Visualization of Prosodic Knowledge Using Corpus Driven MEMOInt Intonation Modelling*

David Escudero-Mancebo and Valentín Cardeñoso-Payo

Univerisity of Valladolid, Valladolid 47014 Spain
descuder@infor.uva.es,
http://www.infor.uva.es

Abstract. In this work we show how our intonation corpus driven intonation modelling methodology MEMOInt can help in the graphical visualization of the complex relationships between the different prosodic features which configure the intonational aspects of natural speech. MEMOInt has already been used successfully for the prediction of synthetic F0 contours in the presence of the usual data scarcity problems. Now, we report on the possibilities of using the information gathered in the modelling phase in order to provide a graphical view of the relevance of the various prosodic features which affect the typical F0 movements. The set of classes which group the intonation patterns found in the corpus can be structured in a tree in which the relation between the classes and the prosodic features of the input text is hierarchically correlated. This visual outcome shows to be very useful to carry out comparative linguistic studies of prosodic phenomena and to check the correspondence between previous prosodic knowledge on a language and the real utterances found in a given corpus.

1 Introduction

The study of intonation in speech technology is important because it brings information about the structure of the message and about the pragmatics of the discourse. The research field of intonation modelling has grown significantly during the last two decades due mainly to its interest in text-to-speech applications. Despite this fact, the state of the art in intonation modelling is still characterized by a diversity of paradigms and approaches, revealing a lack of consensus. In this context the availability of tools shedding light in the many aspects that are a matter of discussion is important. This article presents a methodology for modelling intonation devised attending to two main goals: to be efficient in text-to-speech applications and to offer contrastable information about some of the controversial aspects in this field of research.

The availability of large speech corpora is the main reason for the spectacular advance in the quality of nowadays text-to-speech systems [1]. The high degree of naturalness is achieved by training learning procedures that use the data of the corpus. Although this approach is acceptable from an engineering point of view, results can be frustrating from a scientific point of view: a neural network permits to learn intonation from corpus, but it is not obvious to transfer the knowledge of the neural network to a book on phonetics. Here

* This work has been partially sponsored by Spanish Government (MCYT project TIC2003-08382-C05-03) and by Consejeria de Educacion (JCYL project VA053A05).

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 645–652, 2006.

we defend a methodology that is efficient in learning intonation from corpus and for visually displaying the information obtained from the corpus so that it is possible to contrast results with previous observations or theoretical models.

The problem of intonation modelling consists of finding a matching between the prosodic features of the intonation units found in the text, and the corresponding F0 contours patterns (see [4] for a review). This problem is difficult because of the high number of prosodic features affecting intonation: accent, type of sentence, structure of the sentence, emotions, social culture of the speakers. Furthermore, there is not a consensus about the best technique to use to parameterize the F0 contours: Tilt[15], Fujisaky[9], templates[14]. . . . The correspondence between the prosodic features and the acoustic parameters is also a matter of discussion where many different approaches have been tested: neural networks[11], decision trees [15], regression trees[2], rules[3]. We have to add to these difficulties the intrinsic variability of intonation where speakers can utter the same sentence in many different ways. As a result, even huge corpora get undersized, so that it is necessary to devise a strategy to cope with data scarcity. Here we propose a technique that assumes all these difficulties, modelling and representing the intonation in terms of the prototypical patterns and its variability observed in the corpus and displaying the relationship of these aspects with respect to the prosodic factors affecting them.

The proposed technique is called MEMOInt, and it is based on a combination of agglomerative clustering and sequential feature selection [16]. The combination of both techniques permits to train efficiently intonation models from corpus useful to predict synthetic intonation and also to display as a decision tree the information of the corpus as it has been show in [8][5] Here we show MEMOInt capabilities to display graphically the prosodic information of the corpus. In section 2 we review the MEMOInt fundamentals; in section 3 the experimental results and in 3.2 the results on visualization.

2 MEMOInt: Methodology for Modelling Intonation

Figure 1 describes schematically the fundamentals of MEMOInt. In this section we focus on the modelling stage to explain the procedure that outputs the visual information mentioned above.

