THE 2004 BBN/LIMSI 20xRT ENGLISH CONVERSATIONAL TELEPHONE SPEECH SYSTEM

R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, G. Thattai, O. Kimball, R. Schwartz
BBN Technologies, 10 Moulton St., Cambridge, MA 02138

J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, F. Lefevre
LIMSI-CNRS BP133, 91403 Orsay, France

ABSTRACT

In this paper we describe the English Conversational Telephone Speech (CTS) recognition system jointly developed by BBN and LIMSI under the DARPA EARS program for the 2004 evaluation conducted by NIST. The 2004 BBN/LIMSI system achieved a word error rate (WER) of 13.5% at 18.3xRT (real-time as measured on Pentium 4 Xeon 3.4 GHz Processor) on the EARS progress test set. This translates into a 22.8% relative improvement in WER over the 2003 BBN/LIMSI EARS evaluation system, which was run without any time constraints. In addition to reporting on the system architecture and the evaluation results, we also highlight the significant improvements made at both sites.

1. INTRODUCTION

This paper reports on the English Conversational Telephone Speech (CTS) recognition system jointly developed by BBN and LIMSI under the DARPA EARS (Effective, Affordable, Re-usable, Speech-to-Text) program for the 2004 Rich Transcription evaluation (RT04) conducted by NIST. In the 2003 evaluation (RT03) there was no constraint on computation, whereas for the RT04 English CTS condition, we were required to submit a system that had an execution time of less than 20xRT (real-time). The 2004 BBN/LIMSI system uses both cross-site adaptation and system combination employing ROVER [1] to get a result that is better than either system by itself, but still stays within the allotted time of 20xRT.

In addition to presenting the system architecture and the evaluation results, we also highlight the improvements made at both sites on English CTS. For RT04, about 2000 hours of transcribed conversational speech [2, 3] were made available to the speech recognition community. We describe the large acoustic training corpus in section 2. In section 3, we give a detailed description of the system development effort at BBN and the components used in the combined system. Section 4 details the system development effort at LIMSI and the components used in the combined system. In section 5 we present the system architecture for the 20xRT BBN/LIMSI 2004 EARS system and also the results achieved on the 2004 evaluation test set.

2. LARGE CTS TRAINING CORPUS

Under the DARPA EARS program a major effort was initiated in 2002 to collect a large amount of training data for telephone conversations. Thousands of hours of speech were collected by the Linguistic Data Consortium (LDC) and the collection is now popularly called the Fisher collection [2]. BBN oversaw the quick transcription of 1750 hours of Fisher data and post-processed the resulting transcripts. This data, with 180 additional hours that were quickly transcribed by LDC, were made available to the EARS research community in the beginning of 2004. Therefore, together with the Switchboard I, Switchboard II, CallHome, and Cellular corpora, a total of 2300 hours of English conversational speech were available for acoustic training.

3. BBN SYSTEM DEVELOPMENT

3.1. BBN System Highlights

At BBN our focus was to improve our acoustic and language models by effectively utilizing the large CTS training data. In this section we highlight the improvements we have made to our system.
Compute and Storage Efficient Acoustic Training: The BBN Byblos speech recognition system uses phonetic Hidden Markov Models (HMMs), with State-Clustered-Tied Mixture (SCTM) models. The states of each phonetic model are clustered based on the quinphone context into different "codebooks" (groups of 24-64 Gaussian components). Typically we create about 10,000 codebooks, and the mixture weight distributions are clustered into about 100,000 distinct distributions. We use both within-word (non-crossword) quinphone and triphone models, as well as more detailed between-word (crossword) quinphone models. Parameters for the quinphone models are first estimated in the Maximum-Likelihood (ML) framework using the forward-backward EM algorithm with time constraints provided by "fuzzy labels" (constrained set of active states per frame) [4]. The ML models serve as an initial estimate for discriminative training using Maximum Mutual Information (MMI) [5] and Minimum Phone Error (MPE) [6] objective functions.

Research with large amounts of data requires a fast turnaround in acoustic training and also efficient storage of intermediate data. We first focused on improving the Speaker Independent (SI) ML acoustic training, which has the following major steps: cepstral feature analysis, fuzzy-label generation, state clustering for tying parameters, feature projection estimation (using LDA variants), Gaussian splitting for model initialization, and forward-backward EM training.

For better input/output (I/O) throughput we typically distribute the features and fuzzy labels to the local disks of the compute servers. Therefore, it is critical for such data to be compressed. We explored quantizing the cepstral features using an 8-bit linear scalar quantizer in each dimension. The effect of feature quantization was measured on the 2001 evaluation test set (Eval01) by training SI models on 370 hours of speech from Switchboard

I, II, and Switchboard cellular, with Perceptual Linear Predictive (PLP) [7] features. A small 0.1% WER degradation was observed for feature quantization compared to the baseline WER of 26.6%. Next, we explored reducing the size of the fuzzy labels by pruning more while generating the state alignments. We originally generated and stored fuzzy labels for both non-crossword and crossword models. We experimented with generating labels for the non-crossword models on the fly using the crossword labels. There was no loss in accuracy and together with feature quantization, the storage requirements were reduced by a factor of 3.

