The 1999 BBN BYBLOS 10xRT Broadcast News Transcription System

Long Nguyen, Spyros Matsoukas, Jason Davenport, Jay Billa, Rich Schwartz, John Makhoul

BBN Technologies
70 Fawcett Street
Cambridge MA 02138
ln@bbn.com

ABSTRACT

In this paper, we describe the BBN BYBLOS system used for the 1999 Hub-4E 10xRT evaluation benchmark, and discuss the improvements made to the system in 1999. We focus on the techniques that were new in this year's system to achieve an optimal tradeoff between accuracy and speed for the evaluation benchmark test. Overall, we improved the recognition accuracy on the 1998 Hub-4E evaluation test by 14% relative to our 1998 10xRT system (from 17.1% to 14.7%), or equivalently we sped up the 1998 Primary system 24 times (from 240xRT to 10xRT) while maintaining the same word error rate (14.7%). This progress was attributed to improvement in fast segmentation using dual-band and dual-gender phone-class models based on RASTA-normalized features, supervised MLLR adaptation of band-limited models to real telephone training data, adaptation between decoding passes, and various adaptation speedups.

1. INTRODUCTION

The 1999 BBN BYBLOS 10xRT broadcast news transcription system was based on both the 1998 BYBLOS Primary System [1] and the 1998 BYBLOS 10xRT system [2] with substantial algorithmic improvement as well as system change. Automatic transcription of broadcast news is a challenging speech recognition problem because of the frequent and unpredictable changes that occur in speaker, speaking style, topic, channel, and background conditions. A successful transcription system not only requires to have robust models to deal with these variability, but also needs to have an efficient segmentation strategy to break the continuous audio stream into manageable smaller segments. In contrast to the slow segmentation scheme deployed in our 1998 10xRT system, this year's system used an improved segmentation algorithm that not only took much less time but also could produce better segments which eventually resulted in lower recognition word error rate.

Faster segmentation also provided opportunity (within the 10xRT limit) to have multiple decoding stages with refined models that could lead to better recognition accuracy. Instead of having only one decoding stage in which speaker/channel adapted models could only be used once in the N-Best rescoring pass in the 1998 10xRT system, we could do two decoding stages in this year's system with fast between-pass adaptation during the first decoding stage and full adaptation in the second stage.

Similar to last year's system, we had another set of band-limited acoustic models to handle the telephone speech portion of the evaluation test set. However, these acoustic models were further refined in this year's system. After obtaining the models trained on all acoustic training data analyzed with reduced bandwidth, we applied a supervised MLLR [9] adaptation to these models using the subset of real telephone speech data.

The paper is organized as follows. Section 2 gives an overview of the system used for the 1999 10xRT Hub-4E evaluation. In section 3 we discuss the improvements made to the system since the 1998 benchmark, along with experimental results. We finish with a description of our 1999 Hub-4E evaluation results and the computational resources used during the evaluation in section 4.

2. SYSTEM DESCRIPTION

We used 200 hours (nominal - 140 hours actual) of Broadcast News training data from the 1996, 1997, and 1998 LDC releases plus 5 hours of Marketplace data from 1995. The data was partitioned by gender to create two sets of gender-dependent (GD), speaker-independent (SI) models, without regard for speech condition or signal bandwidth. Two corresponding sets of reduced bandwidth (125-3750 Hz) GD, SI models were also created using the same training data.

For each gender, we created three SI models to be used in our multiple-pass recognizer:

- PTM: 512 Gaussians per phone, within-word triphones
- SCTM NX qph: 64 Gaussians per state, 3.7K states, within-word quinphones
- SCTM XW qph: 64 Gaussians per state, 4K states, cross-word quinphones

We also created reduced bandwidth PTM, SCTM NX qph, and SCTM XW qph models for each gender. (A detailed description of the acoustic models and how each model is trained can be found in [3].)

We used a total of 600 million words to train the language model. The data were from the following sources:

- 556 million words selected from the LDC official releases *North American News Text Corpus, North American News Text Corpus (Supplement) and AP Worldstream English*, and the previous release in 1997. Data prior to 1994 were excluded. We also excluded data in the previous year's test epochs (1996/10/15 to 1996/11/15, and after 1998/02/28).
- 40 million words from in-house data previously obtained through Primary Source Media, and
- 4 million words from the LDC-released acoustic training data (weighted by a factor of 20).

The resulting language model had 13M bigrams and 43M trigrams.

The 1999 BBN BYBLOS 10xRT system was run in three stages: segmentation, first decoding stage, and second decoding stage.

