

Impact of the Selection of the Constructive Type of Intonation Unit in a Data-Driven Intonation Modelling Technique

D. Escudero-Mancebo, V. Cardeñoso-Payo

{descuder, valen}@infor.uva.es

Department of Computer Science
University of Valladolid, Spain

Abstract

In this work we present a methodology for modelling intonation from corpus that operates with alternative types of intonation units. We compare prediction results obtained using a set of different ones. Results permit to select most suitable one depending on the corpus and to obtain information about the relative importance of different prosodic features in characterizing the intonation of the corpus in every case.

1. Introduction

The goal of modelling intonation techniques is to find the correspondence between function of intonation (represented by prosodic features depending on the message and on the context) and its form (F0 contours) (see [1] for an excellent review). Data-driven techniques build this correspondence from corpus with automatic methods like neural networks, decision trees... In this context, one of the most important decision to take is the selection of the type of intonation units that will configure the modelling technique.

In the state of the art, we find that there is not a consensus about the type and boundaries of the intonation unit to use: we find accentual phrase [2], accent groups [3], ToBI transcriptions [4], syllables [5], intonation events in [6]. . . . In this communication, we defend a procedure to illustrate objectively the relative goodness of the selection of a type of unit in out of a set of alternatives in data-driven modelling intonation.

This work is a continuation of our previous activities in modelling intonation. We have defined a framework for modelling intonation from corpus with applications in Text-To-Speech (TTS) systems (from now, we will refer to it as MEMOInt- Methodology for MOdelling Intonation-). In [7] we presented the basic methodology and the parameterisation technique, based on fitting Bézier functions of the F0 contours of the stress groups. In [8] we used MEMOInt to compare different alternatives in the characterization of the stress group. In [9] we applied data mining techniques to measure the relative importance of a set of prosodic features to characterize the stress group. In this work, we test alternative types of intonation units to use in MEMOInt and compare them with the stress group.

The aim is to show that MEMOInt permits to model intonation using alternative intonation units bringing useful prosodic information about the relative relevance of the prosodic features and about the shape of the typical pitch patterns. We show that MEMOInt permits to select the best intonation unit (among a

This work has been partially supported by Junta de Castilla y León under research grant VA083/03 and by MCYT contract TIC2003-08382-C05-03

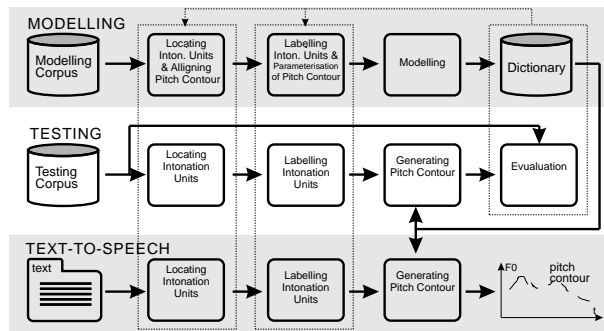


Figure 1: Functional diagram of MEMOInt.

set of them) to be used in data-driven intonation modelling. Although results only refers to Spanish, MEMOInt can potentially be applied to other language if the corpus is available.

First we briefly review MEMOInt. Then, we report the different types of intonation units and its prosodic features that have been analysed here. Experimental procedure is described and results are discussed. We end with the conclusions and future work.

2. MEMOInt

Figure 1 represents the three stages defining MEMOInt: one for modelling, another one for testing and the last one for using results in TTS systems. The output of the modelling stage is a dictionary of models, representing the intonation of the corpus. The representativeness of this dictionary is tested in the second stage. Finally, the models of the dictionary are used to generate synthetic pitch contours in TTS applications.

Initial tasks determine the results of MEMOInt: locating intonation units, its labelling and its parameterisation. Locating intonation units requires to establish a priori the type of intonation unit to be used (syllable, stress group . . .). Then, the utterances of the corpus are divided into its sequence of such intonation units. The labelling task results in assigning values to a set of prosodic features (accent, position etc. . .) characterizing the intonation units. The parameterisation task computes a set quantitative parameters representing the shape of the pitch contour of the intonation units.

