ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Proposal to Quantitatively Select the Right Intonation Unit in Data-Driven Intonation Modeling

D. Escudero-Mancebo, V. Cardeñoso-Payo

{descuder,valen}@infor.uva.es
Department of Computer Science
University of Valladolid, Spain

10.21437/Interspeech.2004-286

Abstract

In this work, we provide a procedure for the systematic evaluation of the quantitative impact of the selection of the basic intonation unit for data-driven intonation modeling. Taking advantage of the corpus based modeling technique previously developed, we show how the number of prosodic features selected and the kind of basic unit determines the final prediction RMSE of the synthesized F0 profiles. This provides a means to fully characterize a given corpus and, also, a procedure to test how 'a priori' linguistic knowledge corresponds with real samples in that corpus. Although all the results shown are for Spanish, the methodology can be readily applied to any other similar language for which a similar annotated corpus is available.

1. Introduction

The main goal of intonation modeling techniques is, from the engineering point of view, providing an accurate correspondence between intonation form and function. Functional information is symbolically represented by means of vectors of prosodic features chosen from feature sets defined under a number of linguistic knowledge constraints and assigned in terms of the conveyed message and the contextual information. Depending on the specific technique, intonation form is modeled either in terms of a closed set of F0 contour shapes or in terms of sets of parameters controlling a production model, be it physiologically based or not. Data-driven techniques aim to provide this function-form mapping by means of automatic trainable methods which can build the correspondence directly from samples stored in an annotated corpus. The reader is referred to [1] for an excellent review of these and many other reported in the state of the art.

A key initial step of any modeling technique is to decide which type of basic intonation units are to be used, since they provide the building blocks both for the prosodic function labeling process and for the compositional building of the associated form of the final intonation profile of a complex utterance. Accentual phrase[2], accent groups[3], ToBI transcriptions[4], syllables[5], or intonation events[6] are just examples of relevant proposals which have been supported across different models for different languages. In most cases, selection criteria are based in well established 'a priori' linguistic knowledge for the studied language or prosodic family of languages.

In this work, we give an alternative view of this problem. We will provide a means to quantitatively measure the impact of the selection of the intonation unit in the framework of our corpus based intonation modeling methodology MEMOInt (MEthodology for MOdelling Intonation) [7]. In [8, 9], we presented a technique to measure the relative importance of prosodic features, in order to properly characterize stress groups when they are used as the basic units. Here, we propose an extension of this technique which can be readily applied to quantitatively evaluate which is the most suitable intonation unit to be used for a given corpus, chosen among a set of alternatives.

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

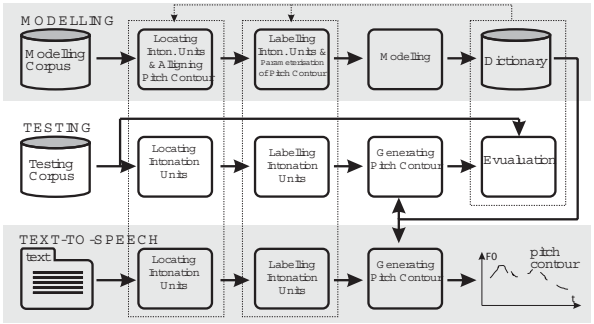


Figure 1: Functional diagram of MEMOInt.

This paper will be structured as follows. In section 2 we will briefly review MEMOInt. Then, we report a discussion of the different types of intonation units used for this work and their set of prosodic features. The experimental procedure will be described in section 4 and a discussion of the main results is presented in section 5. Conclusions and future work proposals appear in the final section.

2. Overview of MEMOInt

MEMOInt aims providing a methodology for intonation modeling based on corpus. The three main functional stages are depicted in figure 1. In the modeling phase, a dictionary of prosodic classes is automatically built, each one representing a set of given annotated prosodic features. In the testing phase, this database of models is validated against unseen corpus samples in order to extract a quality measure of the representativeness of each entry in the dictionary of models. Finally, when applied to TTS, every class is interpreted as a model in order to generate the synthetic pitch contours associated to a given prosodic function in the production phase.

When using MEMOInt, decisions along three key axes affect the overall performance of the methodology: selection and location of intonation units, set of prosodic labels to be asso-

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

Table 1: Cardinality of the prosodic features. SE: sentence, SG: Stress Group, IG: Intonation Group, Syl: Syllable and Phon: 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.

ciated to every initial sample in the corpus and the parameterization of each intonation unit. In order to locate and segment spoken utterances into annotated intonation units, an intonation unit has to be initially chosen. As will be shown in the following sections, in this paper we propose a systematic procedure to guide the selection process among a given set of alternatives for the intonation unit.

