Statistical Modelling of Stress Groups in Spanish

V. Cardeñoso & D. Escudero

Departmento de Informática
Universidad de Valladolid, Spain
{valen; descuder}@infor.uva.es

Abstract

A statistical corpus-based synthesis strategy has been developed for fundamental frequency contours ($F0$) of Spanish sentences. Input text is assumed to be made of a sequence of intonation groups, each one containing one or more stress groups. The stress group is taken as the basic prosodic unit at acoustic level. For every kind of acoustic unit, we get a set of statistical distributions for the parameters of a Bézier function that generates the $F0$ contour of the unit. This distributions are obtained directly from the corpus and two different models are annalyzed. In one of them, linguistic knowledge forces the grouping of stress groups, while in the other an unsupervised clustering is carried out. Comparison of the results of both methods show the relative importance of different prosodic features found in real data.

1. Introduction

The field of prosodic modelling in speech synthesis is being highly influenced by the development of new corpus-based approaches which lead to better generation of prosodic features[8, 9, 6]. The neccesity of large-scale training corpus is reported as the main drawback of these methods since it is usually the case that speech corpora do not include enough prosodic information. However, new generation TTS are to be designed that try to overcome some of the general drawbacks of present systems: automatic speaker adaptation, multilingual TTS, speaker emotion modelling, better naturalness. Corpus-based methods bring new means to learn from real examples how speakers utter their messages in different situations even when a linguistic theory is not yet at hand, if one is needed.

Prosodic modelling could be formulated as the problem of mapping discourse information into acoustic level prosodic information. To solve this complex problem, three fundamental issues have to be addressed. Modelling discourse information should be faced first, forms of prosodic representation at acoustic level should be found and, finally, the number of steps to get the overall mapping into pieces should be decided.

In this work, we have experimented with a statistical modelling of prosodic units extracted from a corpus in order to get a reasonably good and simple model of intonation patterns at acoustic level. First, we have attempted to get statiscical models corresponding to 'a priori' prosodic clases stated from linguistic knowledge. Second, we have carried out an unsupervised clustering of corpus data in terms of their acoustic prosodic features in order to test if the 'a priori' classification is found in real speech and, also, to get better knowledge of the set of linguistic features that are useful to properly label speech segments for prosodic purposes. As a side result, we lay the basis to quantitatively measure speaking characteristics of a language or a speaker and, thus, open new possibilities to establish prosodic corpus standarization.

The main difference between our approach and the ones found in the literature is that the model unit associated to a given input segment at synthesis time is given by a statistical distribution instead of a given corpus unit, which could correspond with one the units found in the training corpus or with some average of them. With this approach, we try to reproduce the speaking style of the corpus donor but at an ergodic level. This means that we are not interested in mapping a given segment of input text to a given acoustic prosodic profile, but with a profile that is randomly generated acording to a given probability distribution function (pdf). Two separate realizations of a given text segment could be mapped to different acoustic profiles. This is not far from reality and gives an interesting alternative when synthesizing long text fragments, for which monotony is a typical drawback.

As a second difference, we use Bézier functions to parameterize the intonation units instead of simple stylization patterns. The main advantage of our parameterization technique is that it can be also applied to intonation units longer than a syllable without any loss of information. For our purposes, we have used stress groups (SG) as intonation units, defined as the sequence of words starting after a given stressed word and ending after the next stressed word in the phrase.

The following sections are organised as follows. In section 2, we briefly describe all the modelling aspects. In section 3 we present results for the two different versions of statistical treatment carried out. Section 4 summarizes the main contributions of the paper.

2. Modelling issues

There are many linguistic features that could be taken into account and have an influence on the final prosodic profile of acoustic generated data[4]. In this first study, we have considered the kind of stress of SGs, the position within a given intonation group (IG) and the kind of IG as relevant prosodic features. Although it is a simplification, it will bring enough information to test our treatment.

