CORPUS BASED EXTRACTION OF QUANTITATIVE PROSODIC PARAMETERS OF STRESS GROUPS IN SPANISH

David Escudero and Valentín Cardeñoso

Universidad de Valladolid (Spain)
{descuder,valen}@infor.uva.es

Antonio Bonafonte

Univ. Politécnica de Cataluña, Spain
antonio@gps.tsc.upc.es

ABSTRACT

We introduce a new corpus-based technique to model the prosodic information contained in spoken utterances. Taking the stress groups and intonation group as the structural building blocks and Bezier parametric functions to approximate F0 contours, we propose a statistical modeling of the relevant categories of stress groups. These models can be directly exploited in speech synthesis tasks in order to get more natural intonation patterns, specially for text reading applications. Suggestions are also made as for the utility of these statistical models in classification and recognition tasks.

1. INTRODUCTION

Generation of intonation patterns from input text is a crucial step of any Text to Speech Synthesis (TTS) system because is one of the most important keys to success in nowadays TTS commercial systems. Traditionally, it has been divided in two separate processes. The first one is mainly related to linguistic knowledge and has to do with accent location and characterization. The second one is obtaining a reasonably good realization for the evolution of F0 over time. Within this two-stage approach, our contribution is essentially related to the second process.

In this paper, we propose a quantitative treatment of intonation facts which is both easy to tune to intonation patterns found in real speech while trying to reflect all the essential structural characteristics of these patterns known to linguists. This aim is not new, and several quantitative approaches to intonation modeling already exist [1] [2] which have been successfully tested in commercial TTS systems. As pointed out in those works, we argue that one of the main advantages of this quantitative approach stems from their suitability to automatic task and speaker adaptation if an intonation corpus labeled with prosodic information is at hand. A reasonable set of desired features of any good intonation model have been discussed by Taylor[3] in his paper about Tilt model. Although one could perfectly follow this recommendations, perhaps with some specific details added in which could be language dependent, it would be most interesting to design a modeling technique which can easily be used as a decision tool for the selection of the optimum structural intonation features to be included in the final model. The technique we present in this work suits these needs also, since it separates parameter extraction phase from the intonation features selection mechanism. For a given set of categories of intonation units, we will be able to extract the corresponding set of statistical models which can be then used in the synthesis phase or as a decision criteria on the best set of intonation units to be chosen. Under this view, the technique could also be conceived as a first step to a classification or recognition task which incorporates prosodic information. Since the main purpose of this paper is to present the modeling technique and its use in the context of the realization of F0 contours in TTS, we will not discuss these possibilities any further.

We have chosen stress groups as the basic intonation unit (a set of unstressed words followed by a stressed word). Spanish intonation is described under the basis of stress groups in several studies (see for example [4] [5]), and it brings the practical advantage that the amount of work to process the corpus for training models is reduced.

First, we will describe the architecture of the system and the corpus used for training. Then the parameterizing technique and the procedure used for modeling intonation patterns are presented. We finish giving some experimental results and conclusions.

2. SYSTEM ARCHITECTURE

The scheme of the modeling technique proposed in this paper and its use in a TTS system are shown in figure 1. Starting from a corpus which includes phonetic and orthographic labels beside pitch contour values and the syllables and words boundaries, the *prosodic segmentation module* locates intonation units and classifies them using 'a priori' linguistic knowledge for Spanish. Every stress group will be associated with a label which reflects a given evolution

This work has been partially supported by Junta de Castilla y León under research contract nº VA-16/00A.

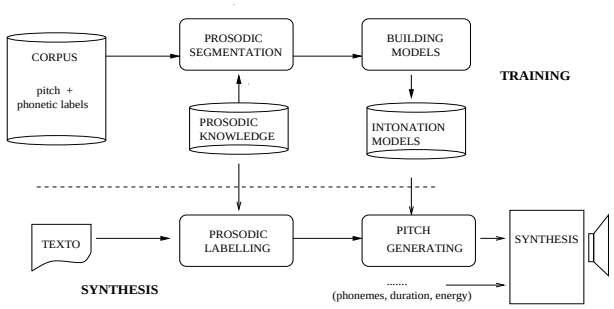

Fig. 1. Scheme of the modeling and synthesis technique

of F0 inside it. The label will also take into account the kind of intonation group in which the stress group is located, the relative position inside the intonation group (initial, inner or final) and the position of the stressed syllable (final: AC1, penultimate: AC2 or antepenultimate: AC3). This prosodic labeling model was already used in TTS by López[5].

