Mining Intonation Corpora Using Knowledge Driven Sequential Clustering*

David Escudero-Mancebo and Valentín Cardeñoso-Payo

Department of Computer Science, University of Valladolid, Valladolid 47071 Spain
descuder@infor.uva.es
http://www.infor.uva.es

Abstract. This work presents a mining methodology designed to cope with the usual data scarcity problems of intonation corpora which arises from the high variability of prosodic information. The methodology is an adaptation of a basic agglomerative clustering technique, guided by a set of domain constraints. The peculiarities of the text-to-speech intonation modelling problem are considered in order to fix the initial configuration of the cluster and the criteria to merge classes and stopping their splitting. The scarcity problem poses the need to apply a sequential selection mechanism of prosodic features, in order to obtain the initial set of classes in the cluster. A searching strategy to select the best class among a set of alternatives is proposed, which provides useful prediction models for accurate synthetic intonation. Visualization of final classes by means of a modified decision tree brings graphical cues about contrastable prosodic information of the intonation corpus.

1 Introduction

Intonation is an important attribute of the human speech which brings relevant information about many linguistic, emotional and social aspects. Despite of its importance and of the number of different approaches which can be found in the bibliography, the huge number of factors which have an effect on intonation make its modelling a very difficult challenge (see [3] for a review). The availability of recorded speech corpora opens way to data mining techniques in order to automatically extract information and generate models of intonation. Nevertheless, the nature of the problem makes it difficult to apply conventional techniques. In this work, we introduce a knowledge driven clustering technique which outcomes useful information about some intonation aspects, such as the relevance of the features and its relation with the typical patterns of intonation.

Intonation has been a matter of interest for long time in linguistics (e.g [13] for Spanish intonation). Intonation is related to the different kinds of intonation information: linguistic information (e.g. interrogative vs. declarative sentences); emotional information (e.g. the mood of the speaker) and sociolinguistic information (e.g. social and geographical origin of the speaker). In the speech technology

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

J.S. Sichman et al. (Eds.): IBERAMIA-SBIA 2006, LNAI 4140, pp. 360–369, 2006.

domain, the primary use of intonation has been related to the improvement of naturalness of text-to-speech systems [18]. In speech recognition tasks, intonation information could provide valuable clues to find sentence boundaries or to identify the kind of sentence [16]. Speaker recognition systems have also benefited from the inclusion of intonation information [15]. Nevertheless, the huge number of variability dimensions of intonation phenomena and the difficulty to represent them adequately, justify the lack of consensus on the best way to model intonation information. The relative importance and even the right number of factors which affect intonation is still a subject of debate in the bibliography.

Time evolution of the F0 value of a speech waveform is recognized as a valuable source of information in the intonation literature. Although the first algorithm to extract F0 appeared on the sixties, it is still a subject of improvement (see [8] for a review). Moreover, there is no overall agreement on the way F0 contours should be best parameterized from the extracted F0 magnitude. The goal of Text-to-Speech (TTS) applications is to automatically obtain a mapping between a set of prosodic features affecting the shape of F0 contours and a set of parameters representing the shape of the F0 contour. This mapping could be adequately obtained using data mining techniques to intonation corpora. In the modelling stage, a mapping is inferred from the samples in the corpus. In the prediction stage, this correspondence is applied to get the synthetic contour parameters from the prosodic features derived from the labelled text. A variety of mapping techniques can be found in the bibliography, from simplified basic rule-based systems[2] to corpus based systems approaches using Neural Networks[12], Decision Trees [19], Regression Trees [1] ....

Two main limitations affect traditional learning techniques. First, they do not provide contrastable linguistic information about the intonation movements. Second, they usually lack enough robustness to cope with data scarcity problems which, as a consequence of the high number of possible combinations of potentially important prosodic features, heavily affect feature covering capabilities of the corpora (using D prosodic factors with an average number of V values each, would require unrealistic corpus sizes for typical situations in which $D > 10$ and $V > 5$, leading to more than 10^5 different units). The scarcity problem could cause unrealistic prediction of F0 contours when the input is labelled with a combination of prosodic features rarely observed or not present at all in the corpus. This can dramatically decrease the naturalness of the synthesized speech.

