THE CU-HTK MARCH 2000 HUB5E TRANSCRIPTION SYSTEM

T. Hain, P.C. Woodland, G. Evermann & D. Povey

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
e-mail: {th223,pcw,ge204,dp10006}@eng.cam.ac.uk

ABSTRACT

This paper describes the Cambridge University HTK (CU-HTK) system developed for the NIST March 2000 evaluation of English conversational telephone speech transcription (Hub5E). A range of new features have been added to the HTK system used in the 1998 Hub5 evaluation, and the changes taken together have resulted in an 11% relative decrease in word error rate on the 1998 evaluation test set. Major changes include the use of maximum mutual information estimation in training as well as conventional maximum likelihood estimation; the use of a full variance transform for adaptation; the inclusion of unigram pronunciation probabilities; and word-level posterior probability estimation using confusion networks for use in minimum word error rate decoding, confidence score estimation and system combination. On the March 2000 Hub5 evaluation set the CU-HTK system gave an overall word error rate of 25.4%, which was the best performance by a statistically significant margin. This paper describes the new system features and gives the results of each processing stage for both the 1998 and 2000 evaluation sets.

1 INTRODUCTION

The transcription of conversational telephone speech is one of the most challenging tasks for speech recognition technology with state-of-the-art systems yielding high word error rates. The primary focus for research and development of such systems for US English has been the Switchboard/Call Home English corpora along with the regular NIST "Hub5" evaluations. The Cambridge University HTK (CU-HTK) Hub5 system has been developed over several years. This paper describes changes to the September 1998 Hub5 evaluation system [6] made while developing the March 2000 system.

Major system changes include the use of HMMs trained using maximum mutual information estimation (MMIE) in addition to standard maximum likelihood estimation (MLE); the use of pronunciation probabilities; improved speaker/channel adaptation using a global full variance transform; soft-tying of states for the MLE based acoustic models; and the use of confusion networks for minimum word error rate decoding, confidence score estimation and system combination. All of these features made a significant contribution to the word error rate improvements of the complete system. In addition, several minor changes have been made and these include the use of additional training data and revised transcriptions; acoustic data weighting; and an increased vocabulary size.

The rest of the paper is arranged as follows. First an overview of the 1998 HTK system is given. This is followed by a description of the data sets used in the experiments and then by sections that discuss each of the major new features of the system. Finally the complete March 2000 evaluation system is described and the results of each stage of processing presented.

2 OVERVIEW OF 1998 HTK SYSTEM

The HTK system used in the 1998 Hub5 evaluation served as the basis for development. In this section a short overview of its features is given (see [6] for details).

The system uses perceptual linear prediction cepstral coefficients derived from a mel-scale filterbank (MF-PLP) [18] covering the frequency range from 125Hz to 3.8kHz. A total of 13 coefficients, including c_0, and their first and second order derivatives were used. Cepstral mean subtraction and variance normalisation are performed for each conversation side. Vocal tract length normalisation (VTLN) was applied in both training and test.

The acoustic modelling used cross-word triphone and quinphone hidden Markov models (HMMs) trained using conventional maximum likelihood estimation. Decision tree state clustering [20] was used to select a set of context-dependent equivalence classes. Mixture Gaussian distributions for each tied state were then trained using sentence-level Baum-Welch estimation and iterative mixture splitting [20]. After gender independent (GI) models had been trained, a final training iteration using gender-specific training data and updating only the means and mixture weights was performed to estimate gender dependent (GD) model sets. The triphone models were phone position independent, while the quinphone models included questions about word boundaries as well as ± 2 phone context. The HMMs were trained on 180 hours of Hub5 training data.

The system used a 27k vocabulary that covered all words in the acoustic training data. The core of the pronunciation dictionary was based on the 1993 LIMSI WSJ lexicon, but used a large number of additions along with various changes. The system used N-gram word-level language models. These were constructed by training separate models for transcriptions of the Hub5 acoustic training data and for Broadcast News data and then merging the resultant language models to effectively interpolate the component N-grams. The word-level 4-grams used were smoothed with a class-based trigram model using automatically derived classes [12].

