MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Structured Discriminative Models for Speech Recognition

Gales, M.; Watanabe, S.; Fosler-Lussier, E.

TR2012-072 November 2012

Abstract

AbstractAutomatic Speech Recognition (ASR) systems classify structured sequence data, where the label sequences (sentences) must be inferred from the observation sequences (the acoustic waveform). The sequential nature of the task is one of the reasons why generative classifiers, based on combining hidden Markov model (HMM) acoustic models and N-gram language models using Bayes rule, have become the dominant technology used in ASR. Conversely, the machine learning and natural language processing (NLP) research areas are increasingly dominated by discriminative approaches, where the class posteriors are directly modelled. This paper describes recent work in the area of structured discriminative models for ASR. To handle continuous, variable length, observation sequences, the approaches applied to NLP tasks must be modified. This paper discusses a variety of approaches for applying structured discriminative models to ASR, both from the current literature and possible future approaches. We concentrate on structured models themselves, the descriptive features of observations commonly used within the models, and various options for optimizing the parameters of the model.

IEEE Signal Processing Magazine

Structured Discriminative Models For Speech Recognition

Mark Gales, *Fellow, IEEE*, Shinji Watanabe, *Senior Member, IEEE*, Eric Fosler-Lussier *Senior Member, IEEE*

***Abstract*—Automatic Speech Recognition (ASR) systems classify structured sequence data, where the label sequences (sentences) must be inferred from the observation sequences (the acoustic waveform). The sequential nature of the task is one of the reasons why generative classifiers, based on combining hidden Markov model (HMM) acoustic models and N-gram language models using Bayes' rule, have become the dominant technology used in ASR. Conversely, the machine learning and natural language processing (NLP) research areas are increasingly dominated by discriminative approaches, where the class posteriors are directly modelled. This paper describes recent work in the area of structured discriminative models for ASR. To handle continuous, variable length, observation sequences, the approaches applied to NLP tasks must be modified. This paper discusses a variety of approaches for applying structured discriminative models to ASR, both from the current literature and possible future approaches. We concentrate on structured models themselves, the descriptive features of observations commonly used within the models, and various options for optimizing the parameters of the models.**

I. Introduction

The dominant technology for Automatic Speech Recognition (ASR) is based on generative models: Hidden Markov Models (HMMs) [1] are typically used as the acoustic models to derive the likelihood of a particular class generating an observation sequence. This is combined with a prior, e.g., an N-gram language model [2], to yield a posterior probability of the class given the observation. Acceptable performance in generative models is accomplished via refinements to the standard HMM acoustic models, including context-dependent modelling, speaker adaptation, discriminative training, and noise compensation [3].

Though current state-of-the-art systems yield satisfactory recognition rates in some domains, performance is generally not good enough for speech applications to become ubiquitous. In discriminative models the posterior probability of the classes (sentences) given the observations are directly modelled. This type of model has the potential to improve performance as a wider range of features from the observation and word sequences can be used for inference compared to generative models. These discriminative models have started to dominate the area of Natural Language Processing (NLP) [4], [5]. One issue in NLP training is that text data comprises variable length sequences of words yielding a vast number of possible classes. It is thus rarely possible to robustly construct models of complete word sequences (sentences). To handle this, structure must be introduced into the classifier by breaking the sentence into smaller units, typically words.

Applying these forms of discriminative classifiers to ASR adds another level of complexity. The observed data comprises sequences of observations, often continuous valued feature vectors, extracted at a fixed frame rate. The word sequences associated with these observations must then be inferred. Thus the number of labels (the word sequence) and the number of observations (frames) differ. For approaches such as Conditional Random Fields (CRFs) [5], [6], there is an implied assumption that the number of labels and observations are the same.[1] To address this problem it is possible to introduce latent variables into CRFs, yielding Hidden CRFs (HCRFs) [9], [10], and make use of sequence kernels and score-spaces [11], [12]. Models that handle this type of data will be referred to as *structured discriminative*

Mark J.F. Gales is with the Cambridge University Engineering Department, Trumpington St., Cambridge CB2 1PZ, U.K. (email:mjfg@eng.cam.ac.uk)

Shinji Watanabe is with Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge MA, 02139, USA. (e-mail: watanabe@merl.com)

Eric Fosler-Lussier is with the Department of Computer Science & Engineering, The Ohio State University, 2015 Neil Ave, Columbus OH, 43210, USA. (e-mail: fosler@cse.ohio-state.edu)

Manuscript received October 1, 2011; revised ??, 2011.

[1]CRFs (and related approaches) can be applied to ASR by using labels that contain an implicit segmentation; for example, a best-path frame labeling posterior with multiple labels per segment can give rise to a segmentation by collapsing repeated instances of labels together [7]. Single class labels can also be obtained using SVMs and sequence kernels [8]. However this paper will focus on the situations where there are sequences of labels associated with the observations.

models. There are a number of approaches that have been applied to ASR which can be described within this framework: log-linear models [13]–[15], Structured Support Vector Machines (SSVMs) [16], HCRFs [9], Segmental CRFs (SCRFs) [17], Conditional Augmented-Models (CAugs) [18], Maximum Entropy Markov Models (MEMMs) [19], Augmented CRFs (ACRFs) [20]. These models differ from each other in terms of the observation features considered, training criterion and how the latent variables are handled.

In addition to models that directly map from the observation sequence to the word-sequence, probability distributions over the word-sequences can also be represented in the same form, yielding discriminative language models [21]. These can either be used in combination with generative acoustic models via Bayes' rule, or as part of a discriminative model.

This paper gives an overview of a number of discriminative sequence and language models. The following sections will describe the general forms of this type of model, the criteria that can be used to train them, and some example applications to speech recognition.

II. Sequence Models and Classification

ASR can be viewed as a structured sequence classification task: there is a sequence of observations from which a single sentence hypothesis must be inferred. Consider a set of T observations $\boldsymbol{O}_{1:T}$ relating to a single sentence label ω:

$$\begin{aligned}\omega &= \texttt{the dog chased the cat}\\ \boldsymbol{O}_{1:342} &= \{\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3, \boldsymbol{o}_4, \ldots, \boldsymbol{o}_{339}, \boldsymbol{o}_{340}, \boldsymbol{o}_{341}, \boldsymbol{o}_{342}\}\end{aligned}$$

The sentence `the dog chased the cat` has been uttered, taking 3.42 seconds, resulting in the observation sequence $\boldsymbol{O}_{1:342}$ (assuming a frame rate of 10ms).

Inferring the most likely sentence $\hat{\omega}$ uses Bayes' decision rule:

$$\hat{\omega} = \underset{\omega}{\operatorname{argmax}} \{P(\omega|\boldsymbol{O}_{1:T}; \boldsymbol{\lambda})\} \quad (1)$$

where statistical model parameters are indicated by $\boldsymbol{\lambda}$. In this classification, the number of output labels (the sentence identity), is not related to the number of observations T. Statistical models that can handle this form of data will be referred to as *sequence models*. As with standard *static* classifiers, these are often split into two broad classes: generative and discriminative models.

A. Generative Models

For many years HMM generative models [1], [3] have dominated speech recognition. This is partly due to their ability to handle sequence data, combined with elegant training and inference algorithms; they also yield good performance in a range of domains.

Generative classifiers for ASR can be split into two parts: a language model, the prior, $P(\omega)$ that yields a probability of any sentence; and an acoustic model $p(\boldsymbol{O}_{1:T}; \boldsymbol{\lambda}^{(\omega)})$ — the likelihood that a sentence ω generated the observations $\boldsymbol{O}_{1:T}$ with model parameters $\boldsymbol{\lambda}^{(\omega)}$. Classification is based on the sentence posterior obtained using Bayes' rule

$$P(\omega|\boldsymbol{O}_{1:T}; \boldsymbol{\lambda}) = \frac{P(\omega)p(\boldsymbol{O}_{1:T}; \boldsymbol{\lambda}^{(\omega)})}{\sum_{\tilde{\omega}} P(\tilde{\omega})p(\boldsymbol{O}_{1:T}; \boldsymbol{\lambda}^{(\tilde{\omega})})} \quad (2)$$

The HMM acoustic model is defined by its topology and its conditional independence assumptions.

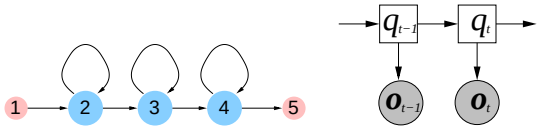

Fig. 1. Example HMM: three emitting states left-to-right topology (left), and DBN (right). Note for the DBN the dependence of the state on the sentence has not been shown.

