INFERENCE ALGORITHMS FOR GENERATIVE SCORE-SPACES

A. Ragni and M. J. F. Gales

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

ABSTRACT

Using generative models, for example hidden Markov models (HMM), to derive features for a discriminative classifier has a number of advantages including the ability to make the features robust to speaker and noise changes. An interesting attribute of the derived features is that they may not have the same conditional independence assumptions as the underlying generative models, which are typically first-order Markovian. For efficiency these features are derived given a particular segmentation. This paper describes a general algorithm for obtaining the optimal segmentation with combined generative and discriminative models. Previous results, where the features were constrained to have first-order Markovian dependencies, are extended to allow derivative features to be used which are non-Markovian in nature. As an example, inference with zero and first-order HMM score-spaces is considered. Experimental results are presented on a noise-corrupted continuous digit string recognition task: AURORA 2.

***Index Terms*—** Structured discriminative model, generative score-space, inference

1. INTRODUCTION

Currently most automatic speech recognition (ASR) systems are based on generative hidden Markov models (HMM). The likelihoods from these are usually combined with the n-gram language model probabilities using Bayes' rule to yield the sentence posterior. Though successful it is widely known that the underlying models are not correct. This has lead to interest in discriminative models which *directly* model the posterior/decision boundaries given a set of features extracted from the observation sequence. Depending on how the structure of sentences is modelled many proposed discriminative models can be divided into *flat* and *structured*. Flat models [1] assume no specific structure which allows to model sentence-wide phenomena. Unfortunately the space of possible sentences is large which makes using these models in ASR complicated. Structured models [2, 1, 3] on the other hand assume partitioning of sentences into basic structural units such as words or phones which is believed to be more appropriate for ASR.

Several options exist to extract features at different structural levels. These include *event detectors* [1] and *generative score-spaces* [2]. Event detectors make use of classifiers and detectors to provide parallel feature (event) streams. The feature streams simultaneously operating at word, phone and subphone levels allow both short and long-spanning dependencies to be incorporated. In order to derive features from the detected events there are operations such as existence, expectation and string edit distance available. Generative score-spaces derive features from generative models and provide systematic approaches to define acoustic features in the form of zero and higher-order derivatives of log-likelihood. The features may inherit or break the underlying model conditional independence assumptions, model complex within and across-state dependencies spanning multiple frames. As the generative models are used to extract features they can be adapted to noise and speaker conditions using model-based techniques. One issue with these features is that they are derived from segmented observation sequences and thus are sensitive to particular segmentation. In previous work with generative score-spaces [3] the segmentation for training was obtained from generative models. However, using the same approach for inference is expected to yield suboptimal performance when the discriminative model trained is sufficiently different from the HMMs used to produce segmentations.

Previous work in this area examined the use of optimal segmentation with a specific type of features [4]. This paper extends that work describing a general inference algorithm suitable for zero and first-order generative score-spaces in particular. This paper shows that for a class of generative models efficient inference with zero and first-order score-spaces is possible even though first-order score-spaces require non-Markovian statistics to be estimated. The rest of this paper is organised as follows. Section 2 describes structured discriminative models and features derived by generative score-spaces. The inference algorithm is presented in Section 3. An example application of the algorithm to ASR is given in Section 4. Finally, Section 5 gives conclusions drawn from this work.

2. STRUCTURED DISCRIMINATIVE MODELS

Given observation sequence $\mathbf{O}$ the discriminative model considered in this work models the posterior probability of sentence $\mathbf{W}$ using log-linear form

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

where $\boldsymbol{\alpha}$ are model parameters, $\boldsymbol{\theta}$ segments observations into structural units and $\boldsymbol{\phi}(\mathbf{O},\mathbf{W},\boldsymbol{\theta})$ is a joint feature vector.

2.1. Structure

For efficiency the dot-product of model parameters with joint feature vector or *sentence score* can be decomposed as a sum of dot-products

Anton Ragni is jointly funded by Toshiba Research Europe Ltd, EPSRC and HTK. The authors would like to thank Dr Federico Flego for providing the adaptively trained HMM system.

at various levels such as word, phone, etc. This paper examines models defined on a single level such as the word level

$$\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O},\mathbf{W},\boldsymbol{\theta}) = \sum_{i=1}^{L}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})},w_i) \tag{2}$$

where w_i is a word, L is the number of words in $\mathbf{W}$ and $t(w_i,\boldsymbol{\theta})$ indexes observations assigned to w_i by $\boldsymbol{\theta}$.