The decoding was performed in stages with successively more complex acoustic and language model being applied in later stages. Initial passes were used for test-data warp factor selection, gender determination and finding an initial word string for unsupervised mean and variance maximum likelihood linear regression (MLLR) adaptation [8, 3]. Word-level lattices were then created using adapted triphone HMMs and a bigram model which were expanded to included the full 4-gram and class model probabilities. Iterative MLLR [17] was then applied using quinphone models and confidence scores estimated using an N-best homogeneity measure for both the triphone and quinphone output. The final stage combined these two transcriptions using the ROVER program [2]. The system gave a 39.5% word error rate on the September 1998 evaluation data.

3 TRAINING AND TEST DATA

The Hub5 acoustic training data is from two corpora: Switchboard-1 (Swb1) and Call Home English (CHE). The 180 hour training set used for training the 1998 HTK system used various sources of Swbd1 transcriptions and turn-level segmentations. For the March 2000 system we took advantage of the January 2000 release from Mississippi State University (MSU) of Swbd1 transcriptions which should provide greater accuracy and consistency. We made a number of changes to these manual corrections and also automatically removed more than 30 hours of silence data at segment boundaries. An important feature of the MSU transcripts is the full-word transcription of false starts and mispronunciations. In order to make use of the extended transcripts a dictionary of false starts and mispronunciations was created for use during training.

Three different training sets were used during the course of development: the 18 hour Minitrain set defined by BBN which gives a fast turnaround; the full 265 hour training set (h5train00) for the the March 2000 system and a subset of h5train00 denoted h5train00sub. The sizes of the training sets are given in Table 1 together with the number of conversation sides that each includes. The h5train00sub set was chosen to include all the speakers from Swb1 in h5train00 as well as a subset of the available CHE sides.

Training Set	Total Time (hrs)	Conversation Sides	
		Swb1	CHE
Minitrain	18	398	–
h5train00sub	68	862	92
h5train00	265	4482	235

Table 1: Hub5 training sets used.

The development test sets used were the subset of the 1997 Hub5 evaluation set used in [6], eval97sub, containing 10 conversation sides of Switchboard-2 (Swb2) data and 10 of CHE; and the 1998 evaluation data set, eval98, containing 40 sides of Swb2 and 40 CHE sides (in total about 3 hours of data). Furthermore results are given for the March 2000 evaluation data set, eval00, which has 40 sides of Swb1 and 40 CHE sides.

Training Set	Clustered States / Gaussians per State	% Word Error Rate		
		Swbd2	CHE	Total
Minitrain	3088 / 12	43.7	57.7	50.6
h5train00sub	6165 / 12	38.7	53.5	46.0
h5train00	6165 / 16	36.4	52.5	44.4

Table 2: % WER on eval97sub using VTLN, GI, MLE triphone models and a trigram language model, different training set sizes.

Basic gender independent, cross-word triphone versions of the system, with no adaptation, were constructed for each training set size. Table 2 shows the number of clustered speech states and the number of Gaussians per state for each of these systems as well as word error rates on eval97sub. An initial 3.5-fold increase in the amount of training data results in a 4.6% absolute reduction in word error rate (WER). However some of this large gain can be attributed to the careful selection of the h5train00sub set to have a good coverage of the full training material. A further approximately 3-fold increase in the amount of training data only brings a further 1.6% absolute reduction in WER.

4 MMIE TRAINING

The model parameters in HMM based speech recognition systems are normally estimated using Maximum Likelihood Estimation (MLE). During MLE training, model parameters are adjusted to increase the likelihood of the word strings corresponding to the training utterances without taking account of the probability of other possible word strings. In contrast to MLE, discriminative training schemes, such as Maximum Mutual Information Estimation (MMIE) take account of possible competing word hypotheses and try to reduce the probability of incorrect hypotheses. The objective function to maximise in MMIE is the posterior probability of the true word transcriptions given the training data.

