A Bilingual Spanish-Catalan Database of Units for Concatenative Synthesis

Ignasi Esquerra, Antonio Bonafonte, Francesc Vallverdú, Albert Febrer

Department of Signal Processing and Communications
Universitat Politècnica de Catalunya
Campus Nord UPC, 08034 Barcelona, Spain
ignasi@gps.tsc.upc.es

Abstract

Different databases of phonetic units are required in multilingual Text-to-Speech systems based on concatenative synthesis. We are currently developing a TTS system able to convert text either in Catalan and Spanish, with some of the modules being used indistinctly by the two languages while others are specific to each language. In order to reduce the total amount of units, a bilingual database has been obtained from two monolingual databases recorded by the same speaker, which contains all possible units for both languages. Common units have been selected according to their phonetic representation. The bilingual database has 1099 units, including diphones and some long units, while the two monolingual databases would result in 1545 units. An analysis of Catalan unit frequencies has been done to select what units should be included in the database. The experiments carried out showed that that synthetic speech has a strong Catalan accent, probably due to the speaker's accent. Some common units, even if they are represented with the same symbol, should be considered separately in a bilingual database in order to cope with acoustically different allophones.

1. INTRODUCTION

Many current text-to-speech (TTS) systems are based on concatenative synthesis to produce high-quality speech. The main disadvantage of this approach is that, to add a new voice to the system, it is necessary to record a completely new unit database from another speaker. This implies a long process of recording, segmenting and labelling all those units, which are usually around 1000 in number. Moreover, for a multilingual TTS system, apart from changes in some of the system's modules, each new language requires also a different database because of the different phonetic set of units, and usually from different speakers.

Catalan is the native language of Catalonia, a bilingual area of Spain where both languages are almost equally spoken according the last statistics. However, the prevalence of Spanish is notorious in many communicative activities, as can be seen by looking at the number of newspapers written in one or the other language. At a personal level, very often both languages can be listened together in a conversation, where one of the participants speaks in Catalan and the other answers in Spanish. Even more, a person can be changing from one language to the other, depending to whom he is talking, or for example to tell a common sentence or a saying.

With respect to written texts, specially in newspapers, a text in Catalan (or Spanish) can include words in Spanish (or Catalan), like personal nouns, places or quotations. In a good TTS system, these words should be detected and pronounced using the adequate set of phonemes and intonation patterns, as a bilingual speaker would normally do. The problem arises in how to detect such words, more than synthesising using a different language database.

Our objective is to develop a full bilingual TTS system, based on concatenative synthesis, that can convert text either in Catalan or Spanish (Bonafonte et al., 1997). The system was initially developed in Spanish, then we added the necessary new modules for synthesising in Catalan, and the system is now being put together so as to work indistinctly in both languages. A SGML-like language is used to mark when a part of text should be spoken in the other language, as well as for changing the speaker characteristics and others speech synthesis controls (Taylor & Isard, 1997).

Four main blocks of processing compose our TTS system: text normalisation, phonetic transcription, prosody generation and acoustic synthesis. The number of common tasks between the two languages has been kept as high as possible. The synthesis module does not need to be rebuilt as it only deals with acoustic units, modifying prosodic values and concatenating them to synthesise the acoustic output. On the other hand, text normalisation and phonetic transcription are totally language-dependent; although the algorithm is basically the same, the rules and data are very different. Finally, even though prosodic rules are surely somewhat different for Catalan and Spanish, prosody generation is currently done using an unique module, applying the same rules but with different parameter values. We hope to be able in the near future to adapt the prosody generation module to the characteristics of both languages by means of intonation and duration analysis of real speech.

Not only it is interesting to have a multilingual system, but to have a system able to produce different regional accents of a language (Williams & Isard, 1997). In both cases, this implies a larger unit database, apart from several modifications in the transcription and prosodic generation modules. Other approaches that also increase the size of databases are those that include different types of units to improve segmental intelligibility (Portele et al., 1997) or that store several realizations of unit in different

prosodic contexts to improve suprasegmental quality (Campbell & Black, 1997).

