Speaker Dependent ASRs for Huastec and Western-Huastec Náhuatl Languages

Juan A. Nolazco-Flores, Luis R. Salgado-Garza, and Marco Peña-Díaz

Departamento de Ciencias Computacionales, ITESM, Campus Monterrey
Av. Eugenio Garza Sada 2501 Sur, Col. Tecnológico
Monterrey, N.L., México, C.P. 64849
{jnolazco,lsalgado,motilio}@itesm.mx

Abstract. The purpose of this work is to show the results obtained when the latest technological advances in the area of Automatic Speech Recognition (ASR) are applied to the Western-Huastec Náhuatl and Huastec languages. Western-Huastec Náhuatl and Huastec are not only native (indigenous) languages in México, but also minority languages, and people who speak these languages usually are analphabetic. A speech database was created by recording the voice of native speaker when reading a set of documents used for native bilingual primary school in the official mexican state education system. A pronunciation dictionary was created for each language. A continuous Hidden Markov Models (HMM) were used for acoustical modeling, and bigrams were used for language Modeling. A Viterbi decoder was used for recognition. The word error rate of this task is below 8.621% for Western-Huastec Náhuatl language and 10.154% for Huastec language.

1 Introduction

Language Technologies, such as ASR and Text-to-Speech Synthesis, is mature in many languages, such as English and Spanish [10] [11] [13]. This allows many computer applications to be developed in these languages, for example educative software. In the same way, Language Technology,when applied to prehispanic, can also be used to develop applications for these languages.

In México there are around 295 native american languages [1]. The total number of persons who speaks this american indian languages is around 8% of the total population in México, this means around 7,000,000 [1]. These 295 languages are grouped in families[1], and some families are grouped in stocks[2] Western-Huastec Náhuatl is in the Uzo-aztec stock and the Huastec Language is inside the Maya family[3],

[1] A family is a group of languages that easily can be shown to be genetically related when the basic evidence is examined [14].) (In México, there are six families, Aztecan, Corachol, Cahita, Tarahumaran, Tepiman, Tubar [1]).

[2] A stock is a group of language families that are genetically related to each other but, because of the time depth involved, the evidence is more difficult to assemble. In México, there are three stocks (Uzo-aztec, otomangue, okano) [14].

[3] The maya family is a language independent of any stock [1]).

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 595–602, 2005.

The southern Uzo-Aztec stock, which comprises around 49 languages, is spoken for around 1,750,000 persons [1]. The Aztecan family which comprises 28 náhuatl languages is one of the most important ones in this stock with around 1,600,000 speakers [1]. Western-Huastec Náhuatl variant is the most popular with 410,000 persons speaking the language, spoken in 1500 communities [1] in the Huastec region of San Luis Potosi, México, where Tamasunchale city is the center of this region [1].

The mayan family, which comprises around 31 languages, is spoken for 3, 381, 300 persons [15], is the most diversified and populous language family of Meso-America. Huastec languages is one of these languages with 101,000 speakers [15]. The Huastec language is separated in time for 2,500 years and physically by more than 1,000 miles from the nearest other Mayan language [15]. The Huastec is spoken in the Huastec region of San Luis Potosi, being Aquismón and Tancahuits de Santos the cities with more speakers [16]. It is also spoken in Veracruz, being Tantoyuca the city with more speakers [1] [16].

Since most of the persons who speaks these languages are analphabet every year the percentage of persons, compared with the total population of the region, who speaks this language is diminishing. Actually, from the 295 native american languages spoken in México 188 are endangered languages [2]. Speaking the majority language better equips children for success in the majority culture than speaking a less prestigious language [2].

However, preserving the language is important because the Language is the most efficient means of transmitting a culture, and it is the owners of that culture that lose the most when a language dies. Every culture has adapted to unique circumstances, and the language expresses those circumstances. Moreover, identity is closely associated with language [2]. The history tied up in a language will go unrecorded; the poetry and rhythm of a singular tongue will be silenced forever. The scientific search for Universal Grammar, the common starting point for all grammars that human children seem to be born with, depends on our knowing what all human languages have in common. The wholesale loss of languages that we face today will greatly restrict how much we can learn about human cognition, language, and language acquisition at a time when the achievements in these arenas have been greater than ever before [2].