For R training observation sequences $\{\mathcal{O}_1, \ldots, \mathcal{O}_r, \ldots \mathcal{O}_R\}$ with corresponding transcriptions $\{w_r\}$, the MMIE objective function is given by

$$\mathcal{F}_{\mathrm{MMIE}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\mathcal{O}_r|\mathcal{M}_{w_r})P(w_r)}{\sum_{\hat{w}} p_\lambda(\mathcal{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w})} \quad (1)$$

where $\mathcal{M}_w$ is the composite model corresponding to the word sequence w and $P(w)$ is the probability of this sequence as determined by the language model. The summation in the denominator of (1) is taken over all possible word sequences $\hat{w}$ allowed in the task and it can be replaced by

$$p_\lambda(\mathcal{O}_r|\mathcal{M}_{\mathrm{den}}) = \sum_{\hat{w}} p_\lambda(\mathcal{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w}) \quad (2)$$

where $\mathcal{M}_{\mathrm{den}}$ encodes the full recognition acoustic/language model.

Normally the denominator of (1) requires a full recognition pass to evaluate on each iteration of training. However as discussed in [16] this can be approximated by using a word lattice which is generated once to constrain the number of word sequences considered. This lattice-based framework can be used to generate the necessary statistics to apply the Extended-Baum Welch (EBW) algorithm [5, 13, 16] to iteratively update the model parameters. The statistics required for EBW can be gathered by performing for each training utterance a forward-backward pass on the lattice corresponding to the numerator of (1) (i.e. the correct transcription) and on the recognition lattice for the denominator of (1). The implementation we have used is rather different to the one in [16] and does a full forward-backward pass constrained by (a margin around) the phone boundary times that make up each lattice arc. Furthermore the smoothing constant in the EBW equations is computed on a per-Gaussian basis for fast convergence and a novel weight update formulation used. The computational methods that we have adopted for Hub5 MMIE training are discussed in detail in [19].

While MMIE is very effective at reducing training set error a key issue is generalisation to test data. It is very important that the confusable data generated during training (as found from the posterior distribution of state occupancy for the recognition lattice) is representative to ensure good generalisation. If the posterior distribution is broadened, then generalisation performance can be improved. For this work, two methods were investigated: the use of acoustic scaling and a weakened language model.

Normally the language model probability and the acoustic model likelihoods are combined by scaling the language model log probabilities. This situation leads to a very large dynamic range in the combined likelihoods and a very sharp posterior distribution in the denominator of (1). An alternative is to scale down the acoustic model log likelihoods and as shown in [19] this acoustic scaling aids

generalisation performance. Furthermore, it is important to enhance the discrimination of the acoustic models without overly relying on the language model to resolve difficulties. Therefore as suggested in [15] a unigram language model was used during MMIE training which also improves generalisation performance [19].

Experiments reported in [19] show that MMIE is effective for a range of training set sizes and model types. Table 3 shows word error rates using triphone HMMs trained on h5train00. These experiments required the generation of numerator and denominator lattices for each of the 267,611 training segments. It was found that two iterations of MMIE re-estimation gave the best test-set performance [19]. Comparing the lines in Table 3 show that, without data weighting, the overall error rate reduction from MMIE training is 2.6% absolute on eval97sub and 2.7% absolute on eval98.

Iteration	eval97sub			eval98		
	Swbd2	CHE	Total	Swbd2	CHE	Total
MLE	36.4	52.5	44.4	42.6	48.6	45.6
MLE/w	35.7	51.8	43.7	42.5	47.7	45.1
1	34.2	50.7	42.4	40.9	46.5	43.7
1/w	34.0	50.2	42.0	40.7	46.2	43.5
2	33.6	50.1	41.8	40.3	45.4	42.9
2/w	33.8	50.0	41.9	40.3	45.1	42.7

Table 3: %WER on eval97sub and eval98 using VTLN GI triphone models and a trigram language model. (w) denotes data weighting.

