Hidden Conditional Random Fields for Phone Classification

Asela Gunawardana, Milind Mahajan, Alex Acero, John C. Platt

Microsoft Research
One Microsoft Way
Redmond, WA 98052
USA
{aselag,milindm,alexac,jplatt}@microsoft.com

Abstract

In this paper, we show the novel application of hidden conditional random fields (HCRFs) – conditional random fields with hidden state sequences – for modeling speech. Hidden state sequences are critical for modeling the non-stationarity of speech signals. We show that HCRFs can easily be trained using the simple direct optimization technique of stochastic gradient descent. We present the results on the TIMIT phone classification task and show that HCRFs outperforms comparable ML and CML/MMI trained HMMs. In fact, HCRF results on this task are the best single classifier results known to us. We note that the HCRF framework is easily extensible to recognition since it is a state and label sequence modeling technique. We also note that HCRFs have the ability to handle complex features without any change in training procedure.

1. Introduction

Recently, there has been a resurgence of interest in discriminative methods for automatic speech recognition (ASR) due to the success of extended Baum-Welch (EBW) based techniques such as maximum mutual information (MMI) and minimum phone error (MPE) training in large vocabulary conversational speech recognition (LVCSR) [1]. However, the methods are poorly understood as they are used in ways in which their convergence guarantees no longer hold, and their successful use is as much art as it is science [1]. The rationale for the use of these EBW based techniques is that general unconstrained optimization algorithms are not well-suited to optimizing generative hidden Markov models (HMMs) under discriminative criteria such as the conditional likelihood [2]. We present a class of models that in contrast to HMMs are discriminative rather than generative in nature, and are amenable to the use of general purpose unconstrained optimization algorithms.

The HMM framework is restrictive in that all states need to model the observations in a uniform way, and that it is difficult to incorporate long-range dependencies between the states and the observations. Maximum entropy Markov models (MEMMs) [3] are direct (non-generative) models that attempt to remedy this – instead of observations being generated at each state, the state sequence is generated conditioned on the observations. The state at each time is chosen with a probability that depends on the previous state as well as the observations. The model does not assign probability to the observations, and the conditional state transition probabilities are exponential ("maximum entropy") distributions that may depend on arbitrary features of the entire observation sequence.

Conditional random fields (CRFs) [4] are generalizations of MEMMs where the conditional probability of the entire state sequence given the observation sequence is modeled as an exponential distribution. Although the CRF framework allows arbitrary dependencies between states, we will impose a Markov structure on the state sequence, but will not insist on normalized conditional transition probabilities at each transition. Thus, while MEMMs use per-state exponential distributions to model the transition probability at each state, CRFs use a single exponential distribution to model the entire state sequence given the observation sequence. In effect, CRFs allow unnormalized or weighted transition probabilities, allowing them to trade off the influence of less informative parts of the observation sequence against that of more informative parts [4]. In both cases, a Markov assumption on the state sequence allows the use of dynamic programming to yield a Viterbi algorithm for decoding [4].

MEMMs and CRFs have been used successfully for tasks such as part-of-speech (POS) tagging and information extraction [3, 4]. MEMMs have also been applied to ASR with some success [5], while recent work on maximum entropy acoustic models [6] can be interpreted as an application of a somewhat constrained CRF to ASR. In ASR, the use of mixture models and multiple state models in modeling the observations means that the training data is incomplete in that the frame by frame state and mixture component alignments are hidden. This is in contrast to POS tagging and information extraction, where each training token is completely labeled. In both previous approaches using MEMMs and CRFs for speech [5, 6], an HMM system is used to reveal the "correct" training state sequence through Viterbi alignment, which is used as ground truth during training. This allows the models to be trained using the generalized iterative scaling (GIS) [7] algorithm and its variants.

We generalize this work and use CRFs with hidden state sequences for modeling speech. We term these models hidden CRFs (HCRFs). HCRFs are able to use features which can be arbitrary functions of the observations without complicating the training. In this paper, we have not taken the advantage of this ability of HCRFs, but instead have limited ourselves to using the standard per-frame MFCC based features which have typically been used in speech recognition. This allows for a careful controlled comparison of the HMM and HCRF model families.

