A MIXED STRATEGY APPROACH TO SPANISH PROSODY *

Juan Manuel Villar Navarro *Eduardo López Gonzalo* *José Relaño Gil*
ETSI Telecomunicación. Universidad Politécnica de Madrid
Ciudad Universitaria s/n. 28040 Madrid SPAIN.
e-mail:juanma@gaps.ssr.upm.es, eduardo@gaps.ssr.upm.es, jrelanio@gaps.ssr.upm.es

Abstract

In this paper we introduce a new method for the synthesis of Spanish prosody suitable for the automatic generation of prosody in a Text-to-Speech system. As some methods we have already proposed our approach is data-based, it models joint F0 contour and segmental durations and its linguistic analysis is rule-based. Unlike previous works it uses a mixture of *a priori* breath group classification (linguistically based) and data-based phonological mapping. This new approach together with the previous ones form a quite open framework for analysis and synthesis of Spanish prosody.
This approach leaves room for growing from a general prosodic model towards application specific prosody. The new method is successfully tested in a particular style of telephone number reading, quite difficult to pick up by previous methods.

10.21437/Eurospeech.1999-411

1 INTRODUCTION

In our previous research we introduced two different approaches to Spanish Prosody that we called "manual" [3] and "automatic"[4]. Both, manual and automatic approaches share many characteristics in common: both perform a joint modeling for F0 contour and segmental durations, both assume a relation between sequences of POS tags and classes of breath groups, both assign prosody in a syllable by syllable basis and both assign the actual F0/duration values from a data-base. Furthermore "manual" and "automatic" methodologies start from an analysis on a recorded corpus, from which some data is obtained to be used in the synthesis of prosody, as depicted in figure 1.

The main phonological difference is that breath group classes in the "manual" methodology were linguistically determined, while breath group classes of the "automatic" methodology were obtained by means of spontaneous clustering of the acoustical features of the sentences in the corpus. In the latter case the method trained a set of rules relating linguistic structures with the breath group classes found in the corpus (The so called "mapping rules"). Figure 2 summarizes the analysis method, both phonological (mapping rules) and phonetic (Prosodic Database).

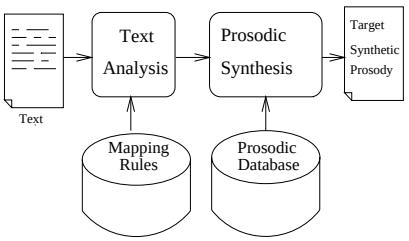

Figure 1: General overview of the synthesis process.

The phonetic difference was in the way of obtaining the intonation contours and duration of the syllables. In the "manual" methodology, a database of syllables covering all relevant variations (for every breath group, for every position, accented or not, etc.) was built by averaging similar appearances in the corpus. The second one used a similar database of syllables but with the inclusion of every (or almost) syllable of the corpus, a pre-selection of candidates and dynamic programming selection between them was performed in order to obtain the actual contours [4].

The main advantage of the first method is its robustness both in the phonological and phonetic levels. In the phonological level the method is less prone to error in the linguistic analysis as its rules perform a well understood set of operations.

This work has been supported by CICYT under the project No. TIC-96-0956-C04-03

6th European Conference on Speech Communication and Technology (EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

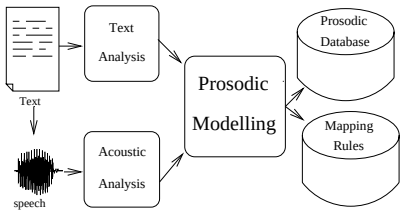

Figure 2: General overview of the analysis process of the "automatic" methodology

In the phonological level, given that its contours were averaged from the corpus, possible errors in segmentation and/or pitch estimation were easily eliminated. In the minus side, the prosody tends to be monotonous and unnatural.

In this paper, trying to improve our previous manual and automatic approaches, we propose a new one which brings together the best of both methods. It is also a framework in which we can inscribe the two previous methods.

The rest of the paper is organized as follows. Section 2 introduces the framework in which we insert old methods and the new one. Our present mixed-strategy approach and experiment over it are presented in section 3 Finally, conclusions and proposals for further research are given at the end of the paper.

2 FRAMEWORK

Almost any prosodic analysis/synthesis method can be seen as a two step process: phonological and phonetic. From the text input you have to derive relevant structures (phonological) which in turn would allow to assign durations, intonation contours, intensities and pauses (phonetic).

