Monolingual and Bilingual Spanish-Catalan Speech Recognizers Developed from SpeechDat Databases

José B. Mariño, Jaume Padrell, Asunción Moreno and Climent Nadeu

Research Center for Language and Speech Technology and Applications (TALP-UPC)
Universidad Politécnica de Cataluña
Jordi Girona 1-3, 08034-Barcelona, Spain.
(canton, jaume, asuncion, climent)@tsc.upc.es

Abstract

Under the SpeechDat specifications, the Spanish member of SpeechDat consortium has recorded a Catalan database that includes one thousand speakers. This communication describes some experimental work that has been carried out using both the Spanish and the Catalan speech material.

A speech recognition system has been trained for the Spanish language using a selection of the phonetically balanced utterances from the 4500 SpeechDat training sessions. Utterances with mispronounced or incomplete words and with intermittent noise were discarded. A set of 26 allophones was selected to account for the Spanish sounds and clustered demiphones have been used as context dependent sub-lexical units. Following the same methodology, a recognition system was trained from the Catalan SpeechDat database. Catalan sounds were described with 32 allophones. Additionally, a bilingual recognition system was built for both the Spanish and Catalan languages. By means of clustering techniques, the suitable set of allophones to cover simultaneously both languages was determined. Thus, 33 allophones were selected. The training material was built by the whole Catalan training material and the Spanish material coming from the Eastern region of Spain (the region where Catalan is spoken).

The performance of the Spanish, Catalan and bilingual systems were assessed under the same framework. The Spanish system exhibits a significantly better performance than the rest of systems due to its better training. The bilingual system provides an equivalent performance to that afforded by both language specific systems trained with the Eastern Spanish material or the Catalan SpeechDat corpus.

I. Introduction

Catalonia[1] is a country whose population share two languages (the Spanish and the Catalan), with three sorts of individuals: speakers of both languages (bilingual population) and speakers of either Spanish or Catalan language (monolingual people), Spanish speakers being the largest monolingual population. The picture of this social-linguistic situation can be completed by noticing that speakers are supposed to use the language of their own preference, since nearly everybody can understand both languages. In such a bilingual situation, it is rather natural to deploy speech recognition technology in both languages simultaneously.

Both Spanish and Catalan languages are distinct enough so that different language speakers can not understand each other, unless they have got some knowledge of the other language. However, since both languages are derived from Latin and have had a long life together, their phonetics have important similarities. As a consequence, even though each language exhibits exclusive sounds, they share an important amount of allophones, as we will see below.

Thus, an alternative to use a specific phonetic description for either Spanish or Catalan can be devised. In this paper a common inventory of allophones for both languages is designed and evaluated against the monolingual counterparts. The paper is organized as follows. The next section describes the available speech databases used to obtain the language specific modeling and the shared phonetic modeling and to test them. It also describes the main features of the recognition system used for this experimental work. In section III the initial sets of sounds for Spanish and Catalan are introduced and the determination of the common inventory is described. The fourth section provides the description of the experimental work carried out to validate and evaluate that bilingual inventory. The paper ends with a discussion section.

[1] Catalonia is a politically autonomous region in the North-East of Spain.

II. Experimental Framework

II.a Speech Databases

Spanish Database

The Spanish speech material used in our experimentation comes from the Spanish corpus of the SpeechDat project (Moreno & Winsky, 1996; Moreno, 1997). The utterances were recorded through an ISDN access to the public telephone fixed network, sampled at 8 kHz and quantified by the A-law at 8 bits per sample. The Spanish SpeechDat database is formed by utterances collected from 5000 speakers: 4500 of them supply the training material and the remaining 500 speakers build up the testing set. As training material we have selected the phonetically balanced sentences. After discarding the utterances with mispronounced or incomplete words and with intermittent noise, we obtained a corpus of 20490 utterances, including more than one million phones (Full set). It exceeds 25 hours and 30 minutes of speech. Picking up the speakers from Catalonia, a subset of this training material was defined (Eastern set). This second training set includes 4952 utterances from 976 speakers providing more than 6 hours of speech with 249,800 phonemes.

