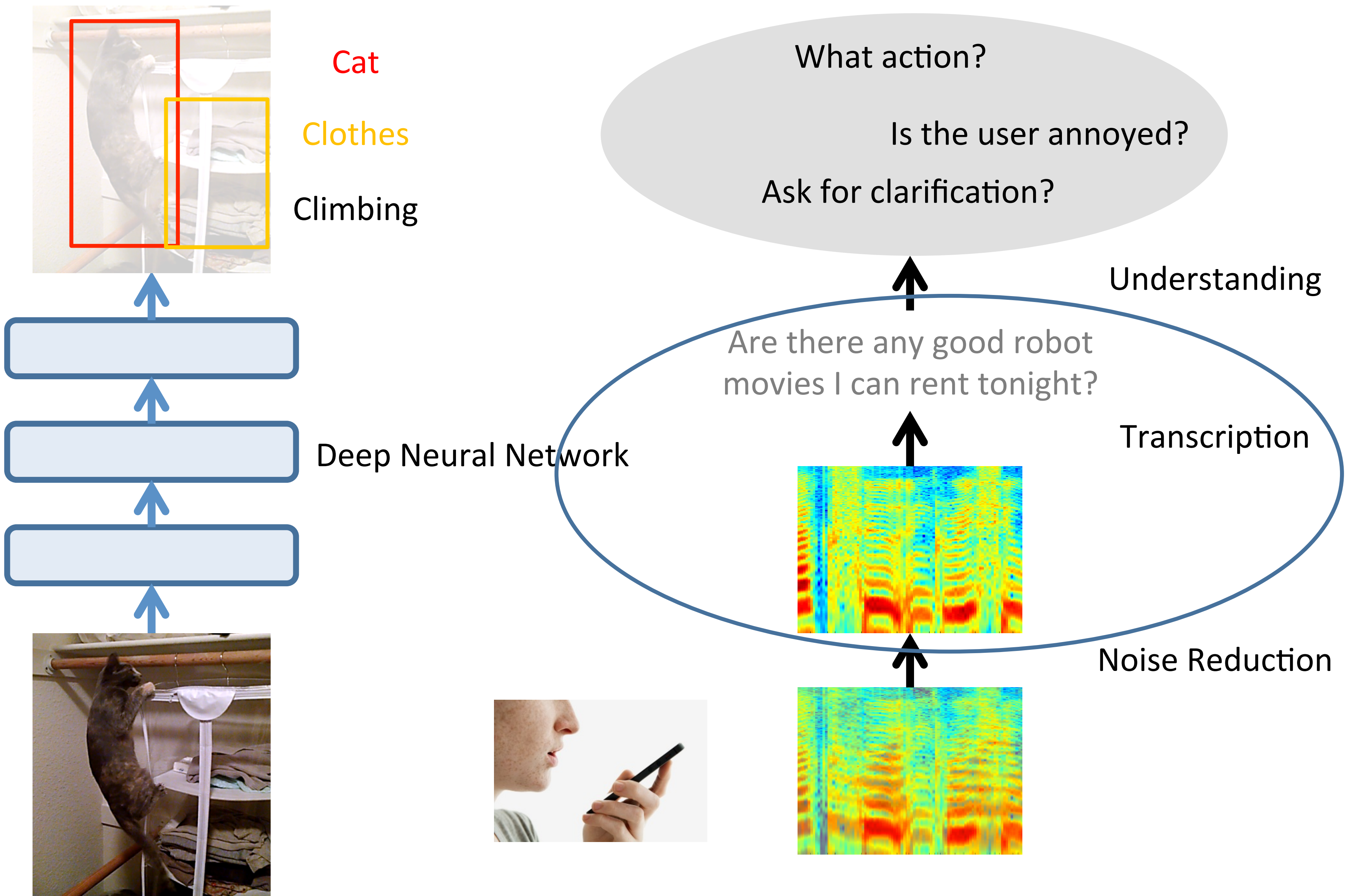
CS224D: Deep Learning for Natural Language Processing

Andrew Maas
Spring 2016

Neural Networks in Speech Recognition

Outline

- Speech recognition systems overview
- HMM-DNN (Hybrid) acoustic modeling
- What's different about modern HMM-DNNs?
- HMM-free RNN recognition



Andrew Maas. Stanford CS224D. 2016

Conversational Speech Data

Switchboard

300 hours

4,870 speakers

but it was really nice to get back with a telephone and the city and everything and you know yeah

well (i-) the only way i could bear it was to (pass) (some) to be asleep i was like well it is not gonna (be-) get over until you know (w-) (w-) yeah it (re-) really i (th-) i think that is what ruined it for us

Outline

- Speech recognition systems overview
- **HMM-DNN (Hybrid) acoustic modeling**
- What's different about modern HMM-DNNs?
- HMM-free RNN recognition

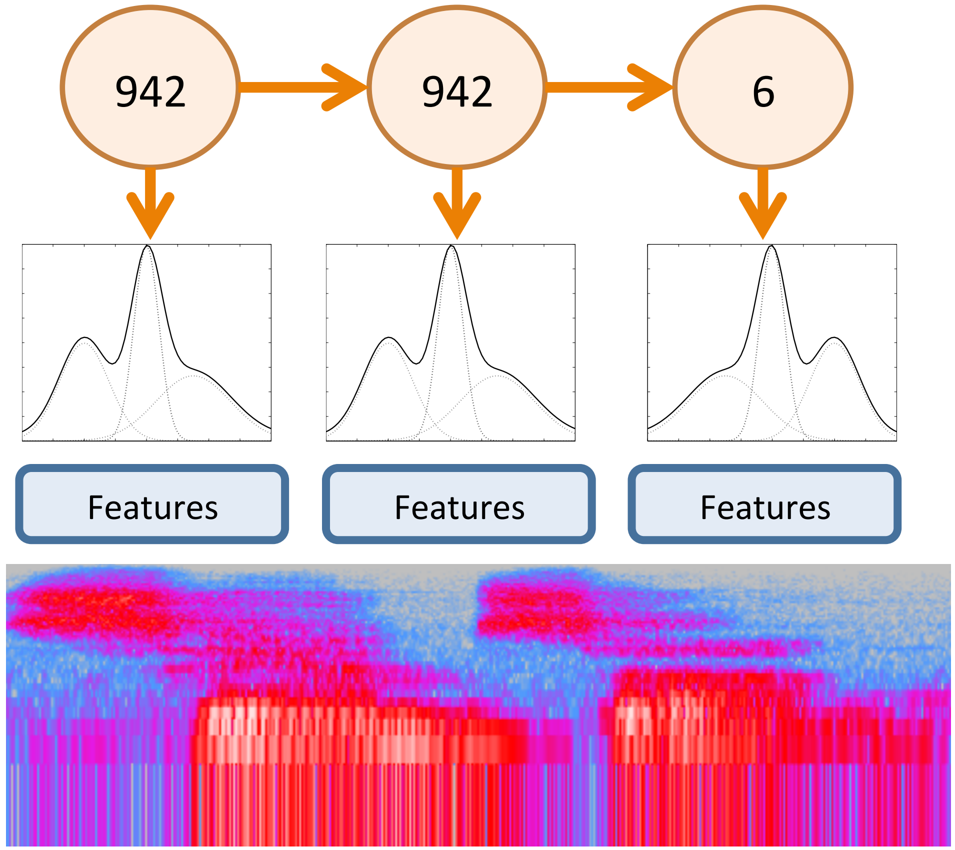
Acoustic Modeling with GMMs

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

Andrew Maas. Stanford CS224D. 2016

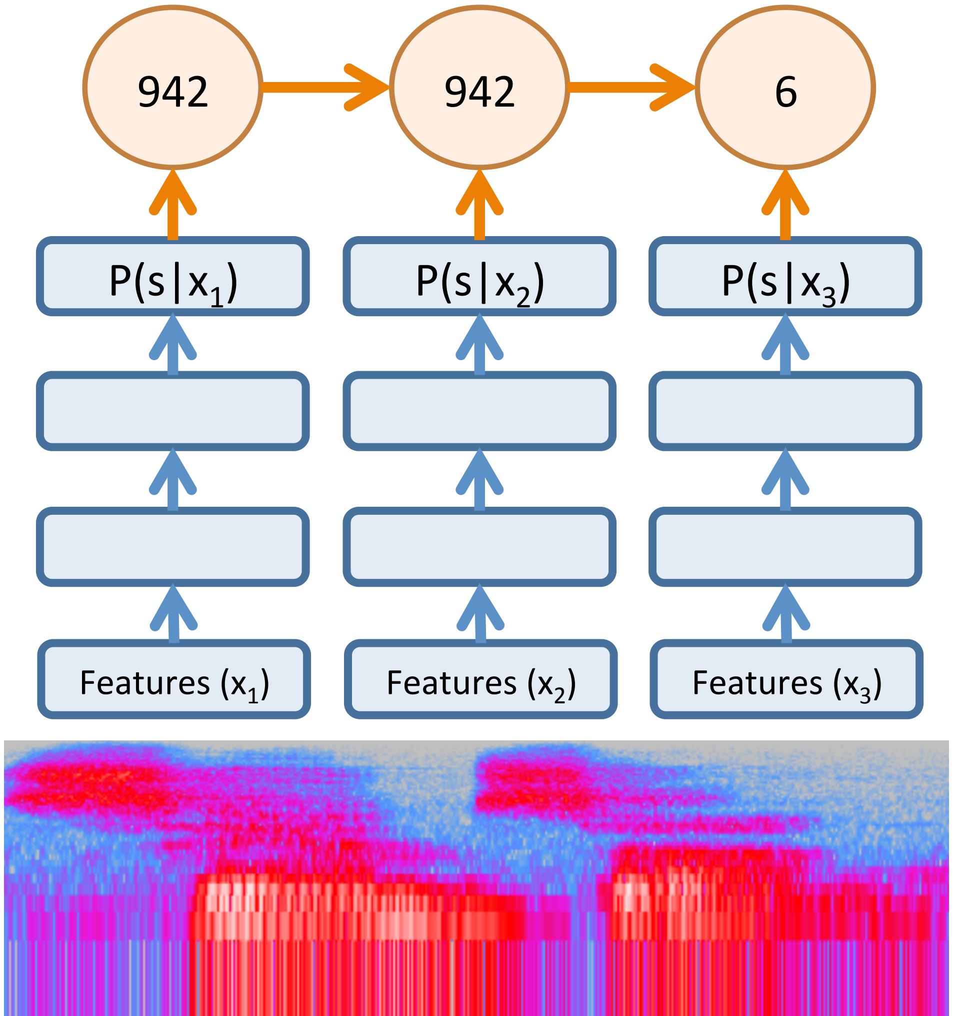
DNN Hybrid Acoustic Models

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

Hidden Markov Model (HMM):

Acoustic Model:

Audio Input:



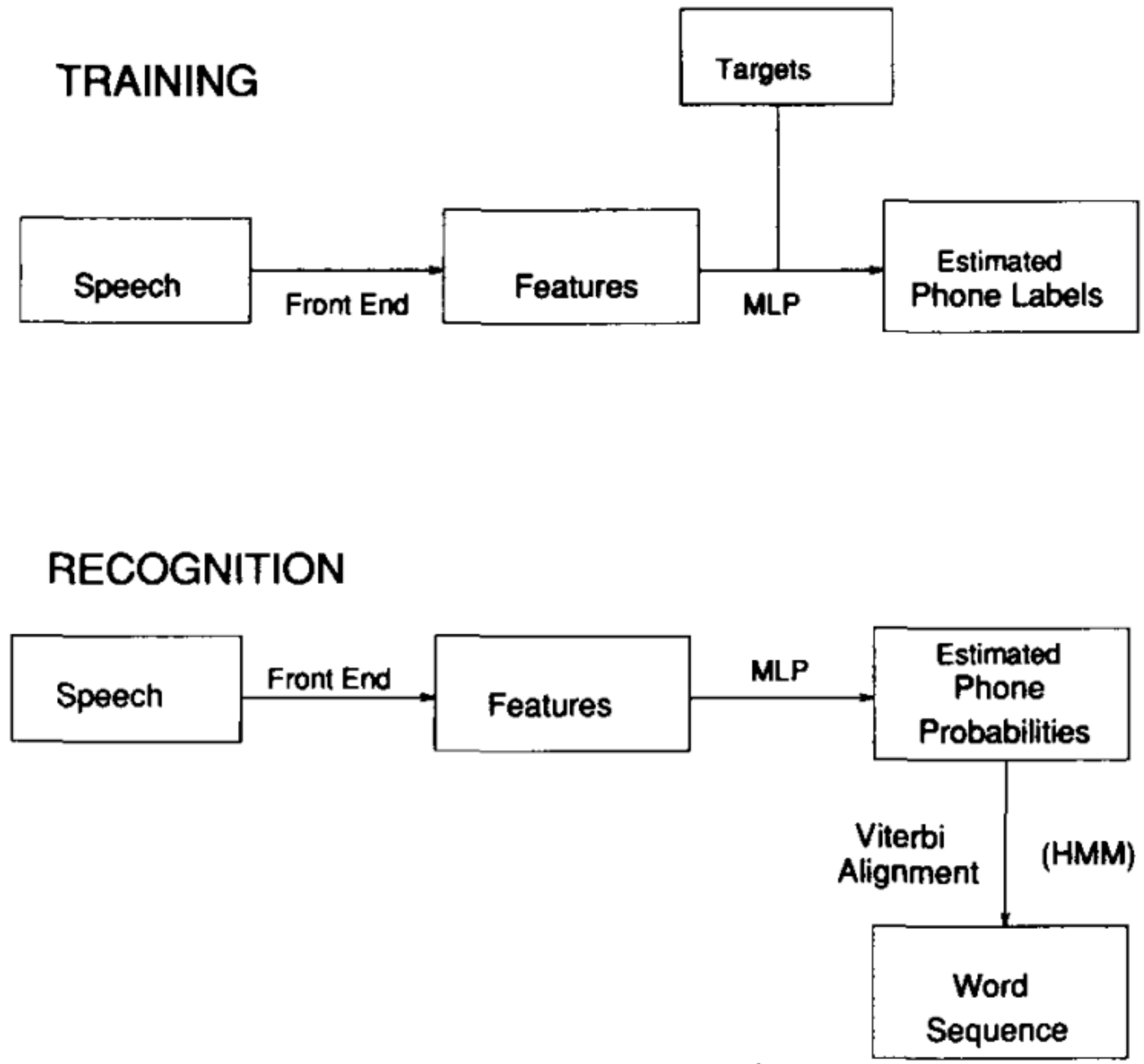
Use a DNN to approximate: P(s|x)

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

DNN * Constant / State prior

Andrew Maas. Stanford CS224D. 2016

Not Really a New Idea



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

Modern Systems use **D**NNs and Senones

COMPARISON OF CONTEXT-INDEPENDENT MONOPHONE STATE LABELS AND CONTEXT-DEPENDENT TRIPHONE SENONE LABELS

# Hidden Layers	# Hidden Units	Label Type	Dev Accuracy
1	2K	Monophone States	59.3%
1	2K	Triphone Senones	68.1%
3	2K	Monophone States	64.2%
3	2K	Triphone Senones	69.6%

Criterion	Dev Accuracy	Test Accuracy
ML	62.9%	60.4%
MMI	65.1%	62.8%
MPE	65.5%	63.8%

Dahl, Yu, Deng & Acero. 2011.

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?

- Fast computers = run many experiments
- Deeper nets improve on shallow nets
- Architecture choices (easiest is replacing sigmoid)
- Pre-training *matters very little*. Initially we thought this was the new trick that made things work
- Many more parameters

Depth Matters (Somewhat)

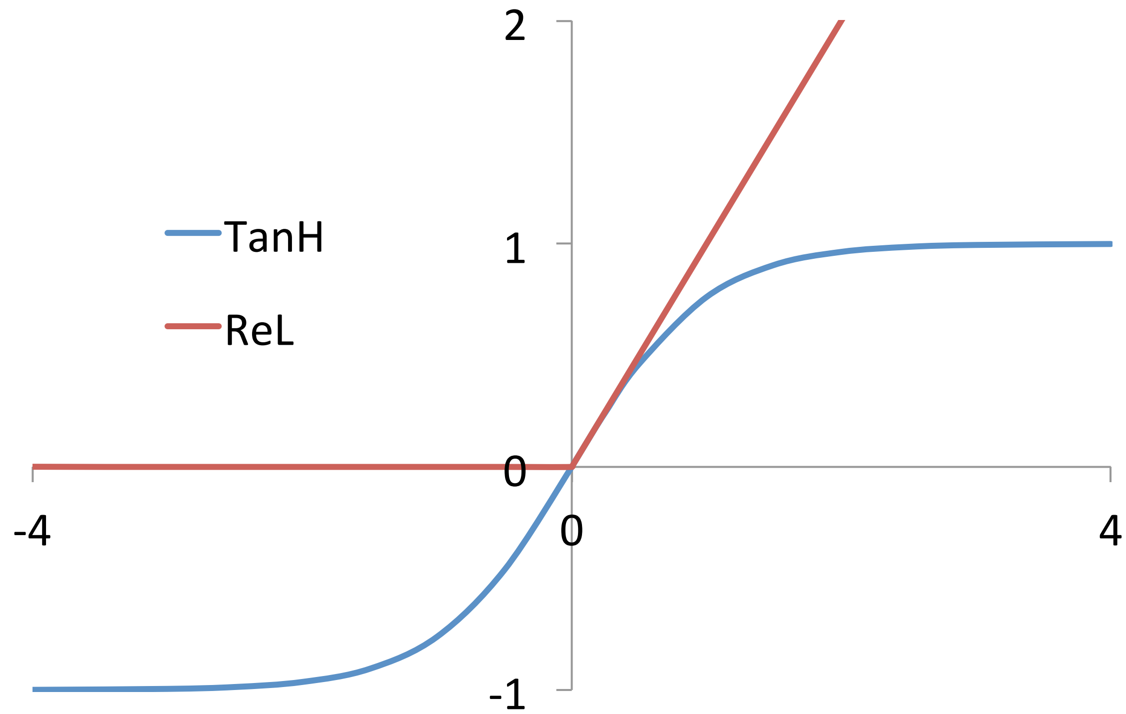
Table 1: Effect of CD-DNN-HMM network depth on WER (%) on Hub5'00-SWB using the 309-hour Switchboard training set. DBN pretraining is applied.

$L \times N$	WER	$1 \times N$	WER
$1 \times 2k$	24.2	–	–
$2 \times 2k$	20.4	–	–
$3 \times 2k$	18.4	–	–
$4 \times 2k$	17.8	–	–
$5 \times 2k$	17.2	1×3772	22.5
$7 \times 2k$	17.1	1×4634	22.6
$9 \times 2k$	17.0	–	–
$5 \times 3k$	17.0	–	–
–	–	$1 \times 16k$	22.1

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.

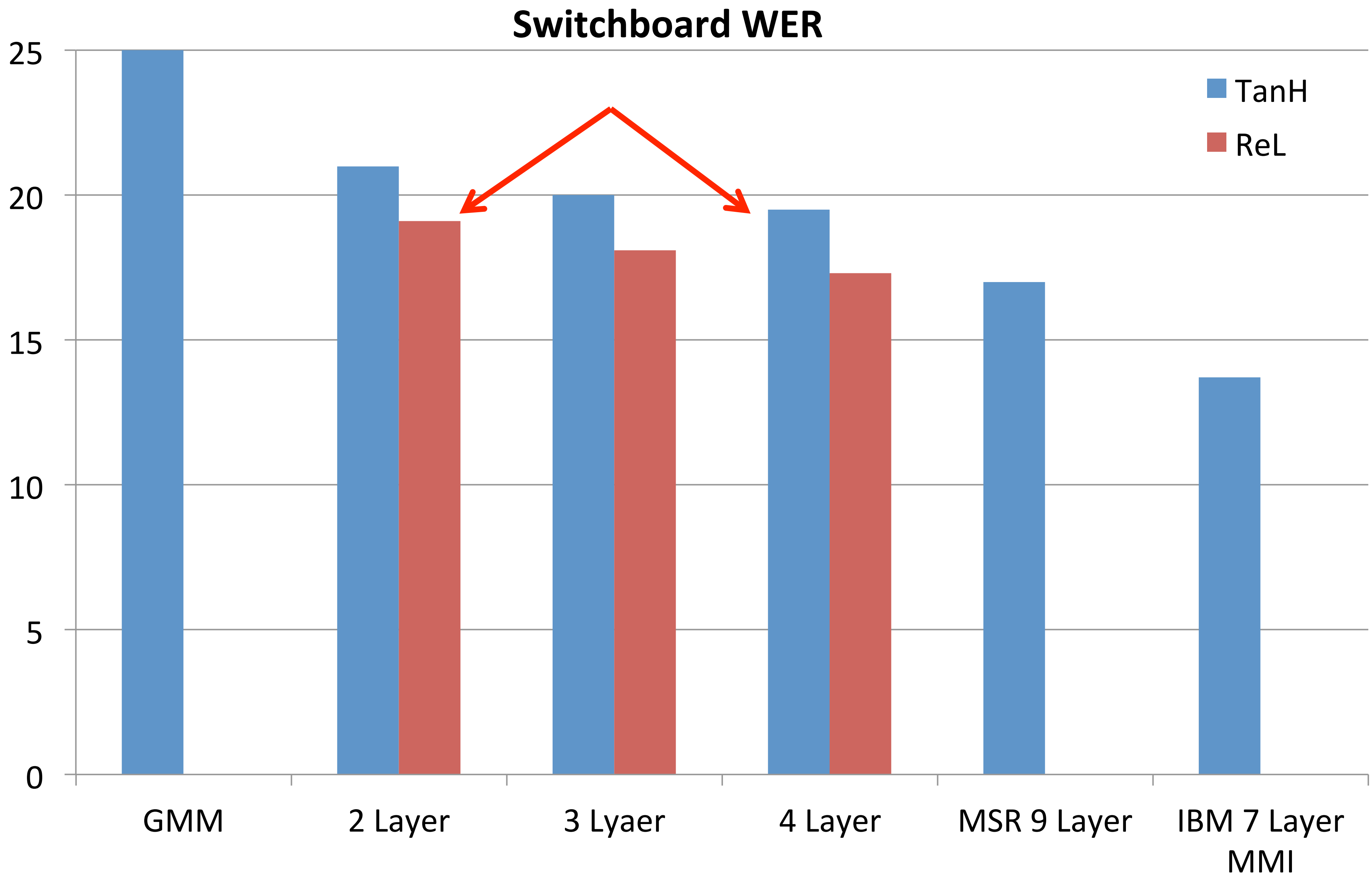
Yu, Seltzer, Li, Huang, Seide. 2013.

