# STATISTICAL METHODS IN DATA-DRIVEN MODELING OF SPANISH PROSODY FOR TEXT TO SPEECH

*E. López-Gonzalo and J.M. Rodríguez-García*

E.T.S.I. de Telecomunicación. Univ. Pol. Madrid
Dep. Señales, Sistemas y Radiocomunicaciones.
Ciudad Universitaria. 28040-Madrid (Spain).
Tel:34.1.5495700. Fax:34.1.3367350. e-mail: eduardo@gaps.ssr.upm.es

## ABSTRACT [1]

In [1], we proposed an automatic data-driven methodology to model both fundamental frequency and segmental duration in TTS converters from a monospeaker recorded corpus. Therefore, it had the advantage that could be adapted to a specific corpus or a particular speaker. The main disadvantage was the size of the obtained prosodic database.

In this paper, we propose to use some statistical methods for reducing the prosodic database required in this methodology. A 50% of reduction can be obtained without compromising the naturalness of the synthetic speech obtained by our previous methodology with the same prosodic corpus. A compromise between variability and reduction in prosodic contours is also discussed.

## 1. INTRODUCTION

Generating proper prosodic information is one of the most important issues for synthesizing speech. Although many approaches of prosody generation had been proposed in the past for text-to-speech (TTS), it still remains a problem to model the variability and fluency of natural speech. For a number of years, the build-up and systematic use of a prosodic corpus (see for example [2] for French) have been recognized as the key to generate more natural synthetic speech. The problem is how to extract the prosodic knowledge from this database.

We proposed in [3] a methodology to model both fundamental frequency and segmental duration in TTS converters for Spanish. The prosodic generation was based on the computation from the text of some independence coefficients between words. These coefficients were the nexus between the linguistic features obtained from text and the prosodic patterns. The prosodic patterns were obtained by averaging some manually labeled data from a single speaker. This "manual methodology" was a subjective and tedious time-consuming work. Other problems appeared for example to adapt the system to a new speaker, where a new prosody should be generated. Therefore, there was a need for automatic methodologies for prosodic modeling.

For this reason, we proposed in [1] an automatic data-driven methodology to model both fundamental frequency and segmental duration in TTS converters with an application to Spanish language. This methodology had several advantages. First, it showed a very good prediction of the two prosodic parameters that reminded the reference speaker, and a greater variability than using the "manual methodology" was obtained. Furthermore, this methodology had the advantage that could be easily adapted to a specific corpus or a particular speaker. This is remarkably good when personalizing a TTS system. Unfortunately, the size of the prosodic database for synthesis grows according to the size of the corpus. The focus of this paper is try to explore some methods for reducing the size of the prosodic database.

The outline of the paper is as follows: first, the methodology for prosodic modeling is presented in a general framework. Second, our previous manual methodology is explained as a particular case of the general framework. Then, we explain in detail how the automatic methodology works in modeling and generation of prosodic contours. The rest of the paper focus on the methods that we have designed to reduce the prosodic database. The paper ends with results and conclusions.

## 2. METHODOLOGY FOR PROSODIC MODELING

As we said above, the use of a prosodic database is the current tendency to obtain a prosodic model. This is what we call a data-driven approach. The prosodic model describes the relationship between some linguistic features extracted from a text corpus and some prosodic features extracted from a related speech corpus. Then, a TTS system relies on a prosodic model to generate the prosodic parameters.

The general scheme for producing a data-driven prosodic model is shown in **Figure 1**. The input to the system is a monospeaker recorded prosodic corpus and its textual representation. It should be noticed that, in general, it is enough in speech synthesis to produce a prosodic model of one speaker, so no multispeaker corpora are needed.

The first thing is to obtain some relevant linguistic features from the text. Here, it is important to define a prosodic phrase structure to be able to classify the acoustic features extracted from prosodic contours. In the case of the Spanish language, we have used (see [1], [3], [4]) a prosodic structure that considers syllables (group of sounds with only one vowel), accent-groups (group of syllables with one lexical stress), breath-groups (group of accent-groups between pauses regardless of the duration of the pause) and sentences.

