Experimental evaluation of the relevance of prosodic features in Spanish using machine learning techniques

David Escudero, Valentín Cardeñoso
Departamento de Informática. Universidad de Valladolid. 47011 Valladolid. Spain.
Contact: descuder@infor.uva.es
Antonio Bonafonte
TALP Research Centre. Universitat Politècnica de Catalunya. 08034 Barcelona. Spain.

Abstract

In this work, machine learning techniques have been applied for the assessment of the relevance of several prosodic features in TTS for Spanish. Using a two step correspondence between sets of prosodic features and intonation parameters, the influence of the number of different intonation patterns and the number and order of prosodic features is evaluated. The output of the trained classifiers is proposed as a labelling mechanism of intonation units which can be used to synthesize high quality pitch contours. The input output correspondence of the classifier also provides a bundle of relevant prosodic knowledge.

1. Introduction

This work is a continuation of the activities we have done in the field of modelling intonation for Spanish text-to- speech(TTS) applications [1, 2, 3]. We proposed an original approach for modelling intonation from corpus, based in quantitative parameters of the pitch contours, and in the definition of a number of classes of intonation with a statistical model associated. Here we propose a new step in the methodology, where different learning techniques are used for obtaining automatically the classes and to establish the matching between prosodic features and pitch contours.

Improvements in the naturalness of TTS systems are still an issue, specially because new fields of application are emerging which require fast adaptation to new speaker voices and speaking styles and a better set of prosodic rules. Although rule based systems can generate high quality prosody, they are difficult to adapt to changing environments. As far as intonation modelling is concerned, corpus based systems could provide the best engineering solution for these challenges. Models of intonation attempt to find out the relationship between linguistic information at a text and prosodic level and pitch contours (characterised by sets of parameters like in TILT or Fujisaki models). Different approaches to represent intonation and to model the relationship between acoustic and linguistic information can be found in state of the art techniques (as reviewed in [5]).

Several machine learning algorithms (neural networks, CART trees, linear regression, ...) have been reported trying to establish direct matching between prosodic features and sets of parameteres controlling intonation. A common problem of these methods is the lack of efficient means to provide prosodic knowledge. This is usually related to either the high number of prosodic features involved or to the huge range of values assigned to the features (generally numeric). Furthermore, the set of parameters to be predicted are highly correlated in most cases and this increases the difficulty to interpret the models obtained, even if the pitch contours are correctly rendered.

In our proposal, an unsupervised intermediate labelling is applied in order to provide a description of sets of intonation patterns which share common behaviours. With this description at hand, several learning algorithms are tested to provide an automatic classification of the label (nominal attribute) in terms of several prosodic features which are known to be relevant in Spanish (type of sentence, type of accent, position of stress group within the intonation group, number of syllables, ...). From this classification, a number of prosodic features can be easily selected as the most relevant candidates, thus providing a knowledge extraction mechanism which can be useful for various applications and, more specifically, to generate satisfactory pitch contours.

Since there is no paramount consensus on the number, nature and relevance of the prosodic features to be used in corpus based TTS in Spanish, the reported solutions either heuristically select the set of features which best serves their needs or simply try to incorporate as many features as possible. In this work, we provide a means to objectively test the relevance of a given set of prosodic features based on entropy measures and we will show that increasing the number of prosodic features not always warrants better pitch prediction.

The rest of the paper is organised as follows. A brief summary of the intonation modelling techniques is given first. Then, we describe the labelling method to obtain classes of intonation patterns and how a relationship between these classes and sets of prosodic features can be established. Finally, we report the main results, some conclusions and future work proposals.

2. Representing and Modelling Intonation

The intonation modelling methodology used in this work is schematically presented in Figure 1. Linguistic analysis of text is used to segment it into intonation units and to provide the set of prosodic features associated with them. Pitch contours are parameterised using Bézier functions. Beside its set of prosodic features, each intonation unit in the corpus is represented by the set of parameters of intonation extracted from the control points of the Bézier fitting function (see [2] for details).

