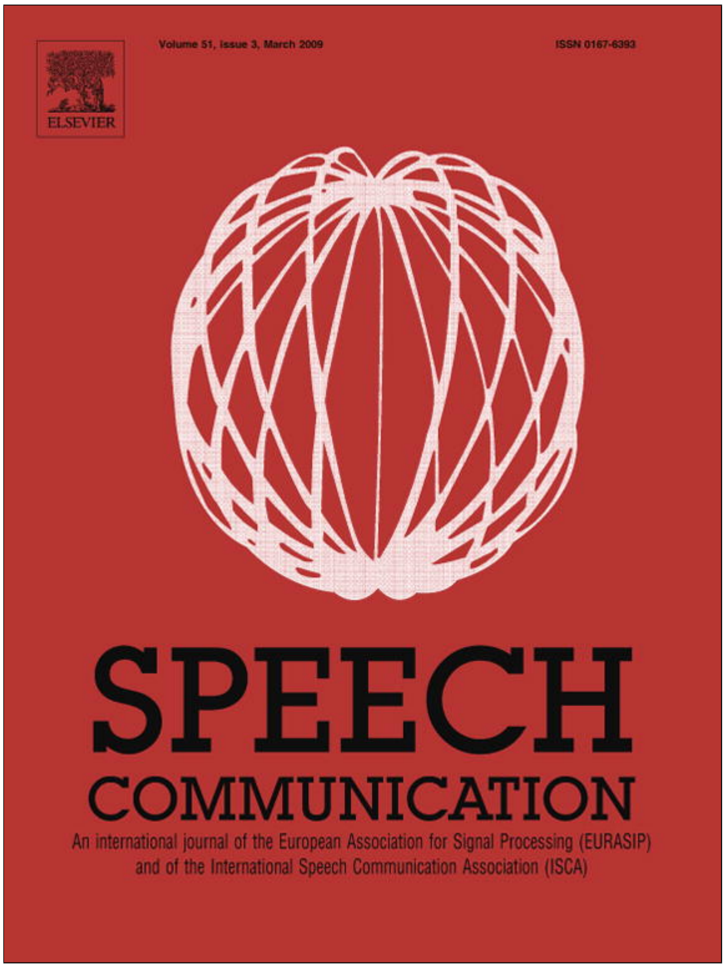


Volume 51, issue 3, March 2009

ISSN 0167-6393

ELSEVIER

SPEECH
COMMUNICATION

An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 217–229
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Multidialectal Spanish acoustic modeling for speech recognition

Mónica Caballero [*], Asunción Moreno, Albino Nogueiras

Talp Research Center, Universitat Politècnica de Catalunya, Spain

Received 30 May 2007; received in revised form 24 July 2008; accepted 27 August 2008

Abstract

During the last years, language resources for speech recognition have been collected for many languages and specifically, for global languages. One of the characteristics of global languages is their wide geographical dispersion, and consequently, their wide phonetic, lexical, and semantic dialectal variability. Even if the collected data is huge, it is difficult to represent dialectal variants accurately.

This paper deals with multidialectal acoustic modeling for Spanish. The goal is to create a set of multidialectal acoustic models that represents the sounds of the Spanish language as spoken in Latin America and Spain. A comparative study of different methods for combining data between dialects is presented. The developed approaches are based on decision tree clustering algorithms. They differ on whether a multidialectal phone set is defined, and in the decision tree structure applied.

Besides, a common overall phonetic transcription for all dialects is proposed. This transcription can be used in combination with all the proposed acoustic modeling approaches. Overall transcription combined with approaches based on defining a multidialectal phone set leads to a full dialect-independent recognizer, capable to recognize any dialect even with a total absence of training data from such dialect.

Multidialectal systems are evaluated over data collected in five different countries: Spain, Colombia, Venezuela, Argentina and Mexico. The best results given by multidialectal systems show a relative improvement of 13% over the results obtained with monodialectal systems. Experiments with dialect-independent systems have been conducted to recognize speech from Chile, a dialect not seen in the training process. The recognition results obtained for this dialect are similar to the ones obtained for other dialects.

Keywords: Multidialectal ASR system; Dialect-independent ASR system; Spanish acoustic modeling; Spanish dialects

1. Introduction

Dialectal variability is a significant degrading factor in automatic speech recognition (ASR) performance. Research shows that a mismatch in dialects between training and testing speakers significantly influences recognition accuracy in several languages like French (Brousseau and Fox, 1992), Japanese (Kudo et al., 1996), Dutch (Diakolukas et al., 1997), German (Fischer et al., 1998) or English (Chengalvarayan, 2001), as an example. Spanish is not an exception, as it has been shown in research (de la Torre et al., 1996; Zissmann et al., 1996; Aalburg and Hoege, 2003). Efforts in dialect ASR technology have followed two different goals: (i) to improve dialectal recognition rates by developing recognition systems tailored to specific dialects and (ii) to design multidialectal ASR systems robust to dialect variation. The primary tools to achieve these goals are lexical and acoustic modeling, while the existence and availability of language resources are the main constraints.

Concerning lexical modeling, a common approach consists in adapting the lexicon to represent dialectal variants, either by adding alternative pronunciations to the lexicon or by adapting the transcription to a given dialect (Beringer et al., 1998; ten Bosch, 2000; Baum et al., 2001). Results show that when using the same set of acoustic models for all dialects, lexical modeling does not lead to a significant improvement. Therefore, lexical modeling alone is not

[*] Corresponding author. Tel.: +34 654 736 452.
E-mail addresses: monica@gps.tsc.upc.edu (M. Caballero), asuncion@gps.tsc.upc.edu (A. Moreno), albino@gps.tsc.upc.edu (A. Nogueiras).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.08.003

enough to achieve good results and has to be combined with acoustic modeling.

Statistical acoustic models have been shown to retain accent and dialect information in a consistent way; they have been widely used in the study of dialectal variation and in the identification and phonetic classification of language variants in a data-driven manner. Different acoustic measures can be applied over dialect-dependent hidden Markov models (HMM) to create dialect maps (Heeringa and Gooskens, 2003; Salvi, 2003), or use the recognition accuracy of dialect-dependent acoustic models to evaluate dialect distances. Training dialect-dependent acoustic models is only possible if dialect data are available, and several approaches can be found in the literature to cope with data scarcity in dialectal ASR applications. If there are enough data, a specific dialect recognizer can be built totally independent of the recognizer developed for the language or standard dialect (Fischer et al., 1998). This approach requires a dialect identification module when a system has to deal with different accents or dialects. More recent approaches are based on sharing data and resources between dialects. Data from one or more dialects can be used to increase the amount of training data of one monodialectal system (Kirchhoff and Vergyri, 2005), or to build a set of multidialectal acoustic models that can be used to recognize speech from several dialects (Chengalvarayan, 2001). The latter approach seems to be the more robust, since the variations in the way the same phone can be pronounced in different dialects cause the resultant acoustic models to provide greater acoustical space coverage.

Adaptation methods can be applied to well-trained acoustic models to obtain a set of models that are specific to a dialect with a limited amount of dialect speech data. In (Diakolukas et al., 1997; Fischer et al., 1998) adaptation is applied to models trained with the standard dialect resources. Alternatively, multidialectal acoustic models could be adapted in the same way as (Schultz and Waibel, 2001) do in a multilingual approach with language-independent models.

