CS224D: Deep Learning for Natural Language Processing

Andrew Maas
Stanford University
Spring 2015

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

Cat

Clothes

Climbing

Deep Neural Network

What action?

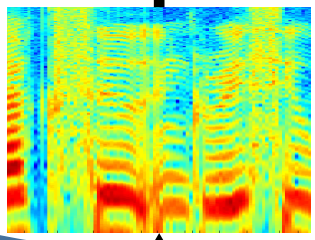
Is the user annoyed?

Ask for clarification?

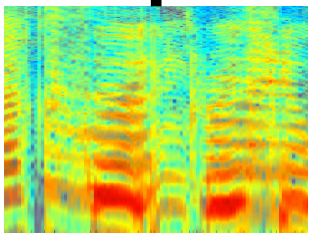
Understanding

Are there any good robot movies I can rent tonight?

Transcription



Noise Reduction



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

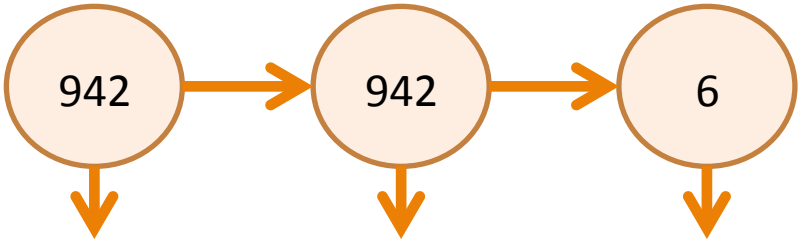
Acoustic Modeling with GMMs

Transcription: Samson

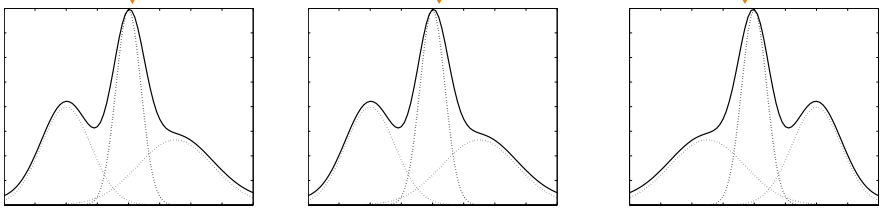
Pronunciation: S – AE – M – S –AH – N

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

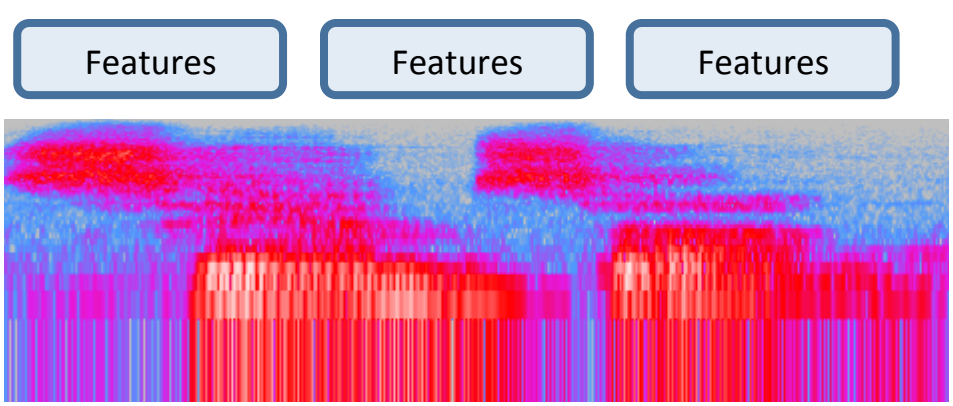
Hidden Markov Model (HMM):



Acoustic Model:



Audio Input:



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

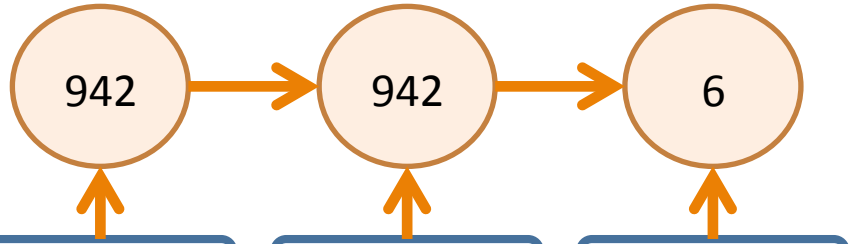
DNN Hybrid Acoustic Models

Transcription: Samson

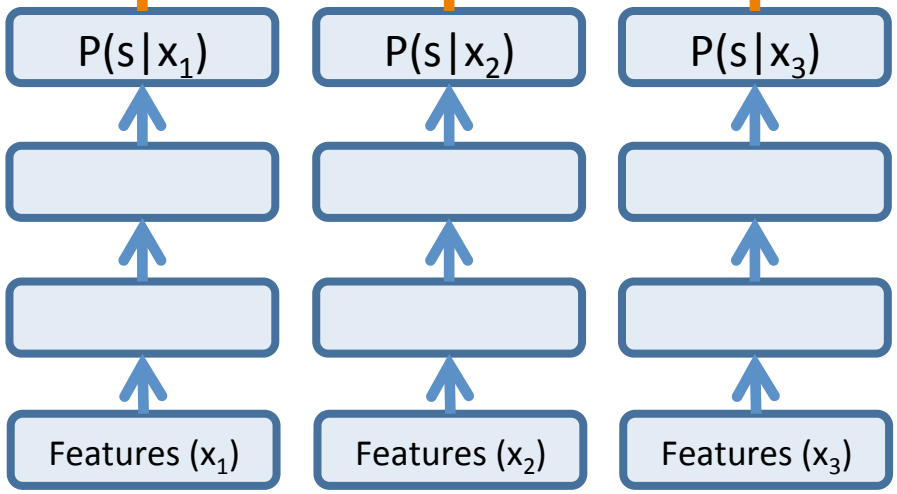
Pronunciation: S – AE – M – S –AH – N

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

Hidden Markov Model (HMM):



Acoustic Model:

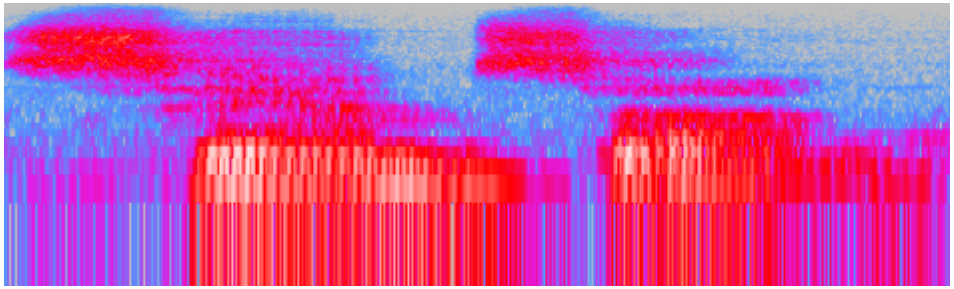


Use a DNN to approximate: $P(s|x)$

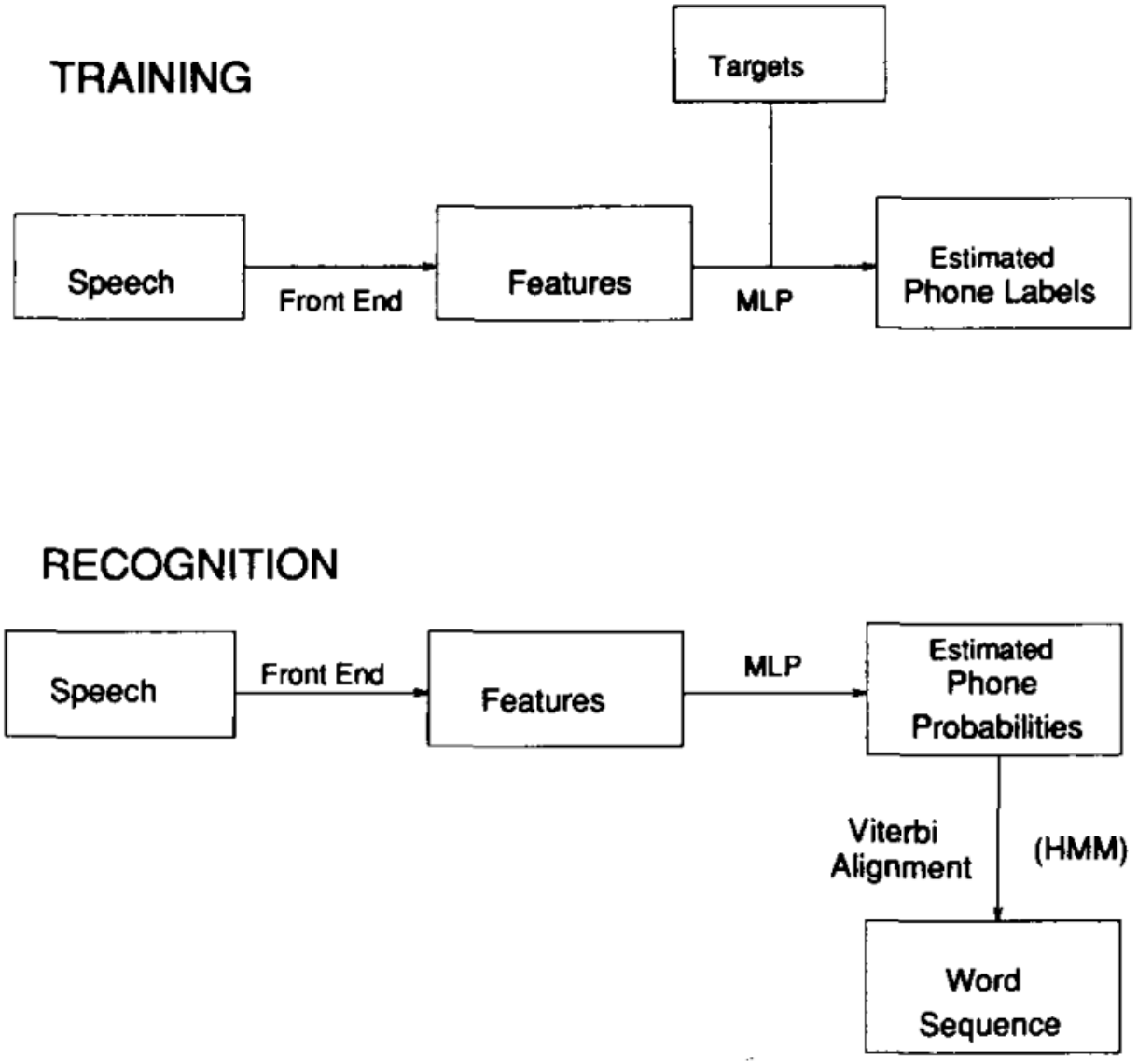
Apply Bayes' Rule: $P(x|s) = P(s|x) * P(x) / P(s)$

DNN * Constant / State prior

Audio Input:



Not Really a New Idea



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

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.

