CS 224S / LINGUIST 285
Spoken Language Processing

Andrew Maas
Stanford University
Spring 2017

Lecture 7: Neural network acoustic models in speech recognition

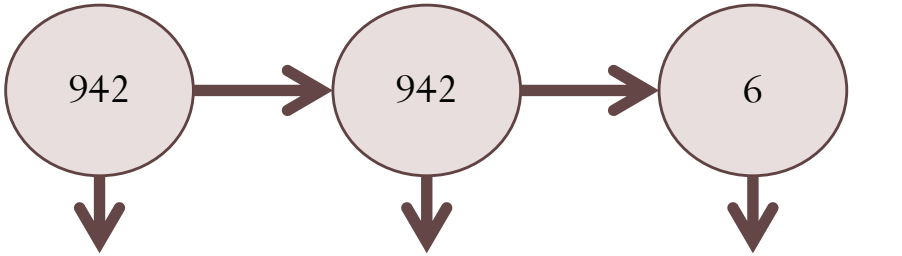
Outline

- Hybrid acoustic modeling overview
  - Basic idea
  - History
  - Recent results
- What's different about modern DNNs?
- Convolutional networks
- Recurrent networks
- Alternative loss functions

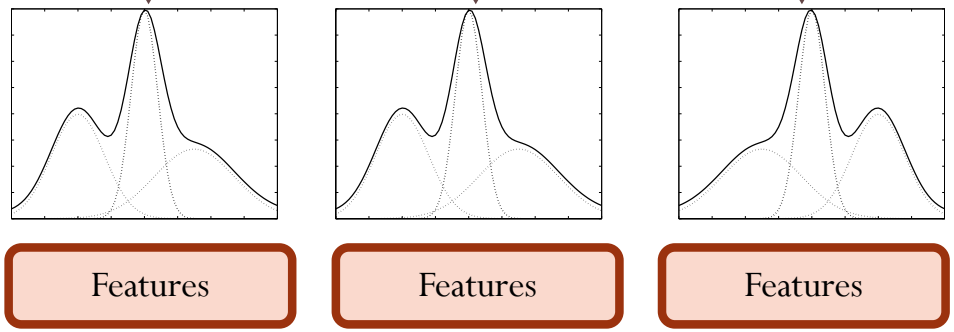
Acoustic Modeling with GMMs

Transcription: Samson
Pronunciation: S – AE – M – S –AH – N
Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

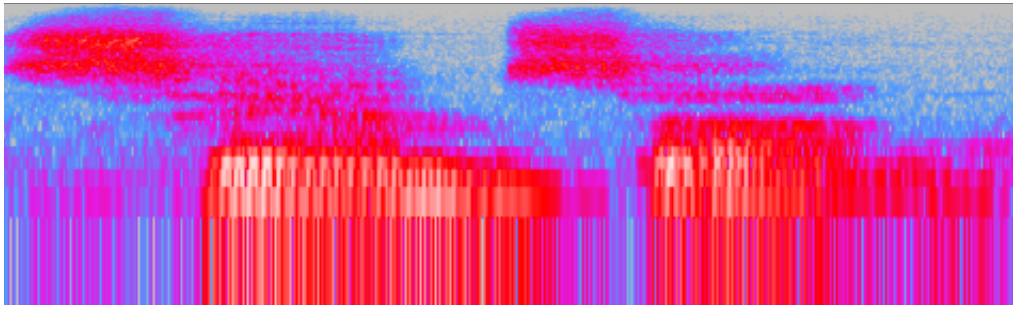
Hidden Markov Model (HMM):



Acoustic Model:



Audio Input:



GMM models:
P(x|s)
x: input features
s: HMM state

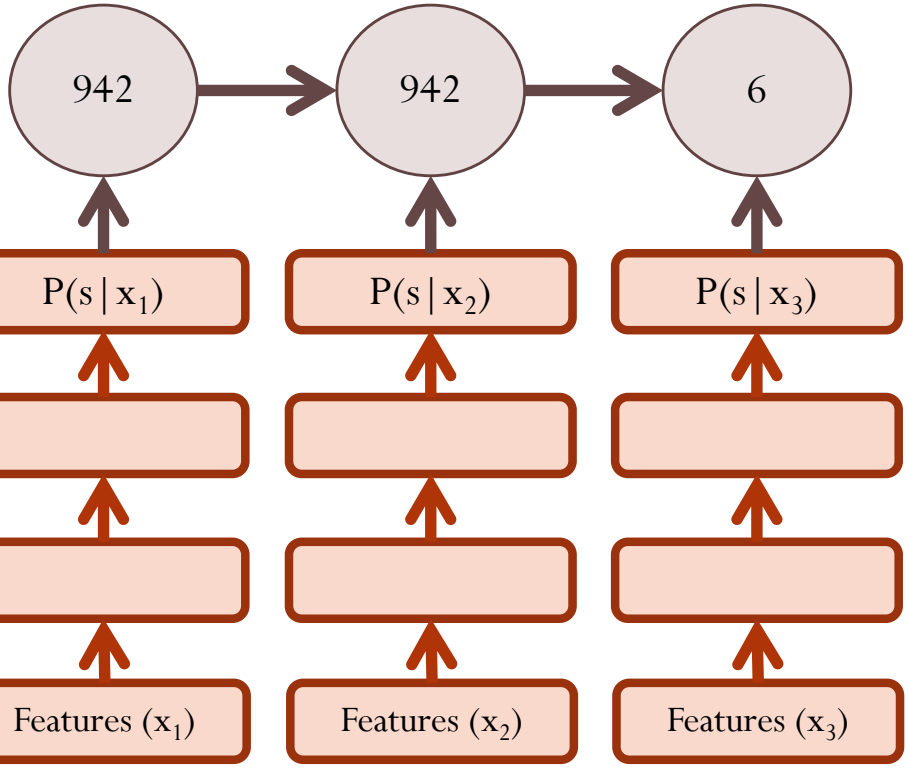
DNN Hybrid Acoustic Models

Transcription: Samson

Pronunciation: S – AE – M – S –AH – N

Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

Hidden Markov Model (HMM):



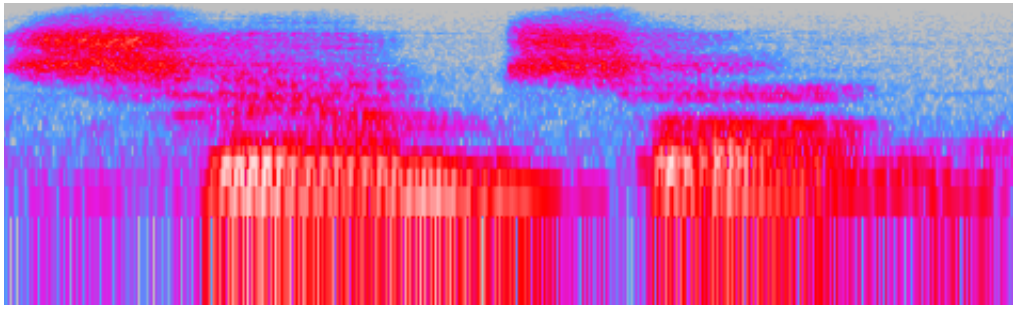
Acoustic Model:

Use a DNN to approximate: P(s|x)

Apply Bayes' Rule: P(x|s) = P(s|x) * P(x) / P(s)

DNN * Constant / State prior

Audio Input:



Noisy Channel Model

- Probabilistic implication: Pick the highest prob S:

$$\hat{W} = \arg\max_{W \in L} P(W \mid O)$$

- We can use Bayes rule to rewrite this:

$$\hat{W} = \arg\max_{W \in L} \frac{P(O \mid W)P(W)}{P(O)}$$

- Since denominator is the same for each candidate sentence W, we can ignore it for the argmax:

$$\hat{W} = \arg\max_{W \in L} P(O \mid W)P(W)$$

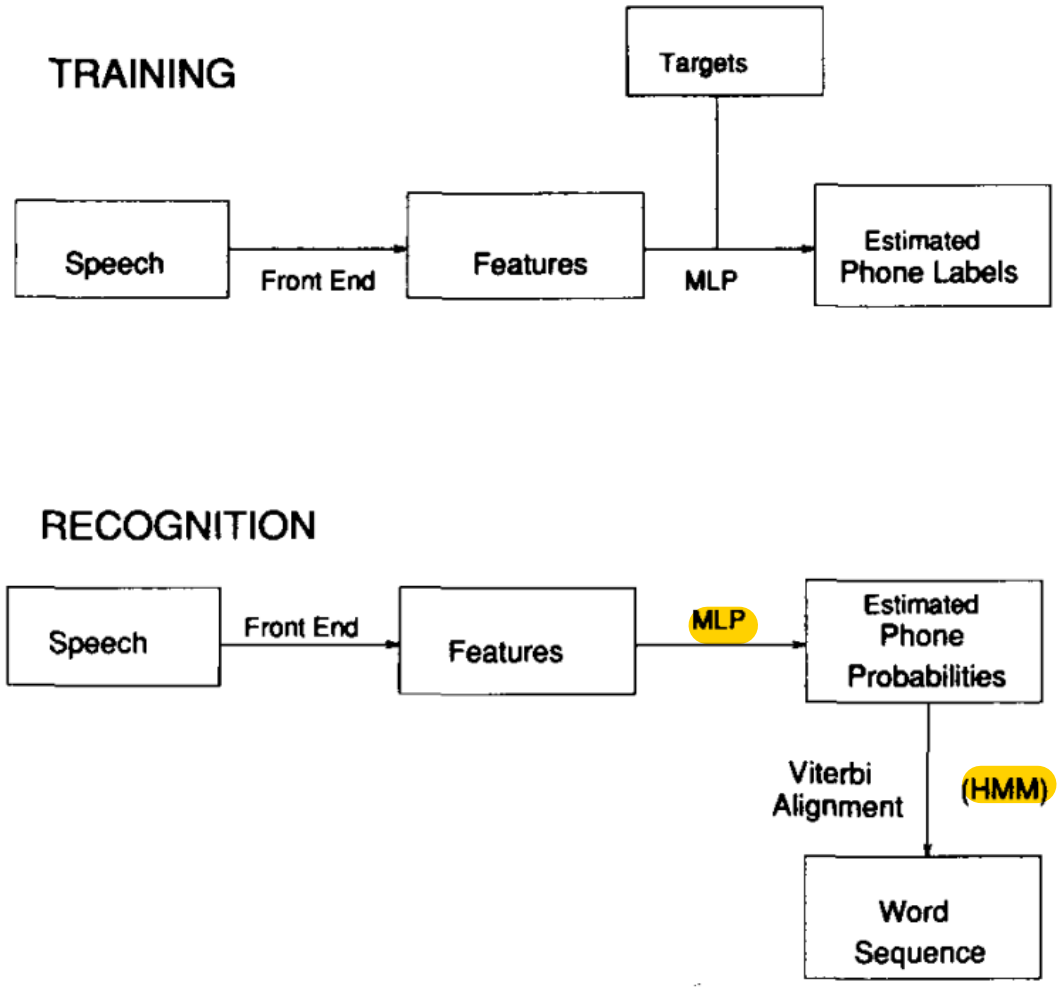
Objective Function for Learning

- Supervised learning, minimize our classification errors
- Standard choice: Cross entropy loss function
  - Straightforward extension of logistic loss for binary

$$Loss(x, y; W, b) = -\sum_{k=1}^{K} (y = k) \log f(x)_k$$

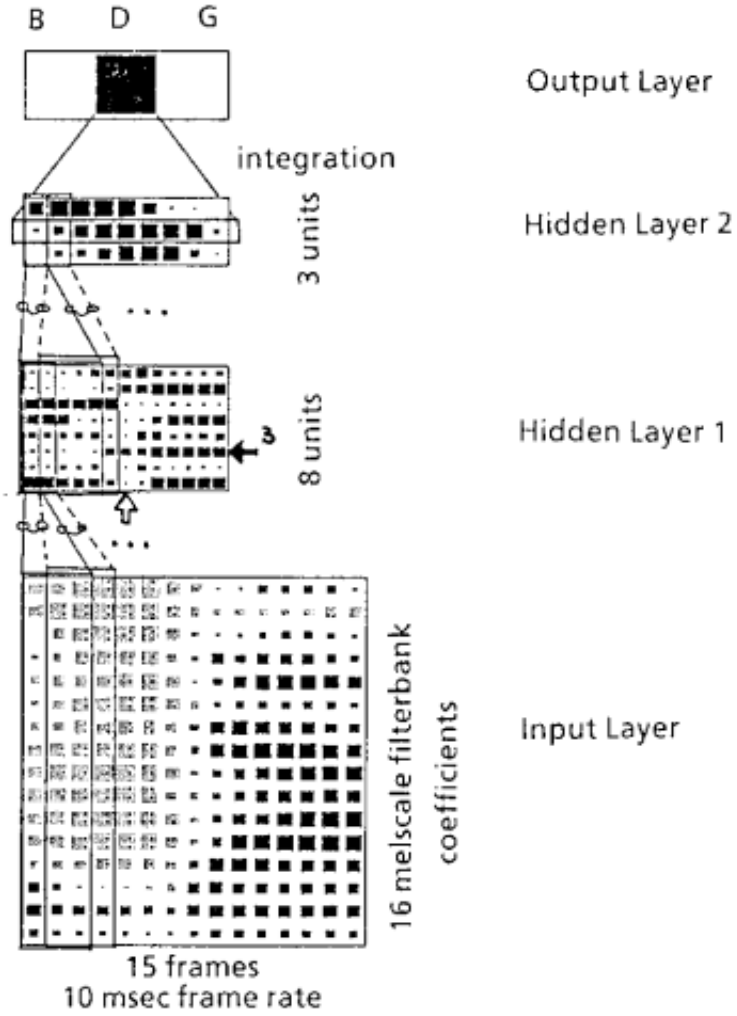
- This is a *frame-wise* loss. We use a label for each frame from a forced alignment
- Other loss functions possible. Can get deeper integration with the HMM or word error rate