Techniques similar to those used in multilingual acoustic modeling research can be used to define a multidialectal set of acoustic models (i.e., each dialect is handled as a different language). In order to define and properly train the multidialectal acoustic models, similar phonetic units have to be identified across dialects. The similarity between the sounds of different languages – or different dialects – can either be defined by an expert, or be estimated by data-driven methods. Expert methods use linguistic knowledge. The most common approach is based on IPA (or SAMPA) alphabet: phones of different languages are considered similar if they map onto the same class as defined by IPA (or SAMPA) (Byrne et al., 2000; Chengalvarayan, 2001). As a result of this procedure, a global phone set is defined for all the languages. In data-driven methods, similarity between phonetic units across languages is commonly estimated by evaluating the distance of their language-dependent acoustic models (i.e. HMMs) using agglomerative (Köhler, 2001; Salvi, 2003; Imperl et al., 2003), decision tree based (Schultz and Waibel, 2001), or a combination of decision tree and agglomerative (Mariño et al., 2000) clustering algorithms. Other data-driven approaches find the similarity between phones by means of a confusion matrix (Byrne et al., 2000). Measuring similarity between language-context-dependent phonetic units, such as demiphones (Mariño et al., 2000), triphones (Imperl et al., 2003) or pentaphones (Schultz and Waibel, 2001) provide better recognition results than measuring similarity between language-context-independent units. In addition, (Imperl et al., 2003) conclude that although an agglomerative clustering algorithm yields a limited number of clusters, the decision tree method gives better recognition results and solves modeling units that are not seen in the training data.

Concerning the structure of the decision tree in context modeling, a distinct decision tree is typically grown for each unit (or each state of each unit) in the phone set. Another approach is to build a single global decision tree structure that allows parameters to be shared by different phones. The single global decision tree structure was used in (Duchateau et al., 1997; Yu and Schultz, 2003) for improving monolingual acoustic modeling. In (Caballero et al., 2004) authors applied this tree structure in the multidialectal acoustic modeling of three Spanish dialects with encouraging results.

These techniques provide robustness in acoustic modeling, but the recognition system has to know the dialect of the test speaker, either because the dialects do not share grapheme-to-phoneme transcription rules and phone sets, or because dialect information is needed to browse the decision tree.

The existence of dialect data resources is a key factor in studying and solving dialectal problems, but it is difficult and expensive to collect new data. With more than 300,000 million speakers worldwide, Spanish is one of the most widely spoken languages and is considered to be one of the global languages in the world. Dialectal variants can be found across Spain and Latin American countries, as well as within countries. Databases for properly training ASR systems for Latin American dialects are appearing. Adding to the former VAHA or CALL HOME databases available in the LDC, the SpeechDat Across Latin America (SALA) project (Moreno et al., 1998) developed a set of telephone databases in most of the Latin American countries for the purposes of training ASR systems.

Some research deals with recognition of Spanish dialects or its influence in a Spanish ASR. Variability due to speakers and data from different dialects is considered to be pronunciation variation; as such, it is modeled by adding alternative pronunciations to the lexicon (Billa et al., 1997; Ferreiros and Pardo, 1999), or by defining a simple phonetic set (Huerta et al., 1997) in order to integrate variability in HMM. Two examples of specific dialectal modeling can be found in (Aalburg and Hoege, 2003; de la Torre et al., 1996). In the first paper, Spanish as spoken in Spain is used to model non-native speech applied to a system trained with Colombian speakers. In the second, Argentinean and Spanish as spoken in Spain are considered. Both studies apply lexical modeling and adaptation techniques in order to improve recognition accuracy for a specific dialect speech. Good results are obtained, but in both cases, the acoustic models are tailored to a single dialect. In (Nogueiras et al., 2002), authors created a multidialectal ASR system for three Spanish dialects that improved monodialectal performances. Authors also showed that testing Latin American dialects in a system trained with data of Spain did not improve the monodialectal performance.

The goal of this paper is to create a multidialectal speech recognition system robust to dialect variations. The intended language is Spanish including dialectal variants from Latin America and Spain. Designing a system that is completely robust to dialectal variations requires total independence to the dialect of the speaker. A number of solutions are possible, such as adding a stage to the recognizer that identifies the dialect of the speaker or having equal representation (transcription) of the recognition vocabulary for all the dialects. This paper focuses on the latter approach. A new overall phonetic transcription technique common to all the Spanish dialects is proposed. In defining an overall transcription, dialect information is used not to adapt the phone set and grapheme-to-phoneme rules to a particular dialect but rather to define a phone set and rules that enable the system to detect similarities and differences between sounds by applying a clustering algorithm in the acoustic modeling stage. This new transcription approach restricts the need for prior design decisions regarding the phone sets for each dialect and decisions regarding whether to transcribe a new dialect that is to be incorporated into the system or that is to be recognized by the system.

To create a robust multidialectal set of acoustic models, different methods for combining training data based on decision tree clustering algorithms are explored. The approaches differ on whether a multidialectal global phone set is defined and in the decision tree structure applied (i.e., multiple roots or one single global decision tree). Both, the multidialectal set of acoustic models and the overall transcription are combined with the aim of finding a robust recognizer for Spanish dialects. The resulting system is designed to be able to recognize any Spanish dialect, even when no training data for a given dialect are available.

The rest of the paper is organized as follows. Section 2 deals with Spanish language and dialects. Sections 3 and 4 describe canonical transcription rules and the overall transcription proposed for Spanish dialects. Section 5 is devoted to the methodology of multidialectal acoustic modeling. Section 6 describes the recognition system used in our research and gives an overview of the experiments carried out and the results obtained. Finally, our conclusions are presented in Section 7.

2. Spanish dialects

2.1. Spanish dialects across the world

As mentioned above, Spanish is one of the global languages in the world and is also one of the most widely spoken languages. It is the official language of Spain and of nearly all Latin America countries except Brazil, Guyanas and some Caribbean islands. Furthermore, its use is growing rapidly in the United States and Australia.

Spanish dialectal variants have been described in the literature and they include phonetic, lexical, semantic and cultural variations. Within Spain, one can roughly distinguish between the standard Castilian and southern dialects. With regard to Latin America, many factors prompted the appearance of dialectal variants: the varieties spoken by Spanish settlers, the state of the Spanish language in the time in which these settlers occupied the territories, contact with other languages, and linguistic drift of the dialects (Lipski, 1994). It is difficult to classify dialects or accents of Spanish in Latin America, since there are no clear boundaries between the varieties; one local variety may merge gradually into another, and it is sometimes easier to find dialectal similarities across countries than it is within countries. Spanish as spoken in Latin America is often broadly classified according to whether it is spoken in the highlands and mountains or in the lowlands and coastal areas. This broad division is due to the Andes mountains, which cross South America from north to south and favoured an initial Castilian settle in the high lands and a later southern Spanish settle in the coast lands. In addition, both dialects can be found across the entire continent, from Mexico to Chile. In this section we describe the main phonetics characteristics of the Spanish dialects.

2.2. Phonological differences between Spanish dialects

Most of the variation between Spanish dialects occurs with consonants, particularly in the fricative class. An explanation is that when Spanish settlers took their language to America, the sounds belonging to this phonetic class were still in evolution in Spain. The main variations can be classified into phonetic differences, consonant weakness in the coda position, and lenition. Next sections describe the details of each class. Examples are shown in Table 1.

2.2.1. Phonetic differences

These differences consist in the substitution or use of one phoneme or allophone by another.

- **Seseo**: Seseo affects the use of SAMPA phonemes /T/ (IPA /θ/) and /s/. The name seseo applies to those dialects that do not pronounce /T/ and pronounce /s/ instead. This is the most common effect.
- **Yeismo/lleismo/zeismo**: They affect the use of SAMPA phonemes /L/, /jj/, and /Z/ (/ʎ/, /y/ and /ʒ/ in the IPA alphabet). The most common effects are

Table 1
Examples of phonological differences between Spanish dialects. Table shows for each effect, a word example and its transcription in the absence or the presence of the related effect.

