Acoustic Modeling Based on Deep Conditional Random Fields

Yasser Hifny YHIFNY@FCI.HELWAN.EDU.EG
University of Helwan, Egypt

Abstract

Acoustic modeling based on Hidden Markov Models (HMMs) is employed by state-of-the-art stochastic speech recognition systems. In continuous density HMMs, the state scores are computed using Gaussian mixture models. On the other hand, Deep Neural Networks (DNN) can be used to compute the HMM state scores. This leads to significant improvement in the recognition accuracy. Conditional Random Fields (CRFs) are undirected graphical models that maintain the Markov properties of Hidden Markov Models (HMMs), formulated using the maximum entropy (MaxEnt) principle. It is possible to use DNN to compute the state scores in CRFs. Using CRFs on the top of DNN will lead to an acoustic model known as Deep Conditional Random Fields (DCRFs). In this paper, we present a phone recognition task based on DCRFs. Preliminary results on the TIMIT task show that DCRFs can lead to good results.

1. Introduction

In hybrid ANN/HMM speech recognition systems (Renals et al., 1994), (Morgan & Bourlard, 1995), Artificial Neural Networks (ANN) models are used as flexible discriminant classifiers to estimate a scaled likelihood. In particular, the emission probability score is given by

$$b_j(\mathbf{o}_t) = \frac{P_\Lambda(\mathbf{s}_j|\mathbf{o}_t)}{P(\mathbf{s}_j)} \quad (1)$$

where $b_j(\mathbf{o}_t)$ is the score of state j in the traditional HMM framework, $P_\Lambda(\mathbf{s}_j|\mathbf{o}_t)$ is the posterior probability of a phonetic state estimated by a connectionist estimator (Trentin & Gori, 2001),(Robinson, 1994) and

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

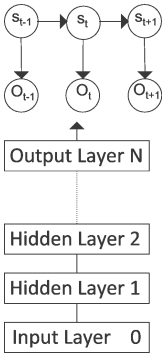

Figure 1. HMM model for phone representation, where the state scores are computed from a DNN.

$P(\mathbf{s}_j)$ is estimated from the labeled data. In addition to discriminative training, if the posterior probability $P_\Lambda(\mathbf{s}_j|\mathbf{o}_t)$ is sensitive to acoustic context, $b_j(\mathbf{o}_t)$ score may help to overcome conditional independence assumption and improve the overall recognition performance without changing the basic HMM framework. HMM is directed graphical model and a graphical representation of the ANN/HMM acoustic model is shown in Figure 1.

DNNs with many hidden layers that are trained using new methods have been shown to outperform Gaussian mixture models in several tasks (Mohamed et al., 2012), (Seide et al., 2011), (Dahl et al., 2012), (Hinton et al., 2012). DNNs are trained in a generative way to learn the structure in the input data. This "pre-training" step provides a good initialization point to the traditional discriminative training using the back-propagation (BP) algorithm. DNN is an active area of research and there is a lot of efforts to improve the training speed of the models (Kingsbury et al., 2012),(Vinyals & Povey, 2012).

Over the last few years, there is an increased interest to develop acoustic models derived from Conditional Random Fields (Lafferty et al., 2001). Hidden Conditional Random Fields(HCRFs) was introduced to score the states based on a mixture of quadratic activation functions (Gunawardana et al.,

2005). In (Yu et al., 2010), a multi-layer CRF model (deep-structured CRF) in which each higher layer's input observation sequence consists of the previous layer's observation sequence and the resulted frame-level marginal probabilities. Deep extensions to HCRFs were developed in (Yu & Deng, 2010),(Mohamed et al., 2010).

In (Hifny, 2006; Hifny & Renals, 2009), a new acoustic modeling paradigm based on Augmented Conditional Random Fields (ACRFs) is investigated and developed. ACRFs paradigm addresses some limitations of HMMs while maintaining many of the aspects which have made them successful. In particular, the acoustic modeling problem is reformulated in a data driven, sparse, augmented space to increase discrimination. Acoustic context modeling is explicitly integrated to handle the sequential phenomena of the speech signal. In the context ANN field , ACRFs can represent CRFs with one hidden layer constructed from scoring a large number of Gaussians.

