REGULARIZATION, ADAPTATION, AND NON-INDEPENDENT FEATURES IMPROVE HIDDEN CONDITIONAL RANDOM FIELDS FOR PHONE CLASSIFICATION

Yun-Hsuan Sung,[1] *Constantinos Boulis,*[2] *Christopher Manning,*[3] *Dan Jurafsky*[4]

Electrical Engineering,[1] Computer Science,[3,4] Linguistics[2,3,4]
Stanford University
Stanford, CA, USA

ABSTRACT

We show a number of improvements in the use of Hidden Conditional Random Fields (HCRFs) for phone classification on the TIMIT and Switchboard corpora. We first show that the use of *regularization* effectively prevents overfitting, improving over other methods such as early stopping. We then show that HCRFs are able to make use of non-independent features in phone classification, at least with small numbers of mixture components, while HMMs degrade due to their strong independence assumptions. Finally, we successfully apply Maximum a Posteriori adaptation to HCRFs, decreasing the phone classification error rate in the Switchboard corpus by around 1% – 5% given only small amounts of adaptation data.

***Index Terms*—** Hidden Conditional Random Fields, Speech Recognition, Phone Classification, Maximum a Posteriori

1. INTRODUCTION

While Hidden Markov Models (HMMs) have proved to be a very successful paradigm for acoustic modeling, they suffer from strong independence assumptions and usually don't work very well with non-independent features. Maximum Likelihood Estimation (MLE) training for HMMs achieves the underlying distributions only if the model assumptions are correct and there is an infinite amount of training data [1]. Since these assumptions are not generally true, researchers have switched to discriminative training methods.

Conditional Random Fields (CRFs) [2] are another widely-used sequence labeling model that are attractive as a potential replacement for HMMs. CRFs don't have strong independence assumptions and have the ability to incorporate a rich set of overlapping and non-independent features. In addition, CRFs are trained discriminatively by maximizing the conditional probability of the label given the observations.

Recently, there has been increasing interest in CRFs with hidden variables, i.e. **Hidden Conditional Random Fields** (**HCRFs**). 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. [3]) and gesture recognition (Quattoni et. al. [4]).

In this paper, we explore a number of extensions to HCRF models for phone classification. Phone classification is one of the simplest tasks in speech recognition, in which we are given a presegmented sequence of observations which must be assigned a single phone label. Gunawardana et. al. [5] have previously shown that HCRFs outperform both generatively and discriminatively trained HMMs on this task.

In our first study we examine the effect of **regularization** on HCRF learning to see if it improves learning. We next explore the use of **multiple overlapping features**. We augment the standard 39 MFCC features with a number of new features and show how HMMs and HCRFs are differently able to make use of this added information. Finally, we look at the important problem of **adaptation**. Adaptation techniques like MLLR and MAP have proved extremely useful in HMM systems for ASR. We explore whether MAP adaptation techniques can be applied to HCRF phone classification to make use of a small amount of adaptation data that comes from the same source as the testing data.

We present the detail on HCRFs in section 2 and 3. The application of HCRFs to phone classification is introduced in section 4. We report our main three studies as follows; applying regularization to remove overfitting (section 5), adding non-independent features (section 6), and MAP adaptation (section 7).

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 as

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

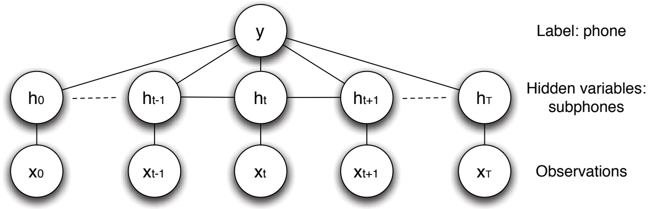

Fig. 1: Hidden Conditional Random Fields

$$p(y|\underline{X};\lambda) = \frac{1}{Z(\underline{X};\lambda)} \sum_{\underline{H}} \exp \sum_{k} \lambda_k f_k(y, \underline{H}, \underline{X}) \qquad (1)$$

where $\underline{H}$ is the sequence of hidden variables, f_k is the k^{th} feature which is a function of the label y, the hidden variable sequence $\underline{H}$, and the input observation sequence $\underline{X}$. λ_k is the corresponding parameter for each feature. The constant Z is called the *partition function* and is defined as

$$Z(\underline{X};\lambda) = \sum_{y'} \sum_{\underline{H}} \exp \sum_{k} \lambda_k f_k(y', \underline{H}, \underline{X}) \qquad (2)$$

