Workshop on Language Resources for European Minority Languages

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, Barcelona, Spain
ignasi@gps.tsc.upc.es

ABSTRACT

A text-to-speech system based on concatenative synthesis requires a different database of units for each language it is able to produce. In our case, the TTS system is being developed to synthesise text either in Catalan or Spanish. Although Catalan is being largely used by people in Catalonia, it still is the minority language and suffers a strong influence form Spanish. In order to reduce the total number of units, a bilingual database has been created which contains all possible units from the two languages. The criteria to select the common units consists in simply comparing their phonetic representation in the computer readable phonetic alphabet SAMPA. The bilingual database has 1099 units, counting diphones and long units, while the sum of units of two monolingual databases would be 1545. An analysis of unit frequencies has been done to select which units should be included in the databases. As a conclusion of the experiments carried out, it can be said that synthetic speech has a strong Catalan accent in all databases, even for the Spanish ones. Since some phones are especially dependent on the language of the corpus from which they were extracted, they should have to be considered as acoustically different.

1. INTRODUCTION

Many of current text-to-speech (TTS) systems use concatenative synthesis to produce high-quality speech. The main disadvantage of this approach is that for every new voice, it is necessary to record a completely new database of units from another speaker. This requires a long process of segmenting and labelling all those units, which are usually around 1000 in number. If the system can convert text in several languages, this increases also the number of databases because each language requires its own database recorded from a different speaker.

Catalonia is a bilingual area, with Catalan as a native language, embedded in a larger region where Spanish is prevalent. In other regions of Spain Catalan is also spoken but there is still even more a minority language. Many often both languages are used indistinctly in a conversation by the same person changing from one another without paying attention. All Catalan speakers can speak in Spanish, but not the other way round. Newspapers and in general any printed text can include some sentences or nouns in the other language that are pronounced using the convenient set of phonemes and intonation patterns. These facts make that Catalan, the minority language, is largely influenced by Spanish in many aspects.

The objective is to develop a full bilingual TTS system, based on concatenative synthesis, that can convert text either in Catalan or Spanish [1]. The system was initially developed for the Spanish language because the transcription module was already available. We are now

expanding the system to Catalan, trying to share the maximum number of modules by the two languages.

Some modules, like the synthesiser, are language-independent since they only deal with units, modifying their prosodic values and concatenating them to build the acoustic output. On the other hand, modules like transcription are totally language-dependent although the algorithm is basically the same. Prosody generation is currently done applying the same rules for Catalan and Spanish but with different parameter values.

In this work we present the development of a joint Spanish-Catalan database of units for the TTS system and its evaluation. Two databases have been created and merged into only one with all the necessary units to convert text in the two languages. In the first section, the phonetic transcription, the set of allophones and some of the problems found in dealing with two languages are introduced. Then, an analysis of unit frequencies is presented before the description of the database generation. The paper finishes with a discussion of perceptive opinions about the quality of the synthetic speech obtained with this database.

2. PHONETIC TRANSCRIPTION

The input text to the TTS system is first normalised, which means that digits, abbreviations, acronyms and other non-readable symbols are expanded into an orthographic form. The normalised text is then segmented in syllables and stress is assigned to vowels in each word; this is of key importance to generate the synthetic prosody. The set of rules applied in those steps depends on the selected language.

Transcription in Spanish is quite straightforward using simple deterministic rules, while in Catalan there are some difficulties that cannot be solved exclusively by rule. For the Spanish language, a set of 31 allophones is currently considered [2]. In Catalan, the number is 36 allophones [3]. In both sets, a symbol representing a pause is included. The SAMPA notation for Catalan is still not approved; many phonemes are common to Spanish, the others have been taken from the French or English SAMPA alphabets.

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

For instance, vowels <e>/<o> in a stressed syllable without orthographic accent can be pronounced either as [e]/[o] or [E]/[O] depending on the word. Another problem is <t> and <r> in word's last position, which some times are pronounced (some monosyllabic words) and others not (infinitives). In these cases, a dictionary containing frequent words presenting these problems is looked up, and a heuristic rule is applied as a default.

One of the problems in developing the Catalan transcription module from the Spanish one is that the data structures and algorithm flow of task were not always appropriate to deal with the new difficulties that appeared in transcribing Catalan texts. On the one hand, more constant strings with different phonetic classes had to be defined because the number of possible situations of a phoneme in different contexts were greater. On the other hand, to solve the ambiguity in the transcription of vowels, it was necessary to put forward some rules during the phase of syllabification.

