AUGMENTED STATISTICAL MODELS FOR SPEECH RECOGNITION

M.I. Layton and M.J.F. Gales

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ, U.K.
Email: {ml362,mjfg}@eng.cam.ac.uk

ABSTRACT

Recently there has been significant interest in developing new acoustic models for speech recognition. One such model, that allows complex dependencies to be represented, is the augmented statistical model. This incorporates additional dependencies by constructing a local exponential expansion of a standard HMM. Unfortunately, the resulting model often has an intractable normalisation term, rendering training difficult for all but binary classification tasks. In this paper, conditional augmented (C-Aug) models are proposed as an attractive alternative. Instead of modelling utterance likelihoods and inferring decision boundaries, C-Aug models directly model the posterior probability of class labels, conditioned on the utterance. The resulting model is easy to normalise and can be trained using conditional maximum likelihood estimation. In addition, as a convex model, the optimisation converges to a global maximum.

1. INTRODUCTION

In recent years, a wide range of acoustic models have been applied to the speech recognition task; the most popular of these is the hidden Markov model (HMM). Unfortunately, HMMs are based upon a series of assumptions that are known to be poor, in particular, successive frames of speech are assumed to be independent given the state that generated them. In order to overcome these limitations, many extensions, such as: segmental models, switching linear dynamical systems (S-LDSs) and buried Markov models, have been proposed. Despite significant differences in structure, these models all share a common goal: to better model speech by incorporating (or providing a framework for incorporating) additional (possibly long-range) dependencies. Unfortunately, none of these techniques provide any indication as to which dependencies should be modelled.

In [1, 2] augmented statistical models were proposed as a systematic technique for both specifying and modelling additional dependencies. This is achieved by constructing a local exponential approximation (using a Taylor series expansion) to a standard – typically HMM or Gaussian mixture model (GMM) – base, statistical model. The latent-variable structure of the augmented model is determined by the base model; the sufficient statistics for the exponential model are given by its derivatives, and are dependent on all observations and states, breaking the conditional independence assumptions of the base model. This allows augmented models to represent highly complex distributions. Unfortunately, the price for this flexibility is a statistical model with an intractable normalisation term. Direct training – maximum likelihood (ML) or maximum mutual information (MMI) [3] – is therefore prohibitively expensive for practical tasks. Instead a binary maximum-margin (MM) criterion can be used [2].

Martin Layton would like to thank the Schiff Foundation for funding. Extensive use was made of equipment supplied to the Speech Group at Cambridge University by IBM under an SUR award.

In this paper, conditional augmented (C-Aug) models are proposed. These are defined similarly to standard (generative) augmented models except that, instead of modelling utterance likelihoods, they directly model the posterior probabilities of class labels. C-Aug models therefore have all the modelling advantages of standard augmented models with the added benefit that the normalisation is calculated as the expectation over the class labels instead of over observation sequences. In addition, as (highly complex) members of the exponential family, C-Aug models are similar to conditional random fields (CRFs) [4] (though without the Markov dependency), allowing CRF training techniques, such as conditional maximum likelihood (CML), to be used. C-Aug models overcome one of the greatest drawbacks of standard CRFs: determining which statistics to include.

In this paper, generative augmented models, their properties and maximum-margin (MM) training are all briefly reviewed. Then, in section 3, conditional augmented models are introduced. Basic properties, maximum likelihood estimation and inference of C-Aug models are discussed. Finally, preliminary results on a large vocabulary rescoring task and on the TIMIT [5] phone classification task are presented.

2. AUGMENTED STATISTICAL MODELS

2.1. The Exponential Family

Many standard statistical models are based upon the exponential family. Common examples are the Gaussian, Poisson and Bernoulli distributions. In terms of observations $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$, $\boldsymbol{o}_t \in \mathbb{R}^d$, sufficient statistics $\boldsymbol{T}(\boldsymbol{O})$ and natural parameters $\boldsymbol{\alpha}$, these models can be written in the form,

$$p(\boldsymbol{O};\boldsymbol{\alpha}) = \frac{1}{\tau(\boldsymbol{\alpha})} h(\boldsymbol{O}) \exp\left[\left\langle \boldsymbol{\alpha}, \boldsymbol{T}(\boldsymbol{O}) \right\rangle\right] \quad (1)$$

where $h(\boldsymbol{O})$ is a reference distribution and $\langle \cdot, \cdot \rangle$ denotes the inner-product between two vectors using an appropriate metric. The normalisation constant, $\tau(\boldsymbol{\alpha})$, ensures that the axioms of probability are satisfied and is calculated as the expectation of (1) over the observations,

$$\tau(\boldsymbol{\alpha}) = \mathcal{E}_{\boldsymbol{O}}\{p(\boldsymbol{O};\boldsymbol{\alpha})\} = \int_{\boldsymbol{O}} h(\boldsymbol{O}) \exp\left(\left\langle \boldsymbol{\alpha}, \boldsymbol{T}(\boldsymbol{O}) \right\rangle\right) d\boldsymbol{O} \quad (2)$$

