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

This paper presents a joint extraction and prediction framework for intonation modeling. The intonation model is based on a superpositional approach using Bézier curves. The components are attached to minor phrase and accent group. A greedy algorithm performs succesive partitions on training data using linguistic information. The parameters related to each partition are obtained using a global optimization procedure. In this way, the extraction process is closely related to the prediction step, and the final performance is higher. Several experiments are performed to test the hypothesis using a two-step intonation modeling procedure for comparison. Results reveal that the prediction accuracy is higher than the reference method. This approach avoids some parameter extraction steps that can produce additional noise, such as the interpolation step used in some intonation models.

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 smoothed.

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 with 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. For TTS it is not useful to preserve in the stylization step information that is not possible to predict from the text.

Fujisaki intonation model [2] represents the fundamental frequency contour as the linear combination of two components, that are associated to 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 a poorer prediction accuracy. Some researchers avoid

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

this effect applying additional knowledge in the extraction process [8] [9].

Tilt [4], Bezier [7] and INTSINT [6] perform successful intonation modeling with the two-step process. However, 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 to obtain the optimal parameters must be used.

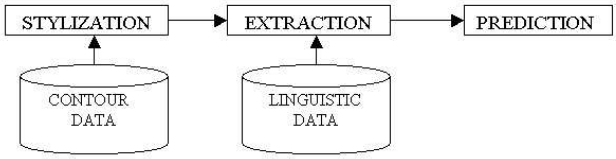

Fig. 1. Classical scheme where extraction and prediction are two separate steps.

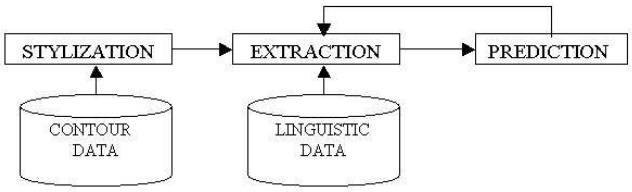

Fig. 2. Proposed scheme where extraction and prediction steps interact.

In this paper this joint process will be explored using a superpositional approach. There are proofs 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 Bézier curves. This 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.

Next section gives a detailed explanation of our approach, with a brief explanation of the mathematical model and training process. Section 3 shows experimental results of this approach, compared with a two-step approach based on Bézier curves. Section 4 gives the conclusions about this work.

2. INTONATION MODEL

2.1. Bézier curves

The parameter representation of the model is based on Bézier curves. The polynomial formulation is shown in equation 1 and the shape of the base polynomials for a fourth order curve are shown in Figure 3. Bézier coefficients allow a meaningful representation compared with the final polynomial coefficients, which are more sensitive.

$$P(t) = \sum_{n}^{N} \alpha_n \begin{pmatrix} N \\ n \end{pmatrix} t^n \, (1-t)^{(N-n)} \qquad (1)$$

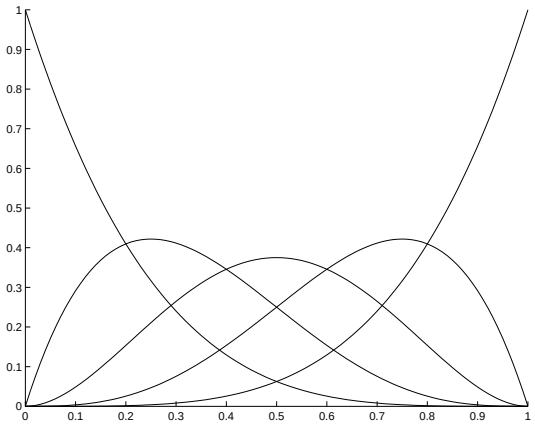

Fig. 3. Bézier polynomials

Figure 4 shows an approximation of a fundamental frequency contour using Bézier curves for accent groups, with continuity constraints up to the first derivative.

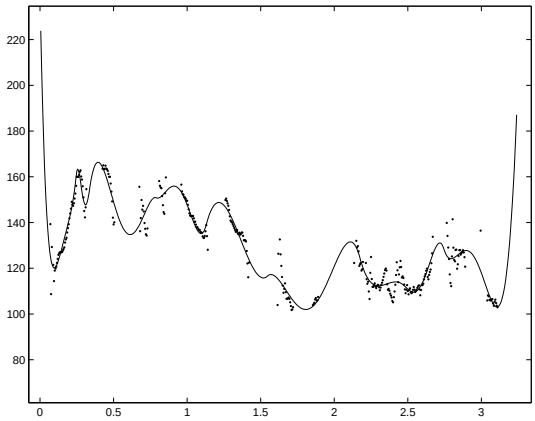

Fig. 4. Fundamental frequency contour approximated using Bézier curves with five coefficients

2.2. Reference system

The joint parameter extraction and prediction framework is compared with other approach, proposed by Escudero [11].