HCRFs, unlike HMMs, do not have normalized probability distributions for transitions or output probabilities. This makes it unnecessary to use special purpose algorithms such as the EBW algorithm used in MMI and MPE estimation. CRFs are typically trained using iterative scaling methods or quasi-Newton methods such as L-BFGS [8]. It is possible to train HCRFs using Generalized EM (GEM) where the M-step is an

iterative algorithm such as GIS or L-BFGS, rather than a closed form solution. As an alternative to (G)EM, direct optimization of the conditional log likelihood using a general optimization technique such as L-BFGS is possible and probably desirable since it avoids the indirection involved in the use of the EM auxiliary function. We have successfully used direct optimization techniques such as L-BFGS and stochastic gradient descent [9] to estimate HCRF parameters. We note that this approach is generalizable to other smooth discriminative criteria such as the conditional expectation of the raw phone or word error rate [10], or the smoothed empirical error of the training data [11].

We compare the performance of the novel HCRF models for speech to that of ML trained HMMs and maximum mutual information (MMI) trained HMMs on the TIMIT phone classification task and show that HCRFs outperform both types of HMMs using the same feature set and the model structure. The performance of HCRFs is the best single classifier results we know of on this task – including techniques such as support vector machines [12] and neural networks [13]. The advantage of HCRFs is that the model is a state sequence probability model, even when applied to the phone classification task, and can easily be extended to recognition tasks where the boundaries of phonetic segments are unknown.

2. HCRFs as a generalization of HMMs

The HCRF model gives the conditional probability of a segment (phonetic) label w given the observation sequence $\mathbf{o} = (o_1, \cdots, o_T)$:

$$p(w|\mathbf{o};\, \lambda) = \frac{1}{z(\mathbf{o};\lambda)} \sum_{\mathbf{s}\in w} \exp\left\{\lambda \cdot f(w, \mathbf{s}, \mathbf{o})\right\}. \qquad (1)$$

If the hidden state sequence $\mathbf{s} = (s_1, \cdots, s_T)$ is not marginalized out, we would have a CRF $p(w, \mathbf{s}|\mathbf{o};\, \lambda)$ rather than an HCRF. The marginalization is over state sequences that belong to the model for w. λ is the *parameter vector* and $f(w, \mathbf{s}, \mathbf{o})$ is a vector of sufficient statistics referred to as the *feature vector*. Note that in this context, the term feature vector refers to the vector of sufficient statistics used by the model, and not to the output of the acoustic front-end. The latter will be referred to as an *observation vector*. The *partition function* $z(\mathbf{o}; \lambda)$ ensures that the model is a properly normalized probability, and is given by $z(\mathbf{o}; \lambda) = \sum_{w,\mathbf{s}\in w} \exp\left\{\lambda \cdot f(w, \mathbf{s}, \mathbf{o})\right\}$.

The choice of sufficient statistics determines the dependencies modeled by the HCRF. In order to compare the performance of HCRFs with that of discriminatively trained Gaussian emission HMMs, we restrict our attention HCRFs with the same sufficient statistics. Namely, we use the vector of sufficient statistics f with components

$$\begin{aligned}
f^{(LM)}_{w'}(w, \mathbf{s}, \mathbf{o}) &= \delta(w = w') && \forall w' \\
f^{(Tr)}_{ss'}(w, \mathbf{s}, \mathbf{o}) &= \sum_{t=1}^{T} \delta(s_{t-1} = s)\delta(s_t = s') && \forall s, s' \\
f^{(Occ)}_{s}(w, \mathbf{s}, \mathbf{o}) &= \sum_{t=1}^{T} \delta(s_t = s) && \forall s \qquad (2) \\
f^{(M1)}_{s}(w, \mathbf{s}, \mathbf{o}) &= \sum_{t=1}^{T} \delta(s_t = s) o_t && \forall s \\
f^{(M2)}_{s}(w, \mathbf{s}, \mathbf{o}) &= \sum_{t=1}^{T} \delta(s_t = s) o_t^2 && \forall s,
\end{aligned}$$

where $\delta(s = s')$ is equal to one when $s = s'$ and zero otherwise. Each (unigram) language model feature $f^{(LM)}_{w}$ triggers on the occurrence of the label w. The transition features $f^{(Tr)}_{ss'}$ count the number of times the transition ss' occurs in $\mathbf{s}$, while the occupancy features $f^{(Occ)}_{s}$ count the occurrences of the state s. The first and second moments $f^{(M1)}_{s}$ and $f^{(M2)}_{s}$ are the sum and sum of squares of observations that align with the state s. These sufficient statistics may be recognized as the ones that are commonly accumulated in order to estimate HMMs. Since all components of f are sums of terms that involve at most pairs of neighboring states, the state sequence is Markov given the observation sequence, which allows the use of dynamic programming algorithms such as Forward-Backward and Viterbi as with HMMs. Note that for simplicity, we have only given expressions for using scalar observations and single Gaussian emission densities, although the arguments hold for vector valued observations and mixture densities. In fact, all experiments were performed with the familiar vector valued observations and diagonal covariance Gaussian mixture emissions.

