

Humans versus Machines: The Case of Conversational Speech Recognition

Andreas Stolcke
Speech & Dialog Research Group
Microsoft AI & Research

anstolck@microsoft.com



INTERSPEECH 2018

Speech research for emerging markets
in multilingual societies

General Chair

B. Yegnanarayana

Local Organizing Chairs

Suryakanth V Gangashetty

V. Kamakshi Prasad

C. Krishna Mohan

General Co-Chairs

C. Chandra Sekhar

Shrikanth Narayanan

S. Umesh

S. R. M. Prasanna

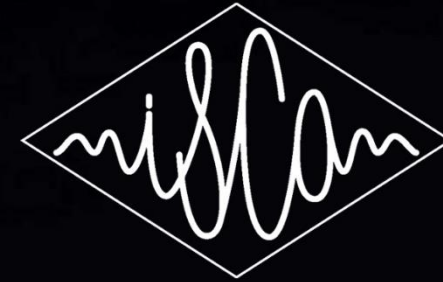
Technical Chairs

Prasanta Kumar Ghosh

Hema A Murthy

Preeti Rao

Paavo Alku



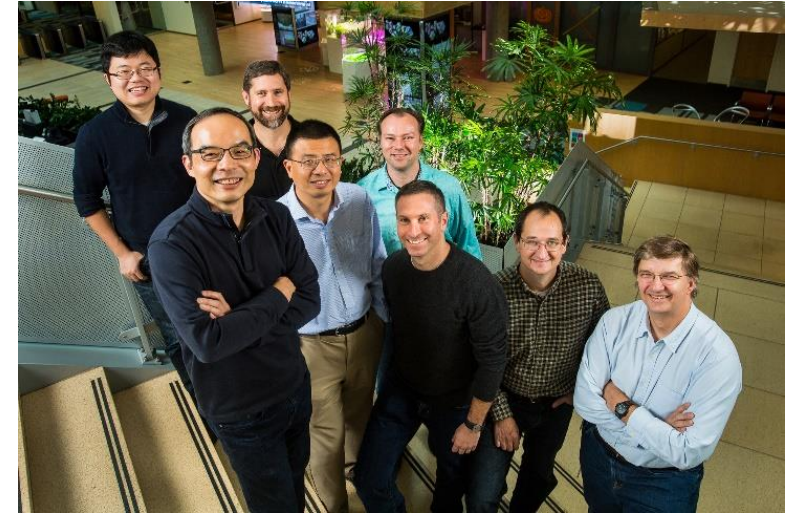
2nd - 6th September 2018

@ HICC Hyderabad

Hyderabad International Convention Centre



Acknowledgments



ACHIEVING HUMAN PARITY IN CONVERSATIONAL SPEECH RECOGNITION

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig

Microsoft Research
Technical Report MSR-TR-2016-71

A great team effort!

ABSTRACT

Conversational speech recognition has served as a flagship speech recognition task since the release of the DARPA Switchboard corpus in the 1990s. In this paper, we measure the human error rate on the widely used NIST 2000 test set, and find that our latest automated system has reached human parity. The error rate of professional transcriptionists is 5.9% for the Switchboard portion of the data, in which newly acquainted pairs of people discuss an assigned topic, and 11.3%

collections of the 1990s and early 2000s provide what is to date the largest and best studied of the conversational corpora. The history of work in this area includes key contributions by institutions such as IBM [12], BBN [13], SRI [14], AT&T [15], LIMSI [16], Cambridge University [17], Microsoft [18] and numerous others.

In the past, human performance on this task has been widely cited as being 4% [19]. However, the error rate estimate in [19] is attributed to a “personal communication,” and the actual source of this number is unknown. The history

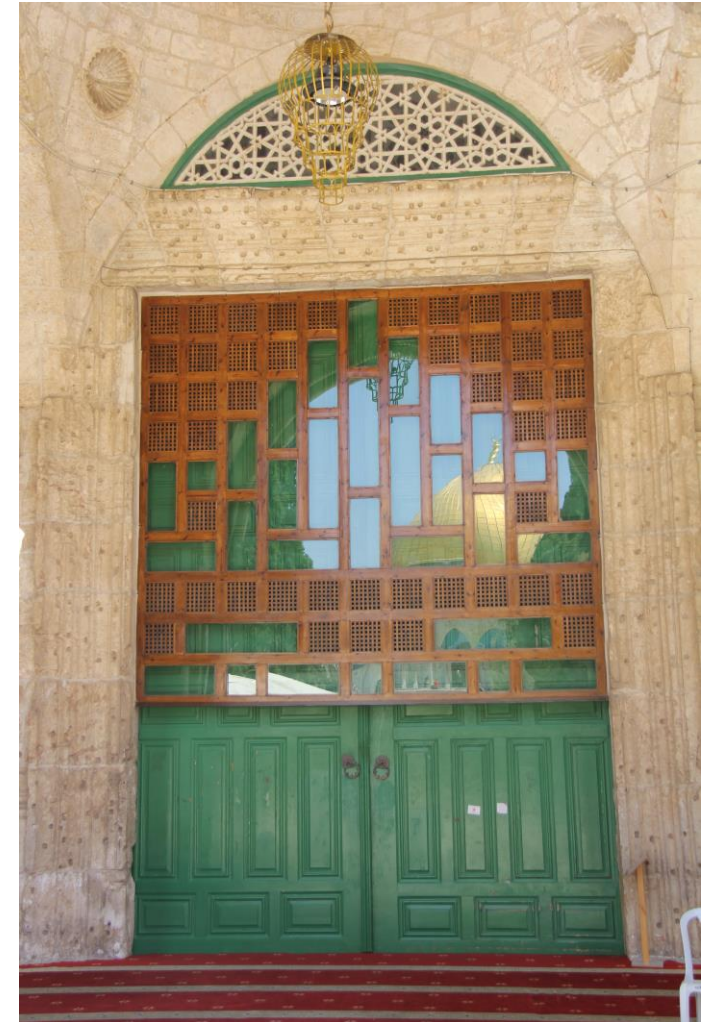
17 Oct 2016

Roadmap

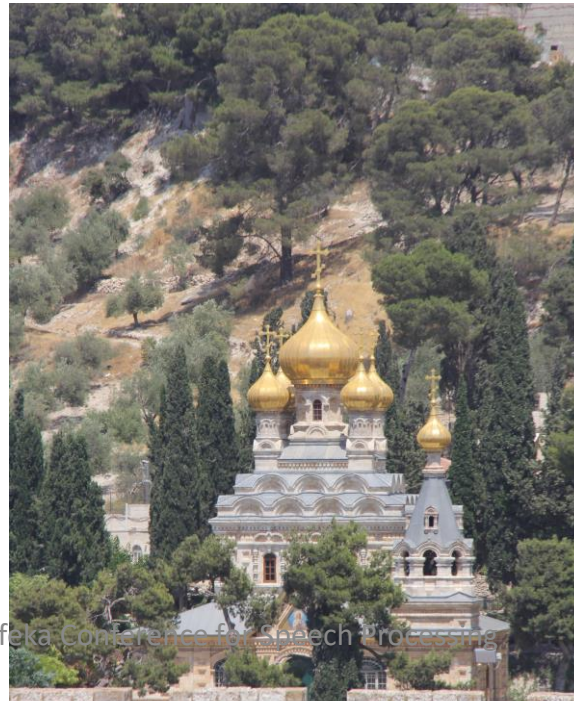
- History of conversational speech transcription
- The Human Parity experiment
- What is human performance?
- Recognition system
- Human vs. machine error comparison
- Conclusions

The Human Parity Experiment

- Conversational telephone speech has been a benchmark in the research community for 20 years
- Can we achieve human-level performance on conversational speech?
- Top-level tasks:
 - Measure human performance
 - Build the best possible recognition system
- Analyze results
 - Inform future research
 - Pick the next challenge ...



The History



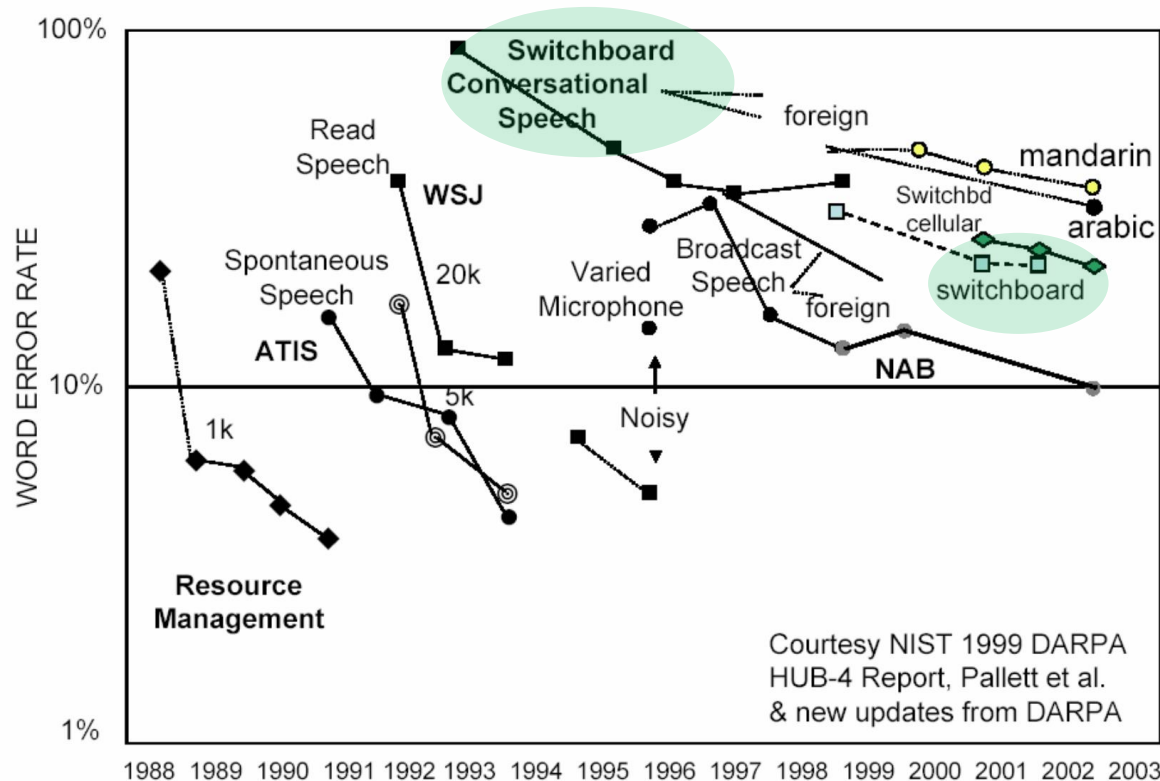
A Community Effort



30 Years of Speech Recognition Benchmarks

For many years, DARPA drove the field by defining public benchmark tasks

DARPA Speech Recognition Benchmark Tests





Read and planned speech:

