Deep Learning: Methods and Applications

Li Deng
Microsoft Research
One Microsoft Way
Redmond, WA 98052; USA
deng@microsoft.com

Dong Yu
Microsoft Research
One Microsoft Way
Redmond, WA 98052; USA
Dong.Yu@microsoft.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

L. Deng and D. Yu. *Deep Learning: Methods and Applications.* Foundations and Trends® in Signal Processing, vol. 7, nos. 3–4, pp. 197–387, 2013.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-815-7

Foundations and Trends® in Signal Processing
Volume 7, Issues 3–4, 2013

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing

Information for Librarians

Foundations and Trends® in Signal Processing, 2013, Volume 7, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in Signal Processing
Vol. 7, Nos. 3–4 (2013) 197–387
© 2014 L. Deng and D. Yu
DOI: 10.1561/2000000039

Deep Learning: Methods and Applications

Li Deng
Microsoft Research
One Microsoft Way
Redmond, WA 98052; USA
deng@microsoft.com

Dong Yu
Microsoft Research
One Microsoft Way
Redmond, WA 98052; USA
Dong.Yu@microsoft.com

Contents

Endorsement

"In the past few years, deep learning has rapidly evolved into the de-facto approach for acoustic modeling in automatic speech recognition (ASR), showing tremendous improvement in accuracy, robustness, and cross-language generalizability over conventional approaches. This timely book is written by the pioneers of deep learning innovations and applications to ASR, who, as early as 2010, first succeeded in large vocabulary speech recognition using deep learning. This was accomplished using a special form of the deep neural net, developed by the authors, perfectly fit for fast decoding as required by industrial deployment of ASR technology. In addition to recounting this remarkable advance which ignited the industry-scale adoption of deep learning in ASR, this book also provides an overview of a sweeping range of up-to-date deep learning methodologies and its application to a variety of signal and information processing tasks, including not only ASR but also computer vision, language modeling, text processing, multimodal learning, and information retrieval. This is the first and the most valuable book for "deep and wide learning" of deep learning, not to be missed by anyone who wants to know the breath taking impact of deep learning in many facets of information processing, especially ASR, all of vital importance to our modern technological society."

— Sadaoki Furui, President of Toyota Technological Institute at Chicago, and Professor at the Tokyo Institute of Technology

Abstract

This monograph provides an overview of general deep learning methodology and its applications to a variety of signal and information processing tasks. The application areas are chosen with the following three criteria in mind: (1) expertise or knowledge of the authors; (2) the application areas that have already been transformed by the successful use of deep learning technology, such as speech recognition and computer vision; and (3) the application areas that have the potential to be impacted significantly by deep learning and that have been experiencing research growth, including natural language and text processing, information retrieval, and multimodal information processing empowered by multi-task deep learning.

L. Deng and D. Yu. *Deep Learning: Methods and Applications*. Foundations and Trends® in Signal Processing, vol. 7, nos. 3–4, pp. 197–387, 2013.

DOI: 10.1561/2000000039.

1

Introduction

1.1 Definitions and background

Since 2006, deep structured learning, or more commonly called deep learning or hierarchical learning, has emerged as a new area of machine learning research [20, 163]. During the past several years, the techniques developed from deep learning research have already been impacting a wide range of signal and information processing work within the traditional and the new, widened scopes including key aspects of machine learning and artificial intelligence; see overview articles in [7, 20, 24, 77, 94, 161, 412], and also the media coverage of this progress in [6, 237]. A series of workshops, tutorials, and special issues or conference special sessions in recent years have been devoted exclusively to deep learning and its applications to various signal and information processing areas. These include:

- 2008 NIPS Deep Learning Workshop;
- 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications;
- 2009 ICML Workshop on Learning Feature Hierarchies;

- 2011 ICML Workshop on Learning Architectures, Representations, and Optimization for Speech and Visual Information Processing;
- 2012 ICASSP Tutorial on Deep Learning for Signal and Information Processing;
- 2012 ICML Workshop on Representation Learning;
- 2012 Special Section on Deep Learning for Speech and Language Processing in IEEE Transactions on Audio, Speech, and Language Processing (T-ASLP, January);
- 2010, 2011, and 2012 NIPS Workshops on Deep Learning and Unsupervised Feature Learning;
- 2013 NIPS Workshops on Deep Learning and on Output Representation Learning;
- 2013 Special Issue on Learning Deep Architectures in IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI, September).
- 2013 International Conference on Learning Representations;
- 2013 ICML Workshop on Representation Learning Challenges;
- 2013 ICML Workshop on Deep Learning for Audio, Speech, and Language Processing;
- 2013 ICASSP Special Session on New Types of Deep Neural Network Learning for Speech Recognition and Related Applications.

The authors have been actively involved in deep learning research and in organizing or providing several of the above events, tutorials, and editorials. In particular, they gave tutorials and invited lectures on this topic at various places. Part of this monograph is based on their tutorials and lecture material.

Before embarking on describing details of deep learning, let's provide necessary definitions. Deep learning has various closely related definitions or high-level descriptions:

- ***Definition 1***: A class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification.

- ***Definition 2***: "A sub-field within machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture. Most of these models are based on unsupervised learning of representations." (Wikipedia on "Deep Learning" around March 2012.)
- ***Definition 3***: "A sub-field of machine learning that is based on learning several levels of representations, corresponding to a hierarchy of features or factors or concepts, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts. Deep learning is part of a broader family of machine learning methods based on learning representations. An observation (e.g., an image) can be represented in many ways (e.g., a vector of pixels), but some representations make it easier to learn tasks of interest (e.g., is this the image of a human face?) from examples, and research in this area attempts to define what makes better representations and how to learn them." (Wikipedia on "Deep Learning" around February 2013.)
- ***Definition 4***: "Deep learning is a set of algorithms in machine learning that attempt to learn in multiple levels, corresponding to different levels of abstraction. It typically uses artificial neural networks. The levels in these learned statistical models correspond to distinct levels of concepts, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts." See Wikipedia http://en.wikipedia.org/wiki/Deep_learning on "Deep Learning" as of this most recent update in October 2013.
- ***Definition 5***: "Deep Learning is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals: Artificial

> Intelligence. Deep Learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text." See https://github.com/lisa-lab/DeepLearningTutorials

Note that the deep learning that we discuss in this monograph is about learning with deep architectures for signal and information processing. It is not about deep understanding of the signal or information, although in many cases they may be related. It should also be distinguished from the overloaded term in educational psychology: "Deep learning describes an approach to learning that is characterized by active engagement, intrinsic motivation, and a personal search for meaning." http://www.blackwellreference.com/public/tocnode?id=g9781405161251_chunk_g97814051612516_ss1-1

Common among the various high-level descriptions of deep learning above are two key aspects: (1) models consisting of multiple layers or stages of nonlinear information processing; and (2) methods for supervised or unsupervised learning of feature representation at successively higher, more abstract layers. Deep learning is in the intersections among the research areas of neural networks, artificial intelligence, graphical modeling, optimization, pattern recognition, and signal processing. Three important reasons for the popularity of deep learning today are the drastically increased chip processing abilities (e.g., general-purpose graphical processing units or GPGPUs), the significantly increased size of data used for training, and the recent advances in machine learning and signal/information processing research. These advances have enabled the deep learning methods to effectively exploit complex, compositional nonlinear functions, to learn distributed and hierarchical feature representations, and to make effective use of both labeled and unlabeled data.

Active researchers in this area include those at University of Toronto, New York University, University of Montreal, Stanford University, Microsoft Research (since 2009), Google (since about 2011), IBM Research (since about 2011), Baidu (since 2012), Facebook (since 2013), UC-Berkeley, UC-Irvine, IDIAP, IDSIA, University College London, University of Michigan, Massachusetts Institute of

Technology, University of Washington, and numerous other places; see http://deeplearning.net/deep-learning-research-groups-and-labs/ for a more detailed list. These researchers have demonstrated empirical successes of deep learning in diverse applications of computer vision, phonetic recognition, voice search, conversational speech recognition, speech and image feature coding, semantic utterance classification, natural language understanding, hand-writing recognition, audio processing, information retrieval, robotics, and even in the analysis of molecules that may lead to discovery of new drugs as reported recently by [237].