		Effect absence		Effect presence	
Effect	Ex. word	SAMPA	IPA	SAMPA	IPA
Seseo	caza (hunting)	K a T a	K a θ a	K a s a	
Yeismo	halla			a jj a	a y a
Lleismo	(find) /	a L a /	a ʎ a /		
	haya	a jj a	a y a	a L a / a jj a	a ʎ a / a y a
Zeismo	(beech)			a Z a	a ʒ a
Aspiration /s/	más (more)	m a s		m a h / m a	
Velarization /n/	manta (blanket)	m a n t a		m a N t a	m a ŋ t a
Distinction	carta (letter)	k a r t a		k a l t a	
/l/-/r/	bolsa (bag)	b o l s a		b o r s a	
Velar /x/	cojín (cushion)	–		k o x i n	
Glottal /h/	cojín (cushion)	–		k o h i n	
Palatal /C/	cojín (cushion)	–		k o C i n	k o ç i n
Elision of /D/ b.v.	lado (side)	l a D o	l a ð o	l a o	

– Yeismo: Only /jj/ is pronounced.
– Lleismo: /L/ and /jj/ are pronounced. It typically occurs in bilingual areas were /L/ is a phoneme of the other language (Quechua, Catalan).
– Zeismo: Only /Z/ is pronounced. It occurs in east Argentina and Uruguay.

2.2.2. *Consonant weakness in the coda position*

Variations between dialects in the coda position (at the end of a word or at the end of a syllable) include the following:

- **Preservation of /s/ in the coda position**: The most distinctive feature of variants of Spanish is the pronunciation of /s/. The main division between Latin American Spanish dialects is characterized by the preservation of /s/ at the end of a syllable or a word in highland areas. In lowlands, it is elided or aspirated, becoming /h/.
- **Velarization of /n/**: Nasal consonants in the coda position are velarized and become the allophone /N/ (/ŋ/ in IPA). This effect is common in lowland variants, mostly in the Caribbean area.
- **Distinction of /l/ and /r/ at the end of a syllable**: In lowland variants, especially in the Caribbean dialect, the distinction between laterals and vibrants in the coda position tends to be eroded. Few speakers exhibit complete neutralization.

2.2.3. *Lenition*

Lenition or softening occurs when a consonant that is considered strong becomes weak. The following are the most common cases:

- **Pronunciation of the fricative /x/**: The fricative voiceless velar /x/ is pronounced in Argentina, Chile, Mexico and Spain. In Chile, it is produced as a fricative voiceless palatal /C/ (IPA /ç/) when it precedes the vowels [e] and [i]. In the Caribbean area and in Colombia, it is aspirated and becomes a fricative voiceless glottal /h/.
- **Elision of /D/ between vowels**: The elision of /D/ is characteristic of lowlands, although it has become very common in informal speech in all dialects.

2.2.4. *Dialect phonological characteristics*

The above mentioned phonetic effects can be grouped into regions to perform a dialect map. We identified a small number of dialectal regions and for each region we found the predominant dialect (normally the variant spoken in the capital) or the variant spoken by the majority of the population:

- Mexico (ME): Represents Mexico and part of Central America. The most populated area is Mexico DF.
- Caribbean (CA): Includes Caribbean Islands, Venezuela and the Atlantic Coast of Central America and Colombia. The most populated area is Caracas (Venezuela).
- High land (CO): Represents the high land dialect of Colombia, Ecuador and Peru. The most populated area is Bogota (Colombia).
- Chile (CH): The most populated area is Santiago.
- Argentina (AR): Represents Argentina and Uruguay. The most populated area is Buenos Aires.
- Spain (SP): The most common dialect is spoken in Madrid.

Differences between dialects can be explained easily by the absence or presence of the effects discussed above. Table 2 summarizes the main differences between the variants.

3. Canonical transcription: following dialectal characteristics

It is well known that Spanish grapheme-to-phoneme transcription can be done with rules with few exceptions. In this study, phonetic transcription is based on rules. It is carried out automatically using SAMPA symbols. Llisterri and Mariño (1993) proposed a set of rules for transcribing Spanish as spoken in Spain. Based on that work,

Table 2
Phonological differences between dialects

Effect	AR	CA	CH	CO	ME	SP
Seseo	•	•	•	•	•	–
[Y/Ll/Z]eismo	Z	Y	Y	Y	Ll	Ll
Aspiration /s/	•	•	•	–	–	–
Velarization /n/	–	•	–	–	–	–
Distinction /l/-/r/	•	–	•	•	•	•
Pronunciation of /x/ (/x/, /h/ or/C/)	x	h	C	h	x	x
Elision of /D/ between vowels	–	•	–	–	–	–

A bullet marks the presence of an effect.

Moreno and Mariño (1998) developed a set of canonical transcription rules for Latin American Spanish dialects according to the specific phonetics of each dialect. To enable dialectal pronunciation to be accurately represented and to cope with all the Latin American dialects, the symbols /h/, /C/ and /Z/ were added to the standard SAMPA symbol set for Spanish (Gibbon et al., 1997).

In order to transcribe Latin American dialects, the (Moreno and Mariño, 1998) canonical transcription rules are applied:

- In all Spanish dialects, only /jj/ is considered to exist (non-existence of /L/), except in Argentina, where both are transcribed as /Z/.
- Because of seseo, across Latin America /T/ is not used and becomes /s/ in the transcription.
- The velar fricative /x/ is transformed into /h/ in Colombia and in the Caribbean. In Chile, /x/ is transformed into /C/ when it precedes the vowels /e/, /i/ and /j/.
- In Argentina, Chile and the Caribbean, [s] in the coda position is transcribed as /h/.
- Nasal consonants in the post-nuclear position are the velar /N/ in the Caribbean.

Table 3
Shared phones across dialects using canonical transcriptions

Vowels	a e i o u
Semivowels	j w
Plosives	p t k b B d D g G
Affricates	tS
Fricatives	f s z
Nasals	m n N J
Liquids	l r rr R l_CG r_CG

Table 4
Phones not shared across all dialects using canonical transcriptions

Allophone	Phonetic attributes	Dialects
jj	Voiced, palatal, fricative	CA CO ME SP
x	Voiceless, velar, fricative	AR ME SP
h	Voiceless, glottal, fricative	AR CA CO
Z	Voiced, palatoalveolar, fricative	AR
T	Voiceless, interdental, fricative	SP
C	Voiceless, palatal, fricative	CH

Right column indicates dialects where phone is present.

Additionally, in this research, /R/ was added to represent the post-nuclear [r], and the phonemes /l/ and /r/ were specifically tagged as 'CG' when they belonged to a consonant group (they followed a plosive or a [f]).

Tables 3 and 4 show SAMPA symbols used in canonical transcriptions of Spanish dialects. Table 3 contains phones shared across dialects, while Table 4 shows phones not present in all variants.

4. Overall transcription common to all dialects

Canonical transcriptions should not be followed blindly. Foldvik and Kvale (1998) found that traditional dialect maps may be of limited use in ASR and that dialectal boundaries are never clear-cut; however, statistical models for speech recognition retain accent information and that information may be useful for the purpose of improving ASR performance. Actually, there are no exclusive dialect rules, only phenomena that may be present in dialects or not. A question that springs to mind is, can these effects be reflected in a single overall transcription for all dialects? An overall transcription would prevent new rules from having to be designed for every new dialect added to the recognition system and would allow any variant of Spanish to be transcribed.

The overall transcription proposed in this work uses phonetic knowledge related to changes across dialects. The process consists in modifying the canonical transcription of the standard Spanish in certain situations by marking phones that are liable to be different in different dialects. In this study, Spanish as spoken in Spain is considered to be the standard variant.

Considering the effects that change across dialects, as explained in Section 2, it is easy to identify and separate special cases. Note that some approaches in the acoustic modeling stage allow dialect-dependent models to be separated or joined.