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 major difference between HCRFs and CRFs is the introduction in HCRFs of some hidden variables corresponding to hidden structure. For speech, these hidden variables correspond to subphones (the states in an HMM model). Since we do not observe these hidden variables directly from input data, we need to marginalize over them in both learning and inference. This makes the training and inference of HCRFs more compute-intensive than traditional CRFs. Introducing hidden variables also makes the log-conditional likelihood non-concave, causing us to face problems with local maxima in training. Therefore, finding a good initialization becomes an important issue for learning in HCRFs.

3. LEARNING AND INFERENCE

When learning HCRFs, we want to maximize the conditional probability of the label y given the observation sequence $\underline{X}$. To simplify calculation, we maximize the log-conditional distribution instead of equation (1) directly. The objective function for optimization becomes

$$\begin{aligned}\log p(y|\underline{X};\lambda) = {} & \log \sum_{\underline{H}} \exp \sum_{k} \lambda_k f_k(y, \underline{H}, \underline{X}) \\ & - \log \sum_{y'} \sum_{\underline{H}} \exp \sum_{k} \lambda_k f_k(y', \underline{H}, \underline{X}) \\ & - \sum_{k} \frac{\lambda_k^2}{2\sigma^2} \qquad (3)\end{aligned}$$

The last term is used to represent Gaussian prior knowledge for **regularization**. Regularization has been shown to be useful at reducing overfitting in learning in CRFs [6].

The learning problem is formulated as an unconstrained optimization problem. As the optimization technique for training, we use Stochastic Gradient Descent (SGD), which has been shown to outperform quasi Newton methods such as Limited-memory BFGS for training HCRFs [5]. In each pass, we randomly draw one utterance from the training set with replacement and calculate the gradient based on that utterance. The parameters are then updated in the direction of the gradient with step size η as shown in equation (4).

$$\lambda_k^{(n+1)} = \lambda_k^{(n)} + \eta^{(n)} \frac{\partial \log p(y^{(n)}|\underline{X}^{(n)};\lambda^{(n)})}{\partial \lambda_k^{(n)}} \qquad (4)$$

The step size η is gradually decreased as the pass number increases by equation (5); τ is used to determine how fast the step size decreases:

$$\eta^{(n)} = \frac{\tau}{\tau + n} \qquad (5)$$

The corresponding gradient with respect to λ_k can be derived as follows

$$\begin{aligned}\frac{\partial \log p(y|\underline{X};\lambda)}{\partial \lambda_k} & = \sum_{\underline{H}} f_k(y, \underline{H}, \underline{X}) p(\underline{H}|y, \underline{X}) \\ & - \sum_{y'} \sum_{\underline{H}} f_k(y', \underline{H}, \underline{X}) p(y', \underline{H}|\underline{X}) - \frac{\lambda_k}{\sigma^2} \qquad (6) \\ & = E_{\underline{H}|y,\underline{X}}[f_k(y, \underline{H}, \underline{X})] - E_{y',\underline{H}|\underline{X}}[f_k(y', \underline{H}, \underline{X})] - \frac{\lambda_k}{\sigma^2} \qquad (7)\end{aligned}$$

When a local maximum is reached, the gradient equals zero. As equation (7) shows, if we do not include a regularization term, the expectation of features by the distribution of hidden variables given the label and observation variables is equal to the expectation of features by the distribution of hidden and label variables given observation variables. Sutton and McCallum [6] have given the corresponding derivation for CRF training. In CRFs, the empirical count of the features is equal to the expectation of features given the model distribution when the maximum is achieved. We can get the same result if we remove the hidden variables, $\underline{H}$ from equation (7). The gradient can be computed efficiently via the forward-backward algorithm.

Because SGD only considers one sample or a small number of samples in calculating the gradient, the gradient calculation becomes much faster than Limited-memory BFGS [7]. Instead of updating the parameters via a very accurate gradient, SGD updates the parameters several times during the same time period using a roughly estimated gradient. Hence SGD works better than other batch training methods when the calculation of gradient is highly time-consuming, as it is for HCRFs, which need to marginalize over all possible hidden variables.

