THE LIMSI 2006 TC-STAR EPPS TRANSCRIPTION SYSTEMS *

Lori Lamel, Jean-Luc Gauvain, Gilles Adda, Claude Barras, Eric Bilinski,
Olivier Galibert, Agusti Pujol, Holger Schwenk, Xuan Zhu
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, France

ABSTRACT

This paper describes the speech recognizers developed to transcribe European Parliament Plenary Sessions (EPPS) in English and Spanish in the 2nd TC-STAR Evaluation Campaign. The speech recognizers are state-of-the-art systems using multiple decoding passes with models (lexicon, acoustic models, language models) trained for the different transcription tasks. Compared to the LIMSI TC-STAR 2005 EPPS systems, relative word error rate reductions of about 30% have been achieved on the 2006 development data. The word error rates with the LIMSI systems on the 2006 EPPS evaluation data are 8.2% for English and 7.8% for Spanish. Experiments with cross-site adaptation and system combination are also described.

Index Terms – Speech recognition

1. INTRODUCTION

The TC-STAR project, financed by the European Commission under the Sixth Framework Program, is envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation. The project objectives are to significantly reduce the gap between human and machine translation performance. The second evaluation of speech recognition technologies was carried out in Jan-Feb 2006. As in the first year evaluation held in March 2005, speech recognition systems were tested for 3 languages (English, Spanish, Mandarin) and multiple tasks (European Parliament, Spanish Parliament, broadcast news). This second evaluation had several new evaluation conditions. First, automatic segmentations of the audio data were used (last year the machine translation systems imposed the use of manual segmentations). Second, although the Spanish test data came from two sources, EPPS and Spanish Parliament (Cortes), it was required that the same system be used to process all data. Thirdly, a requirement of translation systems was that the recognizer produce a case sensitive, punctuated output.

This paper describes the improvements made to the LIMSI EPPS systems in preparation for the 2006 TC-STAR evaluation, and reports on experiments carried out with system combination.

2. DATA DESCRIPTION

About 90 hours and 100 hours of audio recordings are available respectively for English EPPS and Spanish EPPS and Parliament training data, dating from 2004 and 2005. Between 3 and 4 hours of data were reserved for use as a development set (see Table 1, right). The English development data are from June 2005 and the English test data from September 2005; the Spanish EPPS development data are from June-July 2005 and the Spanish Cortes development data are from December 2004, with the test data from September-November 2005. The task-specific text data are comprised of the minutes of the European Parliament also known as the Final Text Editions. The textual training data date from April 1996 through May 2005. Table 1 summarizes the available training and test data for the 2006 evaluation.

The speech recognition evaluation conditions required automatic speech/nonspeech detection and segmentation into sentence-like units. The primary error metric was the case insensitive word error rate (WER) for English and Spanish. Systems were also required to output case-sensitive texts with punctuation marks, which were also scored.

3. SPEECH RECOGNIZER OVERVIEW

The speech recognizer for the Spanish EPPS data uses the same basic modeling and decoding strategy as in the LIMSI English broadcast news system [4]. Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree. The acoustic and language models are language and task specific. Decoding is carried out in four steps (2 more passes than the 2005 system), with unsupervised acoustic model adaptation between each step.

Two variants of the speech segmentation and clustering algorithm based on an audio stream mixture model [4] were developed. Both make use of Gaussian mixture models (GMMs) trained on 1-2 hours of English Hub4 data for speech, speech over music, noisy speech, pure-music and other background conditions. First, the non-speech segments are detected and rejected using the five GMMs representing speech. For the baseline partitioner an iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments. Each segment cluster is assumed to represent one speaker in a particular acoustic environment and is modeled by a GMM. The objective function is the GMM log-likelihood penalized by the number of segments and the number of clusters, appropriately weighted. Four sets of speech GMMs are then used to identify telephone segments and the speaker gender. Segments longer than 30s are chopped into smaller pieces by locating the most probable pause within 15s to 30s from the previous cut. For the second partitioner, the iterative GMM clustering is replaced by BIC clustering, and an additional GMM-based speaker identification clustering stage has been added. This multistage system reduces speaker error by up to 50% relative to BIC alone on French and English broadcast news data [1]. The result of the procedure is a sequence of non-overlapping segments with cluster, gender and tele-

*This work was partially financed by the European Commission under the FP6 Integrated Project TC-STAR.

