Creating a Mexican Spanish Version of the CMU Sphinx-III Speech Recognition System

Armando Varela[1], Heriberto Cuayáhuitl[1], and Juan Arturo Nolazco-Flores[2]

[1] Universidad Autónoma de Tlaxcala,
Department of Engineering and Technology,
Intelligent Systems Research Group,
Apartado Postal #140, 90300 Apizaco, Tlaxcala, Mexico.
{avarela, hcuayahu}@ingenieria.uatx.mx
http://orion.ingenieria.uatx.mx:8080/si/si.jsp
[2] Instituto Tecnológico de Estudios Superiores de Monterrey,
Sucursal de Correos "J", 64849, Monterrey, Nuevo Leon, Mexico.
jnolazco@itesm.mx

Abstract. In this paper we present the creation of a Mexican Spanish version of the CMU Sphinx-III speech recognition system. We trained acoustic and N-gram language models with a phonetic set of 23 phonemes. Our speech data for training and testing was collected from an auto-attendant system under telephone environments. We present experiments with different language models. Our best result scored an overall error rate of 6.32%. Using this version is now possible to develop speech applications for Spanish speaking communities. This version of the CMU Sphinx system is freely available for non-commercial use under request.

1 Introduction

Today, building a new robust Automatic Speech Recognition (ASR) system is a task of many years of effort. In the Autonomous University of Tlaxcala - Mexico, we have two goals in the ASR field: Do research for generating a robust speech recognizer, and build speech applications for automating services. In order to achieve our goals in a short time, we had to take a baseline work. We found that the CMU (Carnegie Mellon University) Sphinx speech recognition system is freely available and currently is one of the most robust speech recognizers in English. The CMU Sphinx system enables research groups with modest budgets to quickly begin conducting research and developing applications. This arrangement is particularly pertinent in Latin America, where the financial support and experience otherwise necessary to support such research is not readily available. In the past, few research efforts have been done for Spanish and these includes work from CMU in broadcast news transcription [1, 2], where basically acoustic and language models have been trained. Our motivations for developing this work are due to the fact that many applications require a speech recognizer for Spanish, and because Spoken Dialogue Systems (SDS) require a robust speech recognizer were reconfiguration and retraining is necessary.

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 251–258, 2003.

In this research, we have generated a lexicon and trained acoustic and language models with Mexican Spanish speech data for the CMU Sphinx speech recognition system. Our experiments are based on data collected from an auto-attendant application (CONMAT) deployed in Mexico [3], with a vocabulary of 2,288 entries from names of people and places inside a university, including synonyms. Our speech data used for training and testing was filtered avoiding noisy utterances. Results are given in terms of the well known evaluation metric: Word Error Rate (WER). In the remainder of the paper we first provide an overview of the system in section 2. In section 3 we describe the components of the Sphinx system and how these were trained. In Section 4 we present experimental results. Finally, in section 5 we provide our conclusions and future directions.

2 System Overview

The Carnegie Mellon University Sphinx-III system is a frame-based, HMM-based, speaker-independent, continuous speech recognition system, capable of handling large vocabularies (see Fig. 1). The word modeling is performed based on subword units, in terms of which all the words in the dictionary are transcribed. Each subword unit considered in its immediate context (triphone) is modeled by 5-state left-to-right HMM model. Data is shared across states of different triphones. These groups of HMM states sharing distributions between its member states are called senones [4].

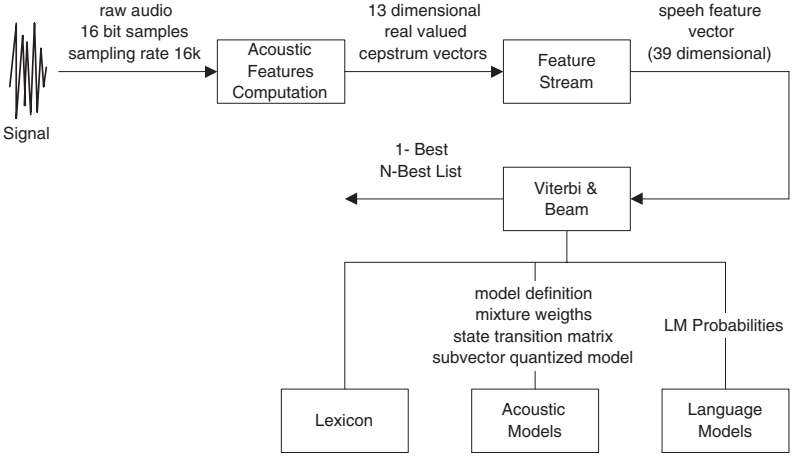

Fig. 1. Architecture of the CMU Sphinx-III speech recognition system. The lexical or pronunciation model contains pronunciations for all the words of interest to the decoder. Acoustic models are based on statistical Hidden Markov models (HMMs). Sphinx-III uses a conventional backoff bigram or trigram language model. The result is a recognition hypothesis with a word lattice representing an N-best list.