Training CRFs on the top of a hidden layer constructed from scoring a large number of sigmoid functions was introduced in (Prabhavalkar & Fosler-Lussier, 2010). One way to improve this approach is the compute the state scores based on a DNN that has many hidden layers. Hence, this improvement will lead to a deep version of CRFs (DCRFs). In this paper, we present a phone recognition task based on DCRFs. Preliminary results on the TIMIT task show that DCRFs can lead to good results.

In Section 2, a mathematical formulation of DCRFs is described. The optimization problem of DCRFs is addressed in Section 3. Section 4 gives experimental results on a phone recognition task. Several issues about the implementation of DCRFs are discussed in Section 5. Finally, a summary of the presented work is given in the conclusions.

2. Deep Conditional Random Fields

Linear chain CRFs can be thought as the undirected graphical twins for HMMs regardless of their training (generative or discriminative) (Lafferty et al., 2001). DCRF acoustic models are a particular implementation of linear chain CRFs where the state scores are computed based on a DNN that has many hidden layers. The feed-forward phase updates the output value of each neuron. Starting from the first hidden layer, each neuron output is computed as a weighted sum of inputs and applying the sigmoid function to it:[1]

$$\mathbf{o}_{tj}^{h} = \mathrm{sigm}(\sum_{i=1}^{n} \lambda_{ij} \mathbf{o}_{ti}^{h-1}) \tag{2}$$

where $\mathbf{o}_t^h$ is an output of a hidden layer, n is the number of inputs, h is an index to a hidden layer, and sigmoid function is computed as follows:

$$\mathrm{sigm}(x) = \frac{1}{1+e^{-x}} \tag{3}$$

The output of an hidden layer is passed to the next layer until the output layer is computed as follows:

$$\mathbf{o}_{tj}^{N} = \sum_{i=1}^{n} \lambda_{ij} \mathbf{o}_{ti}^{N-1} \tag{4}$$

where N is index of the output layer. Hence, the activation of hidden layers is nonlinear based on a sigmoid function and the output layer activation is linear.

A graphical representation of the DCRF acoustic model is shown in Figure 2. The conditional distribution defining DCRFs is given by

$$P_\Lambda(\mathbf{S}|\mathbf{O}) = \frac{1}{Z_\Lambda(\mathbf{O})} \prod_{t=1}^{T} \exp\left(\lambda_{\mathbf{s}_t \mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + b_{\mathbf{s}_t}(\mathbf{o}_t)\right) \tag{5}$$

where

- $P_\Lambda(\mathbf{S}|\mathbf{O})$ obeys the Markovian property:

$$P_\Lambda(\mathbf{s}_t|\{\mathbf{s}_j\}_{j\neq t}, \mathbf{O}) = P_\Lambda(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{O})$$

- $\lambda_{\mathbf{s}_t \mathbf{s}_{t-1}}$ are associated with the characterizing function $a(\mathbf{s}_t, \mathbf{s}_{t-1})$. $a(\mathbf{s}_t, \mathbf{s}_{t-1})$ is a binary function and can be used to define DCRF topology.

- $b_{\mathbf{s}_t}(\mathbf{o}_t) = \mathbf{o}_{ts_t}^N$ is computed from Equation (4). Hence, $b_{\mathbf{s}_t}(\mathbf{o}_t)$ connects DNN output to CRF input.

- $Z_\Lambda(\mathbf{O})$ (Zustandsumme) is a normalization coefficient referred to as the partition function.

HMMs and DCRFs (in general, linear chain CRFs) share the first order Markov assumption, which simplifies the training and decoding algorithms. However, DCRFs do not assume observation independence and causality, as the joint event in this case is factorized as a simple product of exponential functions. Therefore, the observations and the characterizing functions

[1] The bias term is not implemented in neuron activation.

