Optimized Selection of Intonation Dictionaries in Corpus Based Intonation Modelling

David Escudero and Valentín Cardeñoso

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

10.21437/Interspeech.2005-567

Abstract

Data scarcity in corpus-based intonation modelling for TTS applications is addressed. We propose to apply a searching process to a list of dictionaries of classes of intonation patterns previously trained from corpus to avoid problems associated with the scarce number of samples in the classes. Results indicate that better results are obtained in comparison with previous alternatives where the probability of predicting a less representative intonation pattern has been shown to be higher.

1. Introduction

Corpus based systems could provide the best engineering solution for the challenge of fast adaptation to new speaker voices and speaking styles. As far as intonation modelling is concerned, one of the most important problems of corpus based systems is the scarcity of data available in the training corpus. In [1] we defended a strategy to cope with this problem based on the use of a list of dictionaries of classes of pitch patterns and on a process of aggregation of such classes. The list of dictionaries, sorted in terms of informative capabilities, showed to improve significantly the prediction errors in comparison with previous approaches [2]. In this work, we present a new improvement where less informative dictionaries can be selected in the case they offer better prediction results.

Models of intonation attempt to find out the relationship between a set of prosodic features (PFs) of the message and pitch contours (characterised by sets of parameters like in TILT[3] or Fujisaki models[4]). Different approaches to represent intonation and to model the relationship between acoustic and linguistic information can be found in the state of the art (as reviewed in [5]). Scarcity problems arise when corpora don't cover all the possible combinations of PFs (few or no sample for a given combination). Under these situations, it is required to predict intonation for PFs sets which are not properly modelled if they are at all. Of course, there is still a possibility to redesign a corpus and include more samples. However, it is not always possible to design and acquire corpora which include the huge number of required combinations. As an example, there are around 27 millions of possible combinations of PFs reported in [6] (although some of them are obviously absurd) while the corpus used had less than 3000 different combinations. Usually, this scarcity is to be avoided delivering better parameter selection procedures for the classification algorithms (neural networks, regression tress, lineal regression, ...) in those special cases where there are not enough training data or anyone at all. In spite of this, there are situations where the number of possible combinations of PFs which are not properly modelled could represent a high percentage and the naturalness of predicted pitch contours could be highly compromised if a simplistic solution is provided.

This work has been partially supported by MCYT contract TIC2003-08382-C05-03

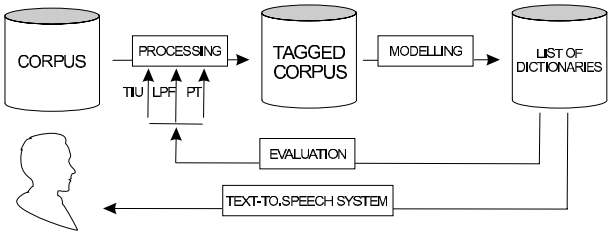

Figure 1: Functional diagram of MEMOInt. TIU (Type of Intonation Unit), LPF (List of Prosodic Features) and PT (Parameterisation Technique) are the parameters of MEMOInt.

The methodology for modelling intonation we defend (MEMOInt) is also sensitive to this problem because the basic idea is to build classes of pitch patterns according to its shape and its prosodic function. When the amount of observations of a given class is limited, the problem is that the associated model won't generalize properly. We show in this communication that to apply a searching process to locate the dictionary of classes to use completes the aggregation process already defended.

The MEMOInt approach is based on the concatenation of sequences of pitch patterns to configure the final pitch contour. We predict pitch contours in sub-units smaller to the sentence. If any of these units is not properly predicted the whole perceived quality can decrease dramatically. This fact is not measured by the standard objective metrics (like RMSE of correlation factor) which usually hide these local deviations. In this communication we show the importance of this effect and how it can be reduced by searching for less specific classes but that generalize better.

We start summarizing MEMOInt and presenting the parameters of the methodology considered in this study. Then we describe the experimental procedure to compare the new searching strategy with previous approaches. Finally, results are discussed and we present some of the conclusions on the benefits of this new approach.

