Microsoft Research

Deep Speech Recognition

New-Generation Models & Methodology for Advancing Speech Technology and Information Processing

Li Deng

Microsoft Research, Redmond, USA

IEEE ChinaSIP Summer School, July 6, 2013

(including joint work with colleagues at MSR, U of Toronto, etc.)

Outline

PART I: Basics of Deep Learning (DL)

--- including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition

PART II: Deeper Substance of DL

--- including connections to other ML paradigms, examples of incorporating speech knowledge in DL architecture, and recent experiments in speech recognition

Deep Learning (DL) Basics

1. **Deep Learning (aka Deep Structured Learning, Hierarchical Learning):** a class of machine learning techniques, where many layers of information processing stages in hierarchical architectures are exploited for unsupervised feature learning and for pattern analysis/classification.
2. **Deep belief nets (DBN):** probabilistic generative models composed of multiple layers of stochastic, hidden variables. The top two layers have undirected, symmetric connections between them. The lower layers receive top-down, directed connections from the layer above. (key: stacked RBMs; Hinton: **Science, 2006**)
3. **Boltzmann machine (BM):** a network of symmetrically connected, neuron-like units that make stochastic decisions about whether to be on or off.
4. **Restricted Boltzmann machine (RBM):** a special BM consisting of a layer of visible units and a layer of hidden units with no visible-visible or hidden-hidden connections. (Key: contrastive divergence learning)
5. **Deep neural nets (DNN, or "DBN" before Nov 2012):** multilayer perceptrons with many hidden layers, whose weights are often initialized (pre-trained) using stacked RBMs or DBN (DBN-DNN) or discriminative pre-training.
6. **Deep auto-encoder:** a DNN whose output is the data input itself, often pre-trained with DBN (Deng/Hinton, interspeech 2010; Hinton, Science 2006)
7. **Distributed representation**: a representation of the observed data in such a way that they are modeled as being generated by the interactions of many hidden factors. A particular factor learned from configurations of other factors can often generalize well. Distributed representations form the basis of deep learning.

Microsoft Research

Distributed Representation

- A representation of the observed data in such a way that they are modeled as being generated by the interactions of many hidden factors. A particular factor learned from configurations of other factors can often generalize well. Distributed representations form the basis of deep learning.
- In contrast to the "atomic" or "localist" representations employed in traditional cognitive science (and in GMM-HMM speech recognition systems), a distributed representation is one in which "each entity is represented by a pattern of activity distributed over many computing element, and each computing element is involved in representing many different entities". (Hinton, 1984)
- **In GMM-HMM, each sound is associated with its own set of parameters. Not so for DNN-HMM.**

Microsoft Research

More on "Deep Learning"

- ***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." see Wikipedia on "Deep Learning" as of this writing in February 2013; see http://en.wikipedia.org/wiki/Deep_learning.

- ***Definition 4****:* "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

Microsoft Research

Data Science 101 (June 2013)

Deep Learning – A Term To Know

Deep Learning is a new term that is starting to appear in the data science/machine learning news.

- Communications of the ACM just published a story on the topic, Deep Learning Comes of Age.
- Deep Learning was named as one of the Top 10 Breakthrough Technologies of 2013 by MIT Technology Review.
- Jeremy Howard, Chief Scientist at Kaggle declared Deep Learning – The Biggest Data Science Breakthrough of the Decade.
- The New York Times published Scientists See Promise in Deep-Learning Programs

What is Deep Learning?

According to **DeepLearning.net**, the definition goes like this:

"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."

Wikipedia provides the following definition:

"Deep learning is set of algorithms in machine learning that attempt to learn layered models of inputs, commonly neural networks. The layers in such 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."

Deep Learning is sometimes referred to as deep neural networks since much of deep learning focuses on artificial neural networks. Artificial neural networks are a technique in computer science modelled after the connections (synapses) of neurons in the brain. Artificial neural networks, sometimes just called neural nets, have been around for about 50 years, but advances in computer processing power and storage are finally allowing neural nets to improve solutions for complex problems such as speech recognition, computer vision, and Natural Language Processing (NLP).

Useful Sites on Deep Learning

- http://www.cs.toronto.edu/~hinton/
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Recommended_Readings
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial (Andrew Ng's group)
- http://deeplearning.net/reading-list/ (Bengio's group)
- http://deeplearning.net/tutorial/
- **http://deeplearning.net/deep-learning-research-groups-and-labs/**

- Google+ Deep Learning community

http://deeplearning.net/deep-learning-research-groups-and-labs/

Deep Learning Research Groups

Some labs and research groups that are actively working on deep learning:

University of Toronto - **Machine Learning Group** (Geoff Hinton, Rich Zemel, Ruslan Salakhutdinov, Brendan Frey, Radford Neal)
Université de Montréal - **Lisa Lab** (Yoshua Bengio, Pascal Vincent, Aaron Courville, Roland Memisevic)
New York University – **Yann Lecun**'s and **Rob Fergus**' group
Stanford University – **Andrew Ng**'s group
UBC – **Nando de Freitas**'s group
Google Research – Jeff Dean, Samy Bengio, Jason Weston, Marc'Aurelio Ranzato, Dumitru Erhan, Quoc Le et al
Microsoft Research – **Li Deng** et al
SUPSI – **IDSIA** (Schmidhuber's group)
UC Berkeley – **Bruno Olshausen**'s group
University of Washington – **Pedro Domingos**' group
IDIAP Research Institute - **Ronan Collobert**'s group
University of California Merced – **Miguel A. Carreira-Perpinan**'s group
University of Helsinki - **Aapo Hyvärinen**'s Neuroinformatics group
Université de Sherbrooke – **Hugo Larochelle**'s group
University of Guelph – **Graham Taylor**'s group
University of Michigan – **Honglak Lee**'s group
Technical University of Berlin – **Klaus-Robert Muller**'s group
Baidu – **Kai Yu**'s group
Aalto University – **Juha Karhunen**'s group
U. Amsterdam – **Max Welling**'s group
U. California Irvine – **Pierre Baldi**'s group
Ghent University – **Benjamin Shrauwen**'s group
University of Tennessee – **Itamar Arel**'s group
IBM Research – **Brian Kingsbury** et al
University of Bonn – **Sven Behnke's** group
Gatsby Unit @ University College London – Maneesh Sahani, Yee-Whye Teh, Peter Dayan

Last modified on April 10, 2013, at 1:27 pm by Caglar Gulcehre

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child?

Additive Manufacturing

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts.

Baxter: The Blue-Collar Robot

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people.

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss.

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket.

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible.

Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave – and even help us understand the spread of diseases.

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical.

ICASSP 2013
Vancouver Convention & Exhibition Centre
May 26 - 31, 2013 • Vancouver, Canada
IEEE
IEEE Signal Processing Society

Plenary Keynote (9:50-10:40am, May 28)

Recent Developments in Deep Neural Networks

Geoffrey E. Hinton

UNIVERSITY OF TORONTO

Google

Host: Li Deng

Geoff Hinton

The New York Times

Scientists See Promise in Deep-Learning Programs
John Markoff
November 23, 2012

Rich Rashid in **Tianjin**, October, 25, 2012

The New York Times

Learning Curve: No Longer Just A Human Trait

By JOHN MARKOFF

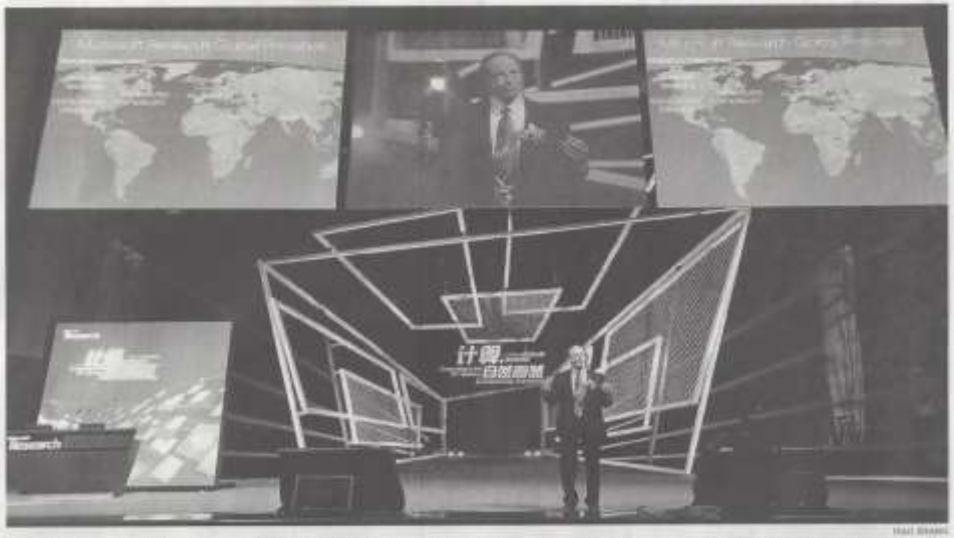
Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

The advances have led to widespread enthusiasm among researchers who design software to perform human activities like seeing, listening and thinking. They offer the promise of machines that converse with humans and perform tasks like driving cars and working in factories, raising the specter of automated robots that could replace human workers.

The technology, called deep learning, has already been put to use in services like Apple's Siri virtual personal assistant, which is based on Nuance Communications' speech recognition service, and in Google's Street View, which uses machine vision to identify specific addresses.

But what is new in recent months is the growing speed and accuracy of deep-learning programs, often called artificial neural networks or just "neural nets" for their resemblance to the neural connections in the brain.

"There has been a number of stunning new results with deep-learning methods," said Yann LeCun, a computer scientist at New York University who did



A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

Scientists See Promise in Deep-Learning Programs

From Page A1

seeing in the accuracy of these systems is very rare indeed."

Artificial intelligence researchers are acutely aware of the dangers of being overly optimistic. Their field has long been plagued by outbursts of misplaced enthusiasm followed by equally striking declines.

In the 1960s, some computer scientists believed that a workable artificial intelligence system was just 10 years away. In the 1980s, a wave of commercial start-ups collapsed, leading to what some people called the "A.I. winter."

But recent achievements have impressed a wide spectrum of computer experts. In October, for example, a team of graduate students studying with the University of Toronto computer scientist Geoffrey E. Hinton won the top prize in a contest sponsored by Merck to design software to help find molecules that might lead to new drugs.

From a data set describing the chemical structure of 15 different molecules, they used deep-learning software to determine which molecule was most likely to be an effective drug agent.

The achievement was particularly impressive because the team decided to enter the contest at the last minute and designed its software with no specific knowledge about how the molecules bind to their targets. The students were also working with a relatively small set of data; neural nets typically perform well only with very large ones.

"This is a really breathtaking result because it is the first time that deep learning won, and more significantly it won on a data set that it wouldn't have been ex-

that organizes data science competitions, including the Merck contest.

Advances in pattern recognition hold implications not just for drug development but for an array of applications, including marketing and law enforcement. With greater accuracy, for example, marketers can comb large databases of consumer behavior to get more precise information on buying habits. And improvements in facial recognition are likely to make surveillance technology cheaper and more commonplace.

Artificial neural networks, an idea going back to the 1950s, seek to mimic the way the brain absorbs information and learns from it. In recent decades, Dr. Hinton, 64 (a great-great-grandson of the 19th-century mathematician George Boole, whose work in logic is the foundation for modern digital computers), has pioneered powerful new techniques for helping the artificial networks recognize patterns.

Modern artificial neural networks are composed of an array of software components, divided into inputs, hidden layers and outputs. The arrays can be "trained" by repeated exposures to recognize patterns like images or sounds.

These techniques, aided by the growing speed and power of modern computers, have led to rapid improvements in speech recognition, drug discovery and computer vision.

Deep-learning systems have recently outperformed humans in certain limited recognition tests.

Last year, for example, a program created by scientists at the Swiss A.I. Lab at the University of Lugano won a pattern recogni-

A student team led by the computer scientist Geoffrey E. Hinton used deep-learning technology to design software.

An advance in a technology that can best human brains.

the images in a set of 50,000; the top score in a group of 32 human participants was 99.22 percent, and the average for the humans was 98.84 percent.

This summer, Jeff Dean, a Google technical fellow, and Andrew Y. Ng, a Stanford computer scientist, programmed a cluster of 16,000 computers to train itself to automatically recognize images in a library of 14 million pictures of 20,000 different objects. Although the accuracy rate was low — 15.8 percent — the system did 70 percent better than the most advanced previous one.

Deep learning was given a particularly audacious display at a

lish on a large screen above his head.

Then, in a demonstration that led to stunned applause, he paused after each sentence and the words were translated into Mandarin Chinese characters, accompanied by a simulation of his own voice in that language, which Dr. Rashid has never spoken.

The feat was made possible, in part, by deep-learning techniques that have spurred improvements in the accuracy of speech recognition.

Dr. Rashid, who oversees Microsoft's worldwide research organization, acknowledged that while his company's new speech recognition software made 30 percent fewer errors than previous models, it was "still far from perfect."

"Rather than having one word in four or five incorrect, now the error rate is one word in seven or eight," he wrote on Microsoft's Web site. Still, he added that this was "the most dramatic change in accuracy" since 1979, "and as we add more data to the training we believe that we will get even better results."

One of the most striking aspects of the research led by Dr. Hinton is that it has taken place largely without the patent restrictions and bitter infighting over intellectual property that characterize high-technology fields.

"We decided early on not to make money out of this, but just to sort of spread it to infect everybody," he said. "These companies are terribly pleased with this."

Referring to the rapid deep-learning advances made possible by greater computing power, and especially the rise of graphics processors, he added:

Keynote Speaker, Vincent Vanhoucke

ICML | Atlanta

International Conference on Machine Learning

16-21 JUNE 2013 ATLANTA

Google

Acoustic Modeling and Deep Learning

June 19th, 2013

Vincent Vanhoucke

Thanks to Vincent for the permission of using his slides & discussions/corrections of information in some slides

Neural Networks for Speech in the 90's

- Time-Delay Neural Networks
  Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. "Phoneme recognition using time-delay neural networks. " IEEE Transactions on Acoustics, Speech and Signal Processing, 37, no. 3 (1989): 328-339.

1989

- Recurrent Neural Networks
  Tony Robinson. "A real-time recurrent error propagation network word recognition system", ICASSP 1992.

1992

1993

- Hybrid Systems
  Nelson Morgan, Herve Bourlard, Steve Renals, Michael Cohen, and Horacio Franco. "Hybrid neural network/ hidden Markov model systems for continuous speech recognition." International journal of pattern recognition and artificial intelligence 7, no. 04 (1993): 899-916.

- Bidirectional Recurrent Neural Networks
  Mike Schuster, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE Transactions on Signal Processing, 45, no. 11 (1997): 2673-2681.

1997

- Hierarchical Neural Networks
  Jürgen Fritsch and Michael Finke. "ACID/HNN: Clustering hierarchies of neural networks for context-dependent connectionist acoustic modeling." ICASSP 1998.

1998

- TANDEM
  Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000.

2000

Speech Recognition

Speech Recognition + Deep Neural Networks?

Speech Recognition + Deep Neural Networks!

3 months - 10%

word error rate
relative reduction
Voice Search

Application Of Pretrained Deep Neural Networks To Large Vocabulary Speech Recognition,
Navdeep Jaitly, Patrick Nguyen, Andrew Senior, Vincent Vanhoucke, Interspeech 2012.

Similar Stories across the Industry

Microsoft

Li Deng
Frank Seide
Dong Yu

IBM

Tara Sainath
Brian Kingsbury

Google

Andrew Senior
Georg Heigold
Marc'Aurelio Ranzato

University of Toronto

Geoff Hinton
George Dahl
Abdel-rahman Mohamed

And many others...

Deep Neural Networks for Acoustic Modeling in Speech Recognition, Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, Brian Kingsbury, IEEE Signal Processing Magazine, Vol. 29, No. 6, November, 2012.

Microsoft Research

Some of Microsoft's Stories…,
Since 2009…

DL Took off in Speech Recognition from MSR

- Speech recognition: the first big (and real-world) success of deep learning
- From MSR (initial collaboration with Hinton et al., 2009-2010) and then to the entire speech industry
- Got out of “local optimum” of GMM-HMM stayed for many years
- Now used by Microsoft, Google, Apple/Nuance/IBM, Baidu, IFlyTech, etc. doing voice search in the cloud for smart phones (plus many other applications.)

Microsoft Research

Renaissance of Neural Network --- "Deep Learning," 2006

-Geoff Hinton invented Deep Belief Networks (DBN) to make neural net learning fast and effective; ***Science, 2006***

- Pre-train each layer from bottom up
- Each pair of layers is an Restricted Boltzmann Machine (RBM)
- Jointly fine-tune all layers using back-propagation

Industry Scale Deep Learning

Started at MSR, 2009

-2008 NIPS: Geoff Hinton & Li Deng reconnected

-Earlier 2009: Initial exploration of DBN/DNN at MSR (image and speech)

-Later 2009: Proof of concept by Mohamed et al.; MSR & Hinton collaborated on applying DBN-DNN to speech feature coding (on spectrogram) and speech recognition

-Dec 2009: NIPS workshop (organizers: Deng, Yu, & Hinton)

Neural Information Processing Systems Foundation

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 b
dominated by a "shallow" architecture --- hidden Markov models (HMMs). Sig
technological success has been achieved using complex and carefully engin
of HMMs. The next generation of the technology requires solutions to remain
challenges under diversified deployment environments. These challenges, no
addressed in the past, arise from the many types of variability present in the
generation process. Overcoming these challenges is likely to require "deep"
with efficient learning algorithms. For speech recognition and related sequen
recognition applications, some attempts have been made in the past to deve
computational architectures that are "deeper" than conventional HMMs, such

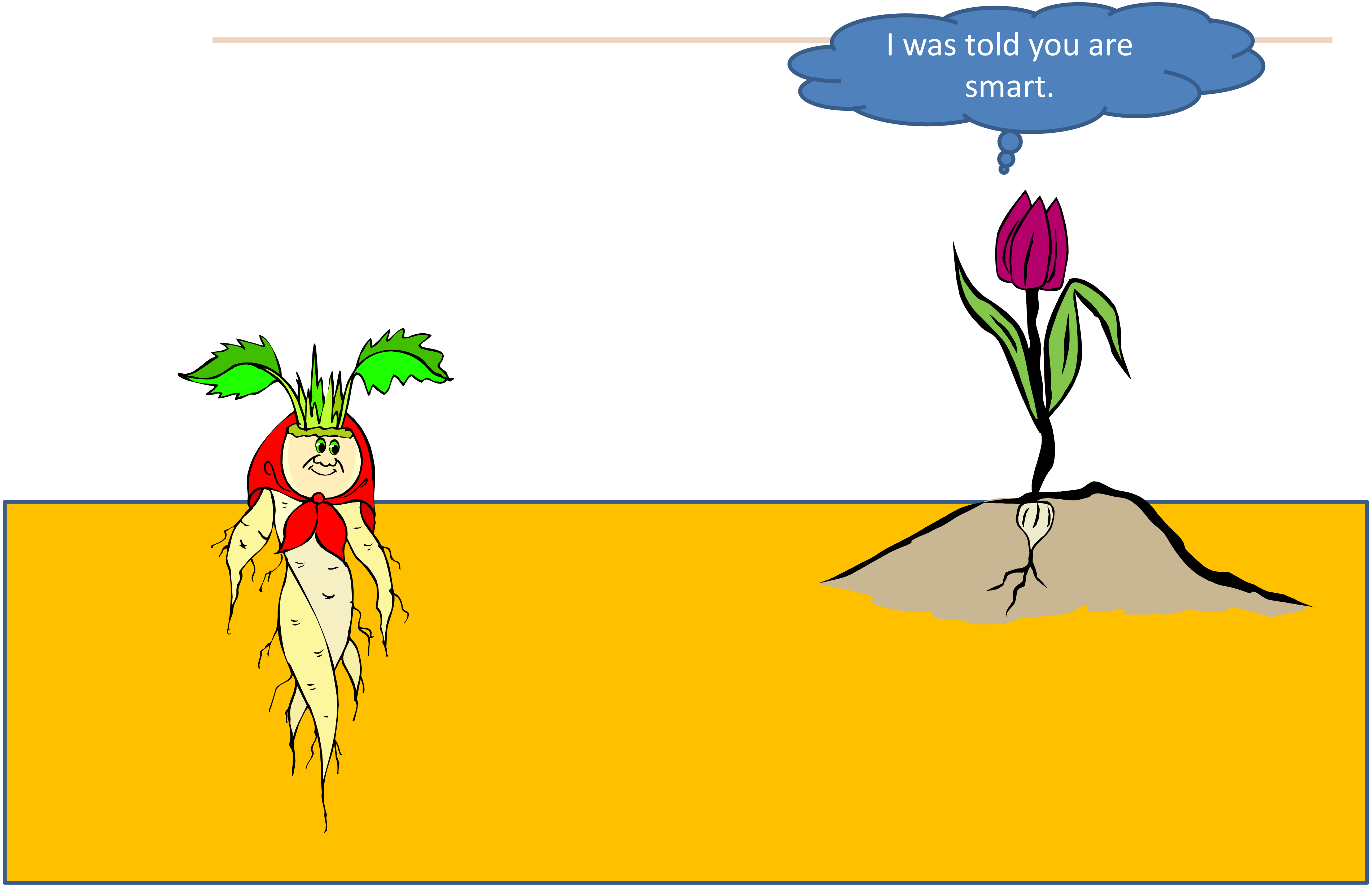
Microsoft Research

Anecdote: Speechless summary presentation of the NIPS 2009 Workshop on Speech

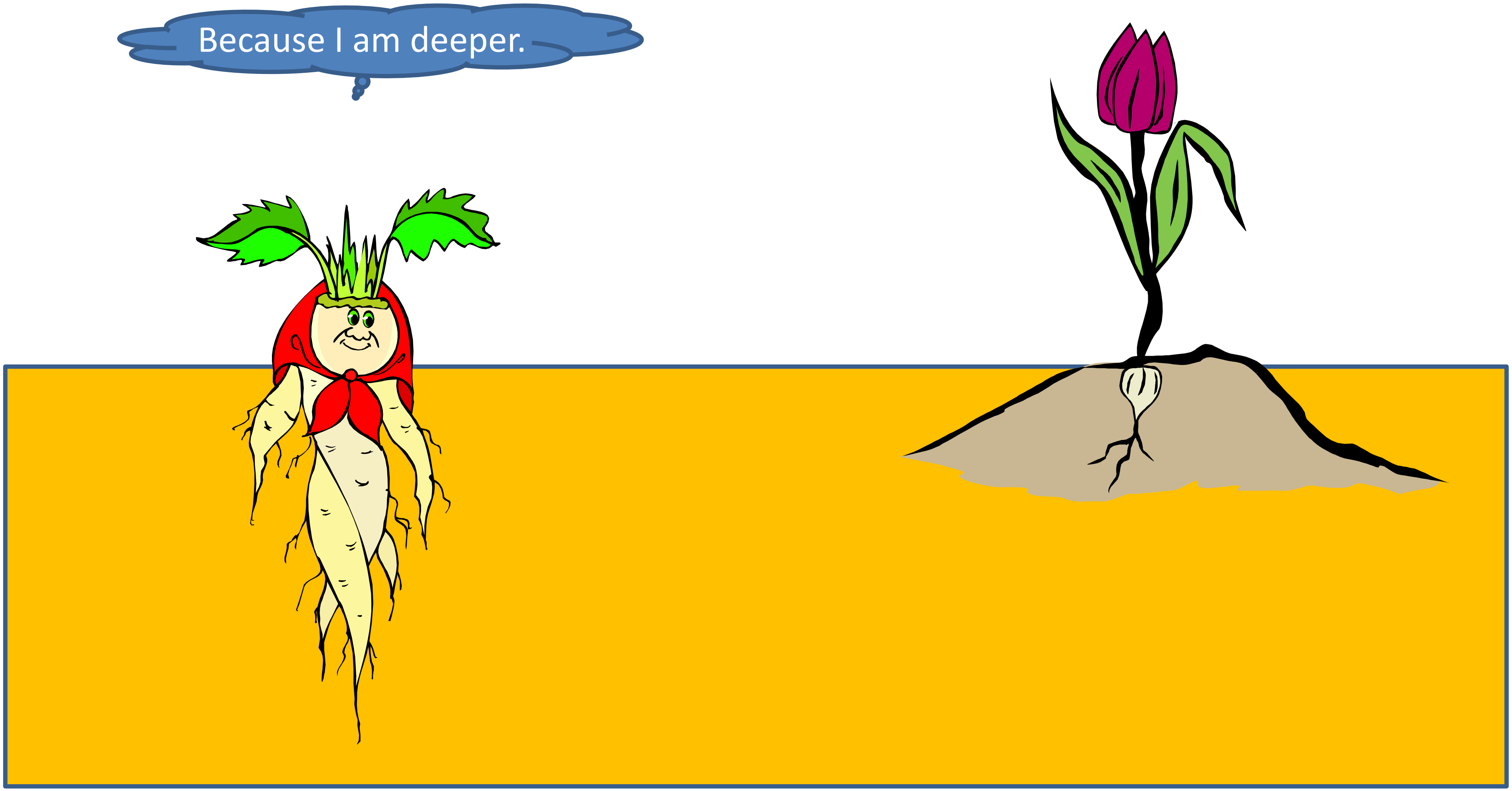
Deep Learning for Speech Recognition and Related Applications

Li Deng, *Dong Yu (Microsoft Research)*
Geoffrey Hinton (University of Toronto)

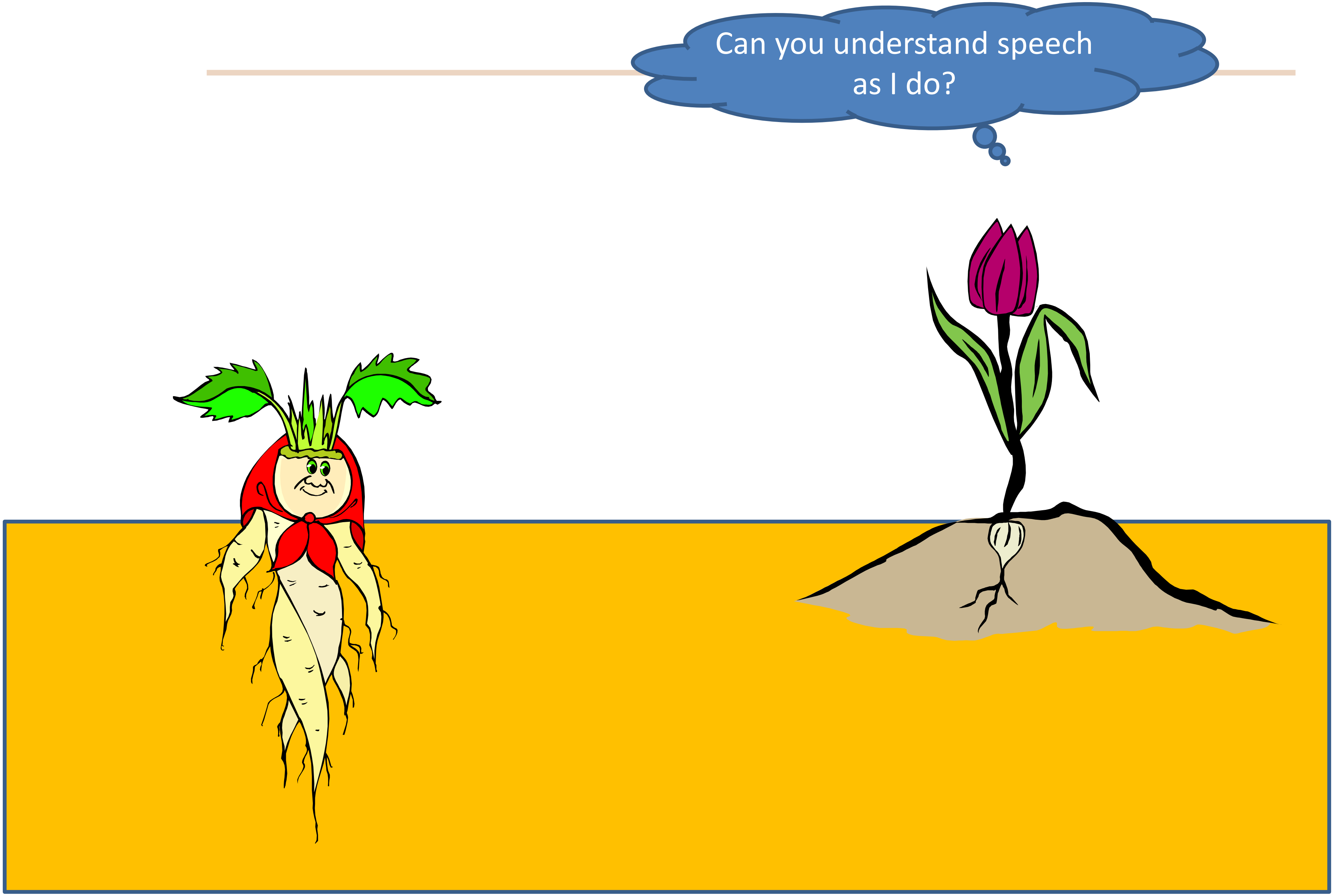
They met in year 2009...


I was told you are smart.

