VOXMEX Speech Database: Design of a Phonetically Balanced Corpus

Esmeralda Uraga and César Gamboa

Departamento de Ciencias de la Computación, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México, México.
Email : {euraga, cgamboa}@leibniz.iimas.unam.mx

Abstract

We present a method for designing a phonetically balanced speech corpus. In this method, we used a phonotactic approach to design the phonetic content of VOXMEX: a phonetically balanced corpus for Mexican Spanish. The transcriptions of VOXMEX contain a complete coverage of phonemes and allophones of Mexican Spanish in every possible context. This corpus is designed for doing phonetic research and acoustic modeling in the speech recognition area. We are recording the readings of the designed text corpus to obtain the speech data of VOXMEX. Our main goal in this project is to construct a phonologically representative speech corpus for Mexican Spanish.

1. Introduction

Spanish is the third spoken language over the world: about 330 million people speak Spanish. Also, Spanish is the first language studied as foreign language in non-Hispanic countries of America and Europe. In Mexico, about 100 million people speak the variety of Mexican Spanish. We note that there are many similarities in Spanish dialects, but there are also differences among these dialects because each one of them has evolved in distinct ways.

Despite this situation, there is no precise phonetic knowledge about the Spanish spoken in Mexico. In fact, without phonetic knowledge, constructing a phonetically balanced corpus is a difficult task. Moreover, in Mexico there are virtually neither suitable nor available spoken language resources for phonetic research. Hence, the development of speech technologies for human-computer interaction for the Mexican community has been limited.

The main goal of this project is to construct a phonologically representative speech corpus. Although each speaker has his/her individual way of speaking, the speech of the speakers of the same language has the same phonological structure. For this reason, we select the phonological level to design our corpus.

In this paper, we describe a method to design a phonetically balanced speech corpus. With this method, we design VOXMEX: a phonetically balanced corpus for Mexican Spanish. This corpus is designed for doing phonetic research and acoustic modeling in the speech recognition area. We used a phonotactic approach to design the phonetic content of VOXMEX. At this stage, The transcriptions of the designed corpus contains a complete coverage of phonemes and allophones of Mexican Spanish in every possible context. We are recording the readings of the designed text corpus to obtain the speech data of VOXMEX. We are obtaining the allophones of Mexican Spanish from the acoustic realization of the phonemes contained in the designed corpus.

2. Corpus Design

Experience has shown that an adequate speech database for training acoustic models is critical for successful deployment of speech recognition systems. Thus, the design of a phonetically balanced corpus for training and evaluating a speech recognition system has become very important (Wang, 1998). In fact, many efforts have been devoted to collecting speech databases in different languages (Angelini et al., 1994; Garofolo, 1993; Radová, 1998; Vaufreydaz et al., 2000).

Sentences containing phonetic events according to their frequency of occurrence in natural speech are called *phonetically balanced sentences* (Gibbon et al., 1997).

The phonetically balanced corpus presented in this paper consists of a set of sentences. The phonetic transcription of this set of sentences includes all recognition units in some phonetic distribution proportional to the phonetic distribution of Mexican Spanish (see Table 3 in section 2). To estimate the frequency distribution of the allophones in natural speech, the frequency distribution of phonetic transcriptions of a speech corpus is computed. Consider, for instance, the statistical study of the frequency of occurrence of the Spanish allophones undertook by Llisterri and Mariño (1993). Since neither an appropriate speech corpus nor data on the frequency distribution of allophones for Mexican Spanish were available, we estimated the frequency distribution of allophones from a pronunciation dictionary of Mexican Spanish.

We base the design of VOXMEX on the following idea: Given a set of sentences whose a phonological transcription contains all the phonemes in all their possible contexts, the reading of these sentences can produce a phonetically rich and balanced speech corpus.