RM  

ATIS 

WSJ 

Conversational Telephone Speech (CTS):

Switchboard  
(strangers, on-topic)

Call Home  
(friends & family, unconstrained)

Prior Work

- DARPA funding ended in 2004 – a collection of papers was published in IEEE Transactions on Speech Audio and Language Processing
 - Best error rate \approx 15% Switchboard, \approx 40% for CallHome
- With the advent of DNNs, significant progress on CTS was reported [Seide et al. 2011]
- More recent papers by IBM group, bringing WER to 6.6%, as of late 2016 [Saon et al., Interspeech]
 - IBM also quoted a 4% human error rate from the literature

Measuring Human Performance

An Early Estimate (1997)

- The 4% rumor



[Lippman, 1997]

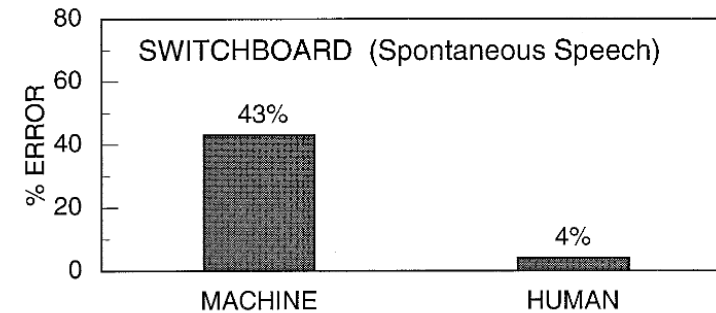


Fig. 7. Word error rates for humans and a high-performance HMM recognizer on phrases extracted from spontaneous telephone conversations in the Switchboard speech corpus (Liu et al., 1996; [Martin, 1996](#)).

1996. Speech recognition on Mandarin Call Home: A large-vocabulary, conversational, and telephone speech corpus. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 157–160.

[A. Martin, 1996. Personal communication.](#)

Miller, G.A., 1962. Decision units in the perception of speech. Institute of Radio Engineers Transactions on Information Theory 8, 81–83.

NIST Study of Transcriber Disagreement (2010)

Language	Genre	Careful Transcription WDR	Quick (Rich) Transcription WDR
English	CTS	4.1-4.5%	9.63% (5 pairs)
	Meeting	-	6.23% (4 pairs)
	Interview	n/a	3.84% (22 pairs)
	BN	1.3%	3.5% (6 pairs)
	BC	n/a	6.3% (6 pairs)

[Glenn et al., LREC 2010]

Significant variability.

Note the bulk of the CTS training data was "quick transcribed."

Our Human Experiment (2015)

- Skype Translator has a weekly transcription contract
 - For quality control, training, etc.
- Initial transcription followed by a second checking pass
 - Two transcribers on each speech excerpt
- One week, we added **NIST 2000 CTS evaluation** data to the pipeline
 - Speech was pre-segmented as in NIST evaluation



The Results

- Applied NIST scoring protocol
- Text normalized to minimize WER (on test set!)
- Switchboard: **5.9%** error rate
- CallHome: **11.3%** error rate
- SWB in the 4.1% - 9.6% range expected
- CH is *difficult for both people and machines*
 - Machine error about 2x higher
 - High ASR error not just because of mismatched conditions

Language	Genre	Careful Transcription WDR	Quick (Rich) Transcription WDR
English	CTS	4.1-4.5%	9.63% (5 pairs)
	Meeting	-	6.23% (4 pairs)
	Interview	n/a	3.84% (22 pairs)
	BN	1.3%	3.5% (6 pairs)
	BC	n/a	6.3% (6 pairs)

History of Human SWB Error Estimates

- Lippman (1997): 4%
 - based on “personal communication” with NIST, no experimental data cited
- LDC LREC paper (2010): 4.1-4.5%
 - Measured on a different dataset (but similar to our NIST eval set, SWB portion)
- Microsoft (2016): 5.9%
 - Transcribers were blind to experiment
 - 2-pass transcription, isolated utterances (no “transcriber adaptation”)
- IBM (2017): 5.1%
 - Using multiple independent transcriptions, picked best transcriber
 - Vendor was involved in experiment and aware of NIST transcription conventions

Recognition System

- Acoustic modeling
- Language modeling
- System combination

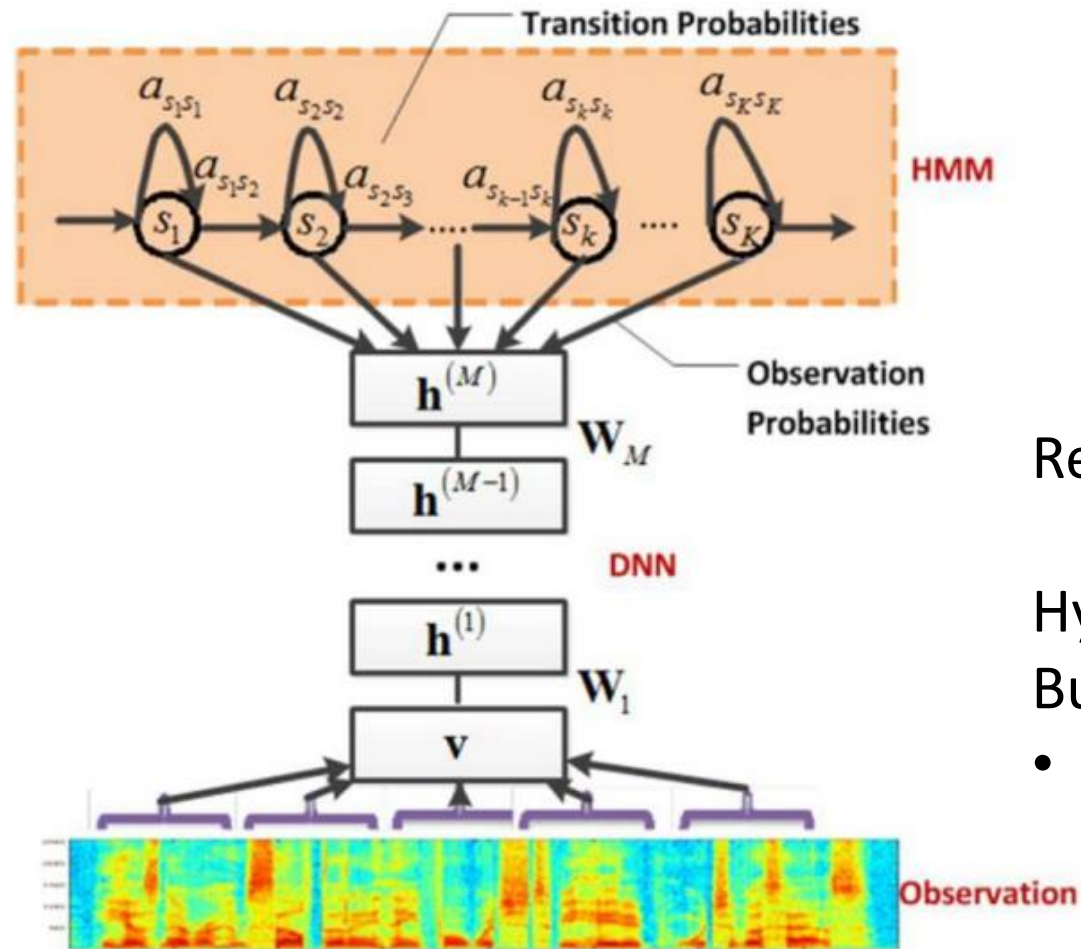
Recognition System: Highlights

- New state of the art in conversational telephone speech transcription accuracy using
- Multiple acoustic model architectures:
 - ResNet, VGG and LACE convolutional nets (CNNs)
 - Bidirectional LSTM nets
 - Speaker-adaptive modeling using i-vectors
 - Lattice-free sequence training
- Forward/backward LSTM-LM rescoring using multiple input representations
- Search for complementary acoustic model
- Confusion-network-based, weighted combination
- System achieves accuracy slightly better than human transcribers: 5.8% WER on Switchboard and 11.0% on CallHome

State of the Art has a Long History

- The current favorites: CNNs, LSTMs
- But building on key past innovations:
 - HMM modeling
 - Distributed Representations [Hinton '84]
 - Early CNNs, RNNs, TDNNs [Lang & Hinton '88, Waibel et al. '89, Robinson '91, Pineda '87]
 - Hybrid training [Renals et al. '91, Bourlard & Morgan '94]
 - Discriminative modeling
 - Speaker adaptation
 - System combination

