A New Clustering Approach for JEMA

Pablo Daniel Agüero and Juan Carlos Tulli

Communications Lab
University of Mar del Plata
Argentina

Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya
Spain

Abstract

This paper focuses on the training process of intonation models for text-to-speech synthesis. In previous papers we concentrated on two key points of intonation modelling: interpolation of fundamental frequency contour in unvoiced segments and sentence-by-sentence parameter extraction. We proposed an alternative approach for model training named JEMA (Joint Extraction and Modeling Approach) using CART. Here we propose a new alternative to obtain the mapping function that relates the linguistic features available in TTS and the fundamental frequency contour space. A clustering algorithm using a distance measure over a variable feature vector dimension space is used to partition the space of fundamental frequency contours in the training data. In this way we seek for important groups of features with specific values that explain the shape of fundamental frequency contours. The proposed technique shows improvements in the experimental results over CART.

1. Introduction

Prosody models play a fundamental role to achieve natural voices in text-to-speech systems. Listening effort measurements show that long synthetic speech fragments do not sound natural for listeners. We need further work to achieve natural and expressive synthetic voices.

The intonation model is one of the key prosodic models, because it conveys important information about the meaning of the message, speaker's attitude, psychological and emotional state, and idiosyncratic and sociolinguistic aspects.

Several intonation models have been proposed in the literature. In general we can differentiate several aspects of intonation models:

- *Mathematical formulation.* Previous papers proposed different mathematical formulations to describe the fundamental frequency contour using a compact representation: exponential (Fujisaki [1]), polynomial (Tilt [2] and Bezier [3]), piecewise lineal (IPO) [4], etc.
- *Parameter estimation procedure*. Parameter extraction procedures are related to the field of curve fitting using optimization algorithms: gradient descent, genetic algorithms, neural networks, etc.
- *Model training algorithm*. Intonation model training consists in finding a mapping function that generates a fundamental frequency contour (f_0) given a set of features (F) extracted from the text available in a text-to-speech system: $G(F) = f_0 + e$. In the case of data-driven approaches this task is done minimizing the prediction error (e) for the training data.

Here we focus in the model training step to find out a better technique to obtain the function G(f), minimizing the error e and improving the generalization.

In previous papers [5, 6, 7] we proposed using trees (CART) to estimate this mapping function. This machine learning technique has the disadvantage of a greedy estimation of the solution and an orthogonal partition of the feature vector space.

In this paper we propose a clustering technique based on feature vector matching using variable dimension (VFVDA: Variable Feature Vector Dimension Approach). In this way we obtain a more precise partition of the fundamental frequency contour space analyzing all possible combinations of feature vector values in the training data. The parameters are estimated using an optimization algorithm over all training data.

In our previous work [8] we only use a few number of combinations obtained from a relevance analysis. Here we perform an exhaustive study of the feature vector space. The chosen mathematical formulation is superpositional. The components are accent group and minor phrase.

In Section 2 we give a description of the parameter estimation technique used in this paper. The intonation model training algorithm is shown in Section 3. Experimental conditions and results are shown in Section 4 using two intonation model training approaches: CART and VFVDA. Finally, conclusions are included in Section 5.

2. Mathematical formulation and parameter estimation procedure

The mathematical formulation chosen in this paper are Bèzier polynomials. They are an adequate representation of the fundamental frequency contour shape. The coefficients can approximate any contour with the desired precision by increasing the degree of the polynomial. Additionally, each coefficient represents a specific location of the contour shape. In this way we can easily analyze the contour shape by studying the coefficients. The approach is superpositional, with accent group and minor phrase components.

The global optimization demands that each accent group and minor phrase must belong to a class in order to decouple correctly the two components. As a consequence, we may obtain the optimal parameters for a given class using complementary information of its constituent contours. The missing fundamental frequency information due to unvoiced segments in a given contour is complemented by the other contours of the same class.

