Microsoft Research

Faculty Summit

2014

15TH ANNUAL

Microsoft Research Faculty Summit 2014 15TH ANNUAL

Deep Learning for Speech Recognition

Dong Yu, Microsoft Research

Microsoft

Debut of DNN ASR

Debut of Deep Neural Network ASR

2009 DNN on Phone Recognition (U Toronto)

TIMIT phone recognition: **22.4%** phone error rate (PER)

Ref: GMM: maximum likelihood training (MLT) 25.6%, sequence-discriminative training (SDT) 21.7%

Same architecture as 1990s but deep: models monophone states, frame-discriminative training, MFCC

Deep network helps; pretraining helps; has potential

GMM MLT
25.6%

CI-DNN FDT
22.4%

GMM SDT
21.7%

Debut of Deep Neural Network ASR

2010 DNN on Large Vocabulary ASR (Microsoft)

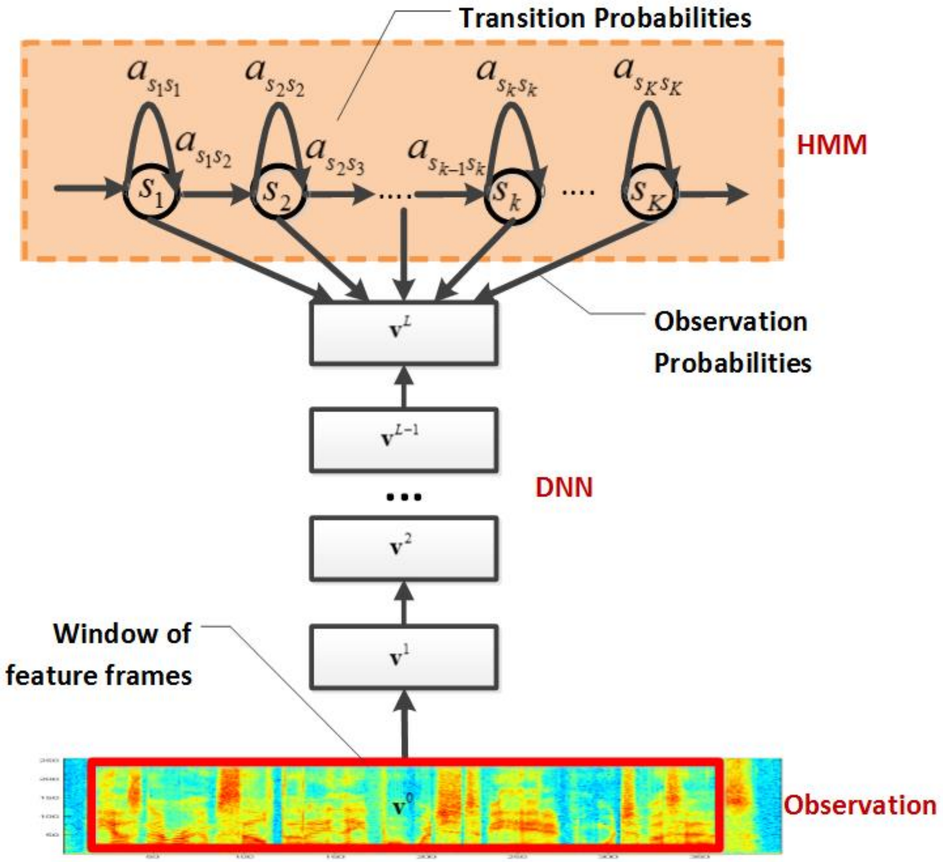
DNN modeling monophone states on voice search (24 hr): **37.3%** word error rate (WER)

Ref: GMM: MLT 39.6%; SDT 36.2%

Context-Dependent DNN-HMM (**CD-DNN-HMM**) frame-discriminative training (FDT) **30.1%**

Different from architectures in 1990s: Models tied triphone states (senones) directly with DNN

Modeling senones is critical; deep is important; input feature with contextual window is important; pretraining sometimes helps; realignment helps; tuning transition probabilities helps a little