Acoustic Modeling Framework: Hybrid HMM/DNN



	CallHome	Switchboard
DNN	21.9%	13.4%

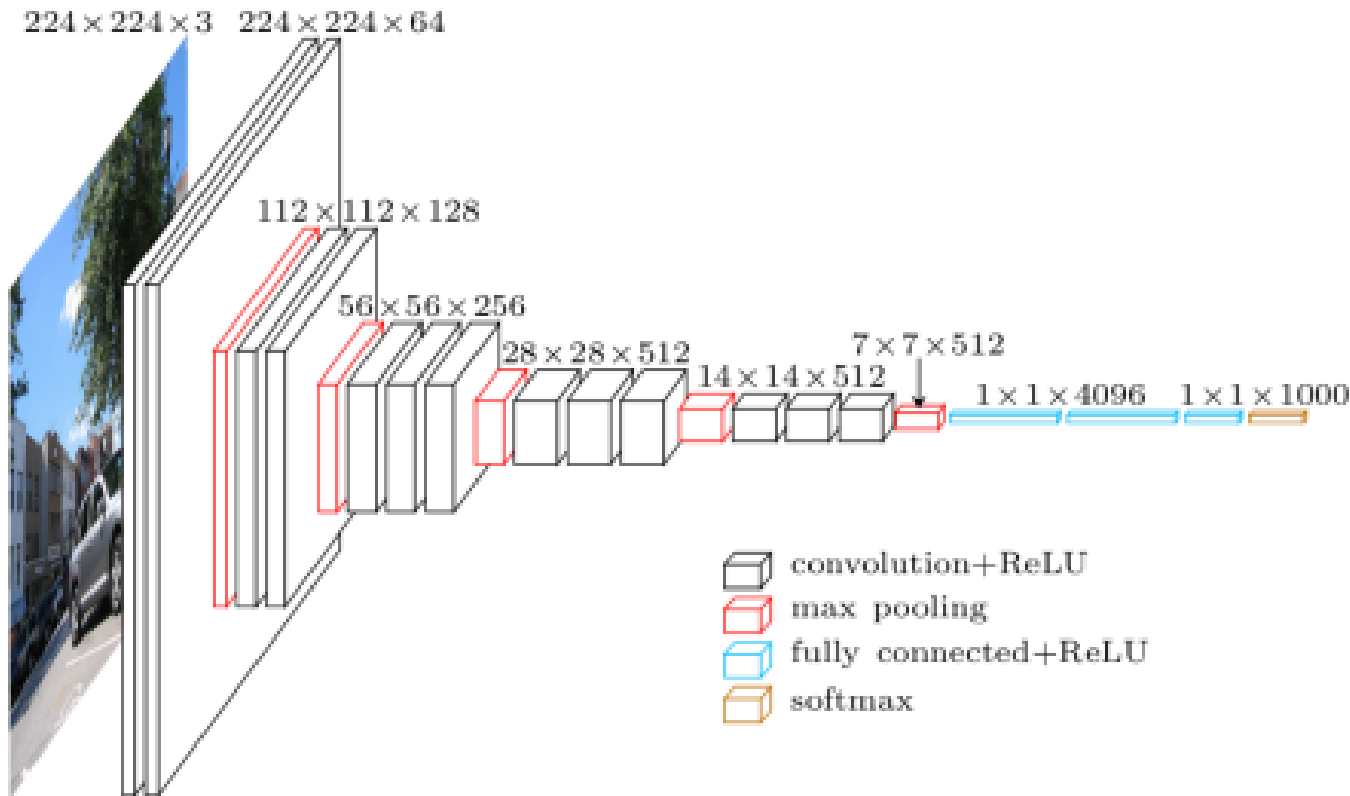
1st pass decoding

Record performance in 2011 [Seide et al.]

Hybrid HMM/NN approach still standard
But DNN model now obsolete (!)

- Poor spatial/temporal invariance

Acoustic Modeling: VGG CNN



Adapted for speech from image processing [Saon et al., 2016]

Robust to temporal and frequency shifts

[Simonyan & Zisserman, 2014; Frossard 2016, Saon et al., 2016, Krizhevsky et al., 2012]

Acoustic Modeling: ResNet CNNs

Adds a non-linear offset to linear transformation of features

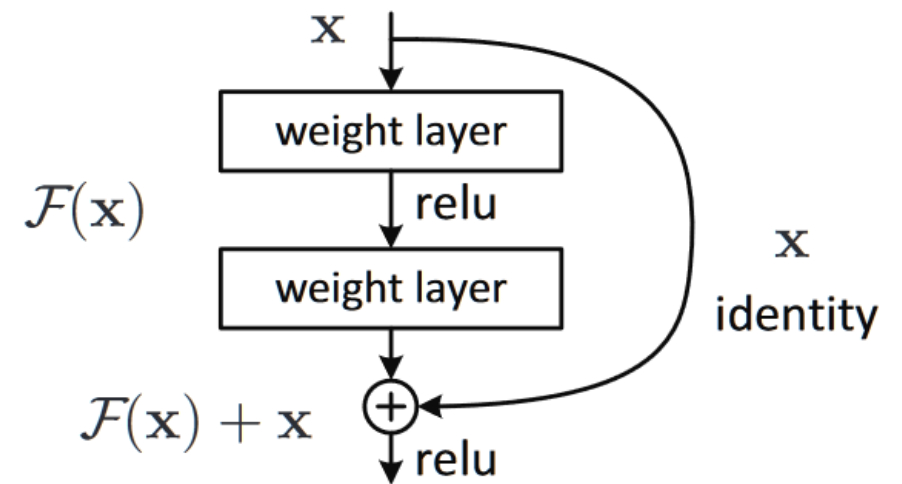
Similar to fMPE in Povey et al., 2005

See also Ghahremani & Droppo, 2016

Our best single model after rescoring

	CallHome	Switchboard
DNN	21.9%	13.4%
ResNet	17.3%	11.1%

1st pass decoding



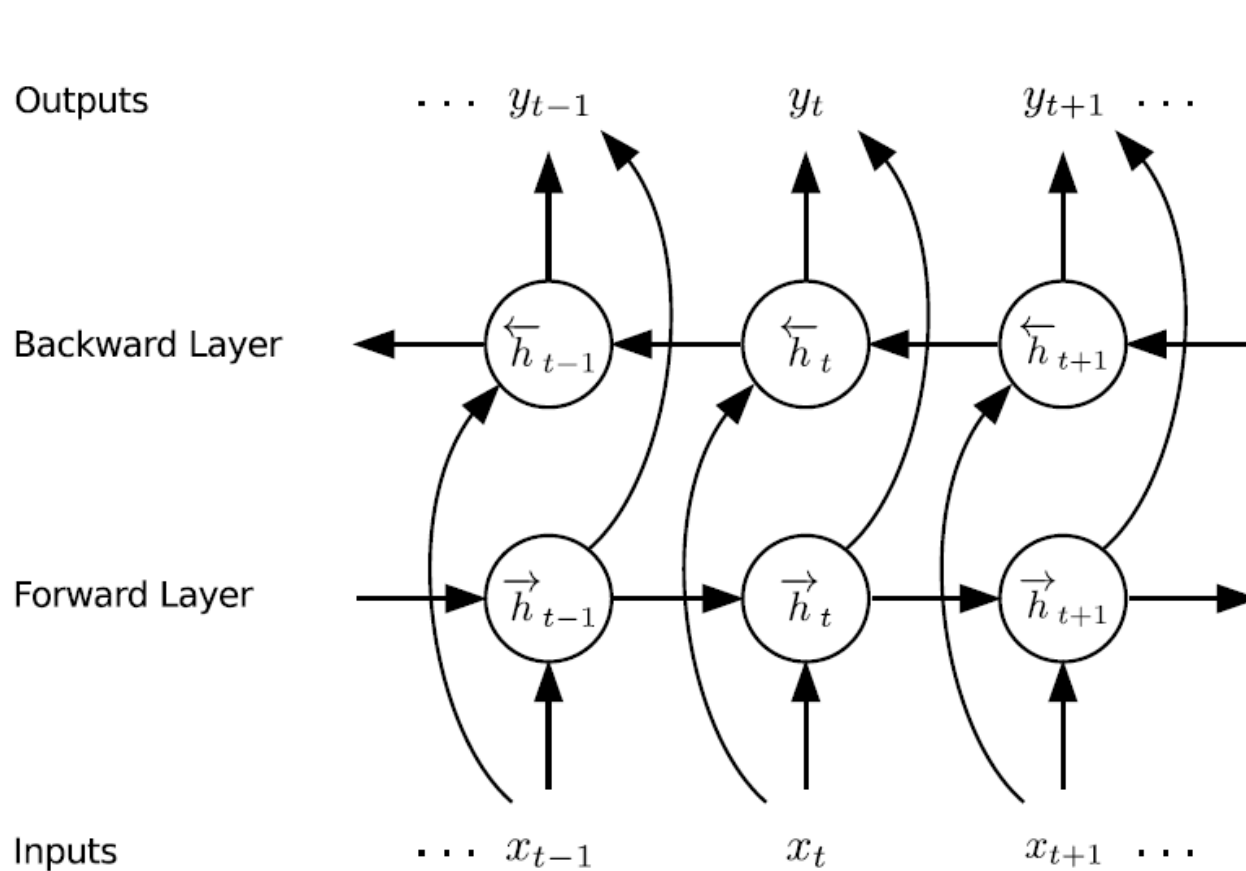
[He et al., 2015]

CNN Comparison

VGG Net (85M Parameters)	Residual-Net (38M Parameters)	LACE (65M Parameters)
14 weight layers	49 weight layers	22 weight layers
40x41 input	40x41 input	40x61 input
3 – conv 3x3, 96	3 – [conv 1x1, 64 conv 3x3, 64 conv 1x1, 256]	5 – conv 3x3, 128
Max pool	4 – [conv 1x1, 128 conv 3x3, 128 conv 1x1, 512]	5 – conv 3x3, 256
4 – conv 3x3, 192	6 – [conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024]	5 – conv 3x3, 512
Max pool	3 – [conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048]	5 – conv 3x3, 1024
4 – conv 3x3, 384	Average pool	1 – conv 3x4, 1
Max pool	Softmax (9000)	Softmax (9000)
2 – FC – 4096		
Softmax (9000)		

