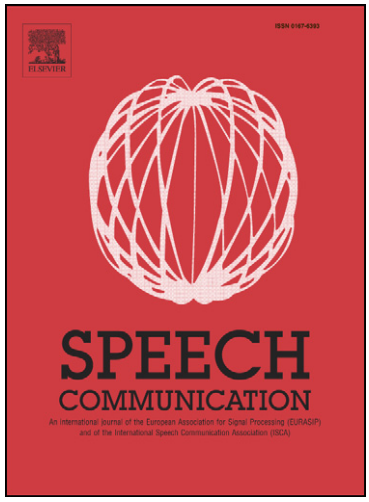
Applying Data Mining Techniques to Corpus Based Prosodic Modeling

David Escudero-Mancebo, Valentìn Cardeñoso-Payo



Applying Data Mining Techniques to Corpus Based Prosodic Modeling

David Escudero-Mancebo, Valentìn Cardeñoso-Payo

PII: S0167-6393(07)00021-0
DOI: 10.1016/j.specom.2007.01.008
Reference: SPECOM 1611

To appear in: *Speech Communication*

Received Date: 19 July 2006
Revised Date: 21 January 2007
Accepted Date: 23 January 2007

Please cite this article as: Escudero-Mancebo, D., Cardeñoso-Payo, V., Applying Data Mining Techniques to Corpus Based Prosodic Modeling, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.01.008

Applying Data Mining Techniques to Corpus Based Prosodic Modeling

David Escudero-Mancebo*, Valentín Cardeñoso-Payo

Departamento de Informática. Universidad de Valladolid.

Campus Miguel Delibes s/n. 47011 Valladolid. Spain.

Abstract

This article presents MEMOInt, a methodology to automatically extract the intonation patterns which characterize a given corpus, with applications in text-to-speech systems. Easy to understand information about the form of the characteristic patterns found in the corpus can be obtained from MEMOint in a way which allows easy comparison with other proposals. A visual representation of the relationship between the set of prosodic features which could have been selected to label the corpus and the intonation contour patterns is also easy to obtain. The particular function-form correspondence associated to the given corpus is represented by means of a list of dictionaries of classes of parameterized `F0` patterns, where the access key is given by a sequence of prosodic features. MEMOInt can also be used to obtain valuable information about the relative impact of the use of different parameterization techniques of `F0` contours or of different types of intonation units and information about the relevance of different prosodic features. The methodology has been specifically designed to provide a successful strategy to solve the data sparseness problem which usually affects corpora as a consequence of the inherent high variability of the intonation phenomenon.

Key words: Prosody, Intonation Modeling, Data Mining, Text-to-speech, F0 Contours

1 Introduction

In the present generation of text-to-speech applications (TTS), there are systems which provide high-quality reproduction of human intonation exploiting their capabilities to adequately extract intonation information from labelled speech corpora (see Aaron et al. (2005) for an excellent review of the current state of the art in TTS technology and products). The availability of huge speech corpora and the use of automatic information analysis techniques is the main reason for this success. Rule-based systems (as the pioneering MITalk (Allen et al., 1987)) have been significantly improved upon, and corpus-based systems are, from an engineering point of view, the best option in terms of quality and adaptability to new speakers and contexts. Nevertheless, corpus-based systems still have the limitation of their vulnerability to the sparse-data problem: huge amounts of data are required to obtain an acceptable quality and additional data are needed to adapt the system to new situations. The root of this problem is that corpora analysis focuses on locating samples to be adapted to the prediction needs and not on obtaining knowledge about the phenomena to be reproduced. The lack of knowledge retrieval makes the solutions acceptable from the engineering point of view, but not solid enough from the scientific point of view and the consequence is the lack of robustness when

* *Corresponding author*: David Escudero-Mancebo
Email address: descuder@infor.uva.es (David Escudero-Mancebo).
URL: http://www.infor.uva.es/~descuder (David Escudero-Mancebo).

corpus data get scarce. In this article we present MEMOInt, a methodology for intonation modeling that aims to obtain information from the corpus to be analyzed, permitting synthetic intonation to be reproduced in a robust way to cope with data scarcity.

The major challenge of MEMOInt is how to obtain the information needed for corpus-based TTS or, in other words, how to extract prosodic information from a given corpus in order to characterize it. This is not an easy task since there are still many open questions in the state of the art on which many methodologies do not shed any light (see Botinis et al. (2001)). The key aspects which have been considered when designing MEMOInt are related to the selection of the basic linguistic unit on which intonation is modeled (intonation unit) and the set of prosodic features, the way to lay out a practical function-form correspondence, and the selection of the F0 contour representation technique to be used.

The selection of the type of intonation unit and the set of associated prosodic features used to characterize the corpus are critical to model intonation properly. The type of intonation units generally considered in studies of intonation include: sentences, intonation groups, stress groups or syllables. Different proposals can be found in the literature when one or several of these units are to be selected as the basic reference unit on which the matching between the text of the message and the F0 contours should be laid: accentual phrase (Sakurai et al., 2003), accent groups (Santen and Möebius, 2000), words (Veronis et al., 1998), syllables (Taylor, 2000; Lee and Oh, 2001; d'Alessandro and Mertens, 1995) and Interperceptual Centre Group (GIPC) (Holm, 2003). Moreover, in superposition models the use of more than one unit is proposed (Fujisaki and Hirose, 1984; Sakai, 2005). With respect to the prosodic features characteriz-

ing these units, the number and type of them also depend on the approach. Some of them are position, number of syllables, number of words, conjugation (Sakurai et al., 2003), part of speech (Sakurai et al., 2003; Veronis et al., 1998; Holm, 2003), number of syllables from previous stress or accented syllable (Taylor, 2000)and structure of the sentence (Lee and Oh, 2001; Holm, 2003). Furthermore, each of these prosodic features can have a different cardinality depending on the proposal. This lack of consensus in the state of the art seems to indicate that the different methodologies are strongly dependent on the selected intonation unit, making it difficult to compare their results. That is the reason why MEMOInt has been designed assuming an abstraction of the kind of intonation unit, which allows easily carrying out comparisons and quality tests both of the impact of the selection of the intonation unit and the associated set of prosodic features just by using several alternatives and comparing the objective quality results brought by MEMOint.

As for the function-form correspondence, several proposals can be found in the state of the art on how to obtain the right relation between the acoustic parameters (representing the `F0` contours) and the prosodic features: stochastic models (Veronis et al., 1998), neural networks (Holm, 2003; Sakurai et al., 2003), linear regression (Sproat and Olive, 1995) and decision and regression trees (Lee and Oh, 2001; Taylor, 2000; Eide et al., 2003). In all these methodologies, two main limitations arise: lack of robustness to cope with data scarcity training conditions and limited capabilities to provide experimentally contrastable prosodic information. MEMOInt introduces the concept of *dictionary of classes* to represent the particular correspondence between the prosodic factors and the classes in a given corpus. In a similar way to pioneer approaches such as (Emerard et al., 1992; Traber, 1992), the dictionary (or

the data base) is a representation of the intonation in the corpus which allows the prediction of an F0 contour from the prosodic features. The main difference with those approaches is that we consider that a class represents not only a prototypical pattern of intonation but also its experimental variability for the given set of prosodic features determining the pattern, as found in the reference corpus we are working with. The use of *lists of dictionaries* will also provide a graphical way to illustrate the correspondence between a set of prosodic features and its associated class.

The discussion mainstream on the representation technique of F0 contours has been usually focused on whether phonetic representations are more appropriate than phonological ones or not (see Botinis-2001 for a review). Classes of intonation patterns are the building blocks of MEMOInt and the set of classes represent both the variety of prototypical intonation movements and the variability of the F0 contour shapes associated to every prototype, within the limited domain of the given corpus. The classes of intonation patterns group metrical representations of the F0 contours (phonetic aspect) and statistically represent the best prototypical patterns (phonological aspect) for a given contour metric. This allows different parameterization techniques to be contrasted and also provides useful information about the characteristic F0 movements found in the corpus and its variability, in an easy readable format.

Concerning the capabilities of MEMOInt to generate accurate synthetic F0 contours in sparse-data conditions, let us recall that data scarcity problems have their origin in the high variability of the intonation phenomena. The final shape of pitch contours is influenced by a high number of factors and it is almost impossible to have a corpus with enough coverage of data associated to the huge combinatorial potential of these factors. As an example, in (Sakurai

et al., 2003) a corpus intonation modeling technique is presented in which the possible combinations of the prosody factors were more than 27 million, but the available samples in the corpus where about 3000. The situation gets even worse when paralinguistic features are considered (see Campbell and Erickson (2004)). Under this condition, it could be possible to record a bigger corpus but when this is not the case, it is necessary to devise a strategy to generate plausible synthetic intonation patterns assuming the corpus has limitations. MEMOInt is specifically designed to cope with this problem and we propose using the corpus to adjust models at different levels of detail, so that it is always possible to select the class of an adequate level of detail, depending on the amount of information provided by the prosodic features labelling, which better predicts the observation of the corpus in terms of the contour metric we are using.

As a result, we find MEMOInt useful not only as an efficient way to predict realistic pitch contours using a data mining technique on the data stored in a corpus, but also as an experimental tool to support corpus-based linguistic research on intonation modeling.

In section 2 we show the architectural scheme of MEMOint, the way the corpus has to be processed, how we can obtain intonation patterns and use them to generate `F0` contours and, finally, how MEMOint manages data scarcity. In section 3 we will discuss an application of MEMOInt to Spanish language, to illustrate the possibilities of our methodology. Specific experiments on the selection of the type of intonation unit and the set of prosodic features, and on different alternatives for the parameterization technique are reported in that section too. Last section of the paper includes conclusions and some proposals for future work.

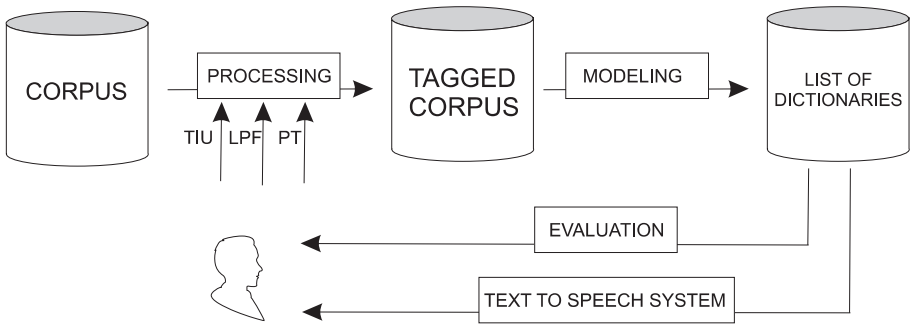


Fig. 1. Functional scheme of MEMOInt.

2 MEMOInt: MEthodology for MOdeling Intonation

Figure 1 shows the functional scheme of MEMOInt. The *Corpus* is processed to obtain the *Tagged Corpus* (*Processing* task) which is the input of the *Modeling* task producing the *List of Dictionaries*. The capabilities of the list of dictionaries are evaluated and it can also be used by a TTS system (*Evaluation* and *TTS* tasks respectively).

Relevant features related to intonation are extracted from the corpus in the **Processing** stage. A representation of the intonation of the corpus is obtained in the **Modeling** stage, which outputs a list of dictionaries of models. In the **Evaluation** stage, the quality of MEMOInt is measured, both in terms of the fidelity of the generated synthetic intonation and also in terms of its capabilities to extract and visualize prosodic information from the corpus. The **text-to-speech** module would make use of the outcomes of the modeling stage in speech synthesis application scenarios. All these basic building blocks will be further detailed in the following subsections 2.1 to 2.4.

The **Parameters** of MEMOInt are: Type of Intonation Unit (`TIU`), List of Prosodic Features (`LPF`), and Parameterization Technique (`PT`). In the introduction we have discussed about the possible values of `TIU` and `LPF`; we discuss

PT in section 2.2.

