Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

From large-scale phonetic studies to speech recognition of Spanish varieties

Master

Natural Language Processing

Speciality :

Multilingual Engineering

par

Nidia Hernández

Thesis Director :

Cyril Grouin

Supervisors :

Bianca Vieru
Ioana Vasilescu

2016/2017

Contents

List of Figures

List of Tables

Abstract

Dialectal variation represents a major challenge for automatic speech processing. The purpose of this research is to improve the performance of a broadcast news transcription system for Latin American Spanish. Automatic speech processing tools were employed to estimate the impact of intervocalic /b/ /d/ /g/ and coda /s/ lenition across Spanish dialects. These findings have been applied to the acoustic model training together with modifications of both the phonemic inventory and lexicon. The effect of extending the training material with dialect-specific data was also studied. Two acoustic model training configurations were compared: an initial set with Peninsular data exclusively and an extended dataset adding Latin American data. The best performing model for Latin American speech includes expert corrections, consonant merge and lenition with the extended dataset. This model obtains a 7% relative gain in WER for Latin American data and remains unchanged for the other Spanish dialects.

Key words: *automatic speech processing - Latin American Spanish varieties - consonant lenition - multi-dialectal pronunciation modeling - dialect-specific extended dataset*

INTRODUCTION

Speech recognition allows spoken language to be automatically converted into written texts. Automatic speech recognition (ASR) is used for transcribing large quantities of audio and video documents and it is also the technology behind muti-modal human-machine interactions such as passing commands to phones and smart home devices. Over the last years speech recognition optimization has been a key objective of major technology companies and the use of digital assistants and voice search continues to grow every day [1].

Most of the ASR systems are language specific so dialectal variation represents a major challenge. The purpose of this research is to improve automatic transcriptions for the Latin American variety of Spanish. This work was done at Vocapia Research, a French R&D company founded in 2000 that is specialized in speech processing. This company has a close working relationship with LIMSI, a French CNRS laboratory with more than 40 years of experience in the field. This special collaboration with LIMSI helps Vocapia to have a rapid take-up of the latest advances in speech recognition research. It is also noteworthy pointing out Vocapia's international profile: it participates at several international projects together with French and foreign partners and its speech-to-text transcription software, VoxSigma, is available for multiple languages.

The Spanish version of VoxSigma has state-of-the-art accuracy when processing speech data from Spain but its performance is less good on Latin American data. In order to improve the results for Latin American Spanish, two complementary directions were followed in this research: first, a study of the impact of Spanish consonant lenition on automatic processing of Spanish spoken data, and second, a specific pronunciation modeling based upon the results of the first study.

This work focuses on Latin American Spanish, therefore Spanish from other countries is not studied with the exception of Peninsular Spanish (i.e., Spanish from Spain) which was taken as a reference point because Vocapia's current system targets mainly this variety. However, the relevance of the Latin American Spanish is undeniable: this variety is spoken by more than 450 million speakers, which is ten times more than the number of Peninsular Spanish speakers.

The rest of this work is organized as follows: Chapter chapter 1 gives an overview of the Spanish language phonetics and phonology and dialectal classifications; Chapter chapter 2 presents the basics of automatic speech recognition systems and examines the main difficulties for automatically transcribing Latin American speech; Chapter chapter 3 is devoted to a large-scale study of consonant reduction in Spanish varieties; the details of the experiments for ASR system adaptation to Latin American Spanish are discussed in Chapter chapter 4. Finally, a summary and discussion of the work carried out as part of this research project can be found in Chapter chapter 5.

[1] https://googleblog.blogspot.fr/2014/10/omg-mobile-voice-survey-reveals-teens.html

CHAPTER 1

Spanish around the world: an analysis of linguistic specificities

Spanish is the mother tongue for more than 470 million people which places it as the second mother tongue of the world. If Spanish as a second language is also considered, the number of Spanish speakers grows to 570 million people. It is the official language of 21 countries[1] and 18% of the population of the United States speaks Spanish. It is the third most used language on the Internet and the second most used language on Facebook and Twitter[2].

As any widespread language, Spanish presents regional varieties, the most important ones for economic, historic and demographic reasons being Peninsular and Latin American Spanish. In spite of the large number of speakers of Latin American Spanish, the literature agrees in emphasizing its homogeneity, specially in standard register [de la Concha et al., 2017] and even more in the media [Ávila, 2001]. Media Spanish (also known as "español neutro", "soap opera Spanish" or "international Spanish") is an artificial variety of Spanish created by Mexican movie producers [López González, 2002]. It must not be confused with academic norm, standard register or general use Spanish (the Spanish taught as a foreign language). Media Spanish is used by TV and radio presenters and for subtitling and its use is very widespread in Latin American media, especially on international TV channels. [Ávila, 2001] made a statistical study on the lexicon of the Spanish-speaking media and concluded that the language used by these media employs few regionalisms (not more than 1.2%) and had a tendency to lexical convergence. As for the phonological characteristics of the Media Spanish, while there is a preference to use Mexican pronunciation in international channels, the national standard pronunciation is used on national channels.

1.1 Spanish phonological system and orthography

According to [Real Academia Española, 2009], the Spanish phonological system has between 22 and 24 phonemes (vowels and consonants included) depending on the subsystem, *seseante* or *distinguidor*. The main difference between subsystems is the presence or absence of the phoneme /θ/: while the *seseante* subsystem neutralizes the distinction between /s/ and /θ/, the *distinguidor* subsystem keeps it. The second most

[1] For more detailed numbers on Spanish-speaking countries see **??**

[2] The data cited on this paragraph were taken from the *Instituto Cervantes'* annual report on the situation of the Spanish language in the world [de la Concha et al., 2017].

	Bilabial	Lab. dent.	Dental	Alveolar	Pal-alveo.	Retroflex	Palatal	Velar	Uvular	Pharyngyl	Glottal
Plosive	p b			t d			c	k g			ʔ
Nasal	m			n			ɲ	ŋ			
Trill				r							
Tap/Flap				ɾ							
Fricative	β	f v	θ ð	s	ʃ ʒ	ʑ	ç ʝ	x ɣ			h
Lat. Fric.											
Approx.							j				
Lat. appr.				l			ʎ				

Table 1.1 – Spanish consonant system. Highlighted symbols represent dialectal phones

important difference is the presence or absence of the phoneme /ʎ/. The absence of /θ/ and /ʎ/ are traditionally known as *seseo* and *yeísmo* respectively. These tendencies are preferred by the majority of the Spanish-speaking countries. Table 1.1 shows the articulatory characteristics of the Spanish consonantic system. The 19 consonants of both subsystems are included in this table as well as the contextual and dialectal allophones that will be mentioned in this work.

The Spanish vowel system is quite simple from a phonological point of view [Real Academia Española, 2009], having only five vocalic phonemes: /i/, /e/, /a/, /o/, /u/. The actual realizations of these vowels can be subjected to nasalization, lengthening, reduction and other phenomena that will not be developed in this work.

The Spanish alphabet derives from the Latin alphabet and has 27 letters. Its most particular characters are *ñ* and the accentuated vowels (*á*, *é*, *í*, *ó*, *ú*). Spanish orthography is overall regular, i.e. there is an almost one-to-one correspondence between letters and phonemes. There are few double letters, which are *ch* (pronounced /c/), *ll* (pronunciation varies dialectally) and *rr* (pronounced /ɾ/). The main spelling difficulties in Spanish are:

- the presence of the grapheme *h* that it is not pronounced: *hotel* /otél/, *habilidad* /abilidád/
- the distinction between *b* and *v* which is merely etymological since they are both pronounced /b/: *cabo* /kábo/, *cavo* /kábo/
- the contextual pronunciation of *g*. The grapheme *u* is required before *e* and *i* for *g* to be pronounced as /g/. Otherwise *ge* and *gi* are pronounced the same as *je* and *ji* (/xe/, /xi/): *guerra* /géra/, *guitarra* /gitára/
- the contextual pronunciation of the digraph *qu*. /u/ is not pronounced before *e* and *i*: *queso* /késo/, *química* /kímica/. On the contrary, /u/ is pronounced when preceding *a* and *o*[3]: *quorum* /kuórum/.
- the transformation of *z* into *c* before *e* and *i*: *vez* → *veces*

[3] These sequences are rare since they only exist in Spanish in loan words.

- in the *seseante* varieties, the non distinction between *z* and *s*, both representing the phoneme /s/

- in the varieties manifesting *yeísmo*, the undifferentiated realization of *ll* and *y*

1.2 Latin American Spanish varieties

Linguists agree to say that the standard register of Latin American Spanish is quite homogeneous. However, on the one hand, it is difficult to find linguistic features that are present in all Latin American countries, on the other hand it is hard to find phenomena that are not also manifested in Spain. The few phonological phenomena common to all Latin American countries, the aforementioned *seseo* or the weakening of /s/ at end of syllable (cf. infra) are also present in the Andalusian region of Spain while the *yeísmo* is not practiced in the Andean variety.

