SIP (2016), vol. 5, e1, page 1 of 15 © The Authors, 2016.

doi:10.1017/ATSIP.2015.22

INDUSTRIAL TECHNOLOGY ADVANCES

Deep learning: from speech recognition to language and multimodal processing

LI DENG

While artificial neural networks have been in existence for over half a century, it was not until year 2010 that they had made a significant impact on speech recognition with a deep form of such networks. This invited paper, based on my keynote talk given at Interspeech conference in Singapore in September 2014, will first reflect on the historical path to this transformative success, after providing brief reviews of earlier studies on (shallow) neural networks and on (deep) generative models relevant to the introduction of deep neural networks (DNN) to speech recognition several years ago. The role of well-timed academic-industrial collaboration is highlighted, so are the advances of big data, big compute, and the seamless integration between the application-domain knowledge of speech and general principles of deep learning. Then, an overview is given on sweeping achievements of deep learning in speech recognition since its initial success. Such achievements, summarized into six major areas in this article, have resulted in across-the-board, industry-wide deployment of deep learning in speech recognition systems. Next, more challenging applications of deep learning, natural language and multimodal processing, are selectively reviewed and analyzed. Examples include machine translation, knowledgebase completion, information retrieval, and automatic image captioning, where fresh ideas from deep learning, continuous-space embedding in particular, are shown to be revolutionizing these application areas albeit with less rapid pace than for speech and image recognition. Finally, a number of key issues in deep learning are discussed, and future directions are analyzed for perceptual tasks such as speech, image, and video, as well as for cognitive tasks involving natural language.

Keywords: Deep learning, Multimodal, Speech recognition, Language processing, Deep neural networks

Received 10 May 2015; Revised 6 December 2015

I. INTRODUCTION

The main theme of this paper is to reflect on the recent history of how deep learning has profoundly revolutionized the field of automatic speech recognition (ASR) and to elaborate on what kind of lessons we can learn to not only further advance ASR technology but also to impact the related, arguably more important, applications in language and multimodal processing. Language processing concerns "downstream" analysis and distillation of information from the ASR systems' outputs. Semantic analysis of language and multimodal processing involving speech, text, and image, both experiencing rapid advances based on deep learning over the past few years, holds the potential to solve some difficult and remaining ASR problems and present new challenges for the deep learning technology.

A message to be conveyed in this paper is the importance of broadening deep learning from deep neural networks (DNNs) to include deep generative models as well. In fact, a brief historical review conducted in Section II will touch on how the development of deep (and dynamic) generative models of speech played a role in the inroads of DNNs into modern ASR. Since 2011, the DNN has taken over the dominating (shallow) generative model of speech, the Gaussian Mixture Model (GMM), as the output distribution in the Hidden Markov Model (HMM). This purely discriminative DNN has been well-known to the ASR community, which can be considered as a shallow network unfolding in space. When the unfolding occurs in time, we have the recurrent neural network (RNN). On the other hand, deep generative models have distinct advantages over discriminative DNNs, including the strengths of model interpretability, of embedding domain knowledge and causal relationships, and of modeling uncertainty. Deep generative and discriminative models represent two apparently opposing approaches yet with highly complementary strengths and weaknesses. The further success of deep learning is likely to lie in how to seamlessly integrate the two approaches in a practically effective and theoretically appealing fashion, and to achieve the best of both worlds.

The remainder of this paper is organized as follows. In Section II, some brief history is provided on how deep learning made inroad into speech recognition, and a number of enabling factors are discussed. Outstanding achievements of deep learning both in academic world and in industry to

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

Corresponding author:
Li Deng
Email: deng@microsoft.com

date are reviewed in Section III, categorized into six major areas where speech recognition technology has been revolutionized within just past several years. Section IV is devoted to more challenging applications of deep learning to natural language and multimodal processing, where active work is ongoing with current progress reviewed. Finally, in Section V, remaining challenges for deep speech recognition are examined, together with much greater challenges for natural-language-related applications of deep learning and with directions for the future development.

II. SOME BRIEF HISTORY OF "DEEP" SPEECH RECOGNITION

Artificial neural networks have been around for over half a century and their applications to speech processing have been almost as long. Representative early work in using shallow (and small) neural networks for speech includes the studies reported in [1–6]. However, these neural nets did not show superior performance over the GMM-HMM technology based on generative models of speech trained discriminatively [7, 8]. A number of key difficulties had been methodologically analyzed, including vanishing gradient and weak temporal correlation structure in the neural predictive models [9, 10]. These difficulties were investigated in addition to the lack of big training data and big computing power in those early days in 1990s. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue (deep) generative modeling approaches for ASR [10–13]. Since mid-1990s, many prominent neural network and machine learning researchers also published their books and research papers on generative modeling [14–19]. This was so in some cases even if the generative models' architectures were based on neural network parameterization [14, 15]. It was not until several years ago with the resurgence of neural networks (with the "deep" form) and with the start of deep learning that all the difficulties encountered in 1990s have been overcome, especially for large vocabulary ASR applications [20–28]. The path towards exploiting large amounts of labeled speech data and powerful GPU-based computing for serious new implementations of neural networks involved extremely valuable academic-industry collaboration during 2009–2010. The importance of making models deep was initially motivated by the limitations of both probabilistic generative modeling and discriminative neural net modeling.

A) A selected review of deep generative models of speech prior to 2009

There has been a long history in speech recognition research where human speech production mechanisms are exploited to construct dynamic and deep structure in probabilistic generative models; see [29] and several presentations at the 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. More specifically, the early work described in [30–33] generalized and extended the conventional shallow and conditionally independent GMM-HMM structure by imposing dynamic constraints, in the form of polynomial trajectory, on the HMM parameters. A variant of this approach has been developed later using different learning techniques for time-varying HMM parameters and with the applications extended to speech recognition robustness [34, 35]. Similar trajectory HMMs also form the basis for parametric speech synthesis [36, 37]. Subsequent work added new hidden layers into the dynamic model, thus being deep, to explicitly account for the target-directed, articulatory-like properties in human speech generation [11–13, 38–45]. More efficient implementation of this deep architecture with hidden dynamics was achieved with non-recursive or finite impulse response filters in more recent studies [46].

Reflecting on these earlier primitive versions of deep and dynamic generative models of speech, we note that neural networks, being used as "universal" non-linear function approximators, have been incorporated in various components of the generative models. For example, the models described in [38, 47, 48] made use of neural networks to approximate the highly non-linear mapping from articulatory configurations to acoustic features. Further, a version of the hidden dynamic model described in [12] has the full model parameterized as a dynamic neural network, and backpropagation algorithm was used to train this deep and dynamic generative model. The key difference between this backpropagation and that used in training the DNN lies in how the error objective function is defined, while the optimization methods based on gradient descent are the same. In the DNN case, the error objective is defined as the label mismatch. In the deep generative model, the error objective is defined as the mismatch at the observable acoustic feature level via analysis-by-synthesis, and the error "back" propagation is towards the top label layer instead of back towards the observations as in the standard backprop. The assumption is that if the speech features generated by this deep model matches well with the observed speech data, then top-layer labels responsible for the speech production mechanism much be correct.

The above 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 [49]. The graphical models [19] 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 due to their capabilities of embedding domain knowledge and of explicitly modeling uncertainty in real-world applications. However, they often suffer from inappropriate approximations in inference, learning, prediction, and topology design, all arising from intractability in these tasks in practical applications.

Indeed, in the history of developing deep generative models of speech, the above difficulties have been found to

seriously hinder the progress in improving ASR accuracy [50, 51]; see a review and analysis in [52]. 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), full or partial factorization of posterior probabilities was assumed while in the M-step rigorous estimation should 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. Only after many additional assumptions were made without sacrificing essential properties of deep and dynamic nature of the generative model (i.e. target-directedness in the phonetic space, smoothness in hidden dynamic variables, adequate representation of phonetic target undershooting, rigorous non-linear relationship between the hidden and observation vectors, etc.), did the model become well performed in inference in both continuous- and discrete-valued latent spaces [46, 53]. In fact, when the hidden layer of the model took the vocal tract resonance vector as its physical variables, the inference algorithm on such continuous-valued vectors produced the best formant tracker then [54, 55]. The resulting estimates actually formed the basis for a standard database of the "ground truth" of formant frequencies to evaluate formant tracking algorithms [56].

B) From deep generative models to deep neural nets

The deep and dynamic generative models of speech, all with probabilistic formulations of the various types discussed above, were closely examined in 2009 during the collaboration between Microsoft Research and University of Toronto researchers. In parallel with the development of these probabilistic speech models characterized by the distribution parameters in the graphical modeling framework, a different type of deep generative models characterized by neural network parameters in terms of connection matrices was developed mainly for image pixels as the observation data. These were called Deep Belief Networks or DBN [15].

The DBNs have 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 of DBNs comes with the simplicity of not modeling dynamics, and thus not directly suitable for speech modeling.