In this work, we will describe a knowledge driven sequential clustering which brings enough robustness to cope with data scarcity problem and provides the core component of an intonation modelling methodology which can be successfully used in TTS applications with a high degree of speech naturalness. In section 2, we formally describe the intonation modeling problem, focusing on the goals, domain constraints and limitations which inspire the decisions to be taken for the clustering procedure, which will be described in section 3. Results and conclusions are reported in sections 4 and 6.

2 Problem Statement

In corpus based intonation modelling, the corpus is considered a set of intonation units IU_i, $C = \{IU_i, \quad 1..N\}$. Every IU_i is a pair $IU_i = (IU_i.PF, IU_i.AP)$. $IU_i.AP$ is an array of numerical values which provide the acoustic parameterization of the F0 contour of IU_i. $IU_i.PF$ is a set of numbers or strings which gather the values of the set of prosodic features which label the prosodic function of IU_i. These features capture several characteristics of intonation like accent, emotion, type of sentence or grammatical structure of the sentence, among others. The prosodic features are either manually labelled in the corpus units or, in a generation stage, are derived from text using a priori linguistic knowledge.

As far as intonation is concerned, the main goal in corpus-based TTS applications is learning the correspondence between AP and PF, given a set C of labelled intonation units. Any model providing such correspondence should adequately predict $IU_i.AP'$ given $IU_i.PF'$ in the generation stage. This set $IU_i.AP'$ of acoustic parameters could then be used to generate a synthetic F0 contour close enough to intonation contours associated to the $IU_i.AP$ of the corpus which are determined to be similar to the $IU_i.AP'$.

Any procedure designed to solve the correspondence learning problem should take into account two main goals and fulfil two main constraints:

Goal 1 (G1) Prediction of synthetic F0 contours is to be as accurate as possible for Text-To-Speech applications.

Goal 2 (G2) Results of the modelling stage should provide linguistically contrastable information.

Domain Constraint 1 (C1) Two different intonation units IU_i and IU_j, are perceptually equivalent iff $IU_i.AP \sim IU_j.AP$, provided the parameterization technique has been properly selected[9].

Domain Constraint 2 (C2) The second constraint concerns to the function of intonation and it establishes that if $IU_i.PF = IU_j.PF$ then $IU_i.AP \sim IU_j.AP$, given the parameterization technique and the prosodic features have been properly selected.

Since the solution has to assume that prosodic knowledge is not complete and that it is usually impossible to gather together in a corpus a set of instances broad enough to cover all the possible prosodic configurations, two fundamental drawbacks have to be taken into account:

Drawback 1 (D1) The ideal set of PF and AP is still an open question, so that it would be desirable to obtain, as a product, information about the number of features to use, their cardinality, and their relative importance.

Drawback 2 (D2) The high number of PF involved makes data scarcity problems a fundamental difficulty to tackle with.

Together, all these goals, constraints and drawbacks drive the design of a new clustering process, adapted from basic clustering techniques [11], which will be described in the next section. As we will show, this technique will also provide visual cues which should be useful to interpret the nature of prosodic phenomena.

3 Clustering Technique

In this section, we describe our proposal for a multilevel clustering technique, driven by a forward sequential feature selection process, which correctly solves the problem described previously. The technique is inspired by classic knowledge-based agglomerative clustering techniques [11] in combination with widely accepted feature selection techniques[20].

Constraint C1 justifies the use of clustering techniques to build sets of classes of IU_i, grouped in terms of the similarity of their APs, since these sets of classes will properly represent the typical movements of F0 contours observed in the corpus. This clustering will also provide a means to find the characteristic intonation profiles, abstracting the intrinsically high variability of the prosodic events.