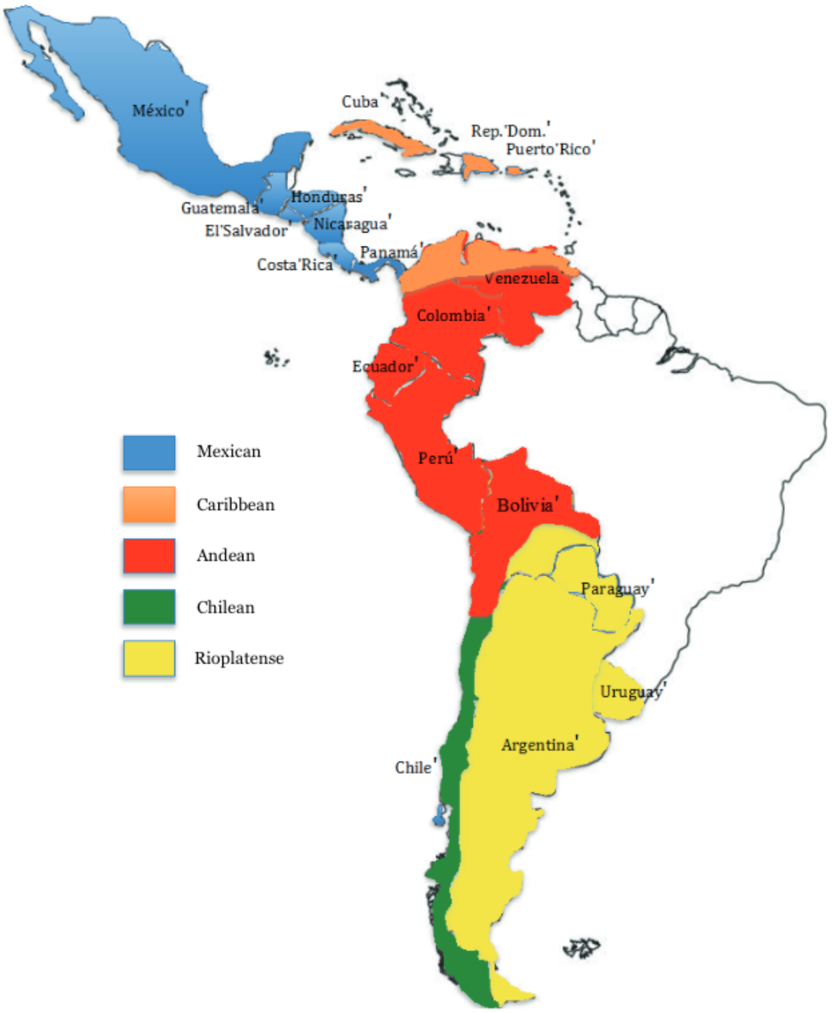

Figure 1.1 – Latin American Spanish varieties. Taken from [Quesada Pacheco, 2014]

[Quesada Pacheco, 2014] presents a historical panorama of Latin American Spanish dialectal classifications. One of the most cited classifications on the literature (illustrated on Figure 1.1) proposes five dialects:

1. Andean: this variety gathers Bolivia, Ecuador, Peru, South of Colombia and South of Venezuela. A typical (but not exclusive) sound change typical of the Andean Spanish variety is the *assibilation*, i.e. the spirantization of the trilled sound and the tap sound after /t/ [Fontanella de Weinberg, 1992] [Real Academia Española, 2009]. For example, the words *carro* and *tres* will change as follows: ['ka.ro] → ['ka.ʐo] and ['tɾɛs] → ['tʐɛs].

2. Caribbean: this group includes Cuba, Dominican Republic, Puerto Rico, Northern Colombia and Northern Venezuela. This dialect is characterized by the *rhotacism*, which consists of exchanging /r/ for /l/ especially on syllable coda [Fontanella de Weinberg, 1992] [Real Academia Española, 2009]: *invierno* → [in'bjelno]. Another typical feature of the Caribbean Spanish is the velarization of /n/[Real Academia Española, 2009]: *canción* → [kaŋ'sjoŋ].

3. Mexican: is the variety spoken not only in Mexico but also in the South of the United States and all continental Central American countries. Mexican Spanish is unanimously appreciated and seen as a "neutral" accent by speakers of all Spanish varieties [Quesada Pacheco, 2014].

4. Rioplatense: is the Spanish spoken in Argentina, Paraguay and Uruguay. It is easily identified by its particular *yeísmo* instantiated as a voiced palatoalveolar fricative [Fontanella de Weinberg, 1992]: *calló* → [ka'ʒo], *cayó* → [ka'ʒo].

5. Chilean: this variety is perceived as a separate kind by its own speakers and by other Spanish-speakers as well [Quesada Pacheco, 2014]. Phonologically, it can be recognized by the palatalization of /x/ before /i/, /e/ or /j/: [Real Academia Española, 2009]: *teje* → [teçe].

Another well known dialectal classification opposes "Low Lands", the Caribbean and the coast of the continent, to the "High Lands", the elevated inland territories [4]. The advantage of this classification is that it allows to account for similarities across countries, namely the realization of /s/ in coda position, which is weakened in the "Low Lands" and conserved in the "High Lands".

1.3 Studies of Spanish consonant reduction

Many corpus linguistics studies have investigated /s/ reduction in Spanish. It should be noted that consonant reduction, also called consonant lenition, is a gradient variation that can be manifested as duration reduction, voicing, aspiration and deletion [Ryant and Liberman, 2016].

In an acoustic study of the variation of /s/ in Peninsular Spanish, [Hualde and Prieto, 2014] find that Spanish word-final intervocalic /s/ is weaker: it is shorter than initial and medial /s/ and 12% of its realizations are voiced. The authors also point out that "in Madrid Spanish aspiration of /s/ is not uncommon before certain consonants, but it is infrequent in the prevocalic /Vs#V/ context". This phenomenon is signaled as a difference with Andalousian and Latin American Spanish. In these varieties, speakers not only aspirate or delete before consonant as in /éste/ [éhte] (*this*) or /dós tóɾos/ [dóhtóɾoh] (*two bulls*) but also in word final before vowel

[4] According to Menéndez Pidal [Quesada Pacheco, 2014], this distribution is the consequence of colonization issues: the Castillan bureaucrats, who kept the /s/, were installed on the inland capitals while the Andalusian sailors, who had a tendency to drop the /s/, generally stayed on the coasts.

/dós animáles/ [dóhanimáleh] (*two animals*). Nevertheless, the authors do not provide statistics on this point as their study analyzes Catalan and Madrilian data only.

[Ryant and Liberman, 2016] investigate Spanish /s/ lenition in a large-scale acoustic study of audiobooks. Their corpus comprises readings by Chilean, Argentinian, Mexican, and Peninsular speakers. The Caribbean variety is also studied but on a broadcast news corpus. The authors find evidence of /s/ reduction in the reading of the Mexican and the Chilean speakers. The first speaker voices /s/ to /z/ routinely before voiced consonants and the second's pre-silence /s/ is much shorter than initial or medial. In spite of the amplitude of their corpus (86h, 760k words, 300k /s/ realizations) and the simplicity to build it, the representativity of this study is mitigated by the following facts: 1. consonant retention is higher in read speech and formal genres, 2. there is only one speaker per variety, and 3. results are not shown for Argentinian, Caribbean and Peninsular varieties.

Another ineluctable phenomenon of Spanish phonetics is the weakening of the voiced stops /b/, /d/ and /g/ in intervocalic and syllable-final position. These phones are systematically realized as /β/, /ð/ and /ɣ/ when they are between vowels and sometimes when in postconsonantal position. This is the result of a historical process and it is extended to all Spanish varieties and all speaking styles [Chitoran et al., 2015].

Despite being a textbook example, there are fewer corpus-based studies of /b/, /d/ and /g/ lenition than for /s/. Chitoran et al. state that "there is substantial variation among speakers in the weakening of intervocalic voiceless stops" but they do not present statistics. [Hualde et al., 2010] study /d/ weakening acoustic and articulatory parameters in detail but acknowledge the need for confirmation of their discoveries on more speakers. An exception to this lack of quantitative evidence is the study of Spanish varieties performed by [Moreno and Mariño, 1998]. These authors report a relative frequency of /b/, /d/ and /g/ as approximants in their Colombian samples almost doubling that of Peninsular samples. However, these measurements are made on a controlled corpus of 9 sentences for each variety.

CHAPTER 2

SPEECH RECOGNITION CHALLENGES

This chapter presents an overview of statistical automatic recognition systems and the main metric for evaluating their performance. The literature dedicated to ASR agrees in the negative influence of dialectal variation in recognition accuracy. The effects of mismatch between train variety and target variety have been studied in several languages such as Arabic [Nallasamy et al., 2011], English [Najafian and Hansen, 2016], [Vergyri et al., 2010] and French [de Mareüil et al., 2013], [Vieru et al., 2011]. A synopsis of different approaches to overcome these issues in Spanish ASR is included in this chapter together with some examples of the main difficulties in transcribing Latin American Spanish speech.

