CHAPTER 1.2

DEEP DISCRIMINATIVE AND GENERATIVE MODELS FOR PATTERN RECOGNITION

Li Deng[1] and Navdeep Jaitly[2]

[1]*Microsoft Research, One Microsoft Way, Redmond, WA 98052*
[2]*Google Research, 1600 Amphitheatre Parkway, Mountain View, CA 94043*
E-mails: deng@microsoft.com; ndjaitly@google.com

In this chapter we describe deep generative and discriminative models as they have been applied to speech recognition and related pattern recognition problems. The former models describe the distribution of data or the joint distribution of data and the corresponding targets, whereas the latter models describe the distribution of targets conditioned on data. Both models are characterized as being 'deep' as they use layers of latent or hidden variables. Understanding and exploiting tradeoffs between deep generative and discriminative models is a fascinating area of research and it forms the background of this chapter. We focus on speech recognition but our analysis is applicable to other domains. We suggest ways in which deep generative models can be beneficially integrated with deep discriminative models based on their respective strengths. We also examine the recent advances in end-to-end optimization, a hallmark of deep learning that differentiates it from most standard pattern recognition practices.

1. Introduction

In pattern recognition, there are two main types of mathematical models: generative and discriminative models. The distinction between them is based on what probability distribution they model. Generally speaking, the main goal of pattern recognition models is to predict some output variable $\mathbf{y}$ given the value of an input variable or pattern $\mathbf{x}$. Discriminative models, including neural networks trained in a way that allows their output to be interpreted as approximate posterior class probabilities, directly compute the probability of an output given an input. On the other hand, generative models provide the joint probability distribution of the input and the output. That is, a discriminative model aims to estimate $p(\mathbf{y}|\mathbf{x})$, and a generative model aims to estimate $p(\mathbf{x}, \mathbf{y})$. For the latter, one can obtain $p(\mathbf{y}|\mathbf{x})$ using Bayes' theorem in order to indirectly perform pattern recognition or classification tasks, while the former, discriminative model performs directly.

The generative-discriminative distinction and their tradeoffs were a popular topic in the pattern recognition and machine learning literature over a decade ago (Ng and Jordan, 2002; Bouchard and Triggs, 2004; McCallum *et al.*, 2006;

Bishop and Lasserre, 2007; Liang and Jordan, 2008). Both theoretical and empirical studies pointed out that while discriminative models achieve lower asymptotic classification error, generative methods tend to be superior when training data are limited. And the bound of the asymptotic error is reached more quickly by a generative model than a discriminative model. While these general conclusions still hold, the dramatic development of deep learning over the past several years (Hinton *et al.*, 2006; Yu and Deng, 2011; Hinton *et al.*, 2012; Krizhevsky *et al.*, 2012; Dean *et al.*, 2012; Bengio *et al.*, 2013; Deng and Yu, 2014; Yu and Deng, 2014; Schmidhuber, 2015) warrants a reexamination of the fundamental issue of generative-discriminative modeling tradeoffs for two reasons. Firstly, the amount of training data (both labeled and unlabeled) and computing power available today is much greater than in previous decades. Secondly, significantly deeper and wider models are commonly used now. This provides the opportunity to embed more domain knowledge into the structure of these models. This was difficult to do in the earlier shallow models because they lacked modeling capacity (Deng, 2014). One main goal of this chapter is to seize these newly surfaced opportunities and to explore ways that deep generative and deep discriminative models can be beneficially integrated to achieve the best of both worlds. Another goal is to explore the recent advances in deep discriminative models with the end-to-end optimization strategy. End-to-end optimization was difficult to carry out for deep models many years ago, but recently many of the difficulties have been overcome. In these explorations, we will focus on the issues related to pattern recognition applied to speech signals.

The remainder of this chapter is organized as follows. In Section 2 we start by reviewing deep generative models of speech from the 1990's that were inspired by properties of speech production by the human vocal apparatus and its motor control driven by phonological units. In Section 3, we describe how understanding some weaknesses of these generative models led to an exploration of another type of generative models, Deep Belief Networks (DBN), in speech recognition towards the end of last decade. The work on DBNs led to subsequent revival of discriminative models for speech recognition. In Section 4 we contrast different aspects of deep generative and discriminative models. In Section 5, we discuss the deep discriminative models that have brought about significant progress in speech recognition accuracy. The tremendous success of these discriminative methods has meant that generative models have taken a back seat for the last several years. However, recently there has been significant progress in generative models and we survey some of these techniques, and outline how they might be used in future speech models. Until now, speech recognition experiments have required the use of traditional HMMs and or language models. Recent progress in deep learning has led to end-to-end methods that do not require traditional models. We explore some of these methods in Section 6. In Section 7 we discuss how generative and discriminative models may come together in the future. We conclude the chapter by a discussion of future avenues for research in speech pattern recognition.

2. Early Deep Generative Models for Speech Pattern Recognition

Prior to 2010 when deep neural nets (DNN) started to be adopted by the speech recognition community, a shallow generative approach based on the Hidden Markov Model (HMM) with Gaussian Mixture Models (GMM) as its state's output distribution had been the dominant method for many years (Baker *et al.*, 2009, 2009a; Deng and O'Shaughnessey, 2003). In the meantime, there had been a long history of research where human speech production mechanisms were exploited to construct deep and dynamic structure in probabilistic generative models (Deng *et al.*, 1997, 2000; Bridle *et al.*, 1998; Picone *et al.*, 1999; Deng, 2006). More specifically, the early work described in (Deng 1993; Deng *et al.*, 1994, 1997; Ostendorf *et al.*, 1996; Chengalvarayan *et al.*, 1998) generalized and extended the conventional shallow and conditionally independent GMM-HMM structure by imposing dynamic constraints on the HMM parameters. Subsequent work added new hidden layers into the dynamic model, giving rise to deep hidden dynamic models, to explicitly account for the target-directed, articulatory-like properties in human speech generation (Deng, 1998, 1999; Bridle *et al.*, 1998; Picone *et al.*, 1999; Togneri and Deng, 2003; Seide *et al.*, 2003; Zhou *et al.*, 2003; Deng and Huang, 2004; Ma and Deng, 2003, 2004). More efficient implementation of this deep architecture with hidden dynamics was achieved with non-recursive or finite impulse response (FIR) filters in more recent studies (Deng *et al.*, 2006, 2006a).

Reflecting on these earlier primitive versions of deep and dynamic generative models of speech, we note that neural networks, being used as "universal" nonlinear function approximators, have been incorporated in various components of the generative models. For example, the models described in (Bridle *et al.*, 1998; Deng and Ma, 2000; Deng, 2003) made use of neural networks to approximate the highly nonlinear mapping from articulatory configurations to acoustic features. Further, a version of the hidden dynamic model described in (Bridle *et al.*, 1998) has the full model parameterized as a dynamic neural network, and backpropagation algorithm was used to train this deep and dynamic generative model. Like DNN training of speech models, this method uses gradient descent for optimization. However, the two methods optimize very different kinds of loss functions. In the DNN case, the loss is defined as label mismatch. In the deep generative model, the loss is defined as the mismatch at the observable acoustic feature level via analysis-by-synthesis using labels to generate the acoustics. These deep-structured, dynamic generative models of speech can be shown as special cases of the more general dynamic network model and even more general dynamic graphical models (Bilmes, 2010), which can comprise many hidden layers to characterize the complex relationship among the variables including those in speech generation. Such deep generative graphical models are a powerful tool in many applications as they can incorporate domain knowledge and model uncertainty in real-world applications quite naturally. However, the approximations in inference, learning, prediction, and topology design

that arise in these intractable problems can reduce their effectiveness in practical applications.

In fact, the above difficulties in generative models have hindered progress in improving speech recognition accuracy (Lee *et al.*, 2003, 2004); see a review and analysis in (Deng and Togneri, 2014). In these early studies, variational Bayes for learning the intractable deep generative model was adopted, with the idea that during inference (i.e. the E step of learning), factorization of posterior probabilities was assumed while in the M-step rigorous estimation is expected to compensate for the approximation errors introduced by the factorization. It turned out that the inference results for the continuous-valued mid-hidden vectors were surprisingly good but those for the continuous-valued top-hidden layer (i.e. the linguistic symbols such as phones or words) were disappointing. Moreover, computation complexity for the inference step was extremely high. However, after additional assumptions were incorporated into the model structure, the inference of both continuous- and discrete-valued latent spaces performed satisfactorily and gave strong phone recognition results (Deng *et al.*, 2006).