Once the spoken utterances are segmented into elementary intonation units, the resulting sequence has to be annotated in the labeling phase. The resulting labels are obtained by assigning values to a set of prosodic features and assuming these sets of values are going to be representative of the shape of the pitch contour of the intonation units. Choosing the best set of prosodic features systematically has been the central contribution of previous works [8]. In the parameterization step, a set of quantitative acoustic parameters which accurately represent the shape of the pitch contour are computed from data for every annotated intonation unit.

Modelling consists of grouping together into the same class the intonation units that share its prosodic features values. The representative model of each class will be the statistical distribution of the acoustic intonation parameters of the intonation units belonging to this class and computed from the modeling corpus. Generalization capabilities of the models increase 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 production for TTS. 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.

A relevant feature of MEMOInt is its ability to provide useful feedback on specific prosodic information about the corpus which can be successfully exploited to improve decisions of previous tasks of the pipeline (as shown by backwards lines in figure 1). In particular, the reports from the testing stage give comparative information for different labeling and segmentation options. Here, we show that this information is also useful to compare the behavior of MEMOInt with different types of intonation units.

3. Prosodic features and intonation units

The set of candidate intonation units stems from well founded definitions of intonation units for Spanish. Syllables can be segmented out using classical Spanish syllabication rules [10]. A

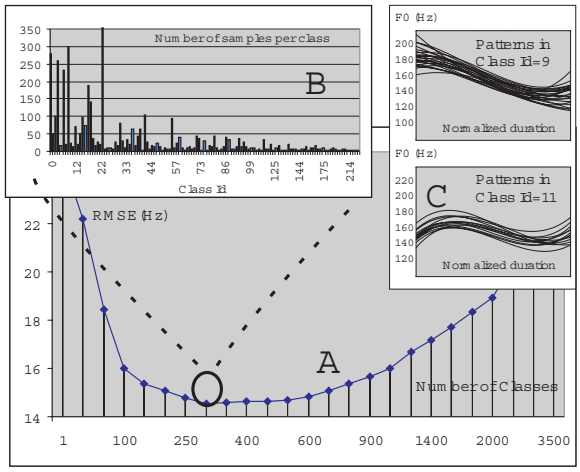


Figure 2: (A) Prediction Error in the iterative grouping. Type of Intonation Unit = Stress Group. All the available prosodic features are used. (B) Number of samples per class at the minimum *Number of Classes=300*. (C) F0 patterns in two of the classes when *Number of Classes=300*

stress group is a set of consecutive words in which only the last one is accented. An intonation group is a set of stress groups separated either by a long pause or by a significant inflexion in the F0 contours (what is meant by long and significant has to be determined experimentally from the corpus itself, of course).

Table 1 shows the number of values of the prosodic features of these three types of prosodic units for Spanish. Obviously, the hierarchical relation among types of intonation units arises in the sets of features. In this study, we have selected only three kinds of prosodic features: linguistic (typeSE, posSTSG, sylRelPos, posSTIniSG, posSTFinSG), position oriented (posIGSE, posSGIG, relPosSyl) and length related (nSGSE, nSGSE, nSylSE, nPhonSE, nSGIG, nSylIG, nPhonSG, nPhonSyl). Features which describe the phonetic [3] or syntactic [2] structure have been left out for the moment because they could not be easily annotated in our corpus. In any case, they could be processed in the same way we describe in the rest of the paper when available.

An uniform coverage of the corpus is chosen as the criterion to fix the cardinality of the set of possible values for every prosodic feature (see table 1) and the same number of possible values is assumed for every unit in the corpus. This is because we try to minimize the impact of cardinality imbalance, since our focus is more related to show the usefulness of the methodological proposal than to get the best set of results for our specific corpus.

