MAXIMUM CONDITIONAL LIKELIHOOD LINEAR REGRESSION AND MAXIMUM A POSTERIORI FOR HIDDEN CONDITIONAL RANDOM FIELDS SPEAKER ADAPTATION

Yun-Hsuan Sung,[1] *Constantinos Boulis,*[2] *Dan Jurafsky*[3]

Electrical Engineering,[1] Linguistics[2,3]
Stanford University
Stanford, CA, USA

ABSTRACT

This paper shows how to improve Hidden Conditional Random Fields (HCRFs) for phone classification by applying various speaker adaptation techniques. These include Maximum A Posteriori (MAP) adaptation as well as a new technique we introduce called Maximum Conditional Likelihood Linear Regression (MCLLR), a discriminative variant of the widely used MLLR algorithm. In previous work, we and others have shown that HCRFs outperform even discriminatively trained HMMs. In this paper we show that HCRFs adapted via MCLLR or via MAP adaptation also work better than similarly adapted HMMs. We also compare MCLLR and MAP adaptation performance with different amounts of adaptation data. MCLLR adaptation performs better when the amount of adaptation data is relatively small, while MAP adaptation outperforms MCLLR with larger amounts of adaptation.

Index Terms— Hidden Conditional Random Field, Speaker Adaptation, Maximum a Posteriori, Maximum Conditional Likelihood Linear Regression

1. INTRODUCTION

The Conditional Random Field (CRF) [1] is a widely used sequence labeling model that has attractive attributes as a replacement for the widely used Hidden Markov Model (HMM). CRFs don't have strong independence assumptions and have the potential to incorporate a rich set of overlapping and non-independent features. Moreover, CRFs are trained discriminatively, i.e. by maximizing the conditional probability of label given the observations.

Recently, there has been increasing interest in CRFs with hidden variables, i.e. **Hidden Conditional Random Fields** (**HCRFs**), introduced below in section 2. Like CRFs, HCRFs are undirected sequence models that incorporate a rich set of features and intrinsic discriminative training, and have proved successful in tasks like string edit distance (McCallum et. al. [2]), gesture recognition (Quattoni et. al. [3]), and phone classification (Gunawardana et al. [4], Sung et al. [5]).

In this paper, we explore techniques for improving HCRF phone classification via speaker adaptation. The first method is **Maximum Conditional Likelihood Linear Regression** (**MCLLR**), a discriminative variant of the widely used Maximum Likelihood Linear Regression (MLLR) method for HMM speaker adaptation [6, 7]. MCLLR resembles MLLR in learning a linear transform to modify the acoustic model parameters, but resembles the discriminative HMM adaptation method of [8] in maximizing the conditional likelihood, hence being a discriminative training method; see section 3 for details. The second method is **Maximum a Posterior** (**MAP**) adaptation which was successfully applied to HMM speaker adaptation by Gauvain and Lee [9], as well as to other models like Maximum Entropy Markov Models (MEMMs) [10], and which we applied to HCRF adaptation in [5].

Unadapted HCRFs have previously been showed to outperform HMMs [4]. We compare adapted HCRFs with adapted HMMs by both MAP and linear regression adaptation methods to see if HCRFs can still work better than HMMs after adaptation. We also compare the performance of MCLLR and MAP adaption for HCRFs with different amounts of adaptation data.

2. HIDDEN CONDITIONAL RANDOM FIELDS

An HCRF is a markov random field conditioned on designated evidence variables in which some of the variables are unobserved during training. The kind of linear chain structured HCRF that we use for speech recognition is simply a conditional distribution $p(y|\underline{X})$ with a sequential structure, as figure 1 shows. Assume that we are given a sequence of observations $\underline{X}$ and we want to give a corresponding label y; HCRFs model the conditional distribution function as:

$$p(y|\underline{X};\lambda) = \frac{1}{Z(\underline{X};\lambda)} \sum_{\underline{H}} \exp\{\lambda^T F(y,\underline{H},\underline{X})\} \qquad (1)$$