3. Inroads of Deep Neural Nets to Speech Pattern Recognition

The above deep and dynamic generative models of speech were critically examined in fruitful collaborations between Microsoft Research and University of Toronto researchers during 2009-2010. While the speech community was developing layered hidden dynamical models outlined in the previous section, the machine learning community made significant strides in the development of a different type of deep generative model. These models were also characterized by layered architectures, similar to neural network. These were the DBN (Hinton *et al.*, 2006), which has an intriguing property: The rigorous inference step is much easier than that for the hidden dynamic model. Therefore, there is no need for approximate variational Bayes as required for the latter. This highly desirable property associated with the DBN, however, comes with the simplicity of not modeling dynamics, and thus making the DBN not directly suitable for speech modeling. In order to reconcile these two different types of deep generative models, an academic-industrial collaboration was formed between Microsoft Research and University of Toronto researchers toward end of 2009, preceding the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, where the first paper on the use of DBNs for phone recognition was presented (Mohamed *et al.*, 2009, 2012). This initial study and the ensuring collaborative work effectively made three simplifying assumptions that turned the deep generative models of speech discussed in Section 2 into DNNs. Firstly, to remove the complexity of rigorously modeling speech dynamics, one can for the time being remove such dynamics but one can compensate for this modeling inaccuracy by using a long time window to approximate the effects of true dynamics. This approximation leaves the task of modeling speech dynamics at the symbolic level to the standard HMM state sequence. Secondly, the direction

of information flow in the deep models can be reversed from top-down in the deep generative models to bottom-up in the DNN. This made inference fast and accurate. Thirdly, a DBN was used to "pre-train" the DNN based on the original proposal of (Hinton *et al.*, 2006) since it was assumed, back then, that neural networks were very difficult to train. However, larger-scale experiments and careful analyses conducted during 2010 at Microsoft (Yu *et al.*, 2010; Seide *et al.*, 2011; Dahl *et al.*, 2011) showed that with bigger datasets and careful weight initialization, generative pre-training DNNs using DBNs became no longer necessary (Yu *et al.*, 2011).

Adopting the above three "tricks" shaped the deep generative models, rather indirectly, into the DNN-based speech recognition framework. The initial experimental results using pre-trained DNNs with DBNs showed rather similar phone recognition accuracy to the deep hidden dynamic model of speech on the standard TIMIT task. The TIMIT data set has been commonly used to evaluate speech recognition models. Its small size allows many different configurations to be tried quickly and effectively. More importantly, the TIMIT task concerns phone-sequence recognition, which, unlike word-sequence recognition, permits very weak "language models" and thus the weaknesses in the acoustic modeling aspect of speech recognition can be more easily analyzed. Such an analysis on TIMIT was conducted at Microsoft Research during 2009–2010 that contrasted the phone recognition accuracy between deep generative models of speech (Deng and Yu, 2007) and deep discriminative models including pre-trained DNNs and deep conditional random fields (Mohamed *et al.*, 2010, 2012, 2009; Yu and Deng, 2009, 2010). A number of very interesting findings surfaced in such detailed analyses, suggesting a need to integrate deep generative and discriminative models.

The second simplifying solution above to the problem is the only one that has not been fixed in today's state of the art speech recognition systems. It is conceivable, however, that an entirely discriminative pipeline, such as that reported by Chorowski *et al.* (2014) may side-step these issues. We will explore these alternative directions of future research in later sections of this chapter.

4. Comparisons Between Deep Generative and Discriminative Models

As discussed earlier, pattern recognition problems attempt to model the relationship between target variables $\mathbf{y}$ (discrete or continuous) given input, covariate data $\mathbf{x}$. A deep discriminative model, such as a DNN, makes use of layered hierarchical architectures to directly optimize and compute $p(\mathbf{y}|\mathbf{x})$.A deep generative model, with examples given in Section 2 and Section 5 later, also exploits hierarchical architectures but the goal to estimate $p(\mathbf{x}, \mathbf{y})$ and then to determine $p(\mathbf{y}|\mathbf{x})$ indirectly via Bayes rule.

It has been well known that all theoretical guarantees associated with generative models are valid only if the model assumptions for data $\mathbf{x}$ are correct. Otherwise, their effectiveness for discriminative pattern recognition tasks is questionable as

incorrect assumptions on the data would lead to incorrect assessment of $p(\mathbf{y}|\mathbf{x})$. Since discriminative methods optimize $p(\mathbf{y}|\mathbf{x})$ directly, even if the model is not as expressive and powerful, the criterion that is optimized at training time may lead to superior pattern recognition performance at test time. However, the parameters of generative models can also be learned discriminatively using the same criterion of $p(\mathbf{y}|\mathbf{x})$. There have not been theoretical results on the degree to which model correctness is essential for discriminatively learned (deep) generative models to be superior or inferior to the purely discriminative models, such as DNNs that compute posterior probabilities directly with no probabilistic dependency among latent variables as common in deep generative models. While rigorous proofs exist for the equivalence between discriminative and generative models for certain shallow-structured models (Heigold *et al.*, 2011), in general, deep generative and discriminative models have different expressive capabilities with no solid theoretical results on their contrasts. Hence, comparisons between these two classes of deep models need be based on empirical ground.

To this end, we make such empirical comparisons in Table 1. Fifteen key attributes are listed in the table, based on which deep generative models (right column) and deep neural networks (mid column), the most popular form of deep discriminative models, are contrasted.

Table 1. High-level comparisons between deep neural networks, a most popular form of deep discriminative models (mid column), and deep generative models (right column), in terms of 15 attributes (left column)

	Deep Neural Networks	**Deep Generative Models**
Structure	Graphical; info flow: bottom-up	Graphical; info flow: top-down
Domain knowledge	Hard	Easy
Semi/unsupervised	Harder	Easier
Interpretation	Harder	Easy (generative "story")
Representation	Distributed	Local or Distributed
Inference/decode	Easy	Harder (but note recent progress in Section 5.2)
Scalability/compute	Easier (regular computes/GPU)	Harder (but note recent progress)
Incorp. uncertainty	Hard	Easy
Empirical goal	Classification, feature learning, etc.	Classification (via Bayes rule), latent variable inference, etc.
Terminology	Neurons, activation/gate functions, weights, etc.	Random variables, stochastic "neurons", potential function, parameters, etc.
Learning algorithm	Almost a single, unchallenged algorithm — Backprop	A major focus of open research, many algorithms, & more to come
Evaluation	On a black-box score — end performance	On almost every intermediate quantity
Implementation	Hard, but increasingly easier	Standardized methods exist, but some tricks and insights needed
Experiments	Massive, real data	Modest, often simulated data
Parameterization	Dense matrices	Sparse (often); Conditional PDFs

Most of these differentiating attributes are obvious. For example, for the attribute of "incorporating uncertainty," deep generative models are designed to capture the distribution of observed variables using a hierarchy of random variables where the variables in the lower layers (child nodes) are modeled conditionally on the variables in the higher layers (parent nodes). Such a model gives rise to "explaining away" in which the posterior over parent variables is a complicated, expressive distribution that cannot be factorized. This is not possible to achieve with DNNs that use softmax layers, because the parents are assumed to be conditionally independent, given the children. Therefore, if the real data and applications require representing such "explaining away" with uncertainty modeling, then deep generative models would do better than their discriminative counterparts.

One common difficulty of DNN models is their general lack of interpretability. Generative models on the other hand are easy to interpret since one can readily use $p(\mathbf{x}|\mathbf{y})$ to analyze what kinds of features $\mathbf{x}$ are associated with each class of $\mathbf{y}$. Such an analysis, however, is difficult to perform for discriminative models, which only compute $p(\mathbf{y}|\mathbf{x})$. Making DNN models interpretable is an active ongoing research.

In our opinion, implementation of learning algorithms for DNNs often involve many tricks known only to experienced researchers. In contrast, for deep generative models, standardized techniques often exist, such as variational EM, MCMC-based, and belief propagation methods. On the other hand, as these are approximation methods, their effectiveness often depends on the insights to the problem at hand which would help select the most appropriate approximation method while making algorithm implementation feasible in practice.

