Progress in the CU-HTK Broadcast News Transcription System

M.J.F. Gales, D.Y. Kim, P.C. Woodland, H.Y. Chan, D. Mrva, R. Sinha and S.E. Tranter

***Abstract*— Broadcast News (BN) transcription has been a challenging research area for many years. In the last couple of years the availability of large amounts of roughly transcribed acoustic training data and advanced model training techniques has offered the opportunity to greatly reduce the error rate on this task. This paper describes the design and performance of BN transcription systems which make use of these developments. First the effects of using lightly-supervised training data and advanced acoustic modelling techniques are discussed. The design of a real-time broadcast news recognition system is then detailed using these new models. As system combination has been found to yield large gains in performance, a range of frameworks that allow multiple recognition outputs to be combined are next described. These include the use of multiple types of acoustic models and multiple segmentations. As a contrast a system developed by multiple sites allowing cross-site combination, the "SuperEARS" system, is also described. The various models and recognition configurations are evaluated using several recent BN development and evaluation test sets. These new BN transcription systems can give gains of over 25% relative to the CU-HTK 2003 BN system.**

***Index Terms*— Automatic speech recognition, Broadcast News transcription, diarisation.**

EDICS Category: SPE-GASR

I. Introduction

THE accurate automatic transcription of broadcast material remains a challenging problem. Broadcast News (BN) transcription is difficult as a range of acoustic conditions and speaking styles must be considered. Over recent years, the performance of BN transcription systems has gradually improved to the stage where, on some "simple" test sets, average word error rates (WERs) of less than 10% can be achieved. In the last couple of years there has been an effort to further dramatically reduce the error on BN transcriptions funded by the DARPA Effective Affordable Reusable Speech-to-text (EARS) programme. As part of this programme large amounts of additional training data were made available for acoustic model training. This paper details the progress made at Cambridge University (CU) in improving BN transcription making use of this additional lightly supervised training data and state-of-the-art modelling techniques. Rather than concentrating on the detailed technical aspects of the techniques, which have previously been published in a range of papers (see for example [2], [3], [4], [5], [6]), this paper gives an overview of the approaches that were examined and how they affected performance. In particular, it describes ways in which additional training data can be used and possible frameworks for system combination. The starting point for this paper is considered to be the CU-HTK 2003 BN system [3].

For BN transcription, the first stage in any recognition process is to segment the audio stream into homogeneous blocks, i.e. blocks associated with the same speaker and possibly acoustic environment. These blocks are then clustered together to, for example, give all the data associated with a particular speaker. This task, *diarisation* [7], has been evaluated as a separate problem within the EARS programme [8], but is also essential for adaptation and normalisation in the speech-to-text (STT) task. This paper considers a number of segmentations and clusterings developed both at CU [9] and at other sites, BBN [10] and LIMSI [11]. In section II these segmentations and clusterings are characterised in terms of the average segment length produced, number of clusters and the diarisation error rates [7]. The performance of individual systems for speech recognition and their potential as sources of diversity for system combination is discussed in section VI.

Section III describes the acoustic model development. Technical advances in acoustic model building, in particular those involved with discriminative training, such as discriminative MAP [4] and modified I-smoothing [12], should improve system performance compared to the models used in the CU-HTK 2003 BN system. Performance gains from these more advanced modelling techniques are briefly described. In addition, significant gains should be possible by making use of thousands of hours of data using "lightly-supervised" techniques. In contrast to the relatively small amounts of carefully transcribed data used to build the CU-HTK 2003 BN system, only rough transcriptions, such as closed-captions, are available for this data. These transcriptions are known to be error-full and this must be taken into account during the training procedure. A range of approaches have been proposed to handle this problem [13], [5], [14]. Motivated by the work in [15], the approach adopted in this paper is to use a "biased" language model during the recognition of the training data to generate a set of transcriptions, which are then directly used for training. Section III-A describes the available acoustic model training sets and the effects that they have on the system performance are given in section III-

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. The authors would like to thank LIMSI and BBN for the use of their segmentation and clustering outputs in these experiments. The authors also thank BBN, LIMSI and SRI for their involvement in developing the SuperEARS system. The full list of people who contributed to the design and implementation of the SuperEARS system is given in [1]. The authors would like to thank all the members of the HTK EARS project team, in particular Gunnar Evermann, Xunying Liu, Khe Chai Sim, Lan Wang and Kai Yu, for their contributions to the development of the systems described in this paper.

All the authors are at the Cambridge University Engineering Dept., Trumpington Street, Cambridge, CB2 1PZ. e-mail: {mjfg, dyk21, pcw, hyc27, dm312, rs460, sej28}@eng.cam.ac.uk, phone: +44 1223 332733, fax +44 1223 332662

C. Additional training data, both audio and text, were also available for language model construction. The effect of this additional data is assessed both in terms of perplexity and WERs in section IV.

The paper then considers various forms of evaluation framework. Two styles of system with different constraints are considered. The first is based on real-time transcription. Though this limits the forms of model that can be used, it is interesting to describe the type of system that can be run in real-time. A two-stage strategy is adopted. The first, very rapid transcription stage, is used to supply an adaptation hypothesis for the second, more precise, recognition stage. The output from this second stage is a set of lattices, to which a more complex 4-gram language model is applied. This will be referred to as a P1-P2 decoding framework in this paper. This form of system is further discussed in section V.

If the real-time constraint is relaxed, then more complex decoding frameworks may be used. In recent years there has been significant interest in combining multiple acoustic model hypotheses together to reduce the error rate. This combination is usually performed using ROVER [16] or confusion network combination [17]. This form of system combination is possible in the $10\times$ real-time constraints considered in this paper (and for the RT04f evaluations). An interesting aspect of such frameworks is how to get diversity into the system so that when the recognition outputs are combined there is a reduction in the WER, i.e. errors in one set of hypotheses do not occur in the other sets of hypotheses. A range of approaches to incorporating diversity are considered. The first, based on the CU-HTK 2003 BN framework [3], is to use multiple types of acoustic models to rescore lattices generated by a P1-P2 style system. The hypotheses from these multiple branches are then combined. Two other forms of framework are also described. One is based on acoustic model diversity in the P1-P2 stage. This yields differences in the 1-best adaptation supervision and lattices for subsequent rescoring. The second is based on diversity in the segmentation and clustering. These frameworks are discussed in section VI. As a further contrast, section VII describes a cross-site system combination framework built in collaboration with BBN, LIMSI and SRI [1]. This framework makes use of systems developed at BBN and LIMSI [10], and SRI [18], in addition to systems built at CU. Results from [1] are quoted to show the effects of this cross-site diversity on performance.

II. Segmentation and Clustering

For Broadcast News transcription, the first stage of processing is to partition the incoming audio data stream into homogeneous segments (the segmentation) and to group these segments into homogenous clusters which can then be used for unsupervised acoustic model adaptation. This is generally done in two separate stages, although it is possible to have more integrated schemes which alter the segmentation during the clustering process [19], [20]. In the segmentation and clustering procedure described in [9], referred to as `CU1`, the basic stages are: removal of music and long periods of silence; initial over-segmentation of the data by detection of acoustic 'change-points'; agglomerative clustering using the likelihood ratio with a penalised likelihood (BIC) stopping criterion; and a final additional gender-dependent clustering based on speaker identification (SID) techniques [20].

When performing segmentation and clustering it is important to consider the task being addressed. If the task is to label "who spoke when", referred to as *diarisation* [7], then the diarisation error rate (DER) is commonly used to measure performance. This is the time-weighted sum of the missed speech (MS), false alarm (FA) and speaker error rates. The DER is very sensitive to splitting the data from one frequently occurring speaker into two clusters. In contrast if the segmentation and clustering is to be used for a STT task, then the final WER is of interest. For STT systems the degradation in WER from splitting data from a single speaker into multiple separate clusters is minimal (provided that clusters are homogeneous and sufficiently large for robust acoustic model adaptation). Furthermore, for speech recognition it may be preferable to split data from the same speaker into different clusters if there are multiple acoustic environments present. Not surprisingly there is little correlation between DER and WER [21].

