Man vs. Machine in Conversational Speech Recognition

George Saon
IBM Research AI

Deep Blue vs. Garry Kasparov, 1997

AlphaGo vs. Lee Sedol, 2016

Watson vs. Jennings and Rutter, 2011

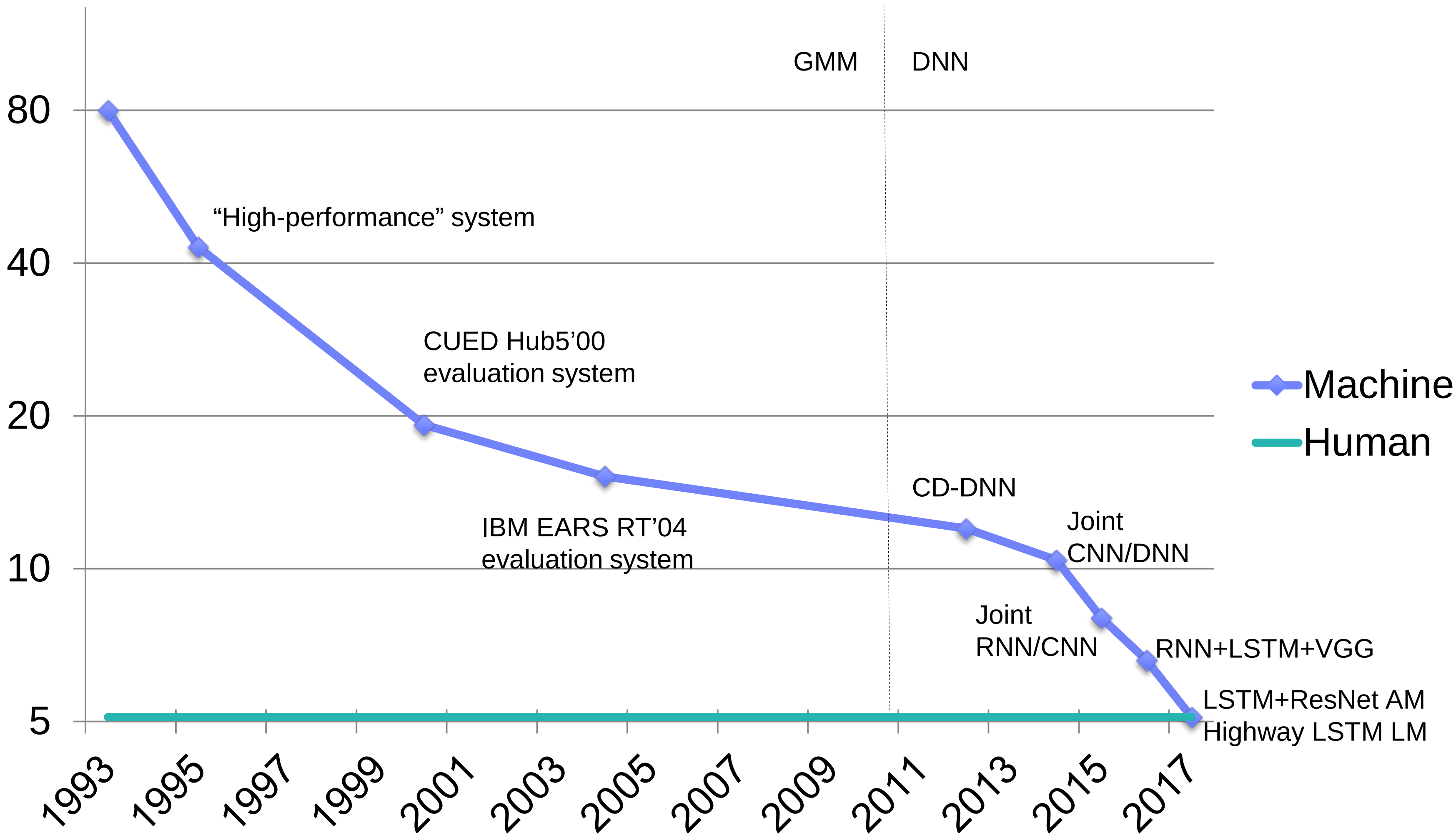
Switchboard and CallHome corpora

- **Switchboard:**
  - Conversations between strangers on a preassigned topic:
  - Each call is roughly 5min in length
  - 2000 hours of training data (300h Switchboard + 1700h Fisher)
  - Representative sample of American English speech in terms of gender, race, location and channel
  - Challenges due to mistakes, repetitions, repairs and other disfluencies
- **CallHome:**
  - Conversations between friends and family with no predefined topic:
  - 18 hours of training data

Why Switchboard?

- **Popular benchmark in the speech recognition community**
- **Largest public corpus of conversational speech (2000 hours)**
- **Has been studied for 25 years**
- **NIST evaluations under the DARPA Hub5 and EARS programs**
  - Companies: AT&T, BBN, IBM, SRI
  - Universities: Aachen, Cambridge, CMU, ICSI, Karlsruhe, LIMSI, MSU