During the modelling phase, classes of intonation units are built according to some classification criterion and the statistical distributions of the four control points of the Bézier function

This work has been partially supported by Junta de Castilla León under Research Contract nº VA083/03 and by the Spanish Ministry of Science and Technology under Research Project nº TIC2000-1669-C0403

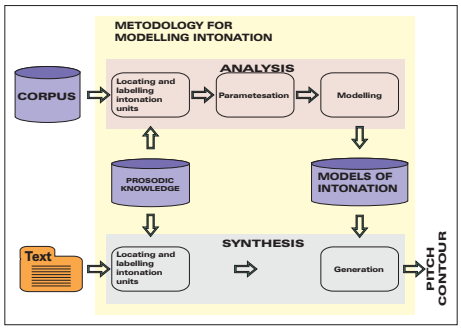

Figure 1: Intonation modelling and generation of synthetic intonation.

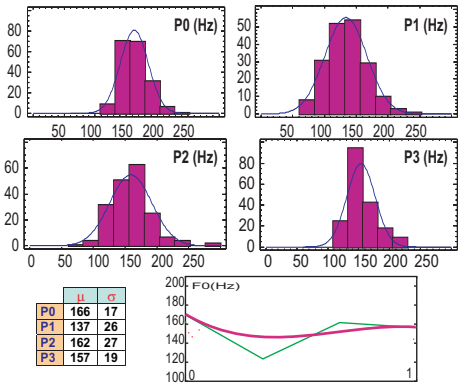

Figure 2: Statistical model corresponding to one the classes of intonation units in the corpus. Histograms show the distribution of the values of the four control points of the Bézier function. Mean values and standard deviations are gathered in the table and the representation of $F_0(t)$ for the representative pattern (mean values) of the class is drawn beside it.

fitting each intonation unit are taken as generation models of the class (see figure 2). While in [3], we used several sets of commonly used prosodic features as the only grouping criterion of intonation units in a class and a comparison of the synthesis results was taken as the validation mechanism, here we attempt to fix the right set of prosodic features to be used from an experimental study of the corpus.

In the generation phase, a pitch contour is generated for each intonation unit using the statistical distributions of the model associated to the given unit. Segmentation and labelling of the intonation units is done separately within the linguistic module of the generator. Objective and subjective evaluations provided inteligibility results comparable to the ones of other approaches and acceptable naturalness qualifications (see [3] for details).

The results of these previous works showed that the selection of an appropriate set of prosodic features to label intonation units is crucial to obtain good results. That's why we try to apply machine learning techniques to provide this information in this work.

3. Learning of Intonation Models

Each of the intonation units of the corpus can be seen as a pair (PF, PI) where PF are the prosodic features and PI are the parameters which control intonation generation. The goal is to find a correspondence between PF and PI which allows predicting PI from PF. As an intermediate step, we decided to group similar intonation patterns under a class of intonation (CI), in order to get a discrete representation of them which becomes the expected output of the machine learning classifier we train later. With this, a proper correspondence between high level prosodic knowledge and unit types is ensured. As a consequence, the correspondence PF->PI is splitted in two separate stages: PF->CI and CI->PI. These two correspondences are determined following the procedural steps described hereafter.

Step 1: *Parameterization of intonation units.* As in previous works, we used a corpus designed for concatenative synthesis which contains 4625 stress groups (sequence of syllables within two consecutive stressed words) and 1615 intonation groups (sequence of stress groups within two relevant changes of F0). The basic intonation unit is the stress group and pitch is evaluated from glottal closing time points. Each stress group is approximated using a third degree Bézier function and the four associated control points are taken as intonation parameters of the unit.