Catalan Databases

The Catalan training material was obtained from a Catalan database (Hernando & Nadeu, 1999) that follows the SpeechDat specifications. The signal format is 8 kHz, 8 bit, A-law and it is recorded through an ISDN access to the public-switched telephone network. The phonetically

balanced sentences were selected as training material. Utterances with mispronounced or incomplete words and with intermittent noise were discarded. Furthermore, only speakers of Eastern Catalan dialect were selected. After this selection, we have available 4978 sentences from 807 speakers that contain about 8 hours of speech with a total of 216,355 phonemes.

The test material was taken from the Catalan VOCATEL speech database (Nadeu, Padrell & Febrer, 1997). This database was recorded by Telefónica I+D and UPC. The recordings were made through an analogue access to the public telephone network, using the same signal format of the SpeechDat database except that the compression A-law was substituted by the mu-law. The corpus was recorded by asking a total of 8000 callers to say a set of 25 prompted sentences (between one word and six words long sentences). This material provides 4567 speakers of the Eastern Catalan dialect.

Table 1 summarises the main figures of the Spanish and Catalan material selected to train the acoustic models for the recognition systems described in this paper.

	speakers	**utterances**	**time**	**phonemes**
Full set	4500	20940	25h30m	1,000,000
Eastern set	976	5952	6h	249,800
Catalan	807	4978	8h	216,335

Table 1. Spanish (Full and Eastern sets) and Catalan training corpus.

II.b Recognition System

The experimental work was carried out with the speech recognition system developed in our laboratories.

The speech is parameterized with mel-cepstrum coefficients. CMS (cepstral mean subtraction) is used. First and second order differential parameters plus the differential energy are employed.

The phonetic unit used is the demiphone (Mariño, Nogueiras & Bonafonte, 1997). It is a contextual phonetic unit that models a half of a phoneme. A left demiphone describes the beginning part of a phoneme and takes into account the coarticulatory effect produced by the previous sound. Accordingly, a right demiphone models the rest of the phoneme and depends on the next phoneme. For instance, the phoneme /s/ between the vowels /o/ and /a/ is modeled by the concatenation of two units o-s s+a, being o-s a left demiphone and s+a a right one. Despite the good coverage of contexts that the demiphone provides, the problem of unseen units during training can happen. Besides, a lack of smoothness in the estimation of Hidden Markov Models (HMM) for some units can also be present. Clustering of models is used to overcome these two drawbacks. A hybrid algorithm that combines decision-tree based (top-down) clustering and agglomerative (bottom-up) clustering is implemented (Mariño & Nogueiras, 1999). This hybrid algorithm shares the advantages of both simpler components: it assigns a model to units that are not present in the training speech material but appear in the target vocabulary, and performs and optimum unrestricted clustering.

The recognition system models the phonetic units by gaussian SCHMM with quantization to the 6 (2 for the energy) closest codewords. The size of the codebooks is 128 (32 for the differential energy).

The languages of tasks are modelled by means of X-gram´s (Bonafonte & Mariño, 1996) whose efficient implementation is described in Bonafonte & Mariño (1998). Every item of the vocabulary is represented by a string of demiphones. A word is provided with a unique transcription.

The optimisation search is sped up by using beam-search and phonetic look-ahead. With a Pentium processor at 200 MHz a real time performance is reached.

II.c The tasks

Two tasks have been chosen to test the performance of the monolingual and bilingual recognition systems. Both tasks are common for the Spanish and Catalan languages.

The first one is an isolated word test. The words correspond to names of people and cities. It is a difficult task because the moderate size of the vocabulary and the great similarity among words.