For the experiments presented in this paper a range of segmentations and clusterings were considered:

LIMSI [22], [11][1]: This was the segmentation and clustering used by LIMSI for the NIST RT04f STT evaluations.
BBN [23][2]: This was the segmentation and clustering used by BBN for the NIST RT04f STT evaluations.
CU [3], [21]: This was the segmentation and clustering used for the CU-HTK 2003 BN evaluation system, and for the baseline acoustic model development results in section III. The segmentation and clustering was different to the other CU systems discussed below. A gender-dependent top-down clustering scheme with an arithmetic harmonic sphericity distance metric and occupancy-based stopping criterion was used. No change-point detection or SID clustering were implemented.
CU1 [9]: This segmentation and clustering was tuned to minimise the DER and used both the BIC and SID clustering stages. It ran significantly slower than the other CU schemes, meaning results based on this segmentation/clustering did not satisfy the time constraints in section VI.
CU2: This segmentation and clustering was taken from the output of the BIC clustering stage of a `CU1`-style system. An increased penalty term was used to reduce the number of clusters to compensate for the omission of the final SID clustering stage. In addition a minimum cluster size was enforced to ensure that there was sufficient data for adaptation. Parameters for this system were tuned for STT performance.

Though the primary interest of this paper is STT, it is

[1]The segmentation and clustering used for these experiments were designed for speech recognition. The `LIMSI` system for diarisation [20], different to that used for STT, achieved a DER of 8.5%, similar to that of the `CU1` segmentation and clustering [9].

[2]The segmentation and clustering used for these experiments were not optimised for diarisation performance.

Segmentation / Clustering	#Segments	#Clusters	Avg. Seg. Length (sec)
LIMSI	1284	313	13.61
BBN	2963	273	5.98
CU	1712	273	10.23
CU1	1316	401	12.69
CU2	1173	558	14.23

TABLE I

THE NUMBER OF SEGMENTS, CLUSTERS AND THE AVERAGE SEGMENT LENGTH FOR LIMSI, BBN AND CU SEGMENTATION FOR eval04.

useful to characterise each of the segmentation and clustering schemes, as this gives an indication of the diversity of the segmentations and clusterings being used. Table I shows the number of segments, clusters and average segment lengths for each of the segmentation/clustering schemes on the eval04 test set (described in detail in section III-A). It is interesting to note that, though both were tuned for minimising WER, the average segment lengths for the BBN and LIMSI were very different. The various CU approaches varied in both the number of segments and the number of "speaker" clusters.

Segmentation/ Clustering	DER (%)		[MS(%)/FA(%)]
	Auto	Ideal	
LIMSI	38.63	3.76	[0.2/1.8]
BBN	24.67	4.10	[0.2/2.3]
CU	58.15	5.61	[0.1/2.8]
CU1	8.58	2.43	[0.3/1.1]
CU2	31.15	2.82	[0.3/1.1]

TABLE II

AUTOMATIC (AUTO) AND "IDEAL" DIARISATION ERROR RATES (DERS) FOR POSSIBLE SEGMENTATION SEGMENTATION AND CLUSTERINGS ON eval04

The DER and percentages of missed and false alarm speech on the eval04 test set are given in Table II. Other than the CU1 segmentation/clustering, all the schemes were designed for WER minimisation. This is reflected in the higher automatic DER for those schemes compared to CU1. In addition to the DER obtained using the automatic clustering schemes, an ideal clustering DER score (Ideal) is also shown. This is the lowest possible DER score for a given segmentation, and gives an indication of the homogeneity of the segments.[3]

The CU scheme has the poorest ideal DER and the CU1 scheme the best ideal DER. If only the speaker error component of the ideal DER is considered (missed speech and false alarm errors are ignored), the LIMSI, BBN and CU2 schemes all perform about the same. As expected the CU1 scheme performs considerably better than the CU scheme. All these numbers are not expected to correlate directly with the speech recognition performance, but they give some indication of the diversity of the segmentation and clustering schemes being considered.

[3]It is hard to directly compare these numbers to one another as over-segmenting the data, relying on the clustering to correctly group segments together, can bias results.

III. ACOUSTIC MODEL DEVELOPMENT

This section describes the data sources and models built during the development of the BN system.

A. Training and Test Data Sets

Earlier work on English broadcast news transcription has relied on the use of acoustic model training data released before 1998 by the LDC, which is known as the Hub4 acoustic training data. There is a total of 144 hours of transcribed Hub4 acoustic training data[4] for which the LDC supplied careful manual annotations. This data was used for acoustic model training in the older CU broadcast news evaluation systems [24] and the more recent CU-HTK 2003 BN system[3]. However, for the system developed for the RT04 evaluation [25] a range of additional broadcast data sources with only closed-caption type transcriptions (of varying quality) were also potentially available for acoustic model training. These new sources consist of two major groups: data prepared originally for the various phases of the Topic Detection and Tracking task (TDT data), and data that the LDC collected in 2003 (BN03 data) for the EARS programme.

The TDT data consists of several phases. The first set of TDT data used in the work reported in this paper is from TDT4 which includes six different broadcast sources (both radio and television) and covers the period October 2000 until January 2001. This contains 235 hours of usable audio. In addition further TDT4 data, from just the four television sources and covering the period March-July 2001, was also made available by the LDC. This second portion of TDT data, which is denoted as TDT4a, contains 375 hours of audio. Experiments were also conducted using the older TDT2 data (broadcast between January and June 1998) which contains about 420 hours of usable data.

In addition to the TDT data, the LDC supplied to the EARS programme participants about 7080 hours of raw BN data collected during March-November 2003, the BN03 data. To help other sites to make use of this large quantity of data, BBN made automatic transcriptions available which were generated using a lightly-supervised recognition/filtering approach. For further details of the method used see [14]. From the BN03 data, three subsets were selected for addition to the acoustic model training pool. The first two subsets each contained about 300 hours of audio and the third around 440 hours. The BN03 data contained a total of 19 separate broadcast sources, some of which were felt to be more applicable to the task being considered than others. The data for the first set was sampled from six major sources: ABC, CNBC, CNN, CNNHL, CSPAN and PBS. The second set comes from CNN and six other sources (CBS, FOX, MSN, MSNBC, NBC, NWI) which were not included in the first set. Finally the third selection was made from the same sources as the first and second along with one new source, WBN.

[4]All training data quantities given in this paper refer to the quantities of audio actually used for training, after removal of commercials, music, and other non-transcribed material.

Training set	Description	Hours
bntr-144h	Hub4 training data	144
bntr-375h	+ TDT4	375
bntr-750h	+ TDT4a	752
bntr-1050h	+ 1st selection of BN03	1050
bntr-1350h	+ 2nd selection of BN03	1350
bntr-1790h	+ 3rd selection of BN03	1790
bntr-2210h	+ TDT2	2210

TABLE III
SELECTED ENGLISH BN ACOUSTIC TRAINING DATA SETS AND SIZES.

Table III lists the training data subsets used in this work[5]. The order of adding the various sources was determined by running preliminary recognition experiments to determine how "close", in terms of WER, each block of data was to the development data available.