GMM MLT 39.6% → CI-DNN FDT 37.3% → GMM SDT 36.2% → CD-DNN FDT 30.1%

DNN Work Started to Show Impact

2011 CD-DNN-HMM on Switchboard (Microsoft)

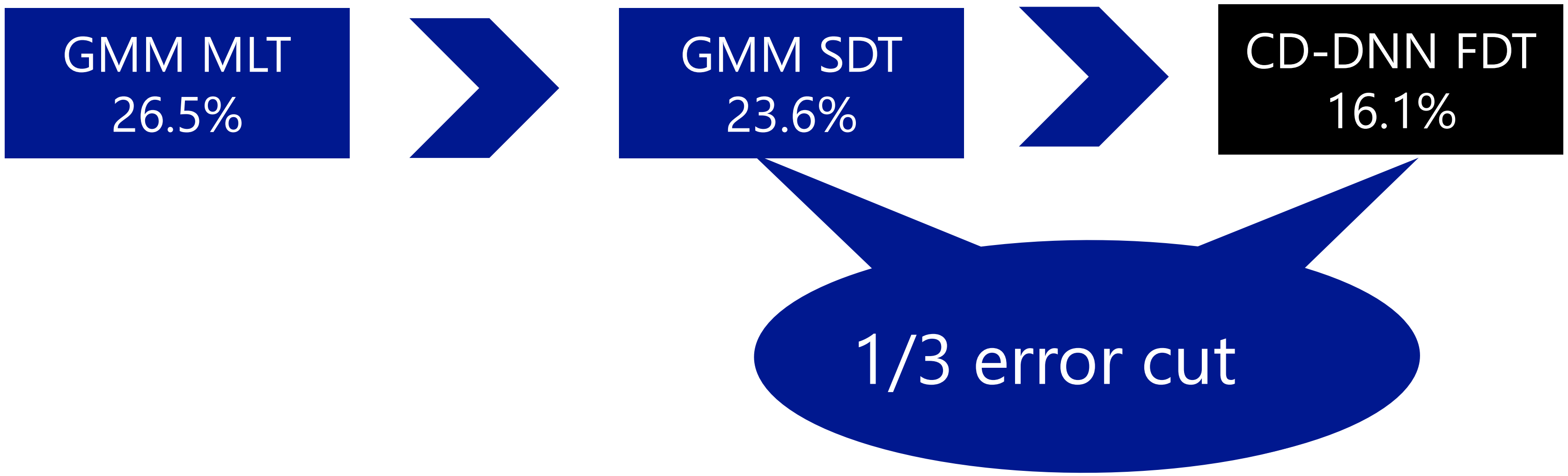
CD-DNN-HMM on Switchboard (309 hr training) with FDT: **16.1%**

Ref: GMM: MLT 26.5%; SDT 23.6% -> 1/3 error cut

Evaluated on a well accepted benchmark task

Same architecture and learning schedule

Scaled to hundreds of hours of speech and thousands of senones



Progress on DNN based ASR Since Then

DNN speed up

DNN sequence-discriminative training

Feature processing and engineering in DNNs

DNN adaptation

Convolution neural network

Recurrent neural network

Multi-task and transfer learning

Recent Progresses

DNN Speed Up

2011 DNN Decoding Speedup (Google)

With engineering optimization: 0.21 real time on single CPU core

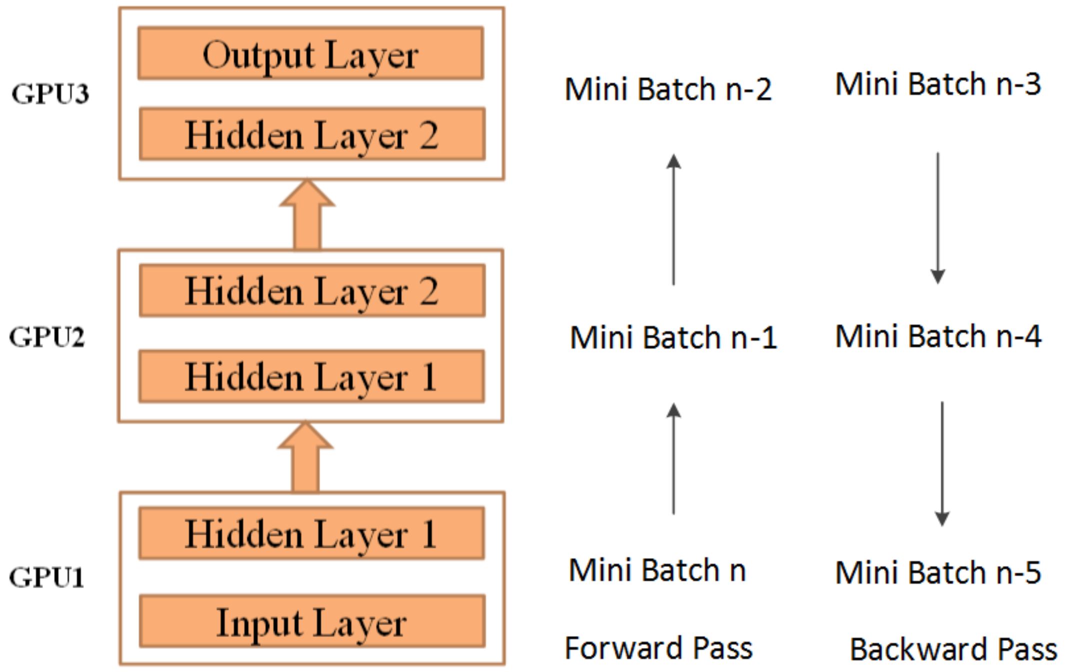
Ref: naive implementation 3.89 real time