The first change we made to improve the compute efficiency was to parallelize the state clustering across phonemes, since each phoneme has it's own decision tree. There was no degradation in WER and the speed-up obtained was proportional to the amount of parallelism. Next, we focused on speeding up estimation of feature projections. In the past, we had adopted Heteroscedastic Discriminant Analysis (HDA) followed by a Maximum Likelihood Linear Transformation (MLLT) [8]. Recently, we incorporated ML based Heteroscedastic Linear Discriminant Analysis (HLDA) [9, 10]. HLDA results in the same performance as HDA+MLLT, but the row-iterative EM estimation is about 20 times faster than the gradient descent estimation for HDA+MLLT.

Gaussian splitting [4] for model initialization was another bottleneck. The splitting procedure starts with one Gaussian per codebook and performs successive Gaussian splits (interleaved with EM passes over the training data) in order to build the final Gaussian mixture with the desired number of components. We were already partitioning the codebooks into groups for parallel estimation on multiple CPUs. The efficiency was further improved by: increasing the number of Gaussians added at each splitting iteration from 2 to 8, computing derivatives and HLDA projections in data preparation only on the relevant segments for each group, partitioning data in a codebook into sub-spaces and splitting each sub-space independently, and changing the model configuration to use fewer codebooks and more Gaussians per codebook. Overall Gaussian splitting time was reduced by a factor of 10 without affecting accuracy, however, the model size was increased by 20%.

The final step in our ML training is to run multiple (typically 6) iterations of EM based on statistics from the forward-backward algorithm over fuzzy labels. We divide the training data into subsets and run forward-backward on each subset. Following each iteration we merge the accumulated statistics from each subset into a single model. The merging process is I/O bound and can take significant time when there are many subsets. To reduce the merging time, we decided to do away with forward-backward for the first 5 iterations. The state alignment is kept fixed and multiple EM iterations are run in parallel for each codebook. At the end, a single iteration of EM with forward-backward is performed. The new procedure speeds up the entire EM training process by 25% with a 0.2% absolute degradation in the WER.

As shown in Table 1, the elapsed time (measured on 40 Pentium 4 Xeon 2.0 GHz processors) for the modern acoustic training on 370 hours of speech is significantly less than the baseline.

Training Procedure	Total Hours	%WER
Baseline	17.7	26.6
Modern	4.2	26.7

Table 1. SI ML training elapsed time measured on 40 Pentium 4 Xeon 2.0 GHz processors. Models trained on 370 hours and WER is measured on the Eval01 test set.

We have also improved our speaker adaptive training (SAT) and lattice-based discriminative training. We reduced the time taken for SAT by a factor of 10 by using "approximate" Constrained Maximum Likelihood Linear Regression (CMLLR) [11]. Discriminative training was made to run more efficiently by on-demand localization of required lattices to the compute servers. The non-crossword quinphone (or triphone) models were trained using the phone-marked lattices from SCTM crossword models, thereby saving the compute for phone-marking. Overall, discriminative training was sped up by a factor of 2.

Improved Automatic Segmentation: Each CTS test file consists of a conversation, typically ten minutes long, between two people talking about a specific topic. The two channels are recorded separately but there is significant cross-talk. For RT03, we had developed a broad-class HMM based segmenter [12] to segment the speech on each channel for decoding. The broad-class HMMs were trained using ML estimation. For the 2004 system, we explored training the broad-class models using the MMI [5] criterion. We decoded the 2002 evaluation test set (Eval02) with the RT03 acoustic models. As shown in Table 2, the MMI segmenter is 0.2% absolute better in WER than the ML segmenter and only 0.1% worse than manual segmentation.

Segmentation	%WER
Manual (Baseline)	23.9
ML Automatic	24.2
MMI Automatic	24.0

Table 2. Comparing MMI and ML based segmentation by adapted decoding on the Eval02 test set using RT03 models.

Fisher Data in Language and Acoustic Modeling: The first experiment we did with the 1930 hours of Fisher data was to train a trigram language model (LM) with the Fisher data added to our 2003 LM training data [13]. We also added 6k new words from Fisher data to our decoding dictionary, however, increasing lexicon size from 55k words to 61k words did not have a significant effect on the out-of-vocabulary (OOV) rate. We decoded (with adaptation) the 2003 evaluation test set (Eval03) with the new trigram LM, and ML acoustic models trained on 370 hours of speech.

AM	LM	#Gaussians	%WER (Eval03)		
			Swbd	Fsh	All
Swbd	Swbd	442k	28.6	20.3	24.6
Swbd	Swbd+Fsh	442k	27.3	19.0	23.3
Swbd+Fsh	Swbd	843k	26.5	19.2	23.0
Swbd+Fsh	Swbd+Fsh	843k	24.9	17.9	21.5

Table 3. Adapted decoding results on the Eval03 test set with additional Fisher data added to both acoustic and language model.

As shown in Table 3, the WER improved on Eval03 by 1.3% absolute, but as one would expect, the relative improvement on the Fisher (Fsh) subset was better than Switchboard (Swbd). Next, we added the 1930 hours of the Fisher data to the 370 hours of acoustic training data and re-estimated the ML acoustic models. The number of Gaussians were increased to 843k from the 442k used in the 370 hours model. The new acoustic model by itself reduced the WER by 1.6% absolute and the relative improvement on Fisher and Switchboard sets was comparable. Adding Fisher data to both

acoustic and language models reduced the overall WER on Eval03 by 3.1% absolute. It is interesting to note that the Fisher collection results in similar reduction in WER for both Switchboard and Fisher test subsets.

