On the Comparison of Catalan-Spanish Intonation Systems using Statistical Corpus Modeling and Objective Metrics

David Escudero-Mancebo, Valentin Cardenoso-Payo [1], *Antonio Bonafonte-Cavez*[2]

[1]Departament of Computer Science University of Valladolid, Spain
[2]Department of Signal Theory and Telecommunications, Polythecnic University of Catalonia, Spain
descuder@infor.uva.es, valen@infor.uva.es, antonio.bonafonte@upc.es

Abstract

This communication presents an ongoing research on the definition of a methodology to compare the intonation of two different corpora. The two corpora that we compare here, try to be representative of the Spanish and Catalan intonation respectively. As a consequence, the comparison reported here, projects the most relevant differences between the Catalan and Spanish intonation systems. First we model the intonation of the corpora using the MEMOInt methodology and we confront the models obtained by confronting the F0 patterns that share its prosodic function. An objective metric guides the identification of the most distant F0 patterns. The differences can be visualized and listened in perceptual tests. Finally we discuss about the weakness of this ongoing research and about future applications.
Index Terms: comparing prosody, modeling intonation, Catalan intonation, Spanish intonation.

1. Introduction

The comparison of the prosody in different characteristic corpora can be an important source of information with benefits in various fields of speech technologies enumerated below. In contrast with other approaches, we consider this problem as a data mining project to retrieve the potential differences between two given corpora. The methodology for modeling intonation named MEMOInt [1] already used in predicting intonation is helping in this challenge.

We enumerate a set of benefits of the availability of a tool for comparing prosody between corpora. (1) Text to speech systems could use the information resulting from the comparison to adapt a given voice to mimic alternative styles or accents [2]. (2) Speaker recognition systems could benefit from the models characterizing the prosody of different languages or different type of speakers to discriminate between them. (3) Educative programs could identify foreign accent utterances contrasting a recorded corpus with respect to a reference one. (4) As a source of information in linguistics where prosody and intonation is still a challenging field of research.

There are already some studies in the state of the art concerning with the comparison of prosody. Some of them limit the study to a concrete aspect such as a given type of ToBI pattern or a part of questioning sentences [3] [4] or are based on examples [5][6]. These approaches have the weakness to renounce to analyze the corpora as a whole discarding a-priori the importance of other aspects to discover. Other approaches limit their scope to get statistics of the F0 contours (raw F0 [7] or Fujisaky parameters [8]) and to contrast results among different corpora. The results obtained with these approaches are insufficient in quantity and in quality to be used in the potential applications mentioned in the previous paragraph (except speaker recognition applications [9])

In this paper we propose to use a methodology that captures the prosodic information of a given corpus modeling the correspondence between form and function of the F0 contours setting up this correspondence in a common representation named the graph of classes. Visual and contrastable information about the differences between corpora can be obtained by confronting graphs of classes. Here we present the comparing technique applied to a Spanish and a Catalan corpus. The results must be considered as a part of an ongoing research to be matured in future works focusing on the aspects enumerated in the discussion section of this communication.

First we briefly review the MEMOInt methodology; second we describe the experimental procedure, presenting the corpora used in this study and the comparison technique; third we present the results of the comparison and we end with conclusions and future work.

2. MEMOInt: Methodology for Modeling Intonation

The MEMOInt methodology is used for data mining a given corpus by the application of the well known techniques *agglomerative clustering*[10] and *sequential learning*[11] with the goal to obtain useful and contrastable information about the intonation of the corpus. The corpus is considered a set $C = \{u_i, \quad 1..N\}$, where u_i is each of the N units of intonation identified in the corpus (the type of intonation could be the syllable, stress groups, intonation group etc...). Every u is a duple $u = (f, p)$. p are a set of acoustic parameters that represent the form of the F0 contour of u. p are obtained automatically from the F0 contours in the parameterization stage. On the other hand, f are a set of prosodic features that represent the prosodic function of u. They reflect different aspects determining the intonation like accent, grammatical structure of the sentences, size of the intonation units, emotions, type of sentence ... These features are to be extracted automatically from text or manually labeled in the corpus. The MEMOInt goal is to infer a matching between f and p, that is, to infer the correspondence between function and shape of the intonation observed in the corpus.