How to reconcile the pros and cons of these two different types of deep generative models? In order to speed up the investigation in the academic-industrial collaborative work during 2009, our collaborators introduced three "quick-fixes". First, 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. Note this first quick-fix used during 2009–2010 has since been made rigorous by adding recurrence to the DNN [57–59]. And the dynamics of speech at the symbolic level can then be approximately captured by the standard HMM.

The second quick-fix was to reverse the direction of information flow in the deep models – from top-down as in the deep generative model to bottom-up as in the DNN, in order to make inference fast and accurate (given the models). However, it was known by 2009 that with many hidden layers, neural networks were very difficult to train. In order to bypass this problem, the third quick-fix was devised: using a DBN to initialize or pre-train the DNN based on the original proposal of [15]. Note this third quick-fix had been automatically resolved after the earlier DNN was subject to large-data training conducted in industry soon after DNNs showed promising results in small tasks [20, 22, 23, 60]. Careful analyses conducted during 2010 at Microsoft showed that with greater amounts of training data, enabled by GPU-based fast computing, and with more sensible weight initialization without generative pre-training using DBNs [24], the gradient vanishing problem encountered in 1990s no longer plagued the training of DNNs. The same results have also been reported by many other ASR groups subsequently (e.g. [61–63]).

Adopting the above three quick-fixes shaped the deep generative models, rather indirectly, into the DNN-based ASR framework. The initial experimental results using pre-trained DNNs with DBNs showed rather similar phone recognition accuracy to the deep generative model of speech on the standard TIMIT task. The TIMIT data set has been commonly used to evaluate ASR 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 ASR can be more easily analyzed. Such an analysis on TIMIT was conducted at Microsoft Research during 2010 that contrasted the phone recognition accuracy between deep generative models of speech [64] and deep discriminative models including pre-trained DNNs and deep conditional random fields [65–68]. There were a number of very interesting findings. For instance, while the overall phone recognition accuracy is slightly higher for the DNN system, it created many test errors associated with short vocalic phone segments. The errors were traced back to the use of long windows (11–25 frames) for the DNN, creating lots of "noise" for these short phones. Further, the acoustic properties of these vocalic phones are subject to phonetic reduction, not captured by the DNN. However, such phonetic reduction arising from articulatory dynamics is explicitly and adequately modeled by the deep generative model with hidden dynamics, accounting for much lower errors in the short vocalic phones than the DNN as well as the GMM systems that do not capture articulatory dynamics. For most

other classes of phone-like segments in TIMIT, the DNN is doing substantially better than the deep generative model. This type of contrastive error analyses shed insights into distinctive strengths of the two types of deep models. With the highly regular computation and the ease of decoding associated with the DNN-based system, the strengths of the DNN identified by the error analysis stimulated early industrial investment onto deep learning for ASR from small to large scales, eventually leading to its pervasive and dominant deployment today.

The second "quick-fix" above is the only one that has not been resolved as in today's state of the art ASR systems. This direction of future research will be discussed later in this article.

C) Summary

Artificial neural networks have been around for over half a century and their applications to ASR have been almost as long, yet it was not until year 2010 that their real impact had been made by a deep form of such networks, built upon part of earlier work on (shallow) neural nets and (deep) generative models developed by both speech and machine learning communities. A well-timed academic-industrial collaboration between Microsoft and University of Toronto played a central role in introducing DNN technology into the ASR industry. As reviewed above, by 2009 the ASR industry had been searching for new solutions when "principled" deep generative approaches could not deliver what industry needed, both in terms of recognition accuracy and decoding efficiency. In the meantime, academic researchers already developed powerful deep learning tools such as DBNs looking for practical applications [15]. Further, with the advent of general-purpose GPU computing and with Nvidia's CUDA library released in 2008, DBN and DNN computation became fast enough to apply to large speech data. And luckily, by 2009 the ASR community, with the government support since 1980s, had been keenly aware of the importance of large amounts of labeled data, popularized by the axiom "no data is like more data," and had collected more labeled data for training ASR systems than any other discipline. All these enabling factors came in at a perfect time when academic and industrial researchers seized the opportunity and collaborated with each other effectively in the industry setting, leading to the birth of the new era of "deep" speech recognition.

III. ACHIEVEMENTS OF DEEP LEARNING IN SPEECH RECOGNITION

The early experiments discussed in the preceding section on phone recognition and error analysis, as well as on speech feature extraction which demonstrated the effectiveness of using raw spectrogram features [69] had pointed to strong promise and practical value of deep learning. This early progress excited researchers to devote more resources to pursue ASR research using deep learning approaches, the DNN approach in particular. The small-scale ASR experiments were soon expanded to larger scales [21, 22, 25, 26, 60], spreading to the whole ASR industry including major companies of Microsoft, Google, IBM, IflyTech, Nuance, Baidu, etc. [59, 61, 62, 70–79]. The experiments 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 monotonically not only in absolute percentages but also in relative percentages. This is the kind of accuracy improvement not seen in the ASR history. In short, for the DNN-based speech recognizers, the more training data are used, the better the accuracy, the greater word error rate reduction over the GMM counterparts in both absolute and relative terms, and further, the less care required to initialize the DNN. Soon after these experiments at Microsoft were reported, similar findings were published by all major ASR groups worldwide.

Since the initial successful debut of DNNs for speech recognition around 2009–2011, there have been huge progresses made. These progresses, as well as future challenging research directions, are elaborated and summarized into six major areas, each dedicated by a separate subsection below.

A) Output representation learning

Most deep learning methods for ASR have focused on learning representations from input acoustic features without paying attention to output representations. The NIPS Workshop on Learning Output Representations held in December 2013 was dedicated to bridging this gap. The importance of designing effective linguistic representations for the output layers of deep networks for ASR was highlighted in [80]. The most straightforward yet most important example is the use of context-dependent (CD) phone and state units as the DNN output layer, originally invented at Microsoft Research as described in [21, 23]. This type of design for the DNN output representations drastically expands the output neurons from the context-independent phone states with the size of 100–200 commonly used on 1990s to the context-dependent ones with the size in the order of 1000–30 000. 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 further found that due to the significant increase of the HMM output weights and thus the model capacity, 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 industry deployment. Importantly, the design of the big CD-DNN within the traditional framework of HMM decoding requires combined expertise in the DNN and in the large-scale ASR decoder. It also requires industry know-how for constructing very large yet efficient CD units ported to the DNN outputs. It further requires knowledge and skills of how

to make decoding of such huge networks highly efficient using HMM technology and how to cut corner in making practical systems.

For future directions, the output representations for ASR can benefit from more linguistically-guided 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 ASR systems, is inadequate in representing such rich structure (e.g. [81–84]). Such inadequacy thus leaves a promising open door to improve ASR systems' performance. Basic theories about the internal structure of speech sounds and their relevance to ASR in terms of the specification, design, and learning of possible output representations of the underlying speech model for speech target sequences have been surveyed in [85]. The application of this huge body of speech knowledge is likely to benefit deep learning based ASR when deep generative and discriminative models are carefully integrated.

B) Moving towards raw features

One fundamental principle of deep learning is to do away with hand-crafted feature engineering and to use raw features. This principle 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 [69]. Over the past 30 years or so, largely "hand-crafted" transformations of speech spectrogram have led to significant accuracy improvements in the GMM-based HMM systems, despite the known loss of information from the raw speech data. The most successful transformation is the non-adaptive cosine transform, which gave rise to MFCCs. The cosine transform approximately de-correlates feature components, important for the use of GMMs with diagonal covariance matrices. However, when GMMs are replaced by deep learning models such as DNNs, deep belief nets (DBNs), or deep autoencoders, such de-correlation becomes irrelevant due to the very strength of the deep learning methods in modeling data correlation.

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 ASR was already explored in early studies on the GMM-HMM-based systems [86–89]. However, greater speech recognition performance gain is obtained only recently in the recognizers empowered by deep learning methods. For example, Mohamed *et al.* [90] and Li *et al.* [91] showed significantly lowered ASR errors using large-scale DNNs when moving from the MFCC features back to more primitive (Mel-scaled) filter-bank features. These results indicate that DNNs can learn a better transformation than the original fixed cosine transform from the Mel-scaled filter-bank features.

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 – e.g. vocal tract length differences across speakers, distinct speaking styles causing formant undershoot or overshoot, etc. – expressed explicitly in the frequency domain. For example, the convolutional neural network (CNN) can only be meaningfully and effectively applied to ASR [25, 26, 92–94] when spectral features, instead of MFCC features, are used. More recently, Sainath *et al.* [74] went one step further toward raw features by learning the parameters that define the filter-banks on power spectra. That is, rather than using Mel-warped filter-bank features as the input features, the weights corresponding to the Mel-scale filters are only used to initialize the parameters, which are subsequently learned together with the rest of the deep network as the classifier. Substantial ASR error reduction is reported.

