STRUCTURED DISCRIMINATIVE MODELS FOR NOISE ROBUST CONTINUOUS SPEECH RECOGNITION

A. Ragni and M. J. F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge, CB2 1PZ, UK
{ar527,mjfg}@eng.cam.ac.uk

ABSTRACT

Recently there has been interest in structured discriminative models for speech recognition. In these models sentence posteriors are directly modelled, given a set of features extracted from the observation sequence, and hypothesised word sequence. In previous work these discriminative models have been combined with features derived from generative models for noise-robust speech recognition for continuous digits. This paper extends this work to medium to large vocabulary tasks. The form of the score-space extracted using the generative models, and parameter tying of the discriminative model, are both discussed. Update formulae for both conditional maximum likelihood and minimum Bayes' risk training are described. Experimental results are presented on small and medium to large vocabulary noise-corrupted speech recognition tasks: AURORA 2 and 4.

***Index Terms*—** Structured model, Noise robustness, Context modelling, Conditional Maximum Likelihood, Minimum Phone Error

1. INTRODUCTION

Most automatic speech recognition (ASR) systems use generative models, in the form of hidden Markov model (HMM), as the acoustic model. Likelihoods from these models are combined with the prior, the language model, using Bayes' rule to yield the sentence posterior. Although successful, it is widely known that the underlying models are not correct. This has lead to the interest in discriminative models, where the posterior probability is *directly* modelled. Depending on how the structure of sentences is modelled, many proposed discriminative models can be divided into *semi-structured* and *structured*. The semi-structured models, e.g. segmental conditional random fields (SCRF) [1], assume a word-level structure. The use of multiple feature streams at the word-level permits a range of events, such as an occurrence of phones, multi-phones and the whole words, to be incorporated. This flexibility enables a wide range of short and long-spanning dependencies. However, the current applications of SCRFs do not attempt to improve the underlying acoustic model, the recognition results from the standard HMM acoustic model are used to derive features and combined with other event detectors.

Structured models, on the other hand, maintain the standard medium to large vocabulary partitioning of words into sub-word, phone, units. This partitioning allows discriminative models to be used as the underlying acoustic model, as all possible words can be modelled, given an appropriate dictionary, which is not possible with the semi-structured models. An example of structured models is a conditional augmented (CAUG) model [2], where the range of possible dependencies is restricted to the phone level. At the phone level dynamic kernels based on generative models (HMM) provide a systematic approach to adding new dependencies through the use of competing likelihoods, and first and high-order derivatives. Extracting features based on generative models has the added advantage that state-of-the-art model-based compensation approaches to adapt features to noise/speaker conditions can be used [3]. The standard approach to training discriminative models is conditional maximum likelihood (CML). However alternative criteria such as minimum Bayes' risk (MBR) [2] and large margin (LM) training [4, 5] have also been proposed.

In previous work with CAUG models a small vocabulary noise-corrupted digit string recognition task based on whole-word HMMs was examined [5]. This paper extends the previous work to handle medium/large vocabulary continuous speech recognition tasks. There are two fundamental issues to handle. First an appropriate score-space is required. Using all possible models, is impractical. Thus context-dependent dynamic kernels are proposed to provide consistent and compact features for context-dependent classes. Second, an appropriate level for clustering the parameters of the discriminative model is required, this does not need to be the same as the generative model. In this paper the use of phonetic decision tree clustering to ensure that sufficient training data exists for robust parameter estimation is investigated. Discriminative model training using both CML and MBR criteria are described.

The paper is organised as follows. Section 2 describes the form of the structured model. Various types of features and the aspect of noise robustness is then detailed in Section 3. Parameter tying is described next. Section 5 provides reestimation formulae for CML and MWE/MPE training. Experimental results are given in Section 6. Finally, Section 7 presents the conclusions.