In addition to the reference list provided at the end of this monograph, which may be outdated not long after the publication of this monograph, there are a number of excellent and frequently updated reading lists, tutorials, software, and video lectures online at:

- http://deeplearning.net/reading-list/
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Recommended_Readings
- http://www.cs.toronto.edu/∼hinton/
- http://deeplearning.net/tutorial/
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

1.2 Organization of this monograph

The rest of the monograph is organized as follows:

In Section 2, we provide a brief historical account of deep learning, mainly from the perspective of how speech recognition technology has been hugely impacted by deep learning, and how the revolution got started and has gained and sustained immense momentum.

In Section 3, a three-way categorization scheme for a majority of the work in deep learning is developed. They include unsupervised, supervised, and hybrid deep learning networks, where in the latter category unsupervised learning (or pre-training) is exploited to assist the subsequent stage of supervised learning when the final tasks pertain to classification. The supervised and hybrid deep networks often have the same type of architectures or the structures in the deep networks, but the unsupervised deep networks tend to have different architectures from the others.

Sections 4–6 are devoted, respectively, to three popular types of deep architectures, one from each of the classes in the three-way categorization scheme reviewed in Section 3. In Section 4, we discuss in detail deep autoencoders as a prominent example of the unsupervised deep learning networks. No class labels are used in the learning, although supervised learning methods such as back-propagation are cleverly exploited when the input signal itself, instead of any label information of interest to possible classification tasks, is treated as the "supervision" signal.

In Section 5, as a major example in the hybrid deep network category, we present in detail the deep neural networks with unsupervised and largely generative pre-training to boost the effectiveness of supervised training. This benefit is found critical when the training data are limited and no other appropriate regularization approaches (i.e., dropout) are exploited. The particular pre-training method based on restricted Boltzmann machines and the related deep belief networks described in this section has been historically significant as it ignited the intense interest in the early applications of deep learning to speech recognition and other information processing tasks. In addition to this retrospective review, subsequent development and different paths from the more recent perspective are discussed.

In Section 6, the basic deep stacking networks and their several extensions are discussed in detail, which exemplify the discriminative, supervised deep learning networks in the three-way classification scheme. This group of deep networks operate in many ways that are distinct from the deep neural networks. Most notably, they use target labels in constructing *each* of many layers or modules in the overall deep networks. Assumptions made about part of the networks, such as linear output units in each of the modules, simplify the learning algorithms and enable a much wider variety of network architectures to be constructed and learned than the networks discussed in Sections 4 and 5.

In Sections 7–11, we select a set of typical and successful applications of deep learning in diverse areas of signal and information processing. In Section 7, we review the applications of deep learning to speech recognition, speech synthesis, and audio processing. Subsections surrounding the main subject of speech recognition are created based on several prominent themes on the topic in the literature.

In Section 8, we present recent results of applying deep learning to language modeling and natural language processing, where we highlight the key recent development in embedding symbolic entities such as words into low-dimensional, continuous-valued vectors.

Section 9 is devoted to selected applications of deep learning to information retrieval including web search.

In Section 10, we cover selected applications of deep learning to image object recognition in computer vision. The section is divided to two main classes of deep learning approaches: (1) unsupervised feature learning, and (2) supervised learning for end-to-end and joint feature learning and classification.

Selected applications to multi-modal processing and multi-task learning are reviewed in Section 11, divided into three categories according to the nature of the multi-modal data as inputs to the deep learning systems. For single-modality data of speech, text, or image, a number of recent multi-task learning studies based on deep learning methods are reviewed in the literature.

Finally, conclusions are given in Section 12 to summarize the monograph and to discuss future challenges and directions.

This short monograph contains the material expanded from two tutorials that the authors gave, one at APSIPA in October 2011 and the other at ICASSP in March 2012. Substantial updates have been made based on the literature up to January 2014 (including the materials presented at NIPS-2013 and at IEEE-ASRU-2013 both held in December of 2013), focusing on practical aspects in the fast development of deep learning research and technology during the interim years.

References

[1] O. Abdel-Hamid, L. Deng, and D. Yu. Exploring convolutional neural network structures and optimization for speech recognition. *Proceedings of Interspeech*, 2013.

[2] O. Abdel-Hamid, L. Deng, D. Yu, and H. Jiang. Deep segmental neural networks for speech recognition. In *Proceedings of Interspeech*. 2013.

[3] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[4] A. Acero, L. Deng, T. Kristjansson, and J. Zhang. HMM adaptation using vector taylor series for noisy speech recognition. In *Proceedings of Interspeech*. 2000.

[5] G. Alain and Y. Bengio. What regularized autoencoders learn from the data generating distribution. In *Proceedings of International Conference on Learning Representations (ICLR)*. 2013.

[6] G. Anthes. Deep learning comes of age. *Communications of the Association for Computing Machinery (ACM)*, 56(6):13–15, June 2013.

[7] I. Arel, C. Rose, and T. Karnowski. Deep machine learning — a new frontier in artificial intelligence. *IEEE Computational Intelligence Magazine*, 5:13–18, November 2010.

[8] E. Arisoy, T. Sainath, B. Kingsbury, and B. Ramabhadran. Deep neural network language models. In *Proceedings of the Joint Human Language Technology Conference and the North American Chapter of the Association of Computational Linguistics (HLT-NAACL) Workshop*. 2012.

[9] O. Aslan, H. Cheng, D. Schuurmans, and X. Zhang. Convex two-layer modeling. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[10] J. Ba and B. Frey. Adaptive dropout for training deep neural networks. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[11] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy. Research developments and directions in speech recognition and understanding. *IEEE Signal Processing Magazine*, 26(3):75–80, May 2009.

[12] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy. Updated MINS report on speech recognition and understanding. *IEEE Signal Processing Magazine*, 26(4), July 2009.

[13] P. Baldi and P. Sadowski. Understanding dropout. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[14] E. Battenberg, E. Schmidt, and J. Bello. *Deep learning for music, special session at International Conference on Acoustics Speech and Signal Processing (ICASSP)* (http://www.icassp2014.org/special_sections.html#ss8), 2014.

[15] E. Batternberg and D. Wessel. Analyzing drum patterns using conditional deep belief networks. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*. 2012.

[16] P. Bell, P. Swietojanski, and S. Renals. Multi-level adaptive networks in tandem and hybrid ASR systems. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[17] Y. Bengio. Artificial neural networks and their application to sequence recognition. Ph.D. Thesis, McGill University, Montreal, Canada, 1991.

[18] Y. Bengio. New distributed probabilistic language models. Technical Report, University of Montreal, 2002.

[19] Y. Bengio. Neural net language models. *Scholarpedia*, 3, 2008.

[20] Y. Bengio. Learning deep architectures for AI. in *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[21] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 27:17–37, 2012.

[22] Y. Bengio. Deep learning of representations: Looking forward. In *Statistical Language and Speech Processing*, pages 1–37. Springer, 2013.

[23] Y. Bengio, N. Boulanger, and R. Pascanu. Advances in optimizing recurrent networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[24] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38:1798–1828, 2013.

[25] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Global optimization of a neural network-hidden markov model hybrid. *IEEE Transactions on Neural Networks*, 3:252–259, 1992.

[26] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2000.

[27] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[28] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2006.

[29] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5:157–166, 1994.

[30] Y. Bengio, E. Thibodeau-Laufer, and J. Yosinski. Deep generative stochastic networks trainable by backprop. arXiv 1306:1091, 2013. also accepted to appear in *Proceedings of International Conference on Machine Learning (ICML), 2014*.