Discriminative Training with Large Corpus: Discriminative training of HMM parameters has been shown to be significantly better than ML estimates [5]. Our RT03 system [13] included MMI trained models. This year, we trained MMI models on 2300 hours of data using unigram lattices with the ML SAT models from Table 3 serving as an initial estimate. Lattices were generated by decoding the training data with the ML SAT acoustic models and a bigram LM. We decoded the 3-hour Fisher development set (Dev04), using the 2300-hour MMI acoustic model and LM used in the last row of Table 3. As shown in Table 4, the WER with MMI models was 16.2% as compared to the 18.4% obtained with ML models.

Estimation	%WER
ML	18.4
MMI	16.2
MPE	15.7

Table 4. Summary of discriminative training on 2300 hours of acoustic data by adapted decoding on the Dev04 set.

Recently we have implemented lattice-based MPE [6] in our system. We have also adopted I-smoothing [6] using MMI prior statistics [14]. MPE models were trained with the same lattices as the ones used for MMI training. We experimented with different acoustic scale factors for MPE training and a scale factor value of $\frac{1}{15}$ was found to be optimal. The MMI models were trained with an acoustic scale factor of $\frac{1}{10}$, which could be sub-optimal. Adapted decoding with the MPE models resulted in a WER of 15.7%, i.e. a 0.5% improvement over the MMI models.

Long Span Features: In Byblos we typically use 14 cepstral features and energy as base features. Together with their first, second and third derivatives we end up with a 60 dimensional feature vector. Finally, we project the features to 46 dimensions with HLDA. Recently we have considered adding information from a wider context by concatenating *n* successive frames and then projecting the concatenated features to a lower dimensional space. We trained acoustic models on 2300 hours of data with the "long span" features and found the optimal configuration to be concatenating 15 frames and projecting the concatenated features to a 60 dimensional space using LDA followed by MLLT [15]. The SAT models were trained with a modified HLDA-SAT procedure [15], where we apply CMLLR-SAT [16] to the base cepstra and energy features, and then concatenate the transformed base features before applying the global projection down to 60 dimensions. In Table 5, we compare models trained using long span features with models trained using derivative features (both were trained with MPE). We rescored (with adaptation) lattices for the Dev04 set. The models using long span features were 0.5% absolute better than the models trained with derivatives.

Features	#Dimensions	%WER
Derivatives	46	15.4
Concatenated Frames	60	14.9

Table 5. Lattice rescoring (with adaptation) on Dev04 lattices, for comparing models trained using long span features with models trained using feature derivatives.

Held-Out MPE training: Following our long span feature exploration, we tried a novel procedure for MPE training. We split the 2300 hours of acoustic training data into two subsets of 800 hours and 1500 hours respectively. First, we estimated MMI models on the 800 hours subset using long span features and unigram lattices generated with ML SAT models. The 1500-hour subset was treated as unseen data and decoded with the 800-hour MMI model and a trigram LM (trained with 2004 LM training data, excluding the 1500-hour subset) to generate lattices. Finally, we trained MPE models with the trigram lattices generated on the 1500-hour subset using the 800-hour MMI model as an initial estimate. No smoothing was used during MPE training, therefore the model size was kept small to avoid over-fitting. In Table 6 we compare the held-out MPE training with the baseline procedure, once again rescoring lattices for Dev04. Although the WER is 0.2% worse for the held-out MPE training, the number of Gaussians is significantly smaller compared to the regular MPE procedure. Therefore, we used these models for a fast initial pass in the 2004 BBN/LIMSI system. After the evaluation, we trained a 360k Gaussian MPE model with unigram lattices on the entire 2300 hours and found that the performance was about the same as the held-out MPE.

MPE Procedure	#Gaussians	%WER
Conventional	855k	14.9
Held-Out	365k	15.1

Table 6. Lattice rescoring to compare held-out MPE training with conventional MPE training on the Dev04 set.

State-Tied Mixtures in Forward Decoding: We experimented with using a more detailed State Tied Mixture (STM) triphone model instead of Phonetic Tied Mixture (PTM) model in the forward decoding pass of our 2-pass N-best decoder [17]. In STM, all triphones of a given phoneme and state position share the same set of Gaussian components (512 on average), while the mixture weights are shared based on linguistically-guided decision tree clustering. We compared using the STM models in the forward pass instead of PTM models by performing a two pass adapted decoding followed by SCTM crossword N-best rescoring. As shown in Table 7, there was a 0.3% absolute gain for using the STM models on the Dev04 set.

Forward Pass Model	#Gaussians	%WER
PTM	25k	15.7
STM	123k	15.4

Table 7. Comparing STM forward pass with the PTM forward pass, by adapted 2-pass decoding followed by SCTM crossword N-best rescoring on the Dev04 test set.

Word Duration Modeling: Motivated by the results in [18], we implemented word duration rescoring. The vector of the component phone durations for a word (obtained by time-aligning the sequence of words) was used as a feature to train Gaussian Mixture Models (GMM) for each word in the training data. If the number of training samples for a word was below a minimum threshold, no word-specific GMM was trained and a back-off model was used instead. The back-off consists of the vector of durations from the triphone models for the word; if any of the triphones had insufficient training, the corresponding phone model was used instead. During N-best rescoring a duration score for each hypothesis in the

N-best list was computed by summing the duration log-likelihood for each word in the hypothesis. The duration score was combined with other scores such as acoustic, language etc. to reorder the list. We compare adapted decoding followed by word duration rescoring of N-best lists in Table 8 using models trained with PLP derivative features. The word duration rescoring resulted in a 0.3% absolute improvement in the WER.