2.1 Speech recognition systems

Although current speech-to-text systems achieve a high level of accuracy, challenges for reaching human proficiency remain: the variability in the speech signal, the lexical creativity and the synchronous processing [Huang et al., 2014] are among the most cited factors. State-of-the-art speech recognition systems are based on statistical modelization of speech. Learning from large corpora, today systems are speaker-independent and capable of large-vocabulary continuous speech recognition (LVCSR).

2.1.1 Components and functioning

Figure 2.1 schematizes the main features of the structure of an automatic speech recognition system. The upper section shows the training stage which takes as learning source large-scale audio, associated transcriptions and text corpora. The lower part represents the resulting ASR system that will be capable of decoding new audio data. ASR systems have three main components: a language model, a lexicon and acoustic models.

The language model estimates the probability of a sequence of words. This probability is obtained from large normalized text corpus using probability distribution. This means that grammatical constraints are represented in the system as n-grams with associated frequency probability.

The lexicon is a set of word entries linked to phonetic pronunciations. Words are stored in the lexicon as orthographic representations that will be used for the output

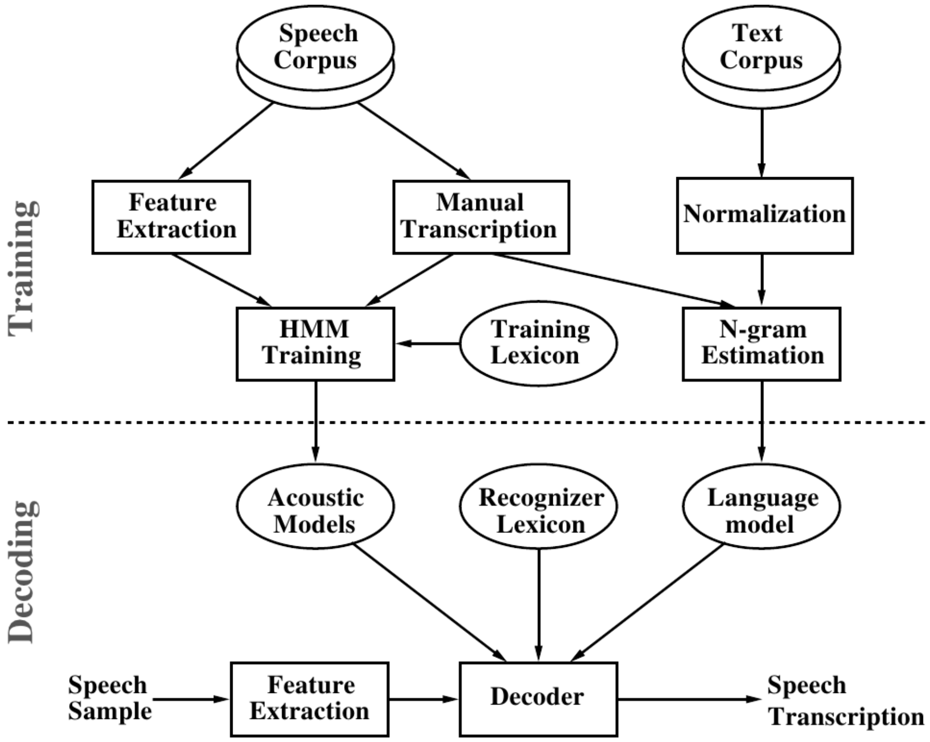

Figure 2.1 – Speech recognition system. Taken from [Lamel and Gauvain, 2003]

transcriptions. Pronunciations are represented using phone sets specific for each language.

The acoustic models represent the acoustic characteristics of a given language. The most common representation is phone based. Other representations are possible but "phone based offer the advantage that recognition lexicons can be described using the elementary units of a given language, and thus benefit from many linguistic studies" [Lamel and Gauvain, 2003]. Phones are modeled on their phonetic context and can be adapted to speaker. Each model is represented by a Markov chain.

The training stage aims to reduce audio signal variability while getting most relevant linguistic information. Prior to acoustic training, the portions of the audio signal containing speech are differentiated from those containing music or noise, that are discarded. This step is known as audio partitioning. Once the speech segments have been identified, acoustic features can be extracted. The speech signal is cut into 20-30ms window frames and in each of these frames, feature coefficients are extracted into an acoustic vector. The HMM phone models are obtained with Deep Neural Network from these acoustic vectors.

The decoding stage consists of finding the most probable word sequence given an acoustic signal. The system extracts the acoustic features from new speech data, identifies the phones and searches for the most probable words for the phone sequence. In this manner, the decoder produces a transcription hypothesis based on the trained language and acoustic models.

It becomes clear that speech processing is a complex task. Despite the steady progress in the domain, there is still place for improvement in aspects that the human listeners manage easily. Dialectal variations is one of them. In respect of Spanish language in particular, the performance of an ASR system can degrade more than 40% relative when exposed to Latin American varieties [Nogueiras et al., 2002] [Elfeky et al., 2015].

2.1.2 Evaluation metric

Automatic speech recognition systems are evaluated in terms of Word Error Rate (WER), a measure based on the number of substitutions, insertions, and deletions produced on the automatic transcription (hypothesis) compared to a human transcription (reference).

A substitution error (S) is produced when the targeted word is replaced by another word. Insertions (I) are words proposed by the system that do not have a correspondence in the reference. Finally, deletion errors (D) occur when one or more words of the reference are missing in the hypothesis. The counting of these errors is used, together with the total number of words in the reference (N), to calculate the WER as follows:

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \tag{2.1}$$

The formula underlines that a lower WER corresponds to a good performance of a given system.

2.2 A state-of-the-art of ASR for Spanish dialects

Several papers study the mismatch in speaker-dependent parameters between training and testing data. [Vergyri et al., 2010] and [Najafian and Hansen, 2016] find a negative influence of dialectal speech on English recognizers, the first for a British English based system and the second for a system trained on US data exclusively. Both studies present significant reduction of errors after extending the training set on accented data. Non-native speech recognition also represents an issue for such systems [Gruhn et al., 2011]. [Mihaylova, 2011] reports over 25% improvement through lexical and acoustic adaptation to Bulgarian accented English. As for French, a characterization of foreign accents using ASR techniques is proposed by [Vieru et al., 2011] to improve pronunciation modeling. [de Mareüil et al., 2013] analyzes the impact of French dialectal pronunciations on automatic speech processing.

Facing the challenge of multi-dialectal languages, several solutions can be adopted: conceiving specific recognizers for each dialect, trying the suitability of a dialect specific system on another dialect (cross-dialect recognition) or building a single system capable of recognizing all dialects.

The first option is explored by [Elfeky et al., 2015], who analyzes the performance of Google's multi-dialectal speech recognition systems for voice search on Arabic, English and Spanish. Google speech recognition system has five recognizers for Spanish: Latin American (multi-dialectal), Mexican, Argentine, Spanish and US Spanish. After testing each system on cross-dialect recognition, the authors conclude that the Latin American recognizer should be discarded and advocate for cross-dialect processing of voice search in Spanish using the Mexican recognizer.

[Nogueiras et al., 2002] and [Caballero et al., 2009] provide results in favor of the opposite approach. Based on the quantitative analysis of a previous study [Moreno and Mariño, 1998], [Nogueiras et al., 2002] conceive specific pronunciation models for four Spanish variants (Peninsular, Andean from Bogota, Caribbean from Caracas and Rioplatense from Buenos Aires). Among other dialectal pronunciations, the authors consider the following phenomena for pronunciation modeling:

- all Latin American variants implement the *seseo* and the *yeismo*
- /s/ lenition in postnuclear position for Caribbean and Rioplatense
- voiced stops lenition in post-nuclear position or in the onset of a syllable following a vowel for Andean variety

They also propose a single phonetic inventory composed of phones common to all dialects and completed by the phones that are not shared between dialects. In this manner, 4 mono-dialectal recognizers are trained with the dialectal-specific pronunciation models and a multi-dialectal model system is trained with the global set of 32 phones. Decoding experiences performed by the authors show that: 1) in both mono-dialectal and cross-dialectal experiences the best performance is obtained by the system of Spain since it has been trained with more data than the others; 2) a single multi-dialectal system can process all the Spanish dialects satisfactorily.

Robustness of Spanish multi-dialectal systems is also studied by [Caballero et al., 2009]. This work focuses on acoustic modeling exploring different methods for combining training data based on decision tree clustering algorithms. It also measures the impact of knowledge-based adaptation testing specific and global phone sets. Acoustic models are trained with audio data from Argentina, the Caribbean, Colombia, Mexico and Spain with two transcription options (dialect-specific transcriptions and an overall transcription). The mono-dialectal experiments show best results for Spain, again thanks to a bigger training dataset. The data sharing analysis shows that Colombian, Mexican and Argentinian variants appear in the majority of the clusters, which means they share data with the rest of the variants while the variant of Spain shares less clusters with the rest of variants. As for the multi-dialectal experiences, the average performance is better but the score for Spain is slightly degraded and the global phone set allowed more data sharing between 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.