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 an iterative grouping of pairs of classes. A maximum similarity criterion is applied in each step using an euclidean distance between centroids. Grouping two classes implies a precision loss

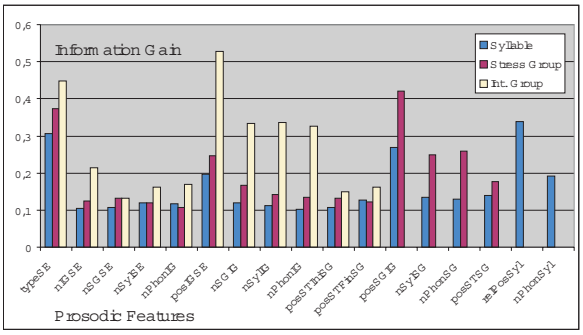


Figure 3: *Information Gain* of every prosodic feature using a Kmeans clustering of the samples of the corpus (see [9] for the *Information Gain* metric)

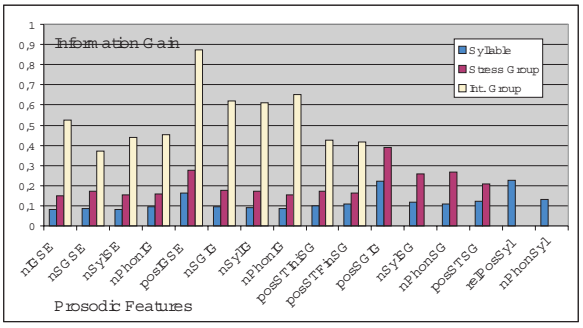


Figure 4: *Information Gain* of every prosodic feature with respect to the best clustering obtaining with the iterative grouping classes method.

but brings a generalization gain.

Grouping two classes implies building a new dictionary. If the prediction error obtained with this new dictionary when synthesizing pitch contours gets smaller, then the new classification is preferred. Repeating the process, we come to a balance between precision and generalization, to obtain an optimum dictionary. The merging process is stopped when the loss of precision forces unwanted prediction results.

Figure 2 illustrates the results of this iterative merging process. The minimum of the RMSE plot (figure 2.A) gives the best dictionary when the merging of classes proceeds from right to left in the plot (*number of Classes=300*). The number of samples per class in the best dictionary of models, which provides the correspondence between prosodic features and F0 patterns, is shown in figure 2.B. Patterns of F0 for samples belonging to two different classes of the dictionary are shown in figure 2.C.

Since some of the classes of the dictionary of models could be void, we have proposed a strategy to cope with this lack of correspondence in the production phase. Several dictionaries are kept and the most informative (higher number of classes) is used first. When a unit with prosodic labels corresponding to a void class appears, a dictionary selection mechanism is triggered which provides the one with the highest number of classes for which a non void class for the same label can be found. This clearly ensures that the synthetic pitch contour is associated with the observations of the corpus which share with it the highest amount of information.

A relevance ranking of the prosodic features (see [9]) was applied to build the collection of dictionaries for each of the three types of intonation units compared in this paper. Figures 3 and 4 provide a comparison of the information gain generated by each prosodic feature when a conventional Kmeans clustering technique (figure 3) and our iterative merging procedure (figure 4) are used. They show that conventional Kmeans can be confidently used to rank the relative importance of prosodic feature when orderly incorporating them to build the ranked collection of dictionaries. The number of clusters in the Kmeans case were: 800 for syllables, 250 for stress groups and 100 for intonation groups. This ensures a balance in the number of samples per class in all cases.

The corpus was the same than in previous works[1]. It contains 14971 syllables, 4665 stress groups and 1747 intonation groups. Since the number of interrogative and exclamatory sentences was below 5%, we restricted to declaratives for this study. As a difference with previous works, we used raw F0 contours to build models and compute prediction errors. Slightly better results could have been obtained if the same smoothing mechanism had been applied but this was not considered relevant in this case.

[1] Gently provided by TALP group of UPC university.

5. Results and discussion

We have applied our iterative grouping method to each element of the ordered set of classifications built for every one of the three intonation units compared in this study. For each type of intonation unit, prosodic features are orderly incorporated to the classification in order to get a new (more informative) element of the set of dictionaries for that intonation unit. Results for the prediction RMSE as a function of the number of classes for each set of prosodic features and type of intonation unit are gathered in figure 5. The minimum of each curve gives the optimum number of classes for the specific set of prosodic features and it is for this number of classes that the results on Information Gain presented in figure 4 were computed. For each type of intonation unit, an additional prosodic feature is incorporated to the current set of features to get each RMSE curve, starting from the initial sets reported in the figure.

As a general rule, higher number of prosodic features gives better prediction results. An exception arises when syllables are used, where incorporating more than 12 features gives worse results, which could be explained in terms of overspecialization in the modeling, leading to a higher impact of scarcity problems. Although the best overall results are obtained when the syllable is used as the basic unit, they are comparable to the ones with stress group, which could not compensate the higher complexity in terms of number of classes. The most relevant features are those that refer to the type of intonation unit which is being studied, except for typeSE and posIGSE which are important for every type of intonation unit. Overall, the most relevant features are typeSE and posIGSE, posSGIG, nSGIG, nSylSG and relPosSyl, which give information on the relative position of the units with respect to their contained units. Although these results have been derived from our specific corpus, they match with the ones reported in the state of the art bibliography for Spanish intonation.

