Facing data scarcity using variable feature vector dimension

Pablo Daniel Agüero and Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain

Abstract

This paper focuses on three key points of intonation modelling: interpolation of fundamental frequency contour, sentence by sentence parameter extraction and data scarcity. In some cases, they introduce noise and inconsistency on training data reducing the performance of machine learning techniques.

We consider that the F0 contour is segmented into prosodic units (such as accent groups, minor phrases, etc). Each segment of F0 contour has a corresponding feature vector with linguistic and non-linguistic components.

We propose to face the limitations mentioned above using a technique based on clustering using different feature vector dimensions. The clustering of feature vectors produces also a partition in the F0 contour space. The proposal consists on a procedure to select the dimension that contributes to predict the best fundamental frequency contour from a RMSE sense compared to a reference contour. Experimental results show an improvement compared to other approaches.

1. Introduction

During last decade, text-to-speech systems have experienced a formidable quality improvement. The quality provided by the acoustic generation module with speech segments selected from a large speech database has been an important factor. In addition, the improvement of prosodic modules using data-driven models also contributed to such quality enhancement.

However, we still have no doubt whether a long speech fragment is synthetic or natural. We need further work to achieve the final goal of natural voices. Prosodic models play a fundamental role for such goal.

Several intonation models have been proposed in the literature, such as Fujisaki [1], Tilt [2], Bézier [3] and INTSINT [4]. In general the training of those models consists of two stages. First, a compact representation is obtained for each sentence, e.g.: Fujisaki commands or Tilt events (step 1). After that, machine learning techniques are used to infer a mapping from the linguistic features (available during speech synthesis) to the parameters (e.g.: Tilt parameters) (step 2). Such models are named in this paper as two-stage algorithms. In this way a set of linguistic features extracted from a given text are used to infer the corresponding set of parameters and predict its intonation contour.

The described two-stage approach presents some limitations that can cause training problems:

- **Interpolation of fundamental frequency contour.** An initial interpolation of F0 in the unvoiced regions is required. This interpolation is somehow arbitrary and may introduce *noise* in the extracted parameters: contours with same F0 contour in voiced parts may be represented by different parameters. This introduces dispersion in parameters reducing the accuracy of the machine learning techniques.
- **Sentence by sentence parameter extraction.** In some intonation models different sets of parameters can represent the F0 contour with the same accuracy. Sentence by sentence parameter extraction may produce that similar contours are represented by different parameters due to the multiple possible solutions (e.g: Fujisaki's intonation model). It increases the variance of the parameters and again reduces the accuracy of machine learning techniques.
- **Data scarcity**. Intonation events that occur with low frequency in training database are not correctly modelled. It is desirable to model events with high and low frequency in order *to capture* such *exceptions* with low frequency. In addition, it is necessary to take advantage as much as possible of the features we extract from text to predict the fundamental frequency contour. In some cases a combination of values of some features convert into irrelevant the rest of features. Such cases should also be modelled even if they have a low frequency in training data.

In previous papers [5, 6] we presented *JEMA: Join feature extraction and modeling approach*, a new approach to train intonation models that combines the parameter extraction (step 1) and model generation (step 2) into a single loop. This approach avoids requirements of continuity of fundamental frequency contours and increases the consistency of parameterization avoiding sentence-by-sentence parameter extraction.

In this paper we go a step forward. We present a training procedure that intends to overcome the previously mentioned limitations: interpolation of fundamental frequency contour, sentence-by-sentence parameter extraction and data scarcity. This goal is faced using an intonation model based on clustering of linguistic features combined with a selection of the dimension of the feature vector.

The paper is organised as follows. In Section 2 we describe the intonation model and in Section 3 we explain our proposal. Then, in Section 4 we show the experimental databases and results. Finally, in Section 5 we present conclusions and future work.

2. Intonation model

Our intonation model uses a superpositional approach combining the influence of two prosodic units: accent groups and minor phrases. Accent group models local effects at the level of the stressed syllable and minor phrase models a long-term effect of the intonation contour.

Each component is modelled considering that there is a limited number of pitch movements (classes). Each class corresponds to a pitch movement which has an approximation error with respect to the contours of the same class in the database.