The process starts building an initial classification of the IU_i from a single prosodic feature $LPF_1 = \{PF^1\}$. Each class corresponds to a given value of this initial prosodic feature PF^1. An agglomerative clustering technique is iteratively applied to this cluster using maximum similarity as the merging criterion and prediction accuracy of 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 PF^1. An additional prosodic feature is added to LPF_1 to construct the next set of prosodic features $LPF_2 = \{PF^1, PF^2\}$ and a new cluster is build repeating the previously described process. Again, the same criterion applies for the selection of PF^2, resembling the typical forward sequential feature selection process. The clustering process stops when all the possible prosodic features have been included into $LPF_{fin} = \{PF^i | i = 1, \ldots, N_{pf}\}$ and it results into a multilevel set of clusters, each one corresponding to an increasingly more specific set of prosodic features.

Constraint C2 implies that two different intonation units sharing the same set of prosodic features are to be in the same class, since they should be similar in a way consistent with the similarity measure used to merge classes. That justifies why we choose the set of classes induced by LPF as the initial set for any cluster level.

Since the main application of this clustering will be TTS generation (see goal G1), it is clear that the stopping condition for the agglomerative process should be related to the prediction accuracy of the clusters when used to generate F0 profiles: agglomeration should stop when the prediction results using a set of clusters after merging are worse than using the present ones.

The agglomeration still provides a correspondence between APs and PFs, if we keep track of the different values of the PF associated to a class after merging. The list $PL_j = \{PF_k | k = 1, \ldots, K_{max}(j)\}$ associated to a class C_j provides an index to it which can be used in TTS to retrieve the APs which correspond to the given sequence of PF annotated in the input text. The APs retrieved sequence will be used to generate F0 contour (G1). We call *dictionary* the set of pairs $D_k = \{(PL_j, C_j), j = 1, \ldots, N(k)\}$. A dictionary is the explicit representation of the correspondence between function of intonation ($PF \in PL_j$) and its shape ($AP \in C_j$ in the class) and bring a way to fulfil goal G2.

As the number of prosodic features increases, the data scarcity problem gets worse (drawback D2). The multilevel clustering technique provides a different clustering for every $LPF_0, LPF_1, \ldots, LPF_{fin}$ and each of them has been optimally adapted to cover the IU_i set in the corpus for a given level of detail in the set of prosodic features. Since the lists LPF_j are orderly enlarged adding the next best predicting feature at each stage, we can use the corresponding ordered set of dictionaries $LD_k = \{D_j | j = 1, \ldots, k\}$ to guide a searching strategy for alternatives to unseen (or infrequent) PF_j combinations, selecting the best predicting dictionary which subsumes PF_j (refer to [4] [7] for details)

The ordered list of dictionaries provides a way to build a decision tree which gives visual information easy to contrast that schematically represents the intonation patterns found in the corpus (see section 5). The ordered set of lists $LPF_0, LPF_1, \ldots, LPF_{fin}$ provides a ranking of importance of the various prosodic features, which besides the previous visualization capabilities provides good fulfilment of goal G2 and adequately copes with drawback D1.

4 Experimental Results

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. Sentences has been segmented and labelled following a semiautomatic process. We selected only the declarative sentences, which represent about 95% of the whole corpus. The sentences has been segmented into different types of intonation units: intonation groups (IG), stress groups (SG) and syllables (see [13] 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 Bzier curves of degree 3 fitting the F0 contours in the intonation units (more details in [6]). 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 metric to guide the merging process. The prediction

