Prosodic Modelling for a Text-to-Speech System in Spanish

Eduardo López-Gonzalo and Luis A. Hernández-Gómez.
Dpt. SSR, E.T.S.I.Telecomunicación. Univ. Politécnica de Madrid.
Ciudad Universitaria s/n, 28040 MADRID.
e-mail: eduardo@gaps.ssr.upm.es luis@gaps.ssr.upm.es

ABSTRACT

In this contribution, we present the results in modeling Spanish prosody for a text-to-speech system based on units concatenation using a TD-PSOLA synthesizer. For this purpose, we have developed a new methodology to transfer to a text-to-speech system the prosody of one speaker considering both fundamental frequency and duration jointly, trying to take into account their interactions. The results on a corpus test over the system shows very good intelligibility and naturalness.

INTRODUCTION

Our aim was to produce a prosodic model of the speaker who recorded the acoustic database for a TD-PSOLA (F. Emerard et. al. 1992) synthesizer trying to capture all his characteristic features. For this task we designed a data-driven methodology that is shown in Figure 1.

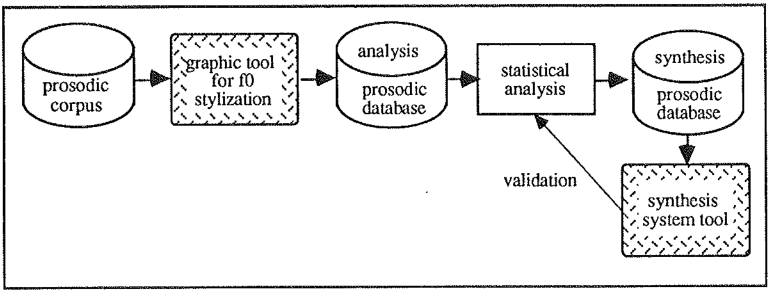

Figure 1. Methodology for prosodic modeling.

This methodology is guided by the main objective of producing a model of natural prosody allowing its artificial simulation. This is a more delimited task that prosodic modeling of the Spanish language but we think it is sufficient in the context of high quality text-to-speech synthesis. The rest of the paper is organized to describe this methodology.

CORPUS DESIGN

First of all, we designed a prosodic structure for Spanish that is an adaptation to the one proposed by V. Aubergé (1990). Based on a first analysis of our data and on the general linguistic knowledge of Spanish we considered two prosodic units: the "prosodic word" that is defined as the sequence of syllables associated with one accent and the "prosodic proposition" that is defined by the sequence of syllables between two pauses (regardless of its duration).

In order to constitute an organized corpus of minimal pairs of attributes, we defined a set of attributes for the two prosodic units. The "prosodic proposition" was characterized by its number of syllables, its number of prosodic words and its prosodic type depending on the syntactic relation between consecutive propositions (9 defined types, two of them complementary to put a contrast prosody in consecutive propositions of the same type). The "prosodic words" were defined by three attributes: the number of syllables, the position of the accent (3 classes) and the ordinal position inside a proposition.

This special corpus should take into account all the different types of propositions formed by all different prosodic words (V. Aubergé 1992). In order to limit the number of sentences to analyze, we impose some restrictions to emphasize the more important characteristics to be modeled. We mainly take into account the position of first and last accent of propositions, and study relations in sentences with only two propositions. Relations between sentences in a paragraph were not considered. The other parameters were kept as broad as possible. The resulting corpus was 144 sentences long, ranging from very short sentences to very large sentences taken from newspapers.

CORPUS STYLIZATION

This corpus was recorded at the CNET by a selected Spanish native speaker to produce a linguistic procesor for ELAN INFORMATIQUE. Then we analyzed it in order to extract for each sentence the pitch contour with marks representing the location of the vowels. For this task, we programmed a variation of a super resolution pitch detector (Y. Medan et.al. 1991).