In this work, we propose to develop an ASR systems for the Huastec and Western-Huastec Náhuatl language using continuous HMM and bigrams. We use this technology because continuous HMM is the most successful acoustic modeling technique and bigrams is also a very successful language modeling technique. Moreover, we believe that native people will be more interested in their own language, when, thanks to this and other studies, they knew that other people is interested in his their languages.

This paper is organized as follows. In section 2, the náhuatl language features are explained. In section 3, the huastec language features are explained. In section 4, the database features are explained. In section 5, the system architecture description is defined. In section 6, the experiments and results are given. Finally in section 7, the comments and conclusions are given.

2 Náhuatl Language

The Náhuatl is well know because it was the language of the Aztecs Empire of central México when spanish arrived. However, is less known that there are 28 types of Náhautl, some of them with less than 1000 speakers [1]. This work is concern with the Western-Huastec Náhuatl, which is the one with more speakers of the Uzo-Aztec languages.

Originally the Náhuatl language writings were a mixture of pictures of three classes: pictogram, ideograms and phonograms [5][4]. When Spanish arrived to mexican culture, one of their first task was to adapt the náhuatl language to the spanish alphabet. Therefore, now the bilingual education in Mexico is with alphabet writing.

Náhuatl language is highly agglomerative, which means that words are formed by a root and a high number of prefixes and suffixes [5]. Therefore, the words in this kind of languages includes a lot of information and potentially each word can be very long and the number of words in the language is very high. As an example, the following written in English [6]:

> “This book is for indigenous boys and girls who are studying basic school with the aim to help them how to read and write the indigenous language spoken in its community”

will look as follows in Western-Huastec Náhuatl language:

> “Ni amochtli tijtlaliaj inmako ockichpilmej uan siuapilmej ankij momachtiaj Tlen eyi uan tlen naui xiuitl tlen se ixelka tlamachtilistli, pampa moneki kiyekosej tlaixpouasej uan tlajkuilosej ika inineltlajtol tlen ika tlajtouaj ipan inchinanko”

Table 1 shows the Western-Huastec Náhuatl language’s phonemes used in this work [5]. There are some important pronunciation rules. First, the j not pronounced when it is at the end of a sentences, and some speaker ignore it even it is inside of the sentence. The rest of the letter are pronounced as they are written, with the following exceptions: when letter C is before letters E and I, then it is pronounced as the phoneme /S/.When the letters H followed by U is located before A, E and I, then it is pronounced as /W/. Given the text the pronunciation rules and the list of phonemes, and the labels in the speech database and the phonemes we create a pronunciation dictionary.

3 Huastec Language

This language is also known as Tenek language. Originally the huastec language writings were a mixture of pictures of three classes: pictogram, ideograms and

[4] In a pictogram an object is represented with one picture, in a ideogram something or an idea is represented with a picture, phonogramas a ?syllable? or phone represented with a picture.

Table 1. Phonemes used in this work for Western-Huastec-Náhuatl and Huastec languages.

Manner	WH Náhuatl	Huastec	Example
Vowels	a	a	h**o**od
	e	e	h**ea**d
	i	i	h**ee**d
	o	o	h**oe**d
	u	u	h**oo**d
Plosives	b		**b**oot
	p	p	**p**ea
	t	t	**t**ea
	k	k	**k**ick
Fricatives	s	s	**s**o
	S	S	**s**how
Nasals	m	m	**m**om
	n	n	**n**oon
Semivowels Glides	w	w	**w**ant
	y	y	**y**ard
Semivowels Liquids	l	l	l
Affricatives	C	C	**ch**urch (written with letter x.)
Others	tl		t and l pronounces as one sound
	tz		t and z pronounces as one sound
		dh	d and h pronounces as one sound

phonograms [15]. It is believed that the first book written in a language different to Náhuatl was in Huastec, and it was called ”Doctrina Cristina en Lengua Guasteca” [16].

Huastec language is highly agglomerative, which means that words are formed by a root and a high number of prefixes and suffixes [16]. Therefore, the words in this kind of languages includes a lot of information and potentially they can be very large, and the number of words in the language is very high. As an example, the following written in English [6]:

> “This book is for indigenous boys and girls who are studying basic school with the aim to help them how to read and write the indigenous language spoken in its community”

will look as follows in huastec language:

> “Axé xi dhuchadh úw, jats abal ka pidhanchik an ts’ik’ách ani an kwitól axi k’wátchik ti exóbal ti al an k’a’aál pejach tin k’a’ál exobintal; axé, jats abal kin ne’ets with’a’chik ti dhuchum ani ti ajum tin tének káwintal.”

