Google™

Acoustic Modeling and Deep Learning

June 19th, 2013

Vincent Vanhoucke

A quick introduction

- Former Technical Lead of the Speech Recognition Quality Research team.
- Now a member of the Deep Learning Infrastructure team.

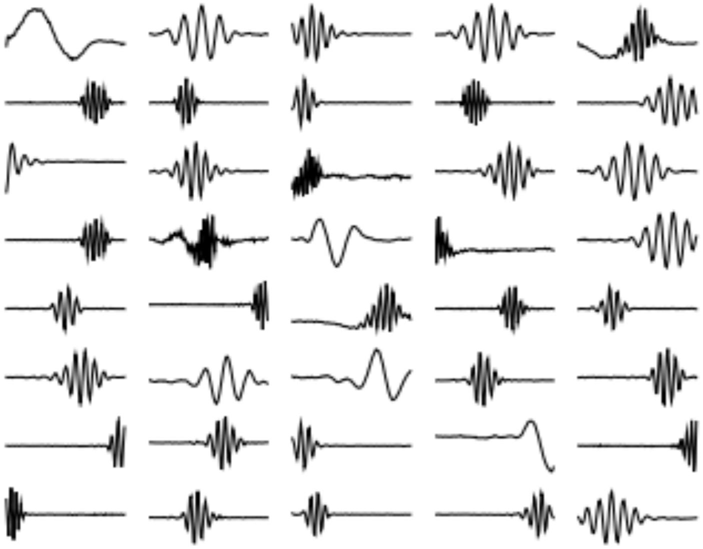
A personal story

Summer 2011…

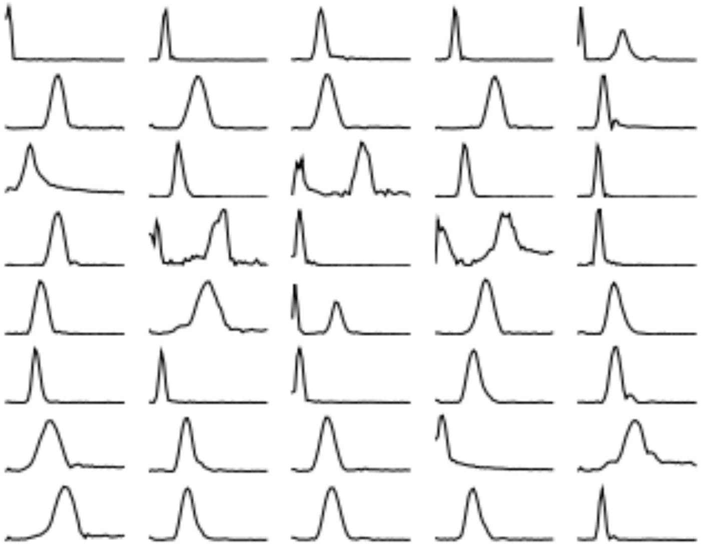
The (better) Internship Story

- Geoff Hinton Student:

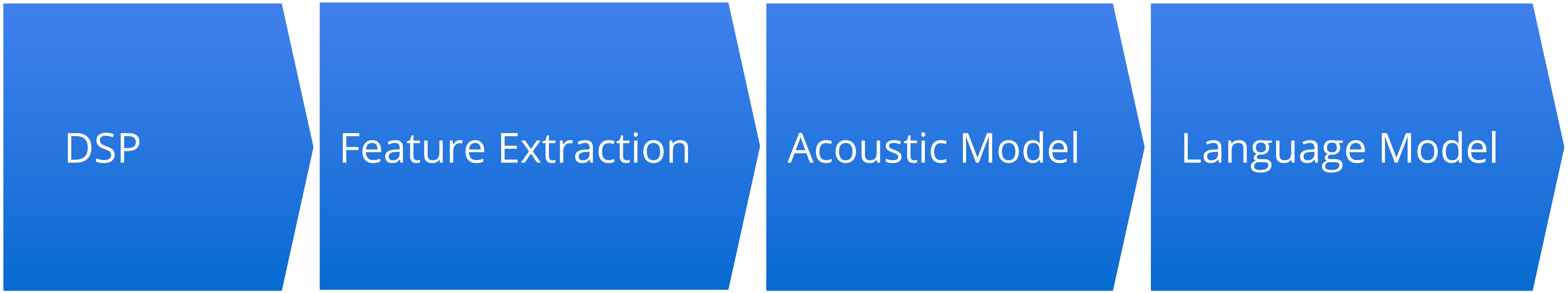
  Navdeep Jaitly, Geoffrey E. Hinton. Learning a better representation of speech soundwaves using restricted Boltzmann machines. ICASSP 2011



- Competitive results on TIMIT
- Deep Belief Network (**DBN**, generative model) used to pre-train a Deep Neural Network (**DNN**, discriminative model).
- NO feature engineering.



Speech Recognition



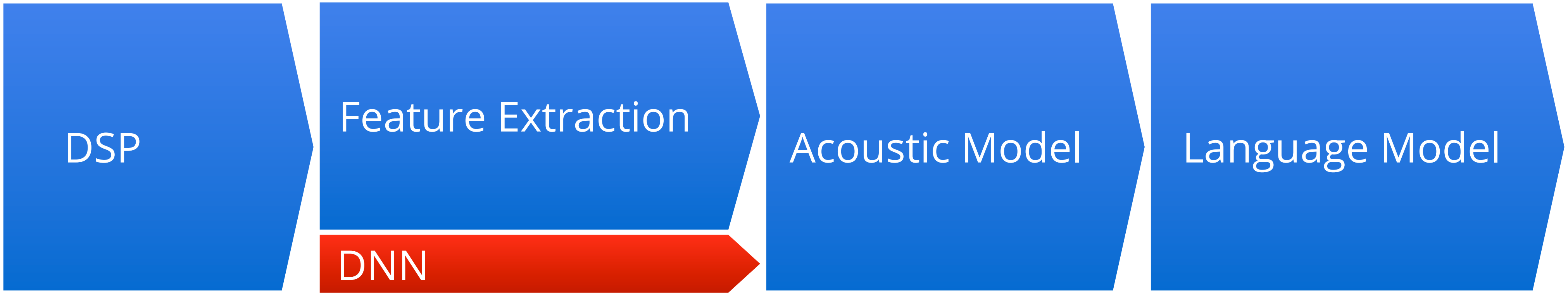
DSP

Feature Extraction

Acoustic Model

Language Model

Speech Recognition + Deep Neural Networks?



DSP

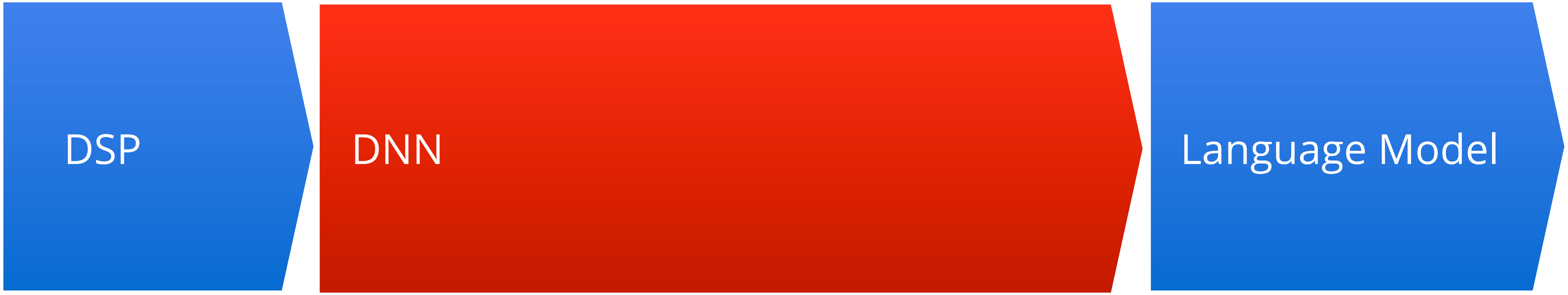
Feature Extraction

DNN

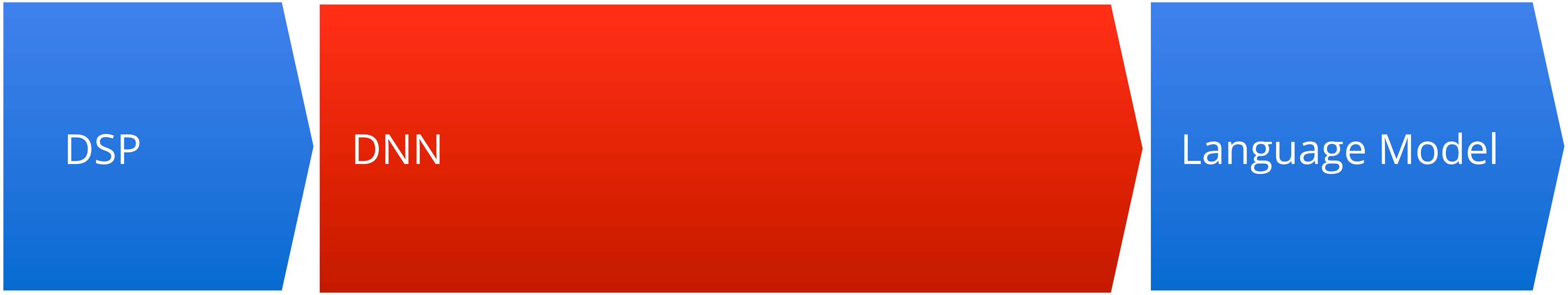
Acoustic Model

Language Model

Speech Recognition + Deep Neural Networks?



Speech Recognition + Deep Neural Networks!



DSP

DNN

Language Model

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.

A historical detour

Everything old is new again...

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.

- Recurrent Neural Networks

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

- 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

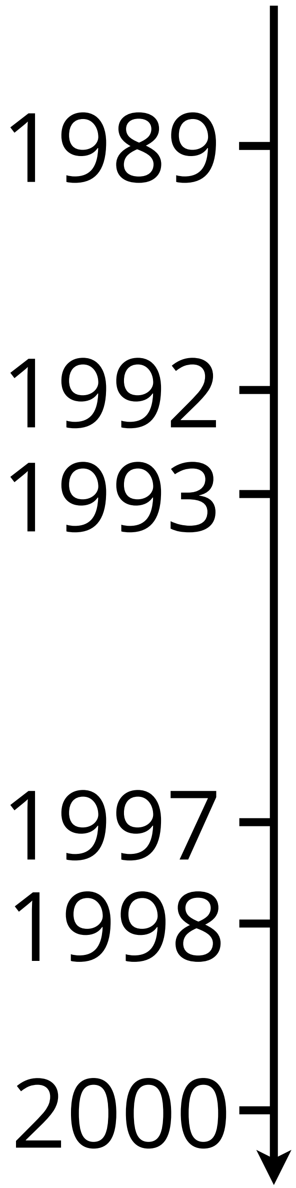
  Mike Schuster, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE Transactions on Signal Processing, 45, no. 11 (1997): 2673-2681.