Ultimately, deep learning would go all the way to the lowest level of raw features of speech, i.e. speech sound waveforms. As an initial attempt toward this goal, the study carried out by Jaitly and Hinton [95] made use of speech sound waves as the raw input feature to a deep learning system. Although the final results were disappointing, similarly for the earlier work on using speech waveforms in generative model-based ASR [96], the work nevertheless showed that more work is needed along this direction. Most recently, the new use of raw waveforms of speech by DNNs (i.e. zero-feature extraction prior to DNN training) was reported by Tuske *et al.* [97]. The study not only demonstrated the same advantage of learning precise non-stationary patterns of the speech signal localized in time across frame boundaries, but also reported excellent larger-scale ASR results. The most recent study on this topic is reported by Sainath *et al.* [98, 99], where the use of raw waveforms produces highest ASR accuracy when combined with the prior state of the art system.

C) Better optimization

Better optimization criteria and methods are another area where significant advances have been made over the past several years in applying DNNs to ASR. In 2010, researchers at Microsoft recognized the importance of sequence training based on their earlier experience on GMM-HMM the [100–103] and started working on full-sequence discriminative training for the DNN-HMM in phone recognition [65]. Unfortunately, we did not find the right approach to control the overfitting problem. Effective solutions were first reported by Kingsbury *et al.* [104] using Hessian-free training, and then by Su *et al.* [105] and by Vesely *et al.* [106] based on stochastic gradient descent training. These authors developed a set of non-trivial techniques to handle the overfitting problems associated with full-sequence

training of DNN-HMMs, including lattice compensation, frame dropping, and F-smoothing, which are widely used today. Other better and novel optimization methods include distributed asynchronous stochastic gradient descent [70, 72], primal-dual method for applying natural parameter constraints [107], and Bayesian optimization for automated hyper-parameter tuning [108].

D) A new level of noise robustness

Research into noise robustness in ASR has a long history, mostly before the recent rise of deep learning. A wide range of noise-robust techniques developed over past 30 years can be analyzed and categorized using five different criteria: (1) feature-domain versus model-domain processing, (2) the use of prior knowledge about the acoustic environment distortion, (3) the use of explicit environment-distortion models, (4) deterministic versus uncertainty processing, and (5) the use of acoustic models trained jointly with the same feature enhancement or model adaptation process used in the testing stage. See a comprehensive review in [109, 110] and additional review literature or original work in [111–114].

The model-domain techniques developed for GMM-HMMs are often not applicable to the new DNN models for ASR. The difficulty arises primarily due to the differences between generative models that GMMs belong to and discriminative models that DNNs belong to. The feature-domain techniques, however, can be more directly applied to the DNN system. A detailed investigation of the use of DNNs for noise robust speech recognition in the feature domain was reported by Seltzer *et al.* [115], who applied the C-MMSE [102, 103] feature enhancement algorithm on 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. More recently, Kashiwagi *et al.* [116] applied the SPLICE feature enhancement technique [117] to a DNN speech recognizer, where the DNN's output layer was determined on clean data instead of on noisy data as in the study reported by Seltzer *et al.* [115].

Recently, a series of studies were reported by Huang *et al.* [118] comparing GMMs and DNNs on the mobile voice search and short message dictation datasets. These data were collected through real-world applications used by millions of users with distinct speaking styles in diverse acoustic environments. A pair of state-of-the-art GMM and DNN models was trained using 400 h of VS/SMD data. The two models shared the same training data and decision tree. The same GMM seed model was used for the lattice generation in the GMM and the senone state alignment in the DNN. Under such carefully controlled conditions, the experimental results showed that the DNN-based system yields uniform performance gain over the GMM counterpart across a wide range of SNR levels on all types of datasets and acoustic environments. That is, the use of DNNs raises the performance of noise-robust ASR to a new level. However, this study, the most comprehensive in the noise-robust DNN-based ASR literature so far, also suggests that noise robustness remains an important research area and techniques such as speech enhancement, noise robust acoustic features, or other multi-condition learning methods need to be further explored in the DNN setup.

In the most recent study on noise-robust ASR using deep learning, Hannun *et al.* [77] reported an interesting brute-force approach based on "data augmentation." It is intriguing to see how deep learning, deep recurrent neural nets in particular, make the problem solution conceptually much easier than other approaches discussed above. That is, simply throw in very large amounts of synthesized or "superpositioned" noisy data that capture the right kinds of variability controlled by the synthesis process. The efficient parallel training system was used to training deep speech models with as many as 100 000 h of such synthesized data and produced excellent results. The challenge for this brute-force approach is to efficiently represent the combinatorially growing size of a multitude of distortion factors known to corrupt speech acoustics under real-world application environments.

Noise robust ASR is raised to a new level in the DNN era. For other notable work in this area, see [119–121].

E) Multi-task and transfer learning

In the area of ASR, the most interesting application of multi-task learning is multi-lingual or cross-lingual ASR, where ASR for different languages is considered as different tasks. 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 [122]. Another successful approach for the GMM-HMM is to map pronunciation units across languages either via knowledge-based or data-driven methods [123]. For the more recent, DNN-based systems, these multi-task learning applications in ASR are much more successful.

In the studies reported by Huang *et al.* [124] and Heigold *et al.* [125], two research groups independently developed closely related DNN architectures with multi-task learning capabilities for multilingual speech recognition. The idea is that the hidden layers in the DNN, when learned appropriately, serve as increasingly complex feature transformations sharing common hidden factors across the acoustic data in different languages. The final softmax layer representing a log-linear classifier makes use of the most abstract feature vectors represented in the top-most hidden layer. While the log-linear classifier is necessarily separate for different languages, the feature transformations can be shared across languages. Excellent multilingual speech recognition results were reported. The implication of this set of work is significant and far reaching. It points to the possibility of quickly building a high-performance DNN-based system for a new language from an existing multilingual DNN. This huge benefit requires only a small amount of training data from the target language, although having more

data would further improve the performance. This multitask learning approach can reduce the need for the unsupervised pre-training stage, and can train the DNN with much fewer epochs. Extension of this set of work would be to efficiently build a language-universal speech recognition system. Such a system will not only recognize many languages and improve the accuracy for each individual language, but also expand the languages supported by simply stacking softmax layers on the DNN for new languages.

More recently, the power of multitask learning with DNN is demonstrated in improved ASR accuracy in difficult reverberated acoustic environments [126].

F) Better architectures

The tensor version of the DNN was reported by Yu *et al.* [127, 128] and showed substantially lower ASR errors compared with the conventional DNN. It extends the DNN by replacing one or more of its layers with a double-projection layer and a tensor layer. In the double-projection layer, each input vector is projected into two non-linear subspaces. In the tensor layer, two subspace projections interact with each other and jointly predict the next layer in the overall deep architecture. An approach is developed to map the tensor layers to the conventional sigmoid layers so that the former can be treated and trained in a similar way to the latter.

The DNN and its tensor version are fully connected. Locally connected architectures, or (deep) CNNs, have each CNN module consisting of a convolutional layer and a pooling layer. The convolutional layer shares weights, and the pooling layer subsamples the output of the convolutional layer and reduces the data rate from the layer below. 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 ASR [25, 26, 62, 92–94, 129]. Note that the time-delay neural network (TDNN, [2]) developed for early days of ASR is a special case and predecessor of the CNN when weight sharing is limited to one of the two dimensions, and there is no pooling layer. It was not until recently that researchers have discovered that the time-dimension invariance is less important than the frequency-dimension invariance for ASR [92, 93].

Another important deep architecture is the (deep) RNN, especially its long short-term memory (LSTM) version. The LSTM was reported to give the lowest error rate on the benchmark TIMIT phone recognition task [57]. 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 [71, 72]. 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 is performed across hundreds of machines in CPU clusters. The best accuracy by year 2014 was obtained by optimizing the frame-level cross-entropy objective function followed by sequence discriminative training [72]. More recently, the use of CTC objective function in the deep LSTM system training further improves the recognition accuracy [59, 130].

When the LSTM model is fed by the output of a CNN and then feeds into a fully connected DNN, the entire architecture becomes very deep, and is called the CLDNN. This architecture leverages complementary modeling capabilities of three types of neural nets, and is demonstrated to be more effective than each of the neural net types including the highest performing LSTM [98, 99].

While the DNN-HMM has significantly outperformed the GMM-HMM, recent studies investigated a novel "deep GMM" architecture, where a GMM is transformed to a large softmax layer followed by a summation pooling layer [131, 132]. Theoretical and experimental results show that the deep GMM performs competitively with the DNN-HMM.

Another set of novel deep architectures, which are quite different from the standard DNN, are reported in [133–135] for successful ASR and related applications including speech understanding. These models are exemplified by the deep stacking network (DSN), its tensor variants [136, 137], and its kernel version [138]. 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 [139].

The initial motivation of concatenating the original input vector with the output vector of each DSN module as the new input vector for each higher DSN module was to avoid loss of information when building up higher and higher modules in this deep model. Due to the largely convex learning problem formulated for the DSN training, such concatenation makes training errors (nearly) always decrease as each new module is added to the DSN. It turns out that such concatenation is also a natural consequence of another type of deep architecture, called deep unfolding nets [140]. These nets are constructed by stacking a number of (shallow) generative models based on non-negative matrix factorization. This stacking process, called unfolding, follows the inference algorithm applied to the original shallow generative model, which determines the non-linear activation function and also naturally requires the original input vector as part of the inference algorithm. Importantly, this type of deep stacking or unfolding models allow the problem-domain knowledge to be built into the model, which DNNs with generic architectures consisting of weight matrices and fixed forms of nonlinear units would have greater difficulties in incorporating knowledge.