In light of these studies, several adaptation methodologies are possible. There is no definitive evidence on the best cross-dialectal performance; depending on the study, the best choice could be Mexico or Spain. Using a single multi-dialectal pronunciation and acoustic modeling for all the Spanish variants spoken in Europe and Latin America seems promising and simpler: this allows one database to train, a single system to decode and eliminates the dialect detection module which is necessary in configurations having multiple mono-dialectal recognizers. Moreover multi-dialectal models are more robust to errors. The main constraint for this approach is data scarcity since quality continuous speech corpora are not available for all Latin American countries. In fact, it should be noticed that [Nogueiras et al., 2002] and [Caballero et al., 2009] WER rates under 5% are obtained on isolated word recognition.

2.3 Error typology

As [Moreno and Mariño, 1998] state, “Dialectal variations influence all the steps in any speech recognition system”. The analysis of errors produced when processing Latin American data can lead to the prioritary aspects to modify. In this purpose, this section is dedicated to an overview of the ASR errors due to dialectal specifities. Decoding errors classification follows two criteria: **general errors** that are dialect independent, and **dialectal origin errors** that are caused by the particularities of the Latin American Spanish.

2.3.1 General errors

A significant amount of decoding errors does not depend on the dialectal particularities of the decoded speech. They can be caused by signal problems (noise, bad audio quality, overlapping speech, etc.), by evaluation restrictions or by inconsistencies in the reference. We can group such errors in three classes:

1. Spelling variation

   Some words have more than one spelling accepted: *mexicano/mejicano*, *transportar/trasportar*, etc. In general, these variations are used in all the Spanish speaking countries, there is not a dialectal distribution as in English. In the example below, the form used in the reference and the one proposed in the system hypothesis are both correct (but the system output corresponds to the form having more resemblance with the actual pronunciation of the speaker). Both forms are equivalent so this output should not be penalized on the evaluation. A specific mapping of these equivalent forms needs to be implemented for these cases.

   ```
   REF: QUIZÁS algunos murieron
   HYP: QUIZÁ   algunos murieron
   ```

2. Normalization problems

   Before evaluating the output of an automatic speech recognition system, the manual transcriptions serving as reference need to be normalized otherwise, automatic transcriptions as the one below will be considered as substitution errors. Normalization consists mainly of recasing words, expanding abbreviations and URLs, removing main punctuation signs and decomposing numbers. These processes are language-specific since each language has its own conventions. For instance, in Spanish ordinal numbers are written in roman but only when they refer to centuries:

   ```
   REF: relación a la celebración del V      centenario
   HYP: relación a la celebración del QUINTO centenario
   ```

3. Transcription inaccuracy

   Accurate transcriptions are crucial for improving speech recognition since they are the reference to measure the performance of the system. Nevertheless manual transcriptions themselves are subject to errors too: spelling mistakes, unsystematic spellings (*Abdallah*, *Abdala, Abdalah*), lack of specific guidelines.

For example, our Latin American monologues corpus is meant to be a support document for Spanish learners thus it tends to privilege the linguistic norm over the fidelity to the real utterances eliminating hesitations, repetitions, anacoluthons, etc. This resulted in an artificially high penalization of the system therefore these transcriptions had to be manually adapted to fit ASR evaluation.

In the example below, for instance the elimination of the stammering resulted in insertions, thus increasing the WER:

```
REF: ** ** ** ** es la gran fiesta ** ** de mendoza
HYP: ES ES ES LA es la gran fiesta DE DE de mendoza
```

2.3.2 Dialectal origin errors

As [Caballero et al., 2009] point out, "dialectal variability is a significant degrading factor in automatic speech recognition" so if an ASR system is trained for Peninsular Spanish, it is not unexpected for it to underperform in recognizing speech belonging to other varieties of Spanish. The decoding errors of an ASR system can be more related to its acoustic component or its linguistic component. The following paragraphs present some samples of errors caused by the dialectal particularities of the Latin American varieties.

1. Phonetic variation

   As seen in the introduction, Latin American Spanish presents specific phonetic processes that may introduce **new sounds**, for instance /ʐ/ and /ŋ/ resulting from the assibilation of /r/ and the velarization of /n/ respectively, or the wide range of fricative sounds: /ç/ (allophone of /x/), /ʃ/, /ʒ/ (allophones of /ʝ/ and /ʎ/) and /h/ (allophone of /s/ and /ð/). These phones are absent in the data used to train the acoustic model so the system is not always able to recognize the words containing them.

   In other cases, the problem is not posed by new sounds but by **neutralizations**, which lead to the presence of one phone where another one is expected. That is the case of the previously described *yeísmo* and *seseo* which affect respectively 3% and 16% of the system vocabulary. These neutralizations are often the source of substitution errors during the automatic transcription of Latin American samples.

   ```
   REF: ALLÁ LAS   COLUMNATAS        REF: ZORROS
   HYP: **** AYALA COLUMNATA         HYP: SORDOS
   ```

   The *seseo* is also at the origin of several spelling mistakes. Sometimes, the output transcriptions present spelling mistakes as the one below. This is a consequence of the statistical building of the language model: since spelling mistakes occur on the source corpus used for the language model construction, when these mistakes have high frequency, they are integrated to the language model and the pronunciation lexicon. Then, given that the Latin American pronunciation is closer to the form having the spelling mistake, this form is selected for the decoding output.

   Finally, the **reduction** or elision of a sound can also have an adverse impact on speech recognition. The cited weakening of /s/ in final or preconsonantic position

Corpus	Substitution frequency (%)
Latin American monologues	7.3
Latin American prepared	3.8
Penninsular Spanish	2.6

Table 2.1 – Substitution errors due to final /s/ lenition

is a widely spread phenomenon in many regions of Latin America. Since in Spanish the plural mark for articles, nouns and adjectives is a suffix *-s*, this reduction has a considerable impact on the meaning of the transcribed speech.

```
REF: PRINCIPIOS que él señaló EN SUS DISCURSOS en SUS CARTAS
HYP: PRINCIPIO  que él señaló ** SU DISCURSO  en SU  CARTA
REF: SUS documentos distintos SUS DIFERENTES escritos
HYP: SU  documentos distintos *** DIFERENTE  escritos
```

It should also be noticed that many high frequency words like pronouns (*lo* / *los*, *le* / *les*), articles (*una* / *unas*) and auxiliary verbs (*ha* / *has*) are affected by this phenomenon, which degrades the results not only in quality but in quantity: 37% of the system's vocabulary is liable to *s* lenition and in fact an average of 5.6% of the substitution errors on Latin American decodings are the consequence of a missing final *s* in the hypothesis (see 2.1).

As explained in the introduction, another phonetic reduction well known in Spanish linguistics literature is the pronunciation of the voiced stops /b/, /d/, /g/ as the fricatives /β/, /ð/, /ɣ/. These weaker variants can be difficult to recognize by an ASR system, as suggest the following errors:

```
REF: con ellos también ESTÁBAMOS en el lugar
HYP: con ellos también ESTAMOS  en el lugar


REF: PUEDES meter el dedo perfectamente
HYP: PUES  meter el dedo perfectamente
```

All Spanish participles contain intervocalic /d/ and, as in the previous example, intervocalic /b/ makes the difference between past and present tense in many verbs. Intervocalic /b/ /d/ /g/ represent 22% of the system's vocabulary.

2. Other levels of variation

   Besides pronunciation, Latin American Spanish differs from Peninsular Spanish in respect of its morphological system and its lexical inventory, amongst other linguistic levels. The most remarkable difference in grammar is the *voseo*, the variant for the 2nd person pronoun and conjugation existing in many Latin American countries: *vos pensás*, *vos querés*, *vos decís* (instead of *tú piensas*, *tú quieres*, *tú dices*). This difference disrupts the verbal paradigm affecting a wide portion of the vocabulary. The systematicity of the errors in the transcription of verbs shows the relevance of this phenomenon:

```
REF: VOS  PRODUCÍS CHIVITOS  PRODUCÍS  conejos o cualquier
HYP: SIGO PRODUCIR QUINIENTOS PRODUCCIÓN conejos o cualquier
```

```
REF: ****** * ANIMALITO PODÉS venir a venderlo
HYP: ANIMAL Y NO        PUEDE venir a venderlo
```

As for lexical variation, it is well known that dialects differ not only in the presence of lexical items exclusive of each variety (regionalisms) but also in the frequency of use of shared words. While the former may probably be out of the ASR system lexical coverage (out-of-vocabulary words) and thus not be recognized, the latter might still produce errors because of their low probability: these words having low frequency in the language model construction corpus, they have low assigned weight so they are less frequently selected for the output. That is the case in the following example, where the word *guaso* is recognized one in two times highlighting that the word is known by the system but is not always considered as the best candidate for the transcription:

```
REF: EL guaso ** el hombre el GUASO chileno
HYP: ** guaso EL el hombre el HUASO chileno
```

Out-of-vocabulary words are critical: not included in the system's vocabulary, they are never recognized. It should be noticed that " on average, each out-of-vocabulary causes more than a single error" [Lamel and Gauvain, 2003]. The Latin American monologue corpus, for example, contains 184 OOVs, which represents 6% of the corpus vocabulary. Among OOVs there are proper names, typographical errors and 3 spelling mistakes. But there are also 4 *voseo* conjugations, an interesting number given the size of the corpus and the monological nature of most of the samples.