- Hierarchical Neural Networks

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

- TANDEM

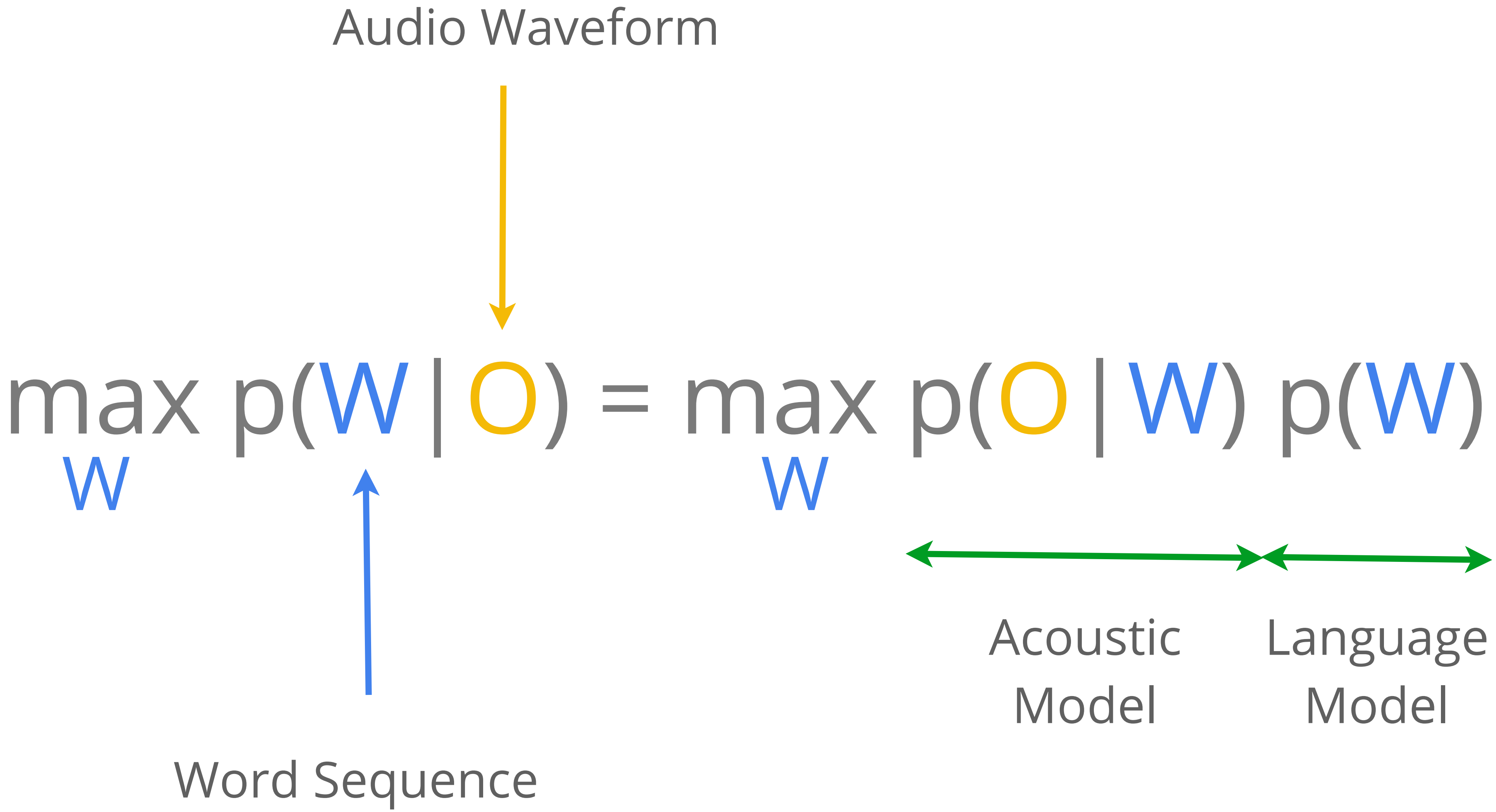
  Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000.



Neural Networks for Speech in the 00's

Speech Recognition

A very quick and inaccurate introduction...



Audio Waveform

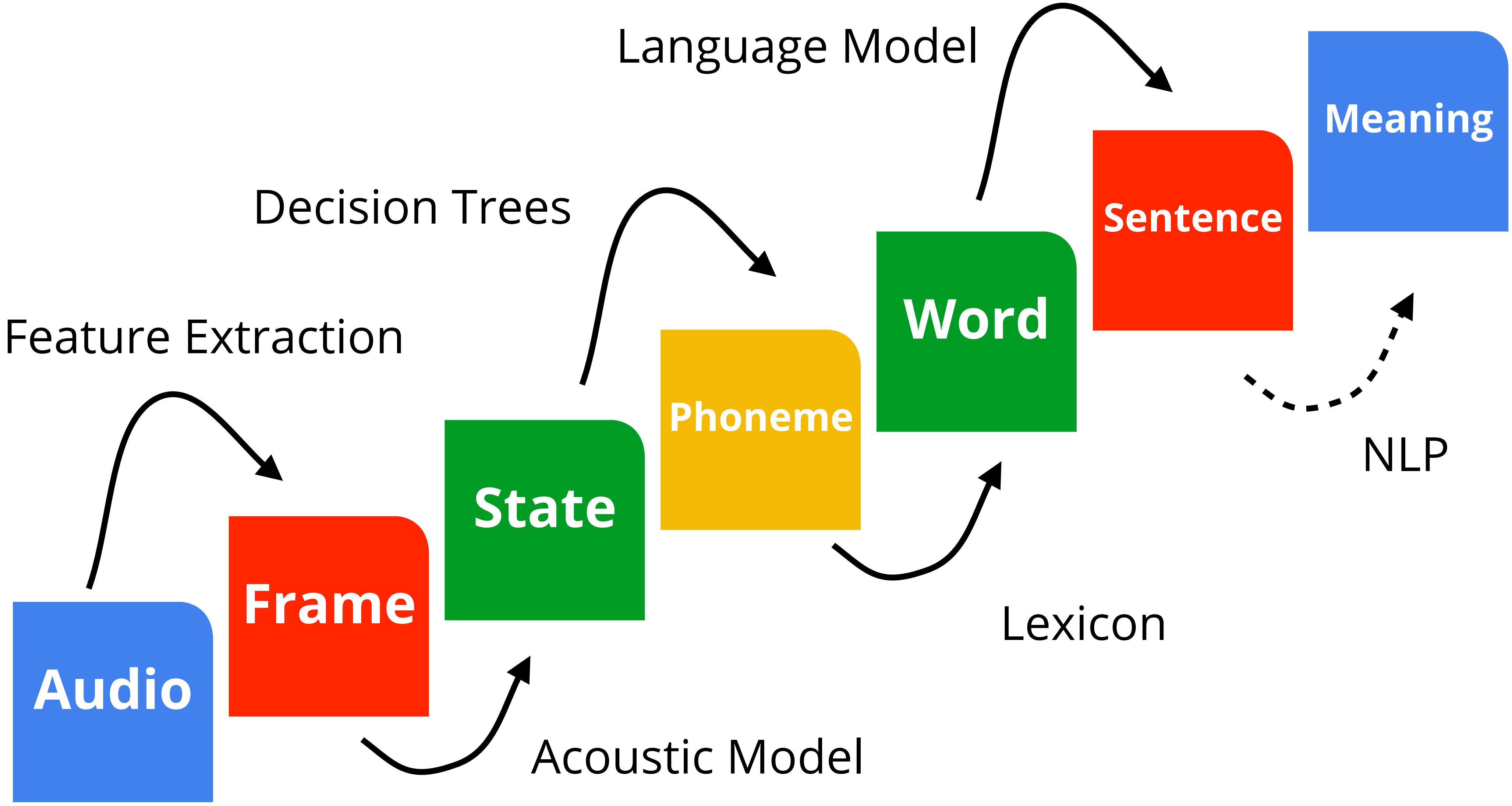
$$\max_{W} p(W|O) = \max_{W} p(O|W)\, p(W)$$

Word Sequence

Acoustic Model

Language Model

Speech Recognition as Probabilistic Transduction



Language Model

Decision Trees

Feature Extraction

Audio

Frame

State

Phoneme

Word

Sentence

Meaning

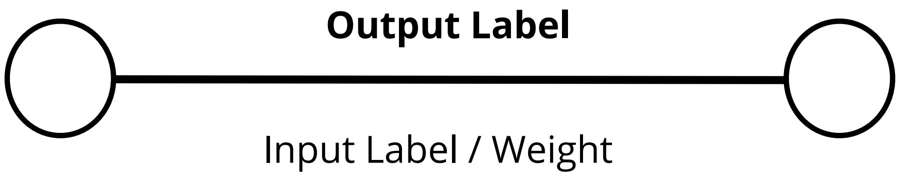
NLP

Lexicon

Acoustic Model

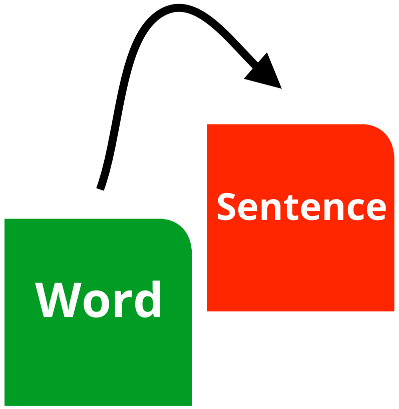
Weighted Finite State Transducers (FSTs)

- One of many ways to express this ‘probabilistic transduction’ idea.
- A mathematically sound way to express probabilistic graphs and algorithms over them. (e.g. Viterbi, forward-backward)
- Powerful algorithms to optimize these graphs.
- Pretty pictures:

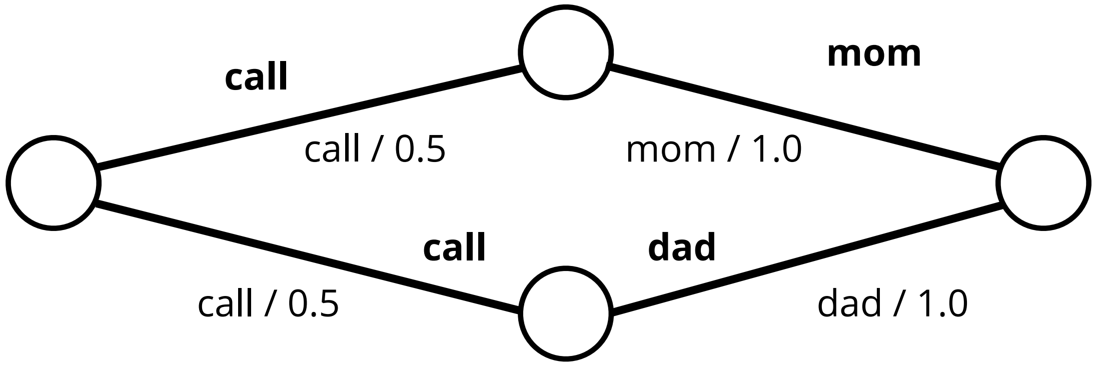


Mehryar Mohri, Fernando Pereira, and Michael Riley. "Weighted finite-state transducers in speech recognition." Computer Speech & Language 16, no. 1 (2002): 69-88.

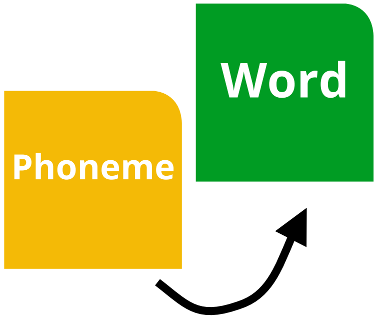
Language Model



- Toy Example:
  - "call mom"
  - "call dad"

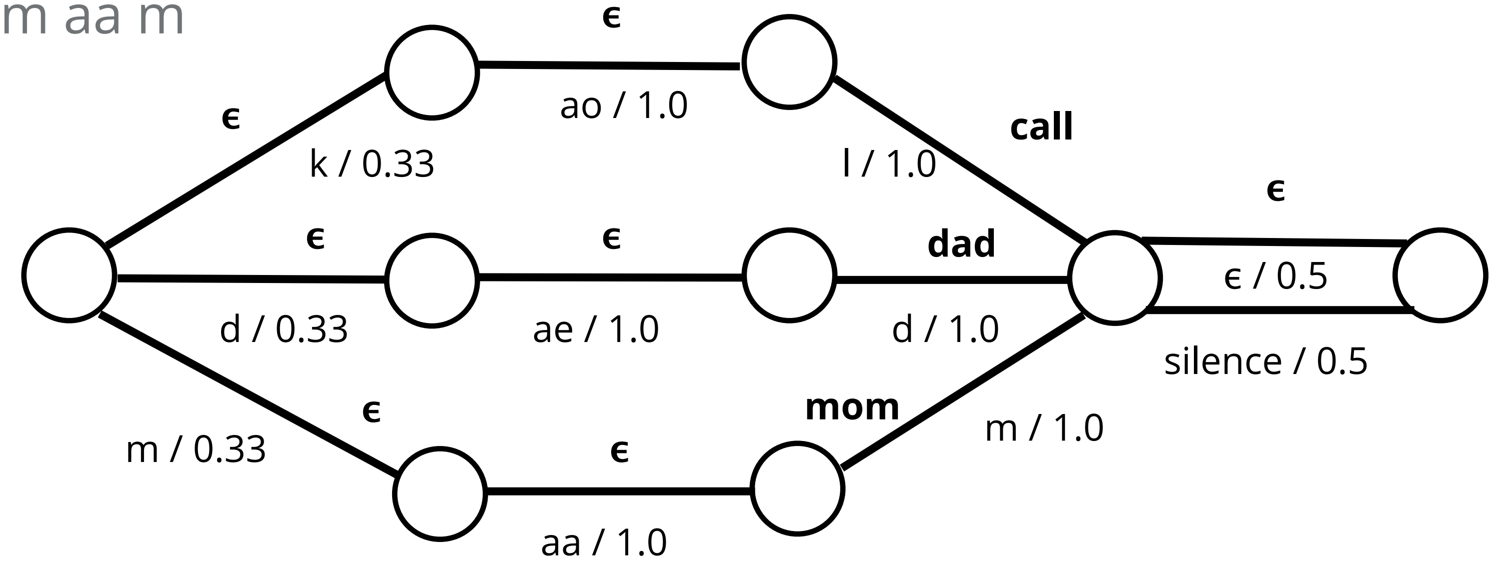


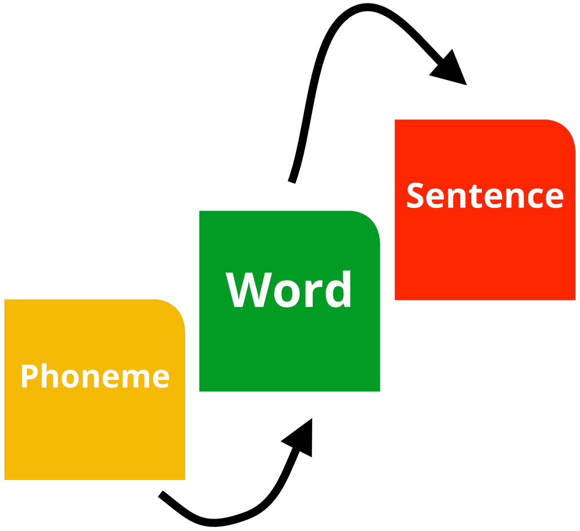
- In reality:
  - 100M+ tokens
  - estimated using billions of training examples.