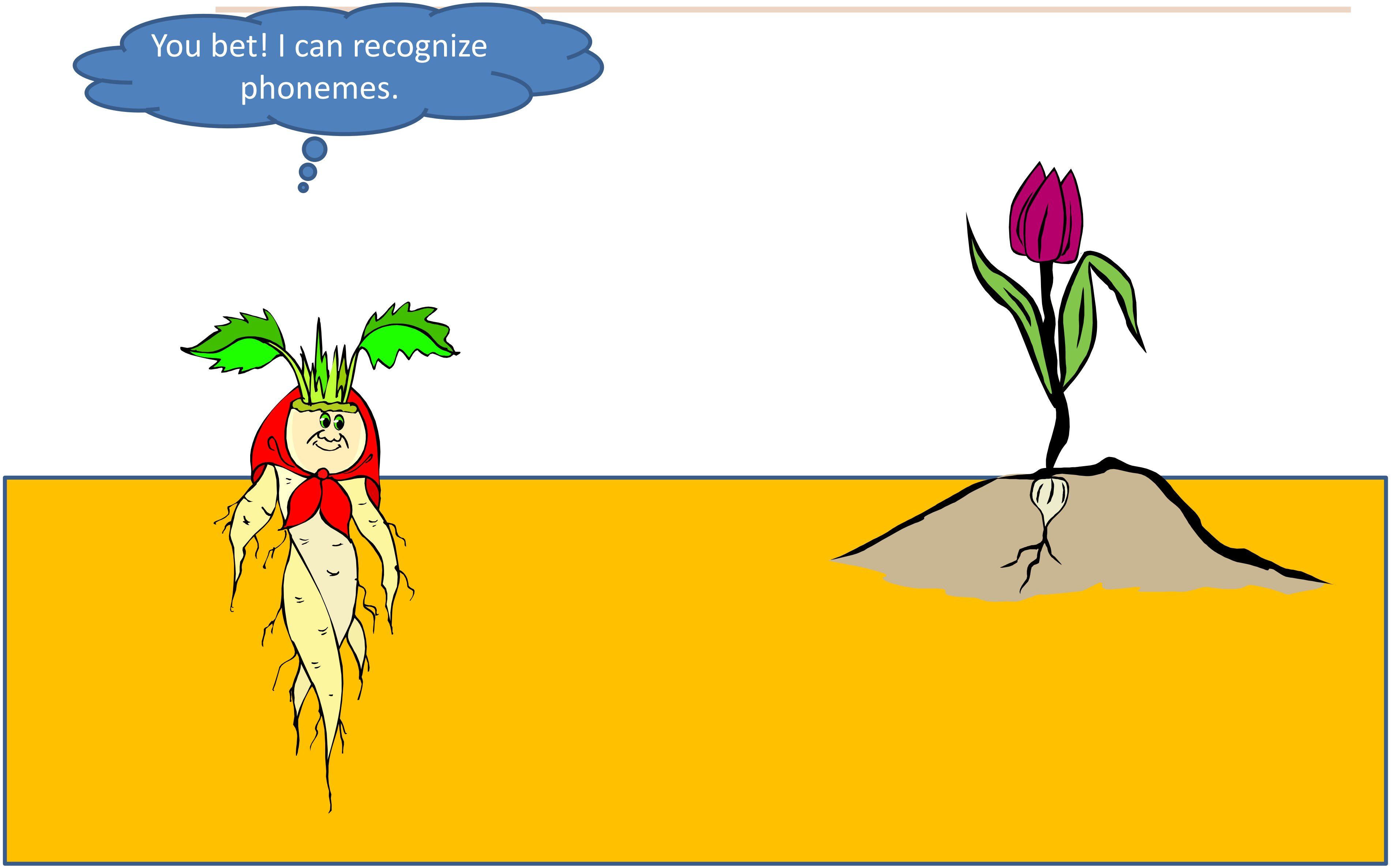
Microsoft Research

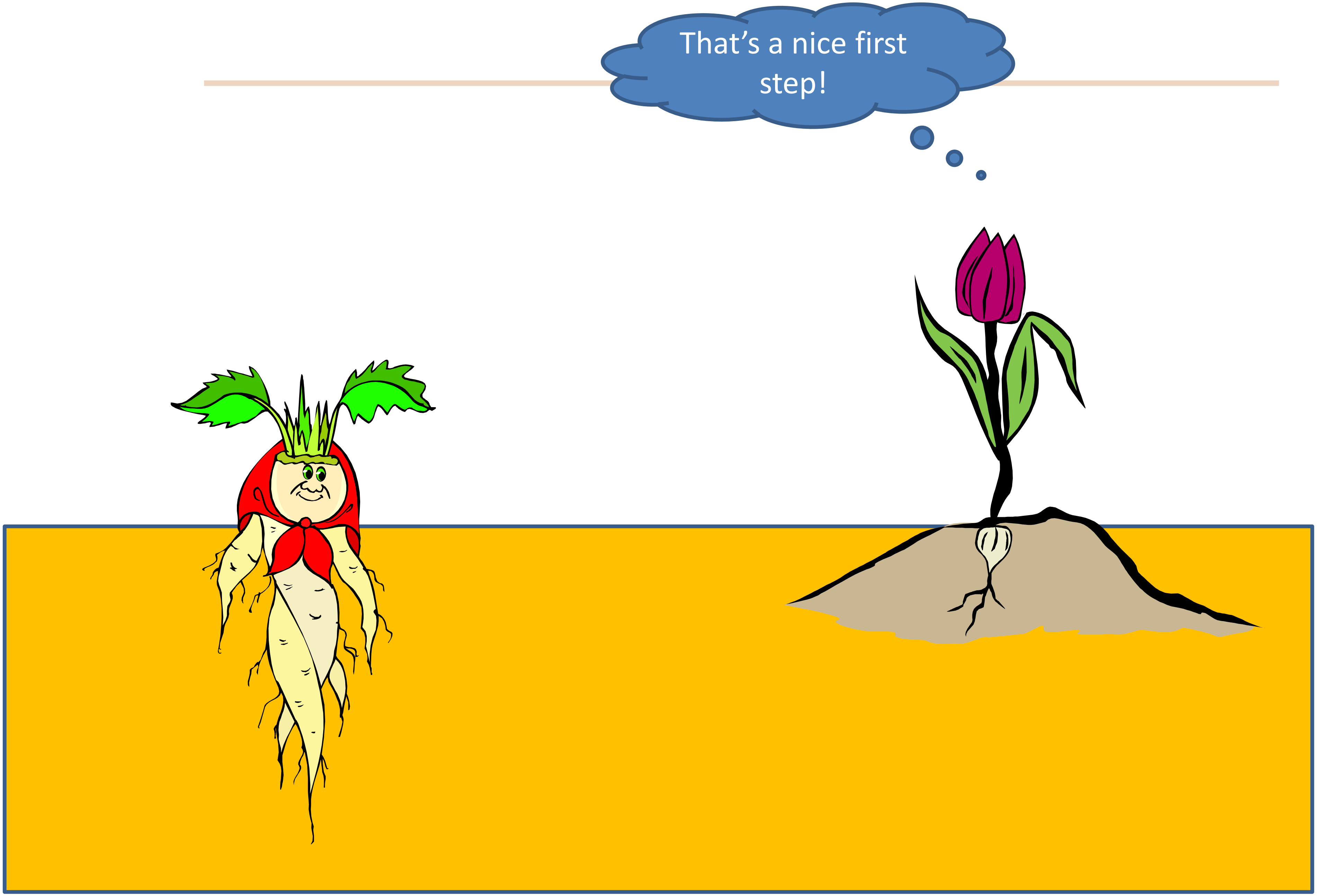


Microsoft Research

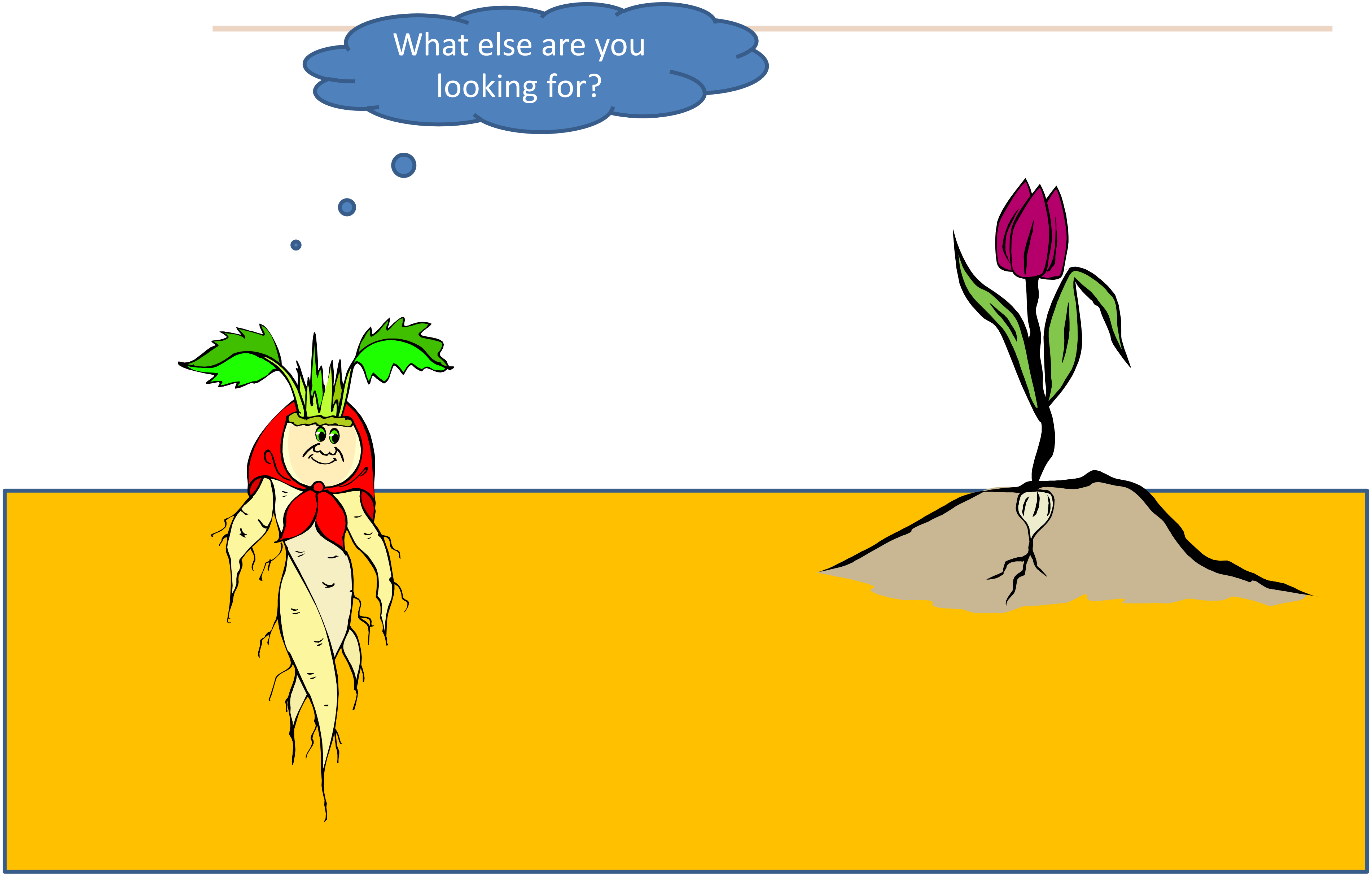


Microsoft Research

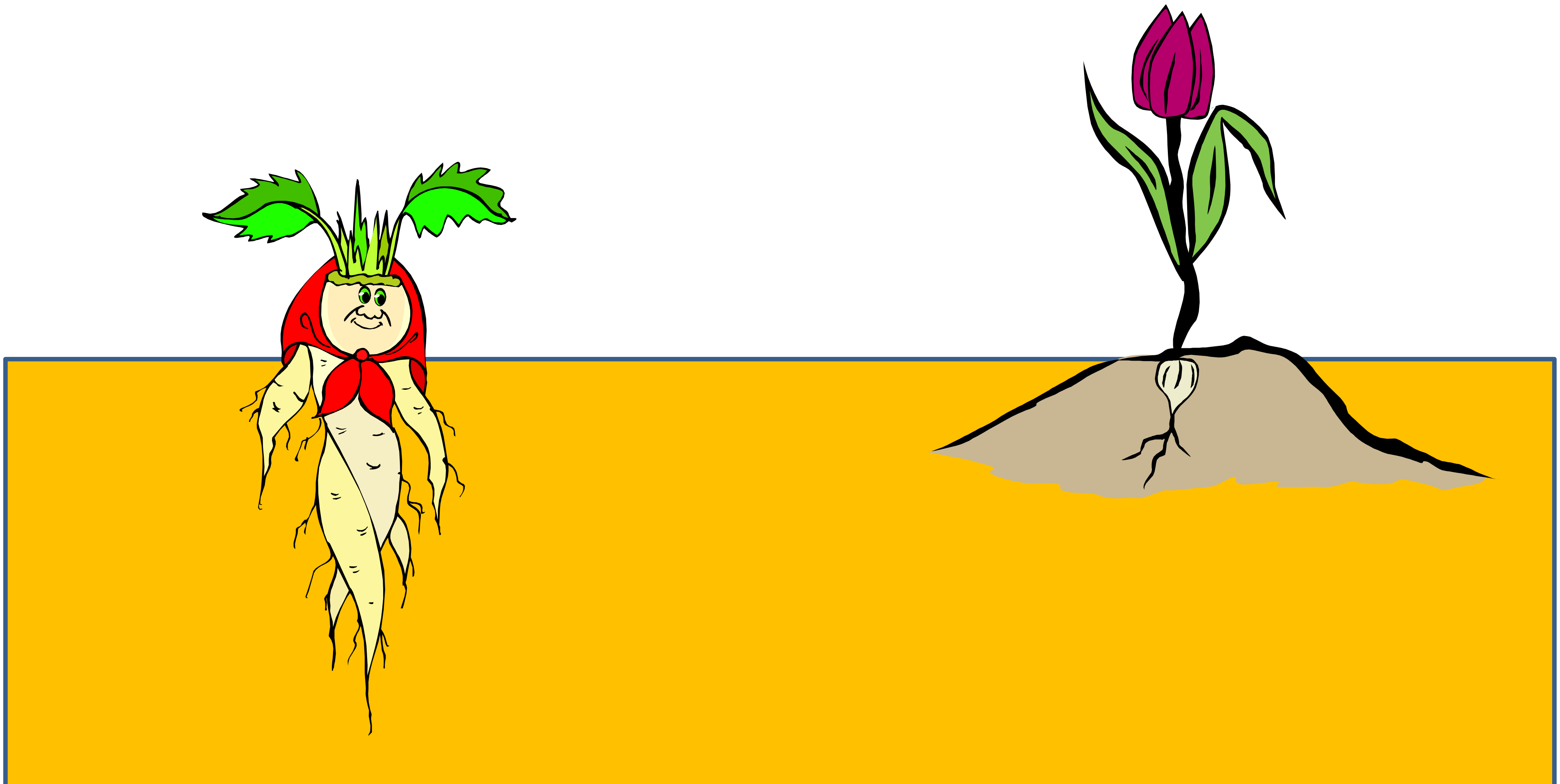



That's a nice first step!

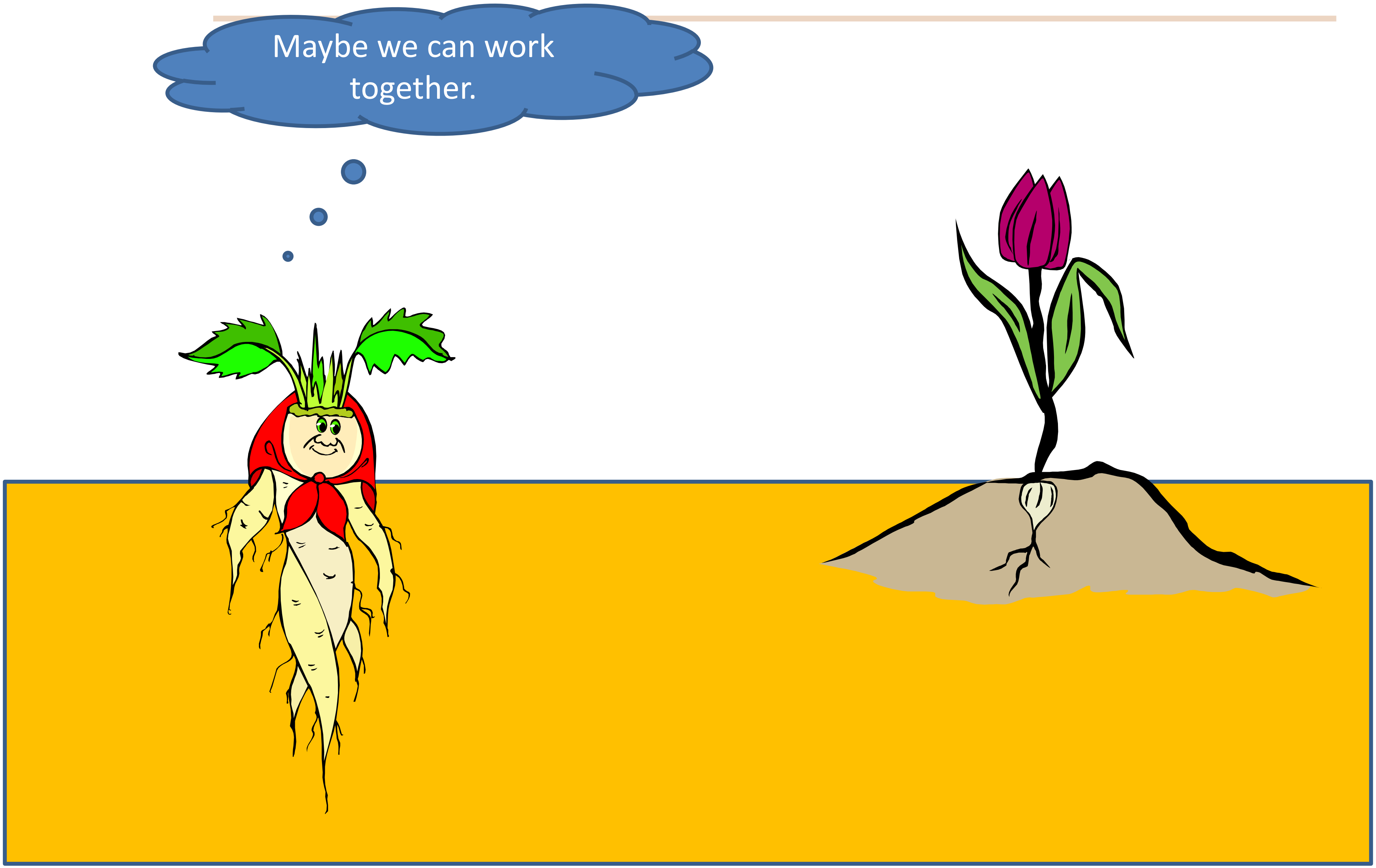
Microsoft Research



Recognizing noisy sentences spoken by unknown people.



Microsoft Research



Deep speech recognizer is born.

Multi-objective

Competitive Learning

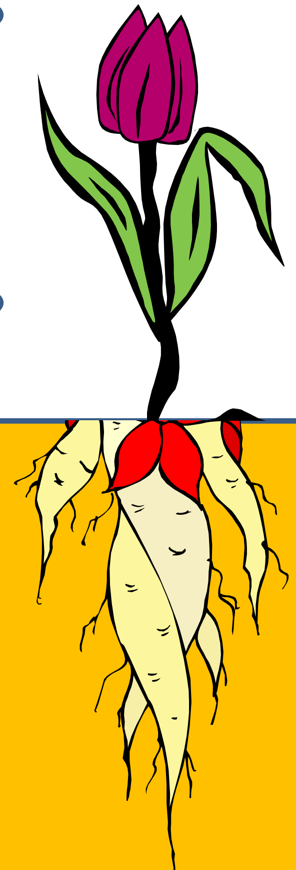
Hierarchical

Deep Belief Net

Conditional

Scalable

Recurrent



Microsoft Research

Industry Scale Deep Learning

Continued at MSR, 2010, 2011…

-2010: slowly more people in MSR-speech joined DBN-DNN research

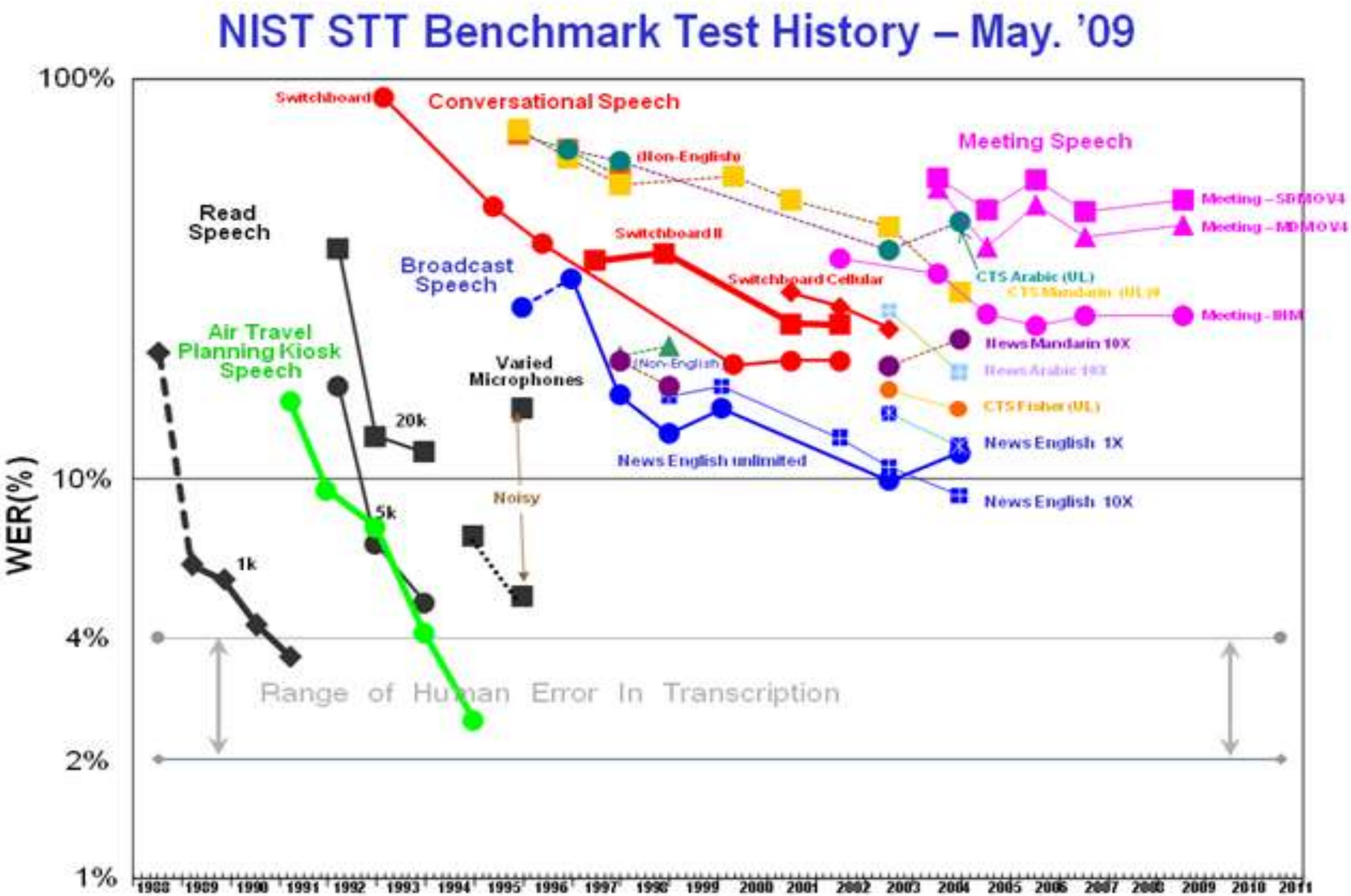
-July 2010: success of bottleneck feature coding using speech spectrogram; Interspeech-2010 paper Deng/Hinton et al.

-**August 2010**: success of DNN in large-vocabulary speech recognition (**voice search**); paper in ICASSP-2011 (Dahl/Yu/Deng)

-Oct 2010: MSR/MSRA collaboration started on Switchboard task

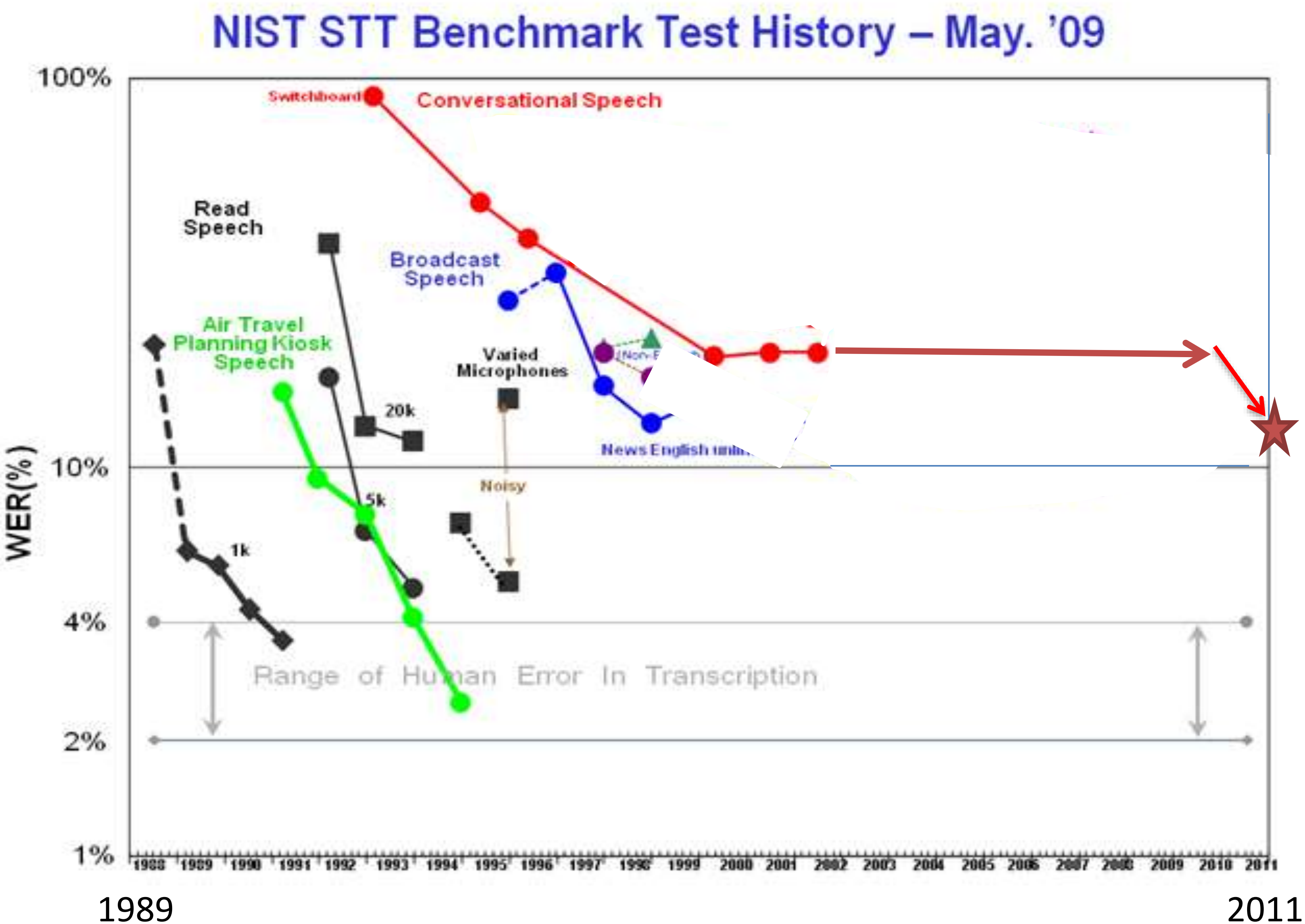
-March 2011: Success in the Switchboard task by MSR/MSRA; Interspeech-2011: Seide/Yu, et al. Success of **deep stacking net**: Deng/Yu/Platt.

The History of
Automatic Speech Recognition Evaluations at NIST



1989 1999 2011

The History of
Automatic Speech Recognition Evaluations at NIST



Deep Learning and Feature Learning Today

Y LeCun
MA Ranzato

- **Deep Learning has been the hottest topic in speech recognition in the last 2 years**
  - A few long-standing performance records were broken with deep learning methods
  - Microsoft and Google have both deployed DL-based speech recognition system in their products
  - Microsoft, Google, IBM, Nuance, AT&T, and all the major academic and industrial players in speech recognition have projects on deep learning
- **Deep Learning is the hottest topic in Computer Vision**
  - Feature engineering is the bread-and-butter of a large portion of the CV community, which creates some resistance to feature learning
  - But the record holders on ImageNet and Semantic Segmentation are convolutional nets
- **Deep Learning is becoming hot in Natural Language Processing**
- **Deep Learning/Feature Learning in Applied Mathematics**

In Many Fields, Feature Learning Has Caused a Revolution (methods used in commercially deployed systems)

Y LeCun
MA Ranzato

- **Speech Recognition I (late 1980s)**
  - Trained mid-level features with Gaussian mixtures (2-layer classifier)
- **Handwriting Recognition and OCR (late 1980s to mid 1990s)**
  - Supervised convolutional nets operating on pixels
- **Face & People Detection (early 1990s to mid 2000s)**
  - Supervised convolutional nets operating on pixels (YLC 1994, 2004, Garcia 2004)
  - Haar features generation/selection (Viola-Jones 2001)
- **Object Recognition I (mid-to-late 2000s: Ponce, Schmid, Yu, YLC....)**
  - Trainable mid-level features (K-means or sparse coding)
- **Low-Res Object Recognition: road signs, house numbers (early 2010's)**
  - Supervised convolutional net operating on pixels
- **Speech Recognition II (circa 2011)**
  - Deep neural nets for acoustic modeling
- **Object Recognition III, Semantic Labeling (2012, Hinton, YLC,...)**
  - Supervised convolutional nets operating on pixels

Microsoft Research

Outline

PART I: Basics of Deep Learning (DL)

(including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

PART II: Deeper Substance of DL

(including connections to other ML paradigms, example of incorporating speech knowledge in DL architecture, and recent experiments in speech recognition)

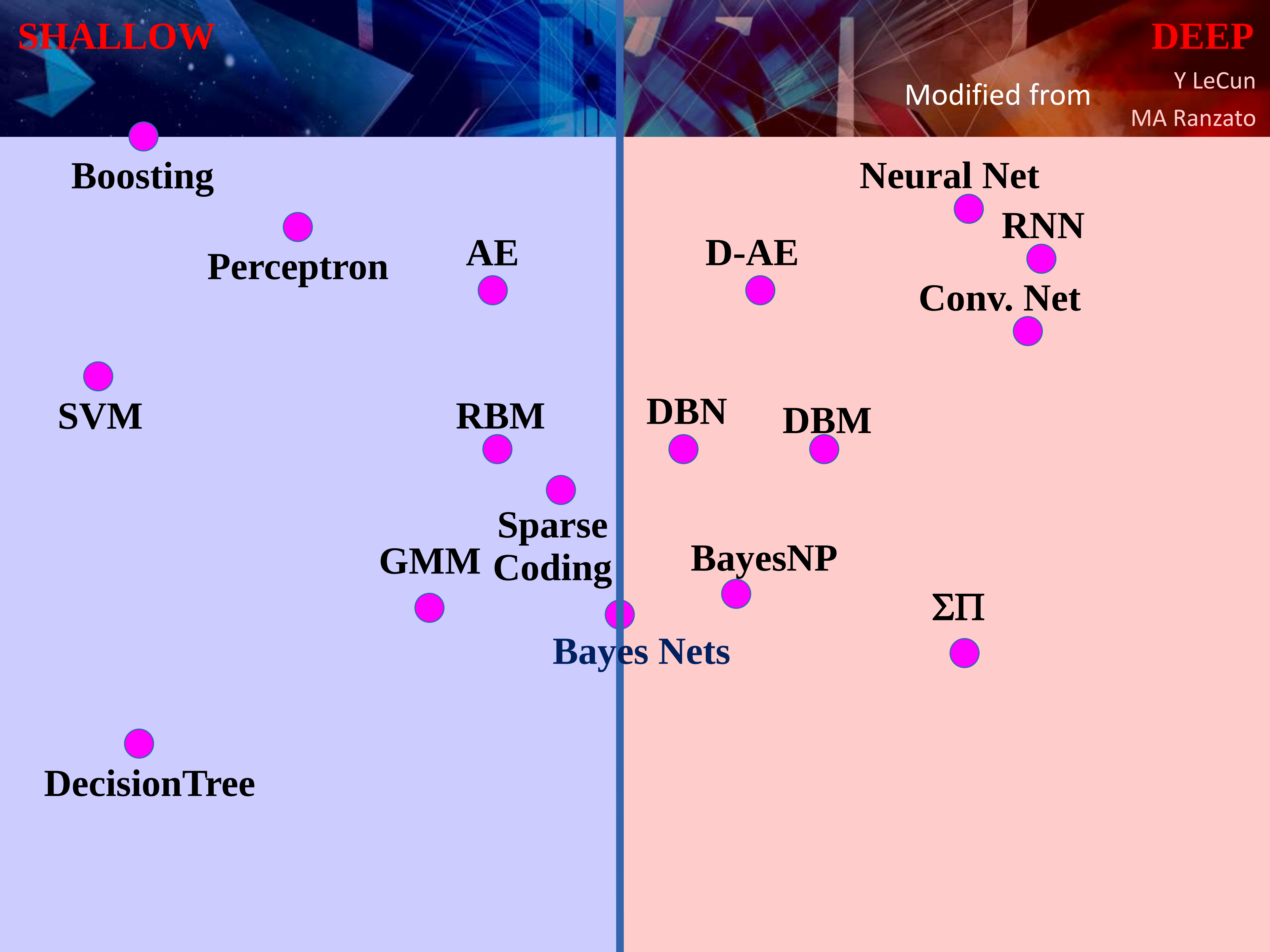
Machine Learning Paradigms for Speech Recognition: An Overview

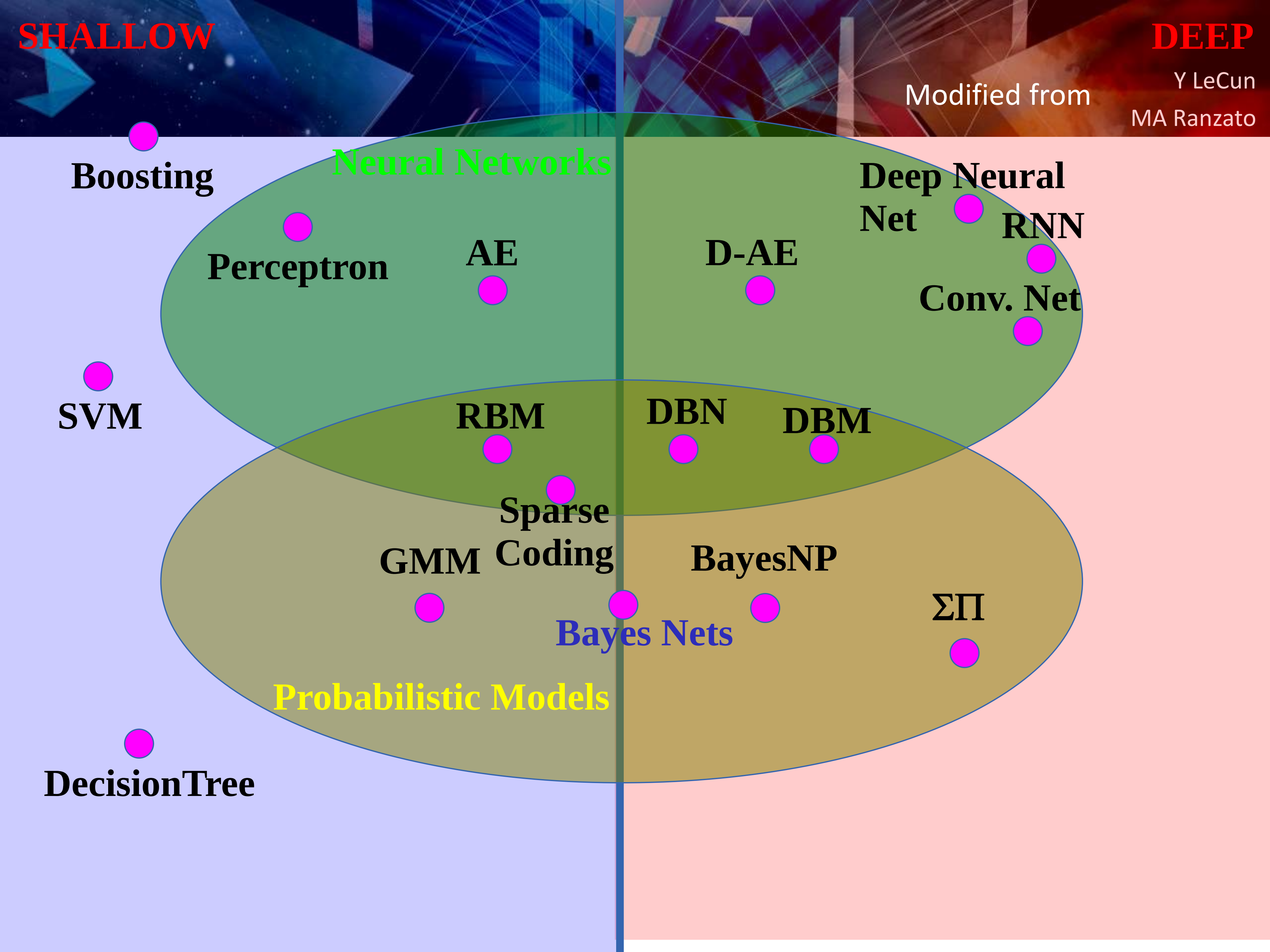
Li Deng, *Fellow, IEEE*, and Xiao Li, *Member, IEEE*

***Abstract*—Automatic Speech Recognition (ASR) has historically been a driving force behind many machine learning (ML) techniques, including the ubiquitously used hidden Markov model, discriminative learning, structured sequence learning, Bayesian learning, and adaptive learning. Moreover, ML can and occasionally does use ASR as a large-scale, realistic application to rigorously test the effectiveness of a given technique, and to inspire new problems arising from the inherently sequential and dynamic nature of speech. On the other hand, even though ASR is available commercially for some applications, it is largely an unsolved problem—for almost all applications, the performance of ASR is not on par with human performance. New insight from modern ML methodology shows great promise to advance the state-of-the-art in ASR technology. This overview article provides readers with an overview of modern ML techniques as utilized in the current and as relevant to future ASR research and systems. The intent is to foster further cross-pollination between the ML**

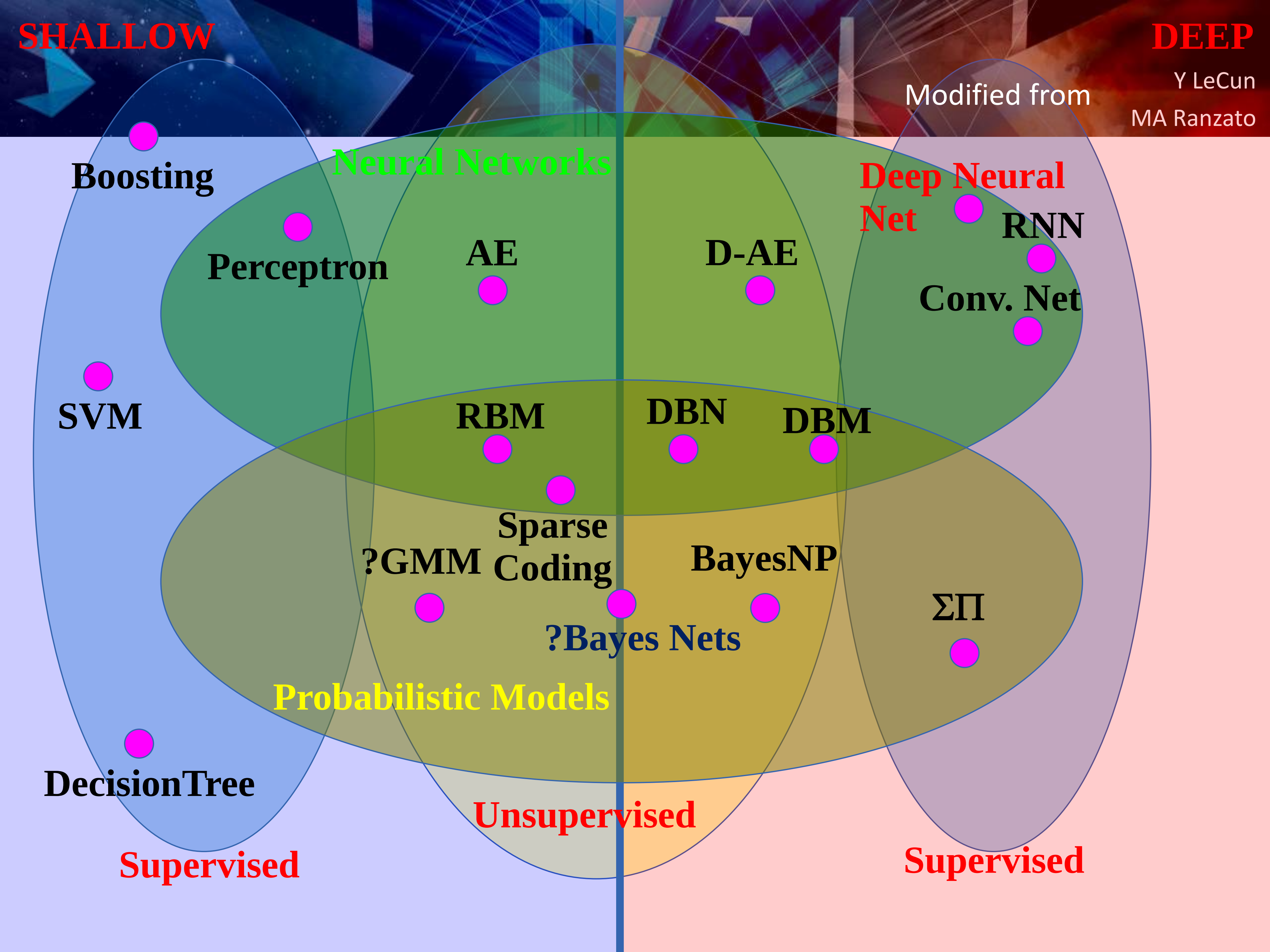
community to make assumptions about a problem, develop precise mathematical theories and algorithms to tackle the problem given those assumptions, but then evaluate on data sets that are relatively small and sometimes synthetic. ASR research, on the other hand, has been driven largely by rigorous empirical evaluations conducted on very large, standard corpora from real world. ASR researchers often found formal theoretical results and mathematical guarantees from ML of less use in preliminary work. Hence they tend to pay less attention to these results than perhaps they should, possibly missing insight and guidance provided by the ML theories and formal frameworks even if the complex ASR tasks are often beyond the current state-of-the-art in ML.