[31] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising autoencoders as generative models. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[32] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal on Machine Learning Research*, 3:281–305, 2012.

[33] A. Biem, S. Katagiri, E. McDermott, and B. Juang. An application of discriminative feature extraction to filter-bank-based speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9:96–110, 2001.

[34] J. Bilmes. Dynamic graphical models. *IEEE Signal Processing Magazine*, 33:29–42, 2010.

[35] J. Bilmes and C. Bartels. Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine*, 22:89–100, 2005.

[36] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. A semantic matching energy function for learning with multi-relational data — application to word-sense disambiguation. *Machine Learning*, May 2013.

[37] A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*. 2011.

[38] L. Bottou. From machine learning to machine reasoning: An essay. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[39] L. Bottou and Y. LeCun. Large scale online learning. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2004.

[40] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling Temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of International Conference on Machine Learning (ICML)*. 2012.

[41] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Audio chord recognition with recurrent neural networks. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*. 2013.

[42] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer, Norwell, MA, 1993.

[43] J. Bouvrie. Hierarchical learning: Theory with applications in speech and vision. Ph.D. thesis, MIT, 2009.

[44] L. Breiman. Stacked regression. *Machine Learning*, 24:49–64, 1996.

[45] J. Bridle, L. Deng, J. Picone, H. Richards, J. Ma, T. Kamm, M. Schuster, S. Pike, and R. Reagan. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. Final Report for 1998 Workshop on Language Engineering, CLSP, Johns Hopkins, 1998.

[46] P. Cardinal, P. Dumouchel, and G. Boulianne. Large vocabulary speech recognition on parallel architectures. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2290–2300, November 2013.

[47] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[48] J. Chen and L. Deng. A primal-dual method for training recurrent neural networks constrained by the echo-state property. In *Proceedings of International Conference on Learning Representations*. April 2014.

[49] X. Chen, A. Eversole, G. Li, D. Yu, and F. Seide. Pipelined back-propagation for context-dependent deep neural networks. In *Proceedings of Interspeech*. 2012.

[50] R. Chengalvarayan and L. Deng. Hmm-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features. *IEEE Transactions on Speech and Audio Processing*, pages 243–256, 1997.

[51] R. Chengalvarayan and L. Deng. Use of generalized dynamic feature parameters for speech recognition. *IEEE Transactions on Speech and Audio Processing*, pages 232–242, 1997a.

[52] R. Chengalvarayan and L. Deng. Speech trajectory discrimination using the minimum classification error learning. *IEEE Transactions on Speech and Audio Processing*, 6(6):505–515, 1998.

[53] Y. Cho and L. Saul. Kernel methods for deep learning. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 342–350. 2009.

[54] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[55] D. Ciresan, U. Meier, L. Gambardella, and J. Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, December 2010.

[56] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. A committee of neural networks for traffic sign classification. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*. 2011.

[57] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 2012.

[58] D. C. Ciresan, U. Meier, and J. Schmidhuber. Transfer learning for Latin and Chinese characters with deep neural networks. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*. 2012.

[59] A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, and B. Catanzaro. Deep learning with COTS HPC. In *Proceedings of International Conference on Machine Learning (ICML)*. 2013.

[60] W. Cohen and R. V. de Carvalho. Stacked sequential learning. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 671–676. 2005.

[61] R. Collobert. Deep learning for efficient discriminative parsing. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2011.

[62] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of International Conference on Machine Learning (ICML)*. 2008.

[63] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal on Machine Learning Research*, 12:2493–2537, 2011.

[64] G. Dahl, M. Ranzato, A. Mohamed, and G. Hinton. Phone recognition with the mean-covariance restricted boltzmann machine. In *Proceedings of Neural Information Processing Systems (NIPS)*, volume 23, pages 469–477. 2010.

[65] G. Dahl, T. Sainath, and G. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[66] G. Dahl, J. Stokes, L. Deng, and D. Yu. Large-scale malware classification using random projections and neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[67] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent DBN-HMMs in large vocabulary continuous speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2011.

[68] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent, pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, & Language Processing*, 20(1):30–42, January 2012.

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

[70] K. Demuynck and F. Triefenbach. Porting concepts from DNNs back to GMMs. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

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

[72] L. Deng. A stochastic model of speech incorporating hierarchical nonstationarity. *IEEE Transactions on Speech and Audio Processing*, 1(4):471–475, 1993.

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

[74] L. Deng. Computational models for speech production. In *Computational Models of Speech Pattern Processing*, pages 199–213. Springer Verlag, 1999.

[75] L. Deng. Switching dynamic system models for speech articulation and acoustics. In *Mathematical Foundations of Speech and Language Processing*, pages 115–134. Springer-Verlag, New York, 2003.

[76] L. Deng. *Dynamic Speech Models — Theory, Algorithm, and Application.* Morgan & Claypool, December 2006.

[77] L. Deng. An overview of deep-structured learning for information processing. In *Proceedings of Asian-Pacific Signal & Information Processing Annual Summit and Conference (APSIPA-ASC)*. October 2011.

[78] L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), November 2012.

[79] L. Deng. Design and learning of output representations for speech recognition. In *Neural Information Processing Systems (NIPS) Workshop on Learning Output Representations*. December 2013.

[80] L. Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. In *Asian-Pacific Signal & Information Processing Association Transactions on Signal and Information Processing*. 2013.

[81] L. Deng, O. Abdel-Hamid, and D. Yu. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[82] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang. High performance robust speech recognition using stereo training data. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2001.

[83] L. Deng and M. Aksmanovic. Speaker-independent phonetic classification using hidden markov models with state-conditioned mixtures of trend functions. *IEEE Transactions on Speech and Audio Processing*, 5:319–324, 1997.

[84] L. Deng, M. Aksmanovic, D. Sun, and J. Wu. Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states. *IEEE Transactions on Speech and Audio Processing*, 2(4):507–520, 1994.

[85] L. Deng and J. Chen. Sequence classification using the high-level features extracted from deep neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2014.

[86] L. Deng and K. Erler. Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units. *Journal of the Acoustical Society of America*, 92(6):3058–3067, 1992.

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

[88] L. Deng, X. He, and J. Gao. Deep stacking networks for information retrieval. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013c.

[89] L. Deng, G. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013b.

[90] L. Deng and X. D. Huang. Challenges in adopting speech recognition. *Communications of the Association for Computing Machinery (ACM)*, 47(1):11–13, January 2004.

[91] L. Deng, B. Hutchinson, and D. Yu. Parallel training of deep stacking networks. In *Proceedings of Interspeech*. 2012b.

[92] L. Deng, M. Lennig, V. Gupta, F. Seitz, P. Mermelstein, and P. Kenny. Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition. *IEEE Transactions on Signal Processing*, 39(7):1677–1681, 1991.

[93] L. Deng, M. Lennig, F. Seitz, and P. Mermelstein. Large vocabulary word recognition using context-dependent allophonic hidden Markov models. *Computer Speech and Language*, 4(4):345–357, 1990.

[94] L. Deng, J. Li, K. Huang, Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. Recent advances in deep learning for speech research at Microsoft. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013a.

[95] L. Deng and X. Li. Machine learning paradigms in speech recognition: An overview. *IEEE Transactions on Audio, Speech, & Language*, 21:1060–1089, May 2013.

[96] L. Deng and J. Ma. Spontaneous speech recognition using a statistical coarticulatory model for the vocal tract resonance dynamics. *Journal of the Acoustical Society America*, 108:3036–3048, 2000.

[97] L. Deng and D. O'Shaughnessy. *Speech Processing — A Dynamic and Optimization-Oriented Approach*. Marcel Dekker, 2003.

[98] L. Deng, G. Ramsay, and D. Sun. Production models as a structural basis for automatic speech recognition. *Speech Communication*, 33(2–3):93–111, August 1997.