Lexicon

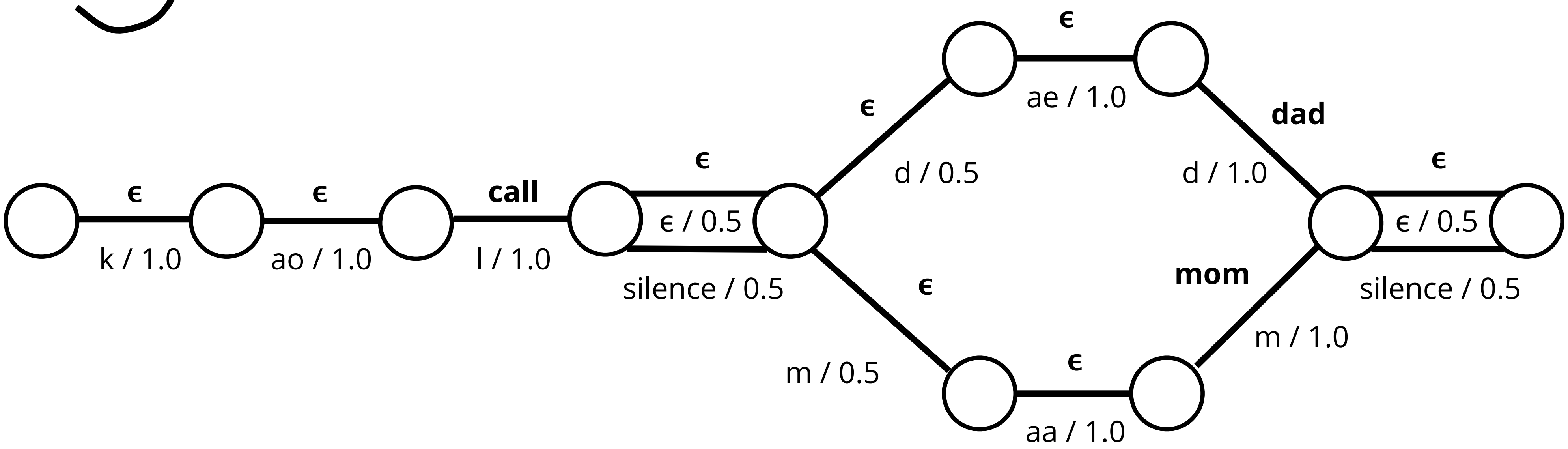
- call: k ao l
- dad: d ae d
- mom: m aa m



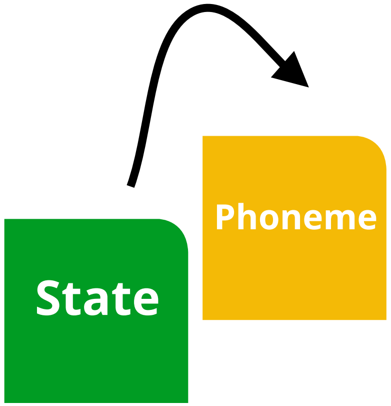


Transduction via Composition

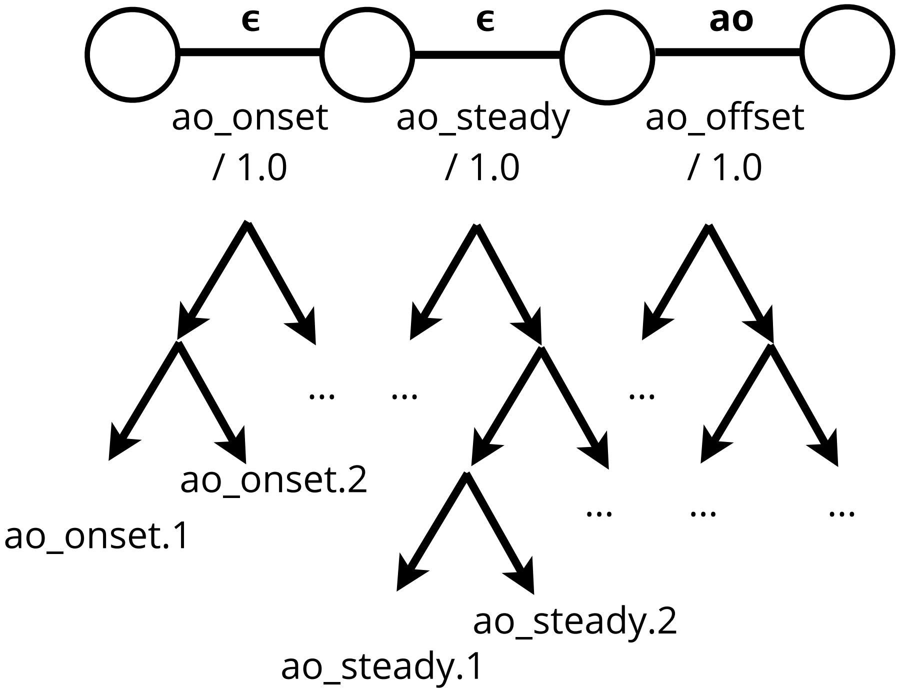
- Map *output* labels of Lexicon to *input* labels of Language Model.
- Join and optimize end-to-end graph.



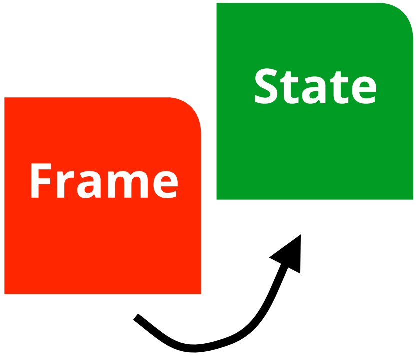
Other operations: Minimization, Determinization, Epsilon removal, Weight pushing.



