Available online at www.sciencedirect.com

Journal of Applied Research and Technology

Journal of Applied Research and Technology 15 (2017) 259–270

www.jart.ccadet.unam.mx

Original

Automatic speech recognizers for Mexican Spanish and its open resources

Carlos Daniel Hernández-Mena [a,*], Ivan V. Meza-Ruiz [b], José Abel Herrera-Camacho [a]

[a] *Laboratorio de Tecnologías del Lenguaje (LTL), Universidad Nacional Autónoma de México (UNAM), Mexico*

[b] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM), Mexico*

Received 21 March 2016; accepted 9 February 2017

Available online 28 April 2017

Abstract

Development of automatic speech recognition systems relies on the availability of distinct language resources such as speech recordings, pronunciation dictionaries, and language models. These resources are scarce for the Mexican Spanish dialect. In this work, we present a revision of the CIEMPIESS corpus that is a resource for spontaneous speech recognition in Mexican Spanish of Central Mexico. It consists of 17 h of segmented and transcribed recordings, a phonetic dictionary composed by 53,169 unique words, and a language model composed by 1,505,491 words extracted from 2489 university newsletters. We also evaluate the CIEMPIESS corpus using three well known state of the art speech recognition engines, having satisfactory results. These resources are open for research and development in the field. Additionally, we present the methodology and the tools used to facilitate the creation of these resources which can be easily adapted to other variants of Spanish, or even other languages.

Keywords: Automatic speech recognition; Mexican Spanish; Language resources; Language model; Acoustic model

1. Introduction

Current advances in automatic speech recognition (ASR) have been possible given the available speech resources such as speech recordings, orthographic transcriptions, phonetic alphabets, pronunciation dictionaries, large collections of text and computational software for the construction of ASR systems. However, the availability of these resources varies from language to language. Until recently, the creation of such resources has been largely focused on English. This has had a positive effect on the development of research of the field and speech technology for this language. This effect has been so positive that the information and processes have been transferred to other languages so that nowadays the most successful recognizers for Spanish language are not created in Spanish-speaking countries. Furthermore, recent development in the ASR field relies on restricted corpora with restricted access or not access at all. In order to make progress in the study of spoken Spanish and take full advantage of the ASR technology, we consider that a greater amount of resources for Spanish needs to be freely available to the research community and industry.

With this in mind, we present a methodology and resources associated to it for the construction of ASR systems for Mexican Spanish; we argue that with minimal adaptations to this methodology, it is possible to create resources for other variants of Spanish or even other languages.

The methodology that we propose focuses on facilitating the collection of the examples necessaries for the creation of an ASR system and the automatic construction of pronunciation dictionaries. This methodology has been concluded on two collections that we present in this work. The first is the largest collection of recordings and transcriptions for Mexican Spanish freely available for research, and the second is a large collection of text extracted from a university magazine. The first collection was collected and transcribed in a period of two years and it is utilized to create acoustic models. The second collection is used to create a language model.

We also present our system for the automatic generation of phonetic transcriptions of words. This system allows the creation of pronunciation dictionaries. In particular, these transcriptions are based on the MEXBET (Cuetara-Priede, 2004)

* Corresponding author.

Peer Review under the responsibility of Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.jart.2017.02.001

Table 1
ASR Corpora for the top five spoken languages in the world.

Rank	Language	LDC	ELRA	Examples
1	Mandarin	24	6	TDT3 (Graff, 2001) TC-STAR 2005 (TC-STAR, 2006)
2	English	116	23	TIMIT (Garofolo, 1993) CLEF QAST (CLEF, 2012)
3	Spanish	20	20	CALLHOME (Canavan & Zipperlen, 1996) TC-STAR Spanish (TC-STAR, 2007)
4	Hindi	9	4	OGI Multilanguage (Cole & Muthusamy, 1994) LILA (LILA, 2012)
5	Arabic	10	32	West Point Arabic (LaRocca & Chouairi, 2002) NetDC (NetDC, 2007)

phonetic alphabet, a well establish alphabet for Mexican Spanish. Together, these resources are combined to create ASR systems based on three freely available software frameworks: Sphinx, HTK and Kaldi. The final recognizers are evaluated, compared, and made available to be used for research purposes or to be integrated in Spanish speech enabled systems.

Finally, we present the creation of the CIEMPIESS corpus. The CIEMPIESS (Hernández-Mena, 2015; Hernández-Mena & Herrera-Camacho, 2014) corpus was designed to be used in the field of the automatic speech recognition and we utilize our experience in the creation of it as a concrete example of the whole methodology that we present at this paper. That is why the CIEMPIESS will be embedded in all our explanations and examples.

The paper has the following outline: In Section 2 we present a revision of corpora available for automatic speech recognition in Spanish and Mexican Spanish, in Section 3 we present how an acoustic model is created from audio recordings and their orthographic transcriptions. In Section 4 we explain how to generate a pronunciation dictionary using our automatic tools. In Section 5 we show how to create a language model, Section 6 shows how we evaluated the database in a real ASR system and how we validate the automatic tools we are presenting at this paper. At the end, in Section 7, we discuss our final conclusions.

2. Spanish language resources

According to the "Anuario 2013"[1] created by the Instituto Cervantes[2] and the "Atlas de la lengua española en el mundo" (Moreno-Fernández & Otero, 2007) the Spanish language is one of the top five more spoken languages in the world. Actually, the Instituto Cervantes makes the following remarks:

- Spanish is the second more spoken native language just behind Mandarin Chinese.
- Spanish is the second language for international communication.
- Spanish is spoken by more than 500 millions persons, including speakers who use it as a native language or a second language.
- It is projected that in 2030, 7.5% of the people of the world will speak Spanish.
- Mexico is the country with the most Spanish speakers among Spanish speaking countries.

This speaks to the importance of Spanish for speech technologies which can be corroborated by the amount of available resources for ASR in Spanish. This can be noticed in Table 1 that summarizes the amount of ASR resources available in the Linguistic Data Consortium[3] (LDC) and in the European Language Resources Association[4] (ELRA) for the top five more spoken languages.[5] As one can see the resources for English are abundant compared to the rest of the top languages. However, in the particular case of the Spanish language, there is a good amount of resources reported in these databases. Besides additional resources for Spanish can be found in other sources such as: reviews in the field (Llisterri, 2004; Raab, Gruhn, & Noeth, 2007), the "LRE Map",[6] and proceedings of specialized conferences, such as LREC (Calzolari et al., 2014).