An example of problem-domain knowledge discussed above in the area of speech processing is how noise speech is formed from clean speech and noise. Another example in the area of language processing is how (hidden) topics can be generated from words in text. Note that in these

examples, the domain knowledge of generative type can be parameterized naturally by matrices in the same way that DNNs are parameterized. This enables similar kinds of DNN learning algorithms to apply to fine-tuning the deep stacking or unfolding nets in a straightforward manner. When the generative models cannot be naturally parameterized by matrices, e.g. the deep generative models of speech with temporal dynamics discussed in Section II.A, how to incorporate such knowledge in integrated deep generative and discriminative models is a challenging research direction. That is, the second "quick-fix" discussed in Section II.B has yet to be overcome in more general settings than those when the deep generative models have not been parameterized by dense matrices and common non-linear functions. Further, when the original generative model moves from shallow to deep, as in the hidden dynamic models discussed in Section II.A, the inference algorithm itself becomes computationally complex and requires various kinds of approximation; e.g. variational inference. How to build deep unfolding models and to carry out discriminative fine-tuning using backpropagation becomes a more challenging task.

G) Summary

Six main areas of achievements and progresses of deep learning in ASR after the initial success of the pre-trained DNN are surveyed in this section. Due to the space limit, several other important areas of progresses are not included here, including adaptation of DNNs for speakers [127, 141], better regularization methods, better non-linear units, speedup of DNN training and decoding, tensor-based DNNs [128, 141], exploitation of sparseness in DNNs [139], and understanding the underlying mechanisms of DNN feature processing.

In summary, large-scale ASR is the first and the most convincing successful case of deep learning in the recent history, embraced by both industry and academia across the board. Between 2010 and 2015, the two major conferences on signal processing and ASR, IEEE-ICASSP and Interspeech, have seen near exponential growth in the numbers of accepted papers in their respective annual conferences on the topic of deep learning for ASR. More importantly, all major commercial ASR systems (e.g. Microsoft Cortana, Xbox, Skype Translator, Google Now, Apple Siri, Baidu and iFlyTek voice search, and a range of Nuance speech products, etc.) nowadays are based on deep learning methods, the best evidence of high achievements of deep learning in ASR.

In addition to ASR, deep learning is also creating high impact in image recognition (e.g. [142]) and in speech synthesis (e.g. [143]), as well as in spoken language understanding [144, 145]. A related major area with perhaps more important practical applications, where deep learning has the potential to make equally strong achievements but where special challenges are lying ahead, will be discussed and analyzed in the next section.

IV. DEEP LEARNING FOR NATURAL LANGUAGE AND MULTIMODAL PROCESSING

ASR involves the inference from low-level or raw speech waves to high-level linguistic entities such as word sequences. Image recognition involves the inference from low-level pixels to high-level semantic categories. Due to the reasonably well understood hierarchical, layered structure of human speech and visual perception systems, it is easy to appreciate why deep learning can do so well in ASR and image recognition.

For natural language processing (NLP) and multimodal processing involving language, the raw signal often starts with words, which already embody rich semantic information. As of this writing, one has not observed as striking achievements of deep learning in natural language and multimodal processing as in speech and image recognition, and huge challenges lie ahead. However, strong research activities have been taking place in recent years. In this section, a selected review is provided on some of these progresses.

A) A selected review on deep learning for NLP

Over the past few years, deep learning methods based on neural nets have been shown to perform well on various NLP tasks such as language modeling, machine translation, part-of-speech tagging, named entity recognition, sentiment analysis, and paraphrase detection, as well as NLP-related tasks involve user behaviors such as computational advertising and web search (informational retrieval). The most attractive aspect of deep learning methods is their ability to perform these tasks without external hand-designed resources or feature engineering. To this end, deep learning develops and makes use of an important concept called "embedding". That is, each linguistic entity (e.g. word, phrase, sentence, paragraph, or a full text document), a physical entity, a person, a concept, or a relation, which is often represented as a sparse, high-dimensional vector in the symbolic space, can be mapped into a low-dimensional, continuous-space vector via distributed representations by neural nets [146–149]. In the most recent work, such "point" embedding has been generalized to "region" embedding or Gaussian embedding [150].

Use of deep learning techniques in machine translation, one most important task in NLP applications, has recently attracted much attention. In [151, 152], the phrase-translation component in a machine translation system is replaced by a set of DNNs with semantic phrase embeddings. A pair of source and target phrases is projected into continuous-valued vector representations in a low-dimensional latent semantic space. Their translation score is then computed by cosine distance between the pair in this new space. In a more recent study, a deep RNN with LSTM cells are used to encode the source sentence into a fixed-length embedding vector, which excites another deep RNN as the decoder that generates the target sentence

[153]. Most recently, Bahdanau *et al.* [154] reported a neural machine translation approach that learns to align and translate jointly, where the earlier encoder-decoder architecture is extended by allowing a soft search, called the "attention mechanism," for parts of source sentence relevant to predicting a target word with no need for explicit segmentation.

Another important NLP-related task is knowledgebase completion, instrumental in question-answering and other NLP applications. In [155], a simple method (TransE) was proposed which models relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. More recent work [149] adopts an alternative approach, based on the use of neural tensor networks, to attack the problem of reasoning over a large joint knowledge graph for relation classification. The most recent work [156] generalizes these earlier models to a unified learning framework, where entities are represented as low-dimensional dense vectors learned from a neural network and relations are represented by bilinear and/or linear mapping functions. For the NLP problem of question answering, a most recent and highly visible deep learning approach is proposed in [157] using memory networks, which use a long-term memory as a dynamic knowledge base. The output of the memory network forms the text response to input questions to the network.

Information retrieval is another important area of NLP applications, where a user enters a keyword or natural language query into the automated computer system that contains a collection of many documents with the goal of obtaining a set of most relevant documents. Web search is a large-scale information retrieval task on largely unstructured web data. Since 2013, Microsoft Research has successfully developed and applied a specialized deep learning architecture, called deep-structured semantic model or deep semantic similarity model (DSSM; [158]) and its convolutional version (C-DSSM; [159, 160]), to web search and related tasks. The DSSM uses the DNN architecture to capture complex semantic properties of the query and the document, and to rank a set of documents for a given query. Briefly, a non-linear projection is performed first to map the query and the documents to a common semantic space. Then, the relevance of each document given the query is calculated as the cosine similarity between their vectors in that semantic space. The DNNs are trained using the clickthrough data such that the conditional likelihood of the clicked document given the query is maximized. The DSSM is optimized directly for Web document ranking exploiting distantly supervised signals, and thus gives strong performance. Furthermore, to deal with large vocabularies in Web search applications, a new *word hashing* method is developed, through which the high-dimensional term vectors of queries or documents are projected to low-dimensional letter-based n-gram vectors.

More recently, the DSSM has been further developed and successfully applied to online ads selection and placement (unpublished), to multitask learning involving both semantic classification and information retrieval tasks [161], to entity ranking in an Microsoft Office application [162], and to automatic image captioning [163]. The latter is a currently trendy multimodal processing task involving natural language, which will be discussed shortly in the next subsection.

B) A selected review on deep learning for multimodal processing

Multimodal processing is a class of applications closely related to multitask learning, where the learning domains or "tasks" cut across more than one modalities for practical applications that embrace a mix of natural language, image/video, audio/speech, touch, and gesture. As evidenced in the successful cases of ASR described in Section III.E, multitask leaning fits very well to the paradigm of deep representation learning where the shared representations and statistical strengths across tasks (e.g. those involving separate modalities of audio, image, touch, and text) is expected to greatly facilitate many machine learning scenarios under low-resource conditions. Before deep learning methods were adopted, there had already been numerous efforts in multimodal and multitask learning. For example, a prototype called MiPad for multimodal interactions involving capturing, leaning, coordinating, and rendering a mix of speech, touch, and visual information was developed and reported in [113, 164]. In [165, 166], mixed sources of information from multiple-sensory microphones with separate bone-conductive and air-born paths were exploited to de-noise speech. These early studies all used shallow models and achieved worse than desired performance. With the advent of deep learning, it is hopeful that the difficult multi-modal learning problems can be solved with eventual success to enable a wide range of practical applications.

The deep architecture of DeViSE (Deep Visual-Semantic Embedding), developed by Frome *et al.* [167], is a typical example of multimodal learning where text information is used to improve the image recognition system. In this system, the loss function used in the training adopts a combination of dot-product similarity and max-margin, hinge rank loss. This is closely related to the cosine distance or maximum-mutual information based loss function used for training the DSSM model in [158] described in Section IV.A. The results show that the information provided by text significantly improves zero-shot image predictions, achieving excellent hit rates across thousands of the labels never seen by the image model.