Pitch contour values and labels of prosodic segments are then passed on to the *model building* module, which carries out the parameterization of the pitch contour of every segment and builds up a statistical representation of the set of parameters associated with all the segments that share the same label. Its output is a set of labeled statistical distributions, each one representing a given prosodic category, according with our specific structural intonation model. Most noticeable, is the fact that any variation of the set of labels does not affect our model extraction technique.

To generate a suitable F0 realization in the synthesis step, a *prosodic labeling* module will take up the building of the sequence of prosodic labels that accompanies the phonetic transcription of the input text. This labels will be used by the *pitch generation* module as an index into the intonation models database in order to get the statistical distribution that could be applied to generate a sequence of synthetic F0 contours.

3. CORPUS DESCRIPTION

The design of the corpus aimed the construction of a long units concatenative TTS system for Catalan and Spanish at the UPC (http://www.gps-tsc.upc.es) [6]. It contains three hours recordings of spoken utterances in both languages. Although it was not specifically designed for prosodic studies, it contains enough data to get significant results.

The corpus was acquired under recording studio conditions in two separate channels at 32 kHz. Speech was recorded in one of the channels and the output of a laryngograph in the other. Data were automatically labeled and manually supervised. As already pointed out, labeling included silences, allophonic transcription, and allophonic boundaries. This information was increased by the additional syllable and word boundaries and stress positions. Pitch was estimated by means of glottal pulses closing time points. It eases the automatic segmentation of stress groups and the selection of the corresponding F0 profiles.

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

Table 1. Description of corpus. IG: Intonation Group. ISG: Initial Stress Group. CSG: Inner Stress Group. FSG Final Stress Group. SG Stress Group.

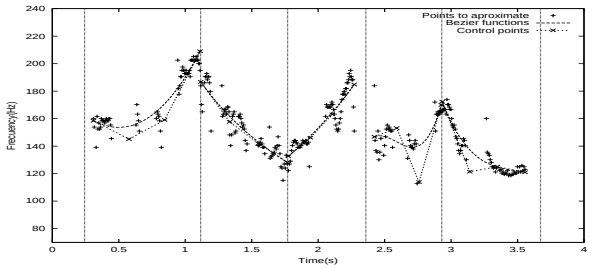

Fig. 2. Example of fitting. Vertical lines are the boundaries between stress groups. At time 2.4s a transition between intonation groups occurs.

A number of 1615 intonation groups and 4625 stress groups were processed. Intonation groups are given by the set of stress groups within two consecutive pauses or F0 steps of 30 or more Hz. Table 1 details the number of occurrences of the different stress groups for each of the six different intonation groups present in the corpus. Only the three first intonation groups have been considered in this study since the rest of them do not include enough data.

4. STRESS GROUPS PARAMETERIZATION

Each stress group (SG) is characterized by: $SG \equiv \{t_{ini}, t_{fin}, \mathbf{P}\}$; where t_{ini} y t_{fin} are the initial and final instants of time respectively, and $\mathbf{P} \equiv \{\bar{p}_i = (T_i, F0_i) | \quad i \in [0:p]\}$ are the $p+1$ points time-frequency of the pitch contour in SG.

The goal is to find a set of parameters P_i with $i \in [0, n]$ showing the evolution in time of the F0 contour and avoiding micro-intonation. The parameters P_i must reflect the various shapes of the contours of F0 in different classes of SG, taking into account that the length of different SG can vary significantly. This section shows how to make use of Bézier curves for this aim.

A Bézier curve $\bar{Q}(t)$ is a parametric polynomial of degree n defined by $n+1$ control points $\bar{P}_i$ with $i \in [0, n]$ 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* (see for example [7]).

A Bézier function in $\mathcal{R}^2$ is a especial 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 $\{(i/n, P_i), \quad i \in [0,n]\}$. The interval $[a,b]$ is the domain of the function.

For representing each SG with a function $\bar{Q}(t)$, the parameters P_i with $i \in [0,n]$ are set by optimizing the fitting of $\bar{Q}(t)$ to the corresponding $\{\mathrm{P}\}$. For doing so, the method of minimum square error is used, which implies to solve the following system:

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

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

Before using this method, F0 contours can be smoothed-filtered so that F0 values in preceding and following SGs are considered. It is important to reduce perturbations on the general tendency of the F0 contour of the intonation group.