Not Really a New Idea

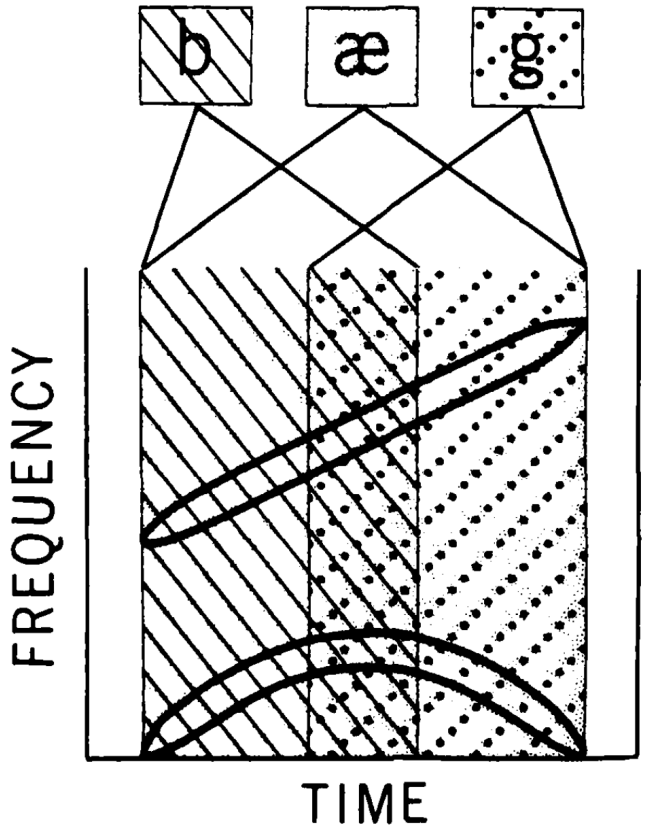


Renals, Morgan, Bourland, Cohen, & Franco. 1994.

Early neural network approaches



(Waibel, Hanazawa, Hinton, Shikano, & Lang. 1988)



(McClelland, & Elman. 1985)

Hybrid MLPs on Resource Management

TABLE I
RESULTS USING THE THREE TEST SETS WITH THE PERPLEXITY 60 WORDPAIR GRAMMAR. (CI-MLP is the context-independent MLP-HMM hybrid system, CD-HMM is the full context-dependent Decipher system, and the MIX system is a simple interpolation between the CD-HMM and the CI-MLP.)

	% error		
Test Set	CI-MLP	CD-HMM	MIX
Feb 91	5.8	3.8	3.2
Sep 92a	10.9	10.1	7.7
Sep 92b	9.5	7.0	5.7

TABLE II
RESULTS USING THE THREE TEST SETS USING NO GRAMMAR (PERLPEXITY 991)

	% error		
Test Set	CI-MLP	CD-HMM	MIX
Feb 91	24.7	19.3	15.9
Sep 92a	31.5	29.2	25.4
Sep 92b	30.9	26.6	21.5

Renals, Morgan, Bourland, Cohen, & Franco. 1994.

Hybrid Systems now Dominate ASR

[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

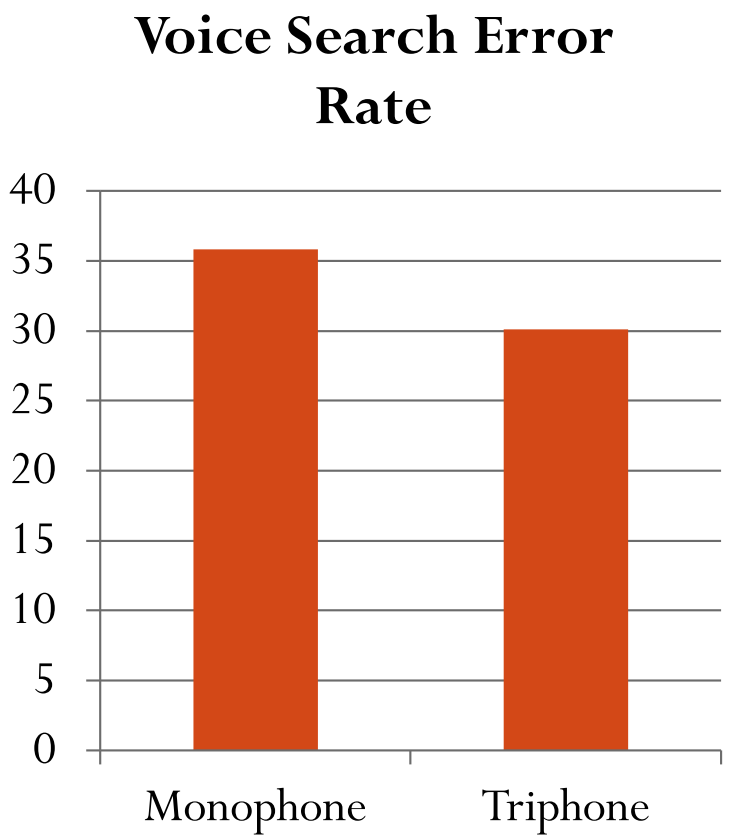
TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

Hinton et al. 2012.

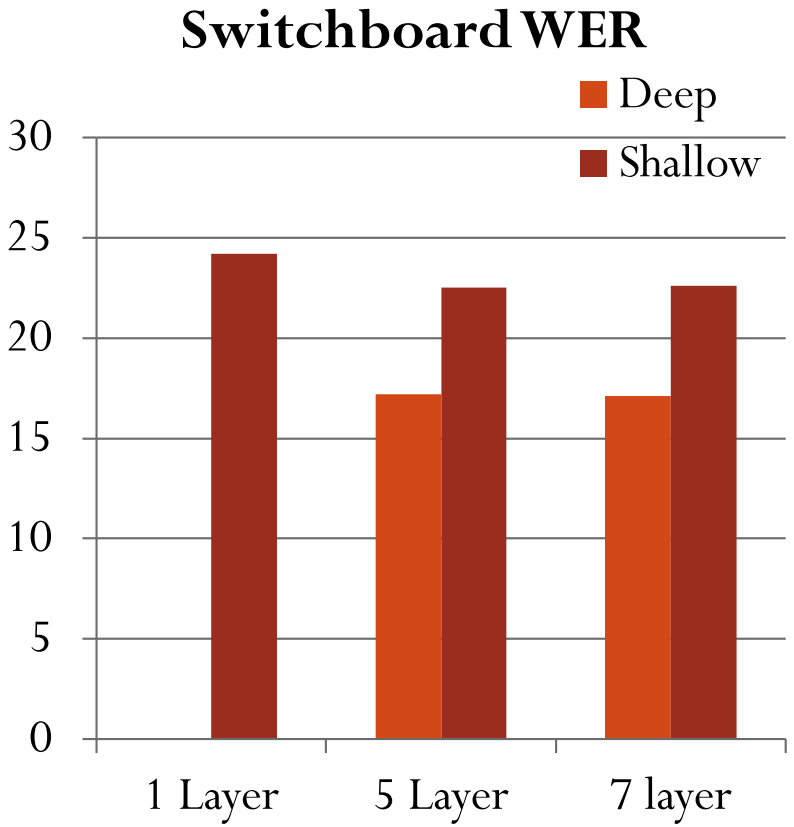
What's Different in Modern DNNs?

- Context-dependent HMM states
- Deeper nets improve on single hidden layer nets
- Hidden unit nonlinearity
- Many more model parameters (scaling up)
- Specific depth (e.g. 3 vs 7 hidden layers)
- Fast computers = run many experiments
- Architecture choices (easiest is replacing sigmoid)
- Pre-training *does not matter*. Initially we thought this was the new trick that made things work

Modern Systems use DNNs and Senones



(Dahl, Yu, Deng, & Acero. 2011)

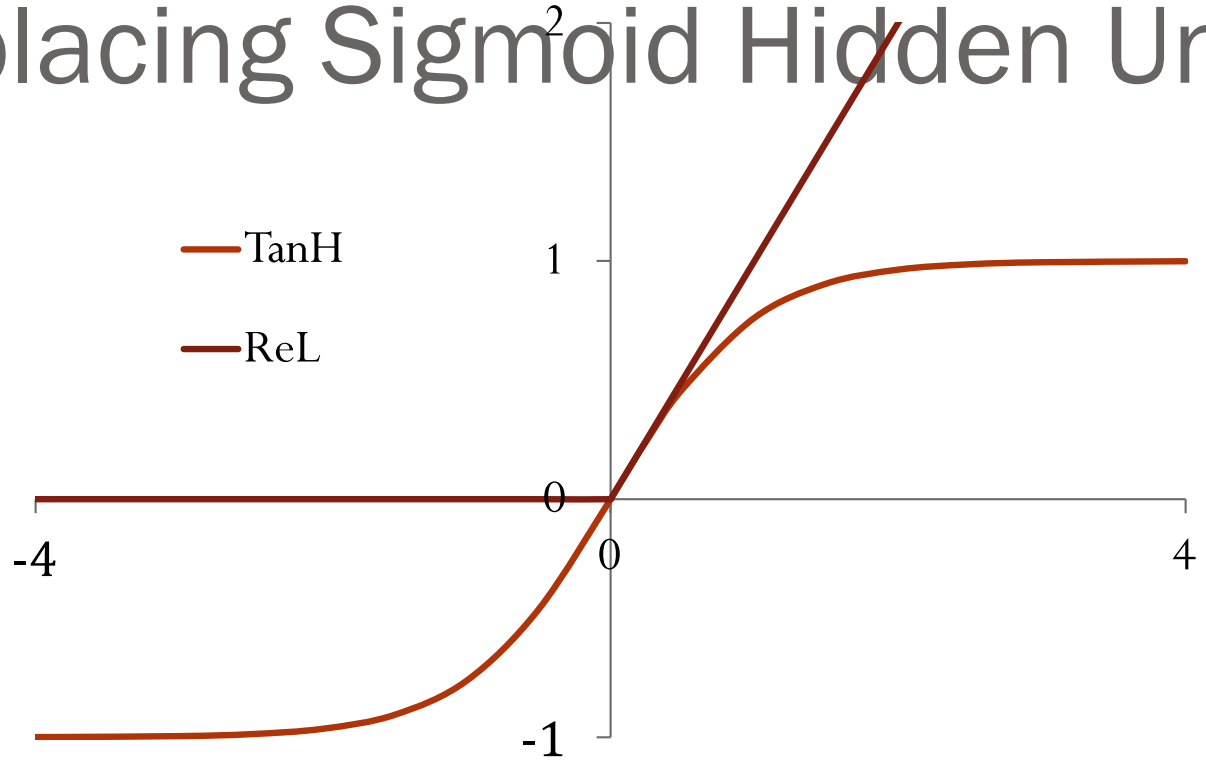


(Yu, Seltzer, Li, Huang, & Seide. 2013)

Many hidden layers

- **Warning!** Depth can also act as a regularizer because it makes optimization more difficult. This is why you will sometimes see very deep networks perform well on TIMIT or other small tasks.