However, due to the small number of samples used in each pass, the results are very unstable in general. Smoothing has been shown to be useful for increasing the convergence rate and stabilizing SGD [8]. The way we do smoothing is as follows:

$$\hat{\lambda}^{(n)} = \frac{\sum_{i=1}^{n} \gamma^i \lambda^{(i)}}{\sum_{i=1}^{n} \gamma^i} \tag{8}$$

where γ is a decay parameter used to determine how important the past parameters are in smoothing. We choose γ to be slightly less than one and $\hat{\lambda}^{(n)}$ is the final model we use for testing.

4. PHONE CLASSIFICATION

We apply HCRFs to the phone classification task, in which we are given a sequence of acoustic observations and must assign a single phone label. The hidden variables we use are the state variables $\underline{S}$, used to model subphones (akin to HMM states), and component variables $\underline{M}$, used to model the feature space structure in each subphone.

The feature functions we apply are the same as those of Gunawardana et. al. [5]. These include phone unigram features $f_{y'}^{(LM)}$, state transition features $f_{y'ss'}^{(Tr)}$, component occurrence features $f_{s,m}^{(Occ)}$, first moment features $f_{s,m}^{(M1)}$, and second moment features $f_{s,m}^{(M2)}$ as follow,

$$\begin{aligned}
f_{y'}^{(LM)} &= \delta(y = y') && \forall y' \\
f_{y'ss'}^{(Tr)} &= \sum_{t=1}^{T} \delta(y = y', s_{t-1} = s, s_t = s') && \forall y', s, s' \\
f_{s,m}^{(Occ)} &= \sum_{t=0}^{T} \delta(s_t = s, m_t = m) && \forall s, m \\
f_{s,m}^{(M1)} &= \sum_{t=0}^{T} \delta(s_t = s, m_t = m) x_t && \forall s, m \\
f_{s,m}^{(M2)} &= \sum_{t=0}^{T} \delta(s_t = s, m_t = m) x_t^2 && \forall s, m
\end{aligned}$$

where $\delta(\cdot)$ is an indicator function. The conditional log-likelihood is not concave, which means we have local maxima problem in learning. In order to find better local maxima, we need to have a good initializations to start the learning procedure. We do this by training an HMM with one Gaussian component by MLE. Then we transform the parameters of the HMM to the corresponding parameters of the HCRF via:

$$\begin{aligned}
\lambda_{y'}^{(LM)} &= \log u_{y'} && \forall y' \\
\lambda_{y'ss'}^{(Tr)} &= \log a_{y'ss'} && \forall y', s, s' \\
\lambda_{s,m}^{(Occ)} &= -\frac{1}{2} \sum_{d} (\log 2\pi\sigma_{s,m,d}^2 + \frac{\mu_{s,m,d}^2}{\sigma_{s,m,d}^2}) && \forall s, m \\
\lambda_{s,m,d}^{(M1)} &= \frac{\mu_{s,m,d}}{\sigma_{s,m,d}^2} && \forall s, m, d \\
\lambda_{s,m,d}^{(M2)} &= -\frac{1}{2\sigma_{s,m,d}^2} && \forall s, m, d
\end{aligned}$$

where $u_{y'}$ is the unigram probability, $a_{y'ss'}$ is the transition probability from state s to s' for phone y', and $\mu_{s,m,d}$ and $\sigma_{s,m,d}^2$ are the mean and variance of the dth dimension of observation vector of the mth Gaussian in the sth state, respectively.

We then do the training for HCRFs with one component. As the training finishes, we clone by splitting the component of each HCRF state into two different components, and adding a small value to $\lambda_{s,m,d}^{(M1)}$ for one, and subtracting it from the other. We use this as the initialization for HCRFs with two components and do the training again. We continue this procedure until the number of components of HCRFs reaches the number of components we want. Our experiments showed that this method gives us a better initialization than simply starting with parameters from an HMM already trained with the same number of Gaussians, especially for HCRFs with large numbers of components and features.

4.1. Task, Corpus, and Methodology

Our first two studies on HCRF phone classification use the TIMIT acoustic-phonetic continuous speech corpus [9]. Our experimental setup follows [10]. We map the 61 TIMIT phones into 48 phones for model training. The phone set is further collapsed from 48 phones to 39 phones for evaluation, replicating the method of Lee and Hon [11].

