QUANTITATIVE EVALUATION OF RELEVANT PROSODIC FACTORS FOR TEXT-TO-SPEECH SYNTHESIS IN SPANISH

D. Escudero-Mancebo, C. González-Ferreras and V. Cardeñoso-Payo

Departamento de Informática. Universidad de Valladolid

Campus Miguel Delibes s/n. 47011 VALLADOLID. SPAIN.
{descuder,cesargf,valen}@infor.uva.es

ABSTRACT

A quantitative comparison of four different proposals for intonation modeling in Spanish is presented. In the framework of a modeling procedure previously introduced by the authors, the stress group is taken as the basic building block and a statistical model is inferred from a corpus for every kind of intonation unit, which is parameterized by means of the four control points of the fitting Bézier function. Applying classical clustering quality assessment metrics to the statistical models predicted under different proposals, an objective comparison is brought among them. From the results, a set of prosodic factors has been taken as the characterization of the stress group and incorporated into a TTS platform, with a reported increase in perceptual and objective quality.

1. INTRODUCTION

In Spanish, there is still no agreement on the set of relevant prosodic factors that should be taken into account from a computational point of view. Although several computational models have been tried in intonation modeling and generation (binary trees, markov models and neural networks among others), it is not always possible to carry out grounded comparisons among all these approximations, and common validation criteria are still to be found in order to lay down the fundamental basis of a good intonation modeling methodology[1].

As pointed out in [2], subjective evaluation techniques still suffer from important practical limitations: availability of listeners, difficult automation of the survey process and very low repetitiveness. Quantitative metrics, on the contrary, give a reasonable alternative allowing fast evaluation of the influence that different modifications or modeling alternatives can have on the final results. Unfortunately, traditional quantitative evaluation based on RMSE, correlation or other perceptual-like techniques[3] are still under debate [4], so it is still interesting trying to introduce new quantitative measuring criteria, which enable direct evaluation of modeling decisions and an adequate selection of best prosodic factors candidates.

In previous works[5, 6], we presented a new parametric modeling technique of intonation patterns in Spanish. In this work, we provide a comparison of four well-known proposals of intonation modeling in Spanish in the framework of our modeling procedure. The comparison is carried out in terms of quantitative statistical measures easily derived from the statistical models obtained in our

modeling technique. As a result, we argue that a common framework can be established for prosodic models assessment, TTS and ASU systems.

The rest of the paper is organized as follows: in section 2, a brief revision of our intonation modeling proposal is made, in section 3 we briefly report on the sets of different prosodic features presented by different authors. In section 4 we discuss the metrics that were used in this work. Section 5 is devoted to results and discussion and, finally, some conclusions and suggestions for future research are presented.

2. INTONATION MODELING

The intonation-modeling framework we presented in [5, 6] can be described as follows. As the first step, a *Prosodic Segmentation* module extracts stress groups (SG) from a corpus and associates them a set of features in terms of their position inside an intonation group, their kind of accent or any other relevant criteria. This module also parameterizes the F0 contour of every stress group using Bézier functions. A labeling, induced by the set of SG features, and a set of four control points of the Bézier function of a given stress group are then passed to the *Model Builder* module which makes use of these parameters to build statistical models of each of the labeled stress groups. In a TTS system, the input text would then be segmented into a sequence of stress groups, each being classified by the *Prosodic Labeling* module and passed to the *Pitch Generation* module, which assigns a F0 contour to every stress group using its label to retrieve its associated statistical model from the *Intonation Models* database.

The originality of our contribution stems from the use of the stress group as the basic building block for intonation modeling, the use of Bézier functions to parameterize individually this linguistic units and the correspondence of a statistical distribution of Bézier function control points to every distinct class of SG. This last distinctive characteristic grounds on the fundamental hypothesis that stress groups of the same kind will have similar, although not necessarily identical, pitch contours. Every kind of SG is labeled by the *Prosodic Knowledge* module, which defines the number and nature of the prosodic characteristics to be considered.

