Chapter 1
Summary and Future Directions

Abstract In this chapter, we summarize the book by first listing and analyzing what we view as major milestone studies in the recent history of developing the deep learning based ASR techniques and systems. We describe the motivations of these studies, the innovations they have engendered, the improvements they have provided, and the impacts they have generated. In this road map, we will first cover the historical context in which the DNN technology made inroad into ASR around 2009 resulting from academic and industry collaborations. Then we select seven main themes in which innovations flourished across-the-board in ASR industry and academic research after the early debut of DNNs. Finally, our belief is provided on the current state of the art of speech recognition systems, and we also discuss our thoughts and analysis on the future research directions.

1.1 Road Map

There are many exciting advancements in the field of automatic speech recognition (ASR) in the past five years. However, in this book we are able to cover only a representative subset of these achievements. Constrained by our limited knowledge, we have selected the topics we believe are useful for the readers to understand these progresses that were described with much more technical detail in many preceding chapters of this book. We feel one way to reasonably summarize these advancements is to provide an outline of major milestone studies achieved historically.

1.1.1 Debut of DNNs for ASR

The application of neural networks on ASR can be dated back to late 1980s. Notable work includes Waibel et al.'s time delay neural network (TDNN) [93, 50] and Morgan et al.'s artificial neural network (ANN)/hidden Markov model (HMM) hybrid system [64, 65]

The resurgence of interest in neural network-based ASR started in The 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications [21], where Mohamed et al. from University of Toronto presented a primitive version of a deep neural network (DNN)[1]-HMM hybrid system for phone recognition [61]. Detailed error analysis and comparisons of the DNN with other speech recognizers were carefully conducted at Microsoft Research (MSR) jointly by MSR and University of Toronto researchers and relative strengths and weaknesses were identified prior to and after the workshop. See [20] for discussions and reflections on this part of early studies that were carried out with "a lot of intuitive guesses without much evidence to support the individual decisions." In [61], the same type of ANN/HMM hybrid architecture was adopted as those developed in early 1990s [64, 65] but it used a DNN to replace the shallow multi-layer perceptron (MLP) often used in the early ANN/HMM systems. More specifically, the DNN was constructed to model monophone states and was trained using the frame-level cross entropy criterion on the conventional MFCC features. They showed that by just using a deeper model they managed to achieve a 23.0% phone error rate (PER) on the TIMIT core test set. This result is significantly better than the 27.7% and 25.6% PER [73] achieved by a monophone and triphone Gaussian mixture model (GMM)-HMM, respectively, trained with the maximum likelihood estimation (MLE) criterion, and is also better than 24.8% PER achieved by a deep, monophone version of generative models of speech [30, 27] developed at MSR but with distinct recognition error patterns (not published). Although their model performs worse than the triphone GMM-HMM system trained using the sequence-discriminative training (SDT) criterion, which achieved 21.7% PER[2] on the same task, and was evaluated only on the phone recognition task, we at MSR noticed its potential because in the past the ANN/HMM hybrid system was hard to beat the context-dependent (CD)-GMM-HMM system trained with the MLE criterion and more importantly because the DNN and the deep generative models were observed to produce very different types of recognition errors with explainable causes based on aspects of human speech production and perception mechanisms. In the mean time, collaborations between MSR and University of Toronto researchers which started in 2009 also looked carefully into the

[1] Note that while the model was called deep belief network (DBN) at that time, it is in fact a deep neural network (DNN) initialized using the DBN pretraining algorithm. See Chapters **??** and **??** as well as [29] for discussions on the precise differences between DNNs and DBNs.

[2] The best GMM-HMM system can achieve 20.0% PER on the TIMIT core test set [73].

use of raw speech features, one of the fundamental premises of deep learning advocating not to use human-engineered features such as MFCCs. Deep autoencoders were first explored on speech historically, during 2009-2010 at MSR for binary feature encoding and bottleneck feature extraction, where deep architectures were found superior to shallow ones and spectrogram features found superior to MFCCs [24]. All the above kind of insightful and exciting results and progress on speech feature extraction, phone recognition, and error analysis, etc. had never been seen in the speech research history before and have pointed to high promise and practical value of deep learning. This early progress excited MSR researchers to devote more resources to pursue ASR research using deep learning approaches, the DNN approach in particular. A series of studies along this line can be found in [22][29].

