Industrial Impact of Deep Learning

from speech to language/multimodality

Li Deng

Deep Learning Technology Center, Microsoft Research, Redmond, WA. USA

November 19, 2014 at BAMMF

Acknowledgements: Geoff Hinton, Dong Yu, Xiaodong He, Jianfeng Gao, Yelong Shen, Xinying Song, Jinyu Li, Yi-fan Gong, Frank Seide, Mike Seltzer, Qiang Huo, Alex Acero, George Dahl, A. Mohamed, Po-sen Huang, Vincent Vanhoucke, Andrew Senior, Tara Sainath, Brian Kingsbury, John Bridle, Nelson Morgan, Hynek Hermansky, Paul Smolensky, Chris Manning, Eric Xing, Chin-Hui Lee, John Hershey, Mari Ostendorf, Les Atlas

Main message

Deep Learning

≈

Neural Networks, Deep

in space & **time (recurrent** LSTM)

+

Generative Models, Deep

in space & **time (dynamic**)

+

…, …, …

	Deep Neural Nets	Deep Generative Models
Structure	Graphical; info flow: **bottom-up**	Graphical; info flow: **top-down**
Incorp constraints & domain knowledge	Hard	**Easy**
Semi/unsupervised	Harder	Easier
Interpretation	Harder	**Easy** (generative "story")
Representation	**Distributed**	Localist (mostly)
Inference/decode	Easy	Harder (but note recent progress)
Scalability/compute	**Easier (regular computes/GPU)**	Harder (but note recent progress)
Incorp. uncertainty	Hard	**Easy**
Empirical goal	Classification, feature learning, …	Classification (via Bayes rule), latent variable inference…
Terminology	Neurons, activation/gate functions, weights …	Random vars, stochastic "neurons", potential function, parameters …
Learning algorithm	A single, unchallenged, algorithm -- BackProp	A major focus of open research, many algorithms, & more to come
Evaluation	On a black-box score – end performance	On almost every intermediate quantity
Implementation	Many untold-tricks	More or less standardized
Experiments	Massive, real data	Modest, often simulated data

J. Acoust. Soc. Am. 121 (2), February 2007

Speech production knowledge in automatic speech recognition

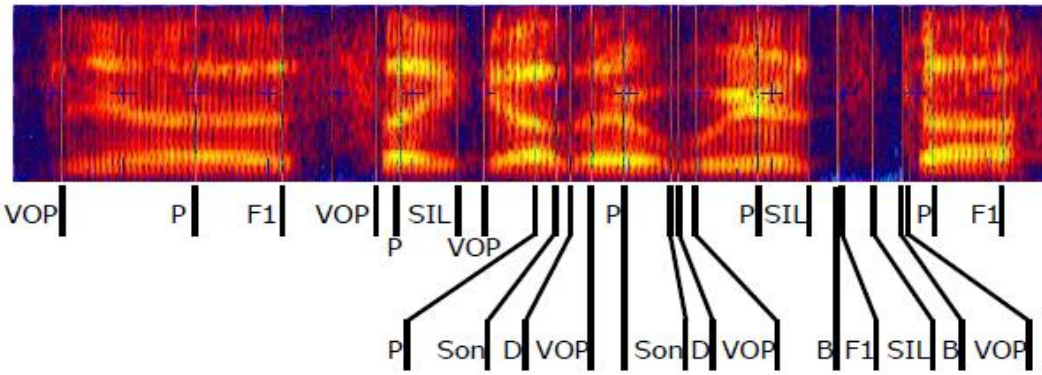
Simon King[a] and Joe Frankel
Centre for Speech Technology Research, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom

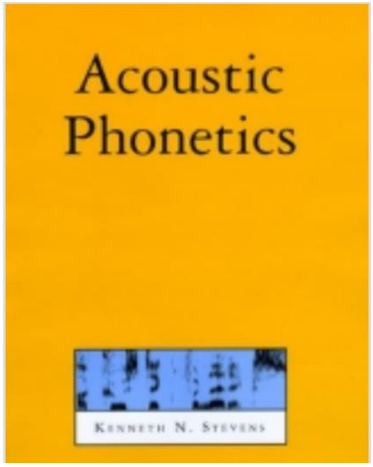
Karen Livescu
MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Room 32-G482, Cambridge, Massachusetts 02139

Erik McDermott
Nippon Telegraph and Telephone Corporation, NTT Communication Science Laboratories, 2–4 Hikari-dai, Seika-cho, Soraku-gun Kyoto-fu 619-0237, Japan

Korin Richmond and Mirjam Wester
Centre for Speech Technology Research, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom

(Received 13 October 2005; revised 1 November 2006; accepted 6 November 2006)





MOVING BEYOND THE 'BEADS-ON-A-STRING' MODEL OF SPEECH

ASRU-1999

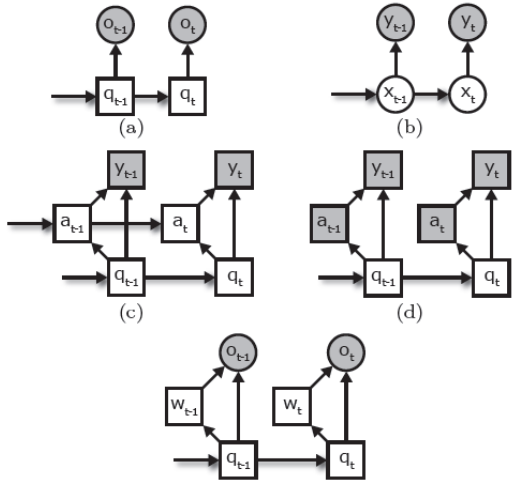
M. Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, WA 98195

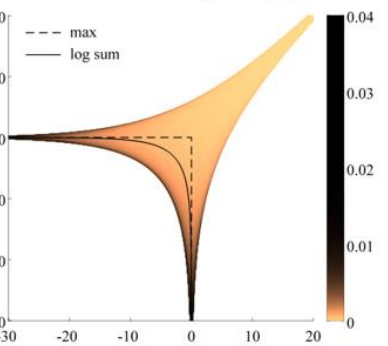
Computational Models of Speech Pattern Processing
NATO ASI Series Volume 169, 1999, pp 199-213

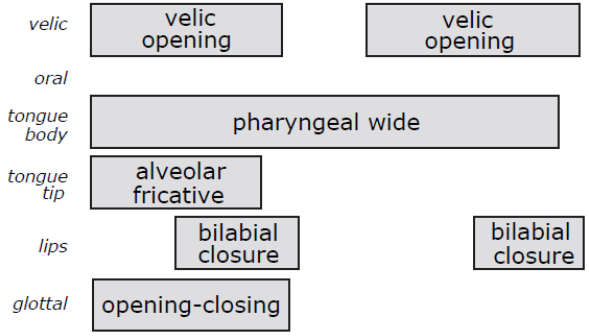
Computational Models for Speech Production

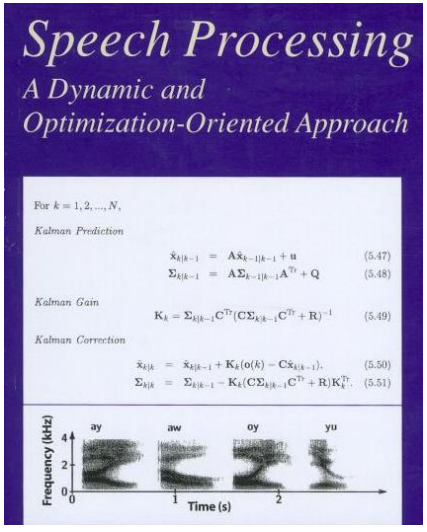
$$\frac{d}{dt}\begin{pmatrix}\mathbf{z}(t)\\ \dot{\mathbf{z}}(t)\end{pmatrix}=\begin{pmatrix}0 & 1\\ -\mathbf{S}^2(t) & -2\mathbf{S}(t)\end{pmatrix}\begin{pmatrix}\mathbf{z}(t)\\ \dot{\mathbf{z}}(t)\end{pmatrix}+\begin{pmatrix}0\\ \mathbf{S}^2(t)\mathbf{Z}^0(t)\end{pmatrix}+\begin{pmatrix}0\\ \mathbf{w}(t)\end{pmatrix}$$



$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h})=\frac{1}{2}\left|\mathrm{diag}\left(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}\right)\right|$$
$$\mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}-e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}-e^{-(\mathbf{n}-\mathbf{x}-\mathbf{h})/2});\mathbf{0},\boldsymbol{\Sigma}_\alpha\right]$$







Outline

- **Part I:** A (brief) early history of 'deep' speech recognition & industrial impact of deep learning: **speech** (& vision)

- **Part II:** Rapid progress of deep learning: **Natural language**, **multimodality**, intelligence in the big-data world, etc.

Outline

- **Part I:** A (brief) early history of 'deep' speech recognition
  - (shallow) neural nets in ASR, before 2009
  - (deep) generative models in ASR and ML, before 2009
  - how/why DNN made recent inroad into ASR
  - roles of academic-industrial collaboration

& industrial impact of deep learning

Neural Networks in Speech Recognition
(prior to the rise of deep learning)

Temporal & Time-Delay (1-D Convolutional) Neural Nets

- Atlas, Homma, and Marks, "An Artificial Neural Network for Spatio-Temporal Bipolar Patterns, Application to Phoneme Classification," NIPS 1987. — 1987
- Waibel, Hanazawa, Hinton, Shikano, Lang. "Phoneme recognition using time-delay neural networks." IEEE Transactions on Acoustics, Speech and Signal Processing, 1989. — 1989

Recurrent Neural Nets — 1991

- Bengio. "Artificial Neural Networks and their Application to Speech/Sequence Recognition", Ph.D. thesis, 1991.
- Robinson. "A real-time recurrent error propagation network word recognition system," ICASSP 1992. — 1992

Hybrid Neural Nets-HMM

- Morgan, Bourlard, Renals, Cohen, Franco. "Hybrid neural network/hidden Markov model systems for continuous speech recognition," IJPRAI, 1993. — 1993

Neural-Net Nonlinear Prediction

- Deng, Hassanein, Elmasry. "Analysis of correlation structure for a neural predictive model with applications to speech recognition," *Neural Networks*, vol. 7, No. 2, 1994. — 1994

Bidirectional Recurrent Neural Nets

- Schuster, Paliwal. "Bidirectional recurrent neural networks," IEEE Trans. Signal Processing, 1997. — 1997

Neural-Net TANDEM

- Hermansky, Ellis, Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000. — 2000
- Morgan, Zhu, Stolcke, Sonmez, Sivadas, Shinozaki, Ostendorf, Jain, Hermansky, Ellis, Doddington, Chen, Cretin, Bourlard, Athineos, "Pushing the envelope - aside [speech recognition]," IEEE Signal Processing Magazine, vol. 22, no. 5, 2005. — 2005

← **DARPA EARS Program 2001-2004: Novel Approach I**

Bottle-neck Features extracted from Neural-Nets

- Grezl, Karafiat, Kontar & Cernocky. "Probabilistic and bottle-neck features for LVCSR of meetings," ICASSP, 2007. — 2007

Historical Development and Future Directions in Speech Recognition and Understanding

Janet M. Baker, Li Deng, Sanjeev Khudanpur, Chin-Hui Lee, James Glass, and Nelson Morgan

This report is one of five reports that were based on the MINDS workshops, led by Donna Harman (NIST) and sponsored by Heather McCallum-Bayliss of the Disruptive Technology Office of the Office of the Director of National Intelligence's Office of Science and Technology (ODNI/ADDNI/S&T/DTO). To find the rest of the reports, and an executive overview, please see http://www.itl.nist.gov/iaui/894.02/minds.html.

…, …, …

3. Models, Algorithms, and Search

Machine Learning: This is an exciting time in the machine learning community. Many new machine-learning algorithms are being explored and are achieving impressive results on a wide variety of tasks. Recent examples include graphical models, conditional random fields, (partially observable) Markov decision processes, reinforcement-based learning and discriminative methods such as large-margin or log-linear (max entropy) models. Recent developments in effective training of these models make them worthy of further exploration. The speech community would do well to explore common ground with the machine learning community in these areas.

Deep Generative Models in Speech Recognition
(prior to the rising of deep learning)

Segment & Nonstationary-State Models

- Digalakis, Rohlicek, Ostendorf. "ML estimation of a stochastic linear system with the EM alg & application to speech recognition," IEEE T-SAP, 1993 — 1993
- Deng, Aksmanovic, Sun, Wu, Speech recognition using HMM with polynomial regression functions as nonstationary states," IEEE T-SAP, 1994. — 1994

Hidden Dynamic Models (HDM)

- Deng, Ramsay, Sun. "Production models as a structural basis for automatic speech recognition," Speech Communication, vol. 33, pp. 93–111, 1997. — 1997
- Bridle et al. "An investigation of segmental hidden dynamic models of speech coarticulation for speech recognition," Final Report Workshop on Language Engineering, Johns Hopkins U, 1998. — 1998
- Picone et al. "Initial evaluation of hidden dynamic models on conversational speech," ICASSP, 1999. — 1999
- Deng and Ma. "Spontaneous speech recognition using a statistical co-articulatory model for the vocal tract resonance dynamics," JASA, 2000. — 2000