5. Successes of Deep Discriminative Neural Nets in Speech Recognition

The early experiments on phone recognition and error analysis discussed in Section 2, as well as on speech feature extraction which demonstrated the effectiveness of using raw spectrogram features (Deng *et al.*, 2010), had pointed to strong promise and practical value of deep learning. This early progress excited industrial researchers to devote more resources to pursue speech recognition research using deep learning approaches. The small-scale speech recognition experiments were soon expanded to larger scales (Dahl *et al.*, 2011, 2012; Seide *et al.*, 2011; Deng *et al.*, 2013b), and from Microsoft to other companies including Google, IBM, IflyTech, Nuance, Baidu, etc. (Jaitly *et al.*, 2012; Sak *et al.*, 2014, 2014a, 2015; Bacchiani and Rybach, 2014; Senior *et al.*, 2014; Sainath *et al.*, 2011, 2013, 2013a,b,c; Saon *et al.*, 2013; Hannun *et al.*, 2014). The experiments in speech recognition carried out at Microsoft showed that with increasing amounts of training data over the range of close to four orders of magnitude (from TIMIT to voice search to Switchboard), the DNN-based systems outperformed the GMM-based systems not only in absolute but also in relative percentages. Experiments at Google revealed that this advantage was retained even when the training sizes were expanded to 5000 hours of voice

search data (Jaitly *et al.*, 2012). This level of improvement in accuracy had rarely been achieved in the long speech recognition history.

The initial success of DNNs for speech recognition during 2009-2011 has led to an explosive development of new techniques. The first important development was pioneered by Microsoft Research related to the use of structured output distributions in the form of context-dependent (CD) phone and state units as the targets of DNNs (Yu *et al.*, 2010; Dahl *et al.*, 2011). Context dependent phones had been previously shown to be useful for shallow-net models (Bourlard *et al.*, 1992), but these models decomposed the probability into separate models for the left and right contexts in order to control the number of parameters. The Microsoft Research approach instead involved modeling the entire CD state distribution in the output layer. This type of design for the DNN output representations drastically expanded the output neurons from the context-independent phone states with the size of 100 to 200 commonly used in 1990's to the context-dependent ones with the size in the order from 1,000 to 30,000. Such design follows the traditional GMM-HMM systems, and was motivated initially by saving huge industry investment in the speech decoder software infrastructure. Early experiments at Microsoft further found that due to the significantly increased number of the HMM output units and hence the model capacity, the CD-DNN gave much higher accuracy when large training data supported such high modeling capacity. The combination of the above two factors accounted for why the CD-DNN has been so quickly adopted for deployment by the entire speech recognition industry.

For training CD-DNNs, GMM-HMM systems were used to generate alignments in the training data. However, the CD states used in these models were themselves created from acoustic confusability under the GMM-HMM models and may not be the best CD state inventory for DNN-HMM systems since DNNs may confuse phones differently from GMMs. In addition, it introduces an additional steps in the speech recognition pipeline. Google researchers have developed approaches that no longer require the initial GMM-HMM systems (Senior *et al.*, 2014; In these approaches, the model training starts directly from a DNN-HMM hybrid model on context independent (CI) states. The CI model is used to seed the creation of a CD state inventory based on the confusability of activations. It was shown that the CD state inventory can be grown with an online algorithm, producing improvements in word error rate as the model is trained (Bacchiani and Ryback, 2014).

For future studies in this area, the output representations for speech recognition can benefit from more linguistically-informed structured design based on symbolic or phonological units of speech. The rich phonological structure of symbolic nature in human speech has been well known for many years. Likewise, it has also been well understood for a long time that the use of phonetic or its finer state sequences, even with (linear) contextual dependency, in engineering speech recognition systems, is inadequate for representing such rich structure (e.g., Deng and Erler, 1992; Ostendorf, 1999; Sun and Deng, 2002). Such inadequacy thus leaves a promising open door to improve speech recognition systems' performance.

The second major area where DNNs have made a significant impact in speech recognition is to move from hand-crafted features to automatic feature extraction from raw signals. This was first explored successfully in the architecture of deep autoencoder on the "raw" spectrogram or linear filter-bank features, showing its superiority over the Mel-frequency cepstral coefficient (MFCC) features which contain a few stages of fixed transformation from spectrograms (Deng *et al.*, 2010).

The feature engineering pipeline from speech waveforms to MFCCs and their temporal differences goes through intermediate stages of log-spectra and then (Mel-warped) filter-banks. Deep learning is aimed to move away from separate design of feature representations and of classifiers. This idea of jointly learning classifier and feature transformation for speech recognition was already explored in early studies on the GMM-HMM-based systems (Chengalvarayan and Deng, 1997; 1997a; Rathinavalu and Deng, 1997). However, greater speech recognition performance gain is obtained only recently in the recognizers empowered by deep learning methods. For example, Mohamed *et al.* (2012a) and Li *et al.*, (2012) showed significantly lowered speech recognition errors using large-scale DNNs when moving from the MFCC features back to more primitive (Mel-scaled) filter-bank features. This work was motivated, in part by the experiments on generative models for raw speech signals, which showed that features found from generative models of raw waveforms were better than MFCCs for speech recognition (Jaitly *et al.* 2011).

Compared with MFCCs, "raw" spectral features not only retain more information, but also enable the use of convolution and pooling operations to represent and handle some typical speech invariance and variability expressed explicitly in the frequency domain. For example, the convolutional neural network (CNN) can only be meaningfully and effectively applied to speech recognition (Abdel-Hamid *et al.*, 2012; 2013, 2014; Deng *et al.*, 2013) when spectral features, instead of MFCC features, are used. More recently, Sainath *et al.* (2013b) went one step further toward raw features by learning the parameters that define the filter-banks on power spectra.

Ultimately, deep learning would go all the way to the lowest level of raw features of speech, i.e., speech sound waveforms, as was reported by Sheikhzadeh and Deng (1994). Jaitly and Hinton (2011) showed that a DBN trained on waveforms could discover features that outperform MFCCs, even though the features were learned in an entirely unsupervised task. Although the features did not outperform Mel filterbanks, it is clear that supervised learning of these features should produce better results. In fact, recent work by Sainath *et al.* (2015) shows that with supervised training raw signals can achieve accuracy comparable to filter banks. Similarly Tuske *et al.* (2014) reported excellent results based on raw waveforms for speech recognition using a DNN.

Third, better optimization criteria and methods are another area where significant advances have been made over the past several years in applying DNNs to speech recognition. In 2010, researchers at Microsoft recognized the importance

of sequence training based on their earlier experience on GMM-HMMs (He *et al.*, 2008; Yu *et al.*, 2007, 2008) and started working on full-sequence discriminative training for the DNN-HMM in phone recognition (Mohamed *et al.*, 2010). Unfortunately, the right approach was not found to effectively control the model overfitting problem. Effective solutions were first reported by Kingsbury *et al.* (2012) using Hessian-free training, and then by Su *et al.* (2013) and by Vesely *et al.* (2013) based on stochastic gradient descent training. Other better and novel optimization methods include distributed asynchronous stochastic gradient descent (Dean *et al.*, 2012; Sak *et al.*, 2014a), primal-dual method for applying natural parameter constraints (Chen and Deng, 2014), and Bayesian optimization for automated hyper-parameter tuning (Bergstra *et al.*, 2012).

The fourth area in which DNNs have made a big impact in speech recognition is noise robustness. Research into noise robustness in speech recognition has a long history, mostly before the recent rise of deep learning. See a comprehensive review in (Li *et al.*, 2014), where the class of feature-domain techniques developed originally for GMMs can be directly applied to DNNs. A detailed investigation of the use of DNNs for noise robust speech recognition in the feature domain was reported by Seltzer *et al.* (2013), who applied the C-MMSE (Yu *et al.*, 2008) feature enhancement algorithm onto the input feature used in the DNN. By processing both the training and testing data with the same algorithm, any consistent errors or artifacts introduced by the enhancement algorithm can be learned by the DNN-HMM recognizer. Strong results were obtained on the Aurora4 task. Kashiwagi *et al.* (2013) successfully applied the SPLICE feature enhancement technique developed for GMMs (Deng *et al.*, 2001, 2002) to a DNN speech recognizer.

Fifth, deep learning has been influencing multi-lingual or cross-lingual speech recognition, the most interesting application of multi-task learning. Prior to the rise of deep learning, cross-language data sharing and data weighing were already shown to be useful for the GMM-HMM system (Lin *et al.*, 2009). For the more recent, DNN-based systems, these multi-task learning applications in speech recognition became much more successful. In the studies reported by Huang *et al.* (2013) and Heigold *et al.* (2013), two research groups independently developed closely related DNN architectures with multi-task learning capabilities for multilingual speech recognition.