1. Segmentation: We first separate the test into wide-band and narrow-band material, using a dual-band phoneme decoder [1]. Each channel is then normalized with RASTA [4], and a dual-gender phoneme decoder is applied to detect gender changes and silence locations. Within each channel-gender chunk, we perform speaker change detection [5], so we end up with an automatic segmentation that defines speaker turns, along with their gender and channel labels. Vocal Tract Length Normalization (VTLN) is then employed to select the optimal stretch factor for each speaker turn, and the test material is re-analyzed using LPC smoothing and non-causal cepstral mean subtraction. At this stage the narrow-band-labeled segments are analyzed at a reduced bandwidth (125-3750 Hz). Finally, the speaker turns are chopped into short segments (averaging 4 seconds) based on the detected silence locations.
2. First Decoding Stage: The first decoding is carried out in a sequence of three passes with fast between-pass adaptation as explained below.
   - forward PTM fastmatch [6]
   - constrained MLLR [7] unsupervised adaptation of SCTM NX qph models using the forward-pass hypotheses as transcripts
   - backward adapted SCTM within-word quinphone decoding, producing an N-best list [8]
   - constrained MLLR unsupervised adaptation of SCTM XW qph models using the top-1 hypothesis from the N-best list
   - adapted SCTM cross-word quinphone rescoring of the N-best, along with a trigram language model rescoring, to find the best hypotheses
3. Second Decoding Stage: The best hypotheses produced by the first decoding stage are then used to adapt the PTM and SCTM model means with 8 transformations. Then, a forward-backward-rescore cascade is run again with the adapted models to produce the system final recognition output.

3. RECENT IMPROVEMENTS

3.1. Fast Segmentation

In the 1998 evaluation we used an elaborate segmentation strategy that employed a GI 12-phone dual-band decoding for band detection, followed by a dual-gender word decoding for gender detection within channel turns. This procedure leads to fairly accurate band and gender detection, but is too expensive to incorporate in a real-time recognition system. Even for the 10x evaluation condition, we had to sacrifice accuracy by pruning the word decoding aggressively in order to bring the total segmentation time down to 1.9 xRT.

Besides the timing constraints, there is also the issue of what is the proper cepstral normalization method to use when the segment boundaries are not defined yet. Non-causal Cepstral Mean Subtraction (CMS) performs best when applied to pure channel/speaker segments, and introduces errors when the segments have mixed conditions. Also, dropping CMS altogether results in bad silence detection and a lot of speech deletions. For all these reasons we decided to use the RASTA method for CMS, which is very robust and does not depend on the segmentation boundaries.

To test the efficacy of RASTA phone-class models for segmentation, we performed a series of experiments on the 1997 Hub-4 evaluation test set, using SI gender-dependent models trained on approximately 150 hours of broadcast news. In those experiments we used RASTA not only for the initial test segmentation phone-class models, but for all the models used in later recognition passes [1]. The results are shown in Table 1, where we can see that the segmentation using RASTA context-independent 12-phone dual-gender model is 0.6% better than last year's 10xRT system's segmentation. The last line in Table 1 shows the WER obtained when the true channel/speaker boundaries are used for the segmentation, and when the silence detection is based on forced alignment of the reference transcripts with the best cross-word SCTM models. It is clear that the accuracy of the fast phone-class RASTA segmentation is very close to that of the unfair segmentation.

Segmentation method	xRT	WER
1998 10x system	1.9	18.8
12-phone dual-gender	0.2	18.2
unfair segmentation	N/A	17.7

Table 1: Effect of fast RASTA segmentation (band-independent)

We also trained a 22-phone dual-band dual-gender model in an attempt to detect silence, band and gender simultaneously. This increased the cost of segmentation by 0.1 xRT, but unfortunately introduced more segmentation errors, and the resulting WER was 18.6%. We tried to fix this problem by separating the channel from the gender detection, using two passes of phoneme decoding: a first pass with a 12-phone dual-band model, followed by a second pass with a 12-phone dual-gender model. By examining the output of the dual-band phoneme recognizer, we found that the channel detection was not very accurate, so we concluded that it is better to perform the channel detection on unnormalized features. Table 3 shows the results using unnormalized cepstra for band-detection:

Segmentation method	xRT	BD models	WER
12-ph dual-band + 12-ph dual-gender	0.4	no	18.3
12-ph dual-band + 12-ph dual-gender	0.4	yes	17.6

Table 2: Effect of fast band/gender detection with and without band-specific models in later recognition passes

So the segmentation procedure that works best is:

1. analyze the waveform to generate cepstra for each frame.
2. decode using the 12-phone dual-band phone-class model, in order to obtain channel change boundaries.
3. smooth out phoneme decoder output to eliminate too short channel turns.