Training/ Task		*Audio data* #Sessions	Size	Epoch	*Text data* (words)
English EPPS	texts	63	91h	Apr04 -	34M
	transcripts			Jan05	690k
Spanish EPPS	texts	63	61h	Apr04 -	36M
	transcripts			Jan05	471k
Spanish Cortes	texts	24	38h	Sep04 -	47M
	transcripts			Oct04	268k

Task	*Development and Test Data* Data type	Size	Epoch
English EPPS	Dev	3.2h	Jun05
	Eval	3.2h	Sep05
Spanish EPPS	Dev	2.4h	Jun-Jul05
	Eval	3.3h	Sep-Nov05
Spanish Cortes	Dev	3.9h	Dec04
	Eval	4.0h	Nov05

Table 1: Summary of available audio and textual training data (left) and 2006 development and evaluation data (right).

Language	*English*	*Spanish*
P1	5k / 5k	2.0k / 2.0k
P2	28k / 11.5k	5.6k / 8.1k
P3	18k / 11.7k	6.3k / 8.7k
P4	18k / 11.5k	6.3k / 8.7k

Table 2: Acoustic models used in the different decoding passes. The #contexts and # tied states are given for each model set.

Language	*English*	*Spanish*
#words	60k	65k
#phones	48+3 / 38+3	27+3 / 25+3
#prons	74k / 74k	94k / 78k

Table 3: Language-specific pronunciation lexicons.

phone/wideband labels.

4. ACOUSTIC MODELING

Standard HMM training requires an alignment between the audio signal and the phone models, which usually relies on an orthographic transcription of the speech data and a good phonemic lexicon. It is common to Viterbi align the orthographic transcriptions with the signal using existing models (via the lexicon) to produce a time-aligned phone transcription. This alignment generally also uses manual segmentations into speaker turns or sentence-like units.

In this work a revised acoustic model training procedure is used, which relies on an automatic segmentation and speaker labeling, instead of the manual annotations. This revised method aligns the words in the reference transcripts with an automatic segmentation created by the audio partitioner. This results in a significantly simplified training procedure which is also more coherent with the subsequent decoding steps. This homogeneous (simplified) method has been applied to all tasks and languages, and can optionally allow non-speech events to be inserted during the alignment step.

Table 2 summarizes the characteristics of the various acoustic model sets used in the four decoding stages for the evaluation systems. All acoustic models are MLLT-SAT trained, gender-dependent, tied-state position-dependent triphone models with backoff to right/left context and context-independent models. Separate cross-word and word-internal statistics are used to select the contexts to be modeled, and language-specific decision trees are used to tie the model states using a divisive decision tree based clustering algorithm.

The English acoustic models were trained on about 90 hours of audio training data from the EPPS English distributed by RWTH. The first pass models cover 5k triphones with 5k tied states (32 Gaussians per state). The second pass models use a reduced phone set and were trained on 600 hours of BN data, 150h with manual transcripts, 450h of selected TDT2,3,4 data (via light supervision) and adapted with the EPPS data. The third and fourth pass models are different iterations of MMIE-trained models, each with about 18k triphones and 11.5k states (32 Gaussians per state).

The Spanish acoustic models were trained on about 100 hours of audio training data from EPPS and Cortes corpora. The first fast models cover 2k contexts with 2k tied states. The second pass models use a reduced phone set (merging /s,z/ and the two r's). The third and fourth pass models are different iterations of MMIE-trained models, each with about 6k triphones and 9k tied states (32 Gaussians per state).

5. PRONUNCIATION LEXICA

The English pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). In the reduced phone set, pronunciations are represented with 38 phones, formed by splitting complex phones. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 60k case-sensitive vocabulary contains 59993 words and has 74k phone transcriptions. Compound words for about 300 frequent word sequences subject to reduced pronunciations were included in the lexicon, as well as the representation of 1000 frequent acronyms as words.

