IMPROVED CROSS-TASK RECOGNITION USING MMIE TRAINING

R. Córdoba, P.C. Woodland & M.J.F. Gales*

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {rd263,pcw,mjfg}@eng.cam.ac.uk

ABSTRACT

This paper investigates the cross-task recognition and adaptation performance of HMMs trained using either conventional maximum likelihood estimation or the discriminative maximum mutual information estimation (MMIE) criterion. Initial experiments used models trained on the low noise North American Business news corpus of read speech. Cross-task testing on Broadcast News data showed that the MMIE models yielded lower error rates both across-task as well as within-task. This result was confirmed using models trained on the Switchboard corpus which were tested on Voicemail (VM) data. This setup was also used to investigate the performance of task-adaptation when using a limited amount of VM data for both acoustic and language modelling. The setup that gave the best performance on the VM test data used Switchboard models trained using MMIE and then adapted to VM data using maximum *a posteriori* adaptation techniques.

1. INTRODUCTION

Standard speech recognition systems often perform well when tested on data similar to that used in training, but give much higher error rates when tested on data from a new task. Since collecting a large amount of task-specific data is often impractical, it is necesaary to build *generic* recognition systems which work well over a range of tasks and good cross-task robustness is of great importance.

The main focus of this paper is to investigate the impact of the training objective function on the *genericity* of large vocabulary HMM-based speech recognition systems. In particular the use of conventional maximum likelihood estimation (MLE) is compared to discriminative training using maximum mutual information estimation (MMIE). While MMIE training has been shown to improve performance when tested within-class [11], it is unclear as to how well it generalises to data of a rather different type.

In addition we investigate how to make use of a relatively small amount of task-specific data. Since task-adaptation techniques are normally based on maximum *a posteriori* (MAP) estimation, how adaptation interacts with the choice of generic model training objective function is a key issue.

Our initial experiments on cross-task genericity used recognition systems trained on the low-noise North American Business News (NAB) corpus of read newspaper texts and tested on television and radio Broadcast News (BN) data. These showed that MMIE-trained models could indeed provide improved cross-task performance. Further experiments, which confirmed this effect, used models trained on the Switchboard (SWB) corpus of conversational telephone speech and tested on data from Voicemail (VM) messages. We also used this task setup to investigate how well MMIE-trained models could be adapted to a new task using conventional acoustic adaptation methods as well as the impact of a task-specific language model.

The paper is organised as follows. First a brief review of our approach to MMIE training is given followed by an overview of the experimental setup for both cross-task recognition scenarios. The NAB/BN cross-task experiments are presented next followed by the SWB/VM experiments. Experiments are then discussed that adapt the SWB acoustic models to VM data, as well as the effect of using a task-specific language model.

2. MMIE TRAINING

MMIE training [1] maximises the mutual information between the training word sequences and the observation sequences. When the language model (LM) parameters are fixed during training, the MMIE criterion increases the *a posteriori* probability of the word sequence corresponding to 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

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

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 denominator in (1) can be replaced by the likelihood given by a composite model that encodes the full acoustic and language model used in recognition.

Normally MMIE would require a recognition pass of the training set for each iteration, but by using word lattices computed once, the denominator term can be approximated in a computationally efficient manner [10]. The lattice-based framework used in the current work uses fixed model alignments to increase speed further and is described in detail in [8, 11]. The Gaussian mixture parameters can be updated using the Extended Baum-Welch (EBW) algorithm [4, 6]. Given a suitable method to the choose smoothing parameters, EBW can give rapid convergence [11].

While MMIE training is very effective at reducing the training set error rate, test-set generalisation is a key issue. It has been found that test-set generalisation can be improved by a process of acoustic scaling [11] which increases the quantity of confusable data in training and also by the use of a weak unigram language model [9, 11] to focus on acoustic discrimination.

*Ricardo Córdoba is now at Grupo de Tecnología del Habla. Universidad Politécnica de Madrid, ETSI Telecomunicación, Ciudad Universitaria s/n, 28040-Madrid, Spain. Email: cordoba@die.upm.es

This work was supported by the EU Framework 5 project Coretex and benefited from an SUR award of computer equipment from IBM.