The structure of the model considered can be compactly represented using lattices. Figure 1 shows a typical lattice used for modelling denominator terms in equation (1). Compared to the lattices

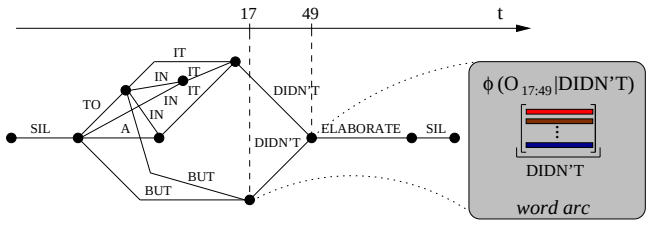

Fig. 1. Sentence structure modelling using lattices

used in discriminative HMM training, each word arc in Figure 1 is augmented by a *set* of acoustic features shown as the column vector. Note that using HMM log-likelihoods as the basic word-level acoustic features and setting the corresponding discriminative model parameters to one and the rest to zero allows to retrieve the performance of the standard HMM classifier. In this work language model was not used though approaches exist to incorporate it [1, 2].

2.2. Features

Generative score-spaces derive features from generative models. The simplest example, zero-order base (b) score-space, is given by

$$\boldsymbol{\phi}^{0}_{\mathtt{b}}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|w_i) = \left[\log\left(p(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|w_i)\right)\right] \tag{3}$$

When the structured discriminative models are based on $\boldsymbol{\phi}^{0}_{\mathtt{b}}$ features the parameters trained could be interpreted as class-specific *acoustic deweighting constants*. The main issue with these features is that they inherit conditional independence assumptions of the underlying generative models. This means that $\boldsymbol{\phi}^{0}_{\mathtt{b}}$ admits only *Markovian statistics* in the features extracted. The first-order score-spaces allow to address this by incorporating *non-Markovian statistics* in the form of derivatives of the log-likelihood. For example, the first-order base score-space has features

$$\boldsymbol{\phi}^{1}_{\mathtt{b}}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|w_i) = \begin{bmatrix}\log\left(p(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|w_i)\right)\\ \nabla_{\boldsymbol{\lambda}}\log\left(p(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|w_i)\right)\end{bmatrix} \tag{4}$$

where $\boldsymbol{\lambda}$ are generative model parameters. Consider the derivatives with respect to HMM mean vectors

$$\nabla_{\boldsymbol{\mu}_{jk}}\log\left(p(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|w_i)\right) = \sum_{t'=\tau}^{t} P(\theta^{jk}_{t'}|\mathbf{O}_{t(w_i,\boldsymbol{\theta})})\boldsymbol{\Sigma}^{-1}_{jk}(\mathbf{o}_{t'}-\boldsymbol{\mu}_{jk}) \tag{5}$$

where $t(w_i,\boldsymbol{\theta})$ is assumed to index observations from time τ to time t. Since these derivatives are functions of component posteriors $P(\theta^{jk}_{t'}|\mathbf{O}_{t(w_i,\boldsymbol{\theta})})$, which depend on the whole subsequence $\mathbf{O}_{t(w_i,\boldsymbol{\theta})}$, the HMM conditional independence assumptions are no longer present in the features extracted. One advantage of using generative score-spaces in ASR is that model-based adaptation and compensation approaches can be applied to compensate generative model parameters and thus the features derived to target speaker and noise conditions [5]. By compensating the features it is possible to train speaker and noise independent discriminative classifiers.

Given the features in equation (3) or (4) the joint word-level feature vector to be used in equation (2) is constructed as follows

$$\boldsymbol{\phi}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})},w_i) = \begin{bmatrix}\delta(w_i,\omega_1)\boldsymbol{\phi}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|\omega_1)\\ \vdots\\ \delta(w_i,\omega_W)\boldsymbol{\phi}(\mathbf{O}_{t(w_i,\boldsymbol{\theta})}|\omega_W)\end{bmatrix} \tag{6}$$

where W is the number of word classes and the use of delta-functions ensures that only one class is active on each word arc.