**1** part of the work was supported by a CNET France Telecom's contract

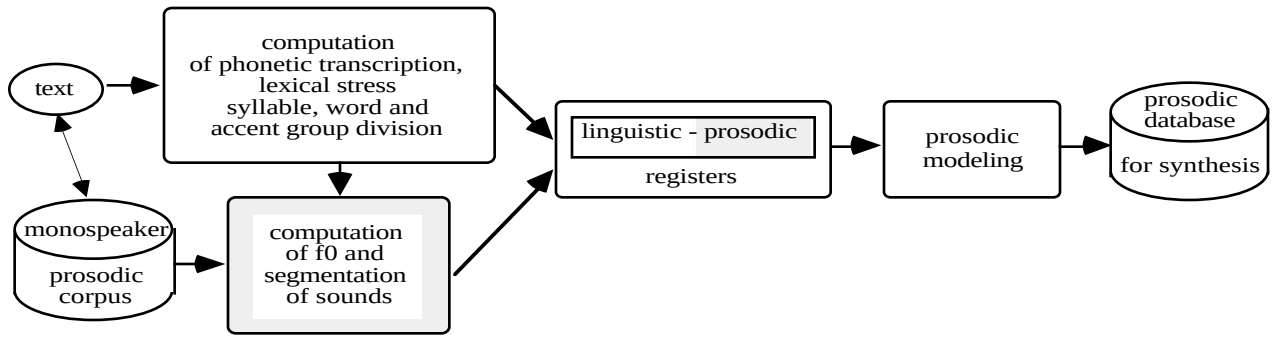


**Figure 1:** Data-driven methodology for prosodic modeling.

As shown in **Figure 1**, the first module obtains from the text its phonetic transcription, its lexical stress marks and its division into syllables, words, accent-groups, breath-groups and sentences.

The second module calculates the prosodic parameters. The prosodic parameters needed for synthesis include the pitch (or f0 contour) for voiced sounds, and the duration and energy level of all sounds. The energy level is generally considered the less important in synthesizers by concatenation of speech units. Then, the output of this module is the f0 contour and the segmental duration.

We have observed that there is some correlation between these two parameters, so we think that a more efficient modeling is obtained with a joint representation of them. Both in [1] and [3], we have used the syllabic prosodic contour shown in **Figure 2**. This contour is represented by 5 parameters (two duration values and three pitch values) obtained from the vowel nucleus of a syllable. This set of five parameters can be seen as a vector in the 5-dimensional space of "prosodic syllabic patterns" (PSP´s).

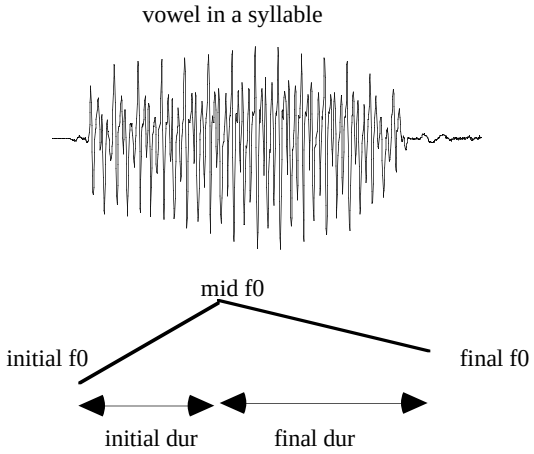


**Figure 2:** Prosodic contour in a syllable defined by 5 parameters to form the prosodic syllabic pattern (PSP).

The output of these two modules is a register of linguistic and prosodic parameters for each syllable in the prosodic corpus. The prosodic model will try to describe the relationship between these two types of parameters. In our methodology, this is explicitly achieved by a database of prosodic patterns available for synthesis. The database procedure is a more flexible mechanism to generate prosody than a rule-based one. The key question is how to process the set of linguistic-prosodic registers to form the prosodic database for synthesis. In the following section, we will describe two particular methods that we have employed to build up this database, namely, the "manual methodology" and the "automatic methodology".