Our interest at MSR was to improve large vocabulary speech recognition (LVSR) for real-world applications. In early 2010 we started to collaborate with the two student authors of the work [61] to investigate DNN-based ASR techniques. We used the voice search (VS) dataset described in Section **??** to evaluate our new models. We first applied the same architecture used by Mohamed et al. [61], which we refer to as the context-independent (CI)-DNN-HMM, to the LVSR. Similar to the results on the TIMIT phoneme recognition task, this CI-DNN-HMM, trained with 24 hours of data, achieved 37.3% word error rate (WER) on the VS test set. This result sits in between the 39.6% and 36.2% WER achieved with the CD-GMM-HMM trained using the MLE and SDT criteria, respectively. The performance breakthrough happened after we adopted the CD-DNN-HMM architecture described in Chapter **??** in which the DNN directly models the tied triphone states (also called senones). The CD-DNN-HMM based on senones achieved 30.1% WER. It was shown experimentally to cut errors by 17% over the 36.2% WER obtained with the CD-GMM-HMM trained using the SDT criterion, and to cut errors by 20% over the 37.3% WER obtained using the CI-DNN-HMM in a few papers published by Yu et al. [98] and Dahl et al. [13] . This is the first time the DNN-HMM system was successfully applied to LVSR tasks. In retrospect, it may be possible that other researchers have also thought about similar ideas and even have tried variants of it in the past. However, due to the limited computing power and training data in early days, no one was able to train the models with the large size that we use today.

As successful as it was in retrospect, the above early work on CD-DNN-HMM had not drawn as much attention from speech researchers and practitioners at the time of publications of the studies in 2010 and 2011. This was understandable as the ANN/HMM hybrid system did not win over the GMM-HMM system in mid-1990s and was not considered the right way to go. To turn over this belief researchers were looking for stronger evidence than that on the Microsoft internal voice search dataset which contains up to 48 hours of training data in those early experiments.

The CD-DNN-HMM work started to show greater impact after Seide et al. of MSR published their results in September 2011 [79] on applying the same CD-DNN-HMMs as that reported by Yu et al. [98] and Dahl et al. [13] to the Switchboard benchmark dataset [35] described in Section ??. This work scaled CD-DNN-HMMs to 309 hours of training data and thousands of senones. It demonstrated, quite surprisingly to many people, that the CD-DNN-HMM trained using the frame cross entropy criterion can achieve as low as 16.1% WER on the HUB5'00 evaluation set — a 1/3 cut of error over the 23.6% WER obtained with the CD-GMM-HMM trained using the SDT criterion. This work also confirmed and clarified the findings in [98, 13, 14]: the three key ingredients to make CD-DNN-HMM perform well are: 1) using deep models, 2) modeling senones, and 3) using a contextual window of features as input. It further demonstrated that realignment in training DNN-DMMs helps improve recognition accuracy and that pretraining of DNNs sometimes helps but is not critical. Since then, many ASR research groups shifted their research focus to CD-DNN-HMM and made significant progresses.

1.1.2 Speedup of DNN Training and Decoding

Right after the work [79] was published, many companies started to adopt it in their commercial systems. The first barrier they need to conquer is the decoding speed. With a naive implementation it takes 3.89 real time on a single CPU core just to compute the DNN score. In late 2011, just several months after [79] was published, Vanhoucke et al. from Google published their work on DNN speedup using engineering optimization techniques [90][3]. They showed that the DNN evaluation time can be reduced to 0.21 real time on a single CPU core by using quantization, SIMD instructions, batching and lazy evaluation - a 20 times speedup over the naive implementation. Their work is a great step ahead since it demonstrated that the CD-DNN-HMM can be used in real time commercial systems without penalty on the decoding speed or throughput.

The second barrier they need to overcome is training speed. Although it has been shown that the CD-DNN-HMM system trained with 309 hours of data outperforms the CD-GMM-HMM system trained with 2000 hours of data [79], additional accuracy improvement can be obtained if the DNN system is trained using the same amount of data as that used to train CD-GMM-HMMs. To achieve this goal some sort of parallel training algorithm needs to be developed. At Microsoft, a pipelined GPU training strategy was proposed and evaluated in 2012. The work [11] done by Chen et al. demonstrated that a speedup of 3.3 times can be achieved on 4 GPUs with this approach. Google, on the other hand, adopted the asynchronous stochastic gradient descent (ASGD) algorithm [67, 51] on the CPU clusters.

[3] Microsoft optimized the DNN evaluation in the internal tool using similar techniques slightly earlier but never published the results.

A different but notable approach to speeding up DNN training and evaluation is the low-rank approximation described in Section ??. In 2013, Sainath et al. from IBM and Xue et al. from Microsoft independently proposed to reduce the model size and training and decoding time by approximating the large weight matrices with the product of smaller ones [71, 96]. This technique can reduce 2/3 of the decoding time. Due to its simplicity and effectiveness it has been widely used in commercial ASR systems.

1.1.3 Sequence Discriminative Training

The exciting results reported in [79] was achieved using the frame-level cross entropy training criterion. Many research groups noticed that an obvious and low-risk way to further improve ASR accuracy is to use the sequence discriminative training criterion widely adopted for training the state-of-the-art GMM systems.

In fact, back in 2009, before the debut of DNN systems, Brian Kingsbury from IBM Research already proposed a unified framework to train ANN/HMM hybrid systems with SDT [46]. Although the ANN/HMM system he tested in his work performs worse than the CD-GMM-HMM system, he did show that the ANN/HMM system trained with the SDT criterion (achieved a 27.7% WER) performs significantly better than that trained with the frame-level cross entropy criterion (achieved a 34.0% WER on the same task). Hence this work did not attract strong attention at the time since even with the SDT the ANN/HMM system still cannot beat the GMM system.