In corpus based modelling intonation the corpus is considered a set $C = \{IU_i, \quad 1..N\}$, where IU_i is each of the N units of intonation identified in the corpus. Every IU is a duple $IU = (PF, AP)$. $IU.AP$ are a set of acoustic parameters that represent the form of the F0 contour of IU. $IU.AP$ are obtained automatically from the F0 contours in the parameterization stage. On the other hand, $IU.PF$ are a set of prosodic features that represent the prosodic function of IU. They reflect different aspects of intonation like accent, grammatical structure of the sentences, size of the intonation units, emotions, type of sentence . . .. Those features are to be extracted automatically from text or manually labeled in the corpus.

MEMOInt applies an agglomerative clustering process following an inter-class maximum similarity criterion that is justified because merging together classes of IU with similar AP is acceptable as they can be perceived the same. Due to the final use of the models in text-to-speech, the agglomerative process must stop when the prediction capabilities of the new

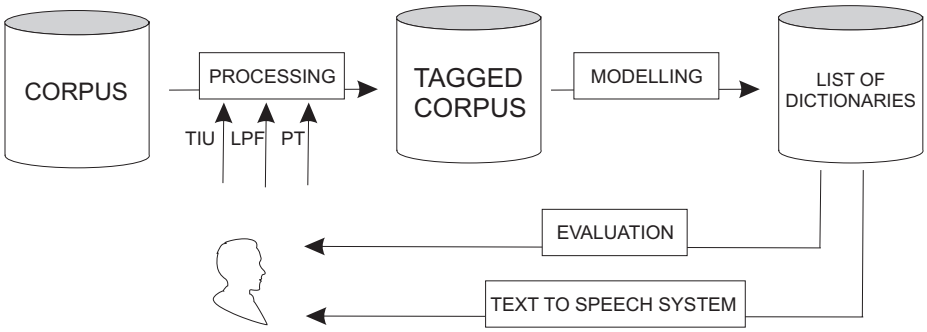

Fig. 1. Functional scheme of MEMOInt. The `CORPUS` is processed to obtain the TAGGED CORPUS which is the input of the modelling stage producing the `LIST OF DICTIONARIES`. The capabilities of the list of dictionaries are evaluated and it can also be use by a TTS system. The parameters of MEMOInt are: Type of Intonation Unit (TIU), List of Prosodic Features (LPF), Parameterization Technique (PT).

configuration after merging classes predicts worst than the previous one. (see [5] and [8] for details)

The agglomerative process determines the correspondence between AP and PF. By keeping track of the different values of the PF merged, it is built an index to assign a class in the final configuration to any PF combination. This can be used in text-to-speech where PF is driven by text and AP can be used to generate a synthetic F0 contour. Let us call *dictionary* to the combination of the index, made of a sequence of PF, and the clustering associated to it.

The more the number of PF 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 PF. In every step MEMOInt selects the PF which inclusion implies better prediction results. As result we obtain N different dictionaries (as many as PF considered). Given any combination of PF we select the cluster that predicts more accurately the sample according to the observations in the training stage (see [8] for details).

The sequential selection of PF allows either to obtain a ranking of importance of different PF or to test new alternatives. The list of dictionaries provides a way to draw a decision tree which gives easy to contrast visual information, which illustrates the intonation patterns of the corpus, as we show in the following sections.

3 Experimental Results

3.1 Building of the Dictionaries

For the experimental validation of the clustering technique, we have used an intonation corpus which contains more than 700 sentences (4363 intonation units) recorded by a professional actress in studio conditions[1]. High quality F0 contours were obtained using a laringograph device. Sentences has been segmented and labelled following a semiautomatic process. We

[1] Gently provided to us by the research group TALP of the Polythecnic University of Catalonia, Spain.