Test set	# Shows	Hours	Period
dev03	6	3	Jan. 2001
eval03	6	3	Feb. 2001
dev04	6	3	Jan. 2001
dev04f	6	3	Nov. 2003
eval04	12	6	Dec. 2003

TABLE IV
SELECTED ENGLISH BN TEST SET SIZES

In order to assess the performance of the various systems developed, a range of development and evaluation data sets were used. The size and epoch of each of these blocks of data are shown in Table IV. eval04 was treated as the evaluation data. All the other sets were treated as development data and were used to help tune model parameters and language model interpolation weights. The dev04f development data is rather different in nature to the other development sets as it contains data from a different set of broadcast sources and typically includes more challenging data with high levels of background noise/music and non-native speakers. The eval04 set consists of two halves: one of which is broadly similar to previous evaluation sets and the 2004 development sets (eval03 and dev04), and one which is more similar to the data found in the dev04f set.

B. Lightly Supervised Training

Detailed transcriptions of the audio data were only available for the bntr-144h training data. All the other data, i.e. the TDT data and the 2003 BN data collection, had only closed-captions (CCs), or similar rough transcripts. These transcriptions are known to be error-full and thus not appropriate for direct use when training detailed acoustic models. To overcome this problem there has been a range of work on "lightly-supervised" training techniques [13], [5], [14]. The procedure used in this work consists of the following general stages.

1) Construct a language model (LM) using only the CC data. This CC LM was interpolated with a general BN LM using interpolation weights heavily weighted towards the CC LM. For the work presented here the interpolation weights were 0.9 and 0.1 for the general BN LM[6]. This yielded a "biased" language model.
2) Recognise the audio data using an existing acoustic model and the biased LM trained in (1). For this work the P1-P2 stages of the evaluation system (including 4-gram expansion) in section VI-A was used. This ran in approximately 5 times real-time ($5\times$RT)[7].
3) Optionally post-process the data. For example only use segments from the training data where the recognition output from (2) is consistent, to some level, with the CCs, or only use segments with high confidence in the recognition output.
4) Use the selected segments for acoustic model training with the hypothesised transcriptions from (2).

A range of options were investigated for post-processing the data, including propagating the confidence scores into the discriminative training stage [15]. However none of these were found to yield significant gains over using all the data so no post processing was performed in the work reported in this paper.

C. Model Training and Evaluation

For all the acoustic models developed in this paper, the same front-end processing as the CU-HTK 2003 BN system [3] was used. Each frame of speech was represented by 13 PLP coefficients based on a linear prediction order of 12 with first, second and third derivatives appended and then projected down to 39 dimensions using HLDA [26] optimised using the efficient iterative approach described in [27]. For initial models where HLDA was not used, the front-end consisted of 13 PLP coefficients with first and second derivatives.

All models were built using the HTK toolkit [28]. State-clustered cross word triphone models [29] were constructed with a total of about 7000 distinct states for the smaller training sets, or 9000 states, for bntr-1050h and larger. Gaussian mixture models with an average of 16 (bntr-375h, bntr-750h), or 32 (bntr-750h and larger) components were used. The distribution of components over the states was determined based on the state occupancy count, referred to in this paper as the *Varmix* process. Models were built initially using maximum likelihood (ML) training and then discriminatively trained using minimum phone error (MPE) training [2], [30]. As some BN data, for example telephone interviews, are transmitted over bandwidth-limited channels, both wide-band and narrow-band spectral analysis variants of each model set were trained.

For all the experiments in this section the segmentation (denoted CU in section II) and tri-gram LM from the CU-HTK 2003 BN evaluation system [3] was used. Possible

[5]No data between 16th January 2001 to the end of February 2001, nor any later than 14th November 2003, was included in training to avoid any time epoch overlap with development test and evaluation test data sets and to respect the epoch restrictions for the RT04f evaluation.

[6]These interpolation weights were chosen to minimise the perplexity of accurately transcribed data from the same source. They were found to be relatively insensitive to the type of source data.

[7]For this work $5\times$RT means that the system ran $5\times$ *slower* than real-time.

alternative segmentations and language models are discussed in sections II and IV respectively.

Training set	Acoustic Training	%WER	
		dev03	eval03
bntr-144h	ML	19.7	–
	+HLDA	17.9	–
	+Varmix	17.8	16.0
	MPE (ML prior)	15.2	13.7
bntr-375h	ML	19.1	17.2
	+HLDA	16.8	15.1
	+Varmix	16.7	14.8
	MPE (ML prior)	13.9	12.6
	MPE (MMI prior)	13.6	12.5
	+GD (GI prior)	13.5	12.3

TABLE V

%WERS FOR dev03 & eval03 WITH THE ML AND MPE TRAINED ACOUSTIC MODELS. SINGLE PASS DECODING WITH THE RT03 TRIGRAM LM.

Table V shows the performance without any adaptation of the Gender-Independent (GI) models built for the CU-HTK 2003 BN system [3], which are considered the baseline models for this paper. These were built using the bntr-144h training set. As previously observed the use of HLDA and discriminative training gave large gains over the baseline system. The final performance of these models on the eval03 test data was 13.7% using MPE training and an ML-prior for I-smoothing as described in [2].

To assess the relative gains of each of the stages using additional lightly supervised training data, a system with approximately the same number of states (around 7000) and Gaussian components per state (16), was built with the larger bntr-375h training data. As expected similar gains from each stage were observed. It is interesting that the performance gains from using MPE training, about 2-3% absolute, were maintained even though lightly-supervised training was required for the additional TDT4 training data. Overall the additional training data gave gains of over 1.0% absolute reduction in WER compared to the bntr-144h trained system. Recently the use of an MMI-prior for I-smoothing, rather than the ML-prior, has been proposed [12]. Using this more complex I-smoothing process gave a small improvement in performance and was therefore used in all subsequent acoustic model generation.

In addition to training GI models for BN systems the use of Gender Dependent (GD) models has been found to be advantageous [24]. Standard MAP training [31] is not appropriate for adapting discriminatively trained models. A modified, discriminative, version, MPE-MAP [4] was therefore used to train GD models. The results for GD models built using MPE-MAP with the GI-MPE system are also shown in Table V. The prior for the MPE-MAP training was the GI system with an MMI-prior for I-smoothing. Again small gains using GD models were observed. This is the form of GD modelling that was used in the experiments in section V and VI.

As discussed in section III-A, large amounts of additional BN training data were made available. For initial evaluation

Training set	#States/Avg Components	eval03		dev04f	
		ML	MPE	ML	MPE
bntr-375h	7K/16	14.8	12.5	–	–
bntr-750h	7K/16	14.8	12.1	–	–
bntr-750h	7K/32	14.2	11.8	26.0	21.6
bntr-1050h	9K/32	13.8	11.4	25.0	20.3
bntr-1350h	9K/32	13.9	11.2	24.8	19.6
bntr-1790h	9K/32	13.7	11.0	24.4	19.3
bntr-2210h	9K/32	13.6	11.1	24.5	19.1

TABLE VI

%WERS WITH THE GI ML/MPE MODELS WITH DIFFERENT TRAINING DATA SIZE. SINGLE PASS DECODING OF WB SEGMENTS WITH THE RT03 TRIGRAM LM. NB HYPOTHESIS USING THE RT03 NB MODELS.

the narrow-band (NB) models built using the bntr-144h data were used for all NB segments, requiring that only wide-band (WB) models were trained. To investigate the effects of the additional data on system performance a range of models were built using the training data-sets described in Table III. These models were evaluated on the eval03 and dev04f test sets as these represented different time epochs and degree of difficulty. Table VI shows the performance of the various acoustic models for both ML and MPE (with an MMI-prior for I-smoothing) training.

A couple of general trends can be observed. For the eval03 test set, consistent gains were obtained for both ML and MPE training as the amount of data, and also the model complexity, was increased as far as the bntr-1050h training set. Beyond this size the gains, expecially for ML training, were significantly less. This may be attributed to the nature of the additional data being added, mainly BN03 data which is not expected to be closely related to the eval03 test data. For the dev04f test set the performance of the system was significantly worse than that on the eval03 test data. This was again expected as "harder" data were included for this test set. In contrast to the eval03 test set, performance on the dev04f test data improved as the amount of data increased for all training sets. Again this can be attributed to the nature of the BN03 data which is more closely matched to the dev04f than the eval03 test set, both in terms of the epoch and data sources.