As an alternative to filtering, the Bézier functions can be forced to join at boundaries superimposing C1 continuity. In order to obtain the parameters of SG_K, the following and preceding SGs (SG_{K-1} and SG_{K+1} respectively) must be considered in this way:

$$\frac{\partial}{\partial P_l^k} \sum_{k=K-1}^{K+1} \left(\sum_{j=0}^{p^k} \left(F_j^k - \sum_{i=0}^{n} P_i^k B_i^n(t_j^k) \right)^2 \right) = 0 \tag{4}$$

where $l \in [0,n], k \in [K-1, K+1]$, P_l^k, p^k, F_j^k and t_j^k are the values of the these parameters in SG_k. Additionally , the constrains for C1 continuity are included: $\bar{Q}^k(1) = \bar{Q}^{k+1}(0)$ $k \in K-1, K$ and $(\bar{P}_1^{k+1} - \bar{P}_0^{k+1})/(\bar{P}_n^k - \bar{P}_{n-1}^k) = (t_{fin}^{k+1} - t_{ini}^{k+1})/(t_{fin}^k - t_{ini}^k)$

5. STATISTICAL MODELLING OF INTONATION PATERNS

Once the parameters P_i with $i \in [0,n]$ of the stress groups of a given class have been obtained following the procedure explained in the previous section, a statistical distribution for each of them has to be generated. A Chi-square test has been carried out in order to determine the normality of these distributions[8]. Since not all of them did satisfy this normality test or did not have enough samples, empirical distributions should be used. Nevertheless, we can reasonably assume that every parameter follows a normal distribution $N(\mu, \sigma)$ when computing time might be compromised. Perceptually, in fact, it is not always easy to find a difference between the synthesized utterances generated by empirical or normal histograms.

Every kind of stress group corresponds to a number of distributions (4 for the results presented later on) from which synthetic pitch profiles can be obtained using a uniform random number generator as input to an inverse transform function. It is specially important to take into account the parameter covariance to avoid generating inadequate pitch profiles. In order to do so, generated samples of a parameter value will be rejected when the relative pitch change with respect to the one of the previously generated samples shows a different trend than the one prescribed by the mean values of the parameters.

On the other hand, the set of distributions associated with every stress group can be taken as the representative information about its nature, derived from real speech acts. This way, they can be exploited as a reference for classification tasks related to the inclusion of prosodic information in ASR systems.

6. EXPERIMENTAL RESULTS

For the experimental validation of the modeling technique proposed in this work, we have approximated every SG with a Bezier curve of third degree. We first compared this kind of parameterization with a simplistic straight line segments stylization made of 3 segments in order to assess the influence of the approximation quality of the segments on the final profiles. We found out that straight line approximation clearly brought worse covariance values than Bezier parameterization. So, the experimental results we report compare just three different possible variants of Bezier fitting. The first one, labeled B3 in the tables, represents isolated fitting of every SG, without taking into account the SG context. The second one, BS3, applies a previous smoothing filter in order to take into account the influence of the global intonation group on the local stress group. The third one, BC3, obtains the control parameters of the Bezier curve of a given SG considering the previous and the next SGs and imposing C1 continuity at the local SG boundary (see eq. 4).

The comparison between the three different versions of Bezier parameterization is based on mean values of the RMS with respect to the original data, standard deviation of the parameters σ and the parameter covariance ρ. RMS gives a measure of the fitting quality, while σ and ρ quantify the similarity of the samples within the same class. The bigger the ρ and the smaller the σ, the most significance can be

assigned to the model.

Table 2 shows that smoothing leads to lower mean values for σ. This could be expected, since smoothing reduces pitch contour variability. However, parameter covariance does not get better and the approximation errors slightly increase. When C1 continuity is superimposed, covariance get considerably bigger but the approximation errors and the σ do too.

	IG1			IG2			IG3		
	$<\sigma>$	$<\rho>$	$<RMS>$	$<\sigma>$	$<\rho>$	$<RMS>$	$<\sigma>$	$<\rho>$	$<RMS>$
B3	28 Hz	0.45	5.7 Hz	35 Hz	0.27	6.5 Hz	30 Hz	0.21	5.7 Hz
BS3	22 Hz	0.21	6.7 Hz	28 Hz	0.23	7.3 Hz	31 Hz	0.24	7.1 Hz
BC3	29 Hz	0.41	6.9 Hz	42 Hz	0.36	7.4 Hz	40 Hz	0.33	6.7 Hz

Table 2. Comparison between the three different parameterization techniques.

For synthesis applications, $B3$ could be selected, since the relative differences both with $BS3$ and $BC3$ are not so relevant and less processing is needed.

As a final result, we include table 3 where all the relevant statistical parameters of every model are gathered. As such, data in this table could be directly used to generate pitch contours from prosodic label text in Spanish using normal distributions. In fact, they have been tested in an experimental version of the UPC TTS system for text reading tasks, where it has been observed a clear increase of naturalness.

7. CONCLUSIONS

A new modeling technique for Spanish intonation has been proposed in which different kinds of stress groups are taken as basic intonation units. This considerably reduces time requirements of corpus processing for prosodic model extraction, since longer speech units are used as building blocks.

