High quality Text-to-Speech System in Spanish for handicapped people

Fernando Lacunza, Yolanda Blanco

Departamento de Ingeniería Eléctrica y Electrónica
ETSII Y IT, Universidad Pública de Navarra
Campus de Arrosadia s/n, 31006 Pamplona - SPAIN

ABSTRACT

This paper describes a high-quality Text-to-Speech system based on the concatenation of diphonemes with the MBR-PSOLA algorithm. Since it was designed as a substitute of natural voice for handicapped people, it must offer a easy to hear speech, with emotional and emphatic information embedded in it. This is obtained with the prosody generator, which uses phonological patterns and a grammatical database to vary three speech parameters: pitch, amplitude and duration. This system accepts plain text, which can be complemented with data about emotions and emphasis.

1. INTRODUCTION

The project commented in this article is a part of the SIVHA (*Síntesis Visual del Habla* – Visual Speech Synthesis) project. It was born to provide help to people who have lost motor abilities, including speech. Being a substitute of natural voice as it is, it must offer a high-quality speech, with good prosody, and give information about emotions and enphasis.

During last years, speech synthesis systems have reached a high degree of intelligibility. Nevertheless, the introduction of prosody int synthesized speech is still a matter under study. So, this is where this project is more open to further improvements.

2. CONSIDERATIONS ABOUT PROSODY

The parameters used to generate prosody in this work are commented below.

2.1. Pitch Profiles

The *phonic group* is defined as the portion of speech between two pauses [1]. In Spanish, phonic groups tend to follow different patterns of pitch. These patterns depend on the type of sentence they represent. In this work we have considered 6 patterns: complete declarative sentence, absolute interrogation, relative interrogation, anticadence sentence, suspension sentence and exclamation.

Although in casual speech the type of sentence does not always correspond to its pattern, sentences with non-matching patterns are rare. Below, the use of each pattern and its shape are explained:

- *Declarative sentence pattern:* it has an upward slope of fundamental frequency (F0) until the first stressed syllable, where it reaches the maximum. Then it goes downward until the end of sentence. This is the most usual pattern. It is applied to declarative sentences without pauses, and to the last phonic groupo of a declarative sentence.

- *Unfinished declarative sentence patterns:* The initial segment is similar to the pattern above. The main difference between these and the complete declarative sentence pattern is the segment from the last stressed syllable to the end of phonic group. In this patterns, fundamental frequency suffers a positive slope (anticadence pattern) or keeps constant (suspension). Both are used in phonic groups ended by an intermediate pause. Anticadence patterns are used in the phonic group before the last intermediate pause, and suspension

patterns are used in the other phonic groups. These are also used in sentences ended in suspension (*Como no vengas...*).

- *Absolute interrogative sentence pattern:* The initial segment until the first stressed syllable has an upward slope, with a higher peak than declarative sentence patterns. Its is followed by a segment with negative slope until just before the last stressed syllable. Final segment has positive slope. This is the pattern for absolute questions, that is, questions that can be answered by 'yes or 'no'. Nevertheless, it can be used in other kind of questions, although this is not usual.

- *Relative interrogative sentence pattern:* The only difference with the one above is its final segment, which has a downward slope. It is used in relative questions and, less frequently, in absolute questions, specially when the speaker knows or suspects the answer

- *Exclamative sentence pattern:* It can be considered a variation of the declarative sentence one. The shape is the same, but the range between maxima and minima in F0 is wider.

There is a phenomenon that affects all patterns: The mean of F0 decreases from the first stressed syllable. This is known as *declination*, and it is one of the caracteristics of spoken Spanish [2]. Declination is also observed along paragraphs [3].

2.2. Stresses in Spanish

Every word in Spanish has stress. A speech synthesis system must recognize and represent them. But how is stress represented in spoken Spanish? In any language, stresses affect one or more of these three parameters: amplitude, pitch and duration of speech sounds. So, in English, stress is mainly expressed as a variation of amplitude, while in French, the most important parameter is duration.