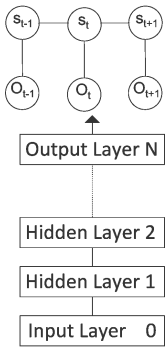

Figure 2. DCRF model for phone representation, where the state scores are computed from a DNN.

can be statistically dependent or correlated and can depend on the past and future acoustic context. The partition function, $Z_\Lambda(\mathbf{O})$, is given by

$$Z_\Lambda(\mathbf{O}) = \sum_{\mathbf{S}} \prod_{t=1}^{T} \exp\Big(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + b_{\mathbf{s}_t}(\mathbf{o}_t)\Big), \quad (6)$$

and it is similar to the total probability $p(\mathbf{O}|\mathcal{M})$ in HMMs, which can be calculated using the forward algorithm (Lafferty et al., 2001).

3. DCRF Optimization

For R training observations $\{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_r, \ldots, \mathbf{O}_R\}$ with corresponding transcriptions $\{W_r\}$, DCRFs are trained using the conditional maximum likelihood (CML) criterion to maximize the posterior probability of the correct word sequence given the acoustic observations:

$$\begin{aligned}\mathcal{F}_{\mathrm{CML}}(\Lambda) &= \sum_{r=1}^{R} \log P_\Lambda(\mathcal{M}_{W_r}|\mathbf{O}_r) \\ &= \sum_{r=1}^{R} \log \frac{P(W_r) \sum_{\mathbf{S}|W_r} \exp \sum_t^T \Psi(\mathbf{O}, \mathbf{S}, c, \Lambda)}{\sum_{\hat{W}} P(\hat{W}) \sum_{\mathbf{S}|\hat{W}} \exp \sum_t^T \Psi(\mathbf{O}, \mathbf{S}, c, \Lambda)} \\ &\approx \sum_{r=1}^{R} \log Z_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{num}}) - \log Z_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{den}}), \quad (7)\end{aligned}$$

where

$$\Psi(\mathbf{O}, \mathbf{S}, c, \Lambda) = \lambda_{\mathbf{s}_t\mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + b_{\mathbf{s}_t}(\mathbf{o}_t) \quad (8)$$

The optimal parameters, Λ^*, are estimated by maximizing the CML criterion, which implies minimizing the cross entropy between the correct transcription model and the hypothesized recognition model. In other words, the process maximizes the partition function of the correct models[2] (the numerator term) $Z_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{num}})$, and simultaneously minimizes the partition function of the recognition model (the denominator term) $Z_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{den}})$. The optimal parameters are obtained when the gradient of the CML criterion is zero.

3.1. Numerical Optimization for DCRFs

DCRFs are can be trained using gradient based approaches. These methods rely on a locally linear or quadratic approximation by expanding the CML nonlinear objective function $\mathcal{F}_{\mathrm{CML}}(\Lambda + \delta)$ using Taylor's expansion around the current model point Λ in the parameter space (Nocedal & Wright, 1999). Such approaches are well established in artificial neural networks research (Bishop, 1995; Haykin, 1998). For example, the CRF training process has been accelerated by using a stochastic meta-descent algorithm which utilizes second-order information to adapt the gradient step sizes (Vishwanathan et al., 2006).

For an e-family activation function based on first-order sufficient statistics, the gradient of the CML objective function for the output layer parameters is given by

$$\nabla\mathcal{F}_{\mathrm{CML}}(\mathbf{O}) = \mathcal{C}_{ji}^{\mathrm{num}}(\mathbf{O}) - \mathcal{C}_{ji}^{\mathrm{den}}(\mathbf{O}) \quad (9)$$

where the accumulators of the sufficient statistics, $\mathcal{C}_{ji}(\mathbf{O})$, for the j^{th} state and i^{th} constraint are calculated as follows:

$$\mathcal{C}_{ji}^{\mathrm{num}}(\mathbf{O}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_j^r(t|\mathcal{M}^{\mathrm{num}}) \mathbf{o}_{rti}^{\mathrm{N}} \quad (10)$$

$$\mathcal{C}_{ji}^{\mathrm{den}}(\mathbf{O}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_j^r(t|\mathcal{M}^{\mathrm{den}}) \mathbf{o}_{rti}^{\mathrm{N}} \quad (11)$$

where r is the utterance index and the frame-state alignment probability γ_j, denoting the probability of being in state j at some time t can be written in terms of the forward score $\alpha_j(t)$ and the backward score $\beta_j(t)$ as in HMMs:

$$\gamma_j(t|\mathcal{M}) = P(\mathbf{s}_t = j|\mathbf{O}; \mathcal{M}) = \frac{\alpha_j(t|\mathcal{M})\beta_j(t|\mathcal{M})}{Z_\Lambda(\mathbf{O}|\mathcal{M})} \quad (12)$$

and to avoid the necessity of building lattices, the $\gamma_j(t|\mathcal{M})$ is approximated with state estimates as follows (Hifny et al., 2005):

$$\gamma_j(t|\mathcal{M}^{\mathrm{den}}) = \frac{\exp\left(\mathbf{o}_{tj}^N\right)}{\sum_{\mathbf{s}} \exp\left(\mathbf{o}_{ts}^N\right)} \quad (13)$$

[2]Since a summation over potential functions is commonly called the partition function in undirected graphical modeling, we coin the notation $Z_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{num}})$ for the summation of all possible state sequences of the correct models.

The delta of the output layer neuron j is given by

$$\delta_{tj}^{N} = \gamma_j(t|\mathcal{M}^{\text{num}}) - \gamma_j(t|\mathcal{M}^{\text{den}}) \tag{14}$$

and the delta of the hidden layers:

$$\delta_{tj}^{h} = \mathbf{o}_{tj}^{h}(1 - \mathbf{o}_{tj}^{h}) \sum_{k \in \text{outputs}} \lambda_{kj}^{h+1} \delta_{kt}^{h+1} \tag{15}$$

and the gradient for the hidden layers parameters is given by:

$$\frac{\partial \mathcal{F}_{\text{CML}}(\Lambda)}{\partial \lambda_{ki}^{h}} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \delta_{rtj}^{h} \mathbf{o}_{rtki}^{h-1} \tag{16}$$

Based on equation (16) and equation (9), a gradient based optimization can be used to estimate the parameters (Nocedal & Wright, 1999). The transition parameters are given by:

$$\lambda_{\mathbf{s}_t \mathbf{s}_{t-1}} = \log a_{\mathbf{s}_t \mathbf{s}_{t-1}}, \tag{17}$$

where $a_{\mathbf{s}_t \mathbf{s}_{t-1}}$ is the transition probability in HMM modeling and is estimated using the maximum likelihood (MLE) criterion.

4. Experiments

We have carried out phone recognition experiments on the TIMIT corpus (Garofolo et al., 1990). We used the 462 speaker training set and testing on the 24 speaker core test set (the SA1 and SA2 utterances were not used). The speech was analyzed using a 25ms Hamming window with a 10 ms fixed frame rate. In all the experiments we represented the speech using 12th order mel frequency cepstral coefficients (MFCCs), energy, along with their first and second temporal derivatives, resulting in a 39 element feature vector. The training data and test data features are pre-processed to have zero mean and unit variance. Hence, acoustic context information is integrated using a window of 9 frames (4 left + current frame+ 4 right) to construct the final frame vector with 351 dimensions.

Following Lee (Lee & Hon, 1989), the original 61 phone classes in TIMIT were mapped to a set of 48 labels, which were used for training. This set of 48 phone classes was mapped down to a set of 39 classes (Lee & Hon, 1989), after decoding, and phone recognition results are reported on these classes, in terms of the phone error rate (PER), which is analogous to word error rate.

The baseline HMMs have three emitting states and the emission probabilities were modeled with mixtures of Gaussian densities with diagonal covariance matrices. The generative HMMs were trained by the maximum likelihood criterion using the conventional EM algorithm (Young et al., 2001).

A DNN with 2 hidden layers was chosen and each layer has 512 neurons. Hence, each phone was represented using a three state left-to-right DCRF, all parameters of DNN were initialized to random values and the transition parameters were initialized either from trained HMM models forcing left to right DCRFs. The training procedure accumulated the $\mathcal{M}^{\text{num}}$ sufficient statistics via a Viterbi pass (forced alignment) of the reference transcription using HMMs trained using maximum likelihood criterion. Several iterations were used to train DCRFs and the language model scaling factor is set to 6.0 during the decoding process. All our experiments used a bigram language model over phones, estimated from the training set.