A starting hypothesis of our proposal is that every kind of stress group will have an associated class of intonation pattern and patterns of intonation corresponding to a given class should bear similar shapes. This can be tested two different sides. First, an assumption can be made about the reasonable set of kind of SGs present in Spanish and a measure of the grouping confidence of corpus samples can be carried out. Second, any asumption about the kind of SGs could be removed beforehand and an automatic classification could be done. Later, the known

This work has been partially supported by Junta de Castilla y León under research contract n VA-16/00A.

prosodic labels of the samples inside a cluster could give information about the viability of our variety of theoretical SG kinds.

To generate a suitable F0 realization in the synthesis step, labels generated from input text by a *prosodic labeling module* are taken as indexes to the *pitch generation* module which gets the pdfs corresponding to them and generates specific profiles. Synthesis results using this method have been succesfully tested[1].

2.1. Segmentation

From a corpus which includes phonetic and orthographic labels, pitch contour values and syllable and word boundaries, the *prosodic segmentation module* locates stress groups and labels them using 'a priori' linguistic knowledge for Spanish.

Labelling includes: the kind of intonation group in which the stress group is located, the number of SGs in the same IG, the relative position inside the intonation group, the position of the stressed syllable, the number of syllables of the SG and the total number of syllables of the IG. Inspired by the suggestions found in [6], we will only consider three different kinds of stressed syllable positions (ultimate, penultimate and antepenultimate) and three different values for the SG position within its IG (initial, inner and final).

Every intonation group is delimited by two consecutive pauses or by a significative jump in F0 value. In our case, we experimentally found that a F0 step of 30 Hz gave good segmentation results.

2.2. Modelling individual stress groups

Once the stress groups have been segmented out from the samples, we carry out a parameterization of every SG in terms of a Bézier function. We have chosen this functions instead of the typical stylizations based on straight line segments or quadratic functions[5][9] because they bring several benefits:

- A fixed and reduced set of parameters is needed for each SG (4 in fact).
- When needed, better fitting quality can be easily obtained if the degree of the polynomials for each parametric component of the curve is increased.
- It is easy to automatically fit a cloud of points to a series of Bézier functions.
- The shape of the functions can be controlled so that they obey specific theoretical patterns.
- The values of the parameters of a given SG have a rather intuitive meaning from the point of view of linguistic interpretation of the SG profile and, most important, altering any individual parameter results in a global change of the SG intonation profile.
- Duration of any SG can be separately modelled, since all the fitting can be carried out in a normalized duration axis and then rescaled to the extent given by any model.

In this work, we have assumed that duration of SGs can be modelled separately of the F0 values. Thus, every SG of the corpus will be represented by a vector of parameters $\{P_i\}, i = 0, ..., 3$ which give control points (in Hz) located at fixed time points (0, $D/3$, $2D/3$, D). The control points and the duration, D, determine the final shape of the Bézier function. Adjusting the generated SGs to a final time scale prescribed by the duration model is straight away, since Bézier functions are controlled by a scalar parameter t which can be resized to the given interval.

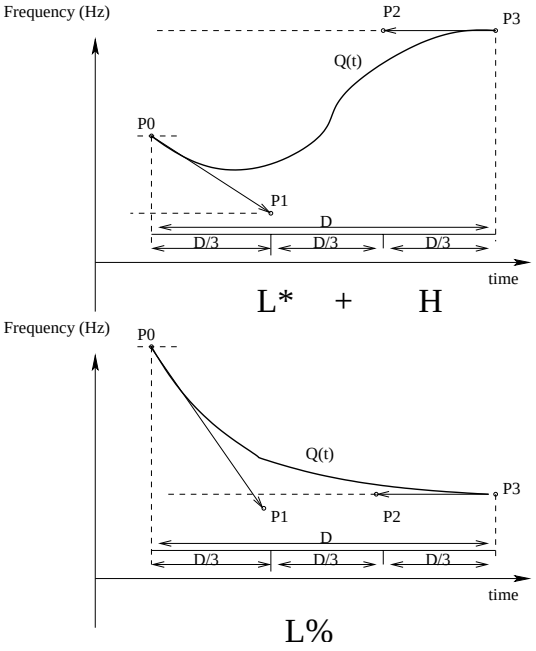

Figure 1: Bézier fitting of two typical intonation patterns.

Using polynomials of higher degree, more than one local maximum could be obtained if needed. A set of P_i parameters controls $F0(t)$ inside the given SG, filtering out any micro-intonation details but giving account of the individual syllable contributions to the overall SG profile.