Although in previous works we used a classification of prosodic units inspired by the ideas of López[7] for illustration purposes, the adequate selection of the classes of stress groups can be the key to success of our methodology. In this way, a quantitative evaluation of other well known proposals of prosodic features classification as the one presented in this work could shed light on what is the most grounded decision.

3. REVIEW OF THE FOUR STUDIED MODELS

There are reference works on modeling of Spanish intonation which introduce several prosodic factors to characterize intonation [7, 8, 9, 10]. Since we have focused in these works for the comparison presented in this paper, we will briefly review their proposals in this section. The four studies reported here were developed for the same kind of applications that ours: TTS or automatic analysis of intonation. In spite of this similarity, it should be pointed out that it has been necessary to perform a careful interpretation in order to match the ideas appearing in this works to our basic intonation unit. In some cases, a projection of the prosodic factors was enough and in some others, the original domain of some of them was cut down in order to adequate it to the size of the corpus we used.

In Lopez [7], intonation is modeled by an stylized representation associated with every type of syllable. Syllables are classified in terms of the kind of intonation group they appear in, their position inside the stress group they belong to, relative position of the stress group within the intonation group and of a Boolean flag indicating if they are stressed or not. The adaptation of this proposal to our modeling framework involves using three different prosodic factors: (1) the type of intonation group (declarative final, non final rising and falling, and neutral); (2) the position of the stress group relative to the intonation group (initial, medial and final) and (3) kind of accent (last, penultimate or antepenultimate syllable).

Garrido [8] carried out a study based on ststylization of intonation contours of different scope. This author enumerates a set of relevant prosodic factors at stress group, intonation group, sentence and paragraph levels. For the purposes of the present study, we have only considered the five following factors: (1) Position and type of stress group (initial, post-initial, medial, final falling, final rising and final rise-fall); (2) Number of syllables in the stress group, taking into account just two extreme situations (less than 2 and 2 or more); (3) Kind of accent, with the same domain as in the proposal by López; (4) Position of the intonation group inside its sentence (initial, medial, final and initial-final (meaning a one IG only sentence)); (5) Number of syllables in the intonation group, taken as a Boolean value indicating whether 7 or more syllables were present or not.

Using a quite different approximation, Vallejo [9] introduces the concept of syllabic nuclei plus a 10 syllables context (5 preceding and 5 following ones). Applying neural networks classifiers, this author concludes that the relevant prosodic factors having an influence on this syllabic nuclei are: the accent, whether the syllable is in the initial or final zone within the intonation group (initial zone starts at the first syllable and ends at the end of the first stressed syllable and final zone takes the rest of the intonation group), the number of syllables of the intonation group, the kind of pause delimiting it and the kind of terminal juncture with the next IG (falling or rising). The adaptation of these prosodic factors to our specific framework implies choosing the 5 following ones: (1) Kind of pause (final declarative or not); (2) Position of the stress group within the intonation group (initial, middle, final and initial-final); (3) Kind of accent; (4) Number of syllables inside the intonation group (1, from 2 to 5, from 6 to 10, from 11 to 15 and more than 15); (5) Kind of terminal juncture (falling vs. rising).

Finally, Alcoba et al. [10] describe Spanish intonation in terms of the stress group, as we do, although they follow the INSINT model. In this way, we can directly consider the same relevant factors proposed by them: (1) the number of stress groups inside an intonation group, (2) the position of the stress group, (3) the kind of accent of the stress group (last syllable versus others), (4) the analytical trend of the terminal juncture (rising vs. falling) and (5) a Boolean flag indicating whether the intonation group is final or not.

In the rest of this work, we will refer to each of these alternatives using the name of the main author, already introduced above.

4. THE METRICS

The grounding hypothesis of all the experiments presented here is that stress groups belonging to the same class will show similar intonation patterns and, thus, similar control point values of the Bézier functions fitting them. Under this assumption, a correspondence can be built between the kind of stress group (defined by a set of relevant factors) and the set of patterns corresponding to it (represented by a class of intonation profiles). Since different proposals introduce alternative classification spaces, they can be seen as different clustering of the same data set. In consequence, a quantitative bundle of quality metrics can be evaluated for this classifiers using well known concepts from clustering theory[11].