## 3. MANUAL AND AUTOMATIC METHODOLOGIES.

In [3] and [4], we described in detail our manual methodology for joint data-driven prosodic modeling of f0 and segmental duration. The segmentation of the speech was done by hand, and the pitch contour was subsequently "stylized" following the IPO methodology [5]. The prosodic database for synthesis was obtained simply doing an average of all linguistic-prosodic registers with the same linguistic parameters. This procedure has the advantages that it produces a very small prosodic database for synthesis and it shows a good performance synthesizing prosody, but it has the disadvantage that is subjective, very time consuming. Besides, it is noted a lack of variability in the synthetic contours due to the averaging procedure for prosodic modeling.

Our next step was to improve the variability in the synthetic prosody. For this purpose, we designed the automatic methodology described in [1]. This methodology performs a total automatic analysis of the prosodic corpus to produce the linguistic-prosodic registers. Each register is formed by 10 features, 5 linguistic features and 5 prosodic features. The linguistic features are name of the nuclear vowel, type of accent-group, type of breath group, distance of the syllable to the lexical accent and place of the syllable in the accent-group. The type of breath group is calculated according to the PSP of the syllable just before the pause.

The prosodic features are:

- the duration of the pause after the syllable (for pause modeling)
- the rhyme lengthening and its difference to the syllable onset lengthening (both calculated as in [6] for consonant modeling)

° two features for vowel modeling: the vector representing the PSP of the current syllable (PSP1) and the pattern of the following syllable in the speech corpus (PSP2).

In the automatic methodology, we do not follow the IPO approach. We reduced the size of the memory to store the prosodic database by means of vector quantization of all PSP's available in the linguistic-prosodic registers. As we said above, the PSP is formed by a 5-component vector. This vector can be efficiently quantizated with only 64 centroids. Therefore, the prosodic database is formed by all linguistic-prosodic registers with quantizated PSP´s. The prosodic contour along an utterance can be seen as a sequence of quantizated patterns or centroids. In the next sub-section, we describe how the prosodic contour is obtained.

## 3.1. Generation of prosody.

For each syllable of the sentence to synthesize, we compute the linguistic features that are stored in the prosodic database. A first approach to generate prosody could be the following: for each syllable, we search in the database for a register with the same linguistic features and retrieve its prosodic features. The concatenation of these prosodic features for each syllable in synthesis would be the synthetic prosody. This is the basic idea but some aspects must be taken into account:

° It may happen that there would be no syllable with the same linguistic features in the database.
° It may happen that there would be more than one syllable with the same linguistic features.
° Some linguistic characteristics may be more relevant than others.
° It is important to consider the sequence of PSP´s in the generation of prosody because large variation of pitch values between syllables can be very harmful.

Therefore, for each syllable we select the most "linguistically" similar vectors (25 in the current implementation) according to an evaluation function of the linguistic features. More weight is given to the stress character and the type of breath group of the syllable than to the other features.

For selecting the best register for each syllable getting a good concatenation of PSP´s, we use the two prosodic features for vowel modeling. These features represent two consecutive PSP´s in the speech corpus. The first one is used for synthesizing the prosody of each syllable and the second one is used to evaluate how fine its register matches to the registers selected for the following syllable. Once we have the selected registers for each syllable, we have a matrix as shown in **Figure 3**. For each register we calculate a distance with all registers of the following syllable. The distance is an euclidean distance between the PSP2 of one syllable and the PSP1 of the following syllable (see **Figure 3**). We have to evaluate the sum of distances through all possible paths ($25^N$ where N is the number of syllables of the sentence). The prosodic contour is formed by the concatenation of PSP´s from registers with the minimal accumulated distance. To compute the best path through the matrix (marked in bold type in **Figure 3**), we use a dynamic-programming search procedure in order to optimize the computing time.