The sixth major area of progress in deep learning for speech recognition is the better architectures. For example, the tensor version of the DNN was reported by Yu *et al.* (2013) and showed substantially lower speech recognition errors compared with the conventional DNN. Another deep learning architecture effective for speech recognition is locally connected architectures, or (deep) convolutional neural networks (CNN). With appropriate changes from the CNN designed for image recognition to that taking into account speech-specific properties, the CNN has been found effective for speech recognition (Abdel-Hamid *et al.*, 2012, 2013, 2014; Sainath *et al.*, 2013; Deng *et al.*, 2013). Further, the deep learning architecture

of (deep) recurrent neural network (RNN), especially its long-short-term memory (LSTM) version, is currently a hot topic in speech recognition. The RNN was reported to give very low error rates on the benchmark TIMIT phone recognition task (Graves *et al.*, 2013; Deng and Chen, 2014). More recently, the LSTM was shown high effectiveness on large-scale tasks with applications to Google Now, voice search, and mobile dictation with excellent accuracy results (Sak *et al.*, 2014, 2014a).

Another set of novel deep architectures, which are quite different from the standard DNN, are reported in (Deng *et al.*, 2011, 2012; Tur *et al.*, 2012; Vinyals *et al.*, 2012) for successful speech recognition and related applications including speech understanding. These models are exemplified by the Deep Stacking Network (DSN), its tensor variants (Hutchinson *et al.*, 2012, 2013), and its kernel version (Deng *et al.*, 2012a). The novelty of this type of deep models lies in its modular design, where each module is constructed by taking its input from the output of the lower module concatenated with the original data input, and in the specific way of computing the error gradient of the weight matrices in each module (Yu and Deng, 2012a).

In addition to the six main areas of recent advances in deep learning for speech recognition summarized above, other important areas of progresses include adaptation of DNNs for speakers (Yao *et al.*, 2012; Yu *et al.*, 2012, 2013a), better regularization methods, better nonlinear units, speedup of DNN training and decoding, and exploitation of sparseness in DNNs (Yu *et al.*, 2012a).

6. Recent Developments of Deep Generative Models

In this chapter we have explored the connections between generative and discriminative models extensively. Discriminative models such as DNNs have the advantage that they can model arbitrarily complex posterior distributions, whereas the posteriors over generative models are defined by the expressiveness of the generative models themselves. As such a simple GMM-HMM system has uninteresting decision surfaces for complicated problems; deep trajectory models with latent dynamic layers such as those described in (Deng, 2006) on the other hand has more constraints built in, and is much more expressive. Recent developments in more powerful generative models thus deserve serious attention since these models have very expressive generative distributions that could lead to posterior distribution of arbitrary expressiveness. Furthermore, models such as the variational autoencoder (Kingma and Welling, 2014), DRAW (Gregor *et al.*, 2015), and Stochastic Generative Networks (Yoshua, *et al.*, 2013) are not associated with difficult inference problems that plagued earlier generative models, and are even applicable to model dynamics.

Deep generative models also deserve consideration as they facilitate principled unsupervised and semi-supervised learning. The models we have discussed so far are largely supervised — during training, we are provided pairs of acoustic data and sequence labels. However, there is a vast amount of unlabeled acoustic and textual

data available in the web that can be used for semi-supervised and unsupervised learning. Generative models that attempt to model the distribution of acoustics and text independently and/or jointly could be used to improve supervised learning in the future.

6.1. *Deep Distributed Generative Models*

Boltzmann Machines can be regarded as the earliest generative models (Hinton and Sejnowski, 1986). Inspired by the computational model of the brain, these models are "distributed" in the sense that they describe the distribution of data in terms of the activities of a population of variables, instead of just individual variables that encapsulate discrete, distinct concepts. That is, the information about data is "distributed" across the activities of a large number of variables, which leads to a compact representation. These models are formally described using principles of statistical physics — an energy function is defined over the states of variables and is used in a Boltzmann distribution which defines a probabilistic generative model over states of the variables. The authors showed how the parameters of these models could be trained using Gibbs sampling and simulated annealing to model interesting distributions. These models were slow both in inference and in learning for reasonably small sized problems, because of the exponentially large state spaces involved. Restricted Boltzmann Machines (RBMs) (Smolensky, 1986) make inference easier by introducing a layered structure where units can be updated in parallel using block Gibbs sampling, but learning is still difficult.

It was not until much later that it was discovered that these models could be trained by a simple algorithm called Contrastive Divergence (Hinton, 2002). Boltzmann machines inspired the development of Sigmoid Belief Networks in which the symmetric connections between variables were replaced by directed connections (Neal, 1992). This model bears similarities with Belief networks that were originally introduced by Pearl (1988) to represent domain knowledge in an expert using a probabilistic graphical structure. However, unlike the earlier models, the parameters of these graphical models were learnt. In Sigmoid Belief Networks the data resides at the lowest layer of the graph, and can be generated from an ancestral pass over the stochastic binary latent variables in the model. Inference in these models can be performed by Gibbs sampling over the latent variables; the partition function of the distribution is local to the units and can thus be computed as part of the inference step itself. However, this computation requires a separate ancestral pass for each variable and thus the computation process is still not suitable for large models. Subsequently Mean Field methods were developed for learning in such networks (Saul *et al.*, 1996). This method used the variational technique for approximating the intractable distribution over latent variables with another distribution that is more tractable and assumes independence amongst the variables. Helmholtz machines extended the intuition of Sigmoid Networks by introducing a model with several layers of directed models for generating data, and solve the diffi-

cult inference problem by coupling the layers to recognition weights that were used to compute approximate posteriors using variational techniques.

The wake-sleep algorithm was used to tune the generative and recognition weights by alternating between wake phase, when hidden unit states were inferred from the recognition weights and the generative weights were modified to generate the data, and the sleep phase, when the generative model was used to fantasize data and the recognition weights were modified to copy the generative process. Subsequently there was also an enormous amount of work done on Gaussian latent variable models, such as mixtures of factor analyzers (e.g. Ghahramani and Hinton, 1996), which can be trained with the Expectation Maximization (EM) algorithm. It was shown by Neal and Hinton (1998) that the EM algorithm could itself be derived from Mean Field methods. Then, an algorithm, called Contrastive Divergence, was invented that could be used to learn parameters of a Product of Experts (PoE) model using approximate gradients of the log likelihoods of data, which were computed from block Gibbs sampling over a small number of steps (Hinton, 2002). Later, it was shown that the CD could even be used on multilayer neural networks with a defined energy function (Hinton *et al.*, 2006a).

Various extensions on these generative models were developed for data with dynamical structure. For example, the Product of HMMs is a marriage of HMMs and Product of Experts (PoE) that uses multiple latent variables at each time step, rather than a single categorical variable at each time step that is used by HMMs. The generative distribution at each time step is a product model of the latent variables at that time step. It was shown that this model could be trained with the CD algorithm (Brown and Hinton, 2001; Taylor and Hinton, 2009). PoE seems to have led to modest gains in speech recognition accuracy (Airey and Gales, 2003) but product of HMMs seem largely to have been untested in the domain of speech recognition.

Other interesting deep distributed generative models for dynamic data have been developed that use the CD algorithm for training (Sutskever *et al.*, 2009; Taylor and Hinton, 2009, 2009a). Here, sequences are modeled using next step prediction. At each time step, a product of experts conditioned on past latent and visible variables is used to model the data at that time step. The models produce very interesting distributions over sequences and can model sequential data from Motion Capture, bouncing balls, etc. However, these models do not seem to have been applied to modeling acoustic data.

6.2. *Variational and Other Methods for Deep Generative Models*

Variational methods were very popular for training probabilistic models in the late 1990's but their use was limited by several factors. In the original formulation, the intractable posterior distributions over latent variables were approximated with simpler distributions such as independent Gaussian distributions over the latent variables, where the KL divergence between the approximating distribution and the

intractable distribution could be analytically treated. As such it could be applied only to a specific class of distributions where such approximations could be computed analytically. Second, such methods are difficult to apply to really large data because of the tricky optimization procedures required. Thus it was difficult to apply them to problems requiring a large number of parameters. Deep learning methods such as DNNs, however, did not suffer this drawback since stochastic gradient descent has proven resilient to massive amounts of training data and large model sizes.