The feature vector computation is a two-stage process. In the first stage, an off-line front-end module is first responsible for processing the raw audio sample stream into a cepstral stream. The input is windowed, resulting in frames of duration 25.625 ms. The output is a stream of 13-dimensional real-valued cepstrum vectors. The frames overlap, thus resulting in a rate of 100 vectors/sec. In the second stage, the stream of cepstrum vectors is converted into a feature stream. This process consists of a Cepstrum Mean-Normalization (CMN) and Automatic Gain Control (AGC) step. The final speech feature vector is created by typically augmenting the cepstrum vector (after CMN and AGC) with one or more time derivatives. The feature vector in each frame is computed by concatenating first and second derivatives to the cepstrum vector, giving a 39-dimensional vector.

3 System Components

3.1 Lexicon

The lexicon development process consisted of defining a phonetic set and generating the word pronunciations for training acoustic and language models.

Table 1. ASCII Phonetic Symbols for Mexican Spanish.

Manner	Label	Example	Worldbet Word
Plosives	p	**p**unto	p u n t o
	b	**b**años	b a ñ o s
	t	**t**ino	t i n o
	d	**d**on**d**e	d o n d e
	k	**c**asa	k a s a
	g	**g**an**g**a	g a n g a
Fricatives	f	**f**alda	f a l d a
	s	mi**s**mo	m i s m o
	x	**j**amas	x a m a s
Affricates	tS	**ch**ato	tS a t o
Nasals	m	**m**ano	m a n o
	n	**n**ada	n a d a
	ñ	**b**año	b a ñ o
Semivowels	l	**l**ado	l a d o
	L	po**ll**o	p o L o
	r(	pe**r**o	p e r(o
	r	pe**rr**o	p e r o
	w	**hu**eso	w e s o
Vowels	i	p**i**so	p i s o
	e	m**e**sa	m e s a
	a	c**a**so	k a s o
	o	m**o**do	m o d o
	u	c**u**ra	k u r(a

Our approach for modeling Mexican Spanish phonetic sounds in the CMU Sphinx-III speech recognition system consisted of an adapted version from the WORLDBET Castilian Spanish phonetic set [5], which resulted in 23 phonemes listed in Table 1. The adaptation consisted in a manual comparison of spectrograms from words including a common phoneme; we found common sounds which we merged in our final list of phonemes. The following are the modifications made to the Castilian Spanish sounds set for generating a Mexican Spanish version:

- Fricative /s/ as in "kasa" and fricative /z/ as in "mizmo" merged into /s/,
- Plosive /b/ as in "baños" and fricative /V/ as in "aVa" merged into /b/,
- Plosive /d/ as in "donde" and fricative /D/ as in "deDo" merged into /d/,
- Plosive /g/ as in "ganga" and fricative /G/ as in "lago" merged into /g/,
- Semi-vowels /j/ as in "majo" and /L/ as in "poLo", and affricate /dZ/ as in "dZugo" merged into /L/,
- Nasal /n/ as in "nada" and nasal /N/ as in "baNko" merged into /n/,
- Fricative /T/ as in "luTes" was deleted due to the fact that this sound does not exist in Mexican Spanish.

The vocabulary size has 2,288 words, which is based on names of people and places inside a university, including synonyms. The automatic generation of pronunciations was performed using a simple list of rules and exceptions. The rules determine the mapping of clusters of letters into phonemes and the exceptions list covers some words with irregular pronunciations. A Finite State Machine (FSM) was used to develop the pronunciations from the word list.

3.2 Acoustic Models

For training acoustic models is necessary a set of feature files computed from the audio training data, one each for every recording in the training corpus. Each recording is transformed into a sequence of feature vectors consisting of the Mel-Frequency Cepstral Coefficients (MFCCs). The training of acoustic models is based on utterances without noise. This training was performed using 3,375 utterances of speech data from an auto-attendant system, which context is names of people and places inside a university.

