NEW FEATURES IN THE CU-HTK SYSTEM FOR TRANSCRIPTION OF CONVERSATIONAL TELEPHONE SPEECH

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 discusses new features integrated into the Cambridge University HTK (CU-HTK) system for the transcription of conversational telephone speech. Major improvements have been achieved by the use of maximum mutual information estimation in training as well as 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. Improvements are demonstrated via performance on the NIST March 2000 evaluation of English conversational telephone speech transcription (Hub5E). In this evaluation the CU-HTK system gave an overall word error rate of 25.4%, which was the best performance by a statistically significant margin.

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. This paper describes changes to the September 1998 Cambridge University HTK Hub5 evaluation system [5] 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. In addition, several minor changes were 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 Hub5 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 Hub5 evaluation system is described and the results of each stage of processing presented.

2. OVERVIEW OF 1998 HTK HUB5 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 [5] for details).

The system uses perceptual linear prediction cepstral coefficients derived from a mel-scale filterbank (MF-PLP) 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 gender independent (GI) and gender dependent (GD) versions of cross-word triphone and quinphone hidden Markov models (HMMs) trained using maximum likelihood estimation. Decision tree state clustering was used to select a set of context-dependent equivalence classes. Mixture Gaussian distributions for each tied state were then trained using iterative mixture splitting. The triphone models were phone position independent, while the quinphone models included questions about word boundaries as well as ± 2 phone context.

The system used a 27k vocabulary that covered all words in the acoustic training data. N-gram word-level language models were constructed by training separate models on transcriptions of the Hub5 acoustic training data and on 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.

The decoding was performed in multiple stages with successively more complex acoustic and language models being applied in later stages. Iterative maximum likelihood linear regression (MLLR) was used for speaker/channel adaptation. The transcription output of two passes was combined 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: Switchboard1 (Swb1) and Call Home English (CHE). The January 2000 release of Swb1 transcriptions from Mississippi State University (MSU) was used for experiments. The complete training data set (h5train00) contains 265 hours of speech and was used for all experiments in this paper. Since only a small proportion of this training set comes from the CHE corpus (17 hours), 3-fold CHE data weighting was used for acoustic model training [6].

The 1998 evaluation data set (eval98) was used as test data for system development, which contains 40 conversation sides of Switchboard2 (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.

4. MMIE TRAINING

The acoustic model parameters in HMM based speech recognition systems are normally estimated using Maximum Likelihood Estimation (MLE), which aims to find model parameters that maximise the likelihood of the correct transcription of the training data. 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

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

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 can be replaced by

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

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

As discussed in [11] the denominator of (1) 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 [9] to iteratively update the model parameters. The implementation we have used here is rather different to the one in [11] and performs 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 is used. The computational methods that we have adopted for Hub5 MMIE training are discussed in detail in [12].

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. 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 approach 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 [12] 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 a unigram language model was used during MMIE training [10] which also improves generalisation performance [12].

Table 1 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 [12]. Comparing the lines in Table 1 show that, the overall word error rate reduction from MMIE training is 2.4% absolute on eval98. MMIE was also used to train quinphone HMMs, where similar gains were found.

Iteration	Swb2	CHE	Total
MLE	42.5	47.7	45.1
1	40.7	46.2	43.5
2	40.3	45.1	42.7

Table 1. %WER on eval98 using VTLN GI triphone models and a trigram language model.

5. SOFT-TYING

The soft tying of states [7] allows Gaussians from a particular state to be used in other mixture distributions with similar acoustics. In a simplified approach [6] for each state the two acoustically most similar states were found based on a single Gaussian version of the model set. 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		
	Swb2	CHE	Total	Swb2	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 2. WER on eval98 using VTLN GI triphone/quinphone models trained on h5train00 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 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.

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. 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. The most successful scheme used three separate dictionary entries for each real pronunciation which differed by the word-end silence type: no silence; a short pause preserving cross-word context; and a general silence model altering context. An estimate for the 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 2 shows that the use of pronunciation probabilities gives a reduction in WER of 1.4-1.7% absolute on eval98. Similar improvements have been found on other test sets.

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 MLLR regression classes 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 [8]. This graph consists of a sequence of confusion sets, which contain competing single word hypotheses with associated posterior probabilities. 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.

The estimates of the word posterior probabilities encoded in the confusion networks can be used directly as confidence scores, but they tend to be over-estimates of the true posteriors. Therefore the posteriors are mapped to confidence scores using a piece-wise linear function based on a decision tree.

The confusion networks with their associated word posterior estimates were also used in an system combination scheme. Confusion network combination (CNC) can be seen as a generalisation of ROVER [2] to confusion networks, i.e. it uses the linear graph and the word posteriors instead of only considering the most likely word hypothesised by each system.

The use of the confusion network output consistently reduced the WER by about 1% absolute. 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 HUB5 EVALUATION SYSTEM

The overall 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.

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. 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 models were all gender dependent and used soft-tying.

9.2. Word List & Language Models

The word list was taken from two sources: the 1998 27k word list [5] and the most frequent 50,000 words occurring in 204 million words of broadcast news (BN) training data. This gave a new combined word list with 54,537 words. This word list reduced the out-of-vocabulary (OOV) rate on eval98 from 0.94% to 0.38%.