Replacing Sigmoid Hidden Units

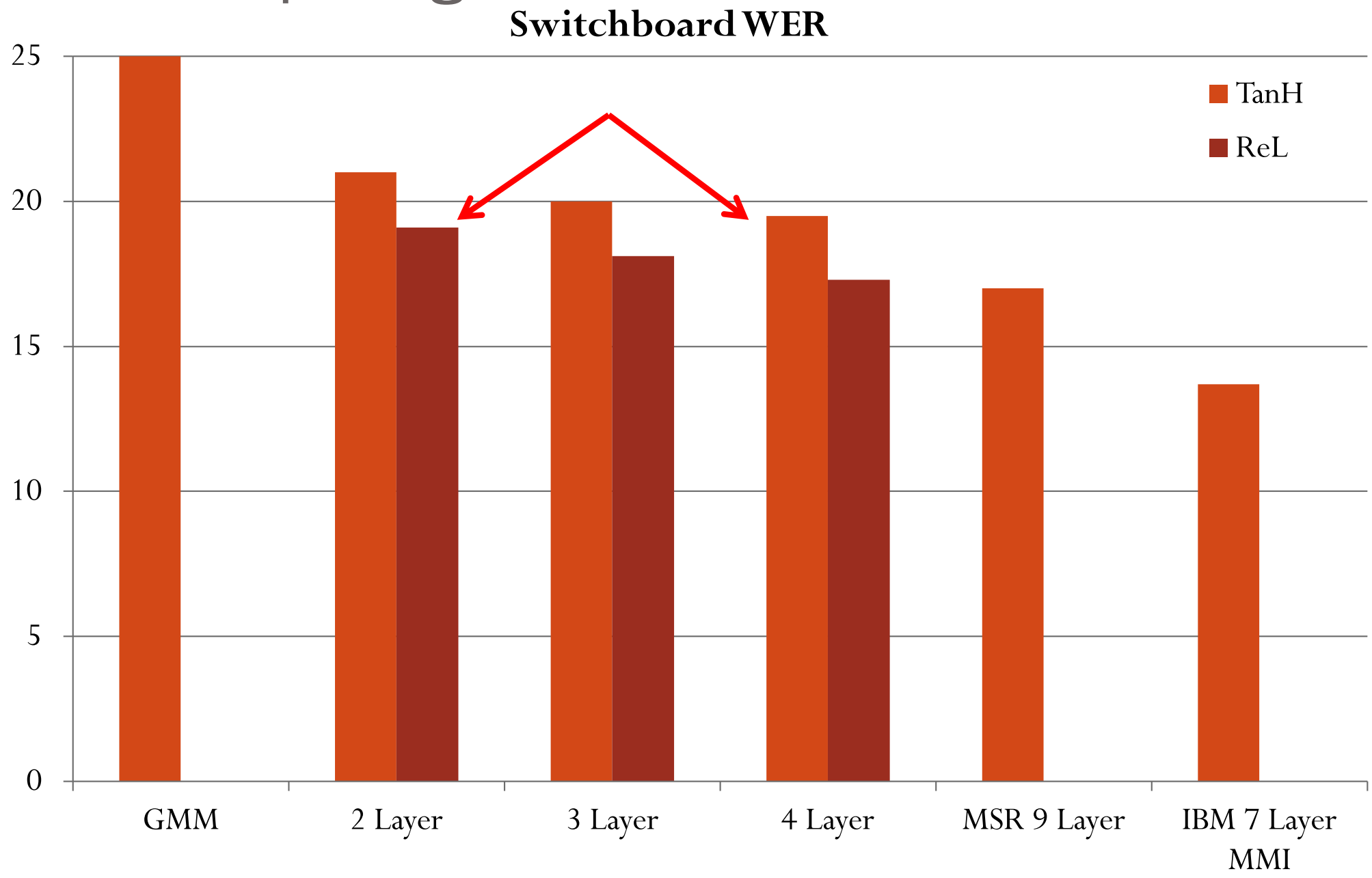


Rectified Linear (ReL)

$$h^{(i)} = \max(w^{(i)T}x, 0) = \begin{cases} w^{(i)T}x & w^{(i)T}x > 0 \\ 0 & \text{else} \end{cases}$$

(Glorot & Bengio. 2011)

Comparing Nonlinearities



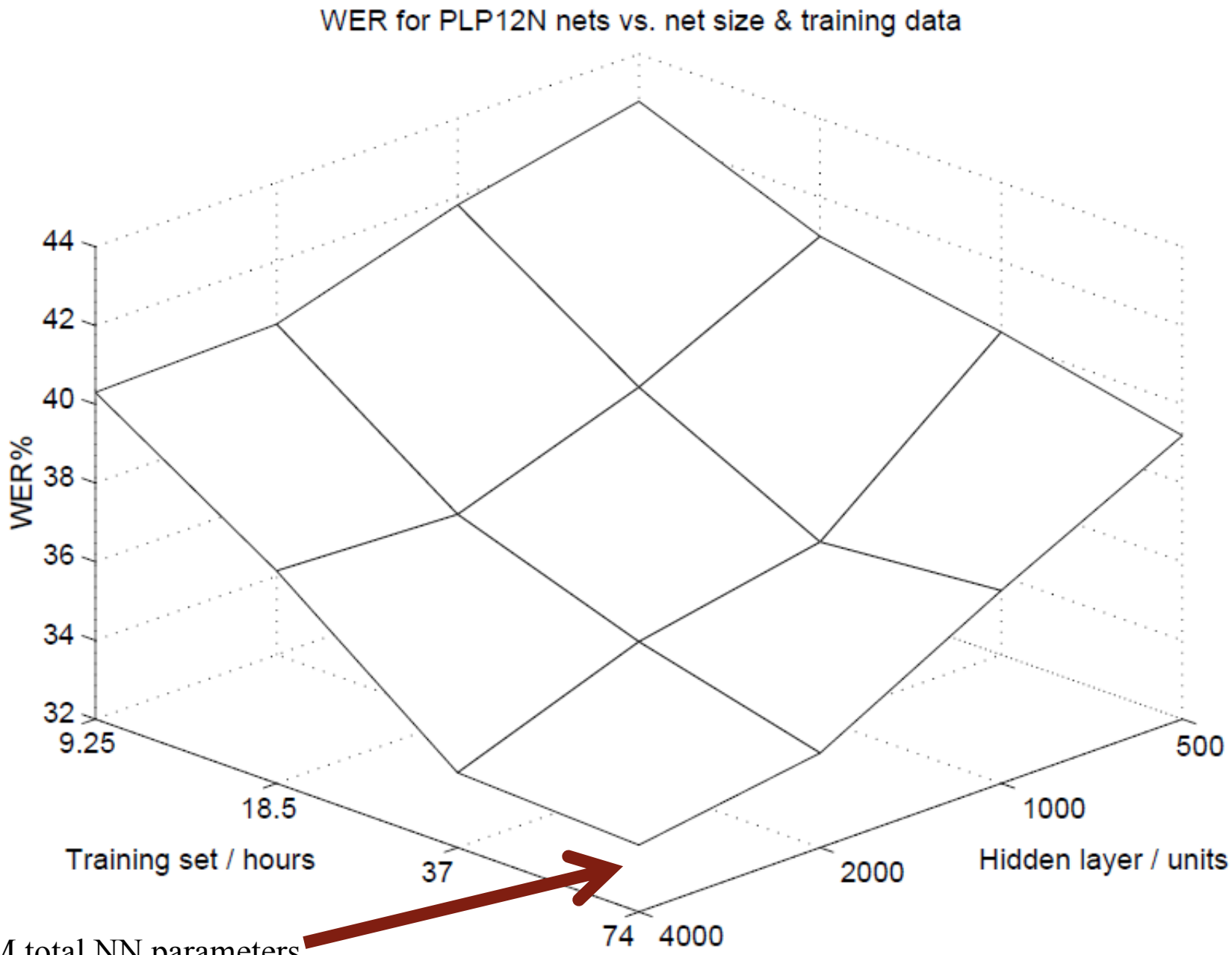
(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Rectifier DNNs on Switchboard

Model	Dev Acc(%)	Switchboard WER
GMM Baseline	N/A	25.1
2 Layer Tanh	48.0	21.0
2 Layer ReLU	51.7	19.1
3 Layer Tanh	49.8	20.0
3 Layer RelU	53.3	18.1
4 Layer Tanh	49.8	19.5
4 Layer RelU	53.9	17.3
9 Layer Sigmoid CE [MSR]	--	17.0
7 Layer Sigmoid MMI [IBM]	--	13.7

Maas, Hannun, & Ng,. 2013.

Scaling up NN acoustic models in 1999



0.7M total NN parameters

[Ellis & Morgan. 1999]

Adding More Parameters 15 Years Ago

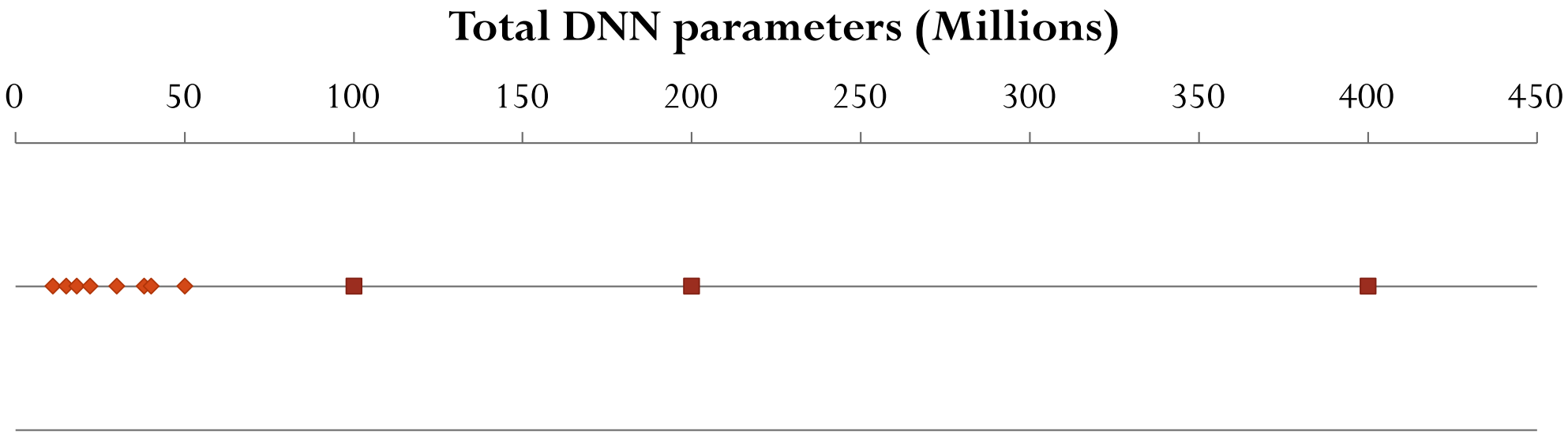
Size matters: An empirical study of neural network training for LVCSR. Ellis & Morgan. ICASSP. 1999.

Hybrid NN. 1 hidden layer. 54 HMM states.

74hr broadcast news task

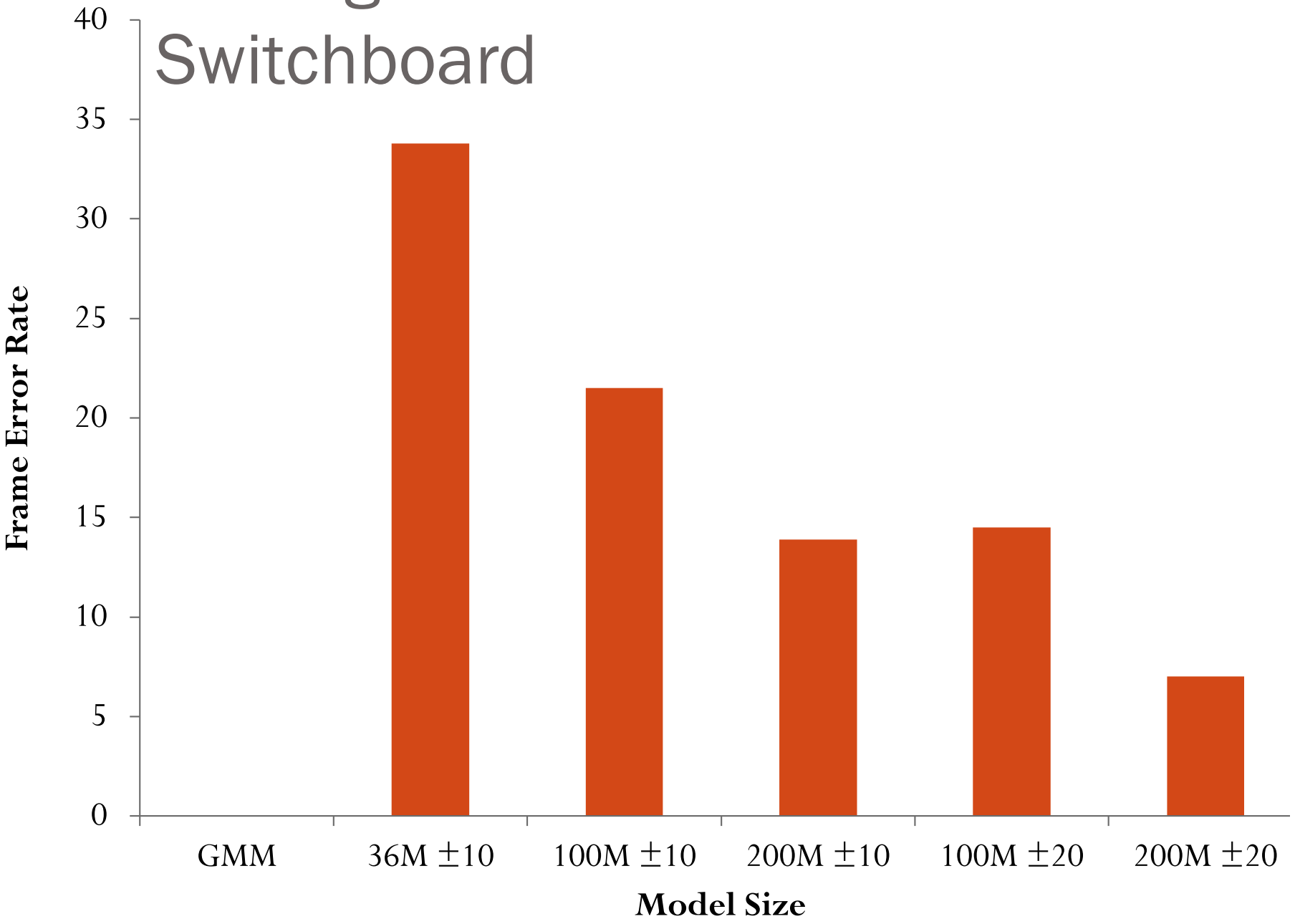
"…improvements are almost always obtained by increasing either or both of the amount of training data or the number of network parameters … We are now planning to train an 8000 hidden unit net on 150 hours of data … this training will require over three weeks of computation."