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

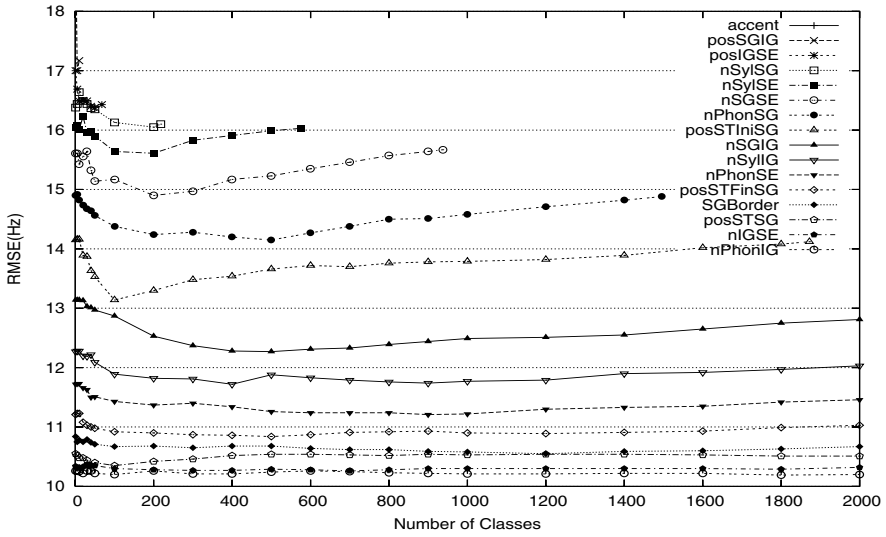

Fig. 1. Building the list of dictionaries: each curve represents the effect of adding dictionary D_i to the list of dictionaries LD_i, $(i = 1, \ldots, N_{pf})$. The name of the PF added to build D_i is the legend of the curve. Each curve represents the prediction error of the training samples as a function 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.

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 1 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. As the number of PF grows the impact of new PF decreases mainly due to the fact that some of the PF are redundant or they introduce few extra information (e.g. $nPhonSG$ and $nSylSG$ are correlated features: when one of them is considered, the incremental procedure rejects the other).

Prediction errors showed in table 1 indicate that the TTS results obtained using our clustering technique are comparable with other approaches found in the bibliography (see [14] for a ranking). Informal listening tests have been done to assess the goodness of the synthetic intonation. The over-training effect observed in the table could be acceptable in TTS applications where imitating the intonation patterns in the corpus does not incur any noticeable loss of naturalness.

Table 2 shows the success of using an multilevel approach. Less specific dictionaries (the ones with less number of PF) are used frequently to predict any IU, both for the testing and training corpora. The difference is more obvious for the testing set, since the number of unseen PF combinations increases for more specific dictionaries.

Table 1. Prediction Errors: RMSE y Pearson Correlation (Corr) versus the number of prosodic features in training and in testing stage. The metrics are computed using all the F0 contours of the training and testing corpus respectively.

	Train		**Test**	
List of Dict.	**RMSE(Hz)**	**Corr**	**RMSE(Hz)**	**Corr**
LD1	21.47	0.59	21.53	0.60
LD2	19.20	0.69	19.66	0.68
LD3	18.55	0.71	18.58	0.72
LD4	18.30	0.72	18.49	0.72
LD5	17.94	0.74	18.49	0.72
LD6	17.23	0.76	18.66	0.72
LD7	16.51	0.78	18.94	0.71

Table 2. Level of use of the dictionaries in a list: each cell contains the percentage of intonation units that are predicted using each of the dictionaries of the list

	Use of the Dictionary (%)						
LD7	D1	D2	D3	D4	D5	D6	D7
Train	0.0	2.8	3.4	14.8	6.1	16.1	56.8
Test	0.2	4.5	6.4	17.4	10.7	18.7	42.2