This approach performs an initial parameter extraction using Bézier curves linked to accent groups. Continuity constraints are applied up to the first derivative.

Escudero proposes a prediction method using vector quantization and classification trees.

In this work, Bézier coefficients are predicted using a regression tree which performs a clustering based on linguistic features related to each accent group. The splitting criteria is the reduction of the mean distance to the centroid.

2.3. 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 2.

$$F_0^k(t) = \sum_i^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \quad (2)$$

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 3. 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.

$$e = \sum_k^{N_s} \left(\sum_t^{T_k} \left(f0^k(t) - \left(\sum_i^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \right) \right)^2 \right) \quad (3)$$

where:

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

2.4. 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 [12] 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.

The linguistic features used to predict minor phrases are: sentence type (declarative, interrogative or exclamative), number of minor phrase in the sentence, position of the minor phrase in the sentence, number of accent groups in the minor phrase, number of words in the minor phrase and number of syllables in the minor phrase.

The linguistic features used to predict accent groups are: sentence type (declarative, interrogative or exclamative), number of minor phrase in the sentence, position of the minor phrase in the sentence, number of accent groups in the minor phrase, number of words in the minor phrase, number of syllables in the minor phrase, number of following accent groups, number of accent groups in the sentence, number of syllables in the accent group and position of the accent group in the minor phrase.

The joint optimization approach has a drawback that does not allow the definition of continuity constraints, because of the global nature of the problem.

In the case of English, accent groups are defined as a sub-sequence of the sequence of syllables contained in a minor phrase, such that the first syllable is accented and the remaining syllables - if any - not accented [10]. As a consequence, discontinuities in the fundamental frequency contour can be produced in an accent group boundary inside a word.

This problem is overcome using a smoothing function in the boundaries of accent groups. This smoothing function performs a linear interpolation in the middle of the discontinuity.

3. EXPERIMENTS

3.1. Experimental conditions

The corpus used in this work is CMU-ARCTIC. This corpus has been recently delivered by Carnegie Mellon University, and can be downloaded from the FestVox site [13].

Pitchmarks are extracted from EGG signal using our own pitch mark extraction application. The fundamental frequency contours are analyzed to remove spurious points and microprosody.

For this work, the corpus has been hand-labelled with minor phrases.

The experiments consisted in intonation modeling of two speakers of ARCTIC database (BDL and STL). 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.

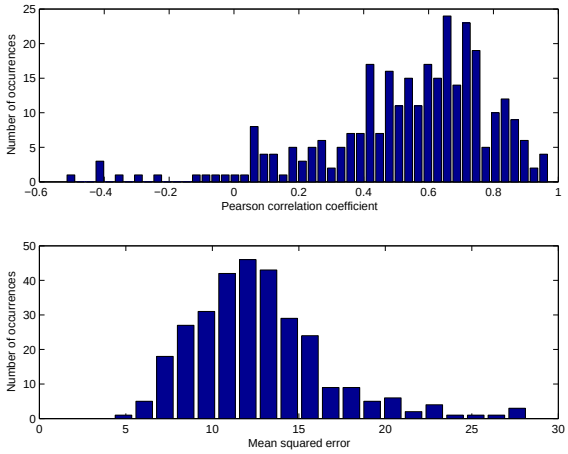

Fig. 5. Histogram of the distribution of Pearson correlation coefficients and RMSE for speaker BDL using smoothing

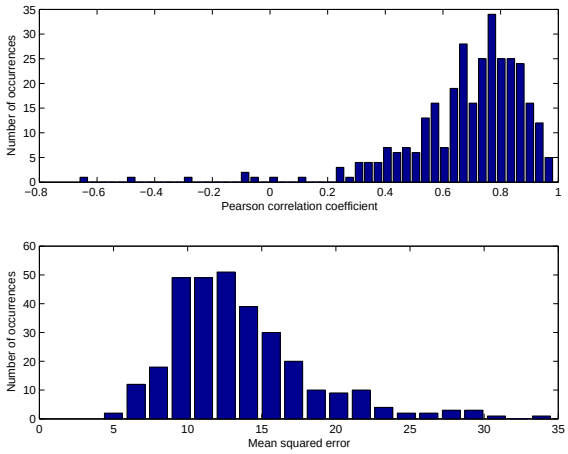

Fig. 6. Histogram of the distribution of Pearson correlation coefficients and RMSE for speaker STL using smoothing

3.2. Experimental results

Results of the different approaches are shown in Table 1. The joint parameter extraction and prediction framework has a superior performance both in MSE and Pearson correlation coefficient than the reference system. Local parameter extraction can cause inconsistent contours for a similar linguistic parameter set, which should have the same shape.

STL speaker has higher correlation than BDL speaker because the frequency range is higher. Then the minor phrase component contributes in a higher degree to the final correlation value. In general, the contribution of the accent group component to the final correlation is lower.

