The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech

Spyros Matsoukas, Rohit Prasad, Srinivas Laxminarayan †, Bing Xiang, Long Nguyen, Richard Schwartz

BBN Technologies, 10 Moulton St. Cambridge, MA 02138
{smatsouk,rprasad}@bbn.com

Abstract

This paper describes the BBN real-time recognition systems used in the 2004 Rich Transcription (RT) benchmark test for the English Conversational Telephone Speech (CTS) and Broadcast News (BN) tasks. We describe the system architecture, along with the algorithms we used in order to reduce computation with minimal impact on recognition accuracy. Particular choices in the design of the final system are analyzed to show the trade-offs between speed and accuracy. We also present recently developed new architecture for the real-time systems, which outperforms the systems we submitted for the RT04 benchmark tests for both domains.

1. Introduction

This paper reports on the real-time systems developed under the DARPA EARS (Effective, Affordable, Re-usable, Speech-To-Text) program for the 2004 Rich Transcription (RT) evaluations. For RT04, BBN submitted Broadcast News (BN) and Conversational Telephone Speech (CTS) English systems that met the required computational limit of 10xRT and 20xRT, respectively [1, 2]. But, given that the ultimate goal under the EARS program is to develop real-time systems in the range of 5-10% WER, we decided to explore less than real-time configurations for both conditions. We submitted real-time systems for both BN and CTS in the RT04 evaluations, which were developed in a very short period of time. Recently, we have updated the real-time architecture and achieved significant improvement over the systems we submitted for the RT04 evaluations.

In section 2, we describe the recognition architecture of the real-time CTS and BN systems we submitted for the R04 benchmark tests. The details of the CTS and BN system development, along with experimental results are given in sections 3 and 4, respectively. In section 5, we present the results on the EARS progress and 2004 evaluation test sets. Section 6 describes the new system architecture we have explored since the RT04 evaluations and the improvements obtained on both domains.

2. Decoding Architecture

Both the CTS and BN real-time systems used the same recognition paradigm, performing essentially three decoding passes (forward, backward, lattice rescoring), as shown in Figure 1.

2.1. Segmentation and Feature Extraction

Automatic Segmentation: The first step was to process the waveforms to find the most likely speaker turns, and for BN, determine speaker clusters. Long speaker turns were divided into a number of short utterances by chopping on detected pauses. In the BN system, several speaker turns were clustered together to ensure sufficient amount of data for unsupervised adaptation.

VTLN: In the CTS system, segmentation was followed by Vocal Tract Length Normalization (VTLN). The optimal VTLN warp factor for each test speaker was selected based on its likelihood score against 21 warp-specific Gaussian Mixture Models (GMMs). To speed up the likelihood calculation, all 21 GMMs were evaluated at each frame, and beam pruning was applied across time, relative to the highest scoring GMM, thus restricting the number of active warps per frame. Note that the BN real-time configuration does not use VTLN.

PLP Analysis: Both CTS and BN systems used Perceptual Linear Predictive (PLP) analysis in order to extract the basic features. The CTS system analyzed at a bandwidth of 125-3750 Hz using 30 (triangular) filters, while the BN system used a bandwidth of 80-6000 Hz and 49 filters. In both cases a 25 msec Hamming window was used, with a 10 msec frame step. The final PLP frames consisted of the first 14 cepstral coefficients, along with normalized energy.

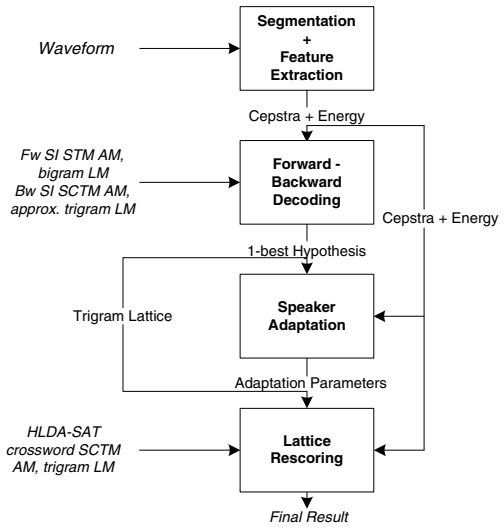

Figure 1: RT04 system architecture for real-time recognition.