Replacing Sigmoid Hidden Units



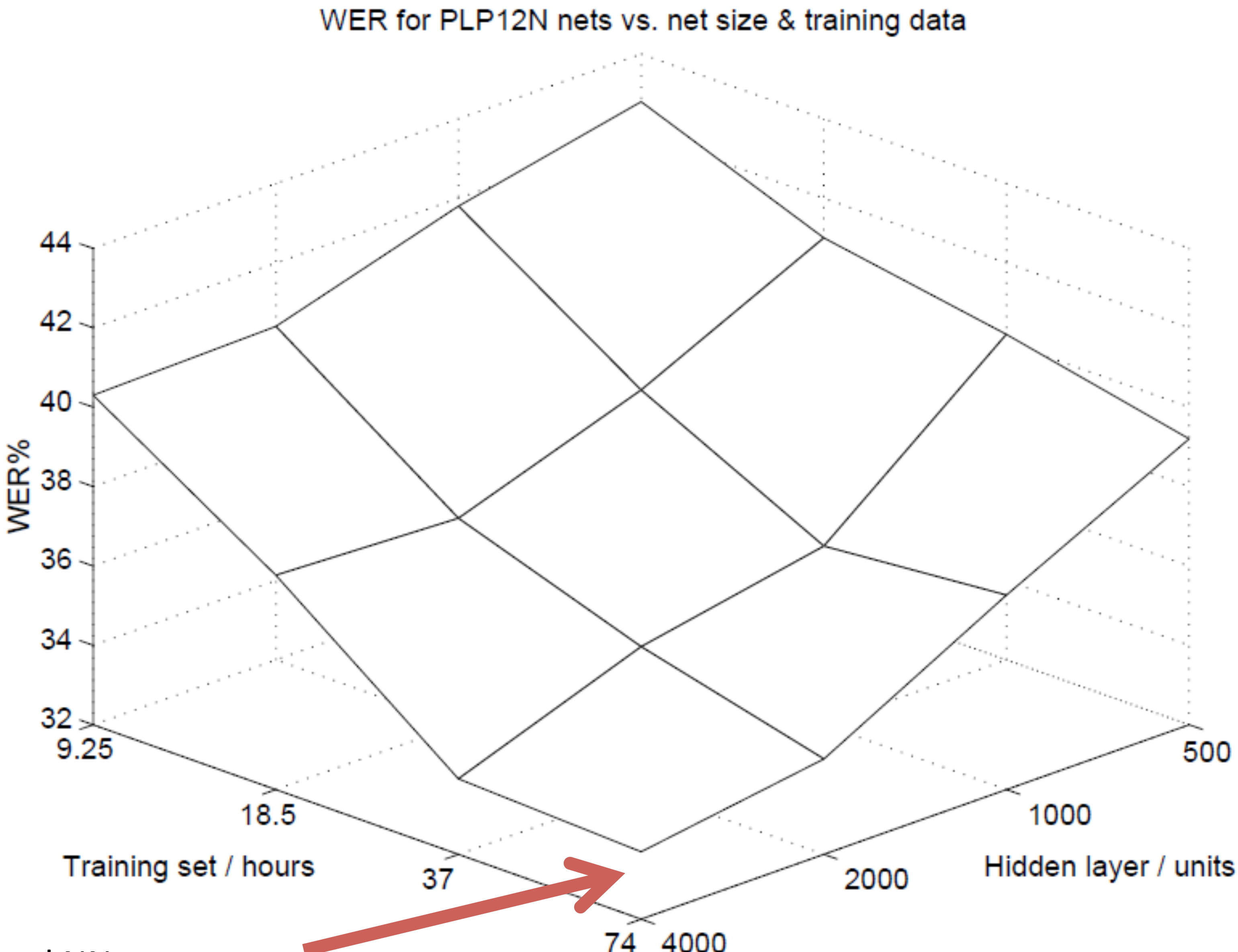
(Glorot & Bengio. 2011)

Comparing Nonlinearities



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

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.

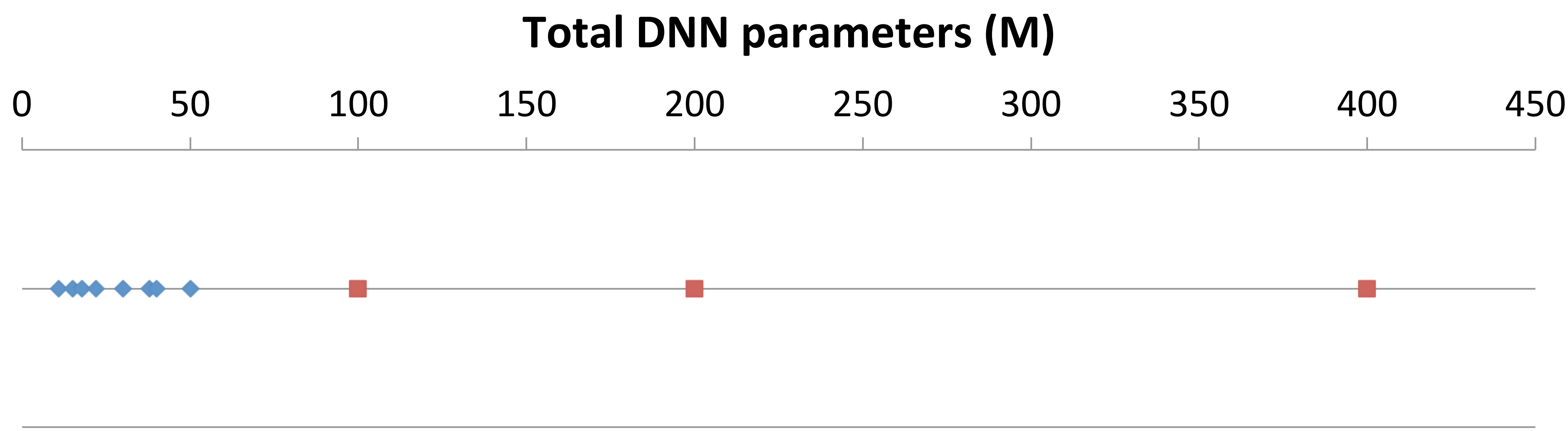
Hybrid NN. 1 hidden layer. 54 HMM states.

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

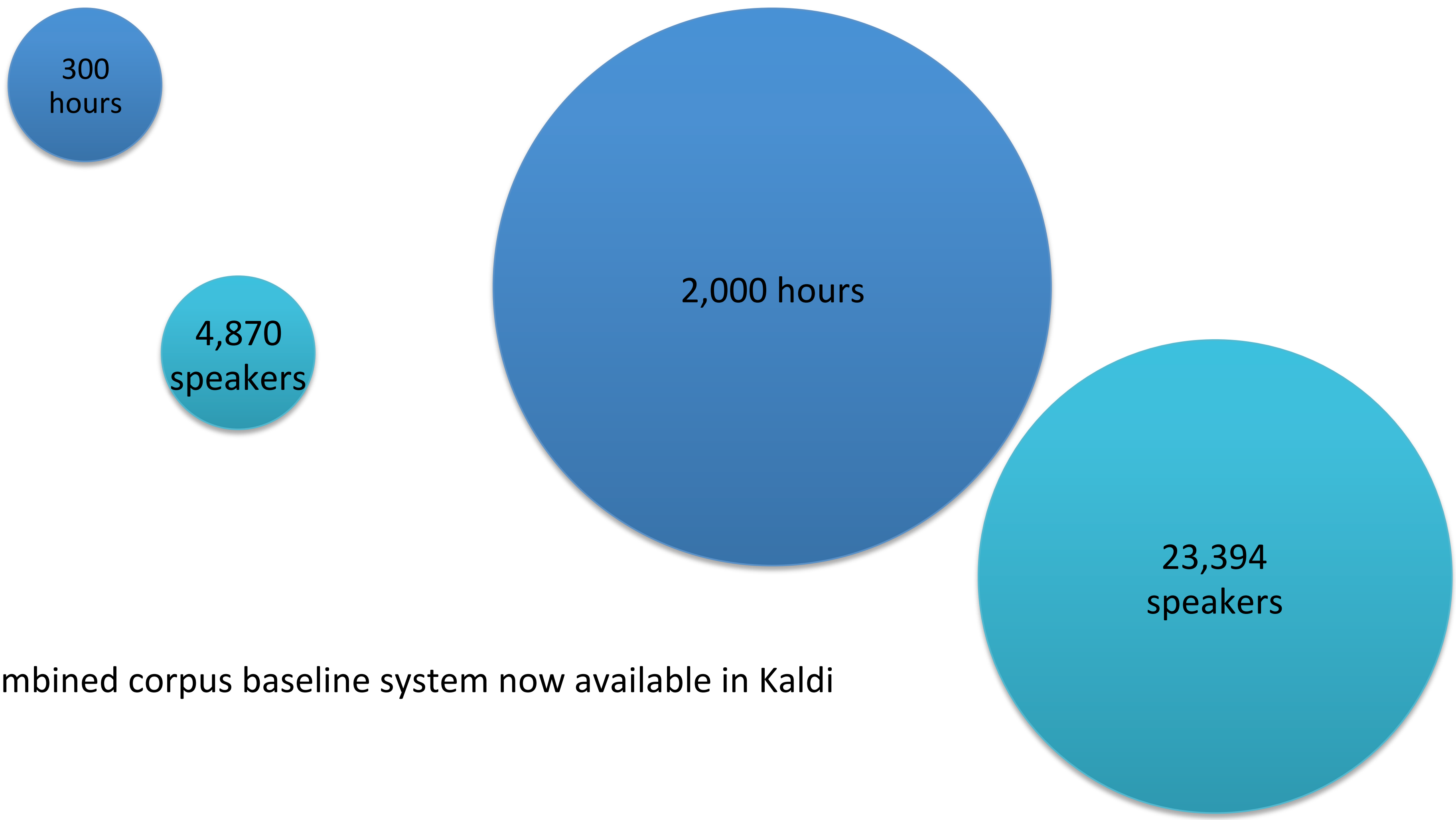
- Comparing total number of parameters (in millions) of previous work versus our new experiments