Corpus and experimental conditions	RMSE	ρ
BDL (joint parameter prediction)	13.612	0.555
BDL (joint parameter extraction and prediction)	13.181	0.593
STL (joint parameter prediction)	15.117	0.685
STL (joint parameter extraction and prediction)	14.403	0.716

Table 1. RMSE and Pearson correlation coefficient of the predicted contours for each training procedure

An alternative representation is proposed to analize the distribution of RMSE and Pearson correlation coefficients. The histograms are shown in Figures 5 and 6. The intonation model performs well on most contours. However, some of the contours have low correlation or high RMSE. It is due to several factors:

- Some linguistic or supra-linguistic features are missing (e.g.: focus, emphasis, etc.). The absence of these features causes more plain contours.
- Several utterances of the same sentence performed by the same person may be completely different. It is possible for a human to produce different contours for sentences with similar linguistic structure.
- Misalignments due to automatic phoneme segmentation. The automatic phoneme segmentation produces errors in the segmentation that cause a misalignment in the position of the points for polynomial stylization.
- Accents are extracted using a dictionary. It is a drawback in the experiment, because some accents are erroneously placed, and other are missing.
- Errors in the extraction of pitch marks. The quality of the pitch marks extracted from laryngograph signal is high, but some errors are still present.

Figures 7 and 8 show the degree of relationship between the value of Pearson correlation coefficient and mean squared error. There is a high correlation between these two parameters, which is a consequence of the joint optimization of mean squared error and Pearson correlation coefficient.

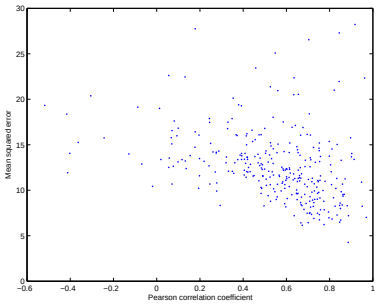

Fig. 7. Correlogram of RMSE and Pearson correlation coefficient for BDL corpus

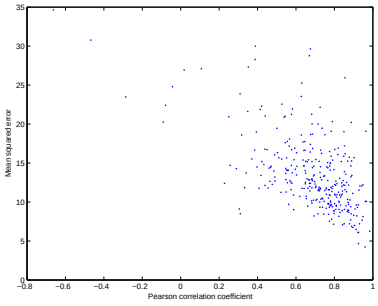

Fig. 8. Correlogram of RMSE and Pearson correlation coefficient for STL corpus

4. CONCLUSIONS

In this paper a new approach for intonation modeling is proposed using a joint extraction-prediction approach. This approach has been presented using a new intonation model. This model represents the contour as the superposition of minor phrase and accent group components. Each component is modelled using Bézier curves.

There are several theoretical advantages of this joint approach:

- The interpolation step is not necessary, and a possible source of errors due to this step is avoided. The set of parameters that are defined to perform this process can cause some ill effects in the contour, that are fit by LMSE methods.
- The effects due to fundamental frequency contour measurement errors and microprosody are reduced.
- The parameters extracted are globally consistent, and provide a higher accuracy in the prediction process.

However, these advantages are obtained with a higher complexity of the training process, with the consequences of a higher computational cost. The speed of the training algorithm is slow due to the number of matrices and its size, taking around four hours to train trees in a single processor machine with 1.2 GHz.

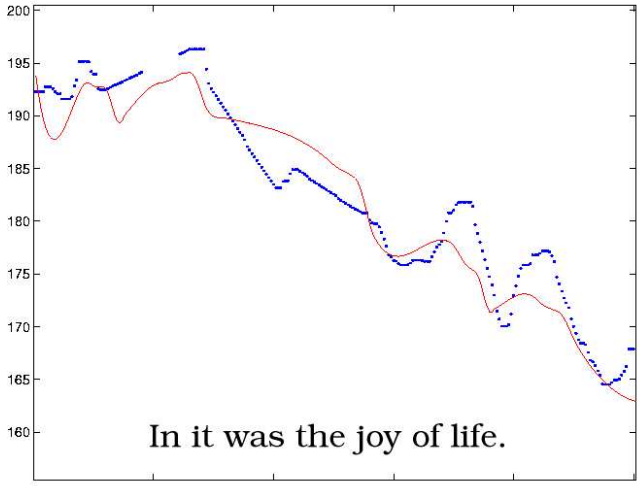

Fig. 9. Predicted fundamental frequency contour (lines) against original contour (dotted)

The comparison with a two-step approach under the same experimental conditions reveals that the joint optimization framework allows a higher prediction accuracy.

The final contours have small discontinuities in the neighborhood of accent groups. They are corrected using a smoothing function. This modification does not change the performance of the model.

5. 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] D. Escudero, V. Cardeñoso, and A. Bonafonte, "Experimental evaluation of the relevance of prosodic features in spanish using machine learning techniques," *Eurospeech 2003*, 2003.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," *Chapman Hall*, 1984.

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