The training set in TIMIT contains 462 speakers and 4620 utterances in total. We use the core test set defined in TIMIT as our main test set (24 speakers and 192 utterances). The remaining 144 speakers (1152 utterances) in the test set are used as a development set for tuning parameters and choosing models.

We extract the standard 12 MFCC features and log energy with their delta and double delta to form 39 dimensional features. The window size and hopping time are 25ms and 10ms, respectively. Hamming window is applied with pre-emphasis coefficient 0.97. The number of filterbank channels is 40 and the number of cepstral filters is 12.

5. STUDY 1: REGULARIZATION

Earlier research suggests that CRFs are subject to overfitting, a problem that has been addressed by adding Gaussian prior

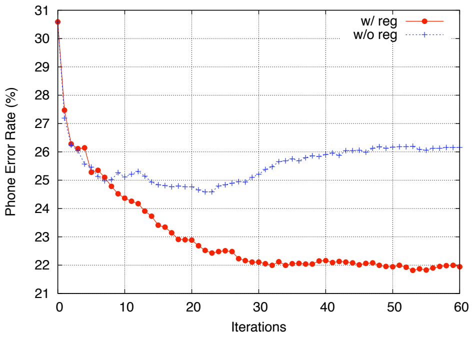

Fig. 2: Comparison between learning with and without regularization (in eight component HCRFs).

knowledge as a regularization term [6]. We found this same overfitting problem in our application of HCRFs to phone classification. We therefore applied the same regularization technique into HCRF learning as are shown in Equation (3).

Figure 2 shows the testing results of HCRFs learning with regularization and without regularization. As the number of iterations increases, the unregularized systems overfits the training corpus. Because of smoothing and the gradual decrease in step size in SGD, the final error rate converges and doesn't overfit too terribly. Generally speaking, it is possible to avoid overfitting by tuning the decrease in step size, but it is extremely difficult to tune perfectly without overfitting.

On the other hand, learning HCRFs with regularization can remove overfitting effectively and is not affected by over-decrease in the step size. As Figure 2 shows, we also converge to a better final results. Finally, regularization makes it possible to choose the models without a development set.

6. STUDY 2: ADDING NON-INDEPENDENT FEATURES

One of the well-known drawbacks for HMMs is that they have very strong independence assumptions among labels and observation. Given the current state, the current features are independent of the previous and next features. This assumption is not generally true in speech. Traditionally, the speech signal is analyzed by a short time Fourier Transform, which extracts the features from a short speech segment. Features are generally calculated in an overlapping window between adjacent features, which shows clearly that they are not independent.

However, HCRFs model the conditional probability directly, which do not explicitly represent the dependencies among the observation variables. Therefore, HCRFs have the potential to add a rich set of features without our caring about the dependency issues between them. In our second study, we incorporate richer features into HCRFs and and HMMs and compare their performance on TIMIT phone classification.

Comps	MFCC	PLP	M+P	M+long
mix01	36.91%	37.08%	37.74%	38.20%
mix02	34.90%	34.76%	35.34%	36.15%
mix04	32.83%	32.46%	33.33%	33.46%
mix08	30.59%	30.20%	31.00%	31.67%
mix16	29.41%	29.08%	29.80%	30.16%

Table 1: Phone error rate for adding non-independent features into MLE-trained HMMs.

Comps	MFCC	PLP	M+P	M+long
mix01	24.05%	24.05%	22.59%	23.41%
mix02	22.90%	22.37%	22.01%	22.25%
mix04	22.19%	22.12%	21.79%	21.91%
mix08	21.75%	21.66%	21.69%	21.75%
mix16	21.84%	21.46%	21.82%	22.12%

Table 2: Phone error rate for adding non-independent features into HCRFs.

6.1. Methods

We added two classes of features. First we combined PLP and MFCC features. In addition to the 39 MFCC features described in the previous study, we also extract Perceptual Linear Prediction (PLP) features, known to be competitive with MFCCs in speech recognition. We use the standard method of extracting 13 PLPs with their delta and double delta for 39 dimensional features. We use the same window size, hopping time and pre-emphasis coefficient as in MFCCs extraction. The Linear Cepstral Coefficient order is 12. We train and test the HCRFs with MFCCs and PLPs alone, respectively. Then, we combine MFCCs and PLPs to form a sequence of 78 dimensional feature vectors as our input observation sequence.