The outcomes of MEMOInt are: (1) a tool to generate synthetic F0 contours, (2) objective and subjective evaluation of its capabilities to produce synthetic intonation of quality, (3) descriptive information about the intonation in the analyzed corpus and (4) information about the suitability of the different tested values of the parameters. The following sections give details about the stages and the parameters enumerated above.

2.1 Corpus Processing

Figure 2 represents the processing stage. The **Splitting** task consists of dividing the sentences of the corpus into intonation units. A corpus can be seen as set of sentences

$$\texttt{Corpus} = \{S_j, \quad j = 1 \dots N_S\} \tag{1}$$

where each sentence S_j is a sorted set of intonation units

$$S_j = \{u_i, \quad i = 1 \dots N_U\}. \tag{2}$$

Each intonation unit u_i involves both a part of the analyzed sentence and the corresponding portion of the F0 contour (see figure 3). The F0 contour of the given u_i can be referred to as u_i.F0 and the corresponding part of the message as $u_i.msg$.

The **Labeling** task consists of assigning values to the prosodic features that are relevant from the point of view of the analysis and synthesis of the F0

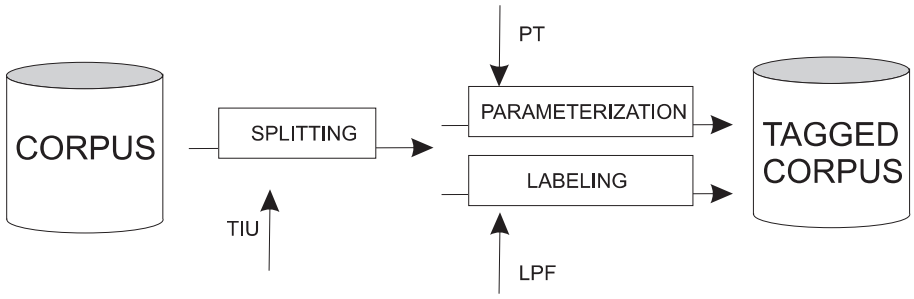


Fig. 2. Processing Task: sentences are first segmented into intonation units. In order to obtain the *Tagged Corpus*, the intonation units are located while in the (*Corpus Splitting* task). Then, the units are labeled (*Labeling* task) and its F0 contour is parameterized (*Parameterization* task).

contours. If we consider a set of prosodic features $\mathcal{F} = \{F_1 \dots F_{N_F}\}$, then each unit has an associated vector of prosodic features

$$u_i.\bar{f} = (f_1 \dots f_{N_F})_i, \quad (f_1 \dots f_{N_F})_i \in F_1 \times \dots \times F_{N_F} \tag{3}$$

where f_i is a value of the feature F_i in the set $\mathcal{F}^i = \{f_i^1 \dots f_i^{N_{F_i}}\}$. The **Parameterization** task permits quantitative parameters to be obtained from F0 contours reflecting its evolution in the intonation unit. After the set of parameters to use has been chosen $\mathcal{P} = \{P_1 \dots P_{N_P}\}$, each u_i can be associated to a vector of acoustic parameters $u_i.\bar{p} = (p_1 \dots p_{N_P})_i$, where p_i represent the possible values of the selected acoustic parameters considered. Acoustic parameters can be obtained from F0 contour by means of a function

$$\mathcal{P}ar : \texttt{F0}(\mathrm{t}) \rightarrow P_1 \times \dots \times P_{N_P}. \tag{4}$$

In which follows, $u_i.\bar{p}$ will be referred as the *intonation pattern* of u_i (see figure 3). The parameterization technique must be reversible, meaning that it exists a function

$$\mathcal{P}ar^{-1} : P_1 \times \dots \times P_{N_P} \rightarrow \texttt{F0}(\mathrm{t}) \tag{5}$$

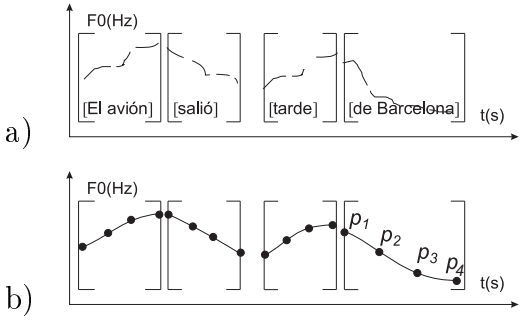


Fig. 3. An example of the application of the *Processing* task. *a)* The figure shows the splitting of an utterance, F0 contour and aligned message divided into the corresponding intonation units (stress groups in this case). *b)* The figure shows one of the possible parameterizations of the last intonation unit: smoothing of the F0 contour and fitting with 4 parameters (p_1, p_2, p_3, p_4).

so that

$$\bar{p} = \mathcal{P}ar(\texttt{F0}) \Longrightarrow \mathcal{P}ar^{-1}(\bar{p}) \sim \texttt{F0}. \tag{6}$$

As a result, we obtain the *Tagged Corpus*,

$$\texttt{TC} = \{u_i = (u_i.\bar{f}, u_i.\bar{p}), \quad j = 1 \ldots N_u\}, \tag{7}$$

where $u_i.\bar{f}$ represents the function of the intonation unit and $u_i.\bar{p}$ its form. TC is divided into three different parts $\texttt{TC} = \texttt{TC}_m \cup \texttt{TC}_t \cup \texttt{TC}_e$, which will be used to model ($\texttt{TC}_m$) and train ($\texttt{TC}_t$) the classes of the dictionaries (see section 2.3) and to test ($\texttt{TC}_e$) the intonation modeling procedure, respectively (see section 2.4.1).

The *Splitting* task requires that the TIU is set a-priori. In a previous work, we provided an experimental comparison of some of these TIU for the Spanish language (Escudero and Cardeñoso, 2004). MEMOInt is independent of the chosen TIU so that it can be used as a tool to test different TIU and to contrast the experimental results.

With respect to the *Labeling* task, MEMOInt can operate with different sets of prosodic features $\mathcal{F}$. In previous works (Escudero et al., 2002; Escudero and Cardeñoso, 2003), we carried out a reviewed comparison of several proposals of $\mathcal{F}$ for Spanish language. *Evaluation* task provides a quality measure of the models as a function of the type of feature and its number of elements in $\mathcal{F}$. Additionally, MEMOInt produces a ranking of these features as will be seen in sections 3.2 and 3.6.

2.2 Parametric Representation of F0 *Contours*

Two main approaches have been followed in F0 contour analysis: the phonological approach versus the phonetic one. The phonological models make use of a code to label the characteristic movements in the contours (the most popular of this type of approaches are the autosegmental-metrical theory of intonation (Pierrehumbert, 1980) and ToBI (Silverman et al., 1992)). On the other hand, the phonetic models consider the F0 contours as a sequence of (*time*, F0(*time*)) points and the aim is to find a suitable quantitative representation of the contours. One of the most popular phonetic models is the one proposed by Fujisaki and Hirose (1984), which is based on physiological arguments. Other phonetic approaches just have the aim to parameterize the F0 contour by tuning a set of acoustic parameters accurately (Tilt (Taylor, 2000) is the most popular supported on RFC representations (Taylor and Black, 1995)). Other phonetic approaches use templates or data bases, where there is a dictionary available containing different prototypical F0 contours, and the goal is to choose the most suitable one in the data base (Emerard et al., 1992; Traber, 1992) or to adapt a template to the situation (Sproat and Olive, 1995; Kochan-

ski and Shih, 2003; Lobanov, 1987). Although some of the models are more popular than others (e.g. the Fujisaki model has been successfully applied to more than 10 languages), there is no consensus on the best methodology to be used and recent approaches do without any parameterization and simply use the points of the training F0 contours (see Tokuda et al. (2000); Eide et al. (2003); Rodríguez and Campillo (2006)).

MEMOInt is mainly focused on the possibility of contrasting different parameterization techniques in terms of predicting results and its capabilities to provide information about the typical movements of F0 contours. The basic technique to be used is the fitting of F0 contours with Bézier functions (see appendix A and (Escudero and Bonafonte, 2002)). Intonation patterns $u.\bar{p}$ are the control points (or variations) of the Bézier function fitting the F0 contour u.F0. MEMOInt permits more sophisticated parameterization techniques to be used, but the aim here is to show that the methodology permits different alternatives to be contrasted. To show this, we use different variations of the basic technique and we contrast results (see section 3.3).

MEMOInt exploits the concept of class of intonation to represent the groupings of the different patterns of intonation found in the corpus and to characterize the typical movements of the F0 contours in terms of the sequences of prosodic features which have been used to label the intonation patterns associated to every single intonation unit in the corpus. In this way, it brings information about the correspondence between the two main levels of description of the intonation information hidden in the corpus data. At the phonetic level, the different exemplars of intonation patterns in a same class convey information about the inherent variability of a prototypical pattern, after the low level variability has been smoothed by means of the F0 contour parameterization

technique. At an abstract level, it finds the best correspondence between the lists of prosodic feature values and the classes of contours, according to the objective quality measure of intonation similarity which has been incorporated into MEMOint. Elucidating the universal adequacy of this measure and of the set of prosodic features used to label the intonation units of the corpus is not a goal of MEMOint, since they serve just as input information to the methodology. Nevertheless, progressive refinement of these important inputs can also be obtained easily following an experimental procedure in which MEMOint is applied for different selections of these parameters to different reference corpora.

2.3 Intonation Modeling

The **Modeling Task** is seen as a Data Mining application into the *Tagged Corpus*. The aim is to automatically obtain the matching between the prosodic features $\bar{f}$ and acoustic parameters $\bar{p}$. As requirements we have: (1) Automatic prediction of $\bar{p}$ from $\bar{f}$, (2) robust modeling against data scarcity and (3) the outcome must provide contrastable information about the prosodic function-form correspondence in the corpus.

All the tools referred to in the introduction to implement the correspondence between $\mathcal{F}$ and $\mathcal{P}$ (neural networks, regression trees, etc.), have shown to be efficient to cope with the first requirement, but it is not clear that they can cope with the three requirements altogether. To do so, we have devised the design of a new approach, which is a multilevel clustering technique driven by a forward sequential feature selection process. The technique is inspired by classic knowledge-based agglomerative clustering (Jain et al., 1999) in combi-

nation with widely accepted feature selection techniques (Webb, 2002). Next, we describe in detail this technique.

The process starts by building an initial classification of the u_i from a single prosodic feature $\mathcal{L}^1 = \tilde{F}_1$ (therefore, $\mathcal{L}$ is a list of selected prosodic features and $\tilde{F} \in \mathcal{F}$ represents the prosodic feature selected at each step). Each class corresponds to a given value of this initial prosodic feature $\tilde{F}_1$. An agglomerative clustering technique is iteratively applied to this cluster using maximum similarity as the merging criterion and prediction accuracy of the `F0` profile as the stopping condition. The prosodic feature which gives the best overall prediction accuracy of `F0` profile over the cluster is selected as $\tilde{F}_1$. An additional prosodic feature is added to $\mathcal{L}^1$ to get the next set of prosodic features $\mathcal{L}^2 = \tilde{F}_1 \times \tilde{F}_2$ and a new cluster is built, repeating the previously described process. Again, the same criterion applies for the selection of $\tilde{F}_2$, resembling the typical forward sequential feature selection process. The clustering process stops when all the possible prosodic features have been included into

$$\mathcal{L}^{N_F} = \tilde{F}_1 \times \ldots \times \tilde{F}_{N_F} \tag{8}$$

and this results in a multilevel set of clusters, each one corresponding to an increasingly more specific set of prosodic features. Let us call

$$\mathcal{C}^k = \{C_i^k \quad | \quad i = 1 \ldots N_c^k\} \tag{9}$$

the set of classes in the cluster performed with $\mathcal{L}^k$.