[1] We later found that there was a small gain for using non-causal CMS on each speaker turn after the segmentation is completed, so the final BYBLOS system used non-causal CMS for the word decoding passes

4. apply RASTA normalization to the cepstra from step (1).
5. segment the RASTA input into channel turns, as specified in step (3).
6. decode each channel turn using the RASTA 12-phone dual-gender phone-class model, in order to obtain gender changes.
7. smooth out phoneme decoder output to eliminate short gender turns.
8. run fast speaker change detection within channel-gender turn, to determine speaker boundaries, and divide into speaker turns. This information can be used later for adaptation.

3.2. Fast Adaptation

Adaptation is very desirable in the design of a real-time system, because it customizes the acoustic models to each test speaker, improving both recognition accuracy and speed. However, when the acoustic model is very large, the cost of adaptation is not negligible, and can be broken down into two parts: the cost of estimating the transformation and the cost of applying the transformation. The estimation stage requires a forward pass to obtain a frame to state alignment and accumulate sufficient statistics. In the case of MLLR, n matrix accumulators are needed, where n is the size of the feature vector ($n = 45$ for our system). In order to speed up the forward pass, we used the same Fast Gaussian Computation (FGC) method that we apply during recognition. In addition, we were able to speed up the accumulation process by using a least squares criterion to estimate the transformation, instead of the usual maximum likelihood. This Least Squares Linear Regression method (LSLR) requires only one accumulator matrix, and suffers only a small degradation in accuracy, as shown in Table 3.

Adaptation Method	FGC in fw pass	xRT	WER
MLLR	no	1.5	17.0
MLLR	yes	1.0	17.0
LSLR	yes	0.6	17.1
constr. MLLR	yes	1.5	17.1
base constr. MLLR	yes	0.4	17.1

Table 3: Effect of FGC, LSLR and constrained MLLR during the estimation of the transformation. The results are on h4e97, using GD RASTA models trained on 140 hours. Timing was done on a Pentium-II 450 MHz machine.

Unfortunately, there isn't much we can do to reduce the cost of applying the transformation. The transformation matrix has to be multiplied with every Gaussian mean in the acoustic model, and this can be very costly when the model is large. It would be much faster if we applied the transformation to the features instead. This is possible with the constrained MLLR adaptation, in which a single transformation matrix is used to adapt both the mean and variance of a Gaussian. In this case, it is equivalent to estimate a transformation matrix that is applied to the observations. This speeds up the application of the transform significantly, but the estimation is very expensive [2]. We found that we can reduce the cost of estimation process by at least a factor of three, by estimating the transformation matrix based on the steady state feature parameters only. In other words, we do not accumulate statistics for the first and second derivatives of the base input features. Then, the resulting base-transform is applied to both the steady-state features and their derivatives. Interestingly, there is no degradation in accuracy from using this approximation, as demonstrated in Table 3.

The constrained MLLR results reported in this table were obtained with a single transformation matrix. Contrary to our expectations, we found no additional gain for using more than one transformations with constrained MLLR. Thus, we decided to use this method only during the first decoding stage. In the second decoding stage we used LSLR adaptation of the model means with 8 transformations.

3.3. Narrow-Band Model Adaptation

The broadcast news acoustic training data contains only a small amount of telephone speech (about 8 hours of male and 3 hours of female), so it is not enough for training gender-dependent telephone-specific models. In last year's system, we band-limited all the training data to make them sound like telephone, and trained a separate set of acoustic models. This year we extended this idea a bit further, and used the real telephone training data for adapting the band-limited models with MLLR. This gave us an extra gain on the telephone conditions (F2 and FX), as shown in Table 4. It is interesting to see that adapting the wideband models to the real telephone data is slightly better than using band-limited models without supervised adaptation.

Band-Specific	Telephone-Adapted	F2	FX	all
no	no	23.8	32.4	16.4
no	yes	21.3	30.3	15.8
yes	no	21.9	30.6	16.0
yes	yes	19.9	29.8	15.6

Table 4: Effect of band-specific models with and without supervised adaptation to real telephone training data, on h4e97. Models were trained on 140 hours.

Note that the second half of the acoustic training transcripts does not contain information about the channel, so we had to detect the channel automatically. We did this using the same 12-phone dual-band phone-class model that was used for channel detection during the test segmentation stage.

4. 1999 HUB-4E RESULTS

Table 5 shows the BBN results on the 1999 10xRT Hub-4E benchmark. [The Hub-4E evaluation test set in 1999 (h4e99) seems to be harder than that of the previous year]. We can see that the recognition accuracy of the first decoding stage (17.9%) is very close to the final system's performance, after two decoding stages (17.3%). In other words, the system can be configured to run in less than half the time (i.e. not running the second decoding stage as illustrated in Table 7) with a tradeoff of 3.4% relative accuracy degradation.