2.3. Parameter estimation

The standard criterion to use with log-linear models is a conditional maximum likelihood. For tasks such as ASR another popular criterion is a minimum Bayes' risk (MBR). In this work the variant maximising *accuracy* (a) of transcriptions [3] will be used

$$\mathcal{F}^{\mathtt{a}}_{\mathtt{mbr}}(\boldsymbol{\alpha}) = \sum_{r=1}^{R}\sum_{\mathbf{W}} P(\mathbf{W}|\mathbf{O}^{(r)};\boldsymbol{\alpha})\mathcal{A}(\mathbf{W},\mathbf{W}^{(r)}_{\mathtt{ref}}) \tag{7}$$

where $\mathbf{W}^{(r)}_{\mathtt{ref}}$ is a reference transcription and accuracy $\mathcal{A}(\mathbf{W},\mathbf{W}^{(r)}_{\mathtt{ref}})$ is defined on the word level. The gradient with respect to parameters of structural units has the following general form

$$\nabla_{\boldsymbol{\alpha}}\mathcal{F}^{\mathtt{a}}_{\mathtt{mbr}}(\boldsymbol{\alpha}) = \sum_{r=1}^{R}\sum_{\mathbf{a}\in\mathbf{L}^{(r)}_{\mathtt{den}}} \mathcal{C}(\mathbf{a})P(\mathbf{a}|\mathbf{O}^{(r)})\boldsymbol{\phi}(\mathbf{O}^{(r)}_{t(\mathbf{a})},\boldsymbol{\omega}) \tag{8}$$

where $\mathbf{a}$ is an arc with class label $\boldsymbol{\omega}$, $\mathbf{L}^{(r)}_{\mathtt{den}}$ is a lattice encoding possible transcriptions including the reference, $P(\mathbf{a}|\mathbf{O}^{(r)})$ and $\mathcal{C}(\mathbf{a})$ are arc posterior probabilities and contributions to the average correctness computed using variants of lattice-based forward-backward algorithms [6]. In this work regularised MBR training is performed by adding a fixed-mean Gaussian prior to the objective function.

3. INFERENCE

It has been assumed so far that segmentation $\boldsymbol{\theta}$ for each sentence $\mathbf{W}$ is known both in training and decoding. Since score-space features are derived from the segmented observation sequences the choice of segmentation is important. The simplest option to obtain it is to use *generative model alignment*

$$\boldsymbol{\theta}_{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\theta}}\left\{P(\mathbf{W})P(\boldsymbol{\theta}|\mathbf{W})p(\mathbf{O}|\boldsymbol{\theta},\mathbf{W})\right\} \tag{9}$$

which maximises the sentence likelihood. An alternative approach is to use *optimal alignment*

$$\boldsymbol{\theta}_{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\theta}}\left\{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{O},\mathbf{W},\boldsymbol{\theta})\right\} \tag{10}$$

which maximises the sentence score and is more directly linked with the objective function. When the discriminative model is initialised with sparse parameters as described in Section 2.1 the alignment produced by the generative model coincides with the optimal. In general, as the discriminative parameters change so does the optimal alignment, $\boldsymbol{\theta}_{\boldsymbol{\alpha}} \neq \boldsymbol{\theta}_{\boldsymbol{\lambda}}$. This means that each estimation step in training or the best path search in decoding should be preceded by *inferring* optimal alignments for all possible sentences.

3.1. General inference algorithm

The *inference problem* with the discriminative model in equation (1) can be formulated by

$$\{\mathbf{W}_{\boldsymbol{\alpha}}, \boldsymbol{\theta}_{\boldsymbol{\alpha}}\} = \arg\max_{\{\mathbf{W},\boldsymbol{\theta}\}} \{\boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{O}, \mathbf{W}, \boldsymbol{\theta})\} \tag{11}$$

which subsumes equation (10) as a special case. In this work the sentence score is decomposed as a sum of word-level scores which yields an efficient, polynomial time, algorithm. For a model m let $\rho_t^{(m)}$ denotes the best score at time t. Then a *recursion* is given by

$$\rho_t^{(m)} = \max_{m',\tau-1} \left\{ \rho_{\tau-1}^{(m')} + c_{m',m} + \boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{O}_{\tau:t}, \omega_m) \right\} \tag{12}$$

where $c_{m',m}$ is a cost on transiting from m' to m, $\tau \in [1, t]$ and maximisation is over all possible segmentations $\boldsymbol{\theta}_{\tau:t}$ and preceding models m'. By running the algorithm from time 1 to time T the optimal sentence and segmentation can be obtained by tracing back the model and time index maximising $\rho_T^{(M)}$. Note that equation (12) is an extension of Viterbi algorithm from frames to sequences and is identical to the one used with semi-Markov conditional random fields [7]. If complexity of computing the dot products $\boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{O}_{\tau:t}, \omega_m)$ was constant the inference algorithm in the worst case would have had *quadratic* complexity $\mathcal{O}(W^2T^2)$, where T and W are the number of frames and models. Since language model is not used in this work $c_{m',m} = 0$ which allows to perform maximisation over m' independently from that over $\tau - 1$. This reduces worst case complexity to $\mathcal{O}(WT^2)$. The next two sections examine inference with zero and first-order generative score-spaces.