The Spanish pronunciations are based on a 27 phone set (3 of them are used for silence, filler words, and breath noises). A second reduced phone set dictionary merges variants for *s/z* and *r/R* which are poorly distinguished by the common word phonetization script,. Pronunciations for the case-sensitive vocabulary are generated via letter to sound conversion rules, with a limited set of automatically derived pronunciation variants. While the rules generate reasonable pronunciations for native Spanish words and proper names, other words are more problematic. The Unitex (www-igm.univ-mlv.fr/unitex/) Spanish dictionary was used to locate likely non-Spanish words, which belong to several categories: typos (which were fixed at the normalization level); Catalan words, borrowed words like *'sir'* or *'von'*, non-Spanish proper nouns which were hand-phonetized by a native speaker; and acronyms. Non-Spanish proper nouns were the most difficult to handle, especially those of Eastern European origin where the variability in the audio data shows that native Spanish speakers do not necessarily know how to pronounce them. The decision taken was to use the perceived phonetization for the names which were represented in the audio data, and use the native speaker's intuition for the rest. Although including non-Spanish phones to cover foreign words was considered, these were too infrequent to estimate reliable models so they were replaced with the closest Spanish phone. Acronyms that tend to be pronounced as words were verified by listening to the audio data or phonetized by a native speaker. The final lexicon has 94871 pronunciations for 65004 entries.

Language	*English*	*Spanish*
Dev06 data	EPPS	EPPS/Cortes
#words	60k	65k
OOV	0.3%	0.6%
Transcripts	690k	471k / 268k
Texts	33.5M	36M / 47M
BN+CNN	293M+180M	
4g ppx	88	80 / 102

Table 4: Summary of EPPS language models.

6. LANGUAGE MODELING

For all systems, n-gram language models were obtained by interpolation of backoff n-gram language models using the modified Kneser-Ney smoothing (as implemented in the SRI toolkit [2, 8]) trained on separate subsets of the available language model training texts. The characteristics of the language models are summarized in Table 4. A neural network LM [7] was trained on the EPPS transcripts and texts, and interpolated with the 4-gram back-off LMs.

Since the text processing is case sensitive, a decision must be taken as to what the true case of each sentence-initial word is. Moreover for some texts the caseing is vague (due to emphasis or segmentation errors), and the caseing of all words needs to be reconsidered. In order to be able to attribute the correct case for the sentence-initial word an interpolated LM was constructed with a set of texts after removing the first word of each sentence. Caseing is added to the original sentence by creating a graph with all possible caseings for all words with multiple caseings, and parsing the graph using the interpolated LM.

Word lists for English and Spanish selected by choosing the n most probable words after linear interpolation of unigram LMs trained on the different text sources so as to minimize the perplexity on the dev data. n is chosen to minimize the OOV ratio while keeping a reasonable size and correctness of the words. For Spanish, a 65k case-sensitive word list was chosen as a good compromise, yielding an OOV rate of 0.6% on the dev06es data. The 2006 English word list is also case-sensitive and contains 60k words, and has an OOV rate of about 0.3%.

The English language models result from the interpolation of component LMs trained on 4 sources: 690k words of audio transcriptions (cut-off 0-0-0); 34M words of Parliamentary texts (cut-off 0-0-1); 180M words of CNN captions [01/2000-31/05/2005]; and 293M words of Broadcast news transcriptions (cut-off 1-1-2). The mixture weights where chosen to minimize the perplexity of the development data. The 4-gram perplexity on the dev06en data is about 88.The LM contains about 8.1M bigrams, 32.8M trigrams and 24.2M 4-grams. The perplexity is reduced to 75 with the NN LM .

For Spanish, component language models were trained on 6 text sources: European Parliament transcriptions (471K words); Spanish Parliament transcriptions (268K words); European Parliament final text editions (FTE) 1996-1999 (15M words); European Parliament FTE 1999-2004 (19M words); European Parliament FTE 2004-2005 (2M words); and Spanish Parliament texts (47M words). The texts were normalized to a common form, and names with multiple written forms were mapped to the most frequent one (*Juncker/Junker, Breshnev/Brezhnev*). Several processing steps were applied to transform the texts closer to a 'spoken' form. (Although originating from speeches, the texts were transformed into a written form for publication on the web sites.) The main normalization steps are similar to those for English [4]: separation of punctuation from words; expansion of abbreviations (*Sr.* → *Señor*); treatment of numerical expressions (*artículo 82.1* → *artículo ochenta y dos uno*, *3.900 millones* → *tres mil novecientos millones*). Acronyms not found in the word list were split into their component letters in order to get an "unknown spelled acronym" model.