Unfortunately standard exponential distributions cannot simply model temporal dependencies or multi-modal distributions. They are therefore unsuitable for many 'real-world' tasks; instead, latent-variable extensions, such as GMMs and HMMs, are used. In many cases, however, the independence and conditional-independence assumptions encoded in these latent-variable models are not correct, potentially degrading classification performance. Improved models can be obtained by adding dependencies through expert-knowledge and hand-tuning, however it is often not clear which dependencies to include. *Augmented statistical models* [2, 1] remove this issue by incorporating additional dependencies in a systematic fashion.

2.2. Augmented Statistical Models

Augmented statistical models (herein referred to as augmented models) are composed of two parts: a base statistical model $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})$ (often a GMM or HMM), and a local exponential expansion (calculated using a ρ-th order Taylor expansion[1]) of that model about each point $\boldsymbol{\lambda}$ [6, 1, 2],

$$p(\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})}\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\exp\left(\boldsymbol{\alpha}^T\nabla_{\boldsymbol{\lambda}}^{(1,\rho)}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\right) \quad (3)$$

where $\boldsymbol{\alpha}$ are the augmented parameters and $\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})$ is a normalisation term (calculated as the expectation over all observation sequences). The sufficient statistics of (3) are given by base model derivatives of orders 1 through ρ [7, 1],

$$\nabla_{\boldsymbol{\lambda}}^{(1,\rho)}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}) = \begin{bmatrix} \nabla_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}) \\ \frac{1}{2!}\text{vec}(\nabla_{\boldsymbol{\lambda}}^2\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})) \\ \vdots \\ \frac{1}{\rho!}\text{vec}(\nabla_{\boldsymbol{\lambda}}^{\rho}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})) \end{bmatrix} \quad (4)$$

At this stage, it is interesting to contrast the nature of the dependencies modelled by augmented models to those of the base model. Since no new statistics are introduced (only new functions of the base model statistics), independence assumptions of the base model are retained. This is not the case, however, with the conditional independence assumptions. In particular, derivatives of latent variable models are a function of the posterior probabilities of the latent states. Since these are dependent on all observations and all latent states, conditional independence is broken.

With their additional modelling power, augmented models can be difficult to train. This is because in general, unlike GMMs and HMMs, no closed-form solution exists for the normalisation term. Maximum likelihood (ML) and maximum mutual information (MMI) estimation of augmented parameters are therefore difficult. Instead a two-stage training algorithm can be used. First, the base model parameters, $\boldsymbol{\lambda}$, are calculated using standard ML or MMI training. Next, the augmented parameters, $\boldsymbol{\alpha}$, are estimated using a distance-based discriminative training criterion.

2.3. Maximum Margin Estimation

Unlike ML and MMI training which try to model underlying source distributions, maximum-margin (MM) estimation is a distance-based technique that tries to directly model the decision surface between classes. It does this by maximising the geometric margin – the distance between the decision boundary and the closest training examples – between classes. This results in a robust classifier that generalises well even when using high-dimensional spaces or limited training data. Furthermore, since it attempts to correctly classify all training examples, it is inherently discriminatory in nature and thus an obvious alternative to discriminative criteria such as MMI. The disadvantage of such an approach is that it is limited to binary classification (although schemes have been proposed to handle the multi-class case [8]).