[99] L. Deng and H. Sameti. Transitional speech units and their representation by regressive Markov states: Applications to speech recognition. *IEEE Transactions on speech and audio processing*, 4(4):301–306, July 1996.

[100] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton. Binary coding of speech spectrograms using a deep autoencoder. In *Proceedings of Interspeech*. 2010.

[101] L. Deng and D. Sun. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, 85(5):2702–2719, 1994.

[102] L. Deng, G. Tur, X. He, and D. Hakkani-Tur. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *Proceedings of IEEE Workshop on Spoken Language Technologies*. December 2012.

[103] L. Deng, K. Wang, A. Acero, H. W. Hon, J. Droppo, C. Boulis, Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and X. Huang. Distributed speech processing in mipad's multimodal user interface. *IEEE Transactions on Speech and Audio Processing*, 10(8):605–619, 2002.

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

[105] L. Deng and D. Yu. Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2007.

[106] L. Deng and D. Yu. Deep convex network: A scalable architecture for speech pattern classification. In *Proceedings of Interspeech*. 2011.

[107] L. Deng, D. Yu, and A. Acero. A bidirectional target filtering model of speech coarticulation: Two-stage implementation for phonetic recognition. *IEEE Transactions on Audio and Speech Processing*, 14(1):256–265, January 2006.

[108] L. Deng, D. Yu, and A. Acero. Structured speech modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1492–1504, September 2006.

[109] L. Deng, D. Yu, and G. Hinton. Deep learning for speech recognition and related applications. *Neural Information Processing Systems (NIPS) Workshop*, 2009.

[110] L. Deng, D. Yu, and J. Platt. Scalable stacking and learning for building deep architectures. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012a.

[111] T. Deselaers, S. Hasan, O. Bender, and H. Ney. A deep learning approach to machine transliteration. In *Proceedings of 4th Workshop on Statistical Machine Translation*, pages 233–241. Athens, Greece, March 2009.

[112] A. Diez. Automatic language recognition using deep neural networks. Thesis, Universidad Autonoma de Madrid, SPAIN, September 2013.

[113] P. Dognin and V. Goel. Combining stochastic average gradient and hessian-free optimization for sequence training of deep neural networks. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[114] D. Erhan, Y. Bengio, A. Courvelle, P. Manzagol, P. Vencent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal on Machine Learning Research*, pages 201–208, 2010.

[115] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory. F0 contour prediction with a deep belief network-gaussian process hybrid model. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6885–6889. 2013.

[116] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32:41–62, 1998.

[117] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[118] Q. Fu, X. He, and L. Deng. Phone-discriminating minimum classification error (p-mce) training for phonetic recognition. In *Proceedings of Interspeech*. 2007.

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

[120] J. Gao, X. He, and J.-Y. Nie. Clickthrough-based translation models for web search: From word models to phrase models. In *Proceedings of Conference on Information and Knowledge Management (CIKM)*. 2010.

[121] J. Gao, X. He, W. Yih, and L. Deng. Learning semantic representations for the phrase translation model. In *Proceedings of Neural Information Processing Systems (NIPS) Workshop on Deep Learning*. December 2013.

[122] J. Gao, X. He, W. Yih, and L. Deng. Learning semantic representations for the phrase translation model. MSR-TR-2013-88, September 2013.

[123] J. Gao, X. He, W. Yih, and L. Deng. Learning continuous phrase representations for translation modeling. In *Proceedings of Association for Computational Linguistics (ACL)*. 2014.

[124] J. Gao, K. Toutanova, and W.-T. Yih. Clickthrough-based latent semantic models for web search. In *Proceedings of Special Interest Group on Information Retrieval (SIGIR)*. 2011.

[125] R. Gens and P. Domingo. Discriminative learning of sum-product networks. *Neural Information Processing Systems (NIPS)*, 2012.

[126] D. George. How the brain might work: A hierarchical and temporal model for learning and recognition. Ph.D. thesis, Stanford University, 2008.

[127] M. Gibson and T. Hain. Error approximation and minimum phone error acoustic model estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1269–1279, August 2010.

[128] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524v1, 2013.

[129] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2010.

[130] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. April 2011.

[131] I. Goodfellow, M. Mirza, A. Courville, and Y. Bengio. Multi-prediction deep boltzmann machines. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[132] E. Grais, M. Sen, and H. Erdogan. Deep neural networks for single channel source separation. arXiv:1311.2746v1, 2013.

[133] A. Graves. Sequence transduction with recurrent neural networks. *Representation Learning Workshop, International Conference on Machine Learning (ICML)*, 2012.

[134] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*. 2006.

[135] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[136] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[137] F. Grezl and P. Fousek. Optimizing bottle-neck features for LVCSR. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2008.

[138] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio. Learned-norm pooling for deep feedforward and recurrent neural networks. http://arxiv.org/abs/1311.1780, 2014.

[139] M. Gutmann and A. Hyvarinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

[140] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan. Transcribing meetings with the AMIDA systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:486–498, 2012.

[141] P. Hamel and D. Eck. Learning features from music audio with deep belief networks. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*. 2010.

[142] G. Hawkins, S. Ahmad, and D. Dubinsky. Hierarchical temporal memory including HTM cortical learning algorithms. Numenta Technical Report, December 10 2010.

[143] J. Hawkins and S. Blakeslee. *On Intelligence: How a New Understanding of the Brain will lead to the Creation of Truly Intelligent Machines*. Times Books, New York, 2004.

[144] X. He and L. Deng. Speech recognition, machine translation, and speech translation — a unifying discriminative framework. *IEEE Signal Processing Magazine*, 28, November 2011.

[145] X. He and L. Deng. Optimization in speech-centric information processing: Criteria and techniques. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[146] X. He and L. Deng. Speech-centric information processing: An optimization-oriented approach. In *Proceedings of the IEEE*. 2013.

[147] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition — a unifying review for optimization-oriented speech recognition. *IEEE Signal Processing Magazine*, 25:14–36, 2008.

[148] G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schluter. Equivalence of generative and log-liner models. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1138–1148, February 2011.

[149] G. Heigold, H. Ney, and R. Schluter. Investigations on an EM-style optimization algorithm for discriminative training of HMMs. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2616–2626, December 2013.

[150] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. Multilingual acoustic models using distributed deep neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[151] I. Heintz, E. Fosler-Lussier, and C. Brew. Discriminative input stream combination for conditional random field phone recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1533–1546, November 2009.

[152] M. Henderson, B. Thomson, and S. Young. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of Special Interest Group on Disclosure and Dialogue (SIGDIAL)*. 2013.

[153] M. Hermans and B. Schrauwen. Training and analysing deep recurrent neural networks. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[154] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2000.

[155] Y. Hifny and S. Renals. Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2):354–365, February 2009.

[156] G. Hinton. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46:47–75, 1990.

[157] G. Hinton. Preface to the special issue on connectionist symbol processing. *Artificial Intelligence*, 46:1–4, 1990.

[158] G. Hinton. The ups and downs of Hebb synapses. *Canadian Psychology*, 44:10–13, 2003.

[159] G. Hinton. A practical guide to training restricted boltzmann machines. UTML Tech Report 2010-003, Univ. Toronto, August 2010.

[160] G. Hinton. A better way to learn features. *Communications of the Association for Computing Machinery (ACM)*, 54(10), October 2011.

[161] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[162] G. Hinton, A. Krizhevsky, and S. Wang. Transforming autoencoders. In *Proceedings of International Conference on Artificial Neural Networks*. 2011.

[163] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[164] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[165] G. Hinton and R. Salakhutdinov. Discovering binary codes for documents by learning deep generative models. *Topics in Cognitive Science*, pages 1–18, 2010.

[166] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv: 1207.0580v1, 2012.

[167] S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. Diploma thesis, Institut fur Informatik, Technische Universitat Munchen, 1991.

[168] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[169] E. Huang, R. Socher, C. Manning, and A. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of Association for Computational Linguistics (ACL)*. 2012.