2.2. Two-Pass Forward Backward Decoding

In our RT04 real-time systems, we used the same 2-pass decoder as described in [3]. The only modification made was to output a trigram word lattice instead of an N-best list.

Forward Pass: The forward pass is a fast-match pass using a composite set bigram language model (LM) along with a

† Srinivas Laxminarayan is a Ph.D. student in the Electrical and Computer Engineering Dept., Northeastern University, Boston, MA

speaker independent (SI), 5-state non-crossword composite triphone Hidden Markov Model (HMM). HMM states are tied via a linguistically-guided decision tree for each phoneme and state position. All triphones of a given phoneme and state position share the same set of Gaussian components, while the mixture weights are shared based on the decision tree clustering. We call this type of model "State Tied Mixture" (STM).

Gaussian computation during decoding is reduced significantly, by pre-computing all Gaussian density values per frame, rather than on demand. We also used short lists to avoid computing all Gaussians within a codebook state, reducing the number of evaluated Gaussians per frame by about a factor of 4.

The output of the forward pass consists of the most likely word ends per frame, along with their partial forward likelihood scores. This set of choices is used in the backward pass to restrict the search space, allowing for less expensive decoding with more detailed acoustic and language models.

Backward Pass: The backward pass is a time synchronous beam search, employing an approximate trigram LM and SI non-crossword quinphone State Clustered Tied Mixture (SCTM) HMMs. State tying in the SCTM model is determined based on a linguistically-guided decision tree. The decision tree is grown in two steps. In the first step, a high threshold on the state cluster occupancy counts is set, and the resulting state clusters determine the sharing of the Gaussian components (codebooks). In the second step, each codebook cluster is divided further by the decision tree, using a lower occupancy threshold, to determine the sharing of the mixture weights.

Gaussian computation is reduced through the use of short lists, as in the forward pass. Additional speedups were obtained for quantizing the Gaussian means and variances jointly to 8 bits, so that the Gaussian distance can be pre-computed quickly in each feature dimension for each frame, before the search. During the search, the likelihood of a particular codebook state is synthesized from the pre-computed 1-dimensional likelihood scores. The overall speedup for this technique depends on the size of the HMM used in the backward pass, but the typical speedup is about 30%. Grammar spreading [4] is used in order to make efficient use of tight pruning beams.

2.3. Speaker Adaptation

Along with the trigram lattice, the backward pass outputs the 1-best hypothesis, which is used for unsupervised speaker adaptation prior to the lattice rescoring pass. The acoustic models used in the lattice rescoring pass were trained with HLDA-SAT [5], an improved speaker adaptive training (SAT) method that is based on speaker dependent (SD) feature projections. In both BN and CTS systems, the original feature space (prior to the HLDA projection) consists of PLP frame concatenated features [6] which range from 135 to 225 dimensions per frame, so instead of estimating feature transforms in the full space, we modified the HLDA-SAT procedure to do the first adaptation on the static cepstra and energy features (15-dimensions). Once we obtained the transformed static cepstra and energy features, we applied the global projection estimated in training, and used the resulting features to estimate a single Constrained Maximum Likelihood Linear Regression (CMLLR) transform per speaker cluster. This transformation matrix was then multiplied with the global projection to provide the final transformed SD features. More details can be found in [6].

We reduced computation in CMLLR estimation by an order of magnitude, by making the assumption that all the Gaussians in the acoustic model have the same covariance matrix, set to the pooled covariance. A similar approach was used to speed up MLLR adaptation. In this case, setting each Gaussian covariance matrix to identity resulted in the optimization of the least squares criterion rather than likelihood and hence it is commonly referred to as Least Squares Linear Regression (LSLR).

Both CTS and BN real-time systems used the approximate CMLLR adaptation method described above. The CTS system also used LSLR adaptation with two regression classes on top of CMLLR. The final BN system did not use LSLR because it was found that it was adding extra computation without offering any improvement in recognition accuracy.