2. STRUCTURED MODEL

The structured model considered has the form of log-linear model

$$P(\mathbf{w}|\mathbf{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O},\boldsymbol{\theta},\mathbf{w};\boldsymbol{\lambda})\right)}{\sum_{\mathbf{w}'}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O},\boldsymbol{\theta},\mathbf{w}';\boldsymbol{\lambda})\right)} \quad (1)$$

where $\boldsymbol{\theta}$ segments the observation. For this work this is obtained from the generative model. This model has two parameter vectors: discriminative $\boldsymbol{\alpha}$ and generative $\boldsymbol{\lambda}$ parameters.

The form of feature vector $\boldsymbol{\phi}(\mathbf{O},\boldsymbol{\theta},\mathbf{w};\boldsymbol{\lambda})$ depends on which knowledge sources are available. Typically, features are extracted

Anton Ragni is funded by Toshiba Research Europe Ltd, EPSRC and HTK. The authors would like to thank Eric Wang for the help in setting up the AURORA 4 task.

from the acoustics, $\phi(\mathbf{O}, \boldsymbol{\theta}|\mathbf{w}; \boldsymbol{\lambda})$, and the language model, $\psi(\mathbf{w}; \boldsymbol{\lambda})$. Other features, for example the alignment posterior, $P(\boldsymbol{\theta}|\mathbf{w}; \boldsymbol{\lambda})$, can also be used. Therefore

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_{\mathtt{am}} \\ \boldsymbol{\alpha}_{\mathtt{lm}} \\ \alpha_{\mathtt{p}} \end{bmatrix} \quad \boldsymbol{\phi}(\mathbf{O}, \boldsymbol{\theta}, \mathbf{w}; \boldsymbol{\lambda}) = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{O}, \boldsymbol{\theta}|\mathbf{w}; \boldsymbol{\lambda}) \\ \boldsymbol{\psi}(\mathbf{w}; \boldsymbol{\lambda}) \\ \log\left(P(\boldsymbol{\theta}|\mathbf{w}; \boldsymbol{\lambda})\right) \end{bmatrix} \tag{2}$$

Note a bias term could also been added for generality.

An important issue is to decide at what level the latent variable $\boldsymbol{\theta}$ segments the data. This determines the level of conditional independence between the features. In previous work [5] the data was segmented at the word level, however this is not useful for medium-large vocabulary acoustic models. Here the data is segmented at the phone level. Thus given the alignment the dot-product in equation (1) evaluates to

$$\begin{aligned} \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O}, \boldsymbol{\theta}, \mathbf{w}; \boldsymbol{\lambda}) = \sum_{i=1}^{L} \boldsymbol{\alpha}_{\mathtt{am}}^{(w_i)\mathsf{T}} \boldsymbol{\phi}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}; \boldsymbol{\lambda}) + \\ \boldsymbol{\alpha}_{\mathtt{lm}}^{\mathsf{T}}\boldsymbol{\psi}(\mathbf{w}; \boldsymbol{\lambda}) + \alpha_{\mathtt{p}} \log\left(P(\boldsymbol{\theta}|\mathbf{w}; \boldsymbol{\lambda})\right) \end{aligned} \tag{3}$$

where $\boldsymbol{\phi}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}; \boldsymbol{\lambda})$ are features extracted from the sub-sequence $\mathbf{O}_{t(w_i,\boldsymbol{\theta})}$ and $\boldsymbol{\alpha}_{\mathtt{am}}^{(w_i)}$ are the associated acoustic parameters.

As this work is primarily interested in the acoustic model, only standard n-gram language models are used. Thus there is only a single dimensional language model feature. Furthermore the value of $\boldsymbol{\alpha}_{\mathtt{lm}}$ was not trained, it was empirically set to the standard language model scale-factor. The alignment posterior features were also not used in this work.

3. SCORE-SPACE FEATURES

This section describes the acoustic feature to be used by the structured model. The discussion focuses on features derived from generative models, as this allows model-based noise and speaker compensation schemes to be applied.