Consider a binary task where class-conditional generative models of the form, $p(\boldsymbol{O};\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})$ and $p(\boldsymbol{O};\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})$ are estimated on the training data. The decision boundary that minimises the probability of training error is given by Bayes' decision rule,

$$\frac{P(\omega_1|\boldsymbol{O};\boldsymbol{\lambda}^{(1)})}{P(\omega_2|\boldsymbol{O};\boldsymbol{\lambda}^{(2)})} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1 \quad (5)$$

When $\tilde{\boldsymbol{\lambda}} = \{\tilde{\boldsymbol{\lambda}}^{(1)},\tilde{\boldsymbol{\lambda}}^{(2)}\}$ are estimated using ML or MMI training, equation (5) can be rewritten as a linear decision boundary with weight $\boldsymbol{w}$ and bias, b, [2]

$$\left(\boldsymbol{w},\boldsymbol{\phi}^{\text{LL}}(\boldsymbol{O};\tilde{\boldsymbol{\lambda}})\right) + b \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0; \quad b = \ln\left[\frac{P(\omega_1)\tau(\tilde{\boldsymbol{\lambda}}^{(2)},\boldsymbol{\alpha}^{(2)})}{P(\omega_2)\tau(\tilde{\boldsymbol{\lambda}}^{(1)},\boldsymbol{\alpha}^{(1)})}\right] \quad (6)$$

where $P(\omega_1)$ and $P(\omega_2)$ are class priors and $\boldsymbol{\phi}^{\text{LL}}(\boldsymbol{O};\tilde{\boldsymbol{\lambda}})$ is a generative score-space (an extension of the Fisher score-space [9]), dependent on only the observations and base model parameters $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(2)}\}$,

$$\boldsymbol{\phi}^{\text{LL}}(\boldsymbol{O};\boldsymbol{\lambda}) = \begin{bmatrix} \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) - \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \\ \nabla_{\boldsymbol{\lambda}^{(1)}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) \\ -\nabla_{\boldsymbol{\lambda}^{(2)}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \end{bmatrix} \quad (7)$$

The task of estimating augmented parameters $\tilde{\boldsymbol{\alpha}} = \{\tilde{\boldsymbol{\alpha}}^{(1)},\tilde{\boldsymbol{\alpha}}^{(2)}\}$ is therefore reduced to finding a linear decision boundary (6) in the score-space (7), where $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{w}$ are related by $\boldsymbol{w} = [1,\ \tilde{\boldsymbol{\alpha}}^{(1)},\ \tilde{\boldsymbol{\alpha}}^{(2)}]^T$. Although many techniques exist for estimating linear decision surfaces, a popular algorithm for MM training is the Support Vector Machine (SVM).

3. CONDITIONAL AUGMENTED MODELS

In the previous section, a MM algorithm was described for training augmented models, allowing the problem of an intractable normalisation term to be mitigated. Unfortunately, this technique is limited to binary tasks. Alternatively, a *conditional augmented (C-Aug) model* can be defined. Here, instead of modelling the observation likelihood, C-Aug models directly model the posterior probability of the class labels, ω,

$$P(\omega|\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\lambda},\boldsymbol{\alpha})}\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(\omega)})\exp\left(\boldsymbol{\alpha}^{(\omega)T}\nabla_{\boldsymbol{\lambda}}^{(1,\rho)}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(\omega)})\right) \quad (8)$$

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(\omega)}\}$ and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^{(\omega)}\}$, $\forall\omega\in\Omega$, the set of all class labels. Although, superficially, this appears identical to the augmented model in (3), it is, in fact, very different. This difference arises from the way that the normalisation term[2], $Z(\boldsymbol{\lambda},\boldsymbol{\alpha})$, is calculated. Since conditional augmented models directly model posteriors, the normalisation is calculated as the expectation of (8) *over the class labels,* $\omega\in\Omega$,

$$Z(\boldsymbol{\lambda},\boldsymbol{\alpha}) = \sum_{\omega\in\Omega}\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(\omega)})\exp\left(\boldsymbol{\alpha}^{(\omega)T}\nabla_{\boldsymbol{\lambda}}^{(1,\rho)}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(\omega)})\right) \quad (9)$$

As this is typically a small number, the summation in (9) is feasible. Direct training of model parameters is therefore possible. It is useful to note that although the conditional distribution is always valid, the generative model associated with it is not necessarily valid.

Although it may seem strange to embed a generative model within a conditional model, this is a perfectly valid operation since the generative model is used only to generate a set of sufficient statistics. Compared to arbitrary statistics (such as $\boldsymbol{o}$ and $\boldsymbol{o}^2$), statistics from generative models are advantageous since they are tuned to match the distribution of the source data, thus providing a better representation of the data [9]. When large feature-vectors are required, the computational cost of considering these features can be mitigated by considering kernelised C-Aug models [10]. These use an implicit mapping to a high-dimensional feature-space whilst performing calculations in the original, lower-dimensional, space.