We next extracted long-distance features. In addition to the original MFCCs analyzed with a 25ms window, we calculate longer-distance MFCCs by applying a longer window length, 75ms, overlapping with the original 25ms window. All the remaining parameters for MFCCs extraction are the same as the one in short window MFCCs. We combine the short and long-term MFCC features to form a 78 dimensional feature vector.

6.2. Results of Study 2

Table 1 shows the results of adding overlapping and non-independent features in HMMs. As the table shows, combining MFCCs with PLPs actually degrades the phone error rate by around 0.7% – 0.8%. Adding long window MFCCs into original MFCCs results in even worse performance, increasing phone error by 1% – 1.2%. This shows the incorrect strong independence assumptions in HMMs.

On the other hand, at least for one, two, and four components, we get obvious improvements for HCRFs by adding non-independent features. In Table 2, we find combining

MFCCs and PLPs decreases the phone error rate for one, two, and four components. For eight and sixteen component HCRFs, the performance of the combined features degrades slightly. We believe this degradation is caused by search problems; adding more features complicates the model space, with the result that it is easy to get stuck in bad local maxima and in general requiring more training data for learning. In current work we are trying to find better initial points and other optimization techniques to solve this problem.

The results on combining short and long windowed MFCCs are very similar to those of the MFCC plus PLP experiments.

In summary, our best error rates (21.5%) are slightly better (lower) than the comparable HCRF results (21.7%) of [5], but just slightly worse (higher) than the current best published results on this task (21.1%) obtained by Large Margin Gaussian Mixture Models [12].

7. STUDY 3: MAP ADAPTATION

Acoustic models are very sensitive to specific speaker characteristics, and adaptation to small amounts of speaker data has been shown to significantly improve ASR performance on that speaker. Maximum a Posteriori (MAP) adaptation is a method that has been successfully applied to HMM speaker adaptation in speech (Gauvain and Lee [13]) as well as to other tasks like text capitalization [14]. In this study, we ask whether MAP adaptation can be used as well in HCRFs for speaker adaptation in phone classification. We trained universal HCRFs on data from various speakers, and adapted these HCRFs to several utterances from individual test speakers.

7.1. Methods

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

$$\begin{aligned}\log p(y|\underline{X};\lambda) = {} & \log \sum_{\underline{H}} \exp \sum_k \lambda_k f_k(y, \underline{H}, \underline{X}) \\ & - \log \sum_{y'} \sum_{\underline{H}} \exp \sum_k \lambda_k f_k(y', \underline{H}, \underline{X}) \\ & - \sum_k \frac{(\lambda_k - \lambda_{ko})^2}{2\sigma^2} \qquad (9)\end{aligned}$$

Equation (3) and (9) 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 universal model, i.e. λ_{ko}. Because the universal 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 (9) which simultaneously considers both the universal models and the new information from the adaptation data.

Comps	mix01	mix02	mix04	mix08
PER	57.56%	59.41%	61.15%	56.46 %
Comps	mix16	mix32	mix64	mix128
PER	50.95%	43.32%	36.59%	32.94%

Table 3: Phone classification error rate on Switchboard.

	HMMs		HCRFs	
Comps	Original	Adapted	Original	Adapted
mix01	88.79%	76.49%	56.42%	51.49%
mix02	79.63%	67.69%	58.18%	52.87%
mix04	73.43%	64.03%	60.67%	55.93%
mix08	65.63%	59.43%	55.39%	51.76%
mix16	57.69%	53.98%	50.00%	45.94%
mix32	48.21%	48.13%	42.09%	39.65%

Table 4: MAP adaptation with different number of components.

Equation (9) is maximized in the same way as the HCRF training described in section 3. SGD is used as the optimization technique and smoothing is also applied to increase the convergence rate of learning.

7.2. Task and Corpus

For MAP adaptation, rather than the TIMIT corpus, we used the part of the Switchboard corpus annotated at ICSI [15]. We used this corpus because we felt that it was important to see how our HCRF phone classification techniques worked on this more difficult corpus of human-human speech, and because the Switchboard corpus includes speakers with sufficient data for adaptation. The corpus, which contains phone boundaries for one hour of Switchboard speech, contains 734 speakers and 1285 utterances. Two of the speakers, 2830A and 2887B, have more than 100 utterances transcribed. We choose the first 60 utterances from those two speakers as the adaptation set and the rest of the utterances, 79 utterances for speaker 2830A and 68 utterances for speaker 2887B, as the test set. The remaining 734 speakers and 1018 utterances are used for training the universal models.