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 *does not matter*. Initially we thought this was the new trick that made things work
- Many more parameters

Depth Matters (Somewhat)

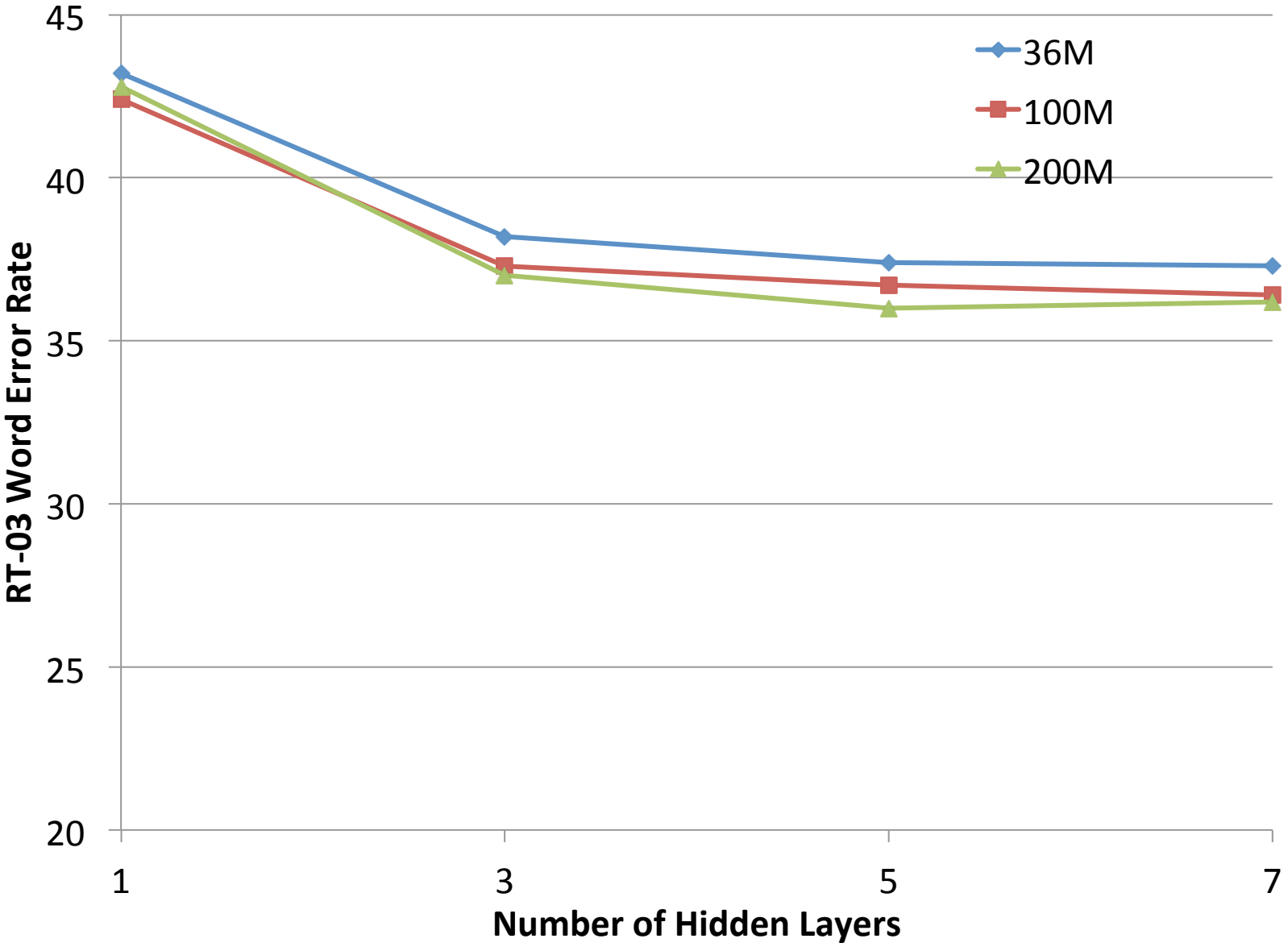
Table 1: Effect of CD-DNN-HMM network depth on WER (%) on Hub5'00-SWB using the 309-hour Switchboard training set. DBN pretraining is applied.

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

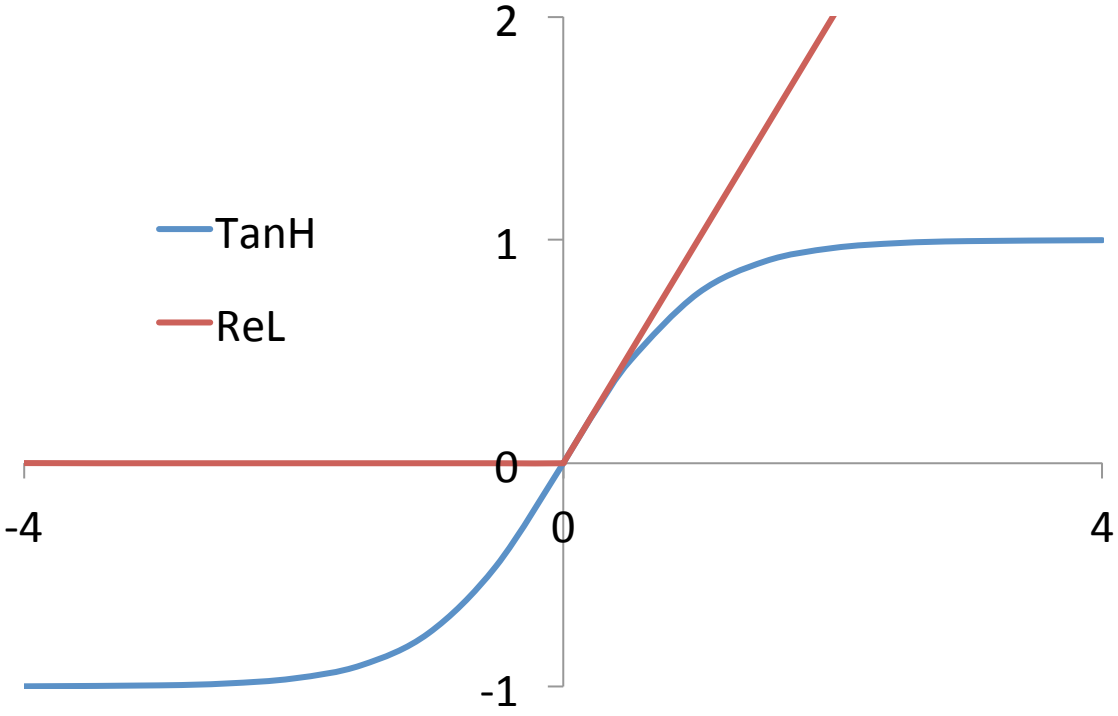
Yu, Seltzer, Li, Huang, Seide. 2013.

