

IMPROVING TEXT TO SPEECH CONVERSION IN SPANISH: LINGUISTIC ANALYSIS AND PROSODY

J.M.Pardo (+), M.Martínez (*), A.Quilis (*) and E.Muñoz (+)

ABSTRACT

In order to improve the naturalness of a text to speech converter for Spanish, a study has been carried out to establish the possibility of designing an algorithm that generates automatically realistic pauses emulating the mechanism that a speaker uses to read a text. Together with it a study on the intonation for Spanish has been carried out to devise a good pitch generator. In this paper an interim report about an automatic pause generator and a pitch generator that we are currently developing is described. The pause generator is based on an automatic word labeler and empirical rules obtained from a database analysis.

INTRODUCTION

Some research efforts are being currently made to implement a high quality text to speech converter for Spanish (ref 1). Although intelligibility of our current systems is as high as intelligibility of real speech, the naturalness of the spoken sentences is still poor. This is due mainly to the simple prosodic algorithms used that do not take into account the variability of the speech according to the syntactic and semantic content of a sentence and the musicality of a human speaker.

The generation of synthetic speech with a good prosody is difficult because the prosody is very dependent on the speaker, emotions, the linguistic content of the sentence, etc... Breath groups are not always bounded by punctuation marks. Moreover, pause duration and pitch patterns do not depend exclusively on punctuation marks. On the other hand, punctuation marks do not always generate a pause nor a pitch change. To generate a good prosody, a syntactic analysis of the sentence is needed. To improve prosody in a preliminary version, a rough syntactic analysis can be used.

In a first phase, we have analyzed a five minutes long text read by ten speakers, totaling an amount of 1127 breath groups. We have studied the pauses which appear without punctuation marks and compared them to the total number of pauses. We have measured the duration of these pauses and breath groups. For every breath group we also measured pitch values and annotated the syntactic patterns of breath groups

(+) Departamento de Ingeniería Electrónica, ETSI Telecomunicación, Ciudad Universitaria, 28040 Madrid, Spain.

(*) Departamento de Lengua Española, Universidad Nacional de Educación a Distancia, Ciudad Universitaria, 28040 Madrid, Spain

without punctuation marks.

Based on the analysis of the above data, we have developed an algorithm to parse the text and locate realistic breath groups automatically. The algorithm calculates first both the number of syllables of the sentence and the category of each word.

From a first empirical analysis of the database we concluded that a complete set of categories was not needed in 75% of the cases. Only a superset of eight word categories, including categories related to verbs and function words, was used in the algorithm.

AUTOMATIC WORD-CLASS LABELER

An automatic word-class labeler has been developed. It automatically classifies the text into one of 34 different categories. Each category represents a part of speech in most cases. However, there are categories that represent a set of parts of speech when necessary. A reduced morphological analysis is performed in a first stage. A 600-word dictionary of exceptions, function words and expressions is included in the system. It uses also explicit knowledge through heuristic rules. For this task, a proprietary syntax for rule writing and a rule compiler has been explicitly designed and used in the algorithm. With this strategy, a linguist can easily check and/or extend the rules so that the word labeler can be continuously refined to improve performance. An example of the syntax of a rule follows:

```
arpr > art /pre,ver,atr +/
```

which can be read as " An ambiguous word that can be an article or a personal pronoun is always an article if it is preceded either by a preposition, an attributive verb or a non-attributive verb (with the exception of verbs "haber" and "ser"). In this example "arpr" stands for the ambiguous word, "art" for an article, "pre" for a preposition, "atr" for an attributive verb and "ver" for a non attributive verb with the exception of "haber" and "ser". The existence of contextual rules implicitly means that a limited syntactic knowledge is included in the categorizer.

Preliminary results of the word categorizer show that from a total of 9963 words tested, 9.8% of the words were considered ambiguous and consequently not categorized, and 1% of the words were incorrectly classified.

AUTOMATIC PAUSE GENERATOR

The analysis of our data shows that 36% of the pauses in the text appear without punctuation marks. This is the reason why a pause generator that inserts automatically pauses in the text has been created. The pause generator (or breath group parser) is based on empirical analysis and heuristic rules. It uses the number of syllables in the text (an algo-

rhythm to count the number of syllables in a text has been also developed) and the category of the word (one out of eight which are supersets of the 34 different categories delivered by the labeler). We are improving the generator by comparing the current output to the analysis of our database.

The generated pauses are classified according to the context where they appear (punctuation mark, no punctuation mark, syntactic content of the sentence, etc.). Different durations are assigned to different groups of pauses (when converting text to speech) based on the data obtained from the analysis. Some of the data obtained and used to design the generator are :

1. The number of phonologic syllables for 75% of the breath groups spans between 3 and 16.
2. The number of stressed syllables for 80% of the breath groups spans between 1 and 4.
3. There are no breath groups with more than 15 stressed syllables.

