The 1998 Carnegie Mellon University SPHINX-3 Spanish Broadcast News Transcription System

Juan M. Huerta, Stanley Chen, and Richard M. Stern

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

This paper describes the 1998 CMU Hub 4 Spanish broadcast news transcription system. We focus on the development and improvements of the system with respect to the 1997 system. Both the 1997 and 1998 systems were developed using exactly the same acoustic and language model training material, thus the improvements obtained resulted from a better utilization and modeling of these corpora and a better decoding configuration and strategy. Specifically, we employed several language models, a larger lexicon and larger acoustic models than our 1997 system. Due to these improvements, we achieved a reduction of 28% in the error rate on this year's development material.

1. INTRODUCTION

For the 1997 DARPA evaluation on Spanish news broadcasts the CMU speech group developed a system based on two decoding passes that utilized a single set of acoustic models and a single set of language models [3]. Compared to the corresponding English-language recognition system [8], the 1997 Spanish-language system was relatively simple. It was defined and developed over a short period of time without a prior system (and no experience in our group in recognition in languages other than English). The system developed for the 1998 Spanish-language evaluation used last year's system as a point of departure. Since no additional training material was made available, it was necessary to refine the systems's organization and its statistical models. In spite of this lack of additional training data we were able to achieve a considerable reduction of the word error rate (WER) for the 1998 development test set (which was also the 1997 evaluation data). This improved WER was obtained by expanding the size of the lexicon and the acoustic model, by increasing the size and number of language models, along with some refinements to the decoding strategy.

In this paper we describe and discuss the improvements to the Spanish language system over the past year, and their motivation. In Section 2 we provide an overview of the decoding process. In Section 3 we describe the various components of system, and we describe how they were trained and developed. In Section 4 we present results and conclusions, along with lessons that have been learned over the year.

2. SYSTEM OVERVIEW

Our 1998 system recognizes speech from Spanish-language news broadcasts using a single recognition lexicon, a single set of acoustic models, and three different language models. Three decoding passes are implemented, with acoustic adaptation performed on the models between passes.

2.1. Parametrization, segmentation and clustering

The acoustic data are converted to 13-dimensional cepstral vectors and segment boundaries are established using CMUseg [9]. A second segmentation pass using tighter parameters is performed on segments longer than 50 seconds. The resulting segments are combined with the segments of the first segmentation pass. The segments are then clustered. No classification with respect to gender or spectral bandwidth is performed on these clusters as a single set of acoustic models is used for all the decoding stages.

2.2. First decoding stage

The first decoding pass is performed using the single set of acoustic models and a trigram language model, producing word lattices. These lattices are then rescored and a single best path is then selected using the same trigram language model with a higher language weight [7].

2.3. Second and third decoding stages

The hypotheses obtained from the first pass are used to adapt the means of the acoustic models for each cluster using a multiple class maximum-likelihood linear regression (MLLR) approach [6], based on six regression classes. These acoustically-compensated new means are then used for a second decoding pass with the trigram-based language model. Lattices are generated during the decoding process and later rescored using a four-gram language model and the best 200 hypotheses for each segment are retained.

The 200 best hypotheses are then rescored using an interpolated four-gram language model and a single best hypothesis for each segment is obtained. The single best hypothesis for each segment was used in a second MLLR compensation pass. A third decoding pass similar to the second decoding pass was performed. The language model weights used in the second and third decoding stages were optimized empirically based on the development data.

3. SYSTEM COMPONENTS

3.1. Lexical Component

The recognition lexicon consists of approximately 60,000 words obtained from the 45,000-word LDC Spanish lexicon plus an additional 15,000 words obtained from the LDC Spanish Newspaper corpus and the BN transcriptions not included in the LDC lexicon. Pronunciations were generated automatically with pronunciations for a few foreign words corrected by hand.

A significant effort was expended in cleaning the text corpus and in continuing the text-conditioning work initiated last year. We focused on substituting words in the text corpora whose spelling could be corrected deterministically . We observed no noticeable difference when the changes in the text were introduced and new language models trained.

We also considered including syllabic, diphthong, and stress information into the pronunciations. The new pronunciations would be based on a much larger number of phonetic units (close to 80) as opposed to the 24 we currently use. This expansion of the number of phonetic units produced a corresponding increase in the number of triphones, along with an explosion in the number of context questions used to develop the clustering trees. Eventually we decided that the 30 hours of available acoustic training material would be insufficient to train the new proposed models. We employed pronunciations similar to those used in 1997: no diphthong, stress, or syllabic information was included.

