SPEECH RECOGNITION IN MULTIPLE LANGUAGES AND DOMAINS: THE 2003 BBN/LIMSI EARS SYSTEM

R. Schwartz, T. Colthurst, N. Duta, H. Gish, R. Iyer, C-L. Kao, D. Liu, O. Kimball, J. Ma, J. Makhoul, S. Matsoukas, L. Nguyen, M. Noamany, R. Prasad, B. Xiang, D-X. Xu
BBN Technologies, 10 Moulton St., Cambridge, MA 02138
J-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen
LIMSI-CNSR BP133, 91403 Orsay, Cedex, France

ABSTRACT

We report on the results of the first evaluations for the BBN/LIMSI system under the new DARPA EARS Program. The evaluations were carried out for conversational telephone speech (CTS) and broadcast news (BN) for three languages: English, Mandarin, and Arabic. In addition to providing system descriptions and evaluation results, the paper highlights methods that worked well across the two domains and those few that worked well on one domain but not the other. For the BN evaluations, which had to be run under 10 times real-time, we demonstrated that a joint BBN/LIMSI system with that time constraint achieved better results than either system alone.

1. INTRODUCTION

In May 2002, DARPA initiated a 5-year research program, called EARS (Effective, Affordable, Reusable, Speech-to-text), a major goal of which is to reduce recognition word error rates (WER) for conversational telephone speech (CTS) and broadcast news (BN) by a factor of 5 in 5 years, down to the 5-10% range, running in real-time on a single processor. The drive to lower WER and to real-time is in several phases, with milestones to be achieved at the end of each phase. Progress is measured on a "Progress Test" in English which remains fixed for the five year duration of the program. In addition, there are "Current Tests" in each of the three languages (English, Arabic, and Mandarin), which change every year. Collaboration across sites was strongly encouraged. BBN and LIMSI have been working closely together and, wherever possible, submitted joint results.

This paper reports on the results of the evaluations that took place in April 2003 and on the lessons learned from working across domains (CTS and BN) and across the three languages, while taking full advantage of trans-Atlantic collaboration. In addition to describing the techniques used, the progress made in the last year of work, and the results achieved, the paper points to those techniques that worked well across the two domains and those few methods that appeared to work well for one domain but not the other. While there were no time constraints for the CTS evaluations, the BN evaluations, all of which had to be run under 10xRT (real-time), presented an interesting challenge in combining systems. We found that it was better to combine the BBN and LIMSI systems, each running at significantly less than 10xRT, than either system alone could achieve at 10xRT.

2. BASIC SYSTEM DESCRIPTIONS

The BBN BYBLOS System [1] and the LIMSI System [2][3] are similar in that they both use HMMs, but there are many differences. Here we describe each system as it existed in April 2002 for the RT02 (Rich Transcription 2002) baseline evaluation. In Section 3 we describe improvements made before the April 2003 (RT03) evaluations.

Front End: The BBN system used Mel-Frequency Cepstra while the LIMSI system used PLP Cepstra [4]. The BBN system computed third derivatives of the features, but then used HLDA [5] to project the space down to 46 dimensions. Both systems scale the frequency axis of the features using Vocal Tract Length Normalization (VTLN). There are also differences between BN and CTS, mostly related to the segmentation of the input into speaker turns and utterances.

HMM Models: The BBN system uses phonetic HMMs, with State-Clustered-Tied Mixture (SCTM) distributions. The states of each phonetic model are distinguished based on 'quinphone' context into several thousand different "codebooks" of 24-64 Gaussians each. Further division results in many sets of mixture weights (~10) for each codebook. The number of codebooks and mixture weights depends on the amount of training data. Each phoneme model has the same 5-state topology, with a minimum duration of two frames for the entire phoneme.

The LIMSI HMM models differ in a few significant ways. The system uses tied state crossword triphones that share both the set of Gaussians and the mixture weights. Each model is a 3-state HMM imposing a minimum duration of 3 frames for a phone. State-tying is based on decision tree clustering with backoff to diphone and monophone models. The gender-dependent models cover about 30k contexts with 10k tied states, and have 16 and 32 Gaussians per state for BN and CTS, respectively.

In the BBN system, training of each acoustic model is performed using the Forward-Backward EM algorithm with time-constraints provided by 'fuzzy labels'. Training in the LIMSI system is performed using discrete time-state labels. This makes the training process very fast.

Recognition Search: The BBN recognition search uses multiple passes, with progressively more detailed information used in each pass [6]. The first pass uses a single phonetic tree with Phonetically Tied Mixtures (PTM) with a fast approximate

search whose sole function is to find those words that might be present. The second pass, which runs backwards, uses non-cross-word SCTM models and an approximate trigram search; it uses the forward pass scores in a "forward-backward search" to greatly reduce the choices of words to consider. The traceback from the second pass is converted to a lattice of alternative words, which are then used to find the N-best hypotheses (N=300). These hypotheses are re-scored using cross-word SCTM models and a trigram or fourgram LM. This 3-pass process is considered one 'decoding'.

