LEARNING ACOUSTIC FRAME LABELING FOR SPEECH RECOGNITION WITH RECURRENT NEURAL NETWORKS

Haşim Sak, Andrew Senior, Kanishka Rao, Ozan İrsoy, Alex Graves, Françoise Beaufays, Johan Schalkwyk

Google

{hasim,andrewsenior,kanishkarao,gravesa,fsb,johans}@google.com

ABSTRACT

We explore alternative acoustic modeling techniques for large vocabulary speech recognition using Long Short-Term Memory recurrent neural networks. For an acoustic frame labeling task, we compare the conventional approach of cross-entropy (CE) training using fixed forced-alignments of frames and labels, with the Connectionist Temporal Classification (CTC) method proposed for labeling unsegmented sequence data. We demonstrate that the latter can be implemented with finite state transducers. We experiment with phones and context dependent HMM states as acoustic modeling units. We also investigate the effect of context in acoustic input by training unidirectional and bidirectional LSTM RNN models. We show that a bidirectional LSTM RNN CTC model using phone units can perform as well as an LSTM RNN model trained with CE using HMM state alignments. Finally, we also show the effect of sequence discriminative training on these models and show the first results for sMBR training of CTC models.

Index Terms— LSTM, CTC, RNN, acoustic modeling.

1. INTRODUCTION

Acoustic modeling with DNNs and RNNs has commonly used the hybrid approach [1], where the neural networks as discriminative models estimate the posterior probabilities of phonetic states — most commonly hidden Markov model (HMM) states. The models are generally first trained using fixed alignments as targets (acoustic frames with the corresponding HMM state labels). These alignments are often obtained from the forced-alignment of the supervised transcript with the acoustic frames using a GMM (Gaussian mixture model)-HMM and can be further refined by realigning with a fully trained neural network and then by retraining the network with the new target alignments. The cross-entropy frame-level loss function is commonly used with a softmax output layer for the labeling of acoustic frames. For speech decoding, the phonetic state posteriors for each acoustic frame are scaled with the state priors to obtain acoustic likelihood scores which are combined with the language model probabilities for phonetic state sequences. LSTM RNNs with the hybrid approach have been recently shown to outperform the state of the art DNNs for acoustic modeling in large vocabulary speech recognition [2, 3, 4, 5].

In this paper, we explore alternative acoustic modeling techniques using LSTM RNNs for large vocabulary speech recognition (described in Section 2). We compare the conventional approach of cross-entropy (CE) training using fixed forced-alignments, with the Connectionist Temporal Classification (CTC) approach [6] which has previously obtained the state of the art results for phoneme recognition on the TIMIT task with deep bidirectional LSTM RNNs [7]. We show the CTC realignment procedure can be easily implemented in finite-state transducer (FST) framework and explain how CTC models can be used in decoding (Section 2.2). We also describe the use of sequence discriminative training with our sequence models (Section 2.3). In Section 4, we describe experiments with two acoustic modeling units – phones and HMM states. We also investigate the effect of acoustic context for LSTM RNN acoustic models by training unidirectional and bidirectional models.

2. ACOUSTIC MODELING WITH LSTM RNN

There are a number of alternative approaches for acoustic modeling with neural networks for automatic speech recognition (ASR). Fundamentally the unit to be modeled by the network must be chosen (e.g. phone, HMM state, context dependent HMM state, diphone, word etc.). Training may use a hard (Viterbi) alignment with a single class label per frame, or a soft (Baum-Welch) alignment giving a probability distribution. Further, a variety of objective functions such as frame discriminative CE or sequence discriminative training criteria may be used. We examine all of these factors in the following sections.