Modelling consists of grouping together into the same class the intonation units that share its prosodic features values. The model of each of the classes is the statistical distribution of the acoustic intonation parameters of the intonation units belonging

	typeSE	nIGSE	nSGSE	nSylSE	nPhonSE	posIGSE	nSGIG	nSylIG	nPhonIG	posSTIniSG	posSTFinSG	posSGIG	nSylSG	nPhonSG	posSTSG	relPosSyl	nPhonSyl
Syllable	3	4	4	4	4	4	5	5	4	4	3	3	6	5	4	3	4
Stress Group	3	4	4	4	4	4	5	5	4	4	3	3	6	5	4	3	4
Intonation Group	3	4	4	4	4	4	5	5	4	4	3	3	6	5	4	3	4

Table 1: Cardinality of the prosodic features. SE is sentence, SG is Stress Group, IG is Intonation Group, Syl is Syllable and Phon is phoneme. posXY is position of X in Y. nXY is number of X in Y. typeSE: type of sentence; posSTIniSG, posSTFinSG is position of the stressed syllable in the initial and final stress group; relPosSyl: position of the syllable with respect to the stressed syllable. typeSE is type of sentence.

to the class observed in the modelling corpus. The generalization capabilities of the models increases after an iterative grouping classes process is applied as will be explained in section 4.

The dictionary of models (dictionary of classes indeed) is used to generate synthetic pitch contours both in the testing stage and in TTS systems. The class identifier of any intonation unit is obtained from its prosodic features. The corresponding synthetic pitch contour comes from the statistical model of the class.

MEMOInt feedbacks prosodic information for improving initial tasks (backwards lines in figure 1). Models of the dictionary and the reports from the testing stage are information to compare different options in the labelling and segmenting stage. This information will be used here to compare the behaviour of MEMOInt with alternative types of intonation units.

3. Prosodic Features and Intonation Units

Stress group was the basic intonation units in all our previous works [7, 8, 9]. Selected prosodic features came from experts in Spanish intonation works (see [8] for reviewing such authors). Here we propose to extend the study to two additional intonation units: the syllable and the intonation group.

We use the typical definitions of the intonation units for Spanish: Syllables are obtained by applying the classical words into syllables division rules (see [10]); a stress group is a set of words, where only the last one is accented; an intonation group is a set of stress groups separated by a pause or by a significant inflexion in the F0 contours.

Table 1 shows the three different types of intonation units, its prosodic features, and its number of values. Note the hierarchical relation between the types of intonation units and its projection into the features. We have only selected three types of prosodic features: linguistic ones typeSE, posSTSG, syl-RelPos, posSTIniSG, posSTFinSG); position ones (posIGSE, posSGIG, relPosSyl) and length ones (nSGSE, nSGSE, nSylSE, nPhonSE, nSGIG, nSylIG, nPhonSG, nPhonSyl). Other features like those considering the phonetic structure [3] or the syntactic structure [2] are left here. They have been not included in this study because they are not yet available, but they will be processed in the same way in future works.

The values of the prosodic features (see table 1) are set to distribute uniformly the number of samples in the corpus per value. Number of values of the features is the same in every intonation unit.

Our aim here is not to show that the prosodic features or that the type of intonation units selected are the best to be used.

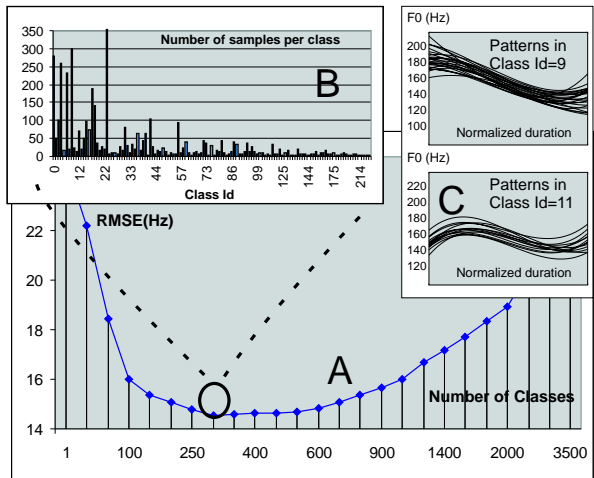


Figure 2: (A) Prediction Error in the iterative grouping. Type of Intonation Unit = Stress Group. All the available prosodic features are used. (B) Number of samples per class at the minimum *Number of Classes=300*. (C) F0 patterns in two of the classes when *Number of Classes=300*

Our aim is to show that MEMOInt is a useful intonation analysis tool to compare alternatives.