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

Prosodic Feature	Acronimus	IG	SG1	SG2	SG3	Syl
Type of sentence	typeSE	3	3	3	3	3
Number of IG in the SE	nIGSE	4	10	10	10	4
Number of SG in the SE	nSGSE	4	6	6	6	4
Number of Syl in the SE	nSylSE	4	6	6	6	4
Number of Phon in the SE	nPhonSE	4	6	6	6	4
Position of IG in the SE	posIGSE	5	7	7	7	5
Number of SG in the IG	nSGIG	5	6	6	6	5
Number of Syl in the IG	nSylIG	4	7	7	1	4
Number of Phon in the IG	nPhonIG	4	6	6	6	4
Position of ST in Initial SG	posSTIniSG	3	3	3	3	3
Position of ST in Final SG	posSTFinSG	3	3	3	3	3
Position of SG in the IG	posSGIG		5	5	5	6
Number of Syl in the SG	nSylSG		9	9	1	5
Number of Phon in the SG	nPhonSG		6	6	4	4
Prominence	Accented		2	2	2	2
Position of ST in the SG	posSTSG		3	3	3	3
World Boundary	SGBorder		1	4	3	1
Position of the Syl in the SG	relPosSyl					6
Number of Phon in th Syl	nPhonSyl					4

Table 1: Prosodic features associated with the different intonation units take into account

is a list of dictionaries of models, representing the intonation of the corpus. The representativeness of this list of dictionaries is evaluated in the second stage. Finally, the models of the dictionary are used to generate synthetic pitch contours in TTS applications.

Processing tasks determine the results of MEMOInt: locating intonation units, its labelling and its parameterisation. Locating intonation units requires establishing a priori the type of intonation unit to be used (syllable, stress group ...). Then, the utterances of the corpus are divided into a sequence of such intonation units. The labelling task results in assigning values to a list of prosodic features (accent, position etc...) characterizing the intonation units. The parameterisation task computes a set of quantitative parameters representing the shape of the pitch contour of the intonation units. This task is strongly dependent on the TIU, LPF and PT parameters (see figure 1) which can be determined by the feedback provided by the evaluation task.

Modelling consists of grouping together into the same class the intonation units that share its prosodic features values. A dictionary is a set of classes. The model of each of the classes is the statistical distribution of the acoustic intonation parameters of the intonation units belonging to the class observed in the modelling corpus. The generalization capabilities of the models increase after an iterative grouping process. This process and the use of list of dictionaries will be briefly described 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.

Because we model a list of dictionaries instead of only one, it is important to have a strategy to select the most suitable one depending on the intonation unit to model. The subject of this communication is to propose a new strategy to do so and to compare it with previous proposals.

3. Corpus and Parameters

In [7] we applied MEMOInt with Stress Groups (SG1), Intonation Groups (IG) and Syllables (Syl) as different types of intonation units. Typical definitions of these intonation units for Spanish (see [8]) were used. Here two additional alternative definitions of the Stress Group are also considered. SG2 definition is inspired in [9] and the stress group is composed by a stressed syllable plus the following unstressed ones. SG3 definition is inspired in [3] and states stress groups as the stressed syllable plus the preceding and the following one.

```
CreateListOfDictionaries(LD, LPFs, LPFu){
    if ( LPFu is not empty){
        PFBest = SelectBestPF(LD, LPFs, LPFu);
        LPFs=LPFs + PFBest;
        LPFu=LPFu - PFBest;
        D = CreateDiccionary(LD, LPFs);
        LD = LD + D;
        CreateListOfDictionaries(LD, LPFs, LPFu);
    }
}
```

Figure 2: Procedure to build the list of dictionaries. `LD` is a List of Dictionaries; `LPFs` and `LPFu` are the list of prosodic features already considered and not yet considered respectively; `SelectBestPF` searches for the best prosodic feature in `LPFu` to add to `LPFs` to build a new Dictionary; `CreateDiccionary` creates a new dictionary `D` based on the features `LPFs` and taking into account **LD**. First invocation to the function must be done passing `LD` and `LPFs` void and `LPFu` assigned to the list of features to consider.