N-best Rescoring	%WER
Baseline Decoding	15.4
Duration Rescoring	15.1

Table 8. Improvement obtained in WER from word duration rescoring of N-best lists on the Dev04 test set.

3.2. BBN Components in the 2004 BBN/LIMSI System

This section describes the details of the specific BBN component systems used in the 2004 BBN/LIMSI system.

Feature Extraction: The base features (14 Cepstral coefficients and normalized energy) were extracted in the same manner as in the RT03 system [13]. We used either PLP or Mel-Frequency Cepstral Coefficient (MFCC) analysis for different systems, following frequency axis scaling using Vocal Tract Length Normalization (VTLN) [19]. Mean removal and covariance normalization [4] were also applied to each conversation side. We used two methods for computing the final feature vectors. The first method was to compute the first, second, and third derivatives using least squares fits to sequences of cepstra. The second method was the long span approach described in section 3.1.

Acoustic Models: Each BBN system comprises of a set of three models: STM non-crossword triphone model, SCTM non-crossword quinphone model, and SCTM crossword quinphone model. All models used gender-independent (GI), 5-state HMMs, trained with MPE estimation on 2300 hours of acoustic training data. Models used in adaptation were estimated via SAT. The following four systems were used for decoding at various stages in the 2004 BBN/LIMSI system (main characteristics are summarized in Table 9):

PLP Long Span Held-Out MPE System (B1): This system used the long span PLP features and was trained with the Held-Out MPE estimation introduced in section 3.1. No smoothing was used in MPE training, therefore a much smaller acoustic model was trained to avoid over-fitting.

PLP Derivative MPE System (B2): This system used PLP derivative features and was trained with MPE with unigram lattices. I-smoothing was used during estimation, with an MMI prior.

PLP Long Span MPE System (B3): This system used long span PLP features like B1, but was trained with conventional MPE training as in B2. Modified HLDA-SAT procedure described in section 3.1 was used to train the SAT models.

MFCC Long Span MPE System (B4): This system is identical to system B3 except for the fact that it was trained with MFCC long span features. We did not train an STM model for this system because it used the forward pass information from another decoding in the 2004 BBN/LIMSI system.

Language Models and Recognition Lexicon: We estimated trigram LMs using modified Witten-Bell smoothing from the following data sources: 20.5M words from the Fisher acoustic training, 3.7M words from Switchboard 1, Switchboard 2, and CallHome, 530M words of web-data released by the University of Washington (UW), 141M words from Broadcast News, 47M words of archived text from CNN and PBS, and 2M words from the TDT4 database. Our LM included the most frequent bigrams and trigrams as compound words, therefore many of the trigrams in the LM were actually higher order n-grams. Two "weighted" grammars were trained where the out-of-domain text resources were weighted using a content-similarity measure. The LM used in decoding used a higher count cutoff threshold to reduce the size of the LM. For N-best rescoring, a "full" grammar with zero count cut-offs was estimated. The LM used for backward decoding consisted of 76M trigrams, whereas the rescoring LM consisted of 173M trigrams.

System	#Gaussians		
	STM	SCTM	SCTM-crossword
B1	121k	586k	365k
B2	123k	786k	843k
B3	120k	788k	855k
B4	-	686k	708k

Table 9. Number of Gaussians in BBN acoustic models used in the 2004 BBN/LIMSI system.

All BBN systems used a lexicon of 61k words (including 2500 compound words). Phonetic word pronunciations were written using a set of 49 phonemes.

Decoding Strategy: We typically used three passes for both unadapted and adapted decodings:

- a forward fast-match pass, using STM model and an approximate bigram LM
- a backward pass, using SCTM within-word quinphone and an approximate trigram LM to produce N-best lists or lattices
- an N-best rescoring pass using SCTM between-word quinphones and full trigram LM

We used several techniques such as fast Gaussian computation using shortlists, pre-computing Gaussian density values, grammar spreading, and Gaussian mean and variance quantization [20, 11] to reduce compute and memory usage during decoding.

For adapted decodings, we first estimated speaker-dependent feature projections via CMLLR with respect to the SCTM crossword quinphone model. Next, all the SAT models were adapted in the new transformed feature space using Least Squares Linear Regression with a maximum of 8 to 16 regression classes depending on the decoding stage in the overall system. In every BBN system, with the exception of B1, a three pass decoding was performed with the adapted models. Models from B1 were used in the framework developed for the 2004 BBN 1xRT system [11], which used lattice rescoring instead of N-best rescoring.

4. LIMSI SYSTEM DEVELOPMENT

4.1. LIMSI System Highlights

The LIMSI systems used for the RT02 [21] and RT03 evaluations have been significantly improved for this evaluation. Some of the main characteristics of the system are: gender-dependent VTLN, which allows us to better estimate the warping coefficients and to make use of all the available training data to train models for each

gender [21]; MAP-adapted [22] gender-dependent acoustic models from speaker-independent seed models; MLLT [8]; SAT [23]; MMI training [5], CMLLR [24]; and multiple regression class MLLR adaptation [25] with a tree organization for the adaptation classes; neural network (NN) language model [26]; two phone sets (a full 48 phone set and a reduced set of 38 phones); lattice-based decoder with Gaussian short lists for efficient decoding; consensus decoding [27] with pronunciation probabilities. Many of the above techniques are new to or have been improved in our RT04 system.