The second task is composed by telephone numbers uttered in a Spanish way. For instance, the number 933216920 is organised in two or three-figure numbers: 93 321 69 20. The telephone number lengths span from 6 or 7 digits of the Catalan utterances to 9 or 10 digits of the Spanish ones. An inventory of 121 words is used to represent the Spanish numbers, while the Catalan lexicon is limited to 39 words. This difference is due to the different approach followed to define either vocabulary. In order to cope with the coarticulation between words, certain Spanish numbers that orthographically are expressed with several words have been represented with a unique word. For instance, “ciento ocho” (108) is expressed as “cientocho”. Although this strategy is easily implemented in Spanish, it is difficult to apply in Catalan. The X-gram that models this task exhibits a test perplexity of 18.5 and 19.4 for Spanish and Catalan, respectively.

Table 2 provides the size of the vocabulary and the number of available utterances for both tasks and both languages. Additionally, the total number of words that compose the telephone number utterances is included. The larger size of the Catalan corpus can be explained in terms of the greater amount of speech material independent of the training corpus available in Catalan than in Spanish.

	names		**phone numbers**		
	size	utt.	size	utt.	words
Spanish	540	1030	121	270	2775
Catalan	743	5597	39	2476	15870

Table 2. Main characteristics of the Spanish and Catalan test tasks (*size* of the vocabulary, number of *utt*erances and total number of *words*).

III. Monolingual and Bilingual Inventories of Sounds

Table 3 summarises the Spanish and Catalan inventories of allophones as they were defined to design the SpeechDat databases. The same table includes for every allophone the attributes considered in the clustering algorithms. It can be observed that both languages share the greatest part of their inventories, with very few specific sounds for each language. Clearly, the Catalan language has the largest set. Its vowel set is composed by eight sounds: the basic set of five vowels (Spanish set) plus the open versions /E/ and /O/, and the neuter or

schwa /@/. The Catalan language contains three affricate consonants (/dz/, /dZ/, /ts/) in addtition to the Spanish /tS/. Catalan substitutes the Spanish fricative unvoiced consonants /T/ and /x/ by /S/ and /Z/, respectively. Finally, it is worth mentioning that three allophones (/j/, /l/, /w/) exhibit different realisations for each language.

ph	Attributes	lang
a	vowel, central, open, voiced.	S, C
@	vowel, central, schwa, voiced	C
b	consonant, bilabial, plosive, voiced.	S, C
B	consonant, bilabial, approximant, voiced.	S, C
d	consonant, dental, plosive, voiced.	S, C
dz	consonant, alveolar, affricate, voiced.	C
dZ	consonant, palatal, affricate, voiced.	C
D	consonant, dental, approximant, voiced.	S, C
e	vowel, front, mid, voiced.	S, C
E	vowel, front, open-mid, voiced.	C
f	consonant, labiodental, fricative, unvoiced.	S, C
g	consonant, velar, plosive, voiced.	S, C
G	consonant, velar, approximant, voiced.	S, C
i	vowel, front, close, voiced.	S, C
j	semivowel, palatal, front, close, voiced.	S
	semiconsonant, palatal, front, close, voiced.	C
jj	consonant, palatal, approximant, voiced.	S, C
J	consonant, palatal, nasal, voiced.	S, C
k	consonant, velar, plosive, unvoiced.	S, C
l	consonant, alveolar, lateral, liquid, voiced.	S
	consonant, alveolar, lateral, liquid, back, voiced.	C
L	consonant, palatal, lateral, voiced.	S, C
m	consonant, bilabial, nasal, voiced.	S, C
n	consonant, alveolar, nasal, voiced.	S, C
N	consonant, velar, nasal, voiced.	S, C
o	vowel, back, mid, voiced.	S, C
O	vowel, back, open-mid, voiced.	C
p	consonant, bilabial, plosive, unvoiced.	S, C
r	consonant, alveolar, tap, rothics, liquid, voiced.	S, C
rr	consonant, trill, alveolar, rothics, vibrate, voiced.	S, C
s	consonant, alveolar, fricative, unvoiced.	S, C
S	consonant,palatal, fricative, unvoiced.	C
t	consonant, dental, plosive, unvoiced.	S, C
ts	consonant, alveolar, affricate, unvoiced.	C
tS	consonant, palatal, affricate, unvoiced.	S, C
T	consonant, dental, fricative, unvoiced.	S
u	vowel, back, close, voiced.	S, C
w	Semivowel, velar, back, close, voiced.	S
	Semiconsonat, velar, back, close, voiced.	C
x	consonant, velar, fricative, unvoiced.	S
z	consonant, alveolar, fricative, voiced.	S, C
Z	consonant, palatal, fricative, voiced.	C