The training process (see Fig. 2) consists of the following steps: Obtain a corpus of training data and for each utterance, convert the audio data to a stream of feature vectors, convert the text into a sequence of linear triphone HMMs using the pronunciation lexicon, and find the best state sequence or state alignment through the sentence HMM for the corresponding feature vector sequence. For each senone, gather all the frames in the training corpus that mapped to that senone in the above step and build a suitable statistical model for the corresponding collection of feature vectors. The circularity in this training process is resolved using the iterative Baum-Welch or forward-backward training algorithm. Due to the fact that continuous density acoustic models are computationally expensive, a model is built by sub-vector quantizing the acoustic model densities (sub-vector quantizing was turned off in our work).

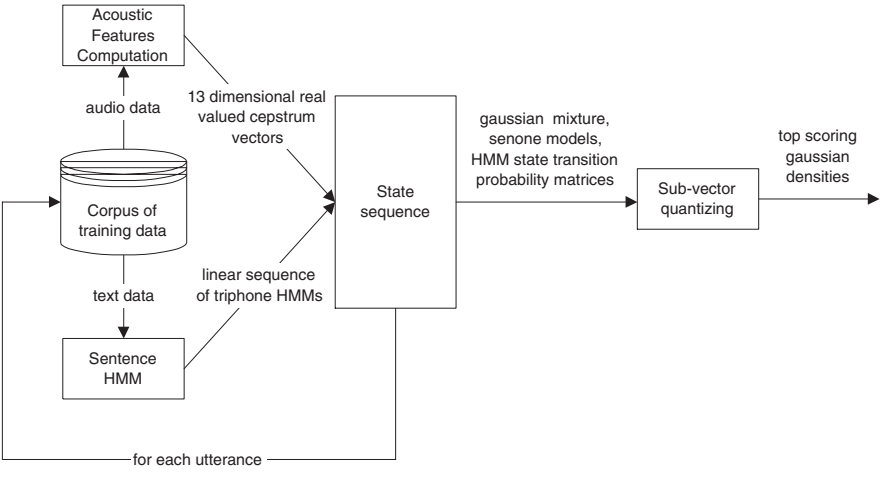

Fig. 2. A block schematic diagram for training acoustic models.

3.3 Language Models

The main Language Model (LM) used by the Sphinx decoder is a conventional bigram or trigram backoff language model. Our LMs were constructed from the 2,288 word dictionary using the CMU-Cambridge statistical language model toolkit version 2.0 [6], see Fig. 3. The training data consisted of 3,375 transcribed utterances of speech data from an auto-attendant system. We trained bigrams and trigrams with four discounting strategies: Good Turing, Absolute, Linear, and Witten Bell. The LM probability of an entire sentence is the product of the individual word probabilities. The output from the CMU-Cambridge toolkit is an ASCII text file, and because this file can be very slow to load into memory, the LM must be compiled into a binary form. The decoder uses a disk-based LM strategy to read the binary into memory. Although the CMU-sphinx recognizer is capable for handling out-of-vocabulary speech, we did not set any filler models. Finally, the recognizer needs to exponenciate the LM probability using a language weight before combining the result with the acoustical likelihood.

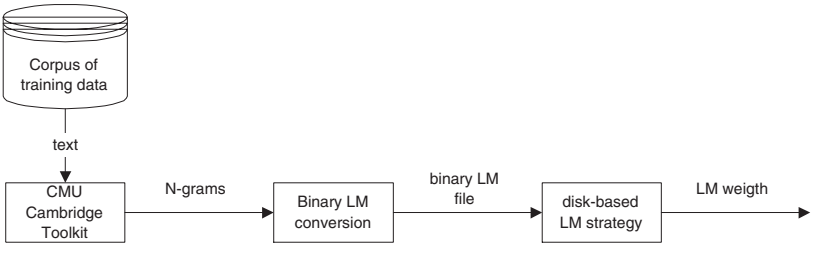

Fig. 3. A block schematic diagram for training language models.

4 Experimental Results

4.1 Experimental Setup

We performed two experiments for evaluating the performance of the CMU Sphinx system trained with Mexican speech data (872 utterances) in the context of an auto-attendant application: the first experiment considered names of people and places as independent words (i.e. any combination of first names and last names was allowed), the second experiment considered names of people and places as only one word. Each experiment was evaluated with two different LMs.

4.2 Evaluation Criteria

The evaluation of each experiment was made according to recognition accuracy and computed using the WER (Word Error Rate) metric defined by the equation 1, which align a recognized word string against the correct word string and compute the number of substitutions (S), deletions (D), and insertions (I) from the number of words in the correct sentence (N).

$$WER = (S + D + I) / N * 100\%. \quad (1)$$

4.3 Results

Recognition results for each decoding stage for the CMU with Sphinx Mexican Spanish test data are shown in Tables 2 and 3. In table 2 (experiment 1), we can observe that the use of Good Turing discount strategy is not convenient, and the use of different n-grams does not make much difference, perhaps bigger training and test sets would yield significant differences. In the mean time, for this experiment the best option is bigrams with Witten Bell discounting strategy, but we observed problems with this approach due that this experiment can yield incorrect hypothesis, i.e. inexistent names of people and places. Thus, another solution was necessary to solve this problem. In table 3 (experiment 2), we observe that due to the conditions of the experiment, would yield no further significant improvements with different n-grams. Despite of this, the best gains are shown in trigrams with Witten Bell discounting strategy.