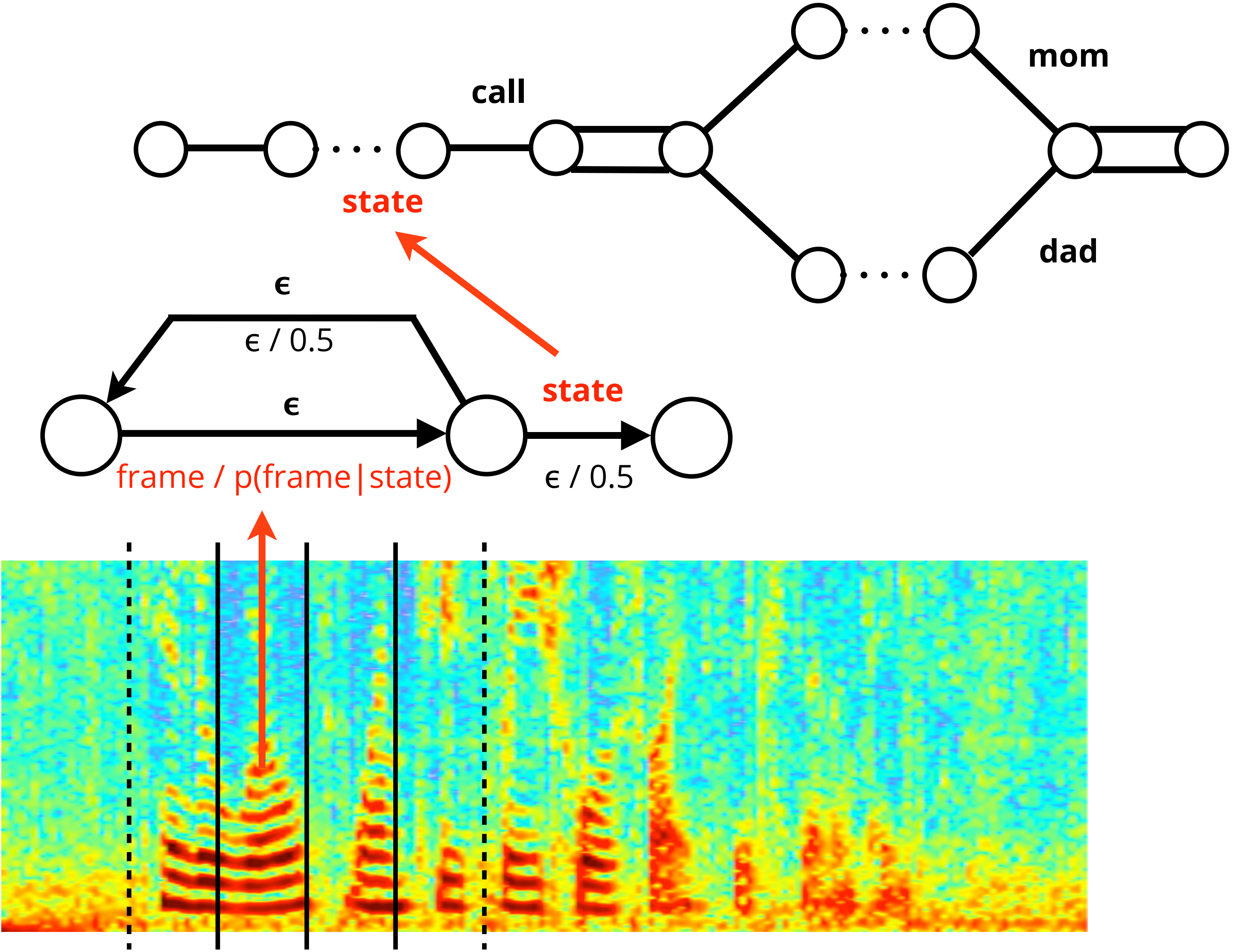
Connecting to the Acoustic Model



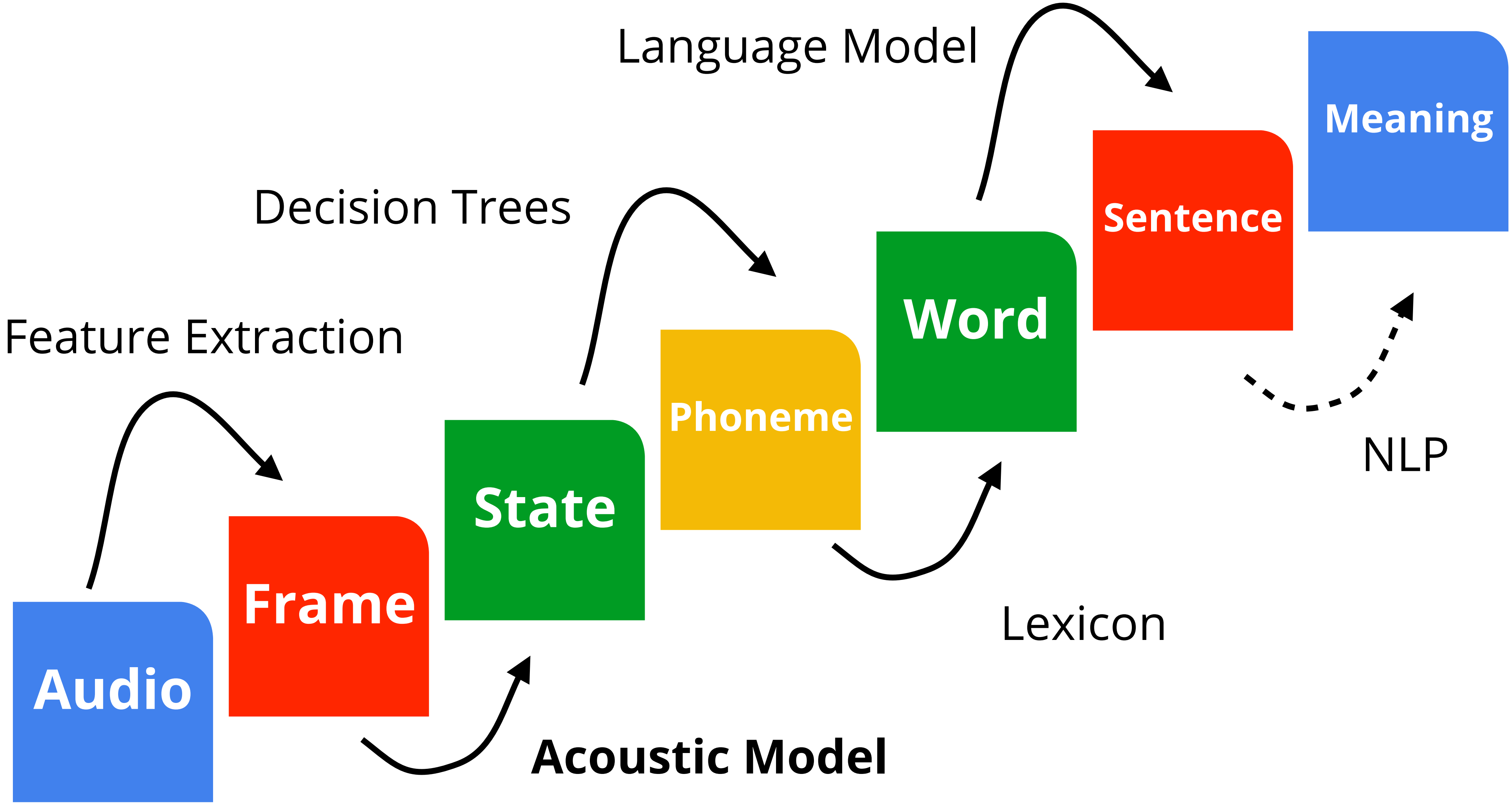
- Phonemes are mapped to a richer inventory of phonetic states.
- ~O(100) states per phoneme.
  - Too few: not enough modeling power.
  - Too many: overfitting.
- Also a (ugly) FST composition operation.
- These 1000 - 10000 states are the targets the acoustic model will attempt to predict, and the hidden states of the Hidden Markov Model (HMM).

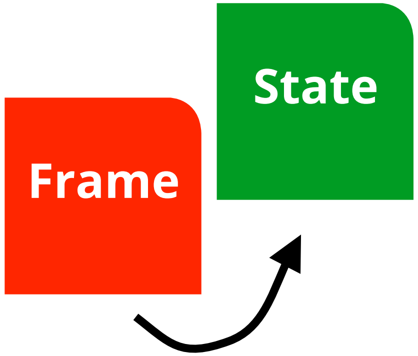


Time Alignment using HMMs



Speech Recognition Pipeline





Acoustic Modeling: p(frame|state)

- HMM is a generative model: need to model the likelihood of the data.
- General prediction setting:
  - Input frames: window of audio, typically in spectral form. ~500 real values.
  - Output states: 1000 - 10000 acoustic states.
- Can be treated as a simple classification problem, but:
  - The whole probability function matters due to combination with language model scores and pruning effects,
  - It is often beneficial to use an utterance-level training criterion (more on this later).

Gaussian Mixture Models

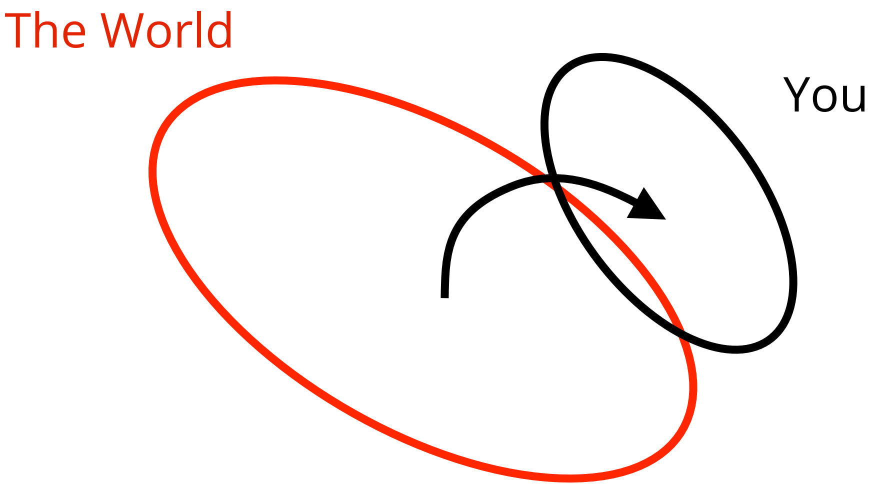
$$p\left(\text{frame}|\text{state}\right) = \sum_i w_i \mathcal{G}\left(\text{Feature}\left(\text{frame}\right), \mu_i, \sigma_i\right)$$

- Why on Earth?
- Training scales very well with both data and compute power. Can easily train 50M+ parameters on 1B frames in a few days.
- Inference is very cheap:
  - Only compute the Gaussians for states that are active in the search.
  - Assuming diagonal covariances, a GMM is a few L2 norms and a set of log-sums (≈ max).
  - Large bag of tricks to find which Gaussians to compute and which to discard.
- Bootstrapping is easy and robust: use a single Gaussian per state, chunk your utterances in equal parts, alternate realignment and parameter estimation.
- There are effective approaches to train these models discriminatively: MMI, bMMI, dMMI, MCE, MPE,...

GMMs and Speaker Adaptation

- Many very effective and very practical techniques:

  MLLR, fMLLR, MAP, CAT, SAT, Eigenvoices, iVectors…
- Large gains. Key to early successful deployments of dictation systems.
- Too successful?
- …

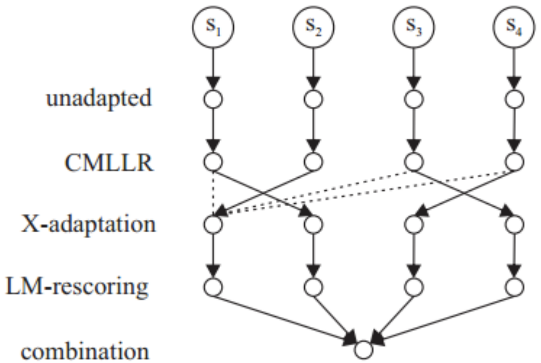


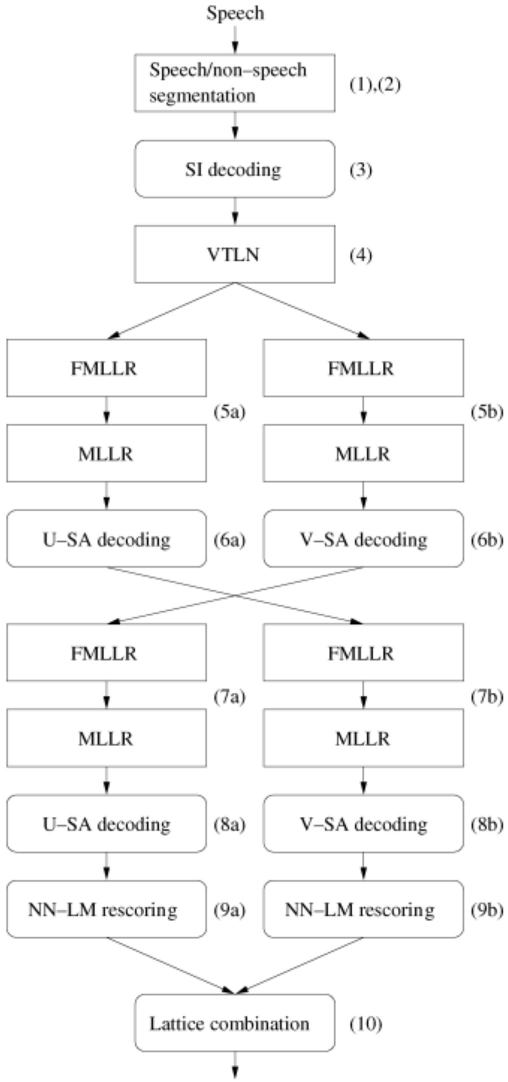
Recognize, adapt, re-recognize, re-adapt...

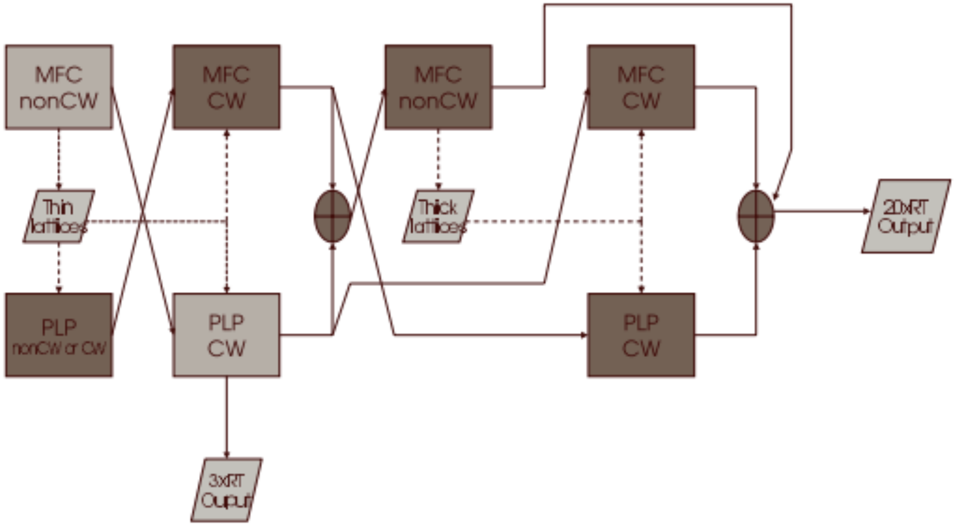
A sample of state-of-the-art systems in the late 200Xs.

Complexity fueled by success of speaker adaptation (and DARPA $$).

Huge barrier to entry.







Gaussian Mixture Models: the Bad

$$p\left(\text{frame}|\text{state}\right) = \sum_i w_i \mathcal{G}\left(\text{Feature}\left(\text{frame}\right), \mu_i, \sigma_i\right)$$

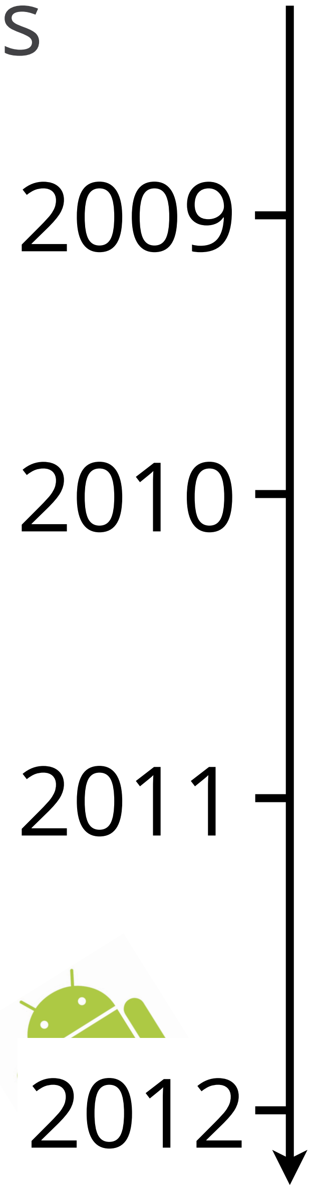
- GMMs as ‘Universal Approximators’ is a lie.
- In practice, features need to be dense, low-dimension, Gaussian-like, uncorrelated.
- Full covariance statistics too expensive: only diagonal Gaussians are used in practice.
- Approximating full covariances using a mixture of diagonal Gaussians: another lie.
- Huge constraints on the feature engineering: MFCCs, PLPs, bottleneck features.
- Many many tricks to compensate for flawed assumptions: STC, subspace methods (my thesis…sigh).

Historical detour, part II

Everything old is new again, …again…

The Second Coming of Neural Networks

- First proof of concept:
  Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. "Deep belief networks for phone recognition." In NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. 2009.
- Li Deng at MSR reviews paper and doesn’t believe a word. Asks for confusion tables to be produced as proof.
  Abdel-rahman Mohamed, and Geoffrey Hinton. "Phone recognition using Restricted Boltzmann Machines." ICASSP 2010.
- Summers 2010/2011: Dahl, Mohamed, Jaitly intern at Microsoft, IBM and Google. Make large vocabulary systems work.
- Summer 2011: Frank Seide publishes the first extensive paper demonstrating the gains:
  Frank Seide, Gang Li, and Dong Yu, Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, Interspeech 2011.
- Summer 2012: Google launches the first user-facing system using DNNs on Android.