2.1. Cross Entropy Training with Fixed Alignments

Let $\boldsymbol{x} = x_1, \ldots, x_T$ denote a sequence of T acoustic feature vectors $x_t \in \mathbb{R}$ for an utterance and $\boldsymbol{w}$ a word sequence. According to the HMM assumption, the acoustic data likelihood is decomposed as follows (using the Viterbi approximation):

$$p(\boldsymbol{x}|\boldsymbol{w}) = \prod_{t=1}^{T} p(x_t|l_t)p(l_t|l_{t-1}),$$

where $l_1, \ldots, l_T$ is the label sequence computed by forced alignment of the utterance with the word sequence $\boldsymbol{w}$. In the hybrid modeling approach, the emission probability is represented as $p(x_t|l_t) = p(l_t|x_t)p(x_t)/p(l_t)$. The label posterior can be modeled by a DNN [8, 9, 10, 11, 12] over asymmetrically windowed acoustic frames, a unidirectional LSTM RNN estimating $p(l_t|x_1^t)$ or a bidirectional LSTM RNN estimating $p(l_t|\boldsymbol{x})$. The label prior $p(l_t)$ is the relative label frequency as observed in the alignments. The data likelihood $p(x_t)$ does not depend on labels and thus can be ignored for decoding/lattice generation and forced alignment [1].

We assume that alignments are available (generated with an existing model such as GMM-HMM or neural network) and fixed. Then, the neural network parameters can be estimated to maximize the CE loss on all acoustic frames of input utterances $\boldsymbol{x}$ with a corresponding frame level alignment $\boldsymbol{l}$ ($|\boldsymbol{x}| = |\boldsymbol{l}|$).

$$\mathcal{L}_{CE} = -\sum_{(\boldsymbol{x},\boldsymbol{l})} \sum_{t=1}^{|\boldsymbol{x}|} \sum_{l} \delta(l, \boldsymbol{l}_t) \log y_l^t. \quad (1)$$

where $\delta(l, \boldsymbol{l}_t)$ is the Kronecker delta and y_l^t is the network output activations. The network output activations for each label and frame, y_l^t as computed with a softmax output layer can estimate the label posteriors $p(l|\boldsymbol{x})$. The gradient of the loss function for each training example $\mathcal{L}(\boldsymbol{x}, \boldsymbol{l})$ with respect to input activations a_l^t of the softmax output layer can be computed as follows:

$$\frac{\partial \mathcal{L}(\boldsymbol{x}, \boldsymbol{l})}{\partial a_l^t} = y_l^t - \delta(l, \boldsymbol{l}_t) \tag{2}$$

Hence, with the CE criterion and softmax activation, the stochastic gradient descent method optimizes the network parameters to make the network predictions y_l^t match the target hard alignments $\delta(l, \boldsymbol{l}_t)$.

2.2. Learning Acoustic Frame Labeling with Connectionist Temporal Classification

The connectionist temporal classification (CTC) [6] approach is a learning technique for sequence labeling using RNNs where the alignment between the inputs and target labels is unknown. CTC can be implemented as a softmax output layer with an additional unit for the *blank* label used to estimate the probability of outputting no label at a given time. Therefore, the output label probabilities from the network define a probability distribution over all possible labelings of input sequences including the blank labels. The network can be trained to optimize the total log probability of correct labelings for training data as estimated using the network outputs and forward-backward algorithm [13]. The correct labelings for an input sequence are defined as the set of all possible labelings of the input with the target labels in the correct sequence possibly with repetitions and with blank labels permitted between separate labels. Hence, CTC differs from the conventional framewise labeling in two ways. First, the additionally introduced *blank* label relieves the network from making label predictions at a frame when it is uncertain. Second, the training criterion optimizes the log probability of state sequences rather than the log likelihood of inputs.

The CTC loss function is defined as the sum of negative log probability of correct labelings for each training example:

$$\mathcal{L}_{CTC} = -\sum_{(\boldsymbol{x}, \boldsymbol{l})} \ln p(\boldsymbol{z}^{\boldsymbol{l}}|\boldsymbol{x}) = -\sum_{(\boldsymbol{x}, \boldsymbol{l})} \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}^{\boldsymbol{l}}) \tag{3}$$

where $\boldsymbol{x}$ is the input sequence of acoustic frames, $\boldsymbol{l}$ is the input label sequence (e.g. phonetic transcription for the utterance), $\boldsymbol{z}^{\boldsymbol{l}}$ is the lattice encoding all possible alignments of $\boldsymbol{x}$ with $\boldsymbol{l}$ which allows label repetitions possibly interleaved with *blank* labels. The probability for correct labelings $p(\boldsymbol{z}^{\boldsymbol{l}}|\boldsymbol{x})$ can be computed using the forward and backward variables estimated with the forward-backward algorithm:

$$p(\boldsymbol{z}^{\boldsymbol{l}}|\boldsymbol{x}) = \sum_{u=1}^{|\boldsymbol{z}^{\boldsymbol{l}}|} \alpha_{x,z^l}(t,u)\beta_{x,z^l}(t,u) \tag{4}$$

where $\alpha_{x,z^l}(t,u)$ is the forward variable representing the summed probability of all paths in the lattice $\boldsymbol{z}^{\boldsymbol{l}}$ starting in the initial state at time 0 and ending in state u at time t, $\beta(t,u)$ is the backward variable representing the summed probability of all paths starting in state u of the lattice at time t and going to a final state, and t can be any time step. The gradient of the loss function with respect to input activations a_l^t of the softmax output layer for a training example can be computed as follows:

$$\frac{\partial \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}^{\boldsymbol{l}})}{\partial a_l^t} = y_l^t - \frac{1}{p(\boldsymbol{z}^{\boldsymbol{l}}|\boldsymbol{x})} \sum_{u \in \left\{u:\boldsymbol{z}_u^{\boldsymbol{l}}=l\right\}} \alpha_{x,z^l}(t,u)\beta_{x,z^l}(t,u) \tag{5}$$

where y_l^t is the network softmax output activation for a label l at time step t, and u represents the lattice states aligned with label l at time t. Contrasting the gradient of the CTC loss with the CE loss of hard alignments, we can see that the CTC computes the CE loss with a soft target alignment computed over a lattice of all possible alignments with the forward-backward algorithm.

The gradients for the CTC output layer can be efficiently and easily computed using finite state transducers (FSTs). In the forward pass of the neural network training, we estimate the label posteriors (including the *blank*) for each acoustic frame of an input utterance using the softmax activations of the CTC output layer. In the backward pass, we calculate the gradient of the loss function with respect to input activations to the output layer by computing the forward and backward variables using the shortest path algorithm on an FST representation of all possible alignments for an input utterance. To build the FST representation of possible alignments, we construct a transducer S encoding the label and its posterior as an arc for each time step (hence the sum of the arc probabilities of a path gives the probability of a label sequence $p(\boldsymbol{l}|\boldsymbol{x})$). Then, we compose this score transducer S with a transducer encoding all valid label sequences L. For CTC, L can be built by composing the string transducer for the target phone sequence of an utterance with a simple transducer C converting phone sequences to sequences of phones with repetitions interleaved with optionally repeated *blank* labels (e.g. *abc* $\rightarrow$ *blank blank a a b blank blank c c c blank*).

For decoding with the CTC models using the standard beam search algorithm, we build the search graph by simply composing the language model G, the lexicon L and the CTC transducer C, $C \circ L \circ G$. For decoding, the CTC C transducer is a single state FST mapping labels to itself and blank to epsilon. Then, the decoder can handle observing the same label multiple times without explicit label repetitions in the search graph similar to handling HMM states. The output label predictions from the CTC model is very spiky due to blank label predicted for about 90% of frames. For this reason, in contrast to conventional models, the phone label posteriors from the CTC model do not require normalization with respect to language model scores. In decoding, we only divide the blank label posterior by a constant value (9) which adds a cost for not outputting any labels, other label posteriors are directly used in decoding.

2.3. Sequence Discriminative Training

Cross-entropy as a framewise discriminative training criterion is suboptimal for word error rate (WER) minimization objective in ASR, since it does not consider the lexical and language model constraints used in speech decoding. A number of sequence-level discriminative training criteria have been proposed in the literature to address this, including maximum mutual information (MMI) [14], minimum phone error (MPE) [15]. Sequence discriminative training has been shown to improve performance of DNN and LSTM RNN acoustic models bootstrapped with cross-entropy training [16, 17, 18, 19, 4].