2.1 Design method

In this section, we propose a method for designing a phonetically balanced corpus. In this method, we use a phonotactic approach to design a text corpus with a complete coverage of phonemes and allophones in every possible context. The phonological system for Mexican Spanish proposed by Perissinoto (1975) consists of 22 phonological units: 5 vowels and 17 consonants. To define every possible context for each unit, we considered the successor and the predecessor phonemes in inter- and intra-word contexts, the type of stressed or non-stressed syllable, and the different positions in the syllable, the word and the sentence. The most of the inter-word contexts used in this work were found in (RAE, 1999).

Given *C*, the set of all possible phonological contexts that occur in a particular language *L* (e.g. Mexican Spanish), the method for designing a phonetically balanced corpus is described in four steps as follow:

1) Obtain a set of words for each phonological context: For each phonological context *j* in *C*, obtain W_j, a non empty set of words or multi-words containing *j* in its transcription.

2) Design a base corpus: Select a word w_i from W_i and a word w_j from W_j, and construct a sentence containing at least the words w_i and w_j. Repeat this step until all phonological contexts in *C* have been included in the corpus. The set of sentences designed in this step is called the *base corpus*.

3) Add the complementary corpus: For each recognition unit *r*, define min_r, the desired minimum frequency of *r* in the corpus, and compute f_r, the frequency of *r* in the corpus.

For every recognition unit *r* such that f_r is less than min_r, construct a set of new sentences containing $(min_r - f_r)$ instances of unit *r* in its transcription. The set of sentences constructed in this step is called the *complementary corpus*, and is part of the designed corpus.

4) Balance the corpus: Define δ as a number between 0 and 1. Let *RP* be a reference percentage distribution of recognition units, where RP_r is the percentage of *r* in *RP*. Let *CP* the percentage distribution of the designed corpus computed from its phonetic transcriptions. CP_r is the percentage of *r* in *CP*. Compute *cc*, the correlation coefficient between *RP* and *CP*. While *cc* is less than δ, do the following:

For each unit *r* such that CP_r is less than RP_r, execute one of the following instructions:

a) Insert a sentence containing *r* in its transcription.

b) Insert a word or short phrase containing the unit *r* in a sentence of the *complementary corpus*.

To illustrate the step 1, we present some phonological contexts of Mexican Spanish and some words and multi-words containing such phonological contexts. Consider, for instance, the following contexts for the voiceless palatal affricate phoneme tʃ of Spanish whose corresponding grapheme is *ch* (see Tables 1 and 2): In inter-word contexts, the predecessor phonemes of tʃ are {s, ɾ, l, n, a, e, i, o, u,} and its successor phonemes are {a, e, i, o, u}; in intra-word contexts the predecessor phonemes of tʃ are {b, d, x, t}. Phoneme tʃ is in a stressed syllable in the word *ar**chi**vo* and in a non-stressed syllable in the word ***chi**lorio*. Phoneme tʃ is always at the beginning of the syllable. The phoneme tʃ is in the initial position in the word ***cha**to* and in a non-initial position in the word *no**che***. The symbol # represents the beginning of a sentence. So, the word *#Chonita* must be placed at the beginning of a designed sentence.

For the phonological context /tʃ+o/ we can define the following set of words $W_{\text{tʃ+o}}$= {*ancho, cacho, Chonita, deschongaron, hecho, mucho*}, and for the context /tʃ+i/ we can define the set $W_{\text{tʃ+i}}$= {*Alchichica, archivo, chicano, chino, chilorio*}.

In the step 2, we can select the word *mucho* from the set $W_{\text{tʃ+o}}$ and the word *Alchichica* from the set $W_{\text{tʃ+i}}$. With these words, we can design the following sentence: *En la laguna de Alchichica siempre hace mucho frío.*

The sentences in VOXMEX were obtained in distinct ways: from manual design of sentences, from plain text of Internet newspapers, and from the Spanish corpus described in (CREA, 2003) and (Villaseñor et al., 2001).

With the performance of the steps 1 and 2 of this method, we obtained a text corpus whose phonetic transcription is phonetically rich (with at least one sample of each allophone). Hence, at phonetic level, this corpus is a *base corpus*. To keep the complete coverage of phonemes and allophones in the corpus, no sentence must be deleted.

