THE 2004 BBN/LIMSI 10xRT ENGLISH BROADCAST NEWS TRANSCRIPTION SYSTEM

Long Nguyen, Sherif Abdou, Mohamed Afify, John Makhoul, Spyros Matsoukas,
Richard Schwartz, Bing Xiang
BBN Technologies, 10 Moulton St., Cambridge, MA 02138, USA

Lori Lamel, Jean-Luc Gauvain, Gilles Adda, Holger Schwenk, Fabrice Lefevre
LIMSI-CNRS, BP133, 91403 Orsay Cedex, France

ABSTRACT

This paper describes the 2004 BBN/LIMSI 10xRT English Broadcast News (BN) transcription system which uses a tightly integrated combination of components from the BBN and LIMSI speech recognition systems. The integrated system uses both cross-site adaptation and system combination via ROVER, obtaining a word hypothesis that is better than is produced by either system alone, while remaining within the allotted time limit. The system configuration used for the evaluation has two components from each site and two ROVER combinations, and achieved a word error rate (WER) of 13.9% on the Dev04f set and 9.3% on the Dev04 set selected to match the progress set. Compared to last year's system, there is around 30% relative reduction on the WER.

1. INTRODUCTION

Right after the EARS 2003 Evaluation, BBN and LIMSI developed a novel approach for system combination by employing cross-site adaptation and ROVER [1] using components from the BBN and LIMSI broadcast news (BN) transcription systems [16]. It was shown that the combined system outperformed both of the individual systems. Continuing on that effort, BBN and LIMSI developed an integrated system, the 2004 BBN/LIMSI 10xRT English BN transcription system, for the EARS RT04f evaluation. This system tightly integrates four subsystems, two from each site. In addition to the significant improvements achieved within each site, cross-site adaptation and system combination are employed in the integrated system which further reduce the error rate while maintaining an overall running time of under 10xRT. A relative gain of about 30% has been achieved compared to either of last year's single systems.

This paper is organized as follows. Section 2 summarizes the audio and speech corpora used for training and test. The 10xRT system architecture used in the evaluation system is described in Section 3. Descriptions of the recognizer modules developed at BBN and LIMSI are given in Sections 4 and 5, respectively. Section 6 provides experimental results with the integrated system on different test sets and presents some of the strategies we investigated for system combination. Finally, some conclusions are given in Section 7.

2. TRAINING AND TEST CORPORA

Two types of acoustic data were available for training. The first consists of the carefully transcribed broadcast news training data (a total of about 140 hours from the 1995, 1996, and 1997 official Hub4 training sets). Additionally about 9000 hours of broadcast news audio data were distributed by the LDC. These include the prior distributions of the TDT2 (630 hours, Jan-June 1998), TDT3 (475 hours, Oct-Dec 1998) and TDT4 (300 hours), the extra TDT4 data from March-July 2001 (465 hours) and the 7000 hours EARS BN data collected from March-Nov 2003. Since the remaining audio data do not have time-aligned transcripts, lightly supervised training [7] was carried out making use of the associated closed-captions, when available. The basic idea is to decode the data using an existing system and choose the segments that agree well with the closed-captions.

At BBN about 1600 hours of audio data were selected from the untranscribed audio data [15] and pooled with the Hub4 training data. At LIMSI, the acoustic models were trained on the Hub4 training data and about 450 hours of data selected from the TDT corpora (150h TDT2, 140h TDT3, 250h TDT4). Only audio segments where the error rate between the hypothesized automatic transcription and the associated aligned closed-captions was under 30% were used for training.

The language model training data include the manual transcriptions of the acoustic BN data (1.8M words); and the American English GigaWord News corpus provided by LDC, for a total amount of approximately 1 billion words of texts. At LIMSI, the transcriptions of the CTS data (27.4M words), commercial transcripts purchased directly from PSMedia, and CNN web archived transcripts (112M words from Jan'2000-Nov'2003, excluding 01/15/01-02/28/01) were also used for training. All data predates November 15, 2003, and excludes the period from 01/15/2001 through 02/28/2001.

