RECENT ADVANCES IN BROADCAST NEWS TRANSCRIPTION

D.Y. Kim, G. Evermann, T. Hain, D. Mrva, S.E. Tranter, L. Wang & P.C. Woodland

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ, U.K.
Email: {dyk21,ge204,th223,dm312,sej28,lw256,pcw}@eng.cam.ac.uk

ABSTRACT

This paper describes recent advances in the CU-HTK Broadcast News English (BN-E) transcription system and its performance in the DARPA/NIST Rich Transcription 2003 Speech-to-Text (RT-03) evaluation. Heteroscedastic linear discriminant analysis (HLDA) and discriminative training, which were previously developed in the context of the recognition of conversational telephone speech, have been successfully applied to the BN-E task for the first time. A number of new features have also been added. These include gender-dependent (GD) discriminative training; and modified discriminative training using lattice re-generation and combination. On the 2003 evaluation set the system gave an overall word error rate of 10.7% in less than 10 times real time (10×RT).

1. INTRODUCTION

Broadcast News transcription has been one of the most challenging and interesting tasks in large vocabulary continuous speech recognition over recent years. Significant progress has been made despite the many difficult problems for automatic transcription that are inherent in this type of data. These problems include the presence of various speaking styles (read, spontaneous and conversational); non-native speakers; background noise and/or music; and different audio channel characteristics (wideband and telephone band).

This paper presents technical details and experimental results for the various acoustic models developed for the RT-03 evaluation as well as the actual evaluation system. As the primary condition for the RT-03 BN-E evaluation required the system to operate in less than 10 times real-time (10×RT), we focus on the design and performance of systems running with that constraint. The main areas of development include the use of HLDA; discriminative training using the minimum phone error (MPE) criterion; maximum a posteriori (MAP)-style MPE (MPE-MAP) training for GD modelling; a complementary part of the system using a single pronunciation dictionary (SPRON); and system combination.

The rest of the paper is arranged as follows. First an overview of our previous 10×RT BN-E system is given. This is followed by a description of the data sets used in the experiments and then by sections that discuss the acoustic model training, adaptation, SPRON, and language models, respectively. Finally the complete evaluation system is described and the results of each stage of processing are presented.

2. PREVIOUS 10×RT CU-HTK BN SYSTEM OVERVIEW

The previous HTK 10×RT Broadcast News system was developed in 1998 [12, 18] and runs in a number of stages. The input audio stream is first segmented; a first recognition pass is performed using gender-independent (GI) triphone HMMs and a trigram language model (LM) to get an initial transcription for each segment; the speaker gender for each segment is found automatically; the segments are clustered, and unsupervised maximum likelihood linear regression (MLLR) [8] transforms estimated for each segment cluster. This is followed by generating a lattice for each segment using the adapted GD triphone models with a trigram LM and expanding these lattice using a word 4-gram interpolated with a category trigram LM. The 1-best hypothesis from the lattice represents the final system output. All acoustic model parameters were estimated using maximum likelihood (ML) training.

The audio segmentation aims to generate acoustically homogeneous speech segments and discard non-speech portions such as music. The data is first split into regions of wideband speech, telephone speech, speech with music/noise and pure music/noise using a Gaussian mixture model (GMM) classifier. The music is discarded and the speech with music/noise treated as wideband speech. GD phone recognisers are then run to locate gender-change points and silence portions to enable these regions to be split into smaller segments. Finally similar adjacent segments are merged and combined with the GMM classifier output to produce the final segmentation with bandwidth and putative gender labels.

For recognition, each frame of input speech is represented by a 39 dimensional feature vector that consists of 13 (including c_0) MF-PLP cepstral parameters and their first and second differentials. Cepstral mean normalisation (CMN) is applied on each segment. The HMMs were initially trained on all the wideband analysed training data. Narrow-band sets were estimated by using a version of the training data with narrow-band analysis (125-3750Hz). GD models for each bandwidth were generated. In testing, the reduced bandwidth models are used for transcribing data classified as narrow band.

3. BROADCAST NEWS DATA

3.1. Acoustic training data

For acoustic model training, the BN-E data released by the LDC in 1997 and 1998 was used. The 1997 data was annotated by the LDC to ensure that each segment was acoustically homogeneous but the 1998 data was transcribed only at the speaker turn level without distinguishing background conditions. In total, these amounted to approximately 143 hours of usable data [5].