In 2010, in parallel with the LVSR work, we at MSR clearly realized the importance of sequence training based on the GMM-HMM experience [100, 99, 38] and started the work on sequence discriminative training for CI-DNN-HMM for phone recognition [60]. Unfortunately we did not find the right approach to control the overfitting problem by then and thus only observed very small improvement by using SDT (22.2% PER) over the frame cross entropy training (22.8% PER)

The breakthrough happened in 2012 when Kingsbury et al. from IBM Research successfully applied the technique described in Kingsbury's 2009 work [46] to the CD-DNN-HMM [47]. Since SDT takes longer time to train than the frame-level cross entropy training they exploited the Hessian-free training algorithm [56] on a CPU cluster to speed up the training. With SDT the CD-DNN-HMM trained on the SWB 309 hour training set obtained a WER of 13.3% on the Hub5'00 evaluation set. It cuts error by relative 17% over the already low 16.1% WER achieved using the frame-level cross entropy criterion. Their work demonstrated that SDT can be effectively applied to the CD-DNN-HMM and result in great accuracy improvement. More importantly, the 13.3% WER achieved using the single pass speaker independent CD-DNN-HMM is also much better than the WER of 14.5% achieved using the best multi-pass speaker-adaptive GMM system. Thus, there is obviously no reason not to replace the GMM systems with the DNN systems in the commercial systems given this result.

However, SDT is tricky and not easy to be implemented correctly. In 2013, the work done at MSR by Su et al. [86] and the joint work done by Veselý et al. from Brno University, University of Edinburgh, and Johns Hopkins University [91] proposed a series of practical techniques for making the SDT effective and robust. These techniques, including lattice compensation, frame dropping, and F-smoothing, are now widely used.

1.1.4 Feature Processing

The feature processing pipeline in the conventional GMM systems involves many steps because the GMMs themselves cannot transform features. In 2011, Seide et al. conducted research at MSR on the effect of feature engineering techniques in the CD-DNN-HMM systems [78, 103]. They found that many feature processing steps, such as HLDA [49] and fMLLR [33], that are important in the GMM systems and shallow ANN/HMM hybrid systems are less important in the DNN systems. They explained their results by considering all the hidden layers in the DNN as a powerful nonlinear feature transformation and the softmax layer as a log-linear classifier. The feature transformation and the classifier are jointly optimized. Since DNNs can take correlated inputs many features that cannot be directly used in the GMM systems can now be used in the DNN systems. Because DNNs can approximate complicated feature transformation through many layers of nonlinear operations many of the feature processing steps used in the GMM systems may be removed without sacrificing the accuracy.

In 2012, Mohamed et al. from University of Toronto showed that by using the log Mel-scale filter bank feature instead of MFCC they can reduce PER from 23.7% to 22.6% on the TIMIT phone recognition task using a two-layer network [62]. Around the same time Li et al. at Microsoft demonstrated that using log Mel-scale filter bank features improves the accuracy on LVSR [53]. They also showed that by using log Mel-scale filter bank features tasks such as mixed-bandwidth speech recognition can be easily implemented in the CD-DNN-HMM systems. Log Mel-scale filter bank features are now the standard in most CD-DNN-HMM systems. Deng et al. reported a series of studies on the theme of backing to the deep learning premise of using spectrogram-like speech features [22].

The efforts to reduce the feature processing pipeline never ceased. For example, the work done at IBM Research by Sainath et al. in 2013 [70] showed that the CD-DNN-HMM systems can directly take the FFT spectrum as the input and learn the Mel-scale filters automatically. Most recently, the use of raw time waveform signals of speech by DNNs (i.e., zero feature extraction prior to DNN training) was reported in [89]. The study demonstrates the same advantage of learning truly non-stationary patterns of the speech signal localized in time across frame boundaries by the DNN as the earlier waveform-based and HMM-based generative models of speech [82, 32], but very different kinds of challenges remain to be overcome.

1.1.5 Adaptation

When the CD-DNN-HMM system just showed its effectiveness on the Switchboard task in 2011, one of the concerns back then was the lack of effective adaptation techniques, esp. since DNN systems have much more parameters than that in the conventional ANN/HMM hybrid systems. To address this concern, in 2011 in the work done at Microsoft Research by Seide et al. the feature discriminative linear regression (fDLR) adaptation technique was proposed and evaluated on the Switchboard dataset with small improvement in the accuracy [78].

In 2013, Yu et al. conducted a study at Microsoft [104], showing that by using Kullback-Leibler divergence (KLD) regularization they can effectively adapt the CD-DNN-HMM on the short message dictation tasks with 3-20% relative error reduction over the speaker-independent systems when different number of adaptation utterances are used. Their work indicates that adaptation on CD-DNN-HMM systems can be important and effective.