It is also interesting to see the overall improvement of the BBN BYBLOS system in 1999. Using the techniques described in the previous section, we were able to reduce the word error rate on the 1998 Hub-4E evaluation set (h4e98) by 14% relative to our 1998 10x sys-

[2] A detailed analysis of the computational complexity of this method is given in [7]

Stage	F0	F1	F2	F3	F4	F5	FX	all
1	9.4	17.6	19.1	16.1	15.8	20.6	44.3	17.9
2	9.1	16.8	18.3	15.6	15.5	19.2	43.1	17.3

Table 5: BBN results on the 1999 Hub-4 10xRT evaluation benchmark. The WER is shown for both the first and second decoding stages.

tem, demonstrated in Table 6 by focus conditions. We can also see that the 1999 10x system achieves the same accuracy as our 1998 primary system, but runs about 24 times faster.

Condition	1998 Primary	1998 10x	1999 10x
F0	9.6	10.3	9.1
F1	14.8	17.0	16.1
F2	18.6	24.9	18.9
F3	22.4	22.5	17.7
F4	21.0	16.5	14.1
F5	18.4	21.7	21.7
FX	29.5	29.7	24.2
Overall	14.8	17.1	14.7
xRT	244.0	10.0	10.0

Table 6: Comparison of the BYBLOS 1998 and 1999 systems on h4e98. Timing was performed on a Pentium-II 450 MHz machine.

Computational Resources and Timing Information The computation for this evaluation was done on Intel-based PCs with 600MHz Pentium-III CPUs, 1024MB of RAM, and 2GB of swap space. The operating system was Linux RedHat 4.1 and the compiler was GNU gcc version 2.95.1 from the Free Software Foundation. Table 7 shows timing information for the basic recognition stages of the 1999 10x system on the h4e99 test set.

Stage	xRT
Segmentation	1.1
First Decoding	3.1
Second Decoding	4.8
Total	9.0

Table 7: Timing information of the 1999 BYBLOS system on the h4e99 test set, measured on Pentium-III 600MHz PC's running Linux.

5. SUMMARY

We have described our 1999 BYBLOS 10xRT broadcast news transcription system deployed in the DARPA 1999 Hub-4E benchmark test. Compared to the previous year, we achieved a relative 14% word error rate reduction when running at the same speed (10xRT). Or equivalently, we sped up the BYBLOS transcription system by a factor of 24 while maintaining the same accuracy (14.7%). This optimal tradeoff was achieved through not only low level code optimization, but also on higher level algorithmic and system changes. We developed a faster and more accurate segmentation strategy in which only phone-class decoding is needed. Band detection is best done on un-normalized cepstra while gender and silence detection is more accurate on RASTA-normalized cepstra. We also developed a fast adaptation approach in the feature space to be used between decoding passes. Adaptation transformation matrix can be estimated only from the steady-state features but can be applied to both the steady-state features and their derivatives. Narrow-band acoustic models trained on all training data analyzed with reduced bandwidth can be refined further by applying supervised adaptation using the subset of real telephone speech as adaptation data.

Acknowledgements

This work was sponsored by the Defense Advanced Research Projects Agency and monitored by the Space and Naval Warfare Systems Command under Contract No. N66001-97-D-8501. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. S. Matsoukas, L. Nguyen, J. Davenport, J. Billa, F. Richardson, M. Siu, D. Liu, R. Schwartz, J. Makhoul, "The 1998 BBN BYBLOS Primary System Applied to English and Spanish Broadcast News Transcription," *DARPA Broadcast News Transcription Workshop*, Herndon, VA, Feb. 1999, pp. 255-260
2. J. Davenport, L. Nguyen, S. Matsoukas, R. Schwartz and J. Makhoul, "The 1998 BBN BYBLOS 10x Real Time System," *DARPA Broadcast News Transcription Workshop*, Herndon, VA, Feb. 1999, pp. 261-263.
3. L. Nguyen, T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos, Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System", *ARPA Spoken Language Systems and Technology Workshop*, Austin, TX, Jan. 1995, pp. 77-81.
4. H. Hermansky and N. Morgan, "RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing*, 2(4):578-589, Oct. 1994.
5. D. Liu, F. Kubala, "Fast Speaker Change Detection for Broadcast News Transcription and Indexing", *Eurospeech '99*, Budapest, Hungary, Sep. 99, pp. 1031-1034.
6. L. Nguyen and R. Schwartz, "Single-Tree Method for Grammar-Directed Search," *ICASSP '99*, Phoenix, AZ., Mar. 1999, pp. 613-616.
7. M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", *Technical Report 291*, Cambridge University, England, May 1997.
8. L. Nguyen and R. Schwartz, "Efficient 2-Pass N-Best Decoder", *EuroSpeech '97*, Rhodes, Greece, Sep. 1997, pp. 167-170.
9. C. J. Leggetter, P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 110-115.