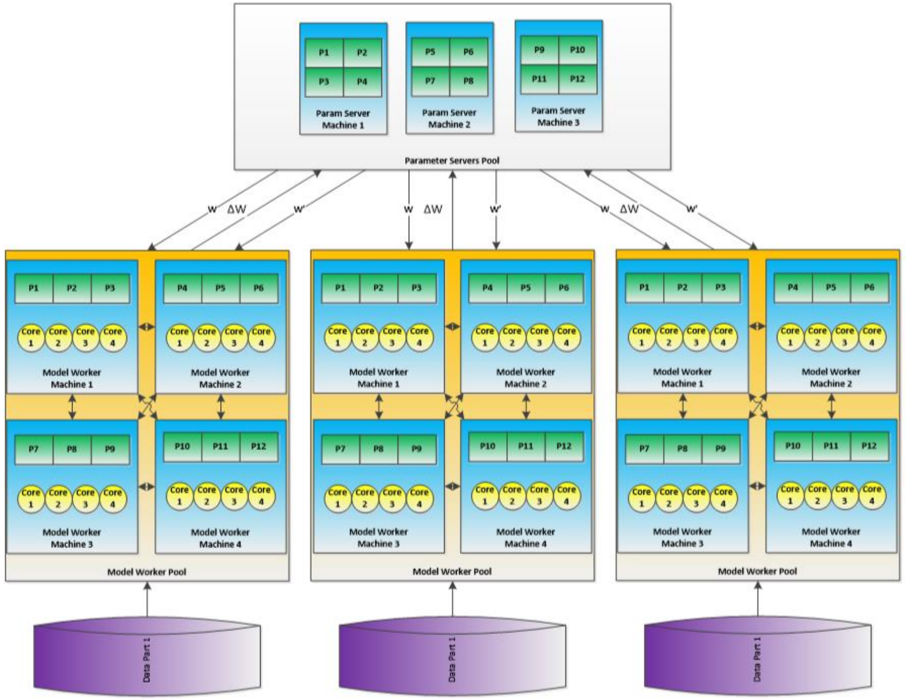
DNN Speed Up

2012 Parallel DNN Training (Microsoft, Google)

Pipelined Training (Microsoft): parallelize across 4 GPUs with 3.3 times of speed up.