In this work we present the creation of a Spanish-Catalan unit database for a bilingual TTS system and its evaluation. Two databases have been recorded and merged into only one with all necessary units for both languages. The objective is not to achieve an important reduction of database units, but to test some aspects of multilinguality applied to speech synthesis, as for instance the influence of regional accents.

In the first section, the set of phonetic units used in this work is introduced, as well as some of the problems found in dealing with two languages during the phonetic transcription process. Next, the second section describes how the monolingual databases were created, including an analysis of Catalan unit frequencies. The bilingual database generation is presented in section four. The paper concludes with a discussion about the quality of synthetic speech obtained with the bilingual database compared to the monolingual ones.

2. PHONETIC TRANSCRIPTION

In a TTS system, the first task is text normalisation, which consists in the expansion of digits, abbreviations and other non-readable symbols into its orthographic form in order to be acceptable by the following modules. In the future, a language detection algorithm is expected to be placed at this point that will be able to automatically assign a language label to each text, or even at sentence or word level, so as to apply the appropriate rules and databases. The full orthographic text is then segmented in syllables and vowel stress is assigned to words. This is of great importance to generate later synthetic prosody.

Phonetic transcription in Spanish is quite straightforward using simple deterministic rules (Mariño, 1995). In Catalan, transcription is not so easy and many times it is not possible to derive a correct phonetic transcription using only deterministic rules (Pujol & Esquerra, 1997). Initially, the Catalan transcription module was developed based on the Spanish one, because the rules for the latter were simpler, which raised the problem of having to dela with some inherited data structures and functions that were not always appropriate for the transcription of the new language.

Symbols for representing Spanish allophones are those defined by the computer-readable phonetic alphabet SAMPA[1]. For Catalan transcription, symbols have been taken mainly from the Spanish one, and some others from French and English alphabets [Table 1].

Spanish (31)	a e i o u p t k b d g B D G f T s z tS x j w jj m n J N l L r rr
Catalan (36)	a e E i o O u @ p t k b d g B D G f s z S Z ts dz tS dZ j w m n J N l L r rr

[1] http://www.phon.ucl.ac.uk/home/sampa/spanish.htm

Table 1: Phonetic symbols in Spanish and Catalan

As it can be seen in the previous table, the amount of vocalic sounds is higher in Catalan (8) than in Spanish (5). The Catalan language, at least the central dialect, has the peculiarity of having two vocalic systems, one for stressed syllables ([a], [e], [E], [i], [o], [O], [u]) and one for unstressed syllables with only three phones ([i], [u], [@]). This fact will be clearly seen in next section where unit frequency will be presented.

A source of errors in Catalan phonetic transcription is the case of mid-vowels in stressed syllables with no orthographic accent, which can be pronounced either as mid-open ([E], [O]) or mid-closed ([e], [o]) depending on the word without any available general rule. For example, *nena* /n'En@/ (little girl) vs. *neva* /n'eB@/ (it snows), or *rosa* /rr'Oz@/ (rose) vs. *rossa* /rr'os@/ (blonde). Another problem is transcription of some word final consonants (as *t* and *r*), which some times are pronounced (some monosyllabic words) and others not (infinitives). In all these cases, a dictionary of known problematic frequent words is looked up at the beginning of linguistic processing. Otherwise, a general grapheme-to-phoneme rule is applied as default.

An evaluation performed on the Catalan transcription program showed that it worked reasonably well (Esquerra, 1997). A dictionary of 1400 words, with a phonetic transcription done by experts, was used as reference. Among the errors, approximately one half correspond to previously said ambiguities. As a whole, the performance of the Catalan transcription module can be considered around the 90% of correctness.

3. UNIT FREQUENCY ANALYSIS AND CORPUS GENERATION

Speech synthesis is done by time-domain diphone concatenation. These units consist in a combination of two phones, spanning from the stationary part of the first phone to the stationary part of the second one. A Spanish diphone database was already available from previous projects. The definition of units was done creating a finite-state grammar of possible phoneme combinations, and it was later validated using the transcription analysis of a large text corpus. However, many of the diphones obtained were extremely unusual, so it was decided to reduce the corpus size by synthesising those units from half phones. As a result, a corpus of 527 diphones is used in our system for the Spanish database.