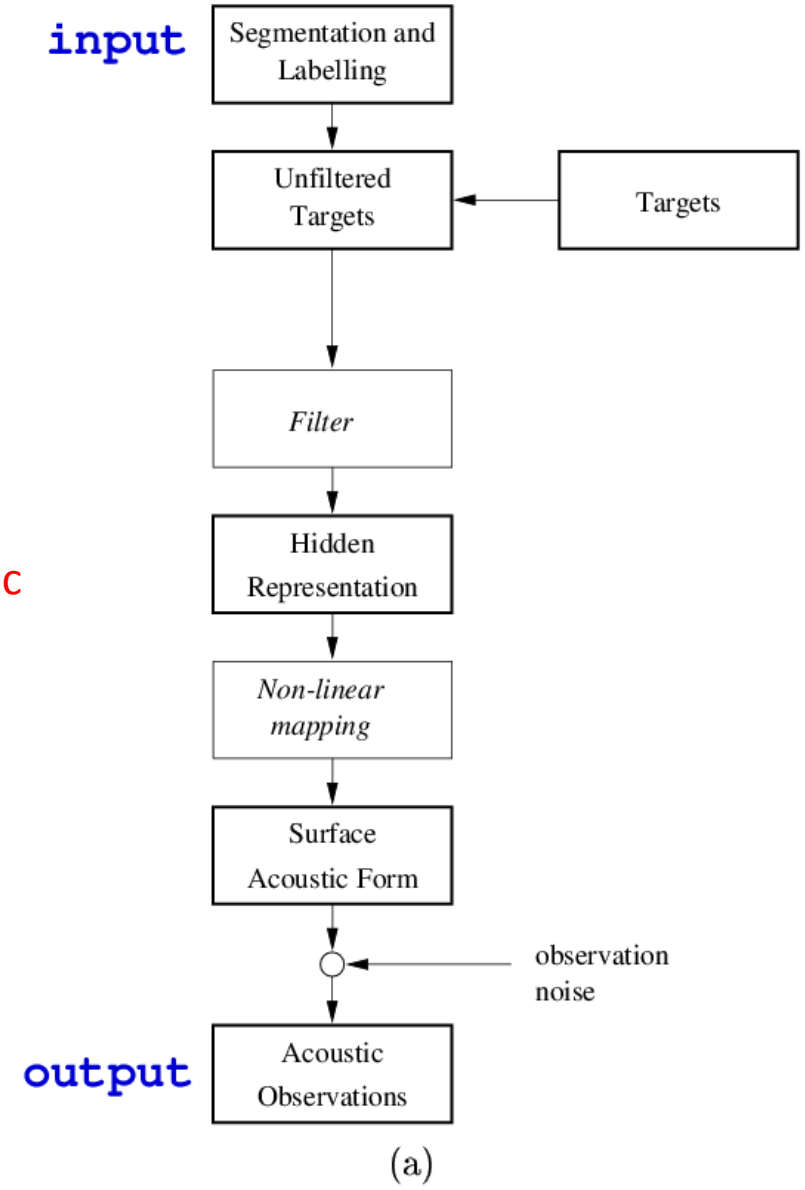
Structured Hidden Trajectory Models (HTM)

- Zhou, et al. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM," ICASSP, 2003. ← **DARPA EARS Program 2001-2004: Novel Approach II** — 2003
- Deng, Yu, Acero. "Structured speech modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006. — 2006

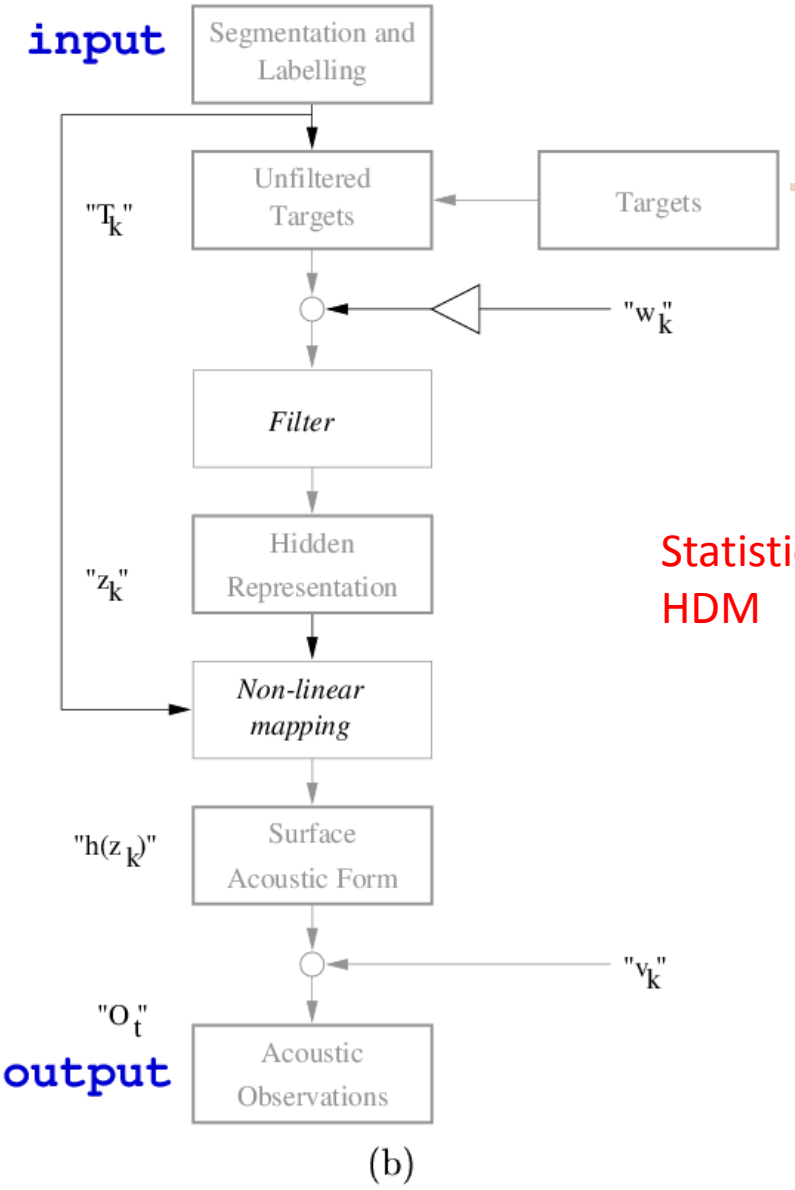
Switching Nonlinear State-Space Models

- Deng. "Switching Dynamic System Models for Speech Articulation and Acoustics," in Mathematical Foundations of Speech and Language Processing, vol. 138, pp. 115 - 134, Springer, 2003.
- Lee et al. "A Multimodal Variational Approach to Learning and Inference in Switching State Space Models," ICASSP, 2004.

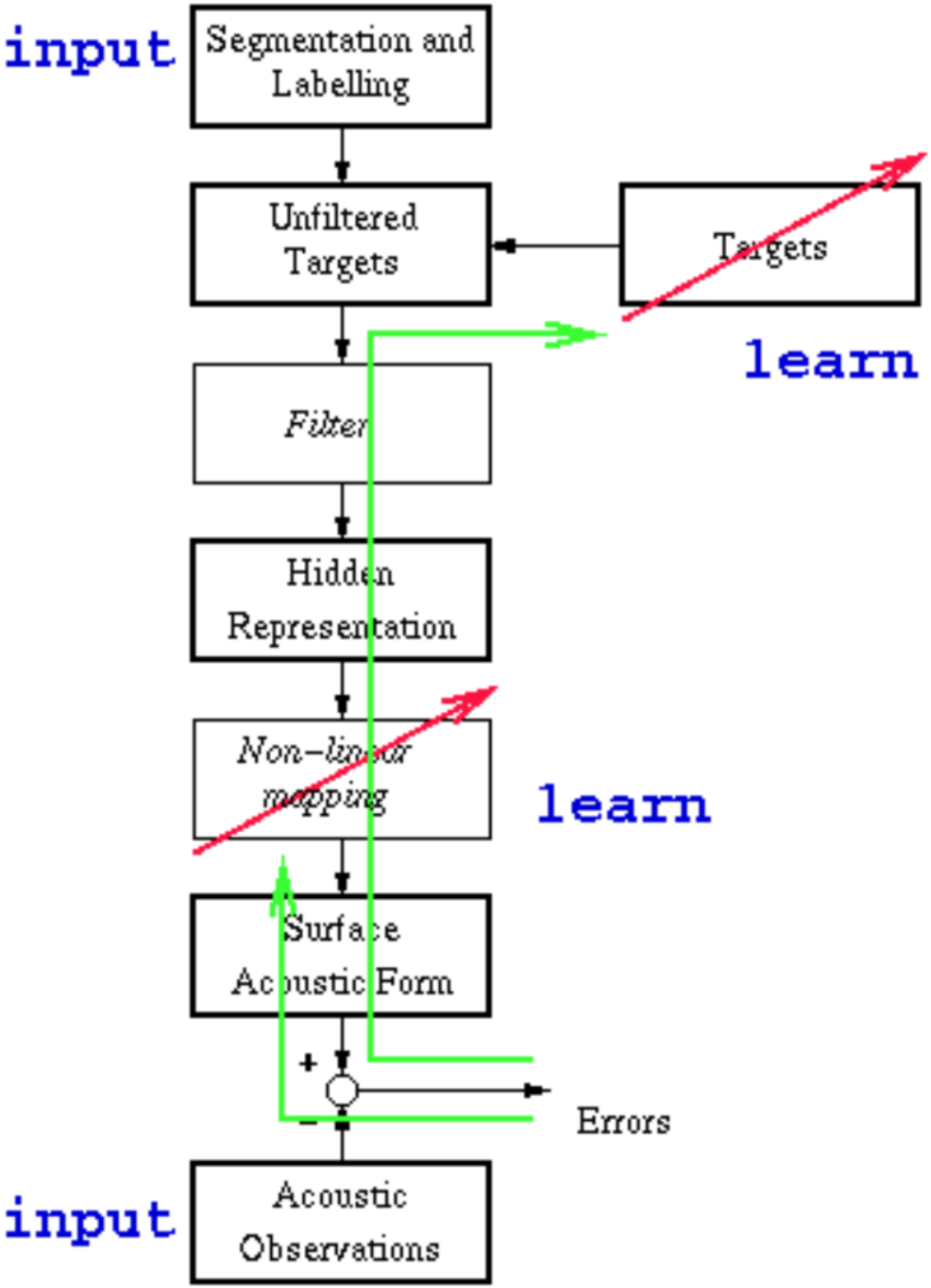
Deterministic HDM



Statistical HDM



Bridle, Deng, Picone, Richards, Ma, Kamm, Schuster, Pike, Reagan. Final Report for Workshop on Language Engineering, Johns Hopkins University, 1998.

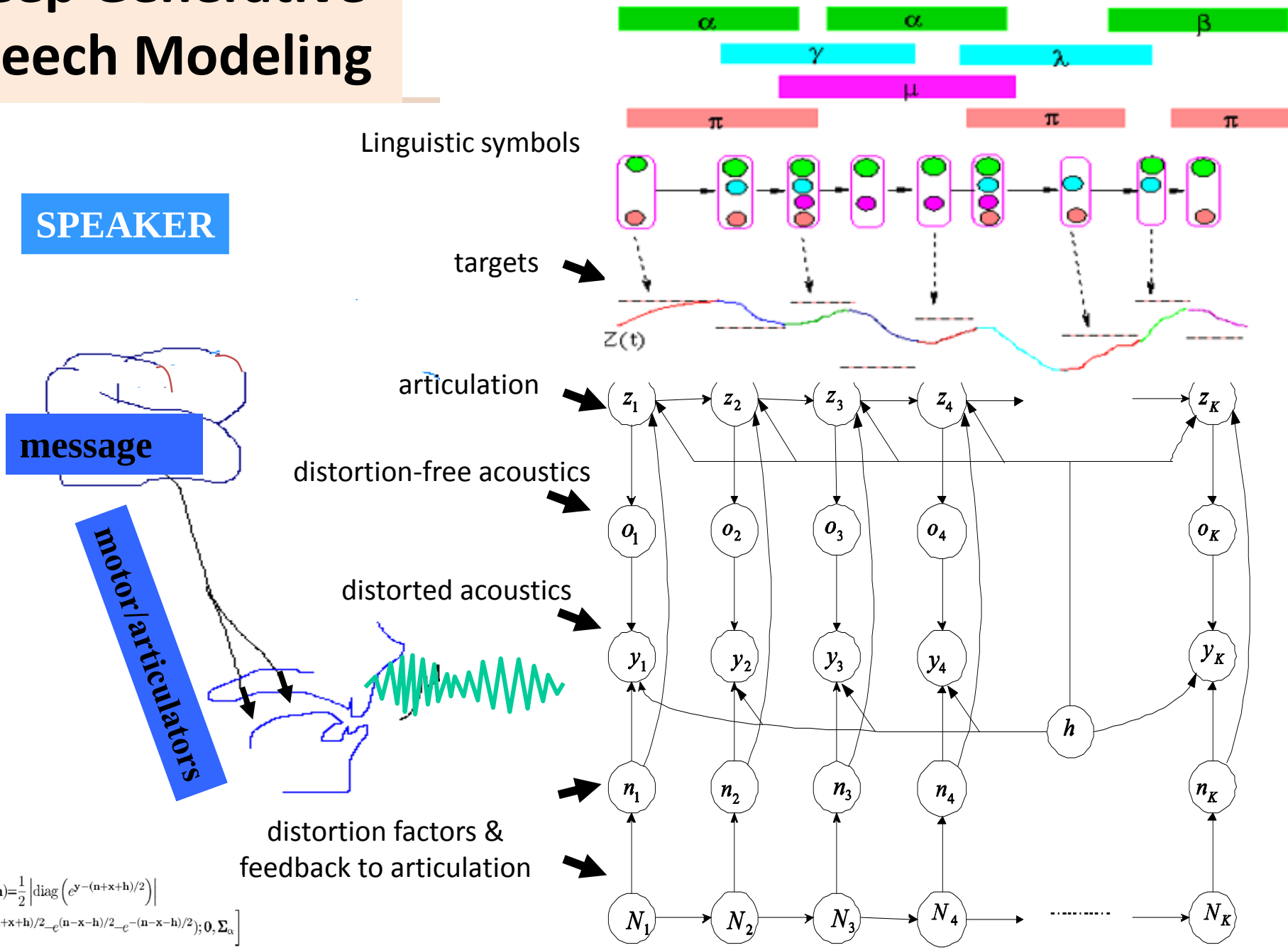


- Learning: Gradient descent
- "Back"-propagation in forward direction (model is generative)
- MSE for BP: acoustic obs, not labels
- Everything is interpretable
- Evaluation on SWBD: disappointment
- We understand why now, not in 1998
- Part II to discuss deep RNN vs. HDM

http://old-site.clsp.jhu.edu/ws98/projects/dynamic/presentations/hywel/presentation/training.html

Deep Generative Speech Modeling

Ostendorf. “Moving beyond the ‘beads-on-a-string’ model of speech,” ASRU, 1999
Deng, Wu, “Hierarchical partitioning of the articulatory state space …,” ICSLP, 1996



$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = \frac{1}{2}\left|\mathrm{diag}\left(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}\right)\right| \mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2} - e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2} - e^{-(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}); \mathbf{0}, \boldsymbol{\Sigma}_\alpha\right]$$

ICASSP-2004

A MULTIMODAL VARIATIONAL APPROACH TO LEARNING AND INFERENCE IN SWITCHING STATE SPACE MODELS

Leo J. Lee[1,2], *Hagai Attias*[2], *Li Deng*[2] *and Paul Fieguth*[3]

University of Waterloo
[1]Electrical & Computer Engineering
[3]Systems Design Engineering
Waterloo, ON, N2L 3G1
Canada

[2]Microsoft Corporation
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6339
USA

Auxiliary function:

$$\mathcal{F}[q] = \sum_{s_{1:N}} \int d\mathbf{x}_{1:N}\; q(s_{1:N}, \mathbf{x}_{1:N}) \cdot [\log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, s_{1:N}) - \log q(s_{1:N}, \mathbf{x}_{1:N})]$$

In the variational approach we approximate the exact posterior $p(s_{1:N}, \mathbf{x}_{1:N} \mid \mathbf{y}_{1:N})$ by a distribution with a tractable structure, denoted by q. Here we choose the following partially factorized structure shown graphically in Fig. 1:

$$p(s_{0:N}, \mathbf{x}_{0:N} \mid \mathbf{y}_{1:N}) \approx q(s_{0:N}, \mathbf{x}_{0:N} \mid \mathbf{y}_{1:N}) = \prod_{n=1}^{N} q(\mathbf{x}_n \mid s_n) q(s_n \mid s_{n-1}) \cdot q(x_0 \mid s_0) q(s_0) \; . \tag{5}$$

E-step: sufficient statistics. As usual, the variational equations above are coupled, with the equations for $\rho_{s,n}$, $\Gamma_{s,n}$ depend on $\eta_{s's,n}$, $\gamma_{s,n}$ and vice versa. These equations are solved iteratively starting from a random or more suitable initialization if available. The solution is the set of sufficient statistics

$$\varphi = \{\rho_{s,n}, \Gamma_{s,n}, \eta_{ss',n}, \gamma_{s,n}\} \tag{16}$$

which are moments of the variational posterior.