The table also shows the effect of giving a factor of three weighting to the CHE training data.[1] This reduced the error rate for the MLE models by 0.5% to 0.7% absolute, but has a much smaller beneficial effect for MMIE trained models. This is probably because while MLE training gives equal weight to all training utterances, MMIE training effectively gives greater weight to those training set utterances with low sentence posterior probabilities for the correct utterance.

MMIE was also used to train quinphone HMMs. The gain from MMIE training for quinphone HMMs was 1.9% absolute on eval97sub from a quinphone MLE system using acoustic data weighting. As shown in [19] the gains from MLLR adaptation are as great for MMIE models as for MLE trained models. Hence the primary acoustic models used in the March 2000 CU-HTK evaluation system used gender-independent MMIE trained HMMs.

5 SOFT-TYING

Soft tying of states [10] allows Gaussians from a particular state, corresponding to a decision tree leaf node, to be also used in other mixture distributions with similar acoustics. Previously, using an implementation from JHU, the technique was investigated using various training set sizes and levels of model complexity [7]. It was found that while consistent improvements were obtained, the improvement in WER was reduced when features such as VTLN and MLLR adaptation were included in the system.

For the March 2000 system, a revised and somewhat simplified implementation of soft-tying was investigated. For a given model set a single Gaussian per state version was created. For each speech state in the single Gaussian system, the nearest two other states were found using a log-overlap distance metric [14], which calculates the distance between two Gaussians as the area of overlap of the two probability density functions. All of the mixture components from the two nearest states and the original state of the original mixture Gaussian HMM are then used in a mixture distribution for the state. Thus the complete soft-tied system has the same number of Gaussians as the original system and three times as many mixture weights per state. After this revised structure has been created all system parameters are re-estimated. This approach allows the construction of both soft-tied triphone and quinphone systems in a straightforward manner.

System Type	Triphones			Quinphones		
	Swbd2	CHE	Total	Swbd2	CHE	Total
GI	42.5	47.7	45.1	42.1	47.3	44.7
ST/GI	42.1	47.4	44.8	41.5	46.9	44.2
ST/GD	41.4	47.0	44.2	41.0	46.1	43.6
ST/GD/PP	40.1	45.5	42.8	39.2	44.6	41.9

Table 4: WER on eval98 using VTLN GI triphone/quinphone models trained on h5train00 (3x CHE) and a trigram LM. ST denotes soft-tied models and PP the use of pronunciation probabilities.

The results of using soft-tied (ST) triphone and quinphone systems on eval98 is shown in Table 4 when data weighting is used.[2]

There is a reduction in WER of 0.3% absolute for triphones and 0.5% for quinphones and a further 0.6% absolute from using GD models. So far, soft-tying has only been used with MLE training, although the technique could also be applied to MMIE trained models.

6 PRONUNCIATION PROBABILITIES

The pronunciation dictionary used in this task contains on average 1.1 to 1.2 pronunciations per word. Unigram pronunciation probabilities, that is the probability of a certain pronunciation variant for a particular word, were estimated based on an alignment of the training data. If words were not seen in the training data a uniform distribution over all pronunciation variants is assumed. However, this straight-forward implementation only brought moderate improvements in WER.

The dictionaries in the HTK system explicitly contain silence models as part of a pronunciation. Experiments with or without inclusion of silence into the probability estimates were conducted [7]. The most successful scheme used three separate dictionary entries for each real pronunciation which differed by the word-end silence type: a no silence version; adding a short pause preserving cross-word context; and a general silence model altering context. The unigram "pronunciation" probability is found separately for each of these entries and the distributions are smoothed with the overall silence distributions. Finally all dictionary probabilities are renormalised so that the pronunciation for each word which has the highest probability is set to one. During recognition the (log) pronunciation probabilities are scaled by the same factor as used for the language model.

Table 4 shows that the use of pronunciation probabilities gives a reduction in WER of 1.4-1.7% absolute on eval98. On other test sets improvements greater than 1% absolute have also been obtained and

[1] The test set is balanced across Switchboard and Call Home data but the training set isn't and so data weighting attempts to partially correct this imbalance.