From the variety of phenomena affecting the system performance, this research focuses on those related to the acoustic component, i.e. the phonetic variations. From this group, the ones having the widest quantitative impact have been selected: /s/ reduction (/s/ is the most frequent consonant in Spanish [1]), voiced stop reduction (concerning 22% of the system's vocabulary), *seseo* (16% of the system's vocabulary), and *yeísmo* (affecting 3%).

The high sophistication of ASR systems architecture implies a wide range of components to act on when working on adaptation. In the same sense, the mixed nature of these systems, having data-driven and knowledge-based components calls for statistical and linguistic approaches.

[1] 9,24% frequency according to [Real Academia Española, 2009]

CHAPTER 3

Lenition of intervocalic stops and of coda /s/ in Spanish: an automatic study of reduction phenomena

In this chapter consonant reduction studies using automatic speech processing tools are presented. More precisely, two cases of lenition are studied:

1. voiced stops /b/, /d/ and /g/ pronounced as fricatives or approximants [β], [ð] and [ɣ] or ∅

2. /s/ pronounced as [z], [h], [^{s}], [ʔ] or ∅

These instances of variation were selected because of their relevance in the literature of Spanish phonetics and because of the considerable rate of vocabulary items liable to these types of lenition. Linguistic studies usually deal with reduced and/or controlled corpora and focus on acoustic and articulatory features [Hualde and Prieto, 2014], [Chitoran et al., 2015], [Hualde et al., 2010].

Figure 3.1 highlights the phenomena analyzed here. There are two realizations of the word *abogado*: one from Spain, where the /b/ /d/ and /g/ are maintained [1] (left) and one from Latin America, where no obstruction phases are visible on the spectrogram (right). As for coda /s/, Figure 3.2 shows high energy concentration corresponding to the final /s/ at the end of the word *precios* for the Peninsular realization (left). In the Latin American example (right), the lack of acoustic information support the deletion of /s/.

These two representations underline the difficulty to associate relevant acoustic information to support the presence or lenition of the sounds investigated here. As for the human perception, the capacity of human perceptual system to rebuild from the surrounding context the missing phonetic cues, might affect an objective decision on the presence or absence of the analyzed consonants [Vasilescu et al., 2014].

In the present study a different approach is adopted which does not rely neither on the acoustic analysis per se nor on the human perception. Here, voiced stops and coda /s/ reduction are estimated in an objective manner through forced alignment using pronunciation variants as in [Adda-Decker and Lamel, 1999], [Renwick et al., 2016]. The main purpose of these experiences is to quantify the lenition phenomena in continuous speech in Spanish in order to verify the hypothesis that these variation patterns are more widespread in Latin America then in Spain, and more specifically in the Caribbean, Rioplatense and Chilean varieties [Hualde and Prieto, 2014]

[1] White zones indicate obstruction of the air flow.

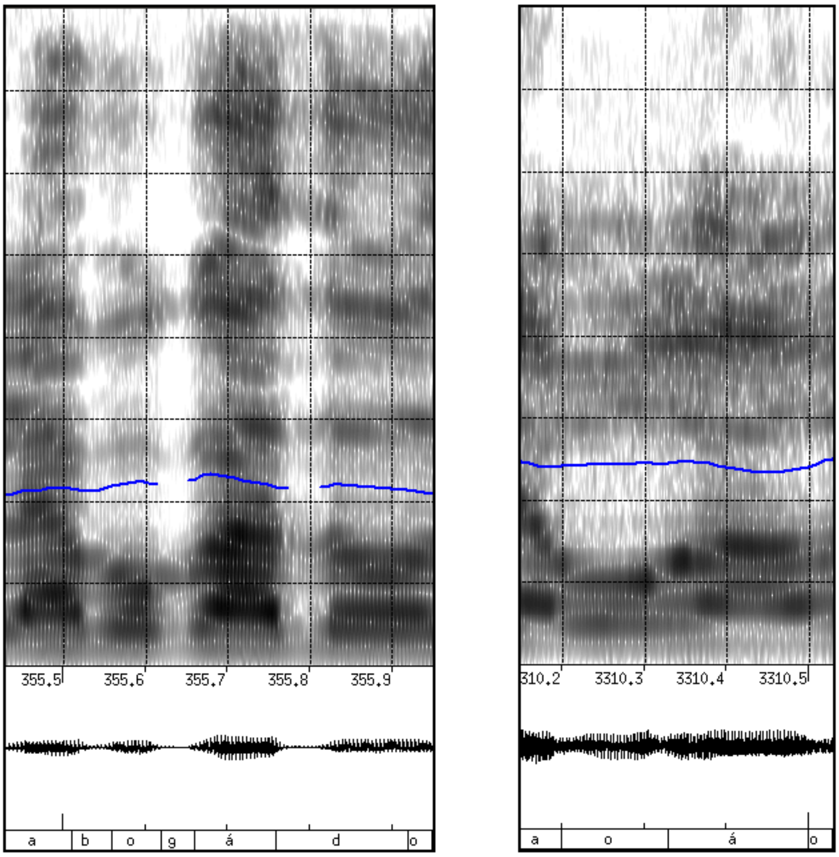

Figure 3.1 – Spectrogram of voiced stops on *abogado*. White vertical zones indicate the obstruction of the air flow by /b/ /g/ and /d/ in [abogádo] (left). No such evidence is found in the weakened pronunciation identified as [aoáo] (right).

[Chitoran et al., 2015]. Additionally, the influence of speaking style is considered. Instances of undetected /b/ /d/ /g/ or coda /s/ by the ASR system are the consequence of a temporal reduction which applies as soon as the expected minimum time span, as is relevant for the ASR system, is not achieved. As a consequence, the alignment leaves out a segment. This can be the result of deletion or of phonetic undershoot. The use of pronunciation variants has the advantage of combining acoustic decoding and the analysis of contextual effects.

3.1 Method

In order to estimate the lenition rate in our Spanish corpora, forced alignment experiences are performed using Vocapia's existing acoustic models for Spanish. The forced alignment system receives as input the speech data, the corresponding reference transcriptions and a lexicon with pronunciation variants, in this case pronunciations with presence or absence of the consonants subject to lenition. Based on the acoustic information, the aligner selects the phonetic representation that resembles most the pronounced word.

The rate of consonant lenition is conditioned by several intra and extra-linguistic factors such as stress, age, gender, education, and socioeconomic status [Ryant and Liberman, 2016]. In the present research, the system preference is cal-

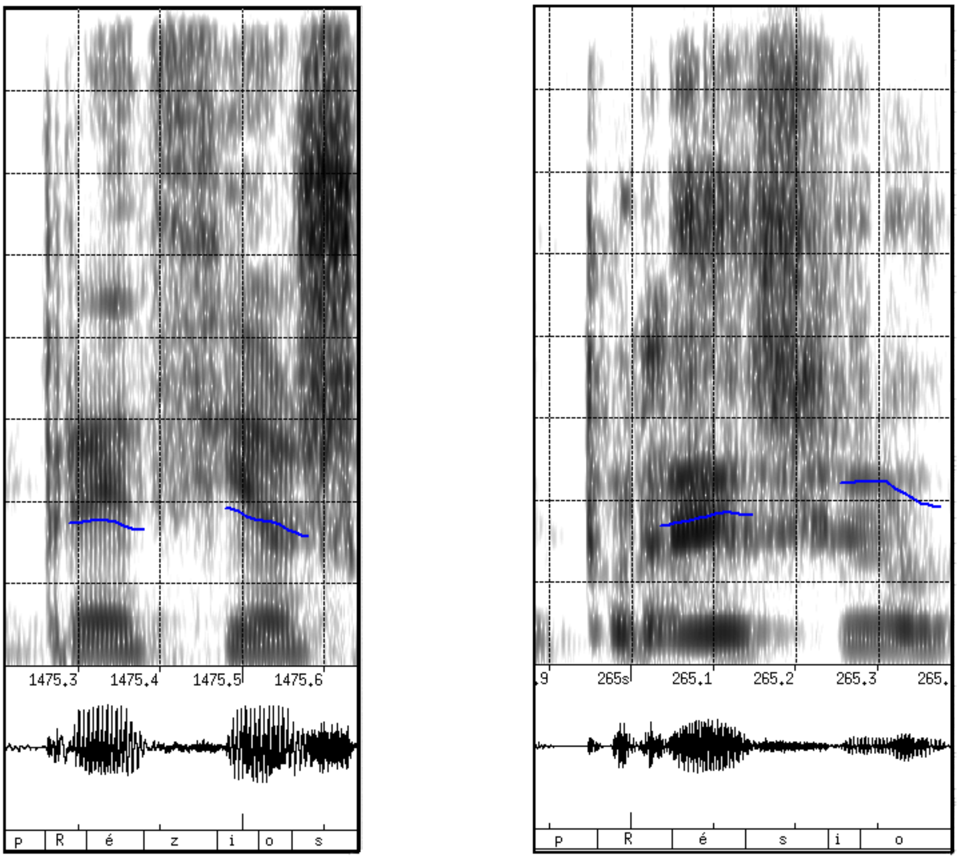

Figure 3.2 – Spectrogram of coda /s/ on *precios*. The black zone on the upper right zone of the spectrogram indicates full pronunciation of /s/ [pRézios] (left). There is no such evidence at the end of the spectrogram identified as [pRésio] (right).