Impact of Depth



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

Replacing Sigmoid Hidden Units

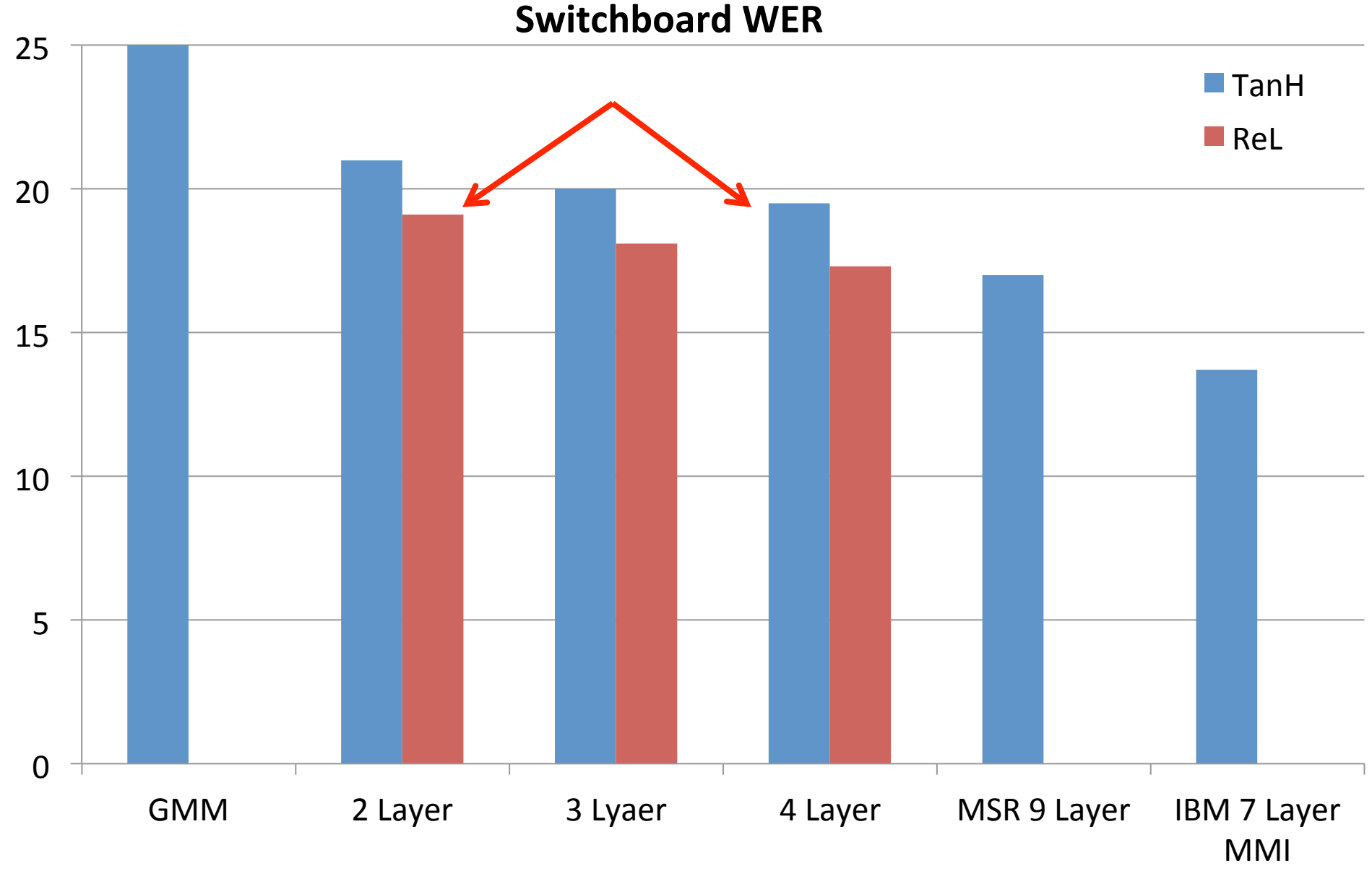


Rectified Linear (ReL)

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

(Glorot & Bengio. 2011)

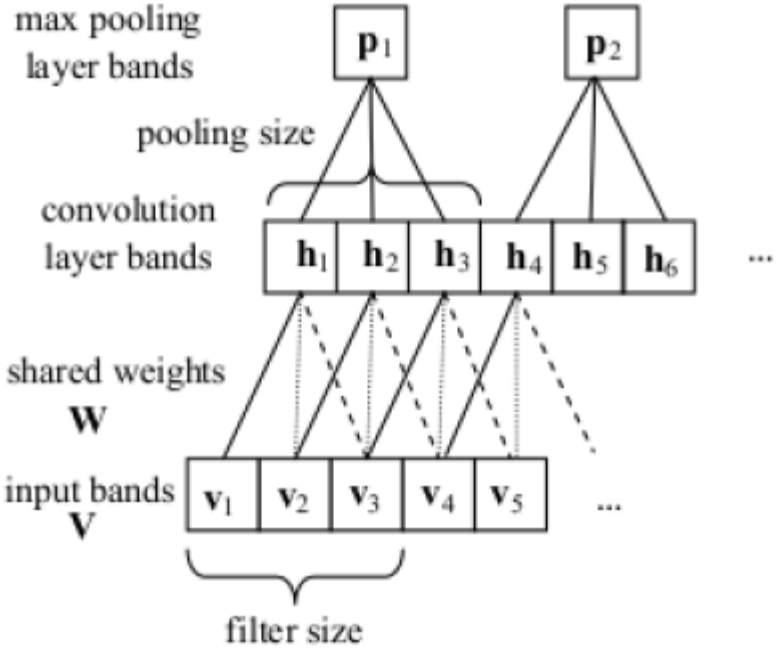
Comparing Nonlinearities

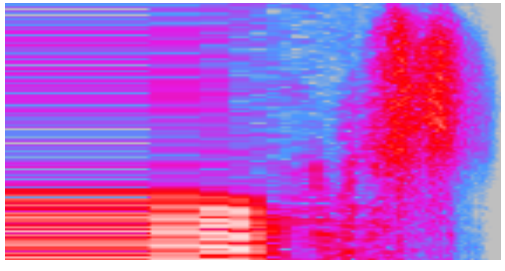


(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In Submission)

Convolutional Networks

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





Sainath, Mohamed, Kingsbury, & Ramabhadran . 2013.

Recurrent DNN Hybrid Acoustic Models

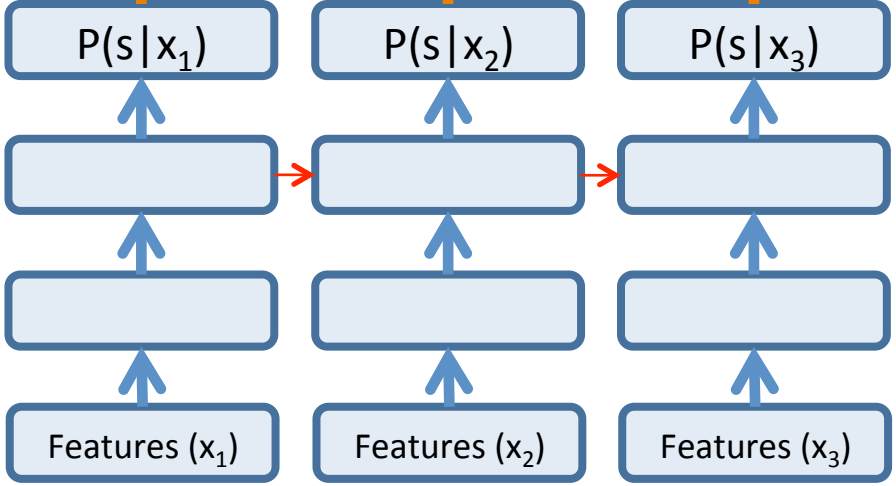
Transcription: Samson

Pronunciation: S – AE – M – S –AH – N

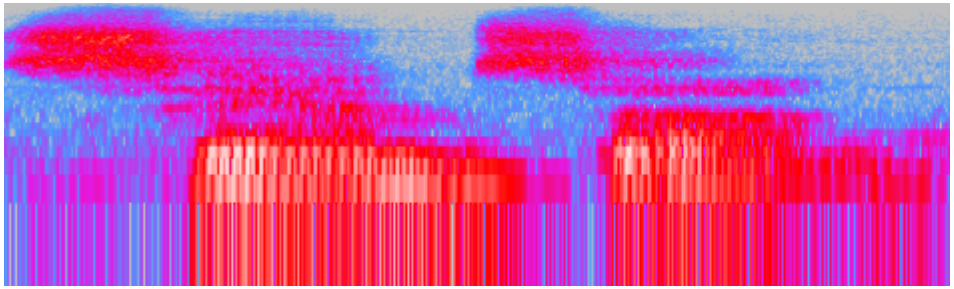
Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

Hidden Markov Model (HMM):

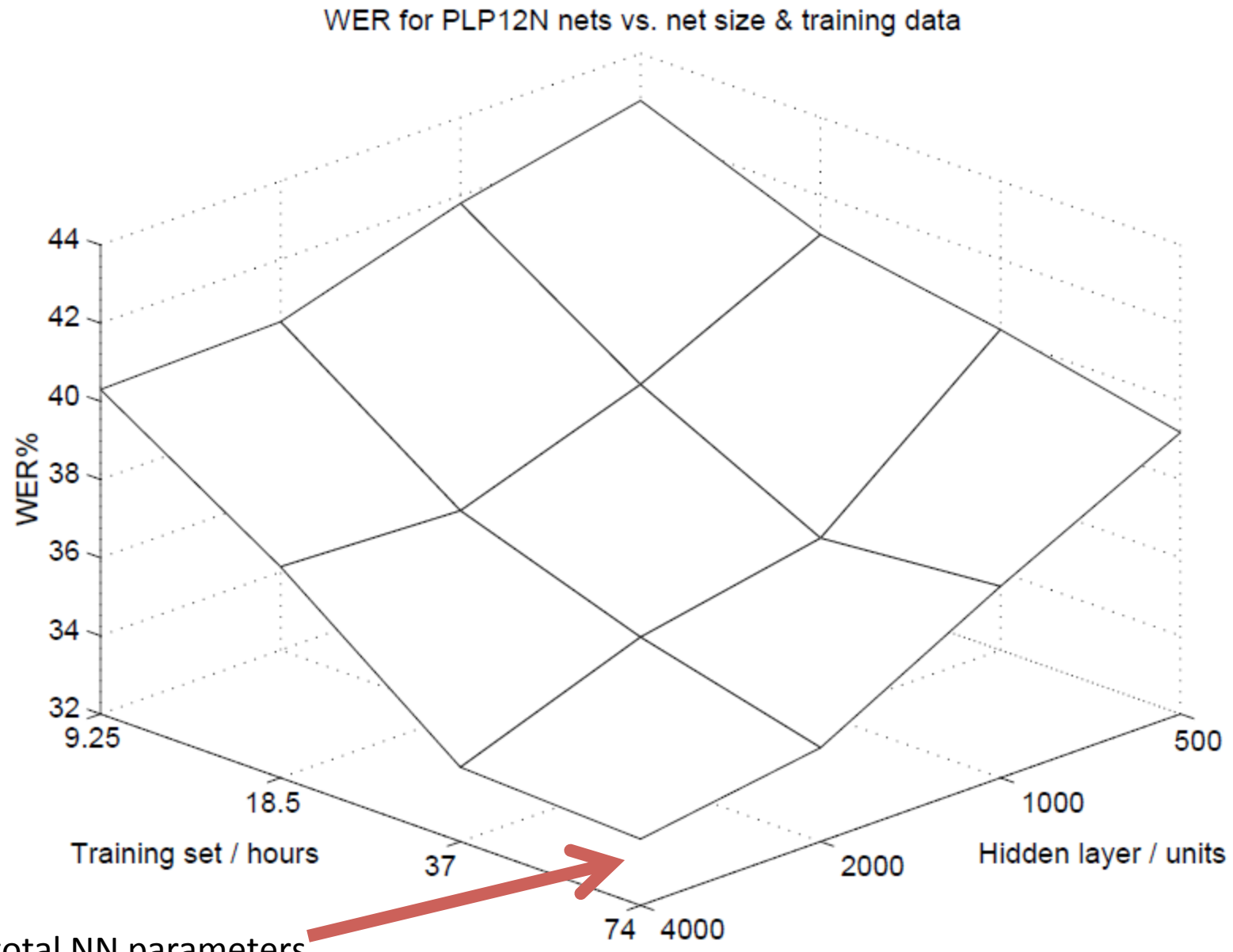
Acoustic Model:



Audio Input:



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.