MEMOInt applies agglomerative clustering to the intonation units of the corpus. The initial cluster is determined by the combination of the prosodic features values. The agglomeration is driven by a given inter-class maximum similarity criterion. The stopping criterion is determined by the prediction capabilities of the new clusters as the classes are to be used in text-to-speech applications (see [1] for further details). The agglomer-

ative process outputs the correspondence between p and f by keeping track of the different values of the merged features. An index is built to assign one class in the final configuration to any f combination. These classes can be used in text-to-speech where f obtained from text and p can be used to generate a synthetic F0 contour. We call *dictionary* to the combination of the index, made of a sequence of f, and the clustering associated to it.

The more the number of f involved, the worst the scarcity problem. To cope with it, we follow sequential learning so that different clusters are constructed by using different number of f. In every step MEMOInt selects the f which inclusion implies better prediction results. As result we obtain N different dictionaries, as many as F considered: the *list of dictionaries* already mentioned. Given any combination of f we select the cluster that predicts more accurately the sample according to the observations in the training stage. The list of dictionaries can be seen as a graph of classes where the values of the features permit to navigate the classes of the dictionaries.

The application of MEMOInt to a given corpus results: (1) A ranking of importance of the different features affecting the intonation (2) Visual and contrastable information of the relationship between F0-patterns and the prosodic features that justify the prototypical F0 movements displayed as a graph of classes (3) A tool to produce synthetic intonation to be used in text-to-speech applications. Next section explains how to use the graph of classes to compare the intonation profiles of two different corpora.

3. Experimental Procedure

MEMOInt is applied sequentially to a given Catalan corpus and to another Spanish one described below. The respective results are to be compared shedding light to the identification of the aspects that make Catalan and Spanish intonation different. First we describe the corpora, second we present the procedure applied to contrast the results and third we describe the MEMOInt parameters used in the experiment.

The aim of the Catalan corpus was to develop the question-answering module of a dialog system to give meteorological information. The aim of the Spanish corpus was to develop a general propose text-to-speech system [12]. The Catalan corpus is about half an hour of reading speech, with 476 sentences (357 declarative ones) with 3447 stress groups (2799 in declarative sentences). The Spanish one is about one hour of reading speech, with 677 declarative sentences (4366 stress groups). Both corpora have been recorded from the same professional actress in studio conditions using a laryngograph device to collect the F0 samples.

MEMOInt is applied to one of the corpus following the procedure explained in [1]. As result we obtain a ranking of prosodic features and the respective graph of classes. This ranking is imposed when MEMOInt is applied to the other corpus so that the two graphs result aligned. Once the two graphs align, there is a correspondence one to one between the nodes of the graphs. A distance measurement is applied to the matching classes to sort the nodes. It is expected that the most distant nodes indicate the differences between the intonation systems of the respective corpora. Distant nodes are visually and perceptually analyzed in terms of their prototypical F0 patterns.

The MEMOInt parameters used here are: **(1)** The reference intonation unit used is the stress group defined as the stressed syllable in combination with the preceding and the following syllables. **(2)** Table 1 shows the prosodic features used to tag the intonation unit, and the different number of values assigned to the prosodic features. We have selected only the three most relevant features among the sixteen available in order to easy the interpretation of the results in this preliminary study. **(3)** The acoustic parameters to be used are the projection of the control points on the Bezier fitting curve of the F0 contours, four parameters per intonation unit (more details about the parameterization technique in [13]). The selection of these parameters is a consequence of the previous works on modeling Spanish. In [1], we show that this parameter combination is the best to represent and predict the intonation of the Spanish corpus.

Prosodic Features	Acronym	Number of Values	Gain Info Catalan	Gain Info Spanish
Prominence	Accented	2	0.120	0.229
Position of SG in the IG	posSGIG	5	0.082	0.128
Position of IG in the SE	posIGSE	7	0.045	0.093

Table 1: Prosodic features characterizing the intonation units of the corpora.