A new word list was developed and optimized on the Dev04 set. With the new word list the OOV rate was decreased from 0.15% to 0.10% on the Dev04 set.

We invested significant effort in order to be able to train acoustic models on the 2300 hours of CTS data, and needed to update our infrastructure, both at the hardware and software levels. We also spent a fair amount of effort in cleaning up the Fisher transcripts. This consisted mainly of correcting major errors (typos), misspellings (often proper place names), and whenever possible, using a consistent spelling for person names within the same call.

One of our first goals after the RT03 evaluation was to speed-up the decoding time for the LIMSI single component system without sacrificing performance. We looked at the computational costs of the various decoding steps and their contribution to the overall performance. Based on this study, we built a single component 13xRT CTS system using models trained on only the data available for the RT03 evaluation. The main changes in the decoding strategy were: speeding up the non-VTLN unadapted decoding which was used in the RT03 system primarily to compute the VTLN warping factors; using these hypotheses for MLLR acoustic model adaptation; generating word lattices using the adapted acoustic models (in the two class adapted decoding) and converting the lattices into word graphs for fast acoustic rescoring. The resulting single component system had a word error rate of 21.1%, which compared favorably to our RT03 single component system running in about 120xRT with a word error rate of 21.9%.

Improvement details	%WER red
Speaker adaptive training	0.9%
MLLT	0.8%
Improved models with Fisher data (LM, large AM, lexicon)	2.5%
Better and faster decoding with AM adaptation with factor of 6 speed-up	0.4%
Multiple phone sets modeling	0.4-0.7%
Overall relative error reduction without Rover	23%

Table 10. Summary of improvements to the LIMSI CTS component system. Absolute WER reductions on the Dev04 set and overall relative word error reduction.

Table 10 summarizes the main improvements in our CTS system from RT03. An absolute error reduction of 1.7% was due to improved acoustic modeling by incorporating SAT and MLLT. An overall improvement of about 2.5% was obtained using the Fisher data after training better (and larger) acoustic and language models, and updating the dictionary. Modifications to and incorporating acoustic model adaptation in a fast decode led to a gain of 0.4% while reducing the computation time by a factor of 6. We also experimented with using multiple phone sets in an attempt to better capture the large differences in individual speaking styles and dialectical variations in CTS. Four alternate representations were investigated, two of which make use of syllable-position dependent phone models. This work is reported in [28].

We investigated two 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 first transformed to be compatible with LIMSI vocabulary. This was done by decompounding the BBN compound words and then applying the LIMSI compounding rules (about 900) 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 the word silence and the lattices were expanded and rescored with the LIMSI 4-gram LM (keeping the BBN 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 trigram LM. The hypothesis generated using the lattice was also used to MLLR-adapt the LIMSI acoustic models. For the full search solution, the BBN hypotheses served for MLLR adaptation prior to decoding and no lattices were exchanged between sites. We found that cross-site adaptation with a fast full search was both a simpler and more efficient solution, when used with ROVER. Therefore, this is the solution adopted for the cross-site combination in the RT04 evaluation. However, lattice rescoring is used to share computing between the LIMSI components.

4.2. LIMSI Components in the 2004 BBN/LIMSI System

This section describes the details of the specific LIMSI component systems in the 2004 BBN/LIMSI system.

Feature extraction: The LIMSI front-end used 39 cepstral features derived from a Mel frequency spectrum estimated on the 0-3.8kHz band every 10ms. For each 30ms frame the Mel scale power spectrum was computed with a VTLN warped filter bank, and the cubic root was taken followed by an inverse Fourier transform. The cepstral coefficients were normalized on a conversation side basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster had a zero mean and unity variance. The 39-component acoustic feature vector consisted of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives. The VTLN warping factors were estimated by alignment of the audio segments with a word level transcription (output of system B1) for a range of warping factors (between 0.8 and 1.25), and the warping factor corresponding to the maximum likelihood was chosen. This was done using single-Gaussian gender-dependent models.

Acoustic Models: The LIMSI acoustic models used in the 2004 BBN/LIMSI system were tied-state position-dependent crossword triphones with Gaussian mixtures. The modeled contexts were automatically selected based on their frequencies in the training data. The most frequent triphone contexts (over 99%) were modeled specifically, with the remaining contexts being modeled by less specific models (right- and left-context phone models and context-independent phone models). In choosing to model right contexts over left contexts, a preference was given to modeling anticipatory co-articulation over perservatory co-articulation. The tied states were obtained by means of a decision tree with questions on the left and right phone contexts, and the phone position within the word. There were on average 32 Gaussians per tied state. Starting from the VTLN cepstrum file, the training procedure included 4

major steps: MLLT estimation, CMLLR SAT estimation for each speaker, ML training, and MMI training.