In order to model both components we assign a class to each minor phrase (minor phrase class) and accent group (accent group class) by means of some criteria. Then the mathematical procedure explained in next section separates the effect of both components solving a set of linear equations. Representative contours for each class are optimal in the RMSE sense for its class.

2.1. Definition of pitch classes

In our previous papers the criteria to assign the class to each minor phrase and accent group was the minimization of the global root mean square error using CART trees. Each prosodic unit is composed of a feature vector and their associated contour. The tree performs a splitting of the F0 contour space by means of questions about the components of the feature vector generating classes. Then the optimization algorithm finds out an optimal contour for each class. Two trees are used to split the minor phrase and accent group space respectively.

In this paper we propose to assign classes to minor phrases and accent groups using a clustering algorithm based on a distance defined for feature vectors. An exhaustive analysis of training data rather than a greedy approach is used.

We work under the assumption that similar feature vectors have similar shapes. Therefore, if we perform a clustering in the feature vector space we are performing also a clustering in the F0 contour space. As a consequence, similar contours are in the same cluster.

The distance used for clustering for continuous and discrete features is defined as follows:

$$distance_{continuous}(V_a, V_b) = e^{-\beta|V_a - V_b|} \tag{1}$$

$$distance_{discrete}(V_a, V_b) = \begin{bmatrix} 1 & V_a = V_b \\ 0 & V_a \neq V_b \end{bmatrix} \tag{2}$$

The clustering is performed using the following steps:

1. **Initialization.** All feature vectors of the database belong to the same cluster. This initial cluster is chosen to be splitted.
2. **Assignation.** Feature vectors are assigned to the closest centroid.
3. **Centroid search.** The centroid of each cluster is the feature vector that is closer to all feature vectors of the same cluster. We avoid using a centroid that is the mean of the cluster.
4. **Iteration.** Step 3 and 4 are repeated until the centrois do not change or a maximum number of iterations are performed.
5. **Cluster selection.** The cluster with the higher number of elements is chosen to be splitted. We continue from step 2. If the selected cluster has a number of elements inferior to a threshold the clustering stops.

Each cluster is associated to a class. The optimization procedure explained in Section 2.2 is used to obtain the corresponding optimal pitch movement for each component (minor phrase and accent group) and each class (cluster).

2.2. Optimization procedure

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

$$P(t) = \sum_{n}^{N} \alpha_n \begin{pmatrix} N \\ n \end{pmatrix} t^n (1-t)^{(N-n)} \tag{3}$$

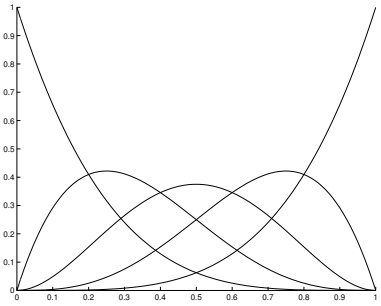

Figure 1: Bézier polynomials

The approximation using Bézier curves is performed minimizing the mean squared error taking into account that:

- The error that is minimized is the global mean squared error (global optimization).
- Two components are combined using Bézier curves (superpositional approach).
- 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 4.

$$\hat{f_0^k}(t) = \sum_{i}^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_{j}^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \tag{4}$$

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 5. 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.

$$e = \sum_{k}^{N_s} \left(\sum_{t}^{T_k} \left(f_0^k(t) - \hat{f_0^k}(t) \right) \right)^2 \quad (5)$$

where:

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

3. Multiple feature vector approach

The use of a distance based on the entire feature vector introduces some limitations.

Some feature vector combinations have a low frequency in the training database. Therefore, intonation events related with such features are bad modelled because of a lack of training data (as pointed out in the work of Escudero et al [7]). A possible solution for such cases is the use of a reduced feature vector which will enable a better modelling.

In addition, in some cases the use of the entire feature vector introduces noise in the measure because some features convert into irrelevant the rest of features for some values. It is desirable to use a reduced feature vector for such cases.

We propose the use of a procedure to select the optimal feature vector dimension based on the behaviour of similar feature vectors at the same dimension of training data. The feature vector can have any number of components. However, we will use only a few number of combinations obtained from a feature relevance analysis. In this way we avoid an explosion in the number of possible combinations for a given number of features, as shown in figure 2. In the future we will evaluate a higher number of combinations in order to avoid this strong assumption. Some combinations of features with the same dimension may be valuable.