List of Dictionaries LD3	Catalan D1	Catalan D2	Catalan D3	Spanish D1	Spanish D2	Spanish D3
Number of classes with more than 10 samples	2	4	8	2	4	17
Number of classes used in training	2	4	8	2	4	17
Number of classes in the final configuration	2	5	16	2	5	40
Initial number of classes	2	10	69	2	10	68
Mean number of samples per class	738	295	137	1235	494	113
Mean RMSE intra-class (Hz)	53	49	43	37	33	31

Table 2: Description of the dictionaries in terms of number of classes, size of the classes and number of samples per class.

4. Results

Table 1 shows the three most representative features to characterize the two corpora. `IG` means intonation group and `SG` means stress group. *Gain Info* gives information about the capabilities of the features to classify the classes resulting from the application of a 60 classes $kmeans$ cluster to the acoustic parameters of the selected intonation unit (see [14] for details about the metric, and [15] for an interpretation of this metric). The similarity of the two languages (with a common root) is probably the reason why the most relevant features in both cases are the same. Furthermore the ranking of relevance is also the same. The differences on the scale of the *Gain Info* values are due to the different size of the corpora.

Table 2 illustrates the configuration of the list of dictionaries after applying MEMOInt to the Catalan and Spanish corpora. Dx means that x features have been sequentially selected to set up the dictionary Dx. The first step to build the dictionary Dx is to index the intonation units in terms of the x most significant features (*Initial number of classes* row in the table) and the second step is to apply agglomerative clustering to set up the final configuration. A class of the dictionary Dx can be discarded if there is any other in the alternative dictionaries that predicts better the training samples (not used classes row in the table). This fact justifies the use of the list of dictionaries configuring a graph of classes. The dictionary $D1$ is set up using the feature $Accented$, $D2$ is set up using $Accented$ and $posSGIG$ and $D3$ using the features $Accented$, $posSGIG$ and $posIGSE$. Although the Spanish corpus has higher number of samples per class and more classes with more than 10 samples per class, the

	Lists of Features		Classes		
	Catalan (Cat)	Spanish (Sp)	Cat	Sp	RMSE(Hz)
1	noAccent,GAFinal,GECentr	noAccent,GAFinal,GECentr	C_1^3	C_{13}^3	71.91
2	noAccent,GAFinal,GEIncia	noAccent,GAFinal	C_1^3	C_3^2	71.24
3	accent,GAFinal	accent,GAFinal,GECentr3	C_2^2	C_2^3	67.05
4	noAccent,GACentr	noAccent,GACentr,GEFinal	C_1^2	C_{20}^3	63.27
5	accent,GAFinal	accent,GAFinal,GEPenul	C_2^2	C_{26}^3	56.44
6	noAccent,GAFinal,GECentr3	noAccent,GAFinal,GECentr3	C_3^3	C_3^3	50.49
7	accent	accent,GAIncia,GESegun	C_0^1	C_{44}^3	50.16
8	noAccent,GASigIn	noAccent	C_1^2	C_1^1	49.77
9	noAccent,GACentr	noAccent,GACentr	C_1^2	C_1^2	49.28
10	accent,GAFinal	accent,GAFinal,GEFinal	C_2^2	C_{19}^3	47.92
11	accent,GAIncia,GEIncia	accent,GAIncia,GEIncia	C_{16}^3	C_{27}^3	43.66
12	noAccent,GAIncia,GESegun	noAccent,GAIncia	C_{13}^3	C_5^2	43.32

Table 3: Most relevant differences between the Spanish and Catalan graph of classes.

intra-class similarity is higher. This fact is justified because the MEMOInt parameters used in this experiment are optimum for Spanish according to the research presented in [1] but it does not guaranty their modeling capabilities for Catalan.