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

Combining Speech Corpora

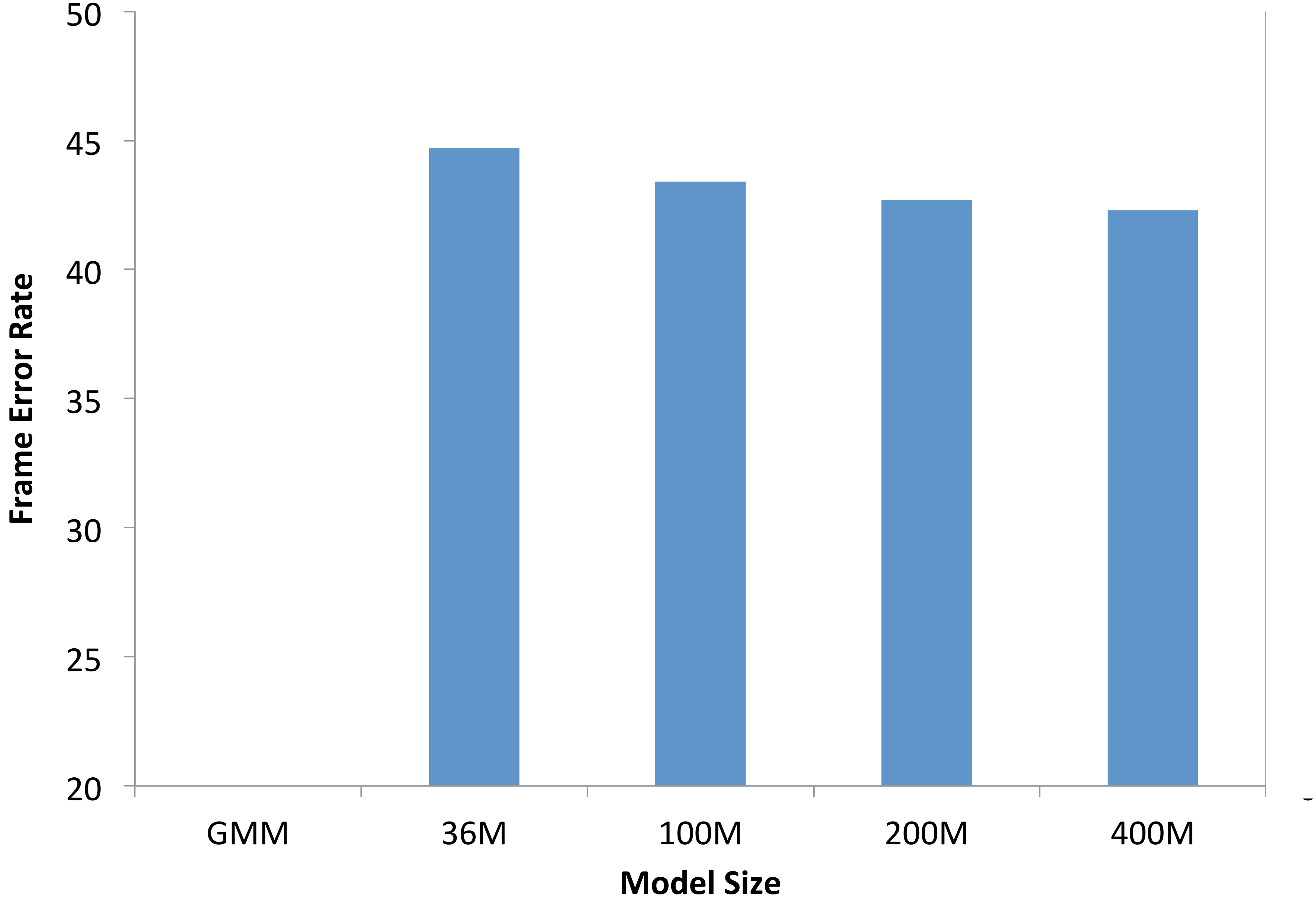
Switchboard **Fisher**



Combined corpus baseline system now available in Kaldi

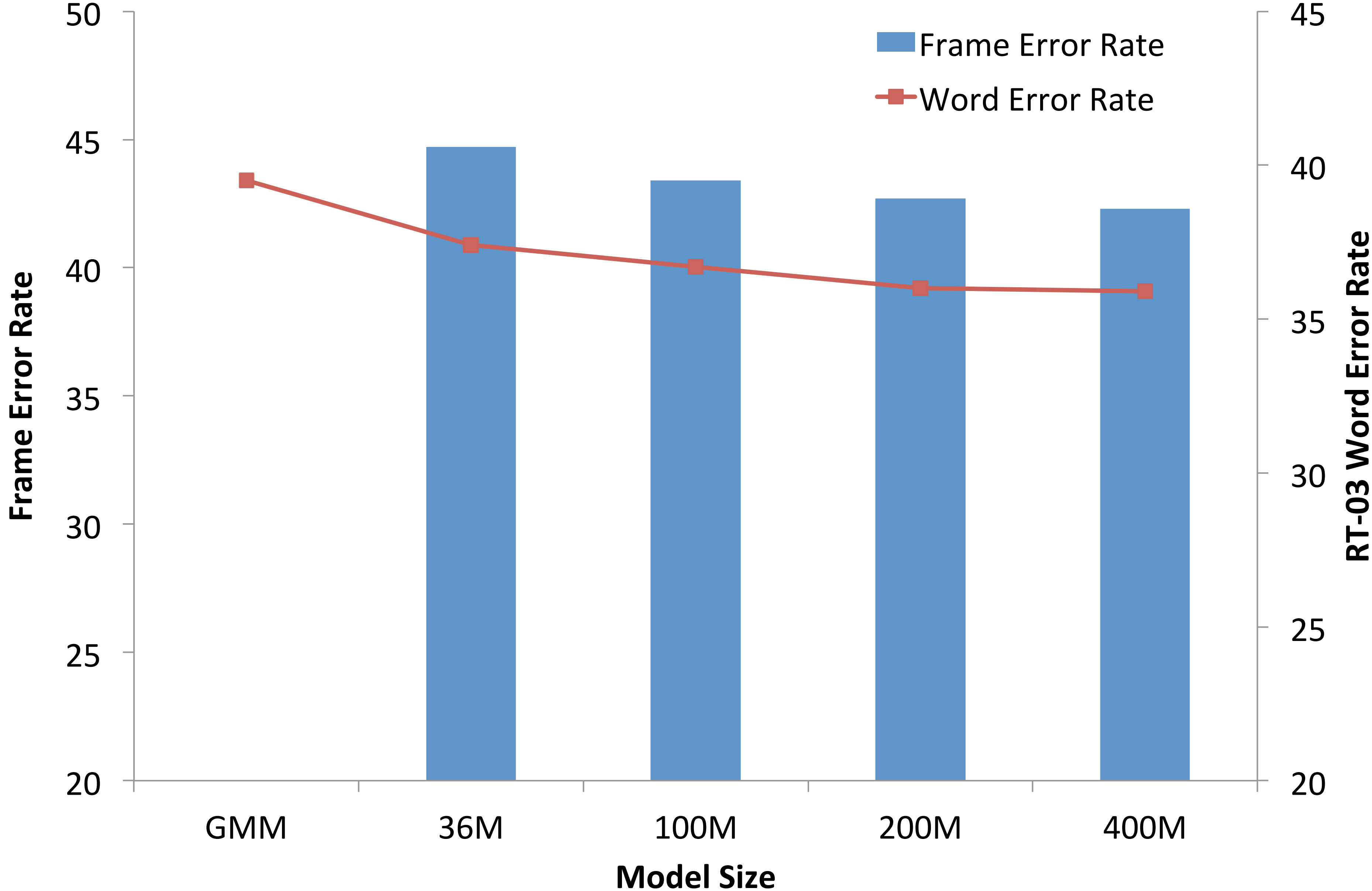
(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In Submission)

Scaling Total Parameters



(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In Submission) Andrew Maas. Stanford CS224D. 2016

Scaling Total Parameters



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

Outline

- Speech recognition systems overview
- HMM-DNN (Hybrid) acoustic modeling
- What's different about modern HMM-DNNs?
- **HMM-free RNN recognition**

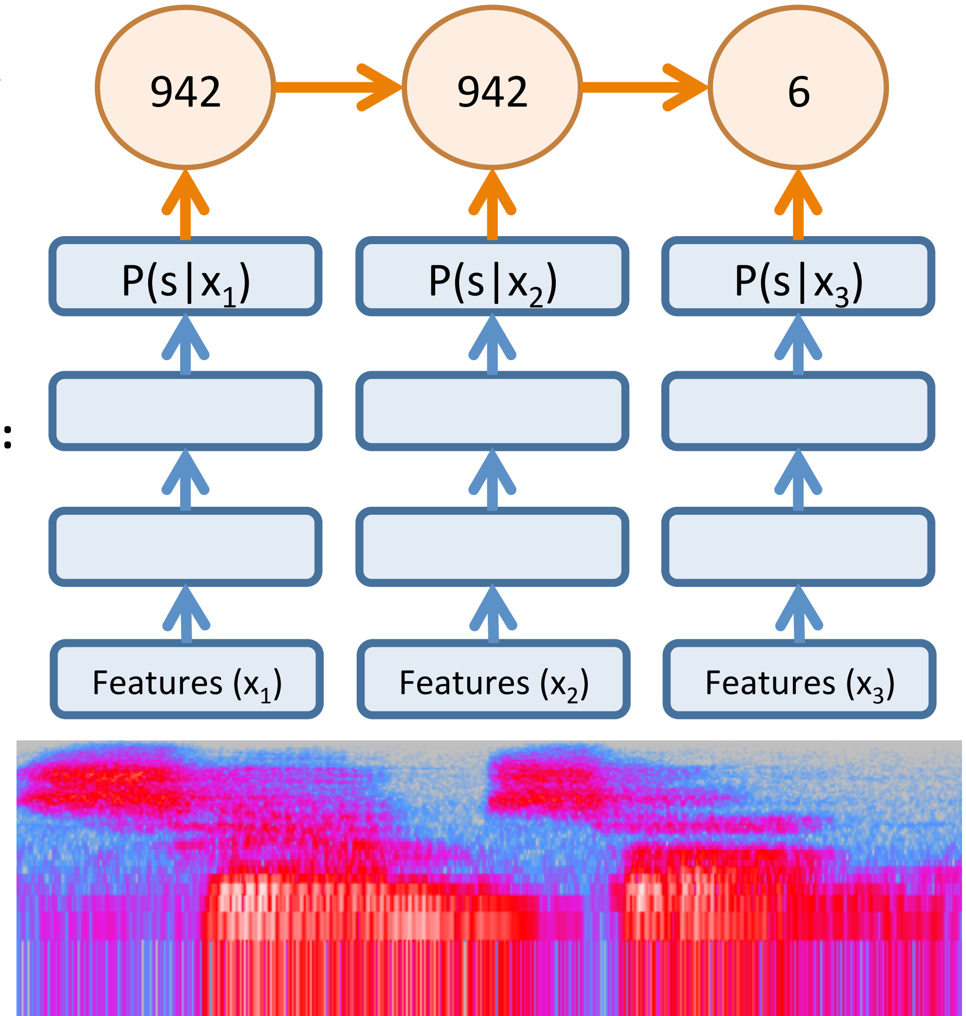
HMM-DNN Speech Recognition

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

Hidden Markov Model (HMM):