A further issue to consider, as the amount of training increases, is how to efficiently build NB models. As less than 10% of the data is usually classified as NB data in the CU system, it is not desirable to rebuild systems from scratch[8]. To overcome this problem a two pass approach was adopted. First, all the data, including the WB data, was parameterised using the NB configuration where the data was band-limited to the range 125-3800 Hz. Then standard single pass retraining (SPR) [28] was used from the non-HLDA ML WB model-set to generate a NB ML model-set on this data. This model was then used to estimate the NB HLDA transform. Using this NB HLDA transform, MPE-SPR [25] was used to generate an initial MPE NB model. Two iterations of MPE training were then used to refine this model set. This procedure dramatically

[8]For all the systems the WB models were trained on all the available data, including the NB data, parameterised using the WB configuration.

reduced the time to train the NB models and gave similar performance to rebuilding the NB models from scratch. This approach was used to build all the NB models for the 1×RT and 10×RT systems.

The test data for the RT04f BN evaluation was known to comprise data related to both the `eval03` and `dev04f` epochs and shows. To balance both training time and performance the `bntr-1350` training data was selected as the primary training data set for use in the RT04f evaluation systems. Unless otherwise stated this will be the training data for all subsequent acoustic models built and evaluated.

IV. Language Model Development

Additional training data to that used for the RT03 LM [3] was also available for training language models. This section briefly details the additional sources available and discusses the performance gains obtained.

Source (additional RT04 sources in bold)	Size(MW) RT03	RT04
PSM's BN transcripts 92-99 `TDT2&TDT3` captions, **BN03 captions**	275	334
Transcripts from CNN's website 99-00, **01-03**	66	147
`TDT4` captions, **TDT4a captions**	2	5
NIST's BN training data from 97/98 Marketplace show transcripts	2	2
Newswire LAT and WP 95-98, NYT 97-00 & **01-02**, Associated Press 97-00 & **01-02**	674	928

TABLE VII
LANGUAGE MODEL TRAINING TEXTS AND THEIR SIZES.

The initial experiments in this paper were carried out using the RT03 LM from the CU-HTK 2003 BN evaluation system [3]. The sources and text sizes used to generate this language model are shown in Table VII. For the RT03 LM a 59K vocabulary, chosen based on word frequency counts was used. Five separate language models were built and interpolated, the partition of sources is indicated in the table. For the 2004 RT04f evaluation, additional sources for constructing the language model were available. The additional sources and combined sizes for the 2004 RT04 LM are also shown in Table VII, indicated in bold. This gave a total text size of about 1.4 billion word tokens. Again a 59k word list was used based on frequency counts from the RT04 LM text corpora. In a similar fashion to the training of the RT03 LM, the text sources were split into five subsets, as shown in Table VII, and 4-gram models were built for each subset. Small to mid-size models were smoothed using modified Kneser-Ney discounting [32] and components trained on large data sets used Good-Turing discounting [28]. The interpolation weights were optimised on text comprising `eval03`, `dev04`, and `dev04f`[9]. After interpolation, the component models were merged and the final 4-gram language model was pruned with entropy-based pruning [33]. The RT04 LM had about 17 million bigrams, 28 million trigrams and 23 million 4-grams compared to the 9 million bigrams, 13 million trigrams and 7 million 4-grams for the RT03 LM.

[9]By tuning the interpolation weights on all the available development sets only a minimal bias in the WER% is expected for these sets.

Language Model	Perplexity [OOV%] `eval03`	`dev04`	`dev04f`	`eval04`
RT03	133 [0.66]	124 [0.57]	153 [0.54]	158 [0.81]
RT04	120 [0.45]	118 [0.49]	132 [0.42]	133 [0.62]

TABLE VIII
PERPLEXITY VALUES FOR THE RT03 AND RT04 4-GRAM LANGUAGE MODELS

Table VIII shows the perplexity scores and out-of-vocabulary (OOV) rates for the RT03 and RT04 LMs for four of the test sets. The OOV rates for the word-list associated with the RT04 LM are consistently lower than those of the RT03 LM. The difference is slightly larger for `dev04f` and `eval04` as the time epoch for these test sets, November/December 2003, is closer to the additional data only used for the RT04 LM. It is not possible to directly compare the perplexities between the two LMs as the word-lists are different. However, looking at the trends over the test sets, the difference between the perplexities for the test sets with harder data, `dev04f` and `eval04`, than the easier sets, `eval03` and `dev04`, is larger for the RT03 LM than the RT04 LM. This again may be attributed to the later epoch data included in the RT04 LM training corpora.

Segmentation/ Clustering	LM	%WER `eval03`	`dev04`	`dev04f`
`CU`	RT03	9.7	12.2	—
	RT04	9.2	11.9	16.2
`LIMSI`	RT04	8.8	11.4	15.8

TABLE IX
%WER USING A P1-P2 FRAMEWORK AND THE `bntr-1050` MODELS AND EITHER THE RT03 OR RT04 LANGUAGE MODELS AND EITHER THE `CU` RT03 OR `LIMSI` SEGMENTER.

The recognition performance of the RT03 and RT04 LMs were then compared on the `eval03`, `dev04` and `dev04f` test sets. For these experiments only the P1-P2 stages of the 10×RT framework were run [3]. These stages run in approximately 5×RT. The results using the `CU` segmentation and clustering are shown in Table IX. For both test sets the RT04 LM outperformed the RT03 LM by between 0.3-0.5% absolute. In addition Table IX compares the performance of the `CU` and `LIMSI` segmentation and clustering discussed in section II. The use of the `LIMSI` segmentation and clustering gave an additional error rate reduction of between 0.4-0.5% absolute. The `LIMSI` segmentation and clustering was used as the baseline for all subsequent experiments.

V. 1xRT Experiments

For the 2004 RT04f evaluation, a system running in less than real-time (1×RT)[10] was developed. The approach adopted was to extend and update the architecture that was first used in

[10]All run-times for RT04 systems were run on a single Intel Xeon 3.2GHz/2MB L3 cache processor with hyperthreading enabled. Note that the compute times refer to data throughput, with no constraints on latency.

developing the Cambridge 10×RT broadcast news system in 1998 [24][11]. Due to the increase in processing power of commodity PCs over the intervening six years, along with the use of improved modelling and tuned decoding, it was feasible to use the same two pass recognition approach when developing the 2004 1×RT system. Hence an initial fast decoding pass is used for unsupervised adaptation, and this is followed by lattice generation, and rescoring. Since it is known that lattice quality is much improved by incorporating an initial adaptation stage [34], and due to the success of the earlier "fast" broadcast news systems, an architecture including two full decoding passes with intermediate adaptation is also generally favoured by other recently-developed 1×RT systems for both broadcast news and conversational speech transcription [35], [36].

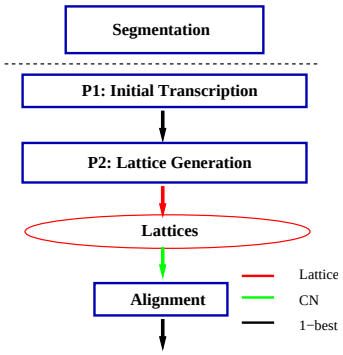

Fig. 1. The 1xRT System Architecture for BN Transcription

The two pass architecture of the 1×RT is shown in Figure 1. Each of the two recognition passes in the 1×RT system is similar to the P1-P2 portion of a single branch 10×RT framework as used in the CU-HTK 2003 BN evaluation system [3], as well as the older "fast" versions of more complex BN evaluation systems developed at CU [24].