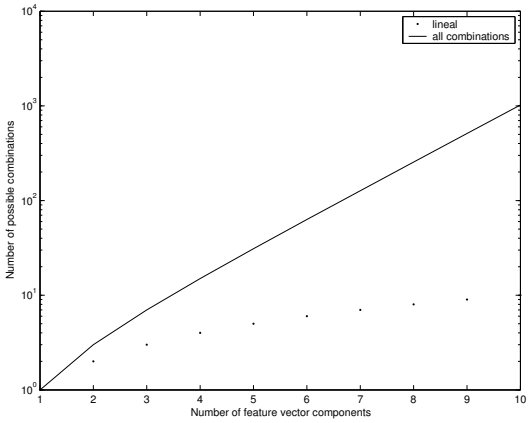

Figure 2: Number of combinations depending on the dimension of feature vector

3.1. Relevance of features

As a first step we need to sort the features by the modelling capabilities.

The relevance of features is analysed to test the generalization capabilities of each feature on the training data using Leave-One-Out.

At the beggining we analyse the most important feature using a feature vector of dimension one by modelling each sentence using the rest of the sentences of the training data (Leave-One-Out). The modelling of the sentence is performed using the clustering technique explained in section 2. We chose as the most relevant feature the one with the minimum approximation error. Then, the analysis is repeated increasing the dimension of the feature vector by one and keeping the most relevant feature detected in the previous iteration. In this way, we analise the incremental relevance of features.

Once all features have been assigned a relevance, we will have the approximation error for each dimension of input feature vector, as shown in table 1.

Feature vector	RMSE
$[F_4]$	$RMSE_{dim1}$
$[F_4, F_2]$	$RMSE_{dim2}$
$[F_4, F_2, F_3]$	$RMSE_{dim3}$
$[F_4, F_2, F_3, F_1]$	$RMSE_{dim4}$

Table 1: Relevance of features

In our approach we use a superpositional model. Therefore, it is necessary to analise the relevance of minor phrase and accent group features. This task is performed alternating minor phrase and accent group relevance analysis. In the first iteration we analyse minor phrase feature relevance. At that iteration all feature vectors of accent groups are assigned the same class because it has dimension zero.

3.2. Optimal feature vector

The procedure to find the optimal dimension of a given feature vector to predict its intonation contour is explained in this section.

We perform the assumption that given a certain dimension the k-nearest feature vectors of the training set have a similar approximation error. Therefore, the optimal dimension of a given feature vector is the dimension with the minimal approximation error for the k-nearest feature vectors of the training set.

In Table 2 we show an example of the procedure to find the optimal dimension of a given feature vector. The dimension with the minimum root mean squared error is the optimal considering the assumption we mentioned before.

Feature vector	N_1	N_2	N_3	N_4	**Total RMSE**
$[F_4]$	e^1_{d1}	e^2_{d1}	e^3_{d1}	e^4_{d1}	$RMSE_{d1}$
$[F_4, F_2]$	e^1_{d2}	e^2_{d2}	e^3_{d2}	e^4_{d2}	$RMSE_{d2}$
$[F_4, F_2, F_3]$	e^1_{d3}	e^2_{d3}	e^3_{d3}	e^4_{d3}	$RMSE_{d3}$
$[F_4, F_2, F_3, F_1]$	e^1_{d4}	e^2_{d4}	e^3_{d4}	e^4_{d4}	$RMSE_{d4}$

Table 2: In this table each column corresponds to the approximation error of the four closest feature vector (e_1, e_2, e_3 and e_4) for each dimension ($d1$, $d2$, $d3$ and $d4$). The last column corresponds to the root mean squared error of the values of the four previous columns.

4. Experiments

Two databases are used for the experiments. They are a parallel speech corpus of Spanish and Catalan from the hotel reservation domain. Two hundred sentences for each language were

produced in a recording room. Fundamental frequency contours were extracted using Praat [8]. The feature vectors for the intonation model were extracted from the text. An additional feature was extracted using a cross-lingual mapping of pitch movements as explained in [9].