Acoustic Model:

Audio Input:



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

Andrew Maas. Stanford CS224D. 2016

HMM-Free Recognition

Transcription: Samson

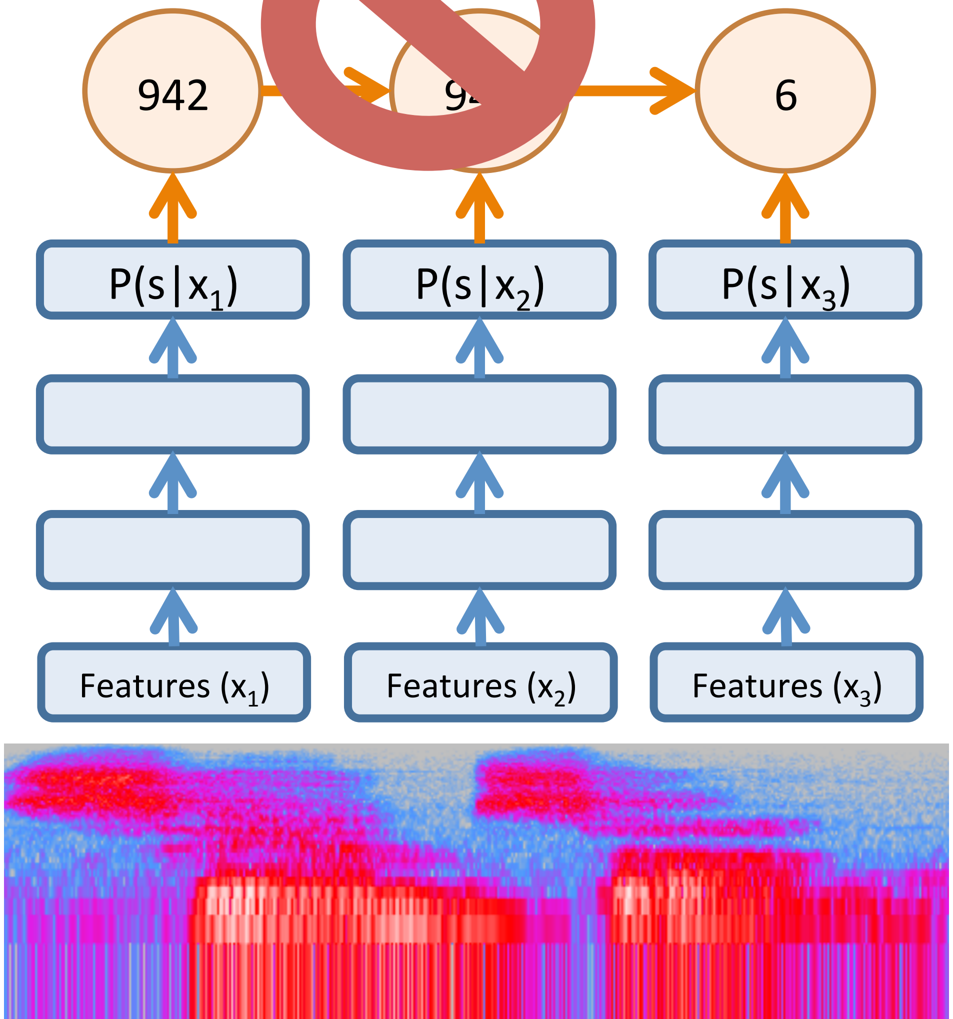
Pronunciation: S – AE – ... N

Sub-phones : 942 – ... 8006 ... 4422 …

Hidden Markov Model (HMM):

Acoustic Model:

Audio Input:



(Graves & Jaitly. 2014)

HMM-Free Recognition

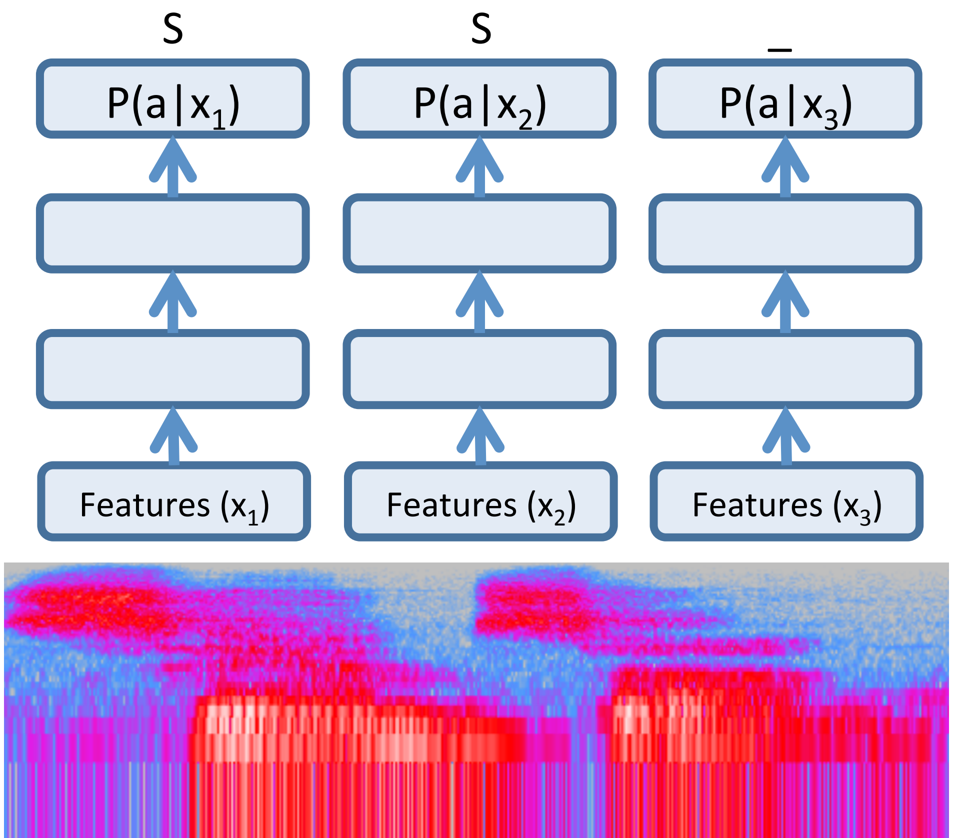
Transcription: Samson

Characters: SAMSON

Collapsing function: SS___AA_M_S___O___NNNN

Acoustic Model:

Audio Input:



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

The distribution over *characters*

(Graves & Jaitly. 2014)

CTC Objective Function

Labels at each time index are conditionally independent (like HMMs)

$$\Pr(\mathrm{a}|x) = \prod_{t=1}^{T} \Pr(a_t, t|x)$$

Sum over all time-level labelings consistent with the output label.

$$\Pr(y|x) = \sum_{a \in \mathcal{B}^{-1}(y)} \Pr(a|x)$$

Output label: AB

Time-level labelings: AB, _AB, A_B, … _A_B_

Final objective maximizes probability of true labels:

$$CTC(x) = -\log \Pr(y^*|x)$$

(Graves & Jaitly. 2014)

Collapsing Example

Per-frame argmax:

```
______________________________________________________________________________y
y__ee_________tt_ __________________________________________a_____
_rr__e________hh__________b___ii_______lll__i_____tt______aa______tt_______iio__n___
___cc_____rrr_u____________________ ________ii___ss
______________o__________nn_____________hhh_a___________________nnddd _________________i__n___
__thh_e_____ ________________________________________bb_uuii_______llllld____ii____nng_____
_________________________________l___o___o_g__g___ii____nng______
____b___rr_ii________ck__s________________________________________p___ll__a________sstt_________eerr__
______a___nnd_ ___b___lll_uu____ee__pp___r___i________nnss_
________________f______oou____________rrr________ ______________f_____oo__rrr__tt_y____
_____t____www_oo__________ ____nn___ew___________________
_____________________________________________________b___e_______t__________i____n___
____e________pp_____aa___rr___tt____mm_ee___nnntss
_______________________________________________________________________________________________
______________________________
```

After collapsing:

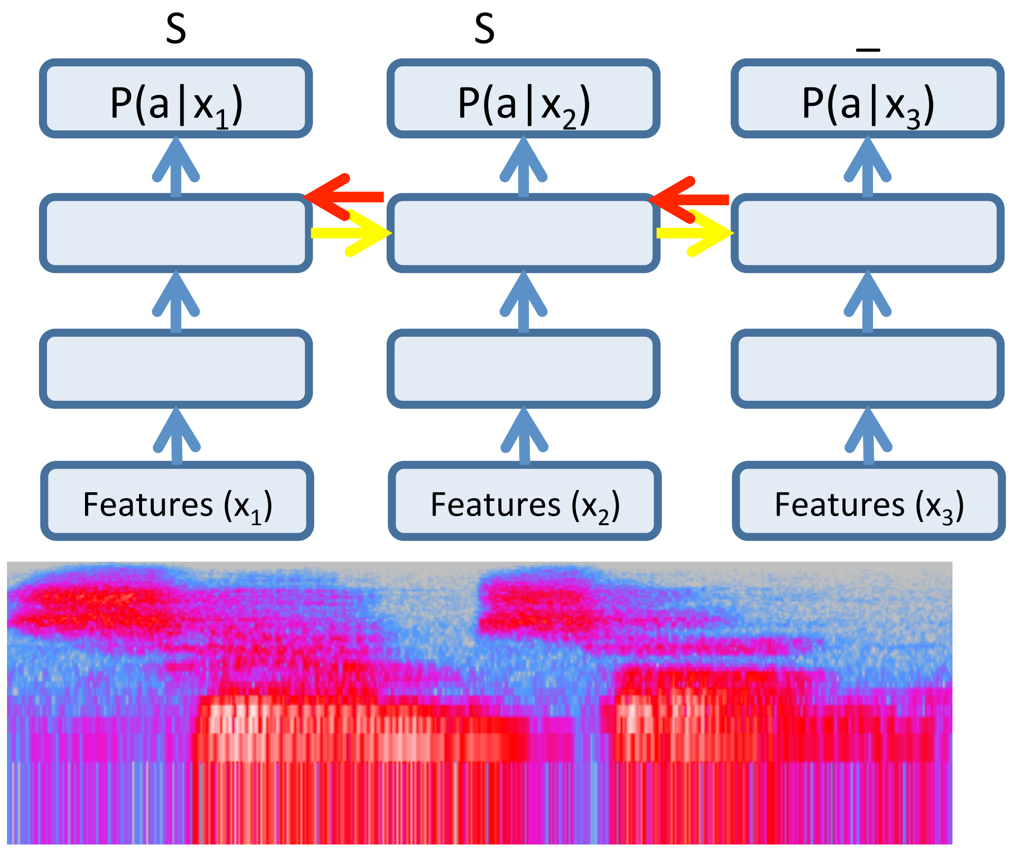
yet a rehbilitation cru is onhand in the building loogging bricks plaster and blueprins four forty two new betin epartments

Reference:

yet a rehabilitation crew is on hand in the building lugging bricks plaster and blueprints for forty two new bedroom apartments

(Hannun, Maas, Jurafsky, & Ng. 2014)

Recurrence Matters!