In this work the the feature-space derived from the generative models will be referred to as a *score-space*. Various score-spaces have been proposed in the literature [2], [6]. Examples include the *appended-all score-space*, $\boldsymbol{\phi}_{\mathtt{A}}^{(0)}(\mathbf{O}; \boldsymbol{\lambda})$

$$\boldsymbol{\phi}_{\mathtt{A}}^{(0)}(\mathbf{O}; \boldsymbol{\lambda}) = \begin{bmatrix} \log\left(p(\mathbf{O}|\omega_1; \boldsymbol{\lambda})\right) \\ \vdots \\ \log\left(p(\mathbf{O}|\omega_K; \boldsymbol{\lambda})\right) \end{bmatrix} \tag{4}$$

which incorporates the log-likelihoods of all models, including the correct class. This form of score-space allows the standard generative model to be obtained, simply by setting the value of $\boldsymbol{\alpha}$ to be one for the correct class, zero otherwise. Thus for class ω_1, this yields the sparse vector $\boldsymbol{\alpha}^{(\omega_1)} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^{\mathsf{T}}$. By appending the scores from competing classes enables a more informative score-space to be derived from the observation sequence.

Alternatively derivative score-spaces can be extracted. In addition to the log-likelihood information, the derivatives of the log-likelihood with respect to the generative parameters are used. The simplest example is the first-order *log-likelihood score-space*

$$\boldsymbol{\phi}_{\mathtt{L}}^{(1)}(\mathbf{O}|\omega; \boldsymbol{\lambda}) = \begin{bmatrix} \log\left(p(\mathbf{O}|\omega; \boldsymbol{\lambda})\right) \\ \nabla_{\boldsymbol{\lambda}} \log\left(p(\mathbf{O}|\omega; \boldsymbol{\lambda})\right) \end{bmatrix} \tag{5}$$

The first-order derivatives for HMMs are a function of component posterior probabilities, $P(\theta_t^{jm}|\mathbf{O}; \boldsymbol{\lambda})$, which depend on the whole sequence of observations. This means that the conditional independence assumptions of the underlying generative model are not maintained in the features extracted.

When medium/large vocabulary speech recognition systems are considered there is an issue with the appended-all score-spaces. The set of generative models comprises all context-dependent phone models. This yields a large score-space. Though in theory this could be used, the number of determinative model parameters becomes large. One option to address this problem is to include a small number of "suitable" models.

A simple approach is adopted in this paper, where the score-space includes every model that shares the same observed context. An example of the score-space with *matched context* is given below

$$\begin{bmatrix} \mathbf{a} - \mathtt{a} + \mathbf{c} \\ \vdots \\ \mathbf{a} - \mathtt{y} + \mathbf{c} \\ \mathbf{a} - \mathbf{z} + \mathbf{c} \end{bmatrix}_{K \times 1} \tag{6}$$

This reduces the dimensionality of the score-space to K.

One advantage of using generative models to define features is that model-based noise compensation can be used to make discriminative classifier robust to changes in noise/speaker conditions [3]. A popular and successful approach is based on vector Taylor series (VTS). In this work the first-order VTS scheme described in [7] is used. Considering just the *static* components the compensated mean and covariance in state j and component m are given by

$$\hat{\boldsymbol{\mu}}_{jm} = \mathbf{C} \log\left(\exp\left(\mathbf{C}^{-1}(\boldsymbol{\mu}_{jm} + \boldsymbol{\mu}_{\mathtt{h}})\right) + \exp\left(\mathbf{C}^{-1}\boldsymbol{\mu}_{\mathtt{n}}\right)\right) \tag{7}$$

$$\hat{\boldsymbol{\Sigma}}_{jm} = \mathbf{J}_{jm}\boldsymbol{\Sigma}_{jm}\mathbf{J}_{jm}^{\mathsf{T}} + (\mathbf{I} - \mathbf{J}_{jm})\boldsymbol{\Sigma}_{\mathtt{n}}(\mathbf{I} - \mathbf{J}_{jm})^{\mathsf{T}} \tag{8}$$

where convolutional noise mean $\boldsymbol{\mu}_{\mathtt{h}}$ and covariance $\boldsymbol{\Sigma}_{\mathtt{h}} = \mathbf{0}$, additive noise mean $\boldsymbol{\mu}_{\mathtt{n}}$ and covariance $\boldsymbol{\Sigma}_{\mathtt{n}}$ are the parameters of the noise model estimated from the data using maximum likelihood (ML) estimation [8]. Other terms in equations (7) and (8) include the discrete cosine transformation matrix $\mathbf{C}$ and the component-specific Jacobians $\mathbf{J}_{jm}$ fully described in [7].