Table 3. Spanish (S) set and Catalan (C) set of allophones as defined for the SpeechDat databases.

The monolingual speech recognition systems were designed with a reduced set of allophones with respect to the full inventories in Table 3. Stop (/b/, /d/, /g/) and approximant (/B/, /D/, /G/) realisations of plosive voiced consonants were assimilated, just as alveolar /n/ and velar /N/ nasals, because the presence of one or another realisation depends only on the context, and this dependence is provided by the demiphone. Furthermore, the Catalan voiced affricates /dz/ and /dZ/, two sounds very similar and with a low incidence in language, were modelled together. Finally, the Spanish palatal lateral consonant /L/ was merged with the palatal approximant /jj/, because most of speakers pronounce /L/ as /jj/ ("yeismo").

As far as the bilingual speech recognition system is concerned, the inventory in Table 3 (with the modifications previously mentioned) could have been taken as the set of allophones for the bilingual recognition system. However, some tests were carried out using a clustering algorithm in order to confirm the acoustic similarity between the Spanish and Catalan sounds transcribed with the same allophone, and to assess whether certain allophones could be grouped or not.

With this aim, the following set of models were estimated:

a) 26 Spanish context independent phones from the Spanish SpeechDat training corpus;
b) 32 Catalan context independent phones from the Catalan SpeechDat training corpus;
c) 690 Spanish demiphones with more than 25 appearances in the Spanish SpeechDat training corpus;
d) 855 Catalan demiphones with more than 25 appearances in the Catalan SpeechDat training corpus.

The applied clustering algorithm is described in Mariño & Nogueiras (1999) under the name of agglomerative clustering. A cluster is represented by an average model obtained from the statistical mean of the models that populate the cluster. The homogeneity function that drives the clustering procedure is related to the entropy of this average model, and the number of available samples in the training corpus of the phonetic units gathered in the cluster. Two clusters are merged when the clustering of the pair provokes the minimum decrement in homogeneity compared to any other pair of clusters. The following paragraphs provide a brief description of our experimentation with clustering and the main conclusions we can extract.

III.a Bilingual Set of vowels

Firstly, the clustering of vowels was carried out. The vowels were forced to gradually reduce the number of groups from twelve to six. Table 4 shows the sequence of clusters that came out. The first column indicates the language, Spanish (S) or Catalan (C), and the rest of columns from left to right show the clusters in the order they appear.

A first glance suggests that Spanish and Catalan vowels behave in a way that could be expected: mid and mid-open versions of Catalan vowels cluster together, and the Spanish and Catalan counterparts of a same vowel join the same group. Besides, the schwa /@/ can not be clustered. Further steps of clustering maintain this vowel alone.

S			i	u	o	e	a
C	O, o	E, e	i	u	O, o	E, e	a

Table 4. Sequence of vowel clustering.