Asynchronous SGD (Google): parallelize across thousands of CPU cores



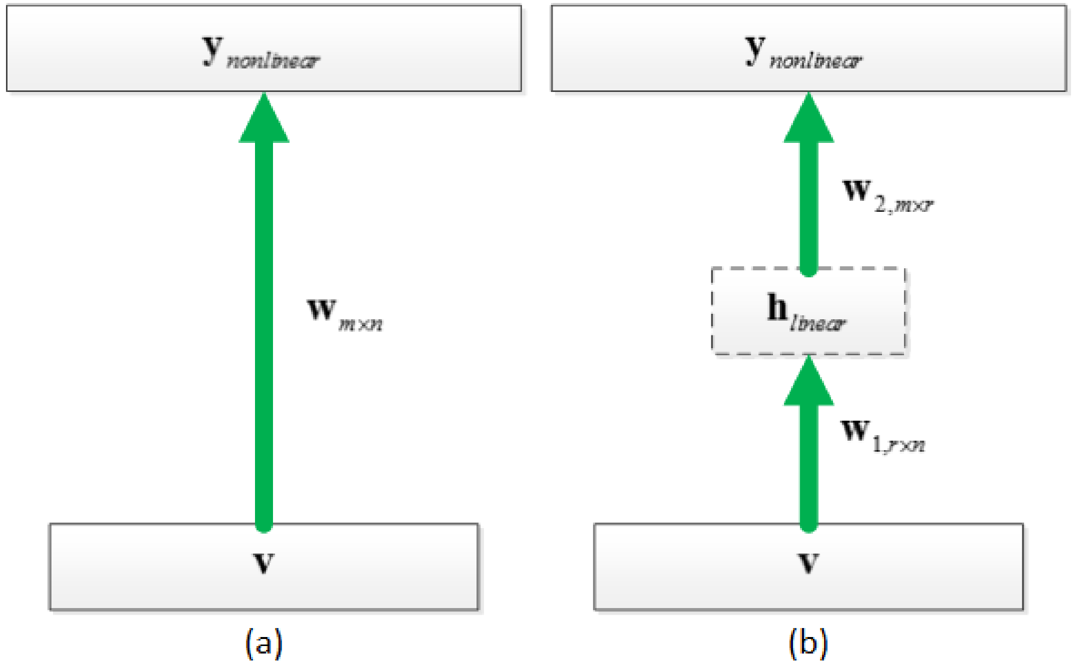


DNN Speed Up

2013 Low Rank Approximation (IBM, Microsoft)

Replace each weight matrix with the product of two smaller matrices by dropping small singular values.

2/3 cut in decoding time and model size



DNN Sequence Discriminative Training

2009 SDT on MLP/HMM Hybrid System for LVSR (IBM)

Multi-layer perceptron (MLP) on Broadcast news (50 hr training, LVSR) **27.7%** WER

Ref: MLP FDT 34.0%; GMM SDT 25.3%

SDT better than FDT on MLP/HMM hybrid system; Unified framework for MLP SDT

MLP FDT 34.0% → MLP SDT 27.7% → GMM SDT 25.3%

2010 SDT on DNN-HMM for TIMIT (Microsoft)

CI-DNN-HMM SDT on phone recognition: **22.2%** PER

Ref: DNN FDT 22.8% (different alignment and label from U Toronto)

DNN Sequence Discriminative Training

2012 SDT on CD-DNN-HMM (IBM)

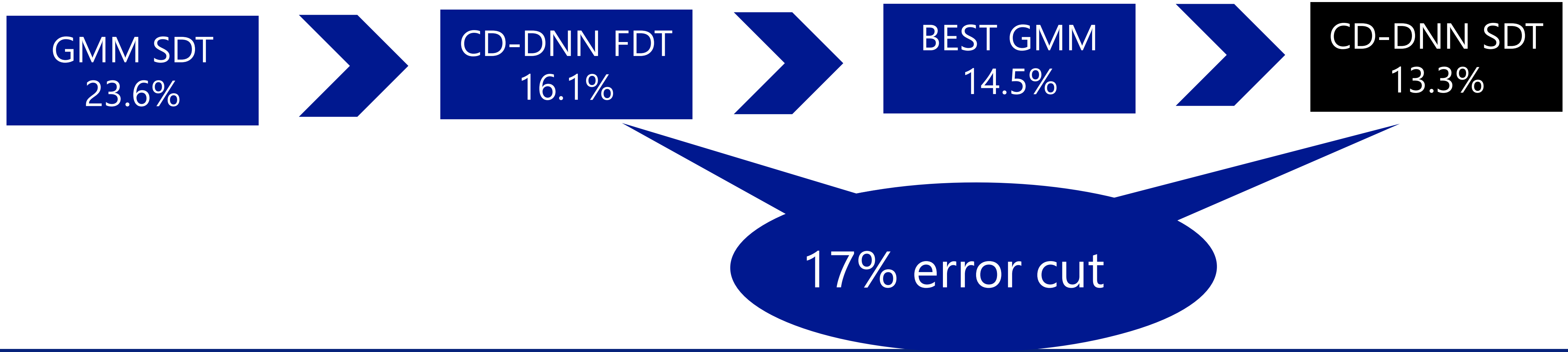
CD-DNN-HMM SDT on Switchboard (309 hr): **13.3%** WER

Ref: CD-DNN-HMM FDT 16.1%   -> 17% WER cut over FDT

Ref: GMM best number with all tricks and adaptation techniques using : 14.5%

State Minimum Bayesian Risk Training Criterion + Hessian-free optimization

CD-DNN-HMM surpasses best CD-GMM-HMM system (with multi-pass, adaptation, etc)



DNN Sequence Discriminative Training

2013 SDT Broader Success With Better Training Recipe (Microsoft, JHU, Google, IFlytech)

Lattice generation: generate lattice with your best system (e.g., FDT CD-DNN-HMM instead of MLT CD-GMM-HMM) or generate lattices during SDT using the current best model

Lattice compensation: handle run-away silence frames, augment lattice with reference transcription, reject bad frames