2.1 Phonological part

The first step is usually done by means of linguistic analysis of the text and some kind of pattern matching which may reveal prosodic patterns (usually breath group classes or BGCs). Most of the times these prosodic patterns reflect actual thinking in linguistic theory (enunciatives vs. interrogatives, rise vs. fall, etc.), although sometimes they can be application oriented (e-mail reading, telephone number, etc.). for our purposes we distinguish between three sources of breath group classes:

1. derived from linguistic analysis,
2. tagged on input text,
3. acoustically derived.

For the first kind you need to provide a rule set capable of finding such BGC from POS tagging, chances are to write down such a rule set or try to train it from a tagged corpus. For the second kind, no computing is needed as they are already tagged. For the third kind you need to train a rule set which maps sequences of POS tags into the obtained BGCs.

This framework leaves room for different linguistic analysis through the tagged text input method. Thus, effects like focal accent [1] semantic correlates of the intonation, etc. can be tagged externally and then processed.

2.2 Phonetic part

The phonetics of intonation remain as an open problem in Spanish. The traditional configurations Model of Navarro Tomás [6] and the three level model of Quilis [7], contrast with Juan María Garrido Model [2] (quite close to IPO theories) or Juan Manuel Sosa [8] study of Spanish intonation according to Pierrelhumbert's model.

All these models concentrate on the description of the general melody of different sentences. All of them admit some variations based on speaker origin, intention, etc. but their effects remain unexplained. Given those facts we have always opted for data-driven approaches for synthesizing intonation, at least in the phonetic side.

We have used, so far, two approaches for intonation assignment. One consisted on looking up in a database containing averaged syllables form the corpus; each instance corresponding to a given combination of: BGC, stressed nature, distance to the end of the breath group, nuclear vowel, etc. In the other approach a database was built with every syllable in the corpus (or a big selection of them). Selection was based on linguistic and acoustic properties, and in a dynamic programming algorithm.

3 EXPERIMENTS

In order to test our framework we built a system enclosing a general part and an application spe-

[1] as stated by Maria Luisa García Lecumberri in her PhD Thesis [1] it is not as important in Spanish as in English, but it still plays a role.

cific one. Input should include free text together with tagged breath groups. The tags lets us experiment with different speaking styles mixed with standard reading style. For the phonetic part we used several syllable databases: the general intonation one was pruned (for robustness), for the speaking-style specific ones we used every syllable in the corpus (as there were not many at all).

The experiments consisted in recording a general corpus and some application specific corpora, of which we present results for an address and telephone reading corpus. Breath groups in the general corpus were derived and classified from linguistic analysis. Application specific breath groups were so tagged in the input text.

Both corpus were recorded for the same speaker. They were then processed in order to find segmental durations, energy and intonation contour (which we describe in terms of initial, middle and final frequency for each vowel). With this data and the linguistic analysis, a database of syllables were formed. Each syllable was classified and ordered according to its linguistic properties (which will serve as the basis for the selection of candidate syllables), the main one being its breath group class (BGC). Acoustic data for each syllable and its original environment is also preserved, this is the information which will serve us to obtain a best alignment of candidates. In order to simplify processing the f0 contours and onset and rhyme lenghtenings are quantified by means of an LBG Vector Quantizer. This step produces a certain degree of stylization, as the acoustical parameters of the centroids are less affected by errors than individual values.

For the prosodic synthesis we perform a linguistic analysis which gives us the pauses, BGCs (these two may come tagged in the input text), syllables, stress and phonetic transcription. Afterwards we perform the phonetic-prosodic pass. In this pass we proceed syllable by syllable selecting the best n^2 candidates in the database. Candidates are selected among the syllables of the same breath group taking into account the stress, distance to the end, and segmental similarity (same vowel and similar consonants). Once all the candidates are selected a dynamic programming is performed for which the distance is obtained by means of similarity of previous/next syllable pitch contour.

[2]Computing time grows as the square of the number of candidates, so a moderate quantity should be taken. We usually take 10 to 20

3.1 Classes of Breath Groups

For the breath groups in the general corpus two approaches have been tested. One is based on prosodic significant structures (yes-no questions, wh-questions, open and closed enumerations, etc.). Possible lists of such phenomena for Spanish have been proposed by several authors, see for instance [6]. In previous works we have employed a set of nine encompassing some general raising or falling patterns whith more specific ones (enumeration, parenthetical, etc.)