Later in the same year and in 2014, a series of adaptation techniques based on a similar architecture were developed. In the noise aware training (NaT) [80] developed at Microsoft by Seltzer et al., a noise code is estimated and used as part of the input. In this work, they showed that with NaT they can reduce the WER on the Aurora4 dataset from 13.4% to 12.4%, beating even the most complicated GMM system on the same task. In speaker aware training (SaT) [76] developed at IBM by Saon et al., a speaker code is estimated as the i-vector of the speaker and used as part of the input. They reported the results on the Switchboard dataset and reduced WER from 14.1% to 12.4% on the Hub5'00 evaluation set, a 12% error cut. In the speaker code approach [2, 97] developed at York University by Abdel-Hamid et al., the speaker code is trained for each speaker jointly with the DNN and used as part of the input.

1.1.6 Multi-task and Transfer Learning

As pointed out in [78, 103, 10, 18] and discussed in Chapter **??**, each hidden layer in the DNN can be considered a new representation of the input feature. This interpretation motivated many studies in sharing the same representation across languages and modalities. In 2012-2013[4] many groups including Microsoft, IBM, Johns Hopkins University, University of Edinburgh, and Google reported results on using the shared hidden-layer architectures for multi-lingual and cross-lingual ASR [88, 42, 39, 34], multi-modal ASR [41], and.multi-objective training of DNNs for speech recognition [55, 81, 9]. These studies indicated that by training the shared hidden layers with data from multiple languages and modalities or with multiple objectives we can build DNNs that work better for each language or modality than those trained specifically for the language or modality. This approach often helps ASR tasks the most for the languages with very limited training data.

1.1.7 Convolution Neural Networks

Using log Mel-scale filter bank features as the input also opens a door to apply techniques such as convolution neural networks (CNNs) to exploit the structure in the features. In 2012 Abdel-Hamid et al. showed for the first time that by using a CNN along the frequency axis they can normalize speaker differences and further reduce the PER from 20.7% to 20.0% on the TIMIT phone recognition task [4].

These results were later extended to LVSR in 2013 with improved CNN architectures, pretraining techniques, and pooling strategies in the studies by Abdel-Hamid et al.[1, 3] and Deng et al. [17] at Microsoft Research and the study by Sainath et al. [69, 72] at IBM Research. Further studies showed that the CNN helps mostly for the tasks in which the training set size or data variability is relatively small. For most other tasks the relative WER reduction is often in the small range of 2-3%. We believe as the training set size continues to increase the gap between the systems with and without CNN will diminish.

1.1.8 Recurrent Neural Networks and LSTM

Since the inroad of the DNN into ASR starting in 2009, perhaps the most notable new deep architecture is the recurrent neural network (RNN), esp. its long-short-term-memory (LSTM) version. While the RNN as well as the related nonlinear neural predictive models saw its early success in small ASR tasks [68, 19], it was not easy to duplicate due to the intricacy in training, let alone to scale them up for larger ASR tasks. Learning algorithms for the RNN have been dramatically improved since these early days, however, and much stronger and practical results have been obtained recently using the RNN,

[4] Some earlier work such as [77] exploited similar ideas but not on the DNN.

especially when the bidirectional LSTM architecture is exploited [37, 36] or when the high-level DNN features are used as inputs to the RNN [10, 18].

The LSTM was reported to give the lowest PER on the benchmark TIMIT phone recognition task in 2013 by Grave et al. at University of Toronto researchers [37, 36]. In 2014, Google researchers published the results using the LSTM on large-scale tasks with applications to Google Now, voice search, and mobile dictation with excellent accuracy results [74, 75]. To reduce the model size, the otherwise very large output vectors of LSTM units are linearly projected to smaller-dimensional vectors. Asynchronous stochastic gradient descent (ASGD) algorithm with truncated backpropagation through time (BPTT) is performed across hundreds of machines in CPU clusters. The best accuracy is obtained by optimizing the frame-level cross-entropy objective function followed by sequence discriminative training. With one LSTM stacking on top of another, this deep and recurrent LSTM model produced 9.7% WER on a large voice search task trained with 3 million utterances. This result is better than 10.7% WER achieved with frame-level cross entropy training criterion alone. It is also significantly better than the 10.4% WER obtained with the best DNN-HMM system using rectified linear units. Furthermore, this better accuracy is achieved while the total number of parameters is drastically reduced from 85 millions in the DNN system to 13 millions in the LSTM system. Some recent publications also showed that deep LSTMs are effective in ASR in reverberant mutltisource acoustic environments, as indicated by the strong results achieved by LSTMs in a recent ChiME Challenge task involving ASR in such difficult environments [95].