2.1. Mexican Spanish resources

In the previous section, one can notice that there are several options available for the Spanish language, but when one focuses in a dialect such as Mexican Spanish, the resources are scarcer. In the literature one can find several articles dedicated to the creation of speech resources for the Mexican Spanish, (Kirschning, 2001; Olguín-Espinoza, Mayorga-Ortiz, Hidalgo-Silva, Vizcarra-Corral, & Mendiola-Cárdenas, 2013; Uraga & Gamboa, 2004). However, researchers usually create small databases to do experiments so one has to contact the authors and depend on their good will to get a copy of the resource (Audhkhasi, Georgiou, & Narayanan, 2011; de_Luna Ortega, Mora-González, Martínez-Romo, Luna-Rosas, & Mu

[1] Available for web download at: http://cvc.cervantes.es/lengua/anuario/anuario_13/ (August 2015).
[2] The *Instituto Cervantes* (http://www.cervantes.es/).
[3] https://www.ldc.upenn.edu/.
[4] http://www.elra.info/en/.
[5] For more details on the resources per language visit: http://www.ciempiess.org/corpus/Corpus_for_ASR.html.
[6] http://www.resourcebook.eu/searchll.php.

Table 2
Corpus for ASR that include the Mexican Spanish language.

Name	Size	Dialect	Data sources	Availability
DIMEx100 (Pineda, Pineda, Cuétara, Castellanos, & López, 2004)	6.1 h	Mexican Spanish of Central Mexico	Read Utterances	Free Open License
1997 Spanish Broadcast News Speech HUB4-NE (Others, 1998)	30 h	Includes Mexican Spanish	Broadcast News	Since 2015 LDC98S74 $400.00 USD
1997 HUB4 Broadcast News Evaluation Non-English Test Material (Fiscus, 2001)	1 h	Includes Mexican Spanish	Broadcast News	LDC2001S91 $150.00 USD
LATINO-40 (Bernstein, 1995)	6.8 h	Several countries of Latin America including Mexico	Microphone Speech	LDC95S28 $1000.00 USD
West Point Heroico Spanish Speech (Morgan, 2006)	16.6 h	Includes Mexican Spanish of Central Mexico	Microphone Speech (read)	LDC2006S37 $500.00 USD
Fisher Spanish Speech (Graff, 2010)	163 h	Caribbean and non-Caribbean Spanish (including Mexico)	Telephone Conversations	LDC2010S01 $2500.00 USD
Hispanic–English Database (Byrne, 2014)	30 h	Speakers from Central and South America	Microphone Speech (conversational and read speech)	LDC2014S05 $1500.00 USD

noz-Maciel, 2014; Moya, Hernández, Pineda, & Meza, 2011; Varela, Cuayáhuitl, & Nolazco-Flores, 2003).

Even though the resources in Mexican Spanish are scarce, we identified 7 corpora which are easily available. These are presented in Table 2. As shown in the table, one has to pay in order to have access to most of the resources. The notable exception is the DIMEx100 corpus (Pineda et al., 2010) which was recently made available.[7] The problem with this resource is that the corpus is composed for reading material and it is only 6 h long which limits the type of acoustic phenomena present. This imposes a limit on the performance of the speech recognizer created utilizing this resource (Moya et al., 2011). In this work we present the creation of the CIEMPIESS corpus and its open resources. The CIEMPIESS corpus consists of 17 h of recordings of Central Mexico Spanish broadcast of interviews which provides spontaneous speech. This makes it a good candidate for the creation of speech recognizers. Besides this, we expose the methodology and tools created for the harvesting of the different aspects of the corpus so these can be replicated in other underrepresented dialects of Spanish.

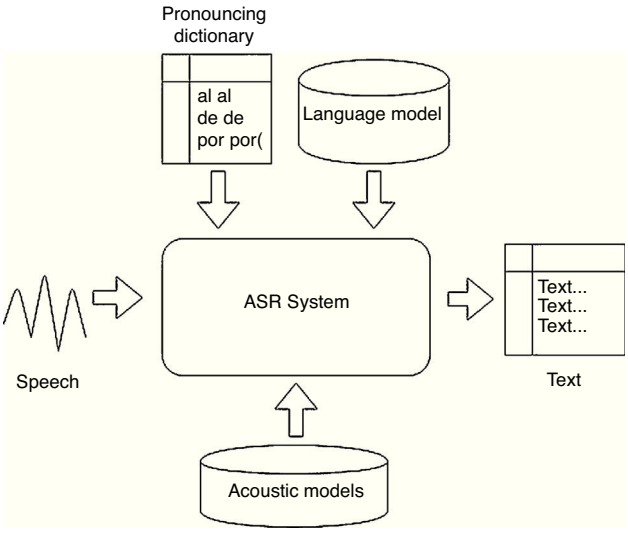

Fig. 1. Components and models for automatic speech recognition.

3. Acoustic modeling

For several decades, ASR technology has relied on the machine learning approach. In this approach, examples of the phenomenon are learned. A model resulting from the learning is created and later used to predict such phenomenon. In an ASR system, there are two sources of examples needed for its construction. The first one is a collection of recordings and their corresponding transcriptions. These are used to model the relation between sound and phonemes. The resulting model is usually referred as the **acoustic model**. The second source of examples of sentences in a language which is usually obtained from a large collection of texts. These are used to learn a model of how phrases are built by a sequence of words. The resulting model is usually referred as the **language model**. Additionally, an ASR system uses a dictionary of pronunciations to link both the acoustic and the language models, since this captures how phonemes compose words. Fig. 1 illustrates these elements and how they relate to each other.

Fig. 2 shows in detail the full process and the elements needed to create acoustic models. First of all, recordings of the corpus pass through a *feature extraction* module which calculates the spectral information of the incoming recordings and transforms them into a format the training module can handle. A list of phonemes must be provided to the system. The task of filling every model with statistic information is performed by the training process.

[7] For web downloading at: http://turing.iimas.unam.mx/ luis/DIME/CORPUS-DIMEX.html.

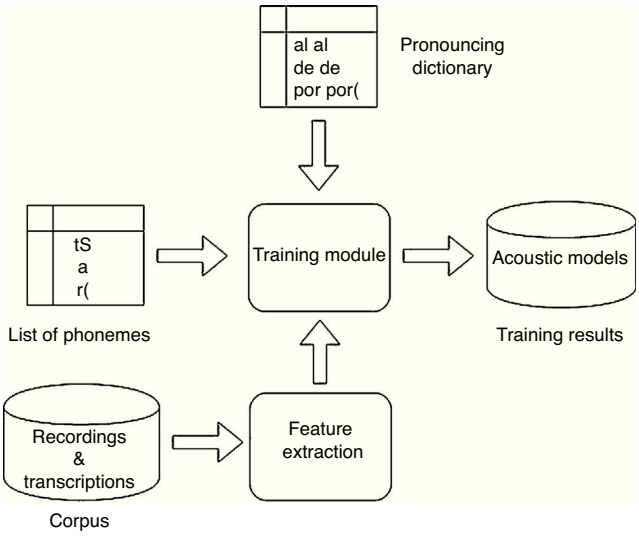

Fig. 2. Architecture of a training module for automatic speech recognition systems.