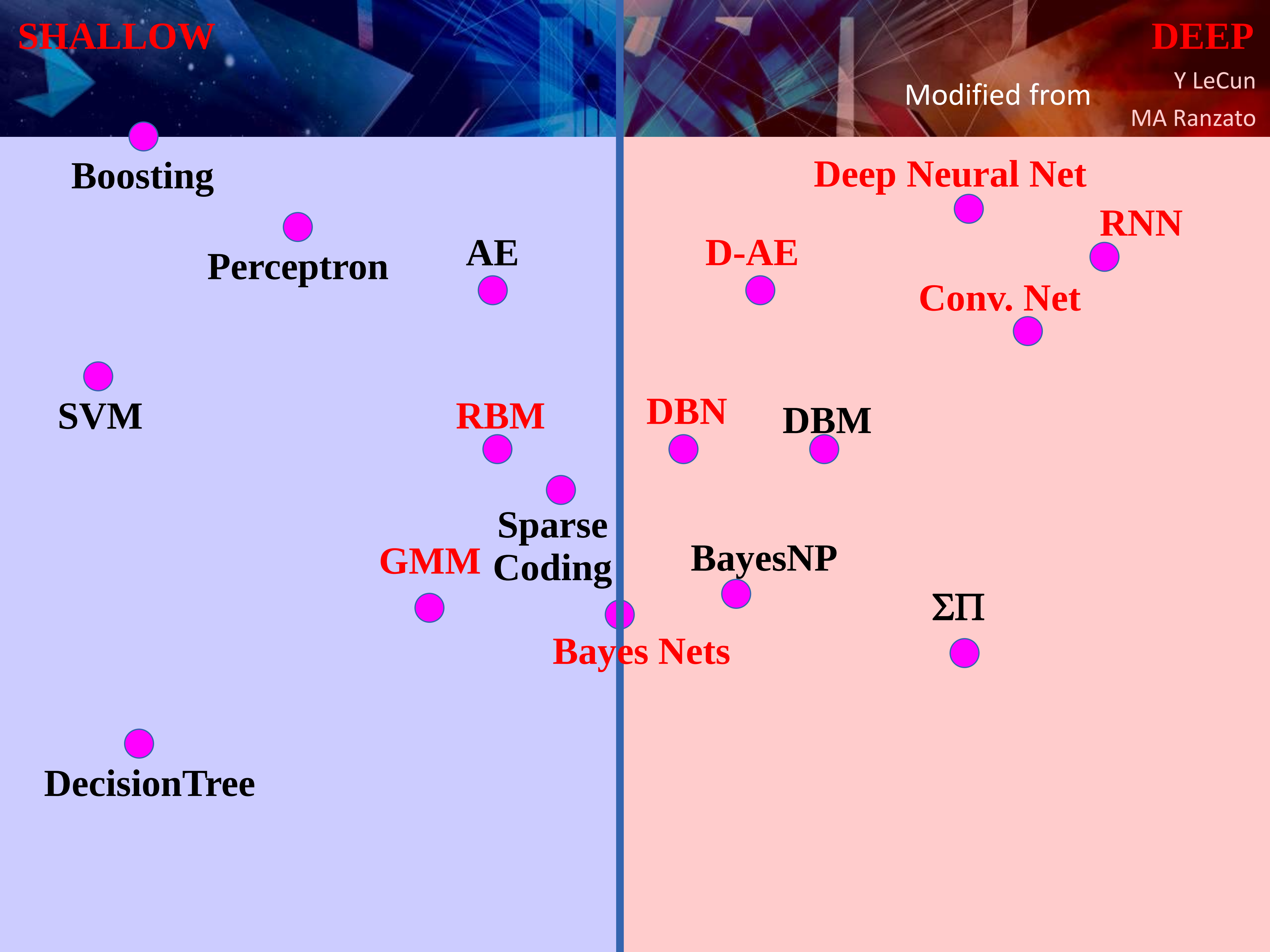
This overview article is intended to provide readers of IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE


SHALLOW
DEEP
Modified from
Y LeCun
MA Ranzato
Boosting
Neural Net
RNN
Perceptron
AE
D-AE
Conv. Net
SVM
RBM
DBN
DBM
Sparse
Coding
GMM
BayesNP
ΣΠ
Bayes Nets
DecisionTree



SHALLOW
DEEP
Modified from
Y LeCun
MA Ranzato
Boosting
Neural Networks
Deep Neural Net
RNN
Perceptron
AE
D-AE
Conv. Net
SVM
RBM
DBN
DBM
Sparse Coding
GMM
BayesNP
ΣΠ
Bayes Nets
Probabilistic Models
DecisionTree


SHALLOW
DEEP
Modified from
Y LeCun
MA Ranzato
Neural Networks
Boosting
Deep Neural Net
RNN
Perceptron
AE
D-AE
Conv. Net
SVM
RBM
DBN
DBM
Sparse Coding
?GMM
BayesNP
ΣΠ
?Bayes Nets
Probabilistic Models
DecisionTree
Unsupervised
Supervised
Supervised


SHALLOW
DEEP
Modified from
Y LeCun
MA Ranzato
Boosting
Deep Neural Net
RNN
Perceptron
AE
D-AE
Conv. Net
SVM
RBM
DBN
DBM
Sparse
Coding
GMM
BayesNP
ΣΠ
Bayes Nets
DecisionTree

Microsoft Research

Outline

PART I: Basics of Deep Learning (DL)

(including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

PART II: Deeper Substance of DL

---**Technical introduction: RBM, DBN, DNN, CNN, RNN**

---Advanced: 2 examples of incorporating domain knowledge (speech) into DL architectures

---Novel DL architectures and recent experiments

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

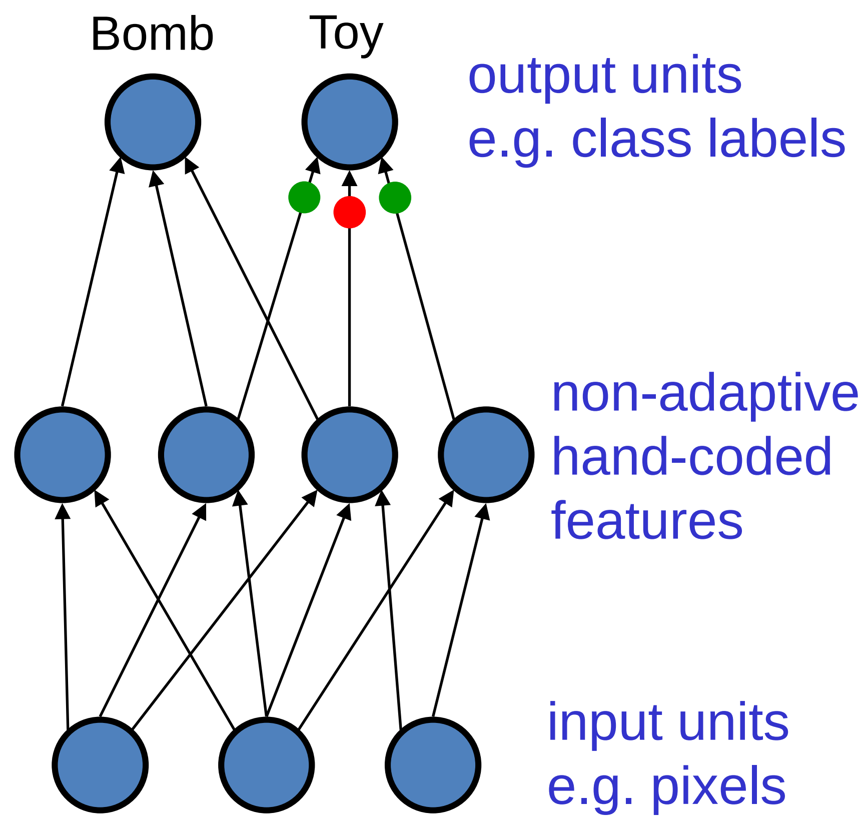
[The shared views of four research groups]

FUNDAMENTAL TECHNOLOGIES
IN MODERN SPEECH RECOGNITION

Microsoft Research

First generation neural networks

- Perceptrons (~1960) used a layer of hand-coded features and tried to recognize objects by learning how to weight these features.
  - There was a neat learning algorithm for adjusting the weights.
  - But perceptrons are fundamentally limited in what they can learn to do.



Sketch of a typical perceptron from the 1960's

(Slide from Hinton)

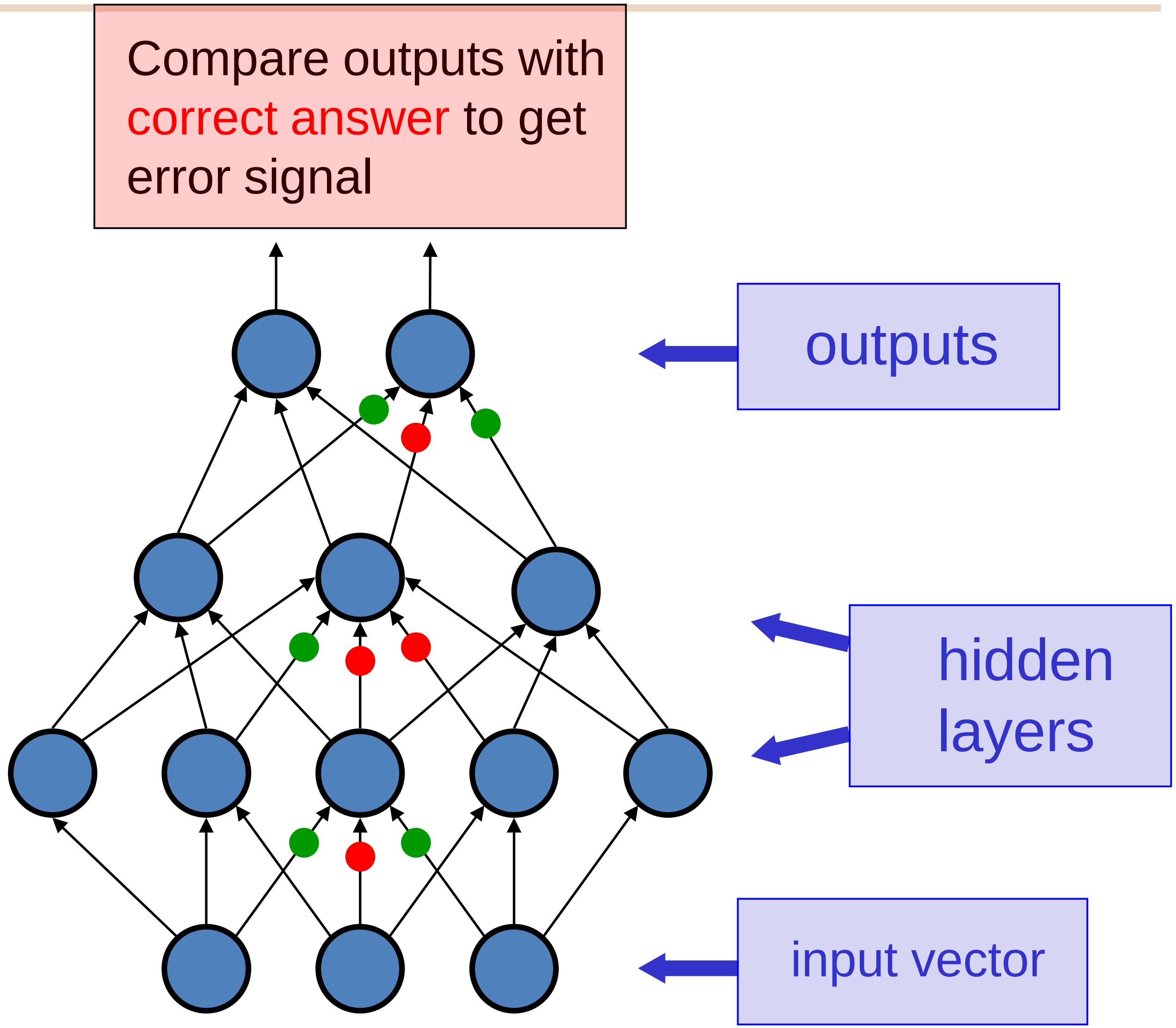
Support Vector Machine is a perceptron

- Vapnik and his co-workers developed a very clever type of perceptron called a Support Vector Machine.
  - Instead of hand-coding the layer of non-adaptive features, each training example is used to create a new feature using a fixed recipe.
    - The feature computes how similar a test example is to that training example.
  - Then a clever optimization technique is used to select the best subset of the features and to decide how to weight each feature when classifying a test case.
    - But its just a perceptron and has all the same limitations.
- In the 1990's, many researchers abandoned neural networks with multiple adaptive hidden layers because Support Vector Machines worked better.

(Slide modified from Hinton)

Second generation neural networks (~1985)

Back-propagate error signal to get derivatives for learning



(Slide from Hinton)

What is wrong with back-propagation?

(a plausible story, but false; Hinton ICASSP-2013)

- It requires labeled training data.
  - Almost all data is unlabeled.
- The learning time does not scale well
  - It is very slow in networks with multiple hidden layers.
- It can get stuck in poor local optima.
  - These are often quite good, but for deep nets they are far from optimal
- **Deep learning (partially) overcomes these difficulties** by using undirected graphical model

Microsoft Research

What was actually wrong with back-propagation?

- We didn’t collect enough labeled data.
- We didn’t have fast enough computers.
- We didn’t initialize the weights correctly

- If we fix these three problems, it works really well.

(Hinton: ICASSP-2013)

Microsoft Research

What has happened since 1985

- Labeled datasets got much bigger.
- Computers got much faster.
- We found better ways to initialize the weights of a deep net using unlabeled data.
- As a result, deep neural networks are now state-of-the-art for tasks like object recognition or acoustic modeling for speech recognition

(Hinton: ICASSP-2013)

Microsoft Research

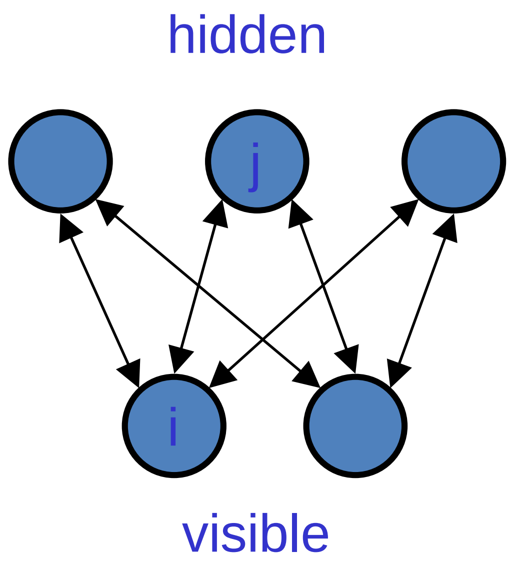
Initializing the weights in a deep neural net using unlabeled data

- This was historically important in overcoming the belief that deep neural networks could not be trained effectively. (Hinton: ICASSP-2013)
  - This was a very strong belief.
  - It prevented papers being published in good conferences and journals.

- For the tasks with small amounts of labeled training data, such initialization is still very useful

Restricted Boltzmann Machines (RBM)

- We restrict the connectivity to make learning easier.
  - Only one layer of hidden units.
  - No connections between hidden units.
- In an RBM, the hidden units are conditionally independent given the visible states.
- So we can quickly get an unbiased sample from the posterior distribution when given a data-vector.



RBM: Weights → Energies → Probabilities

- Joint distribution $\mathrm{p}(\mathbf{v},\mathbf{h};\theta)$ is defined in terms of an energy function $\mathrm{E}(\mathbf{v},\mathbf{h};\theta)$

$$\mathrm{p}(\mathbf{v},\mathbf{h};\theta) = \frac{exp\big(-\mathrm{E}(\mathbf{v},\mathbf{h};\theta)\big)}{Z}$$

- For a Bernoulli-Bernoulli RBM

$$\mathrm{E}(\mathbf{v},\mathbf{h};\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}\, v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j$$

- For a Gaussian-Bernoulli RBM

$$\mathrm{E}(\mathbf{v},\mathbf{h};\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}\, v_i h_j + \frac{1}{2}\sum_{i=1}^{V}(v_i - b_i)^2 - \sum_{j=1}^{H} a_j h_j$$

- $\mathrm{p}(\mathbf{v},\mathbf{h};\theta)$ → generative model!

Restricted Boltzmann Machine (RBM)

- Conditional probabilities are very easy to calculate
- For a Bernoulli-Bernoulli RBM

Inference $$p(h_j = 1|\mathbf{v};\theta) = \sigma\left(\sum_{i=1}^{V} w_{ij}\, v_i + a_j\right)$$

synthesis $$p(v_i = 1|\mathbf{h};\theta) = \sigma\left(\sum_{j=1}^{H} w_{ij}\, h_j + b_i\right)$$

- For a Gaussian-Bernoulli RBM

Inference $$p(h_j = 1|\mathbf{v};\theta) = \sigma\left(\sum_{i=1}^{V} w_{ij}\, v_i + a_j\right)$$

$$p(v_i|\mathbf{h};\theta) = N\left(\sum_{j=1}^{H} w_{ij}\, h_j + b_i, 1\right)$$

- Proof next page. (This is a “neural net” with stochastic units rather than the deterministic MLP that you may be more familiar with)

Microsoft Research

$$
\begin{aligned}
P(\mathbf{h}|\mathbf{v}) &= \frac{e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\tilde{\mathbf{h}}} e^{-E(\mathbf{v},\tilde{\mathbf{h}})}} \\
&= \frac{e^{\mathbf{b}^T\mathbf{v}+\mathbf{c}^T\mathbf{h}+\mathbf{v}^T\mathbf{W}\mathbf{h}}}{\sum_{\tilde{\mathbf{h}}} e^{\mathbf{b}^T\mathbf{v}+\mathbf{c}^T\tilde{\mathbf{h}}+\mathbf{v}^T\mathbf{W}\tilde{\mathbf{h}}}} \\
&= \frac{e^{\mathbf{c}^T\mathbf{h}+\mathbf{v}^T\mathbf{W}\mathbf{h}}}{\sum_{\tilde{\mathbf{h}}} e^{\mathbf{c}^T\tilde{\mathbf{h}}+\mathbf{v}^T\mathbf{W}\tilde{\mathbf{h}}}} \\
&= \frac{\prod_i e^{c_i h_i+\mathbf{v}^T\mathbf{W}_{*,i}h_i}}{\sum_{\tilde{h}_1}\cdots\sum_{\tilde{h}_N}\prod_i e^{c_i\tilde{h}_i+\mathbf{v}^T\mathbf{W}_{*,i}\tilde{h}_i}} \\
&= \frac{\prod_i e^{-\gamma_i(\mathbf{v},h_i)}}{\sum_{\tilde{h}_1}\cdots\sum_{\tilde{h}_N}\prod_i e^{-\gamma_i(\mathbf{v},\tilde{h}_i)}} \\
&= \frac{\prod_i e^{-\gamma_i(\mathbf{v},h_i)}}{\prod_i\sum_{\tilde{h}_i} e^{-\gamma_i(\mathbf{v},\tilde{h}_i)}} \\
&= \prod_i \frac{e^{-\gamma_i(\mathbf{v},h_i)}}{\sum_{\tilde{h}_i} e^{-\gamma_i(\mathbf{v},\tilde{h}_i)}} && (5) \\
&= \prod_i P(h_i|\mathbf{v}). && (6)
\end{aligned}
$$

Since the $h_i \in \{0, 1\}$, the sum in the denominator of equation (5) has only two terms and thus

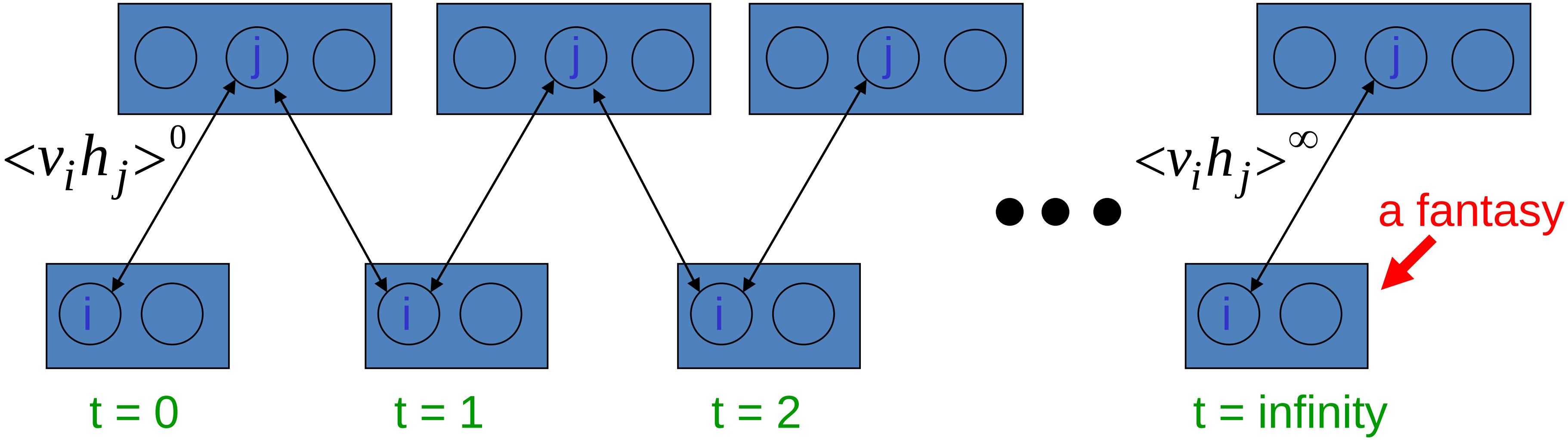
$$
\begin{aligned}
P(h_i = 1|\mathbf{v}) &= \frac{e^{-\gamma_i(\mathbf{v},1)}}{e^{-\gamma_i(\mathbf{v},1)} + e^{-\gamma_i(\mathbf{v},0)}} \\
&= \sigma(c_i + \mathbf{v}^T\mathbf{W}_{*,i}),
\end{aligned}
$$

yielding

$$
P(\mathbf{h} = \mathbf{1}|\mathbf{v}) = \sigma(\mathbf{c} + \mathbf{v}^T\mathbf{W}), \qquad (7)
$$

Microsoft Research

Maximum likelihood learning for RBM



Start with a training vector on the visible units.

Then alternate between updating all the hidden units in parallel and updating all the visible units in parallel.

$$\frac{\partial \log p(v)}{\partial w_{ij}} = <v_i h_j>^0 - <v_i h_j>^\infty$$

Training RBMs

- $\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$
- Approximate $\langle v_i h_j \rangle_{model}$
  i. Initialize $\mathbf{v_0}$ at data
  ii. Sample $\mathbf{h_0} \sim \boldsymbol{p}(\mathbf{h}|\mathbf{v_0})$
  iii. Sample $\mathbf{v_1} \sim \boldsymbol{p}(\mathbf{v}|\mathbf{h_0})$
  iv. Sample $\mathbf{h_1} \sim \boldsymbol{p}(\mathbf{h}|\mathbf{v_1})$
  v. Call $(\mathbf{v_1}, \mathbf{h_1})$ a sample from the model.
- $(\mathbf{v}_\infty, \mathbf{h}_\infty)$ is a true sample from the model. $(\mathbf{v_1}, \mathbf{h_1})$ is a very rough estimate but worked
- Contrastive divergence algorithm (CD)

Microsoft Research

RBM versus GMM

- Gaussian Mixture Model
  - Local representation
  - (In practice,) data vector explained by only a single Gaussian
  - Tend to over-fit
- Bernoulli-Gaussian RBM
  - Distributed representation, very powerful
  - Product of Gaussians
  - Tend to under-fit

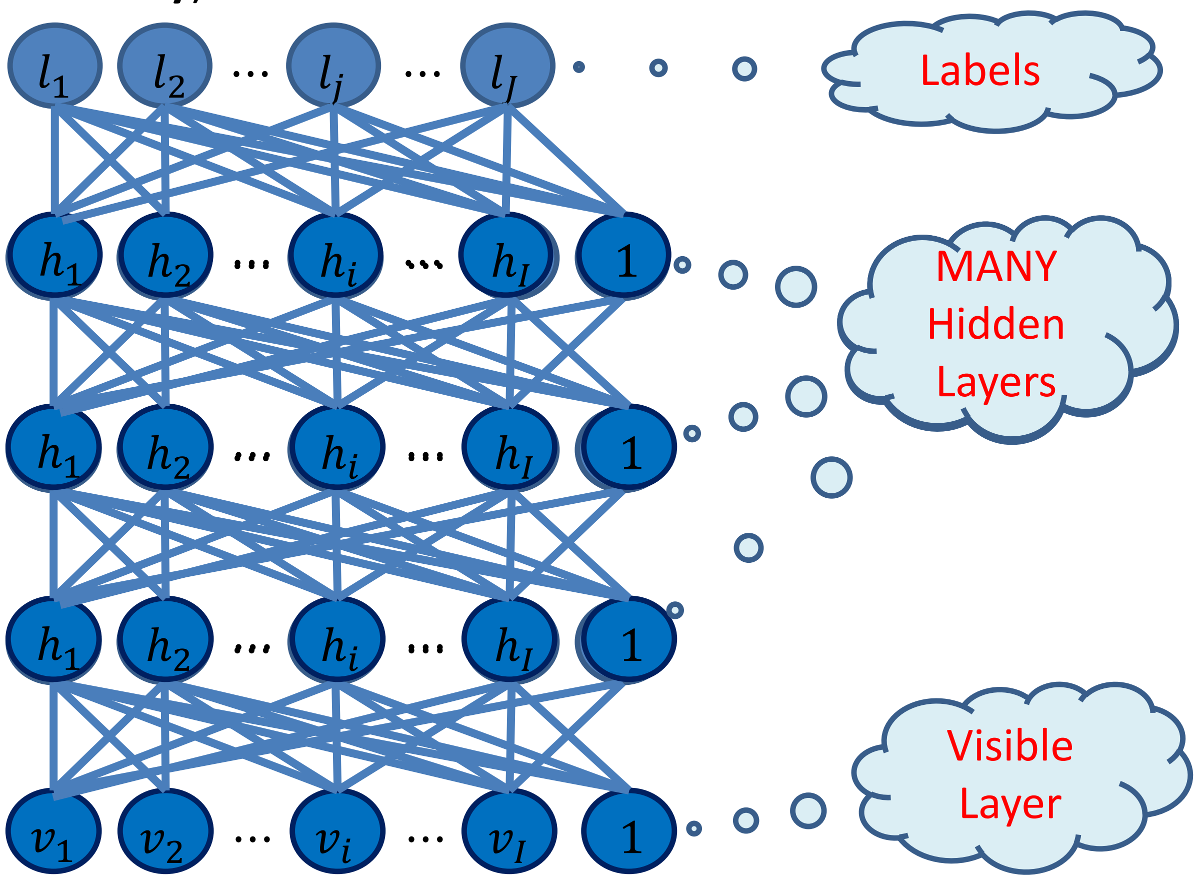
Microsoft Research

Building a Deep Network

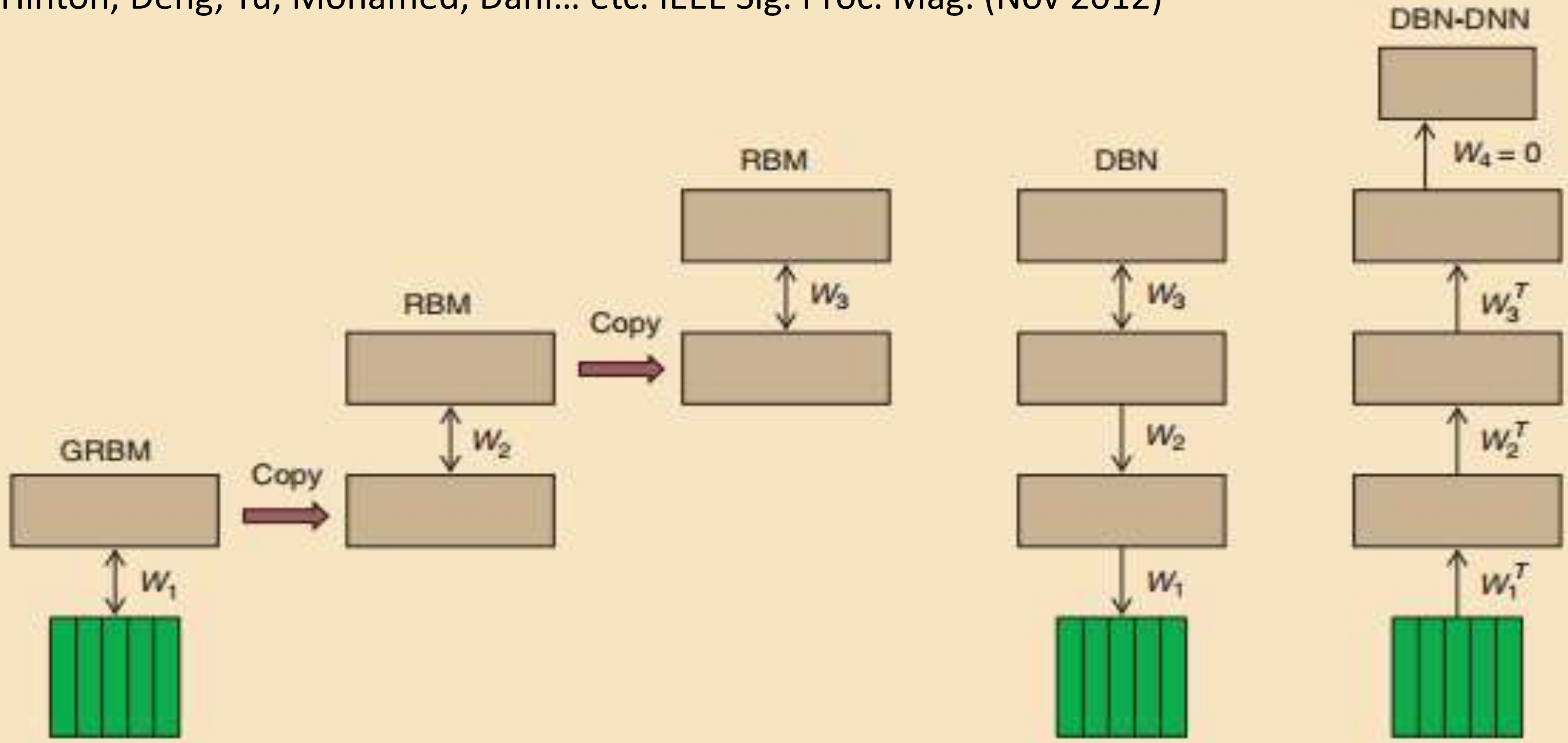
- This is the main reason why RBM's are interesting (as a building block)
- First train a layer of hidden units that receive input directly from the data (image, speech, coded text, etc).
- Then treat the activations of hidden units (the trained "features") as if they were "data" and learn features of features in a second hidden layer.
- It can be proved that each time we add another layer of features we improve a variational lower bound on the log probability of the training data.
  - The proof is complicated (Hinton et al, 2006)
  - Based on an equivalence between an RBM and a deep directed model

Deep Belief Net (DBN) & Deep Neural Net (DNN)

- DBN: Undirected at top two layers which is an RBM; directed Bayes net (top-down) at lower layers (good for **synthesis and recognition**)
- DNN: Multi-layer perceptron (bottom up) + unsupervised pre-training w. RBM weights (good for **recognition only**)



Hinton, Deng, Yu, Mohamed, Dahl… etc. IEEE Sig. Proc. Mag. (Nov 2012)



First train a stack of three models each of which has one hidden layer. Each model in the stack treats the hidden variables of the previous model as data.

Then compose them into a single Deep Belief Network.

Then add outputs and train the DNN with backprop.

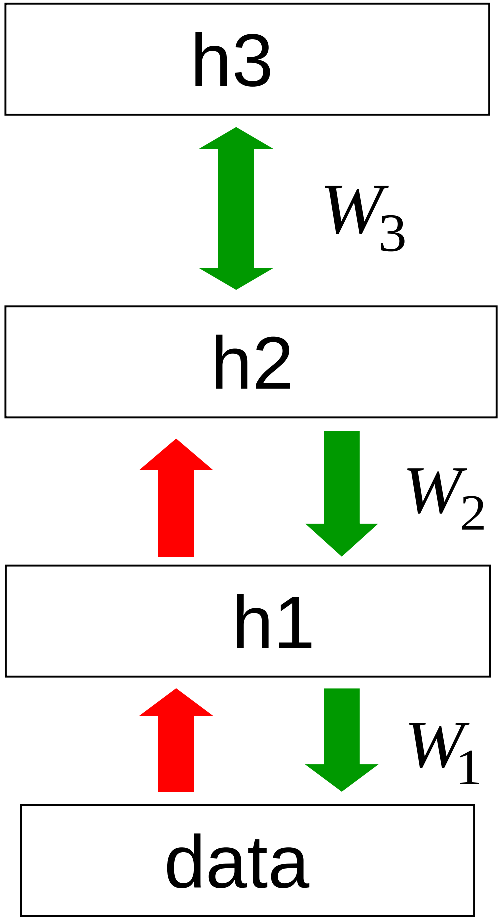
Quiz Questions

1. DBN & DNN: which one is generative? Which one is discriminative?
2. How can a generative model be used for recognition? (Bayes rule as for HMM speech recognition)
3. How does DBN do synthesis?
4. How does DBN do recognition?
5. How does DNN do recognition?
6. For recognition, is RBN or DNN better?
7. What is the difference between DBN and Dynamic Bayes Net (a.k.a. "DBN")?

The Answer to Quiz Question 3:

- To generate data:
  1. Get an equilibrium sample from the top-level RBM by performing alternating Gibbs sampling for a long time.
  2. Perform a top-down pass to get states for all the other layers.

So the lower level bottom-up connections are not part of the generative model. They are just used for inference.

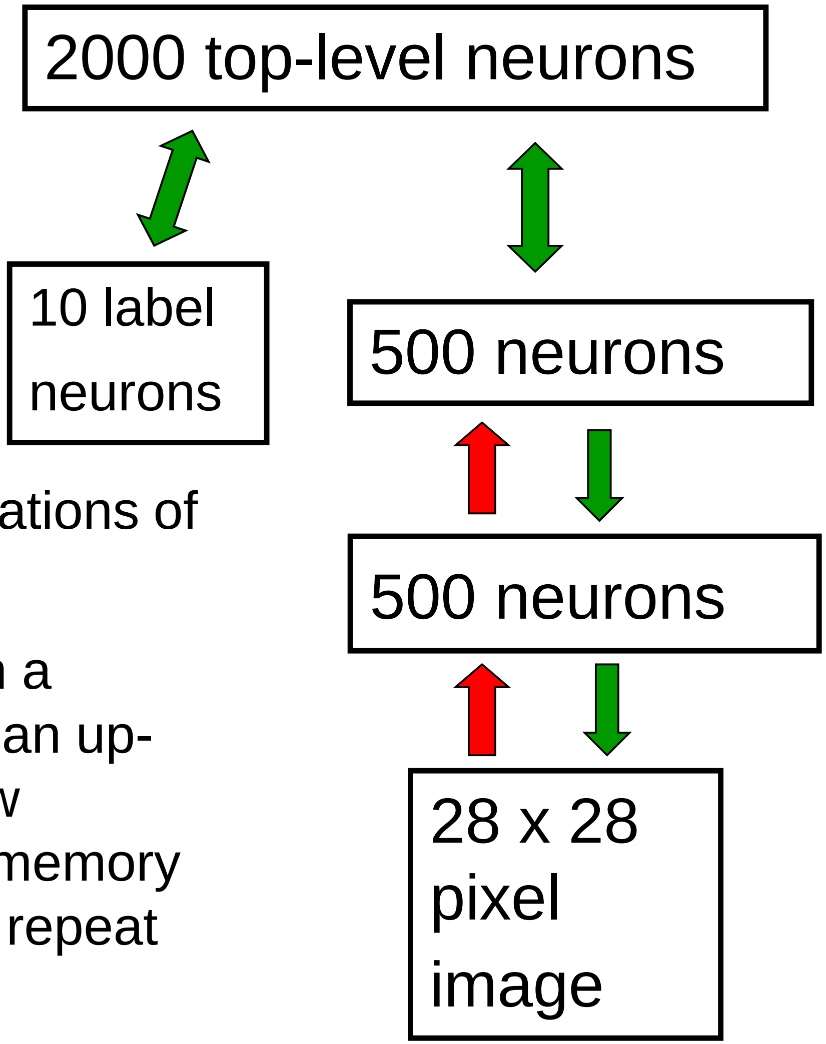


Samples generated by letting the associative memory run with one label clamped. There are 1000 iterations of alternating Gibbs sampling between samples (example from Hinton) .

Answer to Quiz Question 4: Example of digit/image recognition by DBN

The top two layers form an associative memory whose energy landscape models the low dimensional manifolds of the digits

The energy valleys have names →



The model learns to generate combinations of labels and images.

To perform recognition we start with a neutral state of the label units and do an up-pass from the image followed by a few iterations of the top-level associative memory ---> probability of that digit label; then repeat for all digit labels; then compare.

(slide modified from Hinton)

Microsoft Research

DBN & DNN: Fine-tuning for discrimination

- First learn one layer at a time greedily.
- Then treat this as “pre-training” that finds a good initial set of weights which can be fine-tuned by a local search procedure.
- For DBN: Contrastive wake-sleep (see Hinton’s)
- **For DNN: Back-propagation**
  - This overcomes many of the limitations of standard backpropagation (if you do not have large labeled training data).

Fine Tuning DNN after pre-training: Optimization view

- Stacking RBMs one layer at a time scales well to really big networks

- Do not start back-propagation until sensible feature detectors are found by RBM pre-training that should already be very helpful for the discrimination task.

- Back-propagation only needs to perform a local search from a sensible starting point.

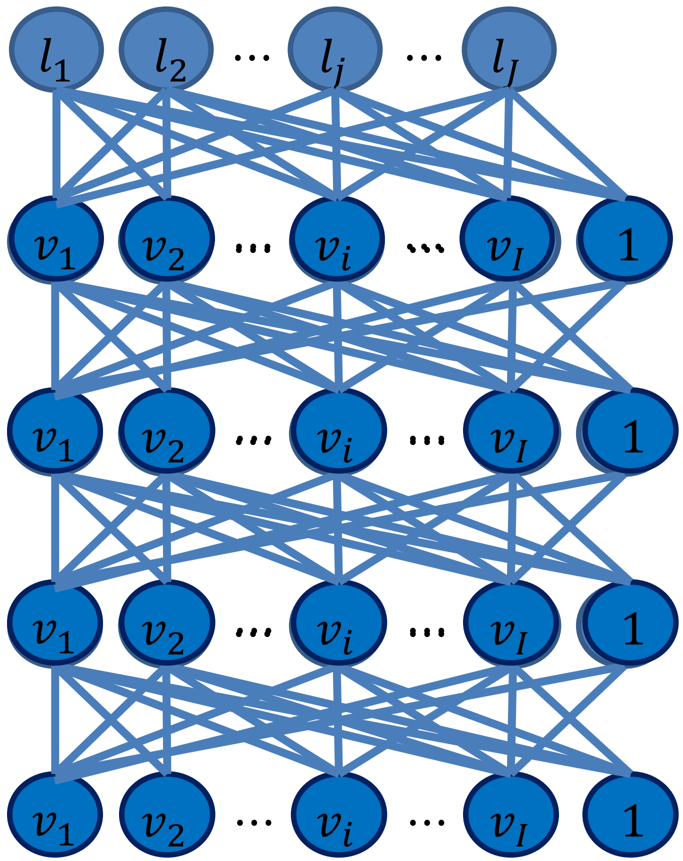
Microsoft Research

Fine Tuning DNN after pre-training: Regularization view

- Information in the pre-trained weights comes from modeling the distribution of input vectors in an "unsupervised" manner.
- The input vectors generally contain a lot more information than the labels.
- The precious information in the labels is only used for the final fine-tuning.
- The fine-tuning only modifies the features slightly to get the category boundaries right. No need to discover "features".
- **Hence less prone to overfit (unlike the old neural nets with typically random weight initialization)**
- This type of backpropagation works well even if most of the training data is unlabeled.
- The unlabeled data is still very useful for discovering good features.

Microsoft Research

DNN with class posteriors (not DBN)



- As stacked RBMs
- Pre-train each layer from bottom up by considering each pair of layers as an RBM.
- Transform the output of the last hidden layer into a multinomial distribution using the softmax operation

$$\mathrm{p}(l = k|\mathbf{h}; \theta) = \frac{exp\left(\sum_{i=1}^{H} \lambda_{ik} h_i + a_k\right)}{Z(\boldsymbol{h})}$$

- Why? Needed for (ASR) sequence recognition (not needed for static or frame-level recognition)
- For ASR: Use GMM-HMM forced alignment to get the label for the final layer when using frame-level training.
- Jointly fine-tune all layers using back-propagation algorithm.