The agglomeration still maintains a correspondence between the features and the parameters, if we keep track of the different values of the $\bar{f}_s^k \in \tilde{F}_1 \times \ldots \times \tilde{F}_k$ associated to a class after merging. The list of values of prosodic features

associated to a class C_j^k:

$$L_j^k = \{u.\bar{f}_s^k \in \mathcal{L}^k \quad | \quad u.\bar{f}_s^k \in C_j^k\} \tag{10}$$

provides an index to it which can be used in TTS to retrieve the $u.\bar{p}$ which corresponds to the given sequence of features annotated in the input text. The retrieved $u_i.\bar{p}$ will be used to generate the `F0` contour.

We call *dictionary*, the set of tuples

$$D_k = \{(L_c^k, C_c^k, w_c^k) \in 2^{\mathcal{L}^k} \times \mathcal{C} \times \mathbb{R}, \quad c = 1 \ldots N_c^k\}. \tag{11}$$

A dictionary is the explicit representation of the correspondence between the function of intonation ($\bar{f}_s^k \in L_j^k$) and its shape ($\bar{p} \in C_j^k$) in the class. w_j^k is the average predicting error over the samples of the training corpus belonging to the class C_j^k.

As the number of prosodic features increases, the sparse-data problem gets worse. The multilevel clustering technique provides a different dictionary D_k for every $\mathcal{L}^k$, each of which has been optimally adapted to cover the u_i set in the corpus for a given level of detail in the set of prosodic features. Since the dictionaries D_k are orderly enlarged, adding the next best predicting feature at each stage, we can use the corresponding prediction results of a training corpus to guide a search strategy for alternatives to unseen (or infrequent) $\bar{f}_s$ combinations, selecting the best predicting dictionary which subsumes $\bar{f}_s$. In Cardeñoso and Escudero (2004) we already defended the multi-dictionaries approach to cope with scarce corpus and in Escudero and Cardeñoso (2005) we analyzed different strategies to select the class to be used. Given $u.\bar{f}_s^{\,k}$ to

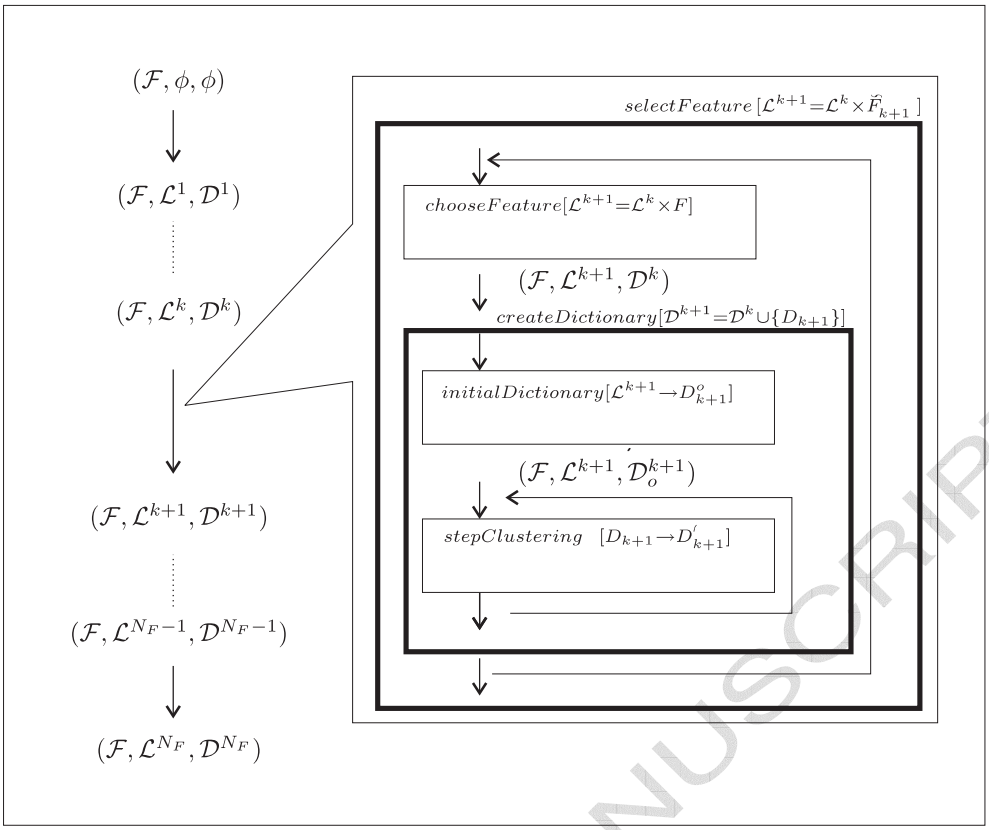


Fig. 4. Creation of the list of dictionaries.

predict $u.\bar{p}$, we have a set of k classes to use and we select the class C_c^l so that

$$c, l = \arg\min_{c,l}(w_c^l); \quad c = 1 \dots N_c^l, \quad l = 1 \dots k, \quad u.\bar{f}_s^{\,l} \in L_c^l. \tag{12}$$

Let us call *list of dictionaries* to the set of dictionaries obtained with different numbers of features:

$$\mathcal{D}^k = \{D_i, \quad i = 1 \dots k\}. \tag{13}$$

Figure 4 schematically shows the process of building this list of dictionaries, which we describe in which follows. The sequence

$$(\mathcal{F}, \mathcal{L}^1, \mathcal{D}^1) \rightarrow \dots \rightarrow (\mathcal{F}, \mathcal{L}^k, \mathcal{D}^k) \rightarrow \dots \rightarrow (\mathcal{F}, \mathcal{L}^{N_F}, \mathcal{D}^{N_F}) \tag{14}$$

is the result of the *Forward Feature Selection* algorithm, where in every step a new feature $\tilde{F} \in \mathcal{F}$ is entered to the list $\mathcal{L}^k$. The feature entered at each step

would be

$$\tilde{F}_{k+1} = \texttt{selectFeature}(\mathcal{F}, \mathcal{L}^k, \mathcal{D}^k), \tag{15}$$

so that

$$\begin{gathered}\tilde{F}_{k+1} = \arg\min_{F} \texttt{PredictionError}(\mathcal{D}^k \cup \texttt{CreateDict}(\mathcal{L}^k, F)) \\ F \in \mathcal{F}, F \notin \{\tilde{F}_1 \dots \tilde{F}_k\}.\end{gathered} \tag{16}$$

Function `PredictionError` is defined in section 2.4.1. This process is described as the iterative composition of the functions `chooseFeature` and `createDict`. Function `chooseFeature` gets one $F \notin \{\tilde{F}_1 \dots \tilde{F}_k\}$ and `createDict` adds this feature to $\mathcal{L}^k$ to create the dictionary D_k. `createDict` is the composition of `initialDict` and the iteration of `stepClustering`.

$$D_k^o = \texttt{initialDict}(\mathcal{L}^k) \mapsto \{(L_c^k, C_c^k, w_c^k)\}, \tag{17}$$

so that

$$\forall u.\bar{f}_s \in \texttt{TC}, \quad \exists L_c^k = \{\bar{f}_s^k\}, \tag{18}$$

meaning that every $\bar{f}_s^k$ observed in the corpus configures a class in the initial configuration of the dictionary. D_k^o is the starting point in the iterative application of the $D' = \texttt{stepClustering}(D)$, where

$$\begin{gathered}D = \{(L_c^k, C_c^k, w_c^k), c = 1 \dots N_c^k\}, \\ D' = \{(L'^k_c, C'^k_c, w'^k_c), c = 1 \dots N_c^k - 1\}, \\ D' = D - (L_i, C_i, w_i) - (L_j, C_j, w_j) + (L', C', w'), \\ i, j = \arg\min_{i,j} dist(C_i, C_j), \quad (i, j \in 1..N_c^k), \\ C' = C_i \cup C_j, \quad L' = L_i \cup L_j.\end{gathered} \tag{19}$$

Procedure `stepClustering` merge two classes iteratively to find the best configuration of the dictionary in terms of prediction results.

The list of dictionaries is to be used also to generate synthetic `F0` contours, both at the final stage, when TTS is the application, and at the intermediate dictionary building steps. The generation process is illustrated in Figure 5. `genF0` is used to compute the w values:

$$\begin{aligned} w_c^k &= \texttt{computeW}(D_k, c) \\ &= \tfrac{1}{N_t^{ck}} \sum_{\substack{u.\bar{f}_s{}^k \in L_c^k \\ u \in \texttt{TC}_t}} \texttt{dist}(u.\texttt{F0}, \texttt{genF0}(u.\bar{f}_s{}^k, D_k)) \\ N_t^{ck} &= |\{c \in \texttt{TC}_t : c \in C_c^k\}|. \end{aligned} \tag{20}$$

`genF0LD` is used to compute synthetic F0 contours in the training and testing stages and in text-to-speech applications.

$$\begin{aligned} &\texttt{Simulation}(C) = \bar{\mu}(\{u.\bar{p}, \quad u \in C\}), \\ &\texttt{Class}(\bar{f}_s{}^k, D_k) = C_c^k \iff \bar{f}_s{}^k \in L_c^k, \end{aligned} \tag{21}$$

$\texttt{SelectClass}(\bar{f}_s{}^k, \mathcal{D}^k) = C_c^l$ by solving equation 12.

2.4 Evaluation

In an ideal scenario, where the contour stylization technique and the contour similarity measurement could be completely derived from perceptual experiments, the grouping of intonation profiles into perceptually disjoint classes would provide a correct and complete model of intonation for the language. Unfortunately, there is no evidence yet that this can be acomplished and we

will have to accept that the grouping we are obtaining from a corpus might be affected by the lack of information associated to an imprecise theoretical model. Nevertheless, we can provide relative indicators of quality of the results at two different levels: the quality of the generated contours, in terms of a reasonable similarity metric, and the soundness of the prosodic information provided by MEMOint on the given corpus, in terms of a comparative reference with well established linguistic models. These are the two main aspects of the *Evaluation Task* in MEMOint, which will be described in the present section.

2.4.1 Quality of synthetic contours

Although subjective perceptual tests are still the best alternative to measure the quality of synthesized speech, their high cost discourages their use and they have to be discarded when the quality assesment has to be integrated algorithmically, as in MEMOint, where we have to test the quality of the selected parameters (`TIU`, `LPF`, `PT`) at each step of the agglomerative process. For this reason, we have to rely on objective quality measurements to test the synthetic contours. As we will illustrate in the experiments for Spanish reported in next section, experiments could be carried out to validate the correspondence between the outputs of perceptual studies and objective measuremts. In consequence, we will use the typical RMSE prediction error as the kernel of the evaluation of the quality of synthetic contours. This is a reference metric well established in the literature and could be, nevertheless, replaced by any other better metric in future, provided additional evidences in favor of that replacement would arise.

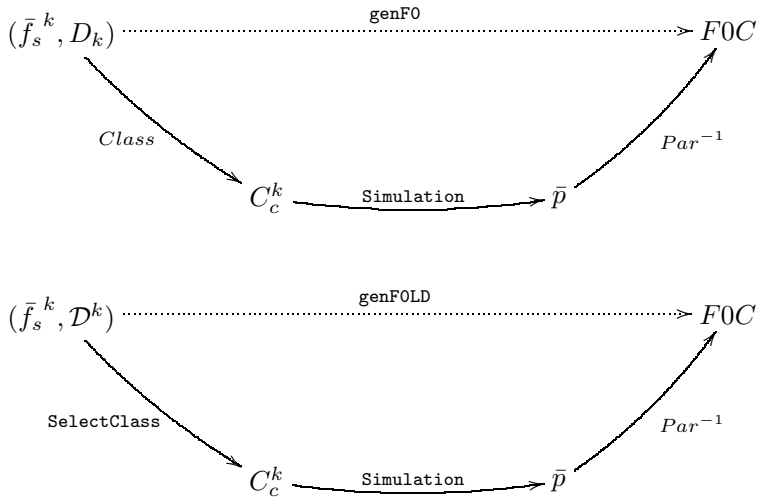


Fig. 5. Generation of synthetic F0 contours.