In this way we avoid the bias of sentence-by-sentence parameter extraction and the interpolation of unvoiced segments (which may also introduce a bias depending on interpolation and smoothing procedures).

Sentence-by-sentence parameter extraction may generate inconsistent parameterizations for similar contours due to the multiple solutions of some approaches, such as superpositional Bèzier.

On the other hand, interpolation of unvoiced segments to complete missing f_0 data may introduce inconsistent information that harms the extraction of parameters.

The mathematical formulation is shown in Equations 1, 2 and 3.

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

$$P_{MP}^{C_{MP_i}^k}(t) = \sum_n^N \alpha_n^{C_{MP_i}^k} \left(\begin{array}{c} N \\ n \end{array} \right) t^n (1-t)^{(N-n)} \quad (2)$$

$$P_{AG}^{C_{AG_j}^k}(t) = \sum_n^N \alpha_n^{C_{AG_j}^k} \left(\begin{array}{c} N \\ n \end{array} \right) t^n (1-t)^{(N-n)} \quad (3)$$

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 parameters are estimated using the cost function shown in equation 4. 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.

$$e = \sum_k^{N_s} \left(\sum_t^{T_k} \left(f0^k(t) - \left(\sum_i^{N_{MP}^k} P_{MP}^{C_{MP_i}^k}(t_{MP_i}^k(t)) + \sum_j^{N_{AG}^k} P_{AG}^{C_{AG_j}^k}(t_{AG_j}^k(t)) \right) \right)^2 \right) \quad (4)$$

where:

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

3. Model training

In Section 2 we shown the procedure to obtain the optimal set of coefficients given a certain number of classes in the training data. In this section we explain how these classes are found using features extracted from the text available in a text-to-speech system.

Model training consists of finding a mapping function G that predicts a suitable fundamental frequency contour (or a set of parameters that are used to synthesize such contour) given a set of features (F). In this expression we included a prediction error (e) that is minimized over the training process: $G(f) = f_0 + e$.

In the literature several algorithms have been proposed to find this mapping function: classification and regression trees, neural networks, instance based learning, etc [9, 10, 11].

In this section we explain two model training algorithms to find the mapping function: CART and VFVDA.

3.1. Baseline algorithm: CART

CART is a training technique that uses trees to represent the partition of the fundamental frequency contour space of prosodic units (accent groups, minor phrases) given particular values of their features. The main limitations are the orthogonal partition of the feature space and the greedy nature of the algorithm.

The process of intonation model training using CART can be summarized into these steps:

- **Initialization.** Initially only one class exists, because the tree has only the root node. In this way, all prosodic units (accent groups, minor phrases) will be represented by the same set of parameters. These parameters are calculated using a global optimization algorithm over all training data.
- **Splitting.** Linguistic and paralinguistic features are used to do questions in the tree to split training data. After a new question is done, the training data will have two new classes obtained from the splitting of the previous class.
- **Optimization.** When the new classes are obtained, a global optimization algorithm is used to find the new optimal parameters. Depending on the parameterization, this optimization step can be time consuming if the optimal solution has not closed-form (e.g.: Fujisaki's intonation model). In such cases hill-climbing algorithms are used to find the optimal solution.
- **Scoring of the splitting.** The new parameterization is used to measure the improvement of the goodness measure compared to its value previous to the splitting.
- **Selection of the highest improvement.** After all possible splittings were tried, the one with the highest improvement is chosen and the tree is updated for the next iteration.
- **Stopping condition.** The decision of another iteration for an additional splitting is performed taking into account a minimum number of elements on each leaf and a minimum improvement of the goodness score.

CART are generally 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.

In the next section we explain VFVDA, another machine learning technique that also provides useful information about the problem.