culated for the following three factors:

1. **phonetic context**: this consists of the previous and/or following phones with respect to the target phone. In the case of voiced stops, the phonetic context analyzed in our experiments is surrounded by vowels as in *lavo* [láβo], *lado* [láðo], *lago* [láɣo]. It occurs exclusively within a word, i.e. cases as *una dama* [unaðama], although susceptible of lenition [Chitoran et al., 2015] were not considered. At this stage of this preliminary investigations. The reason is that intervocalic and intra-word lenition is widely described as concerning **all** varities, Peninsular or Latin American. In the case of /s/, the study focuses on the coda position, that is end of word or before consonant, always within word boundaries, as in *patos* [pátoh], *pasto* [páhto] but not *una señora* [unaseŋóɾa].

2. **dialectal variety**: the dialectal varieties considered in this research are Latin American Spanish and Peninsular Spanish. The varieties within the Latin American group (see section 1.2) are also examined but most of the analysis focuses on the Latin American/Peninsular opposition due to the lack of sufficient data for each Latin American variety.

3. **speaking style**: this factor is considered according to the following oppositions prepared/spontaneous, professional/non-professional. These parameters

are used for characterizing data from radio broadcast, monologues and telephonic conversations.

The lenition study was carried out using exclusively the development corpora.

3.2 Results

It should be pointed out that there can be more than one segment of interest per word: 10% of the V[bdg]V words have more than one vowel-fricative-vowel and 6.5% of the /s/ coda words have more than one /s/ in coda position. As a consequence the different reference unit considered here is the sequence and not the word, which allows counting each lexical item as many times as needed (i.e. *adecuada* counts twice for VdV lenition, *puestos* counts twice for /s/ lenition).

3.2.1 Voiced stop reduction

Considering the phonetic factor, /b/, /d/ and /g/ present similar overall reduction rates. /d/ has the higher percentage of lenition (35.7%), which is favored by its relative frequency, the highest among three voiced stops. The /g/ is in second place with 34.5% reduction rate, in spite of having the lowest frequency.

Weakened realizations are considerable in all the studied corpora, the lowest percentage being 19.8% for /b/ in prepared Peninsular Spanish. Cases of double lenition such as *universidad* [unieRsiád], *jugadores* [xuaóRes], *preguntaba* [pReuntáa] can be found across all the corpora. Double reduction cases represent 5.4%.

Regarding dialectal variety, Latin America shows higher lenition rates than Spain, which confirms previous linguistic findings. In broadcast news speech, an average of 38.5% voiced stops are reduced in Latin America vs. 23.9% in Spain. In particular, the lenition rate of /d/ in Latin American broadcast news corpus doubles the rate observed for Peninsular broadcast news corpora. While the Latin American dialect presents the highest lenition percentages for /d/ in all corpora, the Peninsular dialect is equally affected by /g/ and /d/.

As for Latin American varieties, our results support the tendencies described in linguistic literature: Andean and Mexican dialects (the “High Lands”) present lower values while the Caribbean, Chilean and Rioplatense dialects show higher values. However, the rates for the Caribbean are less important than expected: they are not far from the values for Mexican whereas they were expected to have the highest rate. This may be explained by the small number of speakers of each variety in this corpus.

Regarding the speaking style, linguistic hypotheses are again validated by our experiment based on forced alignement. The system selects the pronunciation option corresponding to the weakening more often in the conversational corpora than in prepared speech corpora: higher percentages are observed in conversational corpora than in broadcast news and monologues. More than half of the /b/ /d/ /g/ occurrencies are weakened for the conversational data, illustrating that reduction phenomena increase with less prepared speaking styles. Professional speech data have lower values than non-professional speakers data.

According to speaking style, voiced stop weakening decreases as follows:

+ conversational non-professional from both dialects
prepared professional from Latin America
conversational professional from Spain
monologue non-professional from Latin America
- prepared professional from Spain

3.2.2 /s/ reduction

Compared to the voiced stops lenition rates, /s/ coda has the lowest overall reduction rate of 27.8%. However, in absolute terms, /s/ coda has the largest number of weakened realizations since it is much more frequent than intervocalic stops.

The coda position includes two contexts, word final /s/ and preconsonantic /s/. Adding the preconsonantic pronunciation variant does not increase the overall lenition rate for /s/ but it allows to register more lenition cases in spontaneous speech. Double lenition of /s/ coda in the same word is not uncommon: *turistas* [tuRíta], *ustedes* [utéde], *estamos* [etámo]. It represents 2.8% of the /s/ lenition cases.

With regard to dialectal variety, as in the case of the voiced stops experiments, previous linguistic hypothesis stating the higher frequency of the phenomena in Latin American varieties are confirmed: Latin American monologues and broadcast news corpora have higher lenition values than Peninsular broadcast news and broadcast conversation. In the case of broadcast news in particular, Latin America presents 27.2% of lenition while less than a half is registered for Spain (12.2%). However, in conversational telephonic speech, the difference between geographical varieties tends to disappear as both varieties are about 60%.

For the Latin American varieties, Andean and Mexican have a preference for /s/ maintenance pronunciation while Caribbean, Chilean and Rioplatense show a tendency towards /s/ weakening. The Chilean and Andean dialects are respectively the most dropping and maintaining dialects.

Finally, regarding speaking style, /s/ reduction in conversational non-professional style speech widely exceeds that of prepared speech, with rates almost doubling.

3.3 Discussion

The study of voiced stops and coda /s/ reduction in Latin American and Peninsular Spanish shows that large-scale analysis based on pronunciation variants are reliable approaches to validate linguistic traditional statements. The main observations concerning the phenomena investigated here are:

- intervocalic /b/ /d/ /g/ and coda /s/ lenition are observed in all Spanish varieties and all speaking styles
- /s/ coda and intervocalic /d/ are the most affected by lenition
- voiced stops and coda /s/ reduction are more frequent in Latin American Spanish than in Peninsular Spanish
- lenition is more present in Latin American speech than Peninsular data across all speaking styles
- in spontaneous speech, Latin America and Spain have similar lenition rates specially for /s/ coda (62%). This suggests that Peninsular Spanish admits more weakening in the informal register

- the opposition of “Low Lands” to “High Lands” is valid for intervocalic voiced stop and and /s/ coda reduction
- in general, spontaneous speech has more consonant reduction than professional speech
- prepared but non professional data (monologues corpus) display higher lenition rates between spontaneous and professional data

From a linguistic point of view, these findings support classical phonetic assumptions with quantitative evidence from continuous speech. Further studies may concern: examining the phonetic context across words boundaries, measuring surrounding vowels duration, studying acoustic parameters as voicing.

As for speech processing, the influence of these findings on automatic transcribing is tested through pronunciation modeling for the Spanish system. These experiments are described in the next chapter.

CHAPTER 4

PRONUNCIATION MODELING FOR ACOUSTIC MODEL ADAPTATION

The pronunciation of a certain word varies depending on diverse factors such as phonetic context, pitch, rhythm, speaker gender, dialect and register, among others, so phonetic knowledge can be applied to encompass the most relevant and frequent of these variations. Taking into account phonetic peculiarities may lead to a modification of the set of phones used to represent the pronunciations. For instance, [Lamel, 2003] points out that using two different phone sets for processing the same English corpus may yield to more than 1% difference in WER.

The pronunciation dictionary draws a mapping between a pronunciation and an orthographic form of a word, however the processus is not necessarily a one-to-one relation since another strategy for modeling phonetic variations is to provide pronunciation options. "If a word is poorly articulated, as is usually the case in relaxed speaking styles, this word may not be recognized and/or near function words may be deleted" [Lamel and Gauvain, 2003]. Adding these relaxed variants as pronunciation options can help to improve alignment and results in more accurate acoustic phone models that may improve the system's performance.

Given the statistical approach of ASR systems, modifications in the pronunciation model entail a retraining of the acoustic model. Each model is a left-to-right 3-state representation of a phone in a context usually corresponding to a triphone. Hence if the phone set is modified, the triphones must be regenerated. Likewise if a phonetic rule producing new phonetic variants is added, the probabilities of each variant must be recalculated.