4. DISCRIMINATIVE MODEL PARAMETER TYING

For small vocabulary systems, where whole-word models are used, the parameters of the discriminative model, $\boldsymbol{\alpha}$, are associated with the individual words. For larger systems, where the data is segmented at the phone-level and often state-level decision tree tying used to determine context-dependent models, the appropriate tying of the parameters is less clear. If there is sufficient training data the parameters could be specified at the context-dependent phone level, as determined by the state-level decision tree. However it is not possible to guarantee that all context-dependent models are observed in the training data, as complete context-dependent models are used for the score-spaces rather than state-level features.

To address this problem model-level parameter tying is performed to determine the appropriate tying of the discriminative model parameters. The standard approach based on phonetic decision trees [9] is used. However, care is required as the generative parameters are themselves tied at the state-level. When using two distinct decision trees, it is possible to get a *tree-intersect* style approach where the effective number of distinct models becomes very large. This can result in robustness issues when training the models.

There are several possible solutions that can be adopted. The one examined in this work consists of clustering only those discriminative parameters where the generative model for the correct class

appears at the leaf nodes of the decision trees created for generative models. The leaves of this model-level tree can be guaranteed[1] to have a minimum occupancy count in the training data and at least one distinct state. A consequence of this approach is that the maximum number of possible classes for the discriminative model is the number of distinct context-dependent models. The system is also sensitive to the context label assigned to each of the context-dependent generative models, this will be investigated in future work.

5. PARAMETER ESTIMATION

This section first provides the details of CML training and then describes a form of minimum Bayes' risk training. For brevity of presentation regularisation terms in the objective functions are omitted.

The CML training is the standard criterion maximising the average log-posterior of training data.

$$\mathcal{F}_{\texttt{cml}}(\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\log\Big(P(\mathbf{w}_{\texttt{ref}}^{(r)}|\mathbf{O}^{(r)};\boldsymbol{\lambda},\boldsymbol{\alpha})\Big) \tag{9}$$

where R is the number of training sentences. Although both $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ can be optimised in this work $\boldsymbol{\lambda}$ is assumed to be trained using, e.g., ML estimation. The standard MMI/MPE lattices [10] are used for efficiency. The gradient with respect to discriminative parameters $\boldsymbol{\alpha}_{\texttt{am}}$ is given by

$$\nabla_{\boldsymbol{\alpha}_{\texttt{am}}}\mathcal{F}_{\texttt{cml}}(\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{\mathbf{a}\in\mathbf{L}_{\texttt{num}}^{(r)}} P(\mathbf{a}|\mathbf{O}^{(r)};\boldsymbol{\lambda},\boldsymbol{\alpha})\boldsymbol{\phi}(\mathbf{O}_{t(\mathbf{a})}^{(r)},w;\boldsymbol{\lambda}) - \sum_{\mathbf{a}'\in\mathbf{L}_{\texttt{den}}^{(r)}} P(\mathbf{a}'|\mathbf{O}^{(r)};\boldsymbol{\lambda},\boldsymbol{\alpha})\boldsymbol{\phi}(\mathbf{O}_{t(\mathbf{a}')}^{(r)},w;\boldsymbol{\lambda}) \tag{10}$$

where $\mathbf{L}_{\texttt{num}}$ and $\mathbf{L}_{\texttt{den}}$ are numerator and denominator lattices respectively, $\mathbf{a}$ is a lattice arc, $P(\mathbf{a}|\mathbf{O};\boldsymbol{\lambda},\boldsymbol{\alpha})$ is arc posterior probability,