Table 3
CIEMPIESS original audio file properties.

Description	Properties
Number of source files	43
Format of source files	mp3/44.1 kHz/128 kbps
Duration of all the files together	41 h 21 min
Duration of the longest file	1 h 26 min 41 s
Duration of the shortest file	43 min 12 s
Number of different radio shows	6

3.1. Audio collection

The original source of recordings used to create the CIEMPIESS corpus comes from radio interviews in the format of a podcast.[8] We chose this source because they were easily available, they had several speakers with the accent of Central Mexico, and all of them were freely speaking. The files sum together a total of 43 one-hour episodes.[9] Table 3 summarizes the main characteristics of the original version of these audio files.

3.2. Segmentation of utterances

From the original recordings, the segments of speech need to be identified. To define a "good" segment of speech, the following criteria was taken into account:

- Segments with a unique speaker.
- Segments correspond to an utterance.
- There should not be music in the background.

Table 4
Characteristics of the utterance recordings of the CIEMPIESS corpus.

Characteristic	Training	Test
Number of utterances	16,017	1000
Total of words and labels	215,271	4988
Number of words with no repetition	12,105	1177
Number of recordings	16,017	1000
Total amount of time of the train set (hours)	17.22	0.57
Average duration per recording (seconds)	3.87	2.085
Duration of the longest recording (seconds)	56.68	10.28
Duration of the shortest recording (seconds)	0.23	0.38
Average of words per utterance	13.4	4.988
Maximum number of words in an utterance	182	37
Minimum number of words in an utterance	2	2

- The background noise should be minimum.
- The speaker should not be whispering.
- The speaker should not have an accent other than Central Mexico.

At the end, 16,717 utterances were identified. This is equivalent to 17 h of only "clean" speech audio. 78% of the segments come from male speakers and 22% from female speakers.[10] This gender imbalance is not uncommon in other corpora (for example see Federico, Giordani, & Coletti, 2000; Wang, Chen, Kuo, & Cheng, 2005 since gender balancing is not always possible as in Langmann, Haeb-Umbach, Boves, & den Os, 1996; Larcher, Lee, Ma, & Li, 2012). The segments were divided into two sets: training (16,017) and test (700), the test set was additionally complemented by 300 utterances from different sources such as: interviews, broadcast news and read speech. We added these 300 utterances from a little corpus that belong to our laboratory, to perform private experiments that are important to some of our students. Table 4 summarizes the main characteristics of the utterance recordings of the CIEMPIESS corpus.[11]

The audio of the utterances was standardized into recordings sampled at 16 kHz, with 16-bits in a NIST Sphere PCM mono format with a noise removal filtering when necessary.

3.3. Orthographic transcription of utterances

In addition to the audio collection of utterances, it was necessary to have their orthographic transcriptions. In order to create these transcriptions we followed these guidelines: First, the process begins with a canonical orthographic transcription of every utterance in the corpus. Later, these transcriptions were enhanced to mark certain phenomena of the Spanish language.

[8] Originally transmitted by: "RADIO IUS" (http://www.derecho.unam.mx/cultura-juridica/radio.php) and available for web downloading at: PODSCAT-UNAM (http://podcast.unam.mx/).

[9] For more details visit: http://www.ciempiess.org/CIEMPIESS_Statistics.html#Tabla2.

[10] To see a table that shows which sentences belong to a particular speaker, visit: http://www.ciempiess.org/CIEMPIESS_Statistics.html#Tabla8.

[11] For more details, see this chart at: http://www.ciempiess.org/CIEMPIESS_Statistics.html#Tabla1.

The considerations we took into account for the enhanced version were:

- Do not use capitalization and punctuation
- Expand abbreviations (e.g. the abbreviation "PRD" was written as: "pe erre de").
- Numbers must be written orthographically.
- Special characters were introduced (e.g. *N* for *ñ*, *W* for *ü*, $ when letter "x" sounds like phoneme /s/, *S* when letter "x" sounds like phoneme /ʃ/).
- The letter "*x*" has multiple phonemes associated to it.
- We wrote the tonic vowel of every word in uppercase.
- We marked the silences and disfluencies.

3.4. Enhanced transcription

As we have mentioned in the previous section, the orthographic transcription was enhanced by adding information of the pronunciation of the words and the disfluencies in the speech. The rest of this section presents such enhancements.

Most of the words in Spanish contain enough information about how to pronounce them, however there are exceptions. In order to facilitate the automatic creation of a dictionary, we added information of the pronunciation of the *x's* which has several pronunciation in the Mexican Spanish. Annotators were asked to replace *x's* by an approximation of its possible pronunciations. Doing this, we eliminate the need for an exception dictionary. Table 5 exemplifies such cases.

Another mark we annotate in the orthographic transcription is the indication of the tonic vowel in a word. In Spanish, a tonic vowel is usually identified by a raise in the pitch; sometimes this vowel is explicitly marked in the orthography of the word by an acute sign (e.g. "acción", "vivía"). In order to make this difference explicit among sounds, the enhanced transcriptions also marked as such in both the explicit and implicit cases (Table 6 exemplifies this consideration). We did this for two reasons, the most important is that we want to explore the effect of tonic and non-tonic vowels in the speech recognition, and the other one is that some software tools created for HTK or SPHINX do not manage characters with acute signs properly, so the best thing to do is to use only ASCII symbols.

Finally, silences and disfluencies were marked following this procedure:

Table 5
Enhancement for "x" letter.

Canonical transcriptions	Phoneme equivalence (IPA)	Transcription mark	Enhanced transcription
sexto, oxígeno	/ks/	KS	sE**KS**to, o**KS**Igeno
xochimilco, xilófono	/s/	$	**$**ochimIlco, **$**ilOfono
xolos, xicoténcatl	/ʃ/	S	**S**Olos, **S**icotEncatl
ximena, xavier	/x/	J	**J**imEna, **J**aviEr

Table 6
Example of tonic marks in enhanced transcription.

Canonical	Enhanced	Canonical	Enhanced
ambulancia	ambulAncia	perla	pErla
química	quImica	aglutinado	aglutinAdo
niño	nINo	ejemplar	ejemplAr
pingñino	pingWIno	tobogán	tobogAn

Table 7
Example of enhance transcriptions.