Figure 1 illustrates how Bézier functions model two typical intonation patterns, corresponding with initial (L*+H) or final (L%) stress groups in an intonation group. Both curves have degree three and their shape is controlled by points $P0$, $P1$, $P2$ y $P3$ located as labeled in the figure.

2.3. Bézier functions

Every SG is characterized by $SG \equiv \{t_{ini}, t_{fin}, \mathbf{P}\}$, where t_{ini} y t_{fin} are the time boundaries of SG, and $\mathbf{P} \equiv \{\bar{p}_i = (T_i, F0_i) | \quad i = 0, ..., p\}$ are the $p+1$ points time-frequency of the pitch contour in SG stored in the corpus.

A Bézier curve $\bar{Q}(t)$ is a parametric polynomial of degree n defined by $n+1$ control points $\bar{P}_i$ with $i = 0, ..., n$ as follows:

$$\bar{Q}(t) = \sum_{i=0}^{n} \bar{P}_i B_i^n(t) \quad t \in [0, 1] \tag{1}$$

where $B_i^n(t) = \binom{n}{i} t^i (1-t)^{n-i}$ are the *Bernstein polynomials* [2].

A Bézier function in R^2 is a special case where:

$$\bar{Q}(t) = \left(a + t(b-a), \sum_{i=0}^{n} P_i B_i^n(t) \right) \quad t \in [0, 1] \tag{2}$$

Control points are now $\{(i/n, P_i), \quad i = 0, ..., n\}$. The interval $[a, b]$ is the domain of the function.

To represent each SG with a function $\bar{Q}(t)$, the parameters P_i with $i = 0, ..., n$ are set by optimizing the fitting of $\bar{Q}(t)$ to the corresponding $\mathbf{P}$. This is done by squared error minimization, which implies solving for P_i the following system of equations[7]:

Code	IG kind	IG	SG	ISG	CSG	FSG
IG1	**Final Declarative**	668	2257	571	1018	668
IG2	Rise Non Final Decl.	400	864	240	224	400
IG3	Fall Non Final Decl.	490	1285	346	449	490
IG4	Questions	47	187	40	100	47
IG5	Exclamative	8	29	8	13	8
IG6	Parenthetical	2	3	1	0	2

Table 1: Corpus sizes. IG: Intonation Group. ISG: Initial Stress Group. CSG: Inner Stress Group. FSG Final Stress Group. SG Stress Group.

$$\frac{\partial}{\partial P_l}\left(\sum_{j=0}^{p}\left(F0_j - \sum_{i=0}^{n} P_i B_i^n(t_j)\right)^2\right) = 0 \quad l = 0, ..., n \tag{3}$$

where $t_j = (T_j - T_0)/(T_p - T_0)$.

Choosing Bézier functions instead of Bézier curves implies fixed t positions for the control points. We have tested that the relaxation of this simplification does not generate sensibly different results, while the computation times are highly increased.

3. Experimental Results

3.1. Corpus description

For the experiments described in this paper, we had access to a corpus designed and adquired at UPC university [3]. It contains read out spoken utterances both in Spanish and Catalan recorded by a professional female speaker under studio conditions. Recordings were made at 32 kHz and a separate channel was recorded with data from a laryngograph, to avoid indeterminacies from pitch estimation algorithms. All the sentences are manually labelled and revised with phonetic and prosodic information including stress information and some ortographic marks. Phrase, word and syllable boundaries are included.

Although it was not specifically designed for statistical modelling, it contains enough data to get significant results. Table 1 describes the number of ocurrences of this corpus. It consists of 1615 intonation groups of six different categories that amount to 4625 stress groups. For the results presented in this paper, we have just considered the final declarative intonation groups because they include the highest amount of data.

3.2. Quality indicators

In order to test the intrinsic quality of the models, we propose several statistical indicators, labelled as $M_i, i = 1, ..., 5$:

- M_1 Radius (Hz) of the class, given by the mean distance to the centroid.
- M_2 Mean value of the absolute values of the elements of the covariance matrix inside the class.
- M_3 Distance (Hz) to the centroid of the nearest class.
- M_4 The index of the nearest class.
- M_5 Mean value of the distance between two different samples of the class.
- M_6 Percentage of samples of a class that result to be closer to its centroid than to the one of any other class.