This representation is known to contain many irrelevant information that is filtered by our auditory system, so we "stylized" these contours (R. Collier 1990) in such a way we keep only pitch movements that are perceived by a common listener. These movements are described following their direction (rise or fall), their slope in semitones per ms, extension in time (one or more syllables) and timing respect the tonic vowel. A frequency-based acoustic module elaborated for the synthesizer (E. Rodríguez, E. López and C. García 1993) was used in an analysis-synthesis mode with a graphical tool developed for the stylization of contours. The tool shows the temporal evolution of the original pitch in a semilogaritmic scale, and let you approximate it by straight lines (Figure 2). Then the acoustic algorithm is used to produce the original speech with the new pitch contour. Whenever a difference is perceived in prosody, we make a better approximation of the pitch contour. The output file of this module codes the fundamental frequency contour and duration aligned with the phonetic transcription of the sentence. The representation of the pitch contour in a vowel is given with a resolution of three points per vowel by means of 5 parameters: the initial, internal and final fundamental frequency value in the vowel and two time durations.

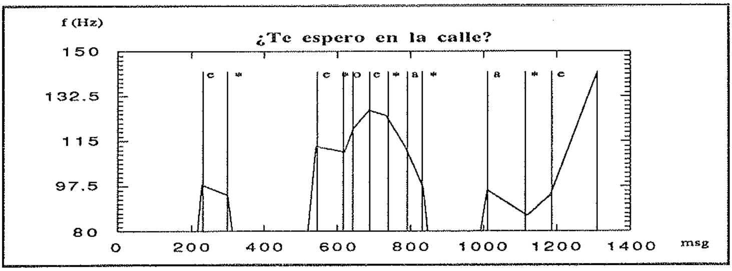

Figure 2. Presentation format of the graphical tool for a stylized contour.

CORPUS ANALYSIS

The result of the stylization process is put in a prosodic database specially designed to measure statistically the real consistency of the suprasegmental data produced by the speaker. A total of 23 fields were needed, the resultant structure is similar to the one proposed by Emerard et. al. (1992) but adapted to our prosodic structure. The database has 3500 vowels, 1700 consonants (including semivowels) and 166 pauses. The duration is stored for the consonants and the pauses. The stylized movement of pitch and the normalized duration is stored for each vowel. Besides that we store all attributes for accessing the prosodic structure (defined above) and phonetic context. The normalization is calculated dividing the duration of each vowel by the mean value of this type of vowel (5 classes in Spanish). Following this approach the data stored have been produced by the speaker. We make a statistical analysis between consistent prosodic words in a certain kind of proposition in order to minimize the interaction between segmental and suprasegmental parameters. The results of this analysis showed us a very good consistency to model absolute frequency contours and normalized suprasegmental duration in each class. As an example in Figure 3, it is shown the last four syllables (a vowel per syllable) of the last prosodic word in a proposition at the end of a declarative sentence for the three positions of the accent considered (accent in the last vowel, accent in the last but one vowel or accent in the last but two vowel).

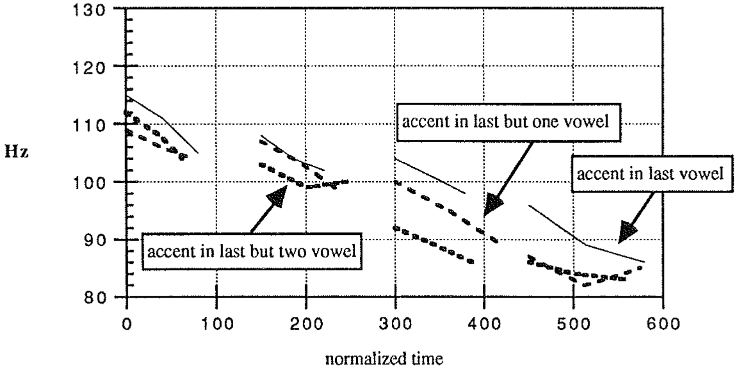

Figure 3. Pitch and duration modeling for the last prosodic word in a declarative sentence, considering the three positions of the accent.

It should be noted that in our model, pitch and duration are modeled in a joint way considering the effect of the position of the accent and position of the pause because of the defined prosodic structure. This model takes into account the possible interaction between duration and pitch in prosody realization. It is observed in Figure 3, that there is a residual effect of suprasegmental duration as predicted by other duration models with the position of the accent. It was also observed from the resultant model that there was a lengthening in the vowel located in a syllable previous to a pause.