Table 1 shows the five different types of intonation units, its prosodic features, and the different number of values of the prosodic features. Note the hierarchical relation between the types of intonation units and its projection into the features. We have only selected four families of prosodic features: linguistic ones typeSE, posSTSG, sylRelPos, posSTIniSG, posSTFinSG, SGBorder); position ones (posIGSE, posSGIG, relPosSyl) and length ones (nSGSE, nSGSE, nSylSE, nPhonSE, nSGIG, nSylIG, nPhonSG, nPhonSyl); Accented is related with a prominent pronountiation (accent) manually set.

The corpus used is the same corpus we have already used in previous works[1]. 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. Corpus is divided in 75% for modelling and 25% for testing. The modelling part is also divided reserving 25% for training the dictionaries.

The acoustic parameters to be used are the projection of the control points on the Bezier fitting curve of the F0 contours (more details in [2]). Prediction errors and statistical models are obtained with raw F0 contours. We test different number of parameters as will be showed in section 5.

4. Experimental Procedure

The process to build a single dictionary is based in an aggregation of classes criteria and it is described in detail in [1]. Briefly, each combination of prosodic features determines one class in the initial dictionary of models. Provided there is few intonation groups of certain class, its model will be not characteristic and its use in prediction can be problematic. Joining classes under a maximum similarity criterion increases the number of samples per class expecting no loss of representativeness. The new dictionary obtained after a class aggregation 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.

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 mod-

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

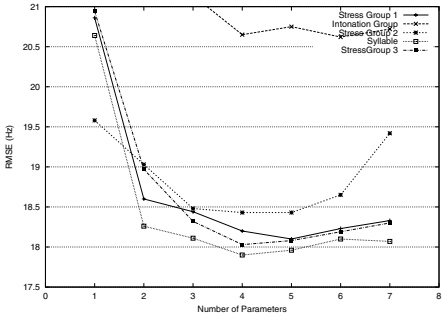

Figure 3: Prediction Error versus the number of parameters. 1 parameter means one single value to fit the F0 points of the Intonation Unit, 2 parameters is a line, 3 parameters a parabola... (see [11])

elling corpus. The more prosodic features in the input the more the number of void classes. But, an intonation unit of any of such void classes can appear when using the dictionary to generate synthetic pitch contours. To cope with this situation we don't use a single dictionary of models but a collection of dictionaries.

The dictionaries in the list differ on the number of prosodic feature that have been used to build them. Figure 2 presents an algorithm to build the list of dictionaries $LD^N = (D^1, .., D^N)$ where N is the number of input prosodic features and the number of dictionaries in the list; D^t is the dictionary built with t prosodic features. If D^t was built with the list of prosodic features LPF^t, then LPF^{t+1} and LPF^t differ only in one prosodic feature: PF^t. PF^t is selected so that the prediction errors of the training corpus samples are minimized.

D^N is the most informative dictionary and it could be the best dictionary to use if the number of samples in the corpus is enough. As this is not the case, we have to explore the list of dictionaries to choose the most suitable. The way to do so determines different strategies compared here. These strategies are: (1) **DP** Default Pattern: a default pattern is assigned when the class of the intonation unit in the D^N is void; (2) **NEMID** Non Empty Most Informative Dictionary: select the biggest t so that D^t classifies the intonation unit into a non void class; (3) **BD** Best Dictionary: select the dictionary which minimizes the average of the predicted errors for the training samples which belong to the class of the predicted intonation unit.

Objective evaluation will be used to compare these three alternatives. Subjective tests have assessed the obtained results.

5. Results

Figure 3 shows the training results depending on the intonation unit and number of parameters to use. No class aggregation process was applied to avoid a possible influence of this process. We select SG3 with 4 parameters because prediction results are similar to the results obtained with Syllables (the best ones). Notice that both can be used to compare the strategies proposed here. We choose SG3 because it has few input prosodic features making it faster to train due to the computational cost of the grouping process.