We have compared three different parameterization techniques that could be used combined with different prosodic units and we have showed its application to a real commercial TTS system, where an increase in naturalness has been observed, specially for text reading tasks, where monotony is always an issue.

8. REFERENCES

[1] A. Syndal, G. Möeler, K. Dusterhoff, A. Conkie, and A. Black, "Three methods of intonation modeling," in *Proceedings of 3rd ESCA Speech Synthesis 1998*, 1998.

[2] K. Silverman, J. Bellegarda, and K. Lenzo, "Smooth contour estimation in data-driven pitch modelling," in *Proceedings of Eurospeech 2001*, 2001.

[3] P. Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *Journal of Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

				P_0			P_1			P_2			P_3		
IG	Pos	St	N	μ	σ	τ	μ	σ	τ	μ	σ	τ	μ	σ	τ
IG1	ISG	AC1	184	157	21	0.85	149	42	0.19	164	36	0.67	168	25	0.15
		AC2	366	161	18	0.70	141	34	0.30	185	41	0.20	170	27	0.21
		AC3	21	163	26	0.10	161	42	0.05	207	43	0.07	172	30	0.29
	CSG	AC1	323	166	22	0.02	137	34	0.27	161	35	0.35	157	22	0.00
		AC2	657	165	22	0.01	129	35	0.91	172	38	0.00	158	25	0.02
		AC3	38	163	23	0.55	126	40	0.34	181	43	0.21	158	26	0.00
	FSG	AC1	119	150	21	0.52	123	36	0.20	130	28	0.35	117	9	0.00
		AC2	499	151	19	0.05	129	26	0.01	126	25	0.55	119	8	0.55
		AC3	50	149	17	0.08	133	31	0.00	121	26	0.44	120	7	0.92
IG2	ISG	AC1	105	156	19	0.04	169	32	0.01	152	30	0.01	192	20	0.05
		AC2	134	159	22	0.05	165	37	0.01	160	24	0.44	188	24	0.92
		AC3	1	162	0	0	276	0		129	0		165	0	
	CSG	AC1	61	172	24	0.93	167	43	0.12	139	27	0.02	170	21	0.00
		AC2	156	168	23	0.01	164	41	0.11	140	32	0.21	166	25	0.00
		AC3	7	172	28	0.03	172	51	0.01	132	29	0.01	164	32	0.12
	FSG	AC1	121	154	20	0.02	158	110	0.00	117	83	0.00	190	34	0.00
		AC2	262	158	21	0.01	142	32	0.13	150	37	0.01	189	39	0.00
		AC3	17	157	21	0.30	182	55	0.00	136	56	0.10	202	38	0.03
IG3	ISG	AC1	115	167	24	0.32	143	38	0.01	203	33	0.02	165	22	0.23
		AC2	223	169	22	0.15	139	35	0.07	216	38	0.03	167	26	0.36
		AC3	8	163	12	0.15	144	36	0.15	219	33	0.26	160	30	0.26
	CSG	AC1	140	175	24	0.37	141	35	0.51	187	31	0.12	154	19	0.24
		AC2	293	170	23	0.22	132	36	0.50	190	35	0.00	149	20	0.00
		AC3	16	182	15	0.17	118	27	0.16	192	33	0.60	148	17	0.33
	FSG	AC1	167	158	25	0.27	135	38	0.31	196	41	0.02	145	30	0.00
		AC2	298	161	22	0.00	127	35	0.97	192	49	0.00	144	38	0.00
		AC3	25	164	22	0.25	125	25	0.32	197	55	0.29	151	38	0.00

Table 3. Statistical models for the various kinds of stress groups considered in this work. SG is stress groups. (IG) is intonation group. Pos is the position of the stress group. St is accent type and N is number of samples. μ and σ give the mean value and the standard deviation of the parameter in Hz and τ is the *P-Value* of the Chi-square test[8].

[4] J. M. Garrido, *Modelling Spanish Intonation for Text-to-Speech Applications*, Ph.D. thesis, Facultat de Lletres, Universitat de Barcelona, España, 1996.

[5] E. López, J. M. Rodríguez, L. Hernández, and J. M. Villar, "Automatic prosodic modeling for speaker and task adaptation in text-to-speech," in *Proceedings of ICASSP 97*, 1997.

[6] A. Ferrer, *Sintesi de la parla per concatenació basada en la selecció*, Ph.D. thesis, Dept. de Teoría del Senyal i Comunicacions, Universidad Politécnica de Cataluña, España, 2001.

[7] G. Farin, *Curves and Surfaces for CAGD*, Cambrige University Press, 4 edition, 1996.

[8] J. E. Gentle, *Random Numbers Generation and Monte Carlo Methods (Statistics and Computing)*, Springer, 1998.