Why the revival?

- **The spark:** Deep Belief Networks.
  Given early successes, DBNs were justification enough to take a second look. They since have been shown to be essentially irrelevant.
- **The catalyst:** GPUs.
  Our early Summer 2011 experiment would have taken a year on a CPU, and nobody at the time was seriously looking at parallel implementations.
- **The fuel:** Data!
  Training neural networks is only forgiving if you don't have to worry about overtraining. Lots of data means you can predictably hit good performance without extensive hyperparameter tuning.
  New techniques such as dropout somewhat alleviate that issue.

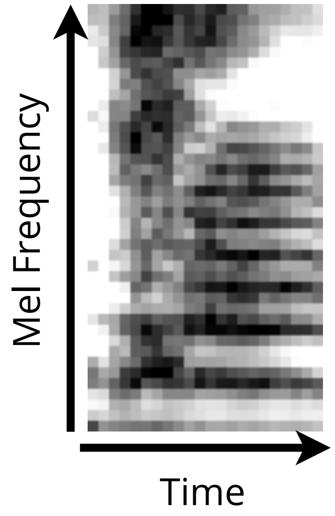
Deep Networks for Speech

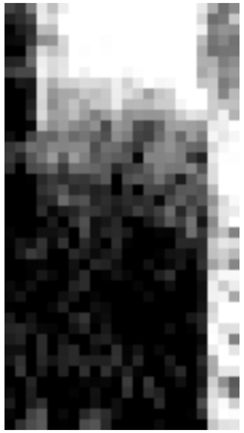
The recipe...

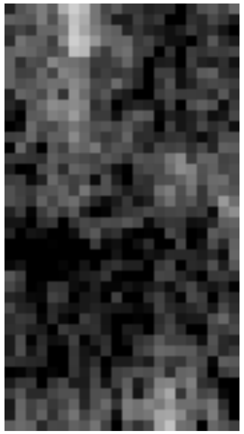
‘Hybrid’ system

$$p\left(\mathrm{frame}|\mathrm{state}\right) \sim \frac{p\left(\mathrm{state}|\mathrm{frame}\right)}{p\left(\mathrm{state}\right)}$$

- Any discriminative classifier can be plugged into p(state|frame).
- A very good baseline:
  - Deep neural network:
    - 4-10 layers, 1000-3000 nodes / layer,
    - Rectified linear activations: y=max(0,x)
    - Full connectivity between layers,
    - Softmax output.
  - Features:
    - 25ms window of audio, extracted every 10ms.
    - log-energy of 40 Mel-scale filterbanks, stacked for 10-30 frames.



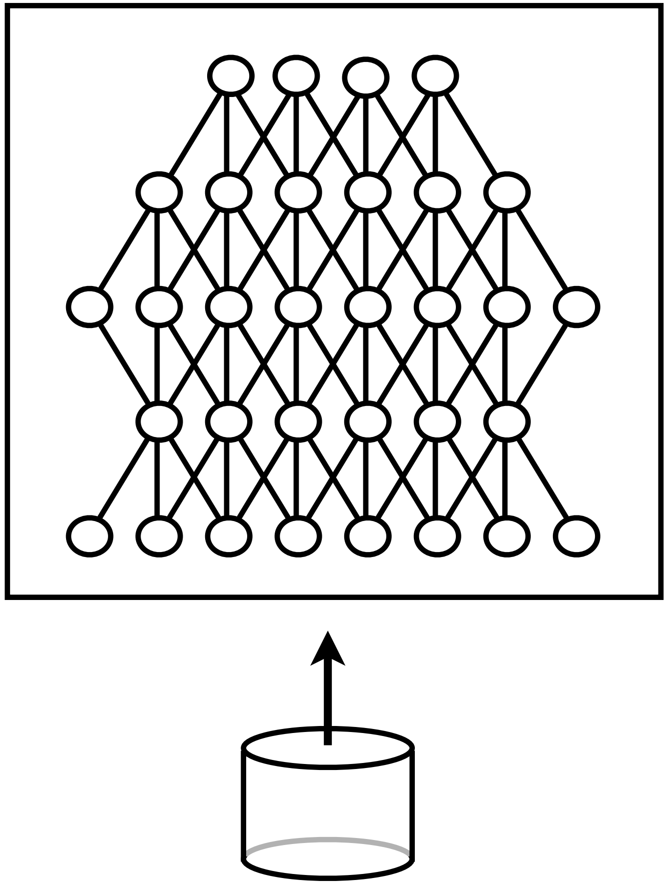




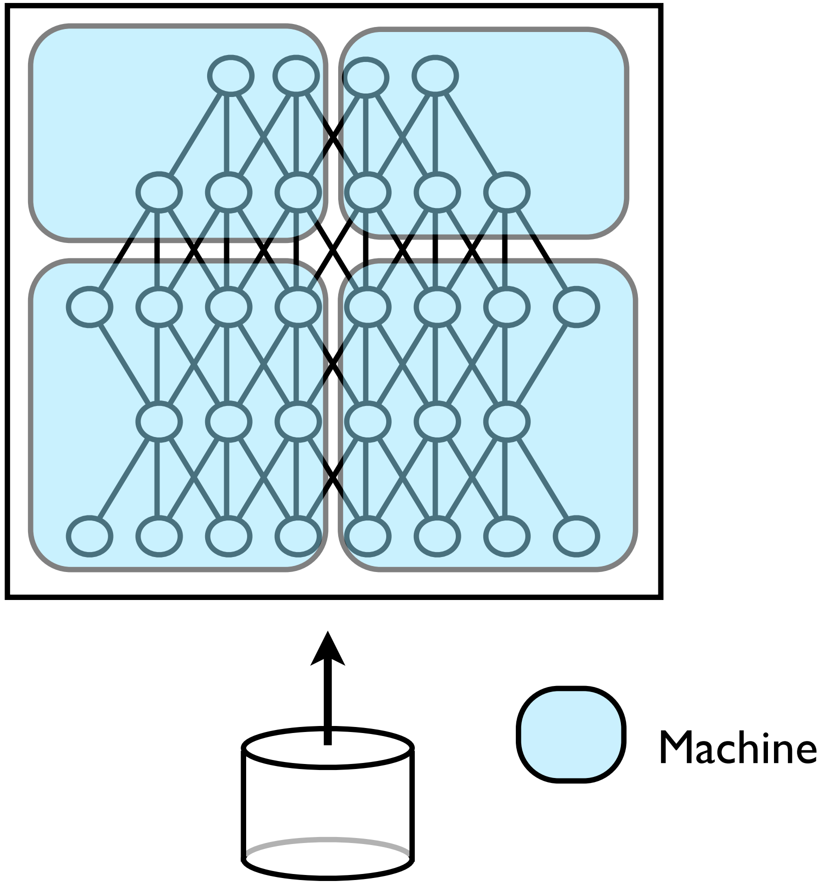
Frame-by-frame Training

- Stochastic Gradient Descent on GPU.
  - Mini-batches (< 256 samples).
  - Exponential or step-wise scheduling.
- Distributed Asynchronous SGD on multiple processors:
  - GPU: Xie Chen, Adam Eversole, Gang Li, Dong Yu, and Frank Seide, Pipelined Back-Propagation for Context-Dependent Deep Neural Networks, Interspeech 2012
  - CPU: Jeff Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc Le, Mark Mao, Marc'Aurelio Ranzato, Antrew Senior, Paul Tucker, Ke Yang, Andrew Ng Large Scale Distributed Deep Networks. NIPS 2012.
- Batch methods using 2nd order optimization: L-BFGS, Hessian-free:
  - James Martens. Deep Learning via Hessian-free Optimization, ICML 2010.
  - Oriol Vinyals, Dan Povey. Krylov Subspace Descent for Deep Learning, AISTATS 2012.