Table 1 shows the huastec language’s phonemes used in this work [5]. There are some important pronunciation rules. First, the *j* is not pronounced when it is at the end of a sentences, and some speaker ignore it even it is inside of the sentence. The rest of the letter are pronounces as they are written, with the following exception: when letter *dh* is read is pronounced as it where one sound.

4 Databases

In order to facilitate our labeling process, for our databases recording we selected some text books used for native language bilingual education in México [6]. Moreover, this was also very convenient to facilitate the labeling process.

In our speech database construction we ask to person to read lessons from the selected textbooks [6]. The number of different words in Western-Huastec Náhuatl are 759,this database contains around 1 hour of recorded data from two speakers, a man and a woman. The number of different words in Huastec are 319,this database contains around 1 hour of recorded data from two speakers, a man and a woman. All the participants with ages between 20 and 25 years old. In both cases, the speech waveform was sampled at 16,000 KHz.

5 System Architecture Description

The CMU SPHINX-III systems is a HMM-based speech recognition system capable of handling large vocabulary. The architecture of this system is shown in Figure 1. As can be observed in this figure the analog signal is sampled, and converted to MFCC coefficients, then the MFCC's first and second derivatives are concatenated [8], i.e. if the number of MFCC is 13 then the total dimension of the feature vector would be 39.

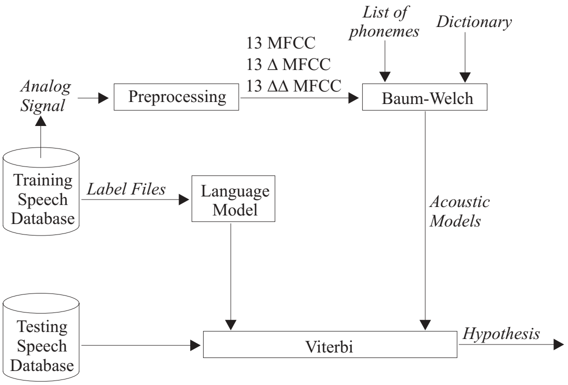

Fig. 1. CMU SPHINX-III ASR architecture.

The acoustic models is also obtained using the SPHINX-III tools. This tools use a Baum-Welch algorithm to train this acoustic models [12]. The Baum-Welch algorithm needs the name of the word units to train as well at the label and feature vectors. The SPHINX-III system allows us to modelate either discrete, semi continuous or continuous acoustic models. In SPHINX-III, system tools allow to select as acoustic model either a phone set, a triphones set or a word set.

The language models are obtained using the CMU-Cambridge statistical language model toolkit version 2.0 [9]. The LM aim is to reduce the perplexity of the task, by predicting the following word based in the words' history. N-grams is the easiest technique with very good results. If all the n-grams are not contained in the language corpus, smoothing techniques need to be applied. In the CMU-Cambridge language model toolkit, unigram, bigrams or trigrams can be configured for this tool, as well as four types of discount model: Good Turing, Absolute, Linear and Witten-Bell.

Using an acoustical model and a language model a Viterbi decoder obtains the best hypothesised text.

6 Experiments

The configuration of the SPHINX-III system is described. Thirteen mel-frequency cepstral coefficients (mfcc) were used. First and Second derivatives were calculated, therefore the feature vector was 39 elements. The speech lower frequency was 300 Hz and the speech higher frequency was 7,000 Hz. The frame rate was set to 50 frames per second. A 30ms Hamming window was used. A 512 samples FFT length was used. The number of filterbanks was set to 40. Five states continuous HMM were used as acoustic modeling technique and bigrams was used as a language modeling technique. Simple phones were used as the word unit. Since our corpus is a small corpus and the number of words is very large, we develop experiments using different smoothing techniques. Table 2 shows the experimental results for Western-Huastec Náhuatl language and Table 3 shows the experimental results for Huastec language. As expected the Witten-Bell discount strategy was the one with better results.

Table 2. WER results for Western-Huastec Náhuatl language over several Gaussian distributions and language model configurations.