Although traditionally amplitude is the parameter which determines stress in Spanish, recent studies have observed relationships between stress an variations of F0 and duration of phonemes. It can be said that a stressed syllable in Spanish has a greater amplitude of sound, a positive slope in F0 along all or nearly all the syllable, and an increase in its duration [3].

In Spanish, words have only one stress. The only exception is the adverbs composed by an adjective and the suffix "*-mente*". This words have two stresses: one belongs to the adjective and the other is in the first syllable of the suffix.

Although every word has stress in Spanish, this is only true for isolated words in spoken Spanish. In speech, some words lose their stress. Below are listed the types of words that lose it:

- Definite articles: *El vestido de la novia.*
- Prepositions: *De mi pueblo hasta mi casa hay un trecho.*
- Conjunctions: *Vino y me dijo que todavía ni lo había pensado.*
- First element in compound numerals: *Diez mil, cincuenta y dos.*
- Unstressed pronouns: *Te dije que se iba a caer.*
- Apocopated possesive adjective: *Mi hermano tiene su propio coche.*
- The words *que, cual, quien, donde, cuando, cuanto, como* when they are not used as interrogative particles: *Te reto a una partida donde sea cuando sea.*

So, a speech synthesis system for Spanish must identify words in order to put stresses when needed.

2.3. Duration of Phonemes

In spoken Spanish the duration of every phoneme depends on its type and where it is placed. Considering the type of phoneme, it can be said that strong vowels ('a', 'e', 'o') use to be longer than weak ones ('i', 'u'), and fricative consonants are longer than occlusive ones, for example. Considering the position of the phoneme, its duration is increased if it is in a stressed syllable, or is decreased if it belongs to a diphthong. Several rules of duration are applied in this manner, obtaining a factor that is

multiplied by a value that indicates the mean duration of phoneme. This parameter can be varied by the user to modify the speed of speech.

2. DESCRIPTION OF THE SYSTEM

The system converts plain text into speech. As an option, with the text we can add information about enphasized words or emotions(like anger or sadness). After being pre-processed in a conditioning phase, text is sent to two modules: one makes the conversion to a phonetic format and the other makes a simple morphosyntactic analysis. Information from both modules is gathered in the prosody generator. It obtains the list of phonemes to be spoken with a series of parameters of pitch, duration and amplitude associated. All these data are sent to the speech synthesizer. The diagram of the system is represented in Figure 1.

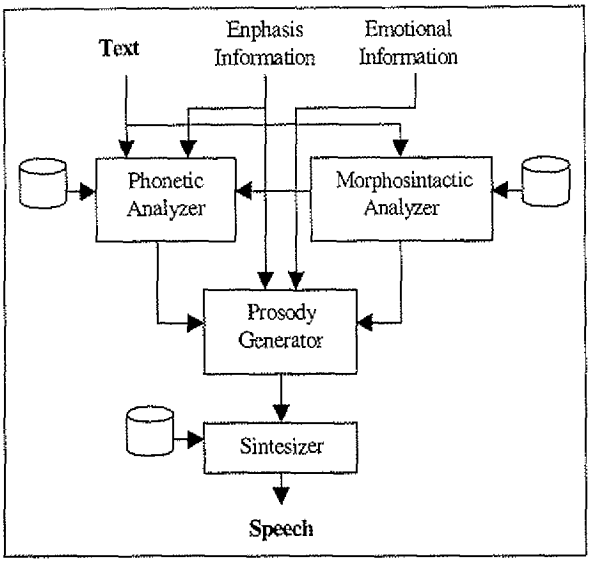

Figure 1: Block diagram showing the Text-to-Speech system elements

Next, a description of each module will be made, indicating its relations with the other modules.

2.1. Text Pre-Processing

In this phase abbreviations and numbers are replaced by plain text. It considers special formats like dates and telephone numbers. All text is made lowercase and divided in sentences. These sentences are sent to the phonetic and morphosyntactic analyzers.

2.2. Phonetic Analysis

In this module a phonetic transcription of the sentence is obtained. This transcription uses the SAMPA phonetic alphabet, with some slight modifications.