Table 3 shows the 12 most relevant differences between the F0 patterns (represented in the classes) aligned in term of its prosodic function (represented in the features). The rows confront matching nodes of the graphs. The lists of features are the tags of the paths and the classes are the nodes of the graphs of classes. RMSE is the distance between the classes. Every list of features is tagged as x, y, z where x, y and z are values of the features $Accented$, $posSGIG$ and $posIGSE$ respectively. C_i^d is the name of the class i of the dictionary d to be displayed in the table 4. We confronted all the classes but we display only the ones with differences over 43 RMSE(Hz). This threshold is the mean RMSE intra-class error observed in table 2 and it is used as an indicator of a potential relevant difference. The value $GAFinal$ means $finalstressgroup$ and it appears in five of the first seven rows of the table. Prieto for Catalan in [16] and Garrido for Spanish in [17] point out that the final part of intonation group is a relevant part of the Catalan and Spanish intonation system. This result seems to indicate that this part has potential capabilities to discriminate these languages. Most of the differences appear in central intonation group ($GECentr$, $GESegun$ and $GEPenul$). We remark row 11 as this feature combination is described in [18] as a characteristic pattern of Catalan intonation (named *primer pic*).

The visual representation displayed in table 4 is useful to detect the particular movements that make the patterns different. Thus, second and third rows display a different trajectory in the F0 contours, the fourth row shows a displacement of the register and the first row seems to be a combination of both effects.

We have applied informal perceptual test by the listening of sentences that include the found distant pattern to corroborate that the objective differences are easily perceived.

5. Discussion

The main objective of this work is the presentation of a procedure to compare the characteristic intonation of two different corpora. In this sense, we remark that although the application of the procedure has been focused here to the Catalan vs. Spanish case, it can be extended to the comparison of other different intonation aspects with practical interest such as it could be the style or emotional attitude projected in the corpus.

Furthermore, the comparison we do here between Catalan and Spanish intonation is weak due to two future experimental tasks to be improved. One of them concerns with the representativeness of the patterns to be confirmed with statistic measurements. Another one is the need to apply a systematic and repetitive perceptual procedure to assess the objective results. We conclude that the value of this communication is not on the comparison between Catalan and Spanish intonation but on the presentation of a procedure to cope with this problem.

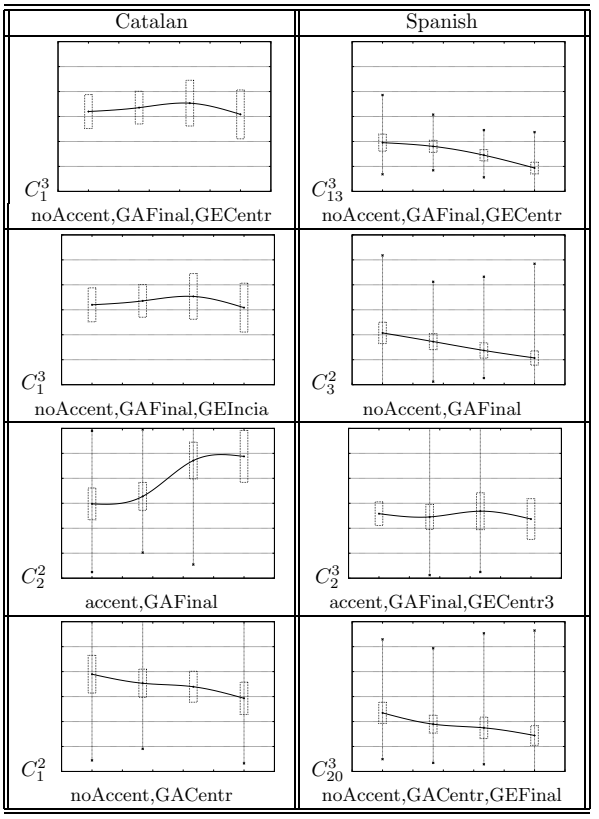

Table 4: Visualization of the four most distant classes. Each row shows the respective nodes of the Catalan and Spanish graphs of classes after confronting them. This table completes the information displayed in table 3 with a visual representation of the typical F0 pattern of the class: X scale is normalized, Y scale is 100-220Hz, the four boxes are the statistics of the acoustic parameters.

On the other hand, we have focused our comparison on the graphs of classes resulting from the application of MEMOInt, but there are other aspects resulting also from the application of MEMOInt which result in further contributions. Thus, the higher intra-class distance of the classes for Catalan (53Hz vs 37Hz for $D1$ in table 2) indicates that the parameters used in this study could be tuned to obtain better results: The optimum set of MEMOInt parameters (parameterization technique, number of acoustic parameters, type of prosodic features) used to model Spanish is not necessary the same to optimize the modeling of Catalan. The different MEMOInt parameters could be another source of information to be explored in order to identify differences between Catalan and Spanish prosody.