$$\boldsymbol{\phi}(\mathbf{O},w;\boldsymbol{\lambda}) = \begin{bmatrix}\delta(w,\omega_1)\boldsymbol{\phi}(\mathbf{O};\boldsymbol{\lambda})\\ \vdots \\ \delta(w,\omega_K)\boldsymbol{\phi}(\mathbf{O};\boldsymbol{\lambda})\end{bmatrix} \tag{11}$$

is a *composite* feature vector, where $\boldsymbol{\phi}(\mathbf{O};\boldsymbol{\lambda})$ is the standard feature vector given, e.g., in equation (4). The arc posterior probabilities can be computed using the standard MMI forward-backward algorithm [10] by replacing the HMM likelihood on the arc $\mathbf{a}$ with the dot-product $\boldsymbol{\alpha}_{\texttt{am}}^{(w)^{\mathsf{T}}}\boldsymbol{\phi}(\mathbf{O}_{t(\mathbf{a})};\boldsymbol{\lambda})$; the language model log-probability, the alignment log-posterior and the bias (if used) weighted by the corresponding discriminative parameters are added to each arc as usual. The combined quantity will be referred to as a CAUG *score*. The gradient expressions for other components of $\boldsymbol{\alpha}$ are similar and omitted here.

Another popular criterion is the MBR training. The objective is to minimise the expected loss in the average sentence accuracy

$$\mathcal{F}_{\texttt{mbr}}(\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{\mathbf{w}} P(\mathbf{w}|\mathbf{O}^{(r)};\boldsymbol{\lambda},\boldsymbol{\alpha})\mathcal{L}(\mathbf{w},\mathbf{w}_{\texttt{ref}}^{(r)}) \tag{12}$$

[1]Some decoders append both the silence (`sil`) and short-pause (`sp`) models to every pronunciation in the dictionary. Since the `sil` model prevents the expansion of the context some of the classes may lose training examples if the correct pronunciation with the `sp` model has been pruned away. In practice with sufficiently large pruning values this rarely happens.

where the loss function, $\mathcal{L}(\mathbf{w},\mathbf{w}_{\texttt{ref}})$, is computed between given $\mathbf{w}$ and the reference sentence $\mathbf{w}_{\texttt{ref}}$. Exactly computing $\mathcal{L}(\mathbf{w},\mathbf{w}_{\texttt{ref}})$ for all paths in a lattice is expensive so various approximations have been developed to allow computation to be carried out on the phone/word level. The approach used in this work [10] computes the accuracy of hypothesised phone/word w attached to arc $\mathbf{a}$ by looking for a reference phone/word $w_{\texttt{ref}}$ maximising

$$\mathcal{A}(w;\mathbf{w}_{\texttt{ref}}) = \max_{w_{\texttt{ref}}\in\mathbf{w}_{\texttt{ref}}}\begin{Bmatrix}-1+2d(w,w_{\texttt{ref}}), & \text{if } w = w_{\texttt{ref}}\\ -1+d(w,w_{\texttt{ref}}), & \text{if } w\neq w_{\texttt{ref}}\end{Bmatrix} \tag{13}$$

where $d(w,w')$ gives the amount of overlap in time between w and w'. In the exact case this expression yields 1, 0 and -1 for correct recognition, substitution and insertion error. In practice it is more convenient to work with phone accuracies. The gradient of the equivalent objective function to be maximised is given by

$$\nabla_{\boldsymbol{\alpha}_{\texttt{am}}}\mathcal{F}_{\texttt{mpe}}(\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{\mathbf{a}\in\mathbf{L}_{\texttt{den}}^{(r)}}\mathcal{C}(\mathbf{a},\mathbf{w}_{\texttt{ref}}^{(r)},\mathbf{L}_{\texttt{den}}^{(r)})P(\mathbf{a}|\mathbf{O}^{(r)};\boldsymbol{\lambda},\boldsymbol{\alpha})\boldsymbol{\phi}(\mathbf{O}_{t(\mathbf{a})}^{(r)},w;\boldsymbol{\lambda}) \tag{14}$$