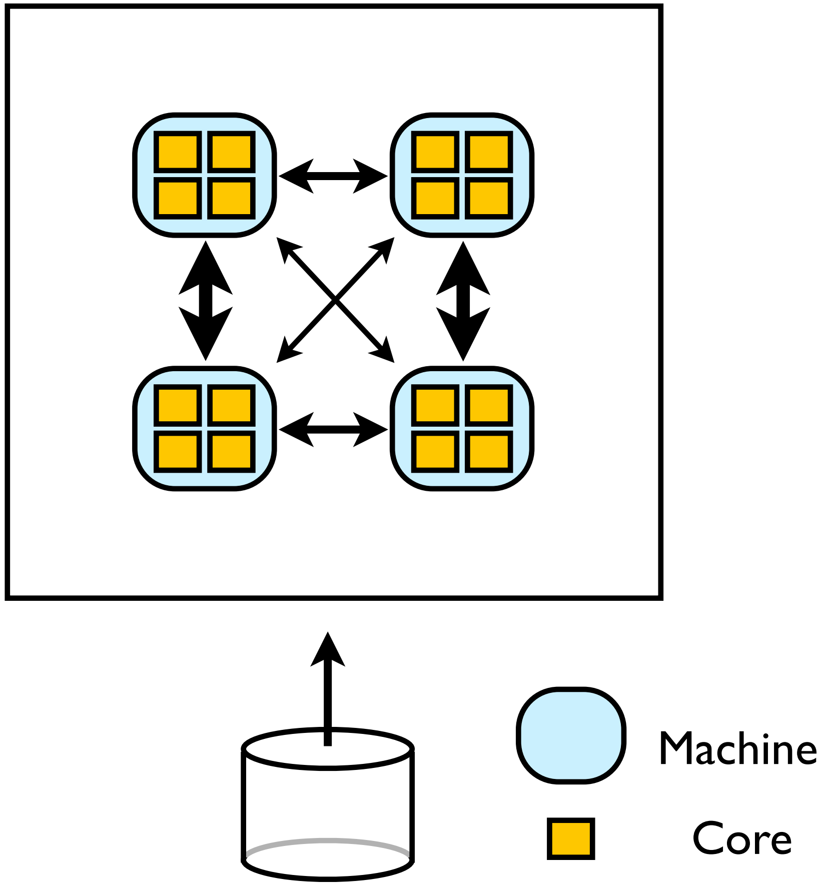
Parallelizing Deep Networks on CPUs



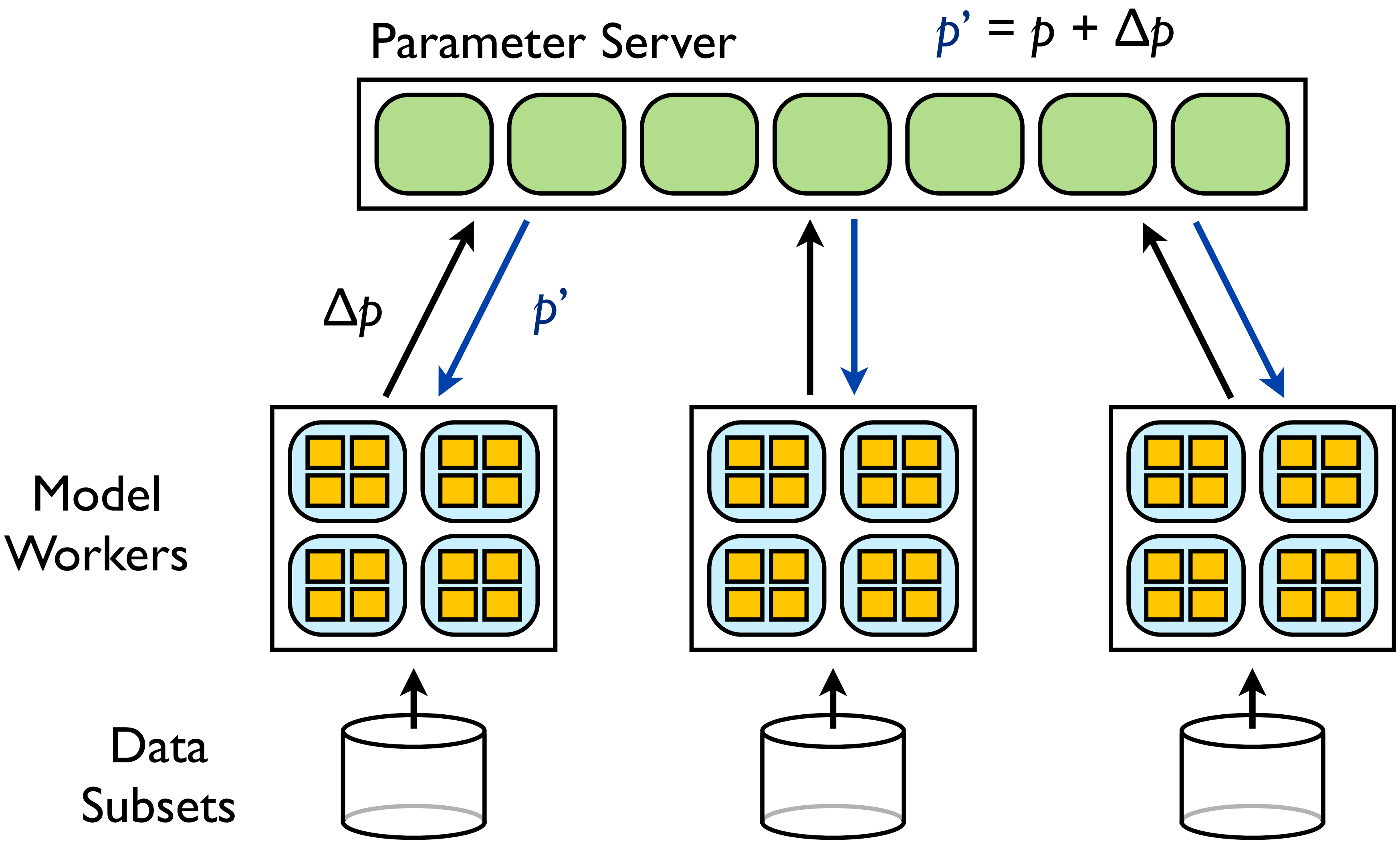
Parallelizing Deep Networks on CPUs



Parallelizing Deep Networks on CPUs



Distributed Asynchronous SGD

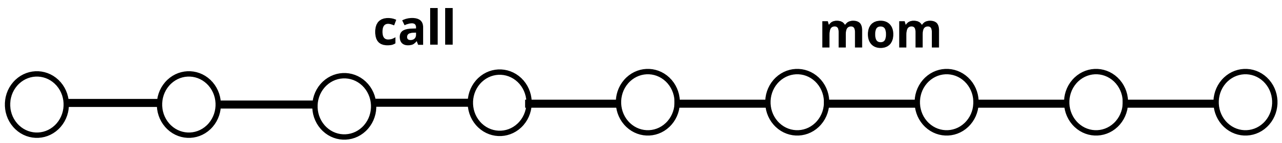


Sequence-based Discriminative Training

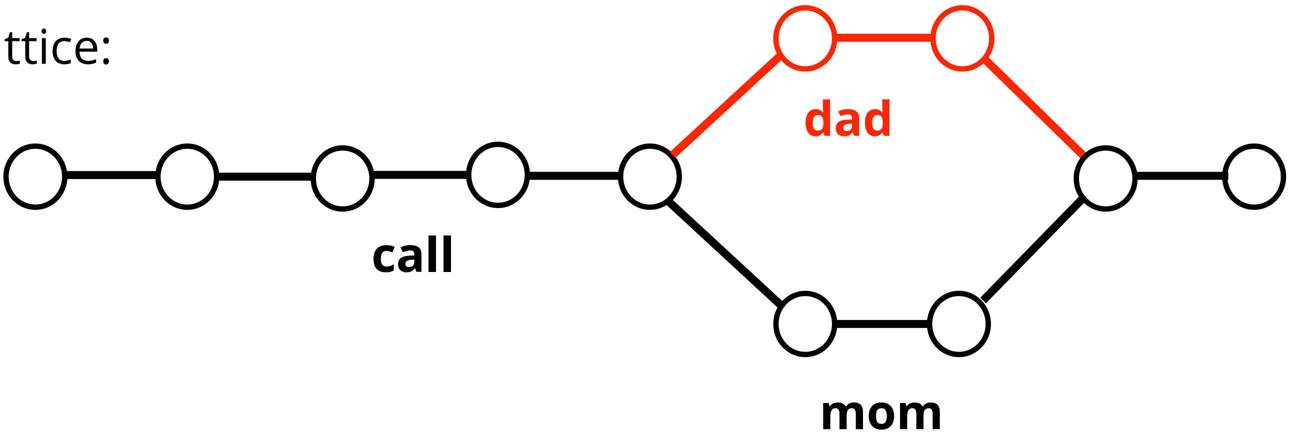
- Glaring issues with frame-based training:
  - 0/1 loss over tied states puts phones in competition with themselves.
  - Very weak connection with actual word error rate.
- Direct optimization of word error rate (edit distance over word strings) not easy: non-smooth, non-additive loss.
- Main ideas:
  - Compare the true alignments to the recognition alignments.
  - Derive a consistent smooth loss that takes into account word, phoneme, or state-level errors.
  - Back-propagate derivative of the loss.

Sequence-based Discriminative Training

Truth based on forced alignment:



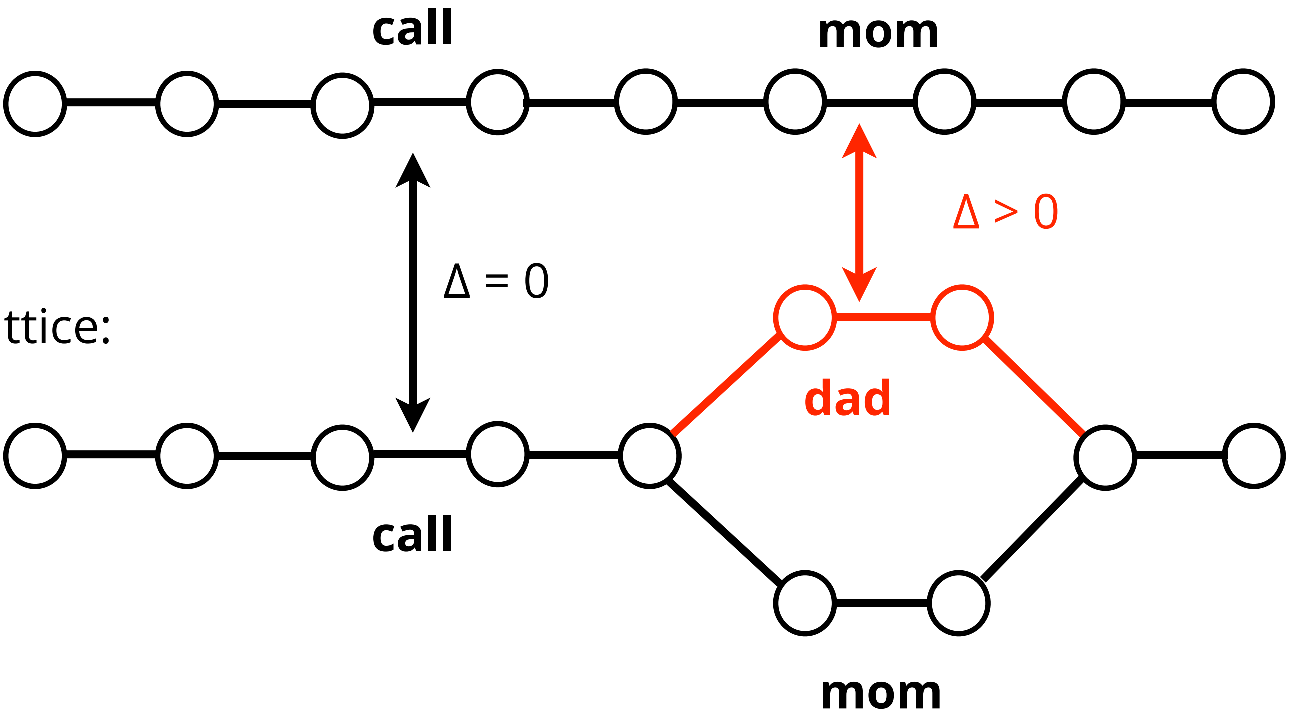
Recognition lattice:



Sequence-based Discriminative Training

Truth based on forced alignment:

Recognition lattice:



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

Speaker Adaptation

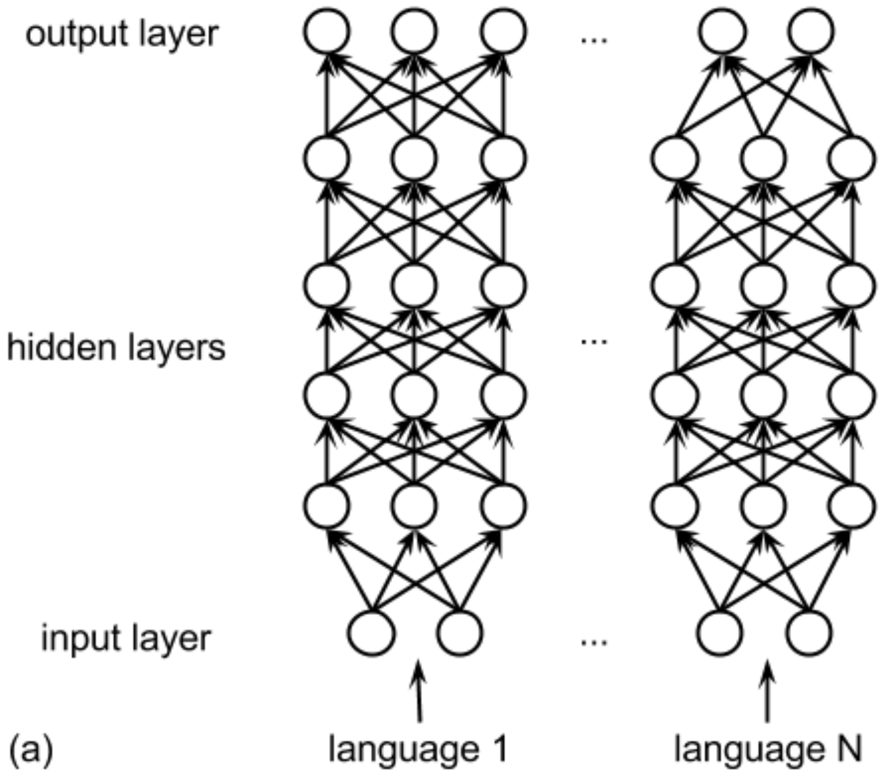
- Improved generalization with DNNs seems to have squashed many of the gains from speaker adaptation.
- Adaptation works well on small nets, but gains vanish as the networks grow.
- Good riddance!

Number of Parameters	Relative WER improvement	Sources (ICASSP 2013)
< 10M	32%, 10%	H. Liao, O. Abdel-Hamid
31M	15%	D. Yu
45M	7%	D. Yu
60M	5%	H. Liao

Speaker Adaptation of Context Dependent Deep Neural Networks, Hank Liao, ICASSP 2013

Multilingual Speech Recognition

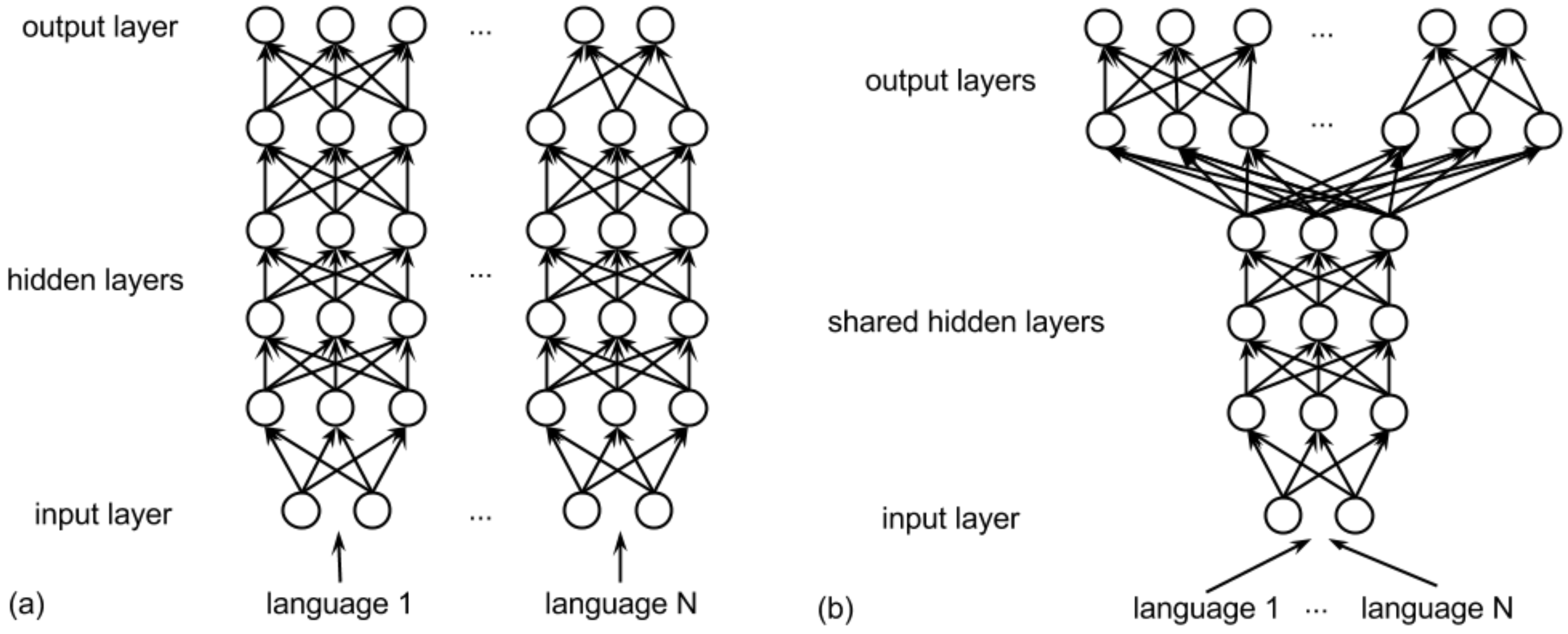
- One of the unexpected successes of deep architectures so far.
- Transfer learning and multitask learning work very well:



Multilingual acoustic models using distributed deep neural networks, Georg Heigold, Vincent Vanhoucke, Andrew Senior, Patrick Nguyen, Marc'aurelio Ranzato, Matthieu Devin, Jeff Dean, ICASSP 2013.

