A BILINGUAL TEXT-TO-SPEECH SYSTEM IN SPANISH AND CATALAN

Antonio Bonafonte, Ignasi Esquerra, Albert Febrer, Francesc Vallverdu
{antonio,ignasi,alfebrer,sisco}@gps.tsc.upc.es

Universitat Politècnica de Catalunya
C/Jordi Girona 1-3
08034 Barcelona, SPAIN

ABSTRACT

This paper summarises the text-to-speech system that has been developed during the last years in the Speech Group of the *Universitat Politècnica de Catalunya* (UPC). The paper emphasises the parts of the system which are language dependent: phonetic transcription, prosodic module, and synthesis units database. One particularity of the system is the fact of being bilingual, i.e., the system is able to speak either in Spanish or in Catalan. Some effort has been done to allow the reading of bilingual texts and to reduce the computational resources needed. In particular, the Spanish and Catalan speech databases are merged to reduce the memory requirements and the development effort. The system is being used by disabled people which suffer from oral disorders. In order to give variability to the voices some experiments have been done in voice transformation using the TD-PSOLA algorithm.

1. INTRODUCTION AND SYSTEM OVERVIEW

Text-to-Speech (TTS) systems produce synthetic voice from unrestricted input text. They are specially relevant for those applications where voice is the most appropriated support (or even the unique possible) for the information. For instance, in telephone applications voice is the medium that have to support the information addressed to people. Other purpose of such systems is to provide people with oral disorders the ability of speaking, using a machine, for communicating with other people.

In this paper we present the Text-to-Speech system developed at the *Universitat Politècnica de Catalunya* (UPC). It has been used in some telephone applications. However, in the following, some details will be explained about a version of the system which has been adapted to assist handicapped people. The work has been developed in co-ordination with the "*Instituto Municipal de Disminuidos de Barcelona*", which assist disabled people and provides equipment adapted to them.

Catalonia is a bilingual region of Spain, where Catalan and Spanish are spoken almost equally by a large part of the population. These two languages are very similar in their set of phonemes. The objective is to develop a bilingual TTS system sharing as much resources as possible, or with the minimum differences between language-dependent modules. Then, each user can choose the desired language or even produce speech in the two languages simultaneously, for instance for proper names.

This research was partially supported by the CICYT under contract TIC95-1022-C05-04

The TTS system is composed of four modules, as can be seen in figure 1:

- Text normalisation module (under development): expansion of acronyms, abbreviations, number sequences, time expressions and dates.
- Phonetic transcription module: conversion from letters to phonetic symbols.
- Prosody generation module: assignment of intonation and duration values to each phonetic symbol.
- Speech synthesiser: generation of the acoustic signal using a database of pre-recorded units.

The TTS system operates in different platforms (Unix, Windows, MacOs), and can be executed from standard word processors as *MS Word 6.0*. The user can configure several parameters of the TTS defining a "Voice" which has the following fields: language (Spanish or Catalan), synthesis-unit database and prosodic pattern file. Other prosodic parameters include the mean pitch value, the pitch range and the articulation speed.

In the following we will explain the different modules of the TTS system and the synthesis unit database. Most of the modules are language independent or can be adapted from a language to the other easily, given the similarities between the Spanish and Catalan languages. Furthermore, as most of the sounds used in Catalan are also present in Spanish, the databases can share a lot of units. In addition, the method used to efficiently label the database will be presented. Finally, some experiments on voice transformation are reported in order to get greater voice variability.

2. TRANSCRIPTION

After receiving the text from the editor, the first processing task consists in normalise the text transforming numbers and other non-readable symbols to its orthographic spelling. Abbreviations, acronyms and irregular words are looked up in an exception table, which can be edited by the user.

Afterwards, the full orthographic text is first segmented in syllables, then accented and finally converted to phonemes. The marks of syllabification and accentuation are of key importance for a correct prosody assignment. The set of rules applied in each of these steps depends obviously on the selected language.

For the Spanish language, a set of 31 allophones is considered [1] and 32 allophones for Catalan [2], taking into account that the sounds */ts/, /dz/, /tS/* and */dZ/* are considered as a combination of two phonemes. The main differences between the two sets are:

- Catalan has 8 vowels: the five Spanish ones (*/a/, /e/, /i/, /o/* and */u/*) plus */@/, /E/* and */O/* (SAMPA notation). Additionally, both sets include two semivowels: */j/* and */w/*