The prosodic features annalyzed in this work are taken from the proposals of previous studies of Spanish intonation carried out by several well recognized authors (see [3]). The following features were used: position of the stress group in the intonation group (`posGAenGE`); type of accent (`aceGA`); number of syllables of the stress group (`nSilGA`); number of stress groups in the intonation group (`nGAenGE`); number of syllables of the intonation group (`nSilGE`); position of the intonation group in the sentence (`posGEenFR`); number of stress groups in the sentence (`nGEenFR`); type of sentence (`tipoFR`). 80% of the sentences were reserved for training and 20% for testing.

Step 2: *Clustering of intonation parameters.* Vector quantization techniques are applied to cluster intonation units in terms of the similarity of their characterizing parameters. The goal of this process is to provide a labelling of the intonation units as a whole, since a high correlation between the values of individual parameters (control points) is found and because we aim providing a relation between a given set of prosodic features (learned in following steps) and a class of intonation units which can be assigned a statistical model for generation purposes. After this step, each intonation unit can be seen as the pair (PF,CI).

To carry out the vector quantization VQ: PI -> CI, with CI= $\{CI_1 ... CI_N\}$ and N the number of classes, we have applied the $Kmeans$ method (using euclidean distance) provided with the free software ML package WEKA[1]. The whole process has been repeated for several values of N in order to evaluate its impact on the final results.

Step 3: *Selection of Prosodic Features.* Instead of using the full set of prosodic features at the same time, we experimented several subsets of prosodic features as input to the classifier (CI->PF). These subsets were built starting with the three most relevant features and progressively adding new prosodic features in their corresponding order of relevance.

To assess the relevance of every prosodic feature to be incorporated to the set, we used the usual entropic information measure considered in ML. Table 1 presents the *information gain* (`FinalGain`) as a function of the prosodic feature. This

[1] http://www.cs.waikato.ac.nz/ ml/weka/

Pr. Feature	Dim.	GainInfo	SplitInfo	Final Gain
posGAenGE	6	0.4806	1.563	0.307
posGEenFR	5	0.3223	1.466	0.219
nSilGA	7	0.3077	1.691	0.181
aceGA	3	0.0855	0.766	0.111
nSilenGE	5	0.1407	1.329	0.105
nGAenGE	5	0.1441	1.485	0.096
nGEenFR	4	0.0856	1.264	0.067

Table 1: Information Gain for the different prosodic features

value is closely related with the entropy of the distribution of samples in terms of their features. The *splitting bias* (`SplitInfo`) factors out the influence of the natural trend to overspecialization of the learning phase. The *total information gain* (`GainInfo`) provides a measure of the potential loss of entropy which would be generated if the splitting of the training set was carried out in terms of the present attribute (see [6] for a description of all these metrics).

As seen in the table, `posGAenGE`, `posGEenFR` and `nSilGA` are the prosodic features that give more information. Other features, like `nGAenGE` and `nGEenFR` have a very low relevance. `TipoFR` doesn't appear in the table because, as a consequence of the small number of sentences of other kinds present in the corpus, we only considered declarative sentences for this study.

Step 4: *Learning Phase.* The pair (PF,CI) will be used to train a set of classifiers, taking PF as the input and CI as the output. A number of learning algorithms have been evaluated, in order to test the influence of this factor and because no 'a priori' information could guide us to the best one.

For the learning phase, classification has been carried out using: C4.5 Decission Trees C4.5 (`j48.J48`), Decision Trees rule learners PART (`j48.PART`), Instance Based Learning (`IB1`), Decision Tables (`DecisionTable`) and Bayesian Decision (`NaiveBayes`) (see [6] for a detailed description of these classifiers).

A matrix with the distances between the centroids of each of the classes has been used as the cost matrix of the learning algorithms. This is done to weight the errors in the classification whenever a class is predicted which is very different from the real class.

Each of these classifiers gives the relationship between prosodic features and a class of intonation patterns. Figure 3 shows the bundle of pitch contours for three of these classes and the associated rule obtained with the `PART` classifier. This figure illustrates to which extent the approach explained here can help the expert to interpret prosodic knowledge present in the corpus she is working with.