2.4. Lattice Rescoring Pass

The final pass in the system architecture of Figure 1 consists of a rescoring of the backward pass lattice using adapted crossword quinphone HLDA-SAT SCTM acoustic models and a trigram LM. The trigram lattices are expanded on the fly for alternate pronunciations and crossword quinphone context. The crossword expansion is optimized in order to avoid making redundant state copies. The resulting state graph is rescored using a backward Viterbi pass with tight pruning beams in order to find the best hypothesis, the system's final recognition result.

3. CTS System Development

3.1. Acoustic Models

All acoustic models were trained on a total of 2300 hours of CTS data, consisting of Switchboard-1, Callhome, Switchboard-2 cellular and Fisher training sets [1]. For the forward decoding pass we trained an SI non-crossword triphone STM model with about 121k Gaussians. This model was estimated with Minimum Phone Error (MPE) [7] on unigram lattices, using PLP frame concatenation (15-frames) and projecting to 60 dimensions via LDA+MLLT [8]. The backward pass used an SCTM non-crossword quinphone HMM with about 586k Gaussians, trained with "held-out" MPE [1] on HLDA-transformed cepstra + derivatives. The lattice rescoring pass ran with an SCTM crossword quinphone HMM, having about 365k Gaussians, also trained with "held-out" MPE. This model, however, was estimated based on 15-frame concatenated cepstra, projected to 60 dimensions through LDA+MLLT.

3.2. Language Models

The LM training data for the 2004 evaluation included the 2003 LM training data, 20.5M words from the Fisher acoustic training, and 530M words of web data released by the University of Washington (UW). The 2003 LM training data consisted of 3.7M words from in-domain data (Switchboard 1, Switchboard 2, and CallHome), and another 300M words from out-of-domain data. Out-of-domain text sources included Broadcast News (141M words), archived text sources from CNN and PBS (47M words), text from the TDT4 database (2M words), and 192M words of web data from UW. A compound word LM using modified Witten-Bell smoothing was estimated using the LM training data described above. The lexicon contained 60k words and the LM included 48M bigrams and 78M trigrams.

3.3. Experimental Results

The CTS real-time system was tuned for speed on the 2004 Fisher development test set (Dev04). The starting baseline system consisted of two recognition stages: an unadapted decod-

ing (forward/backward/rescoring) followed by adaptation of all models, then by a full adapted decoding. The baseline system used all models trained with derivative features. The WER of that system on Dev04 was 15.3%, running at 10.8xRT. As shown in Table 1, the real-time configuration resulted in a WER of 17.5% running at 0.98xRT. All Real Time Factors (RTF) are measured on an Intel Pentium 4 Xeon 3.4 GHz Linux machine using Hyper-threading, unless otherwise noted.

Configuration	%WER	RTF
Slow adapted decoding	15.3	10.8
Real-time system	17.5	0.98

Table 1: Real-time performance compared to the slower baseline on CTS Dev04 test set.

4. BN System Development

4.1. Acoustic Models

The acoustic models for the 2004 BN system were trained on 1700 hours of data. These included 140 hours of Hub-4 data that were used in 2003, while the rest is obtained using light supervision on BN data with captions [2, 9]. For the forward decoding pass we trained an SI non-crossword triphone STM model with about 117k Gaussians. The model was estimated with Maximum Mutual Information (MMI) [10] on unigram lattices, using PLP cepstra + derivatives projected to 46 dimensions via HLDA. The backward pass used an SCTM non-crossword quinphone HMM with about 778k Gaussians, trained with MMI on HLDA-transformed cepstra + derivatives. The lattice rescoring pass ran with an SCTM crossword quinphone HMM, having about 792k Gaussians, also trained with MMI. This model, however, was estimated based on 15-frame concatenated cepstra, projected to 60 dimensions through LDA+MLLT.

4.2. Language Models

The LMs 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. The lexicon contained about 64K words, of which 1945 were frequently occurring compound words. The language models for decoding contained about 33M 2-grams and 70M 3-grams.