If $u.\bar{p}'$ and $u.\texttt{F0}'$ are the synthetic acoustic parameters and F0 contour respectively, then (see figure 5):

$$\begin{aligned} & u.\texttt{F0}' = \texttt{genF0LD}(u.\bar{f}_s{}^k, \mathcal{D}^k) = \\ & \mathcal{P}ar^{-1}(u.\bar{p}') = \\ & \mathcal{P}ar^{-1} \circ \texttt{Simulation} \circ \texttt{SelectClass}(u.\bar{f}_s{}^k, \mathcal{D}^k). \end{aligned} \tag{22}$$

The use of statistical simulation can potentially increase the naturalness of the synthetic speech, because the simulation can reproduce both what is regular in the classes and its variability. Although this has been reported in previous works (Escudero (2002)), in this paper we have decided to use just the mean value because, directly from its definition, it would be the best canonical representative of the class to ensure the minimum RMSE value.

The evaluation consists of computing the distance between the synthetic F0 contours and the real ones in $\texttt{TC}_e$:

$$\texttt{PredictionError}(\mathcal{D}) = dist(\bigcup_{i=1}^{N_{Ce}} \texttt{F0}'_i, \bigcup_{i=1}^{N_{Ce}} u_i.\texttt{F0}) \tag{23}$$

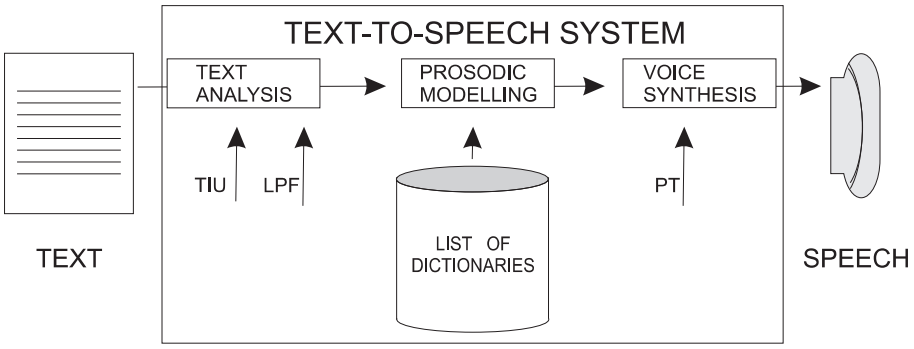


Fig. 6. Use of *List of Dictionaries* in TTS Systems.

so that

$$\texttt{F0}'_i = \texttt{genF0LD}(u_i.\bar{f}_s, \mathcal{D}), u_i \quad i = 1 \ldots N_{C_e}, N_{C_e} = |\texttt{TC}_e|. \tag{24}$$

RMSE and correlation metrics are used to obtain a well known distance, as justified in Hermes (1994).

The interest of intonation models in TTS has lessened due to the use of synthesis techniques based on the speech-unit selection (Eide et al., 2003) or HMM synthesis (Tokuda et al., 2000). Nevertheless, the prediction of realistic target F0 contours is still useful for guiding the search of units in the corpus (Rodríguez and Campillo, 2006; Eide et al., 2003). A list of dictionaries $\mathcal{D}^{N_F}$ can be used on *TTS* systems following the process shown in Figure 6. The *Text Analysis* module identifies the intonation units and it assigns $u.\bar{f}_s^{N_F}$. The prosodic module consults $\mathcal{D}^{N_F}$ to get a synthetic pattern $u.\bar{p}$. The *Synthesis Module* uses $u.\bar{p}$ and the function $\texttt{genF0LD}(\bar{p})$ to perform the synthetic F0 contour and uses it to synthesize the voice.

2.4.2 Assesment of prosodic information

MEMOInt provides information concerning both the behavior of the parameters and the visualization of the shape and function of the intonation. As for

MEMOInt parameters, the quality metrics allow the behavior of the systems to be evaluated when different values for the type of intonation unit, prosodic features and parameterization technique are tested. Also, $\mathcal{L}^{N_F}$ tell us about the relative importance of each of the features in $\mathcal{F}$ in the intonation modeling process.

As for the visualization of the intonation, we should note that $\mathcal{D}^{N_F}$ is an explicit representation of the form of the intonation found in the corpus (the classes of patterns) and of its function (the $\mathcal{L}$ indexing the classes). This information can be checked in different ways, which can be useful to assess the typical patterns from the corpus and the relative importance of the features in $\mathcal{F}$. Every class C_j^i in $\mathcal{D}^{N_F}$ contains the typical `F0` contour movements in the corpus and also its variability. Both aspects are important to properly describe the relation function-form. The different L_j^i in $\mathcal{D}^{N_F}$ contain information about the function of the intonation associated with the corresponding shape C_j^i.

The ordered list of dictionaries provides a way to build a graph of classes which conveys schematic visual information about the intonation patterns found in the corpus and their corresponding labels of prosodic features (see Appendix B for an explanation and section 3.5 for a more detailed discussion of the use of this graph in the experiments for Spanish language). This representation tools can be used to validate the correctness of the particular prosodic information we extracted from the corpus, comparing it with the rules of a reference theoretical model.

3 Experimental Results

The framework shown in previous section is now tested with a corpus in Spanish. First we describe the corpus to be used, the type of intonation units, the set of prosodic features and acoustic parameters considered. Then, we show some peculiarities of the construction of the dictionaries and finally we make comments on the information retrieved from the corpus and we report on the quality tests.

3.1 The Corpus

The corpus used is made up of more than two hours recorded in studio conditions. We select the part of the corpus consisting of the reading of a set of phonetically balanced sentences. The `F0` contour was captured with a laryngograph device. The corpus was recorded to implement a TTS system (MLTTS) based on concatenating units by the TALP research group [1] (see Ferrer (2001) for a more detailed description of MLTTS and the corpus). Although it is not a specific corpus for intonation modeling, its size makes it suitable to be used in MEMOInt.

The corpus contains 1646 declarative intonation groups, 80 interrogative ones and 19 exclamative (4365, 247, 53 stress groups of each type respectively). We focus on declarative sentences because the corpus is scarce in interrogative and exclamative sentences. MEMOInt can also be applied when there are few samples of any type, but in this case we select declarative because our aim is

[1] Research Center for Technologies and Applications in Language and Speech. URL: `http://www.talp.upc.es/talp`

to show the representativeness of the resulting models.

For experimental purposes, 75% of the sentences belongs to training data and the other 25% are testing data ($\texttt{TC}_e$). Training data are also splitted into two sets: 75% for modeling ($\texttt{TC}_m$) and 25% for validation ($\texttt{TC}_t$).

3.2 Intonation Units and Prosodic Features

Three different types of intonation units have been considered and tested in this work: Intonation Group (*IG*), Stress Group (*SG*) and Syllable (*Syl*).

Intonation Group Defined as the parts of the sentence separated by a pause or by a movement in the `F0` contour that is more salient than others in the utterance (Quilis, 1993). This unit has been used to describe Spanish intonation in simplistic approaches like the one in Alarcos (2002) and it has also been used in combination with other units in superpositional models as in Gutierrez et al. (2001). The following set $\mathcal{F}$ is labeled: linguistic features: 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); features measuring the size: 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*; and another feature related to the position of the *IG* in the sentence `posIGSE` (7 values).

Stress Group Defined as a set of syllables where only one is stressed. The *SG* has been used in multiple approaches to describe the Spanish intonation. The most complete study is Garrido (1996) and it has also been used as the

basic unit in Sosa (1999) to apply the autosegmental approach to Spanish intonation.

Three alternative definitions of the SG have been tested:

SG1 Defined as the set of words where only the last one is stressed (Garrido, 1996).

SG2 Defined as the set of syllables where only the first syllable is stressed (Sproat and Olive, 1995).

SG3 Defined as a stressed syllable plus the preceding and/or following ones provided they are not stressed.

We label the following set $\mathcal{F}$: position of the SG in its IG `posSGIG` (6 values), number of syllables `nSylSG` (9 values) and phonemes `nPhonSG` (6 values) in the SG, position of the stressed syllable `posSTSG` (3 values). Furthermore, the SG inherints the $\mathcal{F}$ of the IG it belongs to. SG2 and SG3 versions need to know the position of the SG boundary with respect to the boundary of the stressed word `SGBorder` (3 values).

Syllable Using the classical definition for Spanish (Alarcos, 2002). The syllable has been the unit of reference in some engineering approaches to modeling intonation (López and Rodríguez, 1996; Vallejo, 1998). For every Syllable, we label its number of phonemes `nPhonSyl` (4 values), the position of the syllable in the SG `posSylSG` (4 values) and if it is accented or not `accent` (1 value). Additionally, the syllable inherits the $\mathcal{F}$ of the IG and the SG it belongs to.

`intBez`

RMSE(Hz)	Number of Parameters						
Type of Intonation Unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Intonation Group	24.29	21.5	21.1	20.7	20.68	**20.6**	**20.62**
Stress Group 1	20.83	18.67	18.47	**18.12**	18.16	**18.13**	18.25
Stress Group 2	20.83	19	18.5	**18.24**	**18.29**	18.42	18.46
Stress Group 3	20.56	18.27	**17.98**	**18.02**	18.06	18.06	18.21
Syllable	19.19	18.6	18.32	**18.29**	18.54	18.63	**18.29**

`intLin`

RMSE(Hz)	Number of Parameters						
Type of Intonation Unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Intonation Group	24.29	21.5	21.18	20.84	20.72	**20.58**	**20.64**
Stress Group 1	20.83	18.67	18.2	18.15	**18.05**	**18.04**	18.1
Stress Group 2	20.83	19	18.42	18.29	**18.28**	18.39	**18.24**
Stress Group 3	20.56	18.27	18.13	**18**	**18**	18.08	18.14
Syllable	19.19	18.6	**18.32**	18.57	18.33	**18.21**	**18.32**

`s-intBez`

RMSE(Hz)	Number of Parameters						
Type of Intonation Unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Intonation Group	24.29	21.42	21.11	20.69	20.6	**20.54**	**20.52**
Stress Group 1	20.79	18.54	18.15	18.04	**17.92**	**17.92**	17.99
Stress Group 2	20.8	19.06	18.36	**18.12**	**18.12**	**18.1**	18.23
Stress Group 3	20.51	18.35	18.1	**18.06**	18.1	**17.92**	18.08
Syllable	19.18	18.84	**18.38**	**18.38**	18.44	18.4	18.53

Table 1

Selection of the acoustic parameters: Mean prediction errors of the `F0` contours of the *Training Corpus* using different type of intonation units, different types of parameterization technique and different number of acoustic parameters. Results have been obtained without applying the agglomerative process in order to avoid the impact of merging classes, to be evaluated later. Boldface has been used to highlight minimum values.

3.3 Acoustic Parameters

We test 3 different alternatives that are variations of the same basic technique based on Bézier function fitting:

intBez where the parameters $\mathcal{P}$ are the control points of the fitting Bézier

function. That is $u.\bar{p} = (Y_0 \dots Y_n)$, being n the degree of the curve and $\bar{P}_j = (X_j, Y_j) \quad j = 0 \dots n$ the control point j of the Bézier function (more about Bézier functions in the appendix A). We use a variation where the parameters are equispaced points interpolating the fitting Bézier function. That is: $u.\bar{p} = (y(t_0) \dots y(t_n))$ with, $t_j = j/n \quad j = 0 \dots n$, $t \in [0, 1]$; where $x(0)$ and $x(1)$ are the initial and final time of u respectively, $Q(t) = (x(t), y(t))$ is the fitting function.

intLin where the parameters are the $n+1$ vertex of the fitting polyline $PL = \{Y_j, \quad j = 0 \dots n\}$ following the fitting method presented in the appendix A.2 (straight segments fitting).

s-intBez this method uses `intBez` but the `F0` contour is smooth before parameterizing.

