# Deep-Structured Hidden Conditional Random Fields for Phonetic Recognition

*Dong Yu and Li Deng*

Microsoft Research, One Microsoft Way, Redmond, WA, USA
{dongyu, deng}@microsoft.com

## Abstract

We extend our earlier work on deep-structured conditional random field (DCRF) and develop deep-structured *hidden* conditional random field (DHCRF). We investigate the use of this new sequential deep-learning model for phonetic recognition. DHCRF is a hierarchical model in which the final layer is a hidden conditional random field (HCRF) and the intermediate layers are zero-th-order conditional random fields (CRFs). Parameter estimation and sequence inference in the DHCRF are developed in this work. They are carried out layer by layer so that the time complexity is linear to the number of layers. In the DHCRF, the training label is available only at the final layer and the state boundary is unknown. This difficulty is addressed by using unsupervised learning for the intermediate layers and lattice-based supervised learning for the final layer. Experiments on the standard TIMIT phone recognition task show small performance improvement of a three-layer DHCRF over a two-layer DHCRF; both are significantly better than the single-layer DHCRF and are superior to the discriminatively trained tri-phone hidden Markov model (HMM) using identical input features.

**Index Terms**: hidden conditional random field, conditional random field, deep structure, phone recognition, TIMIT

## 1. Introduction

Recently there have been intense interests in applying deep structure models and deep learning techniques to automatic speech recognition (ASR) and natural language processing (NLP). Mohamed et al. [1][19] applied deep belief network (DBN) to the TIMIT phone recognition task. Deng et al. applied DBN to speech coding [20]. Prabhavalkar and Fosler-Lussier [2] developed the multilayer conditional random fields (MCRFs) with back-propagation training and applied it to phone recognition. Yu et al. proposed the deep-structured conditional random fields (DCRFs) [17] and obtained better than the state-of-the-art results on the query labeling [4] tasks.

In this paper, we extend our earlier work on DCRF to the deep-structured *hidden* conditional random fields (DHCRFs). DHCRF is a hierarchical model in which the final layer is a hidden conditional random field (HCRF) [5] and the intermediate layers are zero-th order conditional random fields (CRFs) that only exploit the observation features. In this work, parameter estimation and sequence inference in the DHCRF are developed, which are carried out in a layer-by-layer manner so that the time complexity is linear to the number of layers. In DHCRF the training label is available only at the final layer and the state boundary is unknown. We address these issues using unsupervised intermediate layer training and lattice-based supervised final layer training as will be described in detail in sections 2 and 3.

HCRF [5][6] was first proposed by Gunawardana et al. for phonetic classification. It is a discriminative model that generalizes both the hidden Markov model (HMM) and the conditional random field (CRF). Similar to the linear-chain CRF, HCRF models the state sequence as being conditionally Markovian given the observation sequence. It differs from the linear-chain CRF in that the segmentations are considered irrelevant and are unavailable in the training phase. HCRF also differs from HMM, where the state sequence is assumed to be Markovian but each observation is assumed to be conditionally independent of all others given the state. HCRF has been successfully applied to phonetic classification [5][6][7][8] and phonetic recognition tasks [9].

The contribution of this work consists of two parts. First, we extend the HCRF and DCRF to DHCRF and propose associated learning algorithms. DHCRF has higher modeling capacity than HCRF as we will show in Section 4. Second, we use tri-phone state sequences in training the DHCRF instead of using the mono-phone sequences as used in [9] for the HCRF, and achieved recognition accuracies that are superior to the tri-phone HMM systems using the identical features.

We evaluated the DHCRF on the standard TIMIT phone recognition task. Experiments show that a three-layer DHCRF can improve the phone accuracy rate (PAR) by 0.3% absolute compared to a two-layer DHCRF, which significantly outperforms the one-layer DHCRF (e.g., HCRF). The three-layer DHCRF is superior to the maximum mutual information estimation (MMIE) trained tri-phone HMM using the identical feature with 0.8% absolute PAR improvement.