Very deep
 Many parameters
 Small convolution patterns
 Processing ~ ½ second per window

Acoustic Modeling: Bidirectional LSTMs



	CallHome	Switchboard
DNN	21.9%	13.4%
ResNet	17.3%	11.1%
LACE	16.9%	10.4%
BLSTM	17.3%	10.3%

Stable form of recurrent neural net
Robust to temporal shifts

2nd best single model

[Hochreiter & Schmidhuber, 1997,
Graves & Schmidhuber, 2005; Sak et al., 2014]

Runtimes

	DNN	BLSTM	ResNet	LACE
AM Training, GPU	0.012	0.022	0.60	0.23
AM eval, GPU	0.0064	0.0081	0.15	0.081
AM eval, CPU	0.052	NA	11.7	8.47
Decoding, GPU	1.04	1.40	1.19	1.38

GPU 10 to 100x
faster than CPU

(Multiples of real-time, smaller is better)

AM Training: Forward, Backward + Update computations

AM eval: Forward probability computation only

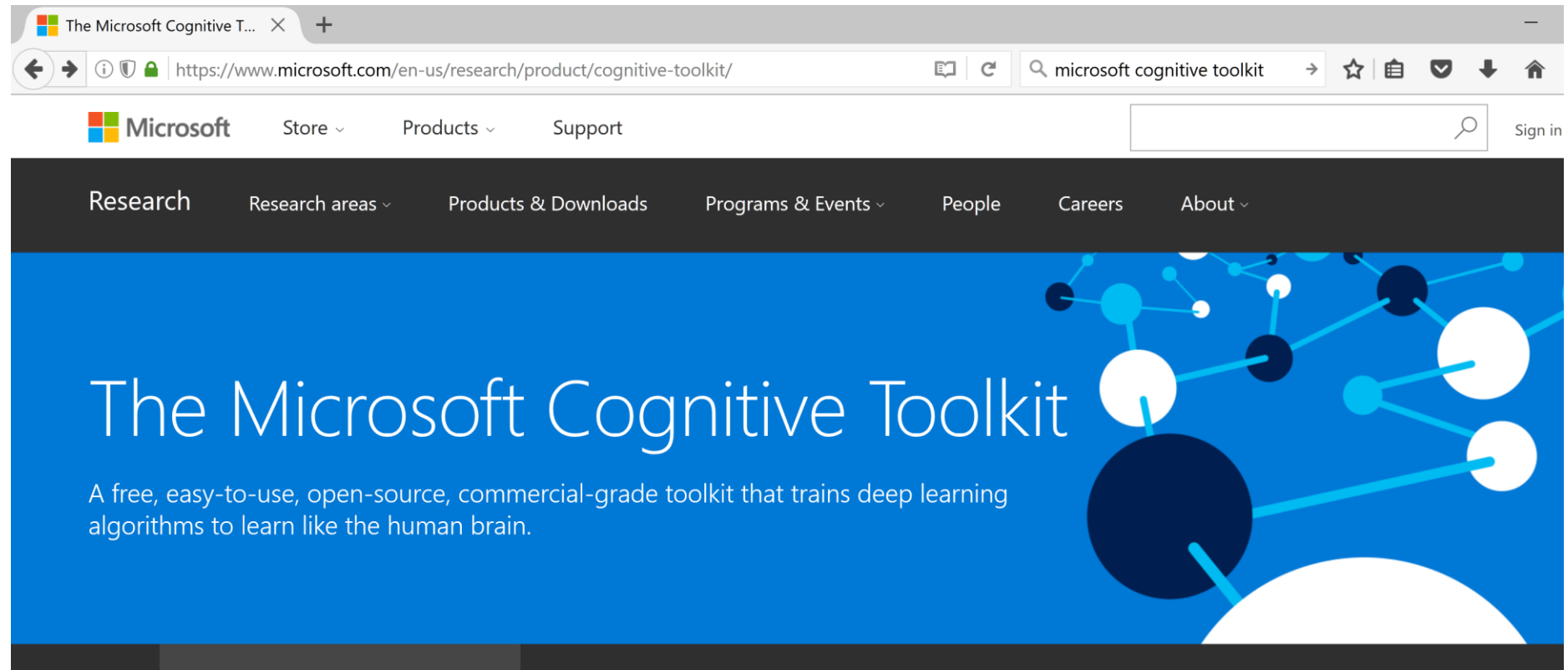
Decoding: Mixed GPU/CPU, complete decoding time with open beams

Titan X GPU & Intel Xeon E5-2620 v3 @2.4GHz, 12 cores

All times are xRT (fraction of real-time required) on Titan X GPU

Cognitive Toolkit (CNTK) Training

- Flexible
- Multi-GPU
- Multi-Server
- 1-bit SGD
- All AM training
- Best LM training



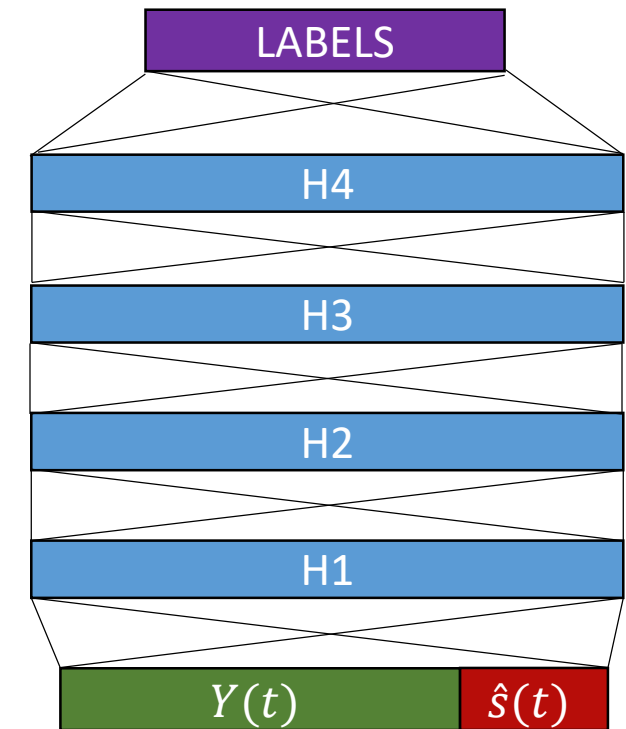
I-vector Adaptation

5-10% relative improvement for Switchboard

Configuration	ResNet		LACE		BLSTM	
	CH	SWB	CH	SWB	CH	SWB
Baseline	17.5	11.1	16.9	10.4	17.3	10.3
i-vector	16.6	10.0	16.4	9.3	17.6	9.9

I-vectors give a fixed-length representation of a speaker's voice [Dehak et al. 2011; Saon et al., 2013]

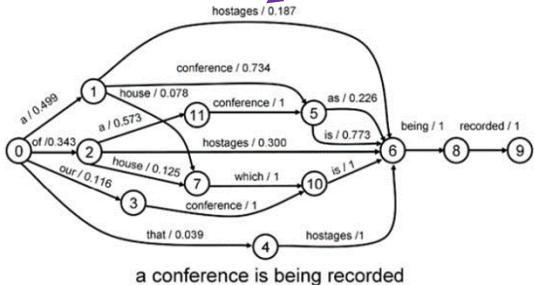
- 100-dim i-vectors computed per conversation side
- CNN models: i-vectors multiplied by weight matrix, serves as additional bias prior to non-linearity
- BLSTM models: i-vectors appended to each input frame



Lattice-free Discriminative Training

$$\begin{aligned}
 & \arg \max_{\Theta} \sum_{w,a \in \text{Data}} \log \frac{P(w,a;\Theta)}{P(w)P(a;\Theta)} \\
 &= \arg \max_{\Theta} \sum_{w,a \in \text{Data}} \log \frac{P(a|w;\Theta)}{P(a;\Theta)} \\
 &= \arg \max_{\Theta} \sum_{w,a \in \text{Data}} \log \frac{P(a|w;\Theta)}{\sum_{w'} P(w')P(a|w';\Theta)}
 \end{aligned}$$

Traditionally approximated by word sequences in lattice (DAG)

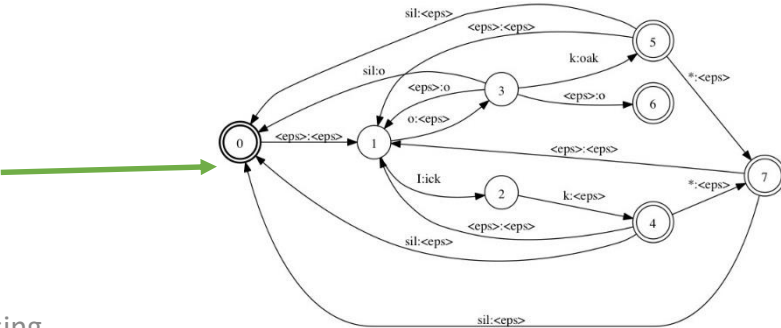


a conference is being recorded
July 3, 2017

- Simple brute force MMI (maximum mutual information criterion)
- Avoids need to generate lattices
- Alignments are always current
- Forward-backward computation can be reduced to matrix operations, run efficiently on GPUs

[Chen et al., 2006, McDermott et al., 2014, Povey et al., 2016]

Instead LFMMI uses all possible word sequences in cyclic FSA



Lattice-free MMI Improvements

Configuration	ResNet		LACE		BLSTM	
	CH	SWB	CH	SWB	CH	SWB
Baseline	17.5	11.1	16.9	10.4	17.3	10.3
i-vector	16.6	10.0	16.4	9.3	17.6	9.9
i-vector+LFMMI	15.2	8.6	16.2	8.5	16.3	8.9