For each of these alternatives we test a different number of parameters and we have computed the RMSE prediction errors as displayed in table 1. In all the cases there are an optimum number of acoustic parameters (e.g. in the case `intBez`, `SG1` the best result is obtained with 4 parameters (RMSE=18.12Hz)). The interpretation to this fact is that it is necessary to have a minimum number of parameters to fit the prototypical movements of `F0` (e.g. a stress group in Spanish can have a maximum and a minimum requiring 4 parameters (a degree 3 Bézier curve)). When this minimum is exceeded, the parameters can fit other micro-intonation effects that are less interesting in our approach, causing the quality of the models to decrease. This can be contrasted in table 1 comparing the results of the tables `intBez` and `s-intBez`: when the micro-intonation is reduced by filtering the `F0` contours, more parameters can be accepted without decreasing the quality.

With respect to the comparison of the parameterization techniques (`intBez` versus `intLin`), it seems that similar results can be obtained but more parameters are required when the straight segments techniques are used. The plasticity of the Bézier curves permits fitting `F0` contours using fewer parameters.

Concerning to the type of intonation unit to use, it seems that `SG3` has some advantages with respect to `SG1` and `SG2`. For `IU` and `Syl`, note that results are difficult to compare because they are highly dependent on the number of prosodic features to be used (different number of classes in each case). Thus, it is necessary to apply the agglomerative process for contrasting the effect of the use of the type of intonation unit as will be seen in the following section.

From these conclusions, we decide to use `4` parameters and `intBez` in the following experiments. The results obtained after the agglomeration process, will be useful to decide about `TIU`.

3.4 Construction of the Dictionary

Figure 7 monitors the building process of the list of dictionaries. The RMSE values are computed given $\texttt{TC}_t = \{u_i, \quad i = 1 \dots N_{C_t}\}$ and the list of dictionaries $\mathcal{D}$ as:

$$\begin{aligned} \texttt{Error}(\mathcal{D}) &= \tfrac{1}{N_{C_t}} \sum_{i=1}^{N_{C_t}} dist(\texttt{F0}'_i, u_i.\texttt{F0}), \\ \texttt{F0}'_i &= \texttt{genF0LD}(u_i.\bar{f}_s, \mathcal{D}). \end{aligned} \tag{25}$$

The minimum values of the lines indicate the optimum number of classes in the dictionaries. The legend of the lines indicates the relevance ranking of the

features. From a certain number of features the results hardly improve. This is due to redundancy in the selections of the features as will be shown in section 3.6.

Table 2 shows that the models in the dictionaries start to be over-trained from a number of features on. In training, the higher the number of features, the better the results, but this is not the case when testing. This is because the agglomeration of classes is guided by the samples in *Training Corpus* and *Modeling Corpus*. This is unacceptable in recognition applications, but in synthesis we should mimic the intonation of the corpus: worse RMSE values do not necessarily mean worse intonation but different intonation.

Table 3 illustrates why we need a contour selection strategy based on multiple levels of dictionaries. Dictionaries $D_i \in \mathcal{D}^N, N < 7$ are selected to predict more than 50% of the testing and training samples. The difference is bigger when testing, because of the higher likelihood to find samples whose $u.\bar{f}_s$ did not appear in the training stage. Table 4 shows that only 23 out of the 119 available classes in $D7$ were used, meaning that MEMOInt could find out that for a given N ($N = 7$ in this case), D_N is not the dictionary which gives the best prediction results for all inputs (as a further difference with a typical tree-based classification procedure).

Table 4 illustrates also the need for the agglomerative process due to the high number of classes in the initial state: D_7 has 2026 classes in the initial configuration and 119 after the agglomeration. Furthermore, the 2026 classes are not all the possible ones: if the corpus had samples to cover all the possible combinations of features, the number of classes would be `2 (accent)`$\times$ `6 (posSGIG)` $\times$ `7 posIGSE` $\times$ `9 (nSylSG)` $\times$ `6 (nSGSE)` $\times$ `6 (nPhonIG)` $\times$

List of Dict.	Train		Test	
	RMSE(Hz)	**Corr**	**RMSE(Hz)**	**Corr**
LD1	22.63	0.53	22.50	0.54
LD2	19.72	0.68	20.00	0.67
LD3	19.00	0.70	19.02	0.70
LD4	18.43	0.72	18.71	0.72
LD5	17.75	0.75	18.74	0.72
LD6	17.06	0.77	19.02	0.71
LD7	16.42	0.79	19.11	0.71

Table 2

Prediction Errors: *RMSE* and *Corr* versus the number of features in the *training* and and test stage. We use `TIU=SG2`, `PT=intBez`, and N_P=4

	Use of the Dictionary (%)						
LD7	D1	D2	D3	D4	D5	D6	D7
Train	0.0	2.0	8.9	3.8	8.3	27.4	49.7
Test	0.0	5.1	16.4	9.4	14.9	23.5	30.7

Table 3

Use of the dictionaries in $\mathcal{D}$: each cell contains the percentage of samples that are predicted using each dictionary in the list. We use `TIU=SG2`, `PT=intBez`, and N_P=4.

`6 (nSylSE) = 163296`. Although some of the combinations are impossible, the figure is illustrative of the magnitude of the corpus required and of the need for having the list of dictionaries to select alternative dictionaries when a combination was not seen during the training.

On the other hand, table 4 also shows that the larger the number of features, the greater the accuracy of the class (fewer intra-class distance), but the less representative it will be (smaller number of samples). Indeed some of the classes have less than 10 samples which can be assumed in synthesis applications, but its use to contrast information in the corpus is problematic.

List of Dictionaries LD7	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Number of classes with more 10 simples	2	5	24	19	24	21	23
Number of used classes	2	5	25	19	25	21	23
Number of classes in the final configuration	2	5	57	55	95	120	119
Initial number of classes	2	10	68	294	785	1449	2026
Mean number of samples per class	1235	494	84	69	36	35	24
Mean RMSE intra-class	37	32	30	27	20	18	18

Table 4

Description of the dictionaries in terms of number of classes and number of samples per class. We use `TIU=SG2`, `PT=intBez`, and N_P=4

3.5 Visualization of Intonation Patterns

The list of dictionaries $\mathcal{D}$ can be used to visualize the association between prototypical patterns in the corpus and the sequences of prosodic features. As showed in appendix B, an intuitive, appealing and easy-to-understand representation of $\mathcal{D}$ can be used which resembles the one provided by classical decision trees. We do this by means of a directed graph in which the classes of the dictionaries at different levels are connected in terms of the prosodic features associated with them. Every $\bar{f}_s^k(l) \in L_j^i, \quad l = 1 \ldots N_{L_j^i}$ labels a path from the root node to a node named $n_j^i(l)$. The node $n_j^i(l)$ is coloured with the corresponding class C_j^i and its w_j^i value. Given any node $n_j^i(l)$ determined by a vector of features $\bar{f}_s^k = (f_1, \ldots, f_k)$, the set of nodes $\{n^{i+1}\}$ linked with it are the ones determined by the arrays $\bar{f}_s^{i+1} = (f_1, \ldots, f_k, f_{k+1})$ with $f_{k+1} \in \tilde{F}_{k+1}$. Nodes which are never used as a consequence of the selection procedure of MEMOint (higher w_j^i value), can be removed from the graph or represented as empty nodes.

The meaning of the relation represented in the edges of this graph is different than the one in a classical decision or regression tree. In decision trees, it just represents class specialization derived from the inclusion of an additional

feature. Here, the full path from the root node to any given node $n_j^i(l)$ contains the set of all classes which could be potentially used to represent the sequence of prosodic features $\bar{f}_s^k(l)$. Each element of the sequence $\bar{f}_s^k(l)$ labels links between related nodes in the graph. For a given graph level k, nodes are labelled with the names of the classes from the dictionary D_k. All the nodes $n_j^k(l), \quad l = 1 \ldots N_{L_j^k}$ are coloured with the same class C_j^k and the fact that a class could be labelling more than a node at the same level is another relevant difference with the standard interpretation of a decision tree.

The visualization of the information in the graph allows to contrast some of the assessments found in the bibliography about Spanish Intonation. In Escudero (2002), an overview of the proposals of several authors can be found. Here we review the main assessments and we contrast them with plots in Figure 8. This figure selects from the graph the branches that are relevant to discuss about the prominence, structure of the stress groups, and junctures observed in the corpus:

- **Prominence** (or relative importance of the stress group with respect to the others) was labeled in the corpus with the prosodic feature `accent`. Observations of the intonation of the corpus projected in figure 8 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**: Face (2001) observed that the prototypical patterns associated to the Spanish stress groups are $L*+H$ pattern and the less frequent $L+H*$ one (using TOBI notation). This fact can be observed in the tree shown in figure 8, where $L*+H$ patterns appear in `C4_104, C4_76, C4_110, C4_144, C4_146`. The pattern $L+H*$ 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 and here are marked by the features `GEFinal` and not `GAFinal`. 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 Navarro-Tomás (1944)).
- **Final boundary**: affecting the last part of the `F0` contour. Typical final juncture of declarative sentences is $L*L\%$ according to Sosa (1999) or $H+L*L\%$ according to Beckman et al. (2000). This pattern is clearly seen in Figure 8 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 8 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.

3.6 Ranking of Prosodic Features

The constructive process of the dictionaries offers a ranking of the prosodic features described in $\mathcal{L}^{N_F}$. This ranking is an objective indicator of the relative relevance of the prosodic features with respect to the shape of the patterns of intonation. This ranking can be validated by measuring the entropy of the different features to classify tree the classes of patterns obtained using a *kmeans* clustering. This process was explained in Escudero and Cardeñoso (2003) and figure 9 shows the results.

The feature rankings obtained by this method and the one obtained from $\mathcal{L}^{N_F}$ are similar: the greater the informative capabilities of a prosodic feature, the sooner it is selected in the building of the list of dictionaries. The exceptions to this rule arise when there are redundant features: correlated features or features that can be obtained as a combination of other. As an example of correlated features in Figure 9, note that the feature `NPhonSG` in the plot `IU=SG1` has an informative value higher than other features that are better in the ranking due to the previous selection of the correlated feature `NSylSG` (*Pearson Correlation* $\rho > 0.9$). As an example of combination of features, `SGBorder` has important informative capabilities in `IU=SG2` but it is not chosen rapidly in $\mathcal{L}^{N_F}$ because it can be deducted from `posSTSG` and `nSylSG` by the

application of a single formula.

With respect to the influence of the stress, SG1 and SG2 do not reflect its influence (posSTSG low value), but this factor is important in SG3 (nSylSG high value). It seems that the parameterization technique filters the effect of the stressed syllable position explaining why SG3 offers better predicting results than SG1 and SG2.

Figure 9 shows that less informative prosodic features are inserted later in the list of dictionaries. Obtaining their informative capabilities can be a good indicator to select or discard a feature prior to the application of MEMOInt reducing the time consumed in the creation of the list of dictionaries.

3.7 Perceptual Validation

Table 5 compares the results of the objective test when it is applied to the different type of intonations units studied. It is important at this point to apply a subjective test in order to get the opinion of a group of evaluators about the quality of the synthetic F0 contours generated by MEMOInt, and also to show that the differences in quality observed in the objective test have a perceptual counterpart.