The rest of the paper is organized as follows. In Section 2 we describe the structure of the DHCRF. In Section 3 we illustrate the strategy of learning the intermediate layers and the final layer of the DHCRF. We report experimental results in Section 4 and conclude the paper in Section 5.

## 2. Formulation of the DHCRF

The DHCRF developed and evaluated in this work is a hierarchical model as shown in Figure 1, where the final layer is an HCRF, and the intermediate layers are zero-th order CRFs that do not use state transition features. This decision was made following our earlier observation [4] that computational cost can be drastically reduced without significantly affecting recognition accuracy by using only the observation features in the intermediate layers.

In DHCRF, the observation sequence $\boldsymbol{o}^{\boldsymbol{j}}$at layer $j$ consists of two parts: the preceding layer's observation sequence $\boldsymbol{o}^{\boldsymbol{j-1}}$ and the frame-level log marginal posterior probabilities $\log p\left(s_t^{j-1}|\boldsymbol{o}^{\boldsymbol{j-1}}\right)$ computed from the preceding layer $j-1$, where $s_t^{j-1}$ is the state value at layer $j-1$. We denote $\boldsymbol{o}=[\boldsymbol{o}_{\boldsymbol{t}}], t=1,\cdots,T$ as the raw observations at the first layer.

Both parameter estimation and sequence inference in the DHCRF are carried out bottom-up layer by layer. The final layer's state sequence conditional probability is

$$p(w|\boldsymbol{o}^N;\boldsymbol{\lambda}^N) = \frac{1}{z(\boldsymbol{o}^N;\boldsymbol{\lambda}^N)}\sum_{s^N\in w} exp\left((\boldsymbol{\lambda}^N)^T\boldsymbol{f}(w,s^N,\boldsymbol{o}^N)\right), \qquad (1)$$

where $N$ is the total number of layers, $(\cdot)^T$ is the transposition of $(\cdot)$, $\boldsymbol{o}^N = (\boldsymbol{o}_1^N, \cdots, \boldsymbol{o}_T^N)$ is the observation sequence at the final layer, $w$ is the phoneme or word sequence, $s^N = (s_1^N, \cdots, s_T^N)$ is a hypothesized state sequence, $\boldsymbol{f}(w, s^N, \boldsymbol{o}^N) = [f_1(w, s^N, \boldsymbol{o}^N), \cdots, f_T(w, s^N, \boldsymbol{o}^N)]^T$ is the feature vector at the final layer, $\boldsymbol{\lambda}^N = [\lambda_1^N, \cdots, \lambda_T^N]^T$ is the model parameter (weight vector), and $z(\boldsymbol{o}^N; \boldsymbol{\lambda}^N) = \sum_{w, s^N \in w} exp\big((\boldsymbol{\lambda}^N)^T \boldsymbol{f}(w, s^N, \boldsymbol{o}^N)\big)$ is the partition function (normalization factor) to ensure probabilities $p(w|\boldsymbol{o}^N; \boldsymbol{\lambda}^N)$ sum to one. Note that in the partition function, we have ruled out invalid sequences by summing over valid phoneme or word sequences only.

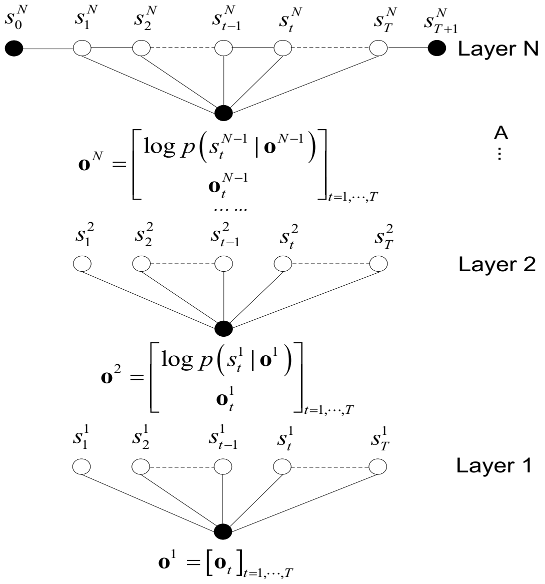
$s_0^N$
$s_1^N$
$s_2^N$
$s_{t-1}^N$
$s_t^N$
$s_T^N$
$s_{T+1}^N$
Layer N
$\mathbf{o}^N = \begin{bmatrix} \log p(s_t^{N-1} \mid \mathbf{o}^{N-1}) \\ \mathbf{o}_t^{N-1} \end{bmatrix}_{t=1,\cdots,T}$
A
⋮
… …
$s_1^2$
$s_2^2$
$s_{t-1}^2$
$s_t^2$
$s_T^2$
Layer 2
$\mathbf{o}^2 = \begin{bmatrix} \log p(s_t^1 \mid \mathbf{o}^1) \\ \mathbf{o}_t^1 \end{bmatrix}_{t=1,\cdots,T}$
$s_1^1$
$s_2^1$
$s_{t-1}^1$
$s_t^1$
$s_T^1$
Layer 1
$\mathbf{o}^1 = [\mathbf{o}_t]_{t=1,\cdots,T}$