The metrics described below have been used under different clustering conditions in the experiments described in section 5. In some sense, all of them give similar information about the quality of a given clustering: the smaller the values, the better the classification.

M1: Sum of the squared classification error.

$$M1 = \sum_{i=1}^{N_c} \sum_{\bar{P} \in C_i} ||\bar{P} - \bar{\mu}_i||^2 \qquad (1)$$

where $\bar{P}$ are the parameters of the stress groups belonging to class C_i, and $\bar{\mu}_i$ with $i = 1..N_c$ is the mean value vector representing C_i; N_c is the number of classes; $||\bar{P} - \bar{P}'||$ represents the Euclidean distance between vectors $\bar{P}$ y $\bar{P}'$ and measures self-similarity of the samples in a same class.

M2: Intra-class samples distance.

$$M2 = \frac{1}{2} \sum_{i=1}^{N_c} n_i s_i, \qquad (2)$$

where

$$s_i = \frac{1}{n_i^2} \sum_{\bar{P} \in C_i} \sum_{\bar{P}' \in C_i} dist(\bar{P}, \bar{P}') \qquad (3)$$

and $n_i, i = 1..N_c$ is the number of elements of the class C_i. If $dist(\bar{P}, \bar{P}') = ||\bar{P} - \bar{P}'||^2$ then $M2 = M1$. As vectors $\bar{P}$ y $\bar{P}'$ are the control points of two Bézier functions (see [6]), we will use the area difference between Bézier functions as the value for the distance.

M3, M4: Scattering measures.
The scatter matrix for cluster C_i is computed as:

$$\mathbf{S}_i = \sum_{\bar{P} \in C_i} (\bar{P} - \bar{\mu}_i)(\bar{P} - \bar{\mu}_i)^t \qquad (4)$$

The intra-cluster scatter matrix would be computed as:

$$\mathbf{S}_W = \sum_{i=1}^{N_c} \mathbf{S}_i \qquad (5)$$

The inter-cluster scatter matrix is:

PROPOSAL	$\#_c$	$\#_{ne}$	$M1_r$	$M2_r$	$M3_r$	$M4_r$
Lopez	36	36	0.241	0.238	0.705	0.225
Garrido	288	181	0.366	0.339	0.854	0.350
Vallejo	240	141	0.286	0.277	0.767	0.275
Alcoba	200	116	0.267	0.271	0.741	0.250

Table 1. Comparative results for the four described metrics when all the possible classes are taken into account. $\#_c$ represents the number of classes and $\#_{ne}$ the number of non-empty classes found from the corpus.

$$\mathbf{S}_B = \sum_{i=1}^{N_c} (\bar{\mu}_i - \bar{\mu})(\bar{\mu}_i - \bar{\mu})^t \qquad (6)$$

where $\bar{\mu}$ is the mean vector over the set of samples. The total scatter matrix $\mathbf{S}_T$ can be computed as $\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$.

To obtain a single scalar indicator of these matrices, we considered the trace and the determinant. The trace of $\mathbf{S}_W$ gives again M1. From the determinant, new metrics are obtained:

$$M3 = |\mathbf{S}_W|, \quad M4 = tr[\mathbf{S}_T^{-1}\mathbf{S}_W] \qquad (7)$$

5. RESULTS AND DISCUSSION

Corpus ESMA-UPC[12] was used to carry out all the experiments presented in this paper. Although it has not been specifically designed for intonation analysis, the number of samples is high enough to obtain representative results, specially for declarative sentences. That's why we have carried out this study only for this kind of sentences, for which 4354 different stress groups are contained in the corpus. Should we have at hand a greater corpus, the results could be extended and rebuilt immediately.