The step 3 is important to guarantee enough samples of each recognition unit. For training acoustic models in a speech recognition system, it is necessary a minimum of min_r samples for each recognition unit *r*, where the recognition units can be phonemes, biphonemes, allophones, syllables, etc. For example, in the work reported by Gamboa (2001) min_r=50 was defined for every phonological unit of Mexican Spanish.

In the step 4 of this method, we took δ=0.9. If the correlation coefficient *cc* is equal to 1, then the percentage distributions *CP* and *RP* are the same. Although option b) implies a manual task, this is a way of increasing the percentage of the unit *r* without considerably modifying the percentage of other phonetic units since, in general, a word has less phonetic units than a sentence. In the next section, we give more details of the results obtained in the step 4 of this method.

Predecessor	Phoneme	Example
s	**tʃ**	de**sch**ongaron
ɾ	**tʃ**	A**rch**ivo
l	**tʃ**	co**lch**a, A**lch**ichica
n	**tʃ**	A**nch**o
Vowels: a, e, i, o, u	**tʃ**	c**ach**o, h**ech**o, L**ich**a, c**och**e, m**uch**o
#	**tʃ**	**#Ch**onita
b	**tʃ**	clu**b ch**icano
d	**tʃ**	Davi**d ch**ocó
x	**tʃ**	relo**j ch**ino
t	**tʃ**	ba**t ch**afa

Table 1. Predecessor phonemes of **tʃ**.

Phoneme	Successor	Example
tʃ	Vowels: a, e, i, o, u	**cha**to, no**che**, **chi**lorio, **cho**que, **chu**eco

Table 2. Successor phonemes of **tʃ**.

2.2 Phonetic analysis of VOXMEX

A statistical analysis of the frequency of occurrence of the Mexican Spanish allophones was carried out to verify whether VOXMEX is phonetically balanced. Each sentence in VOXMEX was automatically transcribed applying a set of rules for grapheme to phone conversion. We used MEXBET, a computational phonetic alphabet for Mexican Spanish, to transcribe the corpus (Uraga, 1999; Uraga & Pineda, 2002).

Since no data on the percentage distribution of the allophones of Mexican Spanish were available, we computed the phonetic percentage distribution from a pronunciation dictionary. This distribution was taken as the reference distribution (see column Dict. in Table 3). Also, we compare the phonetic percentage distribution of

VOXMEX with other distributions of Spanish (Llisterri & Mariño, 1993; Rojo, 1991) (see Table 3).

IPA	MEXBET	Llisterri & Mariño	Rojo	Dict.	VOXMEX
p	P	2.6	2.66	2.26	2.54
b	B	0.45	2.66	1.65	0.44
t	T	4.63	4.48	4.99	4.29
d	D	0.76	4.79	1.22	0.88
k	K	4.04	3.98	4.6	3.90
g	G	0.11	0.95	0.36	0.17
m	M	3.63	3.09	3.02	2.90
n	n	7.02	6.99	5.83	6.51
ɲ	N	0.27	0.19	0.34	0.12
ŋ	nj	0.46	in /n/	0.26	0.30
tʃ	tS	0.4	0.28	0.74	0.34
β	V	2.47	in /b/	1.68	2.14
f	f	0.51	0.68	0.92	1.02
θ[1]	T	1.53	1.68		
ð	D	3.2	in /d/	2.56	4.05
s	s	6.95	7.58	7.74	9.10
z	z	1.33	in /s/	in /s/	in /s/
y	dZ	0.19	0.22	0.66	1.33
x	x	0.63	0.73	0.99	0.80
ɣ	G	0.79	in /g/	0.89	0.91
l	l	4.25	5.08	3.84	5.44
ʎ[1]	L	0.54	0.38		
r	r	0.4	0.79	1.08	0.71
ɾ	r(	4.25	5.67	7.75	6.02
i	i	4.29	7.5	6.16	4.50
j	j	2.6	in /i/	1.98	1.92
e	e	13.72	13.51	9.86	13.32
a	a	13.43	13.4	15.79	13.61
o	o	10.37	9.57	9.94	9.46
u	u	1.98	3.16	2.28	2.42
w	w	1.35	in /u/	0.58	0.86

Table 3. Percentage distribution of Spanish allophones.