3.2. Development data

Three different data sets were used for system development. The first is the 1998 Hub4 evaluation data and consists of two 1.5-hour data files (`eval98`). This is the only test set which allowed measuring the performance by focus condition. The second is the Rich

To appear in *Proc. ASRU 2003*

Transcription 2002 BN-E evaluation data set which is approximately 60 minutes in length (`eval02`). Finally, six 30 minute broadcasts were chosen from the last 2 weeks of the topic detection and tracking (TDT4) data of Jan. 2001, and transcribed manually in conjunction with other speech research sites (`dev03`).

3.3. Text corpora

The following five sets of broadcast and newswire text corpora were used for LM training:.

- Primary Source Media Broadcast News transcriptions (1992-1999) & TDT2+TDT3 closed captions
- CNN show transcriptions (1999-2001)
- TDT4 closed captions
- transcriptions from acoustic training data (1997 & 1998) & acoustic transcriptions of Marketplace shows
- Los Angeles Times and Washington Post newswire service texts (1995-1998) & New York Times newswire texts (1997-2001)

No data produced after 15th January 2001 was used to ensure the training data pre-dated both the `dev03` and evaluation sets. The amount of language model training text is approximately one billion words in total.

4. ACOUSTIC MODEL BUILDING

The basic acoustic model was built using conventional ML with the same front-end as in previous system. Decision tree clustering was used to define cross-word triphone models with about 7000 states. Each speech state was modelled with a 16 component Gaussian mixture distribution.

The experimental results on `eval98` and `eval02` in this section were obtained using a single-pass decoder with a 65k word trigram language model taken from the 1998 CU-HTK BN-E transcription evaluation system [18]. The decoder operated within about $5\times$RT, and no adaptation was used.

The overall word error rate (WER) on `eval98` with the basic ML model was 19.6%. The detailed results broken down by the various focus conditions are given in column (a) of Table 1.

An overview of the complete acoustic model building procedure, described in the following sections, is illustrated in Figure 1.

4.1. HLDA projection

HLDA is an extension of LDA without the restriction that within class covariance matrices have to be identical [7]. By the use of an HLDA projection, an original d-dimensional feature space is divided into p-dimensional *useful* and [d-p]-dimensional *nuisance* subspaces, and only the *useful* subspace is used for actual classification.

In our experiments, a 52-dimensional feature vector was formed by augmenting the basic acoustic representation with 3rd order derivatives, in addition to the usual first and second order derivatives. Acoustic models were built using single-pass re-training in the extended feature space. The HLDA transform is optimised in an iterative fashion using an EM algorithm (i.e. ML estimation). Full covariance statistics were obtained from a system trained on the non-transformed 52-dimensional feature vector and used for the optimisation of the HLDA projection. The nuisance dimensions which contain the least discriminant information are modelled explicitly using a global Gaussian distribution for all acoustic classes during transform optimisation and are eventually discarded. Fisher-ratio values are used to select the nuisance dimensions [9].

F-cond	Ratio	(a)	(b)	(c)	(d)
F0	30.6%	11.1	10.2	9.6	8.8
F1	19.3%	20.1	18.5	17.1	15.5
F2	3.4%	25.8	22.6	22.6	19.6
F3	4.3%	20.9	19.1	17.5	17.3
F4	28.2%	20.1	18.9	16.1	15.3
F5	0.7%	28.1	27.2	21.7	19.1
FX	13.5%	35.0	30.5	27.8	25.7
Overall	100.0%	19.6	17.9	16.2	15.0

Table 1. %WER on `eval98` with (a) ML, (b) HLDA, (c) MPE, and (d) HLDA+MPE (F0: prepared speech, F1: spontaneous speech, F2: speech over telephone channels, F3: speech and music, F4: speech with degraded acoustics, F5: non-native speakers, FX: all other speech).

The experimental results show that the use of an HLDA projection reduced the WER by 1.7% absolute on `eval98` compared with the ML model as shown in column (b) of Table 1. Consistent improvements were observed in various poorly performing conditions as well as for prepared broadcast speech (F0).