[2] We have found that the use of acoustic data weighting reduces the beneficial effect of soft-tying.

size of the gains is found to be fairly independent of the complexity of the underlying system.

7 FULL VARIANCE TRANSFORMS

A side-dependent block-full variance (FV) transformation [4], H, of the form $\hat{\Sigma} = H\Sigma H^T$ was investigated. This can be viewed as the use of a speaker-dependent global semi-tied block-full covariance matrix and can be efficiently implemented by transforming both the means and the input data. In our implementation, the full variance transform was computed after standard mean and variance maximum likelihood linear regression (MLLR). Typically a WER reduction of 0.5% to 0.8% was obtained. However as a side effect, we found that there were reduced benefits from multiple iterations of MLLR when used with a full variance transform.

8 CONFUSION NETWORKS

Confusion networks allow estimates of word posterior probabilities to be obtained. For each link in a particular word lattice (from standard decoding) a posterior probability is estimated using the forward-backward algorithm. The lattice with these posteriors is then transformed into a linear graph, or confusion network (CN), using a link clustering procedure [11]. This graph consists of a sequence of so called confusion sets, which contain competing single word hypotheses with associated posterior probabilities. A path through the graph is found by choosing one of the alternatives from each confusion set. By picking the word with the highest posterior from each set the sentence hypothesis with the lowest overall expected word error rate can be found. This hypothesis is generally more accurate than the one chosen by the normal Viterbi decoding, which minimises the *sentence* error rate.

The estimates of the word posterior probabilities encoded in the confusion networks can be used directly as confidence scores (which are essentially word-level posteriors), but they tend to be over-estimates of the true posteriors. This effect is due to the assumption that the word lattices represent the relevant part of the search space. While they contain the most-likely paths, a significant part of the "tail" of the overall posterior distribution is missing. To compensate for this a decision tree was trained to map the estimates to confidence scores.

The confusion networks with their associated word posterior estimates were also used in an improved system combination scheme. Previously the ROVER technique introduced in [2] had been used to combine the 1-best output of multiple systems. Confusion network combination (CNC) can be seen as a generalisation of ROVER to confusion networks, i.e. it uses the competing word hypotheses and their posteriors encoded in the confusion sets instead of only considering the most likely word hypothesised by each system.

A more detailed description of the use of word posterior probabilities and their application to the Hub5 task can be found in [1].

9 MARCH 2000 EVALUATION SYSTEM

This section gives an overview of the complete system as used in the March 2000 evaluation. The system operates in multiple passes through the data: initial passes are used to generate word lattices and then these lattices are rescored using four different sets of adapted acoustic models. The final system output comes from combining the confusion networks from each of these re-scoring passes. While this architecture results in a complex overall system, this section also reports the results of each of the stages. This allows the performance of many system variants at different levels of complexity to be assessed.

9.1 Acoustic Models

The VTLN acoustic models used in the system were either triphones (6165 speech states/16 Gaussians per state) or quinphones (9640 states/16 Gaussians per state) trained on h5train00. More details on the performance of these models was given in previous sections. It should be emphasised that the MMIE models were all gender independent while the MLE VTLN models were all gender dependent using soft-tying. All the acoustic models used Call Home weighting.

9.2 Word List & Language Models

The word list was taken from two sources: the 1998 27k word list [6] and the most frequent 50,000 words occurring in the 204 million words of broadcast news (BN) LM training data. This gave a new word list with 54,537 words where most of the pronunciations were already available in our broadcast news (Hub4) dictionary. The 54k wordlist reduced the out-of-vocabulary (OOV) rate on eval98 from 0.94% to 0.38%. After the March 2000 evaluation it was found that using the 54k dictionary gave an OOV rate of 0.30% on eval00 compared to 0.69% if the 27k dictionary had been used.