Figure 1: Structure of DHCRF.

Different from the final layer, the state conditional probabilities at the intermediate layer $j$ are simply

$$p(s^j|\boldsymbol{o}^j; \boldsymbol{\lambda}^j) = \frac{1}{z(\boldsymbol{o}^j; \boldsymbol{\lambda}^j)} exp\left( (\boldsymbol{\lambda}^j)^T \boldsymbol{f}(s^j, \boldsymbol{o}^j) \right). \qquad (2)$$

This is different from (1) in two ways. First, transition features are not used in (2) and observation features $\boldsymbol{f}(s^j, \boldsymbol{o}^j)$ can be simplified to $[\boldsymbol{f}(s_t^j, \boldsymbol{o}_t^j)]_{t=1,\cdots,T}$ which is defined in Section 3.1. Second, there is no summation over state sequences with all possible segmentations in (2).

## 3. Learning in the DHCRF

The training supervision of DHCRF is available only at the final layer and is directly determined by the problem to be solved. For example, in the phonetic recognition task, the phoneme sequence $w$ is known at the final layer during the training phase. Parameter estimation at the final layer can thus be carried out in a supervised manner in the same way as explained in [5][6][7][9]. The supervision, however, is not available for the intermediate layers, which play the role of converting original observations to some intermediate abstract representations. For this reason, an unsupervised approach is needed to learn parameters in the intermediate layers. In this section, we will first describe the algorithms we recently developed to learn the intermediate representations. We then describe a lattice based algorithm to estimate the parameters in the final layer so that tri-phone models can be used.

### 3.1. Learning the Intermediate Layers

There are several approaches to learning the intermediate layer representations. For example, in the approach proposed in [3] we cast the intermediate layer learning problem into a multi-objective programming (MOP) problem in which we minimize the average frame-level conditional entropy and maximize the state occupation entropy at the same time. Minimizing the average frame-level conditional entropy forces the intermediate layers to be sharp indicators of subclasses (or clusters) for each input vector, while maximizing the occupation entropy guarantees that the input vectors be represented distinctly by different intermediate states. The MOP optimization algorithm alternates the steps in optimizing these two contradictory criteria until no further improvement in the criteria is possible or the maximum number of iterations is reached. The proposed MOP-based approach has been shown to be effective in a language identification task involving seven language classes [3]. Unfortunately, the MOP optimization becomes difficult when the number of classes in the intermediate layers becomes higher as in the phone recognition task since it is hard to control when to switch to optimize the other criterion given the vastly increased probability of being trapped into a local optimum.

In this study, we have adopted a simpler and more robust Gaussian mixture model (GMM) based algorithm. As discussed in Section 2, DHCRF is trained bottom-up layer by layer. Once a lower layer is trained, the parameters of that layer are fixed and the observation sequences of the next layer are generated using the newly trained lower-layer parameters. This process continues until all the layers are trained.