This year we defined a new set of development data (harder than the Dev03 set) in an attempt to better predict performance on the progress set. While the average word error on the Dev03 data was slightly higher than for the Eval03 shows, the development set is quite a bit easier than the progress set. Also, even though it was decided that the Eval03 data could be used for system development, we thought it best to keep this data for validation test, and to choose the Dev04 data from the latter part of January 2001. Since there are no reference transcripts for the January 2001 TDT4 data, we tried to assess the show difficulties by scoring the recognizer hypotheses against the closed-captions. Four STT sites, LIMSI, BBN, CU, and SRI transcribed the shows from the second half of January using their RT03 systems. Using these scores a subset of 9 shows were selected, two from each source. For the remaining 3 shows, no ROVER results were available. (There are no closed-captions for the MSN_NBW data.) A combined hypothesis was generated by aligning the ROVER output with the captions according to the LIMSI partitioner segments.

The reference transcripts were obtained by manually correction of the ROVER of the 4 recognizer hypotheses. (This correction was led by LIMSI, but shared amongst the four sites.) Once the shows were corrected, the system hypotheses and the ROVER output were rescored and minor adjustments were made. Different combinations of shows, one from each source were then scored, in order to find a balance in difficulty compared to the Dev03 and Eval03 data sets as well as the broadcast dates. The Dev04 set consists of the following shows:

```
    20010125_2000_2100_PRI_TWD,
    20010127_1830_1900_ABC_WNT,
    20010130_1830_1900_NBC_NNW,
    20010130_2100_2200_MSN_NBW,
    20010128_1400_1430_CNN_HDL,
    20010131_2000_2100_VOA_ENG,
```

where the last two shows were also part of the Dev03 set.

A second set of development data for the current test was distributed by LDC, called the Dev04F set. These data consist of 6 shows from the second half of November 2003, one from each of the following sources: ABC, CSPAN, CNN, CNNHL, CNBC, and PBS.

The Eval04F data consist of 12 shows, recorded during December 2003 from the following 7 sources: ABC, CNN, CNNHL, CNBC, CSPAN, PBS and WBN; the last one not covered in the development test. The first show from Eval04F dates from December 2nd, and the last one from December 19th. It should be noted that some of the shows were selected by NIST to be harder than those of Eval03.

3. BBN/LIMSI SYSTEM STRUCTURE

The BBN/LIMSI 2004 English BN evaluation system uses a tightly integrated combination of the BBN and LIMSI speech recognition component systems. The following systems were used in the combined system, more details about each system will be given in next two sections. Systems from BBN are denoted with prefix “B” and those from LIMSI with prefix “L”, also “R” is used to indicate ROVER results.

- **B1**: BBN system running two decoding passes in a relatively fast mode.
- **L1**: LIMSI system that adapts to B1 and does a full 3-pass, 4-gram decode.
- **R1**: ROVER on B1 and L1.
- **B2**: BBN system that adapts to R1 and runs one decoding pass.
- **R2**: ROVER on B1, L1 and B2.
- **L2**: LIMSI system that adapts to R2 and runs another full decode.
- **R3**: The final output is a ROVER of L1, B2, and L2.

Figure 1 shows a schematic diagram indicating the flow of the combined system. The dashed arrow-headed line indicates which hypothesis or ROVER of system hypotheses is used for adaptation.

4. BBN SYSTEM

This section contains a description of the BBN components in the combined system. BBN Byblos speech recognition system consists of three main modules, segmentation and clustering, feature extraction, and decoding, as described below. Acoustic model training, language model training and the improvements achieved at BBN will also be presented.

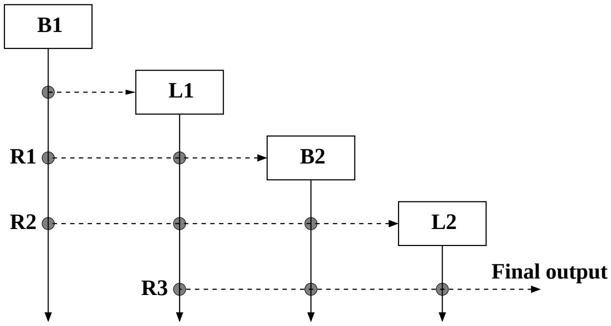

Figure 1: High-level structure of the tightly integrated BBN/LIMSI System. “B” denotes systems from BBN, and “L” systems from LIMSI. “R” is used to indicate ROVER results of the hypotheses joined by the dashed line.