The use of the MSU Swb1 training transcriptions for language modelling purposes raised certain issues. First, the average sentence length was 11.3 words compared to 9.5 words on the LDC transcripts that we previously used. This has the effect that LMs trained on the MSU transcripts have a higher test-set perplexity which is mainly due to the reduced probability of the sentence-end symbol. Since it was not known if LDC-style or MSU-style training transcripts would be more appropriate, both sets of data were used along with the broadcast news data. Bigram, trigram and 4-gram LMs were trained on each data set (LDC Hub5, MSU Hub5, BN) and merged to form an effective 3-way interpolation. Furthermore, as described in [6] a class-based trigram model using 400 automatically generated word classes [12, 9] was built to smooth the merged 4-gram language model by a further interpolation step to form the language model used in lattice rescoring.

9.3 Stages of Processing

The first three passes through the data (P1–P3) are used to generate word lattices. First P1 (GI non-VTLN MLE triphones, trigram LM, 27k dictionary), generated an initial transcription. This P1 pass is identical to the 1998 P1 setup [6]. The P1 output was used solely for VTLN warp-factor generation and assignment of a gender label for each test conversation side. All subsequent passes used the 54k dictionary and VTLN-warped test data. Stage P2 used MMIE GI triphones to generate the transcriptions for unsupervised test-set MLLR adaptation [8, 3] with a 4-gram LM. A global transform[3] for the means (block-diagonal) and variances (diagonal) was computed for each side. In stage P3, the actual word lattices were generated using the adapted GI MMIE triphones and a bigram language model. These lattices were expanded expanded to contain language model probabilities generated by the interpolation of the word 4-gram and the class trigram.

[3] A "global transform" denotes one transform for speech and a separate transform for silence.

Subsequent passes rescored these lattices and operated in two branches: a branch using GI MMIE trained models (branch "a") and a branch using GD, soft-tied, MLE models (branch "b"). Stage P4a/P4b used triphone models with standard global MLLR, a FV transform, pronunciation probabilities and confusion network decoding. The output of the respective branches served as the adaptation supervision to stage P5a/P5b. These were as P4a/P4b but were based on quinphone acoustic models. Finally for the MMIE branch only, a pass with two MLLR transforms was run (P6a). The final system word output and confidence scores was found by using CNC with the confusion networks from P4a, P4b, P6a and P5b.

9.4 System Results on Eval98

Table 5 gives results for each processing stage for the 1998 evaluation set. The large difference (6.8% absolute in WER) between the P1 and P2 results is due to the combined effects of VTLN, MMIE models on the new training set, the larger vocabulary and a 4-gram LM. MLLR adaptation and the smoothing from a class LM results in a further reduction in WER of 2.5% absolute. The second adaptation stage which includes MLLR and a full variance transform (FV), pronunciation probabilities and confusion network decoding reduces the WER by a further 2.9% absolute (P4a), which is 0.8% absolute better than the result of the corresponding MLE soft-tied GD triphone models (P4b).

Stage	Swbd2	CHE	Total	NCE
P1	47.0	51.6	49.3	
P2	40.0	44.9	42.5	
P3	37.5	42.4	40.0	
P4a no FV/CN	36.2	41.4	38.8	
P4a no CN	35.8	40.8	38.3	
P4a	34.5	39.6	37.1	0.238
P4b no FV/CN	37.1	42.2	39.7	
P4b no CN	36.8	41.3	39.0	
P4b	35.5	40.3	37.9	0.235
P5a no CN	35.2	39.5	37.4	
P5a	33.9	38.4	36.2	0.232
P5b no CN	35.6	40.7	38.1	
P5b	34.5	39.5	37.0	0.229
P6a no CN	34.6	39.2	36.9	
P6a	33.6	38.4	36.0	0.224
FINAL/ROVER	32.8	38.0	35.4	
FINAL/CNC	32.5	37.4	35.0	0.225

Table 5: % WER and normalised cross entropy (NCE) values on eval98 for all stages of the evaluation system. The final system output is a combination of P4a,P4b,P6a and P5b. "no FV" denotes system output without full variance transform. "no CN" denotes standard output rather than minimum word error rate output.

