THE LIMSI 2006 TC-STAR 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 evaluated in the TC-STAR Second Evaluation Campaign held in January-February 2006. Systems were developed to transcribe parliamentary speeches in English and Spanish, as well as Broadcast news in Mandarin Chinese. 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 European Parliament Plenary Sessions (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. The character error rate for Mandarin for a joint system submission with the University of Karlsruhe was 9.8%. Experiments with cross-site adaptation and system combination are also described.

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 (SST). SST technology is a combination of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (speech synthesis). 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). In this second evaluation there were 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 two sources of Spanish data were included in the test set, a requirement was that the same system be used to process the EPPS data and the data from the Spanish Parliament (Cortes). Thirdly, the systems were required by the translation systems to produce a case sensitive, punctuated output.

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

2. TASK AND DATA DESCRIPTION

The TC-STAR project is addressing speech-to-speech translation in 3 languages and 3 tasks. For the public European Parliament Plenary Sessions (EPPS) tasks the training cut-off date was set at May 31st, 2005, meaning that no audio or text data after that day could be used for training. 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. 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). 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 Mandarin data are from the LDC Hub4 Mandarin data (27 hours with manual transcripts) as well as portions of data from the TDT2, TDT3, and TDT4 corpora for which only closed-captions were available.

The speech recognition evaluation conditions required automatic speech/nonspeech detection and segmentation into sentence-like units. The primary error metrics were the case insensitive word error rate (WER) for English and Spanish and the character error rate (CER). 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 [1].

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

Training/	*Audio data*			*Text data*		*Development and Test Data*		
Task	*#Sessions*	*Size*	*Epoch*	*(words)*	*Task*	*Data type*	*Size*	*Epoch*
English	63	91h	Apr04 -	34M	English	Dev	3.2h	Jun05
EPPS			Jan05	690k	EPPS	Eval	3.2h	Sep05
Spanish	63	61h	Apr04 -	36M	Spanish	Dev	2.4h	Jun-Jul05
EPPS			Jan05	471k	EPPS	Eval	3.3h	Sep-Nov05
Spanish	24	38h	Sep04 -	47M	Spanish	Dev	3.9h	Dec04
Cortes			Oct04	268k	Cortes	Eval	4.0h	Nov05
Mandarin	350	27h	Jan97 -	600M chars	Mandarin	Dev	3.2h, 6 shows	01-11 Dec98
BN	shows	+ 170h	Dec00	460k chars	BN	Eval	4.0h, 4 shows	23-25 Dec98

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

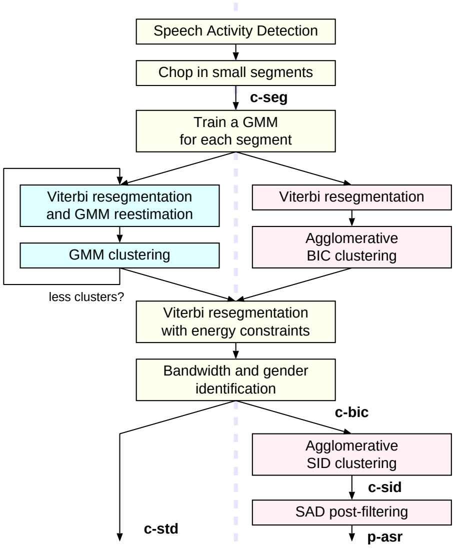
Speech Activity Detection
Chop in small segments
c-seg
Train a GMM for each segment
Viterbi resegmentation and GMM reestimation
GMM clustering
less clusters?
Viterbi resegmentation
Agglomerative BIC clustering
Viterbi resegmentation with energy constraints
Bandwidth and gender identification
c-bic
Agglomerative SID clustering
c-sid
SAD post-filtering
c-std
p-asr

Figure 1: Architecture of the baseline and multi-stage speaker diarization system.

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 [1] 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 (advertisements). 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 [2]. The architecture of the baseline and multi-stage speaker diarization systems are shown in Figure 1. The result of the procedure is a sequence of non-overlapping segments with cluster, gender and telephone/wideband labels.

