The BBN RT04 English Broadcast News Transcription System

Long Nguyen, Bing Xiang, Mohamed Afify, Sherif Abdou, Spyros Matsoukas, Richard Schwartz, and John Makhoul

BBN Technologies
10 Moulton Street, Cambridge, MA, 02138, USA
{ln,bxiang,mafify,sabdou,smatsouk,schwartz,makhoul}@bbn.com

Abstract

This paper describes the BBN English Broadcast News transcription system developed for the EARS Rich Transcription 2004 (RT04) evaluation. In comparison to the BBN RT03 system, we achieved around 22% relative reduction in word error rate for all EARS BN development test sets. The use of additional acoustic training data acquired through Light Supervision based on thousands of hours of *found data* made the biggest contribution to the improvement. Better audio segmentation, through the use of an online speaker clustering algorithm and chopping speaker turns into moderately long utterances, also contributed substantially to the improvement. Other contributions, even of modest size but adding up nicely, include using discriminative training for all acoustic models, using word duration as an additional knowledge source during N-best rescoring, and using updated lexicon and language models.

1. Introduction

The RT03 Evaluation marked the first milestone of the DARPA-sponsored EARS program that was designed to achieve very challenging accuracy targets in producing rich transcription of speech from broadcast news (BN) and conversations over the telephone. Using the momentum of the development towards the RT03 Evaluation, we kept improving our BN transcription system to raise it to a higher level of a state-of-the-art transcription system.

For the BN tasks, especially in English, very large amount of speech data is available since it is fairly easy to capture the data off the air. Matching transcripts can also be captured by decoding the closed caption track encoded in the broadcast signal. This fact created opportunity for fruitful research in light supervision methods to make use of the found data having approximate transcripts. As illustrated later in this paper, we were able to make use of thousands of hours of acoustic training data to improve the transcription engine's accuracy significantly.

The development of the BBN RT04 English BN system also produced some technical achievements that, even though minimal individually, added up nicely. We had improved our audio segmentation module by using an online speaker clustering algorithm and chopping into moderately-long utterances. We had simplified and sped up our discriminative training procedure to make it feasible to train all acoustic models required in our multi-pass recognizer and that also resulted in better performance. We also modeled the duration of words and used it as an additional knowledge source when rescoring the N-best hypotheses. In addition to improvements obtained for individual systems, several sites had joined together to explore various system combination architectures to produce transcriptions much more accurate than any individual system could achieve.

The paper is organized as follows. In Section 2, we describe the composition of the BBN RT04 English broadcast news transcription system. Development and evaluation data is briefly described in Section 3. In the next section, we report the result of the selection of usable data for training the acoustic models from the thousands of hours of found data. Section 5 provides the detailed improvements obtained during the development period leading to the RT04 Evaluation system. Section 6 is dedicated to the evaluation results obtained by participating in the two integrated system combination architectures. Finally, we provide some conclusions in Section 7.

2. System Description

At the core of the BBN RT04 English broadcast news transcription system (or the RT04 system for short) is the Byblos multi-pass recognizer. Various acoustic and language models at different levels of sophistication are deployed at different passes and/or stages.

2.1. Recognizer

The Byblos multi-pass recognizer [1] first does a fast match of the data to produce scores for numerous word endings using a coarse state-tied-mixture[1] (STM) acoustic model (AM) and a bigram LM. Next, a state-clustered tied-mixture (SCTM) AM and an approximate trigram LM are used to generate N-best hypotheses. N-best hypotheses are then re-scored and re-ranked using a cross-word SCTM AM and a 4-gram LM. The top-1 of the re-ranked N-best hypotheses is the recognition result. In other words, the decoding process is a three-step sequence (fast-match, N-best generation, and N-best rescoring) with finer-detailed models being used on narrower search space at later steps [2].