One of the most interesting applications of deep learning methods to multimodal processing appeared recently in November 2014, when several groups almost simultaneously publicized their work on automatic image captioning in ArXiv, all to be revised and officially published at the CVPR-2015 conference. In the Microsoft system [163], the image is first broken down into a number of regions likely to be objects, and then a deep CNN is applied to each region to generate a high-level feature vector to capture relevant visual information. The resulting bag of words

is then put together using a language model to produce a set of likely candidate sentences. They are subsequently ranked by the DSSM which captures global semantics of the caption sentence about the image and produces the final answer. Baidu's approach is based on a multimodal RNN that generates novel sentence descriptions to explain the image's content [168]. Google's paper [169], and Stanford's paper [170] described two conceptually similar systems, both based on multimodal RNN generative models conditioned on the image embedding vectors at the first time step. University of Toronto [171] reported a system pipeline that is based on multimodal neural language models that are unified with visual-semantic embeddings produced by the deep CNN. All these systems were evaluated using the common MSR's COCO database, and thus upon final systems' refinement the results of these different systems can be compared.

C) Summary

The goal of NLP is to analyze, understand, and generate languages that humans use naturally, and NLP is also a critical component of multimodal systems. Significant progress in NLP has been achieved in recent years, addressing important and practical real-world problems. Deep learning based on embedding methods has contributed to such progress. Words in sequence are traditionally treated as discrete symbols, and deep learning provides continuous-space vector representations that describe words and their semantic and syntactic relationships in a distributed manner permitting meaningfully defined similarity measures. Practical advantages of such representations include natural abilities to mitigate data sparseness, to incorporate longer contexts, and to represent morphological, syntactic and semantic relationships across words and larger linguistic entities. The several NLP and multimodal applications reviewed in this section have all been grounded on vector-space embeddings for the distributed representation of words and larger units as well as of the relations among them. In particular, in multimodal processing, all types of signals – image, voice, text – are projected into the same semantic vector space in the deep learning framework, greatly facilitating their comparison, integration, and joint processing. The representation power of such flat vectors based on neural networks in contrast with symbolic tree-like structure in NLP is currently under active investigation by deep learning, NLP, and cognitive science researchers (e.g. [172]).

V. CONCLUSIONS AND CHALLENGES FOR FUTURE WORK

This article reviews part of the history on neural networks and (deep) generative modeling, and reflects on the path to the current triumph of applying deep neural nets to speech recognition, the first successful case of deep learning at industry scale. The roles of generative models have been analyzed in the review, pointing out that the key advantages of embedding knowledge about speech dynamics that are naturally enabled by deep generative modeling have yet to be incorporated as part of the new-generation deep learning framework.

For speech recognition, one remaining future challenge lies in how to effectively integrate major relevant speech knowledge and problem constraints into new deep models of the future. Examples of such knowledge and constraints would include distributed, feature-based phonological representations of sound patterns of language via hierarchical structure based on modern phonology, articulatory dynamics, and motor program control, acoustic distortion mechanisms for the generation of noisy, reverberant speech in multi-speaker environments, Lombard effects caused by modification of articulatory behavior due to noise-induced reduction of communication effectiveness, and so on. Deep generative models are much better able to impose the problem constraints above than purely discriminative DNNs. These deep generative models should be parameterized to facilitate highly regular, matrix-centric, large-scale computation in order to take advantage of modern high-efficiency GPGPU computing already demonstrated to be extremely fruitful for DNNs. The design of the overall deep computational network architecture of the future may be motivated by approximate inference algorithms associated with the initial generative model. Then, discriminative learning algorithms such as backpropagation can be developed and applied to learn all network parameters (i.e. large matrices) in an end-to-end fashion. Ultimately, the run-time computation follows the inference algorithm in the generative model, but the parameters have been learned to best discriminate all classes of speech sounds. This is akin to discriminative learning for GMM-HMMs, but now with much more powerful deep architectures and with more comprehensive ways of incorporating speech knowledge. The discriminative learning will be much more powerful (via backprop through the entire deep structure) than the earlier discriminative learning on shallow architectures of GMM-GMMs that relied on extended Baum–Welch algorithm [100].

The past several years of deep learning research and practical applications have established that for perceptual tasks such as speech and image recognition, DNN-like discriminative models perform extremely well and scale beautifully with large amounts of strongly supervised training data. Some remaining issues would include: (1) What will be the limit for growing recognition accuracy with respect to further increasing amounts of labeled data? (2) Beyond this limit or when labeled data become exhausted or non-economical to collect, what kind of novel unsupervised or semi-supervised deep learning architectures will emerge? Deep generative models which can naturally handle unlabeled training data appear well suited for meeting this challenge. It is expected that within next 4–5 years, the above issues will be resolved and rapid progress will be made to enable more impressive application scenarios, such as analyzing videos and then telling stories about them by a machine.

For the more difficult and challenging cognitive tasks – natural language, multimodal processing, reasoning, knowledge, attention, memory, etc. – deep learning researchers so far have not found as much low-hanging fruit as for the perceptual tasks of speech and image above, and the views for the future development are somewhat less clear. Nevertheless, solid progress has been made over past several years, as we selectively reviewed in Section IV of this paper. If successful, the revolution to be created by deep learning for the cognitive tasks will be even more impactful than the revolution in speech and image recognition we have seen so far. Important issues to be addressed and the technical challenges for future developments would include: (1) Will supervised deep learning, which applies to NLP tasks like machine translation, significantly beat the state of the art currently still held by dominant NLP methods as for speech and image recognition tasks? (2) How do we distill and exploit "distant" supervision signals for (weakly) supervised deep learning in the NLP, multimodal and other cognitive tasks? and (3) Will flat dense-vector embedding with distributed representations, which is the backbone of much of the deep learning methods for language as discussed in Section IV, be sufficient for general tasks involving natural language that is known to possess rich tree-like structure? That is, do we need to directly encode and recover syntactic and semantic structure of natural language?

Tackling NLP problems with the deep learning scheme based on embedding may become more promising when the problems are part of wider big-data analytic applications, where not only words and other linguistic entities but also business activities, people, events, and so on may be embedded into the unified vector space. Then the "distant" supervision signals may be mined with broader context than what we discussed in Section IV for text-centric tasks alone. For example, an email from a sender to a receiver with the email subject line, email body, and possible attachments would readily establish such supervision signals that relate different people in connection with different levels of detail of natural language data. With large amounts of such business-analytic data available including a wealth of weakly supervised information, deep learning is expected to play important roles in a wider range of applications than we have discussed in the current article.

REFERENCES

[1] Toshiteru, H.; Atlas, L.; Marks, R.: An artificial neural network for spatio-temporal bipolar patters: application to phoneme classification, in *Proc. NIPS*, 1988.

[2] Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K.: Phoneme recognition using time-delay neural networks. IEEE Trans. Acoust. Speech Sig. Proc., 37 (1989), 328–339.

[3] Bengio, Y.: Artificial Neural Networks and Their Application to Speech and Sequence Recognition, Ph.D. Thesis, McGill University, Montreal, Canada, 1991.

[4] Robinson, A.: An application of recurrent nets to phone probability estimation. IEEE Trans. Neural Networks, 5 (1994), 298–305.

[5] Bourlard, H.; Morgan, N.: Connectionist Speech Recognition: A Hybrid Approach, Kluwer, Norwell, MA, 1993.

[6] Hermansky, H.; Ellis, D.; Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems, in *Proc. ICASSP*, 2000.

[7] Baker, J. *et al.*: Research developments and directions in speech recognition and understanding. IEEE Sig. Proc. Mag., 26 (3) (2009), 75–80.

[8] Baker, J. *et al.*: Updated MINS report on speech recognition and understanding. IEEE Sig. Proc. Mag., 26 (4), 2009a.

[9] Bengio, Y.; Simard, P.; Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw., 5 (1994), 157–166.

[10] Deng, L.; Hassanein, K.; Elmasry, M.: Analysis of correlation structure for a neural predictive model with application to speech recognition. IEEE Trans. Neural Networks, 7 (2) (1994), 331–339.

[11] Deng, L.; Ramsay, G.; Sun, D.: Production models as a structural basis for automatic speech recognition. Speech Commun., 33 (2–3) (1997), 93–111.

[12] Bridle, J. *et al.*: 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.

[13] Picone, P. *et al.*: Initial evaluation of hidden dynamic models on conversational speech, in *Proc. ICASSP*, 1999.

[14] Hinton, G.; Dayan, P.; Frey, B.; Neal, R.: The wake-sleep algorithm for unsupervised neural networks. Acta Crystallogr. Sect. B, Struct. Sci., 268 (1995), 1158–1161.

[15] Hinton, G.; Osindero, S.; Teh, Y.: A fast learning algorithm for deep belief nets. Neural Comp., 18 (2006), 1527–1554.

[16] Jordan, M.; Bishop, C.: Neural networks. Comp. Surveys, 28 (1996), 73–75.

[17] Jordan, M.; Ghahramani, Z.; Jaakkola, T.; Saul, L.: An introduction to variational methods for graphical models. Machine Learning, 37 (1999), 183–233.

[18] Bishop, C.: Neural Networks and Pattern Recognition, Clarendon Press, Oxford, 1995.