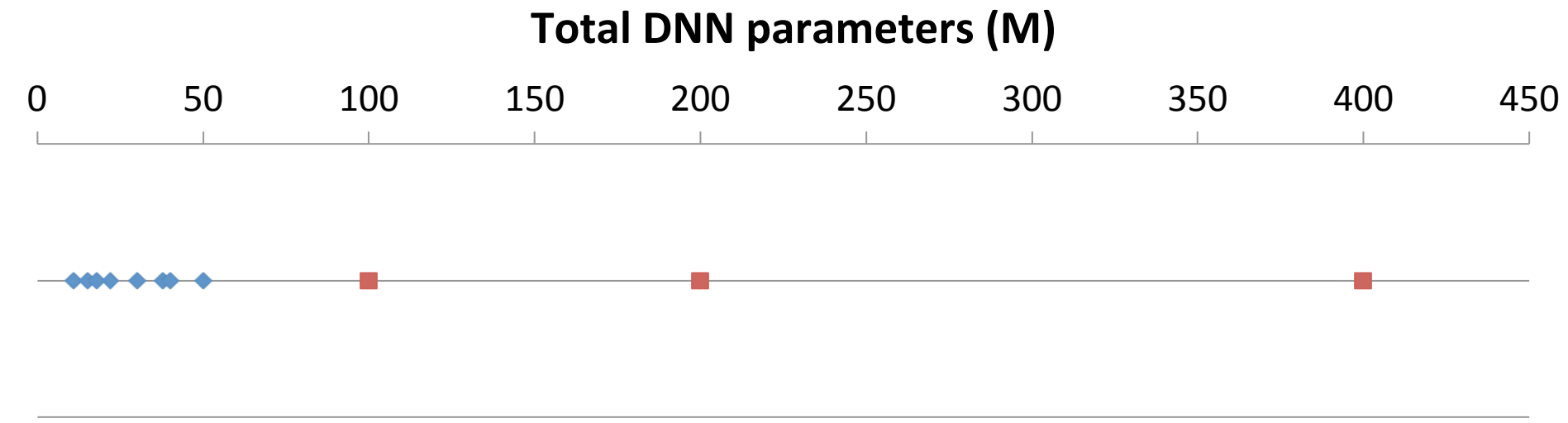
Hybrid NN. 1 hidden layer. 54 HMM states.

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."

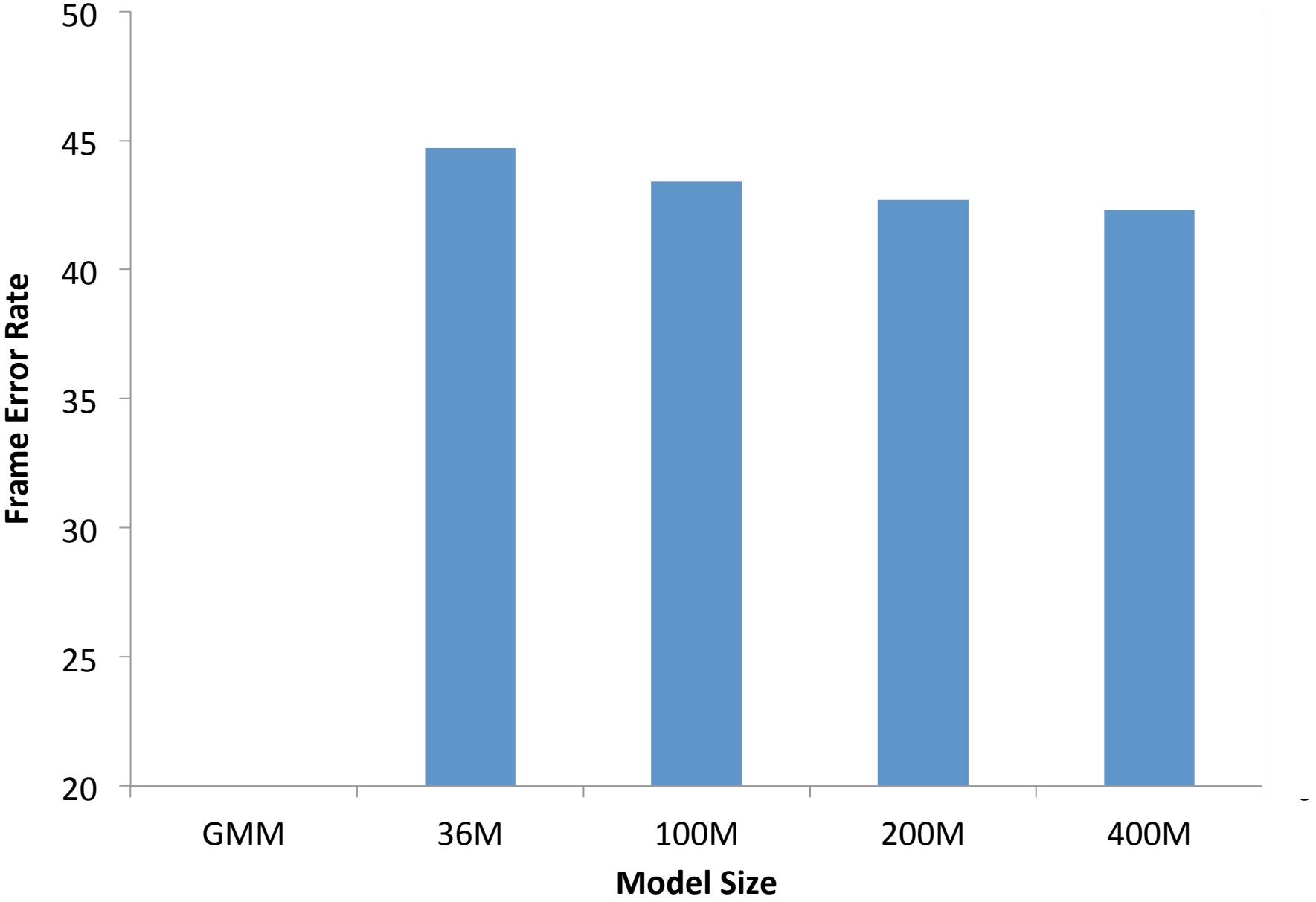
Adding More Parameters Now

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



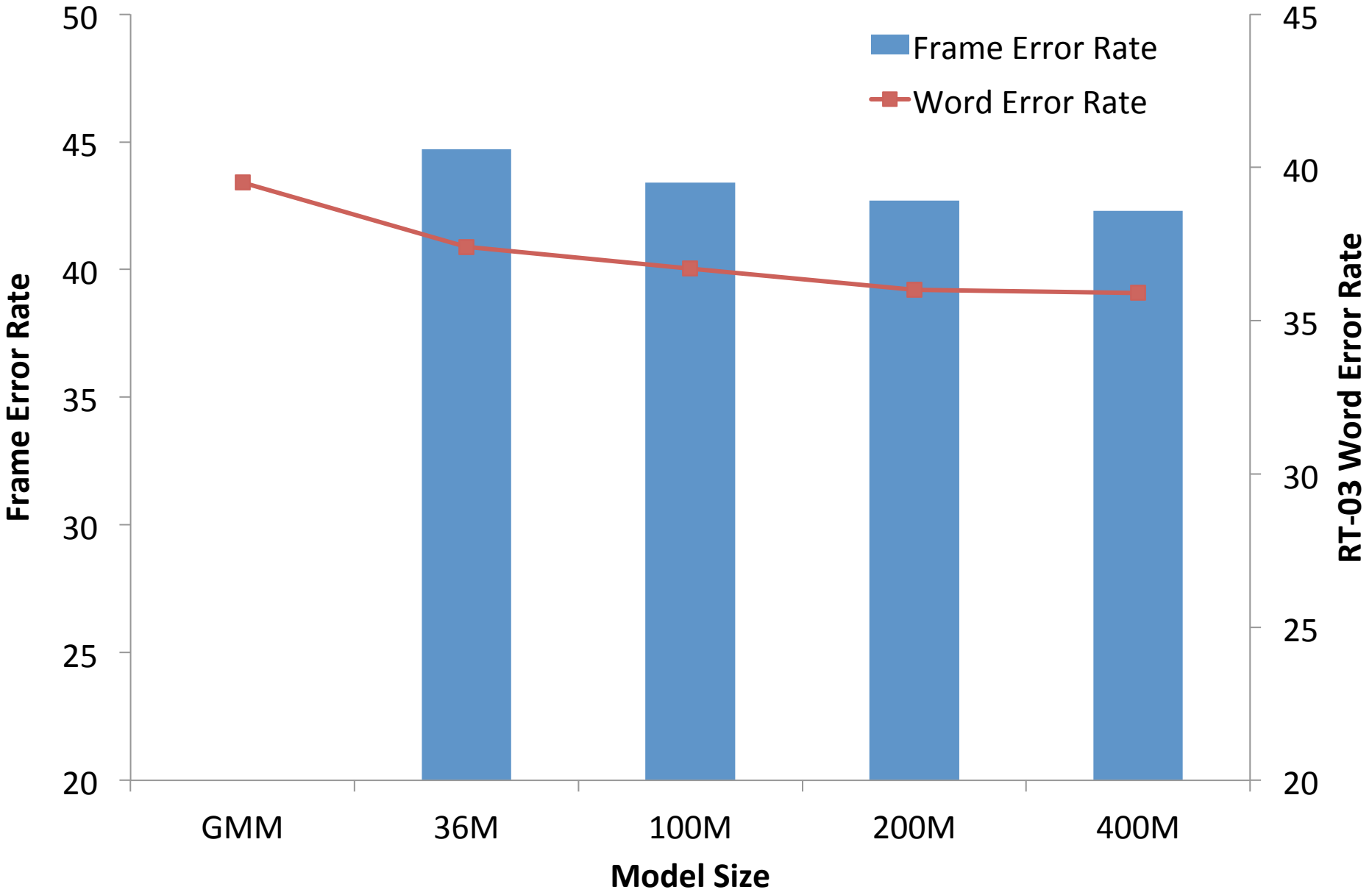
Maas, Hannun, Qi, Lengerich, Ng, & Jurafsky. In submission.