	Decoding Pass			
WER(%)	*Pass1*	*Pass2*	*Pass3*	*Pass4*
English EPPS	15.5	11.6	10.0	9.8
Spanish EPPS	10.0	8.3	7.0	6.9

Table 5: Word error rates (%) after each decoding pass on the English and Spanish EPPS Dev06 data.

System		*Feb05*	*Mar06*	
Language	*Task*	*Dev06*	*Dev06*	*Eval06*
English	EPPS	14.0	9.8	8.2
Spanish	EPPS	9.8	6.9	7.8
	Cortes			13.3

Table 6: Word error rates (%) on the English and Spanish Dev06 and Eval06 data.

Independent models were estimated on each text source and then interpolated with coefficients estimated to minimize the perplexity on the development data. The perplexity of the EPPS dev06 data with the 4-gram model is 79.5, and the perplexity of the Spanish Parliament dev data is 102.4. The perplexities with the NN LM are 71.2 and 92.2 respectively.

7. DECODING

Word recognition is performed in four passes, where each decoding pass generates a word lattice with cross-word, position-dependent, gender-dependent AMs, followed by consensus [6] decoding with 4-gram and pronunciation probabilities. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [5] techniques prior to each decoding pass. The lattices of the last two decoding pass are rescored by the neural network (NN) LM interpolated with a 4-gram backoff LM. The total decoding time is about 6xRT. More specifically, the decoding steps are:

1) Initial hypothesis generation using small cross-word EPPS acoustic models and audio partitioner 1 ($\simeq$ 1.0xRT);

2) 2 class MLLR adaptation of large BN+EPPS acoustic models (AMs) for English and large EPPS+Cortes AMs for Spanish, each with a reduced phone set, and audio partitioner 2;

3) Data driven MLLR adaptation with large EPPS MMIE-trained AMs for English and large EPPS+Cortes MMIE AMs for Spanish, neural network LM interpolated with a 4-gram LM;

4) Data driven MLLR adaptation with large English EPPS MMIE-trained AMs and large Spanish EPPS+Cortes MMIE AMs (the MMIE AMs are different from step 3), NN LM interpolated with a 4-gram LM.

Table 5 gives the word error rates on the EPPS dev06 data for after each decoding pass. The word error after the first real-time decoding pass is 15.5% for English and 10% for Spanish. The largest improvement is obtained in the second pass (25% and 17% relative respectively for English and Spanish), with smaller gains in the subsequent passes.

Table 6 gives the recognition results for the evaluation systems on the TC-STAR Dev06 and Eval06 data sets. The WERs of the Feb05 systems on the Dev06 data are also given. The overall Spanish WER is 10.7%. Relative word error rate reductions of about 30% were obtained for both the English and

Data	*Method*	*Systems*	*WER*	*Rel.Gain*
Dev06en	Rover1	LIMSI06v3, IBM06v3, RWTH3, IRST3, UKA2	9.4	-15%
	Adapt	IBM06v2 + LIMSI06v3	9.1	-15%
	Rover1 + Adapt	+ LIMSI06v3	9.0	-16%
	Rover2	LIMSI06v4, IBM06v4, UKA4, RWTH4, IRST4	8.7	-14%
	Rover2 + Adapt	+ LIMSI06v4	8.7	-14%
Dev06es	Adapt	IRST05, LIMSI05e	8.7	-5%
	Rover1	LIMSI06v2, RWTH06v2, IBM05, IRST05	6.6	-8%
	Rover2	LIMSI06v2, RWTH06v2, IBMv3, IRST06	5.8	-19%

Table 7: Some system combination results on dev06en (top) and dev06es (bottom).

Spanish systems on the Dev06 EPPS data. In a post-evaluation study, the audio partitioner was modified to not throw away music segments, which reduced the overall Spanish WER to 10%.