where $\underline{H}$ is the sequence of hidden variables. F is the feature vector which is a function of the label y, the hidden variable sequence $\underline{H}$, and the input observation sequence $\underline{X}$. λ is the parameter vector whose k^{th} element is the parameter corresponding to the k^{th} element in the feature vector F. The constant Z is called the *partition function* and is defined as:

$$Z(\underline{X};\lambda) = \sum_{y'} \sum_{\underline{H}} \exp\{\lambda^T F(y',\underline{H},\underline{X})\} \qquad (2)$$

1-4244-1484-9/08/$25.00 ©2008 IEEE

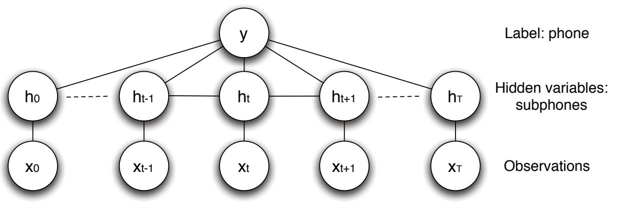

Fig. 1: Hidden Conditional Random Fields

which is used to make sure the conditional distribution summed over all possible labels be one. Due to having to sum over all possible instances of y and $\underline{H}$, the partition function is the main source of computation in learning.

The elements in the feature vector are the same as those mentioned in [4]. They include phone unigram features, state transition features, component occurrence features, first moment features, and second moment features as follow,

$$\begin{aligned}
f^{(LM)}_{y'} &= \delta(y = y') && \forall y' \\
f^{(Tr)}_{y'ss'} &= \sum_t \delta(y = y', s_{t-1} = s, s_t = s') && \forall y', s, s' \\
f^{(Occ)}_{s,m} &= \sum_t \delta(s_t = s, m_t = m) && \forall s, m \\
f^{(M1)}_{s,m} &= \sum_t \delta(s_t = s, m_t = m) x_t && \forall s, m \\
f^{(M2)}_{s,m} &= \sum_t \delta(s_t = s, m_t = m) x_t^2 && \forall s, m
\end{aligned}$$

We train HMMs for each phone as initial models for HCRF learning because the conditional log-likelihood is not concave and good initialization is important for finding a better optimum. In order to reduce overfitting, we add a Gaussian prior with the origin as center for regularization [5]. The regularization term gives the learning process prior knowledge about the parameters and the σ in the Gaussian prior is used to decide the degree of regularization. We used $\sigma = 10$ throughout the experiments in this paper. After taking the logarithm and adding the regularization term, we can reformulate equation (1) as:

$$\begin{aligned}
\log p(y|\underline{X}; \lambda) = \log \sum_{\underline{H}} \exp\{\lambda^T F(y, \underline{H}, \underline{X})\} \\
- \log \sum_{y'} \sum_{\underline{H}} \exp\{\lambda^T F(y', \underline{H}, \underline{X})\} - \frac{\lambda^T \lambda}{2\sigma^2} \quad (3)
\end{aligned}$$

We apply Stochastic Gradient Descent (SGD) as the optimization technique to optimize the conditional log-likelihood, equation (3), because computing the gradient over all the training data is tremendously expensive.

3. MAXIMUM CONDITIONAL LIKELIHOOD LINEAR REGRESSION

We introduce here a new method called Maximum Conditional Likelihood Linear Regression, similar to the MLLR [6] method used in HMM adaptation. MCLLR assumes the adapted parameters are a linear combination of the original parameters and describes the adaptation in matrix form. Unlike MLLR, MCLLR maximizes the conditional likelihood and is a discriminative training method, which improves the correct model and degrades the competitive models at the same time.