Number of	Language Model discounting strategy			
Gaussians	**Good-Turing**	**Linear**	**Absolute**	**Witten-Bell**
4	6.80%	8.43%	6.80%	4.50%
8	6.71%	8.62%	6.80%	4.31%
16	6.80%	8.53%	6.80%	4.22%
32	6.90%	8.53%	6.80%	4.12%
64	6.90%	8.53%	6.80%	4.02%

Table 3. WER results for Huastec Language and over several Gaussian distributions and language model configurations.

Number of	Language Model discounting strategy			
ans	**Good-Turing**	**Linear**	**Absolute**	**Witten-Bell**
4	9.13%	9.83%	8.23%	6.94%
8	9.13%	10.03%	8.48%	6.56%
16	8.74%	9.90%	7.97%	6.81%
32	8.87%	10.15%	8.36%	6.68%
64	8.87%	10.15%	8.36%	6.94%

7 Conclusions

In this work, we present the development of a prehispanic database for Western-Huastec Náhuatl and Huastec languages. We also show the results obtained when Automatic Speech Recognition technology is applied to these languages. Since people that speak prehispanic language do not usually read, then the main problem to develop speech models for prehispanic languages is the difficult to find people that read its own language.

We think that Speech Technology can be a catalizer in the effort to preserve the minority languages. Therefore, as a future work, in first place the database will be extended to include a larger number of speakers, the recording time will also be extended. Other languages technologies, such as Text-to-Speech technology is also planned to be applied. The goal is to better understand the language to develop educative software in these languages.

We also have to refine the phoneme list and the pronunciation rules. We are also planning to create databases for other minority languages, such as Mixteco and Cora.

Acknowledgements

The authors would like to acknowledge the head of the Secretaría de Educación del Estado de Nuevo León, Maestra María Yolanda Blanco García; to the head of Departamento de Educación Indígena, Lic. Luz Divina Tijerina Luna. The authors also would like to acknowledge CONACyT, project CONACyT-2002-C01-41372, who partially supported this work.

References

1. http://www.ethnologue.com.
2. http://yourdictionary.com/elr/living.html.
3. Constitución Política de los Estados Unidos Mexicanos.
4. Plan y Programa de Estudio para la Educación Primaria, SEP, México, 1993.
5. Sullivan, T.O., Compendio de la Gramática Náhuatl, Ejercicios, UNAM, Instituto de Investigaciones Históricas, Second Edition, 1992.
6. Canales Juarez, G., Mendez González, R., Hernández Miranda, J., Roque Cerroblanco, E., "*Nauatlajtoli tlen uaxtekapaj tlali, Lengua náhuatl, Region Huasteca, Hidalgo*", Third and fourth grade, SEP, 1993.
7. http://www.ldc.upenn.edu.
8. Deller, J.R., Proakis, J.G., Hansen, J.H.L., Discrete-Time Processing of Speech Signals, Prentice Hall, Sec. 6.2, 1993.
9. Clarkson, P., Rosenfeld, R., "Statistical Language Modelling using the CMU-Cambridge Toolkit", Proceedings of Eurospeech, Rodhes, Greece, 1997, 2707-2710.
10. Varela, A., Cuayáhuitl, H., Nolazco-Flores, J.A., "Creating a Mexican Spanish Version of the CMU SPHINX-III Speech Recognition System", CIARP, Springer Verlag, LNCS 2905:251-58.

11. Salgado-Garza, L.R., Stern, R., Nolazco, J.A.,"N-Best List Rescoring using Syntactic Trigrams", MICAI 2004: Advances in Artiticial Intelligence LNAI 2972, Springer Verlag, 2004, LNAI 2972:79-88.
12. Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum likelehood for incomplete data via the EM algorithm", J. Roy. Stat. Soc., Vol. 39, No. 1, 1-38, 1977.
13. Huerta, J.M., Chen, S., Stern, R.M.: "The 1998 CMU SPHINX-3 Broadcast News Transcription System", Darpa Broadcast News Workshop, 1999.
14. Dryer, M. S., Large Linguistic areas and lang. samp. *Studies in Language*, 13:257-92, 1996.
15. "Meso-American Indian Languages". Encyclopedia Britannica. 2004. Encyclopedia Britannica Online. 14 May 2004
    http://0-search.eb.com.millenium.itesm.mx:80/eb/article?eu=118158.
16. Grossner-Lerner, E., Los tenek de San Luis Potosi, INAH, 1991.