PITCH GENERATOR

Our work consists of analyzing the discrete elements of the intonation (in this paper we consider pitch and intonation to be the same) in order to quantify and formalize its substance in the form of rules that can be programmed into a text to speech converter.

The difficulty of our work is closely tied to the degree of intonation arbitrariness (much bigger than that of the phonemes). This implies that a lot of work is needed to establish distinctive units and quantify their characteristics (ref 2). We will not discuss here the adequacy of using analysis of levels or analysis of configurations or how the intonation units are correlated with syntactic characteristics (for the time being). Our work has been pragmatic: we are creating an algorithm for text to speech conversion but we hope to use our data to create a theory about intonation in Spanish.

For about 700 breath groups we have collected the following data:

1. Syntactic class of the breath group.
2. Punctuation mark at the end of the group.
3. Total number of syllables and stressed syllables for each breath group.
4. Pitch value for the first and last syllables and for every stressed syllable.
5. Pitch value for some unstressed syllables.
6. Pitch variation of intonemes.

There is a correlation between the kind of pause, the pause duration and the intoneme. Long pauses are correlated to descending intonemes. Short pauses are correlated to ascending intonemes. The reason for this behavior has a syntactic or semantic origin, because the group of speakers is homoge-

neous and the text is the same.

The characterization of the intoneme is different depending on the position of the stress in the word. In non-monosyllabic words the most frequent scheme for the stress the second starting from the final syllable. This makes Spanish different to French, were the most frequent words have the stress in the last syllable (ref 3 and 4).

We have grouped the breath groups into sets depending on the number of stressed syllables and the speaker. We have made statistics on the average pitch values and standard deviations for the different sets. In Fig 1 an example of the statistics made is shown. The average and typical deviation of pitch values are represented as a function of the number of stressed syllables for a set of 154 breath groups with 3 stressed syllables each. Pitch values are normalized by a multiplicative factor so that in every breath group the first stressed syllable takes on the value of 100 units.

1. J.C.Olabe, A.Santos, R.Martinez, E. Muñoz, M.Martinez, A.Quilis and J.Bernstein, "Real time text-to-speech conversion system for Spanish", Proc. of the ICASSP 84, pp 2.10.1-2.10.3, San Diego, March 1984
2. A. Quilis, "Las unidades de entonación", Revista española de lingüística, n. 5, Madrid 1975, pp 261-280.
3. A. Quilis, "Frecuencia de los esquemas acentuales en español" en Estudios ofrecidos a Emilio Alarcos Llorach V Oviedo 1983, pp 113-126.
4. A. Di Cristo, M.Chafcouloff, "L'intoneme progrédient en Français: Caractéristiques intrinseques et extrinseques" in Pierre L and Rossi M. (eds) Problèmes de Prosodie Vol II, Ottawa 1979, Didier pp39-51.

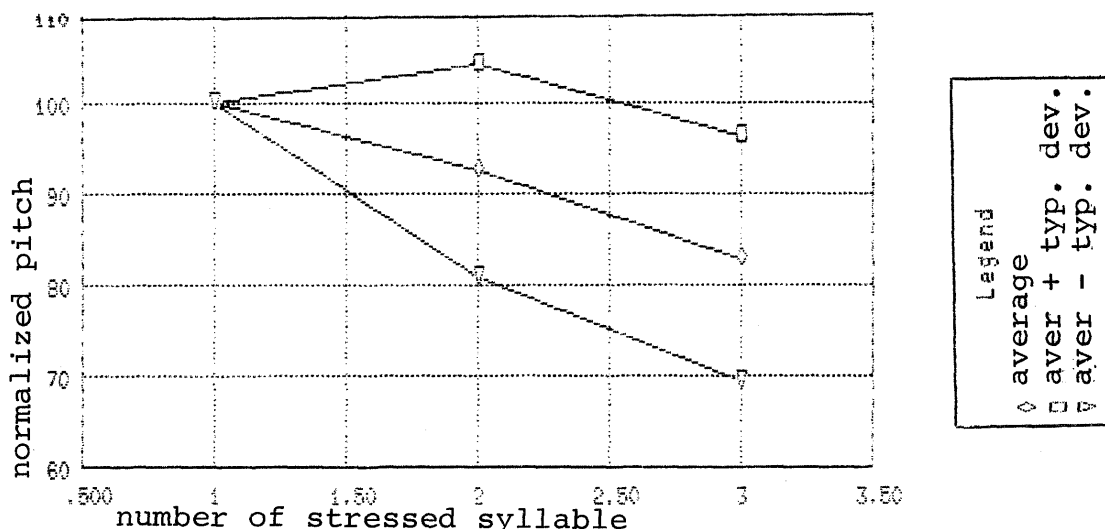


Fig 1. Average and typical deviation of pitch values as a function of the number of stressed syllables for a set of 154 breath groups with 3 stressed syllables. Pitch values are normalized by a multiplicative factor so the 1st stressed syllable has a value of 100 units.