8-14% relative improvement on SWB

- Denominator LM predicts senones based on mixed senone/phone history
- Denominator graph has 52k states and 215k transitions
- GPU-side alpha-beta computation is 0.18xRT, exclusive of NN evaluation

Language Models

- 1st pass n-gram:
 - SRI-LM, 30k vocab, 16M n-grams
- Rescoring n-gram:
 - SRI-LM, 145M n-grams
- RNN LM
 - CUED Toolkit, two 1000 unit layers
 - Relu activations, noise-contrastive estimation (NCE) training
 - Two differently initialized models, plus Ngram LM, interpolated at the word level
- LSTM LM
 - Cognitive Toolkit (CNTK), three 1000 unit layers
 - Interpolated word and letter-trigram encoding models, plus Ngram LM

by Jim Unger



Language Modeling: Results

Other tricks that help:

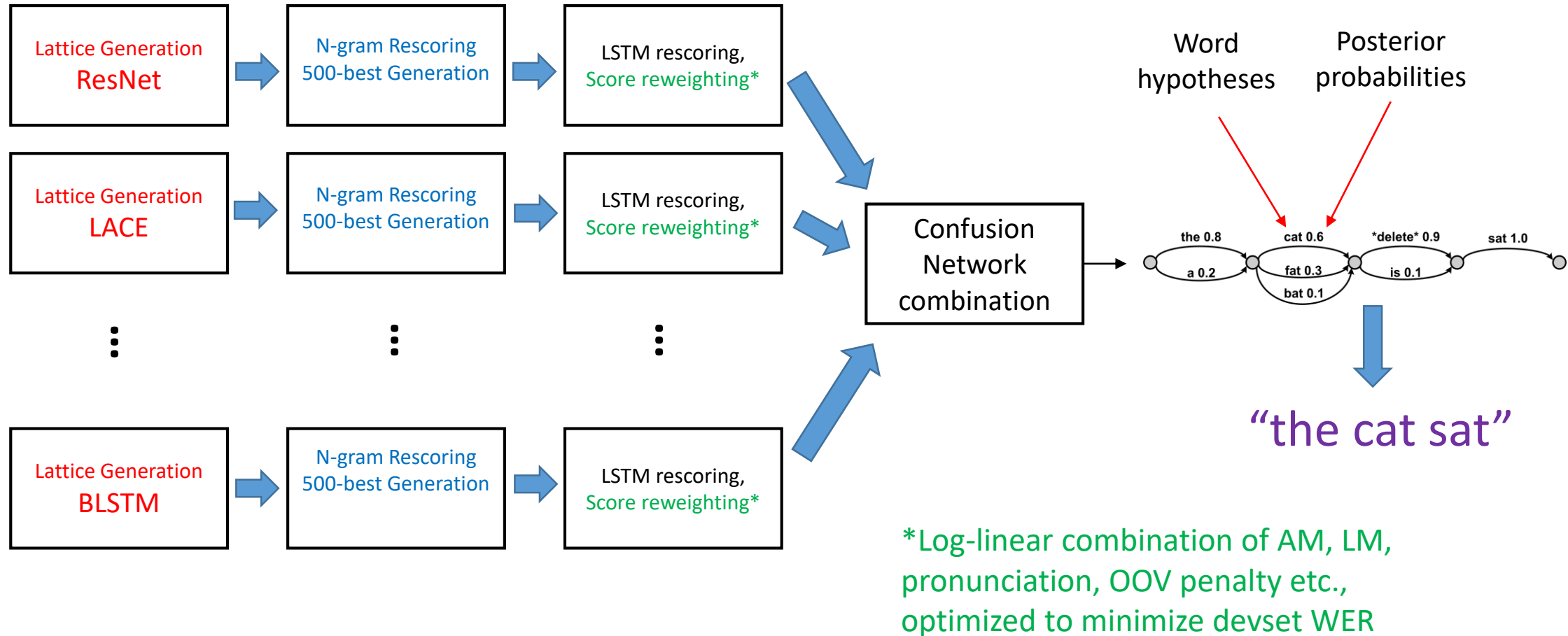
- Train first on in-domain and out-of-domain (Web) data, then tune on in-domain (CTS) data only
- In rescoring, forward and backward running sentence-scores are averaged
- Words outside the NN vocabulary (which is smaller than the N-gram vocab) incur a penalty – magnitude estimated on dev data

WER with ResNet acoustic model
Perplexities on 1997 eval refs

Language model	PPL	WER
4-gram LM (baseline)	69.4	8.6
+ RNNLM, CTS data only	62.6	7.6
+ Web data training	60.9	7.4
+ 2nd hidden layer	59.0	7.4
+ 2-RNNLM interpolation	57.2	7.3
+ backward RNNLMs	-	6.9
+ LSTM-LM, CTS + Web data	51.4	6.9
+ 2-LSTM-LM interpolation	50.5	6.8
+ backward LSTM-LM	-	6.6

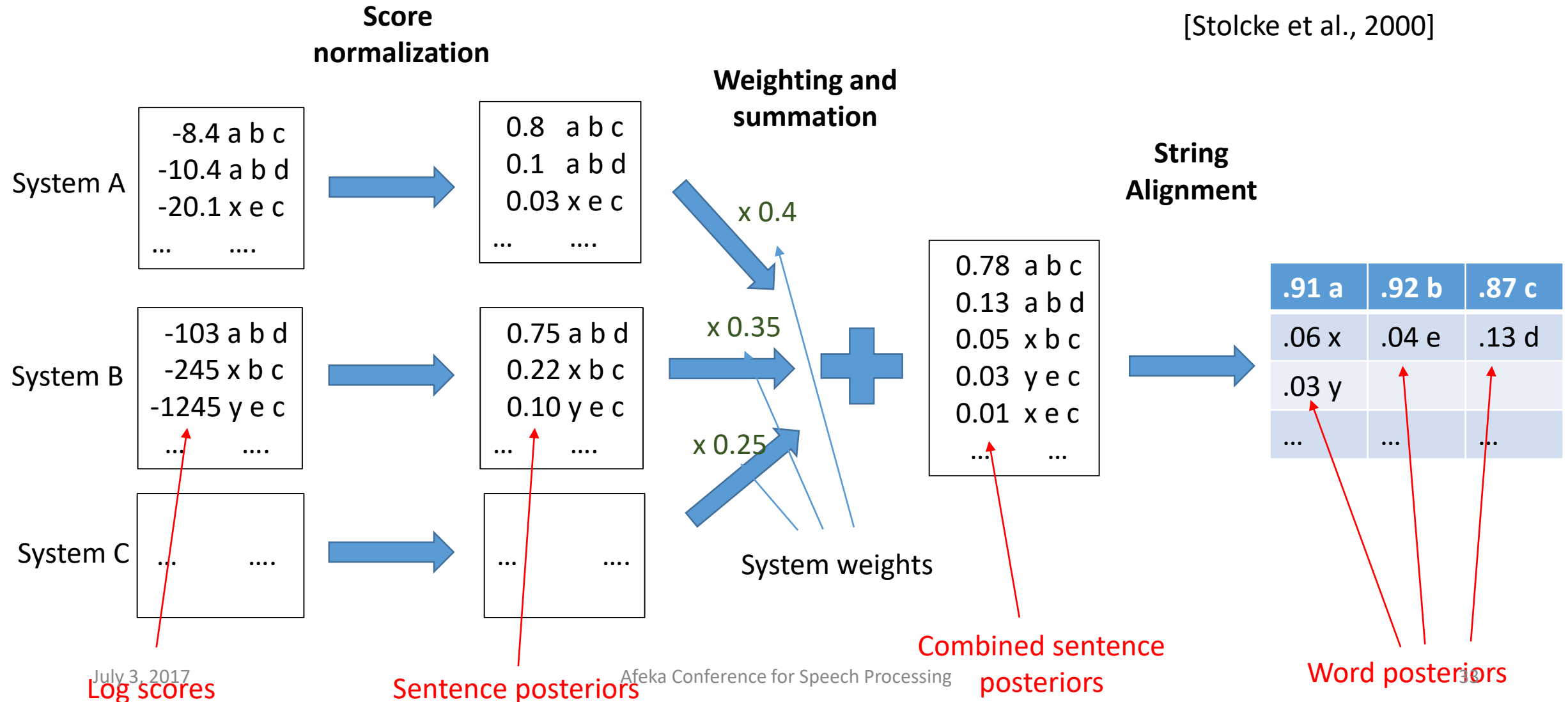
LSTM-LM gives 23% relative improvement over N-gram LM

System Combination



N-best Confusion Network Combination

[Stolcke et al., 2000]



System Selection and Weighting

- Combining all systems is not optimal
- ... and would be way to slow
- **search-rover-combo**: new SRILM tool to find best subset of systems
 - Forward greedy search (always add the system that gives the largest gain)
 - Stop when no more gain can be had
 - Reestimate system weights at each step, using EM
 - Smooth weight estimates hierarchically with previous weights (shrinkage)

Two-level System Combination

- Limited training data for system selection and weighting
 - Using old eval sets, a few thousand utterances)
- Use prior knowledge that helps reduce number of free parameter
- One strategy: two-level combination
 - Search for best subset of BLSTM systems with different meta parameters (number of senones, NN smoothing method, choice of dictionary)
 - Combine those with equal weighting
 - Treat BLSTM combo as a single system in search for all-out system combination
- First-level system selection picks systems that differ along all dimensions
 - BLSTM(1) - Baseline (no smoothing, 9k senones)
 - BLSTM(2) - With spatial smoothing [Droppo, Interspeech 2017], 9 senones
 - BLSTM(3) - With spatial smoothing, 27k senones
 - BLSTM(4) - With spatial smoothing, 27k senones, alternate dictionary