where $\mathcal{C}(\mathbf{a},\mathbf{w}_{\texttt{ref}},\mathbf{L}_{\texttt{den}})$ denotes the average accuracy of sentences passing arc $\mathbf{a}$ minus the average accuracy of all sentence in $\mathbf{L}_{\texttt{den}}$. This is the standard quantity computed during the MWE/MPE training of HMM parameters [10]. The only difference is that similarly to the CML training the HMM likelihood on each arc is replaced by the CAUG score. As in the CML case the gradients with respect to other components of $\boldsymbol{\alpha}$ can be obtain in the similar way.

6. EXPERIMENTS

This section describes experiments with the structured discriminative models in AURORA 2 and 4. The AURORA 2 results are included to contrast the performance of CML and MBR training with large margin training published previously [5]. For all systems the discriminative models are initialised with the sparse parameter vector to yield generative model performance on the first iteration. The first order gradient-based optimisation with increasing step size is used (back-tracking is performed whenever required). To prevent over-training a development set was used to stop training, Set A for AURORA2 and Set C for AURORA4.

AURORA2 is a noise-corrupted connected digit string recognition task. The number of classes is 11 plus the `sil` and `sp` model. The generative model is a whole-word HMM with 16 states and 3 components/mixture trained using ML on the clean data. There are three test sets available for testing. The setup used follows the one described in [3]. The CAUG model is based on the appended-all score-space in equation (4), no language model is used. The number of discriminative parameters is 145. The multi-style data is used for training. The word error-rate (WER) performance of the VTS-compensated HMM and the CAUG models is shown in the following table. In Table 1 all forms of training of the CAUG model achieve gains over the baseline HMM system. The best results were obtained with large margin training. However the gains over MWE training are relatively small and large margin training is not easily mapped to the parallelisation required for training large systems. As a simple contrast a single dimension score-space system (using the correct class) was also constructed. Using the more complicated score-space with all models gave consistent gains over this simpler model.

AURORA 4 is a noise-corrupted medium to large vocabulary task based on the Wall Street Journal (WSJ) data. Two configurations

Classifier	Crit	Test Set			Avg
		A	B	C	
HMM	ML	9.8	9.1	9.5	9.5
CAUG	CML	8.1	7.7	8.3	8.1
	MWE	7.9	7.4	8.2	7.9
	LM	7.8	7.3	8.0	7.7

Table 1. AURORA2 Recognition Results

have been considered.[2] The first repeats the previous setup where the HMM is trained from clean data (SI-84 WSJ0 part, $\sim$14 hours). In the second more advanced VTS-adaptive training (VAT) is used to obtain the canonical HMM [8, 11]. The HMMs are state-clustered triphones ($\sim$3140 states) with $\sim$16 components/mixture. Multiple (4) iterations of VTS compensation are performed for the test data, the supervision hypothesis is updated after each cycle. The CAUG model is based on the context-dependent score-space in equation (6) and trained on the multi-style data. The language model parameters were fixed, only the most likely alignment was considered ($\alpha_{\mathrm{p}} = 0$) and no bias used. Evaluation is performed using the standard 5000-word WSJ0 bigram model on four noise-corrupted test sets[3] based on NIST Nov'92 WSJ0 test set.