Over-fit control: smooth the SDT training criterion with the FDT training criterion

Learning rate control: use 1/5-1/10 of the learning rate used in the FDT

Training criterion: SDT training criterion used does not have huge effect on performance; MMI is simple to implement and thus preferred

Almost all companies deployed CD-DNN-HMM ASR systems since then

Feature Processing and Engineering

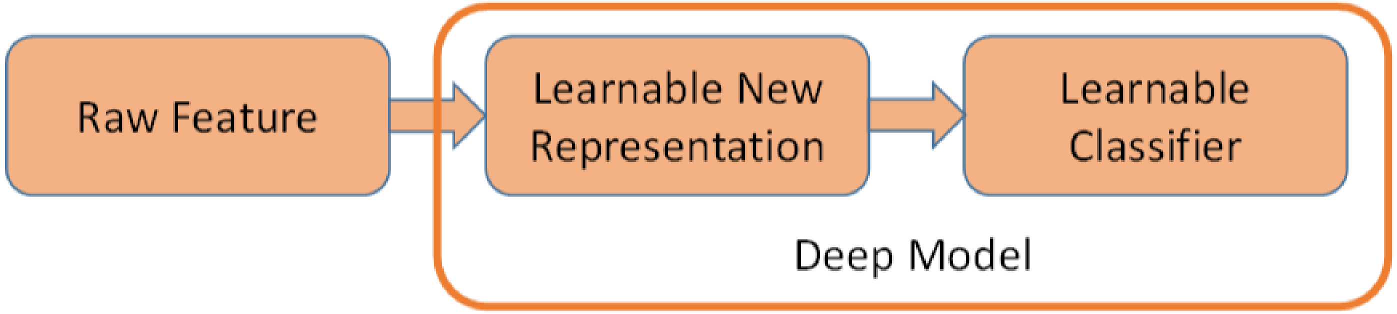
2011 Feature Engineering in DNNs (Microsoft)

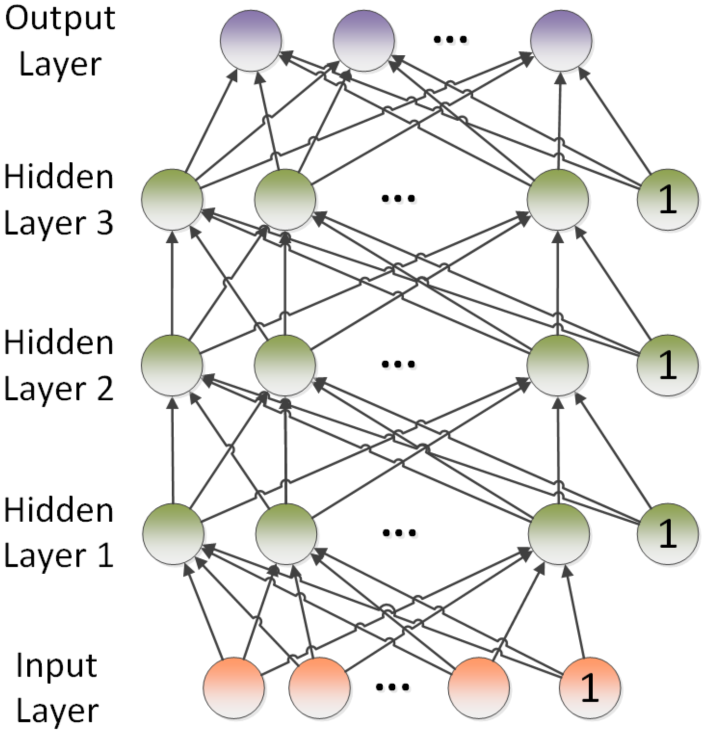
DNN learns the log-linear classifier and the complicated feature transformation jointly

DNN is more robust to speaker variations than shallow models

Feature engineering techniques (e.g., VTLN, fMLLR) help less in deep networks than in shallow models

Hint: can rewind many feature processing steps usually done in the GMM system, has no assumption on input features