3.3. Fixed labeling models

As already pointed out, in this model we use nine different classes for the SGs, which correspond to the possible combinations for the pair (kind of stress, position of stress group). The

		Quality measures					
Class	N^o	M_1	M_2	M_3	M_4	M_5	M_6
F1	186	24	0.38	9	3	17	39
F2	368	24	0.39	8	5	17	10
F3	21	30	0.52	12	1	22	43
F4	331	22	0.21	5	4	16	22
F5	664	23	0.39	4	5	17	7
F6	38	26	0.54	4	4	19	34
F7	122	17	0.61	3	7	13	43
F8	503	15	0.42	3	8	11	11
F9	50	16	0.55	3	7	12	50

a) Quality indicators for every class

	P_0		P_1		P_2		P_3	
Class	μ	σ	μ	σ	μ	σ	μ	σ
F1	156	16	150	30	164	29	168	20
F2	161	14	141	26	185	33	170	23
F3	163	23	161	35	207	37	172	24
F4	166	17	137	26	162	27	157	19
F5	165	18	129	27	171	29	158	20
F6	163	18	126	31	181	32	158	23
F7	149	17	123	24	130	22	117	7
F8	151	15	129	20	126	20	119	6
F9	149	14	133	24	121	21	120	5

b) pdf = $N(\mu, \sigma)$ of SG control points.

Table 2: Results with fixed prosodic labeling of SGs.

stress groups of the corpus corresponding to each class have processed and parameterized following the method described in section 2 and an empirical probability distribution function for each different P_i has been obtained. This pdf's, which can be confidently be considered as normal distributions, are specified by a mean and a standard deviation, which is given in part b) of table 2. Part a) of the table gives the results for the quality indicators for this experiment.

From the results in table 2, we can conclude that better results are obtained for the classes corresponding to stress groups in a final position within their IGs. As for the relative values of P_0 and P_3, they reflect the expectations predicted by the linguistic theories (taking into account that they give the real values of $F0(0)$ and $F0(D)$ respectively).

3.4. Clustering models

In this second model, we try to test if the 9 classes used for the fixed labelling scheme do correspond well with experimental data of the corpus. Although this could be tested by means of statistical properties of the sample grouping characteristics, we have preferred to design a separate test because it could also shed light on the relevant prosodic properties that should be taken into account.

To this end, we have carried an unsupervised clustering of all the stress groups present in the corpus using a standard k-means procedure. Although we have experimented with different number of classes, we present here the results for nine classes, since they should correspond to the data obtained for the previous model.

Results presented in table 3 show that the quality of this classification is better than the previous one, although the values of indicator M_2 are smaller. M_6 obviously amounts to 100%, thus showing the correctness of the clustering procedure. The most striking result is that there is no clear correspondence between the classes automatically obtained and the ones used for fixed labelling.

A correspondence is still found between the group of classes representing final SGs in both models and a general

		Quality measures					
Class	N^o	M_1	M_2	M_3	M_4	M_5	M_6
C1	392	10	0.10	17	2	7	100
C2	298	11	0.01	17	1	8	100
C3	207	14	0.07	17	1	10	100
C4	304	14	0.11	21	2	10	100
C5	306	12	0.07	16	6	9	100
C6	248	14	0.18	16	5	10	100
C7	191	18	0.09	22	6	13	100
C8	199	18	0.08	24	5	13	100
C9	138	20	0.15	27	4	14	100

a) Quality indicators for every class

	P_0		P_1		P_2		P_3	
Class	μ	σ	μ	σ	μ	σ	μ	σ
C1	143	11	129	10	121	11	122	7
C2	160	12	109	11	152	11	120	9
C3	145	14	168	14	104	14	134	13
C4	169	14	102	13	190	13	147	16
C5	162	14	149	12	158	13	147	9
C6	168	16	130	14	157	16	186	12
C7	168	17	153	19	215	18	183	16
C8	151	16	190	18	141	20	173	17
C9	183	17	84	21	242	23	172	17

b) pdf = $N(\mu, \sigma)$ of SG control points.