[19] Bishop, C.: Pattern Recognition and Machine Learning, Springer, Oxford, England, 2006.

[20] Hinton, G. *et al.*: Deep neural networks for acoustic modeling in speech recognition. IEEE Sig. Process. Mag., 29 (6) (2012), 82–97.

[21] Dahl, G.; Yu, D.; Deng, L.; Acero, A.: Context-dependent DBN-HMMs in large vocabulary continuous speech recognition, in *Proc. ICASSP*, 2011.

[22] Dahl, G.; Yu, D.; Deng, L.; Acero, A.: Context-dependent, pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Proc., 20 (1) (2012), 30–42.

[23] Yu, D.; Deng, L.; Dahl, G.E.: Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition, in *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, December 2010.

[24] Yu, D.; Deng, L.; Li, G.; Seide, F.: Discriminative pretraining of deep neural networks, U.S. Patent Filing, November 2011.

[25] Deng, L.; Abdel-Hamid, O.; Yu, D.: A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion, in *Proc. ICASSP*, 2013.

[26] Deng, L. *et al.*: Recent advances in deep learning for speech research at Microsoft, in *Proc. ICASSP*, 2013b.

[27] Deng, L.; Hinton, G.; Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: an overview, in *Proc. ICASSP*, 2013a.

[28] Yu, D.; Deng, L.: Automatic Speech Recognition – A Deep Learning Approach, Springer, London, England, 2014.

[29] Deng, L.: Dynamic Speech Models – Theory, Algorithm, and Application, Morgan & Claypool, December 2006.

[30] Deng, L.: A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. Sig. Process., 27 (1) (1992), 65–78.

[31] Ostendorf, M.; Digalakis, V.; Kimball, O.: From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. Speech Audio Proc., 4 (5) (1996), 360–378.

[32] Deng, L.; Aksmanovic, M.: Speaker-independent phonetic classification using hidden Markov models with state-conditioned mixtures of trend functions. IEEE Trans. Speech Audio Process., 5 (1997), 319–324.

[33] Chengalvarayan, R.; Deng, L.: Speech trajectory discrimination using the minimum classification error learning. IEEE Trans. Speech Audio Process., 6 (6) (1998), 505–515.

[34] Yu, D.; Deng, L.; Gong, Y.; Acero, A.: A novel framework and training algorithm for variable-parameter hidden Markov models. IEEE Trans. Audio Speech Lang. Process., 17 (7) (2009), 1348–1360.

[35] Yu, D.; Deng, D.; Wang, S.: Learning in the deep-structured conditional random fields, in *NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009b.

[36] Ling, Z.; Deng, L.; Yu, D.: Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis, in *ICASSP*, 2013, pp. 7825–7829.

[37] Ling, Z.; Deng, L.; Yu, D.: Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. IEEE Trans. Audio Speech Lang. Process., 21 (10) (2013), 2129–2139.

[38] Deng, L.: A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Commun., 24 (4) (1998), 299–323.

[39] Deng, L.: Computational models for speech production, in Computational Models of Speech Pattern Processing, pp. 199–213, Springer Verlag, Berlin, 1999.

[40] Togneri, R.; Deng, L.: Joint state and parameter estimation for a target-directed nonlinear dynamic system model. IEEE Trans. Sig. Process., 51 (12) (2003), 3061–3070.

[41] Seide, F.; Zhou, J.; Deng, L.: Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM: MAP decoding and evaluation, in *Proc. ICASSP*, 2003.

[42] Zhou, J.; Seide, F.; Deng, L.: Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM: modeling and training, in *Proc. ICASSP*, 2003.

[43] Ma, J.; Deng, L.: A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamical model of speech. Computer Speech Lang., 2000, 101–114.

[44] Ma, J.; Deng, L.: Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model. IEEE Trans. Speech Audio Process., 11 (6) (2003), 590–602.

[45] Ma, J.; Deng, L.: Target-directed mixture dynamic models for spontaneous speech recognition. IEEE Trans. Speech Audio Process., 12 (1) (2004), 47–58.

[46] Deng, L.; Yu, D.; Acero, A.: Structured speech modeling. IEEE Trans. on Audio Speech Lang. Process., 14 (5) (2006), 1492–1504.

[47] Deng, L.; Ma, J.: Spontaneous speech recognition using a statistical coarticulatory model for the vocal tract resonance dynamics. J. Acoust. Soc. Am., 108 (2000), 3036–3048.

[48] Deng, L.: Switching dynamic system models for speech articulation and acoustics. In Mathematical Foundations of Speech and Language Processing (Eds Johnson, M.; Khudanpur, S.P.; Ostendorf, M.; Rosenfeld, R.), Springer-Verlag, New York, 2003, 115–134.

[49] Bilmes, J.: Dynamic graphical models. IEEE Sig. Process. Mag., 33 (2010), 29–42.

[50] Lee, L.; Attias, H.; Deng, L.: Variational inference and learning for segmental state space models of hidden speech dynamics, in *Proc. ICASSP*, 2003.

[51] Lee, L.; Attias, H.; Deng, L.; Fieguth, P.: A multimodal variational approach to learning and inference switching state space models, in *Proc. ICASSP*, 2004.

[52] Deng, L.; Togneri, R.: 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, T.; Togneri, R.; Narasimha, M.S.), Springer, New York, 2014, 153–196.

[53] Deng, L.; Yu, D.; Acero, A.: A bidirectional target filtering model of speech coarticulation: two-stage implementation for phonetic recognition. IEEE Trans. Audio Speech Process., 14 (1) (2006a), 256–265.

[54] Bazzi, I.; Deng, I.; Acero, I.: Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. IEEE Trans. on Audio Speech Lang. Process., 14 (2) (2006), 425–434.

[55] Deng, L.; Attias, H.; Lee, L.; Acero, A.: Adaptive Kalman smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model. IEEE Trans. Audio Speech Lang. Process., 15 (1) (2007), 13–23.

[56] Deng, L.; Cui, X.; Pruvenok, R.; Huang, J.; Momen, S.; Chen, Y.; Alwan, A.: A database of vocal tract resonance trajectories for research in speech processing, in *Proc. ICASSP*, 2006b.

[57] Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks, in *Proc. ICASSP*, 2013.

[58] Deng, L.; Chen, J.: Sequence classification using the high-level features extracted from deep neural networks, in *Proc. ICASSP*, 2014.

[59] Sak, H. *et al.*: Learning acoustic frame labeling for speech recognition with recurrent neural networks, in *Proc. ICASSP*, 2015a.

[60] Seide, F.; Li, G.; Yu, D.: Conversational speech transcription using context-dependent deep neural networks, in *Proc. Interspeech*, 2011, pp. 437–440.

[61] Sainath, T.; Kingsbury, B.; Ramabhadran, B.; Novak, P.; Mohamed, A.: Making deep belief networks effective for large vocabulary continuous speech recognition, in *Proc. ASRU*, 2011.

[62] Sainath, T.; Kingsbury, B.; Soltau, H.; Ramabhadran, B.: Optimization techniques to improve training speed of deep neural networks for large speech tasks. IEEE Trans. Audio Speech Lang. Process., 2111 (2013), 2267–2276.

[63] Jaitly, N.; Nguyen, P.; Senior, A.; Vanhoucke, V.: Application of pretrained deep neural networks to large vocabulary speech recognition, in *Proc. Interspeech*, 2012.

[64] Deng, L.; Yu, D.: Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition, in *Proc. ICASSP*, 2007.

[65] Mohamed, A.; Yu, D.; Deng, L.: Investigation of full-sequence training of deep belief networks for speech recognition, in *Proc. Interspeech*, 2010.

[66] Mohamed, A.; Dahl, G.; Hinton, G.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process., 20 (1), 2012 (the short conference version of this paper was presented at the 2009 NIPS Workshop).

[67] Yu, D.; Deng, L.: Learning in the deep-structured hidden conditional random fields, in *NIPS Workshop on Deep Learning for Speech Recognition*, 2009.

[68] Yu, D.; Deng, L.: Deep-structured hidden conditional random fields for phonetic recognition, in *Proc. Interspeech*, September 2010.

[69] Deng, L.; Seltzer, M.; Yu, D.; Acero, A.; Mohamed, A.; Hinton, G.: Binary coding of speech spectrograms using a deep autoencoder, in *Proc. Interspeech*, 2010.

[70] Dean, J. *et al.*: Large scale distributed deep networks, in *Proc. NIPS*, 2012.

[71] Sak, H.; Senior, A.; Beaufays, F.: Long short-term memory recurrent neural network architecturesfor large scale acoustic modeling, in *Proc. Interspeech*, 2014.

[72] Sak, H. *et al.*: Sequence discriminative distributed training of long short-term memory recurrent neural networks, in *Proc. Interspeech*, 2014a.

[73] Sainath, T.; Mohamed, A.; Kingsbury, B.; Ramabhadran, B.: Convolutional neural networks for LVCSR, in *Proc. ICASSP*, 2013a.

[74] Sainath, T.; Kingsbury, B.; Mohamed, A.; Ramabhadran, B.: Learning filter banks within a deep neural network framework, in *Proc. ASRU*, 2013b.