Recent developments have addressed some of these shortcomings, thus opening up renewed interest in the use of variational methods. Hoffman *et al.* (2013) reported a method for using Stochastic Gradient descent with Variational Bayesian inference. This method allows for online learning of the type used for neural network training, where the model can be progressively trained as more data arrives. The authors show how the model can be used to train LDA models and Hierarchical Dirichlet processes on very large news corpora successfully. While the paper uses stochastic gradient descent with mini-batches, it seems obvious that recent developments in parallel gradient descent algorithms such as asynchronous gradient descent and Hogwild (Dean *et al.*, 2012, Recht *et al.*, 2011) could further help scale up such methods to even larger datasets. These methods nevertheless require traditional variational techniques to compute the gradients for the mini-batches — namely, having an analytical solution for the gradients of the variational parameters, which requires a careful selection of the approximating distributions and limits the use of arbitrary distributions in these settings. Recently, however, there have been breakthroughs that address these problems (Wingate and Weber, 2013; Kingma and Welling, 2014; Ranganath *et al.*, 2013; Mnih and Gregor, 2014). These methods replace the analytical optimization of the Evidence Lower Bound over the variational distribution, with a stochastic gradient optimization step computed by Monte Carlo. The sampling steps can result in gradient estimates that have high variance which must be controlled. Ranganath *et al.* (2013) reduce the variance through the use of Rao-Blackwellization (Casella, 1996) and control variates. Kingma and Welling (2014) use continuous latent variables with prior distributions such as location-scale distributions that can be easily sampled from, and where the gradients of the samples with respect to model parameters can be computed analytically. Mnih and Gregor (2014) use a centering technique that is in essence similar to the methods of control variates described by Ranganath *et al.* (2013). Further, they use conditional gradients for different layers, which is similar to Rao-Blackwellization. Both these methods use neural networks for the posterior distributions and for the generative models. As a result, the two models, Variational Autoencoders and Neural Variational Inference and Learning (NVIL) are very powerful, flexible generative models. See Gregor *et al.* (2015) for an extremely powerful generative model derived from these techniques.

These methods have recently been applied to sequential data with very promising results (Bayer and Osendorfer, 2014; Chung *et al.*, 2015). Future applications to speech domains are likely to follow.

Another really interesting approach to learning generative models comes from Bengio *et al.* (2014). Here the authors approach generative models from the perspective of Markov transition operators that go from corrupted data to clean data. Under certain conditions on the learned transition operators, the authors show that the learnt transition operator can be used to recover the data distribution. Further, the model is easy to sample from and can be trained with backpropagation and sampling. The implications of this approach to generative modeling for speech recognition are yet to be explored.

7. End-to-End Deep Discriminative Models

Deep learning experts have advocated training end-to-end models for pattern recognition systems since the 1990s (LeCun *et al.*, 1998). Originally inspired by discriminative sequence training methods in speech recognition systems, these ideas are being extensively explored in machine learning currently for a variety of tasks, especially those involving sequences, such as machine translation, speech recognition and parsing (Sutskever *et al.*, 2014; Vinyals *et al.*, 2014; Chorowski *et al.*, 2014; Graves and Jaitly, 2014). Part of this revival is fueled by the observation that deep learning systems based on discriminative neural networks often work better when the input data is minimally preprocessed. It is hoped that the same applies as the output loss functions are more directly related to the final objective that the overall system aims to optimize, not a surrogate loss that is correlated with the overall aim of the system. As a side benefit, end-to-end training is simpler as there are no additional complexities arising due to system integration issues. Recent successes in the above applied domains support this assertion. We summarize some of this work here because it is likely these methods will play an important role in speech recognition in the future.

Connectionist Temporal Classification (CTC) is a method for learning to map from a sequence to a shorter sequence of discrete symbols that have a monotonic alignment. It has been applied to handwriting recognition, speech recognition and grapheme-to-phoneme mapping (e.g. Graves *et al.*, 2006). This method has led to improved accuracy in speech recognition, but suffers from the assumption that the state predictions at each frame are independent of the predictions at the other frames.

The sequence to sequence model of (Sutskever *et al.*, 2004) addresses the theoretical shortcoming of these models, by modeling $p(\mathbf{y}|\mathbf{x})$ directly where $\mathbf{y}$ is the transcript and $\mathbf{x}$ is the input utterance using an RNN and the chain rule, i.e. $p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x})\ p(y_2|y_1,\ \mathbf{x}) \ldots p(y_{T-1}|y_{T,}\ \mathbf{x})$. Here, each of the terms in the chain rule is computed using an RNN that first inputs the data, and then the labels, y_i , to perform next step prediction for label y_{i+1}. This method was applied to machine translation and achieved impressive results, even though it was trained with no domain knowledge. Chorowski *et al.* (2014) apply an extension of this idea that uses additional "attention" on input sequence (Bahdanau, *et al.*, 2015) to speech

recognition. This model uses a deep bidirectional LSTM-RNN to process acoustic data to hidden units and a transducer RNN to output the transcript $\mathbf{y}$ conditioned on the hidden codes from the top layer of the acoustic RNN. The transducer RNN uses its hidden state to produce a blending weight over the input acoustic time steps, based on similarity between the hidden state and the hidden states of the acoustic RNN time steps. These weights are used to blend the acoustic RNN hidden states (of the top layer) to a single *context* vector that drives the transducer network to output the next character, by combining it with the hidden state of the transducer. A similar model was recently demonstrated for end-to-end speech recognition on the Google Voice Search Task (Chan *et al.*, 2015) where it was shown that an attention based sequence to sequence model could transcribe voice search data directly to a sequence of English letters without the use of language models or a dictionary. Although not a state of the art model, this model achieves results within 3% absolute of the best production model based on a DNN-HMM pipeline. It raises the exciting possibility that end to end speech recognition may be feasible with a large corpus.

All end-to-end training methods suffer from the problem that acoustic training data is limited compared with the pure text data available, but the model attempts to learn both the acoustic model and language model jointly. However, the techniques of language model blending with acoustic model prediction that are common in speech recognition with GMM-HMM systems are equally applicable here. The only trick is to modify the beam searching routine during inference to incorporate language model probabilities at the right steps of decoding and beam search.

8. Integrating Deep Generative and Discriminative Models

In this section we look at current approaches that blend generative and discriminative approaches, and outline some possible future approaches to speech recognition that use discriminative and generative models together. The advantage of using generative models for processes where the structure of the data is known a priori is that they can add constraints to the discriminative model. It has been shown empirically that for certain generative-discriminative model pairs, such as Naïve Bayes versus logistic regression, generative models can achieve faster convergence on discriminative tasks, with fewer points, but discriminative models achieve better convergence with more points (Ng and Jordan, 2002). Lasserre and Bishop (2007) propose a way of blending these two objectives together by training a discriminative and a generative model whose parameters are jointly described using a prior that encourages the parameters to be the same. They note that a discriminative model performs better than generative models when there is mis-specification of the generative model compared to the true one. One can develop a similar method of sharing parameters between discriminative and generative models below, but using RNNs.

Early approaches to DNN training for neural networks advocated the use of generative DBN pre-training before subsequent discriminative fine-tuning for DNNs. The DBN was trained to maximize the probability of data $\mathbf{x}$ itself, and its parameters were used to initialize the DNN which was trained discriminatively to model the HMM state posteriors, given the input data. However, it was subsequently observed that with very large datasets this pretraining was not necessary.

It should be noted that while GMM-HMM systems are trained generatively with maximum likelihood, the model is conditioned on the label sequence, $\mathbf{y}$, i.e. the objective to be optimized is $p(\mathbf{x}|\mathbf{y})$, rather than $p(\mathbf{x})$ itself, as is done in DBN training. To the best of our knowledge, generative pre-training of these models, akin to DBN pre-training of DNNs, using unsupervised audio data, $\mathbf{x}$**,** alone has not been attempted. One possible way of accomplishing this would be to apply a variational approximation over the (unknown) possible utterances $\mathbf{z}$ given input $\mathbf{x}$ and using these to update the generative models $p(\mathbf{x}|\mathbf{z})$. An unsupervised approach taken by Google resembles this method in principle (Kapralova *et al.*, 2014). In this model good speech recognition models are used without supervision to select utterances that can be decoded with high levels of confidence. These utterances are then added to a new dataset for training speech recognition models.