The `LIMSI` segmenter/clustering was used (which ran in about 0.1×RT), followed by a very rapid first recognition pass (P1). The output of P1 provided the initial transcription hypothesis which was used to adapt the HMMs used in the P2 stage. The P1 stage used smaller acoustic and language models than the P2-stage to reduce computation as well as tighter pruning beam-widths. It was found that the final error rate was relatively independent of that from the P1 stage as shown in Figure 2, and hence the P1 stage search parameters were set at a level which, on the `eval04` data, took only about 0.15×RT.

To reduce the computation in the P1 stage, the P1 HMMs were 16 component per state MPE `bntr-750`-trained models and a single set of acoustic models were used for all data segments independent of bandwidth (and also GI as is usual for P1 models). In addition, the P1 stage used a heavily pruned trigram version of the RT04 LM (3.3 million bigrams and 2.6 million trigrams).

The second stage (P2) used the P1 hypothesis to perform least squares linear regression and diagonal variance adaptation on bandwidth and GD MPE models trained on

[11]The 1998 10×RT system ran on a 450MHz Intel Pentium II processor.

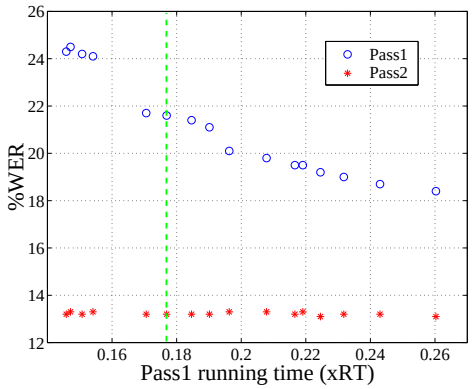

Fig. 2. %WER of `dev04` in P1 (with the trigram LM) and P2 (with 4-gram expansion and applying confusion network decoding) with different P1 decoding times for the 1×RT system. The vertical dotted line shows the operating point chosen for the evaluation system.

`bntr-1350`. Decoding used a more lightly pruned version of the RT04 trigram LM (10.7 million bigrams and 13.8 million trigrams) to produce lattices, which were then expanded with the full RT04 4-gram LM. This was followed by confusion network decoding, estimation of confidence scores [37], and word level alignment. All of the steps involved in adaptation and subsequent P2 decoding required a total of less than 0.7×RT on the eval04 data. As an additional post-processing stage, word tokens with low confidence scores were removed from the recognised results. Note that this final post-processing stage was not found to be beneficial for any systems with less-constrained run-times.

Pass	%WER			
	`eval03`	`dev04`	`dev04f`	`eval04`
P1	17.2	21.7	27.8	25.6
P2-cn	9.9	12.7	17.4	15.4
P2-cn†	9.8	12.5	17.3	15.3

TABLE X
%WER OF THE RT04F 1×RT SYSTEM, P2 USED THE `bntr-1350` MODELS AND PRUNED RT04 LMS AND THE `LIMSI` SEGMENTER. † INDICATES THAT POST-PROCESSING REMOVAL OF LOW-CONFIDENCE WORDS WAS PERFORMED.

Table X shows the performance of the 1×RT system. Comparing the P2-cn output with that given for the same acoustic and language models in a P1-P2 setup (see Table XI), it can be seen that the increase in error rate is between 1.3% for `eval03` and 1.8% for `eval04` without final post-processing, and that the post-processing reduced this gap to 1.2% and 1.7% respectively.

It is also interesting to compare the performance of the 1×RT system with the CU-HTK 2003 BN 10×RT system [3]. That system gave an error rate of 10.6% on the `eval03` data. Thus using the updated acoustic and language models in this 1×RT configuration reduced the error rate by 0.8% absolute while greatly reducing the run-time.

VI. 10×RT Experiments

The previous section described a system that was required to run in less than 1×RT, restricting the recognition framework that could be used. This section examines the form of recognition framework that can be used within a 10×RT constraint[12]. Within this time constraint it is possible to perform multiple recognition runs and combine the outputs. Depending on the framework used, different levels of system diversity can be incorporated, including multiple segmentation and clusterings, acoustic models, adaptation supervision and lattices. All acoustic models evaluated in this section were trained using `bntr-1350h`.

A. Multiple Rescoring (P3) Branch Configuration

The first multiple-pass system combination set-up investigated was based on the 2003 CU-HTK BN evaluation system framework. For these experiments the `LIMSI` segmentation and clustering was used along with the RT04 LM.

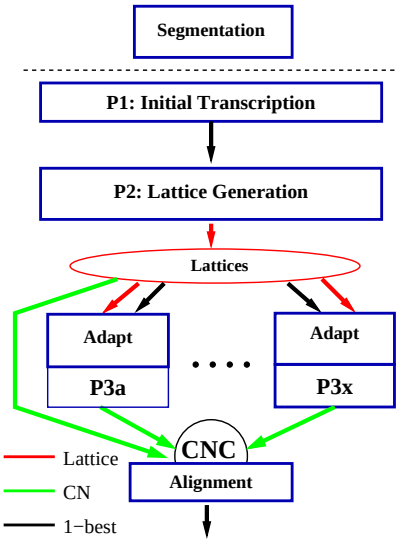

Fig. 3. Multiple Rescoring Branch Framework for BN Transcription.

The overall multiple rescoring branch configuration is shown in Figure 3. As in the 1×RT system, a P1-P2 framework is used to generate lattices and 1-best hypotheses. However, more complex acoustic models and LMs are normally used at the P1 stage compared to those in the 1×RT architecture. For the experiments in this paper 32 component per state acoustic models and the standard RT04 LMs were used. In addition to the P1-P2 stages multiple P3 lattice rescoring branches are run. These branches use the P2 1-best output and lattices for MLLR mean [38] and variance adaptation [39], and lattice-based adaptation [40].

A range of acoustic models were considered for use in the P3 rescoring branches. The baseline P3 acoustic models were the ones used for the CU-HTK 2003 BN evaluation system. The first branch (P3a) system was built using speaker adaptive training (SAT) employing global constrained MLLR (CMLLR) [39] transforms. The second branch (P3c) system was a GD system built using a single pronunciation (SPron) dictionary derived from the standard multiple pronunciation dictionary [41]. Possible alternative acoustic model configurations examined included structured precision matrices [42], cluster adaptive training [43] and a Gaussianised front-end using the scheme described in [44]. These models were initially trained using the `bntr-375h` training set and the performance evaluated within the multi-branch system combination framework shown in Figure 3 [6]. Though some gains over the standard SPron/SAT combination were obtained, it was found that no significant gains over the baseline configuration were obtained with the larger acoustic models and the full `bntr-1350` training data. Hence the same models as those run in the 2003 system, SPron and SAT, were used. In addition to these standard branches, the performance of a GD model using the multiple pronunciation dictionary (MPron), the same model as used in the P2 stage, was also evaluated in the P3 stage (P3b). For all the P3 systems the numbers quoted are after confusion network (CN) decoding [45]. This allows the gains from system combination to be clearly seen. Each of the stages P2, P3a and P3c produce word lattices and these were converted to confusion networks and then combined with CNC. Finally, a forced alignment of the final word-level output was used to obtain accurate word times before scoring.

The recognition framework shown in Figure 3 was also used for the CU-HTK 2003 BN evaluation system [3]. Using this structure, `bntr-144h` trained acoustic models and the RT03 LM gave an error rate of 10.6% on `eval03`[13]. It is interesting to note that for these models a 0.4% absolute gain in performance was obtained using the CNC over the best single branch performance with CN decoding (P3a-cn).