We reconstruct the parameter vector as $\nu = [1, \lambda_1, ..., \lambda_n]$, where $\lambda_1, ..., \lambda_n$ are the original parameters and n is the number of adapted parameters. The constant is added into the parameter vector as an offset. After that, we can describe the adapted parameter λ' as the linear combination of the original parameters by $\lambda' = M\nu$, where M is a n by $n+1$ transformation matrix. The learning process is to find the best M by maximizing conditional probability on the adaptation data.

The conditional probability can be further described as equation (4).

$$\begin{aligned}
p(y|\underline{X}; \lambda') &= \frac{1}{Z(\underline{X}; \lambda')} \sum_{\underline{H}} \exp\{\lambda'^T F(y, \underline{H}, \underline{X})\} \\
&= \frac{1}{Z(\underline{X}; M\nu)} \sum_{\underline{H}} \exp\{\nu^T M^T F(y, \underline{H}, \underline{X})\} \quad (4)
\end{aligned}$$

Instead of maximizing the conditional likelihood, we take the logarithm and add the regularization term to derive equation (5) from equation (4). The first two terms come from the original likelihood and the last term is a regularization term. The regularization is applied to reduce overfitting and is the same as the one mentioned in HCRF learning in section 2.

$$\begin{aligned}
\log p(y|\underline{X}; M) = \log \sum_{\underline{H}} \exp\{\nu^T M^T F(y, \underline{H}, \underline{X})\} \\
- \log \sum_{y'} \sum_{\underline{H}} \exp\{\nu^T M^T F(y', \underline{H}, \underline{X})\} \\
- \frac{\nu^T M^T M \nu}{2\sigma^2} \quad (5)
\end{aligned}$$

Equation (5) is maximized by Limited-memory BFGS, which is a kind of quasi-newton method [11]. The first reason to use Limited-memory BFGS instead of the SGD which we had used in HCRF learning is that the amount of adaptation data is generally small. As a result, it is fast to calculate the gradient over all data and finish one iteration in Limited-memory BFGS. The second reason is that it is not easy to find a good step size for SGD, while Limited-memory BFGS uses line search to decide the best step size in each iteration.

The gradient of conditional log-likelihood with respect to M can be derived as equation (6) and (7). Without considering the regularization term, when we reach the optimal point, the expectation of $F\nu^T$ by the distribution of hidden variables given label and observation variables equals the expectation of $F\nu^T$ by the distribution of label and hidden variables given observation variables, which is similar to the derivation for HCRF learning in [5].

$$\frac{\partial \log p(y|\underline{X}; M)}{\partial M} = \sum_{\underline{H}} F\nu^T p(\underline{H}|y, \underline{X})$$

$$- \sum_{y'} \sum_{\underline{H}} F\nu^T p(y', \underline{H}|\underline{X}) - \frac{M\nu\nu^T}{\sigma^2} \quad (6)$$

$$= E_{\underline{H}|y,\underline{X}}[F\nu^T] - E_{y',\underline{H}|\underline{X}}[F\nu^T] - \frac{M\nu\nu^T}{\sigma^2} \quad (7)$$

In this study, we only adapt the first moment parameters and keep the remaining parameters fixed, which corresponds to adapting the mean of the Gaussian distribution in HMMs. Because we only adapt the first moment parameters and share the transformation over all phones, the total number of free parameters for MCLLR adaptation is much smaller than that for MAP adaptation.

4. MAXIMUM A POSTERIORI ADAPTATION

To explore MAP adaptation for HCRF speaker adaptation, we reformulate equation (3) as:

$$\log p(y|\underline{X}; \lambda) = \log \sum_{\underline{H}} \exp\{\lambda^T F(y, \underline{H}, \underline{X})\}$$

$$- \log \sum_{y'} \sum_{\underline{H}} \exp\{\lambda^T F(y', \underline{H}, \underline{X})\}$$

$$- \frac{(\lambda - \lambda_o)^T (\lambda - \lambda_o)}{2\sigma^2} \quad (8)$$

Equation (3) and (8) differ only in the regularization term. In general HCRF training, we use the origin as the center of the Gaussian prior. In MAP adaptation, we replace the origin by the parameters of the speaker independent model, i.e. λ_o. Because the speaker independent models give us a good idea about what any acoustic model should look like, the last term is used as our general prior on models. The first and second terms are just the conditional log-likelihood given the adaptation data. We learn the new parameters by optimizing equation (8) which simultaneously considers both the speaker independent models and the new information from the adaptation data.