Data

- AM training: 2000h (Fisher, Switchboard, but not CallHome)
 - One system uses 300h (Switchboard only), for diversity
- LM training: Fisher, Switchboard, CallHome, UW Web data, Broadcast News
- Dev-testing, combination tuning: NIST 2002 Switchboard-1 eval set
- Evaluation: NIST 2000 (Switchboard and CallHome portions)

Overall System Results

System	N-gram LM		RNN-LM		LSTM-LM	
	CH	SWB	CH	SWB	CH	SWB
ResNet, 300h training	19.2	10.0	17.7	8.2	17.0	7.7
ResNet	14.8	8.6	13.2	6.9	12.5	6.6
ResNet, GMM alignments	15.3	8.8	13.7	7.3	12.8	6.9
VGG	15.7	9.1	14.1	7.6	13.2	7.1
VGG + ResNet	14.5	8.4	13.0	6.9	12.2	6.4
LACE	15.0	8.4	13.5	7.2	13.0	6.7
BLSTM (1)	16.5	9.0	15.2	7.5	14.4	7.0
BLSTM (2)	15.4	8.6	13.7	7.4	13.0	7.0
BLSTM (3)	15.3	8.3	13.8	7.0	13.2	6.8
BLSTM (4)	14.9	8.3	13.7	7.0	13.0	6.7
BLSTM combination	13.2	7.3	12.1	6.4	11.6	6.0
Full system combination	13.0	7.3	11.7	6.1	11.0	5.8
ICASSP 2017 paper	13.3	7.4	12.0	6.2		
Human transcribers					11.3	5.9

- LSTM-LM gives 15-20% gain over N-gram LM
- BLSTM combination alone is almost as good as the best system!
- System combination 12% relative gain over best single subsystem
- Overall, we edge just past measured human error on the same dataset

Senone-level acoustic model combination (not used in combined system)

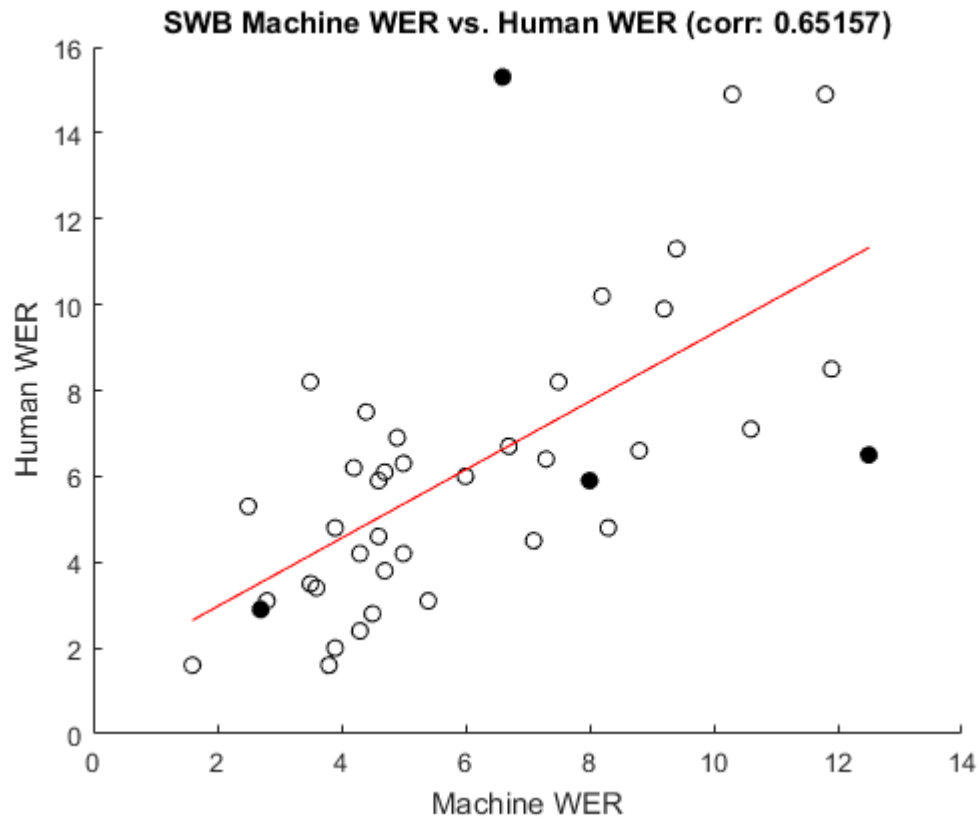
Human/Machine: Analysis

How do human and machine transcripts differ?

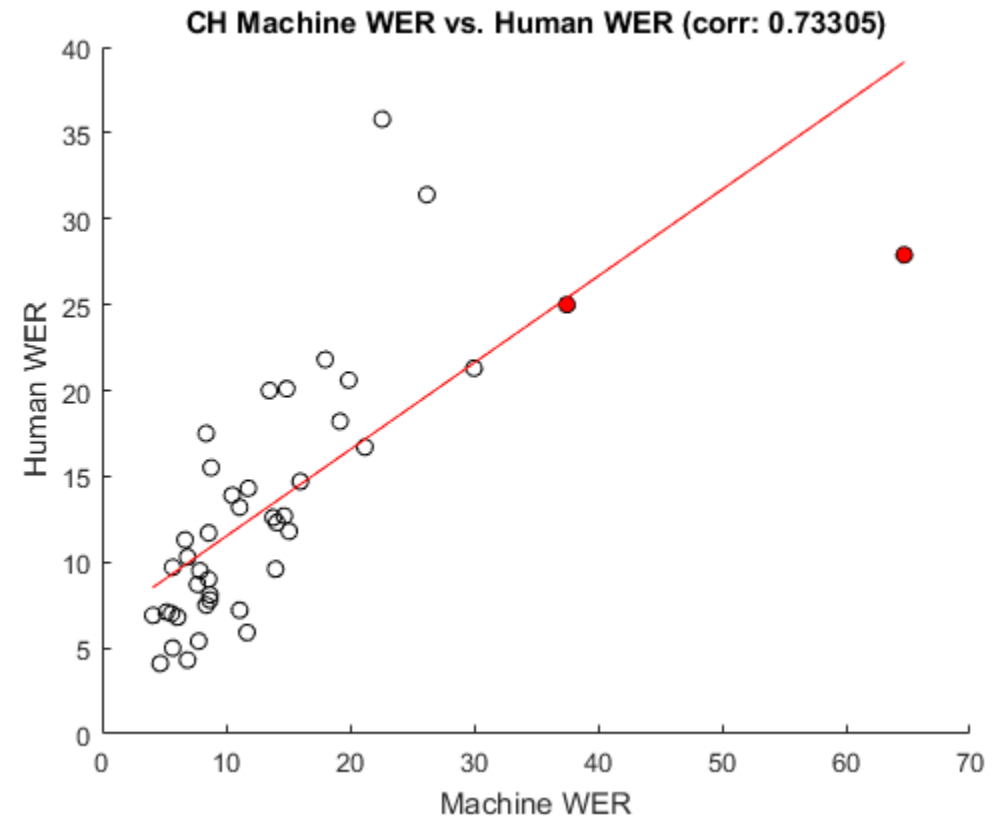
- Transcripts are very close quantitatively, by overall WER
- Research questions:
 - What makes transcription easy or hard for human vs. machine?
 - Does the machine make errors that are *qualitatively* different from humans?
 - Can humans tell the difference?