Multilingual Speech Recognition

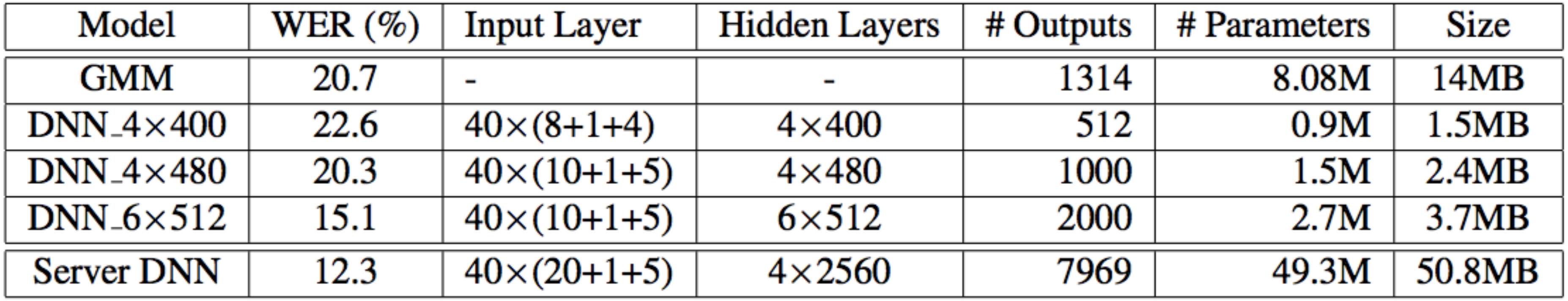
- One of the unexpected successes of deep architectures so far.
- Transfer learning and multitask learning work very well:



Multilingual acoustic models using distributed deep neural networks, Georg Heigold, Vincent Vanhoucke, Andrew Senior, Patrick Nguyen, Marc'aurelio Ranzato, Matthieu Devin, Jeff Dean, ICASSP 2013.

Embedded Large Vocabulary Recognition

- Excellent performance with mere 4MB DNN models.
- End-to-end recognizer takes ~40MB and runs a full dictation model in real-time on a 2011 Android phone.

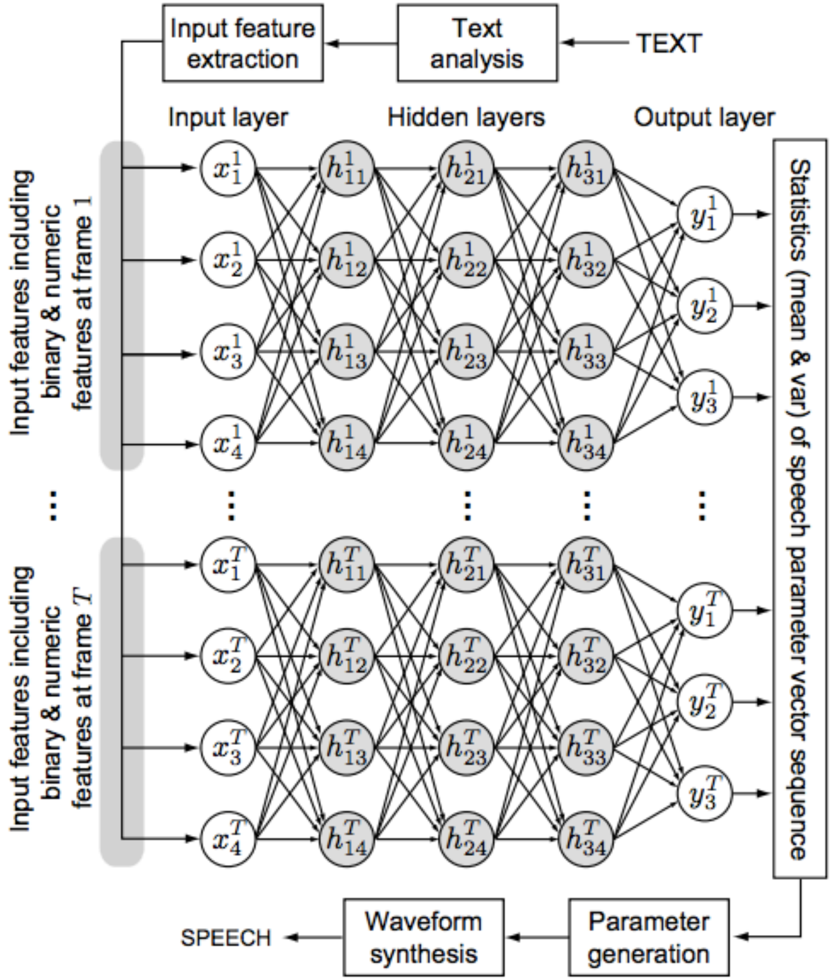


Model	WER (%)	Input Layer	Hidden Layers	# Outputs	# Parameters	Size
GMM	20.7	-	-	1314	8.08M	14MB
DNN_4×400	22.6	40×(8+1+4)	4×400	512	0.9M	1.5MB
DNN_4×480	20.3	40×(10+1+5)	4×480	1000	1.5M	2.4MB
DNN_6×512	15.1	40×(10+1+5)	6×512	2000	2.7M	3.7MB
Server DNN	12.3	40×(20+1+5)	4×2560	7969	49.3M	50.8MB

Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices, Xin Lei, Andrew Senior, Alexander Gruenstein and Jeffrey Sorensen. To appear in Interspeech 2013.

Text-to-Speech

- Use DNNs to generate parameters of a HMM synthesis engine.



Input feature extraction
Text analysis
TEXT
Input layer
Hidden layers
Output layer
Input features including binary & numeric features at frame 1
Input features including binary & numeric features at frame T
x_1^1 x_2^1 x_3^1 x_4^1
h_{11}^1 h_{12}^1 h_{13}^1 h_{14}^1
h_{21}^1 h_{22}^1 h_{23}^1 h_{24}^1
h_{31}^1 h_{32}^1 h_{33}^1 h_{34}^1
y_1^1 y_2^1 y_3^1
x_1^T x_2^T x_3^T x_4^T
h_{11}^T h_{12}^T h_{13}^T h_{14}^T
h_{21}^T h_{22}^T h_{23}^T h_{24}^T
h_{31}^T h_{32}^T h_{33}^T h_{34}^T
y_1^T y_2^T y_3^T
Statistics (mean & var) of speech parameter vector sequence
SPEECH
Waveform synthesis
Parameter generation

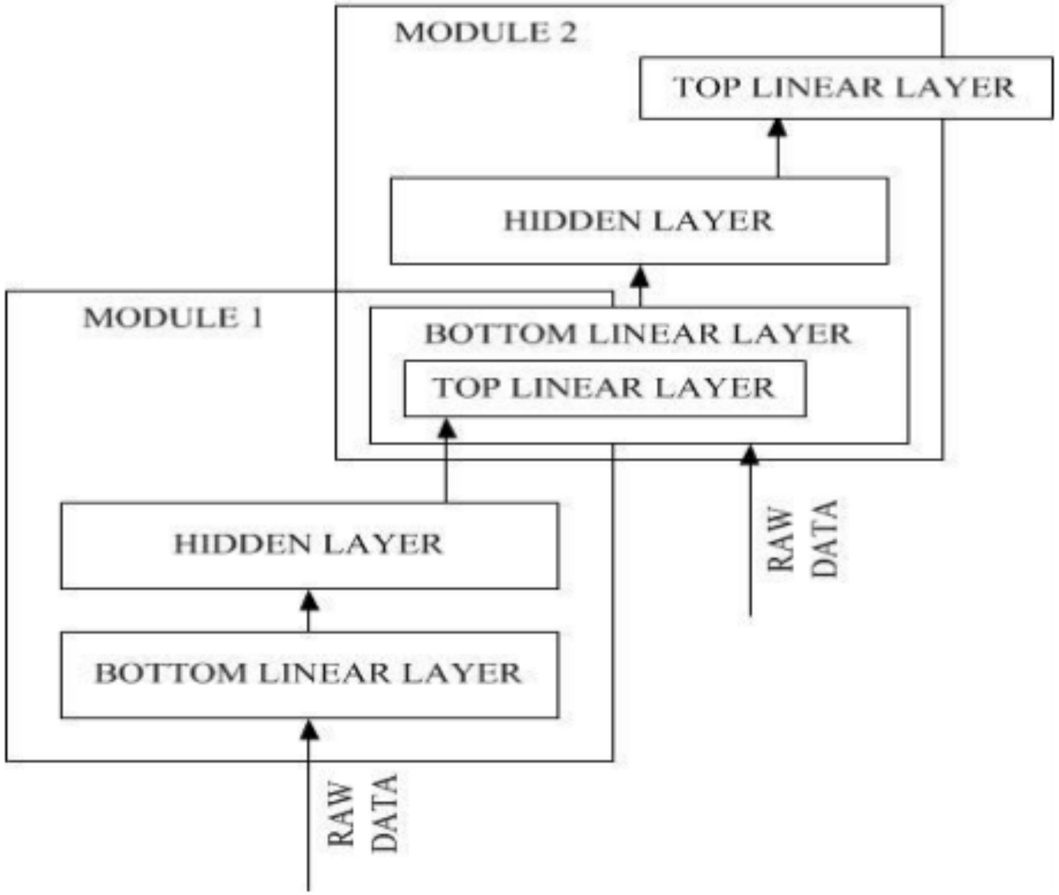
Statistical Parametric Speech Synthesis Using Deep Neural Networks, Heiga Zen, Andrew Senior, Mike Schuster, ICASSP 2013