Figure 4 permits to compare the three different strategies. Different lines in plots refer to different lists of dictionaries. The legend of the lines is the feature selected to add to the previous

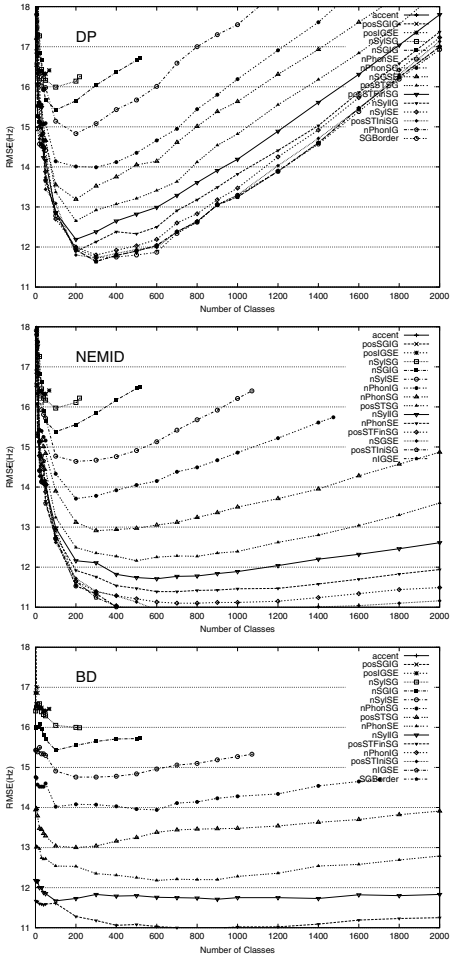

Figure 4: Prediction error obtaining during the process of construction of the list of dictionaries using the aggregation process making use of the three different strategies to select the dictionary:**DP**, **NEMID** and **BD**

list of features to configure the new dictionary. Different points on the lines are the training prediction results obtained during the aggregation process: the number of classes is reduced from right to left as the classes are grouped together. The minimum of each of the lines indicates the best situation.

The trajectory of the prediction results lines differs from one plot to each other. These trajectories lead us to discard **DP** because results are the worst ones. Prediction results are similar in **BD** and **NEMID** around the optimum, but they clearly differ in the rest of the interval: The results obtained with LD^t are always better than the results obtained with LD^u with $t > u$ in **BD** independently of the number of classes. This behaviour is justified by the selection mechanism of the best dictionary: If the new dictionary added to the list does not predict better than the less informative ones it will not be selected, and results will be at least as good as the results obtained with the previous list of dictionaries.

Table 2 shows that **BD** offers better results than **NEMID** both when the dictionary D^N is used and when it is not. **NEMID** does not select D^N if the corresponding class is void. **BD** does not select D^N when another dictionary predicts better. The

	NEMID					BD				
	D^N		$!D^N$			D^N		$!D^N$		
LD^N	RMSE (Hz)	Corr	RMSE (Hz)	Corr	% $!D^N$	RMSE (Hz)	Corr	RMSE (Hz)	Corr	% $!D^N$
LD^1	21.47	0.59	0.00	0.00	0.00	21.47	0.59	0.00	0.00	0.00
LD^2	18.90	0.70	0.00	0.00	0.00	18.42	0.70	23.08	0.67	10.97
LD^3	18.31	0.72	0.00	0.00	0.00	17.68	0.77	19.78	0.58	38.48
LD^4	17.85	0.74	32.07	0.47	0.82	15.83	0.78	20.90	0.68	43.29
LD^5	17.10	0.76	28.93	0.43	1.42	18.67	0.67	16.97	0.78	66.99
LD^6	16.05	0.79	28.10	0.51	3.10	14.35	0.80	18.96	0.73	53.93
LD^7	14.80	0.82	27.33	0.52	2.68	12.80	0.85	19.55	0.71	45.60
LD^8	13.92	0.84	26.53	0.68	3.55	14.71	0.84	15.12	0.81	53.58
LD^9	13.05	0.86	21.18	0.73	7.84	11.71	0.88	16.28	0.81	49.93

Table 2: Training prediction results. D^N refers to prediction errors of samples predicted with the dictionary D^N of the list. $!D^N$ refers to the other case.

percentage of cases where D^N is not used is smaller in the **NEMID** alternative due to the simplicity of this strategy. The importance of these higher RMSE values can be seen in figure 5. Each of the plots zooms one of the lines of figure 4 and separates the results obtained when the D^N is used and when it is not. Although mean results are similar for both cases, the D^N and $!D^N$ lines differ significantly in **BD** and **NEMID**: around the optimum point, the **NEMID** $!D^N$ line grows fast and the **BD** one is bounded. Although the percentage of predicted samples under this situation is smaller in the **NEMID** case, they can be enough to decrease significantly the quality of the perceived intonation as it has been tested perceptually. The reason for this increase on the prediction error can be found in that **NEMID** case select the first dictionary where the class is empty, but it is not taken account that the class can be few representative if the data are scarce.