A one-million word corpus was transcribed and analysed to obtain allophone frequencies [Table 2]. Not surprisingly, the results showed the schwa allophone [@] is the most common sound in Catalan. This is because due to the fact that in the central dialect, spoken by the greatest part of population in Catalonia, all non-stressed vowels *a* and *e* are pronounced with this sound.

Since affricate phones are the allophones that are less frequent, it was decided to regard them as a combination of two phones (plosive+fricative), because considering them as allophones would result in a large quantity of diphones in the database due to the fact that many Catalan words can finish in one of those consonantic sounds.

Frequency analysis also took into account diphones. From the total amount of present units in the corpus, the ones with lower frequency were individually checked. It was found that many of them corresponded to non-existent consonantic combinations, caused either by erroneous text normalisation or transcription. After this reduction of diphones, 798 units are finally considered in the Catalan database.

Apart from diphones, another type of unit, called hereafter long units, has been included in both the Spanish and Catalan databases to improve intelligibility of the synthetic voice. Consonants [r] and [l] after a plosive and before a vowel is phonetically very weak, meaning that it has an important degree of coarticulation with the neighbouring phones. These long units can be viewed as triphones, since they are made of half plosive, the whole lateral or vibrant consonant and half vowel.

Catalan		Spanish	
Allophone	Freq.(%)	Allophone	Freq.(%)
@	18.77	e	13.72
i	7.67	a	13.43
s	6.50	o	10.37
n	6.09	n	6.91
l	5.78	s	6.90
t	5.48	t	4.61
u	5.10	l	4.25
a	4.64	r	4.15
k	4.58	i	4.15
e	3.81	k	4.11
m	3.60	m	3.69
r	3.55	D	3.25
D	3.26	p	2.61
p	3.09	j	2.61
z	3.08	B	2.48
o	2.85	u	1.99
rr	2.43	T	1.54
B	2.21	z	1.39
f	0.99	w	1.35
w	0.82	rr	1.20
d	0.79	G	0.78
G	0.78	d	0.76
Z	0.60	x	0.63
N	0.46	L	0.54
L	0.45	f	0.51
O	0.39	N	0.50
b	0.35	b	0.45
E	0.34	tS	0.40
ts	0.31	J	0.28
S	0.31	jj	0.18
j	0.30	g	0.13
J	0.25		
g	0.18		
dz	0.07		
dZ	0.07		
tS	0.05		

Table 2: Allophone frequency comparison

Two different corpora, one for each language, have been created. Diphones and long units are embedded in a corpus of non-sense words in order to have a low coarticulatory and prosodic influence on the unit we want to get. Words have been automatically created using a phonetic grammar.

4. BILINGUAL DATABASE

The bilingual database was derived from two monolingual databases previously generated. A male speaker was told to read the two unit corpora, first the Spanish one with the indication that it was to obtain a Spanish database of units, and then the one in Catalan. The speaker's mother tongue is Catalan, but he was selected among other candidates because his low Catalan accent when speaking in Spanish, and because his voice gave the best synthetic quality once processed by our TTS system.

Speech signals were recorded in a DAT at a sampling frequency of 48kHz, and later digitally down-sampled to 16kHz. After being split in several files, each containing one word (i.e. one unit), the speech signals were automatically aligned with their phonetic transcription using a HMM-based segmentation tool. A manual validation was performed to correct possible segmentation errors in a task that took approximately 75 hours of work.

To create the bilingual, first it is necessary the determination of common allophones between the two languages. Units are classified into three categories (common, only-Spanish and only-Catalan) according to a very simplistic criteria: common units are those who have an equal phonetic representation [Table 3]. Language specific units are those that are made of allophones that only exist in that language. It must be remembered that affricate phonemes are not considered.