4. Experimental Procedure

Independently of the type of intonation unit selected, each combination of prosodic features determines one class in the initial dictionary of models. If there are few stress groups of certain class, its model will not be characteristic and its use in prediction can be problematic. To avoid this situation, we propose to iteratively group together pairs of classes. Joining two classes implies creating a new class which includes samples of both of them. A maximum similarity criterion is applied in each step. Thus, grouping two classes implies a precision loss but brings a generalization gain. The Euclidean distance between centroids of the classes is the criterion to select the classes to merge.

Grouping two classes implies to build a new dictionary. This Dictionary can be used to produce synthetic pitch contours. If the prediction error obtained with the new dictionary is smaller than the previous one, then the new classification is better. By repeating the process, we can measure the compromise between precision and generalization obtaining an optimum configuration for the dictionary. The grouping process can be stopped when the loss of precision forces unwanted prediction results.

Figure 2 shows this iterative process and its results. The initial clustering is determined by the prosodic features (in figure 2.A *number of Classes=3500*). From right to left the number of classes decreases as classes are merged. The minimum of the RMSE plot (*number of Classes=300*) indicates the best dictionary. The dictionary of models represents the correspondence between prosodic features and pattern of F0 (in figure 2.B we show the number of samples per class in the dictionary and in figure 2.C we show the patterns of two of the classes).

Some of the classes of the dictionary of models can be void. One class is void if there are no units of such class in the modelling corpus. But, an intonation unit of any of such void classes

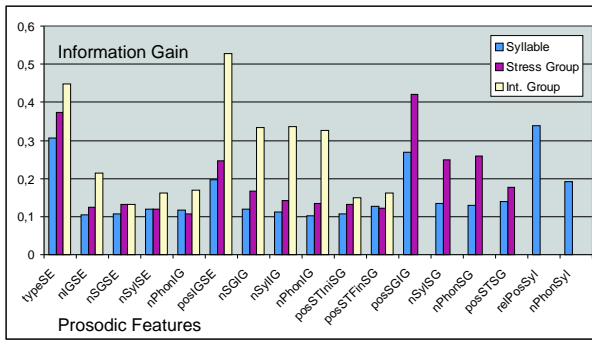


Figure 3: *Information Gain* of every prosodic feature using a Kmeans clustering of the samples of the corpus (see [9] for the *Information Gain* metric)

can appear when using the dictionary to generate synthetic pitch contours. We have observed that, a strategy based on several different dictionaries can be very useful in this situations: most informative dictionary is always used as the first alternative; when a stress group belonging to a void class appears, we recall the dictionary with the higher number of prosodic feature which classifies the stress group into a non void class. Thus, we ensure that the synthetic pitch contour is associated with the right observations in the corpus, at least partially.

Three different collections of dictionaries of models will be built using the three types of intonation units that we are analysing. For every type of intonation unit, each of the dictionaries of the collection will use the N most relevant prosodic feature. All of the dictionaries are built following the grouping criterion explained in the previous section. In the following section we will show that this strategy gives information about the relative importance of the prosodic features. Results in section 5 will show which of the types of intonation units permits to obtain more accurate pitch contours.

For building the collections of dictionaries a ranking of the relevance of the prosodic features is built. Following the procedure explained in [9] such ranking is built for every type of intonation unit considered (figure 3). The number of clusters used are 800 for syllables, 250 for stress groups and 100 for intonation groups. These figures balance the number of samples per class in the three cases.