To appear in *Proc. ICASSP2002, May 13-17, 2000, Florida, USA*

3. DATA SETS & EXPERIMENTAL SETUP

The following sections describe the experimental framework for both the NAB/BN experiments and those on SWB/VM data. In all cases the input data consists of PLP coefficients derived from a mel-scale filter bank (MF-PLP), with 13 coefficients including c_0 and their first and second-order differentials. All the HMMs used were gender independent cross-word triphones built using decision-tree state clustering. Unless otherwise stated, the experiments don't include unsupervised test-set adaptation.

In all cases where MMIE training was used, conventional MLE was used to initialise the HMMs. For all sets of MMIE experiments, word lattices for MMIE training were created using a bigram language model, while unigram probabilities were actually applied to these lattices for MMIE training.

All cross-task recognition experiments used a complete single pass decoding run with a trigram language model to avoid any possible cross-system effects from lattice-rescoring. The pronunciation dictionaries used in training and test were originally based on the 1993 LIMSI WSJ lexicon, but have been considerably extended and modified.

3.1. NAB/BN Systems & Data

The NAB system used HMMs trained on the SI-284 Wall Street Journal database (66 hours of data) and used per-utterance cepstral mean normalisation. This data is low noise and contains read-speech. The HMMs have 6399 speech states and 12 mixture components per state. Two sets of models were created. Initially an HMM set using the close-talking channel 1 (NAB-C1) was created using MLE, and then single-pass retraining was used to create a second version using the far-field desktop microphone channel (NAB-C2). For each NAB MLE model set, a corresponding MMIE-trained version of the HMM set was created. Within-task recognition results for the NAB-C1 MMIE models were presented in [8] in which MMIE yielded a 5.5% relative reduction in word error rate (WER) over MLE models.

The cross-task data consists of the 1996 "partitioned evaluation" Broadcast News development test data (BNdev96pe). This is taken from 6 radio and television shows and was manually segmented according to so-called "F-conditions" which describes the type of data present. This data is particularly interesting for cross-task testing since the impact of different types of data can be established. More details on the BN test and training sets can be found in [12]. The LM used for this task was a 65k word trigram trained primarily on BN text and newswire data. This LM was also used in the 1996 CU-HTK evaluation system [12].

As a contrast, we also present results on the same BN test data with HMMs trained solely on broadcast news data i.e. matched training/test conditions. HMM sets were created with either 36 or 72 hours of BN data. The 36 hour set (BNtrain96) consists of the training data available for the 1996 DARPA BN evaluation, while the 72 hour set was the 1997 training data set (BNtrain97). The MLE HMMs estimated from BNtrain96 had 5628 states while the BNtrain97 had 6684 states. Both HMM sets had 12 Gaussians per state. For BNtrain97 an MMIE version of the model set was also created.

3.2. SWB/VM Systems & Data

For the Switchboard systems, we used HMMs trained on a total of 265 hours of data taken from the Switchboard1 and Call Home English corpora which both consist of telephone conversations. The data had cepstral mean and variance normalisation applied on a conversation side basis, along with vocal tract length normalisation. The HMMs used had 6165 clustered speech states and 16 Gaussians per state. Two SWB HMM sets of this structure were created with either MLE or MMIE training. The SWB model sets are denoted SWB-MLE and SWB-MMIE. The SWB-MMIE models were the best performing triphone models from [11] and included a lattice re-alignment phase. The performance of these model sets was benchmarked on the 1998 Hub5 evaluation data set, eval98. On that data, the SWB-MLE had a 45.6% WER and the the SWB-MMIE models 41.5% [11] (a 9.0% relative reduction in WER). These tests used a 27k word vocabulary and a trigram LM (SWB-LM) formed by the interpolation of language models trained on Switchboard and Broadcast News text data.

The cross-task chosen for the SWB system used the Voicemail (VM) corpus. This data was collected by IBM [7] and made available through the LDC. The VM corpus consists of a set of voice-mail telephone messages. The speaking styles in this data are rather different from those typically found in SWB telephone conversations. Two VM test sets were used: one which is the "devtest" set from the LDC VM corpus and a further test set obtained directly from IBM. Both of these test sets are small (11 minutes and 23 minutes) and so we combined them to form a single VM test set (VMtest) of 34 minutes consisting of 92 separate VM messages[1]. We processed the supplied transcriptions to ensure compatibility with the SWB-LM conventions (removed compound words, expanded capital letter acronynms to a sequence of single letters) and also made some minor corrections.