Microsoft Research

Theoretical Insights of DBN

1. Restricted Boltzmann Machine (RBM) as the building block of DBN
2. RBM can be viewed as infinitely deep directed Bayesian/Belief network with tied weights over layers
3. Complementary prior (Hinton et. al. 2006)
4. Regularization vs. **optimization**
5. Generative vs. discriminative
6. Theory is still weak

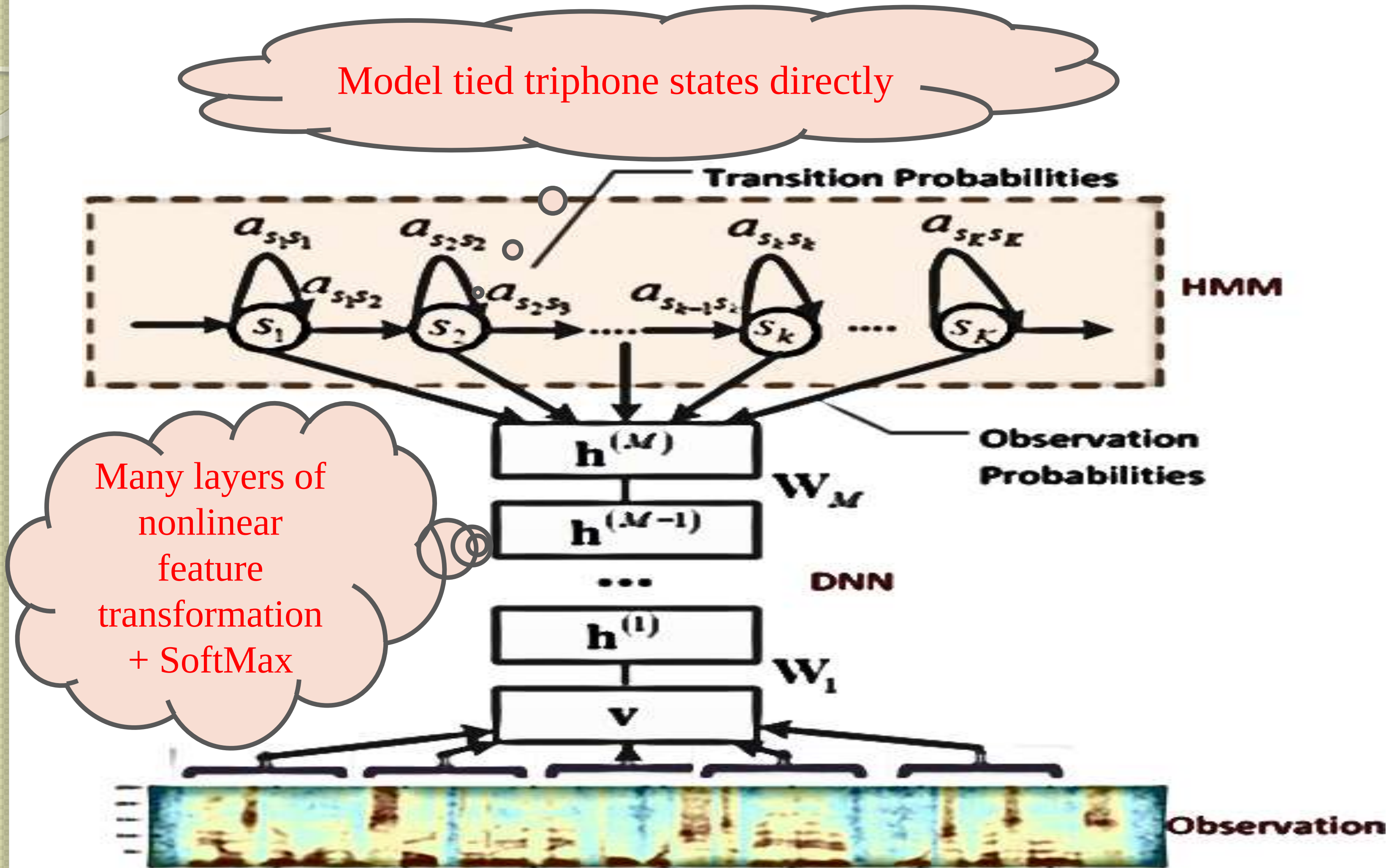
The current wisdom on unsupervised pre-training

- Pre-training achieves two things:
  - It makes optimization easier.
  - It reduces overfitting.
- We now know more about how to initialize weights sensibly by hand.
  - So unsupervised pre-training is not required to make the optimization work.
- Unsupervised pre-training is still very effective at preventing over-fitting when labeled data is scarce.
  - It is not needed when labeled data is abundant.

Hinton: ICASSP-2013

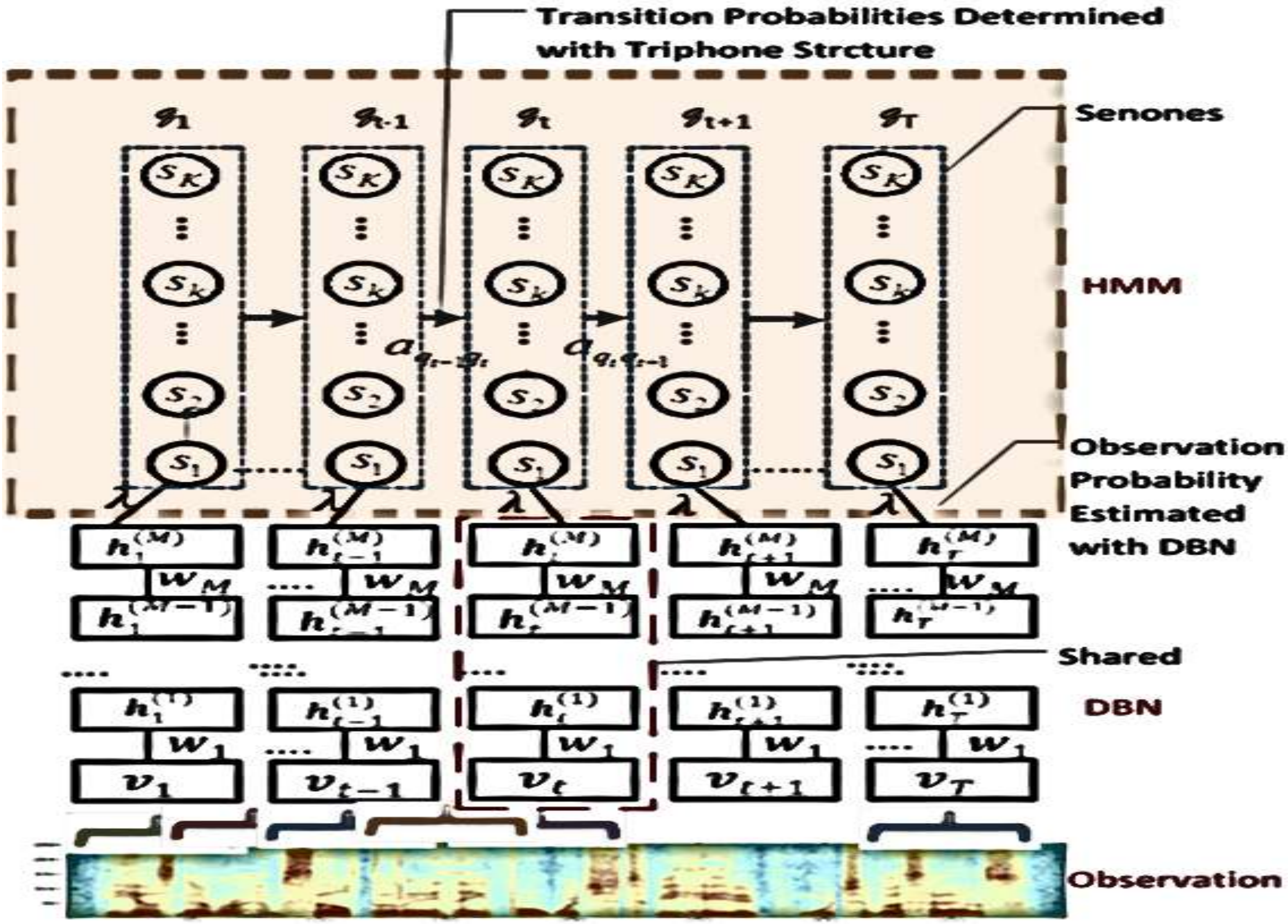
DNN-HMM

(replacing GMM only; longer MFCC/filter-back windows w. no transformation)



Microsoft Research

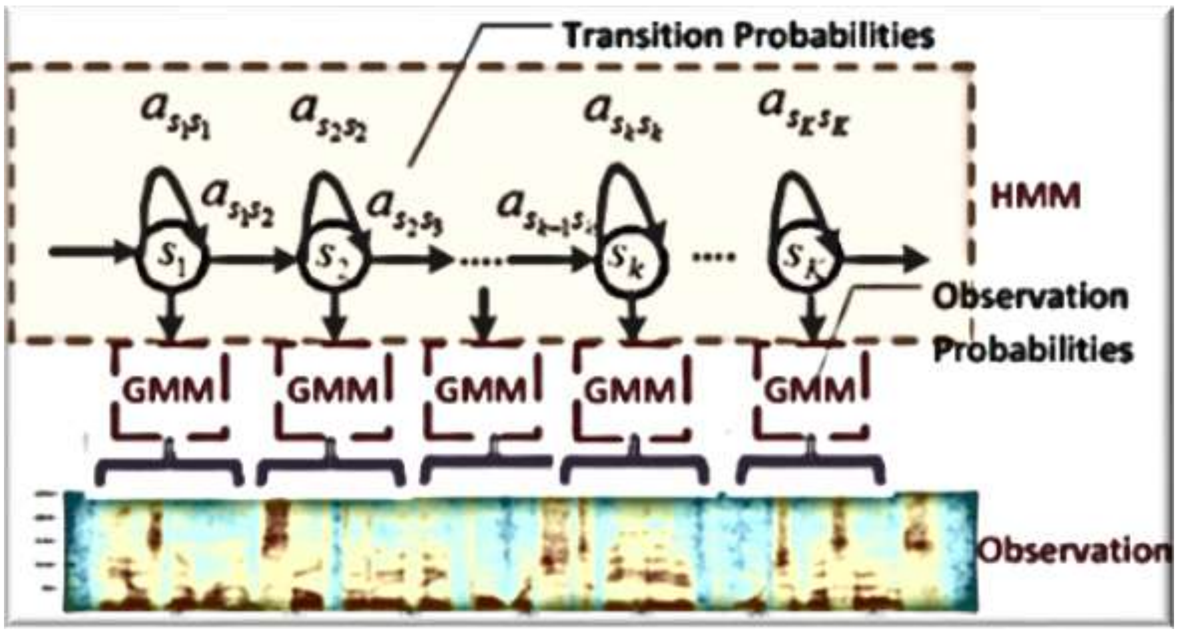
CD-DNN-HMM: Architecture



Microsoft Research

(Shallow) GMM-HMM

- Model frames of acoustic data with two stochastic processes:
  - A hidden Markov process to model state transition
  - A Gaussian mixture model to generate observations
- Train with maximum likelihood criterion using EM followed by discriminative training (e.g. MPE)



Voice Search with DNN-HMM

- First attempt in using deep models for large vocabulary speech recognition (summer 2010)
- Published in ICASSP-2011 & 2012 Special issue of T-ASLP:

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

***Abstract*—We propose a novel context-dependent (CD) model for**

fields (CRFs) [18]–[20], hidden CRFs [21], [22], and segmental

Microsoft Research

MSR Key Innovations (2009-2013)

- Scale the success to large industrial speech tasks
  - Grew output neurons from context-independent phones (100-200) to context-dependent ones (9k-32k)
  - Motivated initially by saving huge MSFT investment in huge speech decoder software infrastructure (e.g. Entropic acquisition)
  - Extremely fast decoder
  - Developed novel deep learning architectures & techniques: DCN/DSN, tensor-DSN, kernel-DCN, tensor-DNN, etc.
- Engineering for large systems:
  - Expertise in DNN **and** speech recognition
  - Close collaboration among MSRR, MSRA, & speech product teams

(Deng, Yu, Seide, Gang Li, Jinyu Li, Jui-Ting Huang, Yifan Gong, etc.)

Microsoft Research

Some Recent News by Reporters

- DNN Research Improves Bing Voice Search (very fast decoder)
- How technology can bridge language gaps: Speech-to-speech translation promises to help connect our world
- Scientists See Promise in Deep-Learning Programs (NYT: speech to speech)
- Microsoft Research shows a promising new breakthrough in speech translation technology
- Bing Makes Voice Recognition on Windows Phone More Accurate and Twice as Fast
- Microsoft revs speedier, smarter speech recognition for phones

Outline

(Including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

PART II: Deeper Substance of DL

---Technical introduction: RBM, DBN, DNN, DNN-HMM, CNN, RNN

---Examples of incorporating domain knowledge (about speech) into DL architectures

1. Hidden/articulatory Speech dynamics into RNN
2. Speech invariance/class-discrim.into deep-CNN

---A few new, promising DL architectures

Outline

~~PART I: Basics of Deep Learning (DL)~~

(including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

PART II: Deeper Substance of DL

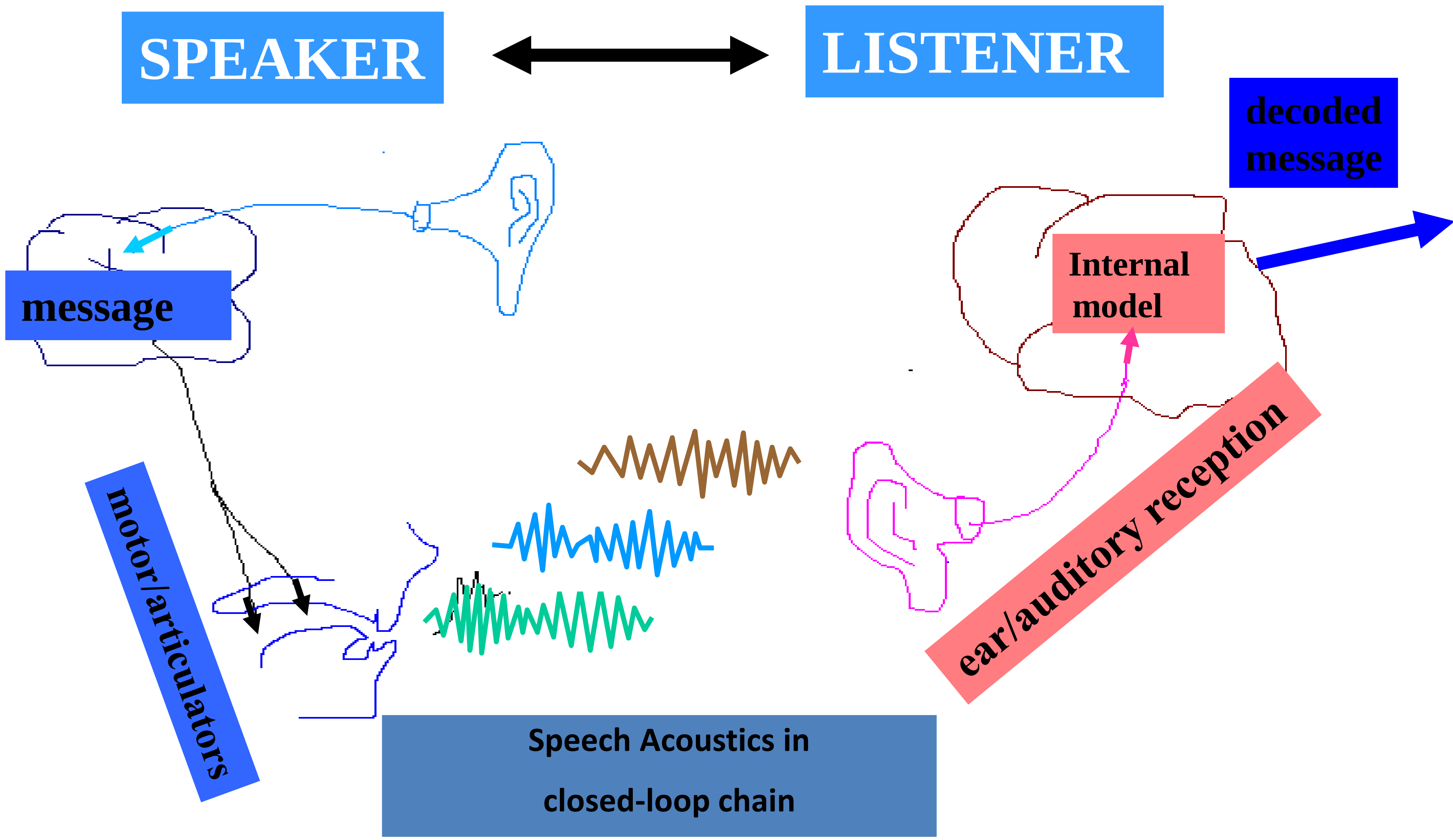
---Example 1: incorporating domain knowledge:

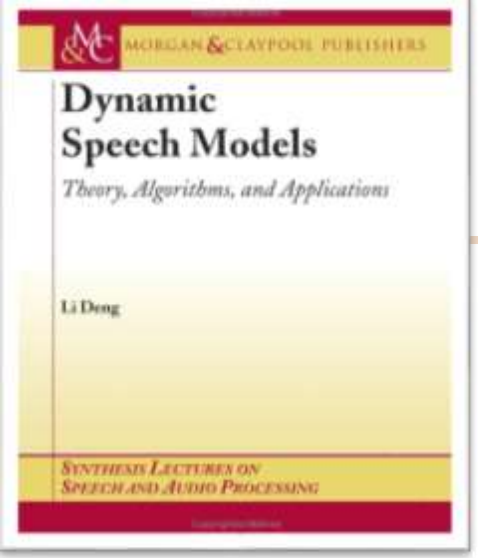
Hidden/Deep Dynamics in Human Speech

Deep/Dynamic Models are Natural for Speech

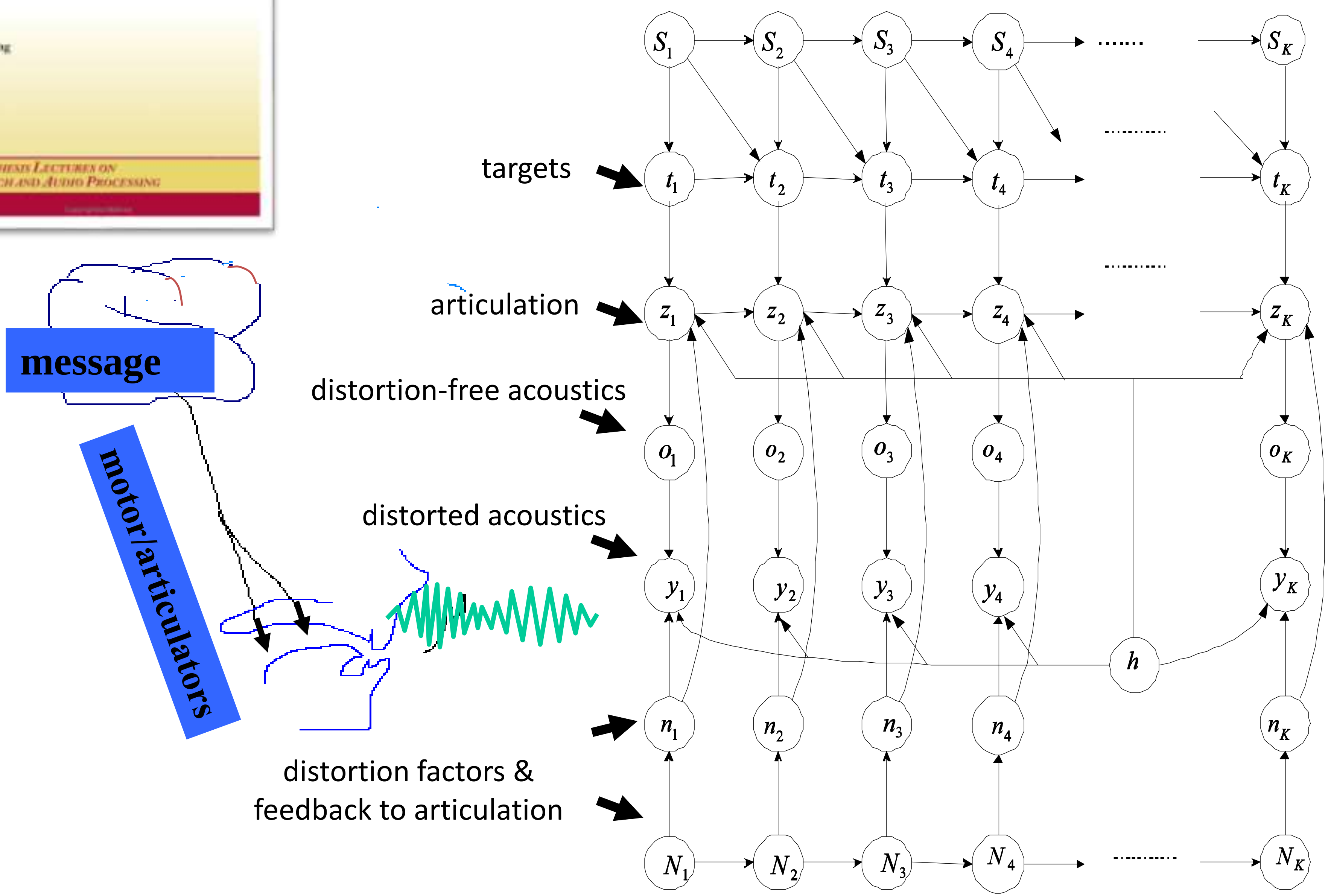
- **Hierarchical structure in human speech generation**
  - **Global concept/semantics formation**
  - **Word sequence formation / prosodic planning**
  - **Phonological encoding (phones, distinctive features)**
  - **Phonetic encoding (motor commands, articulatory targets)**
  - **Articulatory dynamics**
  - **Acoustic dynamics (clean speech)**
  - **Distorted speech**
  - **Interactions between speakers and listener/machine**
- Hierarchical structure in human speech perception
  - Cochlear nonlinear spectral analysis
  - Attribute/phonological-feature detection at higher level(s)
  - Phonemic and syllabic detection at still higher level(s)
  - Word and sequence detection
  - Syntactic analysis and semantic understanding at deeper auditory cortex

Production & Perception: Closed-Loop Chain





(Deep) Dynamic Bayesian Net



Microsoft Research

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

ABSTRACT

An important general model for discrete-time signal processing is the switching state space (SSS) model, which generalizes the hidden Markov model and the Gaussian state space model. Inference and parameter estimation in this model are known to be computationally intractable. This paper presents a powerful new approximation to the SSS model. The approximation is based on a variational technique that preserves the multimodal nature of the continuous state posterior distribution. Furthermore, by incorporating a windowing technique, the resulting EM algorithm has complexity that is just linear in the length of the time series. An alternative Viterbi decoding with frame-based likelihood is also presented which is crucial for the speech application that originally motivates this work. Our experiments focus on demonstrating the effectiveness of the algorithm by extensive simulations. A typical example in speech processing is also included to show the potential of this approach for practical applications.

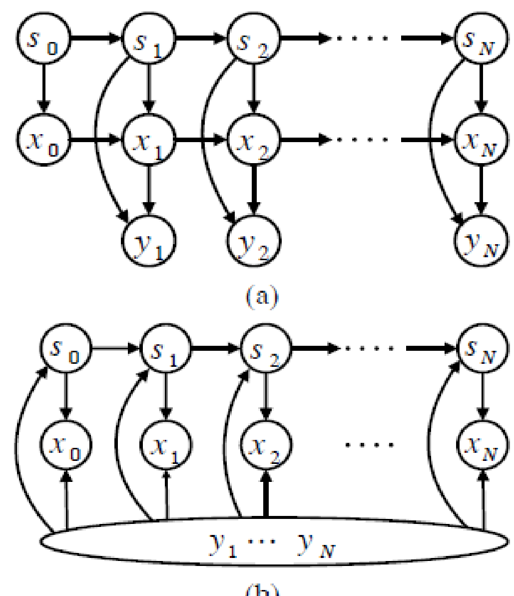


Fig. 1. The model (a) and the variational posterior (b) represented as Bayesian networks.

Microsoft Research

Generative Modeling

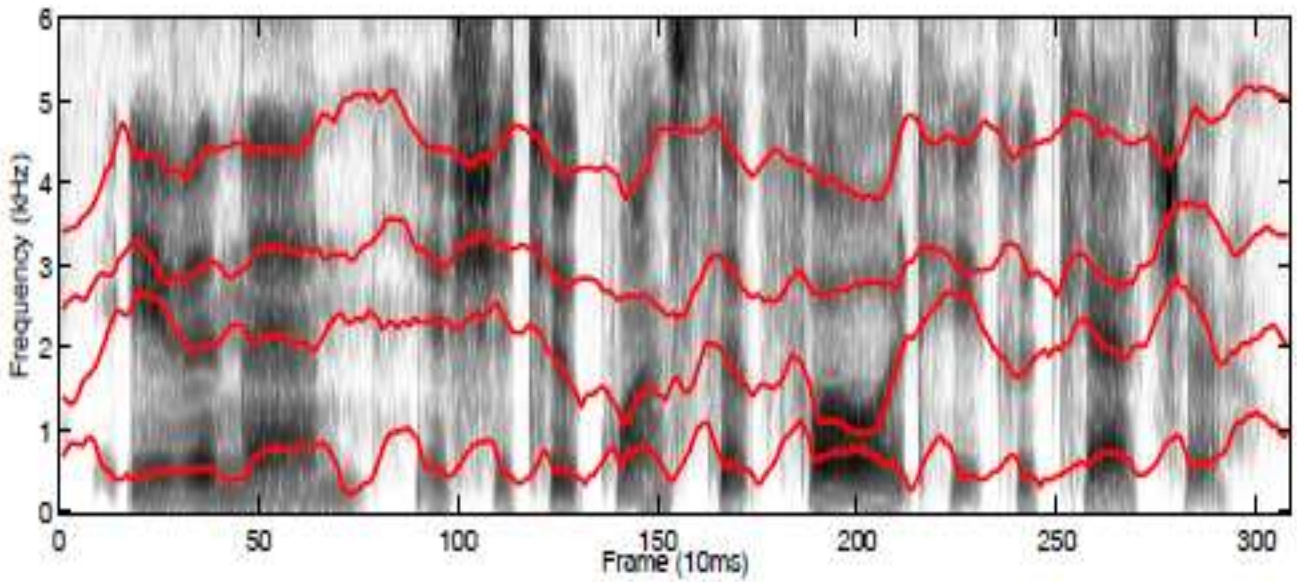


Fig. 5. Tracking VTRs for a speech sentence.

Structured Speech Modeling

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

Abstract—**Modeling dynamic structure of speech is a novel paradigm in speech recognition research within the genera-**

make it indistinguishable with human–human verbal interaction, at present, when users interact with any existing speech recog-

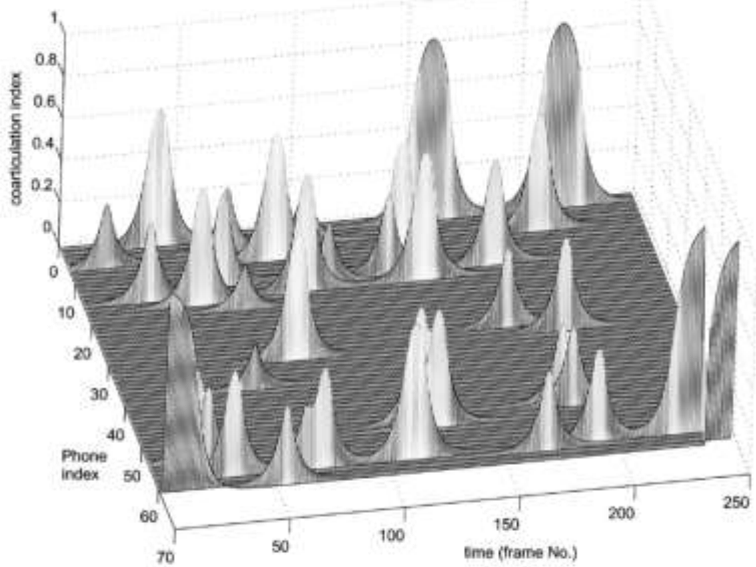
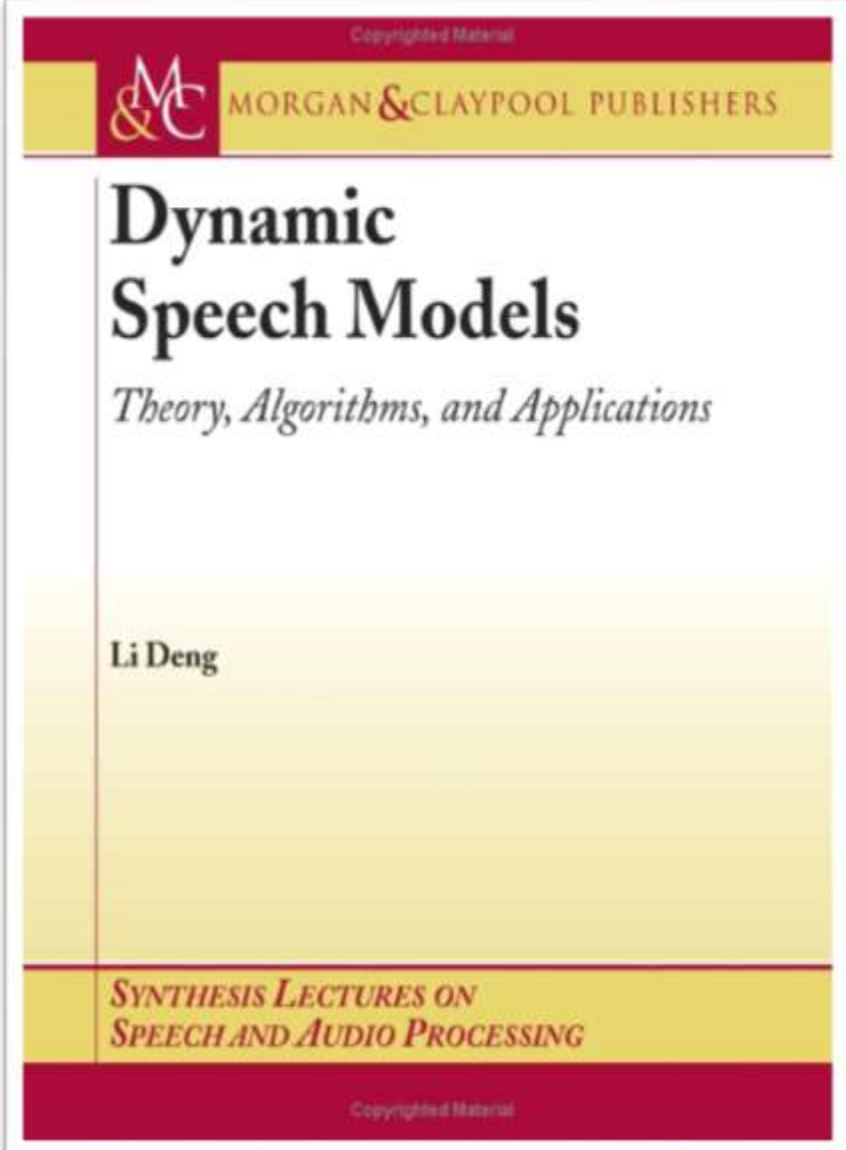


Fig. 1. Illustration of time-varying coarticulatory vectors a_k's for a TIMIT utterance. See text for detailed explanations.

Method	PER
CD-HMM [26]	27.3%
Augmented conditional Random Fields [26]	26.6%
Randomly initialized recurrent Neural Nets [27]	26.1%
Bayesian Triphone GMM-HMM [28]	25.6%
Monophone HTMs [29]	24.8%
Heterogeneous Classifiers [30]	24.4%
Monophone randomly initialized DNNs (6 layers)[13]	23.4%
Monophone DBN-DNNs (6 layers) [13]	22.4%
Monophone DBN-DNNs with MMI training [31]	22.1%
Triphone GMM-HMMs discriminatively trained w/ BMMI [32]	21.7%
Monophone DBN-DNNs on fbank (8 layers) [13]	20.7%
Monophone mcRBM-DBN-DNNs on fbank (5 layers) [33]	20.5%
Monophone convolutional DNNs on fbank (3 layers) [34]	20.0%

(Hidden) Dynamic Models

- Many types of dynamic models since 90's
- Good survey article on earlier work (Ostendorf et al. 1996)
- Hidden Dynamic Models (HDM/HTM) since late 90's
- This is "deep" generative model with >2 layers
- More recent work: book 2006
- Pros and cons of different models
- All intended to create more realistic speech models "deeper" than HMM for speech recognition
- But with different assumptions on speech dynamics
- How to embed such dynamic properties into the DNN framework?



Microsoft Research

DBN (Deep) vs. DBN* (Dynamic)

- DBN-DNN (2009-2012) vs. HDM/HTM (1990's-2006)
- Distributed vs. local representations
- Massive vs. parsimonious parameters
- Product of experts vs. mixture of experts
- Generative-discriminative hybrid vs. generative models
- Longer windows vs. shorter windows

- **A neat way of "pre-training" RNN by HDM and then "fine-tuning" RNN by backprop (non-trivial gradient derivation and computation)**

Building Dynamics into Deep Recurrent Models

- (Deep) recurrent neural networks for ASR: both **acoustic** and language modeling
  - generic temporal dependency
  - lack of constraints provided by hidden speech dynamics
  - Information redundancy & inconsistency: long windows for each “frame” introducing undesirable “noise”
  - Need to go beyond unconstrained temporal dependence and ESN (while easier to learn)
- An active and exciting research area to work on

Outline

PART I: Basics of Deep Learning (DL)

(including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

PART II: Deeper Substance of DL

---Example 2: incorporating domain knowledge:

Speech invariance/variability vs.

Phonetic discrimination in Conv. NN

A Deep Convolutional Neural Net Using Heterogeneous Pooling to Tradeoff Acoustic Invariance w. Phonetic Distinction

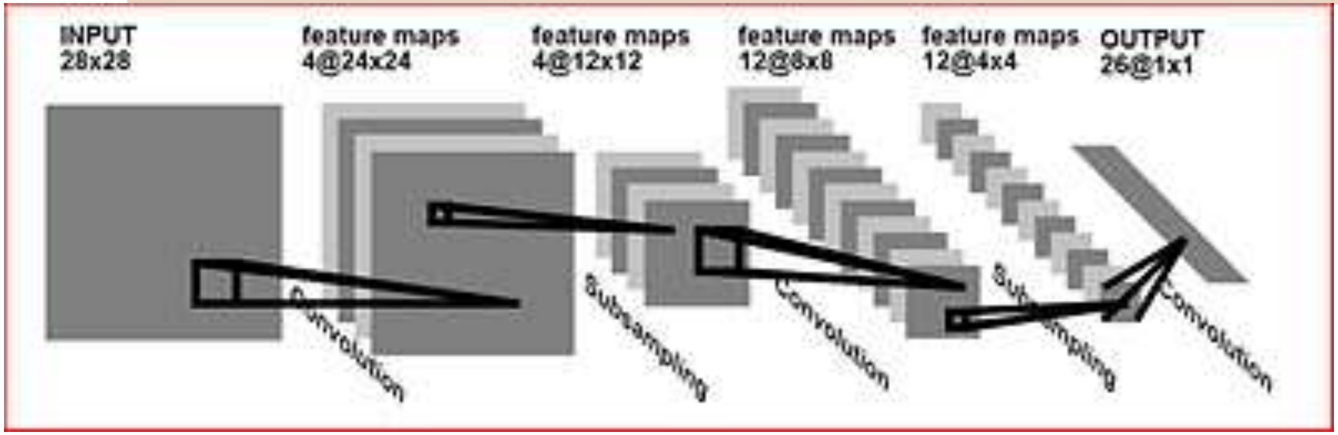
Li Deng, Ossama Abdel-Hamid, and Dong Yu

Microsoft Research, Redmond

York University, Toronto

ICASSP, May 28, 2013

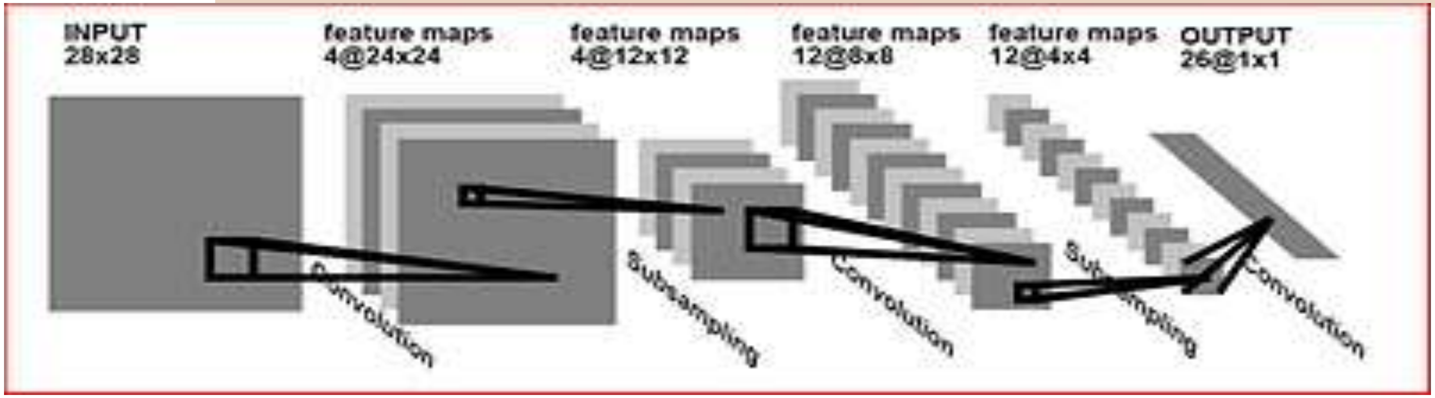
Background: Convolutional Nets (CNN)



LeCun et al. 90's

- **Convolution layer** (w. tying weights): a.k.s. "time/spatial"-invariant FIR filter
- Gives maps of replicated features; neural activities "equivariant" to translation
- **Pooling layer** (max of neighboring units in conv layer): Data reduction & some degree of **invariance.**
- 2D deep-CNN: State of the art in object recognition (Krizhevsky et al., 2012; LeCun et al.; Ciresan et al.)