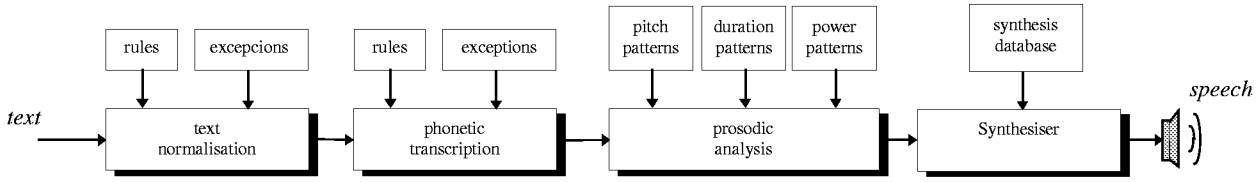

Figure 1: Text to speech system overview

- The Spanish consonants */jj/, /x/* and */T/* are not present in Catalan. However, in Catalan appears the sound */Z/* which does not appear in Spanish.

The Spanish transcription can be done applying simple deterministic rules [1]. In Catalan most of the allophones can be also determined using simple rules [2]. However, Catalan presents some ambiguity in the transcription that are difficult to solve. For instance, in some cases, the pairs of tonic vowels */e/-/E/* and */o/-/O/* can not be disambiguated by deterministic rules. In these cases, a dictionary is used for frequent words and a simple heuristic for the rest.

3. PROSODIC MODEL

This is the principal agent in obtaining a natural sounding quality of the synthetic speech. This module transform the result of the transcription into a string of allophones, each one having associated the values of pitch and duration.

3.1. Intonation

The intonation model assigns a pitch value to each allophone of the message. The intonation model is hierarchical, being the result of the interaction of different levels. At this moment, only the sentence level and the tonic-group level are used, but soon a paragraph level will be added to the model.

Sentence level: the basic intonation patterns considered are: declarative, imperative, interrogative and open (or not finished) sentences [3]. Different patterns exist depending on the number of syllables. Each pattern is composed by straight lines between inflection points and describes the evolution of the pitch along the time axis. Depending upon the number of stress accents, the patterns have one, three or more inflection points. The selection of the pattern is done based on punctuation marks. However, in the cases where very long sentences without punctuation marks are found, a simple heuristic is applied to extract open sentences.

The sentence patterns are represented in a parametric form so that the number of inflection points, position and frequency values of each point can be adjusted to model the intonation characteristics of each language and speaker.

Tonic Level: This level models prosodic variations inside the tonic group, which in Spanish and Catalan roughly correspond to one word plus optional preceding function words. At this moment the model increases a percentage (around 20%) the pitch value in the stressed vowel. Although detailed analysis reveals that the evolution of the pitch at the tonic level is more complicated [4], this simple model improves significantly the intelligibility and naturalness of the speech because it breaks the monotonous long patterns. The rhythm of the resulting speech is positively appreciated.

3.2. Duration

The duration model is composed of two parts: first a reference value for each allophone is assigned and afterwards a adjustment is performed depending of the position of the allophone in the sentence.

Reference value: two simple models have been tried, giving similar subjective quality. In the first one, each phone is assigned the mean value of the phone in the language (computed from speech databases). In the second one, the values of the synthesis units are taken as references. In the first case the statistic is computed from a lot of examples but is independent of the surrounding contexts. In the second one, as will be evidenced in section 5, the duration of each allophone is derived from a single case but depends on the contexts.

Vowel adjustment of the reference value: proper assignment of duration to vowels is of key importance to get natural-sounding synthetic speech. Our systems applies the rules derived in [5,6]. These rules modify the reference value of duration depending on several factors: prepausal/non-prepausal syllable position, stressed/unstressed vowel, syllable structure and voicing and manner of postvocalic consonant.

4. SYNTHESIS

A string of phonemes with their associate intonation and duration parameters is obtained from earlier modules. Synthesis is performed by concatenation of recorded units consisting mainly of diphones, plus some longer units, as will be seen on the next section. The TD-PSOLA algorithm [7] is used to adapt the characteristic of the stored units to the values assigned by the prosodic model. The frequency is linearly interpolated along each allophone.

5. DATABASE OF SYNTHESIS UNITS

The generation of a synthesis database consists of the following steps: determine the synthesis units, record the units inside some word or sentence and extract the unit from the utterance and label the unit. In this section the method followed to generate a bilingual database will be exposed.

5.1 Inventory of synthesis units.

The diphones (in fact di-allophones) are the main units of our database. In order to determine which diphones exist in each language, the following procedure has been followed.

- Get a large dictionary of words with the transcriptions.
- Determine the diphones intra-word.

- Determine the phones which appear at the beginning and at the end of any word and their count frequency. From these information the usual diphones inter-word are determined.
- Define the rest of inter-word diphones using some method for reducing the infrequent diphones.