Figure 1 shows the topology and Dynamic Bayesian Network (DBN) [22] associated with a typical HMM. The left diagram illustrates a standard phone topology, strictly left-to-right with three emitting states, the right diagram the DBN with conditional independence assumptions: the state at time t, q_t, is conditionally independent given the state identity at time $t-1$, q_{t-1}; and the observation at time t is conditionally independent given the state at time t.

For an HMM, the likelihood is found by marginalising over all valid state sequences, $\mathbf{q} = \{q_1, \ldots, q_T\}$. Thus

$$p(\boldsymbol{O}_{1:T}; \boldsymbol{\lambda}^{(\omega)}) = \sum_{\boldsymbol{q}:|\boldsymbol{q}|=T} \prod_{t=1}^{T} P(q_t|q_{t-1}) p(\boldsymbol{o}_t|q_t; \boldsymbol{\lambda}^{(\omega)}) \quad (3)$$

where the model parameters for a particular word sequence $\boldsymbol{\lambda}^{(\omega)}$ defines the set of valid state sequences. Inference with these forms of model can be efficiently achieved using the Viterbi algorithm [23], where the likelihood is approximated using the best-state sequence.

In most state-of-the-art ASR systems, the parameters of the distributions in Equation 3 are trained using discriminative criteria [24]–[26] (see section III-C) rather than maximizing the likelihood of the observations [1]. An alternative approach, discussed next, is to change the *model* to directly discriminate between sentences.

B. Discriminative Models

Discriminative models directly model the sentence (class) posterior given the observation sequence [27]. One of fairly broad-class is the maximum entropy, *max-ent* model [28], also known as a log-linear model. Here

$$P(\omega|\boldsymbol{O}_{1:T};\boldsymbol{\alpha}) = \frac{1}{Z}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{O}_{1:T},\omega)\right) \quad (4)$$

where Z is the normalisation term to ensure a valid probability mass function over all sentences, and $\boldsymbol{\alpha}$ the discriminative model parameters. This form of model can be related to SVMs [29] and the perceptron classifier depending on the form of the training criterion. This relationship will be discussed in more detail in sections III-C and V.

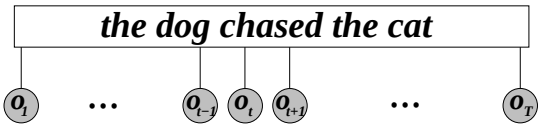

Fig. 2. Graphical model for a simple discriminative model

The form of feature-function $\boldsymbol{\phi}(\boldsymbol{O}_{1:T},\omega)$ is central to the performance of this model. A fundamental requirement of the feature function is that it transforms the variable length observation sequence into a fixed length feature vector as multiple different length segments may be used by the same feature function [2]. The undirected graphical model for the simplest feature-function form is shown in Figure 2. Feature-functions define the relationship between the sequence of observations and the sentence label; for example, a feature-function of the form

$$\boldsymbol{\phi}(\boldsymbol{O}_{1:T},\omega) = \begin{bmatrix} \vdots \\ \delta(\omega, \texttt{the dog chased the cat}) \\ \delta(\omega, \texttt{the dog chased the cat})\sum_{t=1}^{T}\boldsymbol{o}_t \\ \vdots \end{bmatrix} \quad (5)$$

where $\delta()$ is the Kronecker delta-function, provides a simple, first-order relationship between $\boldsymbol{O}$ and ω. More powerful feature-functions are discussed in Section IV.

C. Structured Sequence Data

The sequence models described above have been based on whole sentence models, where the generative model parameters $\boldsymbol{\lambda}^{(\omega)}$, the prior $P(\omega)$, and the discriminative model feature-function, $\boldsymbol{\phi}(\boldsymbol{O}_{1:T},\omega)$, are based on the sentence label ω. For some tasks, predicting the whole sentence ω is reasonable [31], but as the vocabulary size and number of possible sentences increases, this approach becomes impractical. To address this issue *structure* [3] can be introduced into the statistical model, where the sentence hypothesis is broken into a sequences of units such as words or phones. An example decomposition of ω into an L-length word sequence, $\boldsymbol{w}_{1:L}$, or K-length phone-sequence, $\boldsymbol{p}_{1:K}$, where typically $L \neq K \neq T$ is:

$$\begin{aligned} \omega &= \texttt{the dog chased the cat} \\ \boldsymbol{w}_{1:5} &= \{\texttt{the},\texttt{dog},\texttt{chased},\texttt{the},\texttt{cat}\} \\ \boldsymbol{p}_{1:16} &= \{\texttt{/sil/},\texttt{/dh/},\texttt{/ax/},\texttt{/d/},\ldots,\texttt{/t/},\texttt{/sil/}\} \end{aligned}$$

The standard CRF formulation [5] is problematic for representing the match between the label and observation sequence; a model of the word labels, $\boldsymbol{w}_{1:L}$, yields

$$P(\boldsymbol{w}_{1:L}|\boldsymbol{O}_{1:T};\boldsymbol{\alpha}) = \frac{1}{Z}\prod_{\tau=1}^{L}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{O}_{1:T},w_\tau,\tau)\right) \quad (6)$$

The problem arises with the feature-functions $\boldsymbol{\phi}(\boldsymbol{O}_{1:T},w_\tau,\tau)$, as the number of labels, L, and observations, T, are not constrained to be equal so there is no one-to-one matching between observations and labels as in [6].

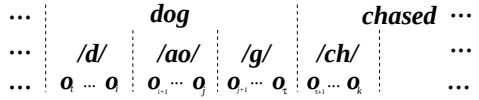

Fig. 3. Word, phone and observation hierarchy with associated possible segmentation of the observations at the phone and word levels.

Rather than considering the complete observation sequence, the observations can be segmented into subsequences each of which are associated with an individual label. This hierarchy of word and phone with a possible observation segmentation is shown in Figure 3. Since this segmentation is not observed, it must be inferred or marginalised over in the final model. Given the structuring of a sentence into labels and associated observation sub-sequences, one can *tie* model parameters

[2]In this presentation, in common with work on CRFs [5] and SSVMs [30] joint feature-spaces involving both features and labels will be used. Even when structure is introduced this requirement to handle variable length data is still necessary.

[3]For some machine learning tasks structured data and sequence data are used inter-changeably. In ASR there are two distinct sequences, the words in a sentence and the observations; the term structured here will be linked with the statistical model. Models like the Flat Direct Model [31], for example, have structured observations, but utilize an unstructured maximum entropy classifier as the statistical model.

together and form the complete sentence "model" by combining multiple sub-sentence labels together.

This structuring of the labels and observations is the standard approach for generative models for ASR. Thus the HMM likelihood and prior (both based on words and phones) can be expressed as

$$p(\boldsymbol{O}_{1:T}|\boldsymbol{w}_{1:L};\boldsymbol{\lambda}) = \sum_{\boldsymbol{a}} P(\boldsymbol{a}^{\mathtt{i}}|\boldsymbol{w}_{1:L}) \prod_{\tau=1}^{|\boldsymbol{a}|} p(\boldsymbol{O}_{\{a_\tau\}};\boldsymbol{\lambda}^{(a_\tau^{\mathtt{i}})}) \tag{7}$$

where $\boldsymbol{a}$ is a set of *segmentations* of the observations, a_τ is the τth segment in the sequence. Each segment specifies a phone/word/sub-unit identity indicated for segment τ as $a_\tau^{\mathtt{i}}$, and range of frames, $\boldsymbol{O}_{\{a_\tau\}}$. The same notation can be used for phone, HMM state, and Weighted Finite State Transducer (WFST) [32] arc sequences. $\boldsymbol{a}^{\mathtt{i}} = \{a_1^{\mathtt{i}}, \ldots, a_{|\boldsymbol{a}|}^{\mathtt{i}}\}$ is the sequence of segment identities. Thus $P(\boldsymbol{a}^{\mathtt{i}}|\boldsymbol{w}_{1:L})$ is the pronunciation probability when the segmentation is associated with phones. An N-gram language model is often used with generative models, for example

$$P(\omega) = P(\boldsymbol{w}_{1:L}) = P(w_1) \prod_{\tau=2}^{L} P(w_\tau|w_{\tau-1}) \tag{8}$$

using a simple bigram language model.

The next section expands upon this idea by examining different forms of structure incorporated into discriminative models.