4. ACOUSTIC MODELING

The LIMSI speech recognizer [1] uses 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

The same general method is used to construct acoustic models for each task/language using the available training data. For English and Spanish, standard supervised training is performed, whereas for Mandarin both supervised and lightly supervised training is used [3]. 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. After normalizing the transcripts and completing the pronunciation dictionary, 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 gener-

Language	English	Spanish	Mandarin
Data type	EPPS	EPPS/Cortes	BN
Audio data	71h	73h	27h+170h
Transcript	manual	manual	man+light
P1	5k / 5k	2.0k / 2.0k	(UKA)
P2	28k / 11.5k	5.6k / 8.1k	24k / 11.5k
P3	18k / 11.7k	6.3k / 8.7k	24k / 11.5k
P4	18k / 11.5k	6.3k / 8.7k	24k / 11.5k

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

ally also uses manual segmentations into speaker turns or sentence-like units.

The acoustic model training procedure has also been completely revised, using an automatic segmentation and speaker labeling, instead of using 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 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.

English models: 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).

Spanish models: 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).

Mandarin BN models: The Mandarin acoustic models were trained on about 27 hours of Hub4-Mandarin training data (from LDC) and about 170 hours of data from the TDT2, TDT3 and TDT4 corpora. Most of these data (about 140 hours) are from VOA. Since time-aligned transcripts are not available for the TDT corpora, models were trained using a lightly supervised training method. [4, 3]. The TDT data from the Mainland China sources (CNR, CTV and VOA) were transcribed with a recognizer using the RT03 BN evaluation system acoustic models and source/show-specific language models estimated on the TDT closed captions for each source/show. Wideband and bandlimited models were trained by pooling the Hub4 Mandarin data and the TDT data. The acoustic models are position-dependent triphones with tied states, obtained using a divisive decision tree based clustering algorithm. Two sets of gender-dependent acoustic models were built using both MAP adaptation [5] of SI seed models for each of wideband and telephone band speech.

5. PRONUNCIATION LEXICA

English: Pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). 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. As done in the past, 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.

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 know 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

Language	English	Spanish	Mandarin
#words	60k	65k	54k
#phones	48 / 38	27 / 25	61
#nonspeech	3	3	4
#prons	74k / 74k	94k / 78k	55k

Table 3: Language-specific pronunciation lexicons.

Language	English	Spanish	Mandarin
Dev06 data	EPPS	EPPS/Cortes	BN
#words	60k	65k	54k
OOV	0.3%	0.6%	$\sim$0
Transcripts	690k	471k / 278k	460k ch
EPPS texts	33.5M	36M	600M
BN+CNN	293M+180M		
Cortes		47M	
4g ppx	88	80 / 102	250

Table 4: Summary of language model development.

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.

Mandarin: The Mandarin lexicon is represented with 61 phones (4 of them are used for silence, filler words, and breath noises). There are 24 consonants and 11 vowels, with the 5 tones for vowels are collapsed into 3 (rising, flat and falling). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 54k vocabulary contains 54025 words with 55377 phone transcriptions.

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 [6]) 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 for English and Spanish, and interpolated with the 4-gram back-off LMs.

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. 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 case-sensitive and contains 60k words, and has an OOV rate of about 0.3%.

English: The English language models result from the interpolation of component LMs trained on 4 sources:
1. Audio transcriptions: 690k (previously 353k words), (cut-off 0-0-0)
2. Parliamentary texts: 34M (previously 32M words), (cut-off 0-0-1)
3. CNN captions [01/2000-31/05/2005]: 180M words
4. Broadcast news transcriptions: 293M words, (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 (compared to 92 with last year's model). 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 .

Since the new 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.