4.1 Method

As seen in section 2.2, several approaches can be adopted for adapting ASR systems conceived for a certain dialect to another dialect of the same language. Dialect-specific or global pronunciation models and datasets can be combined in different configurations and the resulting systems can be applied in dialect-specific, cross-dialectal or multi-dialectal decodings. In the present research, two strategies are adopted in order to integrate Latin American pronunciation variants to the Spanish acoustic models: first, reduction of the Spanish phone set via phone merging; second, adding alternative pronunciations to the lexicon based on the results of the pronunciation studies described in chapter 3.

Each change in the pronunciation lexicon requires a new acoustic training. Besides pronunciation modeling, the impact of dialect-specific train data in the acoustic model is also examined. Two train datasets are employed, one composed of Peninsular speech only and another including Latin American data in addition to the same Peninsular data. The training experiences are performed on broadcast news data exclusively.

An acoustic model using the current pronunciation lexicon was trained to get a baseline. Then six acoustic models combining different pronunciation lexicons and train datasets were trained.

4.1.1 Pronunciation modeling

The pronunciation lexicon can be dialectally customized by extending the phone list (adding phones), reducing the phone list (merging phones), modifying phonetization of specific sequences or adding pronunciation variants. paragraphs. The phonetization of specific sequences was applied for correcting the baseline pronunciation model. Several modifications were added to the **corrections model**. For instance, the *consonant+r* sequence can be pronounced with a trill or a flap vibrant depending on the consonant preceding the vibrant. The sequence *[bcdfgkpqt]+r* forms a cluster hence the vibrant is pronounced /ɾ/ while the sequence *[lmns]+r* is a disjoint group so the vibrant is trilled. This rule and other details were corrected on the original lexicon. It should be noticed that these corrections are general to all Spanish varieties so they should improve Peninsular data processing as well.

The original phone set has been conceived to represent Peninsular Spanish, thus it includes the dialectal-specific interdental phone /θ/. Two modifications of this phone set were implemented to train corresponding acoustic models.

For their multi-dialectal system, [Nogueiras et al., 2002] define a global phone set of 32 phones (+ silence and noise) where specific sounds of all Spanish varieties, such as /ç/, /ʒ/, /θ/ and /ŋ/ among others, are included. This strategy requires available audio and text data for all the dialects. The present study also aims to conceive a global phone set but proceeding with an opposite approach: instead of modeling dialectal allophones separately, they are merged in the same representation. [Renwick et al., 2016] study provides an example of how an ASR system trained with a simplified phone set can perform as well as a system where all phonetic contrasts are preserved. Therefore, reduction of the phone set is privileged expecting for the acoustic learning to be able to generalize over a wider acoustical space.

Two reductions of the phone list were consecutively applied, the first involving vowels and the second involving consonants. [Moreno and Mariño, 1998] observe that "vowels in Spanish represent approximately a 50% of the total allophone counts and there aren't significant differences among dialects".

The **consonant merge model** envisages to represent the *yeísmo* and the *seseo*. It aims to improve the acoustic model accuracy for Latin American speech primarily and to preserve the performance for Peninsular speech.

The final modification of the pronunciation lexicon involves pronunciation variants allowing alternative pronunciations representing coda /s/ and intervocalic /d/. As the analyses indicate, these phenomena are more advanced in Latin America but they are also considerable in Spain. Therefore, the hypothesis is that the **lenition model** will improve Latin American data mainly and Spanish data may also benefit from this modification.

Acoustic models are trained for each of these pronunciation lexicons and the same mappings are applied to the lexicon used for decoding.

4.1.2 Extended dataset

The pronunciation models specifically tailored for Latin American pronunciation are employed for two different training experiences: one with an **initial dataset** of Peninsular Spanish speech and corresponding transcriptions exclusively and an **extended dataset** with supplementary Latin American speech and corresponding transcriptions.

These Peninsular and Latin American data diverge not only in quantity but also in quality: the transcriptions of the Peninsular corpus are the product of careful manual work while the transcriptions of the Latin American corpus were made automatically with only a fast human revision. Such difference may be responsible of a bias in the acoustic training. An initial training with high quality train data may allow to estimate the impact of lowest quality data.

As for blending different dialects in the train set, the positive results obtained for inter-dialect data sharing in [Najafian and Hansen, 2016] for English and [Nogueiras et al., 2002] and [Caballero et al., 2009] for Spanish, provide evidence of the interest to employ a maximum of train data even if such data illustrate another dialect. Therefore, Latin American results are expected to improve with the wider dataset, especially since the additional data belongs to the targeted dialect. It should be noticed that [Caballero et al., 2009] specify that data sharing requires a global phone set.

4.2 Results

The six acoustic models defined in the previous section are applied to transcribe a Caribbean-Latin American broadcast news corpus, a Spanish broadcast news corpora, and a Latin American monologues corpus. Broadcast news corpora from two dialects allow to test the impact of dialectal adaptation. The monologues corpus is employed for cross-dialectal and cross-style evaluation.

4.2.1 Lexicon modifications

The baseline rates show the best results for the Peninsular data and a degradation of the rates when transcribing Latin American broadcast news speech. The speaking-style and the reduced number of speakers per variety may explain the higher WER for the Latin American monologues corpus.

The **corrections model** performance varies across the corpora: Latin American monologues show improvements but small degradations are registered for the Latin American broadcast news. Given these results, it is difficult to decide whether this model improves the performance or not. Since the proposed corrections applied to all Spanish varieties and the scores were improved for some corpora, these modifications were conserved for the following experiments.

The results for the **vowel merge model** suggest that the phonetic contrast between stressed and unstressed vowels is relevant for automatically transcribing Spanish varieties. No significant improvement is observed for this model, as a consequence it was not applied on the subsequent training.

The first Latin American targeted model, the **consonant merge model**, yields a degradation on Peninsular data, which support the *seseo* and the *yeísmo* as pertinent dialectal contrasts. The scores for the Latin American data were less good than expected maybe due to the absence in the train set of audio data representing the changes proposed in the pronunciation lexicon. Adding audio data containing such features may improve the scores of this model.

The performance of the second Latin American-specific model, the **lenition model**, also varies depending on the corpus. This is potentially due to the fact that the lenition does not affect equally the varieties in the corpus. A separate scoring by variety could be done in order to verify if the lenition model underperforms for "Low Lands" and improves for "High Lands".

4.2.2 Dialect-specific extended dataset

Latin American-specific pronunciation models, i.e. consonant merge and lenition, are also employed for a supplementary experiment with an augmented train dataset.

For both consonant merge and lenition models, the extended dataset reduces the error rate with respect to the initial dataset for all corpora. Therefore extending the dataset improves the results even if the new data belongs to a different dialect.

The use of dialect-specific data in the acoustic model training substantially improves the results on matching dialectal data. As for Peninsular corpora, increasing the train dataset even with data from another dialect does not penalize the results and improves the scores as well. This represents evidence in favor of the inter-dialectal data sharing hypothesis and supports the choice for a single multi-dialectal system.

4.3 Discussion

This chapter explored phonetic knowledge-based and data-driven approaches for improving Latin American pronunciation processing.

The best configuration for transcribing Latin American data combines expert corrections, consonant merge and lenition with the extended dataset of Caribbean Spanish. The data are similar to the ones from the test corpus and the pronunciation modifications illustrate linguistic features of the Caribbean variety. This configuration obtains 7% relative gain.

The results for the automatic transcription of the Peninsular corpus suggest that even if the targeted variety is Latin America, the system remains robust: the scores correspond to the baseline level for most of the experiences and never degrade more than 0.2%.

As for cross-dialectal evaluation on other Latin American dialects, the best results are obtained with the extended dataset consonant merge model, since lenition degrades the performance for some varieties. The difference in speech style may also explain the uneven results between corpora.

Overall, results are better for consonant merge over lenition. Lenition with extended data produces a gain on prepared Latin American data without considerable degradation for the rest. Therefore, this configuration could be selected as a robust muti-dialectal recognizer for all Spanish varieties.

CHAPTER 5

SUMMARY AND DISCUSSION

This chapter summarizes the studies accomplished as part of the internship and discusses the applied research methodologies addressing some of the challenges of dialectal variation in Spanish continuous speech processing.

A preliminary study of Spanish phonetics and dialectal classification followed by an analysis of the errors produced on transcriptions of the baseline system were carried out in order to detect the pronunciation variants more liable to affect the system performance. The *yeísmo*, the *seseo* and intervocalic /b/ /d/ /g/ and coda /s/ lenition were detected as the most important variations, concerning 3%, 16%, 22% and 37% of the system vocabulary respectively.

Speech technology experiments (i.e. forced alignment and automatic speech transcription) were conducted on a corpus of continuous speech. A comparative approach between Peninsular and Latin American varieties was adopted for the experimental design and the analysis. Differences among Latin American dialects were also considered.

Progressive modifications of the system lexicon and dataset were proposed based on corpus linguistics analyses. These knowledge-based adaptations combined with extended train data belonging to the targeted dialect have resulted in a relative gain of 7% for Latin American data. Moreover, the proposed models are consistent with the linguistic description of dialectal varieties.