Feature Processing and Engineering

2011 DNN as Feature Extractor (Microsoft)

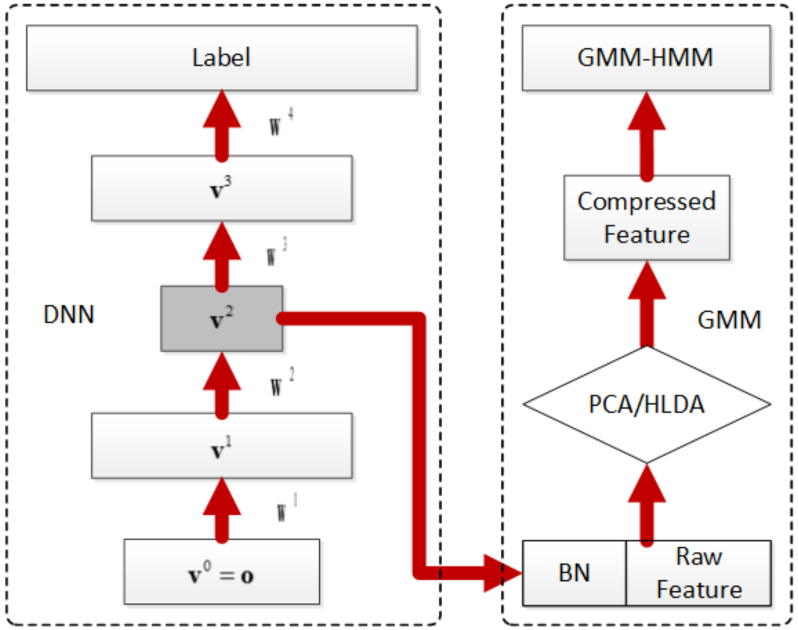
Bottleneck features extracted from CD-DNN-HMM performs better than those from CI-DNN-HMM when used in a GMM-HMM system.

2012 Log Filter Bank Features (U Toronto, Microsoft)

Log filter bank (LFB) feature performs better than MFCC on phone recognition **20.7%** WER (U Toronto)

- Ref: using MFCC (which has one more processing step with loss) 22.4%

Also works better on LVCSR (Microsoft) **29.8%** WER on voice search (24 hr) vs 31.6% using MFCC

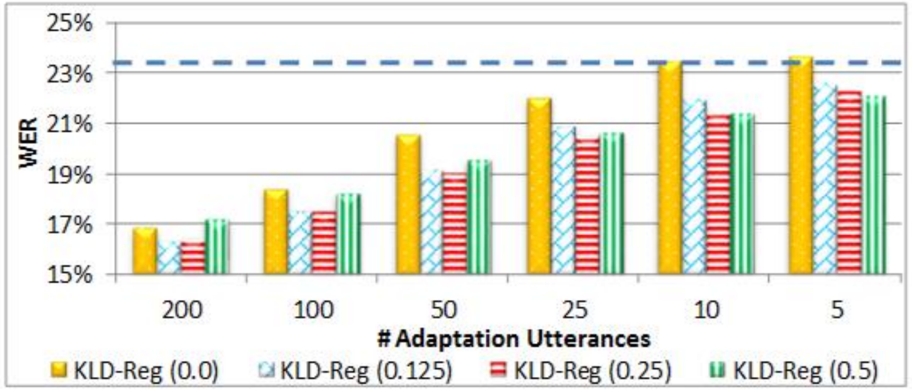


DNN Adaptation

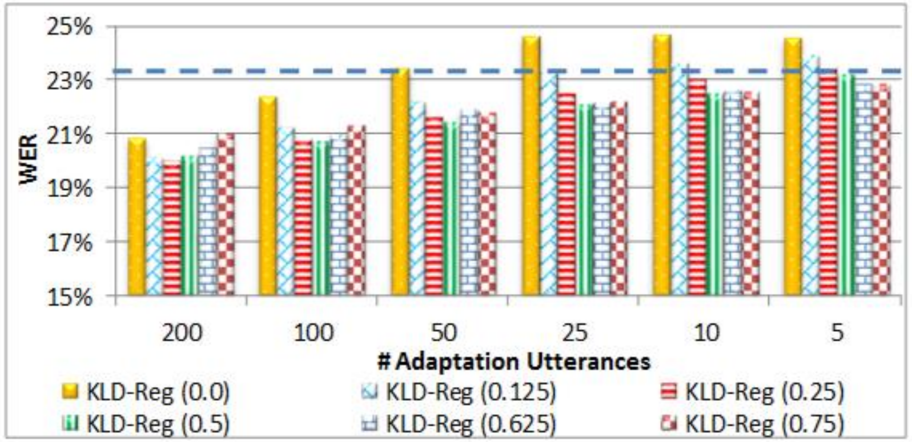
2013 KL-Divergence Regularization (Microsoft)

DNN FDT on short message dictation

3% WER cut with 5 adaptation utterance; 20% WER cut with 100 adaptation utterance