Finally, results suggest that some prosodic features provide almost the same amount of (small) improvement. Careful selec-

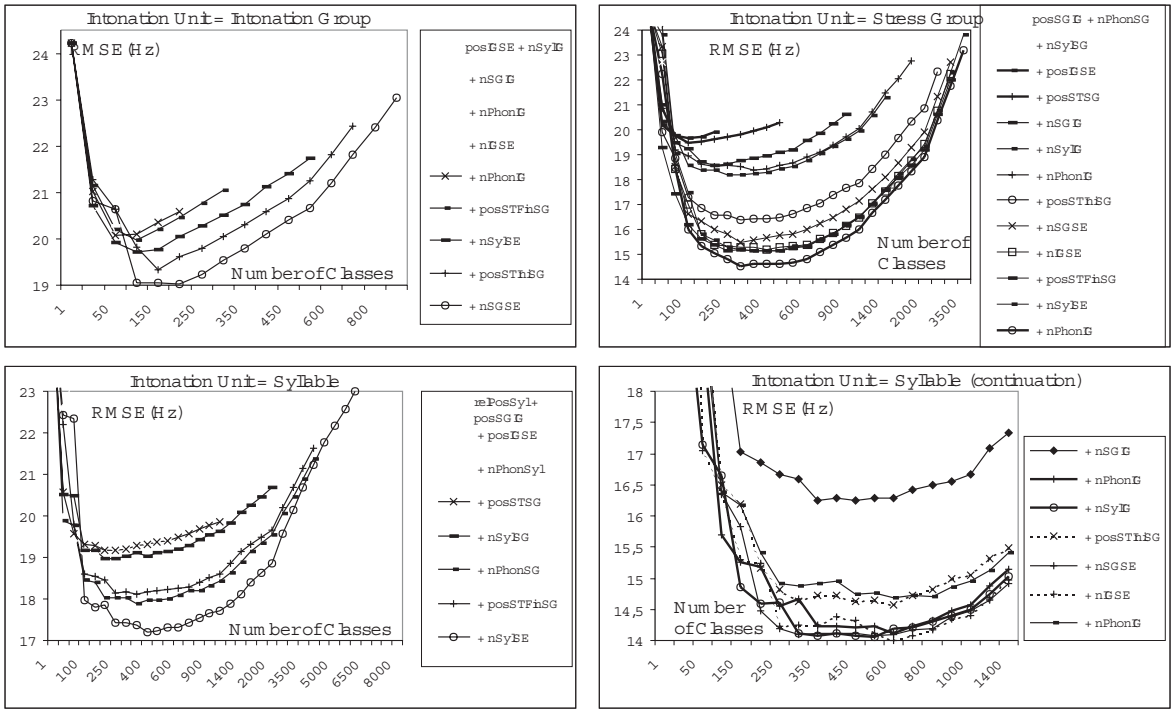


Figure 5: *Prediction Error* of the samples in the test corpus with respect to the number of classes in the iterative grouping process. Three different intonation units are considered. For each one, a family of RMSE is presented, one for every incremental set of prosodic features.

tion of the set of chosen prosodic features could give acceptable results while keeping scarcity problems to a bare minimum.

6. Conclusions and future work

A procedure to analyze quantitatively the impact of the selection of intonation unit on the prediction error has been presented in this work. This has brought a further step towards the validation of MEMOInt as a valuable data-driven intonation modeling methodology. Results for the representative subset of Spanish language contained in our corpus show that both the syllables and stress groups could be most adequately used as intonation units. The ranking of prosodic features derived from the analysis matches well known results for this language. Most important is the fact that MEMOInt can clearly provide a means to evaluate theoretical prosodic hypothesis on the importance of several aspects of intonation if the right corpus is used and, the other way round, is useful to quantitatively evaluate the prosodic quality of a given corpus if the 'a priori' prosodic knowledge is at hand and incorporated.

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 contoures using a superpositional model," *Speech Communication*, vol. 40, pp. 535–549, 2003.

[3] R. Sproat, *Multilingual Text-to-Speech Synthesis*, Kluwert, 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*, May 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*, May 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 Academina Española, 2002.