It can be shown that setting the corresponding components of λ to

$$\begin{aligned}
\lambda^{(LM)}_{w'} &= \log u_{w'} && \forall w' \\
\lambda^{(Tr)}_{ss'} &= \log a_{ss'} && \forall s, s' \\
\lambda^{(Occ)}_{s} &= -\tfrac{1}{2}\left(\log 2\pi\sigma_s^2 + \frac{\mu_s^2}{\sigma_s^2}\right) && \forall s \\
\lambda^{(M1)}_{s} &= \frac{\mu_s}{\sigma_s^2} && \forall s \\
\lambda^{(M2)}_{s} &= -\frac{1}{2\sigma_s^2} && \forall s
\end{aligned}$$

gives the conditional p.d.f. induced by an HMM with transition probabilities $a_{ss'}$, emission means μ_s, emission covariance σ_s^2 and unigram probability u_w.

Note that equation (1) with the feature vector f of equation (2) gives a valid conditional probability for *any* value of the parameter vector λ. However, not every value of λ corresponds to an HMM. In particular, $\lambda^{(M2)}_{s}$ may be non-negative,and $\lambda^{(Occ)}_{s}$ and $\lambda^{(Tr)}_{ss'}$ may include a weight that emphasizes or deemphasizes a particular state or transition. Therefore, even though they model the same dependencies through the same sufficient statistics, the HMMs give a constrained subset of the the set of HCRF conditional probabilities.

3. HCRF Estimation

As noted in Section 1, we have chosen to use direct optimization of the conditional log-likelihood of the training set rather than GEM. We therefore need to find λ to maximize the conditional log-likelihood of the training set

$$\mathcal{L}(\lambda) = \sum_{n=1}^{N} \log p(w^{(n)}|\mathbf{o}^{(n)};\, \lambda).$$

L-BFGS is a well-known low-memory quasi-Newton method which has been applied successfully to the estimation of CRF parameters [14]. L-BFGS approximates the inverse of the Hessian using the history of the changes in parameter and gradient values (known as correction pairs) at previous L-BFGS iterations. Typically, 3 to 20 such most recent correction pairs are stored [8].

L-BFGS is a batch training method which uses the statistics such as $\nabla\mathcal{L}(\lambda)$ computed from the entire training set in

order to make an update to the parameter vector λ. In contrast, stochastic gradient descent (SGD) updates the parameter vector after processing each single training sample using noisy estimates of the gradient $\nabla\mathcal{L}(\lambda)$. More specifically, if $(w^{(1)}, \mathbf{o}^{(1)}) \ldots (w^{(N)}, \mathbf{o}^{(N)})$ is the entire sequence of training samples processed by SGD, then:

$$\lambda^{(n+1)} = \lambda^{(n)} + \eta^{(n)} U^{(n)} \nabla_\lambda \log p(w^{(n)}|\mathbf{o}^{(n)}; \lambda^{(n)})$$

where $\eta^{(n)}$ is the learning rate and $U^{(n)}$ is a conditioning matrix which can be used to speed up the convergence. We used a constant learning rate $\eta^{(n)} = \eta$ and an identity conditioning matrix $U^{(n)} = I$. The training samples processed by SGD can be randomly drawn from the training set and the same sample can be processed multiple times. We also used a parameter averaging technique which is known to benefit robustness of stochastic approximation algorithms like SGD [9, 15]. The averaged parameters are obtained as $\lambda_{avg} = \frac{1}{N}\sum_{n=1}^{N}\lambda^{(n)}$. SGD training can be viewed as a softened extension of perceptron training [15] to hidden variable problems.

Both L-BFGS and SGD require the computation of the gradient of $\log p(\hat{w}|\hat{\mathbf{o}})$. It can be shown that taking the gradient of equation (1) and rearranging gives

$$\nabla_\lambda \log p(\hat{w}|\hat{\mathbf{o}};\, \lambda) = \sum_{\mathbf{s}\in\hat{w}} f(\hat{w}, \mathbf{s}, \hat{\mathbf{o}}) p(\mathbf{s}|\hat{w}, \hat{\mathbf{o}};\, \lambda) - \sum_{w,\mathbf{s}\in w} f(w, \mathbf{s}, \hat{\mathbf{o}}) p(w, \mathbf{s}|\hat{\mathbf{o}};\, \lambda).$$