III. Structured Discriminative Models

Structured discriminative models aim to make use of the same sub-sentence units as the acoustic model (7) and language model (8) of the generative classifier. This section describes some forms for these models, possible approaches to handling latent segmentations, and training criteria. The features (and models) described will focus on the observation sequence. For ASR it is also necessary to have pronunciation-style, $\boldsymbol{\phi}(\boldsymbol{a}^{\mathtt{i}}, \boldsymbol{w})$, and word, $\boldsymbol{\phi}(\boldsymbol{w})$, features. These are discussed in more detail in section IV.

A. Model Structures

The simplest form of structured discriminative model is to make use of graphical models that are closely linked to the DBN of the HMM, Figure 1. The discrete state latent variables can either be introduced in a directed, or undirected, graph. This is the basis of the Maximum Entropy Markov Models (MEMMs) [19] and Hidden Conditional Random Fields [9].

The graphical models associated with MEMMs and HCRFs are shown in Figure 4. For the MEMM, the arrow

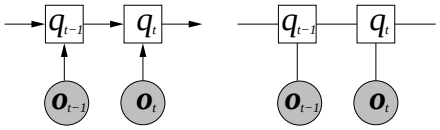

Fig. 4. Graphical models for MEMM (left) and HCRF (right) features. Note the dependence of the state on the word has not been shown.

relating the observations and state is simply reversed compared to the HMM, yielding the discriminative observation state relationship $P(q_t|\boldsymbol{o}_t, q_{t-1};\boldsymbol{\alpha})$ with model parameters, $\boldsymbol{\alpha}$. HCRFs, closely related to other forms of structured ASR models, have a posterior defined as

$$P(\boldsymbol{w}_{1:L}|\boldsymbol{O}_{1:T};\boldsymbol{\alpha}) = \frac{1}{Z}\sum_{\boldsymbol{a}}\sum_{\mathbf{q}\in\mathbf{Q}_{\boldsymbol{a}}} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\left[\begin{array}{c}\sum_{\tau=1}^{|\boldsymbol{a}|}\sum_{t\in\{a_\tau\}}\boldsymbol{\phi}(\boldsymbol{o}_t, q_t, a_\tau^{\mathtt{i}}) \\ \sum_{\tau=1}^{|\boldsymbol{a}|}\sum_{t\in\{a_\tau\}}\boldsymbol{\phi}(q_t, q_{t-1}, a_\tau^{\mathtt{i}})\end{array}\right]\right) \tag{9}$$

where $\mathbf{Q}_{\boldsymbol{a}}$ is the set of all state sequences where $|\mathbf{q}| = T$ and satisfies the segmentation defined by $\boldsymbol{a}$ [33]. If the segmentation of the data is at the word-level then $a_\tau^{\mathtt{i}} = w_\tau$. As there are latent variables (states) in an HCRF it is possible to associate these states with particular words in the feature-functions. These words are then combined together to yield the complete sentence as in an HMM [10]. This is also the underlying form of augmented CRFs [20], where frame-level augmented observations are combined to predict a sentence.

The form presented in (9) implies that the features only depend on the observation and state at time t. It is possible to generalise this to a fixed span of frames and observations - a dynamic undirected graph [6].

This type of model allows the structure to be imposed on the feature-function. However, in (9) the feature-function generates a vector for each frame: while this function can act on a fixed window of observations, or states, it will still generate T vectors for a sequence of T observations. For a particular form of feature function, see (24), HCRFs can be shown be equivalent to discriminative training of HMMs [34]. Segmental feature-functions in models such as Conditional Augmented Models (CAugs) [18], and Segmental CRF (SCRFs) [17], [35] can allow observations across a segment to contribute to the function (similar to generative segmental HMMs [36]; the feature functions relate to the segmentation of the observations $O_{\{a_\tau\}}$:

$$P(\boldsymbol{w}_{1:L}|\boldsymbol{O}_{1:T};\boldsymbol{\alpha}) = \frac{1}{Z}\sum_{\boldsymbol{a}} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\left[\begin{array}{c}\sum_{\tau=1}^{|\boldsymbol{a}|}\boldsymbol{\phi}(\boldsymbol{O}_{\{a_\tau\}}, a_\tau^{\mathtt{i}})\end{array}\right]\right) \tag{10}$$

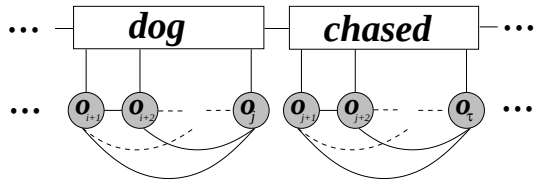

Fig. 5. CAug/SCRF graphical model for a particular segmentation

Figure 5 shows the corresponding graphical model. The feature-function is able to describe all the observations for that segment $\boldsymbol{O}_{\{a_\tau\}}$, requiring the function to convert a variable length set of features to a fixed length feature-vector. (See more detail in section IV.)

B. Handling Latent Variables

The previous section has considered summing over all possible segmentations, $\boldsymbol{a}$, of the data. Though it is possible to define recursions for this task [17], the resulting parameter estimation is no longer convex [37], and the decoding and training time can become slow depending on the exact nature of the feature-extraction process. Also as the optimisation approaches used to train discriminative model parameters are iterative, refining the segmentation every iteration may become impractical.

An alternative approach is to perform the equivalent of Viterbi training and decoding [23]. Using a single segmentation yields a concave maximisation problem [4]. Furthermore it is possible to make use of standard optimisation approaches associated with the perceptron criterion and structured SVMs discussed in section V. For some structured discriminative models, such as the structured SVM [39], this approximation is essential.

With a single segmentation the following posterior is obtained

$$P(\boldsymbol{w}_{1:L}|\boldsymbol{O}_{1:T},\hat{\boldsymbol{a}};\boldsymbol{\alpha}) = \frac{1}{Z}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\left[\begin{array}{c}\sum_{\tau=1}^{|\hat{\boldsymbol{a}}|}\boldsymbol{\phi}(\boldsymbol{O}_{\{\hat{a}_\tau\}},\hat{a}_\tau^{\mathtt{i}})\end{array}\right]\right) \quad (11)$$

The issue now is how the segmentation $\hat{\boldsymbol{a}}$ is obtained. The simplest approach is to use the segmentation derived from a generative model, for example an HMM. This yields efficient training and inference irrespective of the nature of the features. However the optimal Viterbi segmentation for the discriminative model may differ to that of the generative model. The optimal segmentation for the discriminative model is given by

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmax}}\left\{P(\boldsymbol{a}|\boldsymbol{O}_{1:T})P(\boldsymbol{w}_{1:L}|\boldsymbol{O}_{1:T},\boldsymbol{a};\boldsymbol{\alpha})\right\} \quad (12)$$

Since this "best" segmentation is a function of the model parameters the process must be iterated, interleaving the refinement of the segmentation and model parameters during training.

More generally, the segmentation simply enables the feature-function to be clearly associated with units (words) of the sentence. It is not necessary to use a segmentation if this process can be achieved in an alternative fashion. One example of this is based on generative score-spaces [8], discussed in section IV-B. For a derivative score-space, it is possible to write for the features of word w_i of sentence $\omega = \boldsymbol{w}_{1:L}$

$$\boldsymbol{\phi}(\boldsymbol{O}_{1:T},w_i;\boldsymbol{\lambda}^{(\omega)}) = \left[\begin{array}{c}\vdots\\ \delta(w_i,v_j)\boldsymbol{\nabla}_{\boldsymbol{\lambda}^{(w_i)}}\log\left(p(\boldsymbol{O}_{1:T};\boldsymbol{\lambda}^{(\omega)})\right)\\ \vdots\end{array}\right] \quad (13)$$

where v_j is an element in the vocabulary $\mathcal{V}$. There is now no segmentation, $\boldsymbol{a}$, of the observation sequence. Though an interesting theoretical direction, to the authors knowledge this form of approach has not been used. Making the feature extraction stage a function of the whole observation sequence means that decoding can rapidly become impractical, requiring N-best lists.

C. Optimization Criteria

In the same fashion as standard, non-structured, discriminative models and generative models it is possible to use a range of discriminative criteria. One of the most popular is conditional maximum likelihood training [24]. This aims to maximise the probability of the correct sentence label. Though directly linked with the Bayes' decision rule, it may not always be optimal for all tasks. First, the scoring of speech data is not usually at the sentence level, more commonly word error rate (WER) is used. Thus more general minimum Bayes' risk [40] training may be better. Second given the large number of model parameters that are often trained with these forms of system, approaches for improving generalisation performance may be very useful. These same considerations have led to the use of a number of criteria for generative models for ASR [24]–[26]. Similar forms of criteria can also be used for structured discriminative models [41].