8. TC-STAR SYSTEM COMBINATION

Various decoding and system combination methods were studied, based on cross-site adaptation and Rover-like combination. A subset of the results are reported in Table 7. The first entry shows the result of Rover combination [3] of five systems with word error rates ranging from 11 to 16%. The combination results in a 15% gain relative to the best system (10.7%). Cross-site adaptation, i.e. adapting LIMSI models using a transcription from another partner (2nd entry) or from a combination of systems (third entry), is seen to be very efficient as the resulting word error rate is always lower than (or equal to) the WER of the adaptation transcripts, and is considerably lower than the WER of the stand alone system (with relative gains of up to 15%). Even though there were significant improvements for all systems used in Rover2 (WERs ranging from 10.1 to 13%), almost the same relative gain is obtained as with the systems used in Rover1. Similar observations can be made for the Spanish systems, where substantial improvements were made to the systems used in the second Rover.

9. PUNCTUATION

Automatic caseing and punctuation tools have been developed for English and Spanish. These modules use both linguistic and acoustic information (essentially pause and breath noise cues) to add punctuation marks in the speech recognizer output which can be either a single best hypothesis or a word lattice. Separate language models were constructed for speech recognition and punctuation, the former explicitly modeling speech characteristics and disfluencies, and the latter modeling punctuation, but without the disfluencies. Starting with the recognizer hypotheses with time-marks (CTM file), pauses longer than 1.7s are located and a word graph is created for each speech segment. All possible caseings of each word are added to the graph, as well as optional sentence breaks at each pause, and optional punctuation marks (,COMMA and .PERIOD) after each word. The resulting augmented word graph is then decoded with a punctuated, case sensitive LM. (The LIMSI punctuator was not used in the eval submission but was used for SLT).

10. CONCLUSIONS

This paper has summarized the progress made in preparation for the second annual TC-STAR speech recognizer evaluation for the EPPS task in English and Spanish. The baseline performance was that of the Feb'05 systems on the 2006 development data. For English the initial word error rate was reduced from 14.0% to 9.8% and for Spanish the word error rate was reduced from 9.8% to 6.9%. The additional features and improvements to the English and Spanish features include automatic segmentation, four decoding passes with unsupervised adaptation, two phone sets per language (full and reduced), and MLLT, SAT, MMIE training. Large word error rate reductions of about 30% were obtained compared to last year's system.

Innovations contributing to this large performance improvement came from new strategies for unsupervised AM adaptation based on different type of models and different segmentation schemes. Significant improvement is due to the use of more data to build larger and more accurate models, and improved within site and cross-site system combination. One idea growing in popularity is to use alternative models and segmentations in successive decoding passes so as to reduce the impact of the recognition errors, segmentation errors and clustering errors on the adaptation process. Improvements also came from better pronunciation modeling, the use of additional acoustic features, improved SAT model estimation and improved discriminative training methods, and improved neural network LMs.

11. ACKNOWLEDGMENTS

The authors thank the TC-Star partners (IBM, IRST, RWTH, UKA) for collaborating in the combination experiments.

REFERENCES

[1] C. Barras, X. Zhu, S. Meignier, J.L. Gauvain, Multi-stage speaker diarization of broadcast news, *IEEE Trans. Audio, Speech & Language Processing* **14**(5):1505-1512, 2006

[2] S.F. Chen,J. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, *34th Annual Meeting ACL*, Morgan Kaufmann Publishers, San Francisco, A. Joshi and M. Palmer, Eds., 310-318, 1996.

[3] J. Fiscus, A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), *ICASSP'97*, 347-354, Quebec.

[4] J.L. Gauvain, L. Lamel, G. Adda, The LIMSI Broadcast News Transcription System, *Speech Communication*, **37**(1-2):89-108, May 2002.

[5] C.J. Leggetter, P.C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech & Language*, **9**(2):171-185, 1995.

[6] L. Mangu, E. Brill, A. Stolcke, Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeech'99*, 495-498 Budapest.

[7] H. Schwenk, J.L. Gauvain, Building Continuous Space Language Models for Transcribing European Languages, *Eurospeech'05*, 737-740, Lisbon.

[8] A. Stolcke, SRILM - An extensible language modeling toolkit, *ICSLP'02*, **II**:901-904.