4.3. Experimental Results

The BN real-time system was tuned for speed on the 2004 development test set (h4d04). We started from a slower (5xRT) baseline that performed recognition in two stages, a SI decoding followed by adapted (both CMLLR and LSLR adaptation were performed using the SI decoding hypothesis) decoding, using 4-gram language models in both stages. The WER of that system on h4d04 was 10.2% after tuning of the decoding weights (LM exponent, word penalty, silence penalty, etc.). It is important to note that our lattice tools did not support rescoring of 4-gram lattices at the time of the evaluation, so the baseline system incorporated a 4-gram LM via rescoring of an N-best list, generated from the trigram lattice. We found that we couldn't fit this process within the real-time constraint, so we decided to use trigram lattice rescoring instead. This led to a 0.3% absolute degradation in recognition accuracy. Furthermore, use of sub-optimal decoding weights led to another 0.2% loss. So the starting slow trigram baseline for the BN real-time system tuning was at 10.7% on h4d04.

In order to evaluate the effect of adaptation on various stages of recognition, we ran experiments using the following three adaptation paradigms:

a. Adapt (CMLLR+LSLR) based on the hypothesis from the crossword SI lattice rescoring pass.

b. Adapt using the SI backward pass 1-best hypothesis.

c. Same as (b), but no LSLR adaptation

Adapt. Method fw	bw	lat	Prun.	WER (%)	RTF
a	a	a	slow	10.7	4.58
-	a	a	slow	11.0	4.22
-	-	a	slow	11.4	3.16
-	-	b	slow	11.7	2.62
-	-	b	fast	12.7	1.29
-	-	c	fast	12.5	0.98

Table 2: Results on the 2004 BN development test set, showing the effect of sub-optimal adaptation and tighter beam pruning settings on accuracy and speed.

Table 2 shows the results of these experiments. A "-" entry under "adaptation method" indicates that no adaptation was performed in the corresponding decoding pass. We can see that removing adaptation from both forward and backward passes, while keeping slow pruning settings, results in 0.7% absolute degradation in recognition accuracy. An additional 0.3% degradation is incurred for using the less accurate hypothesis from the unadapted backwards pass in order to adapt the crossword model for lattice rescoring. Overall, the loss from sub-optimal speaker adaptation is about 1% absolute. Another 1% degradation is due to aggressive pruning during search in all decoding passes. We found that we had to prune the BN system more aggressively compared to the CTS system because the models where much larger in size and the Gaussian computation was taking a significant part of recognition. Recall that the CTS real-time system used smaller than usual acoustic models (trained with held-out MPE) in the backward and lattice passes, so they could afford less pruning during the search.

The use of large HMMs in the lattice rescoring pass also resulted in increased computation during the estimation and application of the LSLR transformations. In the BN system this computation is incurred more frequently than on CTS, because there are more speaker clusters. Pressed by time, we decided to turn off LSLR adaptation and keep only the approximate CMLLR transforms. Surprisingly, the result showed a 0.2% absolute improvement for not using LSLR. We believe that LSLR degraded accuracy due to using two regression classes per speaker cluster, thus tuning to the recognition errors of the sub-optimal backward pass hypothesis.

5. 2004 Evaluation Results

Table 3 shows the run-time performance of the English CTS and BN systems on the 2004 current test set (Eval04). The CTS system achieved a WER of 19.8% on the 2004 current test set, and 17.8% on the EARS progress set. The BN system's WER on the 2004 current and progress sets was 16.1% and 12.2%, respectively.

Decoding Stage	RTF	
	CTS	BN
Auto. Segmentation	0.097	0.039
Feature Extraction	0.009	0.004
Fw/Bw Decoding	0.510	0.634
Adaptation	0.162	0.101
Lattice Rescoring	0.168	0.206
Total	0.946	0.984

Table 3: Run-time of CTS and BN RT04 systems on the (Eval04) test set at various stages of recognition.

6. New Real-time System Architecture

Analysis in section 4.3 showed that lack of speaker adaptation in the forward and backward decoding passes causes significant degradation in the final WER. Figure 2, illustrates a new architecture we have explored incorporating speaker adaptation in the lattice creation process. A fast first pass (less than 0.1xRT) is used to generate a "rough" transcript for estimating speaker dependent feature projections via CMLLR. Next, we perform a forward decoding pass followed by a backward decoding pass to generate a trigram lattice using the transformed features. The backward pass hypothesis is used to perform another pass of adaptation. In the case of BN, we just estimate a new set of feature projections, whereas in CTS, in addition to estimating a new set of feature projections we also adapt the SCTM crossword quinphone model using LSLR. Finally, the backward pass lattice is rescored with the SCTM crossword model (adapted in the case of CTS) using the transformed features.