The decoding process is repeated three times. First, speaker-independent (and gender-independent) acoustic models are used in the decoding to generate hypotheses for unsupervised adaptation. Then, the decoding is repeated but with speaker-adaptively-trained acoustic models that have been adapted to the hypotheses generated in the first stage. The last decoding is similar to the second but acoustic models are adapted to the second stage's hypotheses using a larger number of regression classes.

2.2. Acoustic Model Training

The typical procedure to train acoustic models at BBN can be logically grouped into these four sequential stages.

[1] Using STM is a new feature of the RT04 system. RT03 and previous systems used PTM instead. In a PTM model, all 5 states of a phoneme share a Gaussian mixture. In an STM model, each of the 5 states of a phoneme has its own Gaussian mixture.

Front-end Processing: 14-dimensional Perceptual Linear Predictive [3] cepstral coefficients are extracted from the overlapping frames of audio data with a frame rate of 10ms. Cepstral mean subtraction is applied for normalization. The normalized energy is used as the 15^{th} component. In addition, the first, second, and third derivatives of the 15 components are also used to form a 60-dimensional feature vector.

ML-SI Training: The 60-dimensional feature vectors are transformed into 46-dimensional vectors by using a global Heteroscedastic Linear Discriminant Analysis (HLDA) [4] and diagonalizing transform. The speaker-independent AMs (i.e. STM, SCTM, and cross-word SCTM) are trained using the EM algorithm. These models are to be used in the speaker-independent (SI) decoding stage.

ML-HLDA-SAT: Speaker-dependent HLDA transforms [5] are then estimated in the original 60-dimensional space to project the feature vectors into another 46-dimensional feature space. The reduced feature space is further refined by using Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation [6]. The speaker-adaptively-trained (SAT) acoustic models are then trained using the Maximum-Likelihood (ML) criterion. These models are subsequently referred to as HLDA-SAT models and to be used only in the adapted decoding stages.

MMI Training: In the last (and optional) stage of acoustic model training, all training data is decoded using the ML models to generate lattices. Then a new set of AMs are estimated using these lattices under the Maximum Mutual Information (MMI) criterion [7]. These models are subsequently referred to as MMI models. In contrast to the RT03 system, all SI and SAT acoustic models of the RT04 system are trained using the MMI criterion.

2.3. Language Model Training

The language models were estimated from a pool of text data with various weightings to emphasize relevant data sources. Typically, BN transcripts are weighted by a factor of 3-6 relative to data from other sources such as newswire or newspaper text. We used a modified Witten-Bell smoothing technique, which we measured to give similar results as the KN smoothing technique when having a substantially large amount of training data.

2.4. Audio Segmentation

BN input is typically a monolithic episode-length waveform of half or an hour long, so audio segmentation is a necessary step to break the input into manageable utterances. The audio segmentation module used in the RT04 system consists of 5 steps: bandwidth detection, gender detection, speaker change detection, speaker-turn clustering, and chopping into utterances. Input speech is first segmented into wideband and narrowband material, using a dual-band phoneme decoder. Each channel is then normalized with RASTA, and a dual-gender phoneme decoder is applied to detect gender changes and silence locations. For each channel-gender chunk, speaker change detection is performed based on the Bayesian Information Criterion and results in a segmentation that defines speaker turns, along with their gender and channel labels. The speaker turns are then clustered using an online algorithm that uses a penalized likelihood measure [8]. Finally, the speaker turns are chopped into sentence-sized utterances.

3. Development and Evaluation Data

To support the research and development for the RT04 Evaluation, there are several test sets. **Dev03** is a collection of 6 TDT4 BN shows aired in January 2001. **Eval03** consists of also 6 TDT4 BN shows aired in February 2001 that was used as the Evaluation set in the RT03 Evaluation. **Dev04** is another collection of 6 TDT4 shows aired in January 2001 with a higher level of difficulty. **Dev04f** comprises 6 shows selected by LDC from the pool of BN data captured in November 2003. **Eval04** includes 12 shows selected from the BN data aired in December 2003 that was used as the Evaluation set in the RT04 Evaluation. Both **Dev04f** and **Eval04** include broader types of broadcast speech such as talk shows.