4.1. Experimental results

Experiments consists on comparing the performance of the proposed approach with other baseline approaches:

- **CART (JEMA)**. The intonation model training procedure proposed in [5] was applied to Spanish and Catalan databases in order to provide a comparison with another machine learning technique.
- **CFV: Complete Feature Vector**. We perform experiments using the complete feature vector without the selection of the optimal dimension. It is considered a baseline of the proposed approach.
- **BDFV: Best Dimension for Feature Vector**. Experimental results using the optimal dimension of feature vector using the proposed approach.
- **BDFVopt: Best Dimension for Feature Vector optimal**. We also show experimental results using the optimal dimension of feature vector having privileged information about the approximation error of such selection. It is considered a ceiling result because it is the best way the system can perform.

The results are shown in Tables 3 and 4. In both tables the results of the proposed approach overcome the result of CART and CFV approaches for the Catalan and Spanish corpus for both RMSE and Pearson correlation coefficient objective measures.

The performance of BDFVopt approach is the highest. This result supports the hypothesis that the selection of the right dimension provides an improvement. We are still far away of chosing the right dimension of the feature vector as shown by the difference between BDFVopt and BDFV approaches. The BDFV proposed in this paper should be considered a first approach. However, BDFVopt is too optimistic because in some cases different dimensions are chosen for the same feature vector. This contradictory behaviour is proper of the privileged information used by BDFVopt.

Corpus	**Train RMSE**	**Test RMSE**	**Train** ρ	**Test** ρ
CART	0.0993	0.1286	0.7758	0.6288
CFV	0.1231	0.1346	0.6155	0.5418
BDFV	0.1099	0.1173	0.7211	0.6811
BDFVopt	0.0778	0.0783	0.8717	0.8555

Table 3: RMSE and Pearson correlation coefficient for Catalan corpus

5. Conclusions

In this paper we show a procedure that intends to overcome three limitations of current intonation model training procedures: interpolation of fundamental frequency contour, sentence-by-sentence parameter extraction and data scarcity.

We proposed to overcome the first two limitations using a parameter extraction that is optimal for the entire database, as proposed in previous papers.

Corpus	**Train RMSE**	**Test RMSE**	**Train** ρ	**Test** ρ
CART	0.0958	0.1214	0.7675	0.6482
CFV	0.1173	0.1298	0.6391	0.5587
BDFV	0.1017	0.1169	0.7419	0.6718
BDFVopt	0.0749	0.0833	0.8644	0.8096

Table 4: RMSE and Pearson correlation coefficient for Spanish corpus

The latest limitation is faced in this paper by means of an algorithm that finds the optimal dimension of the feature vector. In this way we intend to minimize the effects of data scarcity. In addition, this procedure also considers such cases where a feature is more relevant than the others for some feature vector values. The use of the entire feature vector would introduce noise in the measurement of the distance between feature vectors.

Experimental results support our proposal overcoming the use of the full feature vector and another machine learning technique (CART using JEMA) presented in previous papers.

Further work should be done to improve the selection of the optimal dimension of the feature vector in order to be closer to the results of BDFVopt approach.

6. References

[1] H. Fujisaki, S. Narusawa, and M. Maruno, "Pre-processing of fundamental frequency contours of speech for automatic parameter extraction," *Proceedings of the International Conference on Signal Processing*, pp. 722–725, 2000.

[2] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[3] D. Escudero and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 481–484, 2002.

[4] D. Hirst, A. D. Cristo, and R. Espesser, "Levels of representation and analysis for intonation," *Intonation : Theory and Experiment. Kluwer Academic Press, Dordrecht*, 2000.

[5] P. D. Agüero and A. Bonafonte, "Intonation modeling for TTS using a joint extraction and prediction approach," *Proceedings of the International Workshop on Speech Synthesis*, 2004.

[6] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Joint extraction and prediction of Fujisaki's intonation model parameters," *Proceedings of International Conference on Spoken Language Processing*, 2004.

[7] D. Escudero and V. Cardeñoso, "Optimized selection of intonation dictionaries in corpus based intonation modelling," *Proceedings of Eurospeech 2005*, pp. 3261–3264, 2005.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," *http://www.fon.hum.uva.nl/praat/*.

[9] P. D. Agüero and A. Bonafonte, "Improving TTS quality using pitch contour information of source speaker in S2ST framework." *Proceedings of the 12th International Workshop "Advances in Speech Technology 2005"*, 2005.