The approach described in Lasserre and Bishop (2007) can be used for semi-supervised learning by prescribing a common prior over models of speech and audio, allowing differences between these models. These models would thus be able to leverage large amounts of unsupervised text and audio data together with labeled, supervised pairs ($\mathbf{x}$, $\mathbf{y}$). It is expected that in the near future semi-supervised learning methods will play an important role in speech recognition.

A recent study demonstrated another interesting way of integrating deep generative and discriminative models for prediction problems using text data (Chen *et al.*, 2015). In this study, an iterative inference algorithm is first applied to a generative topic model. Each step of the inference operation is treated as a computational "cell" and multiple iterations give rise to several such "cell" stacking on top of each other.

Then, given target labels for the prediction problem in the training data, back-propagation algorithm can be applied to learn all model parameters in an end-to-end manner.

9. Summary and Future Directions

In pattern recognition literature and practice, both discriminative and generative models are popular. Understanding and exploiting the tradeoffs between these two classes of models have been a long standing research, and we focus on such research for various deep forms of these models in this chapter. Pattern recognition examples discussed are drawn mainly from speech recognition, a field which has recently been revolutionized by the use of deep neural networks, a specific and most successful form of deep discriminative models.

Deep discriminative models hold the promise of learning powerful end-to-end systems given enough labeled training data. However it is conceivable that the performance of these systems will plateau because the discriminative models are either not powerful enough, or not constrained enough by an appropriate discriminative architecture for the task of speech recognition. Generative models offer an easy way of incorporating a "correct" architecture into their models, although inference may be tricky under a powerful generative model. As such it is conceivable that the strengths of generative and discriminative models will both be needed for further progress in speech recognition.

In this vein, an important future challenge lies in how to effectively integrate major relevant speech knowledge and problem constraints into new deep models of the future with "correct" architectures. Deep generative models are much better able to impose the problem constraints above purely discriminative DNNs or their variants including recurrent networks. The deep generative models should be parameterized appropriately to facilitate highly regular, matrix-centric, large-scale computation. The design of the overall deep computational network architecture may be motivated by approximate inference algorithms associated with the initial generative model. Then, powerful discriminative learning algorithms of the type of end-to-end backpropagation can be developed and applied to learn all network parameters. Ultimately, the run-time computation follows the inference algorithm in the generative model, but the parameters will have already been learned to best discriminate all classes of speech sounds.

References

1. O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, Convolutional neural networks for speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(10), 1533–1545, (2014).
2. O. Abdel-Hamid, L. Deng, and D. Yu, Exploring convolutional neural network structures and optimization for speech recognition, *Interspeech*, (2013).
3. O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, *Proceedings of ICASSP*, (2012).
4. S. Airey, M. Gales, Product of Gaussians and multiple stream systems, *Proceedings of ICASSP*, (2003).
5. M. Bacchiani, D. Rybach, Context dependent state tying for speech recognition using deep neural network acoustic models, *ICASSP*, (2014).
6. D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, *ICLR*, (2015).
7. J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, Research developments and directions in speech recognition and understanding, *IEEE Signal Processing Magazine*, **26**(3), 75–80, (2009).
8. J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, Updated MINS report on speech recognition and understanding, *IEEE Signal Processing Magazine*, **26**(4), (2009a).

9. J. Bayer, and C. Osendorfer, Learning stochastic recurrent networks, arXiv:1411.7610, (2014).
10. Bengio *et al.* Deep generative stochastic networks trainable by backprop, *Proceedings of ICML,* (2014).
11. Bengio *et al.* Deep generative stochastic networks trainable by backprop, arXiv:1306.1091, (2013).
12. Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Transactions on PAMI*, **38**, 1798–1828, (2013).
13. J. Bergstra, and Y. Bengio, Random search for hyper-parameter optimization, *Journal of Machine Learning Research*, **3**, 281–305, (2012).
14. C. Bishop, and J. Lasserre, Generative or discriminative? Getting the best of both worlds, *Bayesian Statistics*, **8**, 3–24, (2007).
15. J. Bilmes, Dynamic graphical models, *IEEE Signal Processing Magazine*, **33**, 29–42, (2010).
16. G. Bouchard, and B. Triggs, The tradeoff between generative and discriminative classifiers, *Proceedings of COMPSTAT Symposium*, (2004).
17. H. Bourlard, *et al.* CDNN: A context dependent neural network for continuous speech recognition, *ICASSP*, (1992).
18. J. Bridle, L. Deng, J. Picone, H. Richards, J. Ma, T. Kamm, M. Schuster, S. Pike, and R. Reagan, An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition, *Final Report for 1998 Workshop on Language Engineering*, CLSP, Johns Hopkins, (1998).
19. A. Brown, and G. Hinton, Products of hidden Markov models, *Proceedings of Artificial Intelligence and Statistics*, (2001).
20. G. Casella, and C. Robert, Rao-Blackwellisation of sampling schemes, *Biometrika*, **83**, 81–94, (1996).
21. J. Chen, and L. Deng, A primal-dual method for training recurrent neural networks constrained by the echo-state property, *Proceedings of the International Conference Learning Representations*, (2014).
22. R. Chengalvarayan, and L. Deng, Speech trajectory discrimination using the minimum classification error learning, *IEEE Transactions on Speech and Audio Processing*, **6**(6), 505–515, (1998).
23. R. Chengalvarayan and L. Deng, HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features, *IEEE Transactions on Speech and Audio Processing*, 243–256, (1997).
24. R. Chengalvarayan and L. Deng, Use of generalized dynamic feature parameters for speech recognition, *IEEE Transactions on Speech and Audio Processing*, 232–242, (1997a).
25. R. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, End-to-end continuous speech recognition using attention-based recurrent NN: First results, arXiv:1412, (2014).
26. J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, A recurrent latent variable model for sequential data, arXiv:1506.02216, (2015).
27. G. 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*, **20**(1), 30–42, (2012).
28. G. Dahl, D. Yu, L. Deng, and A. Acero, Context-dependent DBN-HMMs in large vocabulary continuous speech recognition, *Proceedings of ICASSP*, (2011).
29. 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, *Proceedings of NIPS*, (2012).

30. L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, *APSIPA Transactions on Signal and Information Processing*, (2014).
31. L. Deng, and D. Yu, *Deep Learning: Methods and Applications*, Now Publishers, (2014).
32. L. Deng, and J. Chen, Sequence classification using the high-level features extracted from deep neural networks, *Proceedings of ICASSP*, (2014).
33. L. Deng, and R. Togneri, Deep dynamic models for learning hidden representations of speech features, Chapter 6 in the Book: *Speech and Audio Processing for Coding, Enhancement and Recognition* (Eds. Ogunfunmi *et al.*) pp. 153–196, Springer, 2014.
34. L. Deng, O. Abdel-Hamid, and D. Yu, A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion, *Proceedings of ICASSP*, (2013).
35. L. Deng, G. Hinton, and B. Kingsbury, New types of deep neural network learning for speech recognition and related applications: An overview, *Proceedings of ICASSP*, (2013a).
36. L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, Recent advances in deep learning for speech research at Microsoft, *Proceedings of ICASSP*, (2013b).
37. L. Deng, D. Yu, and J. Platt, Scalable stacking and learning for building deep architectures, *Proceedings of ICASSP*, (2012).
38. L. Deng, G. Tur, X. He, and D. Hakkani-Tur, Use of kernel deep convex networks and end-to-end learning for spoken language understanding, *Proceedings of IEEE Workshop on Spoken Language Technologies*, (2012a).
39. L. Deng, and D. Yu, Deep Convex Network: A scalable architecture for speech pattern classification, *Proceedings of Interspeech*, (2011).
40. L. Deng, and D. Yu, Deep convex networks for image and speech classification, *Deep Learning Workshop at ICML*, (2011).
41. L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, Binary coding of speech spectrograms using a deep autoencoder, *Proceedings of Interspeech*, (2010).
42. L. Deng, and D. Yu, Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition, *ICASSP*, (2007).
43. L. Deng, *Dynamic Speech Models — Theory, Algorithm, and Application*, Morgan & Claypool, (2006).
44. L. Deng, D. Yu, and A. Acero, Structured speech modeling, *IEEE Transactions on Audio, Speech and Language Processing*, **14**(5), 1492–1504, (2006).
45. L. Deng, D. Yu, and A. Acero, A bidirectional target filtering model of speech coarticulation: Two-stage implementation for phonetic recognition, *IEEE Transactions on Audio and Speech Processing*, **14**(1), 256–265, (2006a).
46. L. Deng, and X.D. Huang, Challenges in adopting speech recognition, *Communications of the ACM*, **47**(1), 11–13, (2004).
47. L. Deng, D. O'Shaughnessy, *SPEECH PROCESSING — A Dynamic and Optimization-Oriented Approach*, Marcel Dekker, (2003).
48. L. Deng, Switching dynamic system models for speech articulation and acoustics, in *Mathematical Foundations of Speech and Language Processing*, pp. 115–134. Springer-Verlag, New York, 2003.
49. L. Deng, K. Wang, A. Acero, J. Hon, Droppo, C. Boulis, Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and X. Huang, Distributed speech processing in MiPad's multimodal user interface, *IEEE Transactions on Speech and Audio Processing*, **10**(8), 605–619, (2002).
50. L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, High performance robust speech recognition using stereo training data, *Proceedings of ICASSP*, (2001).