The LIMSI system also uses multiple decoding steps (2 for BN, 4 for CTS). Each decoding step generates lattices with a bigram or trigram language model and with cross word triphones. The word lattices are then expanded with a four-gram language model to perform a consensus decoding with pronunciation probabilities. All passes are full forward decodes with no approximation other than the standard pruning strategies.

Speaker Adaptation: Both systems perform unsupervised speaker adaptation using MLLR [7]. The BBN system used Speaker Adaptive Training (SAT) [8] during training of the acoustic models to reduce the variance of the speaker-independent models.

Dictionary and Language Model: The two systems have somewhat different phonetic dictionaries with approximately 48 phonemes each. The BBN CTS system defines 2500 'compound words' that are concatenations of common words, while the other configurations use a smaller number of compound words. The BBN system uses Witten-Bell smoothing and combines all the data with different weights. The LIMSI system uses Kneser-Ney smoothing and estimates separate models for each source and then combines the models with weights estimated to minimize perplexity.

3. GENERAL IMPROVEMENTS

Both BBN and LIMSI made significant system improvements for the RT03 evaluation held in April 2003. Some of them were applied only within one site; others where developed initially by one site and then quickly implemented in the other, after success was demonstrated. Below is a list of improvements.

- Applied to BN systems all technology improvements that were made on CTS from 1999 to 2002. For BBN, this technology transfer included a redesigned HMM initialization procedure based on Gaussian splitting (instead of K-means), HLDA, and improved speaker adaptation.
- Implemented lattice-based Maximum Mutual Information Estimation (MMIE) [9].
- Improved performance of system combination (ROVER) [10] by designing systems that were significantly different from each other but with similar word recognition accuracy. Systems differed in feature extraction (MFCC vs. PLP), and phoneme set/pronunciation dictionary. For English BN and CTS, such systems were generated by both BBN and LIMSI, and a subset of these were combined at various stages of recognition.
- Both sites tuned the performance of their BN automatic segmentation procedure. In addition, BBN developed an automatic method to determine speaker turns on CTS data. The output of the latter process was used by both sites for recognition on the development and evaluation test sets.
- BBN explored the use of constrained MLLR (CMLLR) adaptation for speaker adaptive training, as described in [11]. The resulting CMLLR-SAT procedure allowed for transparent integration with MMIE, due to its simplified model parameter reestimation.
- Developed HLDA-SAT [12], a new speaker adaptive training procedure that estimates speaker dependent HLDA feature projections, based on a small HMM with a single full covariance Gaussian per tied state. HLDA-SAT was used only within the BBN system.
- BBN found improved recognition accuracy for using all counts during the estimation of n-gram language models. Because of limitation in memory, these large LMs were used only in N-best rescoring, compiling the list of needed n-grams on the fly from a large database of counts.
- Added more training data to both acoustic and language model training, and increased the number of model parameters.
- Used a neural net to map n-grams to a continuous space, for improved language modeling [13]. This feature was used only within the LIMSI system.

4. CONVERSATIONAL TELEPHONE SPEECH

For CTS, each test file consists of a conversation, typically ten minutes long, between two people talking about a suggested topic. The two channels are recorded separately, but there is substantial cross talk between the channels. The systems must segment the speech and provide the recognition answers for each channel.

On the EARS CTS Progress Test set, the baseline system for 2002, the BBN RT02 system, achieved a 27.8% WER. On the same test the BBN/LIMSI RT03 system achieved a 17.5% WER, a 37% relative reduction in error. In addition to the general improvements discussed in the previous section, there were a number of CTS-specific changes that contributed to the greatly decreased error.

In NIST benchmark tests prior to the RT03 evaluation, the correct segmentation of the test data was provided along with the audio, so participating systems typically had no need of a segmentation capability. To be able to compare the performance of the baseline BBN RT02 system with later systems on the EARS Progress Test, where the correct segmentation was not given, the RT02 system was run using a segmentation algorithm developed by MIT Lincoln Laboratory at about the time of the RT02 test.

In the year between the RT02 and RT03 tests, the BBN/LIMSI team developed its own segmentation algorithm, which was designed to be robust to cross talk and line noise. The algorithm used a broad-class HMM to model observations consisting of the joint cepstral features from both channels of the conversation [14]. This algorithm performed well, giving a WER degradation of just 0.1%-0.4% compared with careful manual segmentation, depending on the language and test set.

Table 1 gives the absolute reductions in WER as a result of various techniques, on English CTS data, listed for both BBN and LIMSI systems. Gains were computed relative to the RT02 evaluation test set, with manual segmentation. The WER of the baseline systems was about 28-29%.