4.2. MPE training

MPE training [13] is an extension of our previous work on discriminative training in a lattice-based framework [19]. It tries to minimise an estimate of the training-set phone error rate computed in a word recognition context. This phone error estimate is calculated based on lattices generated by recognising the training data. A bigram LM trained on the acoustic transcriptions is used with a fast decoder setup to generate word lattices. In a separate pass these lattices were then aligned to find phone model boundaries with the appropriate model set. The acoustic model log likelihoods were scaled down using the usual language model scale factor during training to increase the effective number of phone alternatives. The I-smoothing scheme was used to improve the generalisation of the discriminatively trained models which smoothes between the discriminative and the ML estimates where the degree of smoothing depends on the amount of data available [13].

As shown in column (c) of Table 1, the MPE model reduced the WER by 3.4% absolute from the ML model. Moreover, MPE trained models on top of the HLDA model, (d) in Table 1, was 1.2% absolute better than the MPE model on non-HLDA data, and both of them showed improvements over all F-conditions. Therefore, HLDA was used in all of the MPE based model sets described below.

4.3. GD discriminative training

A MAP-style adaptation method for MPE training (MPE-MAP) was introduced in [14]. Using the concept of weak-sense auxiliary functions, it is simple to extend the MAP scheme to incorporate discriminative training criteria, and results in smoothing the usual

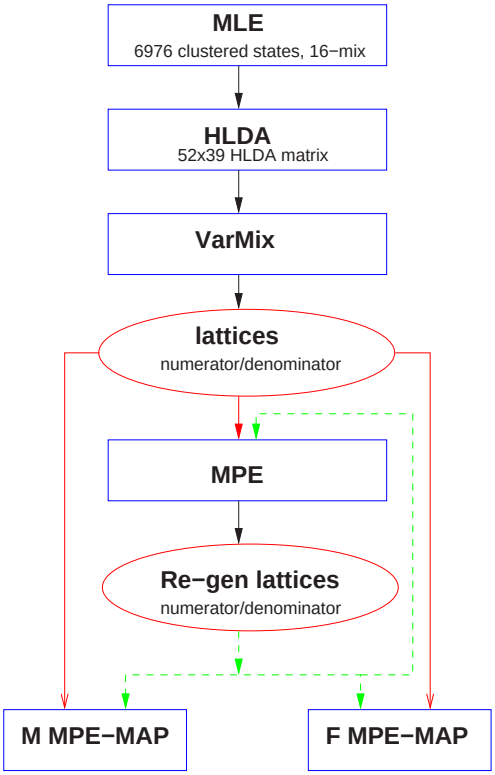

Fig. 1. Stages in final acoustic models building.

	eval98	eval02
GI (MPE)	15.0	13.6
GD (MPE-MAP)	14.5	13.0

Table 2. %WER on eval98 and eval02 with GI MPE and GD MPE-MAP models.

discriminative update counts with the prior counts. The MPE system was used as the original model for adaptation and three iterations of MPE-MAP training were performed for each gender, where only the Gaussian means and mixture weights were updated. The results, given in Table 2, show that the resulting GD models gave 0.5 and 0.6% absolute error reduction on eval98 and eval02, respectively.

As an alternative to MPE-MAP a simple approach to generating GD models was investigated. After GI MPE training, a further MPE iteration was performed on the male and female training data separately. This gave just 0.2% absolute error reduction on eval98.

4.4. Variable number of Gaussians per state

Our previous standard approach was to have a fixed number of Gaussians (N) per speech state and $2N$ for silence states. Here, this was modified to set the number of Gaussians as a function of the number of frames that are available to train each state, while keeping the average number of Gaussians per state at N. This method (VarMix) gave small, but consistent, gains on the development data. Experimental results on eval98 are given in Table 3. Absolute gains of 0.3 and 0.2% were obtained for HLDA and MPE models by allowing the number of Gaussians per state to vary.

	Number of Gaussians	Lattices for MPE	WER (%)
HLDA	fixed	-	17.9
	variable	-	17.6
MPE	fixed	orig	15.0
	variable	orig	14.8
	variable	orig + re-gen	14.4
MPE-MAP	fixed	orig	14.5
	variable	orig + re-gen	13.8

Table 3. %WER on eval98 for variable number of Gaussians and lattice re-generation for MPE training.