M-step: parameter estimation. Given the sufficient statistics φ, the derivation of the M-step is achieved by taking derivatives of $\mathcal{F}$ w.r.t. the model parameters (details omitted).

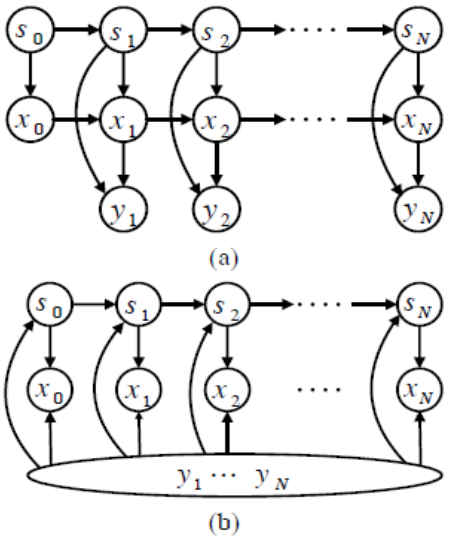
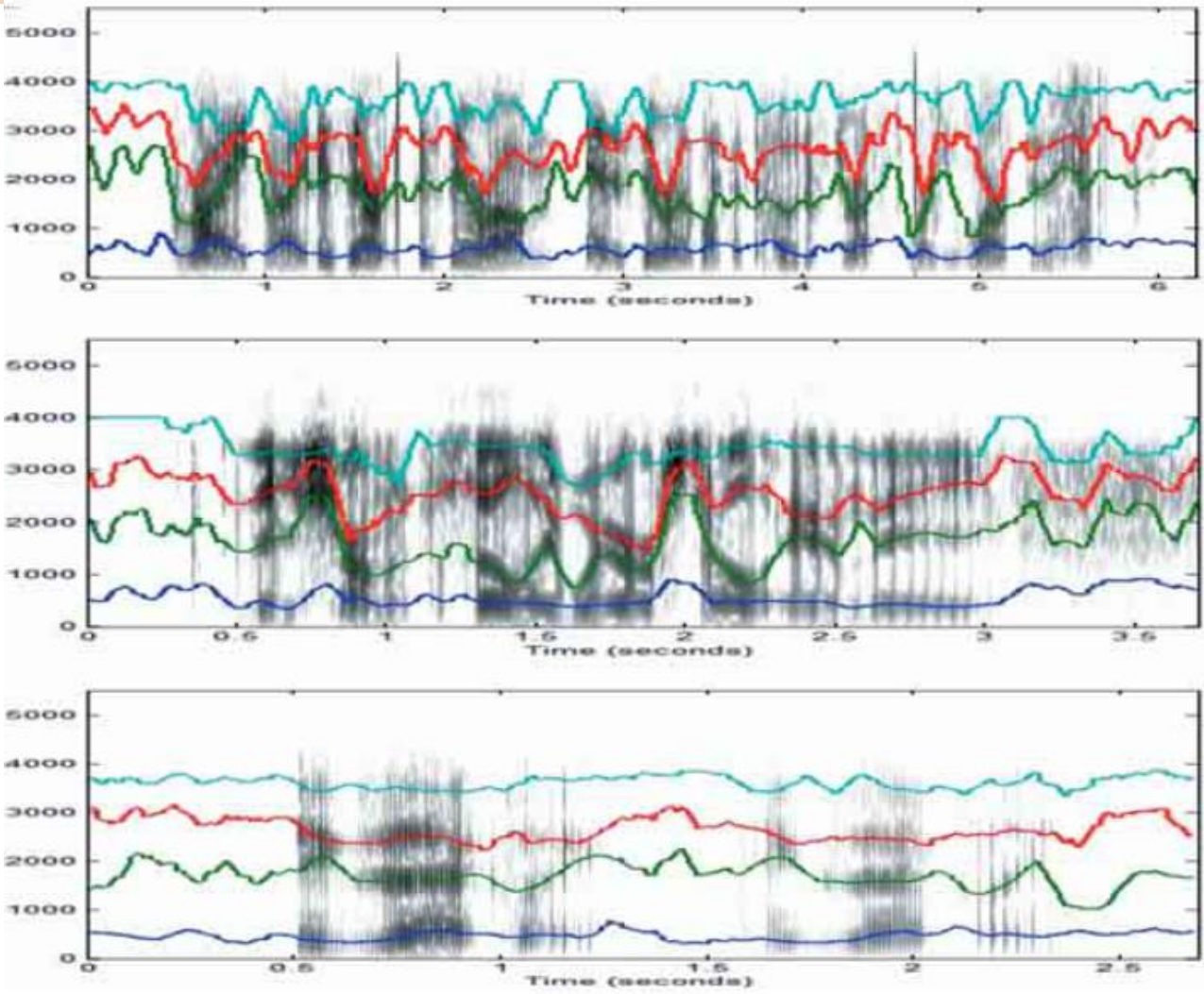


Fig. 1. The model (a) and the variational posterior (b) represented as Bayesian networks.

Surprisingly Good Inference Results for Continuous Hidden States

- By-product: accurately tracking dynamics of resonances (formants) in vocal tract (TIMIT & SWBD).
- Best formant tracker (speech analysis); used as basis to form a formant database as "ground truth"
- We thought we solved the ASR problem, except
- "Intractable" for decoding



Deng & Huang, Challenges in Adopting Speech Recognition, *Communications of the ACM*, vol. 47, pp. 69-75, 2004.

Deng, Cui, Pruvenok, Huang, Momen, Chen, Alwan, A Database of Vocal Tract Resonance Trajectories for Research in Speech , ICASSP, 2006.

Structured Speech Modeling

Li Deng, *Fellow, IEEE*, Dong Yu, *Member, IEEE*, and Alex Acero, *Fellow, IEEE*

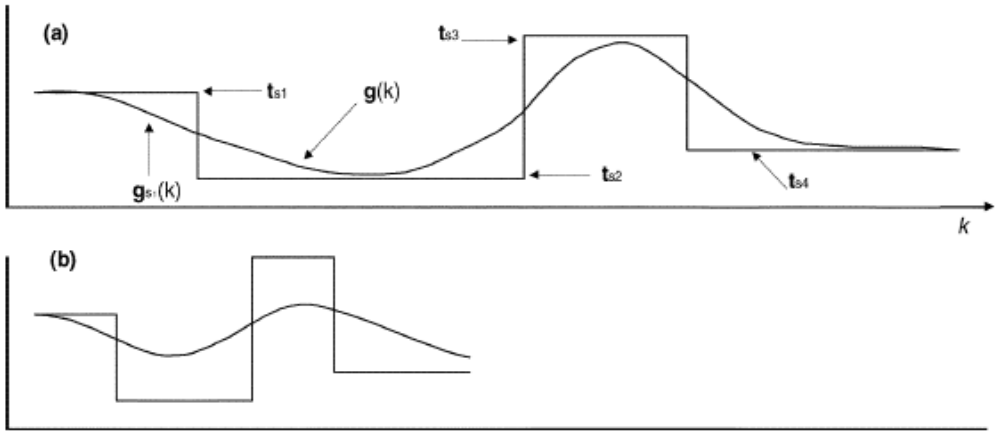


Fig. 1. Illustrations of the various VTR quantities in model Stage-I in utterance with four phone segments. (a) and (b) are for the same four VT targets and their filtered results, but the durations of the four segments a shorter in (b) than in (a).

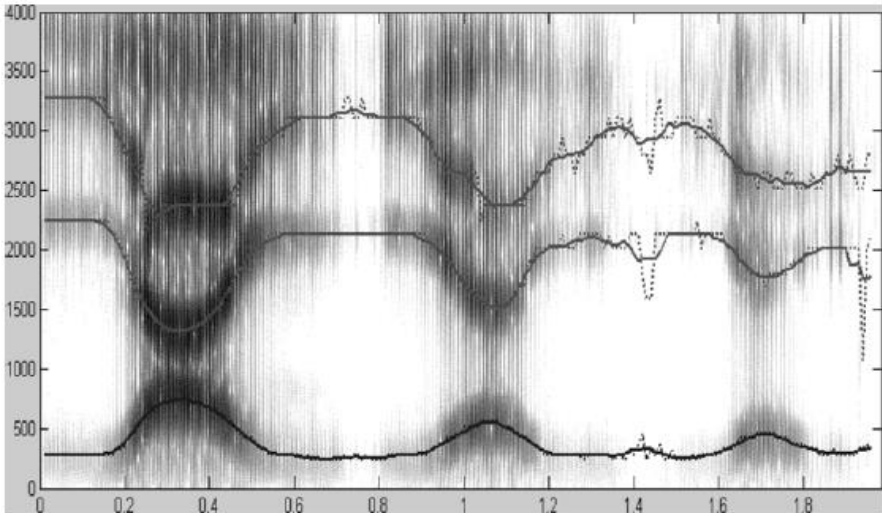


TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRE
WITHIN EACH OF FOUR BROAD PHONE CLASSES

	Sonorants	Stops	Fricatives	Closures
Occurrences	3814	889	1252	1578
HMM	64.05	72.10	75.64	88.72
HTM	72.42	76.27	75.74	90.94

- Elegant model formulation & knowledge incorporation
- Strong empirical results: **96%** TIMIT accuracy with Nbest=1001; **75.2%** lattice decoding w. monophones; fast approx. training
- Still very expensive for decoding; could not ship (very frustrating!)

--corrected many errors: especially “short” phones
--easy to interpret from the “generative” story

Another Deep Generative Model

(developed outside speech)

- Sigmoid belief nets & wake/sleep alg. (1992)
- Deep belief nets (DBN, 2006);

→ Start of deep learning

- Totally non-obvious result:
  Stacking many RBMs (undirected)

→ not Deep Boltzmann Machine (**DBM**, undirected)

→ but a DBN (directed, generative model)

- Excellent in generating images & speech synthesis

- Similar type of deep generative models to HDM
- But simpler: no temporal dynamics
- With very different parameterization
- Most intriguing of DBN: inference is easy
  (i.e. no need for approximate variational Bayes)
  ← "Restriction" of connections in RBM

- Pros/cons analysis→ Hinton coming to MSR 2009

LETTER Communicated by Yann Le Cun

A Fast Learning Algorithm for Deep Belief Nets

Geoffrey E. Hinton
hinton@cs.toronto.edu
Simon Osindero
osindero@cs.toronto.edu
Department of Computer Science, University of Toronto, Toronto, Canada M5S 3G4

Yee-Whye Teh
tehyw@comp.nus.edu.sg
Department of Computer Science, National University of Singapore, Singapore 117543

We show how to use "complementary priors" to eliminate the explaining-away effects that make inference difficult in densely connected belief nets that have many hidden layers. Using complementary priors, we derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory. The fast, greedy algorithm is used to initialize a slower learning procedure that fine-tunes the weights using a contrastive version of the wake-sleep algorithm. After fine-tuning, a network with three hidden layers forms a very good generative model of the joint distribution of handwritten digit images and their labels. This generative model gives better digit classification than the best discriminative learning algorithms. The low-dimensional manifolds on which the digits lie are modeled by long ravines in the free-energy landscape of the top-level associative memory, and it is easy to explore these ravines by using the directed connections to display what the associative memory has in mind.

then Geoff Hinton came to MSR (2009)

- **Kluge 1:** keep the assumption of "frame" independence (i.e. ignore real "dynamics" to facilitate inference/decoding) but use bigger time windows to approximate the effects.
- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets → no problem of "explaining away")
- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.
- **Well-timed** academic-industrial collaboration:
  - ASR industry searching for new solutions when "principled" deep generative approaches could not deliver
  - Academia developed deep learning tools (**DBN**/DNN with hybrid generative/discriminative, 2006) looking for applications
  - Advent of GPU computing (Nvidia CUDA library released 2007/08)
  - Big training data in ASR were available

- NIPS Home
- Overview
- Conference Videos
- Workshop Videos
- Program Highlights
- Tutorials
- Conference Sessions
- Workshops
- Publication Models
- Demonstrations
- Mini Symposia
- Accepted Papers
- Dates
- Committees
- Sponsors
- Awards
- Board

Li Deng, Dong Yu, Geoffrey Hinton

Microsoft Research; Microsoft Research; University of Toronto

Deep Learning for Speech Recognition and Related Applications

7:30am - 6:30pm Saturday, December 12, 2009

Location: Hilton: Cheakamus

Abstract: Over the past 25 years or so, speech recognition technology has been dominated by a "shallow" architecture --- hidden Markov models (HMMs). Significant technological success has been achieved using complex and carefully engineered variants

Invitee 1: give me one week to decide …,…

Not worth my time to fly to Vancouver for this…

there has been virtually no effective communication between machine learning researchers and speech recognition researchers who are both advocating the use of deep architecture and learning. One goal of the proposed workshop is to bring together these two groups of researchers to review the progress in both fields and to identify promising and synergistic research directions for potential future cross-fertilization and collaboration.