4. Light Supervision and Found Data

In the USA, audio with matching closed captions (CC) of most BN programs in English can be captured off the air. It might not be an overstatement to consider these data as *found data*. As described in [9], we had developed an effective *light supervision* method to make use of the found data and significantly improve the recognition accuracy of our BN transcription systems.

As tabulated in Table 1, using all of the found data available to us through the EARS program, we could select 3500 hours of usable data to train our acoustic models. In this sequence of experiments to calibrate the effect of adding more data, we used only models trained within the ML framework for quick turn-around results. It is worthwhile to point out that the strength of the acoustic models seems to reach a saturation point after having 2000 hours of training data. Specifically, acoustic models trained on 3500 hours outperformed models trained on 2100 hours only by 0.1% absolute. Probably, this is a signal to call for a radical modeling technique or a different model structure when a very large amount of training data is available.

Data Set	hours	Gaussians	WER
h4+tdt4	297	354k	12.0
h4+tdt[2,4]	602	720k	11.4
h4+tdt[2,3,4]	843	741k	11.0
h4+tdt[2,3,4,4x]+BN03_r1	2130	867k	10.6
h4+tdt[2,3,4,4x]+BN03_r[1,2]	3573	1459k	10.5

Table 1: Comparison of WERs of the Dev03 test set, using different amounts of training data selected through Light Supervision from thousands of hours of *found data*

Note that 'h4' stands for the 140 hours of Hub4's carefully-transcribed acoustic training data. All the remaining data sources presented in Table 1 are found data: tdt[2-4] represent the 3 different releases of the Topic Detection and Tracking corpora (about 1400 hours), tdt4x is the extra TDT4 data (465 hours) covering the period March-July 2001, and BN03_r[12] the two releases of 7000 hours of BN data captured by LDC in 2003. This collection of found data was designated as training data for the English broadcast news tasks of the EARS program.

5. Improvements

The RT04 system has been significantly improved since the EARS RT03 Evaluation. The contributions of various techniques are listed in Table 2. Compared to the RT03 system, there was 3.0% absolute (22.4% relative) gain on the EARS 2004 development test set Dev04.

Detail of Improvement	WER
1. Baseline (RT03 trained on 200hrs)	13.4
2. 843-hour acoustic training	12.1
3. 1700-hour acoustic training	11.3
4. + MMI for all models	11.0
5. + duration modeling	10.9
6. + online speaker clustering	10.8
7. + longer utterances ($\bar{7}$sec)	10.5
8. + new lexicon and LMs	10.4

Table 2: Improvements in the RT04 system on Dev04 test set.

5.1. More Selective Acoustic Training Data

As mentioned above, we had about 3500 hours of acoustic training data selected from the thousands of hours of found data through light supervision. Given the tiny gain by using 3500 hours in contrast to using 2100 hours, we decided to use a stricter selection criterion. Recall that the original selection imposes only one condition: phrases of three or more contiguous words, that both the CC transcripts and the decoder's hypotheses agree, are selected. We now add one more condition: select only phrases that have sentinel silences; i.e. a short pause is required to be present at the beginning and the end of the phrases.

As shown in the second and third rows of Table 2, using 800 hours and then 1700 hours resulted in significant reduction in WER. In comparison to the baseline RT03 system, adding more data produced 2.1% absolute reduction (11.3% vs. 13.4%).

5.2. Discriminative Training

For the improvement obtained so far by adding more selective data, only some acoustic models were discriminatively trained using MMI. Taking advantage of the simpler and faster MMI training procedure developed after the RT03 Evaluation, all remaining acoustic models were trained using MMI. As shown in the 4^{th} row of Table 2, using MMI models everywhere produced an additional 0.3% absolute gain.