The corpus used is the same corpus we have already used in previous works¹. It contains 14971 syllables, 4665 stress groups and 1747 intonation groups. The number of interrogative and declarative sentences is scarce (only the 5%) so that only declarative sentences are used. All the prosodic features are computed automatically. The acoustic parameters to be used are the control points of the Bézier curves fitting the F0 contours in the intonation units (more details in [7]). Prediction errors and statistical models are obtained with raw F0 contours (note that better results can be obtained applying a smoothing method as we did in previous works).

5. Results and Discussion

Plots in figure 5 are the result of applying the iterative grouping method described in section 4 using the different types of intonation units taken into account. Minimum values represent

¹Gently provided to us by TALP group of UPC university.

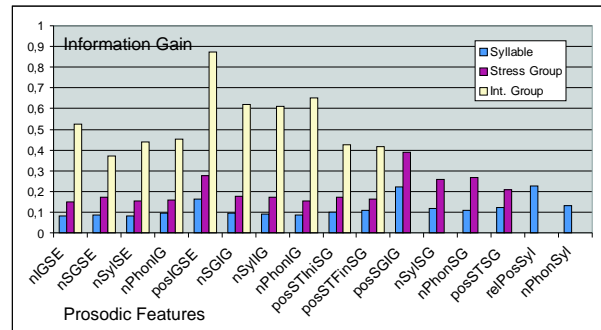


Figure 4: *Information Gain* of every prosodic feature with respect to the best clustering obtained with the iterative grouping classes method.

the optimum number of classes for the different sets of features. Figure 4 shows the relative importance of the prosodic features with respect to these optimum classifications.

From figure 5 we observe that, as general rule, the more prosodic features involved the better the prediction. This rule is broken in the case of syllables when more than 12 features are used (after + nSylIG). This is due to scarcity problems because plot are obtained using a single dictionary and a default value represents void classes.

Best results are obtained with the syllable. Nevertheless, the advantage with respect to the stress group is not so important especially comparing the number of classes.

Comparing figures 3 and 4 it can be observed that the relative rankings are very similar. This result indicates that automatic Kmeans clustering is a good reference to obtain information about the relative importance of the prosodic features.

With respect to the results of the rankings, as general rule, the most relevant features are those that refer to the type of intonation unit that is been studied: In the case of Intonation Groups posIGSE, nSGIG, nSylIG, nPhonIG; In the case of Stress Groups posSGIG, nSylSG, nPhonSG, posSTSG; and for the syllable relPosSyl and nPhonSyl. As remarkable exceptions we have typeSE, posIGSE that are important in every type of intonation unit; nIGSE important to characterize intonation units and posSGIG for syllables. Overall, the most relevant features are typeSE, posIGSE, posSGIG, nSGIG, nSylSG and relPosSyl, that is, relative position of the units with respect to the superior ones and its length. typeSE is also very relevant, but we have been not able to test this results due to the scarcity of the corpus. These results are only valid for our corpus but they are coincident with the results in the bibliography of Spanish intonation.

In the plots of the figure 5 it can be observed that improvements depend on the feature leaving the set. For example, at Intonation Unit = Stress Group the difference between the + nSGIG, + nSylIG, + nPhonIG plots is small. This is due to the fact that this three features refers to the same aspect: the length of the Intonation Group. Information entered by any of them, may have been already taken into account by the others and the results don't change. This leads us to think that very similar results could have been potentially obtained using less prosodic features reducing scarcity problems.