System		%WER			
		`eval03`	`dev04`	`dev04f`	`eval04`
P2-cn	MPron	8.6	11.1	15.9	13.6
P3a-cn	SAT	8.2	10.6	15.3	13.3
P3b-cn	MPron	8.2	10.6	15.4	13.4
P3c-cn	SPron	8.1	10.4	15.2	13.0
P2+P3b		8.3	10.8	15.3	13.3
P2+P3a	CNC	8.0	10.5	15.2	13.2
P2+P3c		8.1	10.3	14.8	12.8
P2+P3a+P3c	CNC	8.0	10.4	14.9	12.9

TABLE XI
%WERs IN P2, VARIOUS P3 BRANCHES IN THE MULTIPLE RESCORING BRANCH FRAMEWORK USING THE `LIMSI` SEGMENTATION/CLUSTERING AND THE RT04 LMs.

Table XI shows the performance of the `bntr-1350h` trained acoustic models using this multiple rescoring branch framework and the RT04 LM. For this configuration the best single branch performance was obtained with the SPron model (P3c-cn). After confusion network decoding, this gave an error rate of 8.1% on `eval03`, a 2.5% absolute reduction in error rate over the CU-HTK 2003 BN system. Combining

[12]The actual run-times for these systems were not measured. However the settings that were used for all configurations were consistent with those used in the RT04 evaluations, which ran in less that 10×RT.

[13]This was the lowest error rate reported for the RT03 evaluation.

the rescoring P3 branches with the initial P2 branch shows slight gains for all systems other than the MPron (P3b) branch. This shows that some acoustic diversity is necessary to obtain combination gains, even if more adaptation and improved supervision is being used. Though the final output (P2+P3a+P3c) gave about a 25% relative reduction in WER compared with the 2003 10×RT system on `eval03`, it performed no better than the best two-way combination (P2+P3c) and little better than the best individual branch (P3c).

B. Dual Acoustic Model Configuration

The system combination gains shown in the previous section were small. To attempt to improve the gains from combination a dual recognition system configuration was examined. One of the limitations of the multiple rescoring branch framework is that the same 1-best supervision and lattices are used for all branches. The dual recognition system considered in this section removes this restriction.

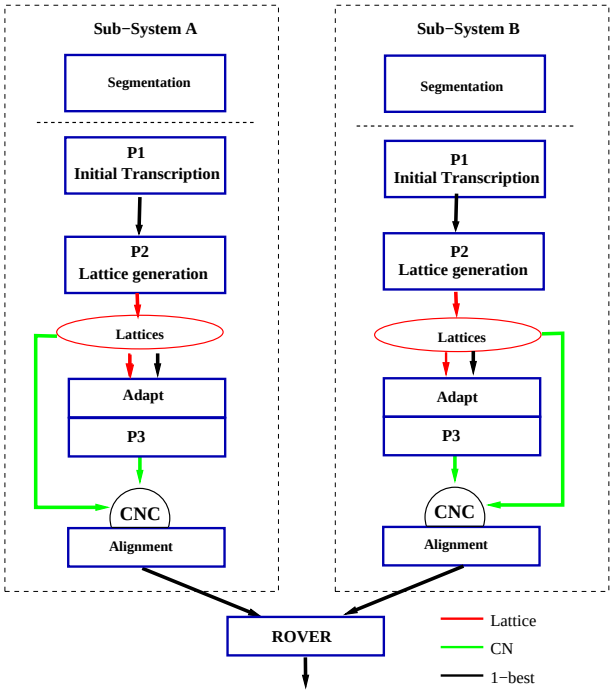

Fig. 4. The Dual System Framework for BN Transcription.

Figure 4 shows the dual system architecture. The structure consists of two completely separate branches where independent P1-P2 stages are used to allow different acoustic models to generate different lattices and 1-best supervision for P3 rescoring. The output from these individual branches are then combined using ROVER [16][14]. Note, confidence scores, derived as described in [37], were used in the ROVER combination in this work. This structure should allow a level of cross-acoustic-model adaptation and combination that was not possible within the previous framework. For these experiments either the SPron or MPron acoustic models were used in the P2 stage. In order to run the dual recognition configuration framework in approximately 10×RT it was necessary to run a slighter faster P1-P2 and P3 rescoring set-up than the multiple rescoring branch system in section VI-A.

[14]CNC could have be used here as the segmentation was consistent.

P2 Models	System		%WER dev04	%WER dev04f	%WER eval04
MPron (P2b)	P2b-cn	MPron	11.1	16.0	13.7
	P3b-cn	MPron	10.7	15.4	13.4
	P3c-cn	SPron	10.4	15.2	13.0
	P2b+P3b	CNC	10.7	15.4	13.4
	P2b+P3c	CNC	10.3	14.9	12.8
SPron (P2c)	P2c-cn	SPron	11.3	15.8	13.7
	P3b-cn	MPron	10.8	15.3	13.3
	P3c-cn	SPron	10.5	15.3	13.2
	P2c+P3b	CNC	10.6	15.0	13.0
	P2c+P3c	CNC	10.7	15.2	13.2
P2b+P3b ⊕ P2b+P3c		ROVER	10.4	15.1	13.0
P2b+P3b ⊕ P2c+P3c		ROVER	10.4	14.9	12.9
P2b+P3c ⊕ P2c+P3b		ROVER	10.3	14.8	12.8

TABLE XII
%WER OF THE DUAL CONFIGURATION SYSTEM USING MPRON AND SPRON ACOUSTIC MODELS AND THE `LIMSI` SEGMENTATION/CLUSTERING

Table XII shows the recognition performance of the dual recognition configuration. Comparing the MPron/SPron branch (P2b+P3c) with the equivalent branch of the rescoring branch configuration (P2+P3c) in Table XI shows that almost no degradation in performance resulted from the faster P1-P2 and P3 stages in the dual configuration. Not surprisingly when using the MPron or SPron models in the P2 stage best performance was obtained by using the other of model type in the P3 stage. However the combination of any two of the individual branches using ROVER [16] (indicated using ⊕) yielded almost no gain over the best single branch. Thus for this configuration incorporating diversity in the form of the 1-best supervision and lattices for the P3 stage gave almost no improvement in performance.

C. Dual Segmentation/Clustering Configuration

Using the dual recognition framework of section VI-B it is possible to add further diversity by using different segmentation and clusterings in each of the individual branches of the system. The use of multiple segmentations has a number of possible advantages. As shown in Table II, all the segmentations available yield some level of missed speech (MS). For those regions it is not possible to hypothesise outputs. Using multiple segmentations reduces the chance of speech being missed. Exploiting different segmentations should also improve the robustness of the system to *end effects*. It has been observed that the numbers of errors is greater at the start and end of segments compared to the middle [46]. Multiple segmentations may lessen this problem as segment boundaries in one segmentation may occur in the middle of segments in another.

Each of the segmentations and clusterings from section II were evaluated within this dual segmentation configuration. The results are shown in Table XIII. For these experiments no diversity in the acoustic processing was used, MPron models were used for the P2 stage and SPron models for the P3 stage.

System	Segmentation/ Clustering	%WER dev04	%WER dev04f	%WER eval04
L0+P3c (P2b+P3c)	LIMSI	10.3	14.9	12.8
B0+P3c	BBN	10.7	15.0	13.0
C0+P3c	CU	10.8	15.5	13.3
C1+P3c	CU1	10.5	15.2	13.0
C2+P3c	CU2	10.4	15.2	12.9
L1+P3c	LIMSI/CU	10.5	15.1	13.0
L0+P3c ⊕ C0+P3c	ROVER	10.0	14.7	12.6
L0+P3c ⊕ B0+P3c	ROVER	10.0	14.4	12.4
L0+P3c ⊕ L1+P3c	ROVER	10.2	14.9	12.8
L0+P3c ⊕ C1+P3c	ROVER	9.8	14.6	12.5
L0+P3c ⊕ C2+P3c	ROVER	9.8	14.7	12.4
B0+P3c ⊕ C0+P3c	ROVER	10.1	14.8	12.6
B0+P3c ⊕ C2+P3c	ROVER	9.9	14.6	12.5
C0+P3c ⊕ C2+P3c	ROVER	10.1	14.9	12.7

TABLE XIII

%WER OF THE DUAL SEGMENTATION SYSTEM USING bntr-1350 TRAINED MODELS (MPRON MODELS IN P2, SPRON IN P3 (P3C)).