Adding More Parameters Now



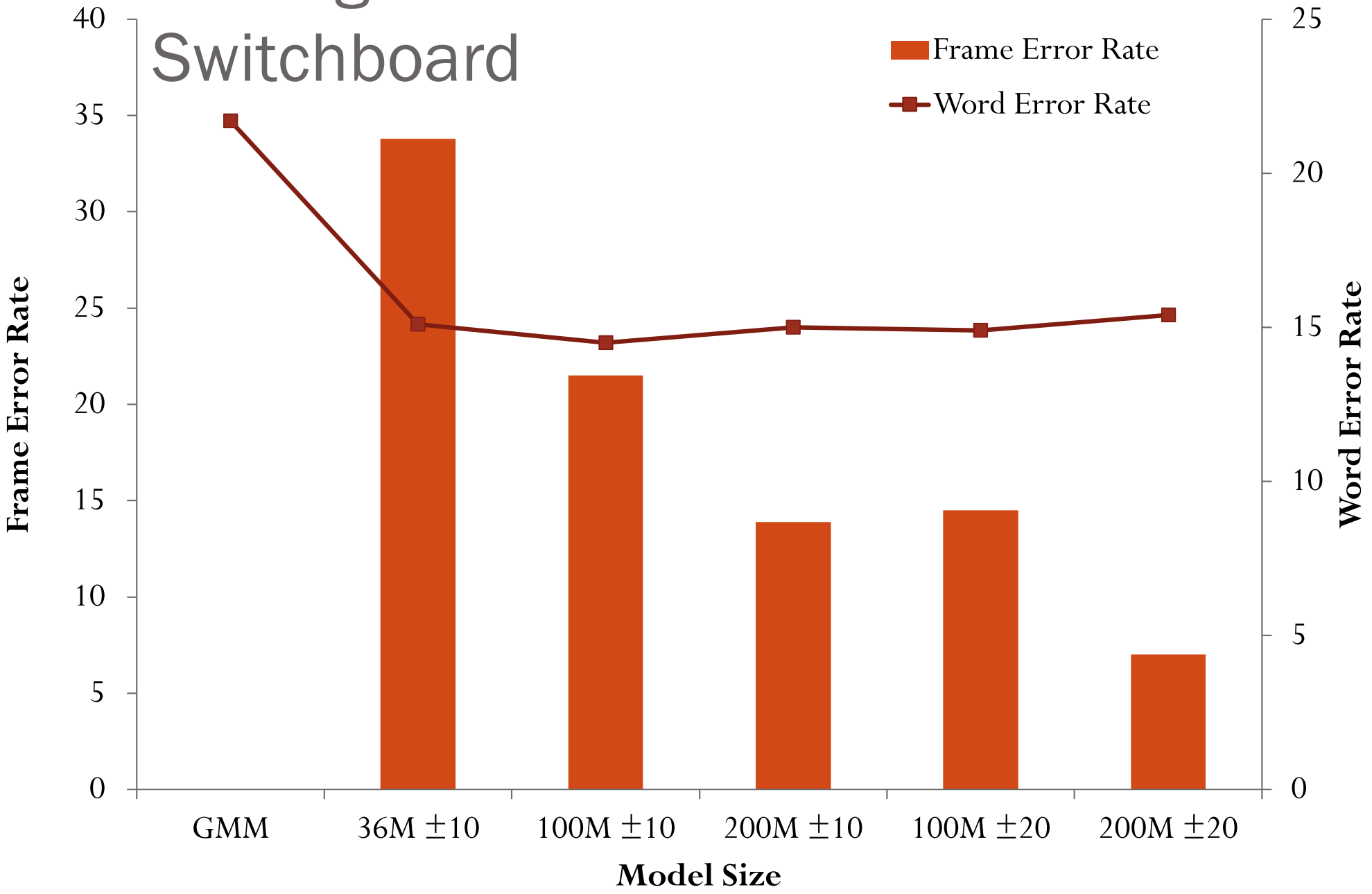
(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Scaling Total Parameters on Switchboard



(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Scaling Total Parameters on Switchboard

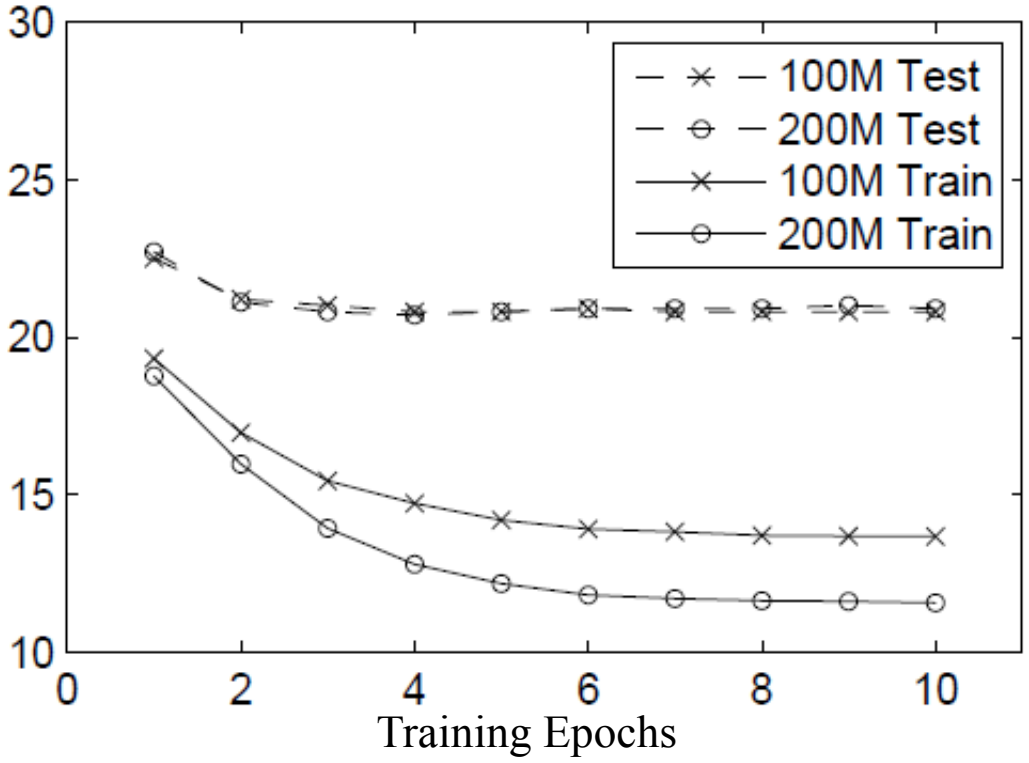


(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

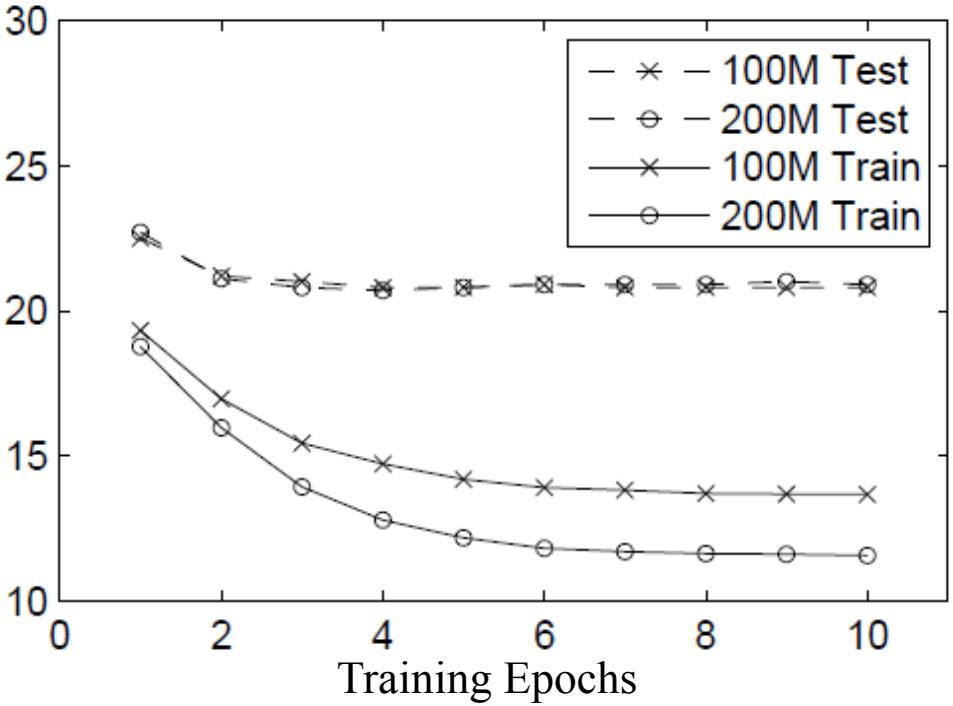
Experimental Framework

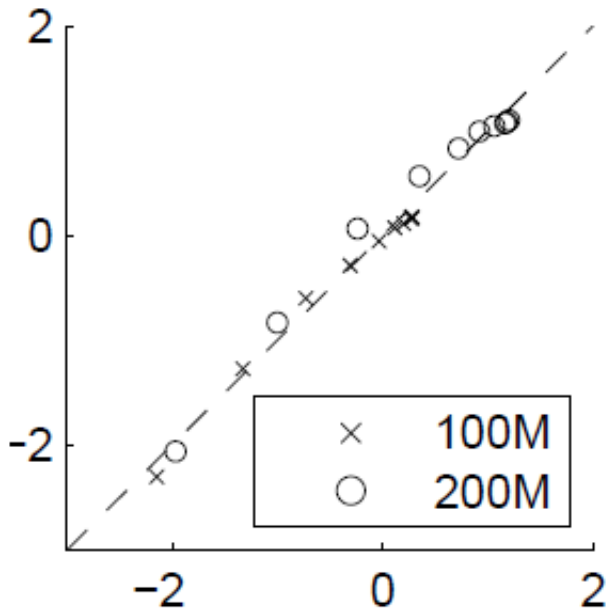
- 1-4 GPUs using the infrastructure of Coates et al (ICML 2013)
- Improved baseline GMM system. ~9,000 HMM states
- Fix DNN hidden layers to 5
- Vary total number of parameters, same number of hidden units in each hidden layer
- Evaluate input context of 21 and 41 frames
- Sizes evaluated: 36M, 100M, 200M
- Layer sizes: 2048, 3953, 5984
- Output layer:
  51% of all parameters for a 36M DNN
  6% of all parameters for a 200M DNN

Word error rate during training



Correlating frame-level metrics and WER





Normalize training set word accuracy and negative cross entropy separately. Strong correlation (r=0.992) suggests that cross entropy is a good predictor of *training set* WER