During our development process we estimated acoustic models on selected subsets of the Fisher corpus, and noticed that the breath word token did not occur in the most recent transcriptions. Given that we use a special phone symbol for breath, existing models were used to automatically add breath tokens to the transcripts. More precisely, the pronunciation dictionary was modified to allow the [silence] word to be the silence phone /./ or the breath phone /H/ during forced alignment. Since there was a tendency for false matches in noisy conditions, a word insertion penalty was used for breath along with a minimal duration requirement (150 ms). New reference transcripts were then created including the automatically detected breath. For model training data was re-segmented with [silence] having the normal /./ pronunciation. We found a slight improvement by allowing breath to be optionally pronounced as with the breath phone or as silence. All of the model sets were estimated on segmentations including the automatically detected breath tokens for the new Fisher data.

Three sets of MLLT-SAT models were trained with MMI on 2300 hours of CTS data. All models contain about 30k tied-states, but there were some differences in their estimation. Table 11 summarizes the characteristics of the three model sets.

L1 Models: For the L1 system, two sets of gender-dependent models were built after dividing the training data into the gender specific subsets, i.e. the two model sets were trained completely independently. Due to a larger proportion of the data being from female speakers, there were slightly more phone contexts in the female models (38.1k) than in the male models (37.8k). These models used 48 phones and included about 30k tied states with 32 Gaussian per state.

L2 Models: The L2 models were reduced phone set models and were trained using the same procedure as used for the L1 models. The reduced phone set differed from the original 48 phone set as follows: the affricates /C,J/ were respectively replaced by the stop-fricative sequences /tS,dZ/; the syllabic consonants /L,M,N/ were replaced by a schwa-nasal sequence /xl,xm,xn/, the diphthongs /W,Y,O/ were replaced by a vowel-glide sequence /aw,ay,cy/; the front and neutral schwas were combined together, as were the retroflex vowel and the retroflex schwa. These models included about 30k tied states for 28k phone contexts.

L3 Models: The L3 gender-dependent models were trained on all the data using a standard MAP estimation procedure from SI seed models [22]. Even though the models used to estimate the VTLN warping factors (for the training and test data) were trained separately on the female and male data, the gender-dependent models used by the recognizer were first trained on all the data, and then adapted with the gender-specific data. These models used 48 phones and included about 31k tied states for 43k phone contexts.

Models	#phones	#contexts	#tied-states
L1	48	38k	30k
L2	38	28k	30k
L3	48	43k	31k

Table 11. Characteristics of the three LIMSI acoustic model sets.

Language Models and Recognition Lexicon: The trigram and four-gram language models used by the decoder were obtained by interpolating backoff n-gram models trained on various data sets, of which the most important were the transcriptions of the CTS training data (27M words), and the transcriptions of broadcast news (370M words). The four-gram backoff LM was also interpolated with a neural network LM trained on only the transcriptions of the CTS data [26]. The data sets used to train each backoff n-gram model were the following:

- CTS transcripts with breath noise (6.1M words): 2.7M words of the swb_ldc transcriptions, 1.1M words from CTRAN transcriptions of Switchboard-II data, 230k words of cellular training data , 215k word of the CallHome corpus transcriptions, 1.7M words of Fisher data transcribed by LDC, transcripts of the evaluation data set from 1997 to 2001.
- CTS transcripts without breath noise (21.2M words): 2.9M words of swb1_isip transcriptions, 18.3M words of Fisher data transcribed by WordWave and distributed by BBN.
- BN transcriptions from LDC (years 1992-95) and from PSMedia (years 1996, 1997, and Jan-Nov 1998): 260.3M words,
- CNN transcripts from the CNN archive (01/2000-31/12/2003): 115.9M words
- the last release of 525M words of web data from the University of Washington

The interpolation coefficients were chosen in order to minimize the perplexity of a development data set containing the Fisher part of the Eval03 test set, and Dev04 (hereafter referred to as the fisher-evaldev set).

A 50k word list was selected from the same text sources (excluding the web data) so as to minimize the OOV rate on the fisher-evaldev set. The word list had an OOV rate of 0.1% on the fisher-evaldev set and 0.13% on Eval03.

In addition a neural network LM [26] was trained on all of the CTS training data transcripts (27M words). The main idea of this approach was to use a neural network to project the word indices onto a continuous space and to estimate the probabilities. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown n-grams can be expected, by these means taking better advantage of the limited amount of representative CTS LM training data. Table 12 summarizes the performance of this LM on the Dev04 set.

	4-gram backoff LM	neural LM
perplexity	48.13	45.00
WER (sys L1)	15.99%	15.51%
WER (sys L2)	14.94%	14.64%
WER (sys L3)	14.71%	14.45%

Table 12. Neural network LM results on the Dev04 set.

The pronunciation dictionary had a total of 59k phone transcriptions for the 50k words. The basic pronunciations were 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 were estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing for unobserved pronunciations. Two versions of the pronunciation lexicon were used, the one represented with the 48 phone set was used in the L1 and L3 systems, and the reduced 38 phone set was used in the L2 system.

Decoding Strategy: For each component, decoding was performed in three steps. CMLLR [24] and MLLR [25] adaptations were performed using the hypothesis of the preceding system component.

Then a word lattice was produced for each speech segment using a dynamic network decoder with a 3-gram language model. This step was a full decode for system L1 (4.8xRT) and a word graph rescoring for systems L2 and L3 (1.2xRT). Finally, the word lattice was rescored with the neural network language model and converted to a confusion network [27] using the pronunciation probabilities. This step took less than 0.05xRT including the NN LM.