Original
< s> a partir del año mil novecientos noventa y siete **< /s>** (S1)
< s> es una forma de expresion de los sentimientos **< /s>** (S2)

Enhanced
< s> < sil> A partIr dEl ANo < sil> mIl noveciEntos **++dis++** novEnta y siEte **< sil> < /s>** (S1)
< s> < sil> Es **++dis++** Una fOrma dE eKSpresiOn < sil> dE lOs sentimiEntos **< sil> < /s>** (S2)

- An automatic and uniform alignment was produced using the words in the transcriptions and the utterance audios.
- Annotators were asked to align the words with their audio using the PRAAT system (Boersma & Weenink, 2013).
- When there was speech which did not correspond to the audio, the annotators were asked to analyze if it was a case of a disfluency. If positive, they were asked to mark it with a **++dis++**.
- The annotators were also ask to mark evident silences in the speech with a <sil>.

Table 7 compares a canonical transcription and its enhanced version.

Both segmentation and orthographic transcriptions in their canonical and enhance version (word alignment) are very time consuming processes. In order to transcribe and align the full utterances, we collaborated with 20 college students investing 480 h each to the project over two years. Three of them made the selection of utterances from the original audio files using the Audacity tool (Team, 2012) and they created the orthographic transcriptions in a period of six months, at the rate of one hour per week. The rest of collaborators spent one and a half year aligning the word transcriptions with the utterances for detecting silences and disfluencies. This last step implies that orthographic transcriptions were checked at least twice by two different person. Table 8 shows a chronograph of each task done per half year.

Table 8
Number of students working per semester and their labors.

Semester	Students	Labor
1st	3	Audio selection and orthographic transcriptions
2nd	6	Word alignment
3rd	6	Word alignment
4th	5	Word alignment

Table 9
Word frequency between CIEMPIESS and CREA corpus.

No.	Words in CREA	Norm. freq. CREA	Norm. freq. CIEMPIESS	No.	Words in CREA	Norm. freq. CREA	Norm. freq. CIEMPIESS
1	de	0.065	0.056	11	las	0.011	0.008
2	la	0.041	0.033	12	un	0.010	0.011
3	que	0.030	0.051	13	por	0.010	0.008
4	el	0.029	0.026	14	con	0.009	0.006
5	en	0.027	0.025	15	no	0.009	0.015
6	y	0.027	0.022	16	una	0.008	0.010
7	a	0.021	0.026	17	su	0.007	0.003
8	los	0.017	0.014	18	para	0.006	0.008
9	se	0.013	0.014	19	es	0.006	0.017
10	del	0.012	0.008	20	al	0.006	0.004

3.5. Distribution of words

In order to verify that the distribution of words in CIEMPIESS corresponds to the distribution of words in the Spanish language, we compare the distribution of the functional words in the corpus versus the "Corpus de Referencia del Español Actual" (*Reference Corpus of Current Spanish*, CREA).[12] The CREA corpus is conformed by 140,000 text documents. That is, more than 154 million of words extracted from books (49%), newspapers (49%), and miscellaneous sources (2%). It also has more than 700,000 word tokens. Table 9 illustrates the minor differences between frequencies of the 20 most frequent words in CREA and their frequency in CIEMPIESS.[13] As one can see, the distribution of functional words is proportional between both corpora.

We also calculate the mean square error ($MSE = 9.3 \times 10^{-8}$) of the normalized frequencies of the words between the CREA and the whole CIEMPIESS so we found that it is low and the correlation of the two distributions is 0.95. Our interpretation of this is that the distribution of words in the CIEMPIESS reflects the distribution of words in the Spanish language. We argue that this is relevant because the CIEMPIESS is then a good sample that reflects well the behavior of the language that it pretends to model.

4. Pronouncing dictionaries

The examples of audio utterances and their transcriptions alone are not enough to start the training procedure of the ASR system. In order to learn how the basic sounds of a language sound, it is necessary to translate words into sequences of phonemes. This information is codified in the pronunciation dictionary, which proposes one or more pronunciation for each word. These pronunciations are described as a sequence of phonemes for which a canonical set of phonemes has to be decided. In the creation of the CIEMPIESS corpus, we proposed the automatic extraction of pronunciations based on the enhanced transcriptions.

4.1. Phonetic alphabet

In this work we used the MEXBET phonetic alphabet that has been proposed to encode the phonemes and the allophones of Mexican Spanish (Cuetara-Priede, 2004; Hernández-Mena, Martínez-Gómez, & Herrera-Camacho, 2014; Uraga & Pineda, 2000). MEXBET is a heritage of our University and it has been successfully used over the years in several articles and thesis. Nevertheless, the best reason for choosing MEXBET is that this is the most updated phonetic alphabet for the Mexican Spanish dialect.

This alphabet has three levels of granularity from the phonological (T22) to the phonetic (T44 and T54).[14] For the purpose of the CIEMPIESS corpus we extended the *T*22 and *T*54 levels to what we call *T*29 and *T*66.[15] In the case of *T*29 these are the main changes:

- For T29 we added the phoneme */tl/* as in *iztaccíhuatl* → / i s. t a k. ˈ s i. u a. t l / → [i̪ s̪ . t a k. ˈ s i. w a. t l]. Even though the counts of the phoneme */tl/* are so low in the CIEMPIESS corpus, we decided to include it into MEXBET becuase in Mexico, many proper names of places need it to have a correct phonetic transcription.
- For T29 we added the phoneme */S/* as in *xolos* → / ˈ ʃ o. l o s / → [ˈ ʃ o. l o̯ s]
- For T29 we considered the symbols /a_7/, /e_7/, /i_7/, /o_7/ and /u_7/ of the levels T44 and T54 used to indicate tonic vowels in word transcriptions.

All these changes were motivated by the need to produce more accurate phonetic transcriptions following the analysis of

[12] See: http://www.rae.es/recursos/banco-de-datos/crea-escrito.

[13] Download the word frequencies of the words of CREA from: http://corpus.rae.es/lfrecuencias.html.

[14] For more detail on the different levels and the evolution of MEXBET through time see the charts in: http://www.ciempiess.org/Alfabetos_Foneticos/EVOLUTION_of_MEXBET.html.

[15] In our previous papers, we refer to the level *T*29 as *T*22 and the level *T*66 as *T*50 but this is incorrect because the number "22" or "44", etc. must reflect the number of phonemes and allophones considered in that level of MEXBET.

Table 10
Comparison between MEXBET T66 for the CIEMPIESS (left or in bold) and DIMEx100 T54 (right) databases.