Generalization: Early realignment

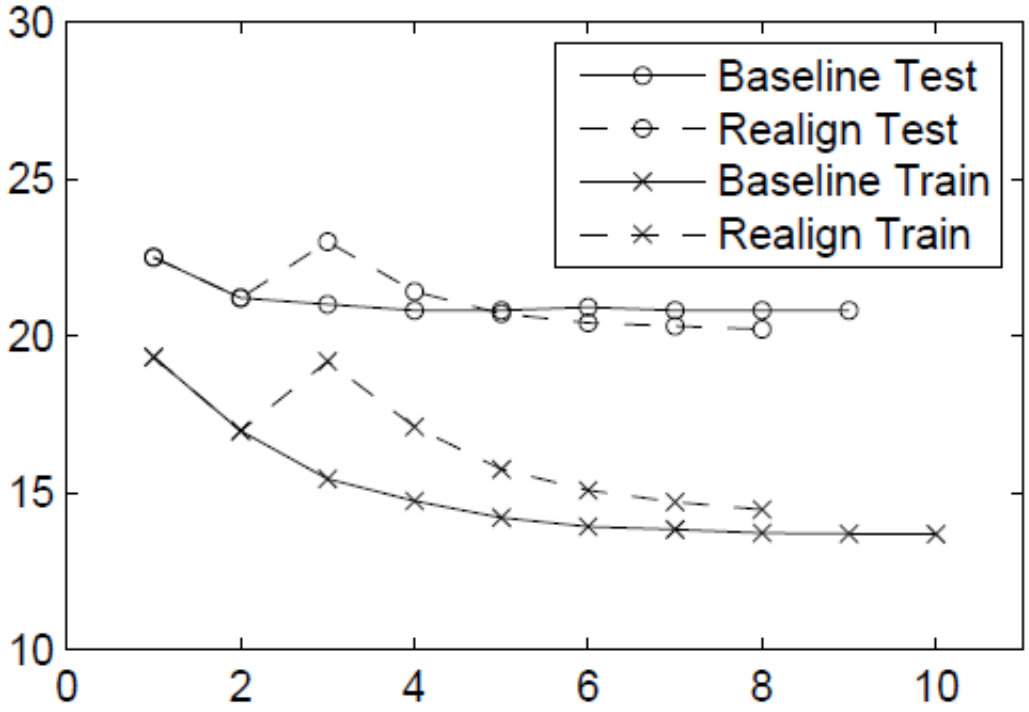
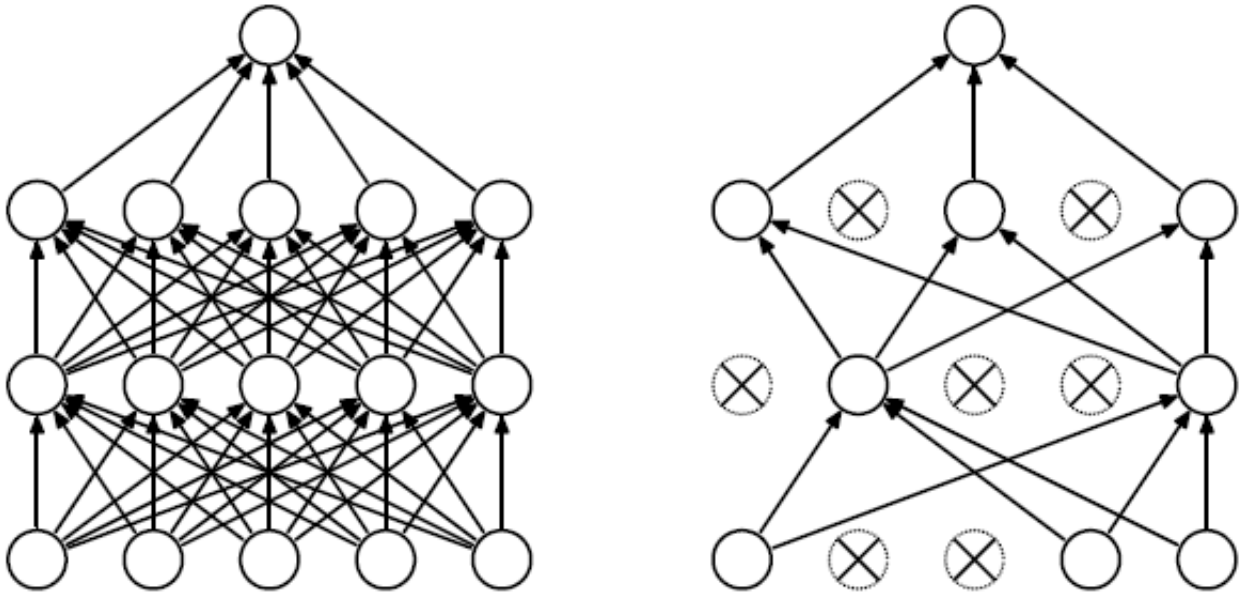


Figure 3. Word error rates on the training and test sets for LVCSR systems with DNN acoustic models trained with and without label realignment after epoch 2. A DNN which re-generates its training labels with a forced alignment early during optimization generalizes much better to test data than a DNN which converges to the original labels.

Generalization: Dropout

- During training randomly zero out hidden unit activations with probability p=0.5
- Cross validate over p in initial experiments
- Previous work found dropout helped on 50hr broadcast news but did not directly evaluate dropout with control experiments (Dahl et al 2013, Sainath et al 2013)

Improving Generalization with Dropout

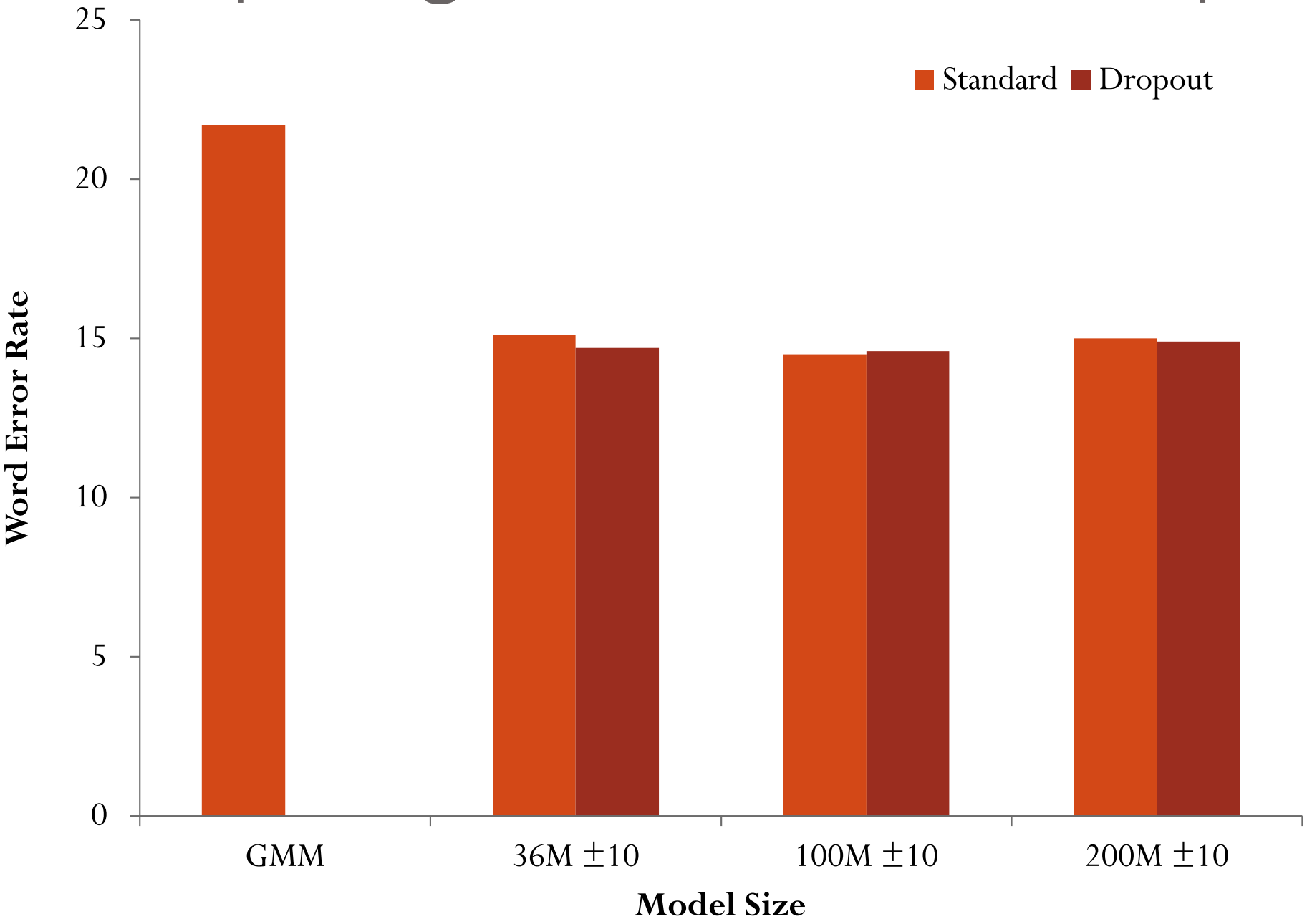


(a) Standard Neural Net

(b) After applying dropout.

(Srivastava, Hinton, Kirzhevsky, Sutskever, & Salakhutdinov. 2014)