To learn the parameters of an intermediate layer, we first train a single GMM with diagonal covariance initialized from the corresponding HMM model which are optimized using the Gaussian splitting strategy. We then assign

$$s_t^j = \underset{i}{\operatorname{argmax}} N(\boldsymbol{o}_t^j; \boldsymbol{\mu}_i^j, \boldsymbol{\Sigma}_i^j) \qquad (3)$$

as the state value to each observation frame $\boldsymbol{o}_t^j$ at layer $j$ by assuming each Gaussian component is a state, where $\boldsymbol{\mu}_i^j$ and $\boldsymbol{\Sigma}_i^j$ are the mean and variance of the $i$-th Gaussian component at layer $j$. We then learn the parameters of the CRF at layer $j$ by maximizing the regularized log-conditional probability

$$J_1(\boldsymbol{\lambda}^j) = \sum_k \sum_{\mathrm{t}} \log p\left(\mathrm{s}_{\mathrm{t}}^{(k),j} | \mathbf{o}_{\mathrm{t}}^{(k),j}; \boldsymbol{\lambda}^j\right) - \frac{\|\boldsymbol{\lambda}^j\|_1}{\sigma_1} - \frac{\|\boldsymbol{\lambda}^j\|_2^2}{\sigma_2}, \qquad (4)$$

where $k$ is the utterance ID, $\|.\|_1$ is L1-norm to enforce sparseness of the parameters associated with each state value, $\|.\|_2^2$ is the square of L2-norm to give preference to smaller weights, and $\sigma_1$ and $\sigma_2$ are positive values to determine the importance of each regularization term. We have used the regularized dual averaging method [21] to solve this optimization problem with L1/L2 regularization terms.

As mentioned in Section 1, transition features are not used in the intermediate layers. Instead, we use only the first- and second-order observation features

$$\boldsymbol{f}_{s'}^{(M1)}(s_t, \boldsymbol{o}_t) = \delta(s_t = s')\boldsymbol{o}_t \qquad \forall s' \qquad (5)$$

$$\boldsymbol{f}_{s'}^{(M2)}(s_t, \boldsymbol{o}_t) = \delta(s_t = s')\boldsymbol{o}_t \circ \boldsymbol{o}_t \qquad \forall s' \qquad (6)$$

where ∘ is element-wise product. We also attempted using the approach described in [10] to construct additional features (which amount to many higher orders) in our experiments but that did not give further accuracy improvement on the test set due to over-fitting. Using only the first- and second-order features provides a special benefit: We can initialize the CRF parameters by converting from the GMM (HMM) model parameters using the procedure described in [5].

### 3.2. Learning the Final Layer

The final layer of DHCRF is trained to optimize

$$J_2(\boldsymbol{\lambda}^{\mathrm{N}}) = \sum_k \log p\left(w^{(k)} | \boldsymbol{o}^{(k),N}; \boldsymbol{\lambda}^{\mathrm{N}}\right) - \frac{\|\boldsymbol{\lambda}^N\|_1}{\sigma_1} - \frac{\|\boldsymbol{\lambda}^N\|_2^2}{\sigma_2} \tag{7}$$

in the supervised manner, where $w^{(k)}$ is the phoneme or word sequence label for the $k$-th utterance without segmentation information. Following [5][6][7], in the final layer we use

$$\boldsymbol{f}_{w^{"}w'}^{(LM)}(w, s, \boldsymbol{o}) = [\delta(w_{i-1} = w")\delta(w_i = w')]_{\mathrm{i}=1,\cdots,\mathrm{I}} \qquad \forall w^{"}w' \tag{8}$$

$$\boldsymbol{f}_{s"s'}^{(Tr)}(w, s, \boldsymbol{o}) = [\delta(s_{t-1} = s")\delta(s_t = s')]_{\mathrm{t}=1,\cdots,\mathrm{T}} \qquad \forall s", s' \tag{9}$$

$$\boldsymbol{f}_{s'}^{(M1)}(w, s, \boldsymbol{o}) = [\delta(s_t = s')\boldsymbol{o}_t]_{\mathrm{t}=1,\cdots,\mathrm{T}} \qquad \forall s' \tag{10}$$