The baseline LIMSI segmentation and clustering (L0+P3c) is thus the same as the P2b+P3c system of Table XII and was the best single individual branch. The performance of all the other single branch systems, other than the original CU segmentation and clustering, were about the same. As previously reported [21], there is little correlation between the DER quoted in Table II and the WER ranking in Table XIII for the individual systems.

ROVER was used to combine the 1-best results from the two branches. For all systems where the segmentation was varied improvement in performance was obtained from combining branches. Three systems all gave similar results, LIMSI (L0+P3c) combined with either BBN (B0+P3c), CU1 (C1+P3c) or CU2 (C2+P3c). These segmentations and clusterings show a range of characteristics as described in section II. This system combination yielded gains of between 0.3-0.5% absolute over the best individual branch error rate.

As an additional contrast the effects of using a fixed segmentation and different clustering schemes was also investigated. This experiment used the LIMSI segmentation with the CU top-down clustering (L1+P3c). For each of the test sets there was a small degradation in performance using the CU clustering compared to the L0+P3 system. However in contrast to the multiple segmentation results, there was almost no gain in performance when using the systems in combination (L0+P3c ⊕ L1+P3c). This indicates that the primary cause of error rate reduction from system combination was due to the use of multiple segmentations rather than any diversity in the clustering schemes (though these may dramatically affect the DER).

Comparing the best performance of this dual segmentation framework on eval04, 12.4%, with that of the final multiple rescoring branch combination in Table XI shows a 0.5% absolute, 4% relative, reduction in WER. This represented a significant difference using the matched-pair sentence-segment word error significance test test [47].

VII. CROSS-SITE COMBINATION: "SUPEREARS"

To further illustrate and explore the performance improvements that can be obtained with a multi-branch system in a combination framework, this section briefly describes the "SuperEARS" system [1]. It provides a contrast to the systems described in previous sections where the various acoustic models, language models and decoders were all implemented at CU.

The SuperEARS system was the result of a cross-site collaboration between research teams at BBN, LIMSI, SRI and CU in the context of the DARPA EARS programme and was designed to still respect the 10×RT constraint for the complete system. It exploits the benefits of both explicit combination via ROVER and implicit combination by using i) hypotheses from one sub-system to adapt another; and ii) using models from one-system to rescore lattices produced by another. As is well-known, in all cases the potential gain from combination is greatest when there are multiple sub-systems with similar average WER but large differences in detailed error patterns.

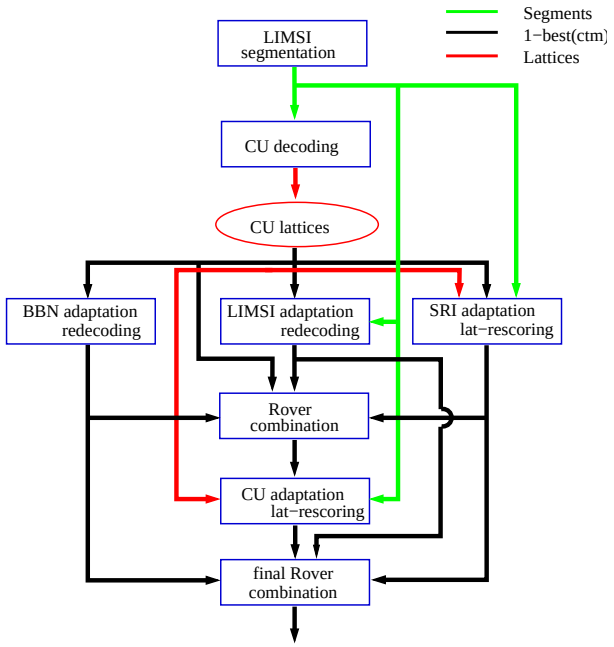

Fig. 5. Cross-Site "SuperEARS" Architecture for BN Transcription.

Figure 5 shows the overall structure of the "SuperEARS" system. The initial stages of the process, including the LIMSI segmentation, and P1-P2 decoding were identical to the corresponding stages of lattice generation in section VI-A, for which the results with multiple rescoring branches are given in Table XI.

The hypotheses, lattices and segmentations were used in different ways by the various teams at BBN, LIMSI and SRI. BBN took only the 1-best hypothesis from the P2-cn stage and used it as adaptation supervision. BBN used the BBN segmentation and clustering described in section II, and then performed a full adapted decoding pass using the BBN acoustic models, language models and decoding system described in [10]. This output is labelled P3B. Similarly LIMSI

used the 1-best output from P2-cn and again performed a full adapted decoding pass with the LIMSI acoustic and language models, with the `LIMSI` segmentation/clustering to give the P3L output. It is interesting to note that while a full decode pass gave a higher WER at the P3L stage than if lattice rescoring was performed, it resulted in a lower overall WER after final combination. While the BBN and LIMSI systems performed full decoding, the SRI sub-system [18] used the lattices from the P2 stage (with the SRI language model applied) to constrain the recognition search with the adapted SRI acoustic models to give the P3S output.

The output from all the P3 stages, which included confidence scores, as well as the P2-cn output were then combined using ROVER. The 1-best output from this was used for adaptation supervision with an increased number of adaptation transforms for the final CU sub-system which used the SPron models in lattice rescoring mode.

The speed of the SuperEARS system was tuned so that the additional P4 pass of adaptation/rescoring with the CU system was possible while still keeping the overall computation within 10×RT. The final output was obtained by combining the output from the P3 and P4 stages together. This configuration combines diversity in terms of segmentation and clustering, front-ends, acoustic models and language models, as well as including both implicit and explicit combination.

System			%WER		
			`dev04`	`dev04f`	`eval04`
P2-cn	CU	MPron	11.1	15.9	13.6
P3B	BBN	decode	9.8	14.3	12.8
P3L	LIMSI	decode	10.5	15.9	14.0
P3S	SRI	rescore	9.7	16.5	14.6
P2⊕P3B⊕P3L⊕P3S		ROVER	8.9	13.9	12.2
P4	CU	SPron	9.6	14.3	12.8
P3B⊕P3L⊕P3S⊕P4		ROVER	8.3	13.4	11.6

TABLE XIV
%WER OF THE SUPEREARS CROSS-SITE COMBINATION SYSTEM.

Table XIV shows the performance of the SuperEARS system. There are a number of interesting contrasts that can be drawn. The best single P3 branch performance was generally obtained by the BBN system (P3B). This is the only system that used a different segmentation. However it made use of supervision hypotheses from the CU system. This again may indicate the advantage of using multiple segmentations when combining/cross-adapting systems. The performance of this single branch on the `eval04` test set is almost as good as the combined BBN/LIMSI system which gave an error rate of 12.7% [10].

It is interesting to compare the output from the initial ROVER stage (prior to P4) with the output of the dual segmentation system using the `BBN` and `LIMSI` segmentation/clusterings (L0+P3c ⊕ B0+P3c in Table XIII). For the `eval04` test data the performance of the dual segmentation system, 12.4%, was only marginally worse than this stage of the SuperEARS system, 12.2%. However for the other two test sets the performance difference is rather larger, in particular on the `dev04` set due to the very good performance of the BBN sub-system on that data. Combining the final P4 CU system with the P3B, P3L and P3S branches gave an additional reduction in WER of around 0.5% over combination with the P2 stage output. This final output is between 0.6% and 1.7% lower than the dual segmentation system using the `LIMSI` and `BBN` segmentations with only CU acoustic and language models throughout (L0+P3c ⊕ B0+P3c in Table XIII).