Consonants	Labial	Labiodental	Dental	Alveolar	Palatal	Velar
Unvoiced Stops	p: p/p_c		t: t/t_c		k_j: k_j/k_c	k: k/k_c
Voiced Stops	b: b/b_c		d: d/d_c			g: g/g_c
Unvoiced Affricate					tS: tS/tS_c	
Voiced Affricate					dZ: dZ/dZ_c	
Unvoiced Fricative		f	s_[	s	**S**	x
Voiced Fricative	V		D **z_[**	z	Z	G
Nasal	m	**M**	n_[	n	**n_j** n~	N
Voiced Lateral			**l_[**	l	**tl l_j**	
Voiceless Lateral				**l_0**		
Rhotic				r(r		
Voiceless Rhotic				**r(_0 r(**-\		

Vowels	**Palatal**	**Central**	**Velar**
Semi-consonants	j		w
	i(		**u(**
Close	i		u
	I		**U**
Mid	e		o
	E		O
Open	a_j	a	a_2

Tonic Vowels	**Palatal**	**Central**	**Velar**
Semi-consonants	**j_7**		**w_7**
	i(_7		**u(_7**
Close	i_7		u_7
	I_7		**U_7**
Mid	e_7		o_7
	E_7		O_7
Open	a_j_7	a_7	a_2_7

Table 11
Example of transcriptions in IPA against transcriptions in MEXBET.

Word	Phonological IPA	Phonetic IPA
ineptitud	i. n e p. t i. ˈ t u d	i. n e̯ p. t i. ˈ t u̯ ð̞
indulgencia	i n. d u l. ˈ x e n. s i a	i̯ n̪ . d u̯ l. ˈ x e n. s j a
institución	i n s. t i. t u. ˈ s i o n	i̯ n s̪ . t i. t u. ˈ s j o̯ n

	MEXBET T29	MEXBET T66
ineptitud	i n e p t i t u_7 d	i n E p t i t U_7 D
indulgencia	i n d u l x e_7 n s i a	I n_[d U l x e_7 n s j a
institución	i n s t i t u s i o_7 n	I n s_[t i t u s i O_7 n

Cuetara-Priede (2004). Table 10 illustrates the main difference between *T*54 and *T*66 levels of the MEXBET alphabets.

Table 11 shows examples of different Spanish words transcribed using the symbols of the International Phonetic Alphabet (IPA) against the symbols of MEXBET.[16]

Table 12 shows the distribution of the phonemes in the automatically generated dictionary and compares it with the DIMEx100 corpus. We observe that both corpora share a similar distribution.[17]

4.2. Characteristics of the pronouncing dictionaries

The pronunciation dictionary consists of a list of words for the target language and their pronunciation at the phonetic or phonological level. Based on the enhanced transcription, we automatically transcribe the pronunciation for each word. For this, we followed the rules from Cuetara-Priede (2004)and Hernández-Mena et al. (2014). The produced dictionary consists of 53,169 words. Table 13 shows some examples of the automatically created transcriptions.

The automatic transcription was done using the *fonetica2*[18] library, that includes transcription routines based on rules for the T29 and the T66 levels of MEXBET. This library implements the following functions:

- *vocal_tonica*(): Returns the same incoming word but with its tonic vowel in uppercase (e.g. cAsa, pErro, gAto, etc.).

[16] To see the equivalences between IPA and MEXBET symbols see: http://www.ciempiess.org/Alfabetos_Foneticos/EVOLUTION_of_MEXBET.html#Tabla5.

[17] To see a similar table which shows distributions of the T66 level of CIEMPIESS, see http://www.ciempiess.org/CIEMPIESS_Statistics.html#Tabla6.

[18] Available at http://www.ciempiess.org/downloads, and for a demonstration, go to http://www.ciempiess.org/tools.

Table 12
Phoneme distribution of the T29 level of the CIEMPIESS compared to the T22 level of DIMEx100 corpus.

No.	Phoneme	Instances DIMEx100	Percentage DIMEx100	Instances CIEMPIESS	Percentage CIEMPIESS
1	p	6730	2.42	19,628	2.80
2	t	12,246	4.77	35,646	5.10
3	k	8464	3.30	29,649	4.24
4	b	1303	0.51	15,361	2.19
5	d	3881	1.51	34,443	4.92
6	g	426	0.17	5496	0.78
7	tS / TS	385	0.15	1567	0.22
8	f	2116	0.82	4609	0.65
9	s	20,926	8.15	68,658	9.82
10	**S**	0	0.0	736	0.10
11	x	1994	0.78	4209	0.60
12	Z	720	0.28	3081	0.44
13	m	7718	3.01	21,601	3.09
14	n	12,021	4.68	51,493	7.36
15	n∼	346	0.13	855	0.12
16	r(	14,784	5.76	38,467	5.50
17	r	1625	0.63	3546	0.50
18	l	14,058	5.48	32,356	4.63
19	**tl**	0	0.0	1	0.00014
20	i	9705	3.78	34,063	4.87
21	e	23,434	9.13	43,267	6.19
22	a	18,927	7.38	41,601	5.95
23	o	15,088	5.88	41,888	5.99
24	u	3431	1.34	13,099	1.87
25	**i_7**	0	0.0	16,861	2.41
26	**e_7**	0	0.0	61,711	8.83
27	**a_7**	0	0.0	39,234	5.61
28	**o_7**	0	0.0	26,233	3.75
29	**u_7**	0	0.0	9417	1.34

Table 13
Comparison of transcriptions in Mexbet T29.

Word	Enhanced-Trans	Mexbet T29
peñasco	peNAsco	p e n∼ a_7 s k o
sexenio	seKSEnio	s e k s e_7 n i o
xilófono	$ilOfono	s i l o_7 f o n o
xavier	JaviEr	x a b i e_7 r(
xolos	SOlos	S o_7 l o s

- *TT*(): "TT" is the acronym for "Text Transformation". This function produces the text transformations in Table 14 over the incoming word. All of them are perfectly reversible.
- *TT_INV*(): Produces the reverse transformations made by the TT() function.
- *div_sil*(): Returns the syllabification of the incoming word.
- *T*29(): Produces a phonological transcription in Mexbet T29 of the incoming word.
- *T*66(): Produces a phonetic transcription in Mexbet T66 of the incoming word.

5. Language model

The language model captures how words are combined in a language. In order to create a language model, a large set of examples of sentences in the target language is necessary.

Table 14
Transformations adopted to do phonological transcriptions in Mexbet.