Spanish: Component language models were trained on 6 Spanish text sources:
1. European Parliament transcriptions (471K words)
2. Spanish Parliament transcriptions (268K words)
3. European Parliament final text editions (FTE) 1996-1999 (15M words)
4. European Parliament FTE 1999-2004 (19M words)
5. European Parliament FTE 2004-2005 (2M words)
6. 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 applied to English [1], such as the separation of punctuation from words, the expansion of abbreviations (*Sr.* → *Seor*), the treatment of numerical expressions (*artculo 82.1* → *artculo ochenta y dos uno, 3.900 millones* → *tres mil novecientos millones*), removing the sentence initial capitalization when appropriate, separating pseudo compound words, and splitting the texts into sentence-like units. After processing there were a total of 84M words (225K distinct) and 2.8M sentences.

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. Acronyms not found in the word list were split into their component letters in order to get an “unknown spelled acronym” model. The models were first estimated independently for each source using the standard Kneser-Ney smoothing as modified by Chen and Goodman [9]. The independent models where 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.

Mandarin: The 4-gram language models were obtained by interpolation [10] of backoff 4-gram language models trained on 8 sources available via the LDC.

1. Hub4 audio transcripts, 460k characters
2. China radio international 1994-96, 87M characters
3. People Daily newspaper, 89.2M characters
4. Xinhua news, 9.9M characters
5. TDT2,3,4 XIN, 12M characters
6. TDT2,3,4 ZBN, 12M characters
7. TDT2,3,4 VOA, 2.3M characters
8. LDC gigaword Mainland texts, 367M characters

The 54k word list was selected from the same text sources so as to minimize the OOV rate on the dev05 data. The word list includes all (about 7000) characters (i.e., there are essentially no OOV characters). The perplexity of the dev06 data was about 250 with the 4-gram LM.

7. DECODING

For English and Spanish 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 [11] decoding with 4-gram and pronunciation probabilities. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [8] techniques prior to each decoding pass. The lattices of the last two decoding pass are rescored by the neural network LM interpolated with a 4-gram backoff LM.

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 backoff LM; 4) Data driven MLLR adaptation with large EPPS MMIE-trained AMs and large EPPS+Cortes MMIE AMs (the MMIE AMs are different from step 3), neural network LM interpolated with a 4-gram backoff LM. Table 5 gives the word error rates on the EPPS dev06 data for English and Spanish 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.

	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 for 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
	All			10.7
Mandarin	BN	10.9	10.7	9.8

Table 6: Word/character error rates on the TC-STAR Dev06 and Eval06 data. Mandarin was a joint submission with UKA.

Table 6 gives the recognition results for the evaluation systems on the TC-STAR Dev06 and Eval06 data sets. Relative word error rate reductions of about 30% were obtained for both the English and 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 [12] of five systems with word error rates ranging from 11 to 16%. The combination results in a 15% gain relative to the best system. 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 signficant improvements for all systems (WERs ranging from 10.1 to 13%), Rover2 obtains almost the same relative gain as 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

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).

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. 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 and the BN task for Mandarin Chinese. 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%. There was no major development effort for Mandarin Chinese. 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.

REFERENCES

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

[2] C. Barras, X. Zhu, S. Meignier, J.L. Gauvain, Multi-stage speaker diarization of broadcast news, *IEEE Trans. on Audio, Speech and Language Processing*. (to appear).

[3] L. Lamel, J.L. Gauvain, G. Adda, Lightly Supervised and Unsupervised Acoustic Model Training, *Computer, Speech and Language*, **16**(1):115-229, Jan. 2002.

[4] T. Kemp, A. Waibel, Unsupervised Training of a Speech Recognizer: Recent Experiments, *Eurospeech'99*, Budapest, **6**:2725-2728, Sept. 1999.

[5] J.L. Gauvain, C.H. Lee, Maximum A Posteriori for Multivariate Gaussian Mixture Observation of Markov Chains, *IEEE Trans. Speech & Audio Proc.*, 291-298, 1994.

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

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

[8] 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.

[9] 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.

[10] L. Chen, J.L. Gauvain, L. Lamel, G. Adda, Unsupervised Language Model Adaptation for Broadcast News," *ICASSP'03*.

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

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