Twenty hours (1801 messages) of VM training data (VMtrain) were also made available. We used this data in several ways. First, using MLE, we created HMMs which were solely trained with VM data. These models had 4626 clustered states and 12 Gaussians per state. Due to the limited amount of training data we did not create an MMIE version of these models, but rather created an MMIE version of models with 6 Gaussians per state. We also used MAP [3] to adapt the SWB HMMs to the VM task and compared the use of 1 hour, 4 hours or 20 hours of VM data for adaptation. Furthermore, we used the transcriptions of VMtrain to augment the 27k SWB-LM vocabulary (to a total of 29k words). An interpolated language model was then created from the VMtrain transcriptions, the SWB training transcriptions and BN text transcriptions which is denoted VM-LM.

4. NAB/BN EXPERIMENTS

The aim of these initial experiments was primarily to compare the effect of MMIE and MLE for cross-task recognition. Since MMIE models are more highly tuned to discriminating their training data it is unclear how well the models will generalise to situations unseen in training. Hence it is particularly interesting to compare the performance of the training schemes across situations which have a severe mismatch.

The results of testing the various NAB channel 1 and channel 2 models described in Sec. 3.1 on the BNdev96pe test set is shown in the first part of Table 1. The results of some of the main F-conditions are also given. While all of the broadcast news data is mismatched to the NAB low noise read-speech training data, the

[1] Ideally a larger VM test set would have been used, but no more suitable test data was available.

prepared native-speaker F0 data would be thought to be the best matched. Other F-conditions shown in Table 1 include F1 (spontaneous speech); F2 (low-fidelity channels, e.g. telephone); F4 (noisy speech) and FX that represents all other speech types (e.g. spontaneous speech from non-natives) and is the most challenging to recognise.

Train Setup	Avg	F0	F1	F2	F4	FX
NAB-C1 MLE	39.8	15.4	37.4	62.0	33.7	64.4
NAB-C1 MMIE	38.0	15.2	35.6	59.1	31.5	61.7
NAB-C2 MLE	36.0	16.3	35.2	51.4	28.6	58.5
NAB-C2 MMIE	34.1	15.8	33.5	49.2	28.1	53.1
BN-36H MLE	31.7	12.8	28.5	42.6	25.4	56.8
BN-72H MLE	29.6	11.6	26.2	38.7	24.6	55.4
BN-72H MMIE	27.8	11.4	24.5	34.9	23.2	51.3

Table 1. %WER on BNdev96pe data using trigram, GI, for NAB channel 1 and channel 2 models trained with either MLE or MMIE. The use of BN training data is also shown for comparison.

There are a number of points to note about the results in Table 1. The NAB-C2 models give a lower WER than NAB-C1 for all conditions apart from F0. However more interestingly the MMIE-trained models are consistently better than the MLE models in cross-task testing (as they are also within-task), and the difference is greatest for the conditions that are most mismatched to the training data e.g. there is a 5.4% absolute reduction in WER for the NAB-C2 MMIE models for FX data. These results imply that MMIE-training is likely to be superior to MLE training for the creation of general-purpose HMMs that operate well across a range of tasks including conditions not seen in training.

The lower portion of Table 1 gives the performance of within-task testing on BN data with varying amounts of training. Compared to the best NAB training setup (NAB-C2 MMIE), MLE training on BN data gives considerable improvements on F0 and F1 (3.0% and 5.5% absolute with 36 hours of BN training data), but surprisingly on the most difficult mismatched FX data the NAB-C2 MMIE models outperform those trained with MLE on even 72 hours of BN data. As would be expected, the use of MMIE training using the 72 hour BN corpus gives an overall error rate reduction of 1.8% absolute and yields the lowest word error rates over all F-conditions. In passing we believe that the results of MMIE BN training are themselves of interest as this is, to our knowledge, the first time that MMIE training on the BN corpus has been reported.