3.2. Zero-order score-spaces

For zero-order score-spaces (3) the recursion involves computing

$$\boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{O}_{\tau:t}, \omega_m) = \alpha^{(\omega_m)} \log\left(p(\mathbf{O}_{\tau:t}|\omega_m)\right) \tag{13}$$

Equations (12) and (13) suggest that each time new segmentation $\boldsymbol{\theta}_{\tau:t}$ is considered the likelihood has to be re-computed. For general generative models this computation is expected to be expensive. However, some models, for example HMMs, can use efficient recursions in the form of forward and backward probabilities. Unfortunately, the use of *utterance-level* forward/backward passes over the lattice is not sufficient as it would only yield a *subset* of likelihoods

$$\left\{ \begin{array}{cc} p(\mathbf{O}_{1:1}|\omega_m), & p(\mathbf{O}_{T:T}|\omega_m) \\ \vdots & \vdots \\ p(\mathbf{O}_{1:T-1}|\omega_m), & p(\mathbf{O}_{2:T}|\omega_m) \\ \underbrace{p(\mathbf{O}_{1:T}|\omega_m),}_{\text{forward pass}} & \underbrace{p(\mathbf{O}_{1:T}|\omega_m)}_{\text{backward pass}} \end{array} \right\} \tag{14}$$

In order to compute the *full set* of likelihoods for each class ω_m, in general, T *segment-level* forward or backward passes must be performed to provide equation (12) with all possible dot-products. In this work the backward pass was integrated into the recursion. For each $\rho_t^{(m)}$ it was run from time t to time 1 to yield t likelihoods required. Since the complexity of the backward algorithm is linear in time T the total complexity remains *quadratic*.

In addition to HMMs, examples of generative models allowing efficient inference are factor-analysed and uncoupled factorial HMMs, buried Markov models. Models not in this category are trajectory HMMs, switching linear dynamic systems, i.e., generative models without state and observation Markov assumptions.

3.3. First-order score-spaces

With first-order score-spaces (4) the dot-product required is

$$\begin{aligned} &\boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{O}_{\tau:t}, \omega_m) = \\ &\quad \alpha^{(\omega_m)} \log(p(\mathbf{O}_{\tau:t}|\omega_m)) + \boldsymbol{\alpha}^{(\omega_m)\mathsf{T}} \nabla_{\boldsymbol{\lambda}} \log(p(\mathbf{O}_{\tau:t}|\omega_m)) \end{aligned} \tag{15}$$

For general generative models efficient inference is not possible since log-likelihood and its derivatives may require segmented statistics to be estimated. However, when the generative model is constrained to be Markovian, such as in the previous section, efficient inference is possible even though the derivatives would require non-Markovian statistics to be used. This is the case considered in this paper. The derivative term in equation (15) for these generative models has the following form

$$\boldsymbol{\alpha}^{(\omega_m)\mathsf{T}} \nabla_{\boldsymbol{\lambda}} \log(p(\mathbf{O}_{\tau:t}|\omega_m)) = \sum_{t'=t}^{\tau} \sum_{j=2}^{N-1} \sum_{k=1}^{K} P(\theta_{t'}^{jk}|\mathbf{O}_{\tau:t}) w_{jkt'}^{(\omega_m)} \tag{16}$$

where $w_{jkt'}^{(\omega_m)}$ is a *discriminative component score* defined by

$$w_{jkt'}^{(\omega_m)} = \boldsymbol{\alpha}_{jk}^{(\omega_m)\mathsf{T}} \nabla_{\boldsymbol{\lambda}} \log(p(\mathbf{o}_{t'}|\theta^{jk})) \tag{17}$$

One issue with equation (16) is that it depends on the segment-level posterior probabilities

$$P(\theta_{t'}^{jk}|\mathbf{O}_{\tau:t}) = \frac{f_{\tau:t'}^{jk} \cdot b_{t':t}^{j}}{p(\mathbf{O}_{\tau:t}|\omega_m)} \qquad t' \in [\tau, t] \tag{18}$$

where the numerator term is a product of segment-level forward $f_{\tau:t'}^{jk} = p(\mathbf{O}_{\tau:t'}, \theta_{t'}^{jk})$ and backward $b_{t':t}^{j} = p(\mathbf{O}_{t'+1:t}|\theta_{t'}^{j})$ probabilities. If recursion is implemented directly the worst case complexity of inference becomes *cubic* in T.