51. L. Deng, and J. Ma, Spontaneous speech recognition using a statistical coarticulatory model for the vocal tract resonance dynamics, *Journal of the Acoustical Society of America*, **108**, 3036–3048, (2000).
52. L. Deng, Computational models for speech production, in *Computational Models of Speech Pattern Processing*, pp. 199–213, Springer Verlag, (1999).
53. L. Deng, A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition, *Speech Communication*, **24**(4), 299–323, (1998).
54. L. Deng and M. Aksmanovic, Speaker-independent phonetic classification using hidden Markov models with state-conditioned mixtures of trend functions, *IEEE Transactions on Speech and Audio Processing*, **5**, 319–324, (1997).
55. L. Deng, G. Ramsay, and D. Sun, Production models as a structural basis for automatic speech recognition, *Speech Communication*, **33**(2–3), 93–111, (1997).
56. L. Deng, M. Aksmanovic, D. Sun, and J. Wu, Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states, *IEEE Transactions on Speech and Audio Processing*, **2**(4), 507–520, (1994a).
57. L. Deng, A stochastic model of speech incorporating hierarchical nonstationarity, *IEEE Transactions on Speech and Audio Processing*, **1**(4), 471–475, (1993).
58. L. Deng, and K. Erler, Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units, *Journal of Acoustical Society of America*, **92**(6), 3058–3067, (1992).
59. Z. Ghahramani, and G. Hinton, The EM algorithm for mixtures of factor analyzers, **60**. Technical Report CRG-TR-96-1, University of Toronto, (1996).
60. A. Graves, and N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, *ICML,* (2014).
61. A. Graves, A. Mohamed, and G. Hinton, Speech recognition with deep recurrent neural networks, *Proceedings of ICASSP*, (2013).
62. Graves *et al.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *ICML*, (2006).
63. Gregor *et al.* DRAW: A recurrent neural network for image generation, arXiv:1502.04623, (2015).
64. A. Hannun, *et al.* Deep speech: Scaling up end-to-end speech recognition, arXiv:1412.5567.
65. X. He, L. Deng, and W. Chou, Discriminative learning in sequential pattern recognition — A unifying review for optimization-oriented speech recognition, *IEEE Signal Processing Magazine*, **25**, 14–36, (2008).
66. G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schluter, Equivalence of generative and log-liner models, *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(5), 1138–1148, (2011).
67. G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, Multilingual acoustic models using distributed deep neural networks, *Proceedings of ICASSP*, (2013).
68. 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, *IEEE Signal Processing Magazine*, **29**, 82–97, (2012).
69. G. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Computation*, 1771–1800, (2002).

70. G. Hinton, S. Osindero, and Y. Teh, A fast learning algorithm for deep belief nets, *Neural Computation*, **18**, 1527–1554, (2006).
71. Hinton *et al.* Unsupervised discovery of nonlinear structure using contrastive backpropagation, *Cognitive Science*, **30**, 725–731, (2006a).
72. G. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Computation*, **14**, 1771–1800, (2002).
73. G. Hinton, P. Dayan, B. Frey, and R. Neal, The wake-sleep algorithm for unsupervised neural networks, *Science*, **268**, 1158–1161, (1995).
74. G. Hinton, and T. Sejnowski, Learning and relearning in boltzmann machines, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Foundations* (Cambridge: MIT Press), pp. 282–317, (1986).
75. Hoffman *et al.* Stochastic variational inference, *The Journal of Machine Learning Research*, **14**, 1303–1347, (2013).
76. J. Huang, J. Li, L. Deng, and D. Yu, Cross-language knowledge transfer using multilingual deep neural networks with shared hidden layers, *Proceedings of ICASSP*, (2013).
77. B. Hutchinson, L. Deng, and D. Yu, A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition, *Proceedings of ICASSP*, (2012).
78. B. Hutchinson, L. Deng, and D. Yu, Tensor deep stacking networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1944–1957, (2013).
79. N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, Application of pretrained deep neural networks to large vocabulary speech recognition, *Proceedings of Interspeech*, (2012).
80. N. Jaitly, and G. Hinton, Learning a better representation of speech sound waves using restricted Boltzmann machines, *Proceedings of ICASSP*, (2011).
81. O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, A big data approach to acoustic model training corpus selection, *Proceedings of Interspeech*, (2014).
82. Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose, Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition, *Proceedings of ASRU*, (2013).
83. D. Kingma, and M. Welling, Auto-encoding variational bayes, *ICLR*, (2014).
84. B. Kingsbury, T. Sainath, and H. Soltau, Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization, *Proceedings of Interspeech*, (2012).
85. A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet classification with deep convolutional neural Networks, *Proceedings of NIPS*, (2012).
86. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of IEEE*, **86**, 2278–2324, (1998).
87. L. Lee, H. Attias, L. Deng, and P. Fieguth, A multimodal variational approach to learning and inference switching state space models, *Proceedings of ICASSP*, (2004).
88. L. Lee, H. Attias, and L. Deng, Variational inference and learning for segmental state space models of hidden speech dynamics, *Proceedings of ICASSP*, (2003).
89. J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, An overview of noise-robust automatic speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**, 745–777, (2014).
90. J. Li, D. Yu, J. Huang, and Y. Gong, Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM, *Proceedings of IEEE SLT*, (2012).

91. P. Liang, and M. Jordan, An asymptotic analysis of generative, discriminative, and pseudo-likelihood estimators, *Proceedings of ICML*, (2008).
92. H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, C.-H. Lee, A study on multilingual acoustic modeling for large vocabulary ASR, *Proceedings of ICASSP*, (2009).
93. J. Ma, and L. Deng, Target-directed mixture dynamic models for spontaneous speech recognition, *IEEE Transactions on Speech and Audio Processing*, **12**(1), 47–58, (2004).
94. J. Ma, and L. Deng, Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model, *IEEE Transactions on Speech and Audio Processing*, **11(**6), 590–602, (2003).
95. J. Ma, and L. Deng, A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamical model of speech, *Computer, Speech and Language*, (2000).
96. A. McCallum, C. Pal, G. Druck, & X. Wang, Multi-conditional learning: Generative/discriminative training for clustering and classification, *Proceedings of AAAI*, (2006).
97. A. Mnih, and K. Gregor, Neural variational inference and learning in belief networks, *ICML*, (2014).
98. A. Mohamed, G. Dahl, and G. Hinton, Acoustic modeling using deep belief networks, *IEEE Transactions on Audio, Speech, & Language Processing*, **20**(1), (2012). (the short conference version of this paper was presented at the 2009 NIPS Workshop).
99. A. Mohamed, G. Hinton, and G. Penn, Understanding how deep belief networks perform acoustic modelling, *Proceedings of ICASSP*, (2012a).
100. A. Mohamed, D. Yu, and L. Deng, Investigation of full-sequence training of deep belief networks for speech recognition, *Proceedings of Interspeech*, (2010).
101. R. Neal, and G. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning In Graphical Models*. Springer, pp. 355–368, (1998).
102. R. Neal, Connectionist learning of belief networks, *Artificial Intelligence*, **56**, 71–113, (1992).
103. A. Ng, and M. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regeression and naïve Bayes, *Proceedings of NIPS*, (2002).
104. M. Ostendorf, Moving beyond the 'beads-on-a-string' model of speech, *ASRU*, (1999).
105. M. Ostendorf, V. Digalakis, and O. Kimball, From HMMs to segment models: A unified view of stochastic modeling for speech recognition, *IEEE Transactions on Speech and Audio Processing*, **4**(5), (1996).
106. J. Pearl, *Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA, 1988.
107. P. Picone, S. Pike, R. Regan, T. Kamm, J. bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster, Initial evaluation of hidden dynamic models on conversational speech, *Proceedings of ICASSP*, (1999).
108. R. Ranganath, S. Gerrish, and M. Blei, Black box variational inference, arXiv:1401.0118, (2013).
109. C. Rathinavalu and L. Deng, Construction of state-dependent dynamic parameters by maximum likelihood: Applications to speech recognition, *Signal Processing*, **55**, 149–165, (1997).
110. B. Recht, *et al.* Hogwild: A lock-free approach to parallelizing stochastic gradient descent, *NIPS*, (2011).
111. T. Sainath, R. Weiss, A. Senior, W. Wilson, and O. Vinyals, Learning the speech front-end with raw waveform CLDNNs, *Proceedings of Interspeech*, (2015).