MLLR adaptation in L1 system used a fixed set of 4 phonemic regression classes. MLLR adaptation in L2 and L3 systems relied on a tree organization of the tied states to create the regression classes as a function of the available data. This tree was built using a full covariance model set with one Gaussian per state. 0n average, 9 regression classes were used for model adaptation.

5. SYSTEM ARCHITECTURE AND RESULTS

The 2004 BBN/LIMSI system uses both cascaded cross-site adaptation and ROVER for combining different systems. Figure 1 shows a block diagram representation of the the joint system. If a system has a single incoming arrow, it indicates that the models were adapted to the previous result before decoding. Multiple incoming arrows into a small circle indicate that the results are combined using ROVER, producing a new hypothesis. The name of the system indicates the site ("B" signifies BBN, "L" LIMSI, and "R" ROVER), and the system number described earlier.

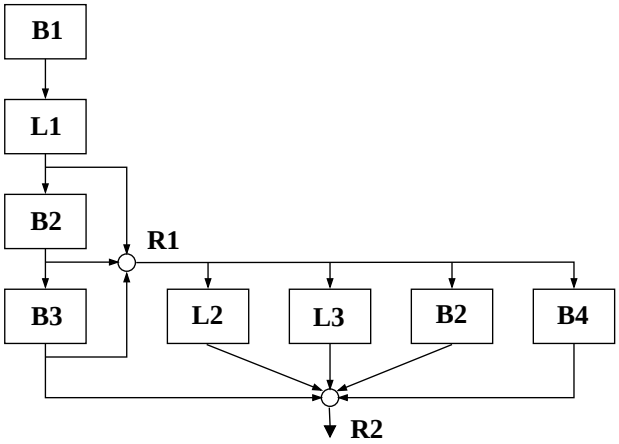

Fig. 1. 2004 BBN/LIMSI CTS Cascade/Rover System Architecture. Systems from BBN are denoted with prefix "B", those from LIMSI with prefix "L", and "R" indicates a ROVER result.

First, the waveforms are segmented using the BBN CTS segmenter described earlier. System B1 (PLP long span Held-Out MPE) is run in slightly over real-time to generate an 18.0% WER hypothesis on the Dev04 test set. The first pass, although identical in framework to the 2004 BBN 1xRT system [11], is slightly worse in WER and also slower than the 1xRT system, because we used an earlier version of that system. The 18.0% WER hypothesis is used to MLLR-adapt LIMSI's L1 (PLP GD MMI) models with 4 fixed regression classes. LIMSI decodes using the same segmentation as BBN with adapted L1 models in about 5xRT to generate lattices and a 15.5% WER hypothesis. Next, BBN adapts the B2 (PLP Derivatives MPE) models to the 15.5% WER LIMSI hypothesis using a maximum of 8 regression classes and performs a three pass decoding to generate a 14.4% WER hypothesis.

The 14.4% WER hypothesis is used to adapt BBN's B3 (PLP long span MPE) models, again using a maximum of 8 regression classes. Decoding with the adapted B3 models results in a WER of 14.2%. The L1, B2 and B3 hypotheses are combined using ROVER resulting in a WER of 13.8% (R1). Next, BBN adapts model sets B2 and B4 (MFCC long span MPE) to the ROVER R1 hypothesis, using a maximum of 16 regression classes for adaptation. BBN performs *partial* decodings with the adapted B2 and B4 models re-using the forward pass output from the B3 run. LIMSI adapts the L2 (PLP GD MMI reduced phone-set) and L3 (PLP GI-MAP MMI) models to the R1 hypothesis and rescores lattices generated from the L1 run. These lattice rescorings are denoted as L2 and L3 in Figure 1, and require slightly more than 1xRT. Finally, hypotheses from five runs: B2, B3, B4, L2, and L3 (all except B3 are adapted to the R1 hypothesis) are combined using ROVER to generate a 13.4% WER hypothesis (R2).

System	Dev04		Eval04	
	%WER	RTF	%WER	RTF
B1	18.0	1.3	21.0	1.2
B1-L1	15.5	4.8	18.3	4.6
B1-L1-B2	14.4	3.1	16.9	3.0
B1-L1-B2-B3	14.2	2.6	16.7	2.5
R1	13.8	0.0	16.2	0.0
R1-L2	14.5	1.2	16.9	1.2
R1-L3	14.6	1.2	17.1	1.3
R1-B2	14.2	2.1	16.4	2.1
R1-B4	14.0	2.1	16.3	2.0
R2	13.4	0.0	16.0	0.0
Overall	13.4	18.5	16.0	18.0

Table 13. Eval04 and Dev04 WER and RTF for each decoding stage in the 20xRT 2004 BBN/LIMSI system.

Table 13 summarizes the WER and real-time factors (RTF) for each decoding stage on both Dev04 and 2004 evaluation (Eval04) test sets. The notation in the table shows the path producing the output, thus the name of the system includes the name of the preceding system, plus the new system that was run. (For example B1-L1-B2 indicates a system that first ran B1, then adapted, then L1, then adapted, then B2.) The RTF was measured on a Pentium 4 Xeon 3.4 GHz Processor. The 2004 BBN/LIMSI system obtained a WER of 16.0% at 18.0xRT on the Eval04 test set. The WER on the EARS progress test set was 13.5% at 18.3xRT, which is 4.0% absolute lower than the WER obtained by the 2003 BBN/LIMSI system. This significant error reduction was obtained while reducing the decoding time by more than a factor of 20.

