The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System

R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, 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, Reusable, 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 NIST ROVER to get a result that is better than either system by itself, but still stays within the allotted time of 20xRT.

In section 2, we describe the large acoustic training corpus made available to the speech recognition community for RT04. In section 3, we describe the system development effort at BBN and the components used in the combined system. Section 4 summarizes 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 and the EARS progress test set.

2. Large CTS Training Corpus

Under the EARS program, thousands of hours of speech were collected by the Linguistic Data Consortium (LDC), and the collection is called the Fisher collection. BBN oversaw the quick transcription of 1750 hours of Fisher data and post-processed the resulting transcripts [1]. This data, with 180 additional hours transcribed by LDC, were made available to the EARS community in the beginning of 2004. Therefore, together with the Switchboard I and II, CallHome, and Cellular corpora, a total of 2300 hours of CTS data were available for acoustic training. The text sources for language modeling included: 27M words of CTS transcripts from the acoustic data, 260.3M words of Broadcast News (BN) transcriptions from LDC, 115.9M words of CNN transcripts, and 525M words of web data from the University of Washington.

3. BBN System Development

3.1. BBN System Highlights

Efficient Acoustic Modeling: The BBN Byblos 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 Gaussian components). Typically we create about 10,000 codebooks, and the mixture weight distributions are clustered into about 100,000 distributions. We use both within-word (non-crossword) quinphone and triphone models, as well as more detailed between-word (cross-word) quinphone models. Parameters for these models are first estimated in the Maximum-Likelihood (ML) framework using the forward-backward EM algorithm with time constraints provided by "fuzzy labels" (probabilistic state alignments). The ML models serve as an initial estimate for discriminative training using Maximum Mutual Information (MMI) or Minimum Phone Error (MPE) [2] objective functions.

We invested significant effort in improving the efficiency of acoustic modeling for facilitating effective research with the large CTS corpus. First, we improved the compute efficiency of speaker-independent (SI) ML acoustic modeling by a factor of 4 without any loss in accuracy [1] by: parallelizing state clustering and Gaussian splitting, adopting row iterative EM estimation in HLDA [3] instead of HDA+MLLT for feature projection, and reducing the number of forward-backward passes during the final EM training. In addition, feature quantization using 8-bit linear scalar quantizers and on-the-fly fuzzy label synthesis for non-crossword models from crossword fuzzy labels were used for reducing the storage requirements by a factor of 3. The compute time taken for speaker adaptive training (SAT) was reduced by a factor of 10 by using "approximate" [4] Constrained Maximum Likelihood Linear Regression (CMLLR).

Fisher Data in Language and Acoustic Modeling: We first trained a trigram language model (LM) with the 1930 hours of Fisher data added to our 2003 LM data. We decoded (with adaptation) the 2003 evaluation test set (Eval03) with the new trigram LM, and ML models trained on 370 hours of speech. As shown in Table 1, the WER improved on Eval03 by 1.3% abso-

[1] Elapsed time for training SI ML models on 370 hours of CTS data measured on 40 Pentium 4 Xeon 2.0 GHz processors

lute, 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 in the acoustic model(AM) were increased to 843k from the 442k used in the 370 hours model. The 2300-hour acoustic model by itself reduced the WER by 1.6% absolute. Adding Fisher data to both AM and LM reduced the overall WER on Eval03 by 3.1% absolute.

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

Table 1: Adapted decoding results on the Eval03 test set with additional Fisher data added to both AM and LM.

Discriminative Training with Large Corpus: Our RT03 system [5] used MMI models. Recently, we have implemented lattice-based MPE in our system. We trained acoustic models with both MMI and MPE criterion using unigram lattices generated by decoding the 2300 hours of training data with the 2300-hour ML SAT model. Next, we decoded the 3-hour Fisher development set (Dev04) with adaptation, using the 2300-hour MMI and MPE acoustic models. The WER with MMI models was 2.2% absolute better than ML models, and MPE resulted in another 0.5% absolute improvement over MMI.

Long Span Features: In Byblos we use 14 cepstral features and their first, second and third derivatives, resulting in a 60-dimensional feature vector. Typically we project the 60-dimensional vector to 46 dimensions using HLDA. We explored 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" [6] 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. Adapted decoding on the Dev04 test set with the long span features resulted in a 0.5% absolute improvement over the derivative features.

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. 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. Adapted decoding on the Dev04 test set with STM models in the forward pass resulted in a 0.3% absolute reduction in WER over the PTM models.

Word Duration Modeling: Motivated by the results in [7], we implemented word duration 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 for a word was obtained by computing the distance of the time-aligned frames against the phonetic duration Gaussian Mixture Models (GMM) of the component phones of the word. Finally, The duration score for each hypothesis was combined with other scores such as acoustic, language etc. to reorder the N-best list. The word duration rescoring resulted in a 0.3% absolute improvement in the WER on the Dev04 test set.

In Table 2, we summarize the significant improvements made since our RT03 system on the Dev04 test set. The 2004 BBN component system has been improved by 25% relative over the RT03 system.