3.2. Proposed algorithm: Variable Feature Vector Dimension Approach (VFVDA)

It is known that some features have an extreme influence in pitch contour shape for certain values, e.g: last accent group in an interrogative sentence, first accent group in an interrogative sentence with an interrogative pronoun, emphasized word in an accent group, etc. Some other features in the feature vector may be ignored in such cases because its influence in the contour shape is insignificant. It is very important to find a training algorithm that focuses in these aspects even when the frequency of occurrence is low in training data. Data scarcity is an important problem in intonation model training and rare events need to be modelled as well as frequent events [12]. The variable feature vector dimension approach is a step in that direction.

Here we propose a training approach using exact matching with variable feature vector dimension. We explore all possible combinations of feature vector values for any possible dimension. Our goal is to find a group of feature vector patterns that define the membership of pitch contours to a given class. In this way we will explain pitch contour shapes related to some particular values in the features.

Given the superpositional mathematical formulation, it is necessary to succesively refine the accent group and minor phrase components in separate steps. In this way we will avoid a significant growth in the number of combinations of features.

The steps of the training procedure are:

1. **Initialization.** In the initial condition we suppose that no class exists.
2. **Generation of all possible combinations of feature vector values and dimensions for accent groups and minor phrases.** All possible feature vector values are explored for any possible dimension. This procedure will show us how complex will be the exhaustive search of the optimal classes. The combinations of values with a number of elements below a threshold are not considered. They do not have enough data to obtain a reliable estimation of the class pattern.
3. **Candidate feature vector pattern (minor phrases)**. Each possible combination generated in the previous step is considered a candidate feature vector pattern (feature vector that represents a pitch shape class). In this step the candidate is added to all previously found feature vector patterns to explore the error reduction produced by its inclusion. The feature vector patterns are explored from dimension one until the maximum dimension. A dimension is only studied when the previous dimension does not produce any improvement in the global RMSE.
4. **Assignation (minor phrases).** Each feature vector in the training data is assigned to the class of a feature vector pattern if the values of the features exactly match. When a feature vector matches the feature vector pattern of several classes, we consider it belongs to the class with the higher dimension (more specific description of feature vector values).
5. **Best feature vector pattern (minor phrases).** Optimal Bezier coefficients for each class are found using the algorithm described in Section 2. RMSE is calculated for all training data. The feature vector pattern with the highest error reduction is added to the feature vector patterns for the next iteration.
6. **Steps 3, 4 and 5 are repeated for accent groups.**
7. **Stop condition.** The algorithm stops when it is not possible to generate classes with a number of samples higher than a predefined threshold (20 in our experiments) or the RMSE reduction is lower than a predefined threshold (0 in our experiments).

The proposed algorithm is exhaustive and the training may take several hours depending on the available training data and the number of features and their combinations. However, the resulting model can be used in real-time situations because it only consists in a lookup table with few entries.

In the next section we will show experimental results of our proposal compared with another intonation model training approach.

4. Experiments

4.1. Experimental conditions

The experiments were performed using the male baseline Spanish voice of the European Project TC-STAR. The data are paragraphs from the parliamentary domain. Due to the restriction of the corpus to ten hours of speech recorded from one speaker, it was decided to focus mainly on coverage of phonetic and prosodic variations. The voice should therefore sound as being uttered by a competent translator speaking in a rather neutral manner.

Two different approaches for intonation model training are compared in these experiments. All of them use the same mathematical formulation: Bèzier coefficients. The difference between them lies in the training procedure:

- **Bezier (CART).** The intonation model is trained using CART proposed in Agüero et al [5].
- **Bezier (VFVDA).** The intonation model is trained using the Variable Feature Vector Dimension Approach proposed in Section 3.

The idea is to study the performance of our current proposal (VFVDA) compared with our previous approach (CART). It is expected that VFVDA will outperform CART due to a more precise analysis of the feature vector space.

4.2. Experimental results

The experimental results are presented in Table 1. We divided the corpus into ten fragments to do 10-fold cross validation experiments. The first columns show the RMSE (using log-Hz) for test data using the different approaches. The last column shows the gain in RMSE of VFVDA over CART for each fold.