Specifically, in Table 4 we show the correlation coefficients between our results and each data reported in Table 3.

3. Recordings

We are getting the phonetically balanced corpus from the acoustic recordings of phonemes. The acoustic realization of phonemes is being obtained from the reading of the designed text corpus. These recordings are still in progress. In the next section, we present some results obtained from the first stage.

3.1 Number and type of speakers

We plan to record 200 speakers in two stages: At the first stage, we have recorded 138 speakers (50% women and 50% men). The age of the speakers ranks between 19 and 48 years and the native language of each speaker is Mexican Spanish. To record a dialectally representative corpus of Mexican Spanish, we are recording speakers from the 5 main dialectal regions of Mexico (see Table 5). At the second stage, we will complete the number of pending speakers. We plan that the number of speakers be regionally proportional to the Mexican population (INEGI, 2004). In Figure 1, we compare the percentage of population with the percentage of speakers per region.

Phonetic Distribution 1	Phonetic Distribution 2	Correlation Coefficient
VOXMEX	Dictionary	0.96
VOXMEX	Llisterri & Mariño	0.98
VOXMEX	Rojo	0.93

Table 4. Correlation coefficients between phonetic distributions.

Region	States and areas
Central	Morelos, Guerrero, Michoacán, Jalisco, Distrito Federal, Puebla, Estado de México, Hidalgo, Tlaxcala, Guanajuato, Querétaro and Nayarit
North	Baja California, Baja California Sur, Sonora, Sinaloa, Chihuahua y oeste de Durango, Coahuila, Oeste de Nuevo León, este de Durango, Aguascalientes, Zacatecas, San Luis Potosí, Tamaulipas
Pacific	Oaxaca, coast of Guerrero, coast of Michoacán, Colima, coast of Jalisco
Gulf	Tabasco and Veracruz
Southeast	Yucatán, Campeche, Quintana Roo and Chiapas

Table 5: States and areas of each dialectal region.

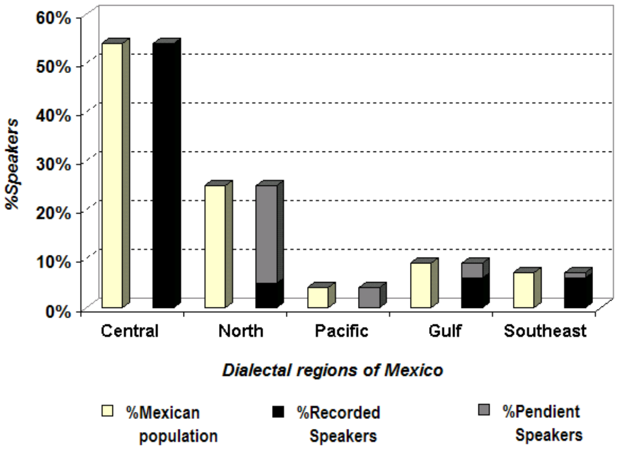

Figure 1. Percentages of Mexican population and speakers.

3.2 Recording Environment and Tools

The recordings have been collected in an office environment and in a house environment. We are using the CSLU Toolkit (CSLU, 2004) and XWaves (Entropics). We are using PC computers with a head-mounted or with a built-in microphone.

[1] In Mexican Spanish, it is used [s] instead of [θ] and [y] instead of [ʎ].

4. Transcriptions

A small portion of the corpus was manually transcribed and was used for training the acoustic models of an automatic transcription tool. With this tool, we transcribed the corpus at orthographic and phonological level.

These transcriptions have also been used to retrain the new acoustic models for the automatic transcription tool. In turn, with this tool we have created a new version of these transcriptions. These transcriptions are in verification process. At the end of the recording stage, we will begin the phonetic transcription process of the speech data.