Some of the more standard criteria are briefly discussed below. Here only supervised training data is considered where the training data, $\mathcal{D}$, comprises (sequence

[4]It can be argued that once the segmentation has been obtained it can be converted into a frame-label sequence that could then be used for CRF training. This is the form examined in the Semi-Markov CRF [38].

length has been dropped for notational simplicity)

$$\mathcal{D} = \left\{ \left\{ \boldsymbol{O}^{(1)}, \boldsymbol{w}^{(1)} \right\}, \ldots, \left\{ \boldsymbol{O}^{(R)}, \boldsymbol{w}^{(R)} \right\} \right\}$$

Discussion about the form of the actual optimisation process or regularisation is deferred to section V.

All the criteria have the same general forms $\mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{O})$, and can be trained using either batch or on-line algorithms. For batch training

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left\{ \frac{1}{R} \sum_{r=1}^{R} \mathcal{F}\left(\boldsymbol{\alpha}, \boldsymbol{w}^{(r)}, \mathbf{O}^{(r)}\right) \right\} \tag{14}$$

and for the mini-batch, on-line, equivalent

$$\hat{\boldsymbol{\alpha}}^{(r)} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left\{ \mathcal{F}\left(\boldsymbol{\alpha}, \boldsymbol{w}^{(r)}, \mathbf{O}^{(r)}; \hat{\boldsymbol{\alpha}}^{(r-1)}\right) \right\} \tag{15}$$

Conditional Maximum Likelihood [24]:

$$\mathcal{F}_{\texttt{cml}}(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{O}) = -\log\left(P\left(\boldsymbol{w}|\mathbf{O}; \boldsymbol{\alpha}\right)\right) \tag{16}$$

This is the form typically used for training discriminative models such as CRFs [5] and is usually the starting point for structured discriminative models [17].

Perceptron Algorithm: [42]

$$\mathcal{F}_{\texttt{per}}(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{O}) = \left[\max_{\tilde{\boldsymbol{w}} \neq \boldsymbol{w}} \left\{ -\log\left(\frac{P(\boldsymbol{w}|\mathbf{O}; \boldsymbol{\alpha})}{P(\tilde{\boldsymbol{w}}|\mathbf{O}; \boldsymbol{\alpha})} \right) \right\} \right]_{+} \tag{17}$$

where $[x]+$ is the hinge-loss function. This can be extended to the averaged perceptron algorithm where the parameters, $\boldsymbol{\alpha}$ are averaged [42].

Minimum Bayes Risk [40]:

$$\mathcal{F}_{\texttt{mbr}}(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{O}) = \sum_{\tilde{\boldsymbol{w}}} P\left(\tilde{\boldsymbol{w}}|\mathbf{O}; \boldsymbol{\alpha}\right) \mathcal{L}\left(\tilde{\boldsymbol{w}}, \boldsymbol{w}\right) \tag{18}$$

where $\mathcal{L}\left(\tilde{\boldsymbol{w}}, \boldsymbol{w}\right)$ is the loss between the word sequence $\tilde{\boldsymbol{w}}$ and the reference $\boldsymbol{w}$. The loss may be measured at the word, phone, or frame level.

Maximum Margin [26]:

$$\mathcal{F}_{\texttt{lm}}(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{O}) = \left[\max_{\tilde{\boldsymbol{w}} \neq \boldsymbol{w}} \left\{ \mathcal{L}\left(\tilde{\boldsymbol{w}}, \boldsymbol{w}\right) - \log\left(\frac{P(\boldsymbol{w}|\mathbf{O}; \boldsymbol{\alpha})}{P(\tilde{\boldsymbol{w}}|\mathbf{O}; \boldsymbol{\alpha})} \right) \right\} \right]_{+} \tag{19}$$

The margin here is the loss between reference and "closest" competing word sequences. This loss may be at the frame level or at a higher level, e.g. word or phone.

One issue that can occur is that the normalisation term can be very expensive, or even intractable, to compute [43]. However for some criteria, the perceptron (17) and maximum margin (19) criteria, it is not necessary to ever compute this term. During training the criterion is a function of the ratio of posteriors (the normalisation term cancels) and the rank ordering for inference is not altered by the normalisation term.

Directly using the above expressions can also cause generalisation issues as the feature-function can result in a very high-dimensional feature-space. To address this, regularisation terms, normally in the form of L1 or L2 regularisation, are introduced [17], [20], [44]. When combined with maximum margin training these regularisation terms result in discriminative models closely related to structured SVMs [14]. Furthermore for some feature-functions one can introduce a more informative prior on the discriminative model parameters by using non zero mean priors for $\boldsymbol{\alpha}$ [16].

D. Adaptation

For generative models adaptation to a particular speaker or environment condition is an essential part of current speech recognition systems [3]. A range of approaches have been developed including: maximum a-posteriori (MAP) adaptation; linear transformation-based approaches; model-based noise compensation; and feature enhancement. For details and references see [3]. Related approaches have been developed for discriminative models [5]. These can be split into three broad categories: general adaptation; linear transformation approaches; and feature adaptation. Note in contrast to the majority of adaptation approaches for generative models which are based on maximum likelihood, discriminative model adaptation is usually based on conditional maximum likelihood.

In [45], two approaches for adapting log-linear models — MAP adaptation and minimum divergence training — are discussed. These approaches yield a general adaptation scheme that makes no assumption about the nature of the features in the model. MAP adaptation has also been applied to HCRFs [33]. Though these general adaptation approaches can be used for discriminative models, they do not take advantage of any structure in the features. Alternatively Linear transformation based approaches for log-linear models are described in [46], [47]. These schemes use approaches similar to the linear transformations for HMMs. Assumptions are made about the relationships between features. To date they have only been applied to models where the features are very similar to those used in standard HMMs. Whether these

[5]In the machine learning literature the problem of handling a mismatch between training and test conditions is sometimes referred to as sample selection bias or covariat shift.

form of approaches can be extended to more general features is an open question.

The final form of adaptation is related to the feature compensation schemes used with generative models. Rather than adapting the model parameters, the features are modified to make them independent of the speaker or environment. This is simplest to do when the feature extraction process is based on generative models [15], [48]. This approach is discussed in more detail in section V-C.

E. Kernel Representations

The discussion of the model-parameters and feature-functions have so far assumed that there is an explicit representation of each of these. It is also possible to consider a more general form that can be highly efficient in dealing with large feature-spaces. Since the model uses an inner-product between model-parameters and features, it is possible to kernelize this operation in the same way as SVMs [29]. This allows the so-called "kernel-trick" to be used where it is not necessary to explicitly operate in the full-feature space. A non-linear kernel function can be applied in the original features-space to yield the results of the inner-product in the full-features space. Here the term in the exponential becomes

$$\begin{aligned}&\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{O}_{1:T},\boldsymbol{w}_{1:L}) \qquad (20)\\ &\quad=\sum_{r=1}^{R}\tilde{\alpha}_r k\left(\{\boldsymbol{O}^{(r)},\boldsymbol{w}^{(r)}\},\{\boldsymbol{O}_{1:T},\boldsymbol{w}_{1:L}\}\right)\end{aligned}$$

where $\tilde{\alpha}_r$ is the equivalent of the Lagrange multiplier for each training utterance in an SVM, and $k(.,.)$ is the kernel. Depending on the nature of the criterion, and the form of regularisation being used, only a small subset of the Lagrange multipliers $\tilde{\alpha}_r$ may be non-zero. For example if the parameters of the discriminative model are trained using the maximum margin criterion with an $L2$ regularisation term, the Lagrange multipliers $\tilde{\alpha}_r$ should be sparse as this form is related to SVM training [14].