Table 3 shows the impact of the agglomerative process in the final number of representative classes: $D7$ has 1795 classes in the initial configuration and 190 at the end of the agglomeration process. Note that the initial number of classes is far away from the maximum: if the corpus would have samples to cover all the possible combinations of PF, the number of classes would be `2 (accent)` $\times$ `6 (posSGIG)` $\times$ `7 posIGSE` $\times$ `9 (nSylSG)` $\times$ `6 (nSylSE)`$\times$ `6 (nSGSE)` $\times$ `6 (nPhonSG) = 163296`. Although some of the combinations are impossible, we can easily figure out the magnitude of the corpus needed to cover them all and the importance of designing a robust strategy to cope with scarcity, as the one proposed in this work. Finally, not all the classes in the final configuration are used since this is decided through the dictionary based selection mechanism for every intonation unit: we see that only 24 out of the 190 available classes are used in $D7$. This significant reduction helps simplifying the visual representation of the clusters which will be presented in the next section.

Table 3. Description of the dictionaries in terms of the number of classes and of the 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

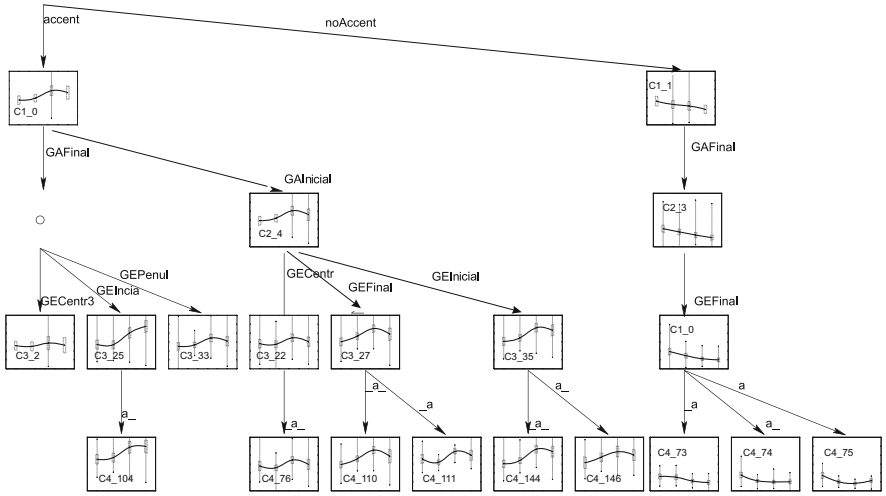

Fig. 2. Models of the dictionary represented as a decision tree. We have selected a part of the whole tree. Normalized x axis, y axis scale: 100-220Hz.

5 Visualization Cues

Figure 2 shows a tree-like graphical representation of the classes in the list of dictionaries. Each node represents a class in the clusters. For every class, we show the Bzier 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 visualization of the information in the tree allows us to contrast some of the assessments found in the bibliography about Spanish Intonation. In [5], 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 2.

– **The importance of the prominence.** We have labelled this function with $PF =$ `accent`. Figure 2 shows that this feature is essential: it is in the top of graph separating two sets of classes clearly different. Patterns with `accent` property, are characterized by high F0 values and by a rising pattern.

- **Position of the stressed syllable:** Prototypical patterns associated to the Spanish stress groups are $L*+H$ and the less frequent $H*$ (using TOBI notation). To analyze this fact, `nSilGA` has 4 possible values: `_a_, _a, a_, a`), where `_` means un-stressed syllable and `a` means stressed one. In this context, F0 contour evolution can be easily aligned with respect to the stressed syllable. The majority of the classes fit with the `L*+H` symbol and some of them with $H*$ according to the observations of [17].
- **Influence of the juncture:** Patterns in the *IU* boundaries have a decreasing trend (node C3_21) of the tree 2), *anticadence* (node C3_25) and *semicadence* (node C3_33) (see [13]).
- **Type of sentence:** affecting mainly the last part of the F0 contour. Typical final juncture of declarative sentences $L*+L\%$ is clearly seen in figure 2.

Finally, we remark that the visualization of figure 2 will probably let the experts to conclude about the intonation phenomena, although a thorough discussion of this is out of the scope of the present paper.