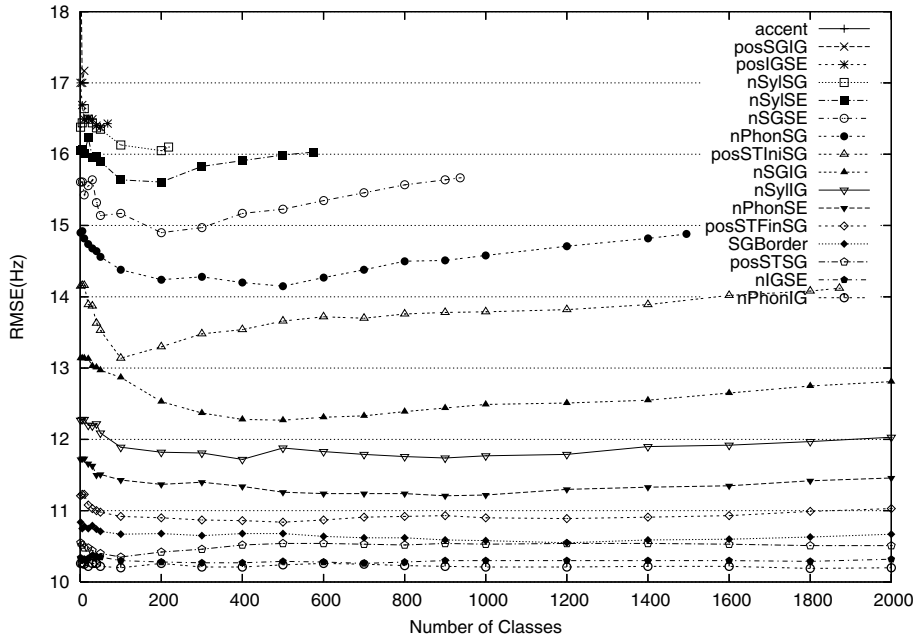

Fig. 2. Building the list of dictionaries: each curve represents the effect of adding dictionary D_i to the list of dictionaries LD_i, $(i = 1, \ldots, N_{pf})$. The name of the PF added to build D_i is the legend of the curve. Each curve represents the prediction error of the training samples as a function of the number of classes at each step of agglomeration, starting at the right end with the maximum number of classes for that set of PF. The optimal number of classes for dictionary D_i corresponds to the minimum of the associated curve.

selected only the declarative sentences, which represent about 95% of the whole corpus. The sentences have been segmented into different types of intonation units: intonation groups (IG), stress groups (SG) and syllables (see [12] for a definition of this units). In this study the basic unit of reference has been the SG, defined as the combination of a stressed syllable of a word plus the preceding and following one. The acoustic parameters are the control points of the Bézier curves of degree 3 fitting the F0 contours in the intonation units (more details in [7]). The following prosodic features were considered: type of sentence `typeSE` (1 value), position of the tonic syllable in the first SG `posSTiniSG` (3 values) and in the last one `posSTfinSG` (3 values), number of IGs `nIGSE` (5 values), SGs `nSGSE` (6 values), syllables `nSylSE` (6 values) and phonemes `PhonSE` (6 values) in the sentence, number of stress groups `nSGIG` (6 values), syllables `nSylIG` (6 values), and phonemes `nPhonIG` (6 values) in the IG, position of the IG in the sentence `posIGSE` (7 values), position of the SG in its IG `posSGIG` (6 values), `SGBorder` indicating the configuration of the SG, number of syllables `nSylSG` (9 values) and phonemes `nPhonSG` (6 values) in the SG, position of the stressed syllable `posSTSG` (3 values). For the experiments, the corpus was split into 3 subsets: modelling, training and testing sets.

We use the centroid to represent the samples of each class in the clusters. The Euclidean distance between the respective centroids of the classes was used as the inter-class similarity

Table 1. Description of the dictionaries in terms of number of classes and number of samples per class

List of Dictionaries LD7	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Number of eligible classes	2	4	17	30	26	21	24
Number of grouped classes	2	5	40	111	83	80	190
Initial number of nlasses	2	10	68	230	631	1068	1795
Mean number of samples per class	1235	494	113	42	35	32	16
Mean intra-class dispersion (Hz)	37	33	31	26	21	20	17

metric to guide the merging process. The prediction error is computed as the distance between the points of the real F0 contour and the points of the corresponding synthetic one. This distance is measured using the recommended RMSE and Pearson Correlations [10].

Figure 2 monitors the building process of the list of dictionaries. Error values were obtained by averaging the prediction error over the set of SG in the training corpus. The quality of the obtained results is comparable with the one obtained following other approaches of the state of the art (see [8] for objective and subjective tests). Table 1 shows the impact of the agglomerative process in the final number of representative classes. Note that for every type of intonation unit, there is a serial of dictionaries to select one: some of the classes are never used. This reduction of classes is helpful to simplify the representation of the clusters that is presented in the following section.