As the length of the observation and word sequences vary over the training and test samples, a sequence kernel is required, a range of which can be described in the rational kernel (both discrete and continuous) framework [12], [49]. More generally when sequence kernels are combined with feature-functions and static kernels the following form can be obtained (assuming segmentation $\boldsymbol{a}^{(r)}$ and $\boldsymbol{a}$ at the word level, $a_i^{\mathtt{i}}=w_i$)

$$\begin{aligned}&k\left(\{\boldsymbol{O}^{(r)},\boldsymbol{w}^{(r)}\},\{\boldsymbol{O}_{1:T},\boldsymbol{w}_{1:L}\}\right)=\sum_{i=1}^{|\boldsymbol{a}^{(r)}|}\sum_{j=1}^{|\boldsymbol{a}|} \qquad (21)\\ &\quad\delta(w_i^{(r)},w_j)k_{\mathtt{st}}\left(\boldsymbol{\phi}\left(\boldsymbol{O}_{\{a_i^{(r)}\}},w_i^{(r)}\right),\boldsymbol{\phi}\left(\boldsymbol{O}_{\{a_j\}},w_j\right)\right)\end{aligned}$$

where $\boldsymbol{\phi}()$ is the score-space associated with the sequence kernel and $k_{\mathtt{st}}(.,.)$ is the static kernel. Here the score-space is the feature-space associated with the sequence kernel. This form of kernel combination has previously been discussed for speaker verification [50].

IV. Model Features

The previous section has assumed the existence of an appropriate feature-function: the selection of this function is central to the performance of these classifiers. Features can be broadly split into observation-features, pronunciation features and word-features

$$\boldsymbol{\phi}(\boldsymbol{O}_{1:T},\boldsymbol{w}_{1:L},\boldsymbol{a})=\begin{bmatrix}\sum_{i=1}^{|\boldsymbol{a}|}\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}},a_i^{\mathtt{i}})\\ \boldsymbol{\phi}(\boldsymbol{a}^{\mathtt{i}},\boldsymbol{w}_{1:L})\\ \boldsymbol{\phi}(\boldsymbol{w}_{1:L})\end{bmatrix} \qquad (22)$$

When the segmentation $\boldsymbol{a}$ is at the word level, the term $\boldsymbol{\phi}(\boldsymbol{a}^{\mathtt{i}},\boldsymbol{w}_{1:L})$ which relates to pronunciation probabilities and variants (i.e. the mapping from segments to words) can be ignored. In this section a word level segmentation is assumed unless otherwise stated. Similar forms of discrete features can be applied for the pronunciation features as the word features described below.

A. Frame-Level Features

The simplest form of feature function is restricted to frame-level features in the same fashion as the HCRF features. The general form of features can be written as

$$\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}},a_i^{\mathtt{i}})=\sum_{t\in\{a_i\}}\boldsymbol{\phi}(\boldsymbol{o}_t,a_i^{\mathtt{i}}) \qquad (23)$$

One of the simplest form of feature-function directly uses the Gaussian sufficient statistics of observations:

$$\boldsymbol{\phi}(\boldsymbol{o}_t,a_i^{\mathtt{i}})=\begin{bmatrix}\vdots\\ \delta(a_i^{\mathtt{i}},v_j)\\ \delta(a_i^{\mathtt{i}},v_j)\boldsymbol{o}_t\\ \delta(a_i^{\mathtt{i}},v_j)\mathrm{diag}(\boldsymbol{o}_t\boldsymbol{o}_t^{\mathsf{T}})\\ \vdots\end{bmatrix}\forall v_j\in\mathcal{V} \qquad (24)$$

where $\mathcal{V}$ is the vocabulary of segment identities. Using these features yields systems related to discriminatively trained HMMs [34], but it can be extended to introduce features of higher-order statistics [13].

An interesting question is what form the observation, $\boldsymbol{o}_t$, takes. Rather than just considering a single frame, frames can be spliced together and optionally transformed (as is adopted in augmented CRFs [20]), much as generative systems use delta and delta-delta parameters (and other generalisations e.g., kernel application).

A slightly different approach is to use classifiers to provide information about the discrimination between

sub-word classes. This can provide bottom-up information to the system on where observations lie in a pseudo-linguistic space. Consider the case of linguistic units $\mathcal{V} = \{v_1, \ldots, v_{|\mathcal{V}|}\}$, which are derived from applying a discriminant function, $g()$, to a span of observations. Thus for frames a to b, $g(\boldsymbol{O}_{a:b}) = \boldsymbol{u}_{a:b}, u_t \in \mathcal{V}$. Discrete features of the form $\delta(u_t, v_i)$ or class posterior features can then used for frame t. Examples of discriminant functions include: multilayer perceptron predictions of the posterior probability of phone units given a fixed span of observations [7]; sparse representations arising from finding the n-closest Gaussians to a single observation from a trained system [20]; and an HMM-based recogniser on the complete observation sequence [35].

B. Acoustic Segment Features

Rather than the feature-function generating a vector for each frame, it can also act on all the observations associated with a segment $\boldsymbol{O}_{\{a_\tau\}}$. Thus, the frame-level features are just one option for extracting features from this segment. It is possible to hypothesise a range of features that could be used. However it is more interesting to consider this process in the context of sequence kernels and score-spaces [18]. These sequence kernels map variable length sequences to a fixed length score-space in which the inner product can be computed. All the acoustic feature extraction schemes for feature extraction satisfy this property. The advantage of discussing acoustic features in this framework is that existing developments from machine learning can be used.

One general form of these sequence kernels, the rational kernel [49], is

$$k\left(\boldsymbol{O}_{\{a_i\}}, \boldsymbol{O}_{\{a_j\}}\right) = \mathcal{C}\left[\mathcal{A}_{\{a_i\}} \circ \left(\mathcal{U} \circ \mathcal{U}^{-1}\right) \circ \mathcal{A}_{\{a_j\}}\right] \quad (25)$$

$$= \boldsymbol{\phi}\left(\boldsymbol{O}_{\{a_i\}}\right)^{\mathsf{T}} \boldsymbol{\phi}\left(\boldsymbol{O}_{\{a_j\}}\right) \quad (26)$$

where $\mathcal{A}_{\{a_i\}}$ and $\mathcal{A}_{\{a_j\}}$ are the acceptors associated with the observation sequences $\boldsymbol{O}_{\{a_i\}}$ and $\boldsymbol{O}_{\{a_j\}}$ respectively, $\mathcal{U}$ is the WFST that determines the form of the rational kernel, $\circ$ is WFST composition, $\mathcal{C}[]$ yields the transducer shortest distance, and $\boldsymbol{\phi}()$ the score-space associated with this kernel [6]. This representation allows operations on sequences of different lengths, i.e there are no constraints that $|\{a_i\}| = |\{a_j\}|$. This form of kernel has been used for both discrete observations [49] and continuous observations [12]. It is able to efficiently represent a range of standard kernels such as string kernels and (gappy) N-gram kernels [11].

If the kernel representation in the previous section is used then rational kernels can be directly applied in structured discriminative models. If the more standard form is used then score-spaces from the kernel can be used as the basis for the feature-function.

One interesting form of score-space for this form of kernel is based on generative models [18], [51] [7]

$$\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}}, a_i^{\mathtt{i}}) = \begin{bmatrix} \log\left(p(\boldsymbol{O}_{\{a_i\}}; \boldsymbol{\lambda}^{(a_i^{\mathtt{i}})})\right) \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}^{(a_i^{\mathtt{i}})}} \log\left(p(\boldsymbol{O}_{\{a_i\}}; \boldsymbol{\lambda}^{(a_i^{\mathtt{i}})})\right) \\ \vdots \\ \boldsymbol{\nabla}^{\rho}_{\boldsymbol{\lambda}^{(a_i^{\mathtt{i}})}} \log\left(p(\boldsymbol{O}_{\{a_i\}}; \boldsymbol{\lambda}^{(a_i^{\mathtt{i}})})\right) \end{bmatrix} \quad (28)$$

where $\boldsymbol{\nabla}^{\rho}_{\boldsymbol{\lambda}}$ represents the (diagonalised) ρ-th order derivative with respect to $\boldsymbol{\lambda}$. If the generative model is an HMM then the resulting features do not have the same underlying conditional-independence assumptions of the HMM [18]. Alternatively if GMMs are used then derivative score-spaces yields frame-level features; the derivative with respect to the component priors, for example, yields sparse GMM posterior features [44]. An interesting aspect of using structured generative models in this fashion is that feature-extraction can be made efficient using an expectation semi-ring within the WFST framework [52].

Similar in spirit to the score-space paradigm are other methods that utilize detections of longer-term acoustic events. In [17], a baseline HMM system hypothesizes linguistic units, which are then evaluated by measuring how consistent the units are with the dictionary pronunciation of a hypothesized word. Another approach is to use template matching to suggest detections of linguistic units that may or may not be consistent with word hypotheses [35].