Architecture	CER
DNN	22

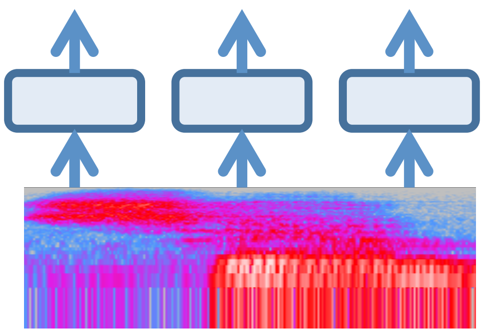
(Hannun, Maas, Jurafsky, & Ng. 2014)

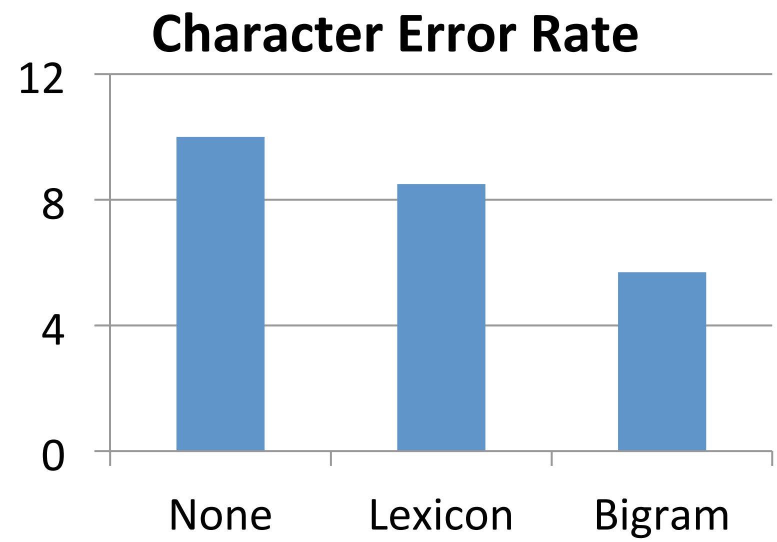
Decoding with a Language Model

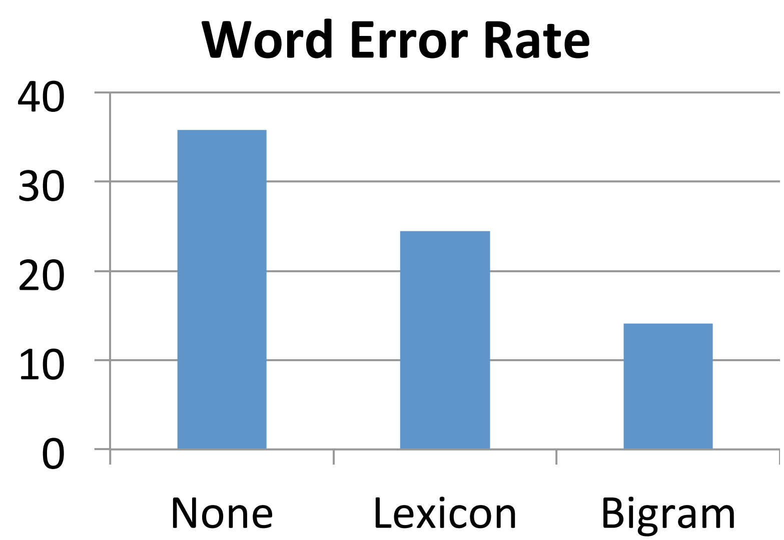
Lexicon [a, …, zebra]

Language Model p("yeah" | "oh")

Character Probabilities __oo_h__y_e_aa_h







(Hannun, Maas, Jurafsky, & Ng. 2014)

Rethinking Decoding

Out of Vocabulary Words

syriza

bae

abo--

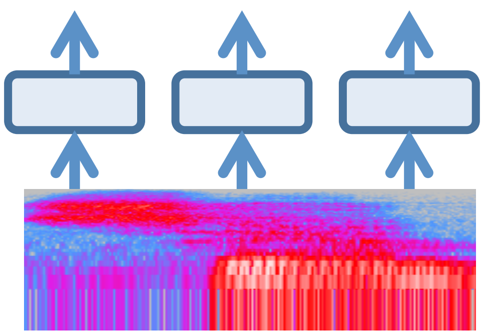
sof--

schmidhuber

Le   [zebra]

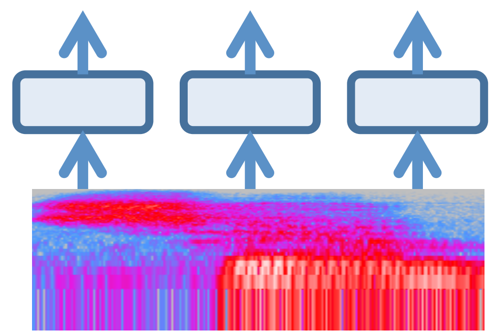
L   uage   (“v   ” | “oh”)
M

Character Probabilities   __oo_h__y_e_aa_h



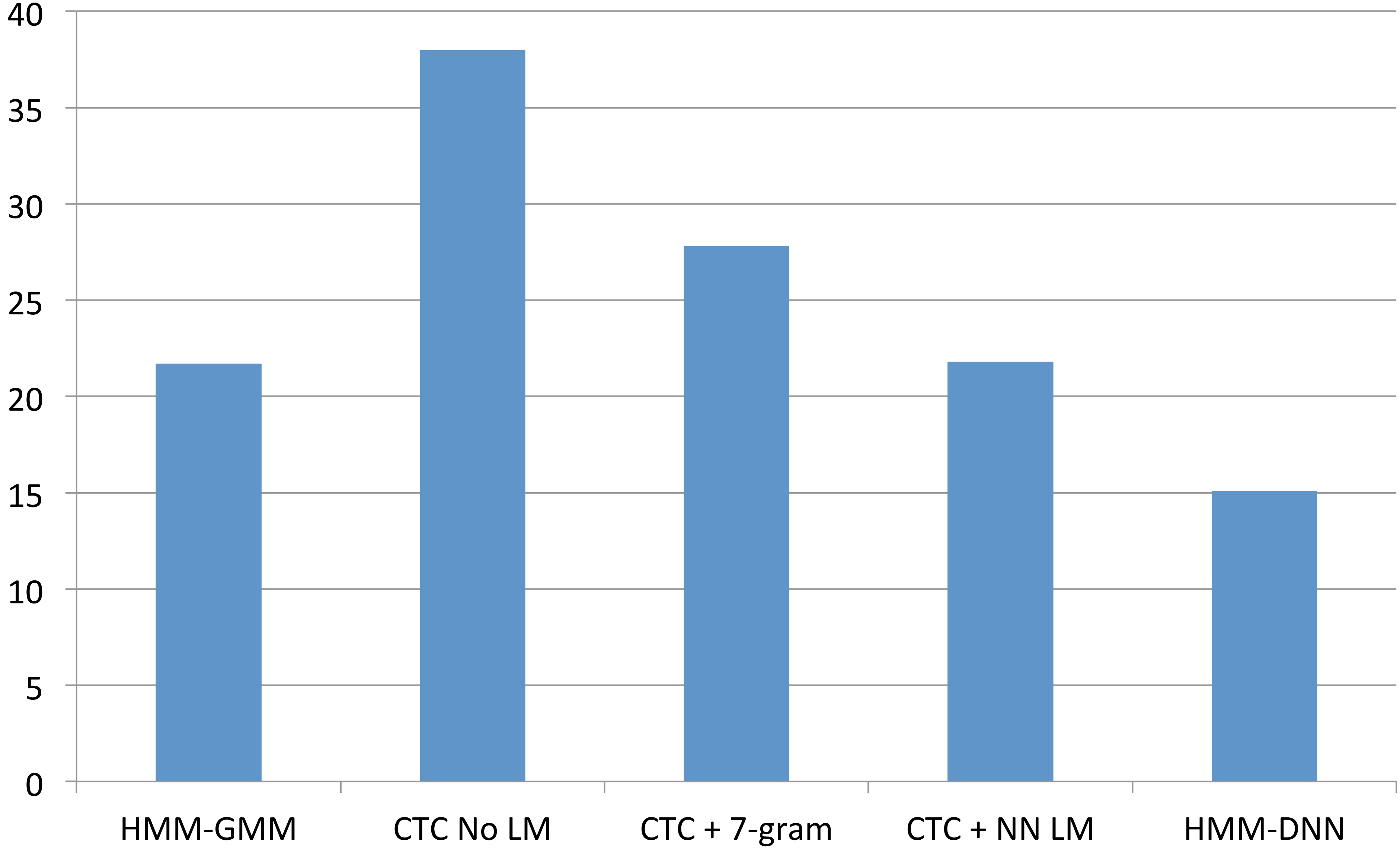
***Character* Language Model**   p(h | o,h, ,y,e,a,)

Character Probabilities   __oo_h__y_e_aa_h



(Maas*, Xie*, Jurafsky, & Ng. 2015)

Lexicon-Free & HMM-Free on Switchboard



(Maas*, Xie*, Jurafsky, & Ng. 2015)