The consonant reduction study provides quantitative evidence supporting sociolinguistic and dialectal descriptions of intervocalic voiced stops and coda /s/ distribution. Both phenomena are attested in all the corpora processed in this work with rates of 20% or more. The presence of both lenition processes is stronger in conversational speech (60% in average) and in Latin American speech (rates 15% higher than Spain), primarily in Caribbean, Chilean and Rioplatense varieties. For /s/ coda in particular, in many Latin American countries the norm accepts preconsonantic /s/ lenition in standard speech (that found in the media and cultivated speakers speech) while in Spain lenition is identified with non-standard speech and dialectally marked speech ([Chitoran et al., 2015]). This may explain why the lenition rate is lower in the Spanish broadcast news corpora (12% vs. 27% in Latin America) and the reason why the scores inter-dialect are closer for the telephone conversation corpora.

These findings have been applied to the broadcast news transcription system for Spanish together with modifications of both the phonemic inventory and lexicon. The lenition was integrated as pronunciation variants. Variants contribute to system versatility: they allow dialectal or speaking-style variations to be recognized while preserving the possibility of recognizing standard speech. Due to scarcity of training data from each of the Latin American varieties, a multi-dialectal phone set representing minimal phonetic contrasts was aimed.

Two acoustic model training configurations were developed: a uni-dialectal configuration with Peninsular data exclusively and a bidialectal configuration with Peninsular and Caribbean data. Cross-dialectal evaluations on Peninsular and Caribbean development corpora were carried out. The robustness of the models was tested on cross-dialectal and cross-register data from the Latin American monologue.

The results were improved for both Latin American and Peninsular data by the bidialectal configuration and the unidialectal respectively. Nevertheless the performance of the bidialectal configuration on the Peninsular data remains close to the best result. The best performing model for Latin America includes expert corrections, consonant merge and lenition with extended training dataset from Caribbean Spanish.

A reliable control over the dialectal origin of the data is not always possible and in the case of Spanish the presence of different dialects on the same broadcast is favored by the high intercomprehension between varieties. The control over the data is even more difficult for a commercially exploitable system therefore, keeping a single system allows robustness facing eventual data of unknown dialect.

5.1 Future studies

Several extensions of this research are possible, both for linguistic experiences and automatic speech processing.

On the linguistic side, the Linguistic Data Consortium data could be employed to validate the lenition study on a larger conversational corpus. The experience should also be extended to a corpus having more speakers for each dialectal variety. The influence of lenition on vowels duration and the evaluation of the impact of vowel stress on the lenition phenomena could also be studied. It would also be interesting to confirm our findings by human perception studies.

Regarding the recognition system, the influence of the dialect data can be evaluated by training specific acoustic models using only Latin America data. If more data for each variety is available, specific acoustic and language models can be build for each one. Moreover, dialect-specific language models should be adapted to Latin American Spanish.

This study made use of data from Caribbean Latin American and Peninsular Spanish varieties. In order to increase dialect robustness, data from different dialects should be added to the train corpus. Adding new data from Latin American radio broadcasts should be employed for ameliorating acoustic model training through unsupervised trainingmethods.

Applying the proposed changes of the pronunciation lexicon on the Spanish conversational telephone speech system seems promising. According to our pronunciation study, the lenition rate is more important in telephonic speech than in broadcast news speech, therefore the lenition model is expected to yield an even more important gain than the one obtained on broadcast data. Furthermore, the gain could be equally important on Latin American and Peninsular telephonic conversations.

Bibliography

[Adda-Decker and Lamel, 1999] Adda-Decker, M. and Lamel, L. (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Commun.*, 29(2):83–98. – Cited page 23.

[Ávila, 2001] Ávila, R. (2001). Los medios de comunicación masiva y el español internacional. pages 16–19. – Cited page 9.

[Caballero et al., 2009] Caballero, M., Moreno, A., and Nogueiras, A. (2009). Multidialectal spanish acoustic modeling for speech recognition. *Speech Communication*, 51(3):217–229. – Cited pages 17, 18, 20 et 31.

[Chitoran et al., 2015] Chitoran, I., Hualde, J. I., and Niculescu, O. (2015). Gestural undershoot and gestural intrusion: From perception errors to historical sound change. In *Proceedings of 2nd ERRARE Workshop-"Errors by Humans and Machines in Multimedia, Multimodal, Multilingual Data Processing", Sinaia, Romania*. – Cited pages 13, 23, 25 et 33.

[de la Concha et al., 2017] de la Concha et al., V. G. (2017). *EL ESPAÑOL: UNA LENGUA VIVA*. – Cited page 9.

[de Mareüil et al., 2013] de Mareüil, P. B., Woehrling, C., and Adda-Decker, M. (2013). Contribution of automatic speech processing to the study of northern/-southern french. *Language Sciences*, 39:75–82. – Cited pages 15 et 17.

[Elfeky et al., 2015] Elfeky, M., Moreno, P., and Soto, V. (2015). Multi-dialectical languages effect on speech recognition: Too much choice can hurt. – Cited page 17.

[Fontanella de Weinberg, 1992] Fontanella de Weinberg, M. B. (1992). El español de américa a partir de 1650. – Cited page 12.

[Gruhn et al., 2011] Gruhn, R. E., Minker, W., and Nakamura, S. (2011). *Statistical pronunciation modeling for non-native speech processing*. Springer Science & Business Media. – Cited page 17.

[Hualde and Prieto, 2014] Hualde, J. and Prieto, P. (2014). Lenition of intervocalic alveolar fricatives in catalan and spanish. 71:109–127. – Cited pages 12 et 23.

[Hualde et al., 2010] Hualde, J., Simonet, M., Shoted, R., and Nadeu, M. (2010). Quantifying iberian spirantization. Keynote of presentation on Linguistic Symposium on Romance Languages-40. – Cited pages 13 et 23.

[Huang et al., 2014] Huang, X., Baker, J., and Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, 57(1):94–103. – Cited page 15.

[Lamel, 2003] Lamel, L. (2003). Pronunciation modeling. Slides of the Spoken Language Processing Group LIMSI-CNRS. – Cited page 29.

[Lamel and Gauvain, 2003] Lamel, L. and Gauvain, J.-L. (2003). *Speech recognition*, chapter 16, pages 305–322. Oxford University Press. – Cited pages 16, 22 et 29.

[López González, 2002] López González, A. M. (2002). La lengua internacional de los medios de comunicación: una convergencia de modelos lingüísticos. pages 522–532. – Cited page 9.

[Mihaylova, 2011] Mihaylova, Z. (2011). *Lexical and Acoustic Adaptation for Multiple Non-Native English Accents*. PhD thesis, Diplomarbeit in Karlsruhe Institute of Technology (KIT). – Cited page 17.

[Moreno and Mariño, 1998] Moreno, A. and Mariño, J. B. (1998). Spanish dialects: phonetic transcription. – Cited pages 13, 17, 18 et 30.

[Najafian and Hansen, 2016] Najafian, M. and Hansen, J. H. (2016). Speaker independent diarization for child language environment analysis using deep neural networks. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 114–120. IEEE. – Cited pages 15, 17 et 31.

[Nallasamy et al., 2011] Nallasamy, U., Garbus, M., Metze, F., Jin, Q., Schaaf, T., and Schultz, T. (2011). Analysis of dialectal influence in pan-arabic asr. – Cited page 15.

[Nogueiras et al., 2002] Nogueiras, A., Moreno, A., and Caballero, M. (2002). Multidialectal spanish speech recognition. 1:841–844. – Cited pages 17, 18, 30 et 31.

[Quesada Pacheco, 2014] Quesada Pacheco, M. Á. (2014). División dialectal del español de américa según sus hablantes. análisis dialectológico perceptual. volume XLIX, pages 257–309. – Cited pages 11 et 12.

[Real Academia Española, 2009] Real Academia Española (2009). *Nueva gramática de la lengua española*, volume Fonética y Fonología. Espasa Libros. – Cited pages 9, 10, 12 et 22.

[Renwick et al., 2016] Renwick, M., Vasilescu, I., Dutrey, C., Lamel, L., and Vieru, B. (2016). Marginal contrast among romanian vowels: evidence from asr and functional load. 2016:2433–2437. – Cited pages 23 et 30.

[Ryant and Liberman, 2016] Ryant, N. and Liberman, M. (2016). Large-scale analysis of spanish /s/-lenition using audiobooks. – Cited pages 12, 13 et 25.

[Vasilescu et al., 2014] Vasilescu, I., Vieru, B., and Lamel, L. (2014). Exploring pronunciation variants for romanian speech-to-text transcription. In *Spoken Language Technologies for Under-Resourced Languages*. – Cited page 23.

[Vergyri et al., 2010] Vergyri, D., Lamel, L., and Gauvain, J.-L. (2010). Automatic speech recognition of multiple accented english data. In *Eleventh Annual Conference of the International Speech Communication Association*. – Cited pages 15 et 17.

[Vieru et al., 2011] Vieru, B., De Mareueil, P. B., and Adda-Decker, M. (2011). Characterisation and identification of non-native french accents. *Speech Communication*, 53(3):292–310. – Cited pages 15 et 17.