- **Phonetic differences**: To deal with seseo, the two phonemes /T/ and /s/ are used in order not to prejudice the dialect of Spain. For apical and sibilant expression of the phoneme /s/, the same SAMPA symbol /s/ is used for all dialects. In the case of yeismo, leismo and zeismo effects, we considered only one expression between /L/ and /jj/ in the canonical transcriptions, as the majority of the population only produces one of them. Linguistic research highlights the fact that there are areas in which these phonemes are kept and /L/ is still pronounced. In order to determine if this affects acoustic models, in overall transcription both phonemes are considered and kept separate.
- **Consonantic weakness in the coda position**: To mark [s], nasal ([n], [m]) and liquid [l] consonants in the coda position, a special tag 'C', is added to their SAMPA symbols. The phone /N/ is kept but not marked, as it is always uttered at the end of a syllable.

Table 5
Phone symbols used in overall transcription

Vowels	a e i o u
Semivowels	j w
Plosives	p t k b B d D **D_C** g G
Affricates	tS
Fricatives	f s **s_C** **T** **T_C** z **x**
Nasals	m **m_C** n **n_C** N J
Liquids	l **l_C** L jj r rr R l_CG r_CG

Symbols in bold face have been added to standard SAMPA symbols in order to reflect the differences across dialects.

Table 6
Comparison between canonical and overall transcription

Word	Dialect	Canonical transcription	Overall transcription
caza (hunting)	SP	k a T a	k a T a
	AR CA CH CO ME	k a s a	
halla (find) / (beech)	AR	a Z a	a L a / a jj a
	CA CH CO ME SP	a jj a	
manta (blanket)	CA	m a N t a	m a n t a
	AR CH CO ME SP	m a n t a	
más (more)	AR CA CH	m a h	m a s_C
	CO ME SP	m a s	
caja (box) / mujer (woman)	CH	k a h a / m u C e R	k a x a / m u x e R
	CA HI	k a h a / m u h e R	
	AR ME SP	k a x a / m u x e R	
red (net)	AR CA CH CO ME SP	r e D	r e D_C

- **Lenition**: Glottal and palatal pronunciations of /x/ are considered as allophonic variations and the same SAMPA symbol is used for the three allophones. In order to have a specific symbol for /D/ between vowels, the approximant /D/ in the coda position is also marked with the tag 'C'.

The final set of phone symbols is summarized in Table 5. All the above-mentioned differences between the dialects are included in the set. Table 6 shows a comparison of the canonical and overall transcription for some example words.

5. Acoustic modeling: data sharing

In this section, we describe various techniques designed to define a set of multidialectal acoustic models combining data of different dialects. These techniques are based on decision trees clustering algorithms. We propose two tree structures and two different starting points in order to ascertain which units could benefit from other data sources.

5.1. Starting point for contextual modeling

Two different approaches for getting the contextual models are applied and compared. The first one is based on the definition of a multidialectal global phone set based on SAMPA alphabet and the second approach avoids a preliminary phone set definition step.

- **Definition of a multidialectal global phone set (GPS)**: The SAMPA and IPA alphabets classify sounds based on their phonetic characteristics. This linguistic knowledge has been used to define which phones could share data in the training process and to define a global phone set. The sounds of different dialects that have the same representation in the SAMPA alphabet are considered to be the same phone. The global phone set is completed by adding phonemes that are not shared between dialects. In this approach, all the material from all the dialects that corresponds to the same SAMPA symbol is used to train the same acoustic model. This is the most common approach in the literature and is very useful if different languages/dialects share a considerable number of symbols, as in Spanish dialects. Once the phone set has been decided, no more dialect information is needed, so this type of measure allows the resultant acoustic models to be used to recognize a dialect that is not present in the training process, whenever no new phone is needed for that variant.
- **Starting with dialect-context-dependent (DCD) acoustic models**: A set of context-dependent acoustic models (i.e. hidden Markov models, HMMs) is trained for each dialect. Dialect-dependent models are marked with a dialect tag in order to be capable to distinguish them in the tree. This approach gives freedom to detect similar context-dependent acoustic units. The decision tree driven by the entropy measured over dialect-dependent models define which units (and from which dialects) are similar enough to share training data. Dialect information is needed in the regression through the decision tree, so it does not allow speech from a dialect that is not considered in the training process to be recognized.

5.2. Tree structures

Two tree structures are studied: a multiroot structure that applies SAMPA restrictions to the clustering algorithm, and an one-root structure with no SAMPA constraints.

- **Multiroot structure (MR)**: A different tree (root) is created for each SAMPA unit. Each root contains all the context-dependent acoustic models belonging to the same phone symbol. This is typically the structure used for context modeling in monolingual systems. Parameter sharing is not allowed between units with different SAMPA representation. When using overall transcription, units belonging to the same SAMPA symbol, even if they are marked, share the same root (e.g., /n/ and /n/_C) in order to keep one tree for each SAMPA symbol.

Table 7
Multidialectal acoustic modeling approaches as a combination of the starting point of the approach and tree structure used

		Starting point	
		Global phone set	Dialect-context-dependent models
Tree	Multiroot	GPS-MR	DCD-MR
Structure	One-root	GPS-OR	DCD-OR

- **One-root structure (OR)**:A single tree is built for all the units in the phone set. Its root contains all the context-dependent acoustic models of all the units. This structure enables data to be shared between different phones.

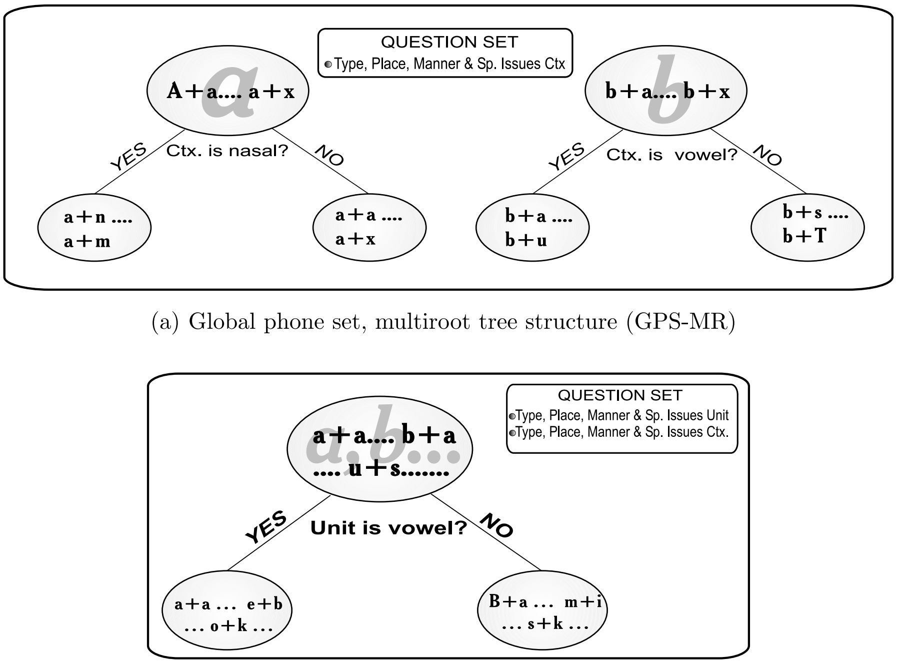
5.3. Multidialectal acoustic modeling approaches

Four approaches for multidialectal acoustic modeling are obtained by combining the types of starting points and decision tree structures presented above, as can be seen in Table 7. These approaches are graphically represented in Figs. 1 and 2.