In addition to having *cubic* worst time complexity the direct inference is inefficient in re-computing forward and backward probabilities in equation (18) once τ or t has changed. In order to reduce the amount of computations required the following *two-pass strategy* was used. During the *first pass* all possible segment-level forward probabilities $\{f_{\tau:t}^{jk}\}$ are computed and cached. These are obtained by starting from each time $1, \ldots, T$. This also yields the segment-level likelihood for any subsequence

$$p(\mathbf{O}_{\tau:t}|\omega) = \sum_{i=2}^{N-1} a_{iN} f_{\tau:t}^{i} \tag{19}$$

where $\{a_{ij}\}$ are transition probabilities and $f_{\tau:t}^{i}$ can be derived from $f_{\tau:t}^{ik}$. The *second pass* is integrated into recursion similar to the inference in Section 3.2 but is more complicated. During this pass all possible segment-level backward probabilities$\{b_{\tau:t}^{j}\}$ are computed. Since the corresponding segment-level likelihoods and forward probabilities at each time $t' \in [\tau, t]$ are already known (see the first pass) they can be combined to yield the posterior in equation (18) for any $b_{t':t}^{j}$. By accumulating the product of posterior with the component score yields the term in equation (16). Since the log-likelihood is known the first term on the right side of equation (15) can be also added. Therefore by reaching time τ the complete dot-product is accumulated. Note that since the set of forward probabilities $\{f_{\tau:t}^{j}\}$ for each starting time τ is different accumulation over time t' in equation (16) has to be treated *independently*.

Note that significantly more efficient inference is possible using *first-order expectation semirings* [8]. These allow to obtain all possible dot-products (16) in a single backward pass from time t to 1. Then similarly to zero-order case inference with first-order score-spaces can be performed in *quadratic time*.

4. EXPERIMENTS

This section describes an application of the algorithms in Section 3 to HMM score-spaces. Both zero and first-order HMM score-spaces are considered which requires inference in the first-order Markovian and non-Markovian feature spaces. For simplicity a small vocabulary task was considered where utterances are sufficiently short so that quadratic and cubic complexities are easy to handle. Additionally optimal inference was only performed during decoding.

AURORA 2 is a noise-corrupted connected digit string recognition task. The number of classes is 11 plus silence and short pause, no language model was used. The generative model of digits is a whole-word HMM with 16 states and 3 components/state. The number of HMM parameters is 46,732. For each utterance model-based vector Taylor series (VTS) compensation was applied using the approach described in [5]. Three HMM setups were considered: clean-trained (VTS), VTS-adaptively trained (VAT) and discriminatively VTS-adaptively trained (DVAT) [9] systems. The discriminative model is based on ϕ_{b}^{0} and $\phi_{\mathrm{b}}^{1\mu}$ score-spaces where μ denotes that derivatives with respect to mean vectors were only used.[1] The number of discriminative model parameters is 13 and 21,554 respectively. The maximum word accuracy criterion was used to train discriminative models on multi-style data using suboptimal alignments. Test set A was used as the validation set to stop training.

The first experiment investigated inference with zero-order score-spaces, ϕ_{b}^{0}, where only first-order Markovian statistics is used. Table 1 shows evaluation results. The first line of each block

HMM	SDM	$\boldsymbol{\theta}$	Test set A	Test set B	Test set C	Avg
VTS	–	–	9.8	9.1	9.5	9.5
	ϕ_{b}^{0}	$\boldsymbol{\theta}_{\boldsymbol{\lambda}}$	8.1	7.4	8.2	7.8
		$\boldsymbol{\theta}_{\boldsymbol{\alpha}}$	7.8	7.3	8.0	7.6
VAT	–	–	8.9	8.3	8.8	8.6
	ϕ_{b}^{0}	$\boldsymbol{\theta}_{\boldsymbol{\lambda}}$	7.6	7.3	7.9	7.5
		$\boldsymbol{\theta}_{\boldsymbol{\alpha}}$	7.1	6.8	7.5	7.1
DVAT	–	–	6.7	6.6	7.0	6.7
	ϕ_{b}^{0}	$\boldsymbol{\theta}_{\boldsymbol{\lambda}}$	6.7	6.5	7.0	6.7
		$\boldsymbol{\theta}_{\boldsymbol{\alpha}}$	6.6	6.5	6.9	6.6