[1] For simplicity, in this paper the *natural* basis and higher-order derivatives are assumed to yield a set of orthogonal basis. It is therefore not necessary to distinguish between covariant and contravariant basis and components [6].

[2] For clarity, the normalisation term of a conditional augmented model is denoted $Z(\cdot)$ instead of $\tau(\cdot)$ to emphasise that the expectation is calculated over the classes instead of over the observation-space.

It is important to note that C-Aug models do not inherit the conditional independence assumptions of the base model since, similarly to generative augmented models, the posterior terms in the derivatives break conditional independence.

3.1. Relationship with CRFs and HCRFs

Before discussing training algorithms, it is helpful to contrast C-Aug models with conditional random fields (CRFs) [4] and hidden conditional random fields (HCRFs) [11]. Consider the case when the base model parameters, $\boldsymbol{\lambda}$, are fixed (such as when the two-stage estimation is used). Equation (8) can be re-expressed in terms of *constant* sufficient statistics (features), $\boldsymbol{T}(\omega, \boldsymbol{O}; \boldsymbol{\lambda})$, as,

$$P(\omega|\boldsymbol{O}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \exp\left(\boldsymbol{\alpha}^T \boldsymbol{T}(\omega, \boldsymbol{O}; \boldsymbol{\lambda})\right) \tag{10}$$

where $\boldsymbol{T}(\omega, \boldsymbol{O}; \boldsymbol{\lambda})$ has elements,

$$\boldsymbol{T}^{LL}_{\omega'}(\omega, \boldsymbol{O}; \boldsymbol{\lambda}) = \delta_{\omega=\omega'} \ln \hat{p}(\boldsymbol{O}; \boldsymbol{\lambda}^{(\omega)}) \qquad \forall \omega \in \Omega$$

$$\boldsymbol{T}^{\nabla}_{\omega'}(\omega, \boldsymbol{O}; \boldsymbol{\lambda}) = \delta_{\omega=\omega'} \nabla^{(1,\rho)}_{\boldsymbol{\lambda}} \ln \hat{p}(\boldsymbol{O}; \boldsymbol{\lambda}^{(\omega)}) \qquad \forall \omega \in \Omega$$

In this form, it is clear that (10) is simply an exponential model[3] with sufficient statistics $\boldsymbol{T}(\omega, \boldsymbol{O}; \boldsymbol{\lambda})$ and natural parameters $\boldsymbol{\alpha}$. Although the similarity to CRFs allows C-Aug models to be regarded as a systematic method for defining CRF feature-vectors, much of the power of both C-Aug models and HCRFs is obtained through the use of latent variables.

HCRFs introduce latent-variables directly into the exponential model resulting in a flexible HMM-style model. Unfortunately there are three main draw-backs to this approach: first, the independence and conditional-independence assumptions are identical to those of HMMs[4]. Second, like CRFs, there is no indication as to which features to include. Third, and perhaps most important, is that by introducing latent-variables directly into the exponential model, the model convexity (and hence global maximum) is lost. This makes HCRFs sensitive to both initialisation and the parameter update algorithm (c.f. L-BFGS versus SGD in [11]).

Conversely, although conditional augmented models also have an HMM-style latent-variable structure, the latent variables are contained solely within the sufficient statistics. When the base model parameters are fixed (such as during two-stage training – section 2), these statistics are constant; model convexity and the existence of a global solution are therefore preserved. A disadvantage of this approach is that, unlike HCRFs, the state-segmentation of utterances is fixed by the base model and cannot be updated during training. However, despite a greater number of segmentation constraints, the complex statistics of C-Aug models (c.f. Gaussian-based statistics of the HCRF) allow highly complex, non-Gaussian, output distributions to be modelled. These distributions need not satisfy the conditional independence assumptions of the base HMM.

3.2. Conditional Maximum Likelihood Estimation and Inference

As discussed above, a significant advantage of C-Aug models over generative augmented models is that the normalisation can be calculated simply. Direct training of model parameters is therefore possible. A practical choice of training criterion is conditional maximum likelihood (CML) estimation. This is a good criterion to use since, in addition to strong theoretical motivations, it is inherently discriminatory in nature.