VFVDA explains better test data because in all cases there is a gain in the RMSE. Gains range from 0.005 to 0.0016 in the sixth fold. These results support the use of VFVDA over CART because it systematically provides better results for unseen data.

Anyway, the proposed technique has some limitations that harm its performance:

- **Speaker variability.** Given a feature vector pattern, all pitch contours of that class are represented by the same pitch contour shape. However, speakers may perform some variations to the contour without a reason. This is

CART	VFVDA	Gain
0.0962	0.0957	0.0005
0.1032	0.1026	0.0006
0.0996	0.0988	0.0008
0.1014	0.1000	0.0014
0.0996	0.0988	0.0008
0.0963	0.0947	0.0016
0.0975	0.0964	0.0011
0.0945	0.0940	0.0005
0.0950	0.0940	0.0010
0.0978	0.0971	0.0008

Table 1: Experimental results for test data using CART and VFVDA.

an intrinsic limitation of intonation model training and evaluation.

- **Missing training features.** It is not possible to generate all necessary features to explain pitch contour shape. Aspects like semantics, pragmatics, dialect, etc., may introduce a variability that may not inferred from the limited training data available in any task of intonation modelling.
- **Limitations in the training procedure.** The training approach proposed in this paper successively finds the best feature vector patterns. This procedure may produce a suboptimal solution because of local decisions (choosing the best feature vector pattern in an iteration) that does not take into account future decisions.

5. Conclusions

This paper describes a new approach for intonation model training. Here we faced problems detailed in previous papers such as fundamental frequency contour interpolation of unvoiced segments, parameter extraction bias, and the important limitation of data scarcity.

We proposed an exhaustive search algorithm for optimal feature vector pattern extraction: Variable Feature Vector Dimension Approach (VFVDA). This algorithm was compared with another training technique: CART.

The proposed algorithm has better performance than CART for the test data, encouraging its use for intonation model training. VFVDA provides a better generalization in our experiments.

The data used in this paper is pronounced in a rather neutral manner. In the future we will focus on using more expressive data in order to study the relevance of the different features in such situation.

6. References

[1] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5, 233-242.

[2] Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(3), 1697-1714.

[3] Escudero, D., 2002. Modelado estadístico de entonación con funciones de Bézier: Aplicaciones a la conversión texto-voz en Español. *PhD Thesis*.

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

[5] Agüero, P.D.; Bonafonte, A., 2004. Intonation modeling for TTS using a joint extraction and prediction approach. *Proceedings of the International Workshop on Speech Synthesis*, 67-72.

[6] Agüero, P.D.; Wimmer, K.; Bonafonte, A., 2004. Joint extraction and prediction of Fujisaki's intonation model parameters. *Proceedings of International Conference on Spoken Language Processing*, 757-760.

[7] Rojc, M.; Agüero, P.D.; Bonafonte, A.; Kacic, Z., 2005. Training the Tilt intonation model using the JEMA methodology. *Proceedings of Eurospeech*, 3273-3276.

[8] Agüero, P.D.; Bonafonte, A., 2006. Facing data scarcity using variable feature vector dimension. *Proceedings of the International Conference on Speech Prosody*, 1-4.

[9] Navas, E.; Hernaez, I.; Sanchez, J.M., 2002. Basque intonation modelling for text to speech conversion. *Proceedings of ICSLP*, 2409-2412.

[10] Mixdorff, H.; Jokisch, O., 2001. Implementing and evaluating an integrated approach to modeling German prosody. *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, 211-216.

[11] Bailly, G.; Holm, B., 2005. SFC: A trainable prosodic model. *Speech Communication*, 46, 348-364.

[12] Cardeñoso, V.; Escudero Mancebo, D., 2004. A strategy to solve data scarcity problems in corpus based intonation modelling. *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 665-668.