- **Global phone set, multiroot tree structure (GPS-MR)**: In this approach, a global phone set is defined based on the SAMPA alphabet, that is, no distinction is made between units across dialects. Context modeling is achieved by applying a decision tree clustering algorithm using a multiroot structure, as in most monolingual systems. The question set only inquires about the context of the unit. When overall transcription is used, questions relative to the unit are added to those trees which have more than one unit in their roots. This is the most immediate approach, and the most intuitive. A major drawback is its dependence on the decisions made at the transcription stage, as it is totally based on the SAMPA alphabet.
- **Global phone set, one-root tree structure (GPS-OR)**: In this approach, a global phone set based on SAMPA is also defined. One-root structure decision tree algorithm is applied. The application of this structure allows models of different phones to be joined if they are similar in certain contexts or situations. The question set contains questions about the phone itself as well as the context.

Both global phone set based approaches determine a set of dialect-independent acoustic models, which can be used for any Spanish dialect, even if there are no data available for it. With canonical transcriptions, this is only possible if all the phones of the new dialect are contained in the global phone set. Using overall transcription does not have this drawback, as all the dialects share the same transcription rules and phone symbol set.

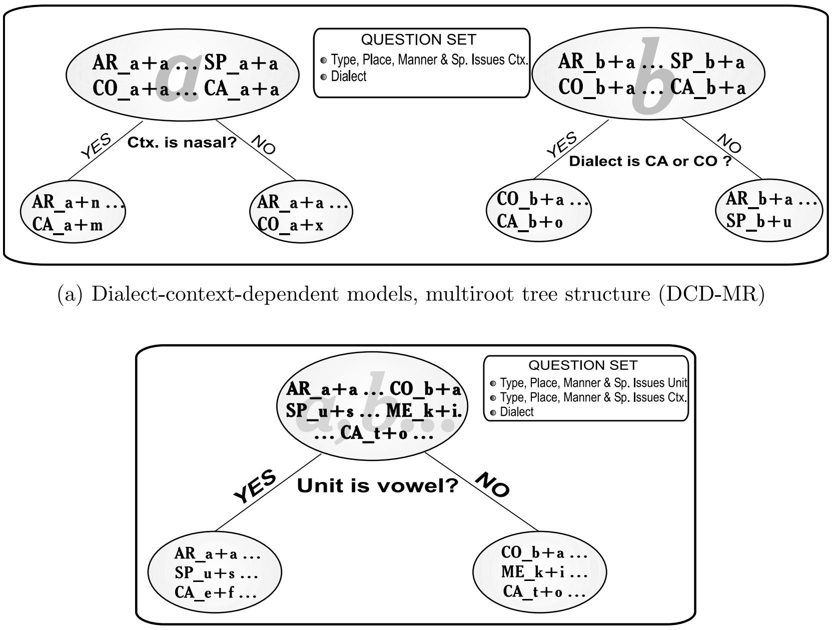
- **Dialect-context-dependent models, multiroot tree structure (DCD-MR)**: Dialect-dependent models are created for each contextual unit. Each root of the decision tree contains all the models whose phone is represented by the same SAMPA symbol. The question set asks for the context unit and the dialect. With this approach, similarity is only evaluated across models whose phone has the same SAMPA representation. It is possible to keep sounds represented by the same SAMPA symbol



(a) Global phone set, multiroot tree structure (GPS-MR)

(b) Global phone set, one-root tree structure (GPS-OR)

Fig. 1. Global phone set based approaches proposed to define a set of multidialectal acoustic models sharing data across dialects. (a) Approach using a multiroot tree structure. (b) Approach using one-root tree structure.



(a) Dialect-context-dependent models, multiroot tree structure (DCD-MR)

(b) Dialect-context-dependent models, one-root tree structure (DCD-OR)

Fig. 2. Approaches proposed to define a set of multidialectal acoustic models sharing data across dialects starting with dialect-context-dependent acoustic models. (a) Approach using a multiroot tree structure. (b) Approach using one-root tree structure.

across different dialects separated if they are not very similar. For example, when overall transcription is used, both /x/ for Spain and /h/ for Colombia are transcribed as /x/. This system is able to separate SP_/x/ from CO_ /x/ if they are really different, as they are assumed to be in canonical transcriptions.

- **Dialect-context-dependent models, one-root tree structure (DCD-OR)**:A single tree is grown with all the dialect-dependent models in the root node. The question set asks for the unit, the context unit, and the dialect. This approach gives maximum freedom to the clustering algorithm because no SAMPA restrictions are applied. As in the DCD-MR approach, models with the same SAMPA representation can be distinguished. The one-root tree structure allows models with distinct SAMPA representation to be joined if they are similar enough. This approach makes the system totally automatic and independent of prior phonetic knowledge. This approach seems to be able to solve errors at the transcription stage. For example, Caribbean canonical transcription assumes that /s/ in the coda position is uttered as /h/. If this is not true or it is superfluous to the acoustic models, some contextual models from Caribbean /h/ can be joined to /s/ contextual models from other dialects, and the error is thus solved. Using overall transcription, in which /L/ and /jj/ are kept separated, it is possible to cluster them if their acoustic models are really similar.

6. Experiments

This section presents the performance of the proposed methods for multidialectal speech recognition. A brief introduction of the in-house speech recognizer is first presented followed by the experimental set-up. A comparison of the four proposed multilingual approaches with both, canonical transcriptions and overall transcriptions is presented. One baseline recognizer was built for each dialect. The purpose was to compare the results to the multidialectal approaches proposed in this work. The performance of the dialect-independent systems is evaluated with a dialect not seen in the training data. This section ends with a discussion.

6.1. Recognition system

Our research was implemented in an in-house speech recognition system. The system is based on semicontinuous hidden Markov models (SCHMMs). Speech signals are parameterized using Mel Cepstrum and each frame is represented by its Cepstrum C, derivatives ΔC and $\Delta\Delta$C, and the energy derivative. The first three features are represented by 512 Gaussians and the energy derivative by 128 Gaussians.

The phonetic units are demiphones (Mariño et al., 1998), which are contextual units that model half a phoneme by taking into account its immediate context. A phone *ph* is modeled by two demiphones, '$l - ph$' '$ph + r$', where l and r stay for the left and the right phone context respectively. Each demiphone is modeled by a two-state left-to-right model. The main advantage of using demiphones instead of other contextual units such as triphones is the lower number of acoustic units that need to be trained, and consequently, they can be properly trained with small databases.

All the systems use decision trees clustering algorithms. For the tree to grow, the entropy of each node is computed. A discrete approximation is used to evaluate partitions. With this approximation, the entropy of a node A can be calculated using Eq. (1), where M is the number of models in the node, S is the number of states in each model, G is the number of Gaussians in the codebook, $f(m)$ is the frequency of the model in the train data, $f(s|m)$ is the quotient between the number of frames of the state s and the total number of frames of the model the state belongs to, and b_{sg}s are the mixture weights for each of the Gaussians in the codebook

$$H(A) = \sum_{m=1}^{M} f(m)\left[\sum_{s=1}^{S} f(s|m) \sum_{g=1}^{G} b_{sg} \log b_{sg}.\right] \tag{1}$$

Splitting stop criteria are defined by a minimum decrease in entropy and/or a threshold of the number of training examples in each cluster (leaf node). To train a demiphone, 100 training examples proved to be enough.

The question set inquires as to the phonetic features of the phonetic unit the model represents (type, place and manner), the dialect of the unit, and optionally non-phonetic questions (i.e. the position in the word, whether the phone is an aspiration, or whether the phone belongs to a consonant group). Compound questions about a single attribute (e.g. manner of articulation) are allowed using a logical OR link (e.g. *Is the manner in which the sound is articulated nasal OR fricative?*). The question set is completely dependent on the approach, and was discussed in Section 5.

6.2. Data

Experiments were carried out with databases recorded in Argentina, Colombia, Chile, Mexico, Spain and Caribbean (Venezuela). The databases consists of fixed network telephone recordings. Except the Caribbean, signals are sampled and recorded from an ISDN line at 8 kHz sampling rate, 8 bits/sample and coded using A-law. The Caribbean database was recorded from an analogue line and μ-law coded. The database from Spain contains speech from 4000 speakers. For the purposes of our research, 3500 speakers were selected for training and 200 for the test. The databases of Latin America contains speech from 1000 different speakers. For training purposes, 800 speakers were selected from each database, and 200 speakers were selected for the test. The canonical phonetic transcription applied to each database coincides with the predominant dialect spoken in each country, as mentioned in Section 2.2.4.