Is in this module where stressing rules for Spanish are applied. To make this, it is necessary to make a previous division in syllables and consult a database to recognize the unstressed words in speech.

Besides stressing, this module also stablishes a duration for every phoneme, following rules related to its position and type. To every duration a small random variation is added, to simulate the slight variations of speed we involuntarily make in speech. The output of this module goes to the prosody generator.

2.3. Morphosyntactic Analysis

The first step is classifying each sentence, determining which type it belongs to (declarative, exclamative, interrogative –absolute or relative-). If the sentence is compound or has enumerations, it is divided into its constituyent elements. For each element the prosody generator will choose the appropiate pitch pattern. Then a database is consulted to apply a label to each word with a grammar cathegory, gender, person and plurality. This label gives information to make a pseudo-syntactic analysis of each element of the sentence.

The last step is dividing the fragments of the sentence that are too long to be spoken without pauses. Now each fragment can be considered as a phonic group.

2.4. Prosody Generation

The prosody generator compiles information obtained by the two analyzers, with data about enphasis and emotions, given by the user.

First, it chooses a pitch pattern for every phonic group. Each pattern has two baselines of values, one showing the peak values and the other showing the valley values of normalized F0. These lines are composed of straight segments, whose junctions are placed in key points in the phonic group, like the first or last stressed syllables.

Once the pattern is obtained, it is denormalized, adapting it to the duration of the phonic group. Emotional information is now used to modify the duration of the phonemes and to denormalize the patterns in frequency. The value of junction points is modified with a small random quantity. This random value avoids the repetition of exact patterns, because a human speaker does not repeat intonations from one sentence to another.

Next step is adapting the phonetic transcription to the pattern, considering the existing stresses. Each stressed syllable will reflect in speech as a positive slope in F0, from the valleys baseline to the peaks baseline, and also as an increase in amplitud and duration of its phonemes

We obtain in this manner the evolution of F0 in speech along the phonic group.

This module also obtains the evolution of amplitude along the phonic group. This is function of the position of stresses and the emotional information.

Finally, some minor adjustments are made on the three parameters in some words of the phonic group. According to its grammar category and enphasis information. All the data obtained in this module goes to the synthesizer in the form of sequences of phonems with associated values of duration, F0 and amplitude.

2.5 Speech Synthesis

The adopted method in this system for speech synthesis is the concatenation of diphonemes, using the MBR-PSOLA algorithm. It uses a database with 676 diphonemes made up of 25 phonemes, normalized in F0 and phase, from a male speaker.

3. RESULTS

A system that converts plain text to speech is the result of this work. It also accepts additional information about enphasis and emotions. The system takes care of obtaining a prosody as natural as posible.

This system obtains a speech whose naturalness is convincing and does not bore the hearer, although it is still somewhat strange. A more powerful syntactic analyzer would increase the quality in the division of phonic groups, specially in long sentences.

About emotional information, it is still limited by now. When the sentence has a neutral emotional meaning, the percentage of errors in recognizing emotions is quite high.

4. REFERENCES

1 Quilis A., Fernández J. A., *Curso de fonética y fonología españolas*, pp. 143-144, 158-160, 163 (1990).

2 Garrido J. M., *Modelización de patrones melódicos del español para la síntesis y reconocimiento del habla*, pp. 94-114 (1991).

3 Garrido J. M., "Modelling Spanish Intonation for Text-to Speech Applications" *Tesis doctoral* (1996).

4 Cahn J. E., "Generating Expression in Synthesized Speech" *Master's Thesis*, Massachusetts Institute of Technology (1989).

5 Dutoit T., Pagel V., Pierret N., Bataille F., Van der Vrecken O., "The MBROLA Project: Towards a Set of High-Quality SpeechSynthesizers Free of Use for Non-Commercial Purposes" *ICSLP'96 Philadelphia*, vol. 3, pp. 1393-1396 (1996).

6 Martí J., Gudayol F., "El ritmo y la entonación en la lectura del castellano" *IX encuentro de la SEPLN de Santiago de Compostela*, p. 278 (1993).