1.1.9 Other Deep Models

A number of other deep learning architectures have been developed for ASR. These include the deep tensor neural networks [101][102], deep stacking networks and their kernel version [28, 26, 31], tensor deep stacking networks [44, 45], recursive perceptual models [92], sequential deep belief networks[5], and ensemble deep learning architectures [23]. However, although these models have superior theoretical and computational properties over most of the basic deep models discussed above, they have not been explored with as much depth and scope, and are not mainstream methods in ASR so far.

1.2 State of the Art and Future Directions

1.2.1 State of the Art — A Brief Analysis

By combining CNN, DNN, and i-vector based adaptation techniques IBM researchers showed in 2014 that they can reduce the WER on the Switchboard Hub5'00 evaluation set to 10.4%. Compared to the best possible WER of 14.5% on the same evaluation set achieved using the GMM systems, DNN systems cut the error by 30%. This improvement is achieved solely through the acoustic model (AM) improvement. Recent advancements in neural network based language model (LM) and large-scale n-gram LM can further cut the error by 10-15%. Together the WER on the Switchboard task can be reduced to below 10%. Also, the LSTM-RNN model developed by Google researchers also demonstrated in 2014 dramatic error reduction in voice search tasks compared with other methods including those based on the feedforward DNN.

In fact, in many commercial systems the word (or character) error rates for the tasks such as mobile short message dictation and voice search are way below 10%. Some companies are even aiming at reducing the *sentence* error rate to below 10%. From the practical usage point of view we can reasonably regard that deep learning has largely solved the close-talk single-speaker ASR problem.

As we relax the constraints we impose on the tasks we are working on, however, we can quickly realize that the ASR systems still perform poorly under the following conditions even given the recent technology advancements:

- ASR with far field microphones; e.g., when the microphone is backgrounded in a living room, meeting room, or field video recordings;
- ASR under very noisy conditions; e.g., when loud music is playing and captured by the microphone;
- ASR with accented speech;
- ASR with multi-talker speech or side talks; e.g., in a meeting or in multiparty chatting;
- ASR with spontaneous speech in which the speech is not fluent, with variable speed or with emotions;

For these tasks, the WER of the current best systems is often in the range of 20%. New technological advances or clever engineering are needed to bring the errors down much further in order to make ASR useful under these difficult yet practical and realistic conditions.

1.2.2 Future Directions

We believe the ASR accuracy under some (not all) of the above conditions may be increased even without substantially new technologies in acoustic modeling for ASR. For example, by using more advanced microphone array techniques we can significantly reduce noise and side-talks and thus improve the recognition accuracy under these conditions. We may also generate or collect more training data for far field microphones and thus improve the performance when similar microphones are used.

However, to ultimately solve the ASR problem so that the ASR system's performance can match or even exceed that of human's under all conditions[5], new techniques and paradigms in acoustic modeling are needed. We perceive that the next-generation ASR systems can be solely described as a dynamic system that involves many connected components and recurrent feedbacks and can constantly make predictions, corrections, and adaptation. For example, the future ASR system will be able to automatically identify multiple talkers in the mixed speech or resolve speech and noise in noisy speech. The system will then be able to focus on and trace a specific speaker by ignoring other speakers and noises. This is a cognitive function of attention that humans effortlessly equipped with yet conspicuously lacking in today's ASR systems. The future ASR system will also need to be able to learn the key speech characteristics from the training set and generalize well to unseen speakers, accents, noisy conditions.

In order to move towards building such new ASR systems, it is highly desirable to first build powerful tools such as computational network (CN) and computational network toolkit (CNTK) we described in Chapter **??**. Such tools would allow large-scale and systematic experimentation on many more advanced deep architectures and algorithms, some of which are outlined in the preceding section, than the basic DNN and RNN. Further, as we discussed in the RNN chapter, new learning algorithms will need to be developed that can integrate the strengths of discriminative dynamic models (e.g., the RNN) with bottom-up information flow and of generative dynamic models with top-down information flow while overcoming their respective weaknesses. Recent progresses in stochastic and neural variational inference shown to be effective for learning deep generative models [59, 15, 40, 7] have moved us one step

[5] Under some constrained conditions, ASR systems can already perform better than humans. For example, in 2008 ASR systems can already beat the human performance on clean digit recognition with a 0.2% error rate [100]. In 2006, IBM researchers Kristjansson et al. reported their results on single-channel multi-talker speech recognition [48]. Their improved system [12] in 2010 can achieve 21.6% WER on the challenge task with very constrained language model and closed speaker set. This result is better than 22.3% WER achieved by humans. In 2014, the DNN-based system developed at Microsoft Research by Weng et al. achieved a WER of 18.8% [94] on the same task and generated far less errors than humans did.

closer to the desired bottom-up and top-down learning algorithms with multi-passes.

We speculate that the next-generation ASR systems may also seamlessly integrate semantic understanding, for example, to constrain the search space and correct semantically inconsistent hypotheses and thus benefit from research work in semantic understanding. In this direction, one needs to develop better semantic representations for word sequences, which are the output of ASR systems. Recent advances in continuous vector-space distributed representations of words and phrases [58, 85, 43, 83, 84], also known as word embedding or phrase embedding, have moved us one step closer to this goal.