No ASCII symbol: example	Transformation: example	Orthographic irregularity: example	Phoneme equivalence: example	Orthographic irregularity: example	Phoneme equivalence: example
á: cuál	cuAl	cc: accionar	/ks/: aksionar	gui: guitarra	/g/: gitaRa
é: café	cafE	ll: llamar	/Z/: Zamar	que: queso	/k/: keso
í: maría	marIa	rr: carro	R: caRo	qui: quizá	/k/: kisA
ó: noción	nociOn	ps: psicología	/s/: sicologIa	ce: cemento	/s/: semento
ú: algún	algUn	ge: gelatina	/x/: xelatina	ci: cimiento	/s/: simiento
ü: güero	gwero	gi: gitano	/x/: xitano	y (end of word): buey	/i/: buei
ñ: niño	niNo	gue: guerra	/g/: geRa	h (no sound): hola	: ola

Table 15
Characteristics of the raw text of the newsletters used to create the language model.

Description	Value
Total of newsletters	2489
Oldest newsletter	12:30 h 01/Jan/2010
Newest newsletter	20:00 h 18/Feb/2013
Total number of text lines	197,786
Total words	1,642,782
Vocabulary size	113,313
Average of words per newsletters	660
Largest newsletter	2710
Smallest newsletter	21

Table 16
Characteristics of the processed text utilized to create the language model.

Description	Value
Total number of words	1,505,491
Total number of words with no repetition	49,085
Total number of text lines	279,507
Average of words per text line	5.38
Number of words in the largest text line	43
Number of words in the smallest text line	2

As a source of such examples, we use a university newsletter which is about academic activities.[19] The characteristics of the newsletters are presented in Table 15.

Even though the amount of text is relatively small compared with other newsletters, it is still being one order magnitude bigger than the amount of transcriptions in the CIEMPIESS corpus, and it will not bring legal issues to us because it belongs to our own university.

The text taken from the newsletters was later post-processed. First, they were divided into sentences, and we filtered punctuation signs and extra codes (e.g, HTML and stylistics marks). The dots and commas were substituted with the newline character to create a basic segmentation of the sentences. Every text line that included any word unable to be phonetized with our *T*29() or *T*66() functions were excluded of the final version of the text. Additionally the lines with one unique word were excluded. Every word was marked with its corresponding tonic vowel with the help of our automatic tool: the *vocal_tonica*() function. Table 16 shows the properties of the text utilized to create the language model after being processed.

Table 17 shows the comparison between the 20 most common words in the processed text utilized to create the language model; the MSE among the two word distributions is of 7.5×10^{-9} with a correlation of 0.98. These metrics point out to a good comprehension of the Spanish language of Mexico City.

6. Evaluation experiments

In this section we show some evaluations of different aspects of the corpus. First, we show the evaluation of the automatic transcription used during the creation of the pronunciation dictionary. Second, we show different baselines for different speech recognizer systems: HTK (Young et al., 2006), Sphinx (Chan, Gouvea, Singh, Ravishankar, & Rosenfeld, 2007; Lee, Hon, & Reddy, 1990) and Kaldi (Povey et al., 2011) which are state of the art ASR systems. Third, we show an experiment in which the benefit of marking the tonic syllables during the enhance transcription can be seen.

6.1. Automatic transcription

In these evaluations we measure the performance of different functions of the automatic transcription. First we evaluated the performance of the *vocal_tonica*() function which indicates the tonic vowel of an incoming Spanish word. For this, we randomly took 1452 words from the CIEMPIESS vocabulary (12% of the corpus) and we predicted their tonic transcription. The automatic generated transcriptions were manually checked by an expert. The result is that 90.35% of the words were correctly predicted. Most of the errors occurred in conjugated verbs and proper names. Table 18 summarizes the results of this evaluation.

The second evaluation focuses on the *T*29() function which transcribes words at the phonological level. In this case we evaluated against the TRANSCRÍBEMEX (Pineda et al., 2004) and the transcriptions done manually by experts for the DIMEx100 corpus (Pineda et al., 2004). In order to compare with the system TRANSCRÍBEMEX, we took the vocabulary of the DIMEX100 corpus but we had to eliminate some words. First we removed entries with the *archiphonemes*[20]: [-B], [-D], [-G], [-N], [-R] since they are not a one-to-one correspondence with a phonological transcription. Then, words with the grapheme *x* were eliminated since TRANSCRÍBEMEX only supports one of four pronunciations. After this, both systems produced the same transcription **99.2%** of the times. In order to evaluate against transcriptions made by experts, we took the pronouncing dictionary of the DIMEX100 corpus and we removed words with the *x* phonemes and the alternative pronunciations if there was any. This shows us that the transcriptions made by our *T*29() function were similar to the transcriptions made by experts **90.2%** of the times.

Tables 19 and 20 summarizes the results for both comparisons. In conclusion, besides the different conventions there is not a noticeable difference, but when compared with human experts, there still room for improvement of our system.

6.2. Benchmark systems

We created three benchmarks based on state of the art systems: HTK (Young et al., 2006), Sphinx (Chan et al., 2007; Lee

[19] Newsletter from website: http://www.dgcs.unam.mx/boletin/bdboletin/basedgcs.html.

[20] An archiphoneme is a phonological symbol that groups several phonemes together. For example, [-D] is equivalent to any of the phonemes /d/ or /t/.

Table 17
Word Frequency of the language model and the CREA corpus.

No.	Words in CREA	Norm. freq. CREA	Norm. freq. News	No.	Words in CREA	Norm. freq. CREA	Norm. freq. News
1	de	0.065	0.076	11	las	0.011	0.011
2	la	0.041	0.045	12	un	0.010	0.009
3	que	0.030	0.028	13	por	0.010	0.010
4	el	0.029	0.029	14	con	0.009	0.010
5	en	0.027	0.034	15	no	0.009	0.006
6	y	0.027	0.032	16	una	0.008	0.008
7	a	0.021	0.017	17	su	0.007	0.005
8	los	0.017	0.016	18	para	0.006	0.010
9	se	0.013	0.015	19	es	0.006	0.008
10	del	0.012	0.013	20	al	0.006	0.005

Table 18
Evaluation of the vocal_tonica() function.

Words taken from the CIEMPIESS database	1539
Number of Foreign words omitted	87
Number of Words analyzed	1452
Wrong accentuation	140
Correct accentuation	1312
Percentage of correct accentuation	90.35%

Table 19
Comparison between TRANSCRÍBEMEX and the T29() function.

Words in DIMEx100	11,575
Alternate pronunciations	2590
Words with grapheme “x”	202
Words with grapheme “x” into an alternate pron.	87
Archiphonemes	45
Number of words analyzed	8738
Non-identical transcriptions	67
Identical transcriptions	8670
Percentage of identical transcriptions	99.2%

Table 20
Comparison between transcriptions in DIMEx100 dictionary (made by humans) and the T29() function.