6. Conclusions and Future Work

This communication presents an ongoing research on the comparison of corpora in terms of their intonation profiles. The

methodology MEMOInt previously used to model intonation with applications in text-to-speech system has been applied to this aim. MEMOInt brings a graph of classes representing the intonation of the input corpus. We present a procedure to compare two aligned graphs of classes representing two different corpora. As result we obtain list of differences that takes into account the form of the prototypical F0 patterns that share the same function.

The procedure has been applied to compare a Catalan and a Spanish corpus. At least twelve prototypical movements show relevant differences. MEMOInt permits display and confront the relevant F0 patterns in association with the prosodic features that justify them.

This is an ongoing research that needs to be reinforced by the application of statistical tests to evaluate the representativeness of the F0 patterns and with the application of systematic perceptual test. The procedure can be extended to other application like comparing styles or emotions.

The application of the different styles in the modification of the intonation of a given voice is a future challenge. In the case reported in this communication, the case is to modify a Spanish voice to speak with Catalan accent or vice versa. Other challenging applications could be the support of speaker recognition systems or the teaching of a foreign language to correct wrong pronunciations.

7. Acknowledgements

This work has been partially sponsored by Consejeria de Educacion (JCYL project VA053A05)

8. References

[1] D. Escudero and V. Cardeñoso, "Applying data mining techniques to corpus based prosodic modeling speech," *Speech Communication*, vol. 49, pp. 213–229, 2007.

[2] M. Jilka, "The contribution of intonation to the perception of foreign accent," Ph.D. dissertation, Universitat Stuttgart, Germany, 2000.

[3] M. Dalton and A. Chasiade, "Modelling intonation in three Irish dialects," in *Proceedings of ICPHS 2003*, 2003.

[4] E. Sardelli and G. Marotta, "Prosodic parameters for the detection of regional varieties in Italian," in *Proceedings of ICPHS 2007*, 2007, pp. 1281–1284.

[5] A. Quilis, *Tratado de Fonología y Fonética*. Editorial Gredos, 1993.

[6] P. Delattre, "Comparing the prosodic features of English, German, Spanish and Frech," *IRAL*, vol. 1, pp. 193–210, 1963.

[7] F. Cummins, F. Gers, and J. Schmidhuber, "Comparing prosody across many languages," Instituto Dalle Molle di Studie sull'Intelligenza Artificiale, Lugano, Switzerland, Tech. Rep., IDSIA-07-99 1999.

[8] H. Mixdorff, H. Pfitzinger, and D. Grauwinkel, "Toward objective measures for comparing speaking styles," in *Proceedings of SPECOM 2005*, 2005, pp. 131–134.

[9] A. G. Adami, "Modeling prosodic differences for speaker recognition," *Speech Commun.*, vol. 49, no. 4, pp. 277–291, 2007.

[10] A. Jain, M. Murty, and P.J.Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, September 1999.

[11] A. Webb, *Statistical Pattern Recognition*, 2nd ed. Wiley, 2002.

[12] A. Ferrer, "Sintesi de la Parla per Concatenació Basada en la Selecció," Ph.D. dissertation, Dpto. de Teoría del Senyal i Comunicacions, Universidad Politécnica de Cataluña, España, 2001.

[13] D. Escudero and V. C. A. Bonafonte, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish," in *Proceedings of ICASSP 2002*, vol. 1, 2002, pp. 481–484.

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

[15] 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, pp. 1165–1168.

[16] P. Prieto, "Entonació," in *Gramàtica del català contemporani*. Barcelona, Empúries, 2002, ch. 12, pp. 393–462.

[17] J. M. Garrido, "Modelling spanish intonation for text-to-speech applications," Ph.D. dissertation, Facultat de Lletres, Universitat de Barcelona, España, 1996.

[18] D. Font, "L'Entonació del Català. Patrons Melòdics, Tonemes i Marges de Dispersiò," Ph.D. dissertation, Universitat de Barcelona, 2005.