The use of quinphone models instead of triphone models gives a further gain of 0.9% for both branches. Whereas the second adaptation stage with two speech transforms for the quinphone MMIE models brings 0.5%, after obtaining CN output the difference is only 0.2%. The final result after 4-fold system combination is 35.0%. This is an 11% reduction in WER relative to the CU-HTK evaluation result obtained on the same data set in 1998 (39.5%).

Note that confusion network output consistently improves performance by about 1% absolute and that combination of the 4 outputs using confusion network combination (CNC) is 0.4% absolute better than using the ROVER approach. Then confidence scores based on confusion networks give an improved normalised cross entropy (NCE) of 0.225 compared to 0.145 from the 1998 CU-HTK evaluation system which used N-best homogeneity based confidence scores.

9.5 March 2000 Evaluation Data Results

Table 6 lists the evaluation system performance on the March 2000 evaluation set. The performance on eval00 gives a similar per stage improvement to that obtained for eval98. However the absolute WER levels are reduced by about 10% absolute.[4]

Stage	Swbd2	CHE	Total	NCE
P1	31.7	45.4	38.6	
P2	25.5	38.1	31.8	
P3	22.9	35.7	29.3	
P4a	20.9	33.5	27.2	0.294
P4b	21.9	33.7	27.8	0.287
P5a	20.3	32.7	26.6	
P5b	21.0	32.8	26.9	0.292
P6a	20.3	32.6	26.5	0.284
P4b+P5b/CNC	20.6	32.4	26.5	0.285
P4a+P6a/CNC	19.5	31.7	25.6	0.278
P4a+P4b+P6a+P5b/CNC	19.3	31.4	25.4	0.271

Table 6: % WER and normalised cross entropy on eval00 for each stage of the CU-HTK Hub5E 2000 evaluation system.

It was again found that there is a fairly consistent 1% absolute reduction in WER from confusion networks. A contrast (not shown in the table) showed that on P2 the use of MMIE models had given a 2.1% absolute reduction in WER over the corresponding MLE models. The combination P4a+P6a denotes a system where only MMIE trained models have been used for decoding which yields a result 0.9% absolute better than the corresponding MLE combination (P5b+P4b). However, the inclusion of the MLE system outputs gives a 0.2% WER absolute improvement. The final error rate from the system (25.4%) was lowest in the evaluation by a statistically significant margin.

9.6 Pure MLE Contrast

A further run on eval98 was performed to investigate the effect of using a combined MMIE/MLE system. For the results in Table 7, MLE models were used to create the lattices and provide the adaptation supervision (Pure MLE) rather than using MMIE based models for P2/P3 and MMIE generated adaptation supervision for P4.

The pure MLE system (MLE models in P2/P3 and MLE lattices) performs 2.1% absolute poorer than the MMIE system on P2. Comparing the performance of MLE models in P4b, they are 0.7% poorer than in the eval setup (MLE models with MMIE lattices and adaptation supervision) without confusion networks but only 0.3% poorer

[4] All participating sites found that the eval00 data was easier to recognise than past Hub5 evaluation data sets.

Stage	Evaluation	Pure MLE
P2	42.5	44.6
P3	40.0	42.0
P4a	37.1	-
P4b no CN	39.0	39.7
P4b	37.9	38.2
P5a	36.2	-
P5b no CN	38.1	38.7
P5b	37.0	37.3
P4b+P5b	36.5	36.8
P6a	36.0	-
FINAL/CNC	35.0	35.0

Table 7: % WER on eval98 for the evaluation system and a completely separate MLE model-based (b) branch (pure MLE).

with confusion networks. An interesting result shows that although the pure MLE branch is poorer than the mixed MMIE/MLE system it is still able to contribute to the 4-way combination by the same amount. Furthermore while the overall performance of the system is significantly enhanced by the use of MMIE models, the complete pure MLE system achieves a 36.8% WER on eval98.