Background: Convolutional Nets (CNN)



- Difficulties of CNN:
  - 2D Images: Information lost about the precise positions of parts → object confusion
  - 2D Speech spectrogram: spectral-temporal information lost about phonetic distinction
  - E.g. 1-D CNN along freq axis (Abdel-Hamid et al., 2012): (TDNN & TF-trajectory CNN)
    local weight sharing + max pooling over a range→ invariance to freq shift
    (VTL normalization)

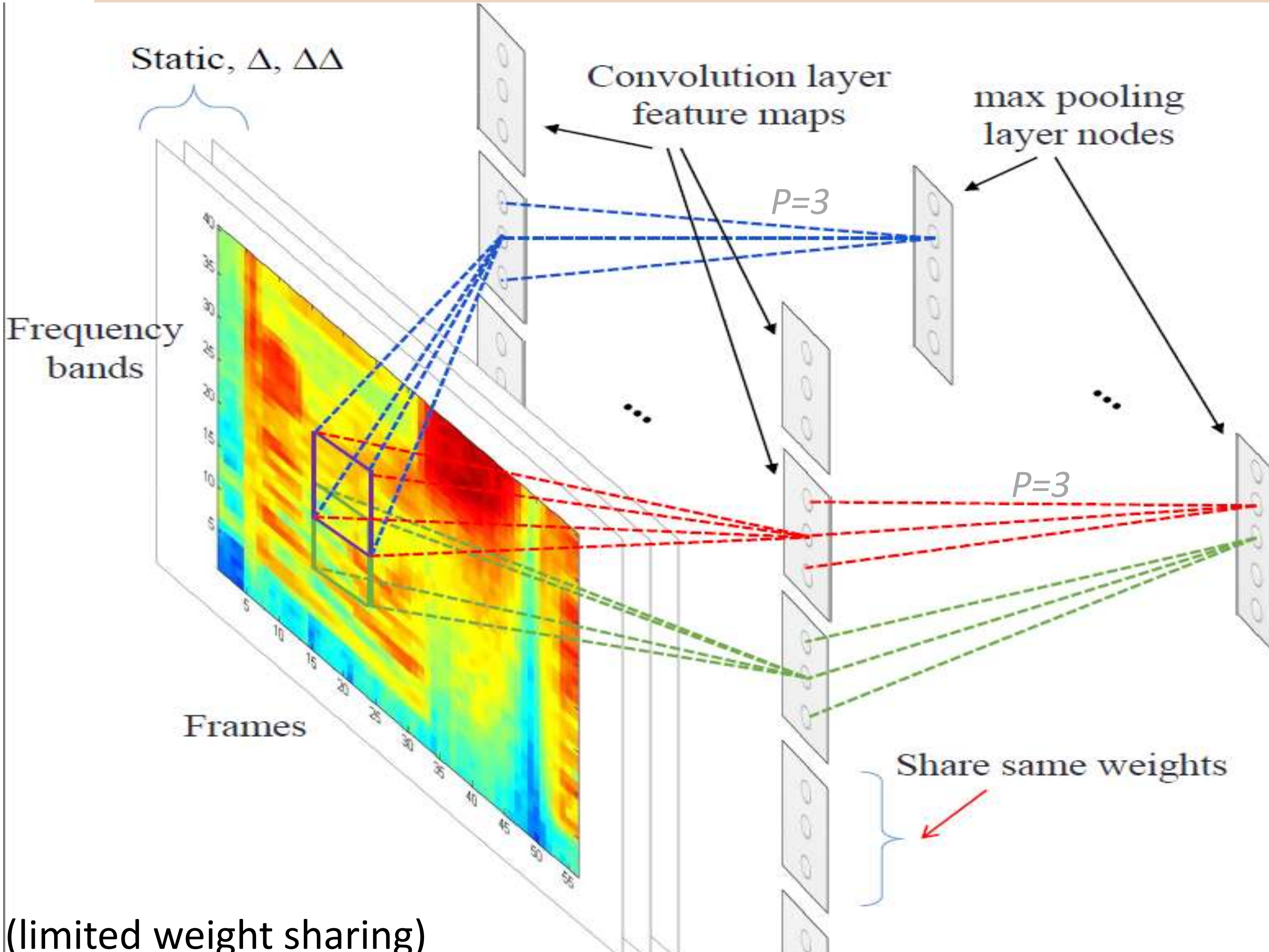
  But if freq range too small → not enough VTL normalization (acoustic invariance)
  too large → formant patterns of a sound shift → phone confusion
- Solutions for image recognition: (tried some for speech, no clear success)
  - Transforming autoencoder (Hinton et al., 2011)
  - Tiled CNN (Le et al., 2012)
  - Deconvolutional nets (Zeiler et al., 2011)
- A **good solution for speech recognition** is surprisingly simple

Main Ideas of This Paper

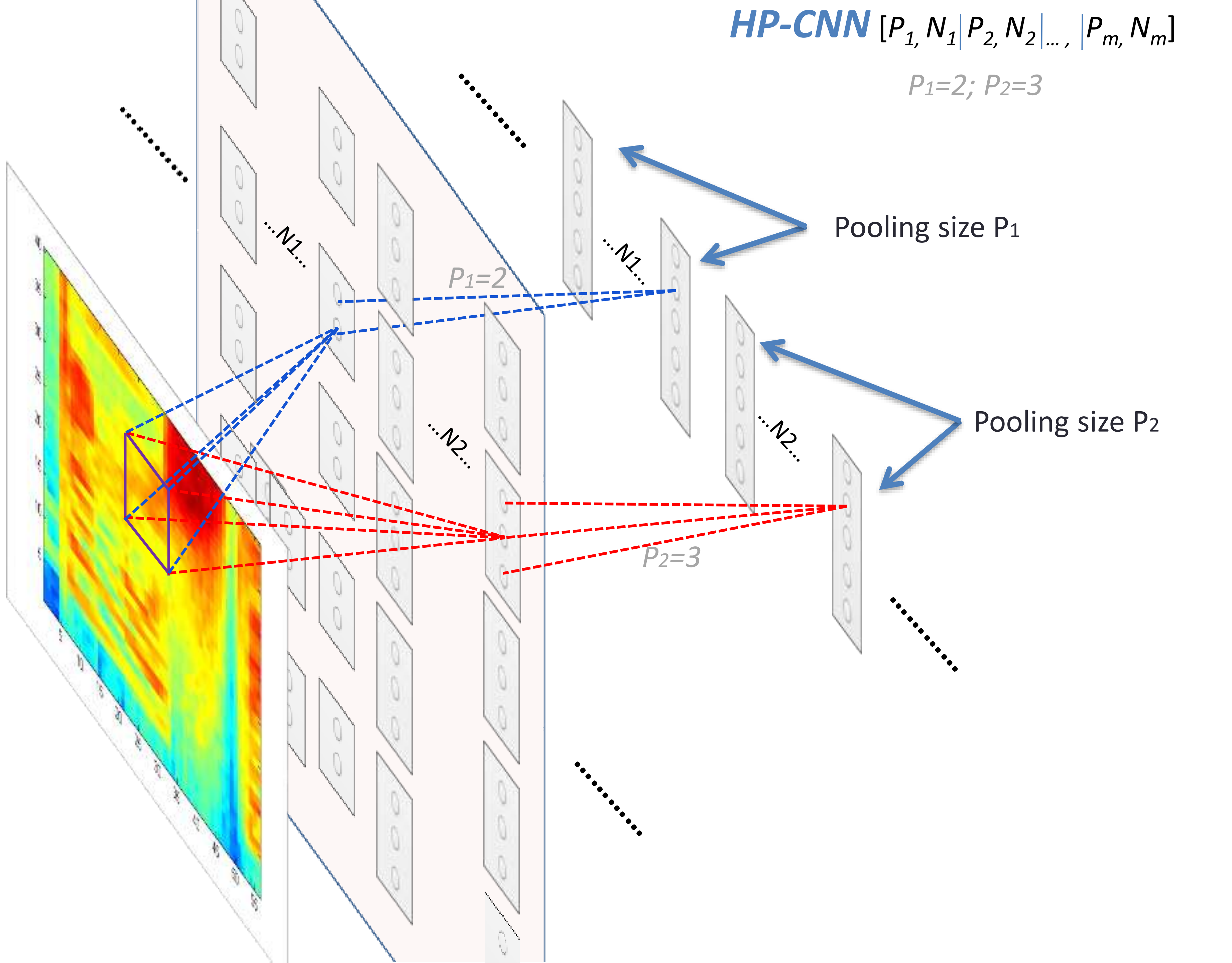
- Bring "confusion" into designing CNN intended for "invariance"
- Exploit the knowledge of how increasing the degree of invariance (to shift along frequency-axis) may reduce phonetic discrimination
- **(Kai Yu this morning: Spatial Pyramid Matching for vision)**
- Examine/predict how the pooling size (i.e. range of freq-shift invariance) affects phonetic classification errors
  – Theoretic guidance possible; e.g.
  – Phonetic reduction (in casual, conversation speech) shrinks formant space →tradeoff towards "distinction" from "invariance"→ smaller pooling size
- Use of many feature maps (afforded by CNN weight tying)
- Different pooling sizes (heterogeneous pooling) for different feature maps
  – Design and use **a distribution of pooling sizes** and randomly sample it.
  – Special case: use a fixed pooling size, optimized by validation or predicted by acoustic-phonetic "theory" (consistent for TIMIT; not as good as HP)

CNN with a Fixed Pooling Size

*(a special case of **HP-CNN** w. P=3)*



(limited weight sharing)



HP-CNN [P1, N1 | P2, N2 | ..., | Pm, Nm]
P1=2; P2=3
Pooling size P1
Pooling size P2
P1=2
P2=3
N1
N2

Regularizing HP-CNN with "Dropout"

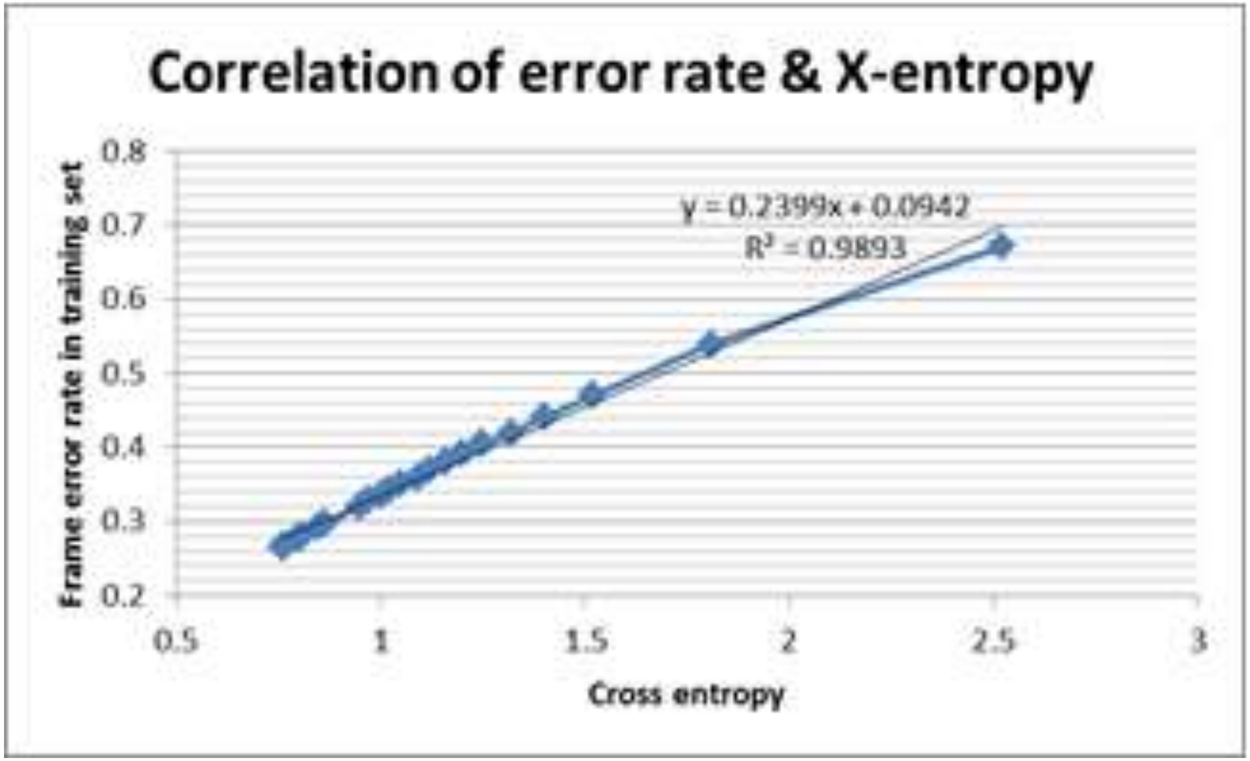
- A variant of the Dropout method for DNN (Hinton et al., 2012)
- Dropout in both conv and pooling layers of CNN is helpful, in addition to fully-connected DNN layers
- Dropout in the input layer (filterbanks) is not helpful
- In TIMIT, for CNN w. N=100 feature maps, and DNN hid=2000, the best dropout rate=0.2
- With dropout rate=0.5 & DNN hid=5000, error rate increases

Standard TIMIT Task: Core Testset Results

Systems	Phone Error Rate
DNN (fully-connected 5 layers)	22.3%
CNN-DNN; P=1 (2 CNN & 3 DNN layers)	**21.8%**
CNN-DNN; P=12	20.8%
CNN-DNN; P=6 (fixed P, optimal)	**20.4%**
CNN-DNN; P=6 (add dropout)	19.9%
CNN-DNN; P=1:m (HP, m=12)	**19.3%**
CNN-DNN; above (add dropout)	**18.7%**

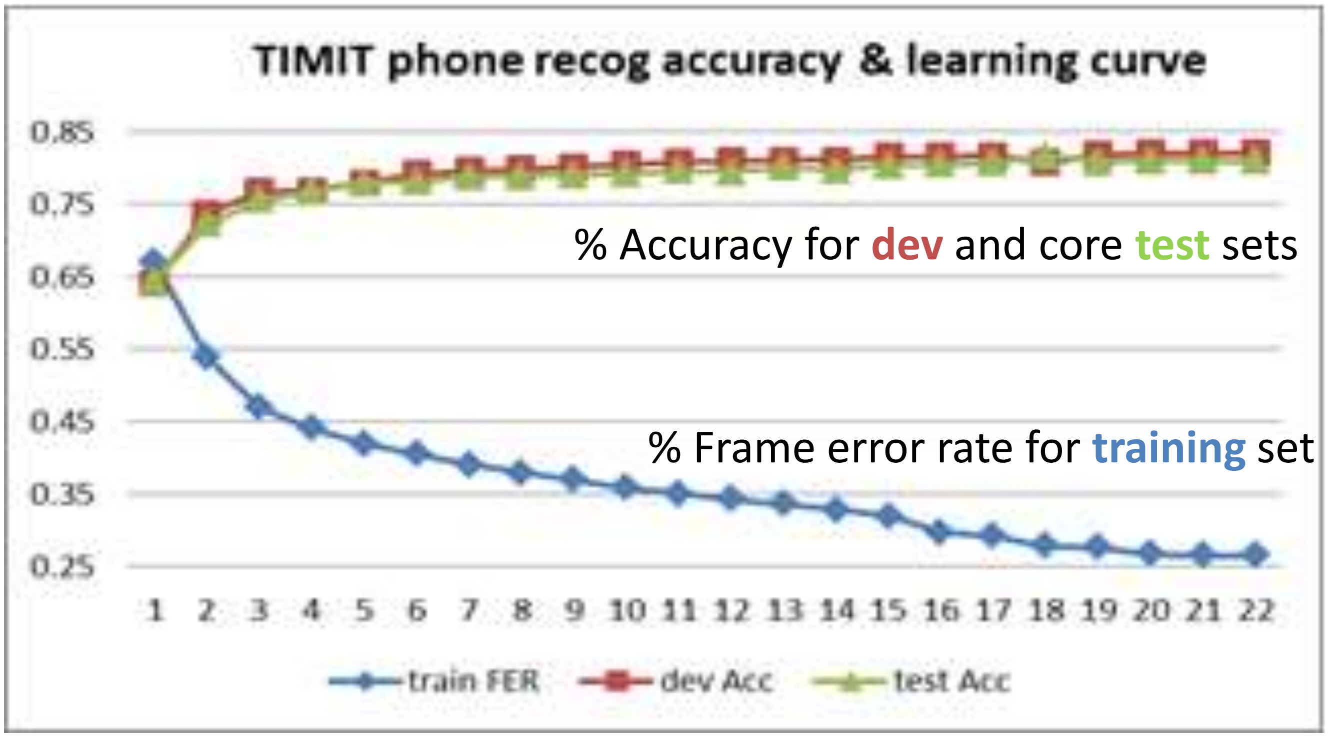
CNN-DNN; P=1 → equivariance: 21.8% > **20.4%** (invariance w. fixed, optimal pooling size=6)
CNN-DNN; P=1:12 → Heterogeneous pooling: **19.3% <** 20.4%
Dropout is always helpful (thanks Geoff!): **18.7% <** 19.3% **;** **19.9% <** 20.4%
18.7% WAS the record low error rate on this standard task (until this morning by LSTM-RNN)

Training Criterion: Cross-Entropy



Effects of Training Epochs (Time)

- Each training epoch (1.12M frames in TIMIT)
  → 2 hrs of CPU time on my PC (no GPU)



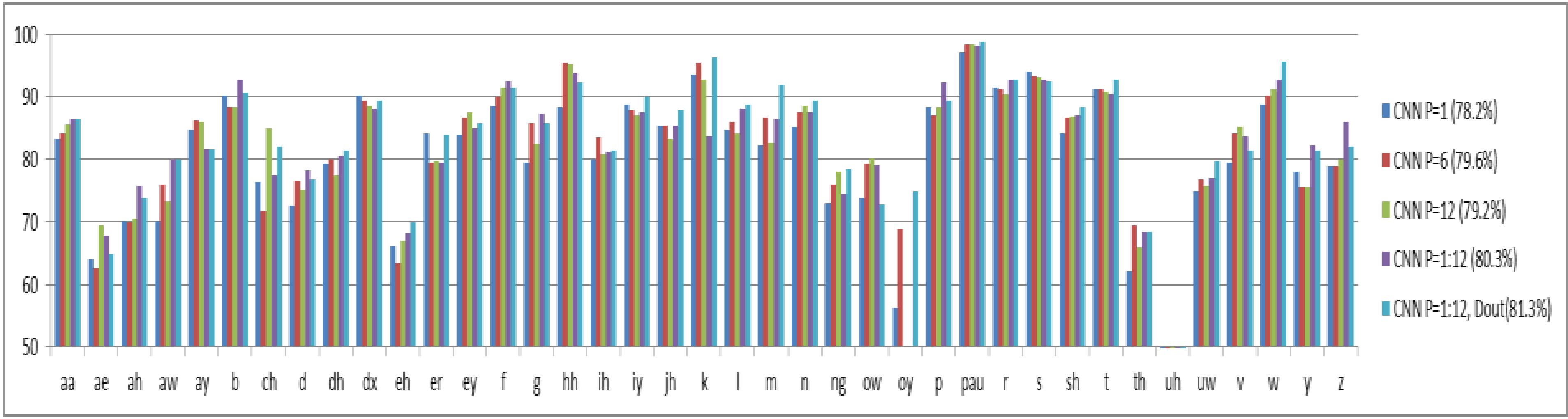
Confusion Matrix (Hresults)

```
------------------------ Overall Results -------------------------
SENT: %Correct=1.04 [H=2, S=190, N=192]
WORD: %Corr=83.98, Acc=81.33 [H=6158, D=338, S=837, I=194, N=7333]
------------------------ Confusion Matrix -------------------------
```

	aa	ae	ah	aw	ay	b	ch	d	dh	dx	eh	er	ey	f	g	hh	ih	iy	jh	k	l	m	n	ng	ow	oy	p	pau	r	s	sh	t	th	uh	uw	v	w	y	z	Del [%c / %e]
aa	190	0	10	2	7	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	3	1	0	0	1	1	0	0	2	0	0	0	0	0	0	0	0	0	0	8 [86.4/0.4]
ae	3	63	2	3	2	0	0	0	0	0	15	0	3	0	0	0	2	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	8 [64.9/0.5]
ah	8	0	218	1	2	0	0	1	0	0	8	8	0	0	0	0	35	0	0	0	0	0	0	0	7	0	0	1	1	0	0	0	0	2	3	0	0	0	0	26 [73.9/1.1]
aw	0	2	0	24	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0 [80.0/0.1]
ay	5	0	4	1	71	0	0	0	0	0	1	0	3	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2 [81.6/0.2]
b	0	0	0	0	0	115	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	2	0	0	5 [90.6/0.2]
ch	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	1 [82.1/0.1]
d	0	0	0	0	0	0	0	79	5	2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	1	0	0	0	10 [76.7/0.3]
dh	0	0	0	0	0	1	0	4	100	3	0	0	0	0	0	1	0	0	0	1	3	1	1	0	0	0	2	0	0	0	0	1	1	0	0	2	0	0	2	5 [81.3/0.3]
dx	0	0	0	0	0	0	0	5	1	76	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5 [89.4/0.1]
eh	1	8	15	0	1	0	0	0	0	0	127	0	2	0	0	0	24	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	7 [69.8/0.8]
er	0	0	3	0	1	0	0	0	0	0	1	192	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	1	0	0	0	0	5 [83.8/0.5]
ey	0	1	0	0	1	0	0	0	0	0	2	0	97	0	0	0	6	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1 [85.8/0.2]
f	0	0	0	0	0	1	0	0	0	0	0	0	0	119	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	1	0	0	2	0	0	3	0	0	0	1 [91.5/0.2]
g	0	0	1	0	0	1	0	3	0	0	0	0	0	0	54	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3 [85.7/0.1]
hh	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	59	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	13 [92.2/0.1]
ih	0	1	32	0	0	0	0	0	0	0	14	10	5	0	0	0	448	23	0	0	0	0	0	0	4	0	0	3	0	1	1	0	0	2	5	1	0	1	0	29 [81.3/1.4]
iy	0	0	1	0	0	0	0	0	0	0	0	0	6	0	0	0	11	214	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	3	0	5 [89.9/0.3]
jh	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1 [87.8/0.1]
k	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	151	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	8 [96.2/0.1]
l	3	0	1	2	0	0	1	0	3	0	0	0	0	1	0	0	0	0	0	0	244	2	0	0	9	0	2	0	3	0	0	1	0	0	0	1	2	0	0	16 [88.7/0.4]
m	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	179	8	0	0	0	1	1	0	0	0	0	0	0	0	0	2	0	0	9 [91.8/0.2]
n	0	0	1	0	0	2	0	2	1	9	0	1	1	0	0	0	2	0	0	0	0	8	324	7	0	0	0	3	1	1	0	0	0	0	0	0	0	0	0	22 [89.3/0.5]
ng	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3	5	40	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1 [78.4/0.2]
ow	10	0	2	2	0	0	0	0	0	0	0	1	1	0	0	0	2	0	0	0	4	0	0	0	64	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1 [72.7/0.3]
oy	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	1	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [75.0/0.1]
p	0	0	0	0	0	5	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	117	0	0	0	0	4	0	0	0	1	0	0	0	5 [89.3/0.2]
pau	0	0	4	0	0	0	0	0	0	3	4	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	1486	0	1	0	1	0	0	0	2	0	0	0	73 [98.7/0.3]
r	0	0	1	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	225	0	0	1	0	0	0	2	0	0	0	27 [92.6/0.2]
s	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	294	0	1	0	0	0	0	0	0	17	3 [92.5/0.3]
sh	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	2	68	0	0	0	0	0	0	1	0	0 [88.3/0.1]
t	0	0	0	0	0	0	1	3	4	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	162	2	0	0	0	0	0	0	12 [92.6/0.2]
th	0	0	0	0	0	0	0	0	3	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	3	26	0	0	0	0	0	0	0 [68.4/0.2]
uh	0	0	3	0	0	0	0	0	0	0	1	1	0	0	0	0	6	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	11	2	0	0	0	0	2 [40.7/0.2]
uw	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	5	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	59	0	0	0	0	0 [79.7/0.2]
v	0	0	0	0	0	4	0	0	2	0	0	0	0	1	0	0	1	0	0	0	0	3	0	0	1	0	2	2	0	0	0	0	0	0	0	70	0	0	0	7 [81.4/0.2]
w	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	130	0	0	8 [95.6/0.1]
y	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	35	0	7 [81.4/0.1]
z	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	2	0	25	1	0	1	0	0	0	0	0	147	2 [82.1/0.4]
Ins	10	3	13	1	4	2	2	12	5	1	4	4	0	2	1	6	14	5	2	6	0	1	5	1	4	1	3	49	9	4	1	3	1	0	3	1	4	6	1	

Recognition Error Breakdown

- Percentage Phone Errors for Each of 39 Classes
- Comparing five different phone recognizers: Effects of HP and dropout



HP-CNN for Large Vocabulary Speech Recognition

- On a voice search task
- Dozens of hours of labeled training data
- Not yet optimized the single fixed pooling size
- No change (yet) from the TIMIT system: m=12, N=104, N_1, N_2, N_3, N_4, N_5, etc.

Systems	Word Error Rate
DNN baseline (fully-connected 5 layers)	**32.4%**
CNN-DNN; P=1 (2 CNN & 3 DNN layers)	32.0%
CNN-DNN; P=1:m (HP, m=12) same distribution as in TIMIT experiment	**30.1%**

Conclusions (of ICASSP-2013 paper)

- Effectiveness of convolution/pooling in image recognition can be ported to speech recognition
- Esp. when speech-specific properties incorporated
- Bring "confusion" into designing CNN intended for "invariance"
- Tradeoffs can be made by adjusting pooling size in CNN
- Optimizing (a single) pooling size provides desirable tradeoffs
- A much better way is to use varying pooling sizes for different feature maps (hence HP)→ record-low TIMIT error rate
- HP-CNN of this paper is limited to convolution along freq-axis
- Can be extended to spectro-temporal patches in spectrograms
- Analogy: Object parts (image) ←→ formant trajectories (speech)
- This is exciting time to integrate speech knowledge into deep learning models

Outline

(including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

PART II: Deeper Substance of DL

---Technical introduction: RBM, DBN, DNN, DNN-HMM, CNN, RNN

---Examples of incorporating domain knowledge (about speech) into DL architectures

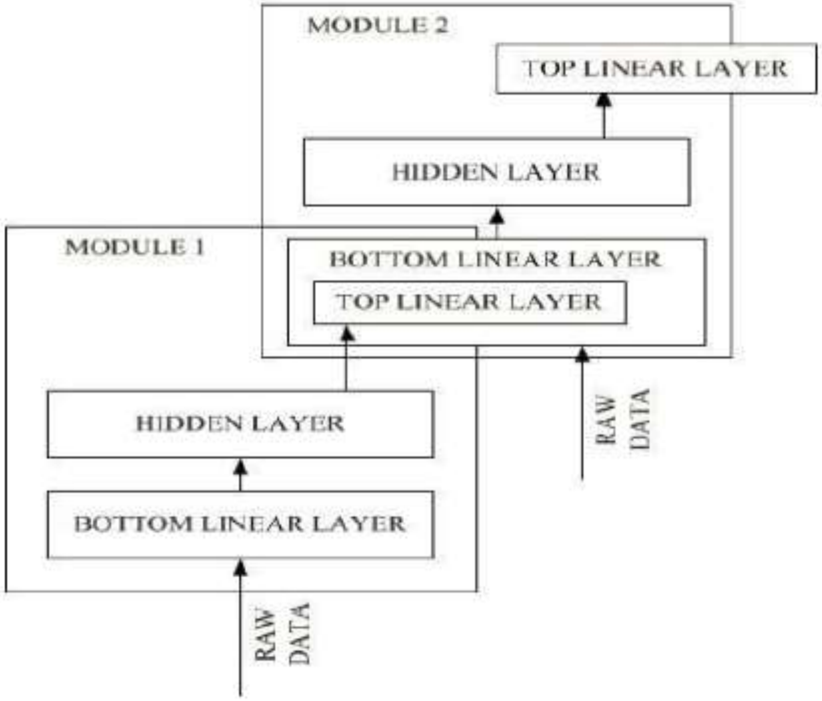
1. Hidden/articulatory Speech dynamics into RNN
2. Speech invariance/class-discrim.into deep-CNN

--- A few new, promising DL architectures

Google

Deep Convex Networks

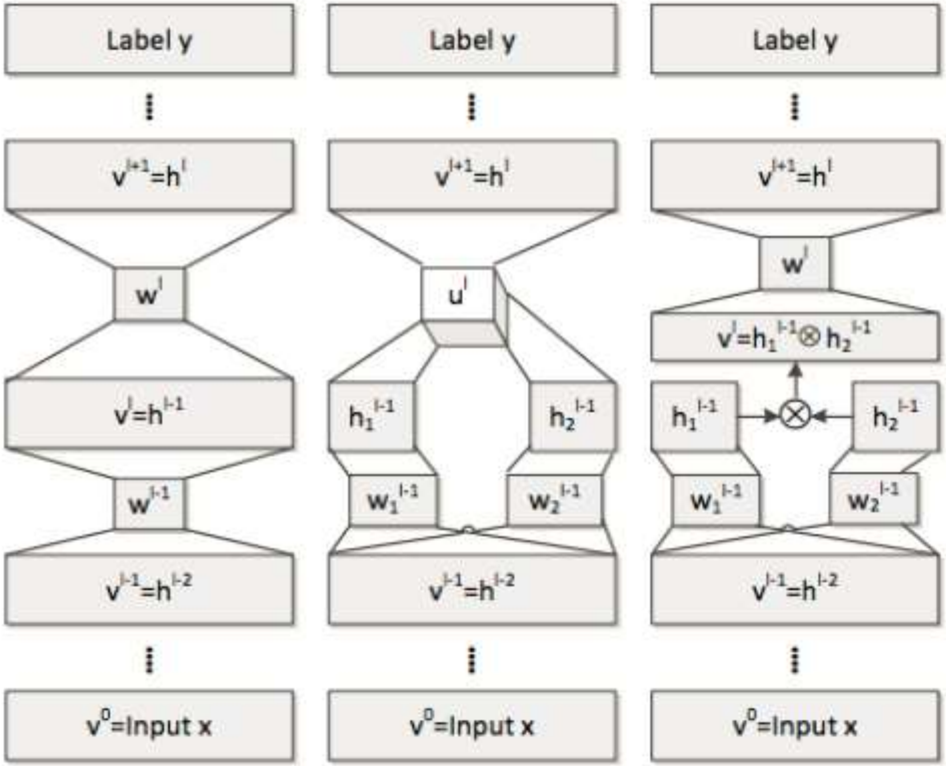
- A simple approach to build a deep model using only convex optimization techniques.
- Successfully 'convexifying' the problem is an interesting line of research.
- Very competitive and fast to train.
- So far, best performance still obtained with non-convex fine tuning and many more layers than DNNs.



Li Deng and Dong Yu. "Deep convex net: a scalable architecture for speech pattern classification." Interspeech 2011.

Deep Tensor Networks

- One example of several attempts at incorporating multiplicative nodes into deep networks.
- Very promising area of research attempting to factor out 'style' (speaker, environment) from 'content' (phonetic label) using multiplicative gating interactions.



Tensor Deep Stacking Networks

Brian Hutchinson, *Student Member, IEEE,* Li Deng, *Fellow, IEEE,* and Dong Yu, *Senior Member, IE*

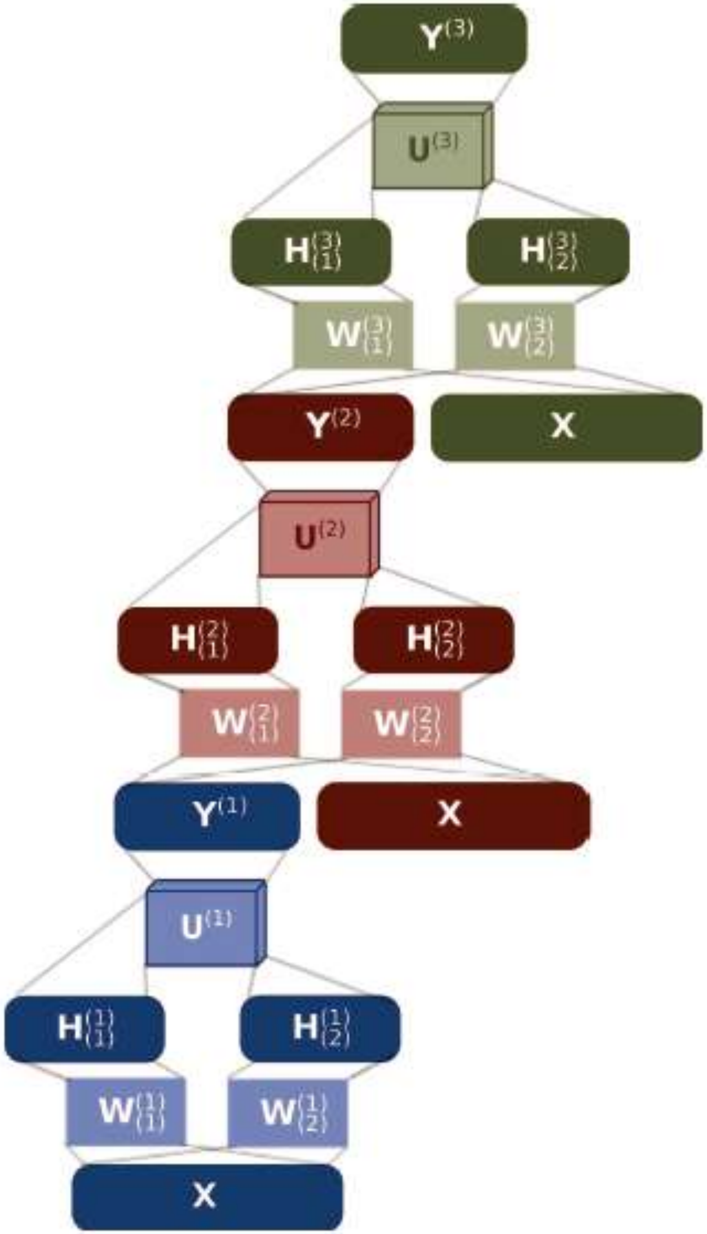
Abstract—A novel deep architecture, the Tensor Deep Stacking Network (T-DSN), is presented. The T-DSN consists of multip stacked blocks, where each block contains a bilinear mapping from two hidden layers to the output layer, using a weight tensor incorporate higher-order statistics of the hidden binary ($[0,1]$) features. A learning algorithm for the T-DSN's weight matrices a tensors is developed and described, in which the main parameter estimation burden is shifted to a convex sub-problem with a close form solution. Using an efficient and scalable parallel implementation for CPU clusters, we train sets of T-DSNs in three popul tasks in an increasing order of the data size: handwritten digit recognition using MNIST (60k), isolated state/phone classification a continuous phone recognition using TIMIT (1.1m), and isolated phone classification using WSJ0 (5.2m). Experimental results in three tasks demonstrate the effectiveness of the T-DSN and the associated learning methods in a consistent manner. In particul a sufficient depth of the T-DSN, a symmetry in the two hidden layers structure in each T-DSN block, our model parameter learni algorithm, and a softmax layer on top of T-DSN are shown to have all contributed to the low error rates observed in the experiments all three tasks.

Index Terms—Deep learning, stacking networks, tensor, bilinear models, handwriting image classification, phone classification a recognition, MNIST, TIMIT, WSJ

INTRODUCTION

ECENTLY, a deep classification architecture built

the T-DSN retains the same linear-nonlinear interl structure as DSN in building up the deep architec



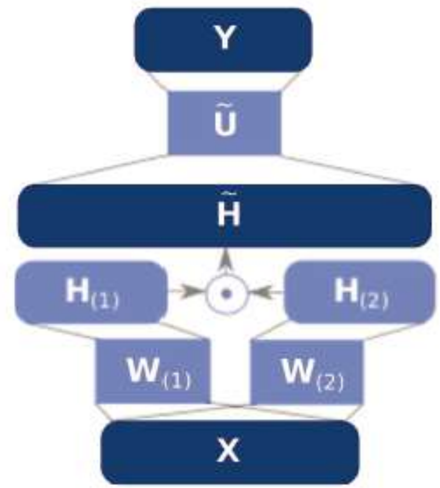


Fig. 2. Equivalent architecture to the bottom block of Fig. 1, where the tensor is unfolded into a large matrix.