For this first model, the duration of each consonant (and semivowel) was considered to be the mean value calculated from the database.

The mayor correlation in the duration of the pauses was found to be with the type of orthographic sign and the number of syllables of the proposition but there was a large dispersion from the mean value. A closer look to the data ordered by the duration value of the pause show us that the duration of the pauses had more to do with the syntactic structure of the sentence.

SYNTHESIS TOOL

In order to validate and refine the model, it seemed sensible to us to use a text-to-speech system as a tool for leading our analysis in the database.

For this purpose, we developed a very flexible text-to-speech system (E. López 1993). The input text is parsed to mark the "prosodic words" and the major “prosodic propositions” based in some independence coefficients between two consecutive words that take into account both rhythmical and syntactical constrains. The same procedure has been integrated in TTS systems for Spanish (E. Rodríguez et.al., E. López et. al. 1993).

The prosody calculation is realized by a look-up in the synthesis database containing the results of the statistical analysis explained before. We choose this data based approach for being more flexible that a rule based system so we could test quickly our hypothesis. This original contour is then post processed with an interpolation of pitch contours between consecutive vowels, and a micromelody for voiced consonants and semivowels is added. The length of the pauses are calculated in function of the independence coefficient.

The model was tested in an implementation of a TTS system based in diphone concatenation with a TD-PSOLA algorithm. We designed a small corpus of isolated sentences considering two types of propositions with all different prosodic words. The naturalness was good recognizing perfectly the peculiar "style" of the speaker, although it was noted some deficiencies in segmental duration of consonant clusters.

For this reason we calculated from the analysis database a better segmental model of consonant duration based in some multiplicative coefficients as a function of phonetic context. This new model was considered in the synthesis tool improving the naturalness of the model.

As a future research, we are going to follow the same methodology to adapt the independence coefficients and consequently produce a better modeling of pauses in large sentences or even paragraphs. We are presently studying a syntactic theory developed by J. Vergne (1992) that can be adapted to our TTS system in order to compute better coefficients.

CONCLUSIONS

We have discussed a data-driven methodology that has been shown very successful to transplant the prosody of a speaker into a text-to-speech system. The complementation between statistical analysis and a synthesis tool was found to be very useful to improve the model.

REFERENCES

F. Emerard et. al. (1992) "Prosodic processing in a text-to-speech synthesis system using a database and learning procedures" in Talking Machines: Theories, Models and Applications" Elsevier 1992

V. Aubergé (1990) "Semi-automatic constitution of a prosodic contour lexicons for the text-to-speech synthesis" Proceedings in ESCA Workshop on Speech Synthesis. Autrans 1990.

E. López Gonzalo (1993). "Técnicas de procesado linguístico-prosódico y acústico para conversión texto-voz mediante concatenación de unidades" Doctoral Thesis. Universidad Politécnica de Madrid. (In preparation).

E. López-Gonzalo, G. Olaszy and G. Nemeth. "Improvements of The Spanish Version of the Multivox Text-To-Speech System". Accepted for publication in Eurospeech 1993. Berlin.

Y. Medan, E. Yair and D. Chazan (1991). "Super resolution pitch determination of speech signals" IEEE Transactions on Signal Processing, vol. 39, nº1. January 1991.

R. Collier "Multi-Lingual Intonation Synthesis: principles and applications" Proceedings in ESCA Workshop on Speech Synthesis. Autrans 1990

G. Bailly "Integration of rhythmic and syntactic constraints in a model of generation of French prosody" Speech Communication 8 1989 pp. 137-146.

E. Rodríguez-Barga, E. López-Gonzalo and C. García-Mateo (1993)."A Text-to-Speech System for Spanish with a Frequency Domain Based Prosodic Modification" International Conference on Acoustics Speech and Signal Processing" (ICASSP) Minneapolis (USA). 1993.

J. Vergne. "Syntax as clipping blocks: structures, algorithms and rules". Jornadas de la Sociedad Española para el Procesado del Lenguaje Natural (SEPLN), Granada 1992.