C. Supra-Segmental Features

The primary form of supra-segmental features are associated with the word (or phone) sequences. Applying log-linear models for language modelling has been an active research area for many years, for example see [43], [53]. These exponential models allow a very rich set of features, for example lexical [21], linguistic, and hierarchical features [54], [55], to be used. For the

[6]To map from this basis representation involves

$$\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}}, a_i^{\mathtt{i}}) = \begin{bmatrix} \delta(a_i^{\mathtt{i}}, v_1)\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}}) \\ \vdots \\ \delta(a_i^{\mathtt{i}}, v_V)\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}}) \end{bmatrix} \quad (27)$$

The trivial generalisation is to allow the features to be dependent on the word. This is closely related to generating the joint feature-space for SVMs [39].

[7]The form of score-space described here can also be related to information geometry and more general forms of generative model [18]. For discrete cases it has also been connected to string-kernels [11].

notation in this paper the segment identity can be used to specify the precise nature of the segment including any hierarchical information.

Similar to the observation features, the number of alignments, $|\boldsymbol{a}|$, and the number of words, $|\boldsymbol{w}|$, are not fixed for all sentences; the same issues as discussed for segment-level features must be addressed. As these supra-segmental features are often discrete, rational kernels and associated score-spaces can be used. Considering word-level features this yields

$$k(\boldsymbol{w}^{(i)}, \boldsymbol{w}^{(j)}) = \boldsymbol{\phi}(\boldsymbol{w}^{(i)})^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{w}^{(j)}) \tag{29}$$

with no constraint that $|\boldsymbol{w}^{(i)}| = |\boldsymbol{w}^{(j)}|$. If the kernel representation discussed in the previous section is used then it is possible to directly use the kernel output, rather than requiring the explicit calculation of the score-space.

One common form of feature-space is based on unigram and higher-order discrete features. Thus one simple form is based on the bag-of-word model [4] (unigram) and higher-order N-grams. For bigram features

$$\boldsymbol{\phi}(\boldsymbol{w}_{1:L}) = \begin{bmatrix} \vdots \\ \sum_{\tau=1}^{L} \delta\left(w_\tau, \texttt{dog}\right) \\ \sum_{\tau=1}^{L-1} \delta\left(w_\tau, \texttt{dog}\right) \delta\left(w_{\tau+1}, \texttt{chased}\right) \\ \vdots \end{bmatrix} \tag{30}$$

It is possible to apply the same concepts to the segmentation and word features which will represent, amongst other things, pronunciation probability. In addition, features can be derived from traversing back-off arcs [17], as well as word-labeled arcs, in the WFST framework.

One of the interesting aspects of supra-segmental features is that they can be easily combined with generative models for classification. The classification of an observation sequence, $\boldsymbol{O}_{1:T}$, with an HMM is based on

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \left\{ \sum_{\boldsymbol{a}} P(\boldsymbol{w}, \boldsymbol{a}^{\mathtt{i}}) \prod_{i=1}^{|\boldsymbol{a}|} p(\boldsymbol{O}_{\{a_i\}}; \boldsymbol{\lambda}^{(a_i^{\mathtt{i}})}) \right\} \tag{31}$$

Rather than using the standard N-gram language model and pronunciation probabilities it is possible to write [43]

$$P(\boldsymbol{w}, \boldsymbol{a}^{\mathtt{i}}; \boldsymbol{\alpha}) = \frac{1}{Z} \exp\left(\boldsymbol{\alpha}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{a}^{\mathtt{i}}, \boldsymbol{w}) \\ \boldsymbol{\phi}(\boldsymbol{w}) \end{bmatrix} \right) \tag{32}$$

Here the feature-function for the observations is a single element, the log-likelihood from the HMM. The N-gram language model log-probability can also be added e.g. [56]. The model parameters for these two elements are sometimes fixed and not updated. This is the basis of discriminative language models in section V-A.

A summary of the features described here can be found in Table I.

Feature type	Example Representation	Example papers
Gaussian sufficient statistics	$\delta(a_i^{\mathtt{i}}, v_j)$ $\delta(a_i^{\mathtt{i}}, v_j)\boldsymbol{o}_t$ $\delta(a_i^{\mathtt{i}}, v_j)\text{diag}(\boldsymbol{o}_t\boldsymbol{o}_t^{\mathsf{T}})$	[9], [13], [34]
Local discriminant functions, e.g. MLP postieriors, closest Gaussians, or HMMs	$\delta(a^{\mathtt{i}}, v_j)P(\boldsymbol{v}\|\boldsymbol{o}_t)$	[7], [19], [20], [35]
Segment-level score spaces	$\delta(a_i^{\mathtt{i}}, v_1)\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}})$	[18], [44], [51], [52]
Segment-level model features	$\delta(a_i^{\mathtt{i}}, v_j)\boldsymbol{\phi}(\boldsymbol{v}, \boldsymbol{O}_{\{a_i\}})$	[17], [35]
Suprasegmental features, e.g. word-level features	$\sum_{\tau=1}^{L} \delta\left(w_\tau, \texttt{dog}\right)$	[17], [21], [43], [56]–[58]

TABLE I
SUMMARY OF FEATURE FUNCTIONS IN COMMON USE

V. EXAMPLE APPLICATIONS

A. Discriminative LMs and WFSTs

As discussed in section IV-C, it is possible to use structured discriminative modelling approaches to train a Discriminative Language Model (DLM) [21] which can then be combined, if desired, with a generative acoustic model for classification. One of the advantages of this form of model, compared to standard N-gram models [2], is that it is simple to combine highly diverse, possibly wide-span, features. For these richer models, ASR is often realized by re-ranking hypotheses rather than direct recognition, or lattice rescoring.

Two important aspects of DLM that make them practical for large training corpora and models are sparse feature representation and convex optimization. As DLMs typically use discrete features, e.g., long context N gram of word/Part-Of-Speech (POS) counts [21], [57], [58], the representation is usually sparse. Furthermore, as there are no latent variables associated with the DLM (or a single segmentation/latent variable value used), it is a convex optimisation problem.

One successful approach to training discriminative language models is to use the perceptron algorithm [21]. Here the parameters of the DLM are estimated using

$$\boldsymbol{\alpha}^{(r+1)} = \boldsymbol{\alpha}^{(r)} + \boldsymbol{\phi}(\boldsymbol{O}^{(r)}, \boldsymbol{w}^{(r)}) - \boldsymbol{\phi}(\boldsymbol{O}^{(r)}, \tilde{\boldsymbol{w}}) \tag{33}$$

where $\tilde{\boldsymbol{w}}$ is the hypothesis for utterance r with parameters $\boldsymbol{\alpha}^{(r)}$. To improve performance the average over all estimated model-parameters is used for classification, the averaged perceptron algorithm. Variants on this form are also possible [56]. An interesting aspect of this form of optimisation is that it is not necessary to compute the normalisation term (in common with the SSVM

section V-C). For classification the normalisation term is not required. It is only needed if a posterior probability is explicitly required from the system.

These forms of model can efficiently represented in a WFST, or lattice, framework. Indeed it is possible to describe the complete ASR process in this framework [32], [56]. WFSTs, or lattices, also provide a framework for efficient training and inference with structured discriminative models. The segmentation of data defined by $\boldsymbol{a}$ is at the arc-level in WFSTs or lattices. To reduce the range of possible segmentations and word-sequences, lattices for training and inference can be generated using a standard generative model [17], [18]. The lattice can be marked at the appropriate level of word, or sub-word to enable training and inference. An interesting aspect of this form of model is the segmentation of the observation sequence, $\boldsymbol{O}_{\{a_i\}}$, for arc a_i. If derived from the generative model it may not be optimal for the structured discriminative model. Thus this segmentation can be refined using the current discriminative model [16].

B. Segmental Conditional Random Fields

As a second application, a fuller description of an implementation of Segmental Conditional Random Fields (SCRFs) is given. SCRFs [17], and the closely related CAug models [18], focus on deriving sequences by marginalising all valid segmentations of the data, as discussed in Section III-A. Thus, the feature functions in this domain revolve around matching observations to segmental-level phenomena – that is, functions of the form $\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}}, a_i^{\mathtt{i}}$. For example, CAug models typically utilize feature-functions based on generative models and continuous features, as in the noise robust ASR work of [15]:

$$\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}}, a_i^{\mathtt{i}}; \boldsymbol{\lambda}) = \left[\begin{array}{c} \log\left(p(\boldsymbol{O}_{\{a_i\}}; \boldsymbol{\lambda}^{(a_i^{\mathtt{i}})})\right) \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}^{(a_i^{\mathtt{i}})}} \log\left(p(\boldsymbol{O}_{\{a_i\}}; \boldsymbol{\lambda}^{(a_i^{\mathtt{i}})})\right) \end{array} \right] \quad (34)$$

which are combined with word and pronunciation feature-functions.