(a) Supervised Adaptation



(b) Unsupervised Adaptation

DNN Adaptation

2013 Noise-Aware Training (Microsoft)

DNN FDT on Aurora4 13.4% WER, + noise-aware training **12.4%**

- Ref: GMM: SDT 22.5%; +adaptive training 15.3%, +VAT+Joint compensation 13.4%

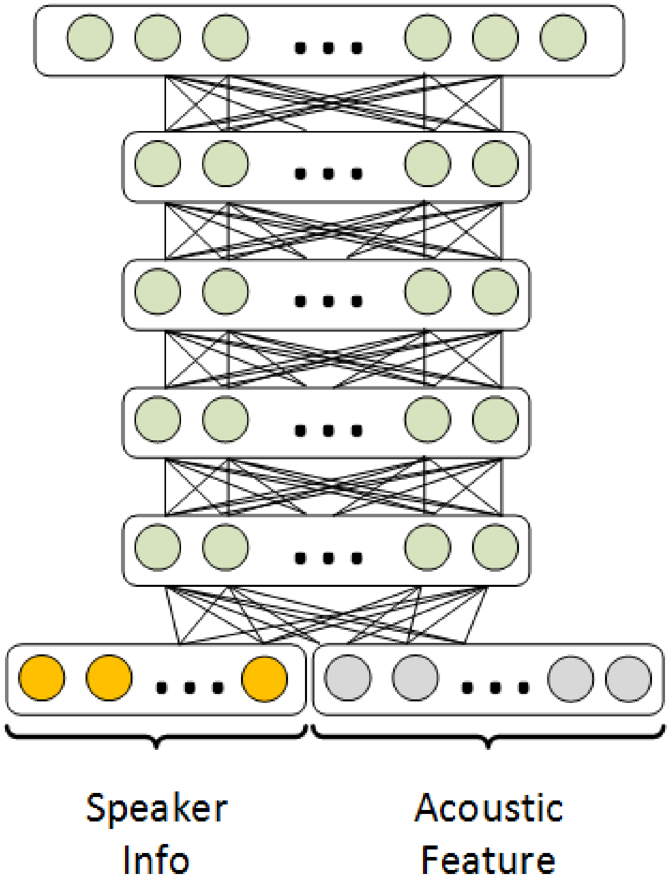
2013 Speaker Code (York U)

10% error cut compared to speaker-independent DNN, speaker code learned in adaptive training way

2013 Speaker-Aware Training (IBM)

DNN SDT 14.1%; + SaT: **12.4%** WER → 12% error cut

Use i-vector to represent speaker and to adjust the bias of each layer



Convolutional Neural Network (CNN)

2012 CNN on Phone Recognition (York U)

On TIMIT with LFB features: **20.0%** PER

- Ref: DNN with LFB features 20.7%

Use CNN at the frequency axis to normalize speaker differences. Only feasible with LFB features

2013 CNN on LVCSR (Microsoft, IBM)

Improved CNN architecture, pretraining techniques, and pooling strategy

CNN works on some LVCSR tasks (no obvious gain on many others)

- Voice search (18 hr training) FDT: **33.4%** WER with CNN vs 35.4% with DNN
- Switchboard (309 hr training) SDT: **11.8%** WER with CNN vs 12.2% with DNN

2014 Combine CNN and DNN (IBM)

Switchboard (309 hr) CNN+DNN+Adaptation+SDT **10.4%**

- Ref: best number with all tricks and adaptation techniques using GMM is 14.5%

Other Advancements

2013 Multi-task and Transfer Learning (Many Groups)

Adopts shared-hidden layer architecture; learned features are shared across tasks
Applied to multi-lingual ASR, low-resource language ASR, and multi-modal ASR

2013 Long-Short Term Memory (U Toronto)

Bidirectional LSTM on TIMIT phone recognition: **18.4%** PER, Ref: CNN 20.0% (U Toronto)
LSTM-HMM FDT: WSJ **11.7%** WER, Ref: DNN 12.3% (U Toronto)

2014 Long-Short Term Memory (Google)

LSTM-HMM SDT: 10% WER reduction over DNN on VS and SMD (detail unknown)

2014 Single-Channel Mixed Speech ASR (Microsoft)

CD-DNN-HMM FDT with joint two-speaker DNN decoder **18.8%** WER
Ref: IBM's superhuman system (factorial GMM) 21.6%, Human 22.3%, next best 34.2%

Moving Forward

Next Frontiers in ASR

Closed Talk Single-Talker ASR Largely Solved

We can achieve 10% or less WER on the difficult Switchboard and many other tasks.