Improvement details	WER red.
Fisher data in AM (with MMI) and LM	3.5%
MPE training	0.5%
Long Span Features	0.5%
STM model in forward decoding	0.3%
Word duration N-best rescoring	0.3%
Overall relative WER reduction since RT03	25%

Table 2: Summary of improvements to the BBN CTS component system. Absolute WER reductions on the Dev04 set and overall relative word error reduction since RT03.

3.2. BBN Components in the 2004 BBN/LIMSI System

Feature Extraction: The base features (14 Cepstral coefficients and normalized energy) were extracted using either Perceptual Linear Prediction (PLP) or Mel-Frequency Cepstral Coefficient (MFCC) analysis after frequency axis scaling using Vocal Tract Length Normalization (VTLN). Mean removal and covariance normalization were also applied to each conversation side. The final feature vectors were either base and derivative features reduced to 46 dimensions or the long span features.

Acoustic Models: Each BBN system comprised 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. Models used in adaptation were estimated via SAT. The following four systems were used for decoding in the combined RT04 BBN/LIMSI system:

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. In this procedure, we first trained an MMI model on 800 hours of training data using unigram lattices generated with an 800-hour SAT ML model. Next, we decoded 1500 hours of the "held-out" training corpus with the 800-hour MMI model and a trigram LM to generate lattices. Finally, MPE models were trained on 1500 hours using trigram lattices. No smoothing was used in MPE training, and a small (365k Gaussians) model was trained to avoid over-fitting.

PLP Derivative MPE System (B2): This system used PLP derivative features and was trained with MPE. The SCTM crossword quinphone model in this system had 843k Gaussians.

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. The crossword quinphone model in this system had 855k Gaussians.

MFCC Long Span MPE System (B4): This system is identical to the system B3 except for the fact that it was trained with MFCC long span features. The SCTM crossword quinphone model in this system had 708k Gaussians.

Language Models: We estimated two trigram LMs using modified Witten-Bell smoothing from the data sources described in section 2. Both LMs included the most frequent bigrams and trigrams as compound words, therefore many of the trigrams in the LM were actually higher order n-grams. 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. 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: In adapted decodings, we first estimated speaker-dependent feature projections via CMLLR and then adapted all the SAT models using Least Squares Linear Regression. With the exception of B1, a three pass decoding was performed: a fast-match forward 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, and finally an N-best rescoring pass using SCTM between-word quinphones and full trigram LM. We used techniques such as Gaussian short lists, pre-computing Gaussian density values, grammar spreading, and Gaussian mean and variance quantization [8] to reduce the compute and memory usage during decoding. Models from B1 were used in the framework developed for the RT04 BBN 1xRT system [4].

4. LIMSI System Development

4.1. LIMSI System Highlights

The LIMSI systems used for RT04 have been significantly improved since RT03. Some of the main characteristics of the system are: gender-dependent (GD) VTLN [9]; MAP-adapted GD acoustic models from SI seed models; MLLT; SAT; MMI training, CMLLR and multiple regression class MLLR adaptation with a tree organization for the adaptation classes; neural network LM [10]; multiple phone sets; lattice-based decoder with Gaussian short lists for efficient decoding; consensus decoding with pronunciation probabilities. Many of the above techniques are new to or have been improved in our RT04 system. We also 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.

One of our first goals for the RT04 evaluation was to speed-up the decoding for the LIMSI single component system. Based on a study of the computational cost at each step, we made the following changes in the decoding strategy: sped-up the non-VTLN unadapted decoding; used these hypotheses for MLLR acoustic model adaptation; generated word lattices using the adapted models and converted the lattices into word graphs for fast acoustic rescoring. The resulting single component system had a WER of 21.1% at 13xRT, which compared favorably to our RT03 single component system running in about 120xRT with a WER of 21.9%. Table 3 summarizes the main improvements in our CTS system from RT03. An absolute WER reduction of 1.7% was due to improved acoustic modeling by incorporating SAT and MLLT. An overall improvement of 2.5% was obtained using the Fisher data after training better (and larger) acoustic and language models. 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 to better capture the large differences in individual speaking styles and dialectical variations in CTS [11].

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 adapt. with factor of 6 speed-up	0.4%
Multiple phone sets modeling	0.4%-0.7%
Overall relative WER reduction since RT03	23%

Table 3: Summary of improvements to the LIMSI CTS component system. Absolute WER reductions on the Dev04 set and overall relative word error reduction since RT03.

4.2. LIMSI Components in the 2004 BBN/LIMSI System

Feature Extraction: The LIMSI front-end used 39 cepstral features (12 cepstral coefficients, log energy, along with first and second order derivatives) derived from a Mel frequency spectrum estimated on the 0-3.8kHz band every 10ms. Cepstral mean removal and variance normalization was performed on each conversations side. VTLN warps were estimated by alignment of audio segments with a word transcript (output of BBN system B1) using single-Gaussian GD models.