Another method of employing generative models as features is to use a first pass HMM to generate detected events; for example, the SCRF work in [17] incorporates what they term as a *baseline* feature: does the hypothesized word w_i appear in the best HMM hypothesis? This allows the SCRF system to benefit from a generative baseline and (hopefully) correct errors made in that system. Other features that can be provided by a generative system include the existence of (N-grams of) subword units detected by an HMM, which can be associated directly with word hypotheses or evaluated with respect to their consistency with the word's pronunciation, as well as the word N-gram language model state, whose inclusion allows joint discriminative training of the language and acoustic models.

A richer set of functions for SCRFs was investigated in [35], which essentially break down into four classes of information: phoneme-based detection, word-based detection, template-based features, and durational scoring. In the first type of information, phoneme detectors are used as complementary information to the HMM-based subword-unit detections; these phonemes can be derived by several means (experiments in [35] ranged from dynamic time warping templates, to MLP or Deep Neural Network based hybrid ANN-HMM detectors). These types of detections can be used in place of, or in addition to, associational features between subword units and word hypotheses. [35] also investigated using point-process word detectors and maximum entropy word detectors trained using a novel demodulation feature; here the relationship between $\boldsymbol{O}_{\{a_i\}}$ and w_i is direct (assuming word segmentation $a_i^{\mathtt{i}} = w_i$): the feature fires if $\boldsymbol{O}_{\{a_i\}}$ is a valid representation of w_i. It is also possible to integrate features derived from exemplar-based systems. In [59] features derived from a k-NN template list is used to derived a range of features based on the a DTW match including common word positions and counts and average template duration (warping factor). Durational features serve as a confirmation of a hypothesis; for example, if $L = |\boldsymbol{O}_{\{a_i\}}|$, they included measures of $P_c(L|w_i)$, the probability of the observed length when w_i is a correct hypothesis, versus $P_i(L|w_i)$, the corresponding probability when the hypothesized word is incorrect.

For the studies in [17], [35], the SCRFs were trained to optimize a regularized CML criterion (16); in particular, both an L1 and L2 regularizer were included in the CML term which penalizes solutions with large weights, and will tend to prefer sparser solutions. In this particular implementation, the Rprop algorithm was used for gradient ascent in the regularized CML space; this allows for relatively fast convergence. Another implementation detail to note is that theoretically the system must investigate all possible segmentations of the data. To cut down on the number of possible segmentations, a fast-match approach is used to restrict the possible set of hypothesized segmentations. In the cited studies, the fast match was achieved by restricting possible segmentations to those found in the lattice produced by the generative baseline system.

C. Structured Support Vector Machines

The theory behind binary SVMs [29] and multi-class SVMs have been well established in the machine learning literature. More recently structured SVM

(SSVMs) [39] have been proposed to handle situations where there is structured data to classify. This section briefly describes the application of SSVMs to noise robust speech recognition [16].

The SSVM criterion can be expressed as minimising

$$\frac{1}{2}||\boldsymbol{\alpha}||_2^2 + \frac{C}{R}\Big[-\max_{\boldsymbol{a}}\Big\{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{O}^{(r)},\boldsymbol{w}^{(r)},\boldsymbol{a};\boldsymbol{\lambda})\Big\} + \max_{\boldsymbol{w}\neq\boldsymbol{w}^{(r)},\boldsymbol{a}}\Big\{\mathcal{L}(\boldsymbol{w},\boldsymbol{w}^{(r)}) + \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{O}^{(r)},\boldsymbol{w},\boldsymbol{a};\boldsymbol{\lambda})\Big\}\Big]_+ \quad (35)$$

This can be related to large-margin training of structured log-linear models with a Gaussian prior of the form $\mathcal{N}(\mathbf{0}, C\mathbf{I})$ and the best segmentation (12) [14].

One of the standard problems encountered in speech recognition is changing background noise environments. Rapidly modifying the parameters of a discriminative model to reflect these changes is challenging. As discussed in section III-D there are a number of approaches that can be applied. The approach that has been adopted with SSVMs is to generate noise-independent features by appropriately compensating generative models [48]. This allows an environment independent discriminative models to be used. The same approaches have also been used for discriminative log-linear models with first-order derivative features [15]. For the work on SSVMs a log-likelihood score-space was used as the basis for the feature-function

$$\boldsymbol{\phi}(\boldsymbol{O}_{\{a_i\}};\boldsymbol{\lambda}) = \begin{bmatrix} \log\left(p(\boldsymbol{O}_{\{a_i\}};\boldsymbol{\lambda}^{(v_1)})\right) \\ \vdots \\ \log\left(p(\boldsymbol{O}_{\{a_i\}};\boldsymbol{\lambda}^{(v_V)})\right) \end{bmatrix} \quad (36)$$

where $V = |\mathcal{V}|$. Two forms of task have been examined with different forms of acoustic model "vocabulary". Both tasks are from the AURORA framework: AURORA2 a continuous digit recognition task with whole word models; and AURORA4 a medium vocabulary continuous speech recognition task with phone-level models.

Having specified the nature of the features, the parameters of the SSVM must be trained. As discussed in section III-B to handle SSVMs it is necessary to only use the one-best alignment. Initially this can be obtained from the compensated HMMs used to derive the features. The parameters can then be found using the *cutting-plane algorithm* [60], which has been found to be an efficient method for training these forms of model. This has been used to train models for speech recognition in [14].

The initial segmentation from the compensated HMM will not be optimal, but it can be refined using (12). For the log-likelihood score-space this expression is related to inference for factorial HMMs [16]. This optimal segmentation can then be integrated into the overall training procedure using *concave-convex* optimisation [61].

Extending SSVMs to larger vocabulary tasks is non-trivial. The number of possible constraints to be satisfied can becomes very large, impacting both the computational load and memory requirements. This is the reason for selecting a log-likelihood based score-space for SSVMs, rather than the derivative forms that have been successfully applied to conditional augmented models for larger tasks [15]. Some of these issues are addressed in [16] where the following techniques are applied: 1-slack variable optimisation; score-space caching; and improved priors.

VI. Summary

This paper has presented a brief overview of structured discriminative models for speech recognition. For general ASR tasks, the model structure must handle both variable-length observations and word (or sub-sentence) sequences. Typical discriminative approaches used in natural language processing tasks do not need to account for segmentation, which can be introduced for ASR by means of latent variables. Segmentation of the sentence and observations sequence enable model-parameters to be tied together and robustly estimated.

Feature-functions play a central role in the model; this work explores both a large number of different kinds of feature-functions; the relation of feature-functions to sequence kernels and score-spaces allows a wide-range of existing approaches to be applied. The combination of latent variables and sequence kernels permits general classification of speech with discriminative models.

Given the rich variety of possible features that can be extracted from the observation and word sequences, the full potential of these discriminative models has barely been touched. The hope is that by incorporating a full range of various features, speech recognition systems will achieve the levels of performance that enable their use as a part of everyday life.

VII. Acknowledgements

The authors would like to thank Rogier van Dalen, Anton Ragni, Austin Zhang and Yotaro Kubo for fruitful discussions. The third author gratefully acknowledges support by NSF grant IIS-0643901 (CAREER).

References

[1] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *in Proc. the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] F. Jelinek, *Statistical methods for speech recognition*, MIT Press, 1997.

[3] M.J.F. Gales and S.J. Young, "The application of hidden Markov models in speech recognition," *Foundation and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[4] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *in Proc. ECML'98*, pp. 137–142, 1998.

[5] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML'01*, 2001, pp. 282–289.

[6] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, pp. 693–723, 2007.

[7] J. Morris and E. Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 617–628, 2008.

[8] N. Smith and M.J.F. Gales, "Speech recognition using SVMs," in *Advances in Neural Information Processing Systems*. 2002, pp. 1197–1204, MIT Press.

[9] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech'05*, 2005, pp. 1117–1120.

[10] Y.-H. Sung and D. Jurafsky, "Hidden conditional random fields for phone recognition," in *Proc. ASRU'09*, 2009, pp. 107–112.

[11] C. Saunders, J. Shawe-Taylor, and A Vinokourov, "String kernels, Fisher kernels and finite state automata," in *Advances in Neural Information Processing Systems*. 2002, pp. 633–640, MIT Press.