Transcribing Out of Vocabulary Words

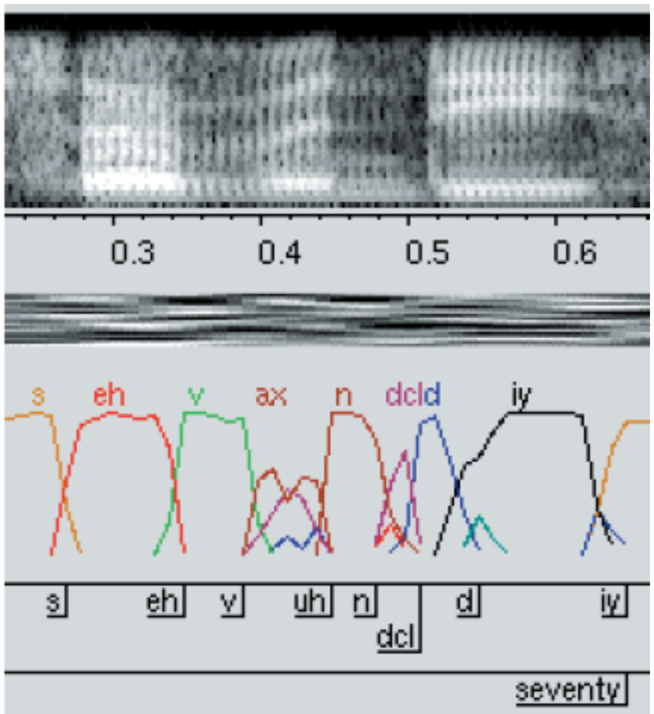
Truth: yeah i went into the i do not know what you think of *fidelity* but
HMM-GMM: yeah when the i don't know what you think of **fidel it even them**
CTC-CLM: yeah i went to i don't know what you think of **fidelity but um**

Truth: no no speaking of weather do you carry a altimeter slash *barometer*
HMM-GMM: no i'm not all being the weather do you uh carry a **uh helped emitters last brahms her**
CTC-CLM: no no beating of whether do you uh carry a **uh a time or less barometer**

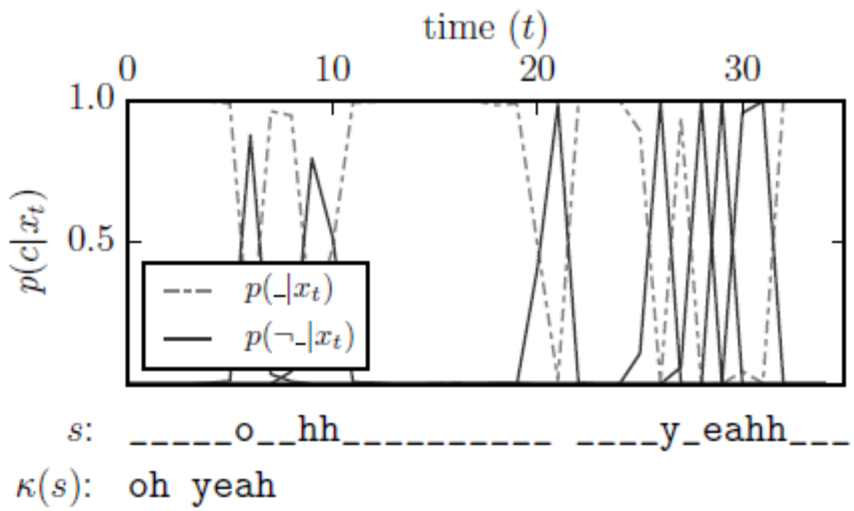
Truth: i would ima- well yeah it is i know you are able to stay home with them
HMM-GMM: i would **amount** well yeah it is i know um you're able to stay home with them
CTC-CLM: i would **ima-** well yeah it is i know uh you're able to stay home with them

(Maas*, Xie*, Jurafsky, & Ng. 2015)

Comparing Alignments



HMM-GMM phone probabilities

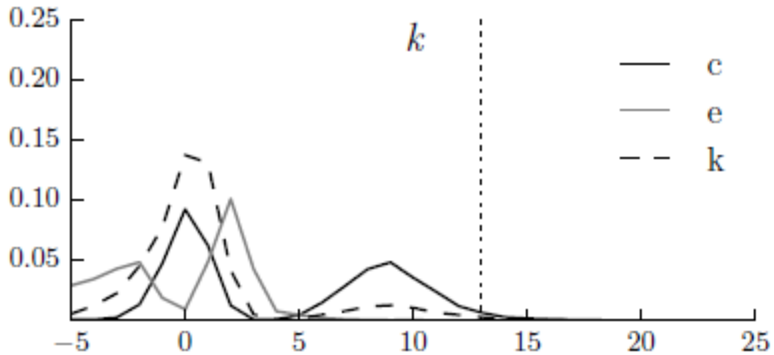


CTC character probabilities

(HMM slide from Dan Ellis)

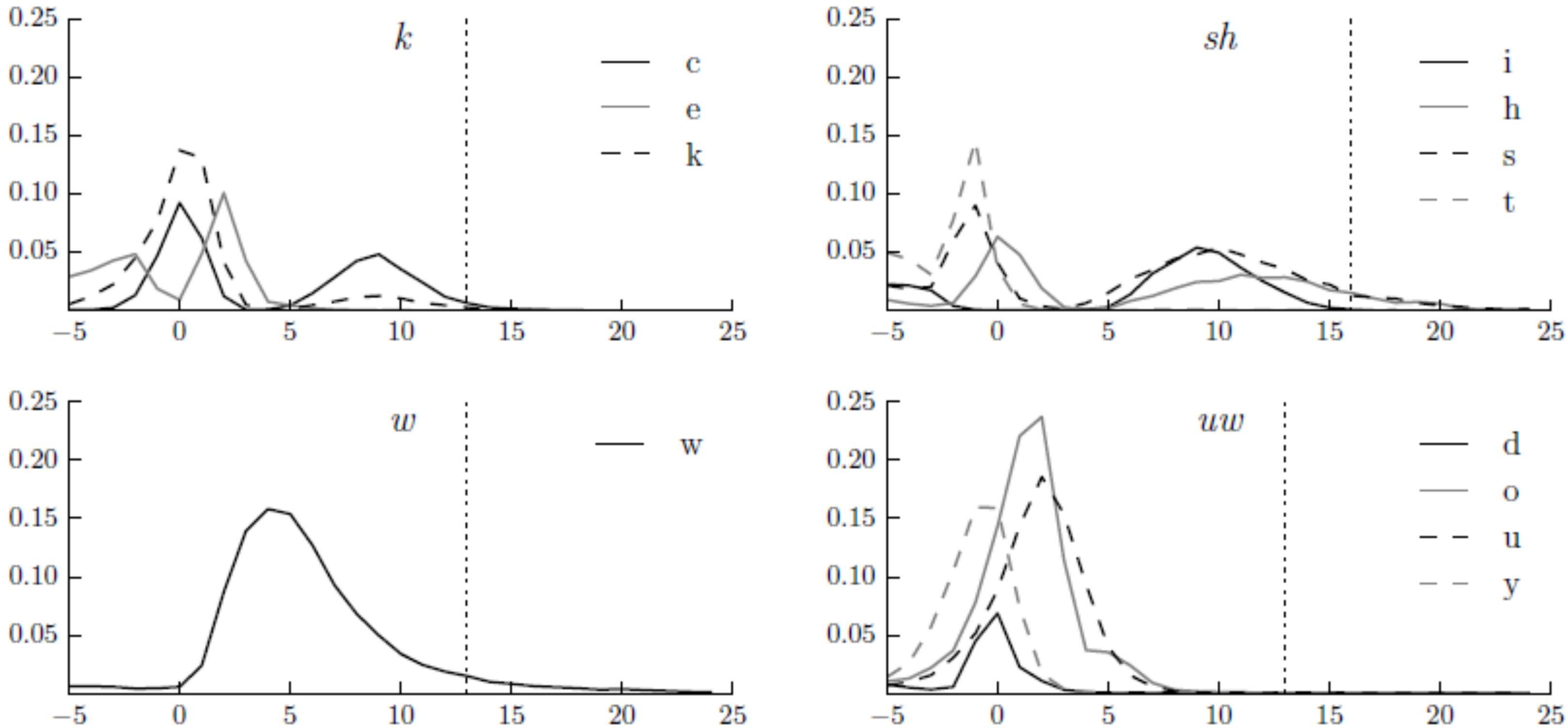
Andrew Maas. Stanford CS224D. 2016

Learning Phonemes and Timing



(Maas*, Xie*, Jurafsky, & Ng. 2015)

Learning Phonemes and Timing

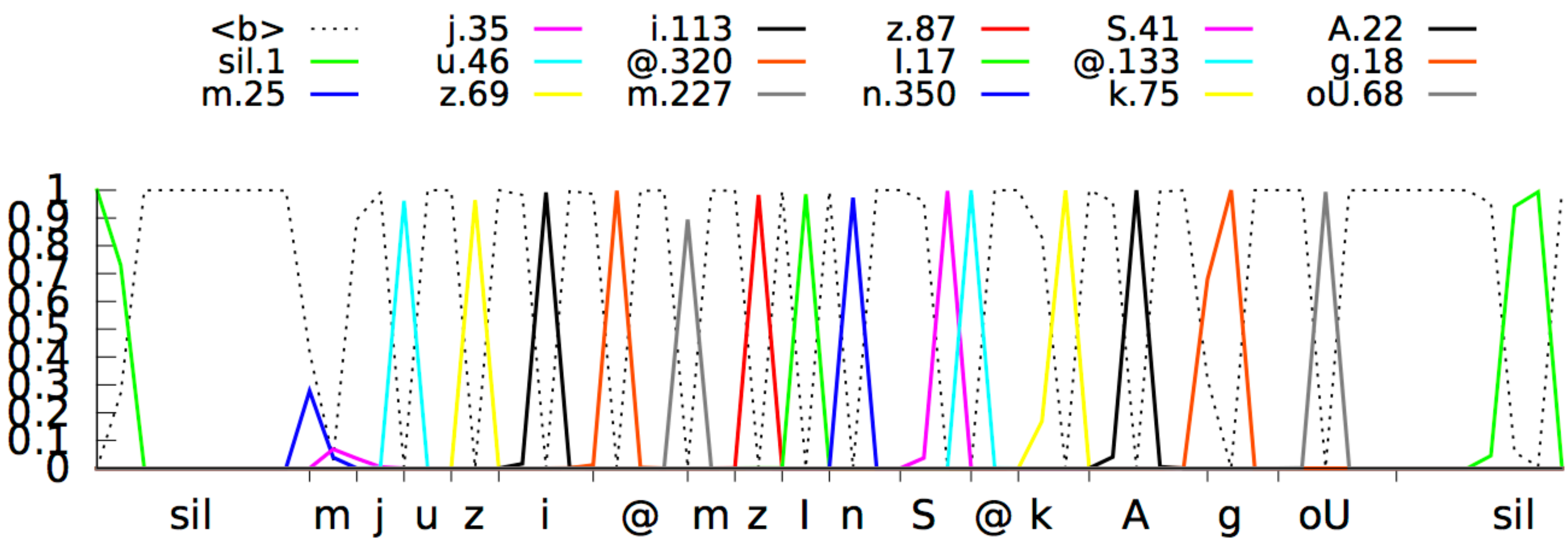


(Maas*, Xie*, Jurafsky, & Ng. 2015)