If a phone occurs rarely at the beginning or at the end of the words, then it can be part of a lot of diphones (but rarely) in the junture of words. This is the reason why it is convinient to use some method to reduce the number of synthesis units. In such cases the rare phone has been recorded in the context of silence. Afterwards, if a rare diphone is requiered it is artificially on-line generated from two diphones. For instance, in Spanish the sound */x/* appears almost exclusively at the begining or inside of a word, and always preceding vowels. However, it can appear at the end of a word in a few words, as in "reloj" (clock): */rre-l'ox/*. Therefore, the diphone */x$/* is recorded (*$* means silence) and afterwards, diphones */xC/*, (being */C/* any consonantic phone) are artificially generated from */x$/* + */$C/* In Spanish this method allows an important save on the database because the ending of words is quite regular. However, in Catalan the number of endings is much more higher so the method is not so effective. Generation of diphones has also been applied to reduce the number of needed units for some phones preceding plosive consonants. For instance, */rt/* can be generated from */r$/* + */$t/* without degrading the quality.

To reduce a little more the number of synthesis units some substitutions are allowed for similar diphones: for instance diphtongh */wi/* is assimilated to */uj/*

The number of selected diphones is 511 for Spanish and 812 for Catalan. The higher number of Catalan diphones is due to the higher number of vowels and to the higher number of ending consonants which require a large number of inter-word diphones.

On the other hand, in order to account the influence of the vowel in preceding vibrant phones, some long units are added to the database, basically (plosive + {*/r/,/l/*} + vowel). This implies 105 units for Spanish and 150 for Catalan.

In order to have a bilingual system, the Spanish and Catalan synthesis units have to be uttered by the same speaker. We have selected speakers with Spanish and Catalan as mother tongue, so that they speak perfectly both languages without accent. As most phones are shared by Spanish and Catalan, the units can be merged so that the database is reduced. The chosen merging criteria is quite simple: if the same SAMPA notation is used for a Catalan and a Spanish allophone, then they are merged. In fact, some allophones are a little different depending on the language: for instance, the articulation point of Catalan */l/* is posterior to the Spanish one. However, we believe that they are quite similar when uttered by the same speaker. If the speaker has not accent, the result is quite acceptable. Synthesis experiments are being performed to validate this hypothesis.

The number of Spanish and Catalan common diphones is 362. Therefore, a bilingual database requires 961 diphones. In contrast, the recording of separate databases would require 1323 diphones. With respect to the longer units, all the Spanish ones are included on the Catalan database. Therefore, 255 unit are needed for two separated databases but only 150 for a bilingual database. The comparison of a bilingual database and two separated databases is presented on table I.

	SP	**CT**	**SP+CT**	**SP ∪ CT**
Diphones	511	812	1323	961
Longer Units	105	150	255	150

Table I. Number of diphones and longer units defined for a Spanish TTS database (SP), Catalan TTS database (CT), two separated databases (SP+CT) and a bilingual database (SP ∪ CT).

5.2 Recording

To record the selected units, they have to be uttered either inside a word or inside a sentence. In previous works, the authors have recorded databases using natural sentences and words and using non-sense artificial words (*logotomas*). The *logotomas* are designed so that the units are not contaminated by the context, basically using plosives with the same articulation point and vowels with the same degree of opening. The result of the comparison is the following:

- If *logotomas* are used, the diphones appear clearly, the extraction of the diphone is very easy, and the pitch of the whole database is very uniform. This is not the case when sentences are used.
- The articulation speed of the uttered diphones is significantly smaller in *logotomas* than in continuous speech. Therefore, in order to have natural speech, the duration of the concatenated units has to be significantly smaller than that of the recorded units. The result is that, if the TD-PSOLA method is used, some significant parts of the speech segments have to be deleted and the quality is degraded.

Based on this comparison, the bilingual database has been recorded using *logotomas* but imposing a high and constant articulation speed.

The speech is recorded with a DAT (at 48 kHz) and decimated to 16 kHz. Once the database is labelled, a 8kHz version is created for telephone applications. A FIR filter of order 37 was used so that the labels can be delayed 9 samples ((37-1)/4)). If the A-law is used to code the samples, the quality is not degraded significantly. The size of the bilingual A-law 16kHz database is around 1.8Mbytes.

5.3 Labelling

The TDPSOLA synthesis algorithm [7] requires a pitch-synchronous marking of the signal. We have used an automatic supervised method which has happened to be very effective.

A segmentation of the non-sense words into phones is performed using a Hidden Markov Model based segmentation tool. From the result, the waveform and the spectrogram of the word, focused on the phones to be labelled, is presented to the supervisor of the segmentation.