Scaling Total Parameters



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

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

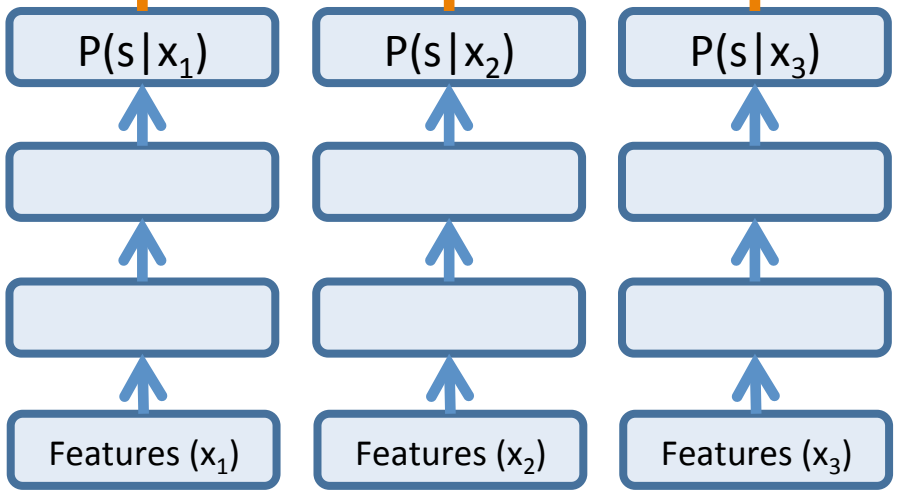
Transcription: Samson

Pronunciation: S – AE – M – S –AH – N

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

Hidden Markov Model (HMM):

Acoustic Model:

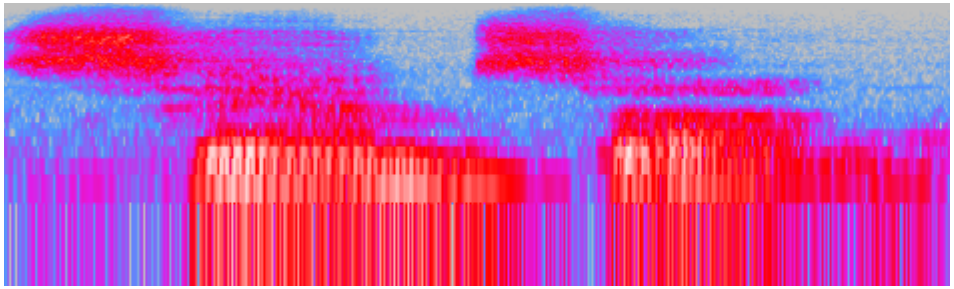


Use a DNN to approximate: $P(s|x)$

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

DNN * Constant / State prior

Audio Input:



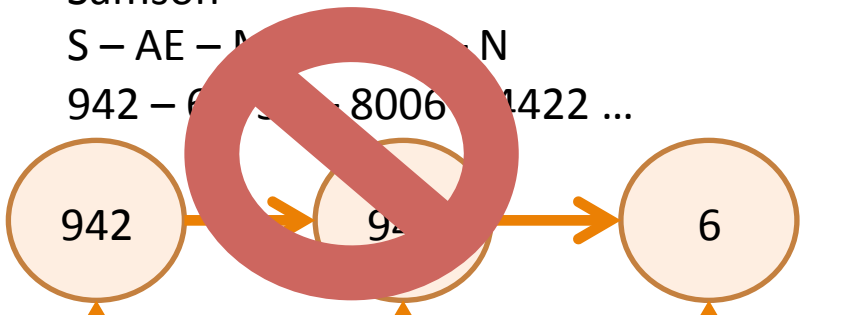
HMM-Free Recognition

Transcription: Samson

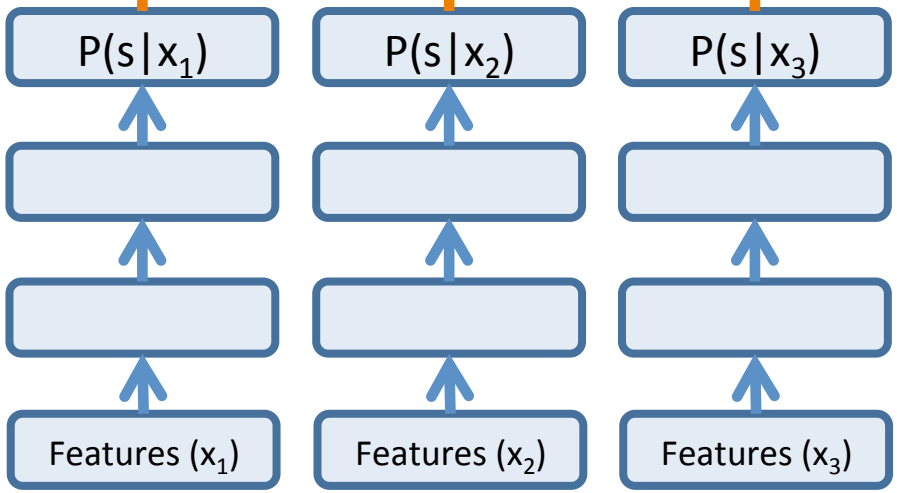
Pronunciation: S – AE – M – N

Sub-phones : 942 – 6 – 8006 – 1422 …

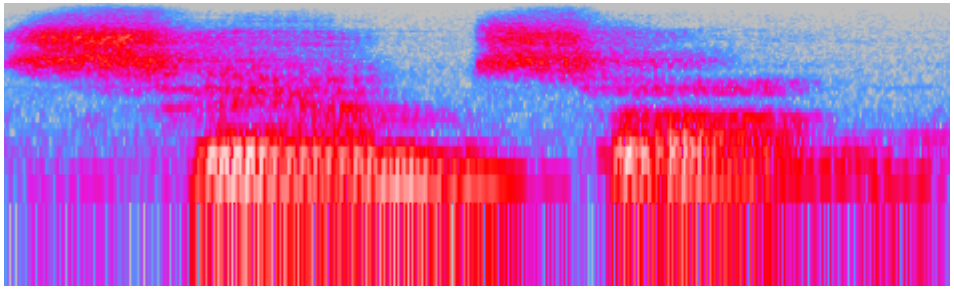
Hidden Markov Model (HMM):



Acoustic Model:



Audio Input:



(Graves & Jaitly. 2014)

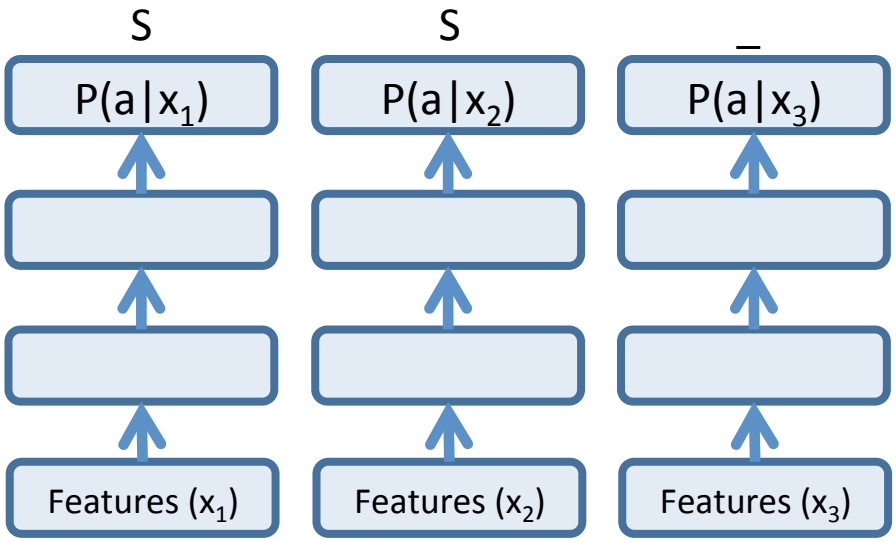
HMM-Free Recognition

Transcription: Samson

Characters: SAMSON

Collapsing function: SS___AA_M_S___O___NNNN

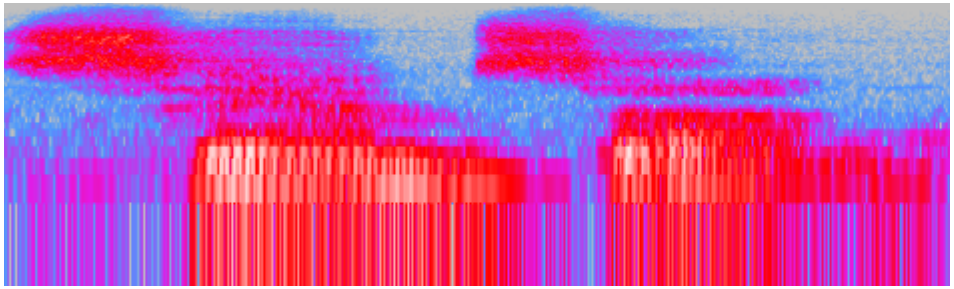
Acoustic Model:



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

The distribution over *characters*

Audio Input:



(Graves & Jaitly. 2014)

CTC Objective Function

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

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

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, ICML 2014)

Collapsing Example

Per-frame argmax:

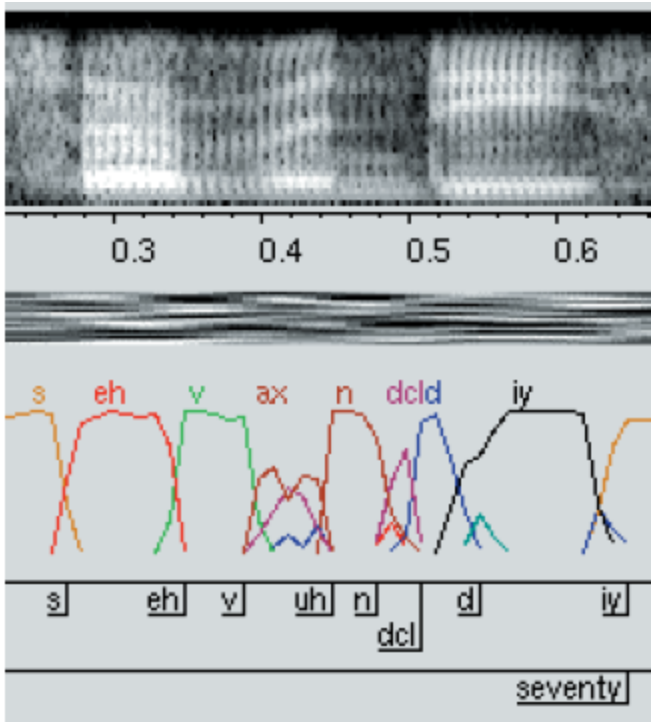
_____y
y__ee_____tt_ _____a_____
_rr__e_____hh_____b___ii_____lll__i_____tt_____aa_____tt_____iio__n___
___cc_____rrr_u_____ _____ii___ss
_____o_____nn_____hhh_a_____nnddd _____i__n___
__thh_e_____ _____bb_uuii_____llldd____ii____nng_____
_____l___o___o_g__g___ii____nng_____
____b___rr_ii_____ck__s_____p___ll__a_____sstt_____eerr__
_____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

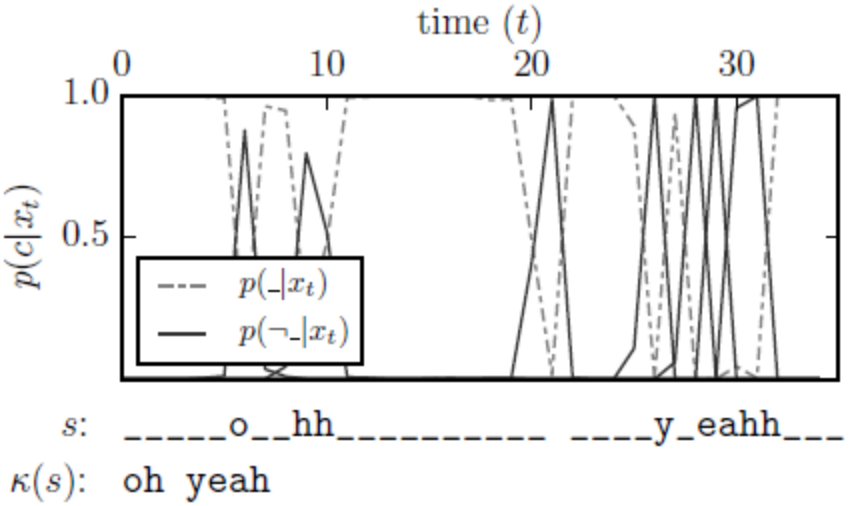
Reference:
yet a rehabilitation crew is on hand in the building lugging bricks plaster and blueprints for forty two new bedroom apartments

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

Comparing Alignments



HMM-GMM phone probabilities



CTC character probabilities

(HMM slide from Dan Ellis)

Example Results (WSJ)

YET A REHBILITATION CRU IS ONHAND IN THE BUILDING LOOGGING BRICKS PLASTER AND BLUEPRINS FOUR FORTY TWO NEW BETIN EPARTMENTS

YET A REHABILITATION CREW IS ON HAND IN THE BUILDING LUGGING BRICKS PLASTER AND BLUEPRINTS FOR FORTY TWO NEW BEDROOM APARTMENTS

THIS PARCLE GUNA COME BACK ON THIS ILAND SOM DAY SOO

THE SPARKLE GONNA COME BACK ON THIS ISLAND SOMEDAY SOON

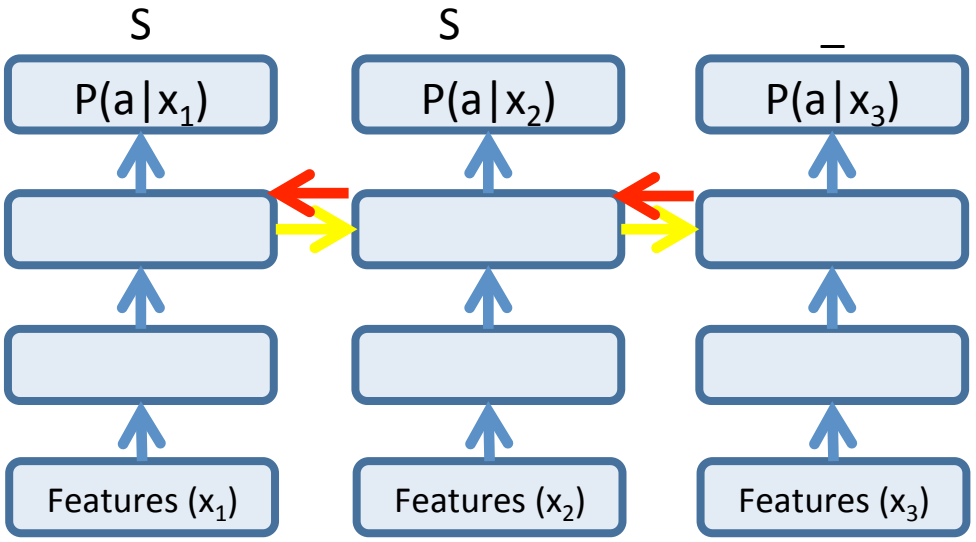
TRADE REPRESENTIGD JUIDER WARANTS THAT THE U S WONT BACKCOFF ITS PUSH FOR TRADE BARIOR REDUCTIONS

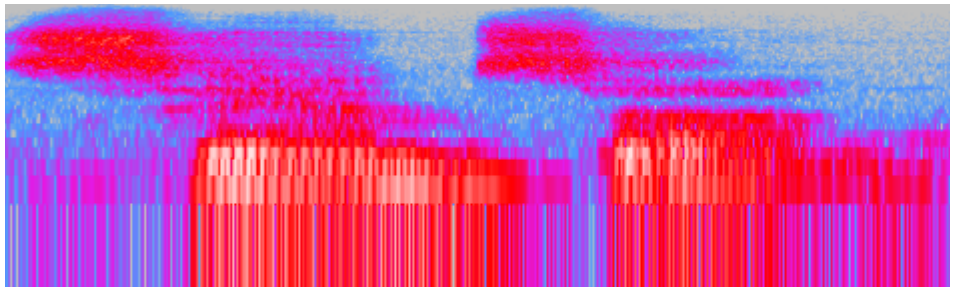
TRADE REPRESENTATIVE YEUTTER WARNS THAT THE U S WONT BACK OFF ITS PUSH FOR TRADE BARRIER REDUCTIONS

TREASURY SECRETARY BAGER AT ROHIE WOS IN AUGGRAL PRESSED FOUR ARISE IN THE VALUE OF KOREAS CURRENCY

TREASURY SECRETARY BAKER AT ROH TAE WOOS INAUGURAL PRESSED FOR A RISE IN THE VALUE OF KOREAS CURRENCY

Recurrence Matters!





Architecture	CER
DNN	22

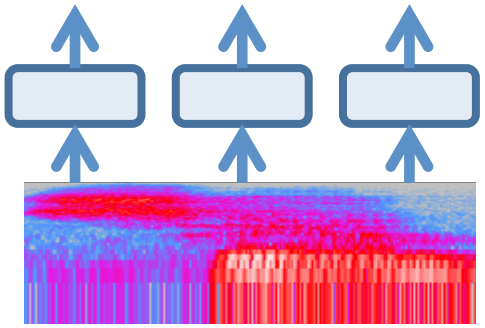
(Hannun, Maas, Jurafsky, & Ng. 2014)

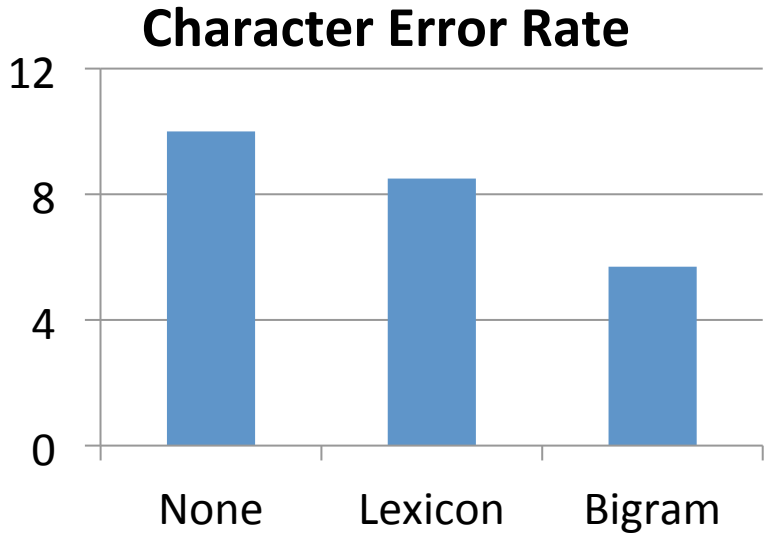
Decoding with a Language Model

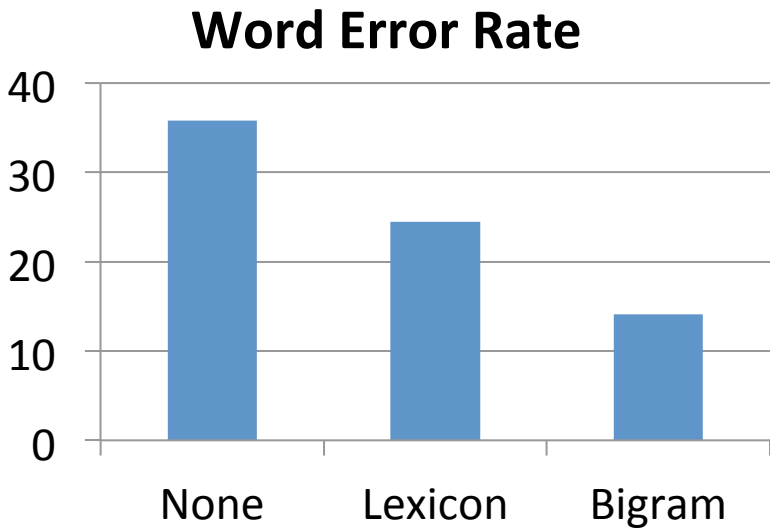
Lexicon [a, …, zebra]