Step 5: *Generation of Synthetic Intonation.* The linguistic module of the TTS system provides the boundaries of the stress groups and assigns prosodic features in terms of linguistic rules for Spanish. This information is the input of the predictors trained in step 4 and the label of the class of intonation pattern is the output.

Given the class label, we can generate a smooth pitch contour using any of our simulation methods for the control points of the Bézier function already discussed in [7].

Step 6: *Evaluation of Synthetic Intonation.* In this step, the classifiers obtained in step 4 are used to predict the class of any stress group. Again, given the class, a synthetic pitch contour is to be generated using the same procedure as above. For evaluation purposes, however, we have used the mean values of the control points of the intonation units in the class to generate the contour.

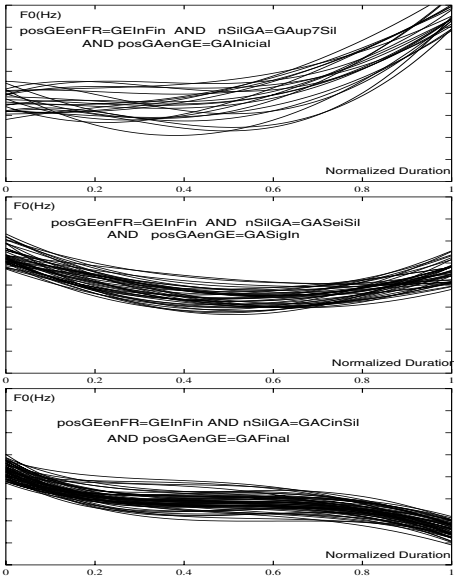

Figure 3: Set of patterns of intonation of the same class, and the rule associated with the class using the PART learning algorithm.

Proposal	RMS	Pearson Correlat.	Number of Clases	Non Empty Clases
Rep0	22.65	0.25	1	1
Lopez	18.58	0.66	36	36
Garrido	17.74	0.68	288	185
Vallejo	20.02	0.59	240	145
Alcoba	20.17	0.61	200	119
NaiveBayes	17.66	0.70	100	61
j48.PART	17.52	0.70	100	74
Id3	17.75	0.69	100	80
DecisionTable	17.40	0.70	100	81
j48.J48	17.43	0.70	100	80
IB1	17.76	0.69	100	80

Table 2: Quality measures of the predicted pitch countours, number of classes and non empty classes for each case studied in this work.

Objective evaluation is done by measuring the distance between the real pitch contours and the predicted ones (RMSE and Pearson Correlation).

4. Results and discussion

Table 2 compares the results obtained using automatic techniques presented in this work with the ones obtained using the knowledge based proposals of previous works. The pitch contour predicted in proposal *Rep0* is the same for every class and it consists of the mean value of the parameters of intonation of all the stress groups in the training corpus (one class world). The proposals *Lopez*, *Garrido*, *Vallejo* and *Alcoba* were obtained applying the method explained in the previous section and using the sets of prosodic features which were considered most relevant for Spanish by experts of the same name. The rest of the data are obtained using the implementations of the learning

algorithms available provided in WEKA.

Results show that the feature sets corresponding to smaller numbers of classes provide better predictions, in general. It is important to take into account that the `Garrido`, `Vallejo` and `Alcoba` proposals used a feature `TrayGA` which is related with the overall shape of the pitch contour and which hasn't been considered here since it should be inferred by the learning algorithm. In terms of the result presented in table 2, we can conclude that this prosodic feature they used is a good candidate to globally describe the combinations of the rest.

In the sets of *Garrido*, *Alcoba* and *Vallejo* void classes appear because of the size of the corpus. For the rest of the cases, they appear because some of classes are never predicted. A specific study might be done of these infrequent, never selected patterns because it could shed light about some intonation particularities of the studied corpus.