An evaluation performed on the Catalan transcription program showed that 75% of the words were perfectly converted [4]. A dictionary of 1400 words, with the phonetic transcription done by experts, were used as a reference. Among the errors, approximately one half correspond to aforementioned ambiguities with vowels and consonantic word final phonemes. As a conclusion,

the performance of the Catalan transcription can be situated around the 90% of correctness.

3. FREQUENCY ANALYSIS OF UNITS

Synthesis is done by diphone concatenation. These units consist in a combination of two phones, having approximately one half of each. A previous database in Spanish was available so that the necessary units were already known. To select the units of the Catalan database, a corpus of 66000 words was transcribed and analysed to obtain the frequencies of phones and diphones. In the following table, the results for the Spanish and Catalan allophone frequencies are presented.

Spanish		Catalan	
Allophone	Freq. (%)	Allophone	Freq. (%)
e	13.25	@	18.27
a	12.97	i	7.31
o	10.01	s	6.45
n	6.67	n	5.97
s	6.66	l	5.64
t	4.45	t	5.33
i	4.15	u	4.84
l	4.10	a	4.49
r	4.01	k	4.25
k	3.97	m	3.63
m	3.56	e	3.57
/	3.47	D	3.55
D	3.14	r	3.46
p	2.52	z	2.95
j	2.52	p	2.91
B	2.39	o	2.89
u	1.92	/	2.85
T	1.49	rr	2.47
z	1.34	B	2.16
w	1.30	f	1.01
rr	1.16	w	0.84
G	0.75	G	0.79
d	0.73	Z	0.63
x	0.61	L	0.55
L	0.52	d	0.41
f	0.49	b	0.40
N	0.48	E	0.35
b	0.43	S	0.31
S	0.39	N	0.31
J	0.27	O	0.30
jj	0.17	j	0.25
g	0.13	J	0.24
		ts	0.22
		g	0.18
		dZ	0.09
		dz	0.06
		tS	0.06

Not surprisingly, the results showed that 18% of sounds in Catalan correspond to schwa [@]. This is 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 seldom appear, it was decided to split them into two phonemes (occlusive+fricative),

because this would result in a large quantity of diphones when making the combinations of affricates with all possible consonantic contexts.

With respect to diphone frequencies in Catalan, 870 different diphones were found in the corpus. Since it is impossible to list all diphones, only the number of left and right contexts for each allophone is presented. It can be observed that only the schwa [@] exists in all possible contexts (37), while the other vocalic sounds have less combinations but with a large number of consonantic contexts. On the other hand, some consonantic allophones have few contexts that result basically from combinations with vowels.

Allophone	# of left contexts	# of right contexts
i	37	36
e	37	35
E	26	29
a	37	34
o	34	34
O	23	24
u	37	35
@	37	37
j	16	24
w	13	29
p	26	26
t	25	29
k	29	28
B	30	18
D	32	18
G	26	20
b	11	15
d	15	17
g	7	12
f	26	19
s	31	25
z	27	22
S	15	22
Z	28	8
ts	8	21
dZ	9	9
tS	12	18
dz	6	6
m	32	30
n	30	32
N	9	19
J	7	18
l	32	29
L	24	24
r	22	8
rr	26	31
/	29	28

4. UNIT DATABASE

From the results of the previous analysis, a set of diphones for Catalan was selected. The most rare diphones are often caused by the concatenation of words beginning and ending in a consonant. Although they can actually be in the transcription output, some can be substituted by

other similar and more frequent units. Alternatively, they can be created from the concatenation of short segments from phone+silence units. Apart from them, databases include some long units that contain a consonantic cluster, made of a plosive, a liquid [r]/[l] and a vowel.

Two different corpus have been recorded from the same male speaker, one telling him to pronounce words with a Catalan accent, the other with a Spanish accent. Non-sense words have been used to carry the units. Different rules have been applied between the two languages to create those corpora because the set of units and coarticulation rules are also different.

Speech signals were recorded in a DAT at a sampling frequency of 48kHz and later digitally down-sampled to 16kHz. After being cut in several files, each containing one non-sense word, they were aligned with their phonetic transcription applying a HMM segmentation tool. For each file, the waveform and the spectrogram, together with the phone labels and pitch marks, were presented to a supervisor who moved, added or removed marks validating the previous automatic segmentation in a process that took around 75 hours of manual work.

Units are classified into three categories (common, only-Spanish and only-Catalan) according to a very simplistic criteria: common units have an equal phonetic representation. In fact, although some allophones are represented with the same SAMPA symbol, they are slightly different in their articulation point. This effect can be perceived very easily for lateral consonants and mid-closed vowels.

