation coefficient. The first table shows results using Bézier intonation model, and the second shows results using Fujisaki's intonation model.

In both cases, the joint extraction and prediction approach outperforms the two-stage approach. It supports the idea that this framework can be applied to other intonation models as well.

Corpus and experimental conditions	RMSE	ρ
BDL (two-stage approach)	13.6	0.55
BDL (joint parameter extraction and prediction)	13.1	0.59
STL (two-stage approach)	15.1	0.68
STL (joint parameter extraction and prediction)	14.4	0.71

Table 1. RMSE and Pearson correlation coefficient of the predicted contours for each training procedure (Bézier intonation model)

Corpus and experimental conditions	RMSE	ρ
ESMA (two-stage approach)	21.79	0.6820
ESMA (joint extraction and prediction)	18.67	0.7315

Table 2. RMSE and Pearson correlation coefficient of the predicted contours for each training procedure (Fujisaki's intonation model)

The RMSE and Pearson correlation coefficient values are superior in the case of female speaker in the ARCTIC corpus. It is due to the higher pitch range, that increases the RMSE, but improves the ρ because of a higher influence of the phrase component, which can be easily predicted.

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 and Bézier 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 two-stage approach for intonation modeling.

6. REFERENCES

[1] K. Silverman, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling english prosody," *Proceedings of ICSLP92*, 1992.

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

[3] J. Hart, R. Collier, and A. Cohen, "A perceptual study of intonation. An experimental approach to speech melody," *Cambridge University Press*, 1990.

[4] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of Acoustical Society of America*, 2000.

[5] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," *3rd ESCA Workshop on Speech Synthesis*, 1998.

[6] D.J. 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*, 1994.

[7] D. Escudero and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish," *ICASSP 2002*, 2002.

[8] B. Möebius, "Components of a quantitative model of German intonation," *Proceedings of ICPhS 95*, 1995.

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

[10] R. Sproat, "Multilingual Text-to-Speech Synthesis," *KLUWER academic publishers*, 1998.

[11] P. D. Agüero and A. Bonafonte, "Intonation modeling for tts using a joint extraction and prediction approach," *5th ISCA Speech Synthesis Workshop*, 2004.

[12] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Joint extraction and prediction of fujisaki's intonation model parameters," *accepted to be presented at ICSLP 2004*, 2004.

[13] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," http://festvox.org/cmu_arctic/.