[170] J. Huang, J. Li, L. Deng, and D. Yu. Cross-language knowledge transfer using multilingual deep neural networks with shared hidden layers. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[171] P. Huang, L. Deng, M. Hasegawa-Johnson, and X. He. Random features for kernel deep convex network. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[172] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. *Association for Computing Machinery (ACM) International Conference Information and Knowledge Management (CIKM)*, 2013.

[173] P. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng. Predicting speech recognition confidence using deep learning with word identity and score features. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[174] S. Huang and S. Renals. Hierarchical bayesian language models for conversational speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):1941–1954, November 2010.

[175] X. Huang, A. Acero, C. Chelba, L. Deng, J. Droppo, D. Duchene, J. Goodman, and H. Hon. Mipad: A multimodal interaction prototype. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2001.

[176] Y. Huang, D. Yu, Y. Gong, and C. Liu. Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration. In *Proceedings of Interspeech*, pages 2360–2364. 2013.

[177] E. Humphrey and J. Bello. Rethinking automatic chord recognition with convolutional neural networks. In *Proceedings of International Conference on Machine Learning and Application (ICMLA)*. 2012a.

[178] E. Humphrey, J. Bello, and Y. LeCun. Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*. 2012.

[179] E. Humphrey, J. Bello, and Y. LeCun. Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems*, 2013.

[180] B. Hutchinson, L. Deng, and D. Yu. A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[181] B. Hutchinson, L. Deng, and D. Yu. Tensor deep stacking networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1944–1957, 2013.

[182] D. Imseng, P. Motlicek, P. Garner, and H. Bourlard. Impact of deep MLP architecture on different modeling techniques for under-resourced speech recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[183] N. Jaitly and G. Hinton. Learning a better representation of speech sound waves using restricted boltzmann machines. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2011.

[184] N. Jaitly, P. Nguyen, and V. Vanhoucke. Application of pre-trained deep neural networks to large vocabulary speech recognition. In *Proceedings of Interspeech*. 2012.

[185] K. Jarrett, K. Kavukcuoglu, and Y. LeCun. What is the best multistage architecture for object recognition? In *Proceedings of International Conference on Computer Vision*, pages 2146–2153. 2009.

[186] H. Jiang and X. Li. Parameter estimation of statistical models using convex optimization: An advanced method of discriminative training for speech and language processing. *IEEE Signal Processing Magazine*, 27(3):115–127, 2010.

[187] B. Juang, S. Levinson, and M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, 32:307–309, 1986.

[188] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Transactions On Speech and Audio Processing*, 5:257–265, 1997.

[189] S. Kahou et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of International Conference on Multimodal Interaction (ICMI)*. 2013.

[190] S. Kang, X. Qian, and H. Meng. Multi-distribution deep belief network for speech synthesis. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 8012–8016. 2013.

[191] Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose. Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[192] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierarchies for visual recognition. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2010.

[193] H. Ketabdar and H. Bourlard. Enhanced phone posteriors for improving speech recognition systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1094–1106, August 2010.

[194] B. Kingsbury. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2009.

[195] B. Kingsbury, T. Sainath, and H. Soltau. Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In *Proceedings of Interspeech*. 2012.

[196] R. Kiros, R. Zemel, and R. Salakhutdinov. Multimodal neural language models. In *Proceedings of Neural Information Processing Systems (NIPS) Deep Learning Workshop*. 2013.

[197] T. Ko and B. Mak. Eigentriphones for context-dependent acoustic modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1285–1294, 2013.

[198] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[199] Y. Kubo, T. Hori, and A. Nakamura. Integrating deep neural networks into structural classification approach based on weighted finite-state transducers. In *Proceedings of Interspeech*. 2012.

[200] R. Kurzweil. *How to Create a Mind*. Viking Books, December 2012.

[201] P. Lal and S. King. Cross-lingual automatic speech recognition using tandem features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2506–2515, December 2013.

[202] K. Lang, A. Waibel, and G. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3(1):23–43, 1990.

[203] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of International Conference on Machine Learning (ICML)*. 2008.

[204] D. Le and P. Mower. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[205] H. Le, A. Allauzen, G. Wisniewski, and F. Yvon. Training continuous space language models: Some practical issues. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 778–788. 2010.

[206] H. Le, I. Oparin, A. Allauzen, J. Gauvain, and F. Yvon. Structured output layer neural network language model. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2011.

[207] H. Le, I. Oparin, A. Allauzen, J.-L. Gauvain, and F. Yvon. Structured output layer neural network language models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):197–206, January 2013.

[208] Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Ng. On optimization methods for deep learning. In *Proceedings of International Conference on Machine Learning (ICML)*. 2011.

[209] Q. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *Proceedings of International Conference on Machine Learning (ICML)*. 2012.

[210] Y. LeCun. Learning invariant feature hierarchies. In *Proceedings of European Conference on Computer Vision (ECCV)*. 2012.

[211] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, Cambridge, Massachusetts, 1995.

[212] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[213] Y. LeCun, S. Chopra, M. Ranzato, and F. Huang. Energy-based models in document recognition and computer vision. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*. 2007.

[214] C.-H. Lee. From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next-generation automatic speech recognition. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 109–111. 2004.

[215] H. Lee, R. Grosse, R. Ranganath, and A. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of International Conference on Machine Learning (ICML)*. 2009.

[216] H. Lee, R. Grosse, R. Ranganath, and A. Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the Association for Computing Machinery (ACM)*, 54(10):95–103, October 2011.

[217] H. Lee, Y. Largman, P. Pham, and A. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2010.

[218] P. Lena, K. Nagata, and P. Baldi. Deep spatiotemporal architectures and learning for protein structure prediction. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[219] S. Levine. Exploring deep and recurrent architectures for optimal control. arXiv:1311.1761v1.

[220] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/Association for Computing Machinery (ACM) Transactions on Audio, Speech, and Language Processing*, pages 1–33, 2014.

[221] J. Li, D. Yu, J. Huang, and Y. Gong. Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In *Proceedings of IEEE Spoken Language Technology (SLT)*. 2012.

[222] L. Li, Y. Zhao, D. Jiang, and Y. Zhang etc. Hybrid deep neural network–hidden markov model (DNN-HMM) based speech emotion recognition. In *Proceedings Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 312–317. September 2013.

[223] H. Liao. Speaker adaptation of context dependent deep neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[224] H. Liao, E. McDermott, and A. Senior. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[225] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee. A study on multilingual acoustic modeling for large vocabulary ASR. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2009.

[226] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and SVM training. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 2011.

[227] Z. Ling, L. Deng, and D. Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio Speech Language Processing*, 21(10):2129–2139, 2013.

[228] Z. Ling, L. Deng, and D. Yu. Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 7825–7829. 2013.

[229] Z. Ling, K. Richmond, and J. Yamagishi. Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21, January 2013.

[230] L. Lu, K. Chin, A. Ghoshal, and S. Renals. Joint uncertainty decoding for noise robust subspace gaussian mixture models. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1791–1804, 2013.

[231] J. Ma and L. Deng. A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamical model of speech. *Computer, Speech and Language*, 2000.

[232] J. Ma and L. Deng. Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model. *IEEE Transactions on Speech and Audio Processing*, 11(6):590–602, 2003.

[233] J. Ma and L. Deng. Target-directed mixture dynamic models for spontaneous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 12(1):47–58, 2004.

[234] A. Maas, A. Hannun, and A. Ng. Rectifier nonlinearities improve neural network acoustic models. *International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[235] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and P. Ng. Recurrent neural networks for noise reduction in robust ASR. In *Proceedings of Interspeech*. 2012.

[236] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[237] J. Markoff. Scientists see promise in deep-learning programs. *New York Times*, November 24 2012.

[238] J. Martens. Deep learning with hessian-free optimization. In *Proceedings of International Conference on Machine Learning (ICML)*. 2010.

[239] J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. In *Proceedings of International Conference on Machine Learning (ICML)*. 2011.