Improving Generalization with Dropout

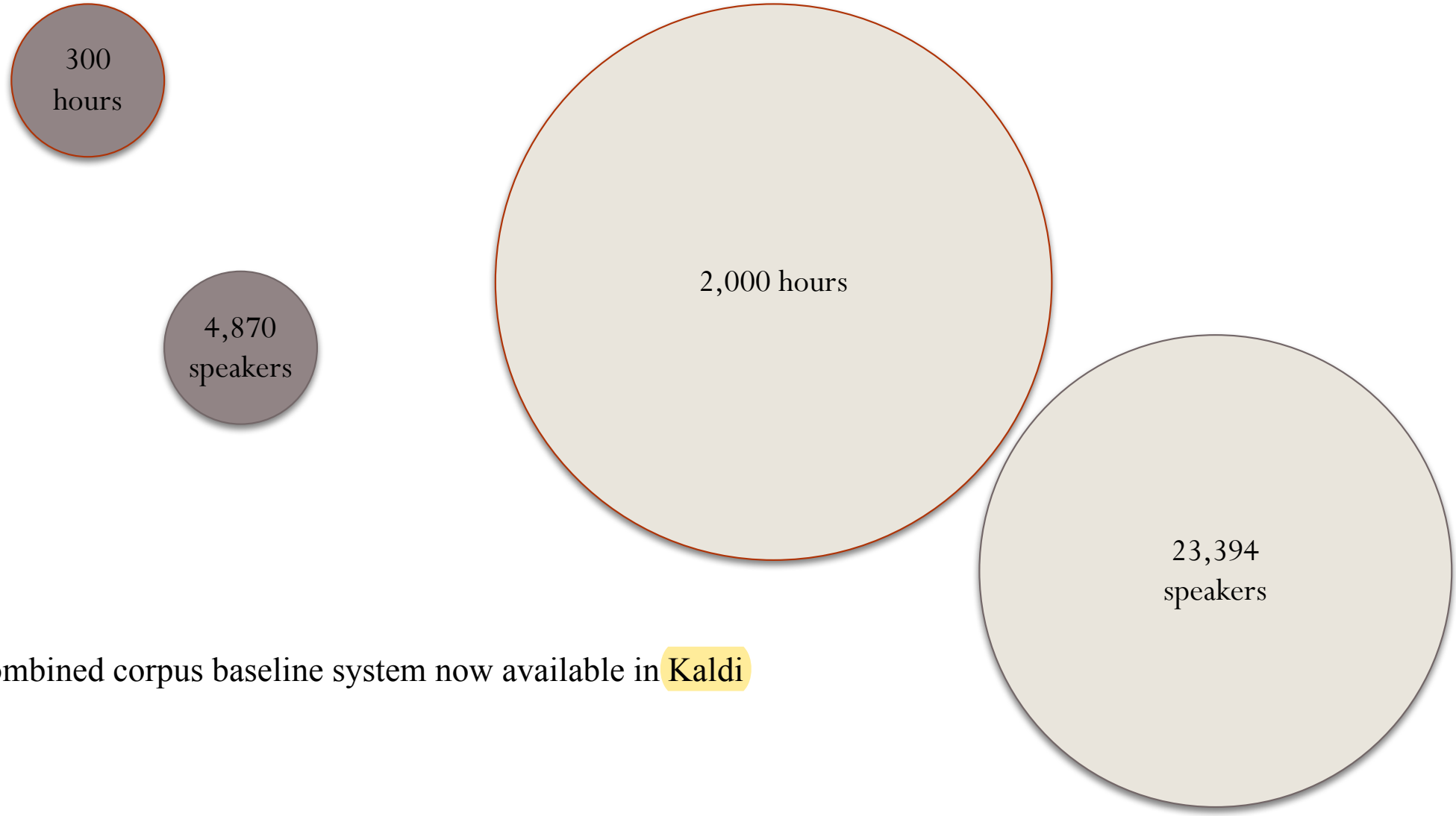


(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Combining Speech Corpora

Switchboard

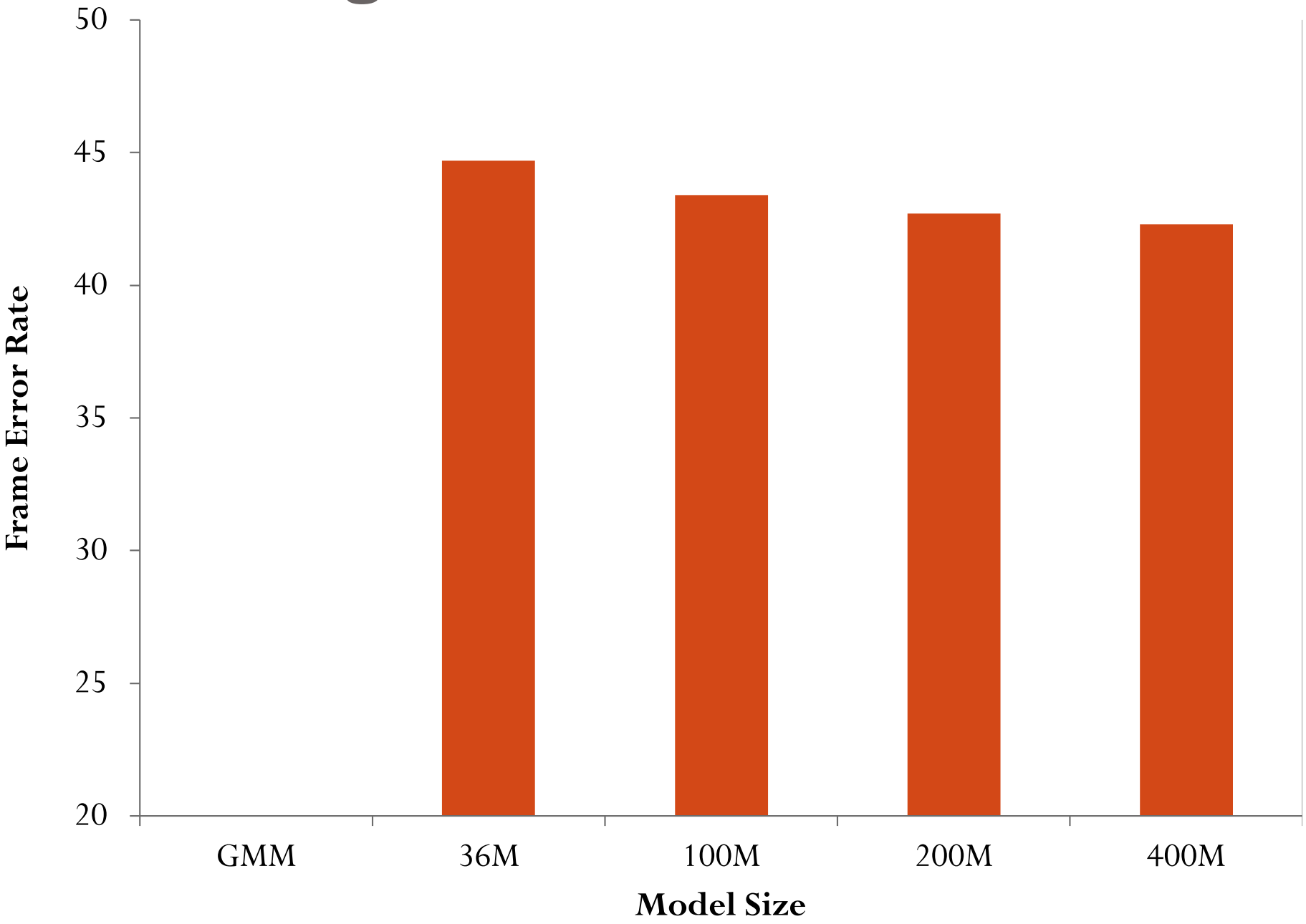
Fisher



Combined corpus baseline system now available in Kaldi

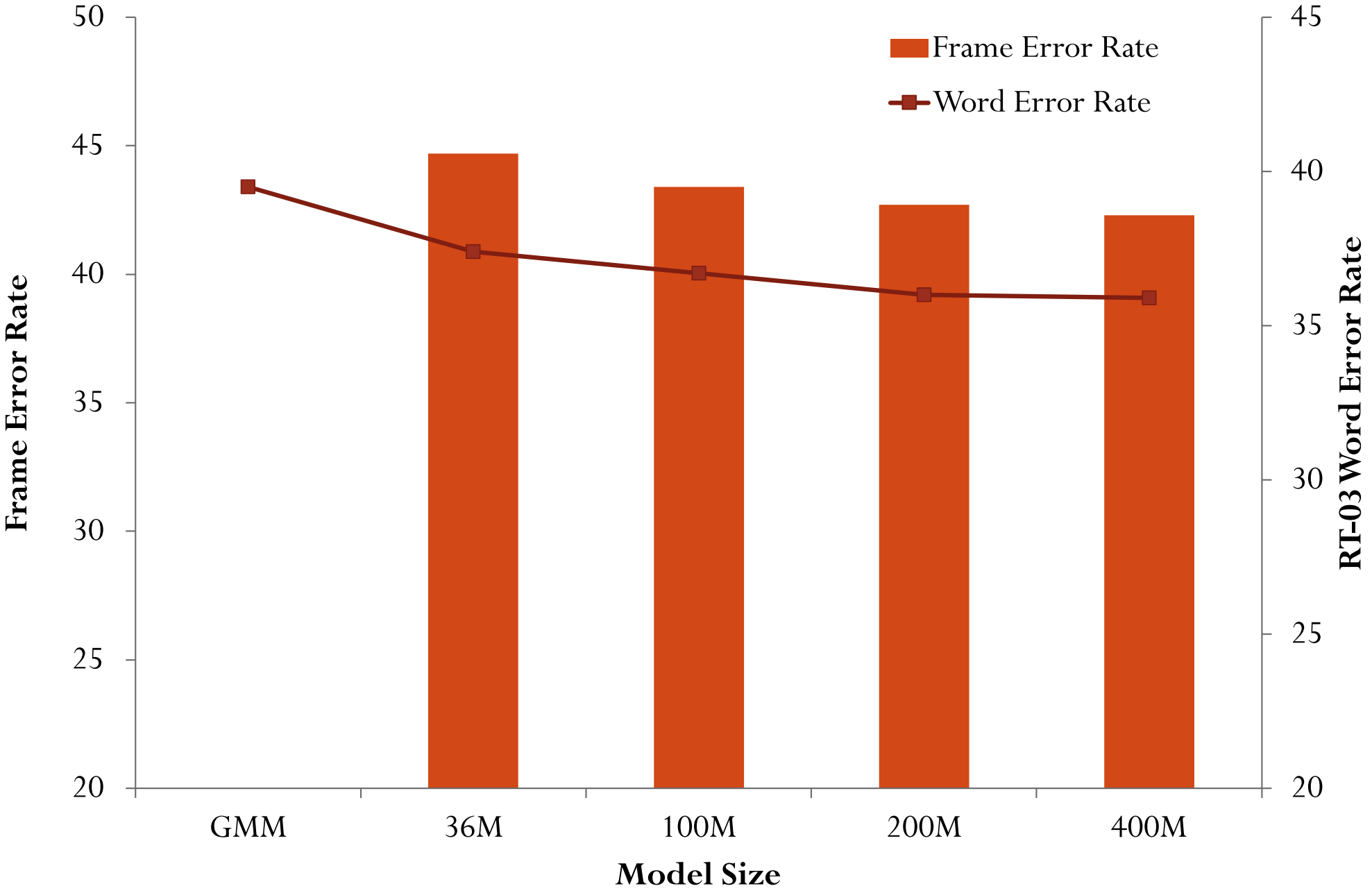
(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In

Scaling Total Parameters Revisited



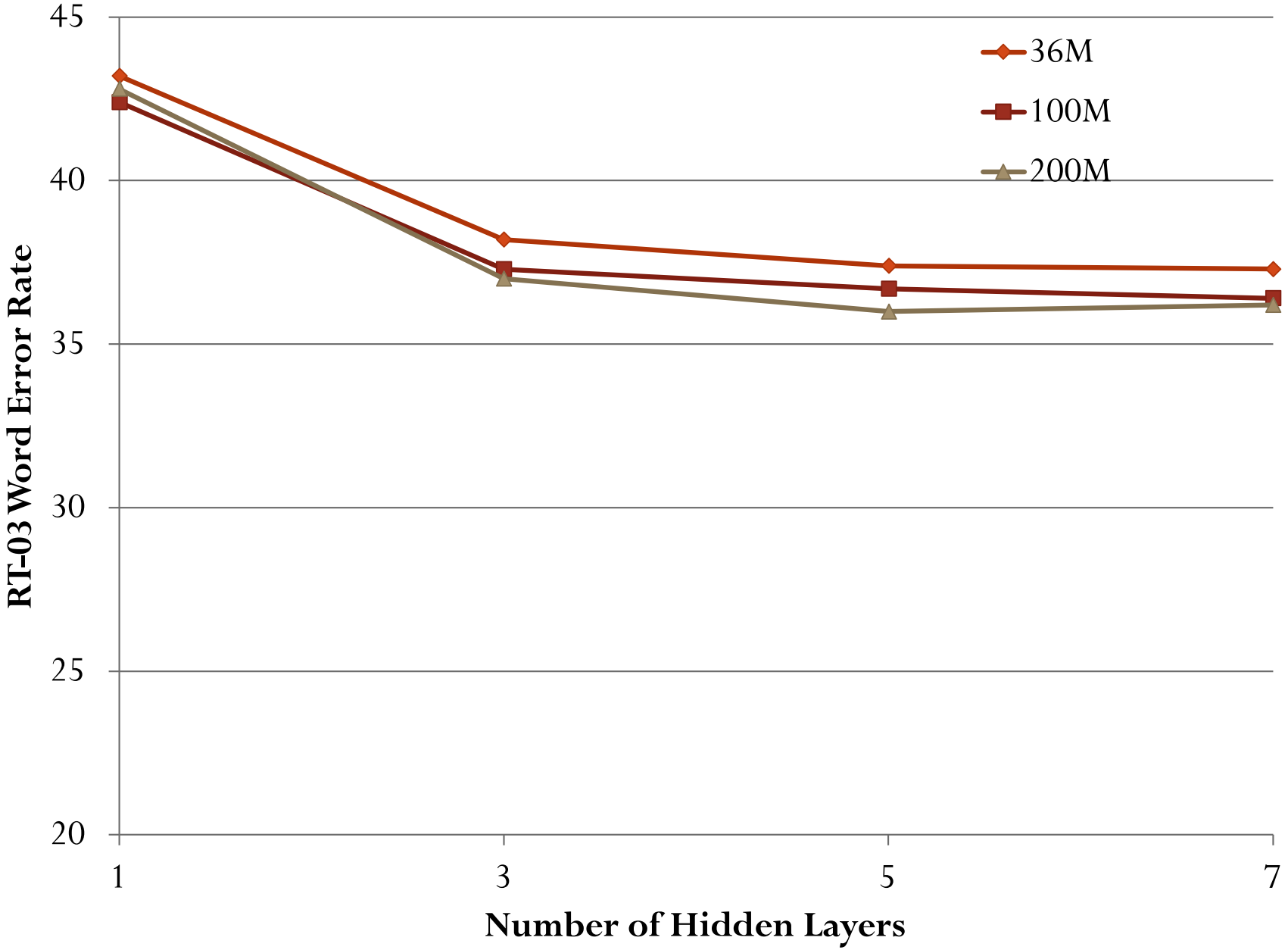
(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Scaling Total Parameters Revisited



(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Impact of Depth



(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Frame accuracy analysis

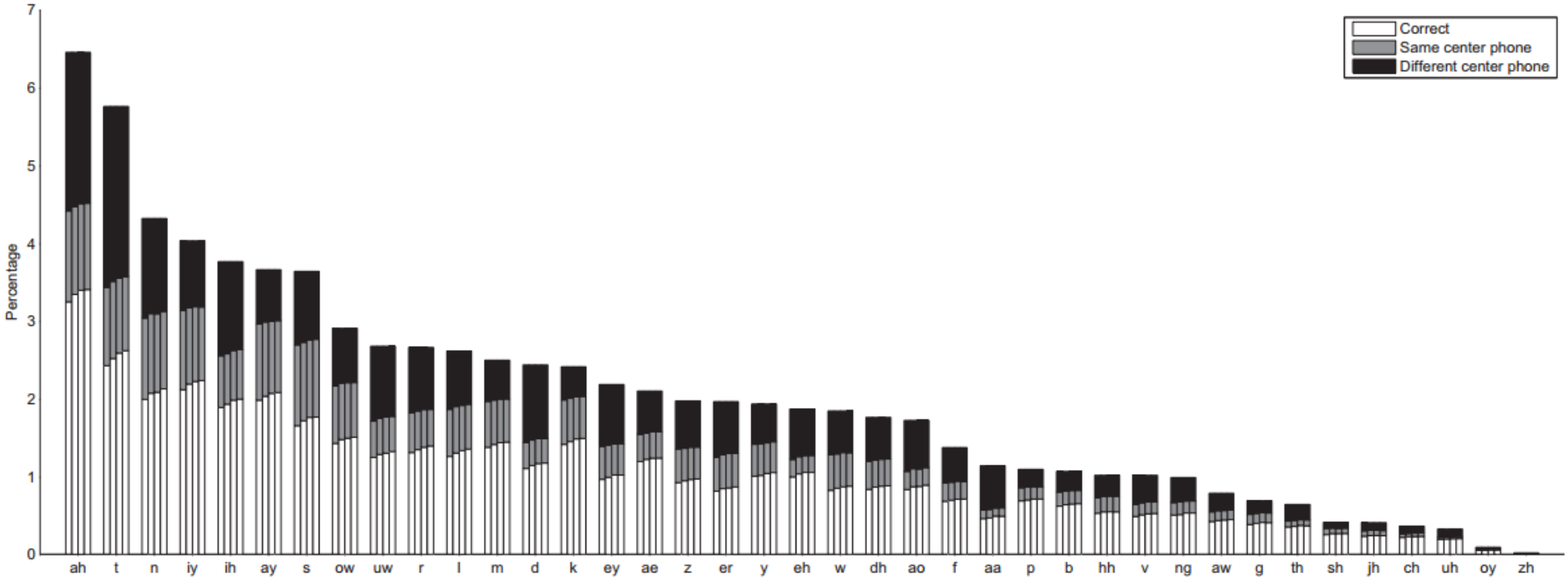


Fig. 7. Senone accuracy of 5 hidden layer DNN systems of varying total parameter count. Accuracy is grouped by base phone and we report the percentage correct, mis-classifications which chose a senone of the same base phone, and mis-classifications which chose a senone of a different base phone. The total size of the combined bar indicates the occurrence rate of the base phone in our data set. Each base phone has five bars, each representing the performance of a different five layer DNN. The bars show performance of DNNs of size 36M 100M 200M and 400M from left to right. We do not show the non-speech categories of silence, laughter, noise, or OOV which comprise over 20% of frames sampled.