It is interesting to gain insight into the cluster entropy of the Catalan /e/ and /E/ and the Spanish /e/. In Table 5 the entropy (H) of both the models and some clusters are reported. Some facts deserve to be remarked:

a) Vowel /E/ exhibits by far the least entropy. It can be explained in terms of the phonetic transcription procedure. Initially, every stressed /E/ without orthographic mark was considered an /e/ and, after looking the word up in a dictionary, the final transcription was decided. Because this dictionary is not exhaustive, the actual set of /E/ is a subset of the real set.
b) Clustering the vowels /e/ and /E/ does not increase the entropy over that of the /e/.
c) The cluster built up by the Catalan front vowels shows less entropy than the Spanish /e/. This result, added to the previous one, seems to suggest a unique transcription /e/ for both vowels in a speech recognition framework. In fact, an informal test provided better performance when either vowel was represented by /e/ than when a distinct transcription was assumed. However, a doubt remains on the effect of the dictionary in this result. So, the point deserves more attention in the future.

Table 5 also points out that putting Spanish and Catalan front vowels together does not imply a noticeable increase of entropy. Similar results can be obtained for mid and mid-open back vowels. Consequently, the inventory of vowels used in the bilingual speech recognition system is formed by /@/, /a/, /e/, /i/, /o/ and /u/.

SC	/e/	/E/	/e/, /E/	/e/	/e/, /E/
H	50.4	47.2	50.4	51.5	51.7

Table 5. Entropy of different clusters of Spanish and Catalan front vowels.

III.b Bilingual Inventory of Consonants

Before clustering the consonant sounds, they were divided into three groups: nasals, voiced (no nasalised) and unvoiced. Except in one case, any fact suggesting a clustering of sounds different from that provided by the phonetic transcription was not observed. The exception was the alveolar, lateral, liquid consonant /l/. In this case, the clustering increases the entropy significantly with respect to the increment produced when the rest of consonants are grouped. Additionally, when the demiphones of both Spanish and Catalan phoneme /l/ are clustered into only two groups, Spanish and Catalan demiphones join in separate ways: Spanish demiphones form one group and Catalan demiphones set up another. On the contrary, when a similar experiment was tried with the rest of consonants, the influence of context prevailed over the influence of language in the cluster composition. Therefore, only the lateral consonant /l/ was provided with a specific model for each language.

As far as the glides /j/ and /w/ are concerned, they do not show a different behaviour than the rest of sounds. So, both languages share the same model.

As a result of this study, 33 allophones were considered in the bilingual recognition system: 6 vowels, 2 glides and 25 consonants.

IV. Evaluation

The following recognition systems were trained and evaluated:

a) A monolingual *Spanish* system, trained from the whole set of Spanish SpeehDat speakers (Full set in Table 1). A total of 550 hidden Markov models of demiphones were estimated.
b) A monolingual *Eastern Spanish* system, whose 350 demiphones have been estimated from the Spanish SpeechDat speakers corresponding to the Eastern region (Eastern set in Table 1).
c) A monolingual *Catalan* system, with 350 demiphones trained from the Catalan SpeechDat material.
d) A *Bilingual* system, trained from the corpus composed gathering the Eastern set of the Spanish SpeechDat database and the Catalan SpeechDat database. Since either part of this training corpus includes a similar amount of speech, a balanced training of Catalan and Spanish was achieved. Two sets of demiphones were obtained: one with 575 models and other with 500 models. The performances of these two designs were practically equivalent.

The Eastern Spanish system was built as a Spanish counterpart of the Catalan system. Both Eastern Spanish and Catalan systems were estimated from corpora of similar size and supply acoustic-phonetic modelling with the same number of hidden Markov models. Thus, both systems provide equivalent references to the bilingual system.

Tables 6 and 7 show the performance of these four systems when applied to the previously described test tasks. Table 6 points out the results obtained with the Spanish material and Table 7 exhibits the performance reached in the Catalan test. The accuracy of words is indicated in bold characters and its range of a 95% of confidence is included in brackets.

System	Names	Phone Numbers
Spanish	**82.6** (80.2, 84.8)	**93.5** (92.5, 94.4)
Eastern Spanish	**81.0** (78.5, 83.3)	**92.0** (90.9, 93.0)
Bilingual	**81.5** (79.0, 83.8)	**91.5** (90.4, 92.5)

Table 6. Word accuracy reached with the Spanish tests.