As a matter of fact, although some allophones are represented with the same SAMPA symbol, they are slightly different. For instance, the correspondence between cardinal vowels is quite valid, but it is not so clear what happens with mid-vowels, as it can be seen by the fact that the Spanish vowel [e] is acoustically between the Catalan mid-closed [e] and the mid-open [E]. Likewise, lateral consonant [l] is an alveolar-dental consonant in Spanish, while in Catalan its articulation point is more palatal. This is, by the way, one of the phonetic cues used to detect Catalan people's origin when speaking in Spanish.

Common	a e i o u p t k b d g B D G f s z S j w m n J N l L r rr
only-Spanish	T x jj
only-Catalan	E O @ Z

Table 3: Phonetic classification of common allophones in the bilingual database

Depending on which monolingual database has been used to extract common units, two different bilingual databases can be considered (BE, BC). The total amount of diphones for this database is 953 (+147 long units), while separate databases would result in 1325 units (+220 long units) [Table 4]. Therefore, with a bilingual database we can achieve a reduction of 446 units in comparison to the monolingual databases.

	Diphones	Long units	Total
ME	527	103	630
MC	798	117	915
ME+MC	**1325**	**220**	**1545**
common	372	74	446
only-E	155	29	184
only-C	426	43	469
BE/BC	**953**	**146**	**1099**

Table 4: Number of units in the monolingual and bilingual databases

5. DATABASE EVALUATION

As said before, classification in common, only-Spanish and only-Catalan units for the bilingual database has been done comparing the phonetic symbols used to represent allophones. Therefore, both monolingual databases are included in the bilingual databases, except for the fact that common units from the latter have been extracted using different corpora of non-sense words, one spoken in Catalan, the other in Spanish.

Synthesising with the different databases created, the first effect that can be perceived is a strong Catalan accent in all sentences, even for the Spanish ones. This is quite logical since the speaker's mother tongue is Catalan. However, listening to that speaker talking in Spanish, he didn't have a strong accent. The reasons why this occurs are probably related to prosody patterns rather than units themselves.

Bilingual databases represent an important reduction of units with respect to monolingual databases, without a significant loss of quality. Some common units are not very well synthesised using a bilingual database because they are not acoustically equivalent in both languages although they are represented with the same phonetic symbol. This leads to the conclusion that they should be discriminated in order to achieve a better quality at the cost of increasing the number of units.

Experiments are currently being carried out with larger bilingual databases. We expect to improve quality of synthetic speech in our bilingual TTS system without having to have two complete monolingual databases. From the comparison between bilingual and monolingual databases, it can be concluded that it is possible to reduce database size by merging common units, but they have to be carefully selected so as not to have a strong influence of the language from which they were extracted.

ACKNOWLEDGEMENTS

Authors want to thank researchers from the Departament de Filologia Espanyola (Universitat Autònoma de Barcelona) for their valuable discussion in many linguistics aspects of our TTS system. This research was partially supported by the CICYT under contract TIC95-1022-C05-04.

REFERENCES

Bonafonte A., Esquerra I., Febrer A., Vallverdú F. (1997), "A Bilingual Text-to-Speech System in Spanish and Catalan", Proceedings of EUROSPEECH'97, Rhodes, sept.97, vol.5, pp. 2455-2458

Campbell N., Black A. (1997), "Prosody and the Selection of Source Units for Concatenative Synthesis", in *Progress in Speech Synthesis*, chap. 22

Mariño J.B. (1995), "Reglas para la transcripción fonética aplicadas a RAMSES", Internal Research Report UPC

Portele T., Höfer F., Hess W.J. (1997), "A mixed Inventory Structure for German Concatenative Synthesis", in *Progress in Speech Synthesis*, chap. 21

Pujol A, Esquerra I. (1996), "Regles de transcripció fonètica del català", Internal Research Report UPC

Esquerra I. (1997), "Avaluació del transcriptor en català", Internal Research Report UPC

Taylor P., Isard A., (1997), "SSML: A Speech Synthesis Markup language, Speech Communication, 21, pp.123-133

Williams B., Isard S., (1997), "A Keyvowel Approach to the Synthesis of Regional Accents of English", Proceedings of EUROSPEECH'97, Rhodes, sept.97, vol.5, pp. 2435-2438