4.1. Segmentation and Clustering

In the segmentation stage [12] the input speech is first segmented into wideband and narrowband material, using a dual-band phoneme decoder. Each channel is then normalized with RASTA, and a dual-gender phoneme decoder is applied to detect gender changes and silence locations. For each channel-gender chunk, speaker change detection is performed based on the Bayesian Information Criterion (BIC) and results in a segmentation that defines speaker turns, along with their gender and channel labels. Finally, the speaker turns are chopped into short segments (averaging 7 seconds) based on the detected silence locations. The resulting segments are then clustered using an online algorithm that uses a penalized likelihood measure [9]. The obtained clusters are used for adaptation and decoding as described below.

4.2. Feature Extraction

Features are extracted from overlapping frames of speech with a duration of 25 ms at a rate of 100 frames/sec. The current system uses 14 perceptual linear prediction (PLP) [5] derived cepstral coefficients and energy. The features are normalized so that they have zero mean and unit variance for each speaker turn. The static features are augmented with their first, second, and third order derivatives, which leads to an initial 60-dimensional parameter space. The dimension is then reduced to 46 using heteroscedastic linear discriminant analysis (HLDA) [6].

4.3. Decoding

The decoding consists of two passes. The first pass outputs a transcription which is used as supervision to adapt the acoustic models, and the adapted models are used in the second decoding pass. Alternatively, the system can accept recognition hypotheses from any other system and run only one decoding pass. In the following we will describe both the search and adaptation parts of the decoding stage.

A multipass search strategy is employed where each stage is used to constrain the search space of the following pass. In the current system a forward pass and a backward pass are run followed by N-best rescoring.

- The forward pass [14] uses simple acoustic models, State-

Tied-Mixture (STM) models, and a bigram language model. The search algorithm is a Viterbi beam search on a single static tree, where bigram probabilities are applied at the inner nodes of the tree in order to speed up the search. The output of this pass is the set of the most likely words at each frame.

- The backward pass [13] then uses the output of the forward pass to guide a Viterbi beam search with more complex acoustic and language models, i.e. at each time only active words from the forward pass are considered during the search. A state-clustered tied-mixture (using decision trees) within-word quinphone acoustic model (SCTM-NX), and a trigram language model are used in this step. During the backward pass an N-best list, and also a lattice can be generated.
- In the current system an N-best list (N=300) is output by the backward pass. This list is rescored using a state-clustered tied-mixture cross-word quinphone model (SCTM-XW), and a 4-gram language model. The top scoring utterance is output as recognition hypothesis.

Each of the above decoding stages includes adaptation which is applied during the second recognition pass using supervision obtained from the first pass. It is worth noting here that the first recognition pass uses speaker independent (SI) models while the second pass employs speaker adaptively trained (SAT) models. Also in all decoding stages fast Gaussian computation and quantized codebooks are employed for improved efficiency.

Adaptation has the following three stages:

- As stated above the system uses an HLDA transformation. The first adaptation step consists of estimating a speaker-specific HLDA transform [11] instead of the general HLDA transform used in the first pass.
- The second adaptation phase consists of a constrained maximum likelihood linear regression (CMLLR) transform, which is a linear feature space transform [2].
- The final adaptation step, after the above two feature space transforms, amounts to estimating L MLLR [8] transforms of the model parameters based on a tree clustering of the model distributions. L is set to 2 for within-site adaptation and 16 for cross-site adaptation. In the current version, MLLR is implemented using least square estimation for efficiency. MLLR is applied to the three models used in decoding, namely, the STM, SCTM-NX and SCTM-XW.

4.4. Acoustic Model Training

The BBN system uses phonetic hidden Markov models (HMMs) to represent an inventory of 50 phonemes, where each HMM has a left-to-right topology and consists of 5 states. These phonetic models are used as building blocks for words from a 64K pronunciation lexicon. There are a total of 250 (50 by 5) states but due to the use of context dependent modeling (triphone and quinphone) there are a huge number of different instances of the same state. Different instances of the same state are clustered using a decision tree algorithm, and each leaf node is modeled using a Gaussian mixture model. In the current system a two-level tying structure is used for the means and variances (referred to as codebooks), and the mixture weights (referred to as pdfs), i.e. different tree depths are used for the means and variances, and the mixture weights.