Language Model p(“yeah” | “oh”)

Character Probabilities __oo_h__y_e_aa_h

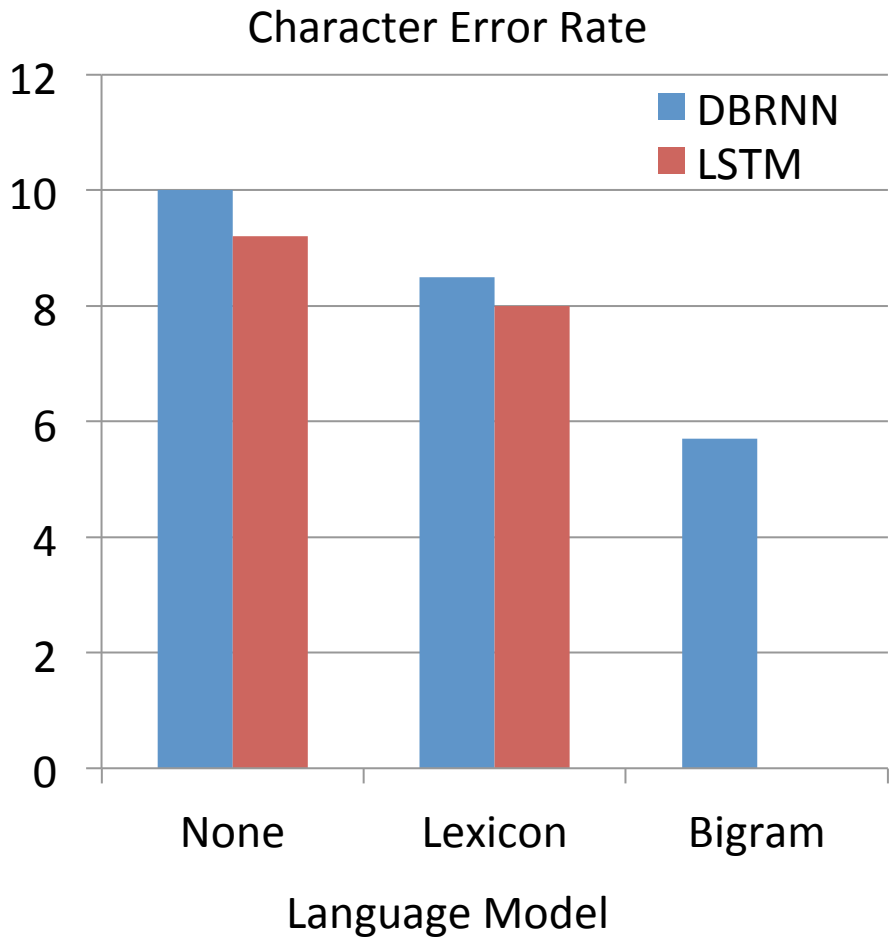


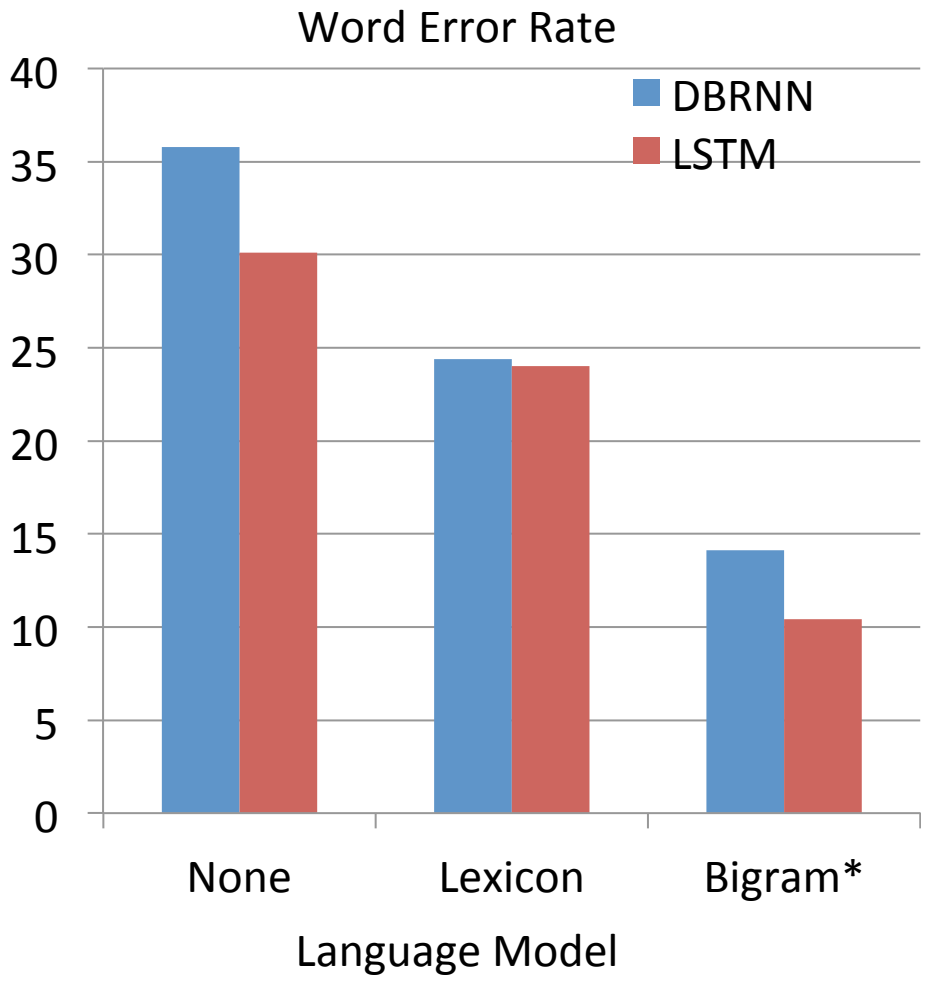




(Hannun, Maas, Jurafsky, & Ng. 2014)

DBRNN vs LSTM





* Graves & Jaitly use lattice rescoring

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

Rethinking Decoding

Out of Vocabulary Words

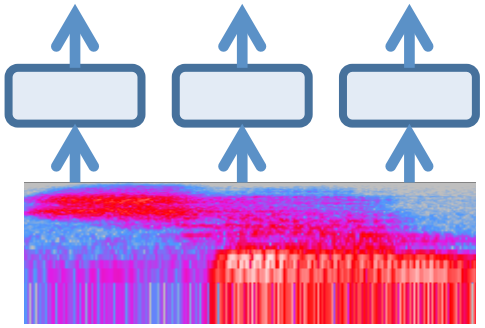
syriza bae abo-- sof-- schmidhuber

Lex [zebra]

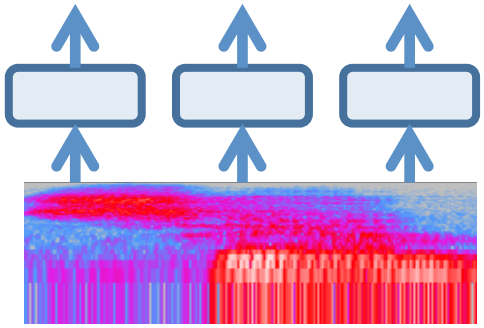
L uage M “ ” | “oh”)