[240] D. McAllester. A PAC-bayesian tutorial with a dropout bound. ArXive1307.2118, July 2013.

[241] I. McGraw, I. Badr, and J. R. Glass. Learning lexicons from speech using a pronunciation mixture model. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):357,366, February 2013.

[242] G. Mesnil, X. He, L. Deng, and Y. Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Proceedings of Interspeech*. 2013.

[243] Y. Miao and F. Metze. Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training. In *Proceedings of Interspeech*. 2013.

[244] Y. Miao, S. Rawat, and F. Metze. Deep maxout networks for low resource speech recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[245] T. Mikolov. Statistical language models based on neural networks. Ph.D. thesis, Brno University of Technology, 2012.

[246] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*. 2013.

[247] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky. Strategies for training large scale neural network language models. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2011.

[248] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1045–1048. 2010.

[249] T. Mikolov, Q. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. arXiv:1309.4168v1, 2013.

[250] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[251] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri. A recognition method with parametric trajectory synthesized using direct relations between static and dynamic feature vector time series. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 957–960. 2002.

[252] A. Mnih and G. Hinton. Three new graphical models for statistical language modeling. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 641–648. 2007.

[253] A. Mnih and G. Hinton. A scalable hierarchical distributed language model. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1081–1088. 2008.

[254] A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[255] A. Mnih and W.-T. Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1751–1758. 2012.

[256] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing arari with deep reinforcement learning. *Neural Information Processing Systems (NIPS) Deep Learning Workshop*, 2013. also arXiv:1312.5602v1.

[257] A. Mohamed, G. Dahl, and G. Hinton. Deep belief networks for phone recognition. In *Proceedings of Neural Information Processing Systems (NIPS) Workshop Deep Learning for Speech Recognition and Related Applications*. 2009.

[258] A. Mohamed, G. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, & Language Processing*, 20(1), January 2012.

[259] A. Mohamed, G. Hinton, and G. Penn. Understanding how deep belief networks perform acoustic modelling. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[260] A. Mohamed, D. Yu, and L. Deng. Investigation of full-sequence training of deep belief networks for speech recognition. In *Proceedings of Interspeech*. 2010.

[261] N. Morgan. Deep and wide: Multiple layers in automatic speech recognition. *IEEE Transactions on Audio, Speech, & Language Processing*, 20(1), January 2012.

[262] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos. Pushing the envelope — aside [speech recognition]. *IEEE Signal Processing Magazine*, 22(5):81–88, September 2005.

[263] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language models. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2005.

[264] K. Murphy. *Machine Learning — A Probabilistic Perspective*. The MIT Press, 2012.

[265] V. Nair and G. Hinton. 3-d object recognition with deep belief nets. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2009.

[266] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki. Voice conversion in high-order eigen space using deep belief nets. In *Proceedings of Interspeech*. 2013.

[267] H. Ney. Speech translation: Coupling of recognition and translation. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 1999.

[268] J. Ngiam, Z. Chen, P. Koh, and A. Ng. Learning deep energy models. In *Proceedings of International Conference on Machine Learning (ICML)*. 2011.

[269] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *Proceedings of International Conference on Machine Learning (ICML)*. 2011.

[270] M. Norouzi, T. Mikolov, S. Bengio, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. arXiv:1312.5650v2, 2013.

[271] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96:163–180, 2004.

[272] B. Olshausen. Can 'deep learning' offer deep insights about visual representation? *Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2012.

[273] M. Ostendorf. Moving beyond the 'beads-on-a-string' model of speech. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 1999.

[274] M. Ostendorf, V. Digalakis, and O. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5), September 1996.

[275] L. Oudre, C. Fevotte, and Y. Grenier. Probabilistic template-based chord recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2249–2259, November 2011.

[276] H. Palangi, L. Deng, and R. Ward. Learning input and recurrent weight matrices in echo state networks. *Neural Information Processing Systems (NIPS) Deep Learning Workshop*, December 2013.

[277] H. Palangi, R. Ward, and L. Deng. Using deep stacking network to improve structured compressive sensing with multiple measurement vectors. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[278] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:423–435, 2009.

[279] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. How to construct deep recurrent neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*. 2014.

[280] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*. 2013.

[281] J. Peng, L. Bo, and J. Xu. Conditional neural fields. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2009.

[282] P. Picone, S. Pike, R. Regan, T. Kamm, J. bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster. Initial evaluation of hidden dynamic models on conversational speech. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 1999.

[283] J. Pinto, S. Garimella, M. Magimai-Doss, H. Hermansky, and H. Bourlard. Analysis of MLP-based hierarchical phone posterior probability estimators. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2), February 2011.

[284] C. Plahl, T. Sainath, B. Ramabhadran, and D. Nahamoo. Improved pre-training of deep belief networks using sparse encoding symmetric machines. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[285] C. Plahl, R. Schlüter, and H. Ney. Hierarchical bottleneck features for LVCSR. In *Proceedings of Interspeech*. 2010.

[286] T. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, May 1995.

[287] T. Poggio. How the brain might work: The role of information and learning in understanding and replicating intelligence. In G. Jacovitt, A. Pettorossi, R. Consolo, and V. Senni, editors, *Information: Science and Technology for the New Century*, pages 45–61. Lateran University Press, 2007.

[288] J. Pollack. Recursive distributed representations. *Artificial Intelligence*, 46:77–105, 1990.

[289] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *Proceedings of Uncertainty in Artificial Intelligence*. 2011.

[290] D. Povey and P. Woodland. Minimum phone error and I-smoothing for improved discriminative training. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2002.

[291] R. Prabhavalkar and E. Fosler-Lussier. Backpropagation training for multilayer conditional random field based phone recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2010.

[292] A. Prince and P. Smolensky. Optimality: From neural networks to universal grammar. *Science*, 275:1604–1610, 1997.

[293] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286. 1989.

[294] M. Ranzato, Y. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2007.

[295] M. Ranzato, S. Chopra, Y. LeCun, and F.-J. Huang. Energy-based models in document recognition and computer vision. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*. 2007.

[296] M. Ranzato and G. Hinton. Modeling pixel means and covariances using factorized third-order boltzmann machines. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 2010.

[297] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2006.

[298] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton. On deep generative models with applications to recognition. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 2011.

[299] C. Rathinavalu and L. Deng. Construction of state-dependent dynamic parameters by maximum likelihood: Applications to speech recognition. *Signal Processing*, 55(2):149–165, 1997.

[300] S. Rennie, K. Fouset, and P. Dognin. Factorial hidden restricted boltzmann machines for noise robust speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[301] S. Rennie, H. Hershey, and P. Olsen. Single-channel multi-talker speech recognition — graphical modeling approaches. *IEEE Signal Processing Magazine*, 33:66–80, 2010.

[302] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*. 1993.

[303] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive autoencoders: Explicit invariance during feature extraction. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 833–840. 2011.

[304] A. Robinson. An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5:298–305, 1994.

[305] T. Sainath, L. Horesh, B. Kingsbury, A. Aravkin, and B. Ramabhadran. Accelerating hessian-free optimization for deep neural networks by implicit pre-conditioning and sampling. arXiv: 1309.1508v3, 2013.

[306] T. Sainath, B. Kingsbury, A. Mohamed, G. Dahl, G. Saon, H. Soltau, T. Beran, A. Aravkin, and B. Ramabhadran. Improvements to deep convolutional neural networks for LVCSR. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[307] T. Sainath, B. Kingsbury, A. Mohamed, and B. Ramabhadran. Learning filter banks within a deep neural network framework. In *Proceedings of The Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[308] T. Sainath, B. Kingsbury, and B. Ramabhadran. Autoencoder bottleneck features using deep belief networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[309] T. Sainath, B. Kingsbury, B. Ramabhadran, P. Novak, and A. Mohamed. Making deep belief networks effective for large vocabulary continuous speech recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2011.