Technique	BBN	LIMSI
Gender-Dependent VTLN		0.5
PLP	0.6	
CMLLR-SAT	0.5	
MMIE	1.5	1.2
Larger LM	0.5	
Quickly Transcribed Swbd data	1.3	0.9
WEB+Archived LM data	0.5	0.3
Larger Lexicon	0.3	0.1
Lattice MLLR	0.3	
Neural Net LM		0.4
Revised decoding		1.0
Trans-Atlantic Sys. Comb.	2.7	4.0

Table 1: Reductions in WER for BBN/LIMSI on RT02.

Additional training data provided significant improvements this year. BBN oversaw the quick transcription of 80 hours of data drawn from the Switchboard-II and Switchboard Cellular corpora and distributed this data to the EARS research community prior to the RT03 evaluation [15] for both acoustic and LM training. In addition, we increased our LM training data by adding 60 million words of data that was collected from the web and provided to the community by the University of Washington and about 47 million words of archived transcripts from CNN and PBS. When we increased our LM training data, we also increased the size of our lexicon. BBN also improved performance by adopting lattice MLLR adaptation [16].

For CTS, the BBN/LIMSI system used an expensive but effective method of system combination, dubbed Trans-Atlantic system combination, that was used previously by SRI to combine systems they developed internally [17]. In our case, three BBN systems and two LIMSI systems, each of which first ran its own adaptation, were combined with ROVER to produce a single transcript, which was then used to re-adapt each of the systems. After adaptation, each system re-recognized the test and the results from each system were again combined with ROVER, which became our final system output.

BBN participated in the Mandarin and Arabic RT03 CTS evaluations with systems configured very similarly to the English system (though without combining with LIMSI systems). The principal difference was that training, development testing, and tuning were done with data from the appropriate language. The gains for these systems relative to previous years' results were due to the same set of technology improvements described above.

For Mandarin acoustic model training, BBN used 35 hours of CallHome and CallFriend Mandarin data; for language modeling, we also used 487 million words of Mandarin Broadcast News data to smooth the CTS data. The resulting system achieved a character error rate of 42.7% on the EARS RT03 Current Test set.

BBN's Arabic system was trained with 20 hours of CallHome Arabic data for both the acoustic and language models. The BBN Arabic CTS system achieve a 37.5% WER on the EARS RT03 Current test set.

The English, Mandarin and Arabic CTS systems described in this section each had the lowest error rate on their respective EARS RT03 CTS Current Test sets.

5. BROADCAST NEWS

For BN, each test file consists of the first half hour of a news broadcast, including commercials, but the commercials are not scored. The number of speakers in the show may vary from a few, to a few dozen. The segmentation process for BN remained essentially the same during the past year, with minor tuning.

On the EARS BN Progress Test set, the BBN RT02 system achieved a 18.0% WER. On the same test the BBN RT03 system achieved a 13.8% WER, a 23% relative reduction in error. The LIMSI system improved from 16.4% to 14.0%. In addition to the general improvements discussed in Section 3, there were some BN-specific changes that contributed to the greatly decreased error.

The list of improvements and their respective gains for the BBN and LIMSI BN systems, measured on a development test set (Dev03), are shown in Tables 2 and 3 below.

Detail of Improvements	%WER
0. Baseline (RT-02 mothballed system)	16.3
1. Baseline + New auto segmentation	16.1
2. Modern ML training	15.2
3. + HLDA-SAT	14.6
4. + MMI Training	13.7
5. + 4-gram rescoring	12.9
6. + TDT4 acoustic training data	11.9
7. + speedup options to run < 10xRT	12.2
8. + Updated lexicon and LMs	11.8
9. + One more pass of adaptation	11.6

Table 2: Improvements in BBN BN system on Dev03 test.

Detail of Improvements	%WER
0. Baseline (RT-02 mothballed system)	14.5
1. Baseline (RT-03 Dryrun system)	14.1
2. + MMI Training	13.6
3. + TDT4 included in LM	12.6
4. + TDT4 light supervision in acoustic model	12.2
5. + Optimized LM and decoding	11.8

Table 3: Improvements in LIMSI BN System on Dev03 test.

Most of these improvements are described above in Section 3. However, the use of TDT4 speech data required some new development. The TDT4 data has closed captions rather than careful transcriptions. Lamel [18] has previously shown that it is possible to use closed captions for acoustic training, even though they do not necessarily match the speech. Both BBN and LIMSI used this speech, but with different procedures. At BBN, we bias the LM to the new data by adding the closed captions in with a large weight. Then we decode the data, and keep those utterances where the recognized words match the closed captions exactly. (Since the April '03 evaluation, BBN has changed the procedure to keep any sequence of three or more words that matches exactly [19].) At LIMSI, we use a fair language model to decode the new data, and keep any utterance that differs from the closed caption by a small enough percentage. In either case, the result is significantly more speech training data, which reduces WER – especially in the BBN system.