Consider the set of training examples $\{\boldsymbol{O}_i\}$ with labels $y_i \in \Omega$, $i \in \{1, \ldots, n\}$. The objective of CML estimation is to find $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\alpha}}$ such that the likelihood of the class labels, conditioned on the observations, is maximised,

$$\{\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}\} = \arg\max_{\boldsymbol{\lambda}, \boldsymbol{\alpha}} \sum_{i=1}^{n} \ln P(y_i|\boldsymbol{O}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \tag{11}$$

Unfortunately, simultaneous optimisation of both $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ is difficult since the objective function has many local maxima. Instead, $\tilde{\boldsymbol{\lambda}}$ is first estimated using standard ML or MMI estimation. This simplifies training by allowing the model to be written as in (10). The optimisation thus becomes,

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(\boldsymbol{\alpha}^T \boldsymbol{T}(y_i, \boldsymbol{O}_i; \tilde{\boldsymbol{\lambda}}) - \ln Z_i(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\alpha})\right) \tag{12}$$

This is convex in $\boldsymbol{\alpha}$ and so has a single, global, maximum. Since there is no closed-form solution, a gradient-based iterative update must be used. For this paper, scaled conjugate gradient (SCG) [12] was chosen. Like all conjugate gradient methods, this updates model parameters by taking a step, not in the direction of the gradient,

$$\nabla_{\boldsymbol{\alpha}} \ln p(y_i|\boldsymbol{O}_i; \tilde{\boldsymbol{\lambda}}, \boldsymbol{\alpha}) = \boldsymbol{T}(y_i, \boldsymbol{O}_i; \tilde{\boldsymbol{\lambda}}) - \sum_{\omega \in \Omega} P(\omega|\boldsymbol{O}_i; \tilde{\boldsymbol{\lambda}}, \boldsymbol{\alpha}) \boldsymbol{T}(\omega, \boldsymbol{O}_i; \tilde{\boldsymbol{\lambda}}) \tag{13}$$

but in a direction that, as far as possible, is conjugate to all previous steps taken. However, unlike many standard algorithms, the step-size is selected using a model-trust region based approach instead of a line search (which is extremely expensive for objective functions such as (11)).

Given an utterance $\boldsymbol{O}_i$ and a ML C-Aug model, $\{\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}\}$, inference simply requires selecting the label, y_i, with the largest posterior,

$$y_i = \arg\max_{\omega \in \Omega} P(\omega|\boldsymbol{O}_i; \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}) \tag{14}$$

This can be efficiently implemented by comparing *unnormalised* posteriors (Z_i is constant across all classes).

4. EXPERIMENTAL RESULTS

In this paper, conditional augmented models are applied to two tasks. The first uses the code-breaking approach of [13] to convert a large vocabulary speech recognition task into a series of binary problems, allowing comparison with MM augmented models. The second is based upon the TIMIT phone classification task [5]. Preliminary results are presented. Feature-spaces of all augmented and C-Aug models in this section consist of derivatives with respect to the means, variances and mixture-component priors of the base HMMs.

4.1. Code-breaking: binary classifiers

This is a multi-pass approach that first uses standard Viterbi decoding to generate a word lattice of the most likely word sequences. The lattice is then converted into a confusion network and pruned so that, at each point in time, a maximum of two words appear. These pairs of words are known as confusions. If sufficiently many occurrences of a confusion exist in the training set, a binary classifier can be trained to separate them.

The corpus used for experiments was a 400 hour subset of the Fisher LDC data. Training examples for highly confusable pairs

[3] It is tempting to claim that C-Aug models are forms of CRF. This is not, however, true since the CRF definition requires a Markov dependency between ω_t and ω_{t-1} [4] which C-Aug models do not have.

[4] Although the HCRF framework makes it relatively easy to increase the number of dependencies modelled, there is no guidance on which dependencies (features) are useful.

were extracted from the confusion networks and the number of positive and negative examples equalised by sampling: random selection therefore yields an accuracy of 50%. Performance was evaluated using 8-fold cross-validation on the training data. Although a number of classifiers were trained, only the pair CAN/CAN'T (with an ASR baseline of 21.5%) is described in this paper. All classifiers were trained using 3-state, 4-mixture-component HMMs. Augmented models were constructed using a 640-dimensional feature-space of derivatives with respect to selected means, variances and component priors.