In total six acoustic models are needed for the two decoding passes, namely the STM, SCTM-NX and SCTM-XW, for both SI and SAT. The STM is trained with within-word triphones, while the SCTMs are trained with quinphones. Although the final models are discriminatively trained using maximum mutual information (MMI) [18] estimation, maximum likelihood (ML) models are also trained as initial models, to build the lattices required for MMI training, and to obtain the mixture weights. The SAT training paradigm that we use employs a speaker specific HLDA transform and a CMLLR transform, which are applied in the feature space, facilitating the ML estimation of SAT models and also their extension to MMI. The training of each of these models consists of the following steps.

- Align the data to the states, using some initialization method, and construct a decision tree for each state. Note that the STM has only one Gaussian mixture for each state.
- Initialize a Gaussian mixture for each tree leaf, then refine the parameters by running six iterations of the forward-backward algorithm.

The SAT training is exactly the same except that data of each speaker is first mapped using the speaker dependent HLDA, and CMLLR as described above. If discriminative training is desired (either MMI or MPE) the following steps are applied:

- Decode the training data in forward and backward pass and output lattices.
- Annotate the lattices with phonemes using the cross-word quinphone SCTM.
- Apply a number of iterations of the generalized Baum algorithm for MMI training of the means and variances of the Gaussians. The state transition probability and mixture weights are taken from the corresponding ML models.

This year's system training made use of 1700 hours of data. These include 140 hours of Hub-4 data that were used last year, while the rest is obtained using light supervision on broadcast news data with captions. The basic idea is to decode the data using an existing system and choose the parts that agree well with the captions [15]. Using this data typical model sizes used for this year's evaluation are listed in Table 1.

Model	# Codebooks	# Gaussians
STM	250	119K
SCTM-NX	6320	767K
SCTM-XW	6489	790K

Table 1: Number of codebooks and Gaussians in STM, SCTM-NX and SCTM-X

4.5. Language Model Training

The language models were estimated from the available Broadcast News data and the GigaWord News corpus provided by LDC.

The total amount of data used was approximately 1 billion words. We created a single model, weighting the counts for the data from the TDT programs by a factor of 3-6 relative to data from other sources. We used a modified Witten-Bell smoothing technique, which we measured to give equivalent results to modified KN smoothing for this amount of data. The lexicon contained about 64K words, of which 1945 were frequently occurring compound words. The language models for decoding contained about 12M 2-grams and 28M 3-grams. The models for rescoring included all 4-grams observed in the training data, that is about 730M.

4.6. Improvements in the BBN System

The BBN system has been significantly improved since RT03 Evaluation. The contributions of various techniques are listed in Table 2. Compared to the RT03 system, there was 3% absolute (22.4% relative) gain on the EARS 2004 development test set Dev04.

Detail of Improvement	% WER
0. Baseline (RT-03 system)	13.4
1. 843-hour acoustic training	12.1
2. 1700-hour acoustic training	11.3
3. + MMI SAT PTM	11.2
4. + MMI SI PTM, SCTMs	11.0
5. + duration modeling	10.9
6. + online speaker clustering	10.8
7. + longer utterances	10.5
8. + new lexicon, LM	10.4

Table 2: Improvements in BBN system on Dev04 test set.

Among those improvements, the additional acoustic training data provided the largest gain. Compared to the RT-03 system trained with 214 hours acoustic training data (141 hours of LDC data plus 73 hours of automatically selected TDT4 data), there were around 1600 hours of additional data being selected via light supervision from the TDT and BN 2003 corpora. These additional data contributed a 15.7% relative gain on the Dev04 test set. In our RT03 system, MMI models were only used in the backward pass and rescoring in adapted decoding stage. With MMI models in all passes, a gain of 0.3% was obtained. Duration modeling and online speaker clustering each provided a minor gain of 0.1%. A change in the segmentation stage to cut the speech into longer utterances (7 seconds on average in contrast to the previously used 4 seconds) gave a 0.3% gain. An update of the lexicon and language model also resulted in a small 0.1% gain.