Table 2. Word error rate in the test set after decoding from the experiment 1, which considered names of people and places as independent words.

Discounting Strategy	Bigrams	Trigrams
Good Turing	12.95	12.88
Absolute	7.82	7.63
Linear	7.94	8.07
Witten Bell	7.63	7.75

Table 3. Word error rate in the test set after decoding from the experiment 2, which considered names of people and places as only one word.

Discounting Strategy	Bigrams	Trigrams
Good Turing	6.88	6.44
Absolute	6.38	6.38
Linear	6.50	6.57
Witten Bell	6.38	6.32

5 Conclusions and Future Work

We described the training and evaluation processes of the CMU Sphinx-III speech recognition system for Mexican Spanish. We performed two experiments in which we grouped differently the word dictionary entries. Our best results of this development considered dictionary entries as only one word for avoiding inexistent names of people and places inside a university. Through a simple lexicon and set of acoustic and language models, we demonstrated an accurate recognizer which scored an overall error rate of 6.32% on in-vocabulary speech data. We achieved the goal of this work from which now we have a baseline product for performing research in speech recognition, which is an important component of spoken language systems. Also, with this work we can start development of speech applications with the advantage that we can retrain and adapt the recognizer according to our needs. This work was motivated due to the fact that people around the world needs to develop applications involving speech recognition for Spanish speaking communities. Therefore, the resulted lexicon, acoustic and language models are freely available for non-commercial purposes under request.

An immediate future work is to provide a bridge for invoking the recognizer and see it as a black box, perhaps we can build a dll file or we can provide something similar as SAPI. This is indispensable for programmers who need to develop speech applications from different programming environments. Another important future direction and due that this development considers only in-vocabulary speech, we plan to retrain the recognizer considering Out-Of-Vocabulary (OOV) speech, measuring computational overhead. This is due to the fact that OOV speech is an important factor in spoken dialogue systems and degrades significantly the performance in such systems [7]. Also, we plan to train Sphinx in different domains, as well as optimize configuration parameters. Finally, we plan to train Sphinx release 4 which was implemented in Java, and make a comparison between Sphinx III and Sphinx 4 in Spanish domains. All this work would be performed considering a bigger corpus.

Acknowledgements. This research was possible due to the availability of the CMU Sphinx speech recognizer. We want to thank to the people involved in the

development of the CMU Sphinx-III and of course the formers of the recognizer [8]. Also, we want to thank Ben Serridge for his writing revision on this paper.

References

1. J. M. Huerta, E. Thayer, M. Ravishankar, and R. M. Stern: The Development of the 1997 CMU Spanish Broadcast News Transcription System. Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, Virginia, Feb 1998.
2. J. M. Huerta, S. J. Chen, and R. M. Stern: The 1998 Carnegie Mellon University Sphinx-III Spanish Broadcast News Transcription System. In the proceedigns of the DARPA Broadcast News Transcription and Understanding Workshop, Herndon, Virginia, Mar 1999.
3. Cuayáhuitl, H. and Serridge, B.: Out-Of-Vocabulary Word Modeling and Rejection for Spanish Keyword Spotting Systems. Lecture Notes in Computer Science, Vol, 2313. Berlin Heidelberg New York (2002) 158–167.
4. Hwang, M-Y: Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition. Ph.D. thesis, Carnegie Mellon University, 1993.
5. Hieronymus L., J.: ASCII Phonetic Symbols for World's Languages: worldbet. Technical report, Bell Labs, 1993.
6. P. Clarkson, and R. Rosenfeld.: Statistical Language Modeling Using the CMU-Cambridge Toolkit. In the proceedings of Eurospeech, Rodhes, Greece, 1997, 2707–2710.
7. Farfán, F., Cuayáhuitl H., and Portilla, A.: Evaluating Dialogue Strategies in a Spoken Dialogue System for Email. In the proceedings of the IASTED Artificial Intelligence and Applications, ACTA Press, Manalmádena, Spain, Sep 2003.
8. CMU Robust Speech Group, Carnegie Mellon University. http://www.cs.cmu.edu/afs/cs/user/robust/www/