5.3. Word Duration Modeling

In the RT04 system, we used an additional knowledge source – words' duration – during N-best rescoring. Each N-best hypothesis now has an additional score evaluated as the sum of the log likelihood of the duration of the words of that hypothesis. The word's duration is modeled by a Gaussian mixture model (GMM) based on its duration feature vector represented by concatenating the durations of its phonemes. This is similar to the approach described in [11]. The durations of phonemes were calculated using the alignment of the acoustic training data. For rare or new words, their duration models are backoff models based on allophones' or even phones' duration. As shown in Row 5 of Table 2, using word duration as an additional knowledge source produced a modest 0.1% absolute reduction.

5.4. Better Audio Segmentation

Instead of using an offline speaker clustering algorithm as in the RT03 system, we now use an online speaker clustering algorithm in the RT04 system's audio segmentation module. This change resulted in a much faster execution time when clustering the speaker turns and produced a modest 0.1% gain as shown in Row 6 of Table 2. Another change that led to substantial reduction is the chopping of the speaker turns into utterances. Instead of requiring the average length of the utterances to be 4 seconds as we typically did in the past, we relaxed it to be 7 seconds and that seemed to be the optimal length. As shown in Row 7, we obtained another 0.3% absolute gain.

5.5. Lexicon and Language Models

We increased the lexicon of the RT04 system to have around 64k words instead of 61k words as used in the RT03 lexicon. As shown in Table 3, the new lexicon reduces the out-of-vocabulary (OOV) rate substantially. The language models were trained on essentially the same amount of data (roughly about 1 billion words) as used in the RT03 system. For the ngrams used during decoding, we applied aggressive cutoffs to obtain about 12M bigrams and 28M trigrams. However, we kept all 4grams observed in the training data for the N-best rescoring step. That resulted in about 730M 4grams. Despite its gigantic size, that requires 12GB of storage, we managed to develop efficient tools that loaded only relevant ngrams needed to rescore the N-best hypotheses. However, the reduction thanks to the high-coverage lexicon and the new language models is only 0.1% absolute, as shown in the last row of Table 2.

Lexicon	Size	Dev03	Eval03	Dev04	Dev04f
RT03	61k	0.39	0.79	0.59	0.83
RT04	64k	0.18	0.21	0.23	0.32

Table 3: Out-of-vocabulary rate for the RT03 and RT04 lexicons

5.6. Improvement on Other Test Sets

In addition to the detailed calibration of the improvements measured on the main EARS development test set Dev04, we also validated the gain on all other test sets. As tabulated in Table 4, the RT04 system produced around 22% relative reduction in WER for all test sets. These results show that the techniques that led to the improvements on the Dev04 test set robustly carry over to all other test sets.

System	Dev03	Eval03	Dev04	Dev04F	Eval04
RT03	11.6	11.2	13.4	(na)	(na)
RT04	9.1	8.7	10.4	15.0	14.2

Table 4: Improvement for the RT04 system in comparison to the RT03 system

6. RT04 Evaluation Results

The BBN RT04 English BN transcription system was used in the EARS RT04 Evaluation under two arrangements: together with

LIMSI's system to constitute the EARS Blue Team's submission, and together with CUED, LIMSI, and SRI to produce the SuperEARS Team's result. Under these arrangement, both cross-site adaptation and system combination, aka ROVER [12], were applied. Note that the total running time of the combined system was limited at 10 times realtime. [It's likely that papers describing other systems would be presented in this conference.]

6.1. The Blue Team's Results

The Blue Team's combined system is tightly integrated as depicted in Figure 1. Systems from BBN are denoted with prefix "B" and those from LIMSI with prexif "L". The 3-D squares enclosing the "B" and "L" prefixes represent the systems. The 2-D circles represent the output annotated with WER. The 3-D squares enclosing the plus sign represent the ROVER operation. The arrows show the flow of input and output among the various system. Specifically, on the development test set Dev04, in the first pass, the BBN system B1 generated hypotheses with an 11.0% error rate. Then, the LIMSI system L1, after adapting to the B1's hypotheses, redecoded and produced new hypotheses with 10.1% error rate. A ROVER of B1 and L1 provided hypotheses of 9.8%. The BBN system B2 then adapted to the ROVER's result and redecoded to produce a 9.9% result. Combining B1, L1, and B2 produced a new result of 9.5% error rate. This latest ROVER's result provided supervision for the second LIMSI system, L2, which, in turn, produced a result of also 9.9% error rate. The final ROVER of L1, B2, and L2 produced the final result of 9.3% error rate.