Classifier	Criterion		Train (%)	Test (%)
	λ	α		
HMM	ML	–	10.4	11.0
HMM	MMI	–	9.0	10.4
Aug	ML	MM	7.1	9.2
C-Aug	ML	CML	7.3	9.1

Table 1. Training and test error rates for CAN/CAN'T

Table 1 shows clearly that both generative and conditional augmented models outperform the ML and MMI baselines. Both MM and CML classifiers perform similarly, yielding an absolute improvement of 1.2%. As the number of augmented parameters increased, MM training was found to be more robust to over-training than CML. For a 1897-dimensional score-space, MM models achieved test error of 9.1% compared to CML's 10.5%. Training errors were 6.1% and 5.0% respectively.

One of the issues with applying binary classifiers to multi-class problems is that separate classifiers must be built for all possible pairs of words. In practice, this is not possible. For example, in [2], 15 binary classifiers were used for rescoring a large vocabulary task. Although each classifier achieved reasonable gains, the small number of classifiers led to only a small improvement overall.

4.2. TIMIT

The multi-class performance of C-Aug models was evaluated using the standard TIMIT phone classification task. The experimental setup described in [11] was used. Models were trained with three states and either ten or twenty mixture-components. Acoustic model decoding was performed without the use of a language model. No data or feature whitening was performed.

Classifier	Criterion		Components	
	λ	α	10	20
HMM	ML	–	29.4	27.3
C-Aug	ML	CML	24.2	–
HMM	MMI	–	25.3	24.8
C-Aug	MMI	CML	23.4	–

Table 2. Classification error on the TIMIT core test set

The TIMIT results show a similar pattern to the pairwise experiments. In particular, it is clear from Table 2 that C-Aug models outperform both ML and MMI HMMs. Although one could argue that this is due to the extra parameters (C-Aug models have twice as many parameters as a standard HMM), this was not found to be the case: 10-component C-Aug models outperform 20-component HMMs. A point of particular interest is that despite poorer state segmentation—the sufficient statistics fix the state segmentation—C-Aug models with ML statistics outperformed the 20-component MMI HMM (24.2% versus 24.8% test error).

Despite good performance compared to standard HMMs, C-Aug models do not quite attain the performance of HCRFs [11]. This is believed to be due to three main factors: the fixed state segmentation from the base model, over-training (training error falls to 15.1% for MMI statistics) and lack of a language model (tests on MMI HMMs suggest that this may yield a gain of up to 0.5% absolute). Further research into segment optimisation techniques and regularisation is therefore required.

5. CONCLUSIONS

In this paper, conditional augmented models are proposed as a powerful acoustic model for speech recognition. C-Aug models share all of the benefits of generative augmented models but have the added advantage of a tractable normalisation term. The convex structure and global maxima make direct CML training simple. Initial results demonstrate that C-Aug models outperform both ML and MMI trained HMMs. Future work will examine higher-order dependencies [7], techniques for updating the base model (to allow state segmentation to vary), kernelisation of C-Aug models, and algorithms for performing recognition.

6. REFERENCES

[1] N.D. Smith, *Using Augmented Statistical Models and Score Spaces for Classification*, Ph.D. thesis, University of Cambridge, September 2003.

[2] M.J.F. Gales and M.I. Layton, "SVMs, score-spaces and maximum margin statistical models," in *Beyond HMM workshop, ATR*, 2004, http://mi.eng.cam.ac.uk/~mjfg/BeyondHMM.pdf.

[3] L.R. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Condition random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 591–598.

[5] J.S. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, 1993.

[6] S. Amari and S. Wu, *Methods of Information Geometry*, Oxford University Press, 2000.

[7] M.I. Layton and M.J.F. Gales, "Acoustic modelling using continuous rational kernels," in *MLSP*, 2005.

[8] J. Weston and C. Watkins, "Multi-class support vector machines," Tech. Rep. CSD-TR-98-04, Royal Holloway, University of London, May 1998.

[9] T. Jaakkola and D. Hausser, "Exploiting generative models in disciminative classifiers," in *NIPS*, 1999, pp. 487–493.

[10] J. Lafferty, X. Zhu, and Y. Liu, "Kernel conditional random fields: Representation and clique selection," in *ICML*, 2004.

[11] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden conditional random fields for phone classification," in *Interspeech*, 2005.

[12] M. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, pp. 525–533, June 1993.

[13] V. Venkataramani, S. Chakrabartty, and W. Byrne, "Support vector machines for segmental minimum Bayes risk decoding of continuous speech," in *ASRU 2003*, 2003.