Table 3: Results for automatic clustering of SGs.

agreement is obtained for the grouping of the nearest class in both models. The consequence could be that there's still more prosodic information than just the kind of accent and the position to be taken into account in order to find any definitive correspondence. Most important is the fact, extracted from the corpus data, that the kind of accent of every SG does not bring by itself any relevant information about the class of the SG.

4. Conclusions

We have proposed a corpus-based statistical modelling of prosodic units taking the stress group as the basic unit. Sets of stress groups sharing common linguistic attributes are gathered in classes and each class is assigned a set of four normal distribution functions that can be used to generate $F0(t)$ contours of segmented input text. Each pdf correspond to a control point of the Bézier function representing its acoustic prosodic profile.

An automatic classification procedure uncovers the fact that kind of accent of stress groups is no so relevant as expected. The behaviour of final stress groups inside a given IG is clearly distinguishable of the rest.

Although complementary experimentation is still neccesary, the results presented here show promising possibilities to include quantitative measures to assert model quality and corpus estandarization.

5. Acknowledgements

We gratefully acknowledge fruitful discussions with researchers of the TALP group of UPC university. Special thanks to A. Bonafonte for his contributions and his efforts to make the corpus available to us.

	Initial (ISG)			Central (CSG)			Final (FSG)		
	AC1	AC2	AC3	AC1	AC2	AC3	AC1	AC2	AC3
	F1	F2	F3	F4	F5	F6	F7	F8	F9
C1	6	5	0	26	33	2	53	245	22
C2	7	15	1	26	52	1	40	142	14
C3	6	12	0	20	38	2	21	96	12
C4	15	55	0	54	150	9	6	14	1
C5	32	52	2	84	124	5	1	5	1
C6	32	53	1	52	105	5	0	0	0
C7	29	82	7	25	44	4	0	0	0
C8	48	56	5	30	56	4	0	0	0
C9	11	38	5	14	62	6	1	1	0

Table 4: Number of samples of class C_i with labels corresponding to different SG positions and kind of accent (AC1:ultimate, AC2:penultimate, AC3:antepenultimate). Every SG label corresponds to a given fixed class F_i.

6. References

[1] Escudero, D.; Cardeñoso, V., 2001. Modelo cuantitativo de entonación del español. *Procesamiento del Lenguaje Natural*. V. 27, 233-240.

[2] Farin, G., 1996. *Curves and Surfaces for CAGD*. Cambrigde University Press.

[3] Febrer, A., 2001. Sintesi de la Parla per Concatenació Basada en la Selecció *Phd Thesis, Dpto. Teoría del Senyal i Comunicacions, Universidat Politécnica de Catalunya, Spain*.

[4] Garrido, J.M., 1996. Modelling Spanish Intonation for Text-to-Speech Applications *Phd Thesis, Facultat de Lletres, Universitat de Barcelona, Spain*.

[5] Hirst, D.J.; Ide, N.; Veronis J., 1994. Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project. *Proceeding of 2nd ESCA/IEEE Workshop on Intonation*. 77-81.

[6] López, E.; Rodriguez, J. M., 1996. Statistical Methods in Data-Driven Modeling of Spanish Prosody for Text to Speech In *Proceedings of ICSLP 96.* .

[6] López, E.; Rodriguez, J. M.; Hernández, L.; Villar, J. M., 1997. Automatic Prosodic Modeling for Speaker and Task Adaptation in Text-to-Speech. In *Proceedings of ICASSP 97.* . 927-931.

[7] Plass, M.; Maureen S., 1983. *Curve-Fitting with Piecewise Parametric Cubics*. Computer Graphics, July 229-239.

[8] Saito T.; Sakamoto M., 2001. Generating F0 Contours by Statistical Manipulation of Natural F0 Shapes, In *Proceedings of Eurospeech.* , Escandinavia, 1171-1174.

[9] Taylor, P., 2000. Analysis and Synthesis of Intonation using the Tilt Model. In *Journal of Acoustical Society of America*, Vol. 107 N. 3, 1697-1714.