The systems were trained with a set of phonetically rich words and phonetically balanced sentences. In order to evaluate dialect-independent systems, one dialect, Chilean, was kept out of the training process. Six recognition tests were defined, one for each dialect. The recognition tests are composed of phonetically rich words. Each test speaker pronounces four of these words. Truncated and mispronunciated utterances were discarded from the test set. Each test utterance contains only one word, thus isolated word recognition language model is used. The vocabulary is identical for all dialects and has a size of 4500 words, containing all the words appearing in the tests. Table 8 shows the total amount of training and testing data for each dialect considered. Training data for Chile do not appear in this table as they did not participate in the training process.

6.3. Monodialectal systems: baseline recognizers

For comparison purposes, one baseline recognizer was built for each dialect. A multiroot decision tree based clustering algorithm was used for context modeling.

Table 9 shows the number of models of each monodialectal system. The number of models depends on the amount of training data and the phone set size because of the definition of the tree growing process. The system trained with data from Spain had the largest set of models, due to the larger amount of data available. The total number of models needed for recognizing all dialects was 3596. Table 9 also shows the percentage of word error rate (WER) for the baseline recognizers and their average value calculated as the mean value of the dialect WERs. The system for Spain gave the best result: 3.62%. Caribbean and Argentinean systems obtained around 7%. The Colombian and Mexican systems gave the worst rates.

Lower rates achieved for Latin American dialects shows a problem of data scarcity. In order to prove this, another baseline system was created for Spain with the same number of speakers as the rest of the dialects (800). The number of models created in this case was 736 and the WER obtained for the Spanish test was 6.00%. The results showed that even the WER for Spain was lower than the one obtained for other dialects using a similar amount of

Table 8
Training and testing data amount for dialects considered in this study

Dialect	AR	CA	CO	ME	SP	CH
Train. utterances	9568	9303	8874	11,506	40,936	–
Train. running words	412,859	425,591	476,559	558,884	956,300	–
Test. utterances	722	686	640	624	718	735

The Chilean dialect is not represented in the training process.

Table 9
Number of models and WER (%) for monodialectal systems developed for Argentina, Caribbean, Colombia, Mexico and Spain

Dialect	AR	CA	CO	ME	SP	Average WER%
Number of models	662	688	683	716	847	
WER%	7.34	6.71	9.22	10.10	3.62	*7.40*

data, there was a loss of nearly 50% in its performance comparing when using all the available data.

6.4. Multidialectal acoustic modeling

6.4.1. Using canonical transcriptions

The approaches proposed in this paper – GPS-MR, GPS-OR, DCD-MR and DCD-OR – were developed using a canonical transcription for each dialect. The GPS-MR and GPS-OR approaches used 988 and 981 models, respectively. These figures are comparable to the sizes of monodialectal sets, since the phone sets for these approaches are similar to the monodialectal ones. Dialect querying (the DCD-MR and DCD-OR approaches) made the decision tree grow to 3600 leaf nodes. In order to determine the optimal size of the acoustic model set, experiments with a variety of different task test sets were carried out. Different acoustic model set sizes (from 500 to the total number of lead nodes) were scanned, and the best results were obtained with 2000 models in both cases.

Table 10 summarizes the results of the experiments of the four proposed systems. All the systems improved the baseline average word error rate. In all the approaches presented, the performance of the dialect of Spain was slightly degraded. This result is not surprising since we added variability to a well-trained system. We consider that this degradation is acceptable as a minor drawback in order to achieve a multidialectal system.

The results for both GPS systems were similar. The improvement of baseline results in both cases was caused by the reduction of the WER in the Colombian, Mexican and Caribbean variants.

The DCD-MR approach improved the average performance achieved with the GPS-MR and GPS-OR approaches, as well as the baseline results. Using the one-root tree structure (the DCD-OR approach) led to the best system, as the average word error rate was reduced to 6.63%, a relative reduction of almost 7% over the baseline results. This system outperformed all the Latin American baseline results and the WER for Spanish as spoken in Spain was almost as good as for the dialect-specific system.

In addition to the results shown in Table 10, it has to be said that in most of the experiments carried out with sets of acoustic models of different sizes, the application of the one-root tree structure gave better results than the application of multiple roots. This leads us to conclude that the one-root tree structure, that is, a structure in which models can be shared between units, allows sharing data in a consistent way.

6.4.2. Using overall transcription

The approaches proposed were also developed using one overall transcription. These systems are referred to as OT. The GPS-MR$_{OT}$ and GPS-OR$_{OT}$ systems produced sets of 846 and 954 acoustic models respectively. The optimal size of the acoustic model set for approaches DCD-MR$_{OT}$ and DCD-OR$_{OT}$ was found to be 2000, the same size as using canonical transcriptions.

The recognition results obtained with these new systems are summarized in Table 11. The average result for the GPS-MR$_{OT}$ system was better than the baseline result, but looking at the dialect-specific rates in more detail, we can see that the results for three of the five dialects were worse than the monodialectal ones. Substantial improvement was achieved with the GPS-OR$_{OT}$ system. Using this system for Colombian and Mexican dialects, the WER was reduced by more than one and two points respectively, compared with the baseline results. The Caribbean rate was also improved, while rates for Spain and Argentina variants degraded only slightly. In addition, this system balanced the word error rate between dialects. It is also interesting to remark that this system outperformed the GPS-OR system trained using canonical transcriptions, with a lower number of acoustic models.

DCD-MR$_{OT}$ led to the best average result; it outperformed all of the systems presented in this paper. There was a relative reduction of 13% with respect to the monodialectal results. This system obtained the best results for Argentinean and Colombian dialects, and WERs pretty close to the best results presented in this work for Mexican and Caribbean variants. The result for the dialect of Spain was no better than the result for the well-trained baseline. DCD-OR$_{OT}$ also improved monodialectal rates, but contrary to what occurred with canonical transcriptions, the one-root structure used in combination with a dialect-dependent contextual models failed to improve the results obtained using a multiroot tree structure.

Table 10
Word error rate for multidialectal recognition systems that use canonical transcriptions

Dialect	GPS-MR	GPS-OR	DCD-MR	DCD-OR
No models	988	981	2000	2000
Argentina	8.31	7.76	6.37	6.23
Caribbean	6.27	6.27	6.41	6.41
Colombia	8.28	8.28	7.97	7.81
Mexico	8.01	8.17	9.62	8.65
Spain	4.74	4.6	4.46	4.04
Average	*7.12*	*7.02*	*6.97*	***6.63***

Table 11
Word error rate for multidialectal recognition systems using overall transcription

Dialect	GPS-MR$_{OT}$	GPS-OR$_{OT}$	DCD-MR$_{OT}$	DCD-OR$_{OT}$
No models	846	954	2000	2000
Argentina	7.89	7.76	5.68	6.23
Caribbean	6.85	6.56	6.71	6.71
Colombia	7.19	7.50	7.03	8.28
Mexico	8.50	7.70	8.66	9.29
Spain	5.44	4.04	4.17	4.04
Average	*7.17*	***6.71***	***6.45***	*6.91*

6.5. Evaluation of dialect-independent recognition systems in absence of dialect-specific training data

As commented in Section 5, approaches based on the definition of a global phone set are able to act as dialect-independent systems. However, to be able to recognize a dialect with no training data available using canonical transcriptions, no new phone can appear in the dialect to be recognized.