For the first part of the experiment, a comparison of the four intonation models was carried out evaluating the four metrics described in section 4 for the different number of classes prescribed by each one. The reference point for each metric was its value for a one-class clustering, which comes to represent the worst possible classification strategy. Table 1 shows the results for this comparison. In this table, and in the two following ones, the following notation was used: $\#_c$ represents the number of classes, $\#_{ne}$ the number of non empty classes and $Mi_r, i = 1..4$ are the relative metrics used to compare the proposals, in per unity scale. As for table 1, $Mi_r = (Mi_0 - Mi)/Mi_0, i = 1..4$ where Mi_0 represents the value of metric Mi when it is evaluated over all the samples of the corpus. This means that the higher the value of Mi_r, the greater the quality of the classification, since these relative metrics can be interpreted as the relative improvement obtained over the all-in-one clustering when a specific classification is chosen. As seen in the results, Garrido gives a clear better result than the rest of the proposals, which show similar behaviour.

The numbers in table 1 could also suggest that the results are to be better the higher the number of classes. Thus, we have carried out an alternative comparison experiment in which the total number of non empty classes was kept to a common minimum value of 36, given by Lopez. To do this, an iterative merging procedure was carried out. In each step, two classes were chosen and merged into one if the influence of this merging on the value of $M1$ was the smallest possible one. Again, Mi_0 was used as a normalizing

PROPOSAL	$\#_c$	$\#_{ne}$	$M1_r$	$M2_r$	$M3_r$	$M4_r$
Lopez	36	36	0.241	0.238	0.705	0.225
Garrido	288	36	0.348	0.311	0.828	0.325
Vallejo	240	36	0.278	0.261	0.751	0.250
Alcoba	200	36	0.262	0.260	0.731	0.250

Table 2. Comparative results for the four described metrics when the number of classes is iteratively merged to 36 as in Lopez. Again, the higher the value of the relative metric Mi_r, the greater the clustering quality.

Proposal	$\#_{ne}$	$M1_r$	$M2_r$	$M3_r$	$M4_r$
Lopez	36	0.784	0.455	0.999	0.677
Garrido	181	0.889	0.601	0.999	0.846
Vallejo	141	0.888	0.607	0.999	0.828
Alcoba	116	0.878	0.585	0.999	0.833

Table 3. Values of the four described metrics when they are normalized against their expected values after a K-means clustering with the number of non-empty classes found in table 1.

reference, with the same meaning than before. The results of this second comparison are shown in table 2. Although the values of the metrics are now smaller, the same comments apply: Garrido is still the best alternative.

Since the results in tables 1 and 2 serve only to compare the clustering quality of the different proposals with respect to each other, we thought it would also be interesting to get information about the, so to say, absolute quality of each proposal. As an approximation to this question, we designed a final comparison in which all four proposals were compared to a common ground classification, obtained through a classical K-means clustering algorithm applied to the number of non-empty classes prescribed by each proposal. In this case, the relative metrics shown in table 3 give a measure, in parts of unity, of the improvement that would be necessary within a given classification proposal in order to get the ideal one. So, a relative value of 1 means 'all improvement' still to be made and a value of 0 would mean 'perfect K-means classified'. More precisely, $Mi_r = (Mi - Mi_k)/Mi \quad = 1..4$, where Mi_k is the value of metric Mi obtained after an automatic K-means clustering has been carried out with the samples of the corpus. The most striking result of table 3 is that no single proposal is close enough to the optimal situation. All of them show a noticeable disagreement with the classification scenario that would be expected starting from scratch and adding no linguistic knowledge at all. This is specially true for metric $M3$, which should be the most distinctive one and is close to 1 in all cases. A difference in the relative behaviour of $M1$ and $M2$ with respect to the one found in tables 1 and 2 is also representative, since K-means algorithm tends to minimize $M1$ and fixes a reference far away from the values obtained in the proposed classifications.

A common sense reading of the results in table 3 could be that a lot of simplification is made in every proposal with respect to what can be found in real samples: there is much more behaviour scattering than expected. Furthermore, it could also be argued that a better model can be inferred from prosodic corpus data by trying to find the number and domain of possible influencing prosodic

	RMSE	Pearson Coef.
Test-1	18.93	0.70
Test-2	17.85	0.73

Table 4. RMSE and Pearson Correlation coefficients of the distance between original and synthetic F0 contours. The latter were obtained as the mean value of all the patterns in the class corresponding to a same stress group. In Test 1, 75 % of the corpus samples were used to get the statistical models for the F0 contours and the rest (25 %) to run the regression test. In Test 2, 100 % of the corpus samples were used both to build the models and to run the tests.

factors which lead to a better clustering agreement with real samples. Although there are still no final results to be published, we are working in this direction at the moment.