[75] Sainath, T.; Kingsbury, B.; Sindhwani, V.; Arisoy, E.; Ramabhadran, B.: Low-rank matrix factorization for deep neural network training with high-dimensional output targets, in *Proc. ICASSP*, 2013c.

[76] Saon, G.; Soltau, H.; Nahamoo, D.; Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors, in *Proc. ASRU*, 2013.

[77] Hannun, A. *et al.*: Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567, 2014.

[78] Yu, D.; Seide, F.; Li, G.; Deng, L.: Exploiting sparseness in deep neural networks for large vocabulary speech recognition, in *Proc. ICASSP*, 2012b.

[79] Yu, D.; Deng, L.: Deep learning and its applications to signal and information processing. IEEE Sig. Process. Mag., 28 (2011), 145–154.

[80] Deng, L.: Design and learning of output representations for speech recognition, in *NIPS Workshop on Learning Output Representations*, December 2013.

[81] Deng, L.; Erler, K.: Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: comparison with segmental speech units. J. Acoust. Soc. Am., 92 (6) (1992), 3058–3067.

[82] Kirchhoff, K.: Syllable-level desynchronisation of phonetic features for speech recognition, in *Proc. ICSLP*, 1996.

[83] Ostendorf, M.: Moving beyond the 'beads-on-a-string' model of speech, in *Proc. ASRU*, 1999.

[84] Sun, J.; Deng, L.: An overlapping-feature based phonological model incorporating linguistic constraints: applications to speech recognition. J. Acoust. Soc. Am., 111 (2002), 1086–1101.

[85] Deng, L.; O'Shaughnessy, D.: SPEECH PROCESSING – A Dynamic and Optimization-Oriented Approach, Marcel Dekker, New York, 2003.

[86] Chengalvarayan, R.; Deng, L.: HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features. IEEE Trans. Speech Audio Process., 5 (1997), 243–256.

[87] Chengalvarayan, R.; Deng, L.: Use of generalized dynamic feature parameters for speech recognition. IEEE Trans. Speech Audio Process., 5 (1997a), 232–242.

[88] Rathinavalu, C.; Deng, L.: Construction of state-dependent dynamic parameters by maximum likelihood: applications to speech recognition. Sig. Process., 55 (2) (1997), 149–165.

[89] Biem, A.; Katagiri, S.; McDermott, E.; Juang, B.: An application of discriminative feature extraction to filter-bank-based speech recognition. IEEE Trans. Speech Audio Process., 9 (2001), 96–110.

[90] Mohamed, A.; Hinton, G.; Penn, G.: Understanding how deep belief networks perform acoustic modelling, in *Proc. ICASSP*, 2012a.

[91] Li, J.; Yu, D.; Huang, J.; Gong, Y.: Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM, in *Proc. IEEE SLT*, 2012.

[92] Abdel-Hamid, O.; Mohamed, A.; Jiang, H.; Penn, G.: Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, in *ICASSP*, 2012.

[93] Abdel-Hamid, O.; Deng, L.; Yu D.: Exploring convolutional neural network structures and optimization for speech recognition, in *Interspeech*, 2013.

[94] Abdel-Hamid, O.; Deng, L., Yu, D., Jiang, H.: Deep segmental neural networks for speech recognition, in *Interspeech*, 2013a.

[95] Jaitly, N.; Hinton, G.: Learning a better representation of speech sound waves using restricted Boltzmann machines, in *Proc. ICASSP*, 2011.

[96] Sheikhzadeh, H.; Deng, L.: Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization. IEEE Trans. Speech Audio Process., 2 (1994), 80–91.

[97] Tüske, Z.; Golik, P.; Schluter, R.; Ney, H.: Acoustic modeling with deep neural networks using raw time signal for LVCSR, in *Proc. Interspeech*, 2014.

[98] Sainath, T.; Weiss, R.; Senior, A.; Wilson, W.; Vinyals, O.: Learning the Speech Frontend with Raw Waveform CLDNNs, in *Proc. Interspeech*, 2015a.

[99] Sainath, T.; Vinyals, O.; Senior, A.; Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks, in *Proc. ICASSP*, 2015b.

[100] He, X.; Deng, L.; Chou, W.: Discriminative learning in sequential pattern recognition – A unifying review for optimization-oriented speech recognition. IEEE Sig. Proc. Mag., 25 (2008), 14–36.

[101] Yu, D.; Deng, L.; He, X.; Acero, X.: Large-margin minimum classification error training for large-scale speech recognition tasks, in *Proc. ICASSP*, 2007.

[102] Yu, D.; Deng, L.; Droppo, J.; Wu, J.; Gong, Y.; Acero, A.: Robust speech recognition using cepstral minimum-mean-square-error noise suppressor. IEEE Trans. Audio Speech Lang. Process., 16 (5) (2008a).

[103] Yu, D.; Deng, L.; He, X.; Acero, A.: Large-margin minimum classification error training: a theoretical risk minimization perspective. Computer Speech Lang., 22 (4) (2008b), 415–429.

[104] Kingsbury, B.; Sainath, T.; Soltau, H.: Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization, in *Proc. Interspeech*, 2012.

[105] Su, H.; Li, G.; Yu, D.; Seide, F.: Error back propagation for sequence training of context-dependent deep neural networks for conversational speech transcription, in *Proc. ICASSP*, 2013.

[106] Vesely, K.; Ghoshal, A.; Burget, L.; Povey, D.: Sequence-discriminative training of deep neural networks, in *Proc. Interspeech*, 2013.

[107] Chen, J.; Deng, L.: A primal-dual method for training recurrent neural networks constrained by the echo-state property, in *Proc. Int. Conf. Learning Representations*, April 2014.

[108] Bergstra, J.; Bengio, Y.: Random search for hyper-parameter optimization. J. Machine Learning Res., 3 (2012), 281–305.

[109] Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process., 22 (2014), 745–777.

[110] Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R.: Robust Automatic Speech Recognition: A Bridge to Practical Applications, Elsevier, London, 2015.

[111] Gales, M.: Model-based approaches to handling uncertainty. In Robust Speech Recognition of Uncertain or Missing Data: Theory and Application, Springer, Berlin, 2011, 101–125.

[112] Acero, A.; Deng, L.; Kristjansson, T.; Zhang, J.: HMM adaptation using vector taylor series for noisy speech recognition, in *Proc. Interspeech*, 2000.

[113] Huang, X.; Acero, A.; Chelba, C.; Deng, L.; Droppo, J.; Duchene, D.; Goodman, J.; Hon, H.: MiPad: a multimodal interaction prototype, in *Proc. ICASSP*, 2001.

[114] Deng, L.; Wu, J.; Droppo, J.; Acero, A.: Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. Speech Audio Process., 13 (3) (2005), 412–421.

[115] Seltzer, M.; Yu, D.; Wang, E.: An investigation of deep neural networks for noise robust speech recognition, in *Proc. ICASSP*, 2013.

[116] Kashiwagi, Y.; Saito, D.; Minematsu, N.; Hirose, K.: Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition, in *Proc. ASRU*, 2013.

[117] Deng, L.; Acero, A.; Jiang, L.; Droppo, J.; Huang, X.: High performance robust speech recognition using stereo training data, in *Proc. ICASSP*, 2001.

[118] Huang, Y.; Slaney, M.; Seltzer, M.; Gong, Y.: Towards better performance with heterogeneous training data in acoustic modeling using deep neural networks, in *Interspeech*, 2014.

[119] Abdelaziz, A.H.; Watanabe, S.; Hershey, J.R.; Vincent, E.; Kolossa, D.: Uncertainty propagation through deep neural networks, in *Proc, Interspeech*, 2015.

[120] Chen, Z.; Watanabe, S.; Erdogan, H.; Hershey, J.R.: Integration of speech enhancement and recognition using long-short term memory recurrent neural network, in *Proc. Interspeech*, 2015.

[121] Li, B.; Sim, K.: Noise adaptive front-end normalization based on vector Taylor series for deep neural networks in robust speech recognition, in *Proc. ICASSP*, 2013.

[122] Lin, H.; Deng, L.; Yu, D.; Gong, Y.; Acero, A.; Lee, C.-H.: A study on multilingual acoustic modeling for large vocabulary ASR, in *Proc. ICASSP*, 2009.

[123] Yu, D.; Deng, L.; Liu, P.; Wu, J.; Gong, Y.; Acero, A.: Cross-lingual speech recognition under runtime resource constraints, in *Proc. ICASSP*, 2009a.

[124] Huang, J.; Li, J.; Deng, L.; Yu, D.: Cross-language knowledge transfer using multilingual deep neural networks with shared hidden layers, in *Proc. ICASSP*, 2013a.

[125] Heigold, G. *et al.*: Multilingual acoustic models using distributed deep neural networks, in *Proc. ICASSP*, 2013.

[126] Giri, R.; Seltzer, M.; Droppo, J.; Yu, D.: Improving speech recognition in reverberation using a room-aware deep neural network and multitask learning, in *Proc. ICASSP*, 2015.

