THE DEVELOPMENT OF THE 1997 CMU SPANISH BROADCAST NEWS TRANSCRIPTION SYSTEM

Juan M. Huerta, Eric Thayer, Mosur Ravishankar, Richard M. Stern

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

ABSTRACT

This paper describes the 1997 CMU DARPA Hub 4 Spanish Broadcast News Transcription system. The system we present is based on the CMU SPHINX-III recognizer and uses a single set of acoustic and language models. The decoding process is performed in two passes: a Viterbi search and a directed acyclic graph (DAG) search are performed on the first recognition stage. The second recognition stage is similar to the first stage except that it is performed using models adapted through maximum-likelihood linear regression (MLLR). We describe the issues relating to the design and development of the acoustic models, language models and lexicon. Developmental results and an analysis are presented.

1. INTRODUCTION

As advances in speech recognition technology allow shifting the focus of research into less restrictive domains such as the open transcription of broadcast news shows, further areas of application become feasible present new research challenges. The extension of current systems into foreign languages is such challenge. While large vocabulary multilingual speech recognition has been an area of intensive work at many research centers (*e.g.* [6, 8]), comparative tests in spontaneous multilingual speech recognition had been possible previously primarily in the context of the DARPA Hub 5 "Switchboard" evaluation. The 1997 DARPA Hub 4 non-English (NE) Evaluation provided for the first time a context for the evaluation of speech recognition systems on foreign language broadcast news material. In addition to the challenges we face in the English broadcast news domain, new issues related to the development of systems that recognize specific foreign languages must be addressed. In this paper we describe our experience developing the CMU 1997 Spanish Broadcast News Transcription system.

The Spanish component of the 1997 Hub-4 NE evaluation, includes 30 hours of standard speech training material collected from various different sources. One hour of the training material was specified as the development set. The evaluation material consisted of one hour of data recorded from a different set of dates from the training data. The material used for this evaluation reflects the characteristics and challenges of its English counterpart: unknown speakers, various acoustic and channel conditions, the presence of music etc. In addition. the evaluation also reflects the wide range of speaking styles and dialects that exist in Latin American and peninsular Spanish.

The CMU 1997 Spanish Broadcast News system is based on the CMU SPHINX-III speech recognition system. Even though the SPHINX system has previously been used for the development of foreign-language small-vocabulary application-specific systems at CMU [4], and for moderate-vocabulary tasks elsewhere [1], this is the first time that a large-vocabulary continuous-density foreign language system has been developed using the SPHINX-III platform. Since no appropriate acoustic or language models in Spanish, nor a dictionary or lexicon, were directly available to us, this task required that we design and develop a complete system from scratch.

In this paper we address the major issues related to the development of the 1997 Spanish Broadcast News Evaluation. Because of the differences across languages, fundamental system parameters such as the size of the vocabulary and language models and the dimensions of the acoustic models must be reviewed. For example, the Spanish language includes numerous word inflections with respect to change of speaker and tense for verbs with respect to change of gender for nouns and adjectives. This renders the definition of a word list that provides a low out-of-vocabulary (OOV) rate while maintaining high recognition accuracy much more challenging than in English. In Section 2 we provide a short overview of the structure of our decoder. In Section 3 we describe the characteristics of our lexicon. In Sections 4 and 5 we describe the development process and structure of our language and acoustic models, respectively. Finally, Sections 6 and 7 contain our discussion of results from the developmental test set as well as our concluding remarks.

2. SYSTEM OVERVIEW

The CMU 1997 Spanish Broadcast News Hub 4-NE system is a continuous-density senonically-tied HMM speech recognizer based on the CMU SPHINX-III English-language broadcast news system [11]. The main characteristics of the system models and lexicon are:

- A single set of full-bandwidth acoustic triphone models: 2500 senones with 8 Gaussian densities per mixture trained exclusively on the 30-hour Spanish training set provided by LDC.
- A backoff trigram language model containing 1.6 M bigrams and 3.5 M trigrams, trained principally on a prefiltered of the

Spanish Language News Corpus, a database consisting of 157M words of Spanish newspaper text provided by the LDC, and then supplemented by the transcripts from the Hub 4-NE training and development test set data.