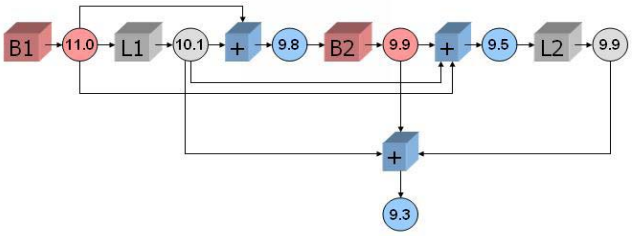

Figure 1: Architecture of the system combination deployed by the EARS Blue Team

This particular integrated architecture of cross-site adaptation and system combination is the result of experimenting with many variations – too many to report here due to space limit. However, this architecture seems to work very well for all of the data sets as listed in Table 5.

6.2. The SuperEARS Team's Results

The BBN RT04 system was also deployed in a 4-way system combination in the SuperEARS Team. It is no surprises to see that the combination of four systems produced the best results. For the Dev04, Dev04f, and Eval04 sets, the SuperEARS combined system produced 8.3%, 13.5%, and 11.6%, respectively.

7. Conclusions

We have presented a description of the BBN RT04 English broadcast news transcription developed for the EARS RT04 Evaluation. Overall, the RT04 system achieved 22% relative reduction in word error rate for most of the EARS BN development and evaluation test sets. We have shown that it was possible to select usable data from thousands of hours of *found data* by way of *Light Supervision* method to improve system's performance significantly. We also presented some of the techniques that individually provided only modest gain but added up nicely. It is worthy to point out that it was still very interesting to see that system combination still works very well as illustrated by the results of either the EARS Blue Team or the SuperEARS Team.

System	Dev04		Dev04f		Eval04	
	WER	xRT	WER	xRT	WER	xRT
B1	11.0	2.6	15.8	2.7	14.4	2.7
L1	10.1	2.7	15.1	2.9	13.6	3.0
B1+L1	9.8	5.3	14.6	5.6	13.2	5.7
B2	9.9	2.1	14.3	2.2	13.4	2.2
B1+L1+B2	9.5	7.4	14.1	7.8	12.8	7.9
L2	9.9	1.8	14.9	1.9	13.5	1.9
L1+B2+L2	9.3	9.2	13.9	9.7	12.7	9.8

Table 5: WERs and execution times of the Blue Team's integrated system on the three BN RT04 test sets

8. References

1. L. Nguyen and R. Schwartz, "Efficient 2-pass N-Best decoder," *Proc. EuroSpeech*, Rhodes, Greece, Sep. 1997, pp. 167-170.
2. L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, 38, pp. 213-230, 2002.
3. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, **87**(4):1738-1752, April 1990.
4. N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication,* **26**(4), Dec. 1998.
5. S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," *IEEE ASRU Workshop*, St. Thomas, Nov. 2003.
6. M. J. F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition," *Tech. Report CUED/F-INFENG/TR291*, Cambridge University Engineering Dept., 1997.
7. P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, (**16**(1):25-47, 2002.
8. D. Liu and F. Kubala, "Online speaker clustering," *ICASSP'04*, Montreal, May 2004, pp. 333-336.
9. L. Nguyen and B. Xiang, "Light supervision in acoustic model training," *ICASSP'04*, Montreal, May 2004.
10. R. Schwartz, et al., "Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system," *ICASSP'04*, Montreal, May 2004.
11. V. Gadde, "Modeling Word Duration," *ICSLP '00*, Bejing, Oct. 2000, pp. 601-604.
12. J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," *IEEE ASRU Workshop*, 1997.