3.2. Acoustic Model Component

Acoustic training was performed using the SPHINX-3 system. The evaluation system used a fully-continuous, diagonal-covariance mixture Gaussian configuration.

The acoustic training was performed using the 30 hours of broadcast news training data distributed for the 1997 DARPA Hub-4 Spanish Evaluation. We constructed a single set of acoustic models consisting of 2800 senonically-tied states [4] and 24 Gaussians per mixture. Our models were based on a set of 25 phonemes plus silence.

The models used for the actual evaluation were trained starting from a flat distribution using one Gaussian per state. Mixture-splitting was performed until 24 Gaussian densities per state were obtained. Table 1 below shows the word error rates for the development data after a 1-pass Viterbi decoding using last year's lexicon and language models, when using different number of senones (columns) and Gaussians per mixture (rows).We cam see that WER can be minimized by using more senones or gaussians. Using only 2800 tied states and 24 Gaussians per mixture, seemed the way to achieve this minimization of the WER with the smallest number of acoustic model parameters, using the available training data.

		Number of Tied States		
		2800	**3000**	**4000**
Gaussians per Mixture	**16**	23.9	22.2	22.3
	24	21.7	21.9	22.0
	28	21.7	–	–

Table 1. Word error rate for different combinations of number of gaussians per mixture and number of tied-states.

3.3. Language Model Component

Our language models were trained using a combination of the Hub-4 Spanish transcriptions and the Spanish newspaper corpus, using the recognition lexicon as the vocabulary. The broadcast news text was weighted four times as much as the newspaper corpus in order to increase its presence when mixed with the newspaper corpus. As mentioned in Section 3, an unsuccessful effort was made to clean the text further.

Eventually, three different language models were generated and used at different stages of the decoding. Having a set of language models that are considerably larger and more complex than last year models allowed us to extract more accurate hypotheses from the word lattices and N-best lists. We now describe these language models in somewhat more detail.

3.3.1 LM 1: trigram language model

The first language model, referred to as LM1, is a Witten-Bell discounted trigram language model constructed using the CMU-Cambridge language modeling toolkit [2]. This model is similar to the language model used in last year's system, with the only difference being that we employed lower cutoffs this year. We used 0-2-2 as cutoffs, excluding all bigrams and trigrams with two or fewer counts. LM1 was used to generate lattices in each pass.

3.3.2 LM 2: four-gram language model

The second language model, LM2, is a 4-gram language model with 0-2-2-3 cutoffs, smoothed using modified Kneser-Ney smoothing [1]. This language model, which is considerably larger than LM1, was used to generate N-best lists from the word lattices generated by the LM1.

3.3.3 LM 3: interpolated four-gram language class model

We partitioned our vocabulary into 1000 classes using an automatic clustering technique [5]. We then constructed a 4-gram class language model with 0-0-1-2 cutoffs using modified Kneser-Ney smoothing. This model is used to predict the class of the current word given the classes of the previous three words, and a unigram model is used to predict the current word given the current class. This class-based model was linearly interpolated with LM 2, the 4-gram word model, to form LM 3. The class-based and word-based models were weighted equally.

Table 2 shows some development results observed when building the Language Models. LMs A and B are 4-gram word LMs (and LM A became LM 2). We note that while reducing the cutoffs for LM B provides further reduction in WER, we could not use LM B for interpolation, because it is evaluated directly from the word counts. Best results were obtained by interpolating the class-based LM C with the word-based LM A.

LM	Type	Cutoffs	WER
A	4-gram word	0-2-2-3	17.06%
B	4-gram word	no cutoffs	16.95%
C	4-gram class	0-0-1-2	16.8%
D	4-gram class	0-0-2-4	16.9%
E	Interp. A & C	N.A.	16.67%
F	Interp. A & D	N.A.	16.69%

Table 2. Effect of applying different language models on N-best rescoring.

3.3.4 Class-based LMs: grammatical versus data-driven categories

Using N-best rescoring, we compared the performance of the interpolated language model described above with a different interpolated language model that was derived from part-of-speech (POS) information obtained from the LDC lexicon. The rationale for using a POS-based class language model was that it would reflect the probabilistic occurrences of POS bigrams and trigrams, thus providing a certain degree of syntactic constraint that could be used to evaluate person/gender agreement in a probabilistic framework. We observed that the data-driven class language model described in Section 3.3.3 produced better results. We subsequently increased the size of the N-best list in order to try to increase the dynamic range of the rescoring process, but even then the data-driven model outperformed the POS-based model (with a WER 17.7% WER vs. 19.1% using a given set of development lattices).