Chilean test data was used in this research to evaluate the performance of the dialect-independent systems developed. Even most of the phenomena of Chilean dialect are seen in other dialects, the Chilean test includes one new phone /C/, not included in the training data. This phone is present in 77 utterances of the Chilean phonetically rich words test, which means a percentage of 10.5% over the whole test. The presence of the /C/ phone did not allow to use GPS approaches developed using canonical transcriptions to recognize Chilean speakers. In systems that use overall transcription, /C/ phone is modeled with /x/ models trained with all the /x/ and /h/ realizations of other dialects. Thus, only the systems that use overall transcription (the GPS-MR$_{OT}$ and GPS-OR$_{OT}$ approaches) were able to recognize Chilean speakers.

The WERs obtained for the Chilean test with GPS-MR$_{OT}$ was 8.03% and with GPS-OR$_{OT}$ was 7.35%. These rates were similar to those obtained for Argentina, Colombia and Mexico using those systems. This result points out that these recognizers generalize across dialects; they provide models that are able to recognize speech in dialects that are not present in the training process.

6.6. Discussion

In this section, we compare the systems trained and analyze them in terms of data sharing and tree behavior.

6.6.1. Data sharing between dialects

Table 12 shows, for each approach, the percentage of full multidialectal (clusters containing data of all dialects) and semi-multidialectal (clusters containing data of more than one dialect) nodes. The percentages for the DCD systems were calculated for the 2000 acoustic model set. Maximum data sharing was provided by approaches that defined a global phone set. The total percentage of full multidialectal models for both GPS-MR and GPS-OR approaches was 70%. The GPS-OR approach slightly increased the data sharing percentage as it allowed dialect-specific models to be joined with other models in the same cluster. When overall transcription is used, all the units of the phone set are shared between dialects. Thus, the GPS-MR$_{OT}$ and GPS-OR$_{OT}$ approaches allowed to share 100% of the data.

The DCD systems allowed expressions of the same unit to be separated for different dialects. Opening the decision tree up to 2000 clusters decreased the percentage of full multidialectal nodes. When the one-root system was applied the percentage of semi-multidialectal nodes increased substantially, while the percentage of full multidialectal nodes decreased. The DCD-MR$_{OT}$ system had the highest percentage of data sharing between all the dialects and the best recognition results.

Table 12
Percentage of clusters that share data between dialects for multidialectal approaches

Approach	Full multidialectal clusters (%)	Semi-multidialectal clusters (%)
GPS-MR	69.23	20.65
GPS-OR	69.72	21.61
GPS-MR$_{OT}$	100.00	0.00
GPS-OR$_{OT}$	100.00	0.00
DCD-MR	6.80	11.20
DCD-OR	6.20	14.85
DCD-MR$_{OT}$	9.25	11.35
DCD-OR$_{OT}$	7.2	13.5

Full multidialectal clusters contain data of all dialects. Semi-multidialectal clusters contain data of few dialects.

To analyze the multidialectal clusters we calculated, for each pair of dialects, the percentage of clusters $P(d_i, d_j)$ that share data from those dialects. $P(d_i, d_j)$ as Eq. (2), where $N_c(d_i, d_j)$ is the number of clusters containing data from dialects d_i and d_j, and N_t is the total number of clusters

$$P(d_i, d_j) = \frac{1}{N_t} N_c(d_i, d_j) 100. \tag{2}$$

When these percentages were measured, all the clusters containing data from the two dialects were counted, even if there were other variants in it. Table 13 shows, for each DCD approach and dialect d_i, the top two scored dialects d_j and the calculated score $P(d_i, d_j)$. Colombian, Mexican and Argentinean variants appear in most of the cells of the table, pointing out that those dialects were present in the majority of the clusters, sharing data with the rest of the variants. On the other hand, the variant of Spain shares less clusters than the rest of variants. This fact indicates that Latin American dialects did not borrow an excessive amount of data from speakers of Spain, but rather just what they need. An interesting result is that overall transcription seems to overcome the barrier to data being shared established by canonical transcriptions, which specify different transcription rules for different dialects. The behavior of data sharing between Colombian and Caribbean dialects is an example. Table 13 shows that overall transcription allowed more data sharing between those dialects than canonical transcriptions. Actually, both dialects have speakers of the other variant. This behavior agrees with (Foldvik and Kvale, 1998), and validates the usefulness of the overall transcription.

6.6.2. Tree behavior

Concerning how the trees treat the models belonging to different phones, multiroot approaches begin with all the

Table 13
Percentage of clusters that share data from each pair of dialectal variants

Dialect	DCD-MR		DCD-OR		DCD-MR$_{OT}$		DCD-OR$_{OT}$	
AR	CO$_{16.1}$	ME$_{14.2}$	CO$_{15.6}$	ME$_{13.4}$	CO$_{20.2}$	ME$_{17.5}$	CO$_{16.8}$	ME$_{14.6}$
CA	CO$_{15.6}$	ME$_{13.2}$	CO$_{14.8}$	ME$_{12.4}$	CO$_{20.2}$	ME$_{17.3}$	CO$_{17.4}$	ME$_{14.2}$
CO	AR$_{16.1}$	ME$_{15.5}$	AR$_{15.6}$	ME$_{15.4}$	ME$_{20.4}$	CA$_{20.2}$	CA$_{17.4}$	ME$_{17.0}$
ME	CO$_{15.5}$	AR$_{14.2}$	CO$_{15.4}$	AR$_{13.4}$	CO$_{20.4}$	AR$_{17.5}$	CO$_{17.0}$	AR$_{14.6}$
SP	AR$_{10.8}$	ME$_{8.3}$	AR$_{10.6}$	ME$_{7.6}$	AR$_{12.6}$	CO$_{9.5}$	AR$_{11.0}$	CO$_{7.6}$

For each dialect and approach, the table shows the top two dialectal variants.

Table 14
Percentages of clusters shared by more than one phone for one-root tree structure approaches

GPS-OR	DCD-OR	GPS-OR$_{OT}$	DCD-OR$_{OT}$
6.6	10.80	8.5	15.3

units separated into different trees. The one-root tree structure allowed clusters to be shared by different phones. Table 14 shows the percentage of clusters shared by more than one phone. The approaches that use overall transcription allowed more clusters to be shared between different phones than canonical transcriptions based approaches. In all the one-root tree structure approaches, the units that were most frequently tied together were semivowels (/j/,/w/) and their corresponding vowel (/i/,/u/) when there was a lack of data for a given context. In the GPS-OR trees, these type of clusters were the most common. The DCD-OR approach also had clusters shared by fricative units /s/, /h/, and /T/, in coda position. Beside fricative units, GPS-OR$_{OT}$ also joined /n/_C with /N/ and /D/ with /R/ in coda position. It is remarkable that this tree separates /jj/ clusters from /L/ clusters, and /T/ clusters from /s/ clusters. DCD-OR$_{OT}$ had the largest percentage of clusters with more than one unit in them, which can cause a loss in acoustic resolution.

On the other hand, the experiments validate the use of the same symbol for allophonic variations across dialects in overall transcription if dialect information is given to the tree. With DCD-MR$_{OT}$, the system that obtained the best recognition results, the decision tree detects and clusters allophone variation between dialects. For example, the tree clearly separates velar /x/ from glottal /x/ and /s/ from aspirated /s/.

7. Conclusions

In this paper we compared several approaches to build a robust multidialectal set of context-dependent acoustic models for Spanish. The acoustic models were achieved by applying a decision tree clustering algorithm. Two tree structures were tested, multiroot, where there is a root for each considered SAMPA symbol and one-root, where the tree starts with a single root. To train the trees, two approaches were considered: dialect-independent models trained with data of all dialects and defined from a global phone set, and dialect-dependent models where each model is defined and trained with data from a single dialect.