Mohamed, Dahl, Hinton, Deep belief networks for phone recognition, *NIPS 2009 Workshop on Deep Learning,* 2009

Yu, Deng, Wang, Learning in the Deep-Structured Conditional Random Fields, *NIPS 2009 Workshop on Deep Learning,* 2009

…, …, …

NIPS Home
Overview
Conference Videos
Workshop Videos
Program Highlights
Tutorials
Conference Sessions
Workshops
Publication Models
Demonstrations
Mini Symposia
Accepted Papers
Dates
Committees
Sponsors
Awards
Board

Li Deng, Dong Yu, Geoffrey Hinton

Microsoft Research; Microsoft Research; University of Toronto

Deep Learning for Speech Recognition and Related Applications

7:30am - 6:30pm Saturday, December 12, 2009

Location: Hilton: Cheakamus

Abstract: Over the past 25 years or so, speech recognition technology has been dominated by a "shallow" architecture --- hidden Markov models (HMMs). Significant

Invitee 2: A crazy idea…
Waveform for ASR is not like pixels for image recognition. It is more like using photons!!!

theoretical guidance to facilitate the development of these deep architectures. Further, there has been virtually no effective communication between machine learning researchers and speech recognition researchers who are both advocating the use of deep architecture and learning. One goal of the proposed workshop is to bring together these two groups of researchers to review the progress in both fields and to identify promising and synergistic research directions for potential future cross-fertilization and collaboration.

Mohamed, Dahl, Hinton, Deep belief networks for phone recognition, *NIPS 2009 Workshop on Deep Learning,* 2009

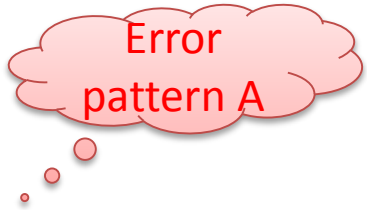
Yu, Deng, Wang, Learning in the Deep-Structured Conditional Random Fields, *NIPS 2009 Workshop on Deep Learning,* 2009

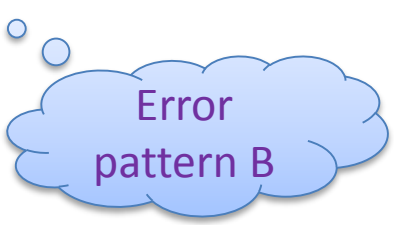
…, …, …

A key discovery at MSR, 2009-2010

Error rates for the TIMIT phone recognition task:

Hidden Dynamic/Trajectory Model	**24.8%**
Deep Neural Network	23.4%





Error-pattern-A

IS VERY DIFFERENT FROM

Error-pattern-B

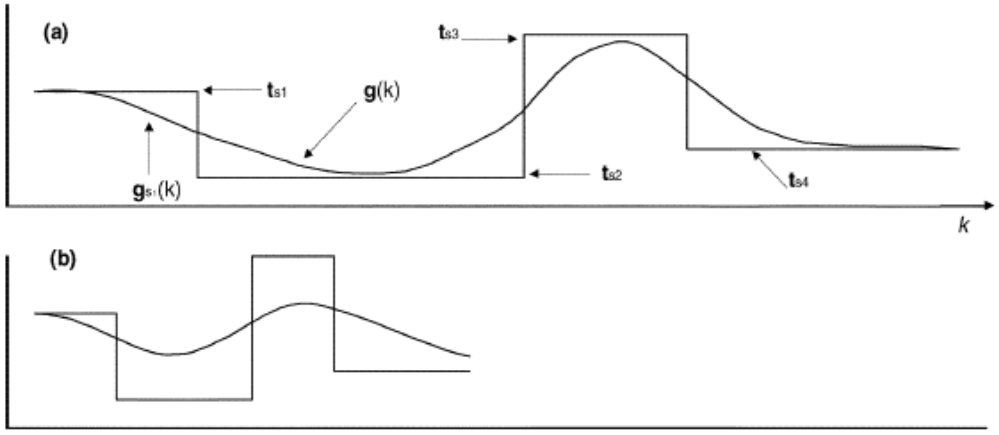


Fig. 1. Illustrations of the various VTR quantities in model Stage-I in an utterance with four phone segments. (a) and (b) are for the same four VTR targets and their filtered results, but the durations of the four segments are shorter in (b) than in (a).

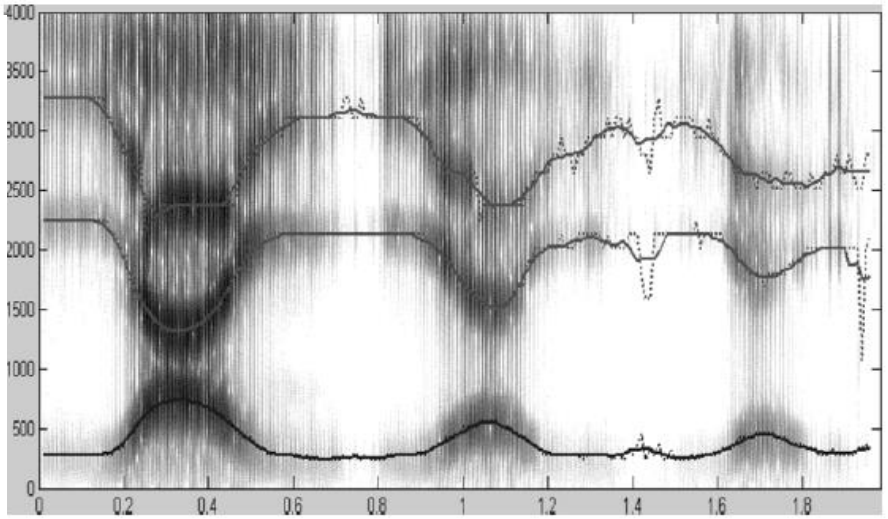


Fig. 2. Spectrogram of three renditions of /iy aa iy/ by one author, with an increasingly higher speaking rate and increasingly lower speaking efforts. The horizontal label is time, and the vertical one is frequency.

TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRECT) WITHIN EACH OF FOUR BROAD PHONE CLASSES

	Sonorants	Stops	Fricatives	Closures
Occurrences	3814	889	1252	1578
HMM	64.05	72.10	75.64	88.72
HTM	72.42	76.27	75.74	90.94

-- DNN made many new errors
on short, undershoot vowels
-- 11 frames contain too much “noise”

Expanding DNN at Industry Scale

- **Scale DNN's success to large speech tasks (2010-2011)**
  - Grew output neurons from context-independent phone states (100-200) to context-dependent ones (1k-30k) → CD-DNN-HMM for Bing Voice Search and then to SWBD tasks
  - Motivated initially by saving huge MSFT investment in the speech decoder software infrastructure (several choices available: **senones**, symbolic articulatory "features", etc.)
  - CD-DNN-HMM gave much higher accuracy than CI-DNN-HMM
  - Earlier NNs made use of context only as appended inputs, not coded directly as outputs

- **Engineering for large speech systems:**
  - Combined expertise in DNN (esp. with GPU implementation) **and** speech recognition
  - Collaborations among MSRR/MSRA, academic researchers

- Yu, Deng, Dahl, Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition, in *NIPS Workshop on Deep Learning*, 2010.
- Dahl, Yu, Deng, Acero, Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS, in *Proc. ICASSP,*
- Dahl, Yu, Deng, Acero, Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing (2013 IEEE SPS Best Paper Award)* , vol. 20, no. 1, pp. 30-42, January 2012.
- Seide, Li, Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", Interspeech 2011, pp. 437-440.
- Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, in *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, November 2012

What Enabled CD-DNN-HMM?

- Industry knowledge of
  - how to construct very large CD output units in DNN
  - how to make decoding of such huge networks highly efficient using HMM technology
  - how to cut corner in making practical systems
  - how to reduce run-time latency, etc.
- GPUs are optimized for fast matrix multiplications, major computation in CD-DNN **training**
- Nvidia's CUDA library for GPU computing released in 2008

DNN vs. Pre-DNN Prior-Art

- **Table:** TIMIT Phone recognition (**3 hours** of training data)

Features	Setup	Error Rates
Pre-DNN	Deep Generative Model	24.8%
DNN	5 layers x 2048	23.4%

~10% relative Improvement (2009-2010)

- **Table:** Voice Search SER (**24-48 hours** of training data)

Features	Setup	Error Rates
Pre-DNN	GMM-HMM with MPE	36.2%
DNN	5 layers x 2048	30.1%

~20% relative Improvement (2010)

- **Table:** SwitchBoard WER (**309 hours** of training data)

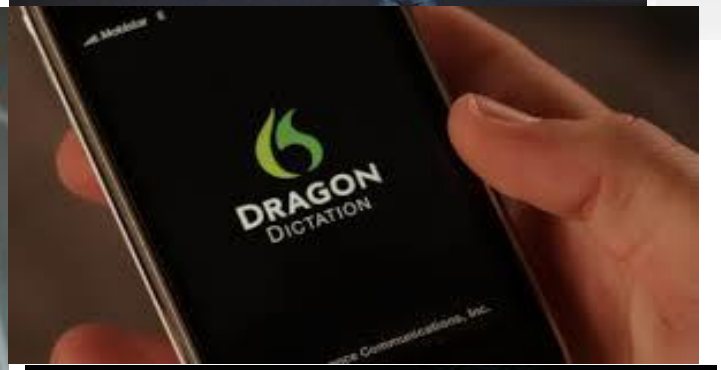
Features	Setup	Error Rates
Pre-DNN	GMM-HMM with BMMI	23.6%
DNN	7 layers x 2048	15.8%

~30% relative Improvement (2011)

For DNN, the more data, the better!

Across-the-Board Deployment of DNN in ASR Industry

(+ university labs & DARPA program)



Many, but NOT ALL, limitations of early DNNs have been overcome

- **Kluge 1:** keep the assumption of frame independence (ignore real "dynamics" to speed up decoding) but use bigger time windows

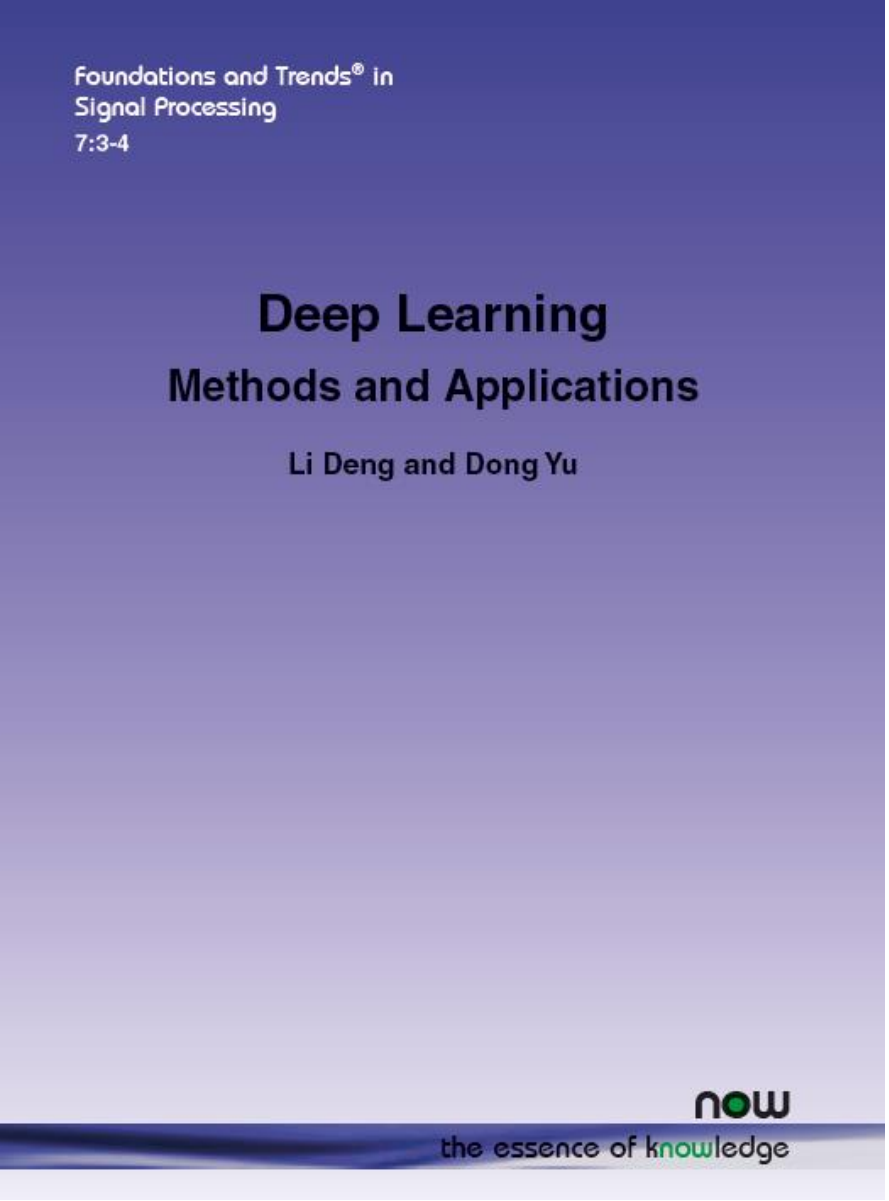
→ LSTM-RNN (single-frame input features)

- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets)

→ NOT YET: integrating deep generative model and DNN

- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.

→no need for DBN pre-training if you have big data; this is well understood now



July 2014

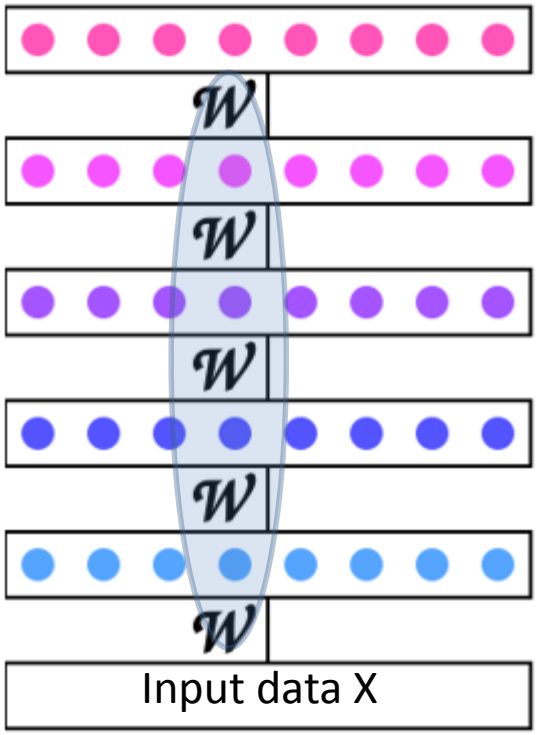
Chapter 7

Selected Applications in Speech and Audio Processing

7.1 Acoustic modeling for speech recognition

As discussed in Section 2, speech recognition is the very first successful application of deep learning methods at an industry scale. This success is a result of close academic-industrial collaboration, initiated at Microsoft Research, with the involved researchers identifying and acutely attending to the industrial need for large-scale deployment [68, 89, 109, 161, 323, 414]. It is also a result of carefully exploiting the strengths of the deep learning and the then-state-of-the-art speech recognition technology, including notably the highly efficient decoding techniques.