4. EXPERIMENTAL RESULTS

Table 3 summarizes the results obtained using our system on the 1998 development data set (which was also the 1997 evaluation set). We show the word error rates for each of the three decoding passes at every stage of each pass. As can be seen, the imposition of the various language models, along with the several adaptation and rescoring stages, provide a reduction in WER from 20.1% (Viterbi first pass) to 16.9% (N-best rescoring third pass) after using the various language models and performing the adaptation and rescoring stages.

	Pass 1	Pass 2	Pass 3
Viterbi Search	20.1	18.4	18.2
Lattice rescoring using 3-grams (LM1)	18.8	18.3	17.8
Lattice rescoring using 4-grams (LM2)	–	17.5	17.4
N-best rescoring using interpolated LM (LM3)	–	–	16.9

Table 3. 1998 development test set results at different stages of the decoding process.

Table 4 compares results obtained using the 1997 and 1998 Spanish broadcast news systems on the 1997 and 1998 evaluation sets. Continued system development resulted in a reduction in WER of more than 28% for the 1997 evaluation data (which was used as the 1998 devel-

opment set). This improvement was achieved through the use of more complex acoustic and language models, rather than further training material.

Data	System	WER
1997 Eval	1997	23.5%
1997 Eval	1998	16.9%
1998 Eval	1997 (1 pass)	29.8%
1998 Eval	1998 (1 Pass)	24.3%
1998 Eval	1998	22.4%

Table 4. Comparison of WER obtained by the 1997 and 1998 CMU Spanish broadcast news systems.

The 1998 system makes better use of the training data through larger and more sophisticated language models and larger acoustic models. Also better decoding configuration (using more passes) were useful. Roughly, one third of the observed improvement was due to enhanced Acoustic modeling and the remaining two thirds were due to a better decoding strategy using more and larger Language Models.

We observed that every component of the English system recognizer was ported or applied to the Spanish language system successfully without the need of major modifications. Relatively little language-specific knowledge was needed for the improvements put in place for the 1998 evaluation, especially when compared to the language-specific knowledge that was utilized in 1997 when the initial system was developed. We believe that dependence on language-specific knowledge is considerably greater when starting a system or developing a system with very small quantities of training material. We expect that further use of this type of knowledge will be of help, particularly when dealing with the challenges and pecularities of the Spanish language.

5. ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

The authors wish to thank the SPHINX group for its many contributions to this work, especially Ravi Mosur and Eric Thayer, who helped in many ways with decoder and trainer-related issues, and Rita Singh, who improved the baseline recognition accuracy of the SPHINX system.

6. References

1. Chen, S.F., and Goodman, J., "An Empirical Study of Smoothing Techniques for Language Modeling", *Harvard University, Computer Science technical report TR-10-98*, 1998.

2. Clarkson, P. and Rosenfeld, R., "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proc. of Eurospeech-97*, September, 1997.

3. Huerta, J. M., Thayer E., Ravishankar, M. K., and Stern, R. M., "The Development of the 1997 CMU Spanish Broadcast News Transcription System", *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, Feb. 1998

4. Hwang, M-Y. "Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition", Ph.D. thesis, Carnegie Mellon University, 1993. (Also published as CMU Computer Science Department Tech Report CMU-CS-93-230.)

5. Ney, H., Essen, U., and Kneser, R., "On Structuring Probabilistic Dependences in Stochastic Language Modeling", *Computer Speech and Language*, **8**:1-38, 1994.

6. Leggetter, C. J., and Woodland, P. C., "Speaker Adaptation of HMMs using Linear Regression", Cambridge University Eng. Dept., F-INFENG, Tech Report 181, June, 1994.

7. Ravishankar, M. K., "Efficient Algorithms for Speech Recognition", Ph.D. thesis, Carnegie Mellon University, 1996. (Also published as CMU Computer Science Department Tech Report CMU-CS-96-143.)

8. Seymore, K., Chen, S., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M.A., Stern, R. M., and Thayer, E. (1998). "The 1997 CMU Sphinx-3 English Broadcast News Transcription System", *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February, 1998.

9. Siegler, M. A., Jain, U., and Raj, B. "Automatic Segmentation, Classification, and Clustering of Broadcast News Audio", *Proc. of the DARPA Speech Recognition Workshop*, Feb. 1997