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

Character Probabilities __oo_h__y_e_aa_h

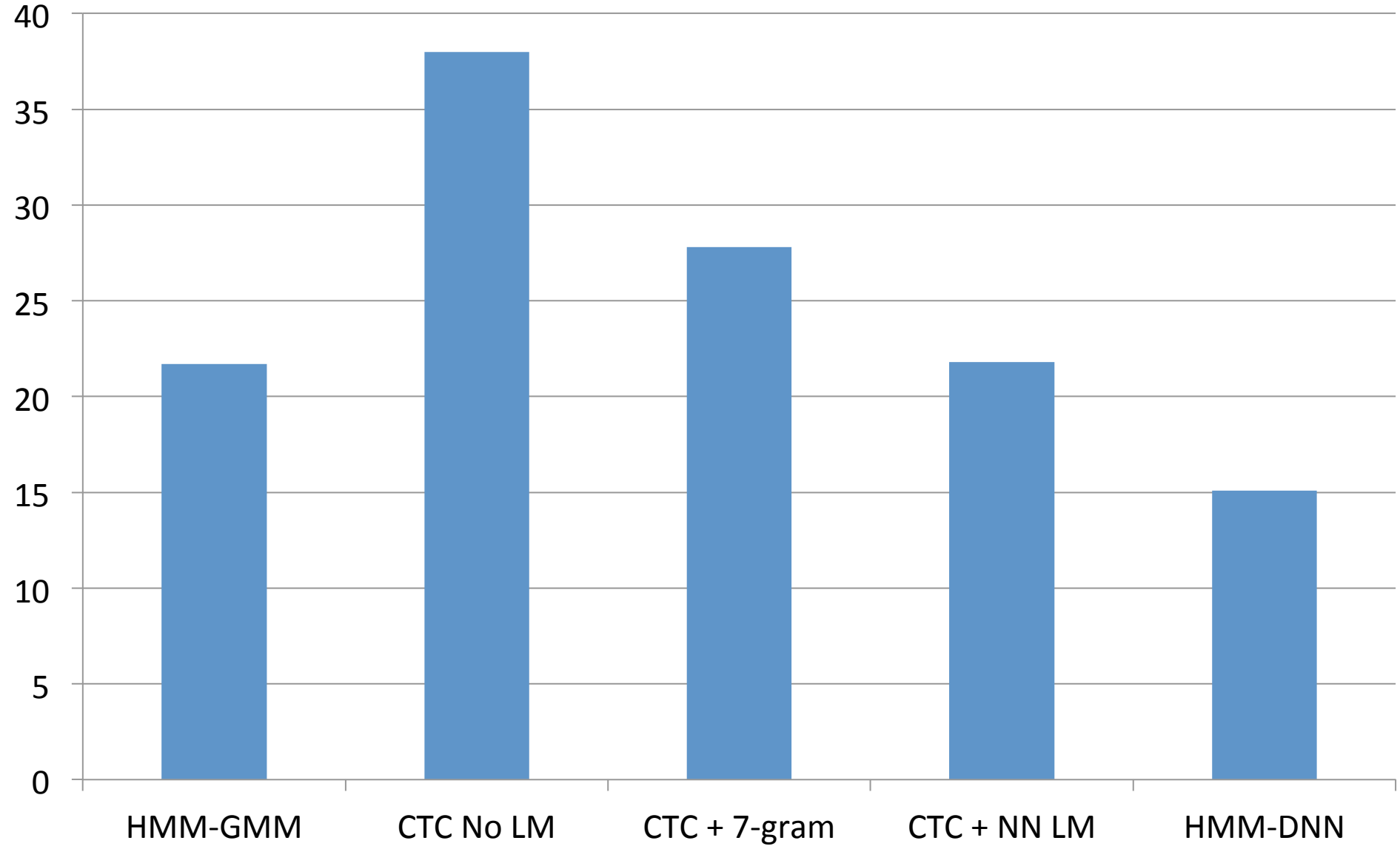


Character Probabilities __oo_h__y_e_aa_h



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

Lexicon-Free & HMM-Free on Switchboard



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

Transcribing Out of Vocabulary Words

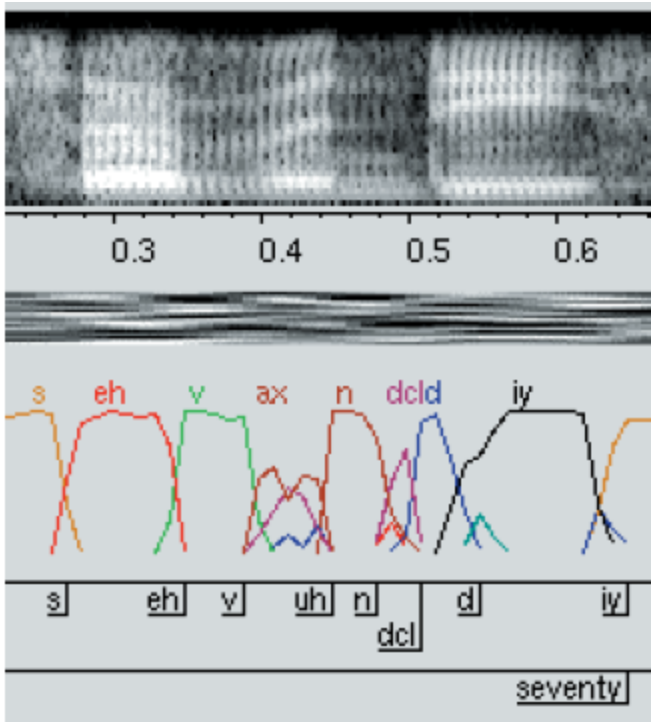
Truth: yeah i went into the i do not know what you think of *fidelity* but
HMM-GMM: yeah when the i don't know what you think of **fidel it even them**
CTC-CLM: yeah i went to i don't know what you think of **fidelity but um**

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**

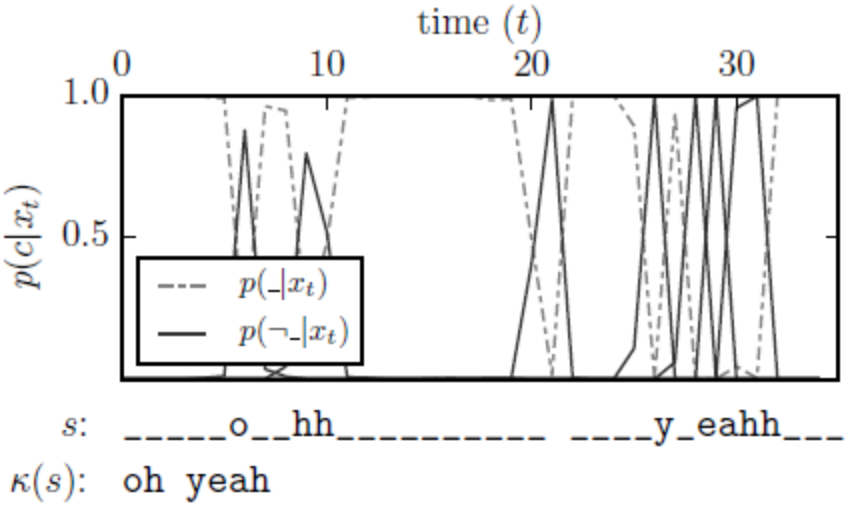
Truth: i would ima- well yeah it is i know you are able to stay home with them
HMM-GMM: i would **amount** well yeah it is i know um you're able to stay home with them
CTC-CLM: i would **ima-** well yeah it is i know uh you're able to stay home with them

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

Comparing Alignments



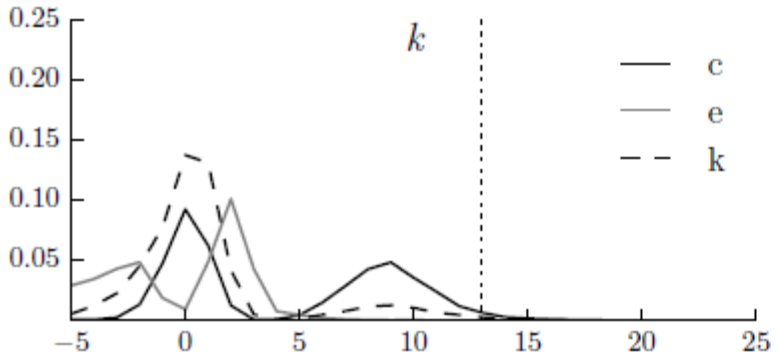
HMM-GMM phone probabilities



CTC character probabilities

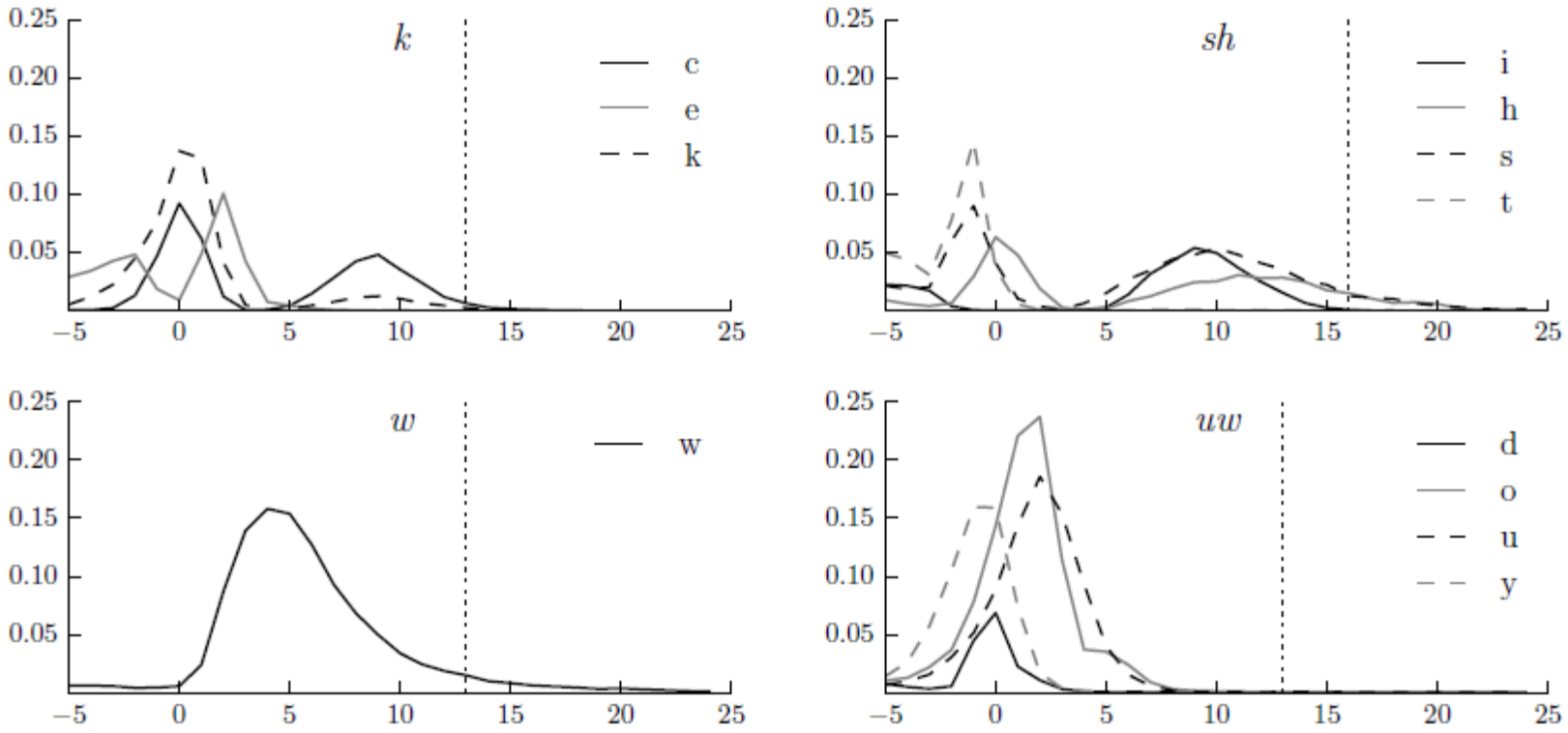
(HMM slide from Dan Ellis)

Learning Phonemes and Timing



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

Learning Phonemes and Timing



(Maas*, Xie*, Jurafsky, & Ng. NAACL 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
- Recent work demonstrates feasibility of pure NN-based speech recognition systems (no HMM)

End

- More on spoken language understanding:
  - cs224s.stanford.edu
  - MSR video: youtu.be/Nu-nlQqFCKg
- Open source speech recognition toolkit (Kaldi):
  - Kaldi.sf.net