Innovation: Better Optimization



- **Sequence discriminative training:**

- Mohamed, Yu, Deng: "Investigation of full-sequence training of deep belief networks for speech recognition," Interspeech, 2010.
- Kingsbury, Sainath, Soltau. "Scalable minimum Bayes risk training of DNN acoustic models using distributed hessian-free optimization," Interspeech, 2012.
- Su, Li, Yu, Seide. "Error back propagation for sequence training of CD deep networks for conversational speech transcription," ICASSP, 2013.
- Vesely, Ghoshal, Burget, Povey. "Sequence-discriminative training of deep neural networks, Interspeech, 2013.

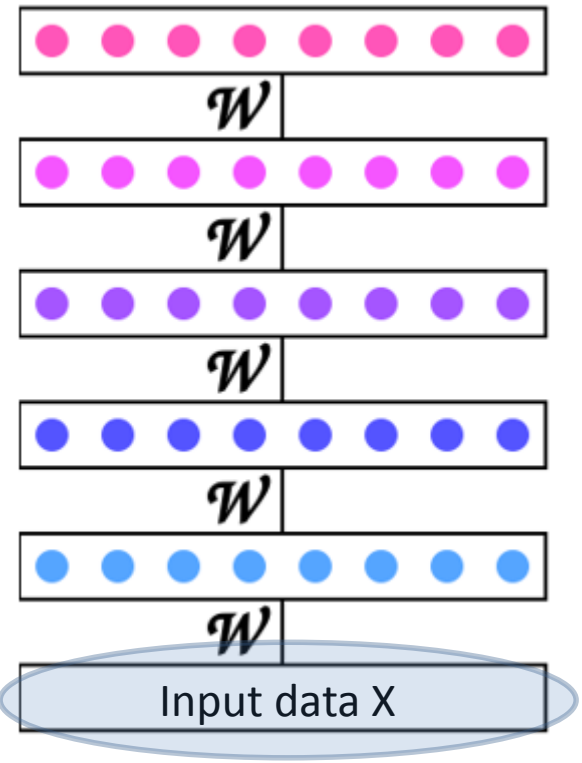
- **Distributed asynchronous SGD**

- Dean, Corrado,…Senior, Ng. "Large Scale Distributed Deep Networks," NIPS, 2012.
- Sak, Vinyals, Heigold, Senior, McDermott, Monga, Mao. "Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks," Interspeech,2014.

- **Primal-dual method**

- Chen, Deng. "A primal-dual method for training recurrent neural networks constrained by the echo-state property," ICLR, 2014.

Innovation: Towards Raw Inputs



- **Bye-Bye MFCCs (Mel-scaling & cosine transform) !**

- Deng, Seltzer, Yu, Acero, Mohamed, Hinton. "Binary coding of speech spectrograms using a deep auto-encoder," Interspeech, **2010**.
- Mohamed, Hinton, Penn. "Understanding how deep belief networks perform acoustic modeling," ICASSP, 2012.
- Li, Yu, Huang, Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM" SLT, 2012
- Deng, Li, Huang, Yao, Yu, Seide, Seltzer, Zweig, He, Williams, Gong, Acero. "Recent advances in deep learning for speech research at Microsoft," ICASSP, 2013.
- Sainath, Kingsbury, Mohamed, Ramabhadran. "Learning filter banks within a deep neural network framework," ASRU, 2013.

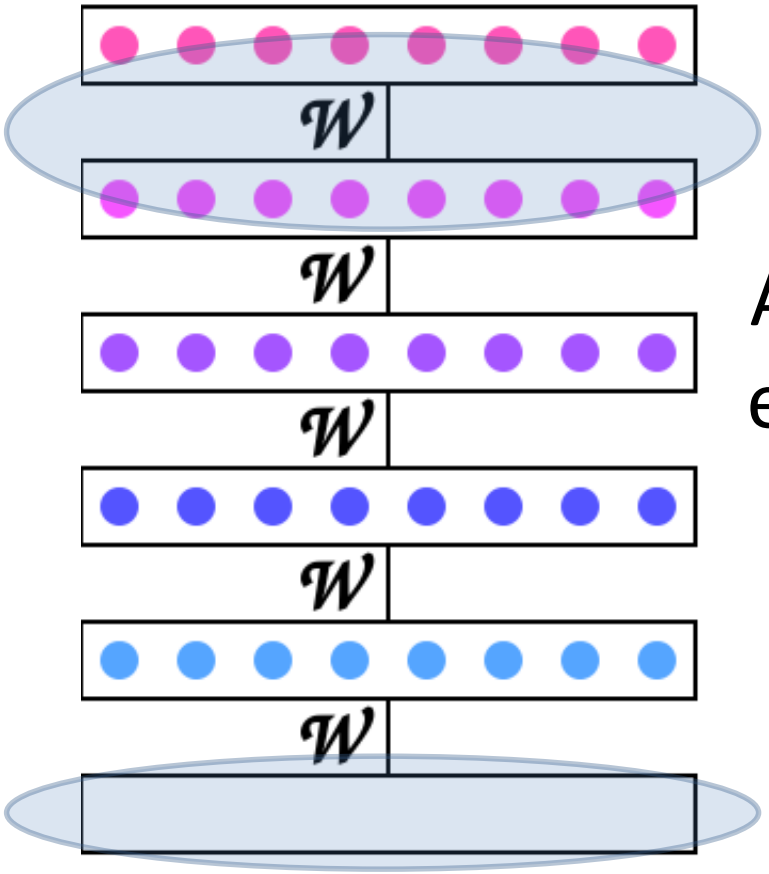
- **Bye-Bye Fourier transforms ?**

(- Sheikhzadeh, Deng, "Waveform-based speech recognition using hidden filter models," *IEEE T-SAP, 1994.)*
- Jaitly and Hinton. "Learning a better representation of speech sound waves using RBMs," ICASSP, 2011.
- Tuske, Golik, Schluter, Ney. "Acoustic modeling with deep neural networks using raw time signal for LVCSR," Interspeech, **2014**.

- **DNN as hierarchical nonlinear feature extractors:**

- Seide, Li, Chen, Yu. "Feature engineering in context-dependent deep neural networks for conversational speech transcription, ASRU, 2011.
- Yu, Seltzer, Li, Huang, Seide. "Feature learning in deep neural networks - Studies on speech recognition tasks," ICLR, 2013.
- Yan, Huo, Xu. "A scalable approach to using DNN-derived in GMM-HMM based acoustic modeling in LVCSR," Interspeech, 2013.
- Deng, Chen. "Sequence classification using high-level features extracted from deep neural networks," ICASSP, 2014.

Innovation: Transfer/Multitask Learning & Adaptation



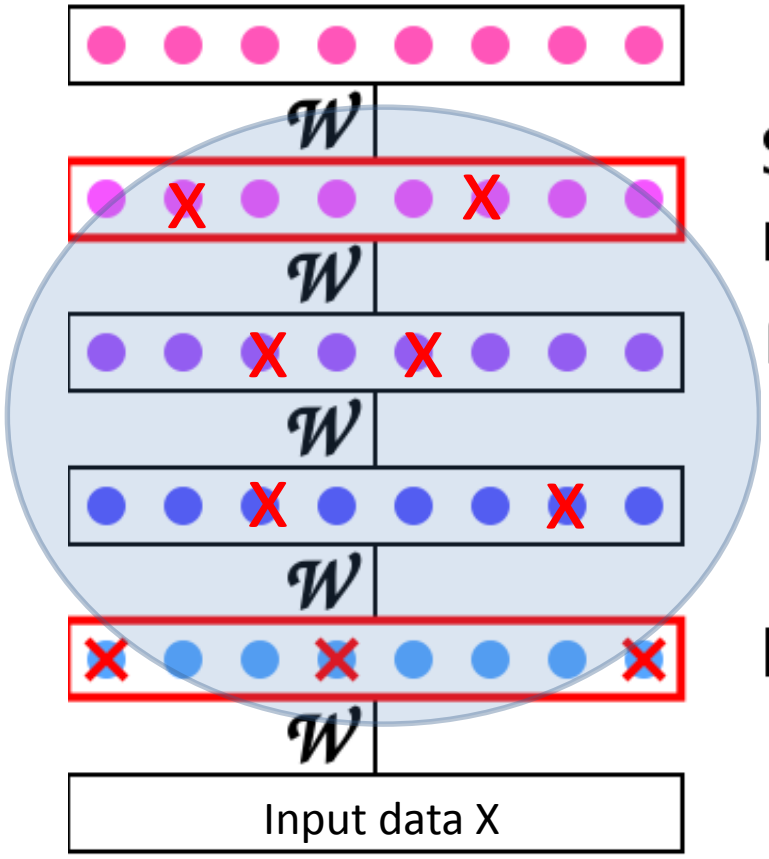
Multi-lingual acoustic modeling

Adaptation to speakers & environments

Mixed-bandwidth acoustic modeling

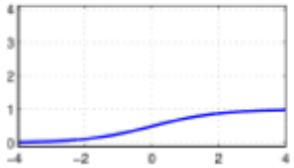
- Too many references to list & organize

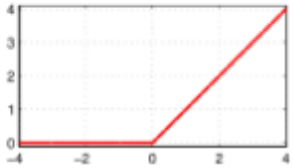
Innovation: Better regularization & **nonlinearity**



Sparsity in hidden representations

logistic ⟶ ReLU , MaxOut,



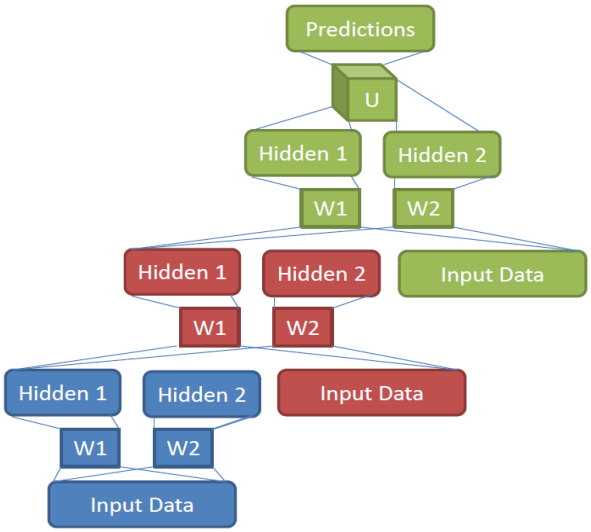


Dropout

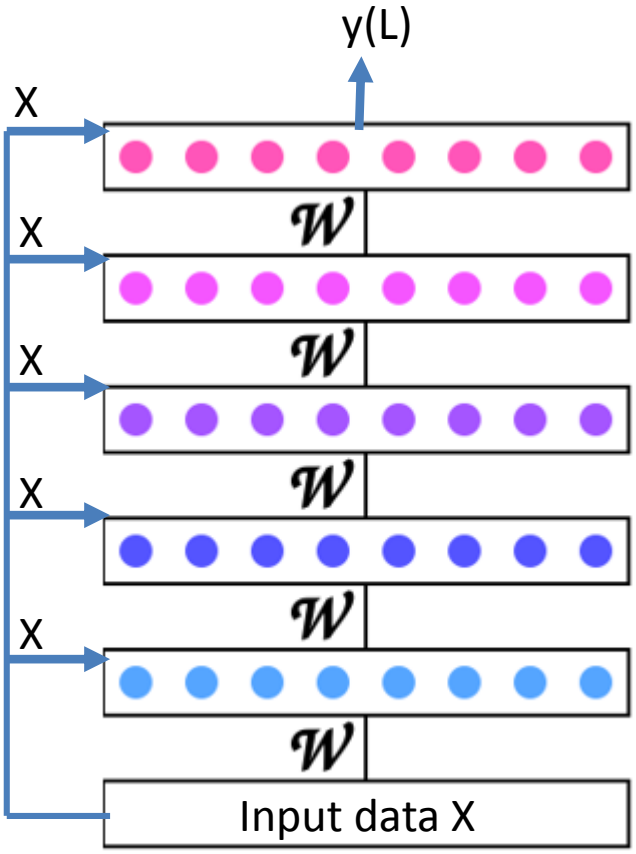
Innovation: Better architectures

Example: **Deep Stacking Network** & **Tensor DSN**

Deep Convex Network: A Scalable Architecture for Speech Pattern Classification, Interspeec
Scalable stacking and learning for building deep architectures, *ICASSP-2012*
Tensor Deep Stacking Networks, *IEEE T-PAMI,* 2013.
DEEP LEARNING --- Methods and Applications, 2014



Stacking w. hidden (or output) layer



$$y(\ell)=\mathcal{F}[y(\ell-1),\mathcal{W}(\ell-1),X]$$

$$y(\ell-1)=\mathcal{F}[y(\ell-2),\mathcal{W}(\ell-2),X]$$

$$\vdots$$

$$y(1)=\mathcal{F}[\mathcal{W}(1),X]$$

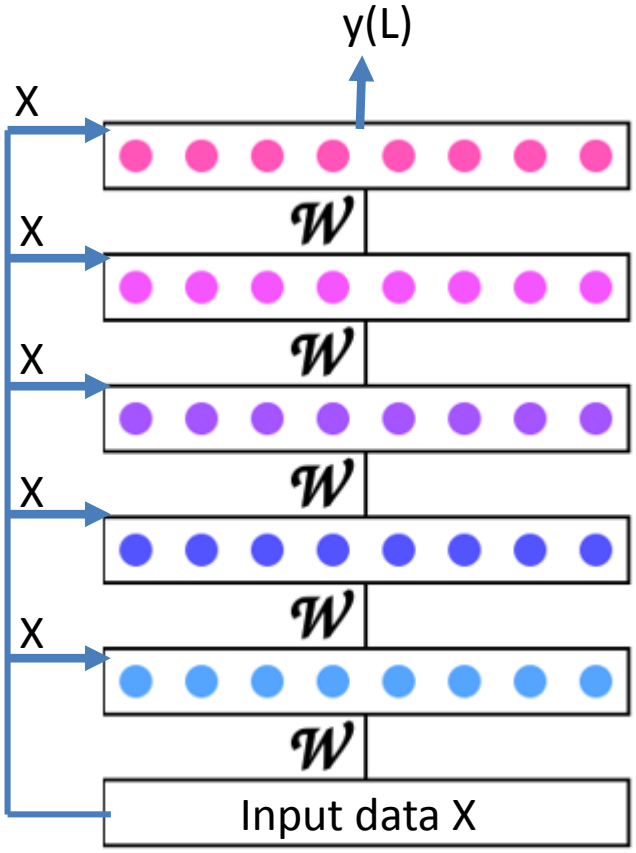
Activation function $\mathcal{F}[\bullet]$

- Pre-fixed
- Logistic or
- ReLu

Innovation: Better architectures

Yet another example: **Deep unfolding network (inspired by generative modeling)**

-- Hershey, Le Roux, and Weninger, "Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures," MERL TR2014-117 & ArXiv 2014.



$$\mathcal{F}[\bullet] = \mathbf{H}_t^k \circ \frac{(\mathbf{W}^k)^T \frac{\mathbf{M}_t}{\mathbf{W}\mathbf{H}_t^k}}{(\mathbf{W}^k)^T \mathbf{1} + \mu}.$$

$y(\ell)=\mathcal{F}[y(\ell\text{-}1), \mathcal{W}(\ell\text{-}1), X]$

$y(\ell\text{-}1)=\mathcal{F}[y(\ell\text{-}2), \mathcal{W}(\ell\text{-}2), X]$

$y(1)=\mathcal{F}[\mathcal{W}(1), X]$

- Activation function $\mathcal{F}[\bullet]$ derived from inference method in a generative model, not fixed in DSN
- The generative model embeds **knowledge/constraint** about how noisy speech is composed from clean speech & noise
  This (shallow) generative model, **non-negative matrix factorization**, unfolds in inference steps to form a DNN after untying weights
- Application: enhancement of speech & source sep. (demo)
- Example of how to integrate DNN with natural constraints in a shallow generative model
- How about deep generative model?