- A lexicon consisting of approximate 40K entries, trained with a mixture of Spanish newspaper text and broadcast news transcripts in the same way as the language model

2.1. Decoder configuration

The parametrization, segmenting, and recognition phases of the system are performed in the following way: audio input is converted into cepstra, delta cepstra, and delta delta cepstra, each with 13 coefficients, and segment boundaries are automatically generated using the technique described in [12]. The segmenter parameters were modified to decrease the number of segments and increase the average segment duration. This has two implications: fewer errors are introduced by avoiding the generation of boundaries in the middle of utterances, and longer segments can be used directly for adaptation without having to cluster the segments. Acoustic classification is not performed on the resulting segments; a single set of acoustic models is employed for all incoming speech. A Viterbi beam search and DAG lattice rescoring is performed on each segment after the parametrization and segmentation. For each segment a Viterbi beam search is then performed, producing a word lattice and a best-path hypothesis. A global best path is generated from the word lattice according to the trigram grammar [3]. Single class maximum-likelihood linear regression (MLLR) [7] matrices are computed for each segment in an unsupervised fashion utilizing the best hypothesis for each segment obtained in the fist decoding pass.

A second Viterbi beam search and DAG lattice rescoring pass is then performed. This second decoding pass, similar in structure to the first decoding one, is performed using the compensated models obtained from the MLLR regression matrix and the uncompensated acoustic models. The global best path of this pass is the final hypothesis for each segment.

The fact that a single set of acoustic models is used for the whole process results in a system that does not rely on gender, environmental, or channel classification. This simplifies the topology of the overall decoding process. However, for this kind of system to produce good results, the acoustic models must perform reasonably well under the different acoustic conditions that it encounters.

3. LEXICON

The 40,000-word vocabulary consists of the most common 35,000 words observed in the Spanish Language News Corpus distributed by the LDC, supplemented by an additional 5000 words from the Hub 4-NE transcripts not observed in the first corpus. The development process consisted of defining the level of detail the pronunciations would have, defining the size and composition of the lexicon, and generating the word pronunciations.

3.1. Phonetic set definition

An important initial decision to be made was the specification of the level of detail desired for the word pronunciations. The pronunciations associated with each word should have sufficient information to allow for good recognition accuracy. On the other hand, only limited time was available for developing the system, so we needed to obtain pronunciations automatically from text transcription without providing dialect information about the speaker. Furthermore, the pronunciations should be constructed using a phonetic set that contains enough information for recognition but is at the same time sufficiently compact to allow robust training of the acoustical models. Given these constraints we opted for a set of phonemes and a pronunciation specification that would not include stress information, syllabic division, or dipthongs. The phoneme /θ/, which is uttered primarily by speakers from northern and central Spain, was not included and its possible realizations were mapped to the phoneme /s/. The resulting compact and robust phoneme set and pronunciation specification facilitated the rapid development of an automatic pronunciation generator that did not depend on external dialect information.

Consonant type	Phonemes
Labial	p,b,f,m
Dental	t,d,v
Alveolar	s,n,l,r,$\bar{r}$
Palatal	$\bar{c}$,$\underline{n}$,l,y
Velar	k,g,x

Table 1. Consonants used in the phonetic set adopted.

The consonant set that we employed is described in Table 1. It was derived from the standard set of Spanish phonemes [2, 10], and was augmented by the five vowels (/a/,/e/,/i/,/o/,/u/) to form the phonetic set used by the system. This produced a total of 24 phonemes plus silence. For comparison, in our English recognizer, the number of base phonemes used for recognition is 50 (out of which 18 are vowels and diphthongs) plus silence and hesitation noises. While our English recognizer does not include syllabic or stress information, the set of diphthongs is included.

	English	Spanish
Number of Phonemes	50	24
Number of Senones	6000	2500

Table 2. Comparison of numbers of phonemes and tied states (senones) for the CMU English and Spanish broadcast news systems.

3.2. Automatic generation of pronunciations

Automatic generation of pronunciations is possible because Spanish is a phonetic language with a very small number of irregularities and exceptions in word pronunciation. 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 the most common morphemes with irregular pronunciations.

A `perl` script using a finite-state algorithm was used to develop initial pronunciations from the word list. The total number of productions is approximately 120, including some acronym-handling rules. A final manual pass on the most common words in the lexicon was performed to modify the pronunciations of some foreign words.

3.3. Lexicon development

The lexicon was constructed by taking the 35,000 most frequent words in the prefiltered Spanish Language News Corpus. Figure 1 depicts the coverage of the HUB 4-NE Broadcast news transcripts as a function of the number of words extracted from the Spanish Language News Corpus. An additional 5,000 words extracted from the broadcast news transcripts were added to the original list of 35,000 words for the evaluation.