This behaviour is also observed when the test corpus is predicted and the results have been contrasted with perceptual tests.

6. Conclusions

Better intonation prediction results can be obtained using a searching strategy to find the most suitable dictionary in a list. The class of the intonation unit to predict must be the searching criteria. Main reason why results improve with respect to previous strategies is that we bound the distance between predicted and real F0 contours.

The fact of selecting classes belonging to less informative dictionaries doesn't imply worst prediction results in some cases. The searching strategy permits to select the most suitable dictionaries depending of the class to predict.

Next step is to study the shape of the patterns and the relation with the prosodic features exploring the relation between different levels of the dictionaries. This will give us information about the intonation in the corpus and about the real influence of the prosodic features in the patterns shape.

7. References

[1] V. Cardeñoso and D. Escudero, "A strategy to solve data scarcity problems in corpus based intonation modelling," in *Proceedings of ICASSP 2004*, 2004.

[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] P. Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *Journal of Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

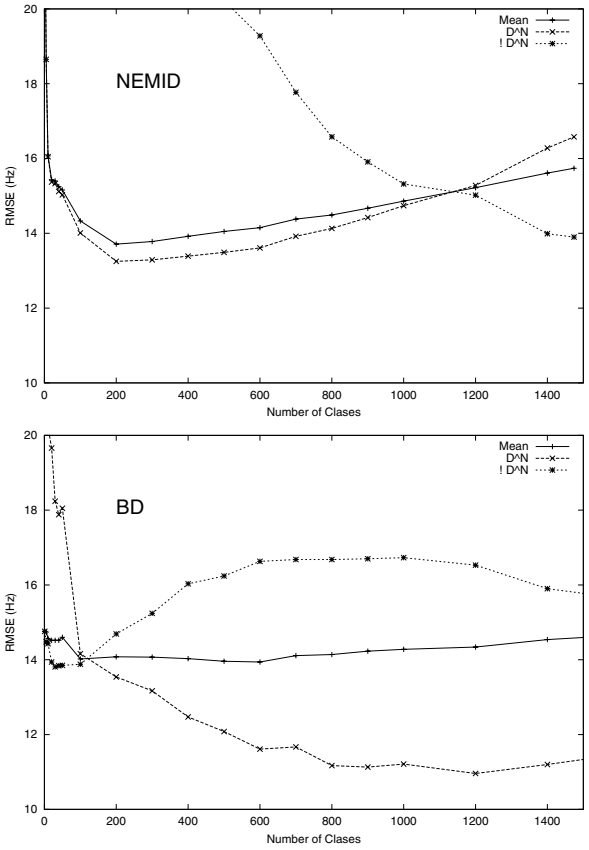

Figure 5: Prediction results obtained using LD^7. We separate the results obtained when D^7 is selected from the results obtained when it is not. The minimum value of the line titled *mean* determines the number of classes that will be used.

[4] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of Acoustics Society of Japan*, vol. 5, no. 4, pp. 233–242, 1984.

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

[6] 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.

[7] D. Escudero and V. Cardeñoso, "Impact of the selection of the constructive type of intonation unit in a data-driven intonation modelling technique," in *Proceedings of ICSLP 2004*, September 2004.

[8] E. Alarcos, *Gramática de la Lengua Española*, Real Academina Española, 2002.

[9] J. van Santen, "Quantitative modeling of pitch accent alignement," in *Proceedings of ISCA Prosody 2002*, 2002.

[10] R. Sproat, *Multilingual Text-to-Speech Synthesis*, Kluwert, 1998.

[11] M. Plass and M. Stone, "Curve-Fitting with Piecewise Parametric Cubics," *Computer Graphics*, pp. 229–239, July 1983.