[310] T. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[311] T. Sainath, B. Kingsbury, H. Soltau, and B. Ramabhadran. Optimization techniques to improve training speed of deep neural networks for large speech tasks. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2267–2276, November 2013.

[312] T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran. Convolutional neural networks for LVCSR. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[313] T. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky. Exemplar-based sparse representation features: From TIMIT to LVCSR. *IEEE Transactions on Speech and Audio Processing*, November 2011.

[314] R. Salakhutdinov and G. Hinton. Semantic hashing. In *Proceedings of Special Interest Group on Information Retrieval (SIGIR) Workshop on Information Retrieval and Applications of Graphical Models*. 2007.

[315] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2009.

[316] R. Salakhutdinov and G. Hinton. A better way to pretrain deep boltzmann machines. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[317] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[318] R. Sarikaya, G. Hinton, and B. Ramabhadran. Deep belief nets for natural language call-routing. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5680–5683. 2011.

[319] E. Schmidt and Y. Kim. Learning emotion-based acoustic features with deep belief networks. In *Proceedings IEEE of Signal Processing to Audio and Acoustics*. 2011.

[320] H. Schwenk. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of Computional Linguistics*. 2012.

[321] H. Schwenk, A. Rousseau, and A. Mohammed. Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the Joint Human Language Technology Conference and the North American Chapter of the Association of Computational Linguistics (HLT-NAACL) 2012 Workshop on the future of language modeling for Human Language Technology (HLT)*, pages 11–19.

[322] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. On parallelizability of stochastic gradient descent for speech DNNs. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2014.

[323] F. Seide, G. Li, X. Chen, and D. Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 24–29. 2011.

[324] F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of Interspeech*, pages 437–440. 2011.

[325] M. Seltzer, D. Yu, and E. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[326] M. Shannon, H. Zen, and W. Byrne. Autoregressive models for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, Language Processing*, 21(3):587–597, 2013.

[327] H. Sheikhzadeh and L. Deng. Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization. *IEEE Transactions on on Speech and Audio Processing (ICASSP)*, 2:80–91, 1994.

[328] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings World Wide Web*. 2014.

[329] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep fisher networks for large-scale image classification. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[330] M. Siniscalchi, J. Li, and C. Lee. Hermitian polynomial for speaker adaptation of connectionist speech recognition systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2152–2161, 2013a.

[331] M. Siniscalchi, T. Svendsen, and C.-H. Lee. A bottom-up modular search approach to large vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech, Language Processing*, 21, 2013.

[332] M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee. Exploiting deep neural networks for detection-based speech recognition. *Neurocomputing*, 106:148–157, 2013.

[333] M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee. Speech recognition using long-span temporal patterns in a deep network model. *IEEE Signal Processing Letters*, 20(3):201–204, March 2013.

[334] G. Sivaram and H. Hermansky. Sparse multilayer perceptrons for phoneme recognition. *IEEE Transactions on Audio, Speech, & Language Processing*, 20(1), January 2012.

[335] P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46:159–216, 1990.

[336] P. Smolensky and G. Legendre. *The Harmonic Mind — From Neural Computation to Optimality-Theoretic Grammar*. The MIT Press, Cambridge, MA, 2006.

[337] J. Snoek, H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[338] R. Socher. New directions in deep learning: Structured models, tasks, and datasets. *Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2012.

[339] R. Socher, Y. Bengio, and C. Manning. Deep learning for NLP. *Tutorial at Association of Computational Logistics (ACL), 2012, and North American Chapter of the Association of Computational Linguistics (NAACL)*, 2013. http://www.socher.org/index.php/DeepLearning Tutorial.

[340] R. Socher, D. Chen, C. Manning, and A. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[341] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 2010.

[342] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013b.

[343] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. *Neural Information Processing Systems (NIPS) Deep Learning Workshop*, 2013c.

[344] R. Socher, C. Lin, A. Ng, and C. Manning. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*. 2011.

[345] R. Socher, J. Pennington, E. Huang, A. Ng, and C. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2011.

[346] R. Socher, J. Pennington, E. Huang, A. Ng, and C. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. 2011.

[347] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. 2013.

[348] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[349] N. Srivastava and R. Salakhutdinov. Discriminative transfer learning with tree-based priors. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[350] R. Srivastava, J. Masci, S. Kazerounian, F. Gomez, and J. Schmidhuber. Compete to compute. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[351] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel. Preliminary investigation of boltzmann machine classifiers for speaker recognition. In *Proceedings of Odyssey*, pages 109–116. 2012.

[352] V. Stoyanov, A. Ropson, and J. Eisner. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2011.

[353] H. Su, G. Li, D. Yu, and F. Seide. Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[354] A. Subramanya, L. Deng, Z. Liu, and Z. Zhang. Multi-sensory speech processing: Incorporating automatically extracted hidden dynamic information. In *Proceedings of IEEE International Conference on Multimedia & Expo (ICME)*. Amsterdam, July 2005.

[355] J. Sun and L. Deng. An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition. *Journal on Acoustical Society of America*, 111(2):1086–1101, 2002.

[356] I. Sutskever. Training recurrent neural networks. Ph.D. Thesis, University of Toronto, 2013.

[357] I. Sutskever, J. Martens, and G. Hinton. Generating text with recurrent neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*. 2011.

[358] Y. Tang and C. Eliasmith. Deep networks for robust visual recognition. In *Proceedings of International Conference on Machine Learning (ICML)*. 2010.

[359] Y. Tang and R. Salakhutdinov. *Learning Stochastic Feedforward Neural Networks*. NIPS, 2013.

[360] A. Tarralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 2008.

[361] G. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2007.

[362] S. Thomas, M. Seltzer, K. Church, and H. Hermansky. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proceedings of Interspeech*. 2013.

[363] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of International Conference on Machine Learning (ICML)*. 2008.

[364] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, H. Yamagishi, and K. Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

[365] F. Triefenbach, A. Jalalvand, K. Demuynck, and J.-P. Martens. Acoustic modeling with hierarchical reservoirs. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2439–2450, November 2013.

[366] G. Tur, L. Deng, D. Hakkani-Tür, and X. He. Towards deep understanding: Deep convex networks for semantic utterance classification. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[367] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of Association for Computational Linguistics (ACL)*. 2010.

[368] Z. Tüske, M. Sundermeyer, R. Schlüter, and H. Ney. Context-dependent MLPs for LVCSR: TANDEM, hybrid or both? In *Proceedings of Interspeech*. 2012.

[369] B. Uria, S. Renals, and K. Richmond. A deep neural network for acoustic-articulatory speech inversion. *Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[370] R. van Dalen and M. Gales. Extended VTS for noise-robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):733–743, 2011.

[371] A. van den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[372] V. Vasilakakis, S. Cumani, and P. Laface. Speaker recognition by means of deep belief networks. In *Proceedings of Biometric Technologies in Forensic Science*. 2013.

[373] K. Vesely, A. Ghoshal, L. Burget, and D. Povey. Sequence-discriminative training of deep neural networks. In *Proceedings of Interspeech*. 2013.

[374] K. Vesely, M. Hannemann, and L. Burget. Semi-supervised training of deep neural networks. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[375] P. Vincent. A connection between score matching and denoising autoencoder. *Neural Computation*, 23(7):1661–1674, 2011.

[376] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[377] O. Vinyals, Y. Jia, L. Deng, and T. Darrell. Learning with recursive perceptual representations. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2012.

[378] O. Vinyals and D. Povey. Krylov subspace descent for deep learning. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2012.

[379] O. Vinyals and S. Ravuri. Comparing multilayer perceptron to deep belief network tandem features for robust ASR. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2011.

[380] O. Vinyals, S. Ravuri, and D. Povey. Revisiting recurrent neural networks for robust ASR. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[381] S. Wager, S. Wang, and P. Liang. Dropout training as adaptive regularization. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2013.