The supervisor corrects the beginning and end labels of each one of the phones of the synthesis unit. The beginning mark of the first phone and the ending mark of the last phone are used to compute the limits of the synthesis unit. The marks between

phones are used during the synthesis to protect the fragile zones of the speech related with transitions between phones.

The signal is pitch-synchronous marked using a time-frequency based detection method [8]. The method uses a rough approximation of the pitch value, which is supplied for each speaker. The supervisor verify the marks and correct the errors. This would have been the most tedious part of the process but it has been almost eliminated because of the low error percentage of the pitch detection algorithm. One of the reasons to achieve so good performance is the use of *logotomas* so that the initial estimation of the pitch is a close approximation of the pitch values. The supervision of the whole database was performed in around 75 hours.

6. TRANFORMATION OF VOICES

As it was mentioned in the introduction, our TTS system has been adapted to provide handicapped people which suffer from oral disorders (mainly palsy people) with the ability of speaking. One of the environments in which the system has to be used is school to support conversation by different students. It is important to personalise somehow the voices so that each voice can be associated to one student. Our system allows the use of different databases and different prosody patterns. However, in order to offer further variation, some experiments have been done to easily change the characteristic of the voice associated to each database.

The first idea is to change the pitch mean value, so that, e.g., from a woman voice, man and child voices can be obtained. Unfortunately the TD-PSOLA algorithm does not allows to use a wide range of pitch values without degrading the quality. Furthermore, the result happens to be similar to the original one. A best effect is accomplished *playing* with the D/A sampling frequency of standard audio devices. It is well known that if a sound is reproduced at lower speed, the spectrum is compressed. It means lower pitch and lower formant values. A moderate change in the formants produce the subjective effect of speaker change. For instance, for a woman voice, recorded at 16kHz, and with mean pitch value of 176, we can get:

- Male voice: reproduce at 11 kHz (factor = 0.70)
- Child voice: reproduce at 20 kHz (factor = 1.25)

In order not to alter the duration of the speech, the duration values assigned by the prosodic patterns have to be affected by the same factor.

For the range of values between 0.70 and 1.25, the resulting voices are quite good, similar to the quality obtained reproducing at the original rate (16kHz). If compared with changing the mean pitch on the prosodic patterns, the voice sounds better and much more different to the original one.

It should be said that even reproducing at 20kHz, the system works in real time with a personal computer (*Pentium* 90Mhz).

7. ACKNOWLEDGMENTS

The authors want to acknowledge Lourdes Aguilar, David Casacuberta and Rafael Marín from the *Universitat Autònoma de Barcelona* for their valuable discussion on the linguistic aspects of the TTS system.

8. SUMMARY

In this paper the UPC text-to-speech system has been presented. The system allows to produce Spanish and Catalan speech. It is composed of a phonetic transcription module based on deterministic rules, a hierarchical intonation model and a rule-based predictor of the vowel duration. The synthesis module is based on the TD-PSOLA algorithm.

Effort has been done to minimise the effort of development and the computational resources of the bilingual system with respect to the monolingual one: for instance, the prosodic models are represented in the same framework and the sounds which are present in both languages are shared on the database. A method has been presented to select the diphones of the language, with possible reduction of the total number of diphones. Furthermore, a supervised labelling system has been presented which allows a very fast segmentation and pitch labelling of the synthesis database.

Finally, a simple but effective method has been presented to transform a woman voice in man and child voice.

REFERENCES

[1] J.B. Mariño, *"Reglas para la transcripción fonética aplicadas en RAMSES"*, Research Report, UPC, June 95

[2] A. Pujol and I. Esquerra, "*Regles de transcripció fonètica del català*", Research Report, UPC, October 1996

[3] J.M. Garrido, *Modelización de patrones melódicos del español para la síntesis y el reconocimiento del habla*, Universitat Autònoma de Barcelona, UAB, 1991

[4] J.M. Garrido, *Modelling Spanish Intonation for Text-to-Speech Applications*, PhD dissertation, Universitat Autònoma de Barcelona, UAB, 1996.

[5] R. Marín, "La duración vocálica en español", *Estudios de Lingüística de Universidad de Alicante*, vol. 10: pp. 213-226, 1995.

[6[L.Aguilar et al, "Catalan vowel duration", *Proceedings of EuroSpeech'97*, Rhodes 1997.

[7] E. Moulines, F. Charpentier, 'Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones', *Speech Communication,* 9, pp 453-467, 1990

[8] J.L. Navarro, I. Esquerra, "A Time-Frequency Approach to Epoch Detection", *Proceedings of Eurospeech'95*, pp 405-408, Madrid, 1995