Acoustic Models: The LIMSI acoustic models used in the 2004 BBN/LIMSI system were tied-state position-dependent cross-word triphones with Gaussian mixtures. 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). 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. Three sets of MLLT-SAT models were trained with MMI on 2300 hours of CTS data:

L1 Models: For the L1 system, two sets of GD models were built after dividing the training data into the gender specific subsets, i.e. the two model sets were trained completely independently. 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 consisting of 38 phones [11], and were trained using the same procedure as used for the L1 models. These models included about 30k tied states for 28k phone contexts.

L3 Models: The L3 GD models were trained using a standard MAP estimation procedure to adapt ML GD SI seed models (which were trained on all the data VTLN warped to the gender) with the gender-specific data. These models used 48 phones and included about 31k tied states for 43k phone contexts.

Language Models: The trigram and four-gram language models used by the decoder were obtained by interpolating backoff n-gram models trained on data sets described in section 2. 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). In addition a neural network (NN) LM [10] was trained on all of the CTS training data transcripts (27M words). The NN LM gave an additional gain of 0.3% to 0.5% absolute depending on the acoustic model set.

A 50k word list was selected from the text sources 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. The pronunciation dictionary had a total of 59k phone transcriptions for the 50k words. Two versions of the pronunciation lexicon were used, the one represented with 48 phones 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 and MLLR adaptations were performed using the hypothesis of the preceding system. Then a word lattice was produced for each speech segment using a dynamic network decoder with a trigram LM. This step was a full decode for system L1 and a word graph rescoring for systems L2 and L3. Finally, the word lattice was rescored with the NN LM and converted to a confusion network. MLLR adaptation in L1 system used a fixed set of 4 phonemic regression classes, whereas MLLR adaptation in L2 and L3 systems relied on a tree organization of the tied states to create the regression classes (on the average 9) as a function of the available data.

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.

First, the waveforms are segmented using the BBN CTS segmenter. System B1 is run in slightly over real-time to generate an 18.0% WER hypothesis on the Dev04 test set. The 18.0% WER hypothesis is used to MLLR-adapt LIMSI's L1 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 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 models, again using a maximum of 8 regression classes. Decoding with the adapted B3 models results in a WER of 14.2%. Then 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 to the 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 and L3 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 are combined using ROVER to generate a 13.4% WER hypothesis (R2).

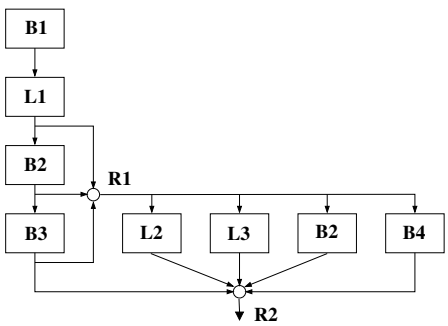

Figure 1: 2004 BBN/LIMSI CTS System Architecture.

System	Eval04		PT	
	%WER	xRT	%WER	xRT
RT03	-	-	17.5	480
RT04	16.0	18.0	13.5	18.3

Table 4: WER and RTF for the 20xRT RT04 BBN/LIMSI system on the Eval04 test set. Also comparing WER on the Progress Test (PT) with the RT03 system.

The 2004 BBN/LIMSI system obtained a WER of 16.0% at 18.0xRT (real-time measured on Pentium 4 Xeon 3.4 GHz) on the 2004 evaluation test set (Eval04). The WER on the EARS progress test set (PT) 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

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. 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] O. Kimball et al., "Quick Transcription and Automatic Segmentation of the Fisher CTS Corpus," *RT04 Workshop*, Palisades, NY, Nov. 2004.

[2] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," *ICASSP*, Orlando, FL, May 2002.

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

[4] S. Matsoukas et al., "The RT04 BBN 1xRT Recognition Systems for English CTS and BN," *RT04 Workshop*, Palisades, NY, Nov. 2004.

[5] R. Schwartz et al., "Speech Recognition in Multiple Language and Domains: The 2003 BBN/LIMSI EARS System," *ICASSP*, Montreal, QC, Canada, May 2004.

[6] B. Zhang et al., "Long Span Features and MPE HLDA," *RT04 Workshop*, Palisades, NY, Nov. 2004.

[7] V.R.R. Gadde, "Modeling Word Durations," *ICSLP*, Beijing, China, Oct. 2000, vol. 1, pp. 601–604.

[8] J. Davenport and R. Schwartz and L. Nguyen, "Towards a Robust Real-time Decoder," *ICASSP*, Mar. 1999.

[9] J.-L. Gauvain et al., "Conversational Telephone Speech Recognition," *ICASSP*, Hong Kong, Apr. 2003.

[10] H. Schwenk and J.-L. Gauvain, "Neural Network Language Models for Conversational Speech Recognition," *ICSLP*, Jeju Island, Oct. 2004.

[11] L. Lamel and J.-L. Gauvain, "Alternate Phone Models for CTS," *ICASSP*, Philadelphia, PA, Mar. 2005.