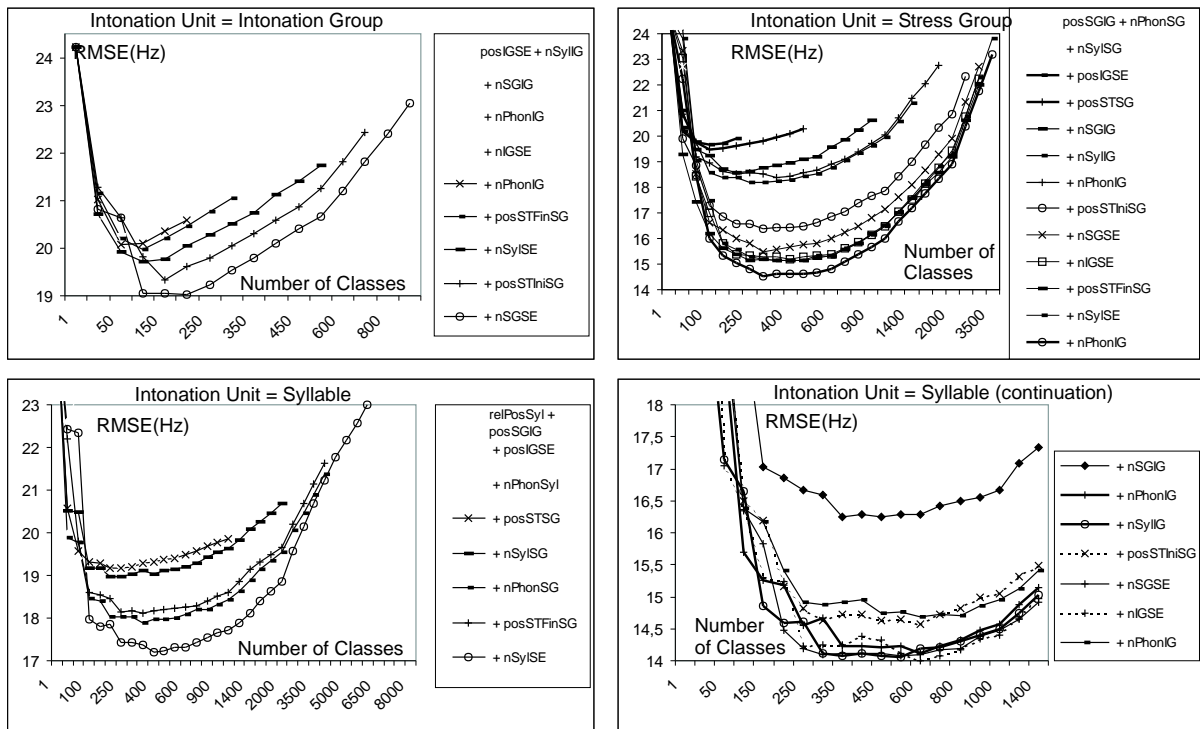


Figure 5: Prediction Error of the samples in the test corpus with respect to the number of classes in the iterative grouping process.

6. Conclusions and Future Work

MEMOInt permits to characterize the intonation of a given corpus using different type of intonation units. The methodology permits to obtain quantitative information about the relevance of the different prosodic features taken into account.

MEMOInt proposes also the optimum set of classes of pitch patterns. In future work, we will study the correspondence of these patterns with the prosodic features to evaluate the coherence of the results.

As future possibilities, MEMOInt permits to value the relevance of any other prosodic feature that could be required to evaluate or any other type of intonation unit to test. This will result in a platform for evaluating hypothesis about the importance of different aspects of intonation.

7. References

- [1] A. Botinis, B. Granstrom, and B. Moebius, "Developments and Paradigms in Intonation Research," *Speech Communications*, vol. 33, pp. 263–296, July 2001.
- [2] A. Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of F0 contours using a superpositional model," *Speech Communication*, vol. 40, pp. 535–549, 2003.
- [3] R. Sproat, *Multilingual Text-to-Speech Synthesis*, Kluwer, 1998.
- [4] J. R. Bellegarda, K. Silverman, and V. Anderson, "Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation," *IEEE Transaction on Speech and Audio Processing*, vol. 9, no. 1, pp. 52–66, January 2001.
- [5] K. Ross and M. Ostendorf, "A dynamical system model for generating f0 for synthesis," in *Proceeding ESCA Workshop On Speech Synthesis*, 1994.
- [6] K. E. Dusterhoff A. W. Black and P. A. Taylor, "Using Tilt Intonation Model: A Data-Driven Approach," in *Data-Driven Techniques in Speech Synthesis*, R. I. Damper, Ed., chapter 9, pp. 199–213. Kluwer Academic Press, 2001.
- [7] 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.
- [8] 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.
- [9] D. Escudero and V. Cardeñoso, "Experimental evaluation of the relevance of prosodic features in spanish using machine learning techniques," in *Proceedings of Eurospeech 2003*, September 2003.
- [10] E. Alarcos, *Gramática de la Lengua Española*, Real Academia Española, 2002.