10 CONCLUSIONS

This paper has discussed the substantial improvements in system performance that have been made to our Hub5 transcription system since the 1998 evaluation. The largest improvement stems from MMIE HMM training, however the MLE model set in their current configuration were shown to still work well. Confusion networks were shown to consistently improve word error rates and yield improved confidence scores. On the 1998 evaluation set a relative reduction in word error rate of 11% was obtained. The system presented here gave the lowest word error rate in the March 2000 Hub5E evaluation. While the overall system is complex, a much simpler setup based on the first few passes of the full system also gives competitive performance.

Acknowledgements

This work was in part supported by GCHQ. Gunnar Evermann has studentships from the EPSRC and the Cambridge European Trust, and Dan Povey holds a studentship from the Schiff Foundation. The authors are grateful to Thomas Niesler and Ed Whittaker for their help in building the class-based language models.

References

1. G. Evermann & P.C. Woodland (2000). Posterior Probability Decoding, Confidence Estimation and System Combination. *Proc. Speech Transcription Workshop*, College Park.
2. J.G. Fiscus (1997). A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–354, Santa Barbara.
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 Transformations for HMM-Based Speech Recognition. *Computer Speech & Language*, Vol 12, pp. 75–98.
5. P.S. Gopalakrishnan, D. Kanevsky, A. Nadas & D. Nahamoo (1991). An Inequality for Rational Functions with Applications to some Statistical Estimation Problems. *IEEE Trans. Information Theory*, Vol. 37, pp. 107–113.
6. T. Hain, P.C. Woodland, T.R. Niesler & E.W.D. Whittaker (1999). The 1998 HTK system for transcription of conversational telephone speech. *Proc. ICASSP'99*, pp. 57–60, Phoenix.
7. T. Hain & P.C. Woodland (1999). Recent Experiments with the CU-HTK Hub5 System. Presentation at June 1999 Hub5 Workshop.
8. C.J. Leggetter & P.C. Woodland (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs. *Computer Speech & Language*, Vol. 9, pp. 171–186.
9. R. Kneser & H. Ney (1993). Improved Clustering Techniques for Class-Based Statistical Language Modelling. *Proc. EUROSPEECH'93*, pp. 973–976, Berlin.
10. X. Luo and F. Jelinek (1999). Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition *Proc. ICASSP'99*, pp. 2044–2047, Phoenix.
11. L. Mangu, E. Brill & A. Stolcke (1999). Finding Consensus Among Words: Lattice-Based Word Error Minimization. *Proc. EUROSPEECH'99*, pp. 495–498, Budapest.
12. 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.
13. Y. Normandin (1991). An Improved MMIE Training Algorithm for Speaker Independent, Small Vocabulary, Continuous Speech Recognition. *Proc. ICASSP'91*, pp. 537–540, Toronto.
14. D. Povey & P.C. Woodland (1999). Frame Discrimination Training of HMMs for Large Vocabulary Speech Recognition. *Proc. ICASSP'99*, pp. 333–336, Phoenix.
15. R. Schlüter, B. Müller, F. Wessel & H. Ney (1999). Interdependence of Language Models and Discriminative Training. *Proc. IEEE ASRU Workshop*, pp. 119–122, Keystone, Colorado.
16. V. Valtchev, J.J. Odell, P.C. Woodland & S.J. Young (1997). MMIE training of large vocabulary speech recognition systems. *Speech Communication*, Vol. 22, pp. 303–314.
17. P.C. Woodland, D. Pye & M.J.F. Gales (1996). Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression. *Proc. ICSLP'96*, pp. 1133–1136, Philadelphia.
18. P.C. Woodland, M.J.F. Gales, D. Pye & S.J. Young (1997). Broadcast News Transcription Using HTK. *Proc. ICASSP'97*, pp. 719–722, Munich.
19. P.C. Woodland & D. Povey (2000). Very Large Scale MMIE Training for Conversational Telephone Speech Recognition. *Proc. Speech Transcription Workshop*, College Park.
20. S.J. Young, J.J. Odell & P.C. Woodland (1994). Tree-Based State Tying for High Accuracy Acoustic Modelling. *Proc. 1994 ARPA Human Language Technology Workshop*, pp. 307–312, Morgan Kaufmann.