# 4. METHODS FOR REDUCING THE SIZE OF THE PROSODIC DATABASE FOR SYNTHESIS

In the described automatic methodology, the number of registers of the prosodic database is equal to the number of syllables of the prosodic corpus. A large speech corpus is desirable to model all possible combinations of linguistic parameters. However, a large database implies high memory and fast processor requirements for real-time generation, so a reduction in the size of the database keeping its prosodic richness is necessary. In our system, the prosodic richness is represented by the variety of PSP´s.

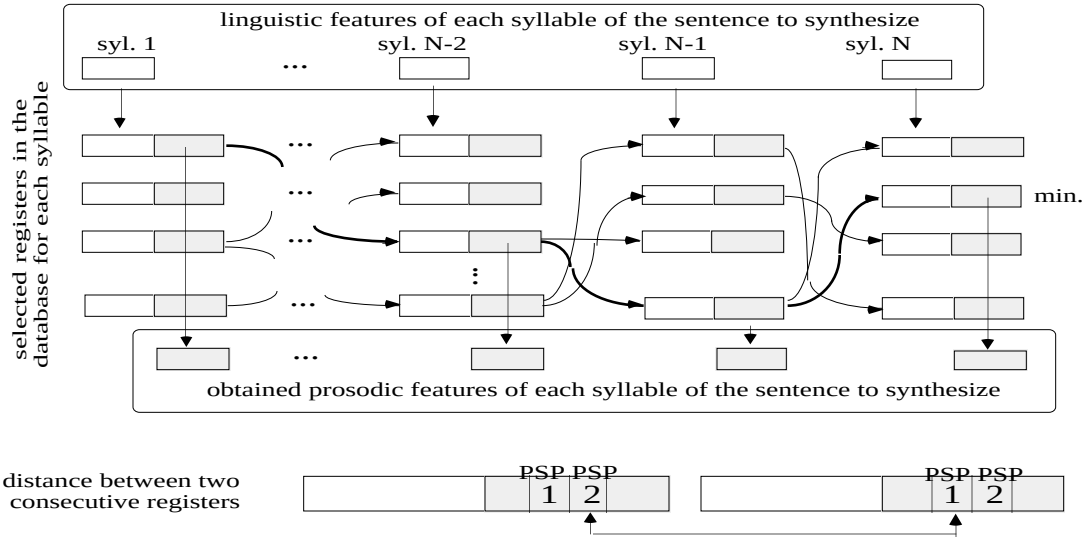


**Figure 3:** Generation of prosody by a dynamic programming search and distance between two consecutive registers.

For this purpose, first we group all registers according to their linguistic features used in selection of registers (see 3.1) and then an algorithm is applied to reduce the number of registers in the same proportion for each group. The algorithm eliminates the registers whose PSP´s are close to the PSP´s of other registers in the same group. Two PSP´s are close if their euclidean distance is small. The distance between registers is a weighted sum of the distances between their PSP1´s and PSP2´s. We attempt to use several reducing algorithms, but the best performance was obtained with the following one:

1. We calculate the distances among all pairs of registers in a group, and choose the pair with the minimal distance.
2. We eliminate one of the registers: the one that has been preserved less times in previous iterations.
3. If number of "survivors" is more than desired, iterate to 1.

This algorithm tries to keep the registers that cover all space, eliminating registers close to each other. This is represented in **Figure 4**.