5. SWB/VM EXPERIMENTS

The aim of the experiments in this section was first to confirm the superiority of MMIE training for cross-task recognition on a rather different task setup, i.e. Switchboard HMMs tested on Voicemail data. We also wanted to investigate how a limited amount of VM training data could best be used. Several avenues have been investigated including MAP acoustic adaptation with varying amounts of VM data and adapting the language model based on the VM transcriptions. In Sec. 5.2 the use of MMIE training on the VM corpus is investigated and compared to adapting the SWB models. Finally in Sec. 5.3 the effect of unsupervised test-set maximum likelihood linear regression (MLLR) [5, 2] adaptation is investigated for the various models.

5.1. Genericty & Task-Specific Adaptation

The results of cross-task training with the SWB HMMs is shown in Table 2. When the SWB-LM is used (i.e. there is no task-specific data) the MMIE trained models show a 2.9% absolute reduction in WER over the MLE models.

HMM Set	Language Model	VM Adaptation Data (hours)			
		None	1	4	20
SWB-MLE	SWB-LM	46.1	45.7	43.5	41.0
SWB-MMIE	SWB-LM	43.2	42.2	40.7	39.5
SWB-MLE	VM-LM	36.9	35.2	34.3	32.7
SWB-MMIE	VM-LM	32.6	32.2	31.8	30.9

Table 2. %WER on the VM test data for supervised MAP adaptation with varying amounts of VM training data. Either MLE or MMIE Switchboard HMM sets were used with either the SWB or the VM interpolated language model.

Table 2 also shows the effect of using various quantities of VM training data for supervised MAP adaptation [3]. All MAP experiments reported use a prior-weight (τ) value of 10. It was found that for this task and the quantity of data used, MAP out-performed supervised MLLR adaptation and gave as good results as MLLR followed by MAP. As expected the reduction in WER increases steadily as more adaptation data is used: 1 hour gives only small improvements while for 20 hours of VM adaptation data, the WER was reduced by 3.7% absolute for SWB-MMIE and by 5.1% absolute for SWB-MLE. The reductions in WER due to MAP with various amounts of VM acoustic data can be related directly to the percentage of Gaussians that are reasonably well adapted. It was found that for 1 hour of data only 5% of the Gaussians receive more than 10 observations, while this is increased to 32% with 4 hours of data and 79% with 20 hours.

The effect of using the VM-LM is also shown in Table 2. With no acoustic adaptation the WER is reduced by 9.2% absolute for SWB-MLE, and by 10.6% absolute for SWB-MMIE with the adapted LM. Hence, in this case, the use of language model adaptation (by itself) is more effective than just acoustic adaptation. It was found that the main improvements in performance with the VM-LM were due to improved recognition of typical “voicemail phrases” such as message openings (e.g. “hi Jane, this is”) and closings (e.g. “talk to you later”, “give me a call”).

When acoustic adaptation is combined with the use of the VM-LM the further gains due to acoustic adaptation are reduced i.e. the gains due to acoustic and LM adaptation are not additive. The total improvements from using 20 hours of data for both language model and acoustic adaptation are 13.4% absolute for the SWB-MLE HMMs and 12.3% for the SWB-MMIE models. This meant that even after 20 hours of task-specific adaptation the MMIE models have a sizable advantage.

It is interesting, and perhaps surprising, to note that conventional MAP adaptation remains effective when applied to models trained with MMIE (in the limit of a very large adaptation set MAP will converge to the MLE solution). To give further insight into this issue we examined how well the the SWB-MMIE models performed after a single iteration of MLE updating using the complete 265 hour Switchboard training set on the eval98 test set. While the performance degraded from 41.5% to 44.6% it is still somewhat better than the baseline MLE system (45.6%). For the MAP adaptation experiments discussed here over an order of magnitude less data is used, and hence a much smaller disturbance occurs to the

MMIE-trained Gaussian parameters. Therefore most of the advantage of the original MMIE training is retained.

5.2. MMIE Training on VM Data

The use of either the 20 hour MAP-adapted SWB-MLE or SWB-MMIE models is compared to using HMMs trained solely on 20 hours of VM data in Table 3 for both the SWB-LM and VM-LM.

HMM Set	Language Model	
	SWB-LM	VM-LM
SWB-MLE + 20hr VM	41.0	32.7
SWB-MMIE + 20hr VM	39.5	30.9
VM-MMIE (6 mix comp)	40.5	34.0
VM-MLE (6 mix comp)	42.5	35.8
VM-MLE (12 mix comp)	41.9	35.5

Table 3. %WER on VMtest for 20 hours of VM acoustic data used either for MAP adaption or for pure VM HMMs. Results are shown for either MLE or MMIE training are shown, with either the SWB-LM or the VM-LM.