Areas Where Performance Not Satisfactory

ASR with far field microphone: living room, meeting room, field video recordings

ASR under very noisy condition: e.g., when music is playing

ASR with accented speech

ASR with multi-talker speech or side talks: meeting, multi-party chat, or when radio is playing

ASR with spontaneous speech

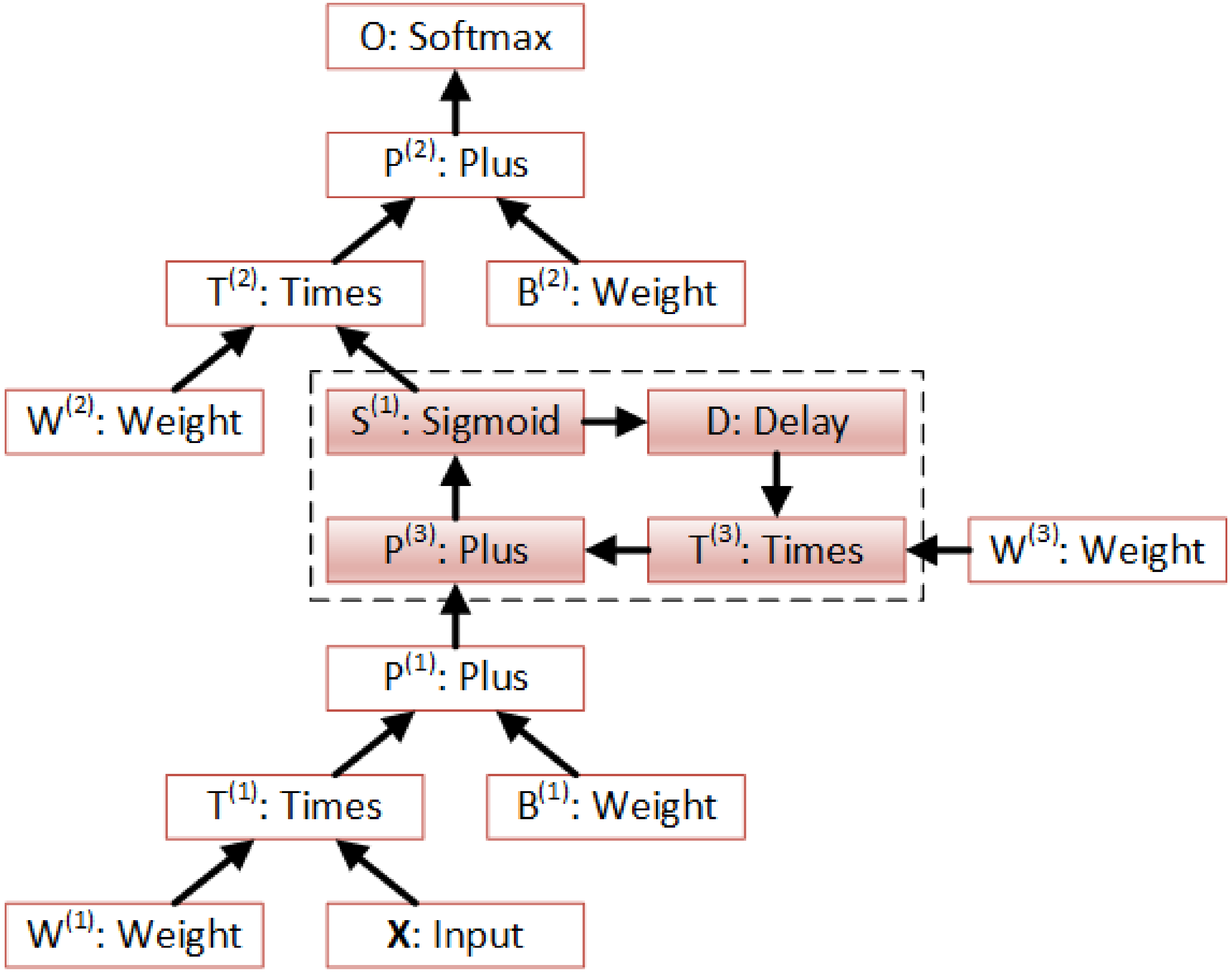
New Model or More Data

More data sufficient to solve the first three problems?

Can the system automatically adapt and constantly learn, e.g., tracing a particular speaker?

Can we take knowledge and semantics as additional constraint?

Computational Network



Save the planet and return your name badge before you leave (on Tuesday)

Microsoft Privacy Policy statement applies to all information collected. Read at research.microsoft.com

Microsoft

© 2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.