The other one is to assign the five configurations (*"tonemas"*) enumerated by Navarro Tomás directly. There are five ones: two rise (*anticadencia* and *semianticadencia*), two fall (*cadencia* and *semicadencia*) and a level one(*suspensión*).

3.2 Intonation of Telephone Numbers

So far adaptation to different speaking styles is done by means of new classes of breath groups and its associated database of prosodic characteristics of syllables in the corpus. These new classes should be annotated in the input text, as there is no easy way to tell them apart based on linguistic analysis. Later, rule training methods could be used in order to try to find mappings as we proposed in [5].

The task of reading telephone numbers poses some interesting problems: how to group the digits, which intonation contours appear, rhythmic questions. For our experiment the grouping of the numbers was decided in advance and the speaker was teach to read them so.

Either way for each new speaking style a new set of BGCs is in order. In the telephone number reading task 2, 3 or 4 BGCs can be devised. In a first approach most people tend to produce an S-shaped contour, for all groups of digits but the last, which usually is an inverted U. From a level viewpoint, the first group of digits tends to finish higher than the other ones. In our corpus the reader produced the highest frequency in the first accent of the second group (see figure 3. This leads us to 4 different BGCs one for the first group, one for the second, one for the last group and one for the rest.

Figure 3: Intonation of a telephone number (91 431 72 97).

We first tried the automatic methodology trained on telephone number breath groups alone. The spontaneous clustering of BGCs gave rise to 3 BGCs, so the second a third groups of digits were in the same BGC (remember that groups are classified according to its last accented syllable). The final effect was related with the number of digits in the group. As the most common digits grouping is 2-3-2-2 the highest first accent was related to the 3 digits grouping, so that groupings of the kind 2-2-3-2 were incorrectly produced with a high first accent in the third breath group.

With the new mixed approach we tagged 4 different BGCs, and built a database with them. The synthesis produce much more similar results. Preliminary evaluation by casual listeners, indicate that the perceived style of intonation produced by synthesis was more similar to the original than the one produced by means of automatic methodology.

4 CONCLUSIONS

Although more experiments are needed in order to be able to have strong conclusions, this new method shows important benefits over past ones. The mixed approach helps to fit better different speaking styles. At the same time it helps to gain understanding of the problem.

The problem of deriving the relevant BGCs from the linguistic analysis remains open. For many speaking styles this may be an impossible task. As it may depend on extra-linguistic clues (such as those found in dialogue systems).

In future works we foresee to research on different methods of integrating "a priori" knowledge with actual data. One such approach may be to separate the intonation contour into macro melody and micro melody. The second one modulating the first one. Thus a general macro melody for each breath group could be derived form the corpus by means of averaging. The syllable database would then be used only for micro-melody.

One possible approach could be to use the same modeling and VQ we use in syllables with breath groups. A moving average of the pitch of each breath group would be computed and subtracted from each syllable. Separate calculations (simplification, and clustering) would be performed on breath groups and syllables.

References

[1] M. L. García Lecumberri. *Intonational Signaling of Information Structure in English and Spanish: A Comparativae Study*. PhD thesis, University of London, 1995.

[2] J. M. Garrido Almiñana. *Modelling Spanish Intonation for Text-to-Speech Applications*. PhD thesis, Universitat Autónoma de Barcelona, Facultat de Lletres, 1996.

[3] E. López Gonzalo and L. A. Hernández Gómez. Data-driven joint f0 and duration modeling in text to speech conversion for spanish. In *Proc. ICASSP*, pages I.589–592.

[4] E. López Gonzalo and L. A. Hernández Gómez. Automatic data-driven prosodic for text to speech. In *Proc. EUROSPEECH '95*, pages I.585–588, Madrid (Spain), 1995. European Speech Communication Association (ESCA).

[5] E. López Gonzalo and J. M. Rodríguez García. Statistical methods in data-driven modelling of spanish prosody. In *Proc. ICSLP*, pages 1373–1376, Philadelphia (USA), 1996. Delaware University.

[6] T. Navarro Tomás. *Manual de Pronunciación Española*. C.S.I.C., Madrid, Spain, (1972) 17 edition, 1945.

[7] A. Quilis. *Fonética Acústica de la Lengua Española*. Editorial Gredos, Madrid, Spain, 1981.

[8] J. M. Sosa. *La Entonación del Español*. Ediciones Cátedra, to be published.