5. LIMSI SYSTEM

This section contains a description of the LIMSI components of the combined system. New additions in the 2004 system are the integration of MLLT and SAT feature transformations, as well as CMLLR adaptation, and MLLR adaptation using a tree organization for the adaptation classes. The acoustic models were trained on 600 hours of BN data, using lightly supervised training for the TDT data. We also use for BN a neural net language model, which had previously only been used in our CTS systems. Although we did not develop a stand-alone 10xRT system for this evaluation, and focused on system combination with BBN and with the SuperEars system, our improvement relative to the LIMSI RT03 system is estimated to be about 20%.

One question that we addressed was is it more efficient, in terms of accuracy and speed, to use lattice rescoring or full decode for system combination. We therefore pursued two decoding strategies for cross-site system combination, one based on acoustic lattice rescoring and the other one relying on a fast full search.

For cross-site lattice rescoring, the BBN lattices were transformed to be compatible with LIMSI vocabulary. This was done by first decompounding the BBN compound words and then applying the LIMSI compounding rules (for 906 compound words and 568 acronymns) to the lattices, i.e. adding a compound link for each link sequence corresponding to a compound. The acoustic scores (log-likelihood) of the new links were obtained by summing the component scores. All words not in the LIMSI vocabulary were then mapped to silence and the lattices were expanded and rescored with the LIMSI 4-gram LM (keeping the original acoustic scores). Each lattice was then pruned and transformed into a consensus graph which served as a grammar for acoustic rescoring by a dynamic network decoder using the LIMSI acoustic models and 3-gram LM. The hypothesis generated using the lattice was also used to carry out MLLR adaptation of the LIMSI acoustic models. For the full search solution, the BBN hypotheses served for MLLR adaptation prior to decoding and no lattices are exchanged between systems. We found that cross-site adaptation with a full decode was both a simpler and more efficient solution, when used with ROVER combination, therefore this is the solution adopted for the RT04 evaluation.

5.1. Segmentation and Clustering

The LIMSI segmentation and clustering is based on an audio stream mixture model [3, 4]. First, the non-speech segments are detected and rejected using GMMs representing speech, speech over music, noisy speech, pure-music and other background conditions. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments. The result of the procedure is a sequence of non-overlapping segments with their associated segment cluster labels. 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 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.

5.2. Feature Extraction

The speech features consist of 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.8kHz 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. LPC-based cepstrum coefficients are then computed. These cepstral coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives. This feature vector is linearly transformed (MLLT) to better fit the diagonal covariance Gaussians used for acoustic modeling.

5.3. Decoding

The L1 and L2 decodes are each performed in three steps. Before decoding CMLLR [2] and MLLR [8] adaptations are performed using the hypothesis of the preceding system component (i.e. the B1 hypothesis for L1 decode and the R2 hypothesis for L2 decode). Then a word lattice is produced for each speech segment using a dynamic network decoder with a 2-gram language model. Finally, the word lattice is rescored with a 4-gram neural network language model and converted to a confusion network [10]. The MLLR adaptation relies on a tree organization of the tied states to create the regression classes as a function of the available data. This tree is built using a full covariance model set with one Gaussian per state. Gaussian short lists and tight pruning thresholds are used to keep the real-time factor under 3xRT for the L1 decode and under 2xRT for the L2 decode.

5.4. Acoustic Models

The acoustic models were trained on about 150 hours of Hub4 training data (the 1995, 1996, and 1997 official Hub4 training sets) and about 450 hours of data from the TDT corpora (150h TDT2, 140h TDT3, 250h TDT4). Since time-aligned transcripts are not available for the TDT data, unsupervised training was used for this data [7]. Only audio segments where the error rate between the hypothesized automatic transcription and the associated aligned closed-captions was under 30% were used for training. The L1 acoustic models include 37k position-dependent triphones with 12k tied states, obtained using a divisive decision tree based clustering algorithm with a 48 base phone set. Two sets of MLLT-SAT gender-dependent acoustic models were built for each data type (wideband and telephone) using MAP adaptation of SI seed models and MMI training. The L2 models include 28k position-dependent triphones with 12k tied states for a reduced 38 phone set.