The MSU Swb1 training transcriptions were used for language modelling but were found to be significantly different to the original transcripts provided. In order to accommodate both transcript styles both sets of data were used along with 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 [5] a class-based trigram model using 400 automatically generated word classes 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. The first pass P1 is identical to the 1998 P1 setup [5] and its output was used solely for VTLN warp-factor estimation 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 [3] with a 4-gram LM. A global transform[1] for the means (block-diagonal) and variances (diagonal) was computed for each side. In stage P3 word lattices were generated using the adapted GI MMIE triphones and a bigram language model. These lattices were expanded to contain language model probabilities generated by the interpolation of the word 4-gram and the class trigram.

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

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

were based on quinphone acoustic models. Finally for the MMIE branch only, a pass with two MLLR speech transforms was run (P6a). The final system word output was found by using CNC with the confusion networks from P4a, P4b, P6a and P5b.

9.4. System Results

Table 3 gives results for each processing stage for both the 1998 and 2000 evaluation sets. Word error rates on eval00 are approximately 10% absolute lower than on eval98. Possible explanations for this difference are a lower disfluency rate on the Swb1 part of eval00 and overall a higher signal-to-noise ratio.

The large difference (6.8% absolute in WER on both test sets) between the P1 and P2 results is due to the combined effects of VTLN, MMIE models, the larger vocabulary and a 4-gram LM. MLLR adaptation and the smoothing with a class LM results in a further reduction in WER of 2.5% absolute. The second adaptation stage which includes MLLR and a FV transform, pronunciation probabilities and confusion network decoding (P4a) reduces the WER by a further 2.9% absolute (2.1% on eval00), which is 0.8% absolute better than the result of the corresponding MLE soft-tied GD triphone models (P4b). The relative performance on eval00 is again similar with 0.6% difference in WER.

	eval98			eval00		
	Swb2	CHE	Total	Swb1	CHE	Total
P1	47.0	51.6	49.3	31.7	45.4	38.6
P2	40.0	44.9	42.5	25.5	38.1	31.8
P3	37.5	42.4	40.0	22.9	35.7	29.3
P4a	34.5	39.6	37.1	20.9	33.5	27.2
P4b	35.5	40.3	37.9	21.9	33.7	27.8
P5a	33.9	38.4	36.2	20.3	32.7	26.6
P5b	34.5	39.5	37.0	21.0	32.8	26.9
P6a	33.6	38.4	36.0	20.3	32.6	26.5
CNC	32.5	37.4	35.0	19.3	31.4	25.4

Table 3. % WER on eval98 and eval00 for all stages of the evaluation system. The final system output is a combination of P4a, P4b, P6a and P5b.

The use of quinphone models instead of triphone models gives a further gain of 0.6-0.9% for both branches. The gain from a second adaptation stage with two speech transforms for the quinphone MMIE model only brings a relatively small gain of 0.1-0.2% WER absolute. The final result after 4-fold system combination is 35.0% on eval98. This is an 11% reduction in WER relative to the CU-HTK evaluation result obtained on the same data set in 1998 (39.5%). The combination of the 4 outputs using confusion network combination (CNC) was found to be 0.4% absolute better than using the ROVER approach. Confidence scores based on confusion networks give an improved normalised cross entropy of 0.225 compared to 0.145 from the 1998 CU-HTK evaluation system which used N-best homogeneity based confidence scores.

On eval00 the combination of the outputs of the MLE systems (P4b+P5b) gave 26.5% WER whereas a combination of outputs generated by the MMIE model sets (P4a+P6a) resulted in an error rate of 25.6% absolute. In spite of the 0.9% difference the inclusion of the MLE system outputs gives a 0.2% WER absolute improvement. The final error rate on eval00 from the system (25.4%) was lowest in the March 2000 Hub5E evaluation by a statistically significant margin.

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 sets in their current configuration were shown to still work well. 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.

11. 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.

12. 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 ASRU Workshop*, 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] 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*, Vol. 1, pp. 57–60, Phoenix.

[6] T. Hain, P.C. Woodland, G. Evermann & D. Povey (2000). The CU-HTK March 2000 Hub5E Transcription System. *Proc. Speech Transcription Workshop*, College Park.

[7] X. Luo and F. Jelinek (1999). Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition *Proc. ICASSP'99*, pp. 2044–2047, Phoenix.

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

[9] Y. Normandin (1991). An Improved MMIE Training Algorithm for Speaker Independent, Small Vocabulary, Continuous Speech Recognition. *Proc. ICASSP'91*, pp. 537–540, Toronto.

[10] R. Schlüter, B. Müller, F. Wessel & H. Ney (1999). Interdependence of Language Models and Discriminative Training. *Proc. IEEE ASRU Workshop*, pp. 119–122, Keystone.

[11] 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.

[12] P.C. Woodland & D. Povey (2000). Large Scale Discriminative Training for Speech Recognition. *Proc. ISCA ITRW ASR2000*, pp. 7-16, Paris.