3.2 Visualization of Intonation Patterns

Figure 3 shows a tree-like graphical representation of the classes in the list of dictionaries (only a selection). Each node represents a class in the clusters. For every class, we show the Bezier curve representing the F0 profile of the centroid and the standard deviation of each control point. The graph at each node provides a visual representation of the prototypical F0 patterns of the IU belonging to that class.

The classes belonging to the level i are the selected classes for dictionary D_i. Only classes which have been effectively used for prediction and contain more than 10 samples have been represented. The labels of tree branches give the values of the PF. The path going from the root to a given node provides one of the sequences of prosodic features which correspond to the node class.

This tree representation differs from a conventional regression tree in many aspects. Here the same class could appear in different nodes if more than one PF combination indexes it. Furthermore, the parent-child relationship does not imply the splitting of the samples of the parent node. Here the hierarchy is determined by the PF and the contents of the nodes by the agglomerative process. The tree is an easy to read representation of the information of the dictionaries. The input is an array of PF and the output is the corresponding class. The criterion to select the output is to choose the most accurate class in the tree: the closer to the root node the less accurate node. Thus, if the input is the array of PF `(noAccent, GAFinal, GEFinal, a)` the output is the class `C4_75`; and if the input is `(accent, GAFinal, GEFinal, a)`, the output is `C1_6`. Navigating the tree, it is observed the impact of the corresponding PF with respect to the shape of the prototypical F0 pattern.

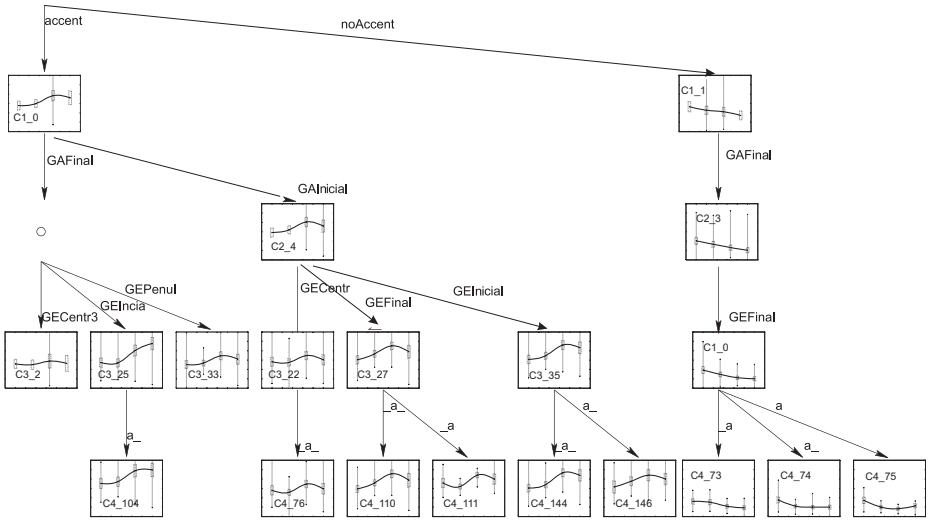

Fig. 3. Models of the dictionary represented as a decision tree. We have selected a part of the whole tree. X scale is normalized. Y scale is 100–220Hz.

The visualization of the information in the tree allows us to contrast some of the assessments found in the bibliography about Spanish Intonation. In [6], an overview of the proposals of several authors can be found. Here we review the main assessments and we contrast them with plots in figures 3.