The basic pronunciations are taken from the LIMSI American English lexicon, for which the most frequent inflected forms have been verified to provide more systematic pronunciations. The pronunciation probabilities are estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing for unobserved pronunciations. The 65523 word lexicon has 78044 pronunciations using the full 48 phone set and 77952 with the reduced 38 phone set.

5.5. Language Models

A single interpolated 4-gram backoff LM was built from 9 component models trained on subsets of the available text materials including the transcriptions of the acoustic BN data (1.8M words); the transcriptions of the CTS data (27.4M words); the TDT2, TDT3 and TDT4 closed captions (14.3M words); commercially produced BN transcripts from LDC and PSMedia (260M words); CNN web archived transcripts (112M words from Jan'2000-Nov'2003, excluding 01/15/01-02/28/01); and newspaper texts (1463M words). All data predates November 15, 2003 with the period 01/15/2001-02/28/2001 being excluded. The word list contains 65523 words and has an OOV rate of 0.48% on the Dev04 set and 0.57% on Dev04f set. The word list also contains compound words for about 300 frequent word sequences and about 1000 frequent acronyms.

Separate LMs were built for Dev04 and Dev04f using the same word list and training data, but different interpolation coefficients. During decoding the date of the show is used to select the LM. The interpolation coefficients were estimated using an EM procedure to optimize the perplexity on the Dev04 data set.

In addition a neural network LM [17] was trained on a subset of about 27M words of data (BN transcriptions, TDT2, TDT3 and TDT4 closed captions and 4 months of CNN transcripts from 2001). The neural network LM is interpolated with the 4-gram backoff LM previously described and then used to rescore word lattices. The perplexity of the Dev04 data is 109.9 for the 4-gram backoff LM alone and 105.4 when interpolated with the neural net LM.

6. EXPERIMENTAL RESULTS

The experimental results of the BBN/LIMSI 10xRT system on development and evaluation test sets are presented in this section. The BBN compute platform is an Intel Xeon (3.4 GHz, 8GB RAM) running Linux RedHat 7.3, with hyperthreading. At LIMSI the compute platform is an Intel Pentium 4 extreme (3.2GHz, 4GB RAM) running Fedora Core 2 with hyperthreading.

6.1. Results on the Dev04 Test Set

Table 3 lists the word error rates and running time at each stage of the integrated system. As described in Section 3, we employed both cross-site adaptation and system combination. In the first pass, the BBN system B1 generated hypotheses with an 11.0% error rate. Then, the LIMSI system L1 adapted to the hypotheses of B1 and reducing the WER to 10.1%. A ROVER of B1 and L1 provided another 0.3% gain. The BBN system B2 adapted to the ROVER results, obtains a word error rate of 9.9%, which when combined in the second ROVER of B1, L1 and B2, resulted in a word error of 9.5% at 7.4xRT. These ROVER results provided supervision for the second LIMSI system L2. The final ROVER between the three best systems produced hypotheses with a 9.3% error rate at 9.2xRT. Compared to the BBN RT-03 system, there is 30.6% relative gain. The gain is 10.6% relative when compared to the BBN RT04 stand-alone system.

System	% WER	xRT
B1	11.0	2.6
L1	10.1	2.7
B1+L1	9.8	5.3
B2	9.9	2.1
B1+L1+B2	9.5	7.4
L2	9.9	1.8
L1+B2+L2	9.3	9.2

Table 3: Results on the development test set Dev04.

6.2. Results on the Dev04f Test Set

The results on the EARS 2004 Fall development set Dev04f are listed in Table 4. The error rate is decreased from 15.8% in the first pass down to the 13.9% in the final ROVER output. The running time is around 9.7xRT, which is a little bit more than that on

Dev04 due to the difficulty of the test data. The BBN RT03 system had 19.7% error rate on this test set, so the relative gain from the BBN/LIMSI system is 29.4%.

System	% WER	xRT
B1	15.8	2.7
L1	15.1	2.9
B1+L1	14.6	5.6
B2	14.3	2.2
B1+L1+B2	14.1	7.8
L2	14.9	1.9
L1+B2+L2	13.9	9.7

Table 4: Results on the development test set Dev04f.