$$\boldsymbol{f}_{s'}^{(M2)}(w, s, \boldsymbol{o}) = [\delta(s_t = s')\boldsymbol{o}_t \circ \boldsymbol{o}_t]_{\mathrm{t}=1,\cdots,\mathrm{T}} \qquad \forall s' \tag{11}$$

as features, where $\delta(x) = 1$ if $x$ is true, and $\delta(x) = 0$ otherwise. $\boldsymbol{f}_{w^{"}w'}^{(LM)}(w, s, \boldsymbol{o})$ are bi-gram language model (LM) features in which each phoneme or word sequence $w$ is consisted of $I$ phonemes or words, $\boldsymbol{f}_{s"s'}^{(Tr)}(w, s, \boldsymbol{o})$ are state transition features, and $\boldsymbol{f}_{s'}^{(M1)}(w, s, \boldsymbol{o})$ and $f_{s'}^{(M2)}(w, s, \boldsymbol{o})$ are the first- and second-order statistics generated from the observations, respectively.

We have the freedom of selecting the phone units for the final layer. For example, in [5][6][7][8] mono-phone units are used and each mono-phone is further split into three states with 144 total number of states for 48 mono-phones mapped from 61 raw phone-like units defined in TIMIT. However, as demonstrated in [8], the best phone accuracy rate achieved with an HCRF with mono-phone units is 71.7%, which is significantly lower than 73.0% that is achievable with a tri-phone HMM trained using the maximum likelihood criterion.

In this study we have chosen to use three-state tri-phone units: Each phone sequence is first converted into the corresponding tri-phone state sequence (without segmentation information) before it is used as the supervision to train the final layer.

Using tri-phone units significantly increases the total number of classes at the final layer and hence the computational cost needed to train the model parameters. For this reason, we have adopted the strategy used in HMM MMIE training [15] to approximate the partition function. In this strategy, we first train a three-state tri-phone HMM system and generate a rich lattice (denoted by $L$) for each utterance in the training set using the HMM system. We then approximate the partition function using the lattice as

$$z(\boldsymbol{o}^N; \boldsymbol{\lambda}^N) = \sum_{w, s^N \in w} exp\left((\boldsymbol{\lambda}^N)^T \boldsymbol{f}(w, s^N, \boldsymbol{o}^N)\right) \cong \sum_{s^N \in L} exp\left((\boldsymbol{\lambda}^N)^T \boldsymbol{f}(w, s^N, \boldsymbol{o}^N)\right). \tag{12}$$

Summation over $w, s^N \in w$ can be simplified to the summation over $s^N \in L$ since the lattice contains the most probable state sequences of the most probable words.

The model parameters in the final layer are initialized by the state sequence generated from forced alignment using the HMM system. Since the segmentation information is given in the state-level alignment, the HCRF is reduced to a CRF and the model parameter estimation problem becomes convex. The parameters are then fine-tuned by optimizing criterion $J_2$ in Eq. (7). This strategy not only greatly improves the efficacy of the training but also provides a good initialization point for the subsequent HCRF training. One downside of the strategy is that the tri-phone state units used in the HCRF model have to be the same as that used in the HMM and the optimized criterion is just an estimate of the true objective function.

After all layers are trained greedily layer by layer, the parameters are jointly fine-tuned following [17].

## 4. Empirical Evaluation

We evaluated the DHCRF model on the TIMIT phonetic recognition task. Different from the phonetic classification task [5][6][7] for which the HCRF has been successfully applied in the past, the boundaries of segments are unknown in the phonetic recognition task. We followed the standard practice [12] of mapping the 61 TIMIT phones into 48 phones for model training, and of collapsing the 48 phones to 39 phones for evaluation. The training stopping point was determined using the MIT development set [11]. The best model parameters discovered were then used to evaluate the core test set. The training, development, and evaluation sets contain 3696, 400, and 192 utterances or 142,910, 15,334, and 7,333 phone-like units, spoken by 462, 50, and 24 speakers, respectively.