System	Names	Phone Numbers
Catalan	**71.5** (70.3, 72.7)	**90.9** (90.4, 91.3)
Bilingual	**70.8** (69.6, 72.0)	**90.8** (90.3, 91.2)

Table 7. Word accuracy obtained with the Catalan tests.

V. Discussion

A first glance at Table 6 shows that the performance of the Spanish (full set) system is clearly higher than the performance yielded by the other two systems. The superior training provided by the full set reasonably accounts for this result. Besides, this behaviour can be considered significant according to the reduced overlap of the confidence intervals in the phone number task. Although the name task shows a greater overlap, the larger size of the phone number test makes its scores to be the most reliable.

As far as the Bilingual system is concerned, its performance can be considered equivalent to that provided

by the Eastern Spanish and Catalan systems according to the scores shown in Tables 6 and 7. Thus, the Bilingual system saves parameters without a loss in performance with respect to these systems.

The Spanish test material covers the different regional accents considered in the SpeechDat project in proportion to their population. Thus, the results in Table 6 describe the average performance for Spanish speakers. However, speakers coming from the Eastern region involve only the 20% of the test material. Table 8 shows the word accuracy provided by the recognition systems when applied only to this part of the test speech. As it can be seen, although the scoring improves (somehow natural for the Eastern Spanish and Bilingual system), the relative behaviour is maintained.

System	Names	Phone Numbers
Spanish	**84.7** (79.3, 88.9)	**96.4** (94.6, 97.6)
Eastern Spanish	**83.0** (77.4, 87.4)	**93.5** (91.3, 95.2)
Bilingual	**82.5** (76.9, 87.0)	**93.7** (91.5, 95.4)

Table 8. Word accuracy reached with the Eastern material of the Spanish tests.

Clearly, the results in Table 6 or 8 for the Spanish mean a better performance than the results in Table 7 for the Catalan. The different recording format of both training and test materials of the Catalan language can be a possible explanation of it. Besides, its isolate word test exhibits a vocabulary nearly a 40% greater than de Spanish test.

Acknowledgements

This research was supported by CICYT of Spanish government under contracts TIC98-0423-C06-01 and TIC98-0685. The authors also wish to acknowledge the access to the VOCATEL database allowed by Telefónica I+D.

References

A. Bonafonte and J.B. Mariño, 1996. Language Modeling using X-grams. In Proc. ICSLP96 (pp. 394-397), Philadelphia, USA.

A. Bonafonte and J.B. Mariño, 1998. Using X-gram for efficient speech recognition. In Proc. ICSLP98 (pp. 2559-2562), Sydney, Australia.

J. Hernando and C. Nadeu, 1999. SpeechDat. Catalan Database for the Fixed Telephone Network. Corpus Design Technical Report, TALP-UPC.

J.B. Mariño, A. Nogueiras and A. Bonafonte, 1997. The demiphone: an efficient subword unit for Continuous Speech Recognition. In Proc. EUROSPEECH'97 (pp. 1215-1218), Rhodes, Greece.

J.B. Mariño and A. Nogueiras, 1999. Top-down bottom-up hybrid clustering algorithm for acoustic-phonetic modeling of speech. In Proc. EUROSPEECH'99 (pp. 1343-1346), Budapest, Hungary.

A. Moreno and R. Winsky, 1996. Spanish Fixed Network Speech Corpus, SpeechDat Project LRE-63314.

A. Moreno, 1997. SpeechDat Spanish Database for Fixed Telephone Networks. Corpus Design Technical Report, SpeechDat Project LE2-4001.

C. Nadeu, J. Padrell and A. Febrer, 1997. Diseño de la Base de Datos Vestel y Preparación de la Captura. Technical report. Fundació Catalana per a la Recerca y Telefónica I+D.