Substituting the vector of sufficient statistics f from equation (2) into the gradient, it can be shown that the first and second terms are the "numerator" and "denominator" counts used in MMI estimation of HMMs [1]. Because the HCRF imposes a Markov structure on the state sequences these statistics can be efficiently computed from the occupancy probabilities $p(s_{t-1} = s, s_t = s'|w, \mathbf{o})$ and $p(s_t = s|w, \mathbf{o})$, which in turn can be computed using a forward-backward algorithm, just as with MMI estimation of HMMs. The forward and backward recursions and the computation of occupancy probabilities are analogous to the case of HMM estimation, with the transition probability $a_{ss'}$ replaced by a transition score $\exp\left(\lambda_{s's}^{(Tr)}\right)$ and the observation probability $\mathcal{N}(o_t;\, \mu_s, \sigma_s^2)$ replaced by an observation score $\exp\left(\lambda_s^{(Occ)} + \lambda_s^{(M1)} o_t + \lambda_s^{(M2)} o_t^2\right)$. For example, the forward recursion for HCRFs is given by

$$\alpha_t(s) = \left(\sum_{s'} \alpha_{t-1}(s') e^{\lambda_{s's}^{(Tr)}}\right) e^{\left(\lambda_s^{(Occ)} + \lambda_s^{(M1)} o_t + \lambda_s^{(M2)} o_t^2\right)}$$

in contrast to

$$\alpha_t(s) = \left(\sum_{s'} \alpha_{t-1}(s') a_{s's}\right) \mathcal{N}\left(o_t;\, \mu_s, \sigma_s^2\right)$$

for HMMs. Thus, the gradient of the log conditional likelihood can be efficiently computed, just as with MMI estimation of HMMs.

Note that the conditional log-likelihood is not convex in λ. Training methods will therefore in general find a local optimum rather than the global optimum. We initialized the HCRF estimation from ML trained HMM parameters.

3.1. Generalizing to multi-component models on vector valued observations

Most state-of-the-art ASR systems use vector valued observations, which are modeled with Gaussian mixture emission densities. In this case, the corresponding HCRF model generalizes to

$$p(w|\mathbf{o};\, \lambda) = \frac{1}{z(\mathbf{o};\, \lambda)} \sum_{(\mathbf{s},\mathbf{m})\in w} \exp\left\{\lambda \cdot f(w, \mathbf{s}, \mathbf{m}, \mathbf{o})\right\}.$$

where $\mathbf{m}$ is a sequence of mixture components. In principle, this can be viewed as the HCRF of equation (1) with a factored state of the form (s, m), with vector-valued first and second moment features. The forward recursions generalize to

$$\alpha_t(s, m) = \left(\sum_{s'} \alpha_{t-1}(s') e^{\lambda_{s's}^{(Tr)}}\right) \cdot e^{\left(\lambda_{sm}^{(Occ)} + \lambda_{sm}^{(M1)} \cdot o_t + \lambda_{sm}^{(M2)} \cdot o_t^2\right)}$$

$$\alpha_t(s) = \sum_m \alpha_t(s, m)$$

where o^2 denotes the vector of per-component squares of the observation vector o, and the first and second moment parameters $\lambda_{sm}^{(M1)}$ and $\lambda_{sm}^{(M2)}$ are now vector valued. The backward recursions and the computation of posterior occupancy probabilities generalize analogously. Note that when an HMM is written in HCRF form, $\lambda_{sm}^{(Occ)}$ will include the logarithm of the mixture weight. If we modeled dependencies between components of the observation vector (i.e. full covariance matrices in the HMM case), there would additional second moment features for cross-terms, rather than just the squared terms as shown above.

4. Experimental Results

In this paper, we validate the ideas described above on the TIMIT phone classification task. We use the experimental setup described in [16]. Results are reported on the MIT development test set [16] and the NIST core test set. The training, development, and evaluation sets have 142,910, 15,334, and 7333 phonetic segments respectively. We follow the standard practice of building models for 48 different phones, and then mapping down to 39 phones for scoring purposes [16]. We use a standard Mel-Frequency Cepstral Coefficient (MFCC) front end. The cepstral analysis uses a 25 msec Hamming window with a frame shift of 10 msec. Frames are aligned so that there is equal overlap at the start and the end of each segment. Spectral analysis is performed using a 40 channel Mel filter bank from 64 Hz to 8 kHz. A pre-emphasis coefficient of 0.97 is used to correct spectral tilt. The first twelve cepstral coefficients as well as the zeroth cepstral coefficient are computed for each frame. The first and second time derivatives of the cepstra are used. The resulting 39-dimensional vectors are normalized so that they have zero mean and unit variance over the training set. The offset and scaling from the training set are used for normalizing the test data. Our baseline HMM system models each of the 48 unmapped phones with a three state left to right model, with 10, 20, or 40 diagonal Gaussians per state. We test HCRF models with exactly the same topologies and feature vectors f.