The 39-dimensional acoustic observations used in the experiments contain the 13-dimensional Mel-frequency cepstral coefficient (MFCC) and its first and second derivatives. Each dimension of the observation is then normalized to follow a zero-mean unit-variance Gaussian distribution. Although normalization should not affect the training result in theory, it has been shown to be effective since it eliminates the requirement to tune optimization parameters (e.g., minimum step size) for different dimensions [7][10]. The RPROP [13] algorithm modified with the regularized dual averaging method [21] is used to train each layer of the DHCRF models in this study.

The model estimation problem of the DHCRF is non-convex and so proper initialization is crucial. In our experiments all layers are initialized from the three-state left-to-right tri-phone HMM system trained with maximum likelihood (ML) criterion. The total number of logical tri-phones in the HMM system is 105,986, which are mapped to 3,114 tied physical tri-phones with a total of 916 senones. Each senone has 16 Gaussian components. This same ML trained HMM system is also used as the initialization point for the MMIE trained HMM system and as the basis to generate the state-level alignments and the lattices used for the DHCRF training as explained in Section 3.

Table 1 shows the phone accuracy rate (PAR) on the TIMIT phonetic recognition task using different models. In the DHCRF (1-layer) setup, there is no intermediate layer. The acoustic observation is directly mapped to 916 states (senones in the HMM system) at the final and single layer. In the DHCRF (2-layer) setup there is one intermediate (hidden) layer which has $916 \times 16 = 14656$ states to be comparable to the HMM system. In the DHCRF (3-layer) setup, we added a second intermediate layer with 916 states. The sparseness constraint is enforced in the 2- and 3-layer setting so that the total number of parameters is close to that used in the HMM system. In all the results reported here, we did not concatenate the raw observation features in the higher layers since we observed over-fitting effect by doing so.

From Table 1 we observe that the baseline ML and MMIE trained tri-phone HMM with the same feature achieved 73.0% and 73.3% PAR respectively on the TIMIT phonetic recognition task. These results are better than the MMIE result reported in [15] and large-margin HMM result reported in [16] but worse than 80%, the by far best result on this task, reported in [14] which used a series of highly complex feature processing, speaker adaptation, and discriminative training techniques.

With one-layer DHCRF we can only achieve 71.2% PAR on the core test set. By adding the second layer we significantly improve the PAR to 73.8%. By using a 3-layer DHCRF, we obtain additional but small PAR improvement of 0.3%. The performance saturated with a 4-layer DHCRF. We believe it saturates at the 4th layer instead of higher layers as in [1] because it makes a 1-of-N soft decision at each layer and thus has low representation power. Overall, the 3-layer DHCRF outperforms the MMIE trained HMM system by 0.8% PAR and performs significantly better than the MCRF reported in [2]. The 0.3% gain on the core test set is statistically significant at significance level of 5%.

Table 1: Phone accuracy rate comparisons on the core test set for the TIMIT phonetic recognition task

| Model | DEV (%) | TEST (%) |
|---|---|---|
| HMM-ML | 74.6 | 73.0 |
| HMM-MMIE | 74.7 | 73.3 |
| DHCRF (1-layer) | 72.8 | 71.2 |
| DHCRF (2-layers) | 75.4 | 73.8 |
| DHCRF (3-layers) | 75.5 | 74.1 |

## 5. Conclusions

We have developed and reported the DHCRF, a hierarchical model in which the final layer is an HCRF and the intermediate layers are zero-th order CRFs. It combines the power of sequential labeling at the final layer and the feature extraction and normalization at the intermediate layers. One key issue for DHCRF is the availability of training supervision information at only the final layer. We provide a solution to this difficulty by learning the final layer and intermediate layers in supervised and unsupervised manners, respectively. The model has been evaluated on the TIMIT phonetic recognition task and moderate PAR improvement has been observed against the MMIE trained HMM model.

In the current study, we have only used single frame as the features to be comparable to the HMM systems. We believe the recognition accuracy can be further improved when multiple frames are used or additional features are introduced as indicated in [1][18].

## 6. Acknowledgements

We would like to thank Dr. Asela Gunawardana and Milind Mahajan at Microsoft Corporation for their valuable discussions and useful helps in conducting experiments.