To do the test, the sentences of the *Testing Corpus* are re-synthesized using the generated synthetic pitch contours. To do so, we use the re-synthesis PSOLA module included in the praat http://www.praat.org software. Before applying the synthetic pitch contour, it is smoothed to reduce F0 jumps in between the intonation units. Gaps are linked with straight segments and the whole contour is filtered, averaging the samples: $\mathtt{F0}_i = \sum_{j=-M}^{N} \mathtt{F0}_{i+j}$ with

TIU	**IG**	**SG1**	**SG2**	**SG3**	**Syl**
RMSE(Hz)	20.89	19.18	18.71	18.49	18.50
Corr	0.58	0.66	0.72	0.72	0.70

Table 5

Objective evaluation in function of the type of intonation unit. We use`PT=intBez`, and N_P=4

		Subjective		**Objective**			
		Correctne(((0-5)		**RMSE**		**Corr**	
Type	**N**	**Mean**	σ	**Mean**	σ	**Mean**	σ
REAL	64	4.64	0.64	0.00	0.00	0.00	0.00
SG3,11	57	4.13	0.88	17.32	4.15	0.76	0.12
Syl,11	70	4.01	0.92	20.56	4.14	0.66	0.13
SG3,2	70	3.81	1.18	20.04	5.56	0.70	0.17
Syl,2	69	2.83	1.19	20.68	5.08	0.59	0.13

Table 6

Perceptual test results. N is the number of evaluations received. *Correctness* is the mark assigned by the evaluation to the synthetic utterance (5-Perfect 4-Very Good 3-Good 2- Acceptable 1-Bad 0-Very Bad). *RMSE* and *Corr* are the distance metrics of the real `F0`-contours of the sentence to test and the synthetic ones. We use `PT=intBez`, and N_P=4

$\texttt{F0}_i$ the point i of the `F0` contour and $M = N = 5$.

Each member of a group of listeners assigns a mark from 0 to 5 to a series of 5 sentences randomly chosen from the *Testing Corpus*. Each of the sentences is uttered 3 times using 3 different versions chosen randomly from 5 possible ones. The 5 possible versions are: (1) REAL, consisting in the PSOLA synthesis of the sentence using the original `F0` contour; (2) The PSOLA synthesis using the `F0`-contour generated by using `SG3` and `2` features; (3) like (2) but using `11` features; (4) like (2) but using `Syl`; and (5) that is like (4) but using `11` features. We do not give references to the listeners (neither the real utterance

nor the worst case with flat F0 contours as, for example, in Bulyko et al. (1999)) but they know a perfect utterance could be shown, so that the qualification of the listener can also be evaluated. We choose SG3 and Syl because they appear to be the best TIU to be used (see table 5) and 2 and 11 prosodic features because table 2 shows that the number of prosodic features cause an important difference in the prediction results.

Table 6 shows the results of the test. With respect to the number of parameters, results are better with 11 parameters than with 3 parameters with statistically significant differences ($H_0 : (\mu_{Syl,11} - \mu_{Syl,3}) = 0; H_a : (\mu_{Syl,11} - \mu_{Syl,3} > 0)$; $P - value = p < 0.05$ and for SG3 $p = 0.042$). The differences with respect to REAL utterances are statistically significant for Syl,11 ($p < 0.05$) and for SG3,11 ($p =, 001$) but the high scores assigned by the listeners (> 4) indicate a satisfactory degree of acceptance. SG3,11 is the best option, but Syl,11 gives comparable results (no statistically significant difference with$p = 0, 41$). The use of SG3 instead of Syl has important advantages from the point of view of the computational cost of the creation of the dictionaries. In view of this evidence of subjective satisfaction, the most interesting result in table 6 is the good correspondence between perceptual validation results and objective RMSE values: the higher the user satisfaction, the smaller the RMSE distance between F0 contours.

4 Conclusions

This article presents a modeling technique that has shown to be able to generate synthetic intonation of an acceptable quality evaluated with objective and subjective tests. The main contribution of this technique is that it offers

a methodological framework that permits the intonation of a given corpus to be analyzed from different points of view. The modeling technique is based on a data mining technique which combines *Sequential Feature Selection* and *Agglomerative Clustering* techniques, and it has shown to be efficient in sparse data conditions. The robustness is increased due to the possibility of selecting a dictionary in a list according to their predicting capabilities.

MEMOInt is able to analyze the corpus using different types of intonation units. This allows the efficiency of the type of intonation units to be compared to characterize the corpus. For Spanish, we have seen that the Syllable is the type of intonation unit which results in the best prediction quality of the synthetic `F0` contours, although Stress Group is a perfect alternative within similar quality results at coarser levels.

The list of dictionaries have been shown to be a useful tool to match the characteristic patterns and the prosodic features associated with them. For Spanish, the visualization of prosodic information derived from the dictionaries provides a good correspondence with the properties found in Spanish phonetics bibliography which reinforces the idea that this visualization feature of MEMOint could provide a valuable research tool for the community.

MEMOInt provides a ranking of the set of prosodic features in terms of their relevance to predict intonation contours. These rankings compare correctly with reference values provided by classical entropy based rankings, both at the dictionary construction stage and at the generation stage.

Preliminary perceptual tests of the synthetic utterances reflect the high quality of the generated `F0` contours. Also, the results of this perceptual test show a good correlation with the objective RMSE metrics which were applied to drive

MEMOint evaluation tasks.

With respect to future work, MEMOInt is a software tool to be applied to different languages and different corpora. MEMOInt is being applied to contrasting different Iberian languages to establish the prosodic features and the patterns that determine the perceptual differences between different languages or dialects. Another aspect to explore in future work is the analysis and synthesis of the intra-class variability. The analysis will permit information to be obtained about the `F0` pattern stability and its relation with the need to include additional features to specialize the classes. In synthesis, this variability could be reproduced to generate more natural speech.

5 Acknowledgments

This work has been partially sponsored by Spanish Government (MCYT project TIC2003-08382-C05-03) and by Consejería de Educación of the Junta de Castilla y León (JCYL project VA053A05).

A Parameterization with Bézier Functions

A.1 Bézier Functions

A Bézier function is a parametric curve given by a set of control points approximating and/or interpolating the curve (see Farin (1996)). In two dimensions, given $\bar{P}_0, \bar{P}_1 \ldots \bar{P}_n \in \mathbb{R}^2$ and $t \in \mathbb{R}$, let the *Bézier curve* $\bar{Q}(t) = \sum_{i=0}^{n} \bar{P}_i \cdot B_i^n(t)$ with $t \in [0,1]$ where $\bar{P}_i = (X_i, Y_i)$ with $i = 0 \ldots n$ are the $n+1$ control points

of the Bézier curve $\bar{Q}(t)$ of degree n. B_i^n, with $i = 0 \ldots n$ are the $n+1$ Bernstein polynomials of degree n, explicitly defined as $B_i^n(t) = \binom{n}{i} t^i (1-t)^{n-i}$.

Bézier curves can be restricted to the case of functional curves called *Bézier functions*. A Bézier function has the form $y = f(x)$, where f is a polynomial. This is written in as a parametric function: $\bar{Q}(t) = (x(t), y(t)) = (t, f(t))$. In terms of the Bernstein polynomials: $f(t) = \sum_{i=0}^{n} P_i \cdot B_i^n(t), \quad t \in [0,1]$ where P_i are real numbers. The control points are now $(j/n, P_j)$; with $j = 0 \ldots n$. Control points are equi-spaced in the axis of abscissas. Considering the interval $[a, b]$, instead of $[0, 1]$, the values of abscissas are $x(t) = a + t(b-a); \quad t \in [0,1]$.

The fitting problem consists of representing a sequence of points (in our case `F0` contours) with a Bézier function. The goal is to minimize the error of approximation between the function and the sequence of points to be fitted. If $u.\texttt{F0} = \bar{p_j} \quad j = 1..p$ are the p points time-frequency $p_j = (t_j, \texttt{F0}_j)$ of the `F0` contour in a intonation unit u, the corresponding acoustic parameters $u.\bar{p}$ are the control points of the fitting Bézier function obtained by minimizing $R = \sum_{j=0}^{p} (Q(t_j), \bar{p}_j)^2$ by using the square minimum method. (see Peña (1999) for minimum squares, and Bartels et al. (1986) Plass and Stone (1983) for Bézier curves fitting).

A.2 Fitting with polylines

The classical stylization consists of approximating `F0` contours with straight segments. The stylization is based on the idea that the original contour and the stylized one are perceptually equivalent Hart et al. (1990). Here we propose a method to stylize `F0` contours of the intonation unit by using a set of N

straight segments defined by $N+1$ vertex. By analogy with the fitting with Bézier functions, the vertex must be equi-spaced in the axis of abscissas.

To implement the fitting we use linear regression in N intervals. The goal is to find the polyline L defined by the set of vertex $\bar{P_i} = (X_i, Y_i)$ with $i = 0 \dots N$ so that $R = \sum_{j=0}^{p} (L(x_j) - y_j)^2$ is to be minimized, (x_j, y_j) with $j = 0 \dots p$ being the sequence of points of `F0` to be fitted and $(x_j, L(x_j))$ the fitting points.

X_i are $X_i = X_0 + i \cdot (X_N - X_0)/N$. The acoustic parameters will be Y_i with $i = 0 \dots N$. The equation of the N intervals of the polyline $L = (X, Y)$ is defined as: $X = X_i + t \cdot (X_{i+1} - X_i) \quad Y = Y_i + t \cdot (Y_{i+1} - Y_i) \quad i = 0 \dots N-1$ with $t \in [0, 1]$

The `F0` contours are divided in N intervals I_i, with $i = 0 \dots N$, where $I_i = \{\bar{p_j} = (x_j, y_j) \quad | \quad X_i \leq x_j \leq X_{i+1}, \quad 0 \leq j \leq p\}$. Given a point $p_j = (x_j, y_j)$ in `F0`, its corresponding value in the polyline is obtained by making $X = x_j$ and solving $y_j = Y$. By minimizing R with respect to the parameters Y_i we have:

$$\begin{aligned}
\frac{\partial}{\partial Y_0} R &= 2 \cdot \sum_{p_j \in I_0} (1 - t_j) \cdot ((1 - t_j) \cdot Y_0 + t_j \cdot Y_1 - y_j) = 0 \\
\frac{\partial}{\partial Y_l} R &= 2 \cdot \sum_{p_j \in I_l} t_j \cdot ((1 - t_j) \cdot Y_{l-1} + t_j \cdot Y_l - y_j) + \\
&\quad + 2 \cdot \sum_{p_j \in I_{l+1}} (1 - t_j) \cdot ((1 - t_j) \cdot Y_l + t_j \cdot Y_{l+1} - y_j) \quad l \in [1, N-1] \\
\frac{\partial}{\partial Y_N} R &= 2 \cdot \sum_{p_j \in I_N} t_j \cdot ((1 - t_j) \cdot Y_{N-1} + t_j \cdot Y_N - y_j) = 0
\end{aligned}$$

These are N equations with N unknown factors. If any of the I_i intervals has no points, the interval is joined with the following or the preceding one.

B Graph of classes

This appendix illustrates the construction of the visualization graph of classes with a simplified example. In this example we use three fictitious prosodic features (feature 1, feature 2 and feature 3), each of them having two possible values (A and B). These features are used to characterize a two dimensional acoustic parameters space. Figure A.1 provides a graphical illustration of the creation of the list of dictionaries. After the agglomerative process, the three dictionaries of the list can be described as the set of classesC_j^i with $i = 1, 2, 3$. In this figure, C_i_j represents C_j^i, `L` represents L_j^i and `w` represents w_j^i, in terms of the nomenclature introduced in section 2.3.

Figure A.2 illustrates the step by step process to build the graph of classes used for visualization of prosodic information in MEMOint. Plot (1) shows a graph representing the classes in the initial configuration. Plot (2) shows the classes which would be grouped and Plot (3) the resulting graph representing the final configuration. The final graph is also represented in Plot (4) but it has now been reduced to a direct graph where some nodes could be replicated. Plot (5) is the graph which results after removing the nodes which were never used because the w values associated with them forced selection of an ancestor. Although plot (5) is not a decision tree, it can be interpreted the same way now, although there could be replicated classes.