Figure 4 shows that, with a fixed number of parameters, the tendency of the RMSE is to get stable as the number of classes of the vectorial quantifier increases. When the number of classes grows, the intra-class similarity gets better, but the combination of the prosodic features is a worse predictor and, as a result, the RMSE does not improve. It can also be observed that when the number of considered prosodic features increases results can get worst for the same number of classes. This is because increasing the number of prosodic features introduces distortion into the classifier. The differences get smaller when the number of classes increases. From that, we can conclude that in order to get benefits of incorporating new prosodic features, it is necessary to increase the number of classes of intonation. This is so because a higher number of classes usually implies more intonation details, and it would be necessary to increase the number of prosodic features to finding a correspondence for this detailed patterns. Thus, the number of classes to use seems to depend heavily on the number of prosodic features involved, and on the size and nature of the corpus.

Subjective evaluation implies carrying out informal tests of the naturalness of the synthetic pitch contours. The test sentences of the corpus are re-synthesised using the generated synthetic pitch contours. To do so, we use the re-synthesis PSOLA module included in the `praat` [2] software tool.

5. Conclusions

This work presents a strategy for corpus based intonation modelling which exploits machine learning techniques to extract sets of prosodic features which guarantee the generation of high quality pitch contours.

Using learning based classification techniques, the relevance of different prosodic features can be assessed. Furthermore, some of the methods of classification (specially the ones based on decision trees) provide prosodic knowledge which offers valuable information on the relationship between prosodic features (linguistic information) and pitch contours (acoustic information). The classifier output can be considered as prosodic information which could be used to compare different languages, speaking styles, emotions or intervening speakers.

It has been shown that incorporating new prosodic features does not necessarily imply better prediction results, the reason being that the excess of input information can distort the classifier, speciallly when the number of training samples has not been enough.

A natural evolution of this work would be to make use of the prosodic knowledge we have obtained. The aim will be to identify the relative impact of the different factors (classes, prosodic features and their relationship) in different environments. This could have applications in areas like speaker, emotions or dialect recognition.

[2] http://www.praat.org

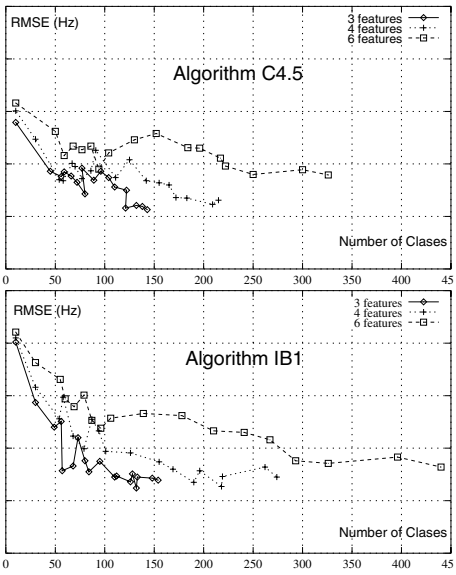

Figure 4: Evolution of the value of RMSE as a function of the number of intonation classes and of the number of prosodic features for two different learning algorithms.

6. References

[1] V. Cardeñoso and D. Escudero, "Statistical modelling of stress groups in spanish," in *Proceedings of Prosody 2002*, 2002.

[2] D. Escudero and V. Cardeñoso A. Bonafonte, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish," in *Proceedings of ICASSP 2002*, Mayo 2002.

[3] D. Escudero, C. González, and V. Cardeñoso, "Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in spanish," in *Proceedings of ICSLP 2002*, Mayo 2002.

[4] A. Botinis, B. Granstrom, and B. Moebius, "Developments and Paradigms in Intonation Research," *Speech Communications*, vol. 33, pp. 263–296, July 2001.

[5] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.

[6] D. Escudero, *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto Voz.*, Ph.D. thesis, Dpto. de Informática, Universidad de Valladolid, España, 2002.