(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Building A Strong DNN Acoustic Model

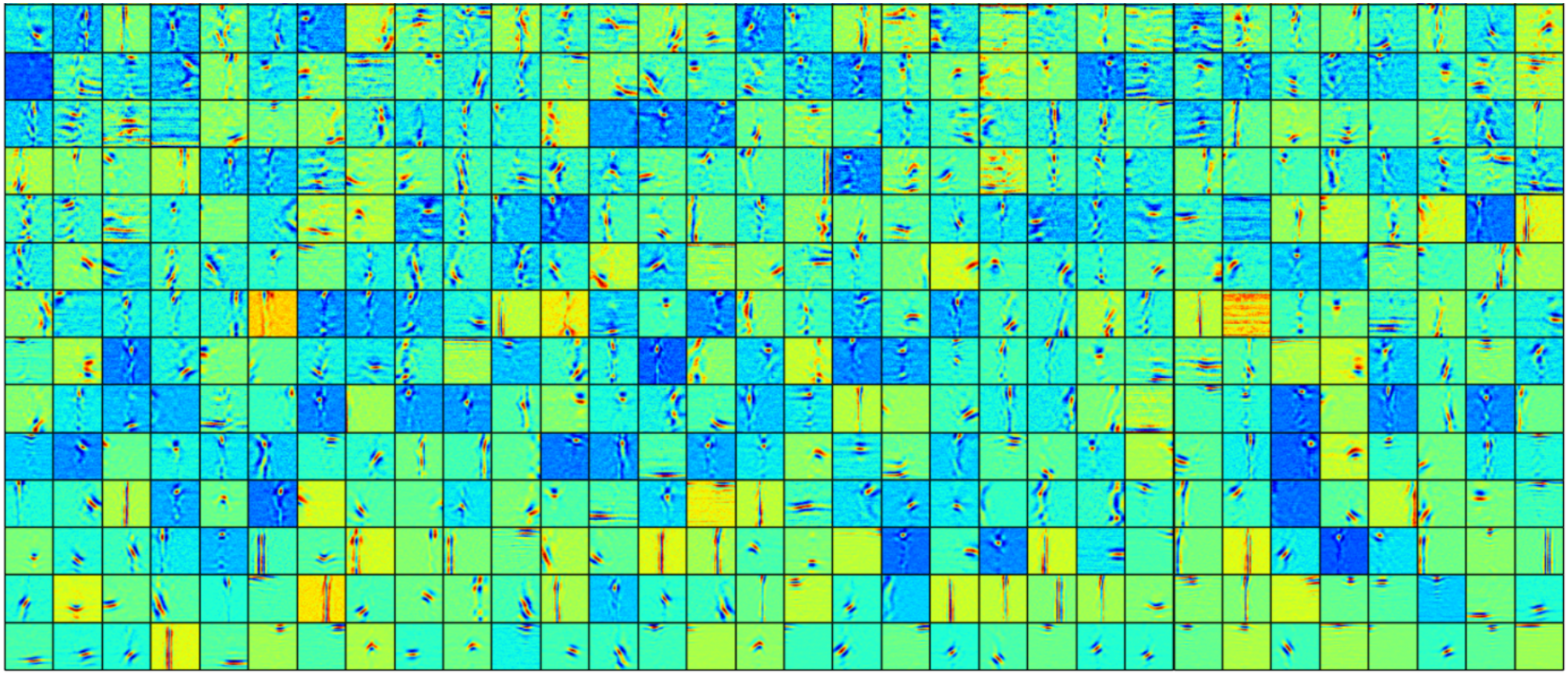
- Large
- At least 3 hidden layers
- Training data
- Less important:
  - Dropout regularization
  - Specific optimization algorithm settings
  - Initialization (we don't need pre-training)

(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. 2017)

Outline

- Hybrid acoustic modeling overview
  - Basic idea
  - History
  - Recent results
- What's different about modern DNNs?
- **Convolutional networks**
- **Recurrent networks**
- **Alternative loss functions**

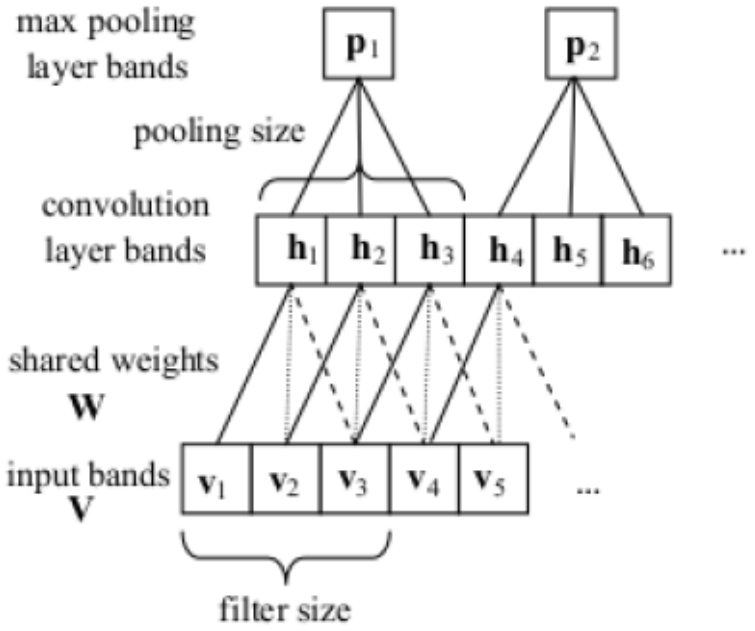
Visualizing first layer DNN features

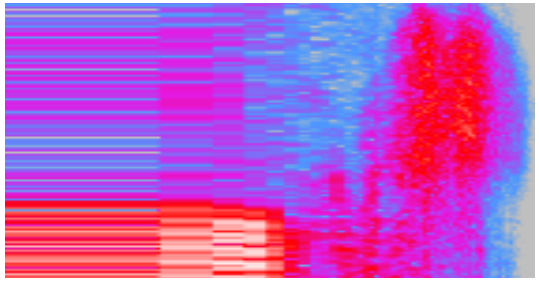


Convolutional Networks

- Slide your filters along the frequency axis of filterbank features
- Great for spectral distortions (e.g. short wave radio)

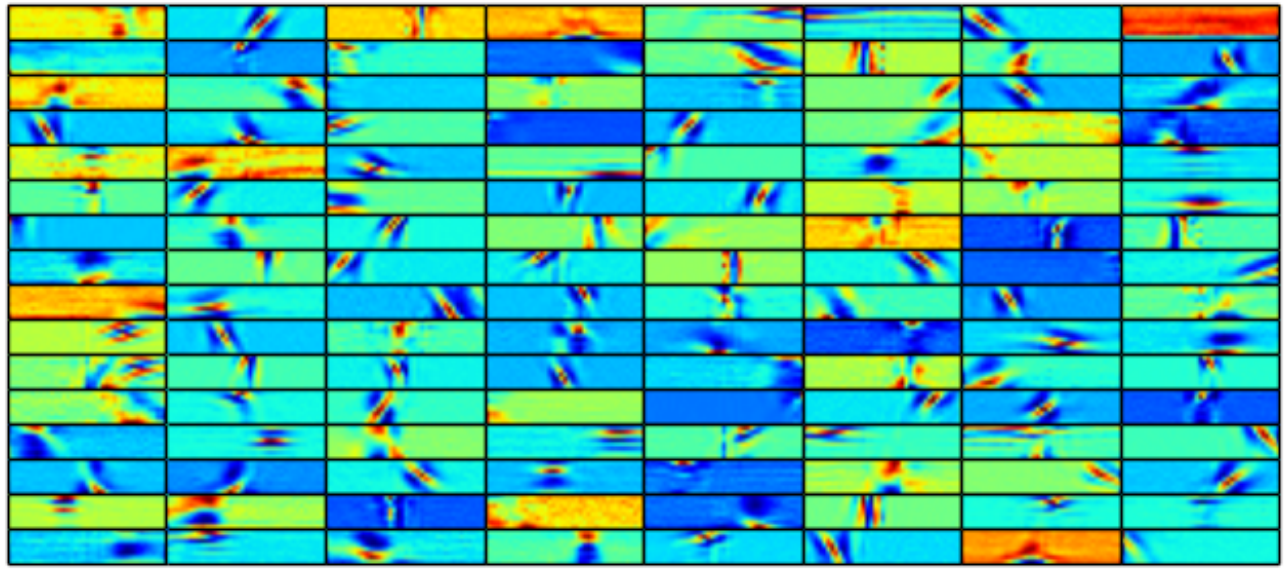
# Conv/Fully Connected Layers	WER
No Conv, 6 full (DNN)	22.6
1 conv, 5 full	22.3
2 conv, 4 full	19.9
3 conv, 3 full	21.2





(Sainath, Mohamed, Kingsbury, & Ramabhadran. 2013)

Learned Convolutional Features



Features for convolutional nets

- Porting VTLN and fMLLR feature transforms important

Feature	WER – 50 hr BN
Log-Mel	19.7
Log-mel + d + dd	19.5
Log-mel + d + dd + energy	19.8
VTLN-warped log-mel+ d + dd	**18.9**

Method	WER
VTLN-warped log-mel+d+dd	18.9
fMLLR+VTLN-warped log-mel+d+dd	18.3
multi-Scale fMLLR (DNN) + VTLN-warped log-mel+d+dd (CNN)	**18.0**

Slide from Tara Sainath

Pooling in time

- Most CNN work in speech pools in frequency only, though people in computer vision pool in both time and space (i.e. frequency)
- Pooling in time, with overlap, is a way to smooth out the signal
- Gains with pooling in time when targets are frame-level context dependent states are minimal

Method	WER – 50 hr BN
Baseline	18.9
Pooling in Time, Max	18.9
Pooling in Time, Stochastic	18.8
Pooling in Time, l_p	18.8

Comparing CNNs, DNNs, & GMMs

- GMM system: speaker adapted, discriminatively-trained, 9300 states and 150K Gaussians
- Baseline DNN trained with fMLLR features
- CNN trained with vtln-warped, log-mel+d+dd features
- All feature-based networks have 512 output targets, cross-entropy +sequence trained