References

Aaron, A., Eide, E., Pitrelli, J. F., 2005. Conversational computers. Scientific American June, 64–70.

Alarcos, E., 2002. Gramática de la Lengua Española. Real Academina Española.

Allen, J., Hunnicutt, M. S., Klatt, D., 1987. From Text to Speech: The MITalk System. Cambridge University Press.

Bartels, R. H., Beatty, J. C., Barsky, B. A., 1986. An Introduction to Splines for use in Computer Graphics and Geometric Modeling. Morgan Kaufmann Publishers, Inc.

Beckman, M. E., Campos, M. D., McGregory, J. T., Morgan, T. A., 2000. Intonation across spanish, in the tones and break indices framework. Tech. Rep. http://www.ling.ohio-state.edu/~tobi/sp-tobi/, University of Ohio.

Botinis, A., Granstrom, B., Moebius, B., July 2001. Developments and paradigms in intonation research. Speech Communication 33, 263–296.

Bulyko, I., Ostendorf, M., Price, P., 1999. On the relative importance of different prosodic factors for improving speech synthesis. In: Proceedings of ICPhs 99. pp. 81–84.

Campbell, N., Erickson, D., 2004. What do people hear? a study of the perception of non-verbal affective information in conversational speech. Journal of the Phonetic Society of Japan 8 (1), 9–28.

Cardeñoso, V., Escudero, D., 2004. A strategy to solve data scarcity problems in corpus based intonation modelling. In: Proceedings of ICASSP 2004. Vol. 1. pp. 665–668.

d'Alessandro, C., Mertens, P., 1995. Automatic pitch contour stylization using a model of tonal perce ption. Computer Speech and Lenguage 9, 257–288.

Eide, E., Aaron, A., Bakis, R., Cohen, P., Donovan, R., Hamza, W., Mathes, T., Picheny, M., Polkosky, M., Smith, M., Viswanathan, M., 2003. Recent improvements to the IBM trainable speech synthesis system. In: Proceedings of ICASSP 2003. Vol. 1. pp. 708–711.

Emerard, F., Montamet, L., Cozannet, A., 1992. Prosodic processing in a text-to-speech synthesis system using a database and learning procedures. In: Bailly, C., Benoit, C., Sawallis, T. (Eds.), Talking Machines: Theories, Models, and Designs. Elsevier Science Publishers, pp. 225–254.

Escudero, D., 2002. Modelado estadístico de entonación con funciones de bézier: Aplicaciones a la conversión texto voz. Ph.D. thesis, Dpto. de Informática, Universidad de Valladolid, España.

Escudero, D., Bonafonte, V. C. A., 2002. Corpus based extraction of quantitative prosodic parameters of stress groups in spanish. In: Proceedings of ICASSP 2002. Vol. 1. pp. 481–484.

Escudero, D., Cardeñoso, V., 2003. Experimental evaluation of the relevance of prosodic features in spanish using machine learning techniques. In: Proceedings of EUROSPEECH-2003. pp. 2309–2312.

Escudero, D., Cardeñoso, V., 2004. A proposal to quantitatively select the right intonation unit in data-driven intonation modeling. In: Proceedings of INTERSPEECH-2004. pp. 745–748.

Escudero, D., Cardeñoso, V., 2005. Optimized selection of intonation dictionaries in corpus based intonation modelling. In: Proceedings of INTERSPEECH-2005. pp. 3261–3264.

Escudero, D., González, C., Cardeñoso, V., Mayo 2002. Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in spanish. In: Proceedings of ICSLP-2002. pp. 1165–1168.

Face, T., 2001. Intonation marking of contrastive focus in madrid spanish. Ph.D. thesis, The Ohio State University, Columbus OH, USA.

Farin, G., 1996. Curves and Surfaces for CAGD, 4th Edition. Cambrige University Press.

Ferrer, A., 2001. Sintesi de la Parla per Concatenació Basada en la Selec-

ció. Ph.D. thesis, Dpto. de Teoría del Senyal i Comunicacions, Universidad Politécnica de Cataluña, España.

Fujisaki, H., Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of Acoustics Society of Japan 5 (4), 233–242.

Garrido, J. M., 1996. Modelling spanish intonation for text-to-speech applications. Ph.D. thesis, Facultat de Lletres, Universitat de Barcelona, España.

Gutierrez, J. M., Montero, J. M., Saiz, D., Pardo, J. M., 2001. New rule-based and data-driven strategy to incorporate Fujisaki's F0 model to a text-to-speech system in Castillian Spanish. In: Proceedings of ICASSP 2001. Vol. 2. pp. 821–824.

Hart, J., Collier, R., Cohen, A., 1990. A Perceptual Study of Intonation. An Experimental Approach to Speech Melody. Cambrige University Press.

Hermes, D. J., February 1994. Measuring the perceptual similarity of pitch contours. Journal of Speech, Language, and Hearing Research 41, 73–82.

Holm, B., 2003. Sfc: Un modèle de superposition de contours multi-paramétriques pour la génération automatique de la prosodie- aprentissage automatique et application à l'Énonciation de formules mathématiques. Ph.D. thesis, Institut National Polythechnique de Grenoble. Grenoble. France.

Jain, A., Murty, M., P.J.Flynn, September 1999. Data clustering: A review. ACM Computing Surveys 31 (3), 264–323.

Kochanski, G., Shih, C., 2003. Prosody modeling with soft templates. Speech Communication 39, 311–352.

Lee, S., Oh, Y.-H., 2001. Tree-based modeling of intonation. Computer Speech and Language 15, 75–98.

Lobanov, B., 1987. The peonemophon test-to-speech system. In: Proceedings

of ICPhs 87. pp. 61–64.

López, E., Rodríguez, J. M., 1996. Statistical methods in data-driven modeling of spanish prosody for text to speech. In: Proceedings of ICSLP-96. pp. 1377–1380.

Navarro-Tomás, T., 1944. Manual de Entonación Española. Madrid, Guadarrama.

Peña, D., 1999. Estadística. Modelos y Métodos. Alianza, Madrid.

Pierrehumbert, J. B., 1980. The phonology and phonetics of English intonation. Ph.D. thesis, MIT.

Plass, M., Stone, M., July 1983. Curve-fitting with piecewise parametric cubics. Computer Graphics , 229–239.

Quilis, A., 1993. Tratado de Fonología y Fonética. Editorial Gredos.

Rodríguez, E., Campillo, F., To be published 2006. A method for combining intonation modelling and speech unit selection. Speech Communication .

Sakai, S., 2005. Additive modeling of English F0 contours for speech synthesis. In: Proceedings of ICASSP 2005. Vol. 1. pp. 277–280.

Sakurai, A., Hirose, K., Minematsu, N., 2003. Data-driven generation of F0 contoures using a superpositional model. Speech Communication 40, 535–549.

Santen, J. P. H. V., Möebius, B., 2000. A qualitative model of F0 generation and alignment. In: Intonation: Analysis, Modelling and Technology. Kluwer Academic Publiser, pp. 269–288.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: A standard for labelling English prosody. In: Proceedings of ICSLP-1992. pp. 867–870.

Sosa, J. M., 1999. La Entonación del Español. Cátedra.

Sproat, R., Olive, J., 1995. An approach to text-to-speech synthesis. In: Speech

Coding and Synthesis. Amsterdam: Elsevier, Ch. 17, pp. 611–633.

Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model. Journal of Acoustical Society of America 107 (3), 1697–1714.

Taylor, P., Black, A., 1995. The Rise/Fall/Connection model of intonation. Speech Communication 15, 169–186.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proceedings of ICASSP 2000. Vol. 3. pp. 1315–1318.

Traber, C., 1992. F0 generation with a database of natural F0 patterns and with a NN. In: Bailly, C., Benoit, C., Sawallis, T. (Eds.), Talking Machines: Theories, Models, and Designs. Elsevier Science Publishers, pp. 287–304.

Vallejo, J. A., 1998. Mejora de la frecuencia fundamental en la conversión de texto a voz. Ph.D. thesis, E.T.S.I de Telecomunicaciones, Universidad Politécnica de Madrid, España.

Veronis, J., Di Cristo, P., Courtois, F., Chaumette, C., 1998. A stochastic model of intonation for text-to-speech synthesis. Speech Communication 26 (4), 233–244.

Webb, A., 2002. Statistical Pattern Recognition, 2nd Edition. Wiley.

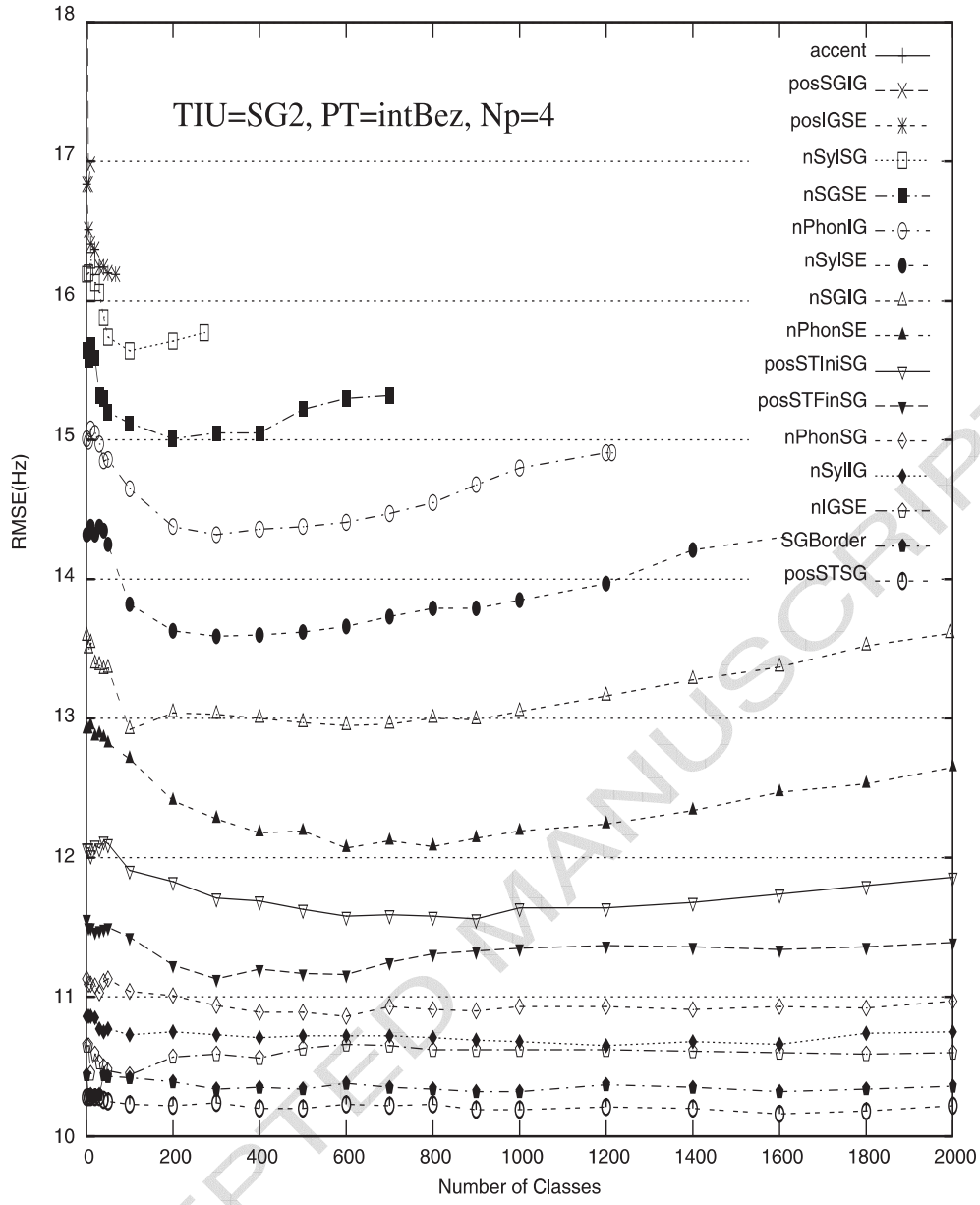


Fig. 7. Creation of the list of dictionaries: each line monitors the construction of the dictionary D_i in the list of dictionaries $\mathcal{D}^{N_F}$. The legend of the line indicates the feature entered to build D_i. Points on the lines are the training errors obtained in the agglomeration of classes. The starting point is at the extreme right of the line, where the number of classes is the maximum. From this initial state, classes are agglomerated and the rest of the points are obtained. Each point is the measurement of the predicting error using the new configuration. The minimum predicting error implies the maximum quality determining the number of classes.