The phone set for the Switchboard transcriptions was quite large since phone labels included context informations. In order to reduce the number of phones to a reasonable number for training, we replace each triphone by its middle phone. The total number of phones we use is 51, which is not exactly the same as the phone set we use for TIMIT phone classification.

7.3. Results of Study 3

As the Switchboard phone classification results in Table 3 show, the error rate decreases as the number of components increase, although there is some fluctuation with small number of components. Interestingly, compared to the situation with TIMIT, in Switchboard we still get obvious improve-

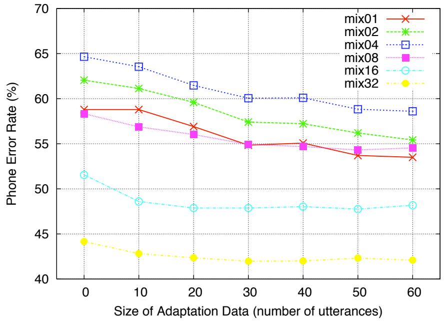

Fig. 3: MAP adaptation results with different amounts of adaptation data for speaker 2830A.

ments by using more than 32 components. We believe this is due to the greater variation in conversational speech than read speech.

We show the MAP speaker adaptation results for both HMMs and HCRFs with 60 utterances as adaptation data in Table 4. As the table shows, we got obvious improvements for all number of components in HCRF adaptation, and the resulting adapted HCRF models also perform significantly better than adapted HMMs. Not surprisingly, the magnitude of the improvement decreases as the number of components increases, presumably since HCRFs with a larger number of components have more parameters and hence need more data for adaptation.

Figure 3 shows how the amount of adaptation data influences the phone error rates. The x-axis is the number of utterances, ranging from no adaptation data (the original models) to 60 utterances. The y-axis is the phone classification error rate. As the figure shows, increasing the amount of adaptation data results in better performance. Even with only 10 utterances, we still get some benefit from MAP speaker adaptation. The improvement is more obvious when the number of components is small.

8. CONCLUSION

In this paper, we have replicated earlier work showing that HCRFs work better than HMMs for phone classification in read speech (TIMIT), and also, not previously shown, in conversational speech (Switchboard). Our work offers a number of augmentations to previous use of HCRFs for phone classification like [5]. We show that regularization can be used effectively to remove overfitting in HCRFs learning. We show preliminary results in HCRFs with small numbers of components suggesting that HCRFs have the potential to incorporate a large set of non-independent features; this result still requires further work to confirm this potential with larger numbers of components. Finally, we show that MAP adaptation can be applied as one adaptation technique for HCRFs, resulting in phone error rate reductions of 1% – 5% in conversational speech.

9. ACKNOWLEDGMENT

We thank Milind Mahajan for helpful conversations, and two anonymous reviewers as well as the members of the Stanford NLP group for useful suggestions. This work was supported by the Edinburgh-Stanford LINK and the ONR (MURI award N000140510388).

10. REFERENCES

[1] 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. on Acoustics, Speech and Signal Processing*, vol. 31, no. 4, pp. 814–817, 1983.

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

[3] A. McCallum, K. Bellare, and F. Pereira, "A conditional random field for discriminatively-trained finite-state string edit distance," in *UAI*, 2005, pp. 388–395.

[4] 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.

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

[6] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, 2006.

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

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

[9] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design an analysis of the acoustic-phonetic corpus," in *the DARPA Speech Recognition Workshop*, 1986.

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

[11] K. F. Lee and H. W. Hon, "Speaker independent phone recognition using hidden markov models," in *Proceedings of ICASSP*, 1980, vol. 37, pp. 1641–1648.

[12] F. Sha and L.K. Saul, "Large margin gaussian mixture modeling for phonetic classification and recognition," in *Proceedings of ICASSP*, 2006, pp. 265–268.

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

[14] 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.

[15] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *Proceedings of ICSLP*, 1996, vol. supplement, pp. S24–S27.