Error Correlation by Speaker

Each data point is a conversation side, $N = 40$



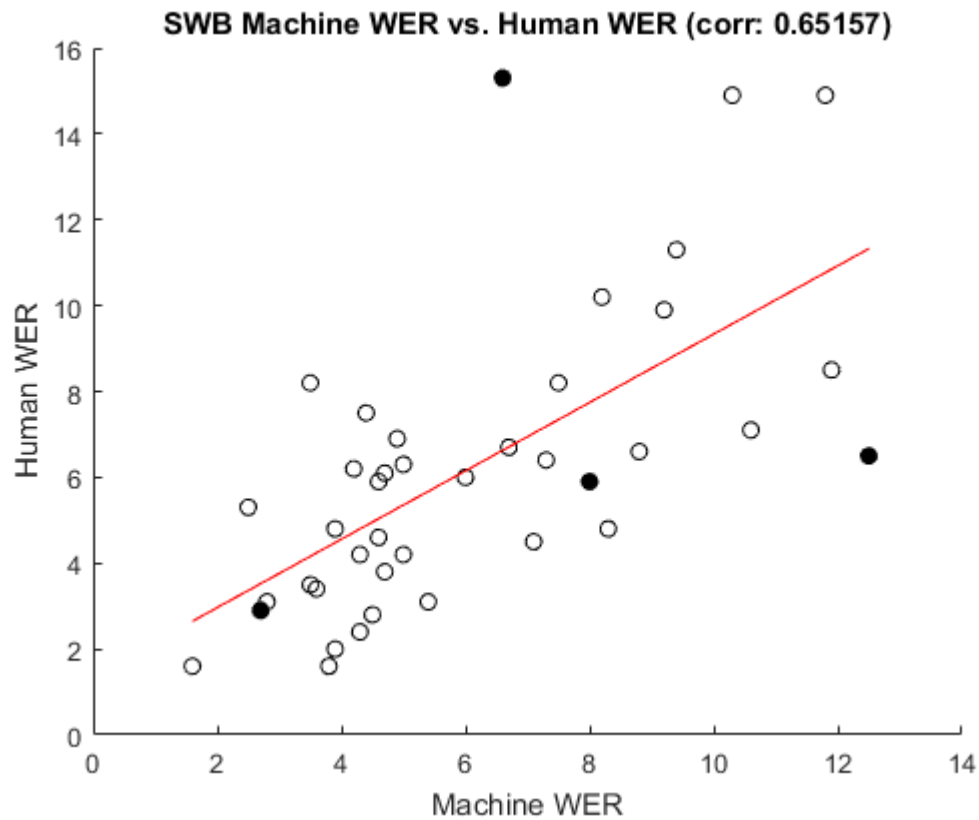
$$\rho = 0.65$$



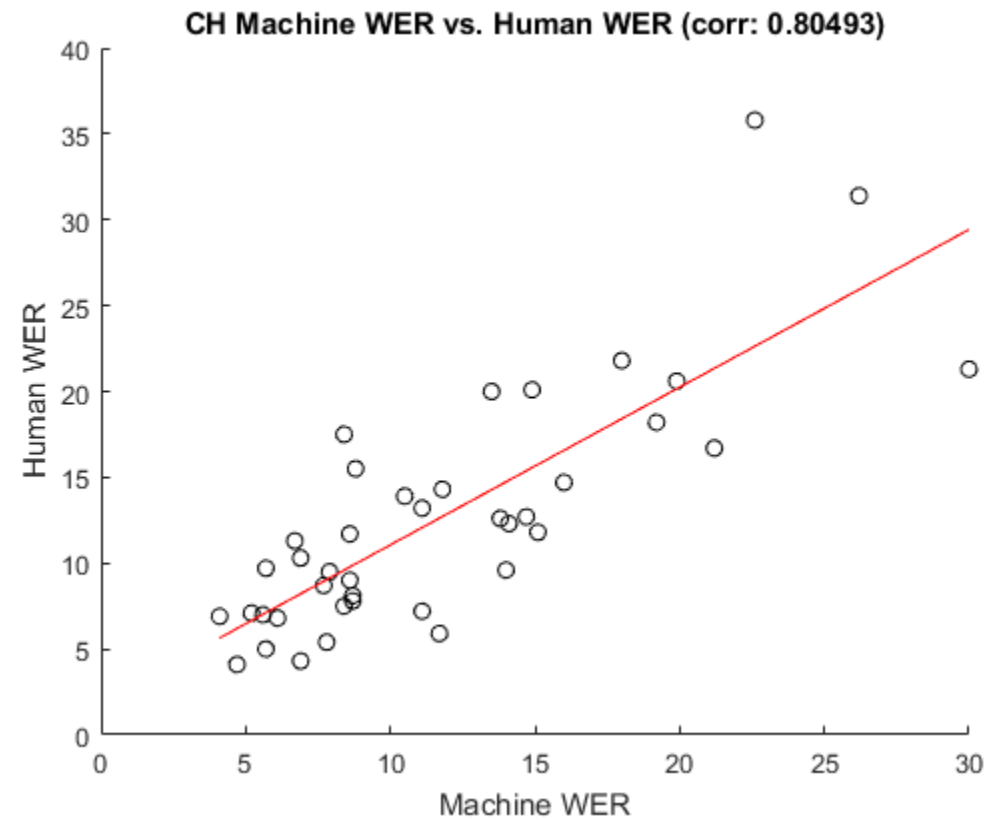
$$\rho = 0.73$$

Error Correlation (without outliers)

Two CallHome conversations have multiple speakers on the same side, resulting in very high WER!



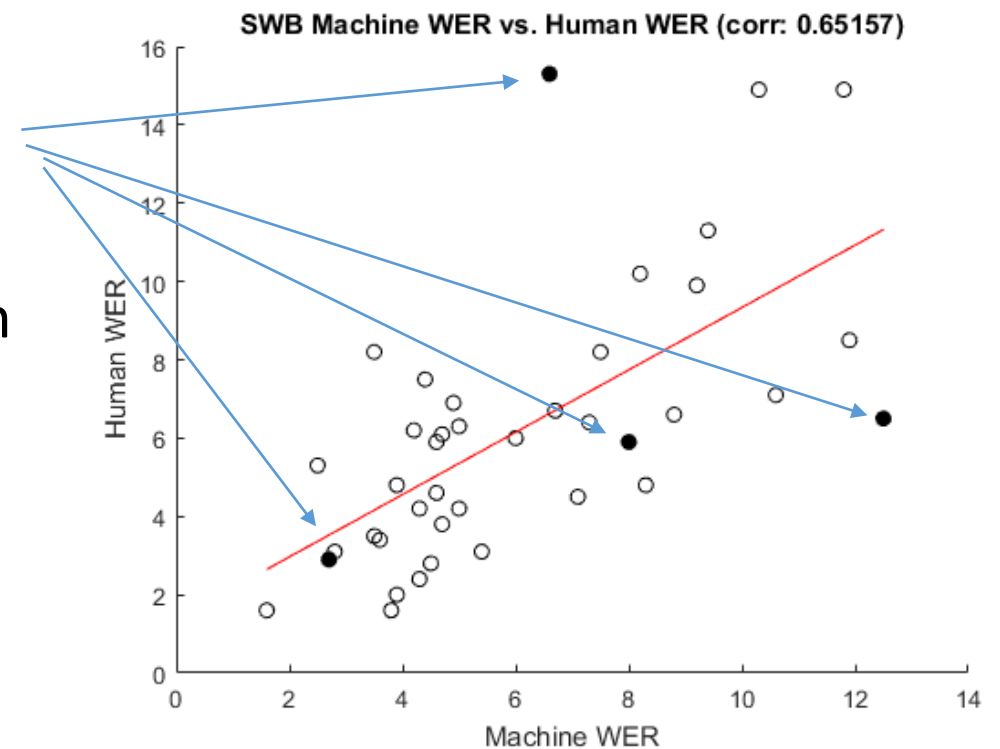
$$\rho = 0.65$$



$$\rho = 0.80$$

Seen vs. Unseen Switchboard Speakers

- It has been suggested that the 2000 Switchboard test set is so “easy” because most of the speakers also occur in the training set (a NIST blunder!)
- The filled dots are the *unseen* speakers
- This doesn't seem to be the case:
 - Machine WER on unseen speakers is within the normal range
 - For the most part (3 of 4), machine WER predicts the human WER



Qualitative differences: Top Error Types

Substitutions (\approx 21k words in each test set)

CH		SWB	
ASR	Human	ASR	Human
45: (%hesitation) / %bcack	12: a / the	29: (%hesitation) / %bcack	12: (%hesitation) / hmm
12: was / is	10: (%hesitation) / a	9: (%hesitation) / oh	10: (%hesitation) / oh
9: (%hesitation) / a	10: was / is	9: was / is	9: was / is
8: (%hesitation) / oh	7: (%hesitation) / hmm	8: and / in	8: (%hesitation) / a
8: a / the	7: bentsy / bensi	6: (%hesitation) / i	5: in / and
7: and / in	7: is / was	6: in / and	4: (%hesitation) / %bcack
7: it / that	6: could / can	5: (%hesitation) / a	4: and / in
6: in / and	6: well / oh	5: (%hesitation) / yeah	4: is / was

Overall similar patterns: short function words get confused

One outlier: machine falsely recognizes backchannel “uh-huh” for filled pause “uh”

- These words are acoustically confusable, have opposite pragmatic functions in conversation
- Humans can disambiguate by prosody and context

Top Insertion and Deletion Errors

Deletions

CH		SWB	
ASR	Human	ASR	Human
44: i	73: i	31: it	34: i
33: it	59: and	26: i	30: and
29: a	48: it	19: a	29: it
29: and	47: is	17: that	22: a
25: is	45: the	15: you	22: that
19: he	41: %bcack	13: and	22: you
18: are	37: a	12: have	17: the
17: oh	33: you	12: oh	17: to

Insertions

CH		SWB	
ASR	Human	ASR	Human
15: a	10: i	19: i	12: i
15: is	9: and	9: and	11: and
11: i	8: a	7: of	9: you
11: the	8: that	6: do	8: is
11: you	8: the	6: is	6: they
9: it	7: have	5: but	5: do
7: oh	5: you	5: yeah	5: have
6: and	4: are	4: air	5: it

Both humans and machines insert “I” and “and” a lot.
Short function words dominate the list for both.

“Spot the Bot”

- Can people tell which transcripts are by machine?
- We ran an informal experiment at the last ICASSP conference
- Inspired by Turing test



Which transcription was created by a human?

Choice One (Click to Select)	Reference Transcription (Click to Play Audio)	Choice Two (Click to Select)
it seems like you know <u>we</u> need furniture <u>then</u> you know <u>a</u> bedroom <u>SUIT</u> then we need to budget it	it seems like you know (u-) (if) we need furniture then you know a bedroom suite then we need to budget it	it seems like you know <u>OH WHEN</u> need furniture <u>AND</u> you know <u>[]</u> bedroom <u>SUITS</u> then we need to budget it