[127] Yu, D.; Chen, X.; Deng, L.: Factorized deep neural networks for adaptive speech recognition, in *International Workshop on Statistical Machine Learning for Speech Processing*, March, 2012.

[128] Yu, D.; Deng, L.; Seide, F.: The deep tensor neural network with applications to large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Proc., 21 (2) (2013), 388–396.

[129] Abdel-Hamid, O.; Mohamed, A.; Jiang, H.; Deng, L.; Penn, G.; Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Proc., 22 (10) (2014), 1533–1545.

[130] Sak, H.; Senior, A.; Rao, K.; Beaufays, F.: Fast and accurate recurrent neural network acoustic models for speech recognition, in *Interspeech*, 2015b.

[131] Variani, E.; McDermott, E.; Heigold, G.: A Gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture, in *Proc. ICASSP*, 2015.

[132] Tüske, Z.; Tahir, M.; Schlüter, R.; Ney, H.: Integrating Gaussian mixtures into deep neural networks: softmax layer with hidden variables, in *Proc. ICASSP*, 2015.

[133] Deng, L.; Yu, D.; Platt, J.: Scalable stacking and learning for building deep architectures, in *Proc. ICASSP*, 2012.

[134] Tur, G.; Deng, L.; Hakkani-Tür, D.; He, X.: Towards deep understanding: deep convex networks for semantic utterance classification, in *Proc. ICASSP*, 2012.

[135] Vinyals, O.; Jia, Y.; Deng, L.; Darrell, T.: Learning with recursive perceptual representations, in *Proc. NIPS*, 2012.

[136] Hutchinson, B.; Deng, L.; Yu, D.: A deep architecture with bilinear modeling of hidden representations: applications to phonetic recognition, in *Proc. ICASSP*, 2012.

[137] Hutchinson, B.; Deng, L.; Yu, D.: Tensor deep stacking networks. IEEE Trans. Pattern Analysis Machine Intelligence, 35 (2013), 1944–1957.

[138] 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 Workshop on Spoken Language Technologies*, December 2012a.

[139] Yu, D.; Deng, L.: Efficient and effective algorithms for training single-hidden-layer neural networks. Pattern Recogn. Lett., 33 (2012a), 554–558.

[140] Hershey, J.; Le Roux, J.; Weninger, F.: Deep unfolding: model-based inspiration of novel deep architectures, MERL TR2014–117 & arXiv 1409.2574, 2014.

[141] Yao, K.; Yu, D.; Seide, F.; Su, H.; Deng, L.; Gong, Y.: Adaptation of context-dependent deep neural networks for automatic speech recognition, in *Proc. ICASSP*, 2012.

[142] Krizhevsky, A.; Sutskever, I.; Hinton, G.: ImageNet classification with deep convolutional neural Networks, in *Proc. NIPS*, 2012.

[143] Ling, Z., *et al.*: Deep learning for acoustic modeling in parametric speech generation, *in IEEE Signal Proc. Magazine*, May, 2015, pp. 35–52.

[144] Mesnil, G.; He, X.; Deng, L.; Bengio, Y.: Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding, in *Proc. Interspeech* 2013.

[145] Mesnil, G. *et al.*: Using recurrent neural networks for slot filling in spoken language understanding, IEEE/ACM transactions on audio. Speech Lang. Process., 23 (2015), 530–539.

[146] Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C.: A neural probabilistic language model. J. Machine Learning Res., 3 (2003), 1137–1155.

[147] Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P.: Natural language processing (almost) from scratch. J. Machine Learning Res., 12 (2011), 2493–2537.

[148] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J.: Distributed representations of words and phrases and their compositionality, in *Proc. NIPS*, 2013.

[149] Socher, R.; Chen, D.; Manning, C.; Ng, A.: Reasoning with neural tensor networks for knowledge base completion, in *Proc. NIPS*, 2013.

[150] Vilnis, L.; McCallum, A.: Word representations via Gaussian embedding, in *ICLR*, 2015.

[151] Gao, J.; He, X.; Yih, W.; Deng, L.: Learning semantic representations for the phrase translation model, in *Proc. NIPS Workshop on Deep Learning*, December, 2013.

[152] Gao, J.; Patel, P.; Gamon, M.; He, X.; Deng, L.: Modeling interestingness with deep neural networks, in *Proc. EMNLP*, 2014a.

[153] Sutskever, I.; Vinyals, O.; Le, Q.: Sequence to sequence learning with neural networks, in *Proc. NIPS*, 2014.

[154] Bahdanau, D.; Cho, K.; Bengio, Y.: Neural machine translation by jointly learning to align and translate, in *ICLR*, 2015.

[155] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, W.; Yakhnenko, O.: Translating embeddings for modeling multi-relational data, in *Proc. NIPS*, 2013.

[156] Yang, B.; Yih, W.; He, X.; Gao, J.; Deng, L.: Embedding entities and relations for learning and inference in knowledge bases, in *ICLR*, 2015.

[157] Weston, J.; Chopra, S.; Bordes, A.: Memory networks. arXiv:1410.3916, 2014.

[158] Huang, P.; He, X.; Gao, J.; Deng, L.; Acero, A.; Heck, L.: Learning deep structured semantic models for Web search using clickthrough data, in *CIKM*, 2013b.

[159] Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G.: A latent semantic model with convolutional-pooling structure for information retrieval, in *CIKM*, 2014.

[160] Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G.: Learning semantic representations using convolutional neural networks for web search, WWW, 2014a.

[161] Liu, X.; Gao, J.; He, X.; Deng, L.; Duh, K.; Wang, Y.: Representation learning using multi-task deep neural networks for semantic classification and information retrieval, in *Proc. NAACL*, May 2015.

[162] Gao, J.; He, X.; Yih, W.; Deng, L.: Learning continuous phrase representations for translation modeling, in *Proc. ACL*, 2014.

[163] Fang, H. *et al.*: From captions to visual concepts and back, arXiv:1411.4952, 2014 and Proc. CVPR, 2015.

[164] Deng, L. *et al.*: Distributed speech processing in MiPad's multimodal user interface. IEEE Trans. Speech Audio Process., 10 (8) (2002), 605–619.

[165] Zhang, Z.; Liu, Z.; Sinclair, M.; Acero, A.; Deng, L.; Droppo, J.; Huang, X.; Zheng, Y.: Multi-sensory microphones for robust speech detection, enhancement and recognition, in *Proc. ICASSP*, 2004.

[166] Subramanya, A.; Deng, L.; Liu, Z.; Zhang, Z.: Multi-sensory speech processing: incorporating automatically extracted hidden dynamic information, in *Proc. ICME*, 2005.

[167] Frome, A. *et al.*: DeViSE: a deep visual-semantic embedding model, in *Proc. NIPS*, 2013.

[168] Mao, J.; Wu, W.; Yang, Y.; Wang, J.; Yuille, A.: Explain images with multimodal recurrent neural networks, arXiv:1410.1090v1, 2014.

[169] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D.: Show and tell: a neural image caption generator, arXiv:1411.4555, 2014.

[170] Karpathy, A.; Fei-Fei, Li.: Deep visual-semantic alignments for generating image descriptions, arXiv 2014 and Proc. CVPR, 2015.

[171] Kiros, R.; Salakhutdinov, R.; Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models, arXiv:1411.2539v1, 2014.

[172] Tai, K.; Socher, R.; Manning, C.: Improved semantic representations from tree-structured long short-term memory networks, arXiv:1503.00075v2, March, 2015.

Li Deng received his Ph.D. from the University of Wisconsin-Madison. He was an assistant and then tenured full professor at the University of Waterloo, Ontario, Canada during 1989–1999. Immediately afterwards he joined Microsoft Research, Redmond, USA as a Principal Researcher where he currently directs R&D at its Deep Learning Technology Center, which he founded in early 2014. His current activities are centered on business-critical applications involving big data analytics, natural language text, semantic modeling, speech, image, and multimodal signals. Outside the main responsibilities, his research interests lie in fundamental problems of machine learning, artificial and human intelligence, cognitive and neural computation with their biological connections, and multimodal signal/information processing. In addition to over 70 granted patents and over 300 scientific publications in leading journals and conferences, he authored or co-authored five books including two latest books: Deep Learning: Methods and Applications (NOW Publishers, 2014) and Automatic Speech Recognition: A Deep-Learning Approach (Springer, 2015). He is a Fellow of the IEEE, a Fellow of the Acoustical Society of America, and a Fellow of the ISCA. He served on the Board of Governors of the IEEE Signal Processing Society. More recently, he was Editors-In-Chief for IEEE Signal Processing Magazine and for IEEE/ACM Transactions on Audio, Speech, and Language Processing, and also served as a general chair of ICASSP and area chair of NIPS. His technical work in industry-scale deep learning and AI has impacted various areas of information processing, with the outcome being used in major Microsoft speech products and text- and big-data related products/services. His work helped initiate the resurgence of (deep) neural networks in the modern big-data, big-compute era, and is recognized by several awards, including 2013 IEEE SPS Best Paper Award and 2015 IEEE SPS Technical Achievement Award "*for outstanding contributions to deep learning and to automatic speech recognition.*"