112. T. Sainath, B. Kingsbury, H. Soltau, and B. Ramabhadran, Optimization techniques to improve training speed of deep neural networks for large speech tasks, *IEEE Transactions on Audio, Speech, and Language Processing*, **21**, 2267–2276, (2013).
113. T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, Convolutional neural networks for LVCSR, *Proceedings of ICASSP*, (2013a).
114. T. Sainath, B. Kingsbury, A. Mohamed, and B. Ramabhadran, Learning filter banks within a deep neural network framework, *Proceedings of ASRU*, (2013b).
115. T. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, Low-rank matrix factorization for deep neural network training with high-dimensional output targets, *Proceedings of ICASSP*, (2013c).
116. T. Sainath, B. Kingsbury, B. Ramabhadran, P. Novak, and A. Mohamed, Making deep belief networks effective for large vocabulary continuous speech recognition, *Proceedings of ASRU*, (2011).
117. H. Sak, A. Senior, and F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, *Proceedings of Interspeech*, (2014).
118. 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, *Proceedings of Interspeech*, (2014a).
119. H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and F. Schalkwyk, Learning acoustic frame labeling for speech recognition with recurrent neural networks, *Proceedings of ICASSP*, (2015).
120. G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, Speaker adaptation of neural network acoustic models using i-vectors, *Proceedings of ASRU*, (2013).
121. L. Saul, T. Jaakkola, and M. Jordan, Mean field theory for sigmoid belief networks, *Journal of Artificial Intelligence Research*, **4**, 61–76, (1996). ks: An overview. *Neural Networks*, **61**, 85–117.
122. J. Schmidhuber, Deep learning in neural networks: An overview, *Neural Networks*, **61**, 85–117, (2015).
123. F. Seide, G. Li, and D. Yu, Conversational speech transcription using context-dependent deep neural networks, *Proceedings of Interspeech*, 437–440, (2011).
124. F. Seide, J. Zhou, and L. Deng, Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM: MAP decoding and evaluation, *Proceedings of ICASSP*, (2003).
125. M. Seltzer, D. Yu, and E. Wang, An investigation of deep neural networks for noise robust speech recognition, *Proceedings of ICASSP*, (2013).
126. A. Senior, G. Heigold, M. Bacchiani, and H. Liao, GMM-free DNN training, *ICASSP*, (2014).
127. H. Sheikhzadeh, and L. Deng, Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization, *IEEE Transactions on Speech and Audio Processing*, **2**, 80–91, (1994).
128. P. Smolensky, Information processing in dynamical systems: foundations of harmony theory, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press. pp. 194–281, (1986).
129. H. Su, G. Li, D. Yu, and F. Seide, Error back propagation for sequence training of context-dependent deep neural networks for conversational speech transcription, *Proceeding of ICASSP*, (2013).
130. J. Sun, and L. Deng, An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition, *The Journal of Acoustical Society of America*, **111**, 1086–1101, (2002).

131. I. Sutskever, O. Vinyals, and Q. Le, Sequence to sequence learning with neural networks, *Proceedings of NIPS*, (2014).
132. I. Sutskever, Hinton, and G. Taylor, The recurrent temporal restricted Boltzmann machine, *NIPS*, (2009).
133. R. Togneri, and L. Deng, Joint state and parameter estimation for a target-directed nonlinear dynamic system model, *IEEE Transactions on Signal Processing*, **51**(12), 3061–3070, (2003).
134. G. Tur, L. Deng, D. Hakkani-Tür, and X. He. Towards deep understanding: Deep convex networks for semantic utterance classification, *Proceedings of ICASSP*, (2012).
135. G. Taylor, and G. Hinton, Products of hidden Markov models: It takes N>1 to tango, *Proceedings of UAI,* (2009).
136. G. Taylor, and G. Hinton, Factored conditional restricted Boltzmann machines for modeling motion style, *ICML*, (2009a).
137. Z. Tuske, P. Golik, R. Schluter, H. Ney, Acoustic modeling with deep neural networks using raw time signal for LVCSR, *Proceedings of Interspeech*, (2014).
138. K. Vesely, A. Ghoshal, L. Burget, and D. Povey, Sequence-discriminative training of deep neural networks, *Proceedings of Interspeech*, (2013).
139. O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, Grammar as a foreign language, arXiv:1412.7449, (2014).
140. O. Vinyals, Y. Jia, L. Deng, and T. Darrell, Learning with recursive perceptual representations, *Proceedings of NIPS*, (2012).
141. D. Wingate, and T. Weber, Automated variational inference in probabilistic programming, arXiv:1301.1299, (2013).
142. K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, Adaptation of context-dependent deep neural networks for automatic speech recognition, *Proceedings of ICASSP*, (2012).
143. D. Yu, and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer, (2014).
144. D. Yu, L. Deng, and F. Seide, The deep tensor neural network with applications to large vocabulary speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(2), 388–396, (2013).
145. D. Yu, K. Yao, H. Su, G. Li, and F. Seide, KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition, *Proceedings of ICASSP*, (2013a).
146. D. Yu, X. Chen, and L. Deng, Factorized deep neural networks for adaptive speech recognition, *International Workshop on Statistical Machine Learning for Speech Processing*, (2012).
147. D. Yu, and L. Deng, Efficient and effective algorithms for training single-hidden-layer neural networks, *Pattern Recognition Letters*, **33**, 554–558, (2012a).
148. D. Yu, F. Seide, G. Li, L. Deng, Exploiting sparseness in deep neural networks for large vocabulary speech recognition, *Proceedings of ICASSP*, (2012b).
149. D. Yu, and L. Deng, Deep learning and its applications to signal and information processing, *IEEE Signal Processing Magazine*, 145–154, (2011).
150. D. Yu, L. Deng, G. Li, and F. Seide. Discriminative pretraining of deep neural networks, U.S. Patent Filing, Nov. 2011.
151. D. Yu, and L. Deng, Deep-structured hidden conditional random fields for phonetic recognition, *Proceedings of Interspeech*, (2010).
152. D. Yu, and L. Deng, Learning in the deep-structured hidden conditional random fields, *NIPS Workshop on Deep Learning for Speech Recognition*, (2009).

153. D. Yu, L. Deng, and G.E. Dahl, Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition, *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, (2010).
154. D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, A. Acero, Robust speech recognition using cepstral minimum-mean-square-error noise suppressor, *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(5), (2008).
155. D. Yu, L. Deng, X. He, and A. Acero, Large-margin minimum classification error training: a theoretical risk minimization perspective, *Computer Speech and Language*, **22**(4), 415–429, (2008).
156. D. Yu, L. Deng, X. He, and X. Acero, Large-margin minimum classification error training for large-scale speech recognition tasks, *Proceedings of ICASSP*, (2007).
157. J. Zhou, F. Seide, and L. Deng, Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM: modeling and training, *Proceedings of ICASSP*, (2003).
158. W. Chan, N. Jaitly, Q. Le and O. Vinyals, Listen, Attend and Speel, arXiv, CoRR, http://arxiv.org/abs/1508.01211, September 2015.
159. J. Chen, J. He, Y. Shen, L. Xiao, X. He, J. Gao, X. Song, and L. Deng. End-to-end Learning of Latent Dirichlet Allocation by Mirror-Descent Back Propagation, accepted and to appear in *Proc. Nips*, December 2015.