4.5. Lattice re-generation for MPE training

In standard lattice-based discriminative training [18], the lattices which represent the confusable hypotheses for each utterance are generated once and the model-level alignment is assumed to be fixed. If the HMM parameters change significantly during discriminative training this may not be a good approximation, so lattice regeneration schemes were investigated.

After four iterations of MPE training the resulting acoustic models were used to regenerate a set of training lattices to ensure that the confusable word alternatives were represented for the subsequent iterations. This lattice generation also used a heavily pruned bigram LM (only about 50k bigrams). In the subsequent iterations of MPE training, statistics based on both sets of lattices were employed.

As shown in Table 3, the WER was reduced both for MPE and MPE-MAP models by lattice regeneration. On eval98, an absolute gain of 0.4% WER was obtained with the GI MPE models. For GD MPE-MAP models, the combination of using variable number of Gaussians and re-generating lattices gave 0.7% absolute improvement.

The MPE-MAP model trained using re-generated lattices (as well as the original lattices) was 5.8% absolute (29.6% relative) better than the basic ML model on eval98 before adaptation. These MPE and MPE-MAP models were used in the actual evaluation system.

5. ADAPTATION AND ADAPTIVE TRAINING

5.1. Adaptation experiments

Based on the MPE model described in section 4.2, several unsupervised transcription-model adaptation experiments were conducted to evaluate the effectiveness of various adaptation techniques for these models and to choose the optimal adaptation strategy. Clustering was performed on the segments for each combination of gender and bandwidth using the method described in [16] with the Gaussian divergence distance metric and a minimum occupancy constraint of 40 seconds.

After global 1-best MLLR adaptation, phone-marked lattices were generated. Using these lattices, 4 iterations of lattice MLLR [17]

	eval98	eval02
unadapted (GI MPE)	15.0	13.6
1-best MLLR	14.2	12.5
lat-MLLR 2trans	13.9	12.3
lat-MLLR 2trans+FV	13.7	12.0
lat-MLLR 4trans+FV	13.6	12.1
lat-MLLR 8trans+FV	13.6	12.0

Table 4. %WER for eval98 & eval02 after adaptation based on the GI MPE model.

	MPE-MAP	SAT
1-best MLLR	14.1	13.4
lat-MLLR 2trans	13.8	13.4
lat-MLLR 2trans+FV	13.6	13.0

Table 5. %WER of SAT models on dev03 in comparison with adaptation results based on GD models. Supervisions for 1-best MLLR were obtained from 4-gram expansion after unadapted single pass decoding using GD MPE-MAP models and trigram.

were performed. On each iteration the number of adaptation transforms was increased using a regression-class tree [8] subject to a threshold on the amount of data per transform. Up to 8 MLLR speech transforms and a global full-variance (FV) transform [4] were estimated. As shown in Table 4, the WER was reduced by 9.3% relative on eval98, and 11.8% on eval02. There were no consistent gains from using more than 2 transforms.

5.2. Speaker adaptive training

Starting from the HLDA ML estimated models, speaker adaptive training (SAT) using constrained MLLR [4] with the same transformation for both the means and variances was applied. Global full-matrix constrained (feature-space) MLLR transforms were estimated for each speaker (one transform for silence, another for speech). These transforms were applied to the acoustic training data during re-estimation.

Starting with the HLDA models with variable number of Gaussians, five iterations of interleaved transform estimation and ML parameter updates were performed. The transforms were then fixed and used with six iterations of MPE training to obtain SAT models. The denominator lattices generated for the previous MPE training were used (without lattice re-generation).

The results in Table 5 show that the SAT model outperformed MPE-MAP models on dev03 after 1-best MLLR and lattice MLLR with two transforms and full variance transforms.[1]

6. SINGLE PRONUNCIATION (SPRON)

SPRON dictionaries for training and testing were generated by selecting pronunciation variants from the multiple pronunciation dictionary using the probabilistic method described in [6]. Here

[1]Experimental results here were obtained with a preliminary version of the 2003 4-gram LM which did not include more recent Broadcast News text data. Also, since we only had a wideband SAT model, NB results from MPE-MAP 1-best MLLR were used to calculate %WER.