[382] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustical Speech, and Signal Processing*, 37:328–339, 1989.

[383] G. Wang and K. Sim. Context-dependent modelling of deep neural network using logistic regression. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2013.

[384] G. Wang and K. Sim. Regression-based context-dependent modeling of deep neural networks for speech recognition. *IEEE/Association for Computing Machinery (ACM) Transactions on Audio, Speech, and Language Processing*, 2014.

[385] D. Warde-Farley, I. Goodfellow, A. Courville, and Y. Bengi. An empirical analysis of dropout in piecewise linear networks. In *Proceedings of International Conference on Learning Representations (ICLR)*. 2014.

[386] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2005.

[387] C. Weng, D. Yu, M. Seltzer, and J. Droppo. Single-channel mixed speech recognition using deep neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2014.

[388] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, 2010.

[389] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. 2011.

[390] S. Wiesler, J. Li, and J. Xue. Investigations on hessian-free optimization for cross-entropy training of deep neural networks. In *Proceedings of Interspeech*. 2013.

[391] M. Wohlmayr, M. Stark, and F. Pernkopf. A probabilistic interaction model for multi-pitch tracking with factorial hidden markov model. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), May 2011.

[392] D. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[393] S. J. Wright, D. Kanevsky, L. Deng, X. He, G. Heigold, and H. Li. Optimization algorithms and applications for speech and language processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2231–2243, November 2013.

[394] L. Xiao and L. Deng. A geometric perspective of large-margin training of gaussian models. *IEEE Signal Processing Magazine*, 27(6):118–123, November 2010.

[395] X. Xie and S. Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 15:441–454, 2003.

[396] Y. Xu, J. Du, L. Dai, and C. Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68, 2014.

[397] J. Xue, J. Li, and Y. Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *Proceedings of Interspeech*. 2013.

[398] S. Yamin, L. Deng, Y. Wang, and A. Acero. An integrative and discriminative technique for spoken utterance classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:1207–1214, 2008.

[399] Z. Yan, Q. Huo, and J. Xu. A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR. In *Proceedings of Interspeech*. 2013.

[400] D. Yang and S. Furui. Combining a two-step CRF model and a joint source-channel model for machine transliteration. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 275–280. 2010.

[401] K. Yao, D. Yu, L. Deng, and Y. Gong. A fast maximum likelihood nonlinear feature transformation method for GMM-HMM speaker adaptation. *Neurocomputing*, 2013a.

[402] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[403] K. Yao, G. Zweig, M. Hwang, Y. Shi, and D. Yu. Recurrent neural networks for language understanding. In *Proceedings of Interspeech*. 2013.

[404] T. Yoshioka and T. Nakatani. Noise model transfer: Novel approach to robustness against nonstationary noise. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2182–2192, 2013.

[405] T. Yoshioka, A. Ragni, and M. Gales. Investigation of unsupervised adaptation of DNN acoustic models with filter bank input. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[406] L. Younes. On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics and Stochastic Reports*, 65(3):177–228, 1999.

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

[408] D. Yu, D. Deng, and S. Wang. Learning in the deep-structured conditional random fields. *Neural Information Processing Systems (NIPS) 2009 Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[409] D. Yu and L. Deng. Solving nonlinear estimation problems using splines. *IEEE Signal Processing Magazine*, 26(4):86–90, July 2009.

[410] D. Yu and L. Deng. Deep-structured hidden conditional random fields for phonetic recognition. In *Proceedings of Interspeech*. September 2010.

[411] D. Yu and L. Deng. Accelerated parallelizable neural networks learning algorithms for speech recognition. In *Proceedings of Interspeech*. 2011.

[412] D. Yu and L. Deng. Deep learning and its applications to signal and information processing. *IEEE Signal Processing Magazine*, pages 145–154, January 2011.

[413] D. Yu and L. Deng. Efficient and effective algorithms for training single-hidden-layer neural networks. *Pattern Recognition Letters*, 33:554–558, 2012.

[414] D. Yu, L. Deng, and G. E. Dahl. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. *Neural Information Processing Systems (NIPS) 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, December 2010.

[415] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero. Robust speech recognition using cepstral minimum-mean-square-error noise suppressor. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5), July 2008.

[416] D. Yu, L. Deng, Y. Gong, and A. Acero. A novel framework and training algorithm for variable-parameter hidden markov models. *IEEE Transactions on Audio, Speech and Language Processing*, 17(7):1348–1360, 2009.

[417] D. Yu, L. Deng, X. He, and A. Acero. Large-margin minimum classification error training: A theoretical risk minimization perspective. *Computer Speech and Language*, 22(4):415–429, October 2008.

[418] D. Yu, L. Deng, X. He, and X. Acero. Large-margin minimum classification error training for large-scale speech recognition tasks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2007.

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

[420] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, and A. Acero. Cross-lingual speech recognition under runtime resource constraints. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2009b.

[421] D. Yu, L. Deng, and F. Seide. Large vocabulary speech recognition using deep tensor neural networks. In *Proceedings of Interspeech*. 2012c.

[422] D. Yu, L. Deng, and F. Seide. The deep tensor neural network with applications to large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):388–396, 2013.

[423] D. Yu, J.-Y. Li, and L. Deng. Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech and Language*, 19:2461–2473, 2010.

[424] D. Yu, F. Seide, G. Li, and L. Deng. Exploiting sparseness in deep neural networks for large vocabulary speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[425] D. Yu and M. Seltzer. Improved bottleneck features using pre-trained deep neural networks. In *Proceedings of Interspeech*. 2011.

[426] D. Yu, M. Seltzer, J. Li, J.-T. Huang, and F. Seide. Feature learning in deep neural networks — studies on speech recognition. In *Proceedings of International Conference on Learning Representations (ICLR)*. 2013.

[427] D. Yu, S. Siniscalchi, L. Deng, and C. Lee. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2012.

[428] D. Yu, S. Wang, and L. Deng. Sequential labeling using deep-structured conditional random fields. *Journal of Selected Topics in Signal Processing*, 4:965–973, 2010.

[429] D. Yu, S. Wang, Z. Karam, and L. Deng. Language recognition using deep-structured conditional random fields. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5030–5033. 2010.

[430] D. Yu, K. Yao, H. Su, G. Li, and F. Seide. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2013.

[431] K. Yu, M. Gales, and P. Woodland. Unsupervised adaptation with discriminative mapping transforms. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):714–723, 2009.

[432] K. Yu, Y. Lin, and H. Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *Proceedings Computer Vision and Pattern Recognition (CVPR)*. 2011.

[433] F. Zamora-Martínez, M. Castro-Bleda, and S. España-Boquera. Fast evaluation of connectionist language models. *International Conference on Artificial Neural Networks*, pages 144–151, 2009.

[434] M. Zeiler. Hierarchical convolutional deep learning in computer vision. Ph.D. Thesis, New York University, January 2014.

[435] M. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*. 2013.

[436] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. arXiv:1311.2901, pages 1–11, 2013.

[437] M. Zeiler, G. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of International Conference on Computer vision (ICCV)*. 2011.

[438] H. Zen, M. Gales, J. F. Nankaku, and Y. K. Tokuda. Product of experts for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):794–805, March 2012.

[439] H. Zen, Y. Nankaku, and K. Tokuda. Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech, and Language Processings*, 19(2):417–430, February 2011.

[440] H. Zen, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 7962–7966. 2013.

[441] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur. Improving deep neural network acoustic models using generalized maxout networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2014.

[442] X. Zhang and J. Wu. Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):697–710, 2013.

[443] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng. Multi-sensory microphones for robust speech detection, enhancement and recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2004.

[444] Y. Zhao and B. Juang. Nonlinear compensation using the gauss-newton method for noise-robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2191–2206, 2012.

[445] W. Zou, R. Socher, D. Cer, and C. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. 2013.

[446] G. Zweig and P. Nguyen. A segmental CRF approach to large vocabulary continuous speech recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2009.