Advances in Inference Algms for Deep Generative Models

Kingma & Welling 2014

ICML-2014 Talk Monday June 23, 15:20

In Track F (Deep Learning II)

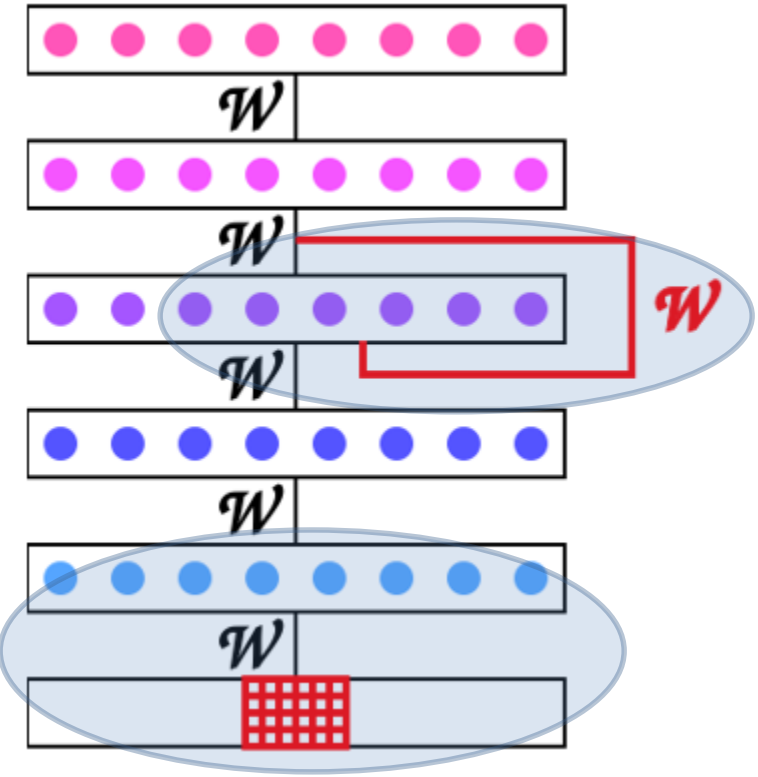
"Efficient Gradient Based Inference through Transformations between Bayes Nets and Neural Nets"

Other solutions to solve the "large variance problem" in variational inference:

-Variational Bayesian Inference with Stochastic Search [D.M. Blei, M.I. Jordan and J.W. Paisley, 2012]
-Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression [T. Salimans and A. Knowles, 2013].
-Black Box Variational Inference. [R. Ranganath, S. Gerrish and D.M. Blei. 2013]
-Stochastic Variational Inference [M.D. Hoffman, D. Blei, C. Wang and J. Paisley, 2013]
-Estimating or **propagating gradients through stochastic neurons**. [Y. Bengio, 2013].
-Neural Variational Inference and Learning in Belief Networks. [A. Mnih and K. Gregor, 2014]

Slide provided by Max Welling (ICML-2014 tutorial) w. revision

Innovation: Better architectures



- **Recurrent Nets (RNN) and Convolutional Nets (CNN) give state-of-the-art ASR results:**

- Sak, Senior, Beaufays. "LSTM Recurrent Neural Network architectures for large scale acoustic modeling," Interspeech,2014.
  → State-of-the-art results: **9.8% WER** for voice search

- Soltau, Saon, Sainath. "Joint Training of Convolutional and Non-Convolutional Neural Networks," ICASSP, 2014.
  → State-of-the-art results: **10.4% WER** for SWBD task (309hr training)

Many, but NOT ALL, limitations of early DNNs have been overcome

- **Kluge 1:** keep the assumption of frame independence (ignore real "dynamics" to speed up decoding) but use bigger time windows

→ Just fixed by LSTM-RNN (single-frame input features)

- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets)

→ NOT YET: integrating deep generative model and DNN

- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.

→no need for DBN pre-training if you have big data;

this is well understood now

Analyzing: RNN (no LSTM) vs. Generative HDM

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}; \mathbf{W}_{hh}, \mathbf{W}_{xh,}, \mathbf{x}_t)$$
$$\mathbf{y}_t = g(\mathbf{h}_t; \mathbf{W}_{hy})$$

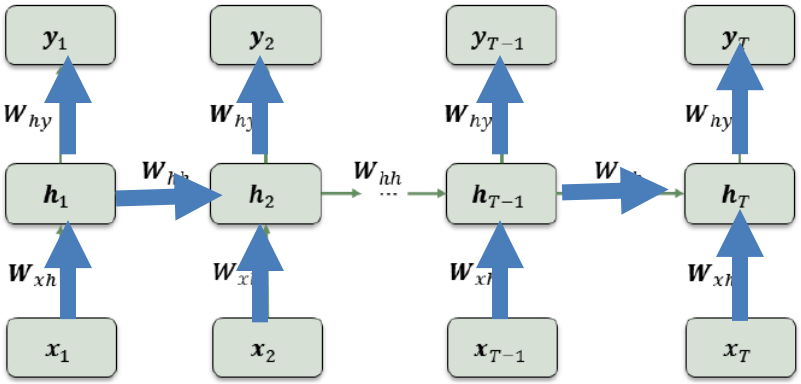
$$\mathbf{h}_t = q(\mathbf{h}_{t-1}; \mathbf{W}_{l_t}, \mathbf{t}_{l_t})$$
$$\mathbf{x}_t = r(\mathbf{h}_t, \Omega_{l_t})$$

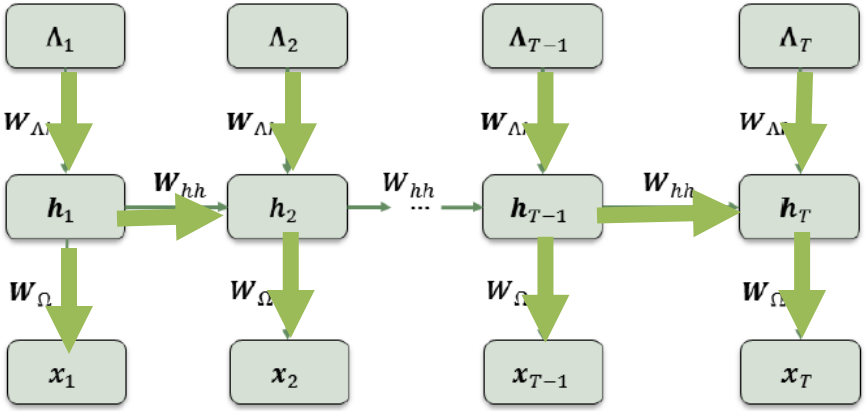
Parameterization:

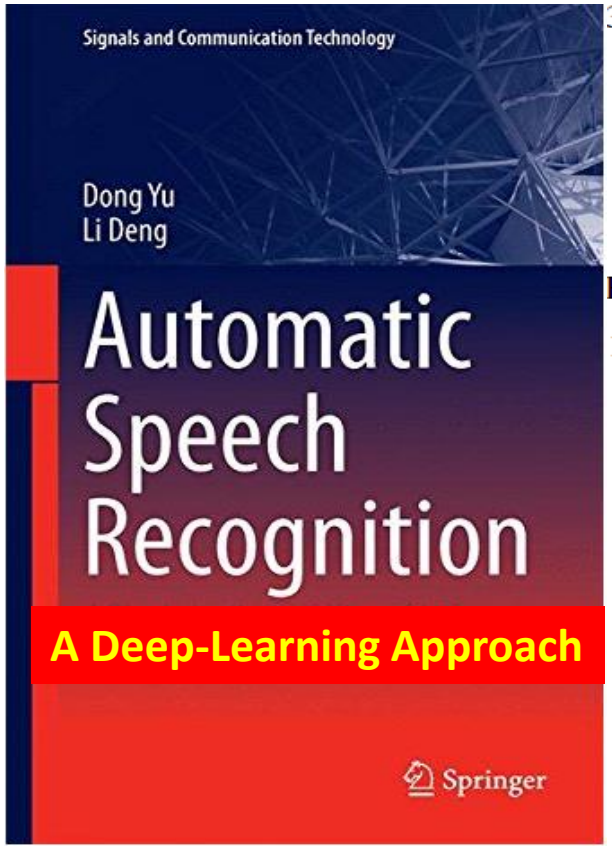
- W_{hh}, W_{hy}, W_{xh}: all unstructured regular matrices

Parameterization:

- W_{hh} =M(γ_l); sparse system matrix
- W_Ω =(Ω_l); Gaussian-mix params; MLP
- $\Lambda = \boldsymbol{t}_l$







November 2014

3.6 The HMM and Variants for Generative Speech Modeling and Recognition . . . 82
3.6.1 GMM-HMMs for speech modeling and recognition . . . 83
3.6.2 Trajectory and hidden dynamic models for speech modeling and recognition . . . 84
3.6.3 The speech recognition problem using generative models of HMM and its variants . . . 86
References . . . 89

13 Recurrent Neural Networks and Related Models . . . 473
13.1 Introduction . . . 473
13.2 State-Space Formulation of the Basic Recurrent Neural Network . . . 475
13.3 The Backpropagation-Through-Time Learning Algorithm . . . 476
13.3.1 Objective Function for Minimization . . . 477
13.3.2 Recursive Computation of Error Terms . . . 477
13.3.3 Update of RNN Weights . . . 478
13.4 A Primal-Dual Technique for Learning Recurrent Neural Networks 480
13.4.1 Difficulties in Learning RNNs . . . 480
13.4.2 Echo-State Property and Its Sufficient Condition . . . 480
13.4.3 Learning RNNs as a Constrained Optimization Problem . . . 481
13.4.4 A Primal-Dual Method for Learning RNNs . . . 482
13.5 Recurrent Neural Networks Incorporating LSTM Cells . . . 485
13.5.1 Motivations and Applications . . . 485
13.5.2 The Architecture of LSTM Cells . . . 486
13.5.3 Training the LSTM-RNN . . . 486
13.6 Analyzing Recurrent Neural Networks - A Contrastive Approach . . . 487
13.6.1 Direction of Information Flow: Top-Down or Bottom-Up . . . 487
13.6.2 The Nature of Representations: Localist or Distributed . . . 490
13.6.3 Interpretability: Inferring Latent Layers or End-to-End Learning . . . 491
13.6.4 Parameterization: Parsimonious Conditionals or Massive Weight Matrices . . . 492
13.6.5 Methods of Model Learning: Variational Inference or Gradient Descent . . . 494
13.6.6 Recognition Accuracy Comparisons . . . 495
13.7 Discussions . . . 495
References . . . 497

Outline

- **Part I:** A (brief) early history of 'deep' speech recognition & industrial impact of deep learning: **speech (& vision)**

- **Part II:** Rapid progress of deep learning: **Natural language**, **multimodality**, intelligence in the big-data world, etc.
  - from DNN to deep semantic modeling
  - **DSSM** developed at MSR for text/multimodal processing

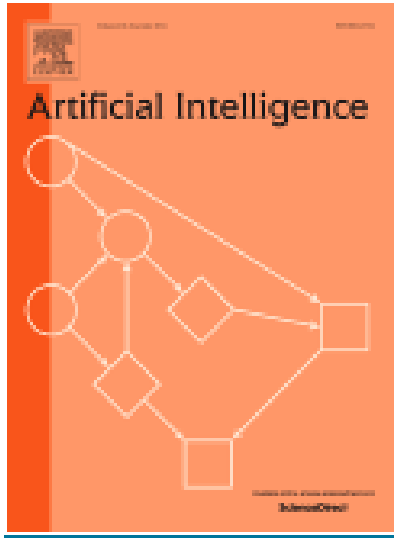
Key Concept of Embedding

(IEEE/ACM Trans Audio Speech Language Proc., Special Issue, Feb. 2015)

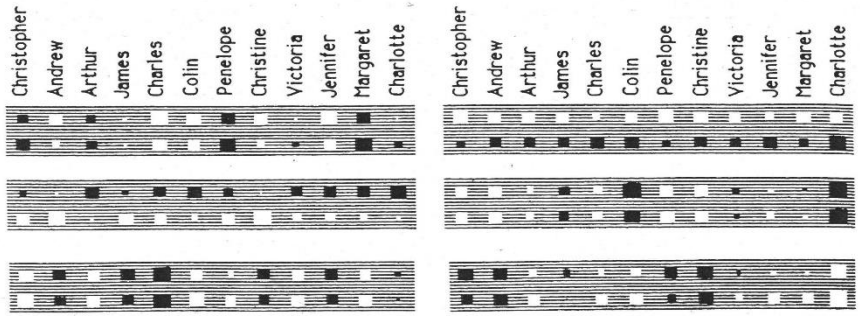
- **A linguistic or physical entity or a simple "relation"**