Models	MPRON	SPRON
ML	20.2	19.7
HLDA, VarMix, MPE	15.3	14.8

Table 6. %WER on dev03 using ML and MPE triphone models with multiple (MPRON) and single (SPRON) pronunciation dictionaries. New trigram LM was used that is presented in section 7.

the necessary pronunciation statistics were obtained from alignment of the Switchboard and Broadcast News training corpora. The SPRON dictionaries were used to train bandwidth-specific, GD triphone acoustic models in the same fashion as described before, including the regeneration of phonetic decision trees. The same word lattices as in MPRON training were used. Four MPE iterations were performed using the denominator lattices generated with the ML models and a further 3 iterations using a combination of the lattices generated with ML and MPE models.

Table 6 shows results using unadapted single pass decoding with GI wide-band triphone models and a trigram language model. For both ML and MPE-HLDA models the improvement was 0.5% absolute. An additional experiment comparing GD versions of the models gave 13.9%, which again is 0.5% absolute better than using the standard multi-pronunciation dictionary model.

7. LANGUAGE MODEL

A 59k entry wordlist was chosen from the most frequent words in the training texts listed in section 3.3 using a weighted sum of frequencies from various subsets of the training corpus. The weights were chosen to minimize the out-of-vocabulary (OOV) rate on the dev03 transcriptions. The resulting vocabulary yields an OOV rate of 0.47% on dev03.

Word-based 4-gram language models were built for each of the 5 data sources separately. All word-based models were merged to form a single model, where the interpolation weights were computed to minimise perplexity. After merging, the resulting language model was pruned [15] to 8.8M bigrams, 12.7M trigrams and 6.6M 4-grams.

A class-based trigram language model was trained which used 1000 classes that were automatically derived based on word bigram statistics [11]. The model contained 0.8M bigrams and 10M trigrams. Finally, the word-based model was interpolated with the class-based model.

Perplexities and WERs on dev03 with the word-based trigram (tg), the word-based 4-gram (fg), and the interpolated word-based 4-gram with the class-based trigram (fgic) are given in Table 7. The WERs for fg and fgic were obtained using lattice re-scoring based on tg lattices. The modified MPE & MPE-MAP models in section 4.5 were used as GI & GD models.

8. RT-03 BN-E EVALUATION SYSTEM

The system structure is shown in Figure 2, and more technical details about fast system design can be found in a companion paper [2].

LM type	Perplexity	%WER	
		GI	GD
tg98	-	16.6	16.1
tg	140.9	14.9	14.4
fg	121.5	14.0	13.5
fgic	119.1	13.7	13.2

Table 7. Perplexities and %WERs on dev03 with various LMs. tg98 is the trigram from the 1998 CU-HTK BN-E system.

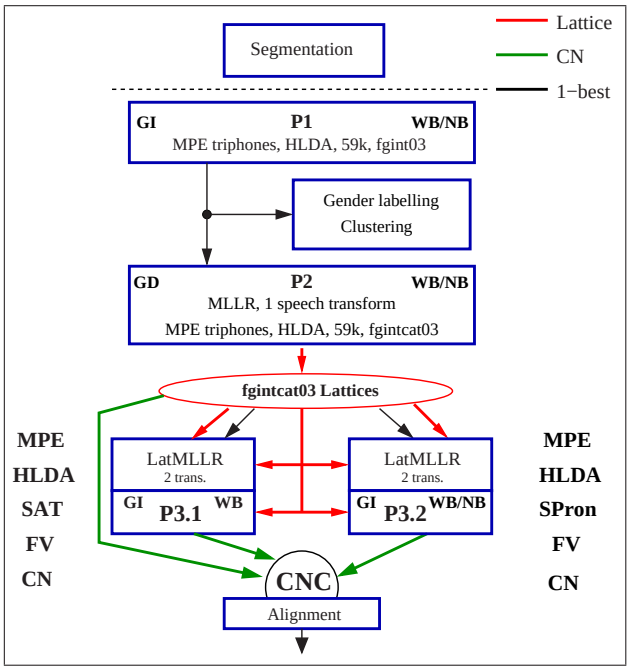

Fig. 2. BN-E evaluation system structure.

8.1. Segmentation

Automatic segmentation was performed using a system similar to that used in the 1998 CU-HTK BN-E 10×RT system [12]. For the RT-03 evaluation system a new music model was built incorporating TDT-4 data, and the clustering/merging procedures within the segmenter were changed to increase a segment purity measure on eval02 data [16].