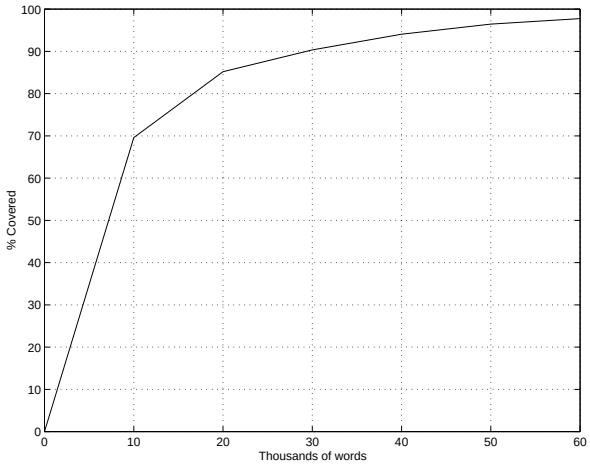

Figure 1. Coverage of the Hub 4-NE Spanish broadcast news training data versus the number of words from the Spanish Language Newspaper Corpus used for the lexicon.

4. LANGUAGE MODELS

The language model used in every stage of the recognition process is a Witten-Bell discounted trigram backoff language model. Language model development consisted of text conditioning of the Spanish Language News Corpus and the determination of language model cutoffs after the newspaper text was mixed with the broadcast transcripts. A small reduction of the perplexity of the language model developed from the news texts (when using the Broadcast News transcripts as a test set) was achieved by using Witten-Bell smoothing instead of Good-Turing.

4.1. Text conditioning

We developed our own set of tools to condition the text used for language model training. This process included the standard type of filters such as conversion of numeric quantities into words and abbreviation expansion. Because of the large number of spelling irregularities in the text corpus, and the necessity of obtaining correct spelling and accentuation, a script that automatically replaces unequivocally-misspelled words was constructed. These efforts were especially important for this system because the word bigrams and trigrams would frequently be the only source of discrimination between two words that are phonetically equivalent (according to our phonetic representation) but which are written in a different way. For example, in the absence of stress markers, the verbs in the phrases *yo acerco* and *el acercó* are identical in the pronunciation dictionary, and can be disambiguated only through the language model.

4.2. Language model development

The language model was constructed from the 40,000-word dictionary using the CMU-CU Statistical Language Model Toolkit Version 2.0 [3]. To provide robustness against spelling irregularities in the corpus, trigrams and bigrams not observed at least five times in the corpus were excluded. The training data consisted of 157 million words from the five sources included in the Spanish Text Corpus distributed by LDC.

Table 3 shows the test-set perplexity (for the training and devtest set data of the Hub 4-NE broadcast news transcripts), and the number of bigrams and trigrams in the language model as a function of the language model bigram and trigram cutoff values. These cutoff values determine how many times a bigram and a trigram must be observed in order to be included in the language model. Higher values will permit the inclusion of fewer N-grams with spelling errors into the model, but at the cost of higher perplexity. Lower values will include many more errors but will result in lower perplexity. The final configuration used cutoff value of 4 for both the bigrams and trigrams.

2G and 3G Cutoffs	Testset Perplexity	Bigrams, Trigrams	Devtest Word Error Rate
5, 5	98.12	1, 2	23.5%
4, 5	97.12	3, 4	23.5%
4, 4	96.3	5, 6	23.2%

Table 3. Testset perplexity, number of bigrams and trigrams, and error rate on the development test set of the Hub 4-NE broadcast news database as a function of the bigram and trigram cutoff values for the language model.

5. ACOUSTIC MODELS

The acoustic model development was based on the 30 hours of training material with the aid of the Latino-40 database. Several models of increasing complexity and state clustering decision trees were obtained align the training process.

5.1. Phonetic classes for state tying

The senonic clustering trees [5] were constructed using 29 acoustic classes (not including word boundaries, noises and silence). These classes were derived from phonetic classifications found in [2] and [6]. In comparison, 46 phonetic classes were used for the English-language evaluation.