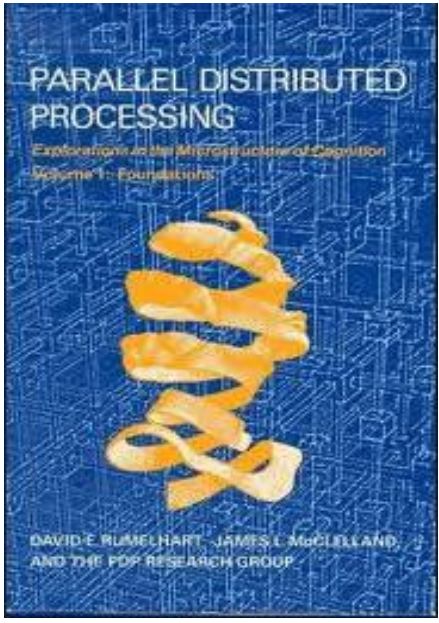
mapping via distributed representations by NN

A low-dim continuous-space vector or embedding



Special Issue, vol. 46 (1990)
Connectionist Symbol Processing
(4 articles)



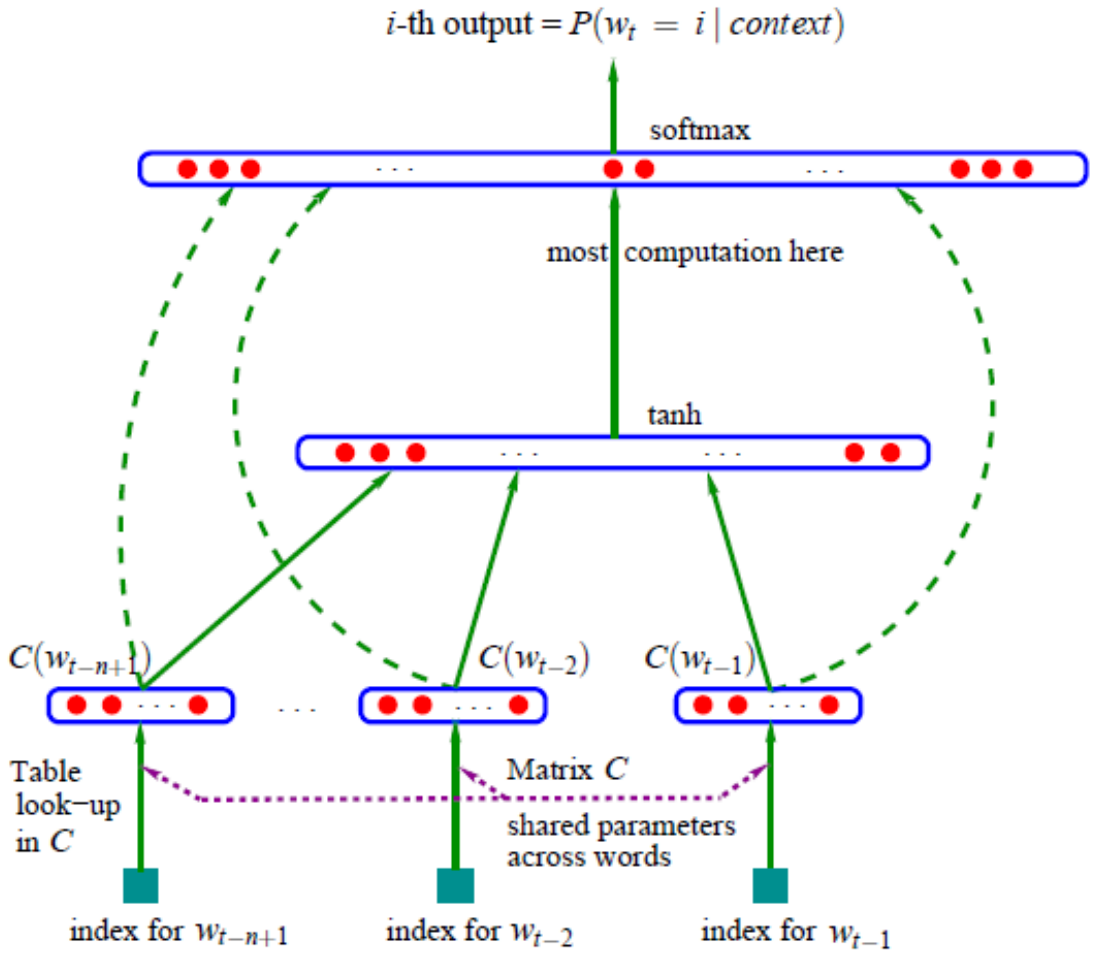


PDP book, 1986

Journal of Machine Learning Research 3 (2003) 1137–1155 also NIPS-2001

A Neural Probabilistic Language Model

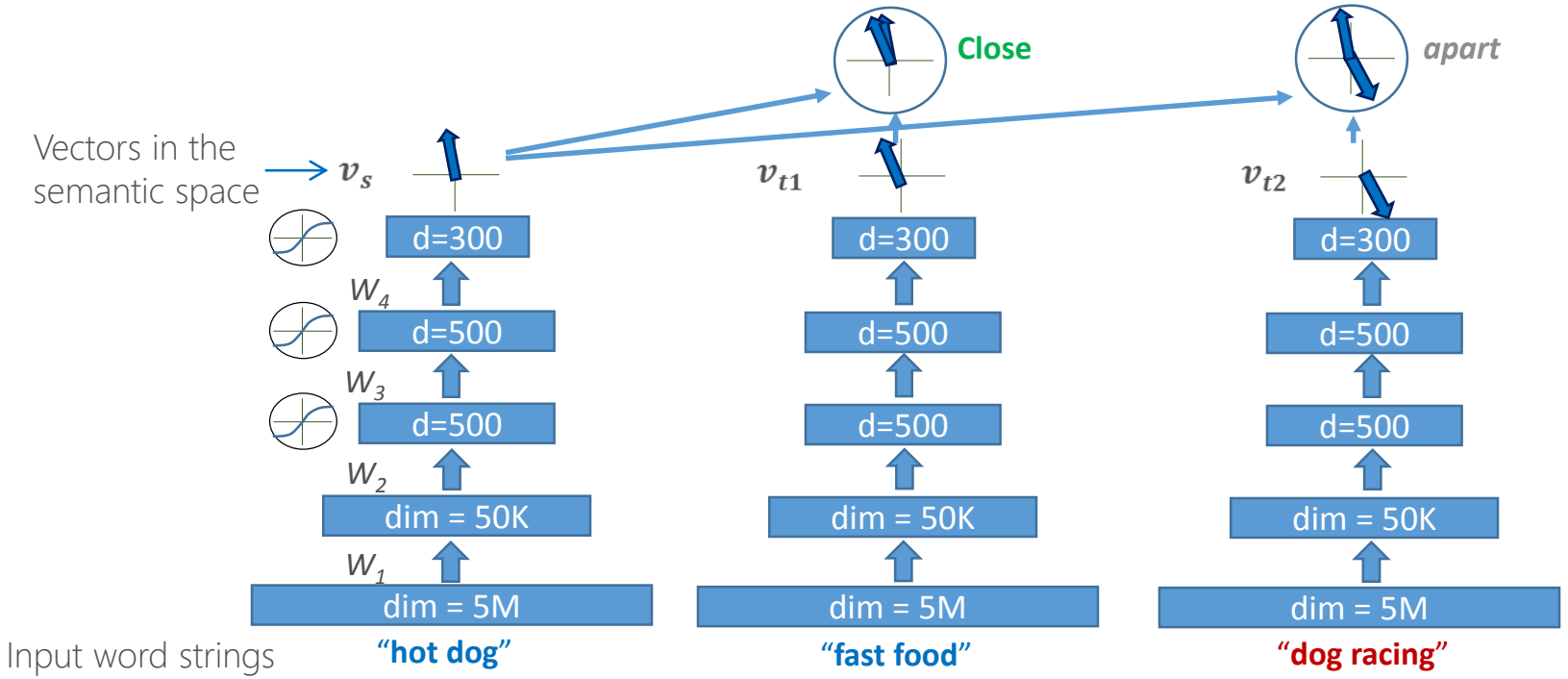
Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Réjean Ducharme DUCHARME@IRO.UMONTREAL.CA
Pascal Vincent VINCENTP@IRO.UMONTREAL.CA
Christian Jauvin JAUVINC@IRO.UMONTREAL.CA



- Feed-forward NN
- Embedding: a by-product of n-gram prediction
- BP to learn embedding embedding vectors
- Output layer huge: |Vocab|
- Overcome by a later method (Collobert/Weston, 2008)

"hot dog" & "fast food": Close in semantic space

"hot dog" & "dog racing": *apart* (although sharing same word "dog")

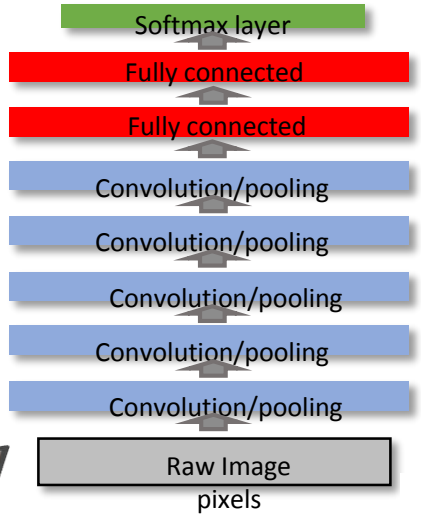


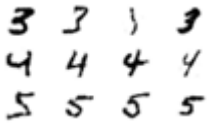
$$L(\Lambda) = -\log \prod_{(Q,D^+)} \frac{\exp[\psi R_\Lambda(Q, D^+)]}{\sum_{D' \in \boldsymbol{D}} \exp[\psi\, R_\Lambda(Q, D')]}$$

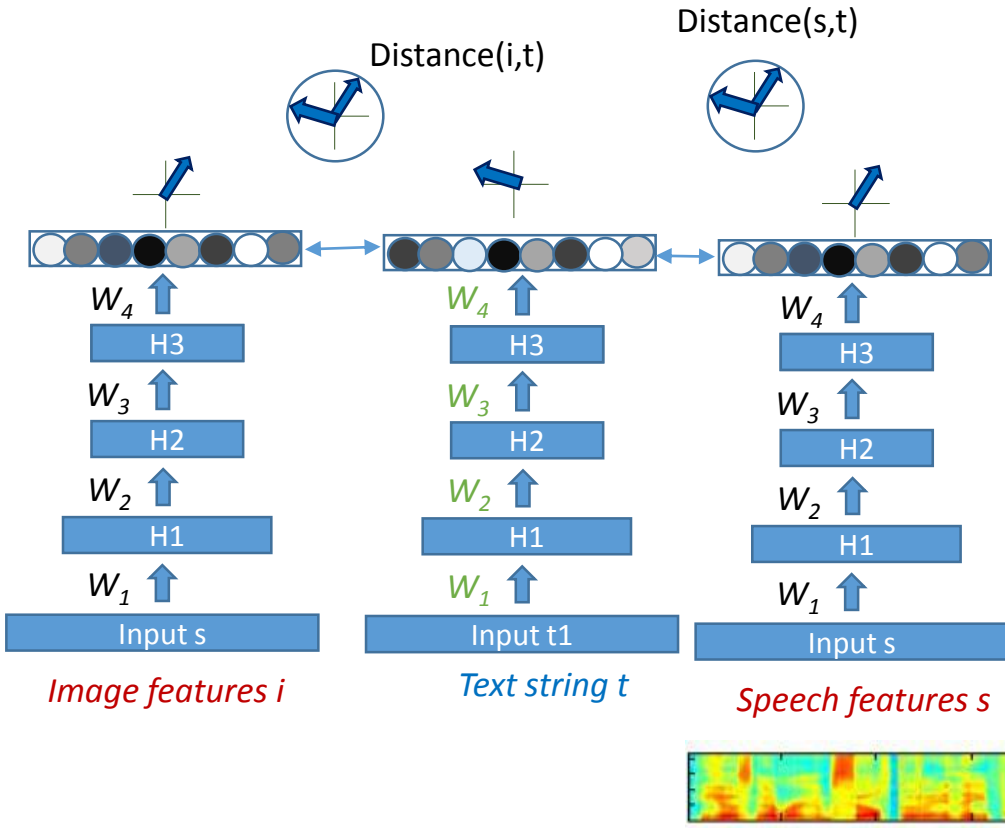
DSSM for **Multi-Modal Learning (text, image, speech)**

--- a "human-like" speech acquisition & image understanding model

- Recall DSSM for text inputs: *s*, t_1, t_2, t_3, …
- Now: replace text s by image *i* & speech *s*
- Using DNN/CNN features of image/speech







Automatic image captioning

Detector Models, Deep Neural Net Features

Computer Vision System

street
signs
light
under
on
pole
stop
red
building
sign
bus
city
traffic

Language Model

Caption Generation System

a red stop sign sitting under a traffic light on a city street
a stop sign at an intersection on a street
a stop sign with two street signs on a pole on a sidewalk
a stop sign at an intersection on a city street
…
a stop sign
a red traffic light

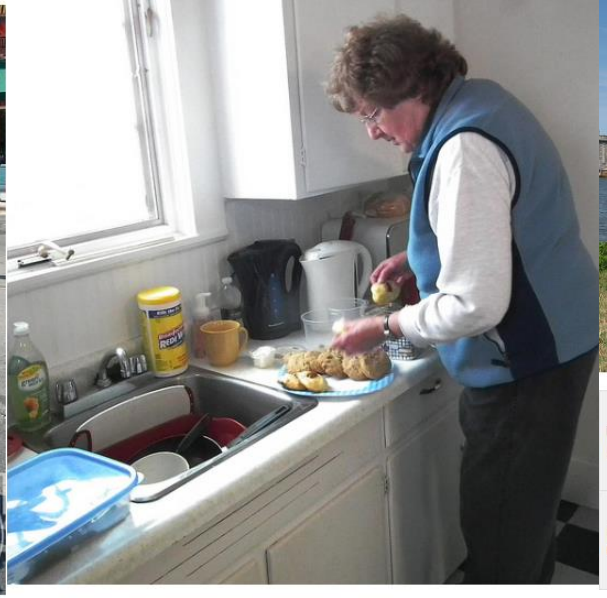
DSSM Model

Semantic Ranking System

a stop sign at an intersection on a city street

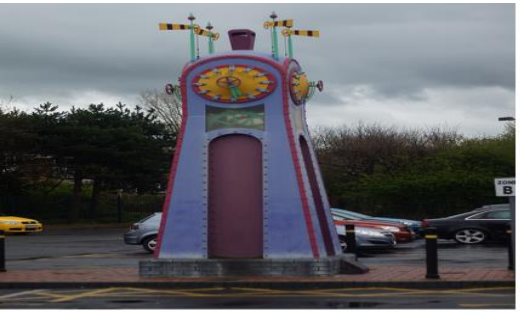
Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "**From captions to visual concepts and back**," submitted to CVPR, 2014; in arXiv

Machine-generated (but turker prefered)	a group of motorcycles parked next to a motorcycle
Human-annotated (but turker not prefered)	two girls wearing are wearing short skirts and one of them sits on a motorcycle while the other stands nearby



Machine-generated (but turker prefered)	a woman in a kitchen preparing food
Human-annotated (but turker not prefered)	woman working on counter near kitchen sink preparing a meal

Machine-generated (but turker prefered)	a bicycle is parked next to a river
Human-annotated (but turker not prefered)	a bike sits parked next to a body of water



Machine-generated (but turker prefered)	a clock tower in the middle of the street
Human-annotated (but turker not prefered)	a statue with a clock on it near a parking lot

Machine-generated (but turker prefered)	a kitchen with wooden cabinets and a sink
Human-annotated (but turker not prefered)	an ornate kitchen is designed with rustic wooden parts

Machine-generated (but turker prefered)	a man holding a tennis racquet on a tennis court
Human-annotated (but turker not prefered)	the man is on the tennis court playing a game

Evaluation: *How far are we from human?*

Training: 400K image/caption pairs as training data

Testing: 20K images, 5 annotators providing 5 captions per image

Hold 1 human as the control system

The other 4 annotations are gold reference for BLEU testing

Entry	BLEU % on 4-ref (higher the better)	Equal to or better than human annotation *
Human (control)	19.3	
Machine	21.1	23.3%

* The percentage that the judgers think the machine's output is equal to or better than the human's annotation.