8.2. Decoding passes

Recognition runs in a number of passes and uses time-synchronous one-pass cross-word triphone decoders. The initial transcription and the lattice generation passes employed a decoder based on that used in [12], and the lattice rescoring passes used the HTK based decoder HDecode.

8.2.1. Pass1: initial transcription

The first pass generates an initial transcription of the data using GI triphone HMMs (MPE) and the word-based trigram with very tight beamwidths. The output trigram lattices are rescored with the 4-gram language model. All segments are gender-labelled by forced-alignment of this transcription with GD HMMs (MPE-MAP). The segments are then grouped gender and bandwidth dependently into clusters comprising at least 40 seconds of data for adaptation purposes in the following passes.[2]

8.2.2. Pass2: lattice generation

Bandwidth-specific, GD triphone HMMs (MPE-MAP) were adapted using transforms estimated based on global least squares regression and MLLR variance transforms [3] with the initial Pass1 transcription as supervision. The data is decoded with the word-based trigram at relatively conservative beamwidths yielding a lattice for each segment. These lattices are expanded using the interpolation of the word-based 4-gram and the class-based trigram.

8.2.3. Pass3: lattice rescoring

Two different models, SAT (Pass3.1) and SPRON (Pass3.2), were used for lattice rescoring. Each model was adapted using a global 1-best MLLR transform and then used for model-marked lattice generation. Based on these model-marked lattices, the following transforms were estimated in stages using lattice MLLR: a global MLLR transform; a full variance transform and upto 2 speech MLLR transforms per cluster. The adapted models were used to rescore the word lattices from Pass2.

8.2.4. Confusion networks and combination (CNC)

In each case the lattice output was converted to a confusion network [10] for later system combination. The word lattices produced by the Viterbi decoder were used to generate confusion networks, which provide a compact representation of the most likely word hypotheses and their associated word posterior probabilities.

The confusion networks produced in Pass2, Pass3.1 and Pass3.2 were combined using a dynamic programming procedure that employed the full set of alternative hypotheses and their posteriors to find the optimal alignment of the outputs from the different stages [1]. Given this alignment the final overall system hypothesis was chosen based on the posterior distribution represented by the corresponding confusion network segments. For this final hypothesis the corresponding word-level confidence scores were generated.

8.3. Performance

The results on the dev03 and eval03 test sets for each of these stages are shown in Table 8. Pass1 ran in 0.9xRT including data coding and segmentation. Very tight beamwidths for fast processing gave 1.6% of absolute loss on dev03 compared to the numbers obtained with the development setup in section 4. GD models adapted using a global 1-best MLLR in Pass2 gave 17-18% relative reduction in WER over Pass1. Lattice MLLR and lattice re-scoring with the SAT model and SPRON system showed clear gains over the Pass2 results, though the gain on eval03 was rather smaller than that on dev03. The CNC effectively combined three different systems and gave another gain. After CNC the WER on

[2] Various minimum occupancy thresholds were tested for adaptation experiments from 25s to 40s in the framework of the RT-03 evaluation system, and it was found that WERs from different thresholds were almost the same after 1-best MLLR. As we use more transformations for lattice-based MLLR, we selected 40s as the threshold.

	dev03	eval03	xRT
Coding & segment	-	-	0.3
Pass1	15.9	14.6	0.6
Pass2	13.1	11.9	3.7
Pass3.1	12.0	11.4	2.5
Pass3.2	12.2	11.4	1.9
CNC	11.6	10.7	0.1

Table 8. %WER on dev03 and eval03 and processing time on eval03 for the RT-03 evaluation system. The system runs on a single processor of a IBM x335 computer with a 2.8GHz Intel Xeon processor/400MHz FSB.

dev03 and eval03 were 11.6% and 10.7%, respectively. The full system on eval03 ran in 9.1xRT and the confidence scores had a Normalised Cross Entropy (NCE) of 0.412.