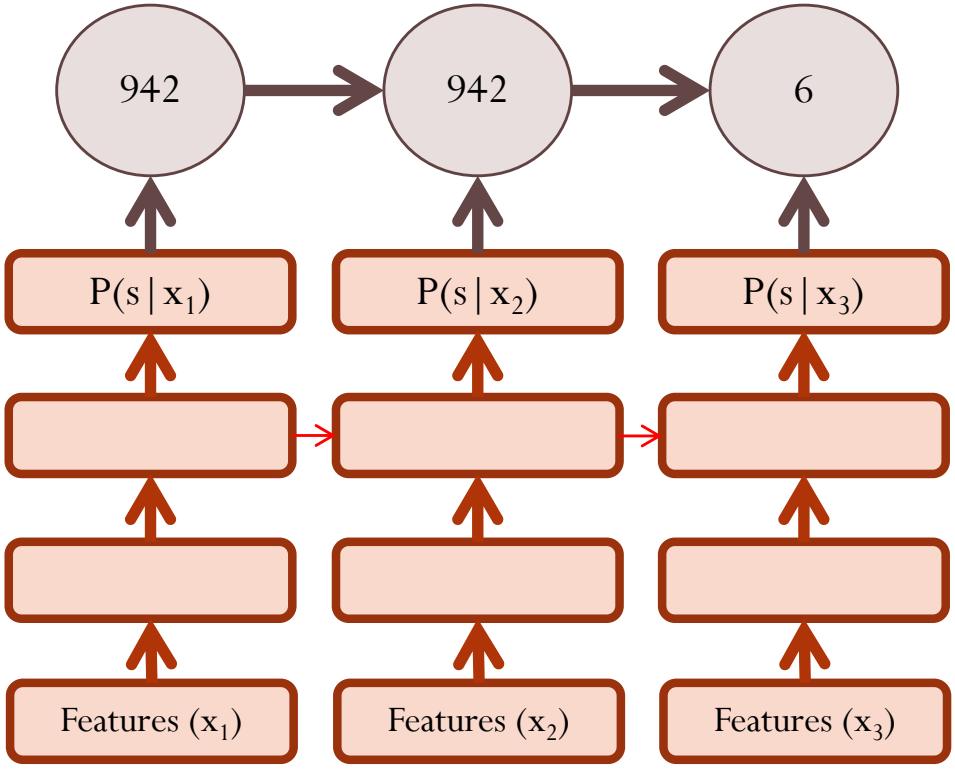
Model	Hub5	rt03 FSH	Rt03 SWB
Baseline GMM/HMM	14.5	17.0	25.2
Hybrid DNN	12.2	14.9	23.5
Hybrid CNN	**11.5**	**14.5**	**22.1**

Slide from Tara Sainath

Recurrent DNN Hybrid Acoustic Models

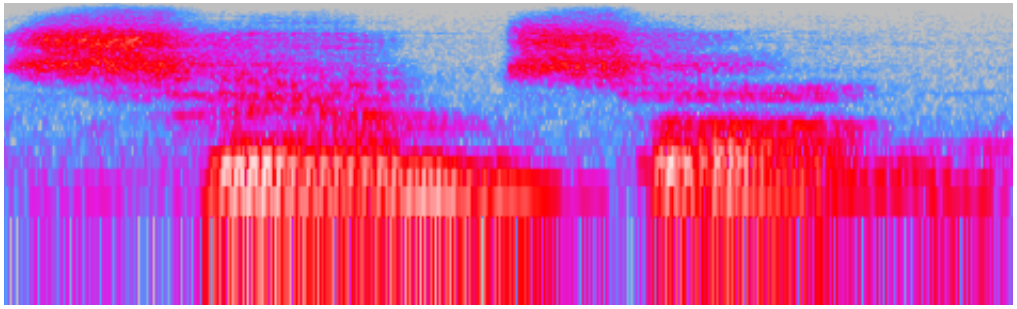
Transcription: Samson
Pronunciation: S – AE – M – S –AH – N
Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

Hidden Markov Model (HMM):

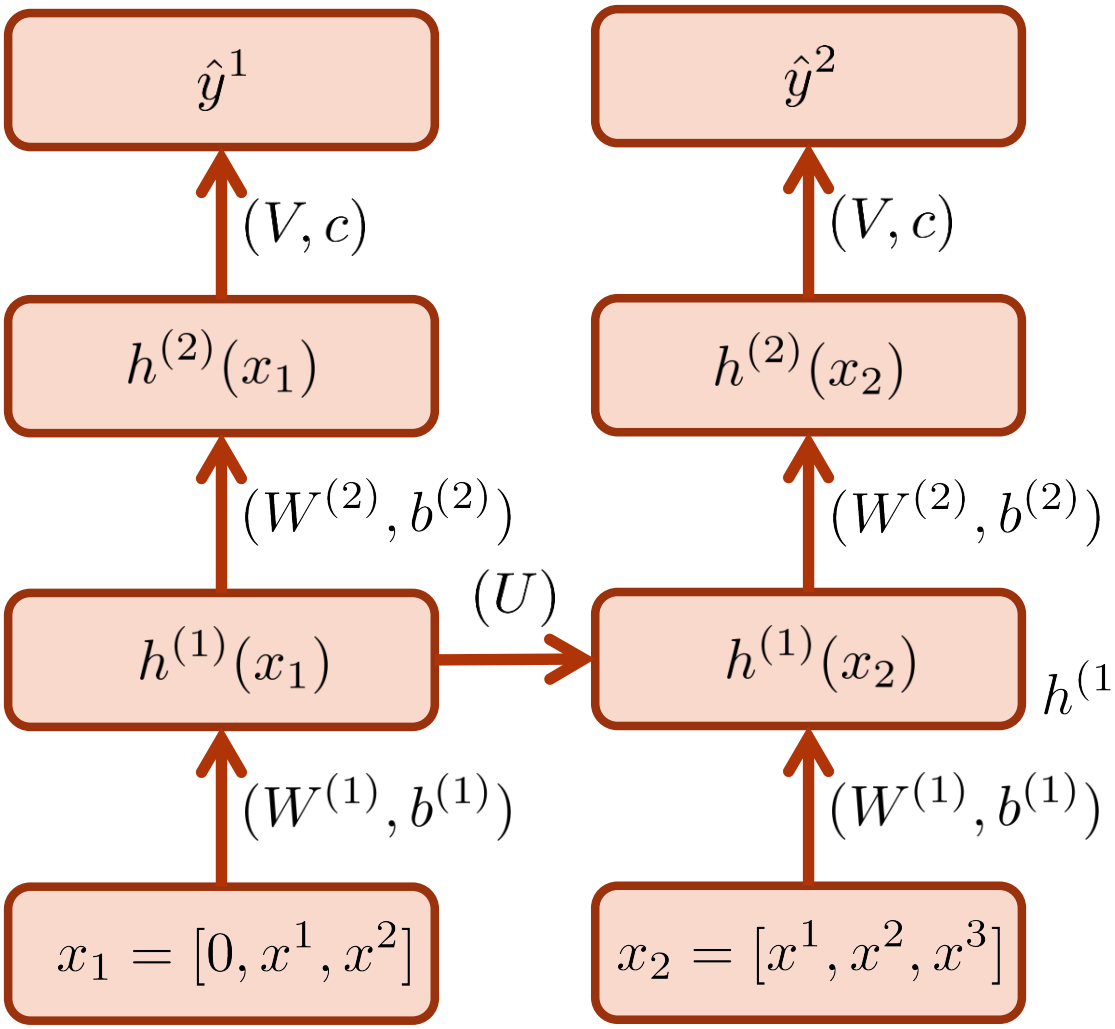


Acoustic Model:

Audio Input:



Deep Recurrent Network



Output Layer

$$\hat{y}^2 = Vh^{(1)}(x_2) + c$$

Hidden Layer

$$h^{(2)}(x_2) = \sigma(W^{(2)}h^{(1)}(x_2) + b^{(2)})$$

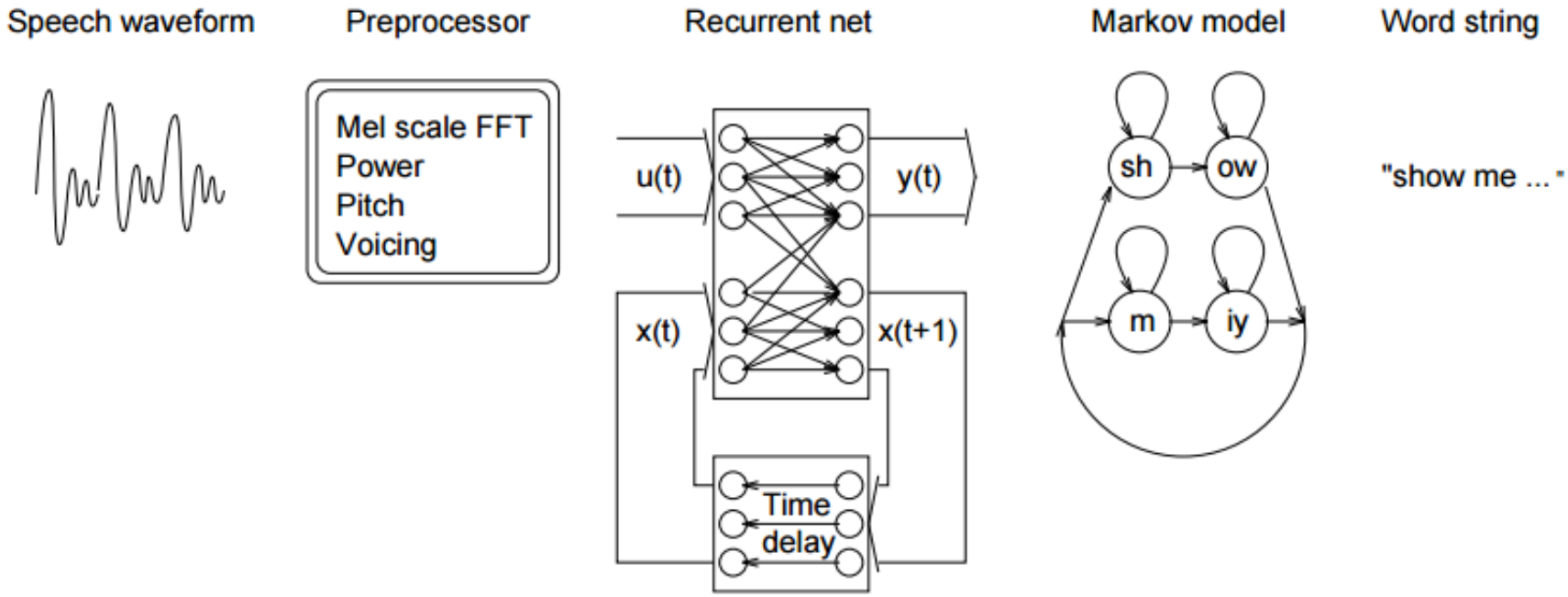
Hidden Layer

$$h^{(1)}(x_2) = \sigma(W^{(1)}x_2 + b^{(1)} + Uh^{(1)}(x_1))$$

Input

$\mathbf{X}$

RNNs for acoustic modeling in 1996



(Robinson, Hochberg, & Renals. 1996)

LSTM RNN on Google voice search

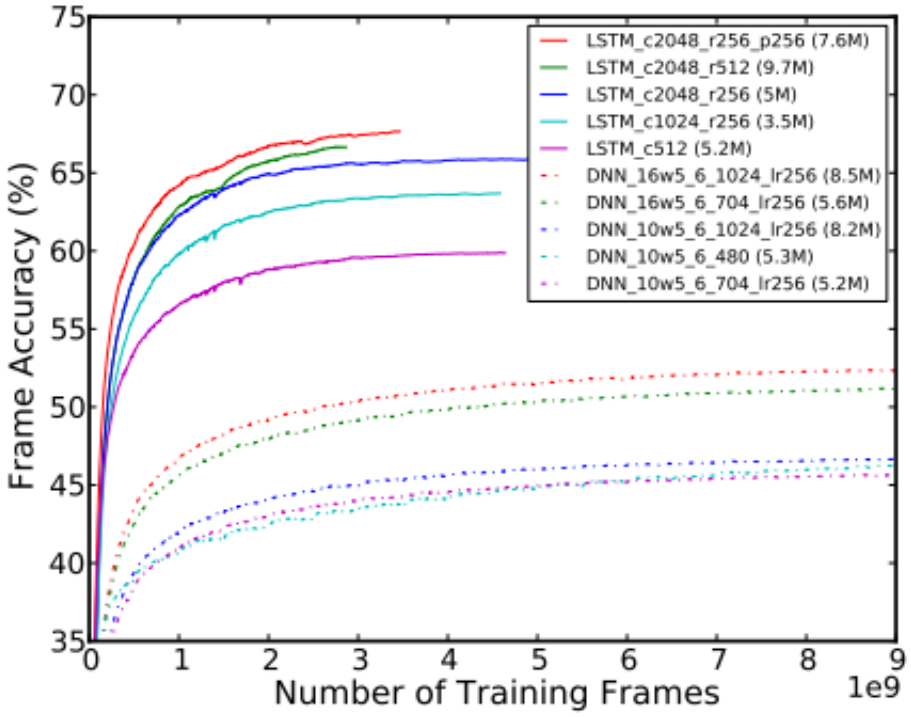


Fig. 4. 8000 context dependent phone HMM states.

- Best LSTM WER: 10.5%
- Best DNN WER: 11.3%
- LSTM required half the training time

(Sak, Senior, & Beufays. 2014)

Alternative loss functions

- Scalable minimum Bayes risk (sMBR)
- Minimum phone error (MPE)
- Maximum mutual information (MMI)
- Move beyond frame classification to decoder outputs

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)},$$

Other Current Work

- Multi-lingual acoustic modeling
- Low resource acoustic modeling
- Learning directly from waveforms without complicated feature transforms