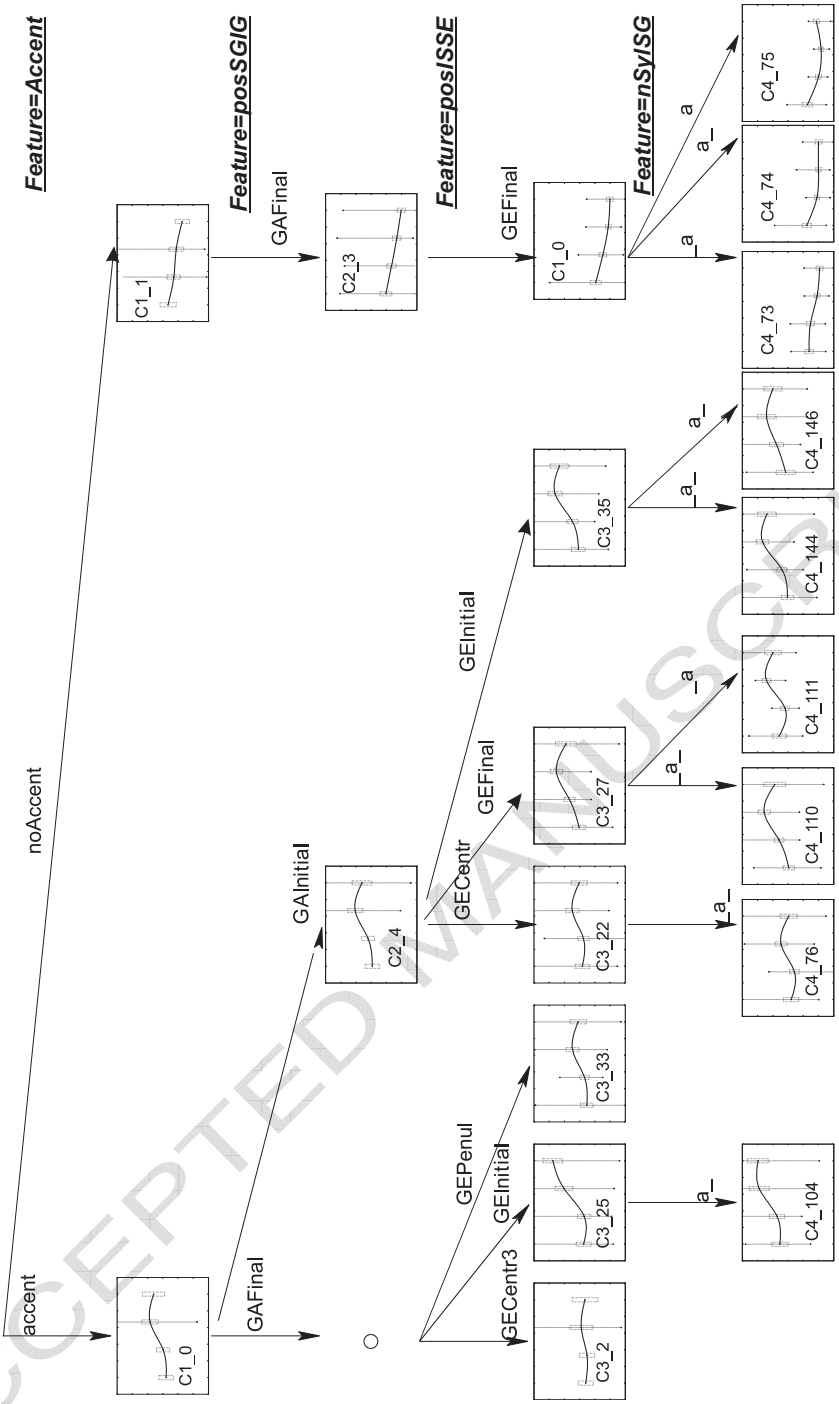


Fig. 8. Models of the dictionary represented as a graph of classes. We have selected a part of the whole tree. X scale is normalized. Y scale is 100-220Hz. We use `TIU=SG3`, `PT=intBez`, and N_P=4. The classes represent the `F0` profile of the centroid and the standard deviation of each control point. The nodes which have a high average prediction error w and are never used for generation, are shown as small circles (like `accent->GAFinal`).

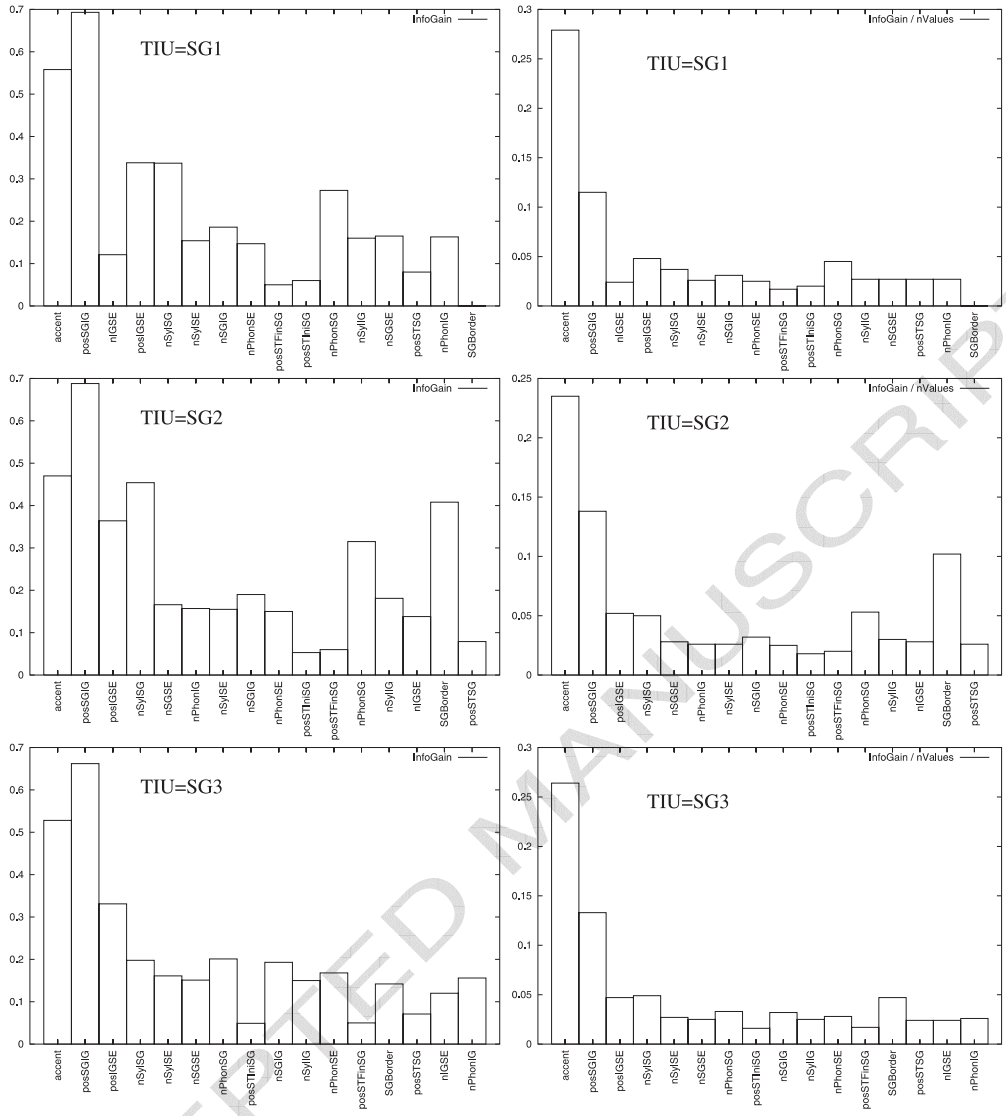


Fig. 9. Relative relevance of the prosodic features. Each row refers to a type of SG according to section 3.2. `Information Gain` is obtained measuring the capabilities of the feature to classify a set of 80 classes obtained applying a *KMeans* clustering with 80 classes. The right column divides `Information Gain` by the number of values of each feature. The features are sorted according to their importance in configuring the list of dictionaries. We use `PT=intBez`, and N_P=4

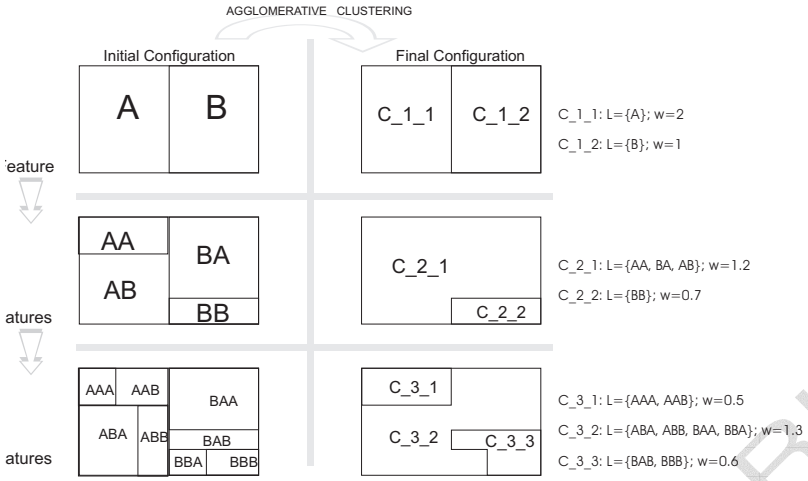


Fig. A.1. Illustration of the process of creation of the list of dictionaries.

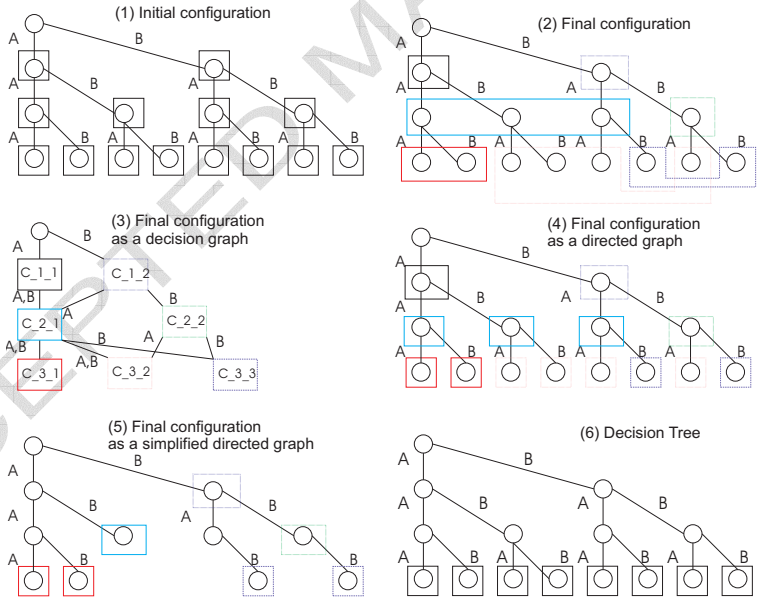


Fig. A.2. Step by step construction of the graph of classes corresponding to the dictionaries displayed in Figure A.1.