CORRECT 1 : 0 INCORRECT

Reset Score

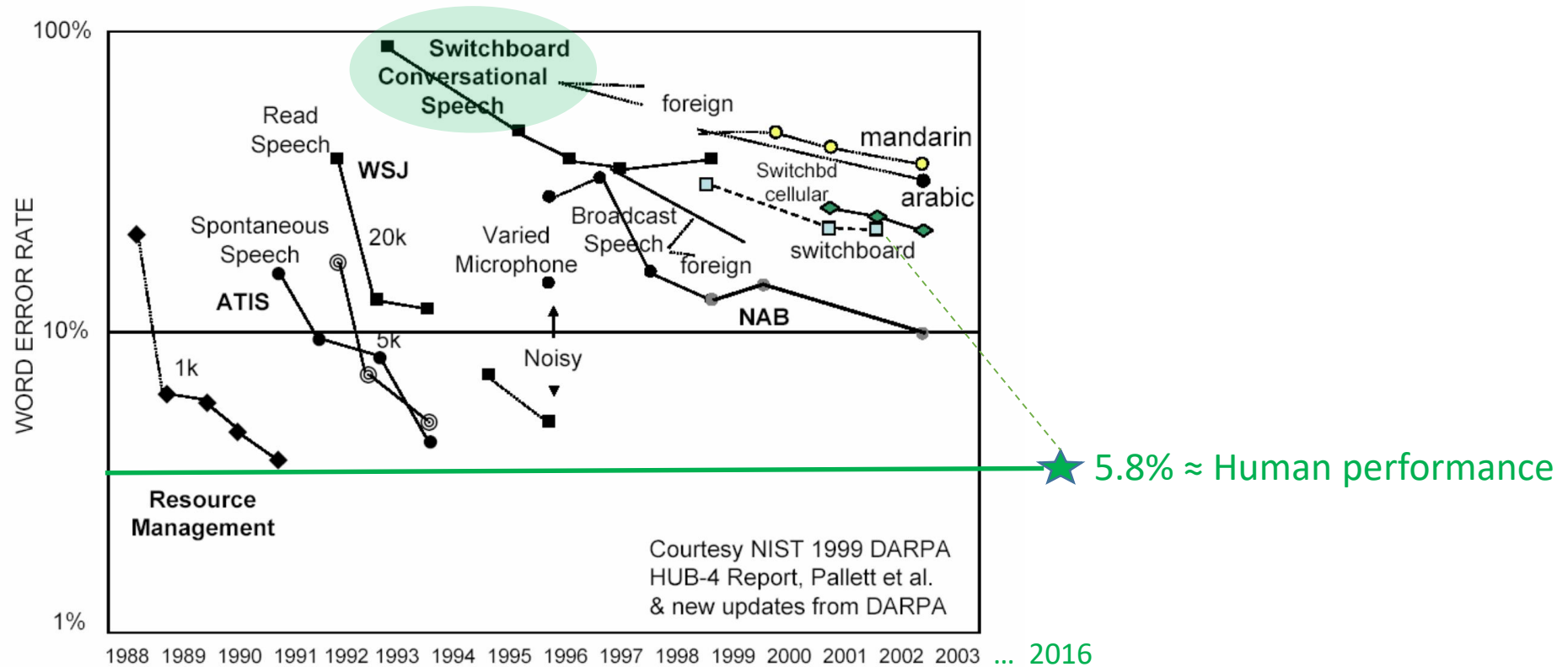
Experiment: Informal results

- Subjects guessed correctly 188 / 353 times (53% accuracy)
- Not different from chance ($p \approx 0.12$, one-tailed)
- Obviously, this was not a rigorous experiment ...
- ... but it gave us a first-hand idea of how difficult it is to tell human from machine transcription

Wrap-up

We've come a long way

DARPA Speech Recognition Benchmark Tests



Conclusions

- Human transcription performance is around 5-6%, but also varies greatly with the function of the amount of effort!
 - Multiple independent transcription passes with reconciliation would lower this further, as done by NIST for their reference transcriptions
- State-of-the-art ASR technology based on neural net acoustic and language models has reached commercial-level accuracy
- Humans and machine transcription performance is highly correlated
 - “Hard” versus “easy” speakers
 - Word types involved in most frequent errors
 - Humans are better at recognizing pragmatically relevant words (“uh” vs. “uh-huh”)

Where to go from here

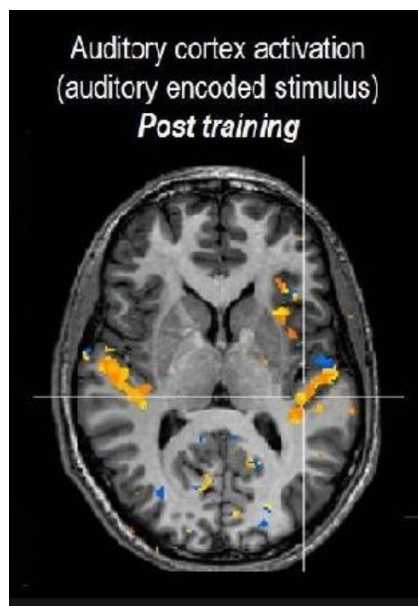
- Pick harder tasks!
- Current focus (again!) = Meeting speech
 - Multiple speakers
 - Overlapping speech
 - Distant microphone capture (background noise, reverberation)

Thank You!

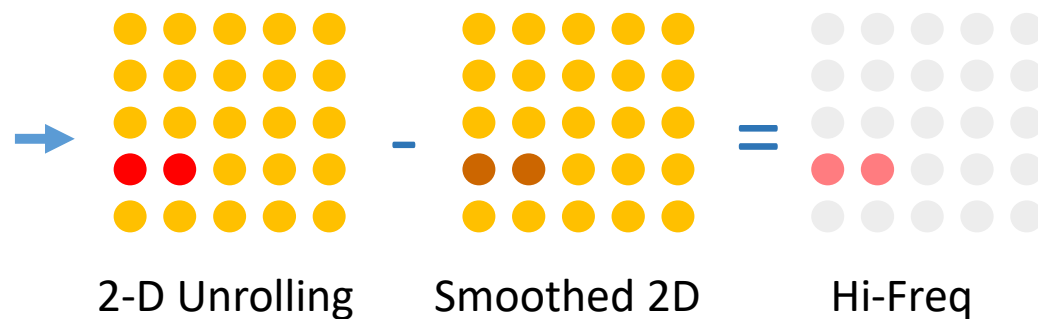
<http://www.microsoft.com/en-us/research/project/human-parity-speech-recognition/>

More Technical Details

BLSTM Spatial Regularization



Regularize with L2 norm of Hi-frequency residual



[Droppo, Interspeech 2017]

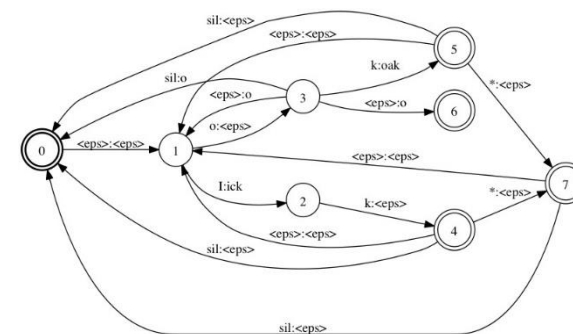
Senones	CallHome WER (%)		SWB WER (%)	
	Baseline	Smoothing	Baseline	Smoothing
9000	21.4	19.2	9.9	9.3
27000	20.5	19.5	10.6	9.2

MMI Denominator GPU computation

- Represent FSA of all possible state sequences as a sparse transition matrix \mathbf{A}
- Implement exact alpha beta computations

$$\alpha_t = (\mathbf{A} \alpha_{t-1}) \cdot o_t$$
$$\beta_t = \mathbf{A}^T (\beta_{t+1} \cdot o_{t+1})$$

- Execute in straight “for” loops on GPU with **cusparseDcsmv** and **cublasDdgm**
- Beautifully simple



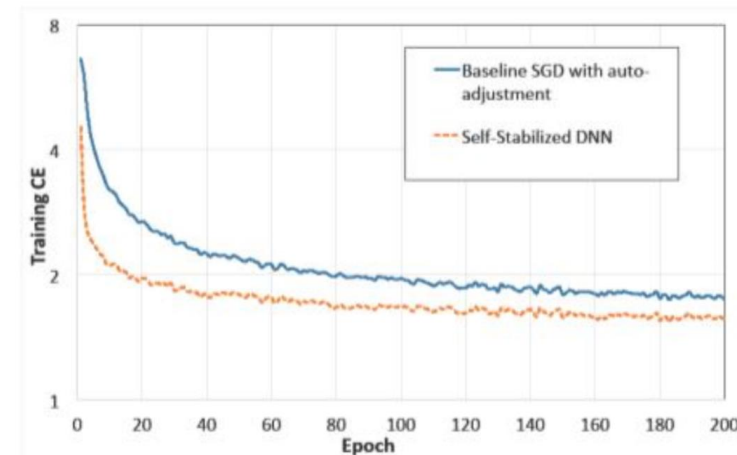
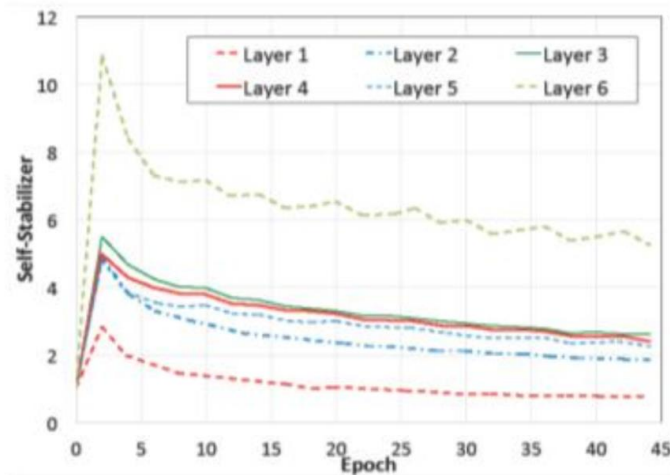
LM Training Trick: Self-stabilization

- Learn an overall scaling function for each layer

$\mathbf{y} = \mathbf{W}\mathbf{x}$ becomes:

$$\mathbf{y} = (\beta \mathbf{W})\mathbf{x}$$

Applied to the LSTM networks, between layers.



Language Model Perplexities

Language model	PPL
Ngram: 4gram baseline (145M ngrams)	75.5
RNN: 2 layers + word input	59.8
LSTM: word input in forward direction	54.4
LSTM: word input in backward direction	53.4
LSTM: letter trigram input in forward direction	52.1
LSTM: letter trigram input in backward direction	52.0

LSTM beats RNN

Letter trigram input slightly better than word input

Note both forward and backward running models

Perplexities on the 1997 eval set