In this paper, we use state-level minimum Bayes risk (sMBR) criterion [16] for sequence discriminative training of LSTM RNN acoustic models first trained with CE or CTC loss functions. The gradient of the sMBR criterion can be estimated from the state occupancy counts of the numerator lattice (generated by forced alignment with the transcript truth) and denominator lattice (generated by decoding with a weak language model) for an utterance using the forward-backward algorithm and can be efficiently computed using the shortest path algorithm [20].

3. LSTM RNN ARCHITECTURES & TRAINING

RNNs can be unidirectional or bidirectional models in terms of modeling over the input context [21]. Unidirectional RNNs estimate the label posteriors $y_l^t = p(l_t|x_t, \overrightarrow{h_{t-1}})$ using only left context of the current input x_t by processing the input from left to right and having a hidden state $\overrightarrow{h_t}$ in the forward direction. This is desirable for applications requiring low latency. One common way to accomplish performance improvement without degrading the latency much is giving the network limited access to the right/future context by having delayed output targets in training. If one can afford the latency of seeing all inputs, bidirectional RNNs can naturally estimate the label posteriors $p(l_t|x_t, \overrightarrow{h_{t-1}}, \overleftarrow{h_{t+1}})$ using separate layers for processing the input in the forward and backward directions. We use deep LSTM RNN architectures simply built by stacking multiple LSTM layers, which have been shown to perform better than shallow models for speech recognition [22, 7, 2, 3]. For bidirectional models, we use two LSTM layers at each depth — one operating in the forward and another operating in the backward direction over the input sequence. Both of these layers are connected to both the forward and backward layers above. The output layer is also connected to both of the final forward and backward layers. The labels used for acoustic frame labeling determine the acoustic modeling units used in speech recognition. We experiment with context dependent HMM states and phones for different acoustic modeling techniques and LSTM RNN architectures. We train the models in a distributed manner using asynchronous stochastic gradient descent (ASGD) optimization technique allowing parallelization of training over a large number of machines on a cluster and enabling large scale training of neural networks [23, 24, 18, 25, 3].

4. EXPERIMENTS

We evaluate the performance of various LSTM RNN acoustic models on a large vocabulary speech recognition task.

4.1. Systems & Evaluation

All the LSTM networks are trained on a 5 million utterance dataset consisting of anonymized and hand-transcribed utterances. The input to the LSTM RNNs is the 40-dimensional log mel filterbank energy features computed every 10ms, with no frame stacking. For the LSTM models requiring an input alignment, the utterances are aligned with an 85 million parameter DNN with 13522 CD HMM states. The weights in all the networks are randomly initialized with a uniform (-0.02, 0.02) distribution. We clip the activations of memory cells to range [-50, 50], and their gradients to [-1, 1] This makes training with CTC models stable, without truncating the errors. The trained models are evaluated in a large vocabulary speech recognition system on a test set of 22,500 hand-transcribed, anonymized utterances. For all the decoding experiments, we use a wide beam to avoid search errors. After a first pass of decoding using the LSTM RNNs with a 5-gram language model heavily pruned to 23 million n-grams, lattices are rescored using a 5-gram language model with 15 billion n-grams. We use an output delay of 5 frames for the unidirectional models trained with the CE criterion using the fixed alignments. The delay is not needed for the CTC or bidirectional models. For bidirectional CTC models, we obtained the best results with LSTM layers of depth 5 with forward and backward layers having 300 memory cells at each depth and for unidirectional CTC models, with 4 LSTM layers of 500 memory cells. For the other models, we got the best results with 2 LSTM layers of 800 cells each with a recurrent projection layer of 512 units.