[12] M.I. Layton and M.J.F. Gales, "Acoustic modelling using continuous rational kernels," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 48, no. 1, pp. 67–82, 2007.

[13] S. Wiesler, M. Nußbaum-Thom, G. Heigold, R. Schlüter, and H. Ney, "Investigations on features for log-linear acoustic models in continuous speech recognition," in *Proc. ASRU 2009*, 2009, pp. 52–57.

[14] 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.

[15] A. Ragni and M.J.F. Gales, "Derivative kernels for noise robust ASR," in *Proc. of ASRU'11*, 2011, pp. 119–124.

[16] S.-X. Zhang and M. J. F. Gales, "Extending noise robust structured support vector machines to larger vocabulary tasks," in *Proc. ASRU'11*, 2011, pp. 18–23.

[17] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU'09*, 2009, pp. 152–157.

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

[19] H.K.J. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 873–881, 2006.

[20] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 354–365, 2009.

[21] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proc. ACL'04*, 2004.

[22] J.A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*, 2003, pp. 191–245.

[23] A. J. Viterbi, "Error bounds for convolutional codes and asymptotically optimum decoding algorithm," *IEEE Transactions Information Theory*, vol. 13, pp. 260–269, 1982.

[24] A. Nadas, "A Decision Theoretic Formulation of a Training Problem in Speech Recognition and a Comparison of Training by Unconditional Versus Conditional Maximum Likelihood," *IEEE Trans Acoustics Speech and Signal Processing*, vol. 31, no. 4, pp. 814–817, 1983.

[25] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP'02*, 2002, vol. 1, pp. 13–17.

[26] F. Sha and L.K. Saul, "Large margin hidden Markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems*. 2007, pp. 1249–1256, MIT Press.

[27] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.

[28] A.L. Berger, S.A. Della Pietra, and Della-Pietra V.J., "A maximum entropy approach to natural language processing," *Comutational Linguistics*, vol. 22, no. 1, pp. 39–72, 1996.

[29] V.N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 2000.

[30] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proc. ICML'04*, 2004.

[31] P. Nguyen, G. Heigold, and G. Zweig, "Speech recognition with flat direct models," *IEEE Journal of Selected Topics in Signal Procesing*, vol. 4, pp. 994–1006, 2010.

[32] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[33] Y.-H. Sung, C Boulis, C abd Manning, and D. Jurafsky, "Regularization, adaptation, and non-independent features improve hidden conditional random fields for phone classification," in *Proc. ASRU'07*, 2007, pp. 347–352.

[34] G. Heigold, R. Schlüter, and H. Ney, "On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields," in *Proc. Interspeech'07*, 2007, pp. 1721–1724.

[35] G. Zweig et al, "Speech recognition with segmental conditional random fields: A summary of the JHU CLSP Summer workshop," in *Proc. ICASSP'11*, 2011, pp. 5044 – 5047.

[36] M. Ostendorf, V. Digilakis, and O. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[37] C.N.J. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1169–1176.

[38] S. Sarawagi and W. W. Cohen, "Semi-Markov conditional random fields for information extraction," in *In Advances in Neural Information Processing Systems*. 2004, pp. 1185–1192, MIT Press.

[39] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, pp. 1453–1484, 2005.

[40] W. Byrne, "Minimum Bayes risk estimation and decoding in large vocabulary continuous speech recognition," *IEICE Special Issue on Statistical Modelling for Speech Recognition*, vol. 89, no. 3, pp. 900–907, 2006.

[41] M.J.F. Gales, "Discriminative models for speech recognition," in *Proc. Information Theory and Applications Workshop*, 2007, pp. 170–176.

[42] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Proc. EMNLP'02*, 2002.

[43] S. F. Chen and R. Rosenfeld, "Efficient sampling and feature selection in whole sentence maximum enropy language models," in *Proc. ICASSP'99*, 1999, pp. 549–552.

[44] S. Wiesler, A. Richards, Y. Kubo, R. Schlüter, and H. Ney, "Feature selection for log-linear acoustic models," in *Proc. ICASSP'11*, 2011, pp. 5324–5327.

[45] C Chelba and A Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *Computer Speech and Language*, vol. 20, no. 4, pp. 382 – 399, 2006.
[46] Y.-H. Sung, C. Boulis, and D. Jurafsky, "Maximum conditional likelihood linear regression and maximum a posteriori for hidden conditional random fields speaker adaptation," in *Proc. ICASSP'08*, 2008, pp. 4293–4296.
[47] J. Loof, R. Schlüter, and H. Ney, "Discriminative adaptation for log-linear acoustic models," in *Proc. Interspeech'10*, 2010, pp. 1648–1651.
[48] M.J.F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Computer Speech and Language*, vol. 24, no. 4, pp. 648–662, 2010.
[49] C. Cortes, P. Haffner, and M. Mohri, "Rational kernels: Theory and algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 1035–1062, 2004.
[50] C. Longworth, *Kernel Methods for Text-Independent Speaker Verification*, Ph.D. thesis, Cambridge University, 2010.
[51] T. Jaakkola and D. Hausser, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*. 1999, pp. 487–493, MIT Press.
[52] R.C. van Dalen, A. Ragni, and M.J.F. Gales, "Efficient decodingwith continuous rational kernels using the expectation semiring," Tech. Rep. CUED/F-INFENG/TR674, Cambridge University Engineering Department, 2012.
[53] S. F. Chen, "Shrinking exponential language models," in *Proc. HLT-NAACL*, 2009, pp. 468–476.
[54] M. Collins, B. Roark, and M. Saraclar, "Discriminative syntactic language modeling for speech recognition," in *Proc. of ACL'05*, 2005, pp. 507–514.
[55] E. Arisoy, M. Saraçlar, B. Roark, and I. Shafran, "Syntactic and sub-lexical features for Turkish discriminative language models," in *Proc. ICASSP'10*, 2010, pp. 5538–5541.
[56] S. Watanabe, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition using WFST-based linear classifier for structured data," in *Proc. Interspeech'10*, 2010, pp. 346–349.
[57] M. Lehr and I. Shafran, "Learning a discriminative weighted finite-state transducer for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1360–1367, 2011.
[58] T. Oba, T. Hori, A. Nakamura, and A. Ito, "Round-robin duel discriminative language models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1244–1255, 2012.
[59] K. Demuynck, D. Seppi, P. van Compernolle, D.and Nguyen, and G. Zweig, "Integrating meta-information into examplar-based speech recognition with segmental conditional random fields," in *Proc. ICASSP'11*, 2011, pp. 5048–5051.
[60] T. Joachims, T. Finley, and C.-N. Yu, "Cutting-plane training of structural SVMs," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.
[61] A. Yuille, A. Rangarajan, and A. L. Yuille, "The concave-convex procedure (CCCP)," in *Advances in Neural Information Processing Systems*. 2003, pp. 915–936, MIT Press.

Mark Gales received the B.A. and PhD degrees from the University of Cambridge, Cambridge, U.K. From 1995 to 1997, he was a Research Fellow at Emmanuel College, Cambridge. He was then a Research Staff Member in the Speech Group at the IBM T. J. Watson Research Center, Yorktown Heights, NY, until 1999 when he returned to Cambridge University Engineering Department as a University Lecturer. He was promoted to the position Professor of Information Engineering in October 2012. He is a Fellow of the IEEE and is currently an Associate Editor of IEEE Transactions on Audio Speech and Language Processing.

Shinji Watanabe received his B.S., M.S., and Dr. Eng. degrees from Waseda University, Tokyo, Japan, in 1999, 2001, and 2006, respectively. From 2001 to 2011, he was a research scientist at NTT Communication Science Laboratories, Kyoto, Japan. From 2011, he has been working at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA. His research interests include Bayesian learning, pattern recognition, and speech and spoken language processing. He is currently an Associate Editor of IEEE Transactions on Audio Speech and Language Processing.

Eric Fosler-Lussier Eric Fosler-Lussier (fosler@cse.ohio-state.edu) received the B.A.S in Computer and Cognitive Studies and the B.A. in Linguistics from the University of Pennsylvania in 1993. He received the Ph.D. degree from the University of California, Berkeley in 1999; his Ph.D. research was conducted at the International Computer Science Institute. As an Associate Professor with the Department of Computer Science and Engineering (and Linguistics by courtesy) at The Ohio State University, he directs the Speech and Language Technologies (SLaTe) Laboratory. He currently serves on the IEEE Speech and Language Technical Committee, and received the 2010 Signal Processing Society Best Paper Award.