A final remark about empty classes is suitable at this point. The fact that we have empty classes is related to the fact that some of the prosodic categories prescribed in the models correspond to highly infrequent stress groups (like antepenultimate accents in Spanish). For TTS purposes, this can cause problems, since no model would be obtained in this cases, the only solution being acquiring a specific and more complete prosodic corpus, which was out of the scope of this study.

Although none of the proposals seems to give the most adequate set of prosodic factors, we have incorporated the one by Garrrido to our intonation modeling procedure for TTS, since it gives the best overall results. Preliminary perceptual tests showed that an intonation quality similar to other commercial systems is obtained. Table 5 shows the results of the conventional RMSE and Pearson correlation tests applied to our TTS system when intonation modeling is carried out in terms of stress group units and the classification proposed by Garrido.

6. CONCLUSIONS

In the framework of an intonation modeling procedure previously introduced by the authors, a method to quantitatively evaluate given sets of prosodic factors has been described. By means of four metrics commonly used for clustering quality assessment, four different intonation modeling proposals for Spanish have been evaluated. The best of these proposals has been incorporated into the *Prosodic Knowledge* module of our TTS platform and its quality has thus been increased.

Nevertheless, there are two reasons why the results presented here cannot be interpreted as a ranking of the four studied models. First, a particular interpretation was necessary in order to adequate three of them to our modeling framework. Second, the results are obtained with a particular prosodic corpus which adequacy for prosodic studies still has to be revised. In order to have a definitive ranking of the models, it would be necessary to design a specific prosodic corpus where all the possible classes foreseen in the proposals would be equally balanced.

Finally, our statistical modeling technique of the control points of the Bézier function that closely approximates the stress group, taken as the basic building block of intonation, brings ways to objectively evaluate the influence of prosodic factors. This opens new possibilities for the automatic extraction of prosodic knowledge from corpora and brings a possible common framework both for TTS applications and prosodic aided Automatic Speech Understanding.

8. REFERENCES

[1] A. Botinis, B. Graanstrom, and B. Mobius, "Developments and paradigms in intonation research," *Speech Communications*, vol. 33, pp. 263–296, 2001.

[2] R Bezooijen V van Heuven, "Quality Evaluation of Synthesized Speech," in *Speech Coding and Synthesis*, chapter 21, pp. 707–738. Amsterdam: Elsevier, 1995.

[3] D. J. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, pp. 73–82, February 1998.

[4] RAJ Clark and KE Dusterhoff, "Objetive methods for evaluating synthetic intonation," in *Proceedings of Eurospeech 99*, September 1999.

[5] V. Cardeñoso and D. Escudero, "Statistical modelling of stress groups in spanish," in *Proceedings of ISCA Prosody 2002*, 2002.

[6] D. Escudero and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish.," in *Proceedings of ICASSP 2002*, 2002.

[7] E. López, *Estudio de Técnicas de Procesado Lingüístico y Acústico Para Sistemas de Conversión Texto Voz en Español Basados en Concatenación de Unidades*, Ph.D. thesis, E.T.S.I de Telecomunicaciones, Universidad Politécnica de Madrid, España, 1993.

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

[9] J. A. Vallejo, *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, 1998.

[10] S Alcoba and J Murillo, "Intonation in Spanish," in *Intonation Systems. A Survey of Twenty Languages*, chapter 8, pp. 152–167. Cambridge University Press, 1998.

[11] R. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[12] A. Ferrer, *Sintesi de la Parla per Concatenació Basada en la Selecció*, Ph.D. thesis, Dpto. de Teoría del Senyal i Comunicacions, Universidad Politécnica de Cataluña, España, 2001.