## 7. References

[1] Mohamed, A.-R. Dahl, G., Hinton, G., “Deep Belief Networks for phone recognition”, in NIPS workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

[2] Prabhavalkar, R., Fosler-Lussier, E., “Backpropagation Training For Multilayer Conditional Random Field Based Phone Recognition”, in Proc. of ICASSP 2010, pp. 5534-5537.

[3] Yu, D., Wang, S., karam, Z., Deng, L., “Language Recognition Using Deep-Structured Conditional Random Fields”, in Proc. of ICASSP 2010, pp. 5030-5033.

[4] Yu, D., Wang, S., Deng, L., “Sequential Labeling Using Deep-Structured Conditional Random Fields”, Journal of Selected Topics in Signal Processing -- Special Issue on Statistical Learning Methods for Speech and Language Processing , 2010 (to appear).

[5] Gunawardana, A., Mahajan, M., Acero, A. and Platt, J. C., “Hidden Conditional Random Fields for Phone Classification”, in Proc. of Interspeech 2005, pp. 1117—1120.

[6] Mahajan, M., Gunawardana, A. and Acero, A., “Training Algorithms for Hidden Conditional Random Fields’, in Proc. of ICASSP 2006, vol. I, pp. 273 – 276.

[7] Yu, D., Deng, L., Acero, A., “Hidden Conditional Random Field with Distribution Constraints for Phone Classification,” in Proc. of Interspeech 2009, pp. 676-679.

[8] Sung, Y.-H., Boulis, B., Manning, C. and Jurafsky D., “Regularization, Adaptation, and Non-independent Features Improve Hidden Conditional Random Fields for Phone Classification”, in Proc. of ASRU workshop 2007, pp. 347-352.

[9] Sung, Y.-H. and Jurafsky, D., “Hidden Conditional Random Fields for Phone Recognition,” in Proc. of ASRU workshop 2009, pp. 107-112.

[10] Yu, D., Deng, L., Acero, A., “Using Continuous Features in the Maximum Entropy Model,” Pattern Recognition Letters. Vol. 30, Issue 14, pp. 1295-1300, October, 2009.

[11] Halberstadt, A. K. and Glass, J. R., “Heterogeneous Acoustic Measurements for Phonetic Classification,” in Proc. of Eurospeech, 1997, pp. 401–404.

[12] Lee, K. F. and Hon, H. W., “Speaker Independent Phone Recognition Using Hidden Markov Models,” in Proc. of ICASSP 1980, pp. 1641–1648.

[13] Riedmiller, M. and Braun, H., “A Direct Adaptive Method for Faster Back-Propagation Learning: The RPROP Algorithm”, in Proc. of IEEE ICNN 1993, vol. 1, pp. 586-591.

[14] Sainath, T. N., Ramabhadran, B., and Picheny, M., “An Exploration of Large Vocabulary Tools for Small Vocabulary Phonetic Recognition”, in Proc. of ASRU workshop 2009, pp. 359-364.

[15] Kapadia, S., Valtchev, V. and Young, S. J. “MMI Training for Continuous Phoneme Recognition on the TIMIT Database,” in Proc. ICASSP, 1993, pp. II-491 - II-494.

[16] Sha, F. “Comparison of Large Margin Training to Other Discriminative Training Methods for Phonetic Recognition by Hidden Markov Models,” in Proc. ICASSP, 2007, pp. IV-313 - IV-316.

[17] Yu, D., Deng, L., and Wang, S., “Learning in the Deep-Structured Conditional Random Fields,” NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications, Dec. 2009.

[18] Yu, D., Deng, L., and Acero, A., "Evaluation of a Long-contextual-span Hidden Trajectory Model and Phonetic Recognizer Using A* Lattice Search," in Proc of Interspeech, 2005, pp. 553-556.

[19] Mohamed, A.-R., Yu, D., Deng, L., "Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition", in Proc of Interspeech, 2010.

[20] Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A-R., and Hinton, H., "Binary Coding of Speech Spectrograms Using a Deep Autoencoder", in Proc of Interspeech, 2010.

[21] Xiao, L., "Dual Averaging Method for Regularized Stochastic Learning and Online Optimization", in Proc. NIPS 2009.