Words in DIMEx100	11,575
Words with grapheme “x”	289
Number of words analyzed	11,286
Non-identical transcriptions	1102
Identical transcriptions	10,184
Percentage of identical transcriptions	90.23%

et al., 1990) and Kaldi (Povey et al., 2011). The CIEMPIESS corpus is formatted to be used directly in a Sphinx setting. In the case of HTK we created a series of tools that can read directly the CIEMPIESS corpus[21] and finally for Kaldi we created the setting files for the CIEMPIESS. We set up a speech recognizer for each system using the train set of the CIEMPIESS corpus

[21] See the “HTK2SPHINX-CONVERTER” (Hernández-Mena & Herrera-Camacho, 2015) and the “HTK-BENCHMARK” available at http://www.ciempiess.org/downloads.

Table 21
Benchmark among different systems.

System	WER (the lower the better)
Sphinx	44.0%
HTK	42.45%
Kaldi	33.15%

Table 22
Best recognition results in learning curve.

Condition	WER (the lower the better)
T29 TONICS	44.0%
T29 NO TONICS	45.7%
T66 TONICS	50.5%
T66 NO TONICS	48.0%

and we evaluated the performance utilizing its test set. Every system were configured with their corresponding default parameters and a trigram-based language model. Table 21 shows the performance for each system.

6.3. Tonic vowels

Given the characteristics of the CIEMPIESS corpus, we decided to evaluate the effect of the tonic vowels marked in the corpus. For this we trained four acoustic models for the Sphinx system. These were tested in the corpus using the standard language model of the corpus. Table 22 presents the *word error rate* (WER, the lower the better) for such cases. We can observe that the distinction of the tonic vowel helps improve the performance.[22] However, the use of phonetic transcriptions (T66 level of MEXBET) does have a negative effect on the performance of the speech recognizer.

Using the same configurations found with the experiment of the tonic vowels, we created four learning curves. Fig. 3 presents the curves, we also can notice that a phonetic transcription (T66)

[22] In Table 22 “TONICS” means that we used tonic vowel marks for the recognition experiment and “NO TONICS” means that we did not.

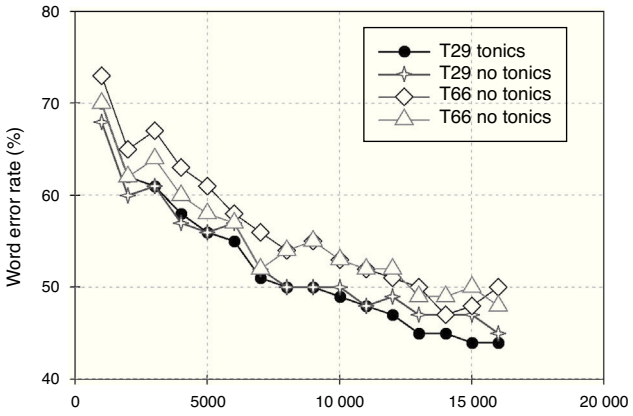

Fig. 3. Learning Curves for different training conditions.

was not beneficial while using a phonological (T29) even with a small amount of data yields to a better performance.

7. Conclusions

In this work we have presented the CIEMPIESS corpus, the methodology, and the tools used to create it. The CIEMPIESS corpus is an open resource composed by a set of recordings, its transcriptions, a pronunciation dictionary, and a language model. The corpus is based on speech from radio broadcast interviews in the Central Mexican accent. We complemented each recording with its enhanced transcription. The enhanced transcription consisted of orthographic convention which facilitated the automatic phonetic and phonological transcription. With these transcriptions, we created the pronunciation dictionary.

The recordings consist of 17 h of spoken language. To our knowledge, it is the largest collection openly available of Mexican Spanish spontaneous speech. In order to test the effectiveness of the resource, we created three benchmarks based on the Sphinx, HTK and Kaldi systems. In all of them, it showed a reasonable performance for the available speech (e.g. Fisher Spanish corpus (Kumar, Post, Povey, & Khudanpur, 2014) reports 39% WER using 160 hours).[23]

The set of recordings were manually transcribed in order to reduce the phonetic ambiguity among *x* letter. We also marked the tonic vowel which is characteristic of Spanish. These transcriptions are important when building an acoustic model. Conventions were essential in facilitating the automatic creation of the pronunciation dictionary. This dictionary and its automatic phonetic transcriptions were evaluated by comparing with both manual and automatic transcriptions finding a good coverage (>1% difference automatic, >10% difference manual).

As a part of the CIEMPIESS corpus we include a language model. This was created using text from a university magazine which focuses on academic and day to day events. This resource was also compared with the statistics from Spanish and we found that is close to Mexican Spanish.

The availability of the CIEMPIESS corpus makes it a great option compared with other Mexican Spanish resources which are not easily or freely available. It makes further research in speech technology possible for this dialect. Additionally, this work presents the methodology and the tools which can be adapted to create similar resources for other Spanish dialects. The corpus can be freely obtained from the LDC website (Hernández-Mena, 2015) and the CIEMPIESS web page.[24]

Conflict of interest

The authors have no conflicts of interest to declare.

Acknowledgements

We thank UNAM PAPIIT/DGAPA project IT102314, CEP-UNAM and CONACYT for their financial support.

References

TC-STAR 2005 evaluation package – ASR Mandarin Chinese ELRA-E0004. DVD, 2006.

NetDC Arabic BNSC (Broadcast News Speech Corpus) ELRA-S0157. DVD, 2007.

TC-STAR Spanish training corpora for ASR: Recordings of EPPS speech ELRA-S0252. DVD, 2007.

CLEF QAST (2007–2009) – Evaluation package ELRA-E0039. CD-ROM, 2012.

LILA Hindi Belt database ELRA-S0344, 2012.

Audhkhasi, K., Georgiou, P. G., & Narayanan, S. S. (2011). *Reliability-weighted acoustic model adaptation using crowd sourced transcriptions*. pp. 3045–3048.

Bernstein, J., et al. (1995). *LATINO-40 Spanish Read News LDC95S28*. Web Download.

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer. Version 5.3.51*.. Retrieved from http://www.praat.org/

Byrne, W., et al. (2014). *Hispanic–English database LDC2014S05. DVD*.

Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., & Piperidis, S. (2014). *Conference on language resources and evaluation (LREC)*. ELRA.

Canavan, A., & Zipperlen, G. (1996). *CALLHOME Spanish speech LDC96S35*. Web Download.

Chan, A., Gouvea, E., Singh, R., Ravishankar, M., & Rosenfeld, R. (2007). *(Third Draft) The hieroglyphs: Building speech applications using CMU Sphinx and related resources*.