To solve the necessity of using a canonical phonetic transcription per each dialect, a new approach, an overall transcription common to all the dialects has been proposed. The overall transcription has two advantages. First, it avoids knowing in advance the dialect of the speaker and, consequently, can be used to develop a multidialectal system to recognize speech from a broad number of dialects. Second, it overcomes errors in the phonetic transcription of the training databases due to a lack of knowledge of the dialect spoken per each speaker. The overall transcription has been successfully tested in four multidialectal approaches and compared with results obtained using a canonical transcription for each dialect. The recognition results with overall transcription overcome results with canonical transcriptions. Overall transcription is simpler and can be applied to recognize dialects non-seen during the training stage.

All the systems proposed improve monodialectal performance. Building the tree with dialect-dependent contextual models shows better performance than using context-dependent models defined from a global phone set. Concerning decision tree structure, in most of the experiments, one-root structure performs better than multiple root structure. The combination of dialect-dependent contextual models, multiroot structure and overall transcription outperforms all the other systems.

Overall transcription, in combination with the definition of a multidialectal global phone set led to a totally dialect-independent system with a reduced set of models. Its performance with one-root structure was nearly as good as the best found and it used half the number of models. This system is suitable for use with all Spanish speakers even if their dialect was not seen in the training phase

Acknowledgement

This work was granted by Spanish Government TEC2006-13694-C03.

References

Aalburg, S., Hoege, H., 2003. Approaches to foreign-accented speaker-independent speech recognition. In: Proc. Eurospeech, Geneva, Switzerland, pp. 1489–1492.

Baum, M., Muhr, R., Kubin, G., 2001. A phonetic lexicon for adaptation in ASR for Austrian German. In: ISCA Workshop Adaptation methods for Speech Recognition, Sophia-Antopolis, France, pp. 135–138.

Beringer, N., Schiel, F., Regel-Brietzmann, P., 1998. German regional variants – a problem for automatic speech recognition? In: Proc. ICSLP, Sydney, Australia, Vol. 2, pp. 85–88.

Billa, J., Ma, K., McDonough, J.W., Zavaliagkos, G., Miller, D.R., Ross, K.N., ElJaroudi, A., 1997. Multilingual speech recognition: the 1996 byblos callhome system. In: Proc. Eurospeech, Rhodes, Greece, pp. 363–366.

Brousseau, J., Fox, S.A., 1992. Dialect-dependent speech recognizers for Canadian and European French. In: Proc. ICSLP, Banff, Alberta, Canada, pp. 1003–1006.

Byrne, W., Beyerlein, P., Huerta, J.M., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D., Wang, W., 2000. Towards language independent acoustic modeling. In: Proc. ICASSP, Istanbul, Turkey, Vol. 2, pp. 1029–1032.

Caballero, M., Moreno, A., Nogueiras, A., 2004. Data driven multidialectal phone set for Spanish dialects. In: Proc. ICSLP, Jeju Island, Korea, pp. 837–840.

Chengalvarayan, R., 2001. Accent-Independent universal HMM-based speech recognizer for American, Australian and British English. In: Proc. Eurospeech, Aalborg, Denmark, pp. 2733–2736.

de la Torre, C., Caminero-Gil, J., Alvarez, J., Martín del Álamo, C., Hernández-Gómez, L., 1996. Evaluation of the Telefónica I+D natural numbers recognizer over different dialects os Spanish from Spain and America. In: Proc. ICSLP, Philadelphia, Vol. 4, pp. 2032–2035.

Diakolukas, D., Digalakis, V., Neumeyer, L., Kaya, J., 1997. Development of a dialect-specific speech recognizers using adaptation methods. In: Proc. ICASSP, Munich, Germany, pp. 1455–1458.

Duchateau, J., Demuynck, K., Van Compernolle, D., 1997. A novel node splitting criterion in decision tree construction for semi-continuous HMMs. In: Proc. Eurospeech, Rhodes, Greece, Vol. 3, pp. 1183–1186.

Ferreiros, J., Pardo, J.M., 1999. Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations. Speech Comm. 29 (1), 65–76.

Fischer, V., Gao, Y., Janke, E., 1998. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer. In: Proc. ICSLP, Sydney, Australia, paper 0233.

Foldvik, A.K., Kvale, K., 1998. Dialect maps and dialect research; useful tools for automatic speech recognition. In: Proc. ICSLP, Sydney, Australia, paper 0470.

Gibbon, D., Moore, R., Winski, R., 1997. Handbook of Standards and Resources for Spoken Language Resources. Mouton de Gruyter, New-York, ISBN 3-11-015366-1.

Heeringa, W., Gooskens, C., 2003. Norwegian dialects examined perceptually and acoustically. Comput. Humanities 37, 293–315.

Huerta, J.M., Thayer, E., Ravishankar, M., Stern, R.M., 1998. The development of the 1997 CMU Spanish broadcast news transcription system. In: DARPA BN Transcription Understanding Workshop, February 1998, Lansdowne, Virginia, USA.

Imperl, B., Kačič, Z., Horvat, B., Žgank, B., 2003. Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones. Speech Comm. 39 (3–4), 353–366.

Kirchhoff, K., Vergyri, D.M., 2005. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Comm. 46 (1), 37–51.

Köhler, J., 2001. Multilingual phone models for vocabulary-independent speech recognition tasks. Speech Comm. 35 (1–2), 21–30.

Kudo, I., Nakama, T., Watanabe, T., Kameyama, R., 1996. Data collection of Japanese dialects and its influence into speech recognition. In: Proc. ICSLP, 1996, Philadelphia, PA, USA, pp. 2021–2024.

Lipski, J.M., 1994. Latin American Spanish. Longmans Linguistics Library, New York, PC4821.l56.

Llisterri, J., Mariño, J.B., 1993. Spanish adaptation of SAMPA and automatic phonetic transcription. Report SAM-A/UPC/001/VI. February 1993.

Mariño, J.B., Pachés-Leal, P., Nogueiras A., 1998. The demiphone versus the triphone in a decision-tree state-tying framework. In: Proc. ICSLP, Sydney, Australia, Vol. I, pp. 477–480.

Mariño, J.B., Pradell, J., Moreno A., Nadeu, C., 2000. Monolingual and bilingual Spanish-Catalan speech recognizers developed from SpeechDat databases. In: Proc. Internat. Workshop on Very Large Telephone Speech Databases, Athens, May 2000, pp. 57–61.

Moreno, A., Mariño, J.B., 1998. Spanish dialects: Phonetic transcription. In: Proc. ICSLP, Sidney, Australia, paper 0598.

Moreno, A., Höge, H., Köhler, J., 1998. SpeechDat Across Latin America. Project SALA. In: Proc. Internat. Conf. on Language Resources and Evaluation (LREC), Granada, Spain, Vol. I, pp. 367–370.

Nogueiras, A., Caballero, M., Moreno, A., 2002. Multidialectal Spanish speech recognition. In: Proc. ICASSP, Orlando, USA, paper 2928.

Salvi, G., 2003. Accent clustering in Swedish using the Battacharya distance. In: Proc. 15th Internat. Congress of Phonetic Sciences, Barcelona, Spain.

Schultz, T., Waibel, A., 2001. Language independent and language adaptative acoustic modelling for speech recognition. Speech Comm. 35 (August), 31–51.

ten Bosch, L., 2000. ASR, dialects, and acoustic/phonological distances. In: Proc. ICSLP, Beijing, China, Vol. 3, pp. 1009–1012.

Yu, H., Schultz, T., 2003. Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition. In: Proc. Eurospeech, Geneva, Switzerland, pp. 1869–1872.

Zissmanm, M.A., Gleason, T.P, Rekart, D.M., Losiewicz, B.L., 1996. Automatic dialect identification of extemporaneous, conversational, Latin American Spanish Speech. In: Proc. ICASSP, Atlanta, GA, USA, pp. 777–780.