Most recently, the concept of word embedding (i.e., with distributed representations of words) is introduced as an alternative to the traditional, phonetic-state-based pronunciation model in ASR, giving improvement in ASR accuracy [6]. This exemplifies an interesting new approach, based on distributed representations in continuous vector space, to modeling the linguistic symbols as the ASR output. It appears to be more powerful than several earlier approaches to distributed representations of word sequences in symbolic vector space - articulatory or phonetic-attribute based phonological models [87, 16, 25, 8, 52, 54]. Further research along this direction may exploit multiple modalities - speech acoustics and the associated image, gesture, and text - all embedded into the same "semantic" space of phonological nature - to support weakly supervised or even unsupervised learning for ASR.

For a longer term, we believe ASR research can benefit from human brain research projects and research in the areas of representation encoding and learning, recurrent networks with long-range dependency and conditional state switching, multi-task and unsupervised learning, and prediction-based methods for temporal/sequential information processing. As examples, effective computational models of attention and of phonetic feature encoding in cortical areas of the human auditory system [57, 66] are expected to help bridge the performance gap between human and computer speech recognition. Modeling perceptual control and interactions between speaker and listener has also been proposed to help improve ASR and spoken language processing performance and its practical use [63]. These capabilities are so far not reachable by current deep learning technology, and require "looking outside" into other fields such as cognitive science, computational linguistics, knowledge representation and management, artificial intelligence, neuroscience, and bio-inspired machine learning.

References

1. Abdel-Hamid, O., Deng, L., Yu, D.: Exploring convolutional neural network structures and optimization techniques for speech recognition pp. 3366–3370 (2013)
2. Abdel-Hamid, O., Jiang, H.: Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7942–7946 (2013)
3. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE Transactions on Audio, Speech and Language Processing (2014)
4. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Penn, G.: Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280. IEEE (2012)
5. Andrew, G., Bilmes, J.: Backpropagation in sequential deep belief networks. Proc. Neural Information Processing Systems (NIPS) (2013)
6. Bengio, S., Heigold, G.: Word embeddings for speech recognition (2014)
7. Bengio, Y.: Estimating or propagating gradients through stochastic neurons. CoRR (2013)
8. Bromberg, I., Qian, Q., Hou, J., Li, J., Ma, C., Matthews, B., Moreno-Daniel, A., Morris, J., Siniscalchi, S.M., Tsao, Y., Wang, Y.: Detection-based ASR in the automatic speech attribute transcription project. In: Proc. Interspeech, pp. 1829–1832 (2007)
9. Chen, D., Mak, B., Leung, C.C., Sivadas, S.: Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014)
10. Chen, J., Deng, L.: A primal-dual method for training recurrent neural networks constrained by the echo-state property. In: Proc. ICLR (2014)
11. Chen, X., Eversole, A., Li, G., Yu, D., Seide, F.: Pipelined back-propagation for context-dependent deep neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2012)
12. Cooke, M., Hershey, J.R., Rennie, S.J.: Monaural speech separation and recognition challenge. Computer Speech and Language **24**(1), 1–15 (2010)
13. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4688–4691 (2011)

14. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech and Language Processing **20**(1), 30–42 (2012)
15. Danilo Jimenez Rezende Shakir Mohamed, D.W.: Stochastic backpropagation and approximate inference in deep generative models. In: Proc. International Conference on Machine Learning (ICML) (2014)
16. Deng, L.: Articulatory features and associated production models in statistical speech recognition. In: Computational Models of Speech Pattern Processing, pp. 214–224. Springer-Verlag, New York (1999)
17. Deng, L., Abdel-Hamid, O., Yu, D.: A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6669–6673 (2013)
18. Deng, L., Chen, J.: Sequence classification using high-level features extracted from deep neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014)
19. Deng, L., Hassanein, K., Elmasry, M.: Analysis of the correlation structure for a neural predictive model with application to speech recognition. Neural Networks **7**(2), 331–339 (1994)
20. Deng, L., Hinton, G., Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: An overview. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada (2013)
21. Deng, L., Hinton, G., Yu, D.: Deep learning for speech recognition and related applications. In: NIPS Workshop. Whistler, Canada (2009)
22. Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., Acero, A.: Recent advances in deep learning for speech research at microsoft. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada (2013)
23. Deng, L., Platt, J.: Ensemble deep learning for speech recognition. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2014)
24. Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., Hinton, G.: Binary coding of speech spectrograms using a deep auto-encoder. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2010)
25. Deng, L., Sun, D.: A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. Journal Acoustical Society of America **85**, 2702–2719 (1994)
26. Deng, L., Tur, G., He, X., Hakkani-Tur, D.: Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In: Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 210–215 (2012)
27. Deng, L., Yu, D.: Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 445–448 (2007)
28. Deng, L., Yu, D.: Deep convex network: A scalable architecture for speech pattern classification. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2011)
29. Deng, L., Yu, D.: Deep Learning: Methods and Applications. NOW Publishers (2014)
30. Deng, L., Yu, D., Acero, A.: Structured speech modeling. IEEE Transactions on Speech and Audio Processing **14**, 1492–1504 (2006)
31. Deng, L., Yu, D., Platt, J.: Scalable stacking and learning for building deep architectures. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2012)
32. Ephraim, Y., Roberts, W.J.J.: Revisiting autoregressive hidden markov modeling of speech signals. IEEE Signal Processing Letters **12**, 166–169 (2005)