9. CONCLUSIONS

This paper has described the development and performance of the 2003 CU-HTK BN-E transcription system. Many useful techniques, including HLDA, MPE training and lattice-based adaptation, have been successfully applied to the Broadcast News transcription task for the first time. Furthermore, a number of new techniques were used including MAP-style GD discriminative training (MPE-MAP) and modified lattice-based discriminative training. The evaluation system was carefully designed to meet the 10×RT time restriction of the primary condition of the RT-03 BN-E evaluation while still including a number of stages of decoding, lattice-based adaptation and system combination. On the RT-03 current test set evaluation data the system gave an overall error rate of 10.7%, the lowest error rate in the evaluation.

10. ACKNOWLEDGMENTS

This work was supported by DARPA grant MDA972-02-1-0013 under the EARS program. 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 all of the member of the HTK STT team, in particular X. Liu, D. Povey, M.J.F. Gales, H.Y.Chan and K. Yu.

11. REFERENCES

[1] G. Evermann & P.C. Woodland (2000). "Posterior Probability Decoding, Confidence Estimation and System Combination." *Proc. Speech Transcription Workshop*, College Park, MD.

[2] G. Evermann & P.C. Woodland (2003). "Design of Fast LVCSR Systems." *Proc. ASRU'03*, St. Thomas.

[3] M.J.F. Gales & P.C. Woodland (1996). "Mean and Variance Adaptation within the MLLR Framework." *Computer Speech & Language*, Vol. 10, pp. 249-264.

[4] M.J.F. Gales (1998). "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition." *Computer speech & language*, Vol. 12, pp. 75-98.

[5] D. Graff (2002). "An Overview of Broadcast News Corpora." *Speech Communication*, Vol.37, pp. 15-26.

[6] T. Hain (2002). "Implicit Pronunciation Modelling in ASR." *ITRW PMLA 2002* (Invited Short Lecture), Estes Park, CO.

[7] N. Kumar (1997). "Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition." Ph.D. Thesis, Johns Hopkins University, Baltimore, MD.

[8] C.J. Leggetter & P.C. Woodland (1995). "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression." *Proc. Eurospeech'95*, pp. 1155–1158, Madrid, Spain.

[9] X. Liu, M.J.F. Gales & P.C. Woodland (2003). "Automatic Complexity Control for HLDA Systems." *Proc. ICASSP'03*, pp. I-132-135, Hong Kong.

[10] L. Mangu, E. Brill & A. Stolcke (1999). "Finding Consensus Among Words: Lattice-Based Word Error Minimization." *Proc. Eurospeech'99*, pp. 495–498, Budapest, Hungary.

[11] T.R. Niesler, E.W.D. Whittaker & P.C. Woodland (1998). "Comparison of Part-of-Speech and Automatically Derived Category-based Language Models for Speech Recognition." *Proc. ICASSP'98*, pp. 177-180, Seattle, WA.

[12] J.J. Odell, P.C. Woodland & T. Hain (1998). "The CUHTK-Entropic 10×RT Broadcast News Transcription System." *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 271-275, Lansdowne, VA.

[13] D. Povey & P.C. Woodland (2002). "Minimum Phone Error and I-Smoothing for Improved Discriminative Training." *Proc. ICASSP'02*, pp. I-105-108, Orlando, FL.

[14] D. Povey, P.C. Woodland & M.J.F. Gales (2003). "Discriminative MAP for Acoustic Model Adaptation." *Proc. ICASSP'03*, pp. I-312-315, Hong Kong.

[15] A. Stolcke (1998). "Entropy-based Pruning of Backoff Language Models." *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp.270-274, Lansdowne, VA.

[16] S.E. Tranter, K. Yu, D.A. Reynolds, G. Evermann, D.Y. Kim, & P.C. Woodland (2003). "An Investigation into the Interactions between Speaker Diarisation Systems and Automatic Speech Transcription." *Tech. Report*, Cambridge University, CUED/F-INFENG/TR-464.

[17] L.F. Uebel & P.C. Woodland (2001). "Speaker Adaptation Using Lattice-Based MLLR." *Proc. ISCA ITRW on Adaptation Methods in Speech Recognition*, pp.57-60, Sophia-Antipolis, France.

[18] P.C. Woodland (2002). "The Development of the HTK Broadcast News Transcription System: An Overview." *Speech Communication*, Vol.37, pp. 291-299.

[19] P.C. Woodland & D. Povey (2002). "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition." *Computer Speech and Language*, Vol. 16 No. 1, pp. 25-47.