Pushing Performance with HMM-Free

CTC now powers Google search ASR

- Context-dependent states rather than characters
- Uni-directional LSTM for faster streaming
- CTC + sequence discriminative loss

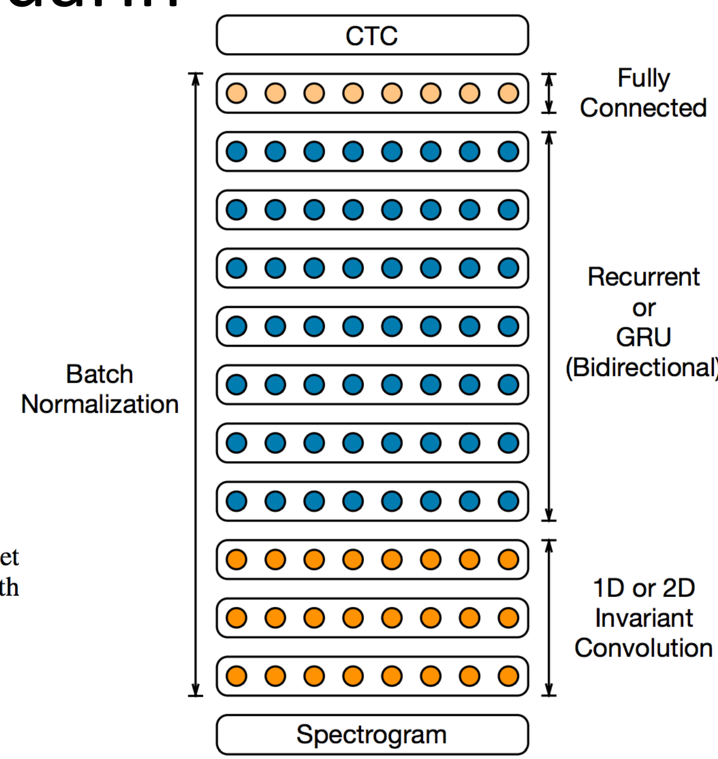


http://googleresearch.blogspot.com/2015/09/google-voice-search-faster-and-more.html

(Sak, Senior, Rao, & Beaufays. 2015)

Deep Speech 2: Scaling up CTC

- Efficient GPU training
- Some recurrent architecture variants
- Data augmentation
- Works on both English and Mandarin



Fraction of Data	Hours	Regular Dev	Noisy Dev
1%	120	29.23	50.97
10%	1200	13.80	22.99
20%	2400	11.65	20.41
50%	6000	9.51	15.90
100%	12000	8.46	13.59

Table 10: Comparison of English WER for Regular and Noisy development sets on increasing training dataset size. The architecture is a 9-layer model with 2 layers of 2D-invariant convolution and 7 recurrent layers with 68M parameters.

(Amodei *et al*. 2015)

Listen, attend, and spell

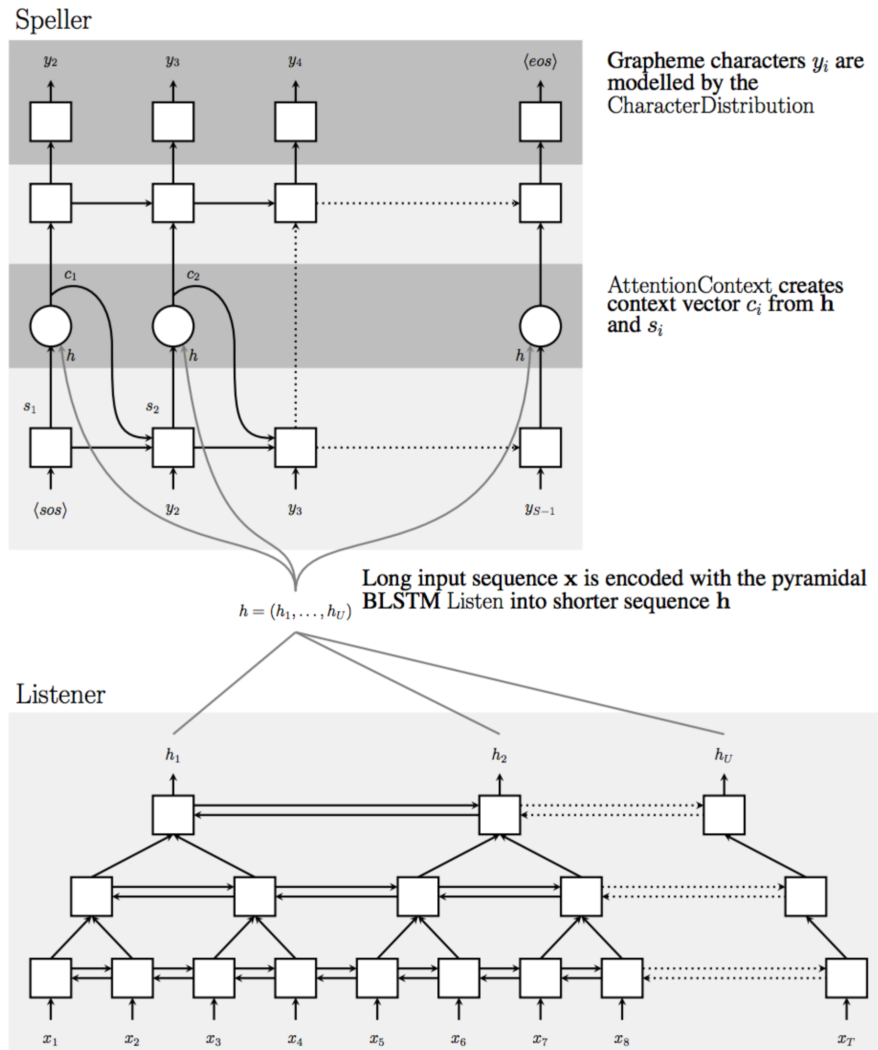


Figure 1: Listen, Attend and Spell (LAS) model: the listener is a pyramidal BLSTM encoding our input sequence $\mathbf{x}$ into high level features $\mathbf{h}$, the speller is an attention-based decoder generating the $\mathbf{y}$ characters from $\mathbf{h}$.

(Chan, Jaitly, Le, & Vinyals. 2015)

Listen, attend, and spell

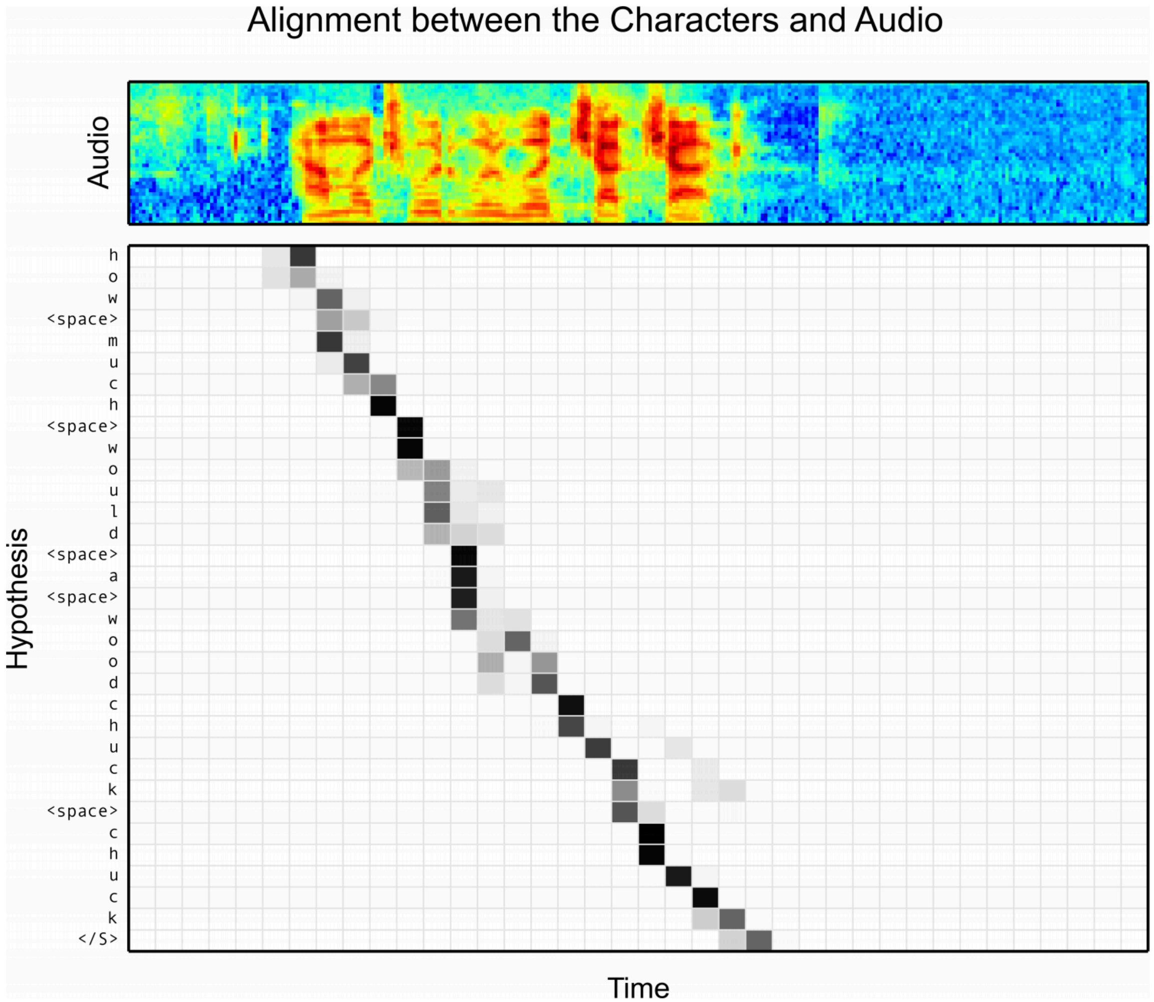


Figure 2: Alignments between character outputs and audio signal produced by the Listen, Attend and Spell (LAS) model for the utterance "how much would a woodchuck chuck". The content based attention mechanism was able to identify the start position in the audio sequence for the first character correctly. The alignment produced is generally monotonic without a need for any location based priors.

(Chan, Jaitly, Le, & Vinyals. 2015)

Conclusion

- HMM-DNN systems are now the default, state-of-the-art for speech recognition
- We roughly understand why HMM-DNNs work but older, shallow hybrid models didn't work as well
- HMM-Free approaches are rapidly improving and making their way to production systems
- *It's a very exciting time for speech recognition*

End

- More on spoken language understanding:
  - cs224s.stanford.edu
- Open source speech recognition toolkit (Kaldi):
  - Kaldi.sf.net
- Multiple open source implementations of CTC available