Fig. 1. An example T-DSN architecture with three stacking blocks, where each block consists of three layers, and

Deep Stacking Networks for Information Retrieval

Li Deng, Xiaodong He, and Jianfeng Gao

Microsoft Research, Redmond

ICASSP, May 30, 2013

Outline

- Motivation: deep learning for Information Retrieval (IR)
  - Learning to rank
  - Semantic feature extraction for ranking
- Deep Stacking Net (DSN)
  - Basic modular architectures
  - Novel discriminative learning algorithm
- Applying DSN for IR --- learning to rank
  - Formulating IR as a classification problem
  - Special role of regularization
- Experiments
  - IR task, data sets, and features
  - Relationship between NDCG score & classification error rate
  - NDCG results on an IR task (Ads selection)

Background of IR

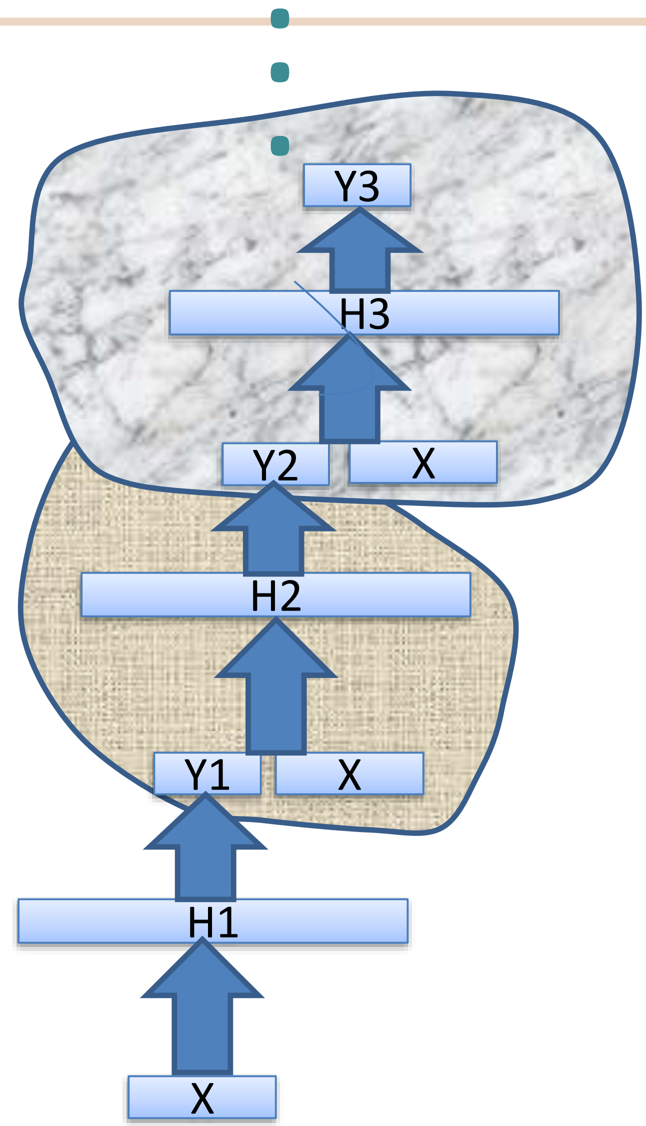
- Goal of IR: ranking text documents (D) for a query (Q)
- Common methods:
  - Lexical matching: suffers from text discrepancy btwn Q and D (e.g. vocabulary, word usage, expression style, etc.)
    - E.g., TF-IDF weighted vector space model
  - Semantic matching: to bridge lexical gaps btw Q and D
    - E.g., Latent Semantic Analysis (LSA), PLSA, LDA, etc.
  - Learning Q-D matching using clickthrough data
    - E.g., translation models, bilingual topic models etc.
  - These linear models suffer from restricted expressive power

Deep Learning for IR

- Multilayers of nonlinearities
  - Greater expressive power
  - Better able to capture semantic contents in Q and D
    - E.g., semantic hashing (Hinton et al, 2007)
  - More effective use of supervised clickthrough data
- Use of (labeled) clickthrough data for IR ranking
  - Shallow linear models: Gao et al., 2010;2011
  - Shallow nonlinear models: Burges et al., 2005;2006

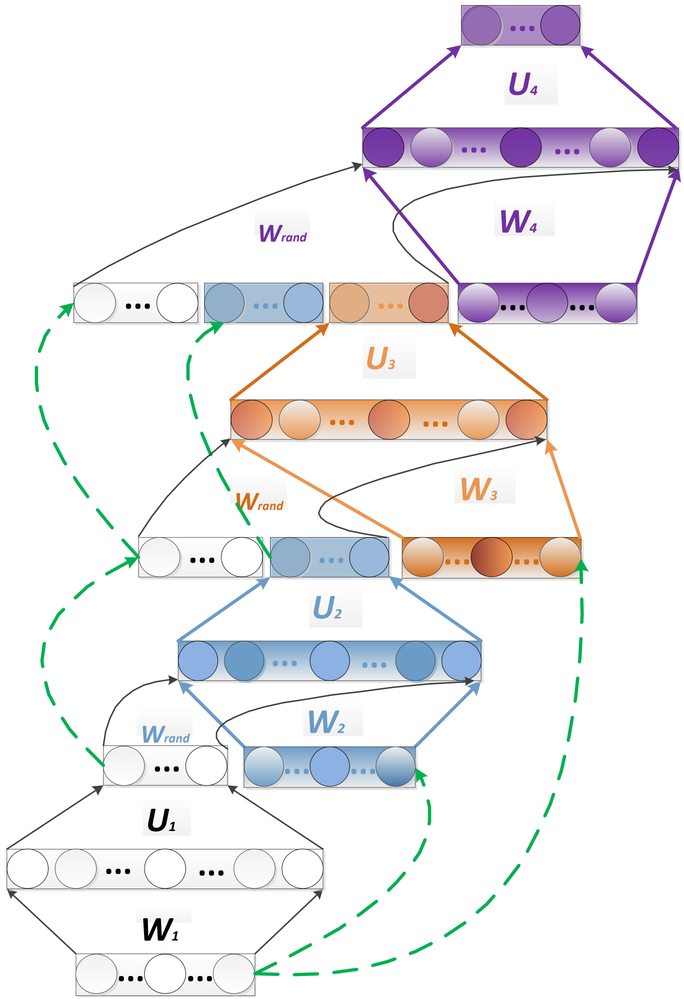
Deep Stacking Net (DSN)

- Deep Stacking Nets (Deng & Yu, Interspeech'10; Deng, Yu, Platt, ICASSP'12)
- Interleave linear/nonlinear layers
- Exploit closed-form constraints among network's weights
- Much easier to learn than DNN
- Naturally amenable to parallel training
- (Largely) convex optimization
- Extended to tensor version (Hutchinson et al, ICASSP'12, TPAMI-2013)
- Extended to kernel version (Deng et al, SLT'12)
- Works very well for MNIST, TIMIT, WSJ, SLU
- This paper: a more recent application to **IR ranking**



Learning DSN Weights --- Main Ideas

- Learn weight matrices U and W in individual modules separately.
- Given W and linear output layer, U can be expressed as explicit nonlinear function of W.
- This nonlinear function is used as the constraint in solving nonlinear least square for learning W.
- Initializing W with RBM (bottom layer)
- For higher layers, part of W is initialized with the optimized W from the immediately lower layer and part of it with random numbers

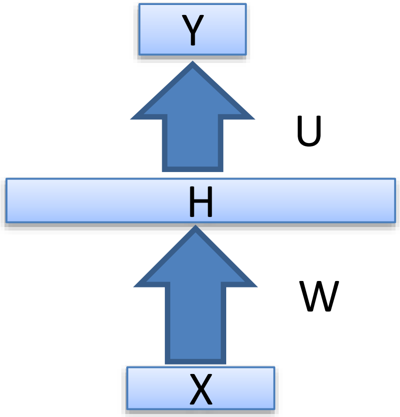


Learning DSN Weights --- Single Module

$$\mathrm{E} = \frac{1}{2}\sum_n ||\boldsymbol{y}_n - \boldsymbol{t}_n||^2, \qquad \text{where } \boldsymbol{y}_n = \boldsymbol{U}^T\boldsymbol{h}_n = \boldsymbol{U}^T\sigma(\boldsymbol{W}^T\boldsymbol{x}_n) = G_n(\boldsymbol{U},\boldsymbol{W})$$

$$\frac{\partial E}{\partial \boldsymbol{U}} = 2\boldsymbol{H}(\boldsymbol{U}^T\boldsymbol{H} - \boldsymbol{T})^T \rightarrow \quad \boldsymbol{U} = \left(\boldsymbol{H}\boldsymbol{H}^T\right)^{-1}\boldsymbol{H}\boldsymbol{T}^T = \mathrm{F}(\boldsymbol{W}), \text{ where } \boldsymbol{h}_n = \sigma(\boldsymbol{W}^T\boldsymbol{x}_n)$$

$\mathrm{E} = \frac{1}{2}\sum_n ||G_n(\boldsymbol{U},\boldsymbol{W}) - \boldsymbol{t}_n||^2$, *subject to* $\boldsymbol{U} = \mathrm{F}(\boldsymbol{W})$,



Use of Lagrange multiplier method:

$$\mathrm{E} = \frac{1}{2}\sum_n ||G_n(\boldsymbol{U},\boldsymbol{W}) - \boldsymbol{t}_n||^2 + \lambda\, ||\boldsymbol{U} - \mathrm{F}(\boldsymbol{W})||$$

to learn ***W*** and then ***U*** → no longer backpropagation

- Advantages found:
  --- less noise in gradient than using chain rule ignoring explicit constraint $\boldsymbol{U} = \mathrm{F}(\boldsymbol{W})$
  --- batch learning is effective, aiding parallel training

Experimental Evaluation

- IR task
  - Sponsored Search: retrieve and rank relevant ads given a query
- Data sets
  - Training: 189K query–ads pairs
  - Testing: 58K query–ads pairs
- Features to DSN
  - A total of 160 features in two categories
    - Text features: TF-IDF, word overlap, length, etc.
    - User click features: clickthrough, clicked queries, etc.
- State-of-the-art baseline system (Burges et al. 2006)
  - LambdaRank, a single-hidden-layer neural network
  - Trained to maximize (a smoothed approximation of) NDCG via heuristic lambda-function

Evaluation Metric

- Metric: Normalized Discounted Cumulative Gain (NDCG)
- DCG at rank p = $relevance_1 + \sum_{i=2}^{p} \frac{relevance_i}{\log_2 i}$; $relevance_i$: human label of doc$_i$, scale 0-4
- IDCG: Ideal DCG, DCG score when assuming docs are ranked by human label
- NDCG = DCG/ IDCG
- 1 NDCG pt (0.01) in our setting is statistically significant

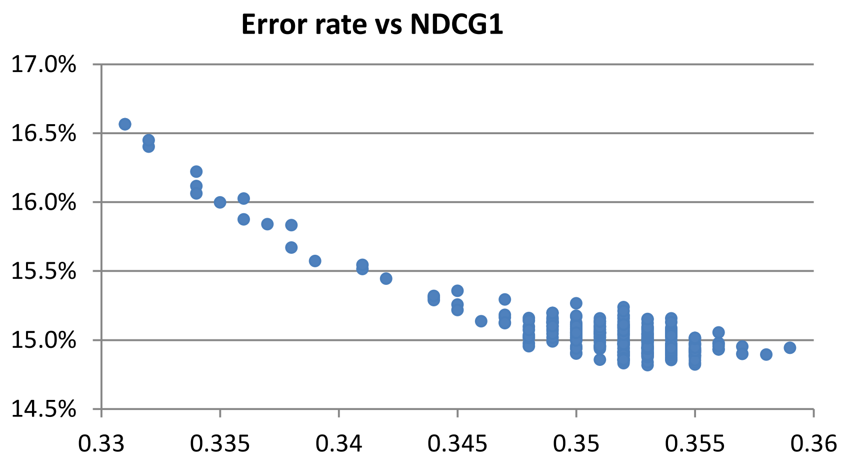
NDCG Results

IR Quality measures (NDCG) for the DSN System vs. Baseline

IR Systems	NDCG@1	NDCG@3	NDCG@10
LambdaRank	0.331	0.347	0.382
DSN system	0.359	0.366	0.402

Analysis

Relationship between classification error rates and NDCG@1 measure)



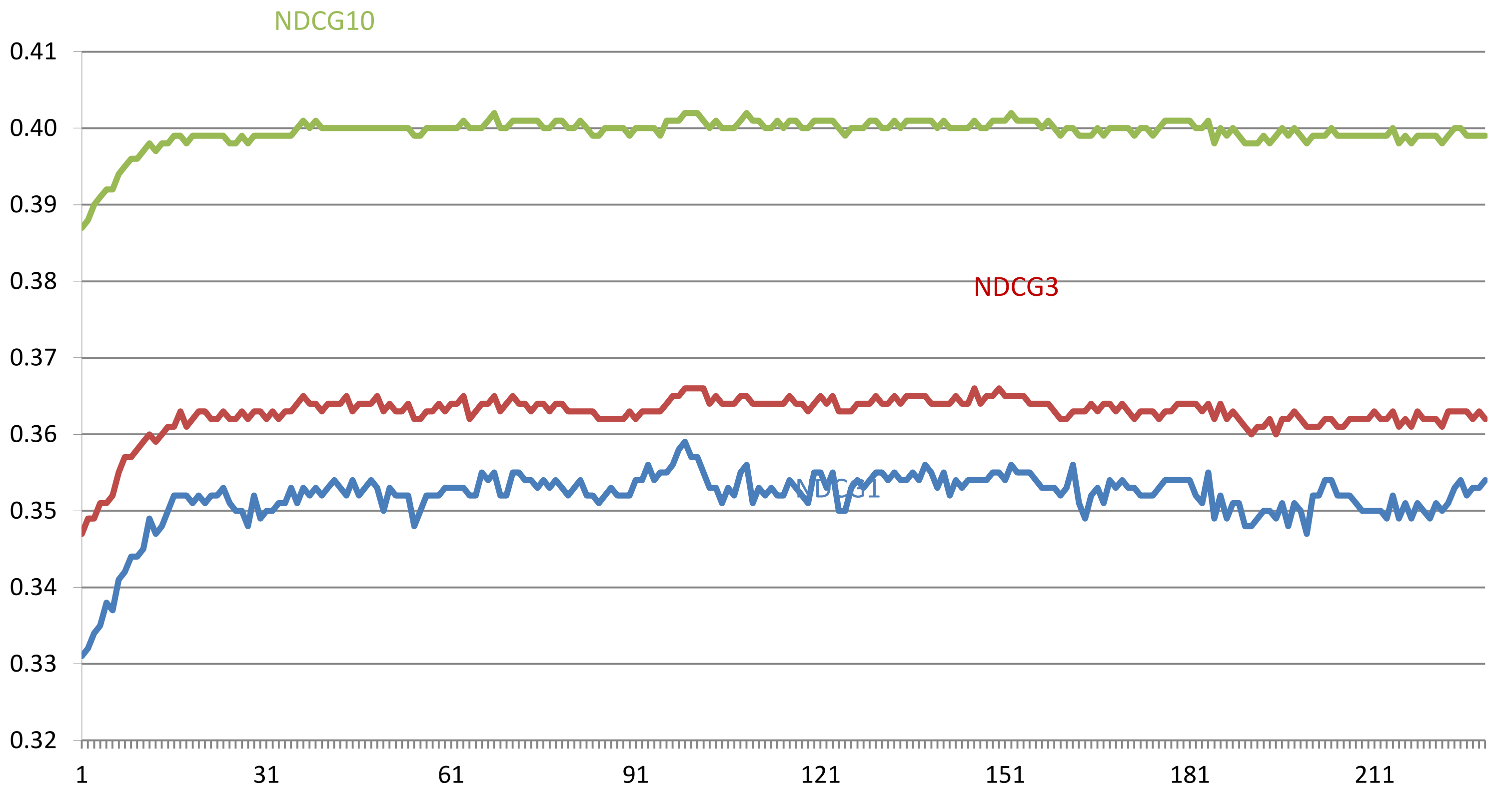
Observations:

- Correlation is clearly evidenced for NDCG1 < 0.35
- Weaker correlation in the high IR-quality region, i.e., NDCG1 > 0.35

Implication:

- Due to the inconsistency between the training objective and the IR-quality measure
- It is desirable to train the model to optimize the end-to-end IR quality directly

Learning Curves



Conclusions (of this ICASSP-2013 paper)

- First study on the use of deep learning techniques for learning-to-rank in IR problems
- Significantly better than shallow neural network
- Model trained by MSE
  - Generally correlated well with the NDCG as the IR quality measure
  - But weaker correlation in the region of high IR quality
- Deep learning using end-to-end IR-relevant metric is a key future direction

(including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition)

Outline

PART II: Deeper Substance of DL

---Technical introduction: RBM, DBN, DNN, DNN-HMM, CNN, RNN

---Examples of incorporating domain knowledge (about speech) into DL architectures

1. Hidden/articulatory Speech dynamics into RNN
2. Speech invariance/class-discrim.into deep-CNN

-A few new, promising DL architectures **(CONTINUED)**

New Types of Deep Neural Network & Learning for Speech Recognition+
An Overview

Li Deng, Geoffrey Hinton, Brian Kingsbury

MSR, U. Toronto/Google, IBM

ICASSP Special Session, May 28, 2013

Microsoft
UNIVERSITY OF TORONTO
Google
IBM Research

Special Session Motivations

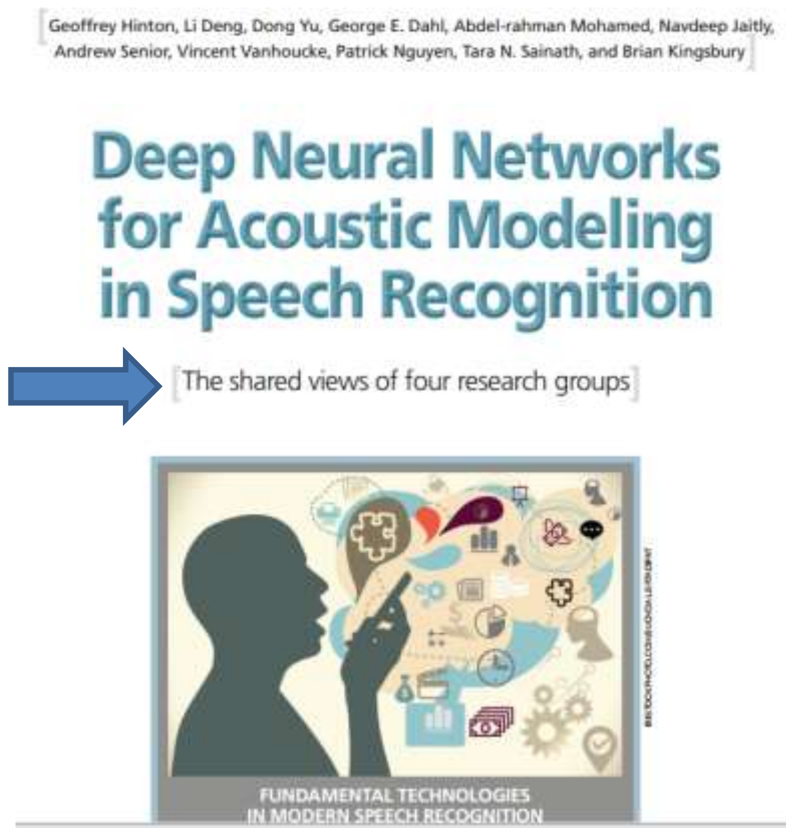
- Huge impact of deep neural nets (DNN) in speech (and vision, language, etc.)

The New York Times

Special Session Motivations

- Review article (2011-2012)
- Key factors:
  - Deeper network
  - Faster hardware
  - Larger network output layer (& hidden, input layers)
  - Better network initialization (not essential with big data)
- Rather standard MLP architecture
- Also standard backprop learning (1980's)

IEEE Sig. Proc. Mag, Nov 2012


Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

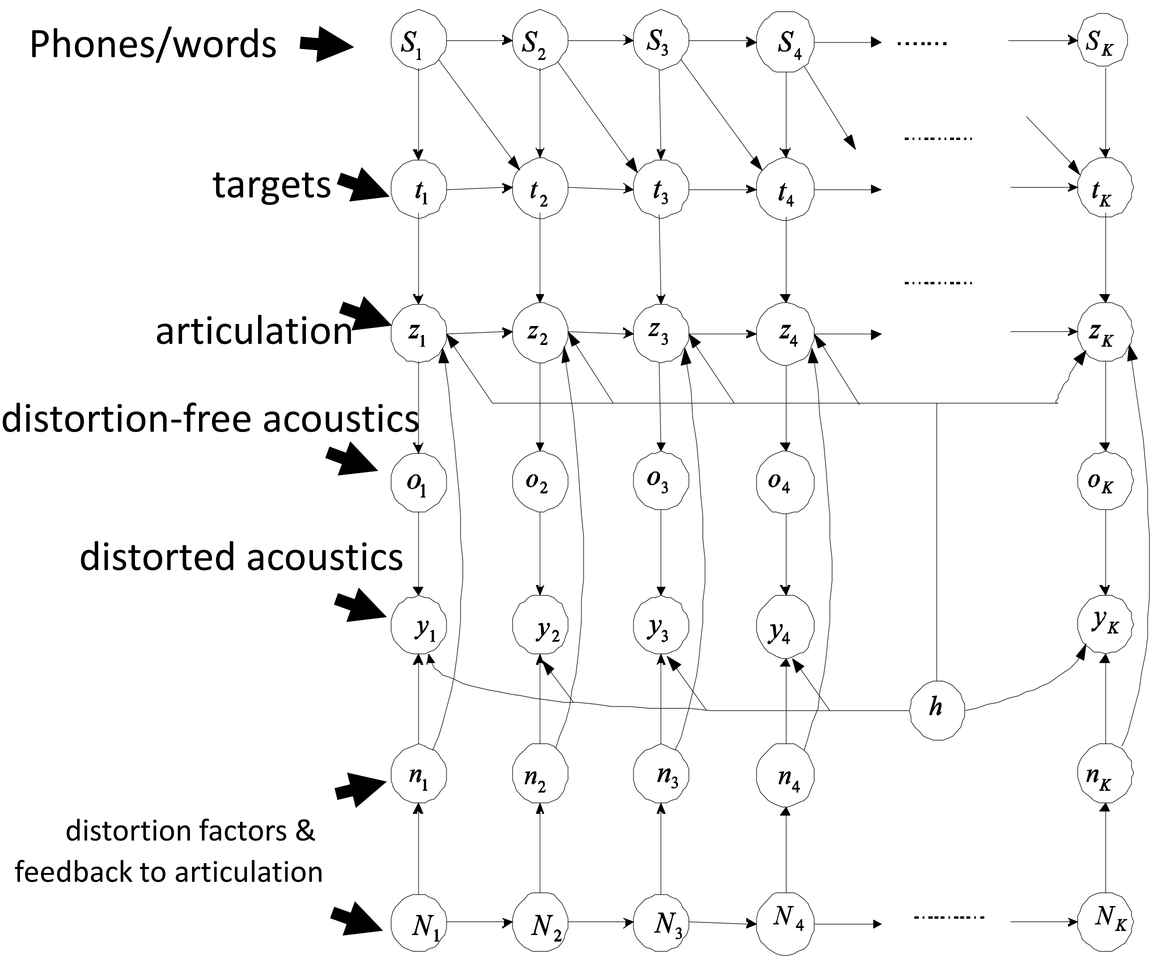
The shared views of four research groups

Take-Away from This Special Session

- New models and new learning methods
- Key capabilities of DNNs in knowledge transfer, learning representations, etc.
- Advances in DNNs since the SPM overview paper

Recent History of "Deep" Models in Speech

- MSR's (deep) Dyn. Bayes Net (2004-2007)
- U Toronto's DBN-DNN (2006-**2009**)



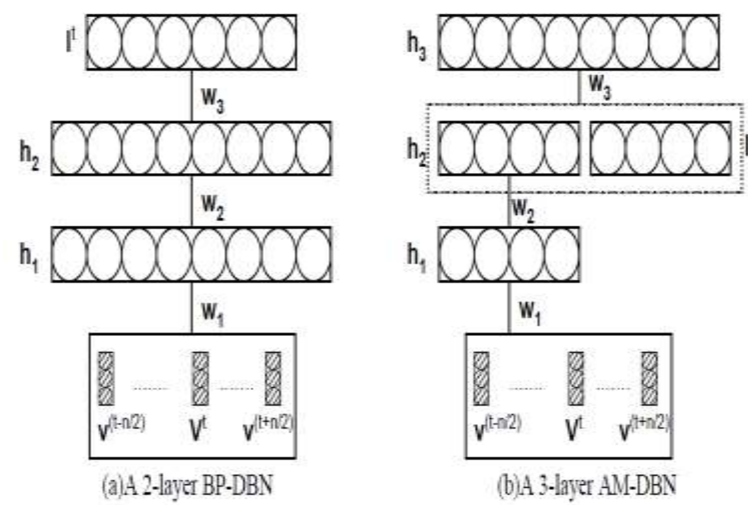
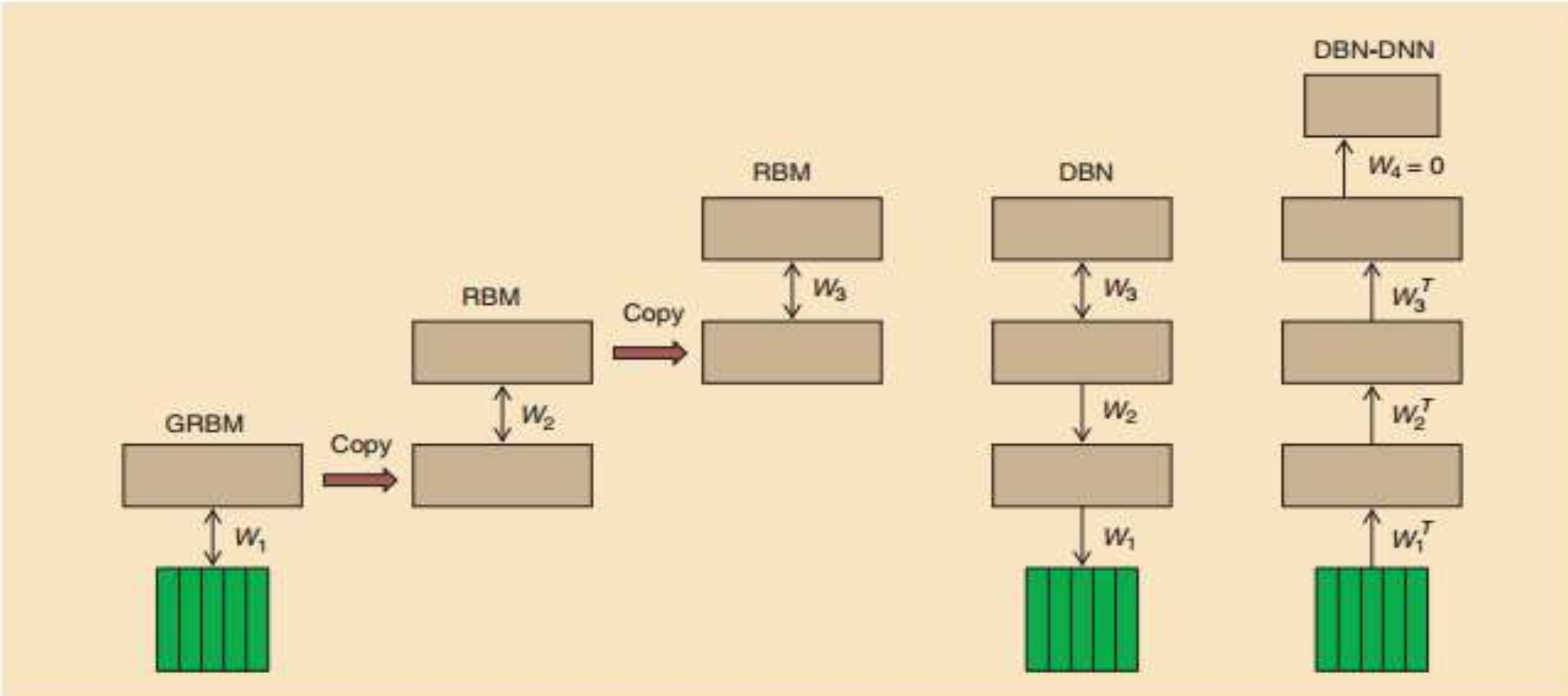


Figure 2: The DBN architectures used in this work.

Mohamed, Dahl, Hinton, NIPS-WS, 2009
(a simple "recipe")

Hinton's 2009"Recipe"



[FIG1] The sequence of operations used to create a DBN with three hidden layers and to convert it to a pretrained DBN-DNN. First, a GRBM is trained to model a window of frames of real-valued acoustic coefficients. Then the states of the binary hidden units of the GRBM are used as data for training an RBM. This is repeated to create as many hidden layers as desired. Then the stack of RBMs is converted to a single generative model, a DBN, by replacing the undirected connections of the lower level RBMs by top-down, directed connections. Finally, a pretrained DBN-DNN is created by adding a "softmax" output layer that contains one unit for each possible state of each HMM. The DBN-DNN is then discriminatively trained to predict the HMM state corresponding to the central frame of the input window in a forced alignment.

IEEE SIGNAL PROCESSING MAGAZINE [6] NOVEMBER 2012

Neural Information Processing Systems Foundation

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 of HMMs. The next generation of the technology requires solutions to remaining technical challenges under diversified deployment environments. These challenges, not adequately addressed in the past, arise from the many types of variability present in the speech generation process. Overcoming these challenges is likely to require "deep" architectures with efficient learning algorithms. For speech recognition and related sequential pattern recognition applications, some attempts have been made in the past to develop computational architectures that are "deeper" than conventional HMMs, such as hierarchical HMMs, hierarchical point-process models, hidden dynamic models, and multi-level detection-based architectures, etc. While positive recognition results have been reported, there has been a conspicuous lack of systematic learning techniques and 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.

http://research.microsoft.com/en-us/um/people/dongyu/NIPS2009/

Deep Learning for Phone Recognition
(a stunning discovery at MSR, 2009)

	METHOD	Error rate
	(Shallow) GMM-HMM (1987-2010)	27.3%
	AUGMENTED CONDITIONAL RANDOM FIELDS [26]	26.6%
	RANDOMLY INITIALIZED RECURRENT NEURAL NETS [27]	26.1%
	BAYESIAN TRIPHONE GMM-HMM [28]	25.6%
	Deep/hidden trajectory model (MSR, 2006)	24.8%
DNN	MONOPHONE RANDOMLY INITIALIZED DNNs (SIX LAYERS) [13]	23.4%
DNN	MONOPHONE DBN-DNNs (SIX LAYERS) [13]	22.4%
	MONOPHONE DBN-DNNs WITH MMI TRAINING (MSR, 2010)	22.1%
	TRIPHONE GMM-HMMs DT W/ BMMI (IBM, 2010)	21.7%
	MONOPHONE DBN-DNNs ON FBANK (EIGHT LAYERS) [13]	20.7%
	MONOPHONE MCRBM-DBN-DNNs ON FBANK (FIVE LAYERS) [33]	20.5%
	Deep convolutional nets w. DropOut & Heter. Pooling (MSR, 2012)	**18.7%**

Error pattern A

Error pattern B

IEEE SIGNAL PROCESSING MAGAZINE [6] NOVEMBER 2012

Deep Learning for Large-Vocabulary Speech Recognition

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

IEEE SIGNAL PROCESSING MAGAZINE [6] NOVEMBER 2012

New Discoveries about the DNN "Recipe" since 2009

- Pre-training not needed when a lot of labeled data are available (2010)
- The recipe works well for LVCSR when DNN output units correspond to CD HMM states (2010)
- **Decoding alg. & infrastructure largely unchanged, enabling industry-scale speech recognition (2010-2013)**
- Filterbank features (closer to waveform) better than MFCCs for DNNs (opposite to GMM systems) (2011-2013)
- DNN works surprisingly well for noisy speech (2012)
- Fully-connected DNN can be modified to include "convolutional" layers to handle speech variability (2012-2013)
- DNN highly effective for multi-task/transfer learning (e.g. multilingual ASR, 2012-2013)
- DNN effective for applications beyond ASR.