In order to be used as a reference, two monolingual databases have been created from the corresponding corpus. Depending on the language of origin of common units, two bilingual databases have been considered. The total amount of diphones for this database is 961 (+150 long units), while separate databases would result in 1323 units (+255 long units).

The characteristics of these databases are summarised in the following table:

	Diphones	Long units
Monolingual		
m_esp	527	103
m_cat	798	117
Bilingual		
b_cat, b_esp	953	146
common	372	74
only-Spanish	155	29
only-Catalan	426	43

5. DATABASES EVALUATION

As said before, classification is done according to the phonetic symbols used to represent each sound. This means that the monolingual databases are included in the bilingual databases, except from the fact that the common units from latter have been created using different corpora of non-sense words, one in Catalan, the other in Spanish.

The first effect that can be acoustically perceived is a strong Catalan accent in all databases, even the Spanish ones. This is quite surprising because listening to that speaker when talking in Spanish, he didn't have almost any accent. Therefore, a carefully selection of a speaker is very important since language accent is amplified by the synthesis process.

Bilingual databases represent an important reduction of units with respect to monolingual databases, without a significant loss of quality. Some phones are synthesised not very well using the bilingual database depending on the corpus from which the common units were extracted. This leads to the conclusion that they probably should have to be discriminated although they share the same phonetic symbol.

Experiments are currently being carried with larger bilingual databases differentiating between

phones with a strong influence on the accent. We expect to improve quality of synthetic speech without having to have two complete monolingual databases.

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

[1] A. Bonafonte, I. Esquerra, A. Febrer, F. Vallverdú, "A Bilingual Text-to-Speech System in Spanish and Catalan", Proc. of EUROSPEECH'97, Rhodes, sep.97, vol.5, pp. 2455-2458

[2] J.B. Mariño, "Reglas para la transcripción fonética aplicadas a RAMSES", Internal Research Report UPC, jun.95

[3] A. Pujol, I. Esquerra, "Regles de transcripció fonètica del català", Internal Research Report UPC, oct.96

[4] I. Esquerra, "Avaluació del transcriptor en català", Internal Research Report UPC, jul.97

Workshop on Language Resources for European Minority Languages

A Bilingual Spanish-Catalan Database of Units for Concatenative Synthesis

Ignasi Esquerra, Albert Febrer, Antonio Bonafonte, Francesc Vallverdú

Universitat Politècnica de Catalunya, Spain

Abstract

Many of the current text-to-speech (TTS) systems use concatenative synthesis to produce high-quality speech. The main disadvantage of this approach is that for every new voice, it is necessary to record a completely new database of units from another speaker. In addition, to extend the system to other languages, not only is required to change the unit database, but to adapt some modules of the TTS system.

Catalonia is a bilingual area, in which all Catalan speakers are able to speak in Spanish. Many often both languages are used indistincly in a conversation. In addition, newspapers and in general any text can have embedded some sentences or proper nouns in the other language that have to be converted using the convenient unit database.

Our objective is to achieve a full bilingual TTS system, being able to convert text either in one or the other language even if they are mixed in the same text. We started from a system developped for the Spanish language and we are now adapting it to Catalan, with some of the modules being shared by the two languages, while others are language specific. The unit database is language-dependent by definition. However, we have created a bilingual Spanish-Catalan database that contains all units necessary to convert text in both the two languages.

Two databases have been recorded from the same speaker, one telling him to pronounce the non-sense words corpus with a Catalan accent, the other with a Spanish accent. Rules to create those corpora are different since the set of units for each language is also different.

Diphones, which are the phonetic units used by the TTS system, have been classed into three categories: common (362 units), only-Catalan (450), and only-Spanish (149). The total amount of units for this bilingual database is 961, not taking into account some consonantic cluster. Two separate databases, one for Catalan (812) and the other for Spanish (511) would result in 1323 units.

As a first experiment, classification is done according to the phonetic symbol used to represent each sound. From a phonetic point of view, this is not totally correct since phones are slightly different in their place of articulation. This effect can be acoustically perceived as a strong accent depending on the corpus from which the common units were extracted.

A second experiment has been carried, collating only phones that acoustically are similar, and leaving apart those that have a strong language dependence even if they are represented with the same phonetic symbol. The resulting voice has less accent than before, although the database size is increased. It is interesting to note that the accent of the synthetic voice is much stronger than the actual speaker's accent.