6.3. Results on the Eval04 and Progress Test Sets

The results on the 2004 Evaluation test set are given in Table 5. The error rate is reduced from 14.4% in the first pass down to the 12.7% in the final ROVER output, with run time of about 9.8xRT.

System	% WER	xRT
B1	14.4	2.7
L1	13.6	3.0
B1+L1	13.2	5.7
B2	13.4	2.2
B1+L1+B2	12.8	7.9
L2	13.5	1.9
L1+B2+L2	12.7	9.8

Table 5: Results on the Eval04 test set.

With the same structure, the BBN/LIMSI system achieved 9.5% error rate at 9.3xRT on the progress test set, which satisfies the EARS target in terms of both recognition performance and decoding time constraints. Compared to the BBN results on the progress test last year, 13.8%, there is a 31.1% relative WER reduction.

	System	WER	xRT
A	BBN two-pass, lattice generation	11.1	3.5
B	LIMSI, rescore lattices	10.5	1.5
R1	ROVER of A,B	10.5	5.0
C	BBN PLPdelta, adapt to B, redecode	10.4	2.5
R2	ROVER of A,B,C	9.9	7.5
D	BBN PLPfr adapt to R2, rescore n-best C	10.3	1.5
R3	ROVER of A,B,C,D	9.6	9.0

Table 6: Alternative system configuration with a slower first pass.

6.4. Discussion

The above subsections have summarized the results of the tightly integrated BBN/LIMSI RT04 broadcast news system. The final configuration was determined after quite a lot of experimentation, exploring a variety of factors, such as: how many decoding passes could be carried out and still fit in the time constraints? which site should go first? what is the best performance/speed tradeoff for the first pass (i.e., is it better to have a faster first pass that mainly serves for cross-site adaptation) and more or slower additional passes? Is it better for latter passes to be full decodes or lattice rescoring? Is it better to adapt to individual system hypotheses or to a ROVER result combining multiple systems. While we were unable to address all combinations of the above questions, we did try to explore the large space of possibilities in an attempt to draw general conclusions about system combination.

	System	WER	xRT
A	BBN one-pass, no adaptation	14.9	1.0
B	LIMSI, adapt to A, redecode	11.4	2.0
C	BBN, PLPdelta adapt to B, redecode	10.4	2.5
D	BBN PLPfr, adapt to B, rescore C n-best	10.4	1.5
R1	ROVER of B,C	10.2	5.5
E	BBN PLPfr adapt to R1, rescore C n-best	10.3	1.5
R2	ROVER of B,C.D	9.9	7.0
R3	ROVER of B,C,E	9.8	7.0

Table 7: Alternative system configuration with a fast (1xRT) first pass.

Tables 6, 7 and 8 report some of the system combinations that were considered, for the Dev04 set. In all of these configurations, the BBN system was run first, which is in contrast to our combined RT03 post-eval system where the first pass decode was done by LIMSI. This decision was taken in part since the BBN was developing a real-time BN system.

In Table 6, a slower first pass was run by BBN, followed by lattice rescoring at LIMSI. In 5xRT a word error of 10.5% is obtained, which is somewhat worse than the 9.8% error at 5.3xRT as reported in Table 3. The ROVER of these two hypotheses does not give any gain. A second decode by BBN after adapting to the LIMSI hypothesis results in a 10.4% word error, which is reduced to 9.9% at 7.5xRT after ROVER of the three hypotheses. A rescoring pass of another BBN system trained with 9-frame-concatenated PLP features (noted as PLPfr) provided 10.3% word error. The final ROVER reduced the WER to 9.6% in 9xRT. In Table 7 a faster first pass decode was first carried out at BBN, with an error rate of 14.9%. LIMSI adapted to these hypotheses and redecoded, producing a hypothesis at 11.4% in 3xRT. Different options (rescoring n-best, redecoding) were tried on the BBN side, with obtaining an error rate of 9.8% in 7xRT. While it was possible to pursue further cross-site experiments to reduce the error rate with the remaining 3xRT available, we took the decision to further explore an intermediary first pass at 2xRT, shown in Table 8.