We provide results both for ML trained HMMs and MMI trained HMMs [1]. Technically, our discriminative HMMs are conditional maximum likelihood (CML) trained rather than MMI trained, as the (unigram) language model is also discriminatively estimated using the EBW algorithm, whereas MMI

Mix Comp.s	HMM (ML)	HMM (MMI)	HCRF (L-BFGS)	HCRF (SGD)
10	27.8%	23.8%	22.1%	20.6%
20	25.8%	23.2%	21.6%	**20.3%**
40	25.1%	23.4%	21.4%	20.4%

(a) Development set results

Mix Comp.s	HMM (ML)	HMM (MMI)	HCRF (L-BFGS)	HCRF (SGD)
10	28.1%	24.8%	23.7%	21.8%
20	26.8%	24.6%	23.2%	**21.7%**
40	26.4%	25.3%	23.3%	22.3%

(b) Evaluation set results

Table 1: Classification error as a function of the number of mixture components. HMMs estimated using ML and MMI criteria are compared to HCRFs estimated using L-BFGS and stochastic gradient descent.

only updates acoustic model parameters. However, we yield to common usage in the sequel and refer to these models as MMI trained. We build exactly comparable HCRFs using L-BFGS [8] and stochastic gradient descent [9]. ML models are used to initialize MMI HMMs and HCRFs. In the HMM case, we searched for the optimal language model weight on the development set. In the case of the HCRFs and MMI HMMs, the training algorithms were observed to automatically scale the acoustic and language components of the model appropriately. Training parameters such as learning rate, MMIE flattening weights, number of training iterations etc are optimized on the development set and held fixed on the evaluation set. The optimal learning rate was 0.003, while the LM scale and MMIE flattening weight were 4 and 0.25 respectively.

The results are compared in Table 1. It can be seen that HCRFs significantly outperform even discriminatively trained HMMs ($p < 0.001$). Our best result is 21.7% classification error, as compared to 22.4% reported in [12]. It should be noted that while MMI estimation of the HMMs and SGD estimation of the HCRFs converged within ten iterations over the training set, L-BFGS convergence was much slower, taking up to fifty iterations. Since all three algorithms make use of exactly the same statistics of the data, estimation time per iteration is comparable. Since the partition function $z(\mathbf{o}; \lambda)$ does not need to be computed during decoding, the decoding costs of HMMs and HCRFs are also comparable.

5. Conclusions

We have proposed HCRFs, which extend CRFs to problems with hidden variables. In particular, the use of hidden state sequences as in HMMs allows the use of HCRFs for the modeling of speech. Our results show that HCRFs can be efficiently trained using direct optimization of the conditional likelihood by stochastic gradient descent, and that they significantly outperform discriminatively trained HMMs at phone classification on the TIMIT database. In fact, the 21.7% classification rate yielded by HCRFs is the best result known to the authors that does not take advantage of combining multiple classifiers. Since the model deals naturally with hidden state sequences, it can easily be extended to recognition.

6. References

[1] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, 2003.

[2] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inf. Thry.*, vol. 37, pp. 107–113, Jan. 1991.

[3] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *ICML*, pp. 591–598, 2000.

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

[5] H.-K. J. Kuo and Y. Gao, "Maximum entropy direct model as a unified direct model for acoustic modeling in speech recognition," in *ICSLP*, 2004.

[6] W. Macherey and H. Ney, "A comparative study on maximum entropy and discriminative training for acoustic modeling in automatic speech recognition," in *Eurospeech*, pp. 493–496, 2003.

[7] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Stat.*, vol. 43, no. 5, pp. 1470–1480, 1972.

[8] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer-Verlag, 1999.

[9] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 1997.

[10] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP*, vol. I, pp. 105–108, 2002.

[11] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Sig. Proc.*, vol. 40, pp. 3043–3054, Dec. 1992.

[12] P. Clarkson and P. Moreno, "On the use of support vector machines for phonetic classification," in *ICASSP*, vol. 2, pp. 585–588, 1999.

[13] S. A. Zahorian, P. Silsbee, and X. Wang, "Phone classification with segmental features and a binary-pair partitioned neural network classifier," in *ICASSP*, vol. 2, pp. 1011–1014, April 1997.

[14] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. HLT-NAACL*, 2003.

[15] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *EMNLP*, pp. 1–8, 2002.

[16] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *Eurospeech*, pp. 401–404, 1997.