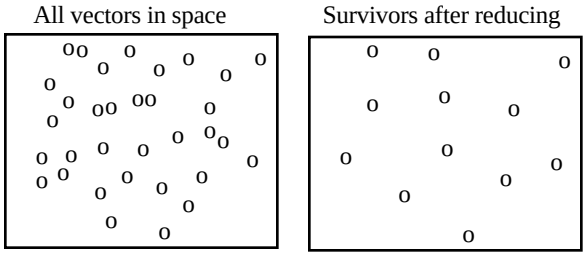


**Figure 4**. Results of the reducing algorithm in a 2-D space.

We have obtained the best performance with a distance that assigns more weight to the PSP2 than to the PSP1; in other words, the best is to keep the diversity in the PSP2, in this way, there is a wider range of possible transitions between registers than in other options.

We have also tried other methods, averaging the components of the compared vectors in step 2, but a worse covering of the space is obtained as a result.

## 5. RESULTS AND CONCLUSIONS

For this methodology with the reducing algorithm, we have used the same designed corpus that we had already used in [3] and [1]. This has allowed us to compare the results. This corpus was recorded at CNET (France Telecom) from a selected speaker. It has 144 sentences, 311 breath groups, 818 accent groups, 3500 vowels, 1700 consonants (including semivowels) and 166 internal pauses.

For evaluation of the reducing algorithm, we designed a test corpus of 21 sentences that was synthesized using HISPAVOC (TTS in Spanish, with the CNET PSOLA/TD[1] synthesizer). We have obtained very good results with a reduction of a 50% of the registers in the database. Here, you can get some links to sound files for one of the sentences:

° synthetic prosody [SOUND A244S01.WAV] with 100% of registers
° synthetic prosody [SOUND A244S02.WAV] with 50% of registers
° synthetic prosody [SOUND A244S03.WAV] with 10% of registers

It can be noted that a reduction of 50% of the database does hardly degrade the quality obtained with the whole database. However, a reduction of 90% shows a worse performance. The size of the original database was not too large (3500 registers). We expect to improve the quality using a larger corpus but keeping the same final number of registers.

This methodology has the advantage that can be easily adapted to a specific corpus or a particular speaker. In the future, we want to develop a better modeling of the relationship between prosody and syntax. Using the proposed automatic methodology with the reducing algorithm over a larger speech corpus, we expect to improve the location and quantization of these prosodic boundaries.

We want also to use the automatic methodology for integration of prosody in speech recognizers and speaker identification systems.

## 6. ACKNOWLEDGMENTS

We would like to acknowledge the work of the speech synthesis group at CNET in collaboration with the phonetic laboratory group at UAB for the recording of the prosodic corpus. We want also to acknowledge to Christel Sorin for allowing us to use the CNET synthesizer.

## 7. REFERENCES

1. E. López-Gonzalo and L.A. Hernández-Gómez "Automatic Data-Driven Prosodic for Text to Speech" in *Proc. EUROSPEECH* pp. I-585 I-588. Madrid (SPAIN). Sep. 1995.
2. F. Emerard et. al. "Prosodic processing in a TTS synthesis system using a database and learning procedures" in *Talking Machines: Theories, Models and Applications* Editors G. Bailly and C. Benoit. Elsevier 1992.
3. E. López-Gonzalo and L.A. Hernández-Gómez "Data-driven Joint $F_0$ and Duration Modeling in Text to Speech Conversion for Spanish" in *Proc. ICASSP*, pp. I-589 I-592. Adelaide (AUSTRALIA). Mar. 1994.
4. E. López Gonzalo . "Técnicas de procesado linguístico-prosódico y acústico para conversión texto-voz mediante concatenación de unidades" *Doctoral Thesis*. Universidad Politécnica de Madrid. Jul. 1993.
5. J. ´t Hart, R. Collier and A. Cohen "A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody" *Cambridge, UK: Cambridge Univ. Press,* 1990.
6. C.W. Wightman and M. Ostendorf "Automatic Labelling of Prosodic Patterns" in *IEEE Trans. on Speech and Audio Processing* pp. 469-481. Volumen 2, Number 4. Oct. 1994.

**1** PSOLA/TD is a registered trademark of France Telecom/CNET. Patent granted in France, Europe, Canada and USA, pending in other countries.