Alignment	Label	Context	CE (%)	+sMBR (%)
Fixed	phone	Uni	13.2	-
Fixed	phone	Bi	11.0	-
Fixed	CD state	Uni	10.0	8.9
Fixed	CD state	Bi	9.7	9.1
CTC	phone	Uni	10.5	9.4
CTC	phone	Bi	9.5	8.5

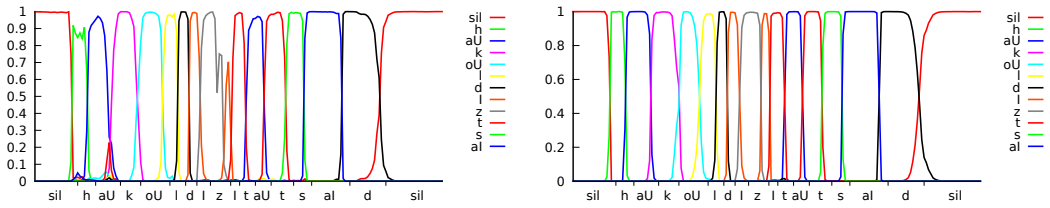
Table 1: *LSTM RNN acoustic models.*

4.2. Results & Conclusions

Table 1 shows the word error rates (WERs) of various LSTM RNN acoustic models on the voice search task. Our best CE + sMBR trained DNN model used to obtain initial alignments gives 10.1% on this test set. The alignment *Fixed* refers to the models trained with CE using the fixed alignments, while *CTC* refers to the CTC models constantly realigning the data during training. The context *Uni* refers to the unidirectional models, while *Bi* refers to the bidirectional models. We report the WERs both after the *CE* training and after the *sMBR* training which always starts after convergence of CE training. The bidirectional CTC phone model performs better than both phone and context dependent HMM state models trained on fixed alignments. The unidirectional CTC phone model is significantly better than phone model trained on fixed alignments. However, it does not perform as good as the context dependent HMM state model. The use of bidirectional context makes a significant improvement for the CTC phone model, while it is much less significant for the CE trained model using the context dependent HMM states. Sequence discriminative training with sMBR criterion improves all the models significantly, but it is unexpected that after sequence training the bidirectional LSTM with HMM states does not perform better than the unidirectional LSTM even though the bootstrapping bidirectional CE model has a better WER.

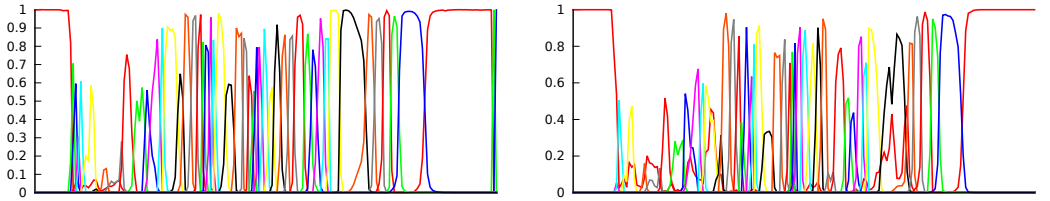
Figure 1 shows the plots of the label posteriors at each time step estimated by various LSTM RNN acoustic models. We observe that bidirectional models make better predictions thanks to use of past and future context. CTC models make only a few spikes for each phone while predicting *blank* label with high probability the rest of the time. Sequence discriminative training seems to be generally increasing the uncertainty by distributing the probabilities to other labels to fix the decoding errors due to disambiguation power of the language model. We observed that training CTC models is not very stable, especially for unidirectional models. The models can sometimes converge to a suboptimal alignment. This is due to the combinational effect of realignment during training and having a model with a memory that enables to learn alignments not necessarily corresponding to a true frame/label alignment. The model can remember the acoustic states it has seen and choose to output the short spikes at any time, without being constrained to output them in synchrony with the corresponding acoustic features. We found bootstrapping CTC models with CE trained models on fixed alignments makes training stable and convergence faster.

Decoding with CTC phone acoustic models is significantly faster than conventional context dependent HMM state models due to spiky predictions of the model. Note that the search graph is also smaller due to context independent acoustic units.