Many possible applications of DSSM: Learning semantic similarity between *X* and *Y*

Tasks	*Source*	*Target*
Word semantic embedding	***context***	***word***
Web search	***search query***	***web documents***
Question answering	***pattern / mention*** **(in NL)**	***relation / entity*** **(in KB)**
Recommendation	***doc in reading***	***interesting things / other docs***
Machine translation	***sentence in language a***	***translations in language b***
Text/Image joint learning	***text / image***	***Image / text***
Ad selection	*search query*	*ad keywords*
Entity ranking	*mention (highlighted)*	*entities*
Knowledge-base construction	*entity*	*entity*
…		

Main message

Deep Learning

=

$$\mathcal{F}[\ldots\mathcal{F}[\text{Deep-Neural-Net};\text{Deep-Generative-Model}]\ldots]$$

Speech recognition	**RNN, LSTM**	**HDM (w. new formulation & paramet.)**
Speech enhancement	**DNN/DSN (feedforw'd)**	**Unfolded non-negative matrix factorization**
Language/multimodal	**DSSM**	**(Hierarchical) topic models**
Algorithms	**BackProp,…**	**BP & BP (BeliefProp, multiplicative units), stochastic variational EM (BackP in E-step)**
Neuroscience	**"Wake"**	**"Sleep"**

Two seemingly opposing approaches: with highly complementary strengths and weaknesses

Thank You

Q/A & discussions

References

- Auli, M., Galley, M., Quirk, C. and Zweig, G., 2013. Joint language and translation modeling with recurrent neural networks. In EMNLP.
- Auli, M., and Gao, J., 2014. Decoder integration and expected bleu training for recurrent neural network language models. In ACL.
- Bengio, Y., 2009. Learning deep architectures for AI. Foundumental Trends Machine Learning, vol. 2.
- Bengio, Y., Courville, A., and Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE Trans. PAMI, vol. 38, pp. 1798-1828.
- Bengio, Y., Ducharme, R., and Vincent, P., 2000. A Neural Probabilistic Language Model, in NIPS.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P., 2011. Natural language processing (almost) from scratch. in JMLR, vol. 12.
- Dahl, G., Yu, D., Deng, L., and Acero, 2012. A. Context-dependent, pre-trained deep neural networks for large vocabulary speech recognition, IEEE Trans. Audio, Speech, & Language Proc., Vol. 20 (1), pp. 30-42.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., and Harshman, R. 1990. Indexing by latent semantic analysis. J. American Society for Information Science, 41(6): 391-407
- Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G., 2010. Binary Coding of Speech Spectrograms Using a Deep Auto-encoder, in Interspeech.
- Deng, L., Tur, G, He, X, and Hakkani-Tur, D. 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding, Proc. IEEE Workshop on Spoken Language Technologies.
- Deng, L., Yu, D. and Acero, A. 2006. Structured speech modeling, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504.
- Deng, L., Yu, D., and Platt, J. 2012. Scalable stacking and learning for building deep architectures, Proc. ICASSP.
- Deoras, A., and Sarikaya, R., 2013. Deep belief network based semantic taggers for spoken language understanding, in INTERSPEECH.
- Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J., 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation, ACL.
- Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T., 2013. DeViSE: A Deep Visual-Semantic Embedding Model, Proc. NIPS.
- Gao, J., He, X., Yih, W-t., and Deng, L. 2014a. Learning continuous phrase representations for translation modeling. In ACL.
- Gao, J., He, X., and Nie, J-Y. 2010. Clickthrough-based translation models for web search: from word models to phrase models. In CIKM.
- Gao, J., Pantel, P., Gamon, M., He, X., and Deng, L. 2014. Modeling interestingness with deep neural networks. In EMNLP
- Gao, J., Toutanova, K., Yih., W-T. 2011. Clickthrough-based latent semantic models for web search. In SIGIR.
- Gao, J., Yuan, W., Li, X., Deng, K., and Nie, J-Y. 2009. Smoothing clickthrough data for web search ranking. In SIGIR.
- Gao, J., and He, X. 2013. Training MRF-based translation models using gradient ascent. In NAACL-HLT.
- Graves, A., Jaitly, N., and Mohamed, A., 2013a. Hybrid speech recognition with deep bidirectional LSTM, Proc. ASRU.
- Graves, A., Mohamed, A., and Hinton, G., 2013. Speech recognition with deep recurrent neural networks, Proc. ICASSP.

References

- He, X. and Deng, L., 2013. Speech-Centric Information Processing: An Optimization-Oriented Approach, in Proceedings of the IEEE.
- He, X. and Deng, L., 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models , ACL.
- He, X., Deng, L., and Chou, W., 2008. Discriminative learning in sequential pattern recognition, Sept. IEEE Sig. Proc. Mag.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97.
- Hinton, G., and Salakhutdinov, R., 2010. Discovering binary codes for documents by learning deep generative models. Topics in Cognitive Science.
- Hu, Y., Auli, M., Gao, Q., and Gao, J. 2014. Minimum translation modeling with recurrent neural networks. In EACL.
- Huang, E., Socher, R., Manning, C, and Ng, A. 2012. Improving word representations via global context and multiple word prototypes, Proc. ACL.
- Huang, P., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In CIKM.
- Hutchinson, B., Deng, L., and Yu, D., 2012. A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition, Proc. ICASSP.
- Hutchinson, B., Deng, L., and Yu, D., 2013. Tensor deep stacking networks, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, pp. 1944 - 1957.
- Kiros, R., Zemel, R., and Salakhutdinov, R. 2013. Multimodal Neural Language Models, Proc. NIPS Deep Learning Workshop.
- Koehn, P. 2009. Statistical Machine Translation. Cambridge University Press.
- Krizhevsky, A., Sutskever, I, and Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks, NIPS.
- Le, H-S, Oparin, I., Allauzen, A., Gauvain, J-L., Yvon, F., 2013. Structured output layer neural network language models for speech recognition, IEEE Transactions on Audio, Speech and Language Processing.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol. 86, pp. 2278-2324.
- Li, P., Hastie, T., and Church, K.. 2006. Very sparse random projections, in Proc. SIGKDD.
- Mesnil, G., He, X., Deng, L., and Bengio, Y., 2013. Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, in Interspeech.
- Mikolov, T. 2012. Statistical Language Models based on Neural Networks, Ph.D. thesis, Brno University of Technology.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space, Proc. ICLR.
- Mikolov, T., Kombrink,. S., Burget, L., Cernocky, J., Khudanpur, S., 2011. Extensions of Recurrent Neural Network LM. ICASSP.
- Mikolov, T., Yih, W., Zweig, G., 2013. Linguistic Regularities in Continuous Space Word Representations. In NAACL-HLT.
- Mohamed, A., Yu, D., and Deng, L. 2010. Investigation of full-sequence training of deep belief networks for speech recognition, Proc. Interspeech.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. 2011. Multimodal deep learning, Proc. ICML.

References

- Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. 2013. Convolutional neural networks for LVCSR, Proc. ICASSP.
- Salakhutdinov R., and Hinton, G., 2007 Semantic hashing. in Proc. SIGIR Workshop Information Retrieval and Applications of Graphical Models
- Sarikaya, R., Hinton, G., and Ramabhadran, B., 2011. Deep belief nets for natural language call-routing, in Proceedings of the ICASSP.
- Schwenk, H., Dchelotte, D., Gauvain, J-L., 2006. Continuous space language models for statistical machine translation, in COLING-ACL
- Seide, F., Li, G., and Yu, D. 2011. Conversational speech transcription using context-dependent deep neural networks, Proc. Interspeech
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search, in Proceedings of WWW.
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. A convolutional latent semantic model for web search. CIKM
- Socher, R., Huval, B., Manning, C., Ng, A., 2012. Semantic compositionality through recursive matrix-vector spaces. In EMNLP.
- Socher, R., Lin, C., Ng, A., and Manning, C. 2011. Learning continuous phrase representations and syntactic parsing with recursive neural networks, Proc. ICML.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng A., and Potts. C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Proc. EMNLP
- Song, X. He, X., Gao. J., and Deng, L. 2014. Learning Word Embedding Using the DSSM. MSR Tech Report.
- Song, Y., Wang, H., and He, X., 2014. Adapting Deep RankNet for Personalized Search. Proc. WSDM.
- Tur, G., Deng, L., Hakkani-Tur, D., and He, X., 2012. Towards Deeper Understanding Deep Convex Networks for Semantic Utterance Classification, in ICASSP.
- Wright, S., Kanevsky, D., Deng, L., He, X., Heigold, G., and Li, H., 2013. Optimization Algorithms and Applications for Speech and Language Processing, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11.
- Xu, P., and Sarikaya, R., 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling, in IEEE ASRU.
- Yao, K., Zweig, G., Hwang, M-Y. , Shi, Y., Yu, D., 2013. Recurrent neural networks for language understanding, submitted to Interspeech.
- Yann, D., Tur, G., Hakkani-Tur, D., Heck, L., 2014. Zero-Shot Learning and Clustering for Semantic Utterance Classification Using Deep Learning, in ICLR.
- Yih, W., Toutanova, K., Platt, J., and Meek, C. 2011. Learning discriminative projections for text similarity measures. In CoNLL.
- Yih, W., He, X., Meek, C. 2014. Semantic Parsing for Single-Relation Question Answering, in ACL.
- Zeiler, M. and Fergus, R. 2013. Visualizing and understanding convolutional networks, arXiv:1311.2901, pp. 1-11.

References

- Atlas, Homma, and Marks, “An Artificial Neural Network for Spatio-Temporal Bipolar Patterns, Application to Phoneme Classification,” NIPS 1987.
- Waibel, Hanazawa, Hinton, Shikano, Lang. “Phoneme recognition using time-delay neural networks.” IEEE Transactions on Acoustics, Speech and Signal Processing, 1989.
- Bengio. “Artificial Neural Networks and their Application to Speech/Sequence Recognition”, Ph.D. thesis, 1991.
- Robinson. “A real-time recurrent error propagation network word recognition system,” ICASSP 1992.
- Morgan, Bourlard, Renals, Cohen, Franco. "Hybrid neural network/hidden Markov model systems for continuous speech recognition,” IJPRAI, 1993.
- Deng, Hassanein, Elmasry. “Analysis of correlation structure for a neural predictive model with applications to speech recognition,” *Neural Networks*, vol. 7, No. 2, 1994.
- Schuster, Paliwal. "Bidirectional recurrent neural networks," IEEE Trans. Signal Processing, 1997.
- Hermansky, Ellis, Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000.
- Morgan, Zhu, Stolcke, Sonmez, Sivadas, Shinozaki, Ostendorf, Jain, Hermansky, Ellis, Doddington, Chen, Cretin, Bourlard, Athineos, “Pushing the envelope - aside [speech recognition],” IEEE Signal Processing Magazine, vol. 22, no. 5, 2005.
- Grezl, Karafiat, Kontar & Cernocky. “Probabilistic and bottle-neck features for LVCSR of meetings,” ICASSP, 2007.
- Digalakis, Rohlicek, Ostendorf. “ML estimation of a stochastic linear system with the EM alg & application to speech recognition,” IEEE T-SAP, 1993
- Deng, Aksmanovic, Sun, Wu, Speech recognition using HMM with polynomial regression functions as nonstationary states,” IEEE T-SAP, 1994.
- Deng, Ramsay, Sun. “Production models as a structural basis for automatic speech recognition,” Speech Communication, vol. 33, pp. 93–111, 1997.
- Bridle et al. “An investigation of segmental hidden dynamic models of speech coarticulation for speech recognition,” Final Report Workshop on Language Engineering, Johns Hopkins U, 1998.
- Picone et al. “Initial evaluation of hidden dynamic models on conversational speech,” ICASSP, 1999.
- Deng and Ma. “Spontaneous speech recognition using a statistical co-articulatory model for the vocal tract resonance dynamics,” JASA, 2000.
- Zhou, et al. “Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM,” ICASSP, 2003. Deng, Yu, Acero. “Structured speech modeling,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006.
- Deng. “Switching Dynamic System Models for Speech Articulation and Acoustics,” in Mathematical Foundations of Speech and Language Processing, vol. 138, pp. 115 - 134, Springer, 2003.
- Lee et al. “A Multimodal Variational Approach to Learning and Inference in Switching State Space Models,” ICASSP, 2004.

References

- Yu, Deng, Dahl, Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition, in *NIPS Workshop on Deep Learning*, 2010
- Dahl, Yu, Deng, Acero, Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS, *Proc. ICASSP, 2011,*
- Dahl, Yu, Deng, Acero, Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing (2013 IEEE SPS Best Paper Award)* , vol. 20, no. 1, pp. 30-42, January 2012.
- Seide, Li, Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", Interspeech 2011, pp. 437-440.
- Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, in *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, November 2012