DNN parameter estimation is based on a variant of the Resilient Propagation (RProp) algorithm (Riedmiller & Braun, 1993), which uses a Manhattan update rule. The Manhattan update rule does not involve the gradient magnitude. Th algorithm is detailed in (Hifny, 2013) and it was shown in (Hifny, 2006) that this algorithm outperforms other gradient based algorithms for CRF parameter estimation.

In Table 1, DCRFs recognition performance is reported in terms of PER on TIMIT task (core test set). The training process is divided into two phases online and batch. The online training phase computes an initial model that used to start the batch training phase. A complete iteration of the online Manhattan update implies ten loops over the training data. The phone error rate is reduced to 42.2 % after the online training is complete. Hence, the final model of the online training phase is used to initialize a model to start the batch training phase. A batch Manhattan update is used to update the models. As shown in the results, the recognition accuracy improves when the number of iteration is high. The batch training phase was executed over a computer grid of 28 cores.

5. Discussions

In this section we address several issues about the implementation of DCRFs.

5.1. Training criterion

In the traditional CRFs, the conditional maximum likelihood (CML) criterion is used to maximize the posterior probability of the correct word sequence given the acoustic observations. However, in Section 3, the $\gamma_j(t|\mathcal{M})$ is approximated with state estimates

Table 1. DCRF decoding results on TIMIT recognition task in terms of PER.

Training Method	#Itr	PER
Online Manhattan	1	42.2%
Batch Manhattan	50	36.9%
	100	35.5%
	150	34.7%
	200	34.4%
	250	34.2%
	300	33.7%
	350	33.6%
	400	33.4%
	450	33.2%
	500	33.2%
	550	33.1%
	600	33.0%

as shown in (Hifny et al., 2005). Hence, the effective training criterion used to train DCRFs is frame level conditional maximum likelihood (CML) criterion. In addition, maximizing frame level CML is equivalent to minimizing the frame level cross-entropy loss. This criterion is identical to the training criterion used to train traditional DNN in ANN/HMM hybrid systems.

5.2. Decoding speed

In hybrid ANN/HMM speech recognition systems, the HMM state scores are computed based on equation (1). This equation implies the calculations of a softmax activation function for each frame to compute the state posteriors. On the other hand, DCRFs state scores are based on a linear activation function in the output layer. Hence, a softmax activation function is not used in DCRFs decoding. Consequently, DCRFs decoders are running faster than traditional DNN/HMM decoders.

5.3. Prior work

Training CRFs on the top of a single hidden layer constructed from scoring a large number of sigmoid functions was introduced in (Prabhavalkar & Fosler-Lussier, 2010). In (Mohamed et al., 2010), state scores are computed based on DNN setup but the output layer has a softmax activation function. In this work, the state scores are also computed based DNN architecture but the output layer has a linear activation function. In addition, we do not estimate state transition parameters or language model parameters within DCRF framework. The state transition parameters were estimated using traditional HMM framework. In addition, Maximum Likelihood (ML) criterion is used to estimate bigram language model. Hence, DCRF architecture may be computationally efficient for training and decoding. During the decoding process, a language model scaling factor is used improve the results. On the other hand, frame level CML criterion is used to estimate DCRFs rather than the full-sequence training.

6. Conclusions

In this paper, we present a method to construct deep conditional random fields. In this approach, the state scores are computed based on a DNN that has many hidden layers. The feed-forward phase updates the output value of each neuron. Starting from the first hidden layer, each neuron output is computed as a weighted sum of inputs and applying the sigmoid function to it. The output is forwarded to the next layer until the output layer is updated as a weighted sum of inputs. DCRF state scores are connects the DNN output layer. Hence, the gradient is computed and a back-propagation algorithm is used to compute the gradient of each parameter in the hidden layers. Any gradient based optimization technique can be used estimate the parameters. Preliminary results on the TIMIT phone recognition task show that DCRFs can lead to good results. DCRFs parameter estimation is slow. Future work will focus on tuning the learning parameters , the number of neurons per hidden layer, and the number of hidden layer.

References

Bishop, Christopher M. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

Dahl, George, Yu, Dong, Deng, Li, and Acero, Alex. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on Deep Learning for Speech and Langauge Processing*, 2012.