Table 8 reports word error rates and compute times for a number of cross-site combinations with the initial word error of 12.5% in 2xRT. At LIMSI we carried out experiments comparing lattice rescoring and full decoding, using the same hypothesis for adaptation (compare E1 and E2, or E3 and E4). Redecoding consistently gives a better error rate, but is about 25% more costly in computation time. However, as can be seen in the Rover results reported in the lower part of the table, the redecode hypotheses are also found to be better for system combination.

	System	WER	xRT
A	BBN two-pass	12.5	2.0
B	LIMSI, adapt to A	10.7	2.1
R1	ROVER of A,B	10.6	4.1
C	BBN PLPdelta, adapt to R1, redecode	10.2	2.5
R2	ROVER of A,B,C	9.9	6.6
D	BBN PLPfr adapt to R2, rescore n-best C	10.2	1.5
E1	LIMSI, adapt to C, rescore B lattices	10.7	1.2
E2	LIMSI, adapt to C, redecode	10.5	1.9
E3	LIMSI, adapt to D, resccore B lattices	10.7	1.2
E4	LIMSI, adapt to D, redecode	10.5	1.9
E5	LIMSI, adapt to C, redecode, red phones	10.2	1.9
	ROVER of B,C,D	9.6	8.1
	ROVER of A,B,C,D	9.8	8.1
	ROVER of B,C,D,E1	9.8	9.3
	ROVER of B,C,D,E2	9.7	10.0
	ROVER of A,B,C,D,E1	9.7	9.3
	ROVER of A,B,C,D,E2	9.6	10.0
	ROVER of A,B,C,D,E3	9.7	9.3
	ROVER of A,B,C,D,E4	9.6	10.0
	ROVER of B,C,D,E3	9.8	9.3
	ROVER of B,C,D,E4	9.7	10.0
	ROVER of C,D,E4	9.6	10.0
	ROVER of B,C,E5	9.6	8.5

Table 8: Alternative system configuration with a 2xRT first pass.

7. CONCLUSIONS

As could be seen in the experimental results section, the development sets (Dev04f and Dev04) were good indicators of the Eval04 and progress test sets. The integrated BBN/LIMSI 10xRT English BN transcription system produced a significantly better result than any subsystem by itself. The running time stays within the alotted time limit, 10xRT. On top of the improvement achieved within each site, the cross-site adaptation and system combination provided further gain. Compared to last year's single system, there is around 30% relative reduction on the WER. Finally, we want to point out that the ROVER-based system combination is less effective when the component systems have already been used for cross-site adaptation.

References

1. J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," *IEEE ASRU Workshop*, 1997.
2. M. J. F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition," *Tech. Report CUED/F-INFENG/TR291*, Cambridge University Engineering Dept., 1997.
3. J. L. Gauvain, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, 1335-1338, Sydney, Dec. 1998.
4. J. L. Gauvain, L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.
5. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, **87**(4):1738-1752, April 1990.
6. N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, **26**(4), Dec. 1998.
7. L. Lamel, J.L.Gauvain and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer, Speech and Language*, 16(1):115-129, Jan. 2002.
8. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, **9**:171-186, 1995.
9. D. Liu and F. Kubala, "Online speaker clustering," *ICASSP'03*, April 2003.
10. L. Mangu, E. Brill and A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeeech'99*, 495-498, Budapest, Sep. 1999.
11. S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," *IEEE ASRU Workshop*, St. Thomas, Nov. 2003.
12. L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, **38**:213-230, 2002.
13. L. Nguyen and R. Schwartz, "Efficient 2-pass N-best decoder," *EuroSpeech'97*, Rhodes, Sep. 1997.
14. L. Nguyen and R. Schwartz, "Single-tree method for grammar-directed search," *ICASSP'99*, Phoenix, March 1999.
15. L. Nguyen and B. Xiang, "Light supervision in acoustic model training," *ICASSP'04*, Montreal, May 2004.
16. R. Schwartz, et al., "Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system," *ICASSP'04*, Montreal, May 2004.
17. H. Schwenk and J. L. Gauvain, "Neural Network Language Models for Conversational Speech Recognition," *ICSLP'04*, Jeju, Oct. 2004.
18. P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, (**16**(1):25-47, 2002.