Alternative Architectures

A sampler…

Deep Convex Networks

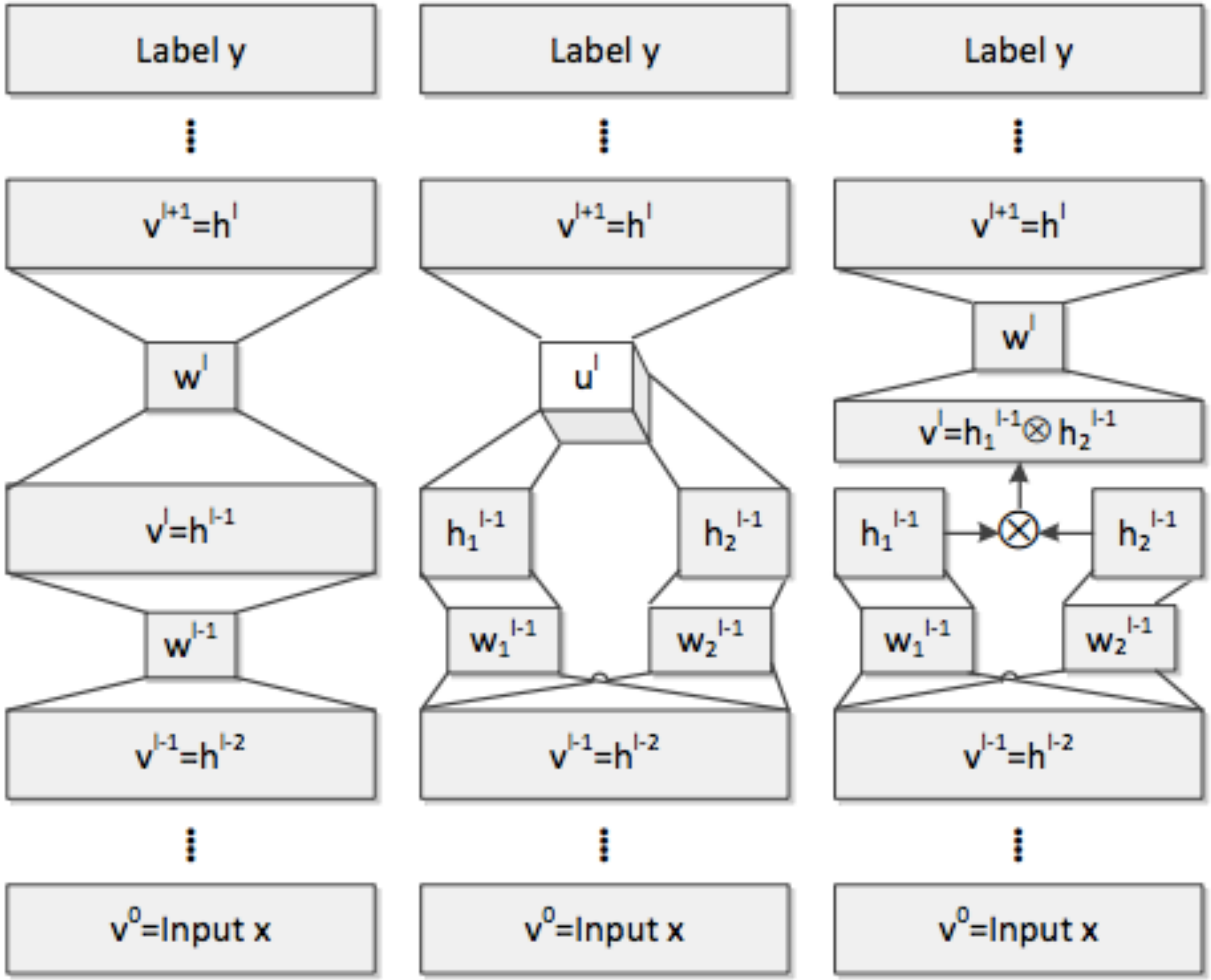
- A simple approach to build a deep model using only convex optimization techniques.
- Successfully 'convexifying' the problem is an interesting line of research.
- Very competitive and fast to train.
- So far, best performance still obtained with non-convex fine tuning and many more layers than DNNs.



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

Deep Tensor Networks

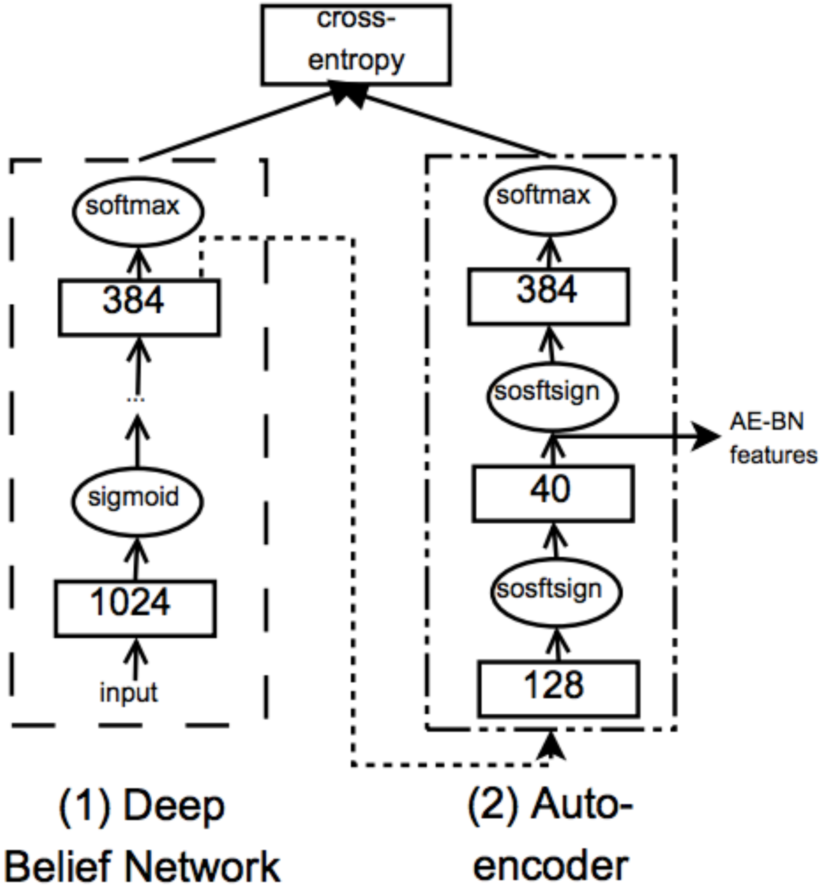
- One example of several attempts at incorporating multiplicative nodes into deep networks.
- Very promising area of research attempting to factor out ‘style’ (speaker, environment) from ‘content’ (phonetic label) using multiplicative gating interactions.



Large Vocabulary Speech Recognition Using Deep Tensor Neural Networks. Dong Yu, Li Deng, and Frank Seide, Interspeech 2012

Bottleneck features

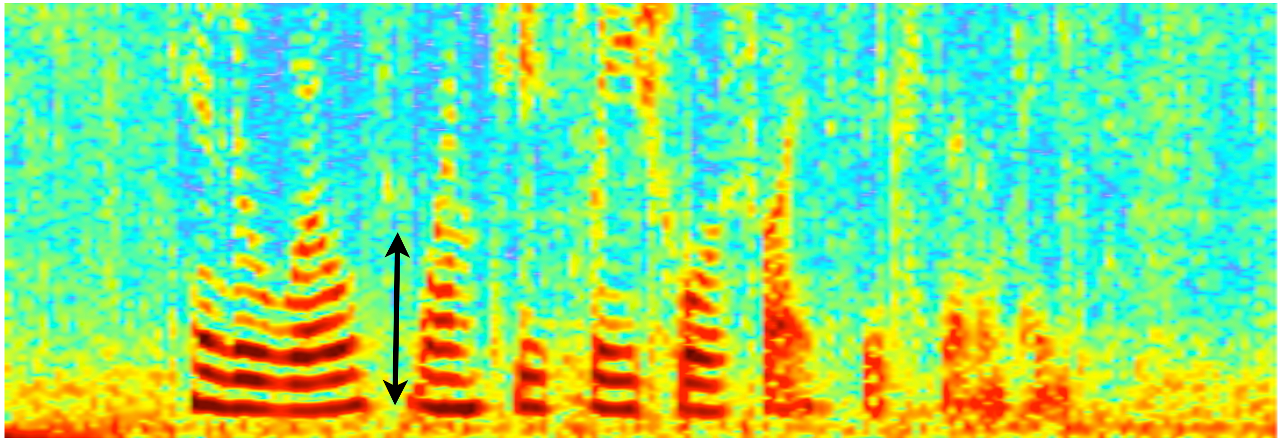
- Very popular alternative to 'hybrid' systems: use the neural network for feature extraction, and simply layer the traditional GMM machinery on top.
- Recently revived by casting the problem as an autoencoder over the output labels.
- Interesting because it attempts to answer whether DNNs are good models of the data or good classifiers or both.
- So far, 'both' seems to be the best answer.



T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-Encoder Bottleneck Features Using Deep Belief Networks," ICASSP 2012.

Convolutional Networks

- On a Mel scale, a pitch change is mostly a shift in the frequency axis.
- Convolutions in frequency seem like a natural way to represent invariance.
- Most significant reported improvement over basic DNNs so far.



T. N. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep Convolutional Neural Networks for LVCSR," ICASSP 2013.

What Next?

- Acoustic Models and Language Models trained independently.
  - Historically, training data for AM and LM came from different sources.
  - No longer necessary: we have lots of matched spoken data.
  - Joint models? Back-propagate through everything?
- Take another stab at the dreaded 'Independence Assumption'.
  - HMMs are a very weak model of the joint probability.
  - Recurrent Neural Networks are on the verge of making their own 'comeback' thanks to more sophisticated optimizers and renewed interest in Long Short-Term Memory (LSTM) architectures:

    Speech Recognition with Deep Recurrent Neural Networks, Alex Graves, Abdel-Rahman Mohamed, Geoffrey E. Hinton, ICASSP 2013
- 'Long-tail' issues: accented speech, kids' speech, non-stationary noise conditions are becoming a very significant factor in overall performance.

In Numbers: Phone Error Rates on TIMIT

Method	PER
Context-Dependent (CD) GMM-HMM	27.3%
Deep Convex Networks	22.0%
Context-Independent (CI) DNN on raw speech	21.9%
CD GMM-HMM, discriminatively trained (bMMI)	21.7%
CI DNN	20.5%
CI Convolutional DNN	20.0%
CI DNN, Rectified Linear Activations, dropout	19.7%
CI Convolutional DNN, dropout	18.7%
CTC Long Short Term Memory bi-recurrent DNN	17.7%

Conclusion

An Era of Convergence?

Speech Research is Accessible Again

- State-of-the-art, simple large-vocabulary baselines are now available:
  - Kaldi Open Source Toolkit: http://kaldi.sourceforge.net/
  - Supports the well-known Wall Street Journal and Switchboard datasets.
- Short of that, the speech community looks at TIMIT results once more! Easy baseline to set up.

Ads by Google

We support (your) academic research!

Google
Faculty Research Awards

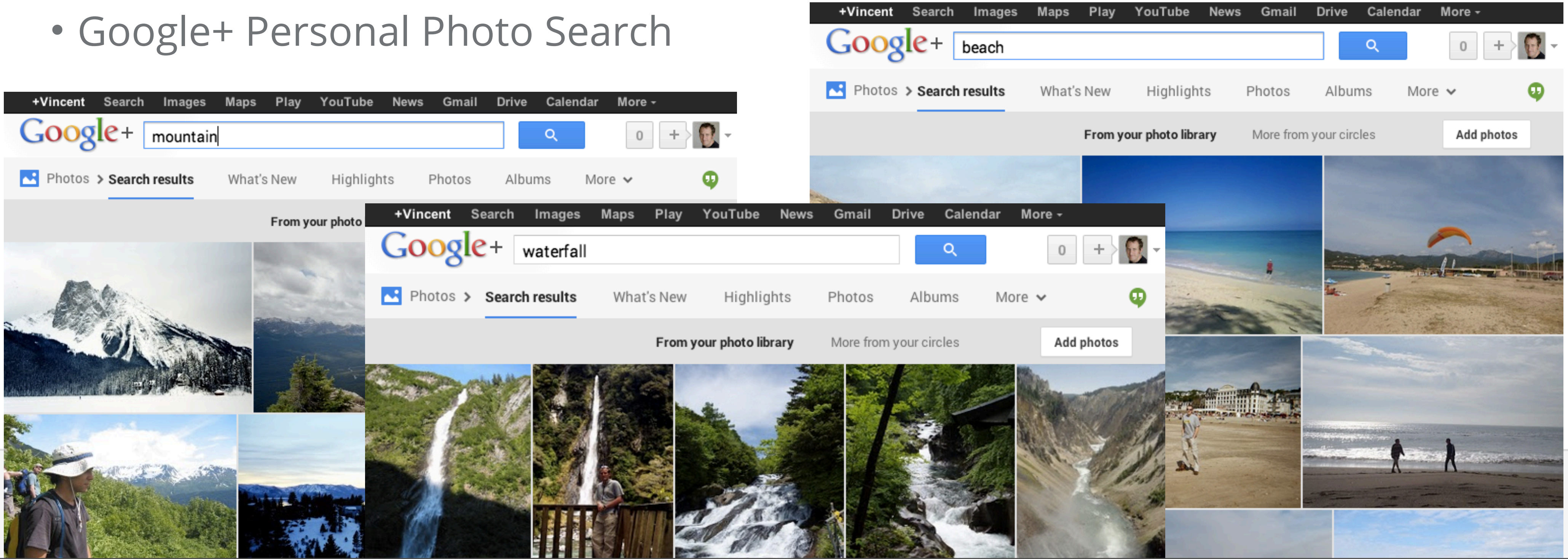
research.google.com

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann et al. "The Kaldi speech recognition toolkit." ASRU 2011.

DNNs + Rectified Linear Activations

+ (dropout) + (convolutions)

- Acoustic Modeling
- Molecular activity
  http://blog.kaggle.com/2012/11/01/deep-learning-how-i-did-it-merck-1st-place-interview/
- Salary prediction
  http://blog.kaggle.com/2013/05/06/qa-with-job-salary-prediction-first-prize-winner-vlad-mnih/
- 2013 ICML Challenges!
- Google+ Personal Photo Search



Putting it all together

http://youtu.be/RH0pYhKTvuc

Thank You

vanhoucke@google.com