Five Technical Papers in Our Special Session

RECENT ADVANCES IN DEEP LEARNING FOR SPEECH RESEARCH AT MICROSOFT

Microsoft

IMPROVING DEEP NEURAL NETWORKS FOR LVCSR USING RECTIFIED LINEAR UNITS AND DROPOUT

UNIVERSITY OF TORONTO

DEEP CONVOLUTIONAL NEURAL NETWORKS FOR LVCSR

IBM Research

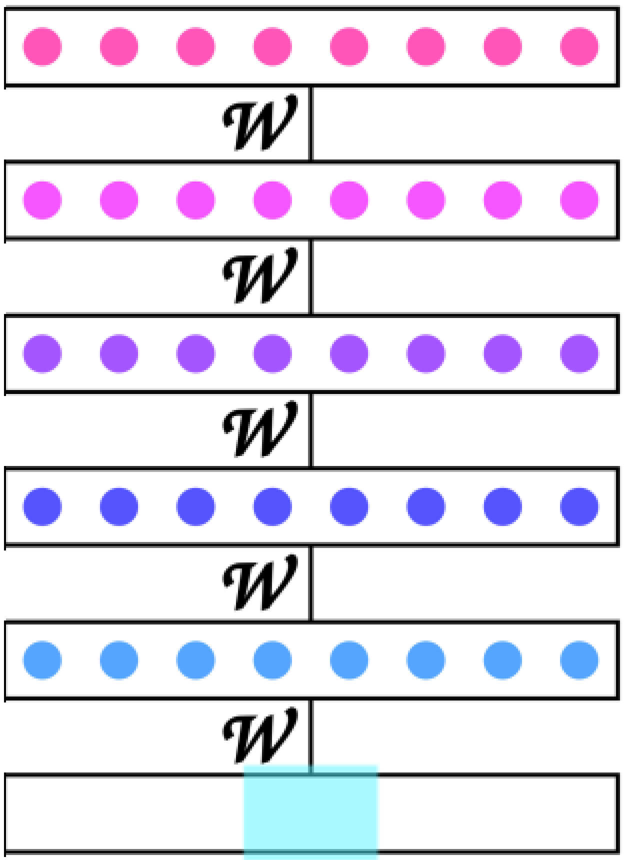
MULTILINGUAL ACOUSTIC MODELS USING DISTRIBUTED DEEP NEURAL NETWORKS

Google

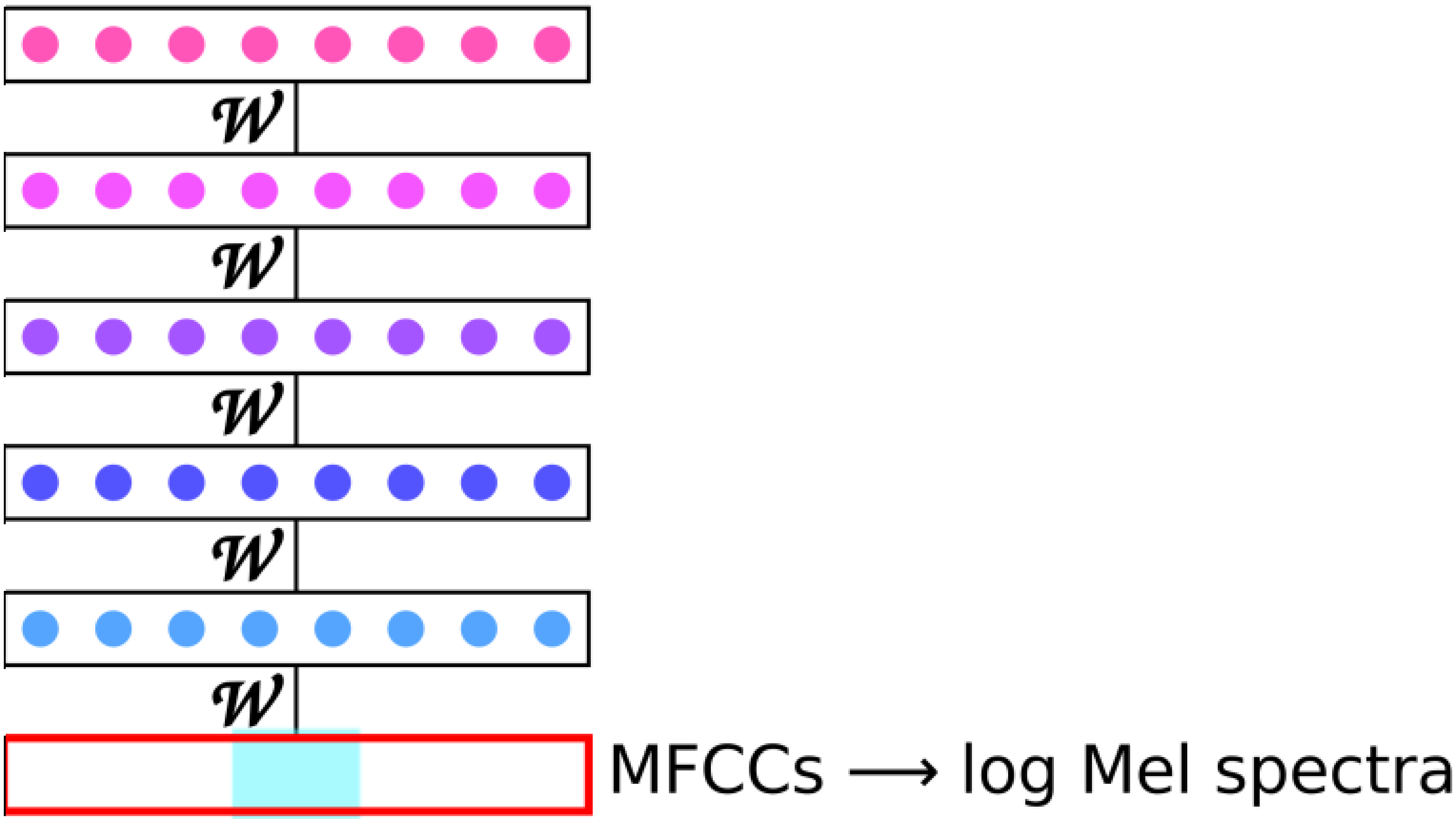
ADVANCES IN OPTIMIZING RECURRENT NETWORKS

Université de Montréal

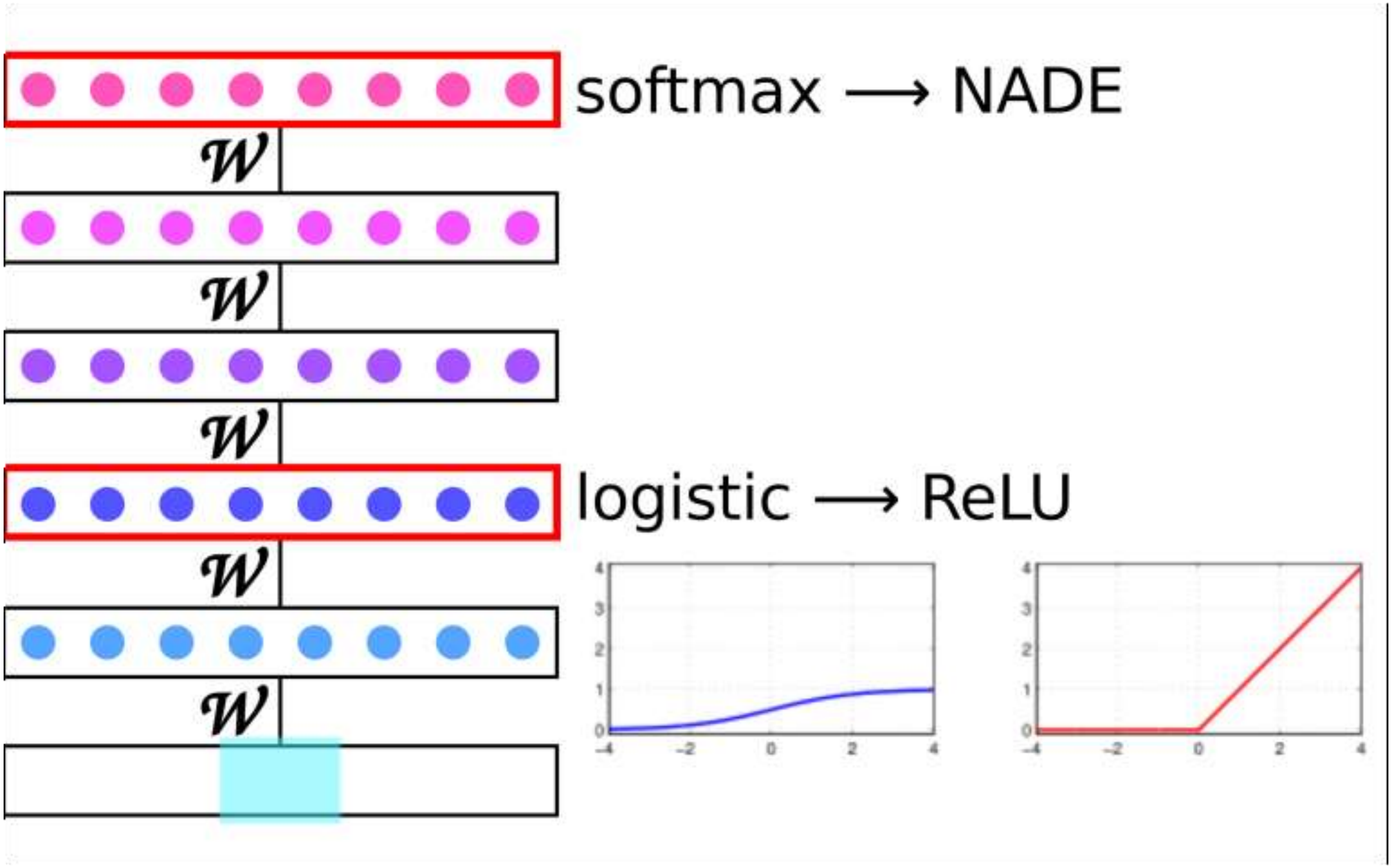
Themes in the Session



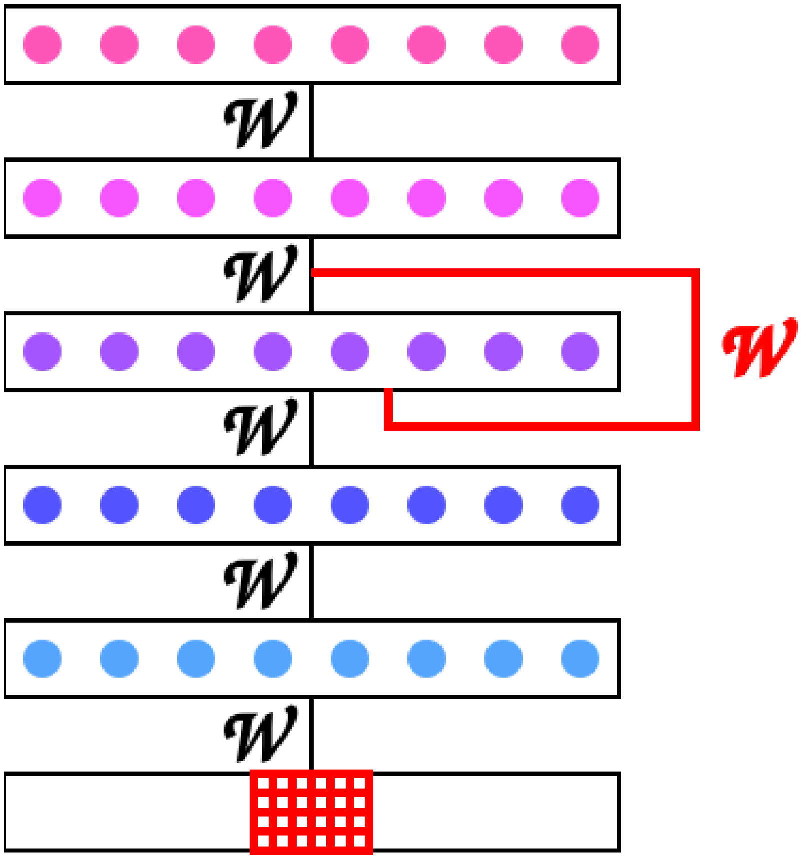
Themes: Better Inputs



Themes: Nonlinearities



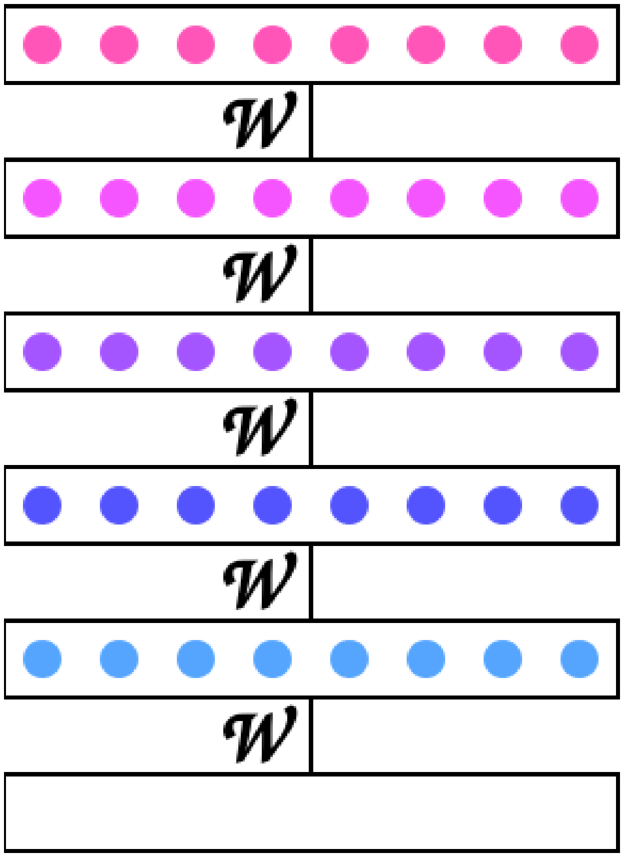
Themes: Architectures



Recurrent NNs

Convolutional NNs

Themes: Optimization



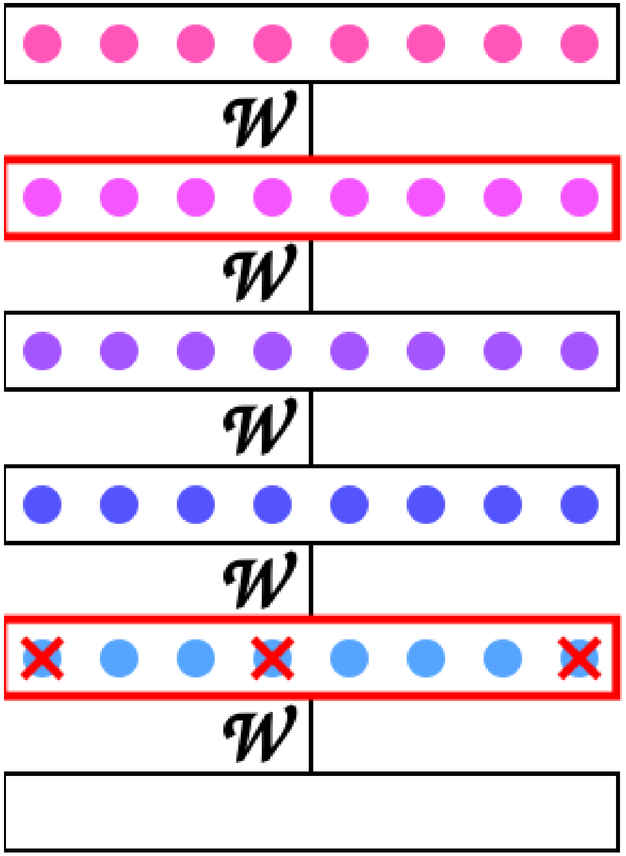
Distributed asynchronous SGD

Distributed Hessian-free optimization

Nesterov's accelerated gradient method

Gradient clipping

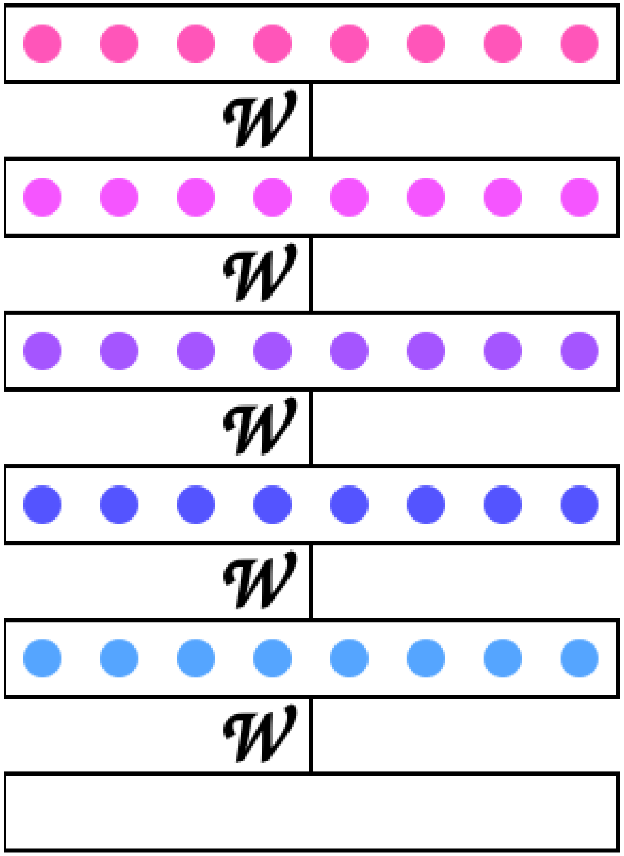
Themes: Regularization



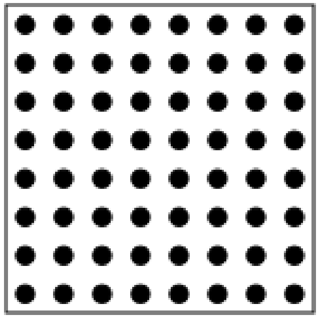
Sparsity in hidden representations

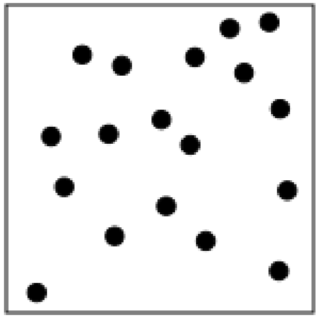
Dropout

Themes: Hyperparameters



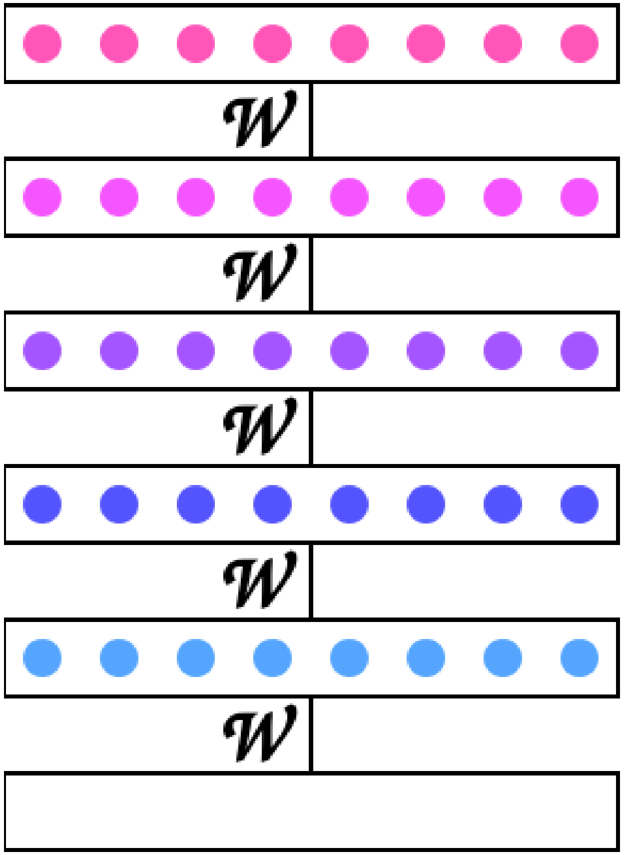
Grid search ⟶ Sampling





Bayesian optimization

Themes: Multi-task Learning



Multi-lingual acoustic modeling

Mixed-bandwidth acoustic modeling

Recent Advances in Deep Learning for Speech Research at Microsoft

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Mike Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, Alex Acero

ICASSP Special Session, May 28, 2013

Outline

- Advances in deep learning for **features**/ representations
- Advances in deep learning for **models/** architectures
- Systems and applications in acoustic modeling, language modeling, dialogue, (and information retrieval/search)

Learning Features/Representations

- Advances in deep learning for **features**/ representations
- Advances in deep learning for **models/** architectures
- Systems and applications in acoustic modeling, language modeling, dialogue, (and information retrieval/search)

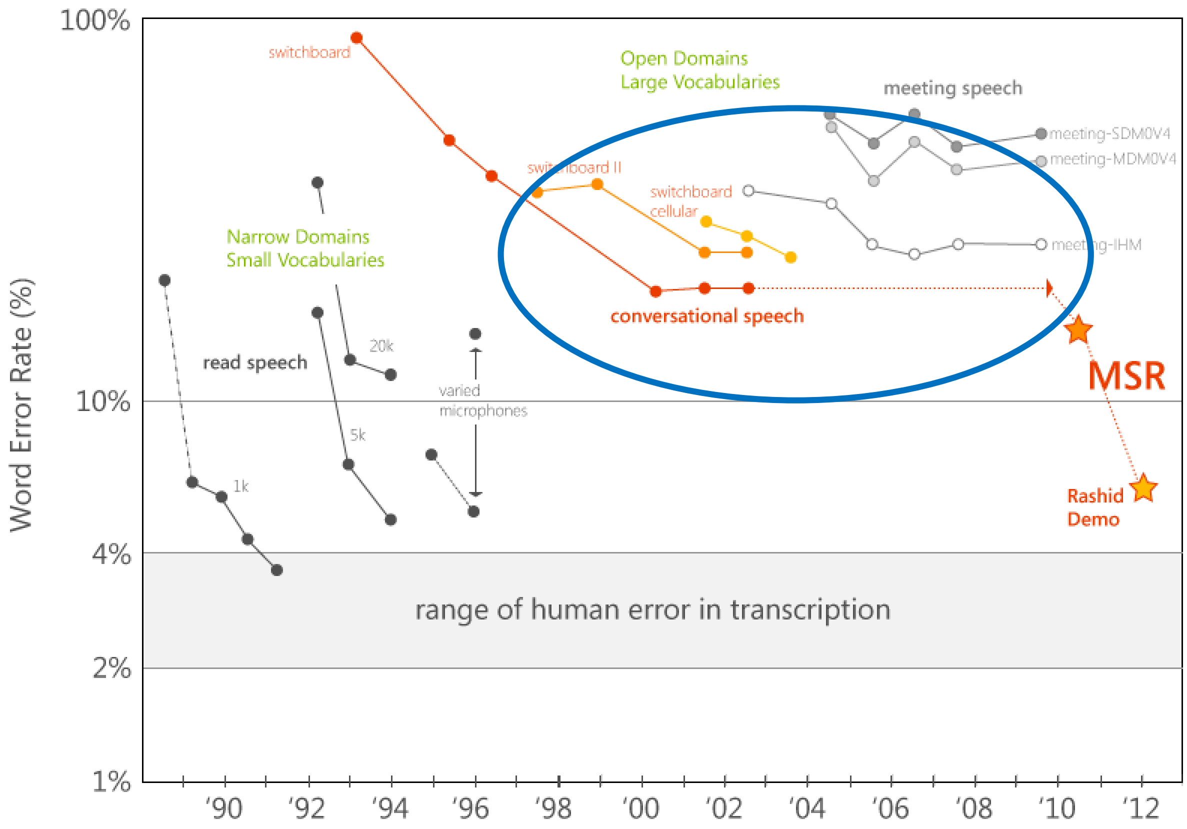
The New York Times

Scientists See Promise in Deep-Learning Programs

John Markoff

November 23, 2012

Speech Recognition Progress:
--- gleaned from NIST evaluations



After no improvement for 10+ years by the research community…

…MSR used deep learning to reduce error rate from **~23%** to **~13%** on SWBD (and under 7% for Rick Rashid's demo)!

Back to Primitive Spectral Features

- Philosophy of deep learning:
  - Learning representations automatically instead of manually engineering/design them (e.g., MFCC, PLP)
- DNN capability in representing correlated feature dimensions
- → eliminate cosine transform in MFCC in favor of filterbanks in spectral domain

Back to Primitive Spectral Features

- Philosophy of deep learning:
  - Learning representations automatically instead of manually engineering/design them (e.g., MFCC, PLP)
- DNN capability in representing correlated feature dimensions
- → eliminate cosine transform in MFCC in favor of filterbanks and spectrograms in the spectral domain

INTERSPEECH 2010

Binary Coding of Speech Spectrograms Using a Deep Auto-encoder

L. Deng[1], M. Seltzer[1], D. Yu[1], A. Acero[1], A. Mohamed[2], and G. Hinton[2]

[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052, US
[2] University of Toronto, Toronto, Ontario, Canada

In early 2010, we discovered:

For deep autoencoding of speech features:

- Both spectrogram/filterbank features are better than MFCCs
- Better to use spectrogram features than filterbanks
- "Better" in terms of coding efficiency (i.e., errors/energy)

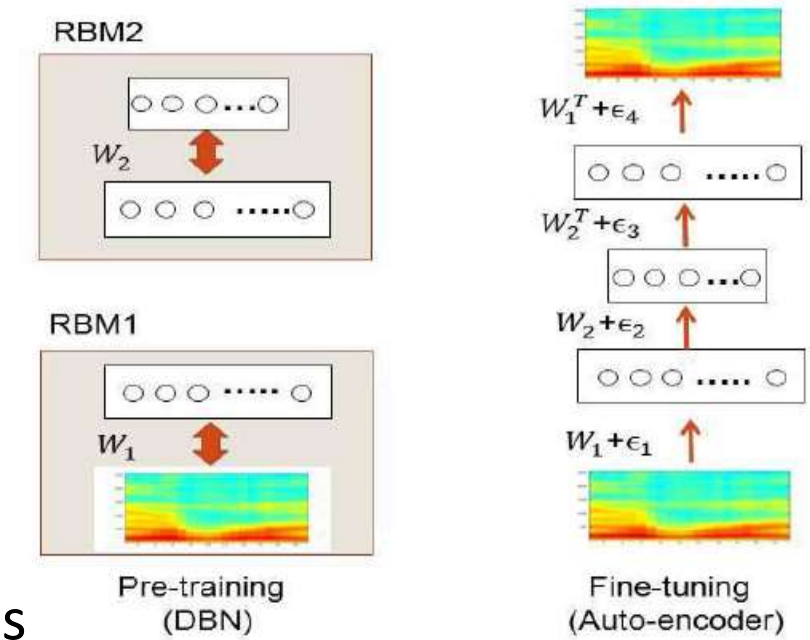


Fig.1. *Left: Illustration of pre-training of the DBN that consists of two RBMs used in this work. Right: Illustration of fine tuning that produces the final deep auto-encoder.*

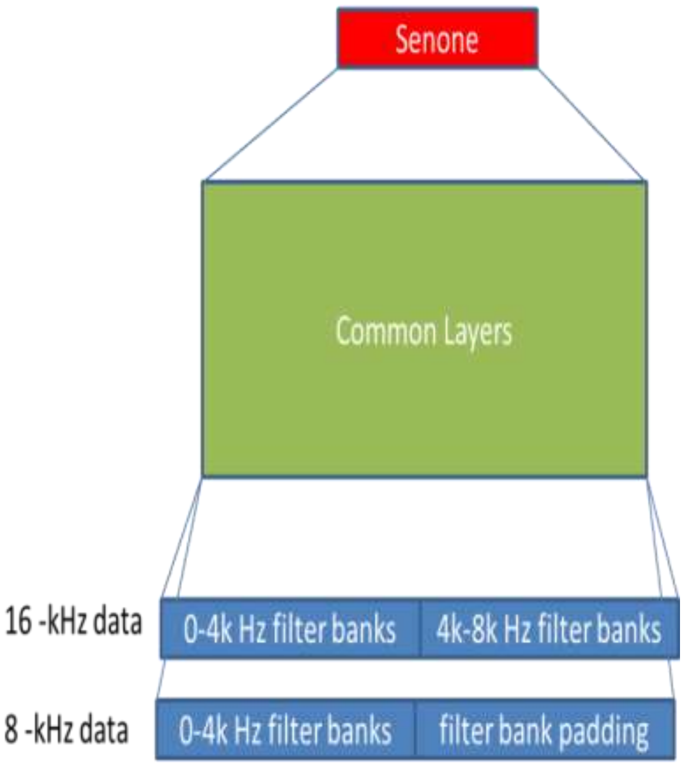
LVCSR Using Spectral Features

LVCSR Systems	Word error rate
Best GMM-HMM (MFCCs; fMPE+BMMI)	34.7%
DNN (MFCCs)	**31.6%**
DNN (Spectrogram --- 256 log FFT bins)	**32.3%**
DNN (29 log filter-banks)	30.1%
DNN (40 log filter-banks)	**29.9%**

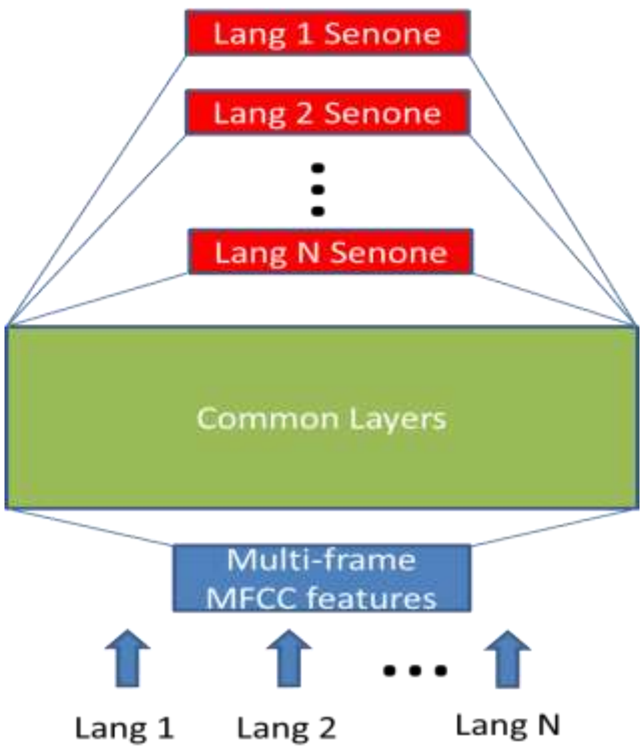
- Filterbanks > MFCC > Spectrograms
- Not quite consistent with deep autoencoder results
- Further research: regularization, online feature normalization at sentence level, etc.

Learning Multi-Task Features

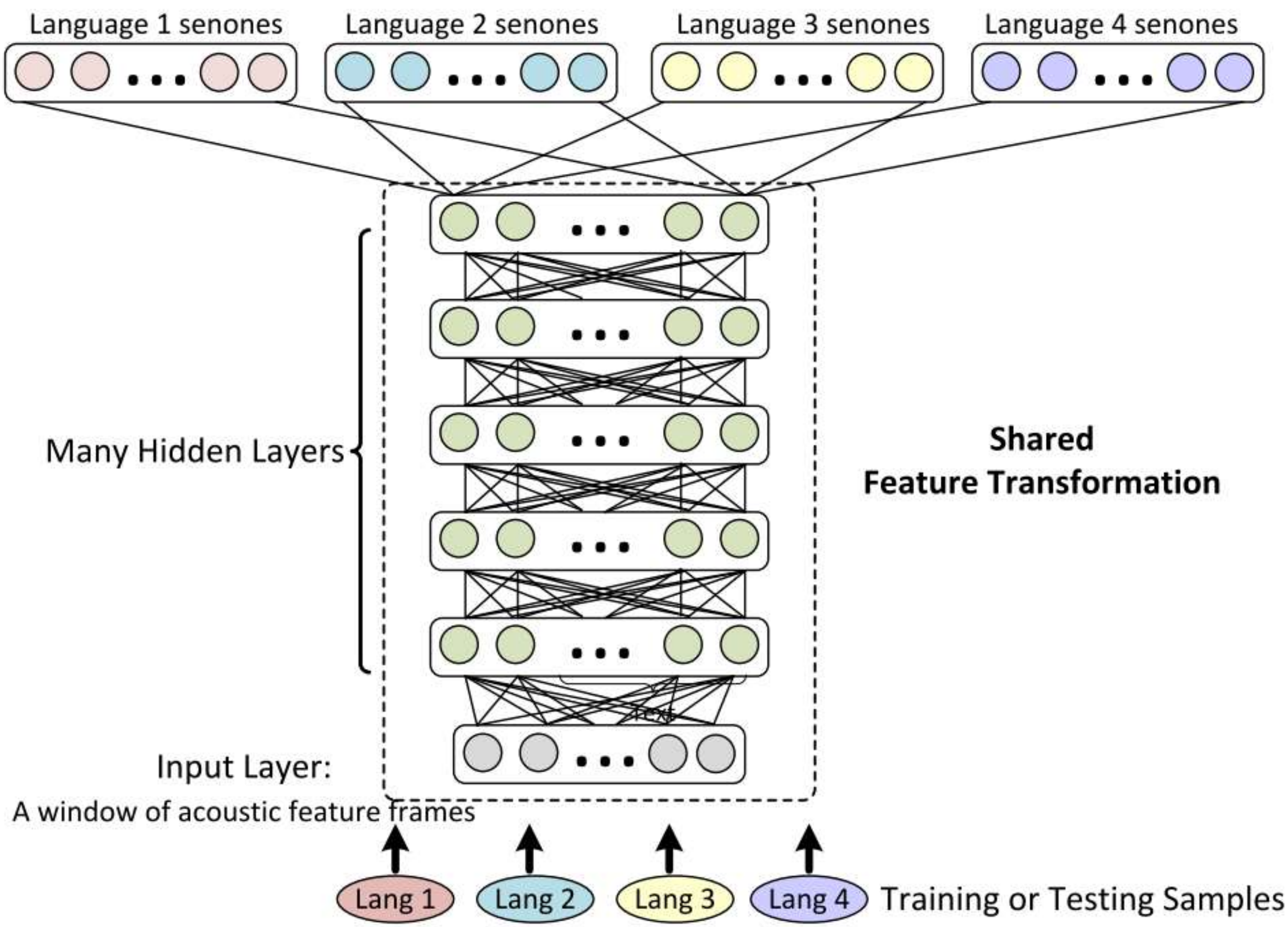
Mixed-Band DNN architecture:



Multilingual DNN architecture:



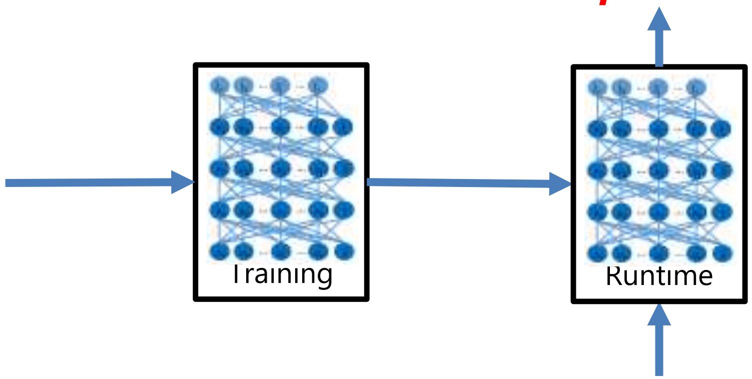
Shared Hidden Layers with Language-Specific Output Layers



Learning Multi-Task Features

English ***or French*** ***or Chinese*** words
with fewer errors

8KHz waveforms
16KHz waveforms
English training data
French training data
Chinese training data



8Khz or 16KHz English ***or French*** ***or Chinese***
speech input

Multi-Band ASR Summary Results

Practical Goal: exploit narrowband labeled data from earlier telephone-based applications

Training Data	Test WER (Wideband)	Test WER (Narrowband)
Wideband only	30.0%	71.2%
Narrowband only	-	29.0%
Wideband+Narrowband	28.3%	29.3%

Multilingual ASR Summary Results

Speech Recognizers	WER on ENU
DNN trained with only ENU data	**30.9%**
+FRA, retrain all layers with ENU	30.6%
or +FRA, retrain the top layer with ENU	27.3%
or +FRA+ DEU+ ESP+ITA, retrain top layer	**25.3%**

Deep Convolutional Net w. Spectral Features

- "Spatial" (freq-domain) invariance of speech due to vocal-tract-length differences across speakers
- Convolution/pooling makes sense for
  - spectral features, not MFCC
  - "spatial" dimension, not (pure) temporal dimension
- Excellent results on TIMIT:

DNN	MONOPHONE RANDOMLY INITIALIZED DNNs (SIX LAYERS) [13]	23.4%
	MONOPHONE DBN-DNNs (SIX LAYERS) [13]	22.4%
	MONOPHONE DBN-DNNs WITH MMI TRAINING (MSR, 2010)	22.1%
	TRIPHONE GMM-HMMs DT W/ BMMI (IBM, 2010)	21.7%
	MONOPHONE DBN-DNNs ON FBANK (EIGHT LAYERS) [13]	20.7%
	MONOPHONE MCRBM-DBN-DNNs ON FBANK (FIVE LAYERS) [33]	20.5%
	Deep convolutional nets w. DropOut & Heter. Pooling (MSR, 2012)	**18.7%**

Disastrous when MFCC is used for convolutional net over "spatial" dimension

Noise Robust DNN Features

- Beating state-of-the-art WER results on Aurora4 task (medium vocabulary task based on WSJ0)
- DNN: not yet exploited explicit noise compensation algorithm
- DNN: no multi-pass decoding allowing for adaptation

ASR Word Error Rate % for Autora4:

	A	B	C	D	AVG
GMM-HMM (Baseline)	12.5	18.3	20.5	31.9	23.9
GMM (MPE+VAT)	7.2	12.8	11.5	19.7	15.3
GMM + Deriv. Kernels	7.4	12.6	10.7	19.0	14.8
DNN (7x2000)	5.6	8.8	8.9	20.0	13.4

•

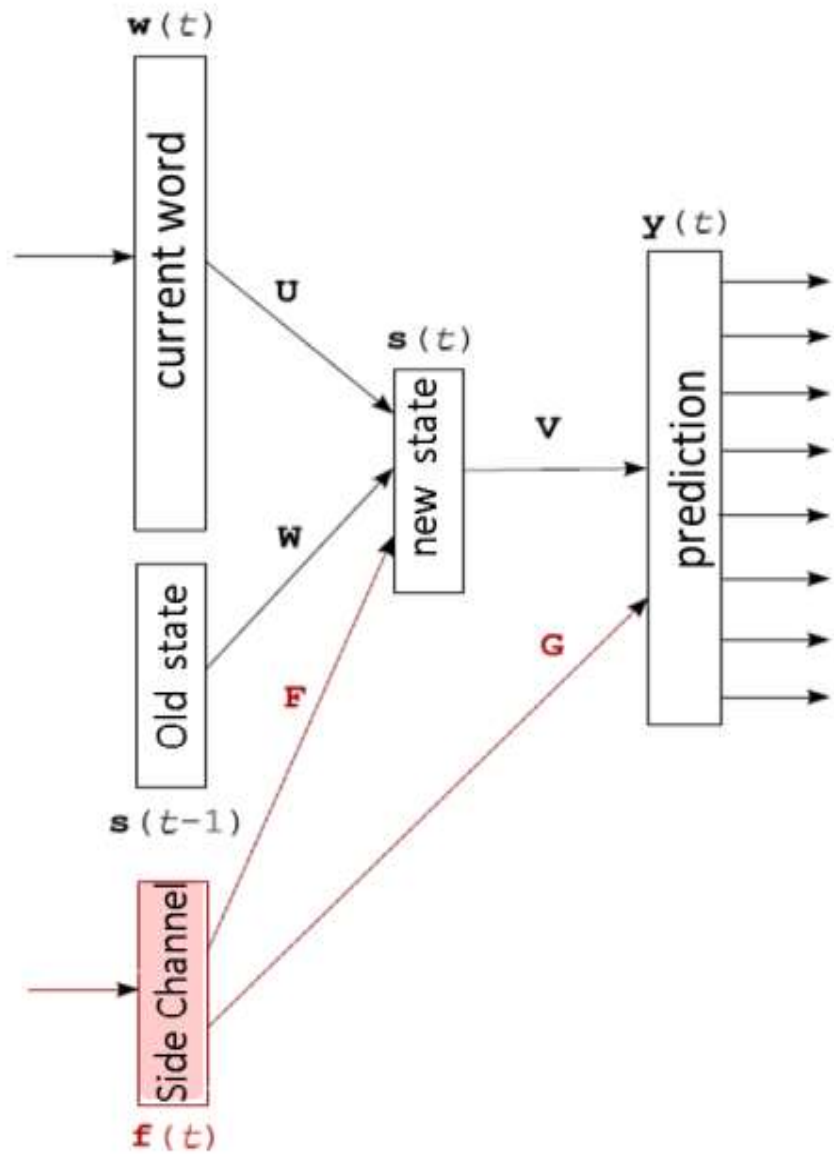
DNN "Model" Adaptation

Speech Recognition Systems	WER
GMM-HMM	43.6%
DNN	**34.1%**
DNN + AdaptSoftMax (SGD)	29.4%
DNN + fDLR (SGD)	**28.5%**

New “Model” Architecture: Recurrent Net

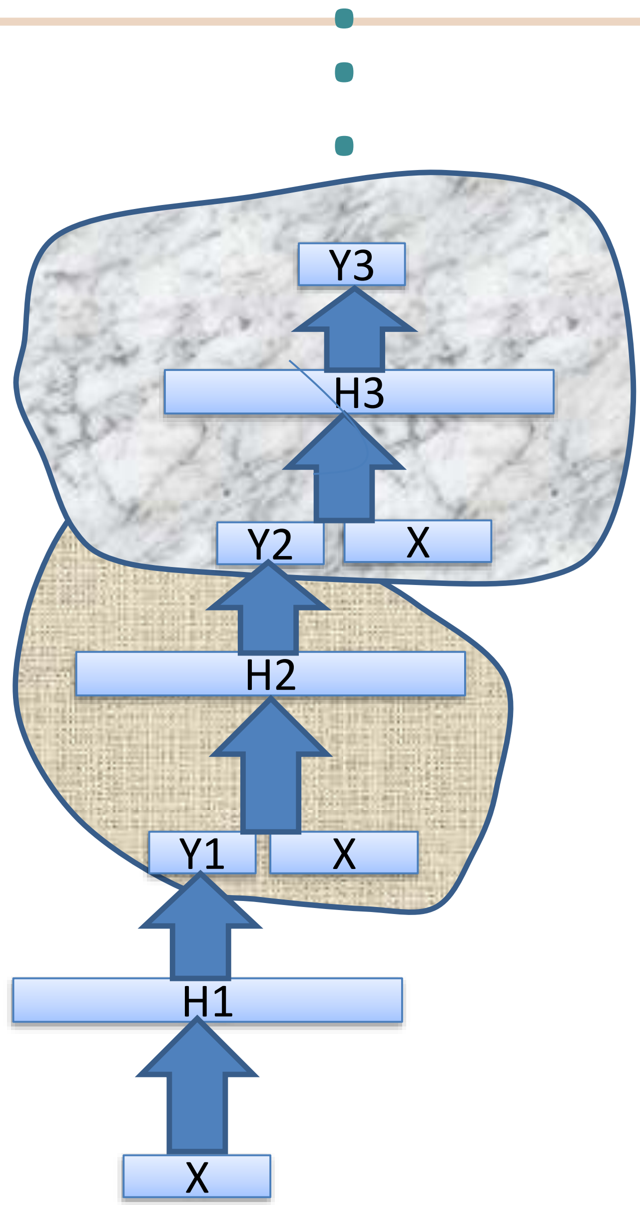
- For language modeling application
- Exploit context dependency (side channel information) in RNN
- Side channel consists of slowly varying LSA vectors of preceding text
- Evaluation on Penn Treebank data
- Baseline (KN 5-gram LM w. cache)
  - perplexity= 126
- RNN w. side channel
  - perplexity= 110

(lowest single-model perplexity for this data)



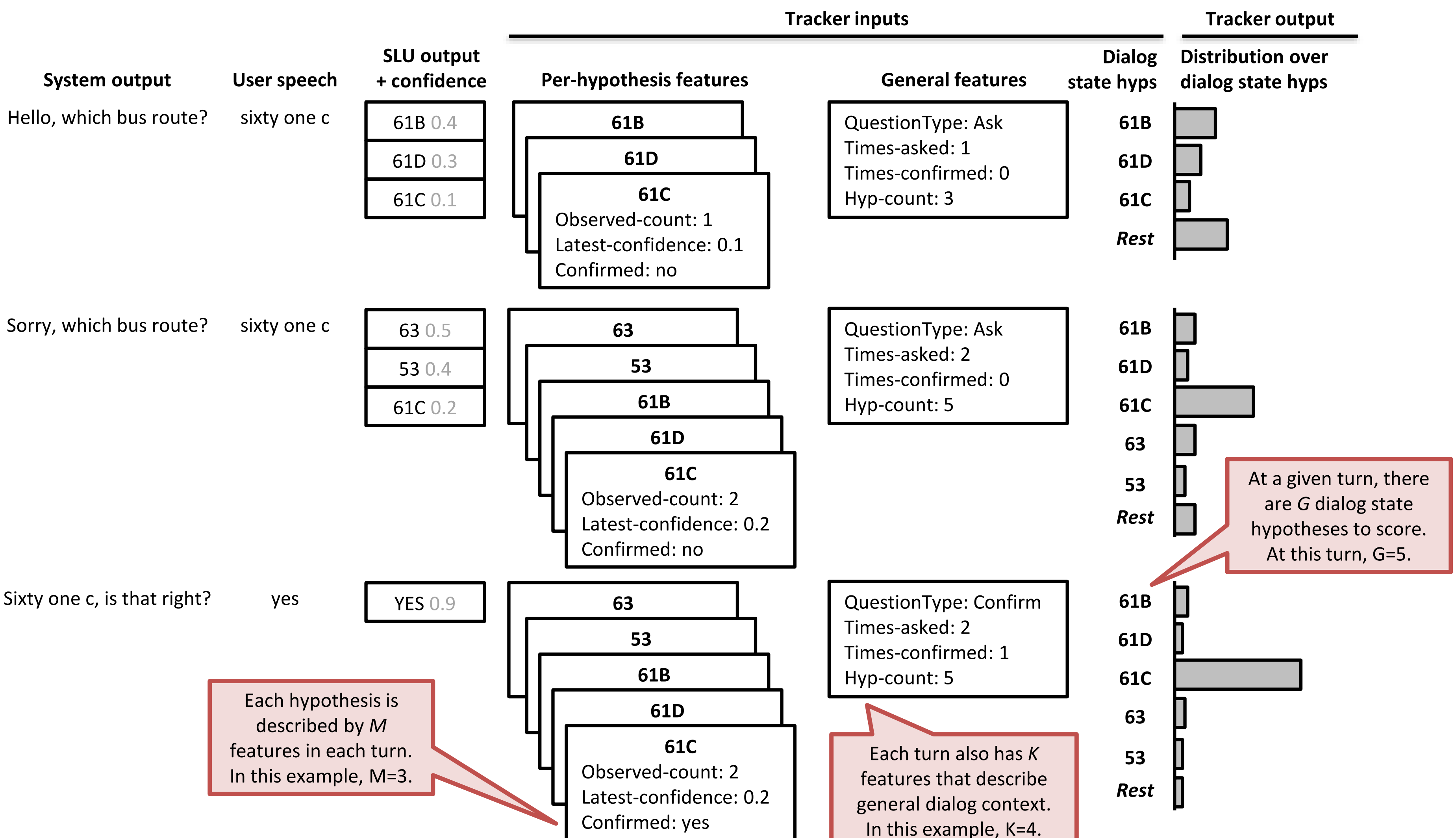
Deep Learning for Dialogue State Tracking

- Fertile area with preliminary exploration
- Use of a new architecture: Deep Stacking Net (Deng & Yu, 2010)
- Interleave linear/nonlinear layers
- Exploit closed-form constraints among network's weights
- Much easier to learn than DNN
- Naturally amenable to parallel training
- (Largely) convex optimization
- Extended to tensor and kernel versions
- Works very well for MNIST, TIMIT, WSJ, SLU, and IR ranking (Deng, He, Gao: ICASSP 2013)
- Here we show a more recent application to **state tracking task** in spoken dialogue systems.