5.2. Acoustic model initialization and training

Even though is possible to obtain initial models by using phonetic mappings across languages, we opted to train Spanish language models using the small Latino-40 database provided by the LDC, and then bootstrap through a series of increasingly more complex acoustic models. The Latino-40 database consists of 40 speakers reading 125 short utterances each. We trained context-independent (CI) Latino-40 models starting from flat distributions. These CI models then were used to segment the Latino-40 database to assign state boundaries to the utterances, which was needed to build our senonic clustering trees. After the trees were obtained, we trained context-dependent (CD) Latino-40 models, and these CD models were used to segment (through

forced alignment) the whole thirty hours of Hub 4-NE training material. Senonic trees were then obtained based on the broadcast news material, and CD models were trained from the broadcast news training set. Because less Spanish-language than English-language acoustical data were available, we used a smaller number of senones, 2500, than for the English-language Hub 4 system. Using density splitting we trained CD acoustic models using 2500 senones and 8 densities per senone, which were employed for recognition (*cf.* Table 4). (In contrast, the non-telephone models for our English-language Hub 4 system consisted of 6000 senones and 20 densities per senones.) The single compact set of models in the Spanish language system provided for faster decoding.

Densities per Senone	Devtest Word Error Rate
2	34.0%
4	28.4%
8	23.2%
10	25.4%

Table 4. Development set word error rate as a function of number of Gaussian densities per tied state.

6. RESULTS

Recognition results for each decoding stage for the Hub 4-NE development test data are shown in Table 5. There is a reduction in word error rate in each stage of the decoding except for the last DAG rescoring stage. MLLR was performed after each DAG rescoring. Experiments on the development test set indicated that additional MLLR passes would yield no further improvement. The word error rate on the final evaluation set data was 23.3%, compared to [FILL IN] for the English language evaluation.

Decoding Stage	Devtest WER
Viterbi Search	23.2%
DAG rescoring	22.4%
Viterbi Search	21.9%
DAG rescoring	21.9%

Table 5. Word error rate for development test set data at different stages of decoding.

7. SUMMARY AND CONCLUSIONS

We described the structure and development process of the CMU 1997 H4-NE Spanish language broadcast news recognition system. Through a single set of acoustic and language models, we demonstrated a simple and robust recognizer which provided good performance in the 1997 DARPA evaluation. This recognizer was developed and tuned without bootstraping our models from non-Spanish models.

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. We thank Kristie Seymore and Roni Rosenfeld for their help with the language models, as well as all members of the CMU speech group for their support for this effort.

REFERENCES

1. Álvarez, J., Tapias, D., Crespo, C., Cortazar, I., and Martínez, F., "Development and Evaluation of the ATOS Spontaneous Speech Conversational System, *Proc. IEEE ICASSP-97*, Munich, Germany, **2**: 1139-1142.

2. Barrutia, R. and Terrel, T. D., *Fonética y Fonologia Españolas*, Wiley, 1982.

3. Clarkson, P. and Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit," Eurospeech 1997 **5:** 2702-2710, Rhodes, Greece.

4. Frederking, R., Rudnicky, A. and Hogan, C., "Interactive Speech Transaction in the DIPLOMAT Project," Working notes of the Spoken Language Translation workshop at ACL 97, Madrid, 1997.

5. Hwang, M.-Y., "Predicting Unseen Triphones with Senones", IEEE Trans. Speech and Audiio Proc., **4**: 412-419, Novemmber, 1996.

6. Lamel, L., Adda-Decker, M. and Gauvain, J. L. "Issues in Large Vocabulary Multilingual Speech Recognition," Eurospeech 1996 Vol.1 pp.185-188, Madrid.

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

8. Pye D., Woodland, P. C. and Young, S. J. "Large Vocabulary Multilingual Speech Recognition Using HTK," Eurospeech 1995, Vol.1 pp.181-184 Madrid.

9. Ravishankar, M. K., *Efficient Algorithms for Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Computer Science Department tech report CMU-CS96-143.

10. Real Academia Española, Comisión de Gramática, *Esbozo de una Nueva Gramática de la Lengua Española*, Espasa-Calpe, Madrid 1973.

11. Seymore, K., Chen, S., Doh, S.-J., Eskenazi, E., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R., and Thayer, E., "The 1997 CMU Sphinx-3 English Broadcast News Transcription System", these Proceedings.

12. Siegler M., Jain U., Raj B. and Stern R., "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," Proc. of the 1997 ARPA Speech Recognition Workshop, pp. 97-99, Feb. 1997.