System	Crit	Class	Test set				Avg
			A	B	C	D	
HMM	ML	-	7.1	15.3	12.2	23.1	17.8
CAUG	CML	47	7.2	14.7	11.1	22.8	17.4
		432	7.1	14.5	11.0	22.4	17.1
		4020	6.7	14.4	10.8	22.1	16.9
	MPE	47	7.3	14.7	11.2	22.7	17.4
		432	7.0	14.4	11.3	22.0	16.9
		4020	6.7	14.3	10.4	21.9	16.7

Table 2. AURORA4 Recognition Results

The first configuration investigates the usefulness of model-level phonetic-decision tree clustering (Section 4) for parameter tying. Table 2 shows the initial AURORA4 results with VTS-compensated HMM and CAUG models trained using CML and MPE. As expected as the number of classes increases performance improves. For all test sets the MPE-trained CAUG outperformed the CML-trained model, though all CAUG models outperformed the baseline average. When the number of distinct classes reaches 4000 the gains over the VTS-compensated HMM is more than 1% absolute. Note even for the largest CAUG configuration the increase in the number of model parameters is less than 5% compared to the baseline HMM system.

System	Crit	Test set				Avg
		A	B	C	D	
HMM (VAT)	ML	8.6	13.8	12.0	20.1	16.0
CAUG	MPE	7.5	13.0	10.9	19.4	15.3

Table 3. AURORA4 VTS Adaptively Trained Recognition Results

The second configuration used an VTS adaptively trained HMM system. The following table shows the baseline performance and MPE-trained CAUG with 4029 classes[4]. As expected the VAT system in Table 3 on average out-performed the baseline clean system, 16.0% compared to 17.8%. Again the use of a CAUG MPE trained model yielded gains for all test sets. Note for this configuration both the generative model and the discriminative model are trained on multi-style data.

[2]The results given here are based on `HDecode` from HTK V3.4. Using HTK 3.4.1 results in small differences in performance. This is due to the improvement introduced to HTK 3.4.1 in the path merging stage.

[3]The test set A is clean, set B has 6 types of noise added, set C has the channel distortion introduced and set D has both the additive noise and the channel distortion. SNR is randomly chosen bewteen 5 and 15 dB for test data and between 10 and 20 dB for multi-style training data.

7. CONCLUSIONS

This paper has described a structured discriminative model suitable for noise-robust medium/large vocabulary speech recognition. Here generative models, which can be compensated to handle speaker and noise changes, are used to extract features from the observation sequence. Previous work using whole-word models are extended to allow context-dependent sub-word model to be used. Context-dependent kernel are used to yield compact feature vectors at the phone level. Additionally model-level phonetic decision tree clustering of the discriminative model parameters, is described. Both conditional maximum likelihood and minimum Bayes' risk training of these models are detailed. The performance of classifier was evaluated on a simple AURORA 2 and more complex AURORA 4 noise-corrupted speech recognition tasks. Consistent gains has been observed over clean trained and VTS adaptively trained VTS-compensated HMM systems.

8. REFERENCES

[1] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *ASRU 2009*.

[2] M. Layton, *Augmented Statistical Models for Classifying Sequence Data*, Ph.D. thesis, Cambridge University, 2006.

[3] M.J.F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Computer Speech and Language*, vol. 24, pp. 648–662, 2010.

[4] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, pp. 1453–1484, 2005.

[5] S.-X. Zhang, A. Ragni, and M.J.F. Gales, "Structured log-linear models for noise robust speech recognition," *IEEE Signal Processing Letters*, vol. 17, pp. 945–948, 2010.

[6] N.D. Smith, *Using Augmented Statistical Models and Score Spaces for Classification*, Ph.D. thesis, Cambridge Uni., 2003.

[7] A Acero, L Deng, T Kristjansson, and J Zhang, "HMM Adaptation using Vector Taylor Series for Noisy Speech Recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[8] H. Liao and M.J.F. Gales, "Joint uncertainty decoding for robust large vocabulary speech recognition," Tech. Rep. CUED/F-INFENG/TR552, Cambridge Uni., November 2006.

[9] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA Workshop HLT*, 1994, pp. 307–312.

[10] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge Uni., 2004.

[11] O Kalinli, ML Seltzer, and A Acero, "Noise adaptive training using a vector Taylor series approach fo noise robust automatic speech recognition," in *Proc. ICASSP*, 2009.

[4]The difference in the number of classes between clean and VAT system comes from the slightly different number of training utterances used due to very large pruning values otherwise required for alignment.