For reasons mentioned in section 3, limited-memory BFGS is used as the optimization technique to maximize conditional log-likelihood. The gradient of conditional log-likelihood with respect to λ can be derived further as:

$$\frac{\partial \log p(y|\underline{X}; \lambda)}{\partial \lambda} = \sum_{\underline{H}} F(y, \underline{H}, \underline{X}) p(\underline{H}|y, \underline{X})$$

$$- \sum_{y'} \sum_{\underline{H}} F(y', \underline{H}, \underline{X}) p(y', \underline{H}|\underline{X}) - \frac{\lambda - \lambda_o}{\sigma^2} \quad (9)$$

$$= E_{\underline{H}|y,\underline{X}}[F(y, \underline{H}, \underline{X})] - E_{y',\underline{H}|\underline{X}}[F(y', \underline{H}, \underline{X})] - \frac{\lambda - \lambda_o}{\sigma^2} \quad (10)$$

Speakers	s17	s20	s21	s22	s24
Adaptation	296	253	471	430	356
Test	336	182	329	1015	323

Speakers	s25	s26	s32	s33	s34
Adaptation	316	324	436	282	271
Test	680	346	309	281	451

Table 1: Numbers of utterances of adaptation and test data for each speaker.

5. THE BUCKEYE SPEECH CORPUS

The corpus we used for phone classification is the Buckeye Speech Corpus [12], which is a wide-band conversational speech corpus recorded in Ohio State University. The corpus contains 20 speakers conversing freely with an interviewer in Columbus, Ohio. The speech was orthographically transcribed and phonetically labeled by hand.

We choose the first 10 speakers (5 males and 5 females) for training speaker independent HMMs and HCRFs. We then use the remaining 10 speakers (5 males and 5 females) for adaptation and testing. For each testing speaker, we preserve 2 – 3 interviews for adaptation and use the remaining interviews for testing. The numbers of utterances of adaptation and test data for each speaker are shown in table 1 with averages 343.5 and 425.2, respectively. The average number of phones per utterance is around 27.03.

6. EXPERIMENT RESULTS

6.1. Comparison between HMM and HCRF adaptation

In the first experiment, we compare MAP and linear regression (MLLR and MCLLR) adaptation for HMMs and HCRFs. Table 2 shows the adaptation results for all adaptation data which has more than 250 utterances for each speaker. In the table, mix01 – mix08 stand for the number of components in both HMMs and HCRFs.

In the speaker-independent case, HCRFs work better than HMMs for all numbers of components by 8% – 15%. After MAP and linear regression adaptation, HCRFs still outperform HMMs with similar differences. For HCRF adaptation, we have a large amount of adaptation data, so MAP adaptation works better than MCLLR adaptation by 4% – 5%. This is because in MCLLR we only adapt the first moment parameters and assume the adapted parameters are just linear transformation of original parameters in MCLLR which constrains the freedom of adaptation. As a result, it performs worse than MAP adaptation with large amounts of adaptation data.

6.2. Comparison between MAP and MCLLR adaptation

In the second experiment, we explore how the amount of adaptation data influences the adaptation results for both MAP and MCLLR adaptation. In figure 2, the x-axis is the number of utterances in the adaptation data, from no adaptation data

HMM	mix01	mix02	mix04	mix08
Spkr-indep	64.95%	58.23%	55.41%	53.29%
MLLR	59.03%	52.86%	50.82%	49.29%
MAP	52.28%	48.47%	46.23%	44.51%

HCRF	mix01	mix02	mix04	mix08
Spkr-indep	49.73%	47.32%	46.25%	45.45%
MCLLR	44.02%	41.67%	41.37%	40.16%
MAP	39.13%	37.58%	36.82%	36.60%

Table 2: Phone Classification Errors for HMM and HCRF Adaptation with all adaptation data.