- **Prominence** (or relative importance of the stress group with respect to the others) was labelled in the corpus with the prosodic feature `accent`. Observations of the intonation of the corpus projected in figure 3 permits to assess that this feature is the most relevant one attending to the shape of the F0 patterns. This is reflected in the fact that this feature has been selected the first one among all the prosodic features taken into account when the learning procedure previously detailed has been applied. Furthermore, the tree shows that the classes in the branches corresponding to the prominent part (`accent` value) are characterized by higher F0 values in contrast with the patterns appearing in the unaccented branch (`noAccent` value). This observation is in consonance with the Phonetics theory that gives to the F0 feature the function of focusing different parts of the sentences.
- **Prosodic structure of the stress group**: Sosa[13] observed that the prototypical patterns associated to the Spanish stress groups are $L*+H$ pattern and the less frequent $H*$ one (using TOBI notation). This fact can be observed in the tree where $L*+H$ patterns appear in `C4_104, C4_76, C4_110, C4_144, C4_146`. The pattern $H*$ appears in the class `C4_111`. Apparently `C4_111` does not differ significantly from the other classes, but it must be taken into account that the duration is normalized so that the peak of the F0 contour is coincident with the stressed syllable without any temporal displacement as it occurs in the $L*+H$ classes already mentioned (note that `nSilGA` has 4 possible values: `_a_, _a, a_, a`), where `_` means un-stressed syllable and `a` means stressed one).

- **Junctures or prosodic boundaries** are very important to arrange the structure of the discourse. The boundaries use to precede or even to substitute the pauses. They are characterized by an abrupt jump in the tendency of the F0 contour. The typical pattern is a rising one called *anticadencia* that can be observed in classes `C3_25, C4_104`. The patterns in `C3_2 and C3_33` are known exceptions called *semicadencia* in the Spanish Phonetics literature (see [12]).
- **Final boundary**: affecting the last part of the F0 contour. Typical final juncture of declarative sentences is $L*+L\%$. This pattern is clearly seen in figure 3 in classes `C1_0, C4_73, C4_74, C4_75`. This final part of the F0 contours has associated the distinctive function to discriminate the type of sentence. When the corpus is enriched with interrogative and exclamative sentences it is expected that the patterns with the prosodic feature values `GAFinal` and `GEFinal` will be determinant.

Finally, we remark that the visualization of figure 3 will surely let experts to get more conclusions about the intonation phenomena, although a thorough discussion of this is out of the scope of the present paper.

4 Conclusions and Future Work

In this communication we have shown the application of MEMOInt to visualize the prosodic information of a given corpus. The classes of F0 patterns correspond with the observations of different authors in Spanish phonetics. The decision tree permits to track the relationship between the typical F0 movements and the set of input prosodic features causing those movements. The observations driven from the tree corroborate the observations found in the bibliography. This converts MEMOInt in a tool to shed light in the many aspects to reveal in this field of research.

Next step in our research will be to measure the capabilities of MEMOInt to analyze the prosodic structure of the utterances with respect to the grammatical structure of the sentence. Furthermore, we expect to use the decision tree to compare the intonation of different corpora.

References

1. A. Aaron, E. Pitrelli, and J. F. Pitrelli. Conversational computers. *Scientific American*, June:64–70, 2005.
2. P.D. Aguado, K. Wimmer, and A. Bonafonte. Joint extraction and prediction of fujisaki's intonation model parameters. In *Proceedings of EuroSpeech 2005*, 2005.
3. J. Allen, M. S. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, 1987.
4. A. Botinis, B. Granstrom, and B. Moebius. Developments and Paradigms in Intonation Research. *Speech Communications*, 33:263–296, July 2001.
5. V. Cardeñoso and D. Escudero. A strategy to solve data scarcity problems in corpus based intonation modelling. In *Proceedings of ICASSP 2004*, 2004.
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.

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 and V. Cardeñoso. Optimized selection of intonation dictionaries in corpus based intonation modelling. In *Proceedings of Eurospeech*, September 2005.
9. H. Fujisaki and K. Hirose. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of Acoustics Society of Japan*, 5(4):233–242, 1984.
10. D. J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41:73–82, February 1994.
11. O. Joskisch, H. Mixdorff, H. Kruschke, and U. Kordon. Learning the parameters of quantitative prosody models. In *Proceedings of ICSLP 2000*, 2000.
12. T. Navarro-Tomás. *Manual de Entonación Española*. Madrid, Guadarrama, 1944.
13. J. M. Sosa. *La Entonación del Español*. Cátedra, 1999.
14. R. Sproat. *Multilingual Text-to-Speech Synthesis*. Kluwer, 1998.
15. P. Taylor. Analysis and Synthesis of Intonation using the Tilt Model. *Journal of Acoustical Society of America*, 107(3):1697–1714, 2000.
16. A. Webb. *Statistical Pattern Recognition*. Wiley, 2nd edition, 2002.