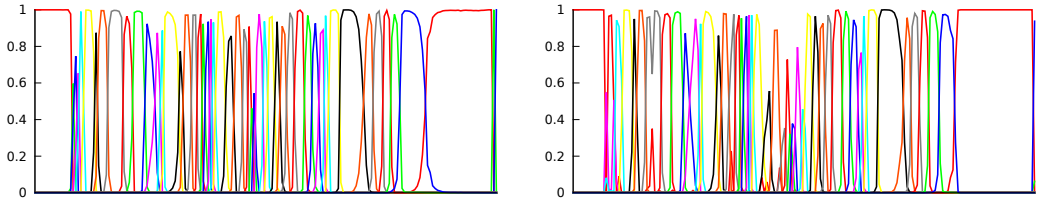
(a) CE unidirectional LSTM with phone labels

(b) CE bidirectional LSTM with phone labels



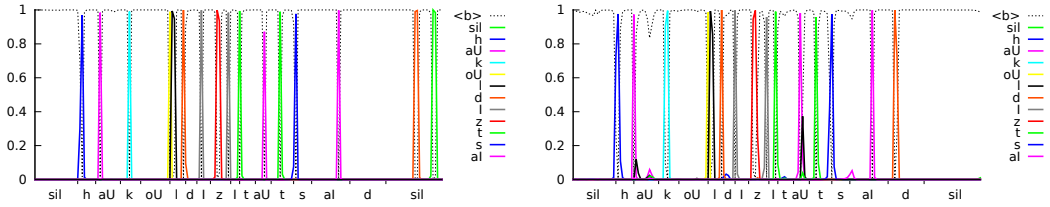
(c) CE unidirectional LSTM with HMM state labels

(d) CE + sMBR unidirectional LSTM with HMM state labels



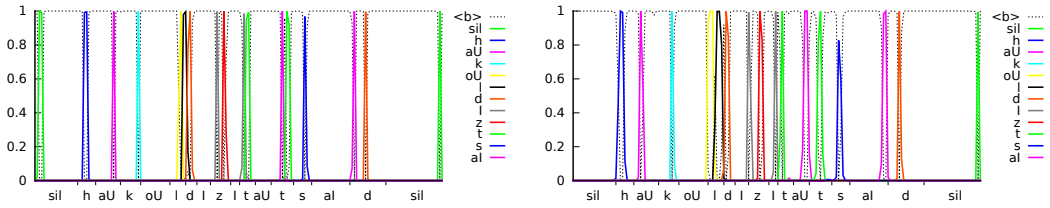
(e) CE bidirectional LSTM with HMM state labels

(f) CE + sMBR bidirectional LSTM with HMM state labels



(g) CTC unidirectional LSTM with phone labels

(h) CTC + sMBR unidirectional LSTM with phone labels



(i) CTC bidirectional LSTM with phone labels (10% PER)

(j) CTC + sMBR bidirectional LSTM with phone labels

Fig. 1: Label posteriors estimated by various LSTM RNN models plotted against fixed DNN frame level alignments on a heldout utterance *'how cold is it outside'*. We plot the posteriors for only the labels in the alignment. <*b*> refers to the *blank* label in CTC models.

5. REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist speech recognition*. Kluwer Academic Publishers, 1994.

[2] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[3] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *INTERSPEECH 2014*, 2014.

[4] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," in *Interspeech*, 2014.

[5] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," *ArXiv e-prints*, Feb. 2014.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[7] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of ICASSP*, 2013.

[8] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *INTERSPEECH*, 2011, pp. 437–440.

[9] G. Dahl, D. Yu, and L. Deng, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[10] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1109/TASL.2011.2134090

[11] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *INTERSPEECH*, 2012.

[12] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[13] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[14] Y. Normandin, "Hidden Markov models, maximum mutual information, and the speech recognition problem," Ph.D. dissertation, McGill University, Montreal, Canada, 1991.

[15] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge, England, 2004.

[16] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3761–3764.

[17] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *INTERSPEECH*, 2012.

[18] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 6664–6668.

[19] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *INTERSPEECH*, 2013.

[20] G. Heigold, "A log-linear discriminative modeling framework for speech recognition," Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, Jun. 2010.

[21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] F. Eyben, M. Wollmer, B. Schuller, and A. Graves, "From speech to letters using a novel neural network architecture for grapheme based ASR," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 376–380.

[23] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng, "Building high-level features using large scale unsupervised learning," in *International Conference on Machine Learning*, 2012, pp. 81–88.

[24] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[25] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Vancouver, Canada, Apr. 2013.