Dialogue State Tracking Example (Jason Williams)

- Prob [CorrectUserGoals_t | DialogueHistory_{1,2,…,t-1}, UserInfo_{1,2,…,t-1}]



DSN Results for Dialogue State Tracking

- Task: Dialog state tracking (defined in Spoken Dialogue Challenge 2010)
- Strong interactions among features → strength of deep networks
- Can be framed as a multiple binary classification problem
- Baseline: carefully tuned, highly optimized Max Entropy classifier (J. Williams)
- Deep Stacking Nets (slightly tuned) achieve similar accuracy% for all 5 slots:

	State of the Art baseline	**Deep Stacking Networks**
Bus route	58.0%	**58.1%**
Origin location	56.4%	**57.1%**
Destination location	66.5%	65.4%
Date	83.9%	**84.6%**
Time	63.1%	62.5%

Conclusions

- Deep learning is a powerful technology
  - Automatic learning of representations
  - Multi-task learning
  - Factorizing/disentangling multiple causes of variations
- Future directions
  - More effective deep architectures and learning algms
  - Scale deep mode training with bigger data
  - Extend applications of deep learning: acoustic models, language models, dialogue, end-2-end language understanding & translation, IR/search, synthesis, music processing, etc.

Summary

PART I: Basics of Deep Learning (DL)

--- including impact and recent history of DL (Deep Neural Net, DNN) in speech recognition

PART II: Deeper Substance of DL

--- including connections to other ML paradigms
--- two examples of incorporating speech knowledge in DL architectures,
---recent experiments in speech recognition with new DL architectures beyond DNN

Microsoft Research Selected References (updated, 2013)

Abdel-Hamid, O., Mohamed, A., Jiang, H., and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," Proc. ICASSP, 2012.

Arel, I., Rose, C., and Karnowski, T. "Deep Machine Learning - A New Frontier in Artificial Intelligence," IEEE Computational Intelligence Mag., Nov., 2010.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shaughnessy, D. "Research developments and directions in speech recognition and understanding," IEEE Sig. Proc. Mag., vol. 26, no. 3, May 2009, pp. 75-80.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shaughnessy, D. "Updated MINS report on speech recognition and understanding," IE Sig. Proc. Mag., vol. 26, no. 4, July 2009a.

Bengio, Y., Boulanger, N., and Pascanu. R. "Advances in optimizing recurrent networks," Proc. ICASSP, 2013.

Bengio, Y., Courville, A., and Vincent, P. "Representation learning: A review and new perspectives," IEEE Trans. PAMI, 2013a.

Bengio, Y. "Learning deep architectures for AI," in Foundations and Trends in Machine Learning, Vol. 2, No. 1, 2009, pp. 1-127.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. "A neural probabilistic language model," Proc. NIPS, 2000, pp. 933-938.

Bengio, Y., De Mori, R., Flammia, G. and Kompe, F. "Global optimization of a neural network—Hidden Markov model hybrid," in Proc. Eurospeech, 1991.

Bergstra, J. and Bengio, Y. "Random search for hyper-parameter optimization," J. Machine Learning Research," Vol. 3, pp. 281-305, 2012.

Bottou, L. and LeCun. Y. "Large scale online learning," Proc. NIPS, 2004.

Bilmes, J. "Dynamic graphical models," IEEE Signal Processing Mag., vol. 33, pp. 29–42, 2010.

Bilmes, J. and Bartels, C. "Graphical model architectures for speech recognition," IEEE Signal Processing Mag., vol. 22, pp. 89–100, 2005.

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

Bouvrie, J. "Hierarchical Learning: Theory with Applications in Speech and Vision," Ph.D. thesis, MIT, 2009.

Bridle, J., 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.

Caruana, R. "Multitask Learning," *Machine Learning*, Vol. 28, pp. 41-75, Kluwer Academic Publishers, 1997.

Cho, Y. and Saul L. "Kernel methods for deep learning," Proc. NIPS, pp. 342–350, 2009.

Ciresan, D., Giusti, A., Gambardella, L., and Schmidhuber, J. "Deep neural networks segment neuronal membranes in electron microscopy images," Proc. NIPS, 2012.

Cohen, W. and R. V. de Carvalho. "Stacked sequential learning," Proc. IJCAI, pp. 671–676, 2005.

Collobert, R. "Deep learning for efficient discriminative parsing," Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

Collobert, R. and Weston J. "A unified architecture for natural language processing: Deep neural networks with multitask learning," Proc. ICML, 2008.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. "Natural language processing (almost) from scratch," J. Machine Learning Research, Vo. 12, pp. 2493-2537, 201

Microsoft Research

Selected References

Dahl, G., Yu, D., Deng, L., and Acero, A. "Context-dependent DBN-HMMs in large vocabulary continuous speech recognition," Proc. ICASSP, 2011.

Dahl, G., Yu, D., Deng, L., and Acero, 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, January 2012.

Dahl, G., Ranzato, M., Mohamed, A. and Hinton, G. "Phone recognition with the mean-covariance restricted Boltzmann machine," Proc. NIPS, vol. 23, 2010, 46 477.

Dean, J., Corrado, G., R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, Yang, K., and Ng, A. "Large Scale Distributed Deep Networ Proc. NIPS, 2012.

Deng, L. and Li, X. "Machine learning paradigms in speech recognition: An overview," IEEE Trans. Audio, Speech, & Language, May 2013.

Deng, L., Abdel-Hamid, O., and Yu, D. "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," Proc. ICASSP, 2013.

Deng, L., Li, J., Huang, K., Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. "Recent advances in deep learning for speech research at Microsoft," Proc. ICASSP, 2013a.

Deng, L., Hinton, G., and Kingsbury, B. "New types of deep neural network leaning for speech recognition and related applications: An overview," Proc. ICASSP 2013b.

Deng, L., He, X., and J. Gao, J. "Deep stacking networks for information retrieval," Proc. ICASSP, 2013c.

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

Deng, L., Yu, D., and Platt, J. "Scalable stacking and learning for building deep architectures," Proc. ICASSP, 2012a.

Deng, L., Hutchinson, B., and Yu, D. "Parallel training of deep stacking networks," Proc. Interspeech, 2012b.

Deng, L. "An Overview of Deep-Structured Learning for Information Processing," Proceedings of Asian-Pacific Signal & Information Processing Annual Summit Conference (APSIPA-ASC), October 2011.

Deng, L. and Yu, D. "Deep Convex Network: A scalable architecture for speech pattern classification," Proc. Interspeech, 2011.

Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G. "Binary coding of speech spectrograms using a deep auto-encoder," Proc. Interspeech, 2010.

DENG, L., YU, D., AND HINTON, G. "DEEP LEARNING FOR SPEECH RECOGNITION AND RELATED APPLICATIONS" NIPS WORKSHOP, 2009.

DENG, L. AND YU, D. "USE OF DIFFERENTIAL CEPSTRA AS ACOUSTIC FEATURES IN HIDDEN TRAJECTORY MODELING FOR PHONETIC RECOGNITION," PROC. ICASSP, 2007.

Deng, L. DYNAMIC SPEECH MODELS – Theory, Algorithm, and Application, Morgan & Claypool, December 2006.

Deng, L., Yu, D. and Acero, A. "Structured speech modeling," IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504, September 2006

Deng, L., Yu, D. and Acero, A. "A bidirectional target filtering model of speech coarticulation: Two-stage implementation for phonetic recognition," IEEE Transactions on Audio and Speech Processing, vol. 14, no. 1, pp. 256-265, January 2006a.

Microsoft Research

Deng, L., Wu, J., Droppo, J., and Acero, A. "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, vol. 13, no. 3, pp. 412–421, 2005.

Deng, L. and O'Shaughnessy, D. SPEECH PROCESSING – A Dynamic and Optimization-Oriented Approach, Marcel Dekker, 2003.

Deng, L. "Switching dynamic system models for speech articulation and acoustics," in Mathematical Foundations of Speech and Language Processing, pp. 115–134. Springer-Verlag, New York, 2003.

Deng, L. "Computational Models for Speech Production," in Computational Models of Speech Pattern Processing, pp. 199-213, Springer Verlag, 1999.

Deng, L., Ramsay, G., and Sun, D. "Production models as a structural basis for automatic speech recognition," Speech Communication, vol. 33, no. 2-3, pp. 93–111, Aug 1997.

Deng, L. and Sameti, H. "Transitional speech units and their representation by regressive Markov states: Applications to speech recognition," IEEE Transactions on speech and audio processing, vol. 4, no. 4, pp. 301–306, July 1996.

Deng, L., Aksmanovic, M., Sun, D., and Wu, J. "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 507-520, 1994.

Deng L. and Sun, D. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," Journal of the Acoustical Society of America, vol. 85, no. 5, pp. 2702-2719, 1994.

Deng, L. "A stochastic model of speech incorporating hierarchical nonstationarity," IEEE Transactions on Speech and Audio Processing, vol. 1, no. 4, pp. 471-475, 1993.

Deng, L. "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," Signal Processing, vol. 27, no. 1, pp. 65–78, 1992.

Deselaers, T., Hasan, S., Bender, O. and Ney, H. "A deep learning approach to machine transliteration," Proc. 4th Workshop on Statistical Machine Translation , pp. 233–241, Athens, Greece, March 2009.

Erhan, D., Bengio, Y., Courvelle, A., Manzagol, P., Vencent, P., and Bengio, S. "Why does unsupervised pre-training help deep learning?" J. Machine Learning Research, pp. 201-208, 2010.

Fine, S., Singer, Y. and Tishby, N. "The hierarchical hidden Markov model: Analysis and applications," Machine Learning, vol. 32, p. 41-62, 1998.

Gens, R. and Domingo, P. "Discriminative learning of sum-product networks," NIPS, 2012.

George, D. "How the Brain Might Work: A Hierarchical and Temporal Model for Learning and Recognition," Ph.D. thesis, Stanford University, 2008.

Gibson, M. and Hain, T. "Error approximation and minimum phone error acoustic model estimation," IEEE Trans. Audio, Speech, and Language Proc., vol. 18, no. 6, August 2010, pp. 1269-1279.

Glorot, X., Bordes, A., and Bengio, Y. "Deep sparse rectifier neural networks," Proc. AISTAT, April 2011.

Glorot, X. and Bengio, Y. "Understanding the difficulty of training deep feed-forward neural networks" Proc. AISTAT, 2010.

Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. "Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks," Proc. ICML, 2006.

Graves, A. "Sequence Transduction with Recurrent Neural Networks," Representation Learning Worksop, ICML 2012.

Graves, A., Mahamed, A., and Hinton, G. "Speech recognition with deep recurrent neural networks," Proc. ICASSP, 2013.

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

Hawkins, G., Ahmad, S. and Dubinsky, D. "Hierarchical Temporal Memory Including HTM Cortical Learning Algorithms," Numenta Tech. Report, December 10, 2010.

He, X., Deng, L., Chou, W. "Discriminative learning in sequential pattern recognition – A unifying review for optimization-oriented speech recognition," IEEE Sig. Proc. Mag., vol. 25, 2008, pp. 14-36.

He, X. and Deng, L. "Speech recognition, machine translation, and speech translation – A unifying discriminative framework," IEEE Sig. Proc. Magazine, Vol. 28, November, 2011.

He, X. and Deng, L. "Optimization in speech-centric information processing: Criteria and techniques," Proc. ICASSP, 2012.

He, X. and Deng, L. "Speech-centric information processing: An optimization-oriented approach," Proc. of the IEEE, 2013.

Heigold, G., Vanhoucke, V., Senior, A. Nguyen, P., Ranzato, M., Devin, M., and Dean, J. "Multilingual acoustic models using distributed deep neural networks," Proc. ICASSP, 2013.

Heigold, G., Ney, H., Lehnen, P., Gass, T., Schluter, R. "Equivalence of generative and log-liner models," IEEE Trans. Audio, Speech, and Language Proc., vol. 19, no. 5, February 2011, pp. 1138-1148.

Heintz, I., Fosler-Lussier, E., and Brew, C. "Discriminative input stream combination for conditional random field phone recognition," IEEE Trans. Audio, Speech, and Language Proc., vol. 17, no. 8, Nov. 2009, pp. 1533-1546.

Hifny, Y. and Renals, S. "Speech recognition using augmented conditional random fields," IEEE Trans. Audio, Speech, and Language Proc., vol. 17, no. 2, February 2009, pp. 354-365.

Hinton, G. and Salakhutdinov, R. "Discovering binary codes for documents by learning deep generative models," Topics in Cognitive Science, pp. 1-18, 2010.

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

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, November 2012.

Hinton, G., Krizhevsky, A., and Wang, S. "Transforming auto-encoders," Proc. Intern. Conf. Artificial Neural Networks, 2011.

Hinton, G. "A practical guide to training restricted Boltzmann machines," UTML Tech Report 2010-003, Univ. Toronto, August 2010.

Hinton, G., Osindero, S., and Teh, Y. "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, pp. 1527-1554, 2006.

Hinton, G. and Salakhutdinov, R. "Reducing the dimensionality of data with neural networks," Science, vol. 313. no. 5786, pp. 504 - 507, July 2006.

Hinton, G. "A better way to learn features," Communications of the ACM," Vol. 54, No. 10, October, 2011, pp. 94.

Huang, J., Li, J., Deng, L., and Yu, D. "Cross-language knowledge transfer using multilingual deep neural networks with shared hidden layers," Proc. ICASSP, 2013.

Huang, S. and Renals, S. "Hierarchical Bayesian language models for conversational speech recognition," IEEE Trans. Audio, Speech, and Language Proc., vol. 18, no. 8, November 2010, pp. 1941-1954.

Huang, E., Socher, R., Manning, C, and Ng, A. "Improving Word Representations via Global Context and Multiple Word Prototypes," Proc. ACL, 2012.

Hutchinson, B., Deng, L., and Yu, D. "A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition," Proc. ICASSP, 2012.

Hutchinson, B., Deng, L., and Yu, D. "Tensor deep stacking networks," IEEE Trans. Pattern Analysis and Machine Intelligence, 2013.

Jaitly, N. and Hinton, G. "Learning a better representation of speech sound waves using restricted Boltzmann machines," Proc. ICASSP, 2011.

Jaitly, N., Nguyen, P., and Vanhoucke, V. "Application of pre-trained deep neural networks to large vocabulary speech recognition," Proc. Interspeech, 2012.

Jarrett, K., Kavukcuoglu, K. and LeCun, Y. "What is the best multistage architecture for object recognition?" Proc. Intl. Conf. Computer Vision, pp. 2146–2153, 2009.

Jiang, H. and Li, X. "Parameter estimation of statistical models using convex optimization: An advanced method of discriminative training for speech and language processing," IEEE Signal Processing Magazine, vol. 27, no. 3, pp. 115–127, 2010.

Juang, B.-H., Chou, W., and Lee, C.-H. "Minimum classification error rate methods for speech recognition," IEEE Trans. On Speech and Audio Processing, vol. 5, pp. 257–265, 1997.

Kavukcuoglu, K., Sermanet, P., Boureau, Y., Gregor, K., Mathieu M., and LeCun, Y. "Learning Convolutional Feature Hierarchies for Visual Recognition," Proc. NIPS, 2010.

Ketabdar, H. and Bourlard, H. "Enhanced phone posteriors for improving speech recognition systems," IEEE Trans. Audio, Speech, and Language Proc., vol. 18, no. 6, August 2010, pp. 1094-1106.

Microsoft Research

Kingsbury, B. "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," Proc. ICASSP, 2009.

Kingsbury, B., Sainath, T., and Soltau, H. "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," Proc. Interspeech, 2012.

Krizhevsky, A., Sutskever, I. and Hinton, G. "ImageNet classification with deep convolutional neural Networks," Proc. NIPS 2012.

Kubo, Y., Hori, T., and Nakamura, A. "Integrating deep neural networks into structural classification approach based on weighted finite-state transducers," Proc. Interspeech, 2012.

Kurzweil R. How to Create a Mind. Viking Books, Dec., 2012.

Lang, K., Waibel, A., and Hinton, G. "A time-delay neural network architecture for isolated word recognition," Neural Networks, Vol. 3(1), pp. 23-43, 1990.

Larochelle, H. and Bengio, Y. "Classification using discriminative restricted Boltzmann machines," Proc. ICML, 2008.

Le, H., Allauzen, A., Wisniewski, G., and Yvon, F. "Training continuous space language models: Some practical issues," in Proc. of EMNLP, 2010, pp. 778–788.

Le, H., Oparin, I., Allauzen, A., Gauvain, J., and Yvon, F. "Structured output layer neural network language model," Proc. ICASSP, 2011.

Le, Q., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. "On optimization methods for deep learning," Proc. ICML, 2011.

Le, Q., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., Ng, A. "Building High-Level Features Using Large Scale Unsupervised Learning," Proc. ICML 2012.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. "Gradient-based learning applied to document recognition," Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.

LeCun, Y. and Bengio, Y. "Convolutional networks for images, speech, and time series," in The Handbook of Brain Theory and Neural Networks (M. Arbib, ed.), pp. 255- 258, Cambridge, Massachusetts: MIT Press, 1995.

LeCun, Y., Chopra S., Ranzato, M., and Huang, F. "Energy-based models in document recognition and computer vision," Proc. Intern. Conf. Document Analysis and Recognition (ICDAR), 2007.

Lee, C.-H. "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next-generation automatic speech recognition," Proc. ICSLP, 2004, p. 109-111.

Microsoft Research

Lee, H., Grosse, R., Ranganath, R., and Ng, A. "Unsupervised learning of hierarchical representations with convolutional deep belief networks," Communications of the ACM," Vol. 54, No. 10, October, 2011, pp. 95-103.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations," Proc. ICML, 2009.

Lee, H., Largman, Y., Pham, P., Ng, A. "Unsupervised feature learning for audio classification using convolutional deep belief networks," Proc. NIPS, 2010.

Lena, P., Nagata, K., and Baldi, P. "Deep spatiotemporal architectures and learning for protein structure prediction," Proc. NIPS, 2012.

Li, J., Yu, D., Huang, J., and Gong, Y. "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," Proc. IEEE SLT 2012.

Lin, H., Deng, L., Yu, D., Gong, Y., Acero, A., and C-H Lee, "A study on multilingual acoustic modeling for large vocabulary ASR." Proc. *ICASSP*, 2009.

Ling, Z., Richmond, K., and Yamagishi, J. "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," IEEE Trans. Audio, Speech, and Language Proc., Vol. 21, Jan, 2013.

Markoff, J. "Scientists See Promise in Deep-Learning Programs," New York Times, Nov 24, 2012.

Martens, J. "Deep learning with Hessian-free optimization," Proc. ICML, 2010.

Martens, J. and Sutskever, I. "Learning recurrent neural networks with Hessian-free optimization," Proc. ICML, 2011.

Mikolov, T. "Statistical Language Models based on Neural Networks," PhD thesis, Brno University of Technology, 2012.

Mikolov, T., Deoras, A., Povey, D., Burget, L., and Cernocky, J. "Strategies for training large scale neural network language models," Proc. IEEE ASRU, 2011.

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., and Khudanpur, S. "Recurrent neural network based language model," Proc. ICASSP, 2010, 1045–1048.

Minami, Y., McDermott, E. Nakamura, A. and Katagiri, S. "A recognition method with parametric trajectory synthesized using direct relations between static and dynamic feature vector time series," Proc. ICASSP, pp. 957-960, 2002.

Mnih, A. and Hinton G. "Three new graphical models for statistical language modeling," Proc. ICML, 2007, pp. 641-648.

Mnih, A. and Hinton G. "A scalable hierarchical distributed language model" Proc. NIPS, 2008, pp. 1081-1088.

Mohamed, A., Dahl, G. and Hinton, G. "Acoustic Modeling Using Deep Belief Networks", IEEE Trans. Audio, Speech, & Language Proc. Vol. 20 (1), January 2012.

Microsoft Research

Mohamed, A., Hinton, G., and Penn, G., "Understanding how deep belief networks perform acoustic modelling," Proc. ICASSP, 2012a.

Mohamed, A., Yu, D., and Deng, L. "Investigation of full-sequence training of deep belief networks for speech recognition," Proc. Interspeech, Sept. 2010.

Mohamed, A., Dahl, G., and Hinton, G. "Deep belief networks for phone recognition," in Proc. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.

Morgan, N. "Deep and Wide: Multiple Layers in Automatic Speech Recognition," IEEE Trans. Audio, Speech, & Language Proc. Vol. 20 (1), January 2012.

Morgan, N., 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, vol. 22, no. 5, pp. 81–88, Sep 2005.

Murphy, K. Machine Learning – A Probabilistic Perspective, The MIT Press, 2012.

Nair, V. and Hinton, G. "3-d object recognition with deep belief nets," Proc. NIPS, 2009.

Ney, H. "Speech translation: Coupling of recognition and translation," Proc. ICASSP, 1999.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. "Multimodal deep learning," Proc. ICML, 2011.

Ngiam, J., Chen, Z., Koh, P., and Ng, A. "Learning deep energy models," Proc. ICML, 2011.

Oliver, N., Garg, A., and Horvitz, E. "Layered Representations for Learning and Inferring Office Activity from Multiple Sensory Channels," Computer Vision and Image Understanding," vol. 96, pp. 163-180, 2004.

Olshausen, B. "Can 'Deep Learning' offer deep insights about Visual Representation?" NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2012.

Ostendorf, V. Digalakis, and O. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," IEEE Trans. Speech and Audio Proc., vol. 4, no. 5, September 1996.

Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," IEEE Trans. Audio, Speech, and Lang. Processing, Vol.17(3), pp. 423-435, 2009.

Peng, J., Bo, L., and Xu, J. "Conditional neural fields," Proc. NIPS, 2009.

Picone, P., 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," Proc. ICASSP, 1999.

Pinto, J., Garimella, S., Magimai-Doss, M., Hermansky, H., and Bourlard, H. "Analysis of MLP-based hierarchical phone posterior probability estimators," IEEE Trans. Audio, Speech, and Language Proc., vol. 19, no. 2, Feb. 2011.

Poggio. T. "How the Brain Might Work: The Role of Information and Learning in Understanding and Replicating Intelligence," In: Information: Science and Technology for the New Century, Editors: G. Jacovitt, A. Pettorossi, R. Consolo and V. Senni, Lateran University Press, pp. 45-61, 2007.

Poon, H. and Domingos, P. "Sum-product networks: A new deep architecture," Proc. Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, 2011. Barcelona, Spain.

Povey, D. and Woodland, P. "Minimum phone error and I-smoothing for improved discriminative training," Proc. ICASSP, 2002, pp. 105–108.

Prabhavalkar, R. and Fosler-Lussier, E. "Backpropagation training for multilayer conditional random field based phone recognition", Proc. ICASSP 2010, pp. 5534-5537.

Ranzato, M., Chopra, S. and LeCun, Y., and Huang, F.-J. "Energy-based models in document recognition and computer vision," Proc. International Conference on Document Analysis and Recognition (ICDAR), 2007.

Ranzato, M., Boureau, Y., and LeCun, Y. "Sparse Feature Learning for Deep Belief Networks," Proc. NIPS, 2007.

Ranzato, M., Susskind, J., Mnih, V., and Hinton, G. "On deep generative models with applications to recognition," Proc. CVPR, 2011.

Rennie, S., Hershey, H., and Olsen, P. "Single-channel multi-talker speech recognition — Graphical modeling approaches," IEEE Signal Processing Mag., vol. 33, pp. 66–80, 2010.

Rifai, S., Vincent, P., X. Muller, X. Glorot, and Y. Bengio, "Contractive autoencoders: Explicit invariance during feature extraction," Proc. ICML, 2011, pp. 833-840.

Robinson, A. "An application of recurrent nets to phone probability estimation," IEEE Trans. Neural Networks, Vol. 5, pp. 298-305, 1994.

Sainath, T., Ramabhadran, B., Picheny, M., Nahamoo, D., and Kanevsky, D., "Exemplar-Based Sparse Representation Features: From TIMIT to LVCSR," IEEE Transactions on Speech and Audio Processing, November 2011.

Sainath, T., Kingbury, B., Ramabhadran, B., Novak, P., and Mohamed, A. "Making deep belief networks effective for large vocabulary continuous speech recognition," Proc. IEEE ASRU, 2011.

Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. "Convolutional neural networks for LVCSR," Proc. ICASSP, 2013.

Salakhutdinov R. and Hinton, G. "Semantic hashing," Proc. SIGIR Workshop on Information Retrieval and Applications of Graphical Models, 2007.

Salakhutdinov R. and Hinton, G. "Deep Boltzmann machines," Proc. AISTATS, 2009.

Salakhutdinov R. and Hinton, G. "A better way to pretrain deep Boltzmann machines," Proc. NIPS, 2012.

Microsoft Research

Sarikaya, R., Hinton, G., Ramabhadran, B. "Deep belief nets for natural language call-routing," Proc. ICASSP, pp. 5680-5683, 2011.

Seide, F., Li, G., Chen, X., and Yu, D. "Feature engineering in context-dependent deep neural networks for conversational speech transcription," Proc. ASRU 2011, pp. 24-29.

Seide, F., Li, G., and Yu, D. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," Interspeech 2011, pp. 437-440.

Shannon, M., Zen, H., and Byrne W. "Autoregressive models for statistical parametric speech synthesis," IEEE Trans. Audio, Speech, Language Proc., Vol. 21, No. 3, 2013, pp. 587-597.

Sheikhzadeh, H. and Deng, L. "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," IEEE Trans. on Speech and Audio Processing, Vol. 2, pp. 80-91, 1994.

Siniscalchi, M., Yu, D., Deng, L., and Lee, C.-H. "Exploiting deep neural networks for detection-based speech recognition," Neurocomputing, 2013.

Siniscalchi, M., Svendsen, T., and Lee, C.-H. "A bottom-up modular search approach to large vocabulary continuous speech recognition," IEEE Trans. Audio, Speech, Language Proc., Vol. 21, 2013a.

Sivaram G. and Hermansky, H. "Sparse multilayer perceptron for phoneme recognition," IEEE Trans. Audio, Speech, & Language Proc. Vol. 20 (1), January 2012.

Snoek, J., Larochelle, H., and Adams, R. "Practical Bayesian Optimization of Machine Learning Algorithms," Proc. NIPS, 2012.

Socher, R. "New Directions in Deep Learning: Structured Models, Tasks, and Datasets," NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2012.

Socher, R., Lin, C., Ng, A., and Manning, C. "Learning continuous phrase representations and syntactic parsing with recursive neural networks," Proc. ICML, 2011.

Socher, R., Pennington, J., Huang, E., Ng, A., and Manning, C. "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions," Proc. EMNLP, 2011a.

Socher, R., Pennington, J., Huang, E., Ng, A., and Manning, C. "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, Proc. NIPS 2011b.

Socher, R., Bengio, Y., and Manning, C. "Deep learning for NLP," Tutorial at ACL, 2012, http://www.socher.org/index.php/DeepLearningTutorial/DeepLearningTutorial.

Stoyanov, V., Ropson, A. and Eisner, J. "Empirical Risk Minimization of Graphical Model Parameters Given Approximate Inference, Decoding, and Model Structure," Proc. AISTAT, 2011.

Srivastava, N. and Salakhutdinov R. "Multimodal learning with deep Boltzmann machines," Proc. NIPS, 2012.

Sutskever. I. "Training Recurrent Neural Networks," Ph.D. Thesis, University of Toronto, 2013.

Sutskever, I., Martens J., and Hinton, G. “Generating text with recurrent neural networks,” Proc. ICML, 2011.
Taylor, G., Hinton, G. E., and Roweis, S. “Modeling human motion using binary latent variables.” Proc. NIPS, 2007.
Tang, Y. and Eliasmith, C. “Deep networks for robust visual recognition,” Proc. ICML, 2010.
Taralba, A, Fergus R, and Weiss, Y. “Small codes and large image databases for recognition,” Proc. CVPR, 2008.
Tur, G., Deng, L., Hakkani-Tür, D., and X. He. “Towards deep understanding: Deep convex networks for semantic utterance classification,” Proc. ICASSP, 2012.
Vincent, P. “A connection between score matching and denoising autoencoder”, Neural Computation, Vol. 23, No. 7, pp. 1661-1674, 2011.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. “Stacked denoising autoencoders: Leaning useful representations in a deep network with a local denoising criterion,” J. Machine Learning Research, Vol. 11, 2010, pp. 3371-3408.
Vinyals, O., & Povey, D. “Krylov Subspace Descent for Deep Learning,” Proc. AISTAT, 2012.
Vinyals, O., Jia, Y., Deng, L., and Darrell, T. “Learning with recursive perceptual representations,” Proc. NIPS, 2012.
Vinyals O., and Ravuri, S. “Comparing multilayer perceptron to deep belief network tandem features for robust ASR,” Proc. ICASSP, 2011.
Welling, M., Rosen-Zvi, M., and Hinton, G. “Exponential family harmoniums with an application to information retrieval,” Proc. NIPS, Vol. 20, 2005.
Wohlmayr, M., Stark, M., Pernkopf, F. “A probabilistic interaction model for multi-pitch tracking with factorial hidden Markov model,” IEEE Trans. Audio, Speech, and Language Proc., vol. 19, no. 4, May. 2011.
Wolpert, D. “Stacked generalization,” Neural Networks, 5(2), pp. 241-259, 1992.
Xiao, L. and Deng, L. “A geometric perspective of large-margin training of Gaussian models,” IEEE Signal Processing Magazine, vol. 27, no. 6, pp. 118-123, IEEE, November 2010.
Yamin, S., Deng, L., Wang, Y., and Acero, A. “An integrative and discriminative technique for spoken utterance classification,” IEEE Trans. Audio, Speech, and Language Proc., 2008.
Yang, D., Furui, S. “Combining a two-step CRF model and a joint source channel model for machine transliteration,” Proc. ACL, Uppsala, Sweden, 2010, pp. 275-280.
Yu, D., Deng, L., and Seide, F. “The deep tensor neural network with applications to large vocabulary speech recognition,” IEEE Trans. Audio, Speech, Lang. Proc., 2013.
Yu, D. and Deng, L. “Efficient and effective algorithms for training single-hidden-layer neural networks,” Pattern Recognition Letters, 2012.

Microsoft Research

Yu, D., Seide, F., Li, G., Deng, L. "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," Proc. ICASSP 2012.

Yu, D., Siniscalchi, S., Deng, L., and Lee, C. "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition", Proc. ICASSP 2012.

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

Yu, D. and Deng, L. "Deep learning and its applications to signal and information processing," IEEE Signal Processing Magazine, January 2011, pp. 145-154.

Yu, D. and Deng, L. "Accelerated parallelizable neural networks learning algorithms for speech recognition," Proc. Interspeech 2011.

Yu, D., Deng, L., Li, G., and F. Seide. "Discriminative pretraining of deep neural networks," U.S. Patent Filing, Nov. 2011.

Yu, D. and Deng, L. "Deep-structured hidden conditional random fields for phonetic recognition," Proc. Interspeech, Sept. 2010.

Yu, D., Wang, S., Karam, Z., Deng, L. "Language recognition using deep-structured conditional random fields," Proc. ICASSP, 2010, pp. 5030-5033.

Yu, D., Wang, S., Deng, L., "Sequential labeling using deep-structured conditional random fields", J. of Selected Topics in Signal Processing, 2010a.

Yu, D., Li, J.-Y., and Deng, L. "Calibration of confidence measures in speech recognition," IEEE Trans. Audio, Speech and Language, 2010b.

Yu, D., Deng, L., and Dahl, G.E., "Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition," NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, Dec. 2010c.

Yu, D., Deng, D., Wang, S., "Learning in the Deep-Structured Conditional Random Fields," NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

Yu, D, Deng, L., Gong, Y. and Acero, A. "A novel framework and training algorithm for variable-parameter hidden Markov models," IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 7, September 2009, pp. 1348-1360.

Yu, D., Deng, L., Liu, P., Wu, J., Gong, Y., and Acero, A. "Cross-lingual speech recognition under runtime resource constraints," Proc. ICASSP, 2009.

Yu, D. and Deng, L. "Solving nonlinear estimation problems using Splines," IEEE Signal Processing Magazine, vol. 26, no. 4, pp. 86-90, July 2009.

Zamora-Martínez, F., Castro-Bleda, M., España-Boquera, S. "Fast evaluation of connectionist language models," Intern. Conf. Artificial Neural Networks, 2009, pp. 144-151.

Zen, H., Nankaku, Y., and Tokuda, K. "Continuous stochastic feature mapping based on trajectory HMMs," IEEE Trans. Audio, Speech, and Language Proc., vol. 19, no. 2, Feb. 2011, pp. 417-430.

Zen, H. Gales, M. J. F. Nankaku, Y. Tokuda, K. "Product of experts for statistical parametric speech synthesis," IEEE Trans. Audio, Speech, and Language Proc., vol. 20, no. 3, March, 2012, pp. 794-805.

Zweig, G. and Nguyen, P. "A segmental CRF approach to large vocabulary continuous speech recognition," Proc. ASRU, 2009.

Microsoft Research

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio, Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, in *Interspeech 2013*, August 2013

Ossama Abdel-Hamid, Li Deng, and Dong Yu, Exploring convolutional neural network structures and optimization techniques for speech recognition, in *Proc. Interspeech, Lyon, France*, August 2013

Ossama Abdel-Hamid, Li Deng, Dong Yu, and Hui Jiang, Deep segmental neural networks for speech recognition, in *Proc. Interspeech, Lyon, France*, August 2013

George Dahl, Jack W. Stokes, Li Deng, and Dong Yu, Large-Scale Malware Classification Using Random Projections and Neural Networks, in *Proceedings IEEE Conference on Acoustics, Speech, and Signal Processing*, IEEE SPS, 26 May 2013

Li Deng, Geoffrey Hinton, and Brian Kingsbury, New types of deep neural network learning for speech recognition and related applications: An overview, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013*, May 2013

Po-Sen Huang, Li Deng, Mark Hasegawa-Johnson, and Xiaodong He, Random Features for Kernel Deep Convex Network, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero, Recent Advances in Deep Learning for Speech Research at Microsoft, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013

Xiaodong He, Li Deng, Dilek Hakkani-Tur, and Gokhan Tur, Multi-Style Adaptive Training for Robust Cross-Lingual Spoken Language Understanding, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013

Jennifer Gillenwater, Xiaodong He, Jianfeng Gao, and Li Deng, End-To-End Learning of Parsing Models for Information Retrieval, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, CROSS-LANGUAGE KNOWLEDGE TRANSFER USING MULTILINGUAL DEEP NEURAL NETWORK WITH SHARED HIDDEN LAYERS, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013

Li Deng, Xiaodong He, and Jianfeng Gao, Deep Stacking Networks for Information Retrieval, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013

Li Deng, Ossama Abdel-Hamid, and Dong Yu, A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013

Po-Sen Huang, Kshitiz Kumar, Chaojun Liu, Yifan Gong, and Li Deng, PREDICTING SPEECH RECOGNITION CONFIDENCE USING DEEP LEARNING WITH WORD IDENTITY AND SCORE FEATURES, in *Proc. ICASSP*, May 2013

Hamid Palangi, Rabab Ward, and Li Deng, Using deep stacking network to improve structured compressive sensing with multiple measurement vectors, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013

Zhen-Hua Ling, Li Deng, and Dong Yu, Modeling Spectral Envelopes Using Restricted Boltzmann Machines For Statistical Parametric Speech Synthesis, in *ICASSP 2013*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013

Xiaodong He and Li Deng, Speech-Centric Information Processing: An Optimization-Oriented Approach, in *Proceedings of the IEEE*, vol. 31 May 2013.

Microsoft Research

"DBN vs DBN" (for fun)

From: Geoffrey Hinton [mailto:geoffrey.hinton@gmail.com]
Sent: Tuesday, January 17, 2012 9:33 AM
To: Li Deng
Subject: DBNs are beating DBNs

http://acronyms.thefreedictionary.com/DBN Acronym	Definition
DBN	1, 5-Diazabicyclo(4.3.0)Non-5-Ene (chemical compound)
DBN	Doing Business - Not
DBN	Dialog Broadband Networks (Dialog Telekom PLC; Sri Lanka)
DBN	De Bonis Non (Legal: appointment of a personal representative to a vacancy)
DBN	Divisible by None (band)
DBN	**Deep Belief Network (machine learning)**
DBN	**Dynamic Bayes Network**
DBN	Data Bus Network
DBN	Dial-Back Number
DBN	Day Beacon
DBN	Domain-Border Node
DBN	Digital Billboard Network (Australia)
DBN	Drunk Before Noon
DBN	District Borough Number (New York City Department of Education school identifier)
DBN	Database Notification
DBN	Directed Bipartite Network