It can be seen from Table 3 that with the SWB-LM the performance of the best 12 mixture component VM MLE-trained acoustic models is a little poorer (0.9% absolute) than the adapted SWB-MLE models, and 2.4% absolute poorer than the VM-adapted SWB-MMIE HMMs. It is interesting to note that this gap in performance becomes noticeably larger when using the VM-LM and there appear to be additional advantages in using an adapted generic HMM set in this case.

As a further contrast we applied MMIE training to the 6 mixture component VM models and obtained on average a 1.9% (5% relative) reduction in WER over the 6 mixture component MLE models. These HMMs give the best performance of those trained purely on VM data, but are still some way behind the performance of the SWB-MMIE models with MAP adaptation.

5.3. Unsupervised Test-Set Adaptation

Finally we investigated the performance of unsupervised test-set MLLR adaptation [5, 2]. The MLLR setup used a regression-class tree and updated both the Gaussian means and variances.

HMM Set	Language Model	No VM Adapt		20 hour Adapt	
		Base	MLLR	Base	MLLR
SWB-MLE	SWB-LM	46.1	45.0	41.0	39.9
SWB-MMIE	SWB-LM	43.2	41.5	39.5	38.2
SWB-MLE	VM-LM	36.9	35.2	32.7	31.8
SWB-MMIE	VM-LM	32.6	31.4	30.9	29.7

Table 4. %WER on the VM test data with (MLLR) and without (Base) unsupervised MLLR adaptation for both the MLE and MMIE SWB model sets with either SWB-LM or VM-LM.

The results of these experiments are shown in Table 4. All the results of unsupervised test set adaptation give reductions in WER in the range 0.9% to 1.7% absolute and MLLR seems to work equally well across the different types of HMM training and language model. We also obtained similar improvements when testing HMMs trained solely on VM. The best result we have obtained on the VM test data to date (29.7%) used the SWB-MMIE models with MAP adaptation to 20 hours of VM training data, the VM-LM and unsupervised MLLR adaptation.

6. CONCLUSIONS

This paper has discussed the creation of HMMs aimed at improved performance when tested on new data types (i.e. cross-task recognition). It has been shown that MMIE training improves cross-task performance, both for the recognition of broadcast news using models trained on low-noise read speech, and for the transcription of voicemail data based on Switchboard-trained HMMs. The advantage of MMIE over conventional MLE training appears to increase as the mismatch between training and test data increases. It was further found that it is advantageous to use the MMIE trained Switchboard HMMs when up to 20 hours of voicemail data is available for both acoustic and language model adaptation since MAP adaptation continues to perform well when applied to large MMIE-trained models.

7. REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. de Souza & R.L. Mercer (1986). Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. *Proc. ICASSP'86*, pp. 49–52, Tokyo.

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

[3] J.L. Gauvain & C.H. Lee (1994) Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. SAP*, Vol. 2, pp. 291-298.

[4] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas & D. Nahamoo (1991). An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems. *IEEE Trans. on Information Theory*, Vol. 37, pp 107-113.

[5] C.J. Leggetter & P.C. Woodland (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. *Proc. ARPA SLT Workshop*, pp. 104-109. Morgan Kaufmann.

[6] Y. Normandin (1991). *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem.* Ph.D. thesis, Dept. of Elect. Eng., McGill University.

[7] M. Padmanabhan, B. Ramabhadran, E. Eide, G. Ramaswamy, L.R. Bahl, P.S. Gopalakrishnan & S. Roukos (1997). Transcription of New Speaking Styles—Voicemail. *Proc. DARPA Hub4 Workshop*.

[8] D. Povey & P.C. Woodland (2001). Improved Discriminative Training Techniques for Large Vocabulary Continuous Speech Recognition. *Proc. ICASSP'2001*, Salt Lake City.

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

[10] 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.

[11] P.C. Woodland & D. Povey (2002). Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. To appear, *Computer Speech & Language*.

[12] P.C. Woodland (2002). The Development of the HTK Broadcast News Transcription System: An Overview. To appear, *Speech Communication*.