Garofolo, John S., Lamel, Lori F., Fisher, William M., Fiscus, Jonathan G., Pallett, David S., Dahlgren, Nancy L., and Zue, Victor. TIMIT acoustic-phonetic continuous speech corpus, 1990. URL http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

Gunawardana, A., Mahajan, M., Acero, A., and Platt, J. Hidden conditional random fields for phone classification. In *Proc. INTERSPEECH*, pp. 1117–1120, Lisbon, Portugal, 2005.

Haykin, Simon. *Neural Networks: A Comprehensive Foundation*. Prentice Hal, 2nd edition, 1998.

Hifny, Yasser. *Conditional Random Fields for Continuous Speech Recognition*. PhD thesis, University Of Sheffield, 2006.

Hifny, Yasser. Deep learning using a Manhattan update rule. *Deep Learning for Audio, Speech and Language Processing, ICML*, 2013.

Hifny, Yasser and Renals, Steve. Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech and Language Processing*, 17(2):354–365, 2009.

Hifny, Yasser, Renals, Steve, and Lawrence, Neil. A hybrid MaxEnt/HMM based ASR system. In *Proc. INTERSPEECH*, pp. 3017–3020, Lisbon, Portugal, 2005.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George, rahman Mohamed, Abdel, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara, , and Kingsbury, Brian. Deep Neural Networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 2012.

Kingsbury, Brian, Sainath, Tara N., and Soltau, Hagen. Scalable minimum bayes risk training of Deep Neural Network acoustic models using distributed hessian-free optimization. In *interspeech*, 2012.

Lafferty, John, McCallum, Andrew, and Pereira, Fernando. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pp. 282–289, 2001.

Lee, K.-F. and Hon, H.-W. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 37(11):1641–1648, Nov 1989.

Mohamed, Abdel-rahman, Yu, Dong, and Deng, Li. Investigation of full-sequence training of Deep Belief Networks for speech recognition. In *Interspeech*, 2010.

Mohamed, Abdel-rahman, Dahl, George, and Hinton, Geoffrey. Acoustic modeling using Deep Belief Networks. *IEEE Transactions on Audio, Speech and Language Processing*, 20:14–22, 2012.

Morgan, N. and Bourlard, H. Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995.

Nocedal, Jorge and Wright, Stephen J. *Numerical Optimization*. Springer, 1999.

Prabhavalkar, R. and Fosler-Lussier, E. Backpropagation training for multilayer conditional random field based phone recognition. In *Proc. IEEE ICASSP*, volume 1, pp. 5534 – 5537, France, March 2010.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. Connectionist probability estimators in hmm speech recognition. *IEEE Transactions on Speech and Audio Processing*, January 1994.

Riedmiller, M. and Braun, H. A direct method for faster backpropagation learning: The RPROP algorithm. In *Proc. IEEE International Conference on Neural Networks*, pp. 586–591, 1993.

Robinson, A. An application of recurrent neural nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.

Seide, F., Li, G., and ., D. Yu. Conversational speech transcription using context-dependent Deep Neural Networks. In *Interspeech*, 2011.

Trentin, E. and Gori, M. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, April 2001.

Vinyals, Oriol and Povey, D. Krylov subspace descent for deep learning. In *AISTATS*, 2012.

Vishwanathan, S. V. N., Schraudolph, Nicol N., Schmidt, Mark W., and Murphy, Kevin P. Accelerated training of conditional random fields with stochastic gradient methods. In *Proc. ICML*, pp. 969–976, 2006. ISBN 1-59593-383-2. doi: http://doi.acm.org/10.1145/1143844.1143966.

Young, Steve, Kershaw, Dan, Odell, Julian, Ollason, Dave, Valtchev, Valtcho, and Woodland, Phil. *The HTK Book, Version 3.1*. 2001.

Yu, Dong and Deng, Li. Deep-structured hidden conditional random fields for phonetic recognition. In *Proc. INTERSPEECH*, 2010.

Yu, Dong, Wang, Shizhen, and Deng, Li. Sequential labeling using deep-structured conditional random fields. *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, 2010.