33. Gales, M.J.: Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language **12**(2), 75–98 (1998)
34. Ghoshal, A., Swietojanski, P., Renals, S.: Multilingual training of deep-neural netowrks. Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
35. Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 517–520 (1992)
36. Graves, A., Jaitly, N., Mahamed, A.: Hybrid speech recognition with deep bidirectional lstm. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada (2013)
37. Graves, A., Mahamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada (2013)
38. He, X., Deng, L., Chou, W.: Discriminative learning in sequential pattern recognition — a unifying review for optimization-oriented speech recognition. IEEE Signal Processing Magazine **25**(5), 14–36 (2008)
39. Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J.: Multilingual acoustic models using distributed deep neural networks. Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
40. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference
41. Huang, J., Kingsbury, B.: Audio-visual deep learning for noise robust speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7596–7599 (2013)
42. Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y.: Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
43. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: ACM International Conference on Information and Knowledge Management (2013)
44. Hutchinson, B., Deng, L., Yu, D.: A deep architecture with bilinear modeling of hidden representations: applications to phonetic recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2012)
45. Hutchinson, B., Deng, L., Yu, D.: Tensor deep stacking networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2013)
46. Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3761–3764 (2009)
47. Kingsbury, B., Sainath, T.N., Soltau, H.: Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2012)
48. Kristjansson, T.T., Hershey, J.R., Olsen, P.A., Rennie, S.J., Gopinath, R.A.: Super-human multi-talker speech recognition: the ibm 2006 speech separation challenge system. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2006)
49. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Communication **26**(4), 283–297 (1998)
50. Lang, K.J., Waibel, A.H., Hinton, G.E.: A time-delay neural network architecture for isolated word recognition. Neural networks **3**(1), 23–43 (1990)
51. Le, Q.V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., Ng, A.Y.: Building high-level features using large scale unsupervised learning. arXiv preprint arXiv:1112.6209 (2011)

52. Lee, C.H.: From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next-generation automatic speech recognition. In: Proc. International Conference on Spoken Language Processing (ICSLP), pp. 109–111 (2004)
53. Li, J., Yu, D., Huang, J.T., Gong, Y.: Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In: Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 131–136 (2012)
54. Lin, H., Deng, L., Yu, D., Gong, Y.f., Acero, A., Lee, C.H.: A study on multilingual acoustic modeling for large vocabulary ASR. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4333–4336 (2009)
55. Lu, Y., Lu, F., Sehgal, S., Gupta, S., Du, J., Tham, C.H., Green, P., Wan, V.: Multitask learning in connectionist speech recognition. In: Proc. Australian International Conference on Speech Science and Technology (2004)
56. Martens, J.: Deep learning via Hessian-free optimization. In: Proc. International Conference on Machine Learning (ICML), pp. 735–742 (2010)
57. Mesgarani, N., Chang, E.F.: Selective cortical representation of attended speaker in multi-talker speech perception. Nature **485**, 233–236 (2012)
58. Mikolov, T.: Statistical language models based on neural networks. Ph.D. thesis, Brno University of Technology (2012)
59. Mnih, A., K. Gregor, .: Neural variational inference and learning in belief networks. In: Proc. International Conference on Machine Learning (ICML) (2014)
60. rahman Mohamed, A., Yu, D., Deng, L.: Investigation of full-sequence training of deep belief networks for speech recognition. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 2846–2849 (2010)
61. Mohamed, A.r., Dahl, G.E., Hinton, G.: Deep belief networks for phone recognition. In: NIPS Workshop on Deep Learning for Speech Recognition and Related Applications (2009)
62. Mohamed, A.r., Hinton, G., Penn, G.: Understanding how deep belief networks perform acoustic modelling. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4273–4276 (2012)
63. Moore, R.: Spoken language processing: Time to look outside? In: Second International Conference on Statistical Language and Speech Processing (2014)
64. Morgan, N., Bourlard, H.: Continuous speech recognition using multilayer perceptrons with hidden Markov models. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 413–416 (1990)
65. Morgan, N., Bourlard, H.A.: Neural networks for statistical recognition of continuous speech. Proceedings of the IEEE **83**(5), 742–772 (1995)
66. N. Mesgarani C. Cheung, K.J.E.C.: Phonetic feature encoding in human superior temporal gyrus. Science **343**, 1006–1010 (2014)
67. Niu, F., Recht, B., Ré, C., Wright, S.J.: Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. arXiv preprint arXiv:1106.5730 (2011)
68. Robinson, A.J.: An application of recurrent nets to phone probability estimation. IEEE Transactions on Neural Networks **5**(2), 298–305 (1994)
69. Sainath, T.N., Kingsbury, B., Mohamed, A.r., Dahl, G.E., Saon, G., Soltau, H., Beran, T., Aravkin, A.Y., Ramabhadran, B.: Improvements to deep convolutional neural networks for lvcsr. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU), pp. 315–320 (2013)
70. Sainath, T.N., Kingsbury, B., Mohamed, A.r., Ramabhadran, B.: Learning filter banks within a deep neural network framework. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU) (2013)
71. Sainath, T.N., Kingsbury, B., Sindhwani, V., Arisoy, E., Ramabhadran, B.: Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6655–6659 (2013)