System	%WER	
	Dev04	Eval04
RT04 BN	12.5	16.1
New BN	11.7	15.5
RT04 CTS	17.5	19.8
New CTS	16.8	19.5

Table 4: New 1xRT architecture WER on the 2004 development (Dev04) and current (Eval04) test sets for both domains.

Table 4 summarizes the improvement in WER obtained by the new architecture over the RT04 systems, while still satisfying the compute constraint of less than 1xRT. The WER on the BN Eval04 test set improved by 0.6% absolute and on the CTS Eval04 test set by 0.3% absolute. Incorporating adaptation before the 2-pass decoding for lattice creation speeds up the decoding significantly, thereby allowing the overall compute requirement for the new systems to be still under 1xRT. We also found that the quality of the first fast pass output does not impact the final WER significantly. Therefore, slowing down the first decoding does not offer any advantage.

7. Conclusions

We have presented the development of the RT04 BBN 1xRT English CTS and BN recognition systems, highlighting the multi-pass architecture and the components used in each decoding pass. These systems achieved competitive results on the 2004 current and progress test sets, even though they were designed in a very short period of time. Since the RT04 evaluations, we have improved the real-time systems significantly by incorporating speaker adaptation in earlier stages. Future work will focus on eliminating redundant computation in each stage. In particular, we are planning to speed up the automatic segmentation process, make more efficient use of Gaussian short lists, and improve lattice rescoring by incorporating higher order language models, Gaussian quantization and forward-backward pruning.

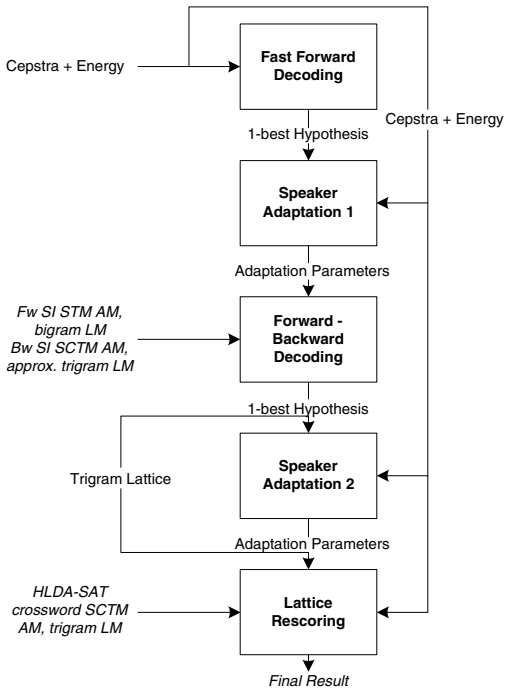

Figure 2: Post-RT04 real-time system architecture.

8. References

[1] R. Prasad et al., "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech System," *Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[2] L. Nguyen et al., "The BBN/LIMSI 10xRT BN English System," *Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[3] L. Nguyen and R. Schwartz, "Efficient 2-pass N-best Decoder," *EUROSPEECH*, Rhodes, Greece, Sept. 1997.

[4] J. Davenport et al., "Towards a Robust Real-time Decoder," *ICASSP*, Mar. 1999, vol. 2, pp. 645–648.

[5] S. Matsoukas and R. Schwartz, "Improved Speaker Adaptation Using Speaker Dependent Feature Projections," *ASRU*, Virgin Islands, U.S., Nov. 2003.

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

[7] D. Povey and P. C. Woodland, "MPE and I-smoothing for Improved Discriminative Training," *ICASSP*, 2002.

[8] G. Saon et al., "Maximum Likelihood Discriminant Feature Spaces," *ICASSP*, Istanbul, Turkey, June 2000.

[9] L. Nguyen and B. Xiang, "Light Supervision in Acoustic Model Training," *ICASSP*, Montreal, Canada, May 2004.

[10] P. C. Woodland, "Large Scale Discriminative Training for Speech Recognition," *ISCA ITRW ASR*, 2000.