Table 1. Inference results with zero-order generative score-spaces

shows HMM word error rate (WER) performance. As expected the use of adaptive (VAT) and discriminative adaptive (DVAT) training improves the performance. The second line in each block shows the performance of structured discriminative models (SDM) when suboptimal alignments were used in decoding. In all configurations considered the SDMs perform at least as good as the HMMs though the number of additional parameters is just 13. Looking at the first (VTS) block small 3% relative improvement can be observed from using optimal $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ rather than suboptimal $\boldsymbol{\theta}_{\boldsymbol{\lambda}}$ HMM alignments. This observation is consistent with [4] where similar gains were reported with another zero-order score-space having 10 times more parameters. Slightly larger 5% relative improvement can be observed in the second (VAT) block which is believed to be due to significantly more data available for training HMMs. When the DVAT HMM setup was considered the use of optimal alignment yielded small 1-2% relative improvement. However, in this case the SDM gives improvement on test set B only.

[1] Addition of covariance derivatives, which doubles the number of parameters, has lead to small improvements on top of $\phi_{\mathrm{b}}^{1\mu}$ in the VTS setup.

In the second experiment more complex first-order score-spaces, $\phi_{\mathrm{b}}^{1\mu}$, with non-Markovian statistics were used. Table 2 shows evaluation results. The SDM based on $\phi_{\mathrm{b}}^{1\mu}$ outperforms the HMM and

HMM	SDM	$\boldsymbol{\theta}$	Test set A	Test set B	Test set C	Avg
VTS	$\phi_{\mathrm{b}}^{1\mu}$	$\boldsymbol{\theta}_{\boldsymbol{\lambda}}$	7.0	6.6	7.6	7.0
		$\boldsymbol{\theta}_{\boldsymbol{\alpha}}$	6.8	6.4	7.3	6.7
VAT		$\boldsymbol{\theta}_{\boldsymbol{\lambda}}$	6.6	6.5	7.0	6.6
		$\boldsymbol{\theta}_{\boldsymbol{\alpha}}$	6.2	6.1	6.8	6.3
DVAT		$\boldsymbol{\theta}_{\boldsymbol{\lambda}}$	6.1	6.2	6.7	6.3
		$\boldsymbol{\theta}_{\boldsymbol{\alpha}}$	6.1	6.1	6.6	6.2

Table 2. Inference results with first-order generative score-spaces

zero-order score-space in each setup considered (Table 1). The results in Table 2 show that similar 4-5% relative improvement can be observed from inference in the VTS and VAT HMM setup. The use of DVAT HMM as the base generative model again reduces the gain from the optimal alignment to 1-2% relative.

5. CONCLUSIONS

This paper has examined inference with generative score-spaces which derive features from generative models (HMMs). When these are used by structured discriminative models the derived features are dependent on the segmentation which is typically obtained from the HMMs. The use of optimal segmentation for general generative models is complicated. This paper has described the inference algorithm suitable for zero and first order score-spaces based on a class of generative models. These generative models are required to be first-order Markovian though derivative features are non-Markovian in nature. An efficient recursion has been presented for the zero and first-order score-spaces. An example application was performed in a noise-corrupted small vocabulary task where 1-5% relative gains over suboptimal segmentation were observed.

6. REFERENCES

[1] G. Zweig and P. Nguyen, "From flat direct to models to segmental CRF models," in *Proc. ICASSP*, 2010.

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

[3] A. Ragni and M. Gales, "Structured discriminative models for noise robust speech recognition," in *Proc. ICASSP*, 2011.

[4] S.-X. Zhang and M. Gales, "Structured support vector machines for noise robust continuous speech recognition," in *Proc. Interspeech*, 2011.

[5] M. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *CSL*, 2010.

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

[7] S. Sarawagi and W. Cohen, "Semi-markov conditional random fields for information extraction," in *Proc. NIPS*, 2005.

[8] R. C. van Dalen, A. Ragni, and M. J. F. Gales, "Efficient decoding with continuous rational kernels using the expectation semiring," Tech. Rep. CUED/F-INFENG/TR.674, 2012.

[9] F. Flego and M.J.F. Gales, "Discriminative adaptive training with VTS and JUD," in *Proc. ASRU*, 2009.