Cole, R., & Muthusamy, Y. (1994). *OGI multilanguage corpus LDC94S17*. Web Download.

Cuetara-Priede, J. (2004). *Fonética de la ciudad de México, Aportaciones desde las tecnologías del habla (M.sc. thesis in Spanish linguistics)*. (in Spanish).

Federico, M., Giordani, D., & Coletti, P. (2000). Development and evaluation of an Italian Broadcast News Corpus. In *LREC*. European Language Resources Association, http://dblp.uni-trier.de/db/conf/lrec/lrec2000.html#FedericoGC00; http://www.lrec-conf.org/proceedings/lrec2000/pdf/95.pdf; http://www.bibsonomy.org/bibtex/2e5569e427c9fffd61769cf12a3991994/dblp.

Fiscus, J., et al. (2001). *1997 HUB4 Broadcast News evaluation non-English test material LDC2001S91*. Web Download.

Garofolo, J., et al. (1993). *TIMIT acoustic–phonetic continuous speech corpus LDC93S1*. Web Download.

Graff, D. (2001). *TDT3 Mandarin audio LDC2001S95*. Web Download.

Graff, D., et al. (2010). *Fisher Spanish speech LDC2010S01. DVD*.

[23] Configuration settings and software tools available at: http://www.ciempiess.org/downloads.

[24] http://www.ciempiess.org/downloads.

Hernández-Mena, C. D. (2015). *CIEMPIESS LDC2015S07*. Web Download.

Hernández-Mena, C. D., & Herrera-Camacho, A. (2015). Creating a grammar-based speech recognition parser for Mexican Spanish using HTK, compatible with CMU Sphinx-III system. *International Journal of Electronics and Electrical Engineering*, *3*, 220–224.

Hernández-Mena, C. D., & Herrera-Camacho, J. A. (2014). CIEMPIESS: A new open-sourced Mexican Spanish radio corpus. In N. C. C. Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 371–375). Reykjavik, Iceland: European Language Resources Association (ELRA).

Hernández-Mena, C. D., Martínez-Gómez, N.-N., & Herrera-Camacho, A. (2014). *A set of phonetic and phonological rules for Mexican Spanish revisited, updated enhanced and implemented*. pp. 61–71. CIC-IPN volume 83.

Kirschning, I. (2001). *Research and development of speech technology and applications for Mexican Spanish at the Tlatoa Group, CHI'01 Extended Abstracts on Human Factors in Computing Systems*. pp. 49–50.

Kumar, G., Post, M., Povey, D., & Khudanpur, S. (2014). Some insights from translating conversational telephone speech. *IEEE*, 3231–3235.

Langmann, D., Haeb-Umbach, R., Boves, L., & den Os, E. (1996). Fresco: The French telephone speech data collection-part of the European SpeechDat (M) project. In *IEEE international conference on volume 3* (pp. 1918–1921).

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2012). *RSR2015: Database for text-dependent speaker verification using multiple pass-phrases*.

LaRocca, S., & Chouairi, R. (2002). *West point Arabic speech LDC2002S02*. Web Download.

Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *Acoustics, Speech and Signal Processing*, *38*, 35–45.

Llisterri, J. (2004). Las tecnologías del habla para el espa nol. In *Fundación Espa nola para la Ciencia y la Tecnología* (pp. 123–141).

Moreno-Fernández, F., & Otero, J. (2007). *Atlas de la lengua espa nola en el mundo*. Real Instituto Elcano-Instituto Cervantes-Fundación Telefónica.

Morgan, J. (2006). *West point heroico Spanish speech LDC2006S37*. Web Download.

Moya, E., Hernández, M., Pineda, L. A., & Meza, I. (2011). Speech recognition with limited resources for children and adult speakers. *IEEE*, 57–62.

Olguín-Espinoza, J. M., Mayorga-Ortiz, P., Hidalgo-Silva, H., Vizcarra-Corral, L., & Mendiola-Cárdenas, M. L. (2013). VoCMex: A voice corpus in Mexican Spanish for research in speaker recognition. *International Journal of Speech Technology*, *16*, 295–302.

de_Luna Ortega, C. A., Mora-González, M., Martínez-Romo, J. C., Luna-Rosas, F. J., & Mu noz-Maciel, J. (2014). Speech recognition by using cross correlation and a multilayer perceptron. *Revista Electrónica Nova Scientia*, *6*, 108–124.

Others (1998). 1997 Spanish Broadcast News Speech (HUB4-NE) LDC98S74. Web Download.

Pineda, L. A., Castellanos, H., Priede, J. C., Galescu, L., Juarez, J., Llisterri, J., Prez-Pavn, P., & Villase nor, L. (2010). The corpus DIMEx100: Transcription and evaluation. *Language Resources and Evaluation*, 44.

Pineda, L. A., Pineda, L. V., Cuétara, J., Castellanos, H., & López, I. (2004). DIMEx100: A new phonetic and speech corpus for Mexican Spanish. In C. Lemaître, C. A. R. García, & J. A. González (Eds.), *IBERAMIA, Vol. 3315* (pp. 974–984). Springer, http://dblp.uni-trier.de/db/conf/iberamia/iberamia2004.html#PinedaPCCL04; http://dx.doi.org/10.1007/978-3-540-30498-2_97; http://www.bibsonomy.org/bibtex/20bb754fd7ab188238a444cb5033f3bd1/dblp.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

Raab, M., Gruhn, R., & Noeth, E. (2007). *IEEE workshop on non-native speech databases*. pp. 413–418.

Team, A. (2012). *Audacity*.

Uraga, E., & Gamboa, C. (2004). *VOXMEX speech database: Design of a phonetically balanced corpus*.

Uraga, E., & Pineda, L. A. (2000). *A set of phonological rules for Mexican Spanish*. México: Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas.

Varela, A., Cuayáhuitl, H., & Nolazco-Flores, J. A. (2003). Creating a Mexican Spanish version of the CMU Sphinx-III speech recognition system. In A. Sanfeliu, & J. Ruiz-Shulcloper (Eds.), *CIARP, Volume 2905 of Lecture notes in computer science* (pp. 251–258). Springer, http://dblp.uni-trier.de/db/conf/ciarp/ciarp2003.html#VarelaCN03; http://dx.doi.org/10.1007/978-3-540-24586-5_30; http://www.bibsonomy.org/bibtex/2d93f813e7f3fd0a990b17e571d39e958/dblp.

Wang, H. M., Chen, B., Kuo, J. W., & Cheng, S. S. (2005). MATBN: A Mandarin Chinese Broadcast News corpus. *International Journal of Computational Linguistics and Chinese Language Processing*, *10*, 219–236.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2006). *The HTK Book (for HTK version 3.4)*.