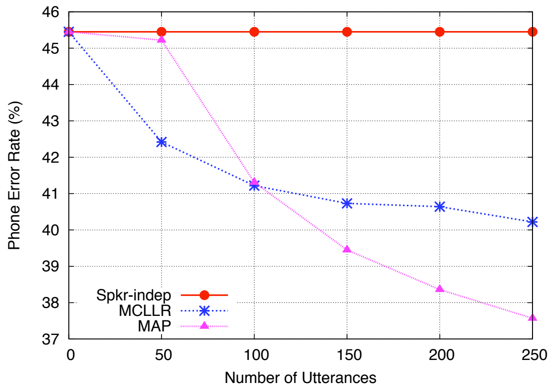

Fig. 2: Comparison between MAP and MCLLR adaptation

to 250 utterances for each speaker. The y-axis is the phone classification error rate. As the number of utterances in adaptation data increases, both MAP and MCLLR adaptation improve over speaker-independent models.

When the speaker-independent models are adapted by less than 100 utterances, MCLLR adaptation clearly works better than MAP adaptation. That is because the freedom for adaptation in MCLLR is much smaller than in MAP adaptation. As a result, MAP can not adjust the models too well when we don't have enough adaptation data. On the other hand, when the number of utterances is increased further, the advantage of the greater amount of freedom in MAP parameters becomes dominant. Therefore, the performance of MAP adaptation is better than that of MCLLR. The results are very similar to HMMs with MAP and MLLR adaptation in [13].

7. CONCLUSION

In this paper, we explore speaker adaptation for HCRF phone classification using two different approaches, Maximum a Posteriori adaptation, and a new discriminative method, Maximum Conditional Likelihood Linear Regression. Previous research found that unadapted HCRFs outperform even discriminatively trained HMMs. We find that the speaker-adaptive HCRFs still outperform the speaker-adaptive HMMs whether using MAP or linear regression methods. We also found that the performance of MAP and MCLLR HCRF adaptation with different amounts of adaptation data resembles the performance of MAP versus MLLR HMM adaptation. When the amount of adaptation is relatively small, we get better adaptation performance in MCLLR adaptation. When we have relatively large amount of adaptation data, MAP adaptation outperforms MCLLR adaptation.

8. ACKNOWLEDGMENT

We thank Chris Manning and the members of the Stanford NLP group for helpful comments. This work was supported by the Edinburgh-Stanford LINK and the ONR (MURI award N000140510388).

9. REFERENCES

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

[2] A. McCallum, K. Bellare, and F. Pereira, "A conditional random field for discriminatively-trained finite-state string edit distance," in *UAI*. AUAI Press, 2005, p. 388.

[3] A. Quattoni, S. Wang, L.P. Morency, M. Collins, and T. Darrell, "Hidden-state conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

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

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

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," in *Computer Speech and Language*, 1995, pp. 171–185.

[7] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," in *IEEE Trans. on Speech and Audio Processing*, 1995, pp. 357–366.

[8] A. Gunawardana and William Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proceedings of Eurospeech*, 2001, pp. 1203–1206.

[9] K. L. Gauvain and C. H. Lee, "Bayesian learning of gaussian mixture densities for hidden markov models," in *Proceedings of the DARPA speech and Natural Language Workshop*, 1991, pp. 272–277.

[10] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot.," *Computer Speech & Language*, vol. 20, no. 4, pp. 382–399, 2006.

[11] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-verlag, 1999.

[12] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," *Columbus, OH: Department of Psychology, Ohio State University*, 2007.

[13] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.