We combined the two results (BBN at 11.6% and LIMSI at 11.8%) using Rover, and the WER decreased to 10.3%, but the combined CPU time was 17.3xRT, so this system did not meet the 10xRT requirement. After the RT-03 evaluation, we reduced the computation by combining the two systems in a novel way as shown in Figure 1. The LIMSI system was run first at 5.8xRT to produce a result that was then used to adapt the BBN system, which ran at 3.4xRT. The adapted LIMSI and BBN results were combined, again achieving 10.3%, with a CPU time of 9.2xRT.

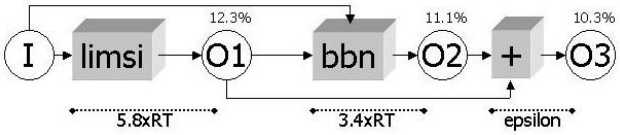

Fig. 1: BBN/LIMSI Integrated System. WER and times shown.

BBN also participated in the BN evaluation for Arabic and both sites participated separately in Mandarin. The techniques used (including the use of TDT4 data) were the same, except for some details. For Arabic, BBN defined the "phonetic spelling" for each word directly from the printed letters (consonant and unknown vowel) so that we could use large amounts of available text data in the LM.

In Mandarin, both BBN and LIMSI use traditional phonemes with distinctions of tone for vowels and final consonants rather than initial-final demi-syllables. Both sites built separate models for the data from Mainland China and Taiwan due to differences in language and acoustic conditions of the recordings. On the Dev03 set, the BBN system improved from 33.8% CER to 17.7% CER, a relative reduction of 47%; the LIMSI system improved from 34.5% to 22.6%, a relative reduction of 36%. On the evaluation, the results were 19.1% and 21.7% respectively.

6. SUMMARY

We were able to run essentially the same system on both BN and CTS in all three languages. The processing of the input, such as segmentation was necessarily different. But the recognition techniques were the same. Methods that worked on English always worked in Mandarin and Arabic, despite the different nature of the languages and the greatly reduced amount of training. There were differences in the phoneme set, but the internal structure of the phonetic models was unchanged. A few methods did not work equally well on the two domains. VTLN did not help for BN and the HLDA-SAT technique did not help for English CTS. Four-grams did not help in the BBN CTS system – perhaps due to the very large number of compound words that we defined.

One encouraging result was that more training continues to help significantly, as long as we increase the number of parameters appropriately.

7. REFERENCES

[1] L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz and J. Makhoul, "Progress in transcription of Broadcast News using Byblos", Speech Communication 38(1 2):213-230, Sep 2002

[2] J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen and F. Lefevre, "Conversational telephone speech recognition", *ICASSP*, I-212-215, Hong Kong, April 2003.

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

[4] H. Hermansky, "`Perceptual linear predictive (PLP) analysis of speech, *JASA*, vol. 87, no. 4, pp. 1738--1752, April 1990.

[5] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition", *Speech Communication*, vol. 26, no. 4, December 1998.

[6] L. Nguyen and R. Schwartz, "Efficient 2-pass N-best decoder," Proc. EuroSpeech, Rhodes, Greece, Sep. 1997.

[7] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Computer Speech and Language*, vol. 9, no. 2, pp. 171-186, April 1995.

[8] T.Anastasakos, J.McDonough, R.Schwartz, and J.Makhoul, "A compact model for speaker adaptive training", *ICSLP*, Philadelphia, PA, USA, October 1996.

[9] P. C. Woodland and D. Povey, "Large Scale Discriminative Training for Speech Recognition", *Proc. ISCA ITRW ASR2000,* 2000.

[10] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)", *ASRU*, pp. 347-354, Santa Barbara, CA, 1997.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", Tech. Rep. 291, University of Cambridge, May 1997.

[12] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections", to appear *ASRU*, St. Thomas, November 2003.

[13] H. Schwenk and J.L. Gauvain, "Using Continuous Space Language Models for Conversational Speech Recognition", Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, April 2003.

[14] D. Liu and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech", to appear in the *Proceedings of the 2003 ASRU Workshop*, St. Thomas, November 2003.

[15] O. Kimball, R. Iyer, C. Kao, T. Arvizo and J. Makhoul, "Using quick transcriptions to improve conversational speech models", submitted to *ICASSP*, Montreal, Canada, May 2004.

[16] L. F. Uebel and P. C. Woodland, "Improvements in Linear Transform Based Speaker Adaptation", *ICASSP* 2001.

[17] http://www.nist.gov/speech/tests/rt/rt2002/presentations/sri+-rt02-stt.pdf

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

[19] L. Nguyen, B. Xiang, "Light supervision in acoustic model training," submitted to ICASSP, Montreal, Canada, May 2004