6 Conclusions and Future Work

The peculiarities of the intonation modelling problem have inspired the definition of an ad-hoc clustering methodology. The methodology provides synthetic F0 contours of a comparable quality to the ones found in the state of the art.

The methodology does not depend on the selection of the prosodic features, acoustic parameters, and type of intonation unit. This could be exploited to experiment the effect of those prosodic factors on quality intonation modelling.

Extracting contrastable information from the corpus of study was also a goal, in a field where many conceptual questions are still open. The proposed clustering procedure provides a ranking of importance of the prosodic features typically used to classify intonation patterns.

Furthermore, the tree-like representation of the result classes provides visual cues which aid contrast relationships between prosodic features and typical F0 contours patterns, as found in the working corpus. This information could be most valuable as an objective means to test the intonation of a given corpus against others or to validate the linguistically correct intonation of a given corpus, with respect to a set of recognizable theoretical prosodic assessments. The results presented for Spanish show good agreement with accepted prosodic knowledge for this lenguage.

References

1. P.D. Aguado, K. Wimmer, and A. Bonafonte. Joint extraction and prediction of fujisaki's intonation model parameters. In *Proceedings of Eurospeech 2005*, 2005.
2. J. Allen, M. S. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System.* Cambridge University Press, 1987.
3. A. Botinis, B. Granstrom, and B. Moebius. Developments and Paradigms in Intonation Research. *Speech Communications*, 33:263–296, July 2001.

4. V. Cardeoso and D. Escudero. A strategy to solve data scarcity problems in corpus based intonation modelling. In *Proceedings of ICASSP 2004*, 2004.
5. D. Escudero. *Modelado Estadstico de Entonacin con Funciones de Bzier: Aplicaciones a la Conversin Texto Voz.* PhD thesis, Dpto. de Informtica, Universidad de Valladolid, Espaa, 2002.
6. D. Escudero and V. Cardeoso A. Bonafonte. Corpus based extraction of quantitative prosodic parameters of stress groups in spanish. In *Proceedings of ICASSP 2002*, Mayo 2002.
7. D. Escudero and V. Cardeoso. Optimized selection of intonation dictionaries in corpus based intonation modelling. In *Proceedings of Eurospeech*, September 2005.
8. D. Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical Report TR-CS 2003-06, Department of Computer Science, University of Regina, Regina, Saskatchewan, CANADA, November 2003.
9. J. Hart, R. Collier, and A. Cohen. *A perceptual study of intonation. An experimental approach to speech melody.* Cambrige University Press, 1990.
10. D. J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41:73–82, February 1994.
11. A.K. Jain, M.N. Murty, and P.J.Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, September 1999.
12. O. Joskisch, H. Mixdorff, H. Kruschke, and U. Kordon. Learning the parameters of quantitative prosody models. In *Proceedings of ICSLP 2000*, 2000.
13. T. Navarro-Toms. *Manual de Entonacin Espaola.* Madrid, Guadarrama, 1944.
14. S. Sakai. Additive modeling of english f0 contours for speech synthesis. In *Proceedings of ICASSP 2005*, 2005.
15. E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling Prosodic Feature Sequences for Speaker Recognition. *Speech Communication*, 46 (3-4):455–472, 2005.
16. E. Shriberg, A. Stolcke, D. Hakkani, and G. Tur. Prosody-Based Automatic Segmentation into Sentences and Topics. *Speech Communication*, 32(1-2):127–154, 2000.
17. J. M. Sosa. *La Entonacin del Espaol.* Ctedra, 1999.
18. R. Sproat. *Multilingual Text-to-Speech Synthesis.* Kluwert, 1998.
19. P. Taylor. Analysis and Synthesis of Intonation using the Tilt Model. *Journal of Acoustical Society of America*, 107(3):1697–1714, 2000.
20. A. Webb. *Statistical Pattern Recognition.* Wiley, 2nd edition, 2002.