The use of the various sub-systems and combination strategies within the final SuperEARS framework produced a system which was robust to sub-system performance differences across test sets. The final system gave low error rates as shown in Table XIV, with additionally a WER of 6.7% on the `eval03` data set, a 38% relative reduction in WER over the CU-HTK 2003 BN system. Finally it is interesting to note that the SuperEARS system gives essentially identical WERs to taking the individual best RT04 10×RT BBN/LIMSI [10], SRI [18] and CU [25] systems (overall run-time of 30×RT) and combining those with ROVER.

VIII. CONCLUSIONS

This paper has described a series of developments associated with the design of a state-of-the-art Broadcast News transcription system. The use of large amounts of lightly supervised acoustic training data for constructing discriminatively trained acoustic models is discussed, along with the performance on a range of standard test sets. The use of these updated acoustic models, along with updated language models, within both a real-time framework and an approximately 10×RT framework is also described. For the real-time system the performance on the 2003 evaluation data was significantly better than that of the 10×RT 2003 CU-HTK BN evaluation system. For the 10×RT systems a number of possible decoding frameworks were described, which allow the hypotheses from multiple systems to be combined. Using the same multiple-rescoring-branch combination-framework as the CU-HTK 2003 BN evaluation system, the new acoustic and language models gave gains of about 25% relative over the 2003 system. However, only small gains in performance over the best single branch system were obtained. Two modifications to this framework were then considered to increase the diversity of the hypotheses to combine. The first used multiple adaptation hypotheses and rescoring lattices, but again little improvement by combination were obtained. The best combination results were obtained by using multiple segmentations. This multiple segmentation system gave additional gains of about 4% relative over the multiple rescoring branch framework. As a contrast a cross-site, combining systems from BBN, LIMSI and SRI, was also described. Using this combination of both the diverse segmentations, and acoustic and language models, gave a 6% relative gain over the best CU acoustic and language model system.

REFERENCES

[1] P. C. Woodland, H. Y. Chan, G. Evermann, M. J. F. Gales, D. Y. Kim, X. A. Liu, D. Mrva, K. C. Sim, L. Wang, K. Yu, J. Makhoul, R. Schwartz, L. Nguyen, S. Matsoukas, B. Xiang, M. Afify, S. Abdou, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, F. Lefevre, D. Vergyri, W. Wang, J. Zheng, A. Venkataraman, R. R. Gadde, and A. Stolcke, "SuperEARS: Multi-site broadcast news system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, November 2004.

[2] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002.

[3] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. E. Tranter, L. Wang, and P. C. Woodland, "Recent advances in broadcast news transcription," in *Proc. IEEE ASRU Workshop*, St. Thomas, U.S. Virgin Islands, November 2003, pp. 105–110.

[4] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Proc. Eur. Conf. Speech Commun. Technol.*, Geneva, Switzerland, September 2003.

[5] H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, Canada, March 2004.

[6] X. Liu, M. J. F. Gales, K. C. Sim, and K. Yu, "Investigation of acoustic modeling techniques for LVCSR systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, March 2005.

[7] NIST Speech Group, "Fall 2004 rich transcription RT-04f evaluation plan," NIST, Tech. Rep., 2004, available at http://www.nist.gov/speech/tests/rt/rt2004/fall/.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarisation systems," *IEEE Trans. on SAP.*, to appear in the Special Issue on Rich Text Transcription.

[9] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Proceedings InterSpeech*, Lisbon, Portugal, September 2005.

[10] L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J. Gauvain, G. Adda, H. Schwenk, and F. Lefevre, "The 2004 BBN/LIMSI 10xRT English Broadcast News transcription system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, November 2004.

[11] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Computer Speech and Language*, pp. 89–108, 2002.

[12] G. Saon, D. Povey, and G. Zweig, "CTS decoding improvements at IBM," in *EARS STT workshop*, St. Thomas, U.S. Virgin Islands, December 2003.

[13] L. Lamel and J.-L. Gauvain, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[14] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, Canada, March 2004.

[15] H. Y. Chan, "Lightly supervised discriminative training for LVCSR," M. Sc. thesis, Cambridge University, Cambridge, UK, 2004.

[16] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. IEEE ASRU Workshop*, 1997.

[17] G. Evermann and P. C. Woodland, "Design of fast LVCSR systems," in *Proc. IEEE ASRU Workshop*, St. Thomas, U.S. Virgin Islands, November 2003, pp. 7–12.

[18] A. Venkataraman, R. Gadde, A. Stolcke, D. Vergyri, W. Wang, and J. Zheng, "SRI's 2004 Broadcast News speech to text system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, November 2004.

[19] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM Approach for Learning and Adapting Sound Models for Speaker Indexing," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Crete, Greece, June 2001, pp. 175–180.

[20] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proceedings InterSpeech*, Lisbon, Portugal, September 2005.

[21] S. E. Tranter, K. Yu, D. A. Reynolds, G. Evermann, D. Y. Kim, and P. C. Woodland, "An investigation into the interactions between speaker diarisation systems and automatic speech transcription," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR-464, 2003.

[22] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. Int. Conf. Spoken Lang. Process.*, vol. 4, Sydney, Australia, December 1998, pp. 1335–1338.

[23] L. Nguyen, B. Xiang, M. Afify, S. Abdou, S. Matsoukas, R. Schwartz, and J. Makhoul, "The BBN RT04 English Broadcast news transcription system," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2005.

[24] P. C. Woodland, "The development of the HTK Broadcast News transcription system: an overview," *Speech Communication*, vol. 37, pp. 47–67, 2002.

[25] D. Y. Kim, H. Y. Chan, G. Evermann, M. J. F. Gales, D. Mrva, K. C. Sim, and P. C. Woodland, "Development of the CU-HTK 2004 Broadcast News transcription systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, March 2005.

[26] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, John Hopkins University, 1997.

[27] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[28] S. J. Young, G. Evermann, , M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.3*. Cambridge, UK: Cambridge University Engineering Department, 2005.

[29] S. J. Young, J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA Human Language Technology Workshop*, 1994.

[30] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, pp. 25–47, 2002.

[31] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[32] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, September 2002.

[33] ——, "Entropy-based pruning of backoff language models," in *Proc. DARPA News Transcription and Understanding Workshop*, 1998.

[34] P. Woodland, D. Pye, and M. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, 1996, pp. 1133–1136.

[35] S. Matsoukas, R. Prasad, S. Laxminarayan, B. Xiang, L. Nguyen, and R. Schwartz, "The 2004 BBN 1xRT recognition systems for English broadcast news and conversational telephone speech," in *Proceedings InterSpeech*, Lisbon, Portugal, September 2005.

[36] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Proc. Eur. Conf. Speech Commun. Technol.*, Geneva, Switzerland, September 2003, pp. 1977–1980.

[37] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, College Park, MD, May 2000.

[38] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[39] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[40] L. F. Uebel and P. C. Woodland, "Speaker adaptation using lattice-based MLLR," in *Proc. ITRW on Adaptation Methods for Speech Recognition*, 2001.

[41] T. Hain, "Implicit pronunciation modelling in ASR," in *ISCA ITRW PMLA*, 2002.

[42] K. C. Sim and M. J. F. Gales, "Basis superposition precision matrix modelling for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 2004.

[43] K. Yu and M. J. F. Gales, "Adaptive training using structured transforms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004.

[44] M. J. F. Gales, B. Jia, X. Liu, K. C. Sim, P. C. Woodland, and K. Yu, "Development of the CUHTK 2004 Mandarin conversational telephone speech transcription system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, March 2005.

[45] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1999.

[46] N. Duta and R. Schwartz, "Error analysis of the BN and CTS results," in *EARS STT workshop*, St. Thomas, U.S. Virgin Islands, December 2003.

[47] D. S. Pallett, W. M. Fisher, and J. G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990.