72. Sainath, T.N., Mohamed, A.r., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for LVCSR. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614–8618 (2013)
73. Sainath, T.N., Ramabhadran, B., Picheny, M.: An exploration of large vocabulary tools for small vocabulary phonetic recognition. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU), pp. 359–364 (2009)
74. Sak, H., Senior, A., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2014)
75. Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., Mao, M.: Sequence discriminative distributed training of long short-term memory recurrent neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2014)
76. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU), pp. 55–59 (2013)
77. Schultz, T., Waibel, A.: Multilingual and crosslingual speech recognition. In: Proc. DARPA Workshop on Broadcast News Transcription and Understanding, pp. 259–262 (1998)
78. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU), pp. 24–29 (2011)
79. Seide, F., Li, G., Yu, D.: Conversational speech transcription using context-dependent deep neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 437–440 (2011)
80. Seltzer, M., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
81. Seltzer, M.L., Droppo, J.: Multi-task learning in deep neural networks for improved phoneme recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6965–6969 (2013)
82. Sheikhzadeh, H., Deng, L.: 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)
83. Shen, Y., Gao, J., He, X., Deng, L., Mesnil, G.: A latent semantic model with convolutional-pooling structure for information retrieval. In: ACM International Conference on Information and Knowledge Management (2014)
84. Socher, R., Huval, B., Manning, C., Ng, A.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2012)
85. Socher, R., Lin, C.C., Ng, A., Manning, C.: Parsing natural scenes and natural language with recursive neural networks. In: Proc. International Conference on Machine Learning (ICML), pp. 129–136 (2011)
86. Su, H., Li, G., Yu, D., Seide, F.: Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
87. Sun, J., Deng, L.: An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition. Journal Acoustical Society of America **111**, 1086–1101 (2002)
88. Thomas, S., Ganapathy, S., Hermansky, H.: Multilingual MLP features for low-resource LVCSR systems. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4269–4272 (2012)

89. Tuske, Z., Golik, P., Schluter, R., Ney, H.: Acoustic modeling with deep neural networks using raw time signal for LVCSR. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2014)
90. Vanhoucke, V., Senior, A., Mao, M.Z.: Improving the speed of neural networks on CPUs. In: Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
91. Veselỳ, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2013)
92. Vinyals, O., Jia, Y., Deng, L., Darrell, T.: Learning with recursive perceptual representations. Proc. Neural Information Processing Systems (NIPS) **15** (2012)
93. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. IEEE Transactions on Speech and Audio Processing **37**(3), 328–339 (1989)
94. Weng, C., Yu, D., Seltzer, M., Droppo, J.: Single-channel mixed speech recognition using deep neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5669–5673 (2014)
95. Weninger, F., Geiger, J., Wollmer, M., Schuller, B., Rigoll, G.: Feature enhancement by deep lstm networks for ASR in reverberant multisource environments. Computer Speech and Language pp. 888–902 (2014)
96. Xue, J., Li, J., Gong, Y.: Restructuring of deep neural network acoustic models with singular value decomposition. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2013)
97. Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L.: Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6389–6393 (2014)
98. Yu, D., Deng, L., Dahl, G.: Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. In: Proc. Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning (2010)
99. Yu, D., Deng, L., He, X., Acero, A.: Large-margin minimum classification error training for large-scale speech recognition tasks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. IV–1137 (2007)
100. Yu, D., Deng, L., He, X., Acero, A.: Large-margin minimum classification error training: A theoretical risk minimization perspective. Computer Speech and Language **22**(4), 415–429 (2008)
101. Yu, D., Deng, L., Seide, F.: Large vocabulary speech recognition using deep tensor neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2012)
102. Yu, D., Deng, L., Seide, F.: The deep tensor neural network with applications to large vocabulary speech recognition **21**(3), 388–396 (2013)
103. Yu, D., Seltzer, M.L., Li, J., Huang, J.T., Seide, F.: Feature learning in deep neural networks - studies on speech recognition tasks (2013)
104. Yu, D., Yao, K., Su, H., Li, G., Seide, F.: Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7893–7897 (2013)