6. CONCLUSIONS AND FUTURE WORK

The 2004 BBN/LIMSI system is 22.8% relative better in WER than the 2003 BBN/LIMSI system on the EARS progress test set and also stays within the alloted time of 20xRT. The combination of cascaded cross-adaptation and ROVER was found to be effective for system design. Both sites have benefited from the large CTS acoustic training corpus and new methods were developed to use the large corpus effectively. Significant effort was also directed toward developing fast and efficient methods to train with such a large amount of data. In the future we will focus on new features and novel modeling techniques to further improve the gains from the large acoustic training corpus.

7. REFERENCES

[1] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Santa Barbara, Quebec, Canada, May 1997, IEEE, pp. 347–354.

[2] C. Cieri, D. Miller, and K. Walker, "From Switchboard to Fisher: Telephone Collection Protocols, Their Uses and Yields," in *Proceedings of Eurospeech*, Geneva, Switzerland, Sept. 2003, ISCA.

[3] O. Kimball, C.-L. Kao, T. Arvizo, J. Makhoul, and R. Iyer, "Quick Transcription and Automatic Segmentation of the Fisher Conversational Telephone Speech Corpus," in *Proceedings of Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[4] S. Matsoukas, T. Colthurst, O. Kimball, A. Solomonoff, and H. Gish, "The 2002 BBN Byblos LVCSR System," in *Proceedings of Rich Transcription Workshop*, Vienna, VA, May 2002.

[5] P. C. Woodland and D. Povey, "Large Scale Discriminative Training for Speech Recognition," in *Proceedings of ISCA ITRW ASR*, 2000.

[6] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2002.

[7] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[8] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum Likelihood Discriminant Feature Spaces," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. IEEE, June 2000, vol. 2, pp. II1 129–II1 132.

[9] N. Kumar and A. G. Andreou, "A Generalization of Linear Discriminant Analysis in Maximum Likelihood Framework," Tech. Rep. JHU-CLSP Technical Report No. 16, Johns Hopkins University, Aug. 1996.

[10] M. J. F. Gales, "Maximum Likelihood Multiple Subspace Projections for Hidden Markov Models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 37–47, Feb. 2002.

[11] S. Matsoukas, R. Prasad, B. Xiang, L. Nguyen, and R. Schwartz, "The RT04 BBN 1xRT Recognition Systems for English CTS and BN," in *Proceedings of Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[12] D. Liu and F. Kubala, "A Cross-Channel Modeling Approach for Automatic Segmentation of Conversational Telephone Speech," in *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, St. Thomas, Virgin Islands, U.S., Nov. 2003, IEEE, pp. 333–336.

[13] R. Schwartz et al., "Speech Recognition in Multiple Language and Domains: the 2003 BBN/LIMSI EARS System," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Montreal, Quebec, Canada, May 2004, IEEE, vol. III, pp. 753–756.

[14] D. Povey et al., "EARS Progress Update," in *Proceedings of EARS Speech-To-Text Workshop*, St. Thomas, Virgin Islands, U.S., Nov. 2003.

[15] B. Zhang, S. Matsoukas, J. Ma, and R. Schwartz, "Long Span Features and Minimum Phoneme Error Heteroscedastic Linear Discriminant Analysis," in *Proceedings of Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[16] S. Matsoukas and R. Schwartz, "Improved Speaker Adaptation using Speaker Dependent Feature Projections," in *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, St. Thomas, Virgin Islands, U.S., Nov. 2003, IEEE, pp. 273–278.

[17] L. Nguyen and R. Schwartz, "Efficient 2-pass N-best Decoder," in *Proceedings of Eurospeech*, Rhodes, Greece, Sept. 1997.

[18] V.R.R. Gadde, "Modeling Word Durations," in *Proceedings of International Conference on Spoken Language Processing*, Beijing, China, Oct. 2000, ISCA, vol. 1, pp. 601–604.

[19] P. Dognin, A. El-Jaroudi, and J. Billa, "Parameter Optimization for Vocal Tract Length Normalization," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, June 2000.

[20] J. Davenport and R. Schwartz and L. Nguyen, "Towards a Robust Real-time Decoder," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Mar. 1999, vol. 2, pp. 645–648.

[21] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational Telephone Speech Recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003, pp. I–212–215.

[22] J.-L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[23] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proceedings of International Conference on Spoken Language Processing*, Philadelphia, October 1996, pp. 1137–1140.

[24] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," Tech. Rep. CUED/FINFENG/TR291, Cambridge Univ., 1997, Tech. Report.

[25] C.J. Legetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[26] H. Schwenk and J.-L. Gauvain, "Neural Network Language Models for Conversational Speech Recognition," in *Proceedings of International Conference on Spoken Language Processing*, Jeju Island, October 2004.

[27] L. Mangu, E. Brill, and A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," in *ISCA Eurospeech*, Budapest, Sept. 1999, pp. 495–498.

[28] L. Lamel and J.-L. Gauvain, "Alternate Phone Models for CTS," in *Proceedings of Rich Transcription Workshop*, Palisades, NY, Nov. 2004.