5. Conclusions

We have presented a method based on a phonotactic approach to design a speech corpus. With this method we obtained a phonetically balanced corpus. Additionally, the transcriptions of this corpus contain a complete coverage of phonemes and allophones of Mexican Spanish. Also, we have presented a phonetic percentage distribution for Mexican Spanish, which is very similar to other percentage distributions of Spanish. For phonetic languages (like Dutch or Italian), our approach could work fairly well. The performance of our acoustic models trained with a part of VOXMEX and evaluated within a continuous speech recognition system, was highly satisfactory (Gamboa, 2001).

6. Future work

At present, recording phase is at an advanced stage. As future work, we intend to complete the recordings and transcriptions of the speech database. Additional experiments with this corpus will be realized to improve the performance of our speech recognition system.

Acknowledgments

The authors gratefully acknowledge the useful discussions and feedback from Javier Cuétara. Also, we thank the following people: Carlos Miranda and Linda Rosales from ITESM Campus Cuernavaca, Juan Pablo Conradis and Pamela López from “Jóvenes hacia la investigación” (UNAM program), Carolina Zazueta and the students from FFyL-UNAM who have collaborated in the recordings of the designed corpus. This work was partially supported by grants 27948A and 499316-5-C092A from the Consejo Nacional de Ciencia y Tecnología (CONACYT).

References

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R. and Omologo, M. (1994). Speaker independent continuous speech recognition using an acoustic-phonetic Italian corpus. In Proceedings of the International Conference of Spoken Language Processing, vol. III, (pp. 1391--1394).

CREA (2003). Corpus de Referencia del Español Actual. http://corpus.rae.es/creanet.html

Cole, R. (1999). Tools for research and education in speech science. In Proceedings of the International Conference of Phonetic Sciences, San Francisco, CA, USA.

Gamboa, C. (2001). Un Sistema de Reconocimiento de Voz para el Español. Tesis de Licenciatura. UNAM. México.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, USA.

Gibbon, D., Moore, R., Winsky, R. (Eds.) (1997). Handbook of Standards and Resources for Spoken Language Systems. Mouton De Gruyter, Berlin, New York.

Llisterri, J. (1993). Spanish adaptation of SAMPA and automatic phonetic transcription, Universidad Autónoma de Barcelona, Barcelona, España.

INEGI (2004). Instituto Nacional de Estadística Geografía e Informática: http://www.inegi.gob.mx.

Perissinoto, G. (1975). Fonología del español hablado en la Ciudad de México. El Colegio de México, México.

Radová, V. (1998). Design of the Czech Speech Corpus for Speech Recognition Applications with a Large Vocabulary. In: Sojka, Matousek, Pala, Kopecek (Eds.), In Proceedings of the First Workshop on Text, Speech and Dialogue, (pp. 299--304). Czech Republic.

Rojo, G. (1991). Frecuencia de fonemas en español actual. In BREA, M., Fernandez Rei, F. (Coord). Homenaxe ó profesor Constantino García. Servicio de Publicación e Intercambio Científico. (pp. 451--467). Universidad de Santiago de Compostela. Santiago de Compostela, España.

RAE (1999). Esbozo de una nueva gramática para el español. Real Academia Española. Editorial Espasa-Calpe. (pp. 9--63). España.

Uraga, L.E. (1999). Modelado fonético para reconocimiento de voz continua en Español. Tesis de Maestría. ITESM Campus Morelos . México.

Uraga, E. & Pineda, L. (2002). Automatic Generation of Pronunciation Lexicons for Spanish. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2002. Springer. México.

Vaufreydaz, D., Bergamini, C., Serignat, J.F., Besacier, L., Akbar, M. (2000). A New Methodology For Speech Corpora Definition From Internet Documents, In Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, ELRA, (pp. 423—426), Athens, Greece.

Villaseñor, L., Massé, A., Pineda, L. (2001). The DIME Corpus. 3er. Encuentro Internacional de Ciencias de la Computación ENC-01. Aguascalientes, México.

Wang, H-M. (1998). Statistical analysis of mandarin acoustic units and automatic extraction of phonetically rich sentences based upon a very large chinese text copus. Computational linguistics and chinese languague processing, vol. 3, no. 2, (pp. 93—114). Computational linguistics society of R.O.C.