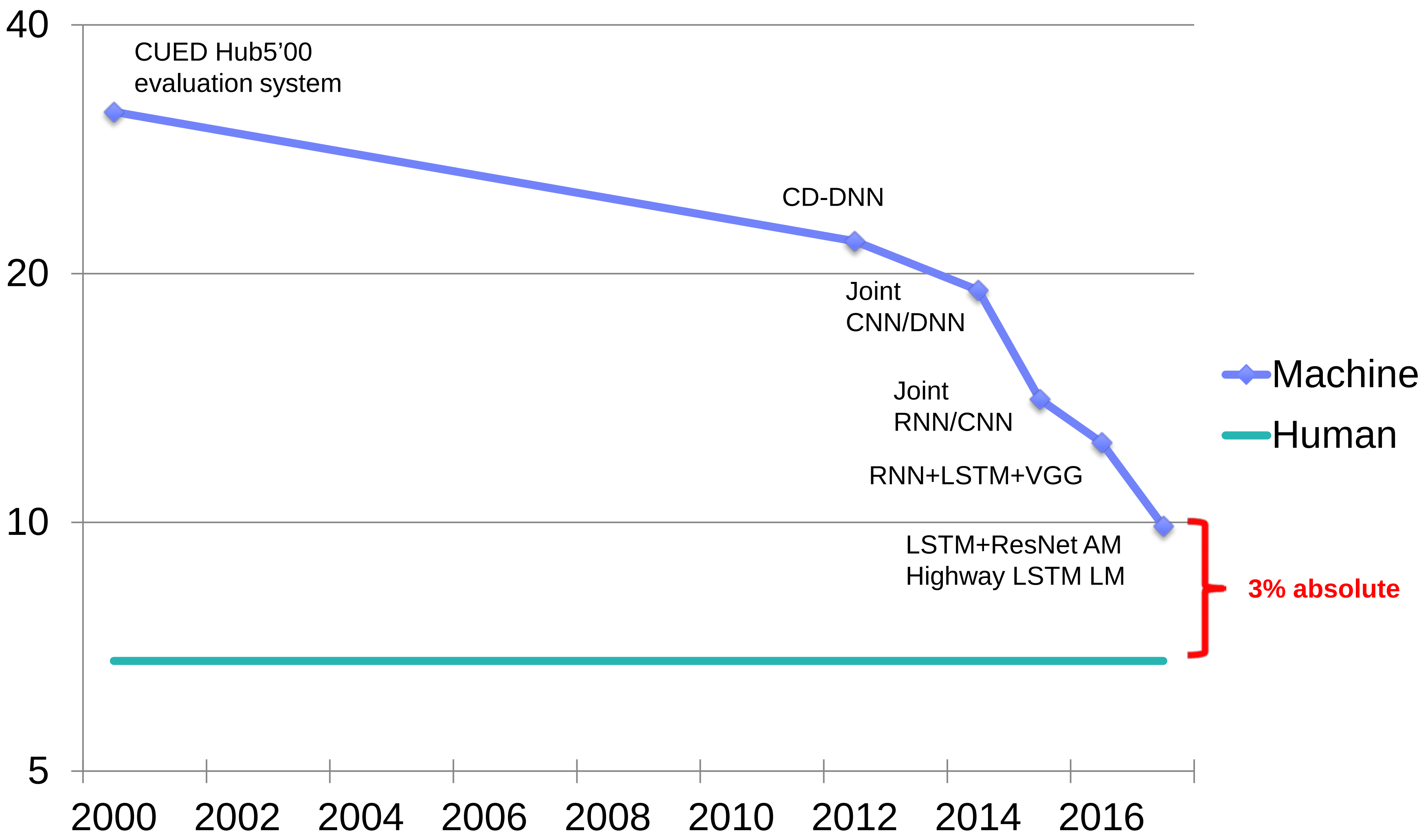
Progress on Switchboard (Hub5'00 SWB testset*)



*Except for 1993,1995,2004

Is conversational speech recognition solved?

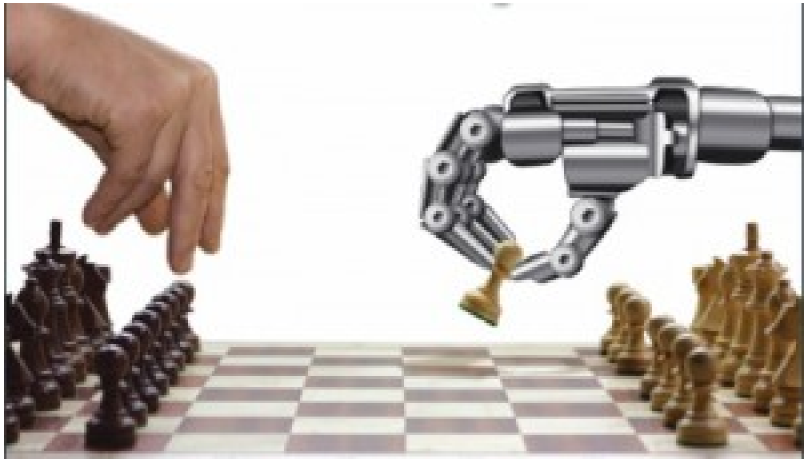
Progress on CallHome (Hub5'00 CH testset)



SWITCHBOARD WARS

	Hub5'00 SWB	Hub5'00 CH
IBM Interspeech'15	8.0	14.1
STC Interspeech'16	7.8	--
IBM Interspeech'16	6.6	12.2
MSR ArXiv'16 (a)	6.2	12.0
MSR ArXiv'16 (b)	5.8	11.0
BBN Interspeech'17	6.1	10.4
IBM Interspeech'17	5.5	10.3
Capio.ai Interspeech'17	5.3*	10.1*
MSR ArXiv'17	5.1	--
IBM ASRU'17	5.1	9.9

IBM Switchboard ASR systems 2015 - 2017



2015 system

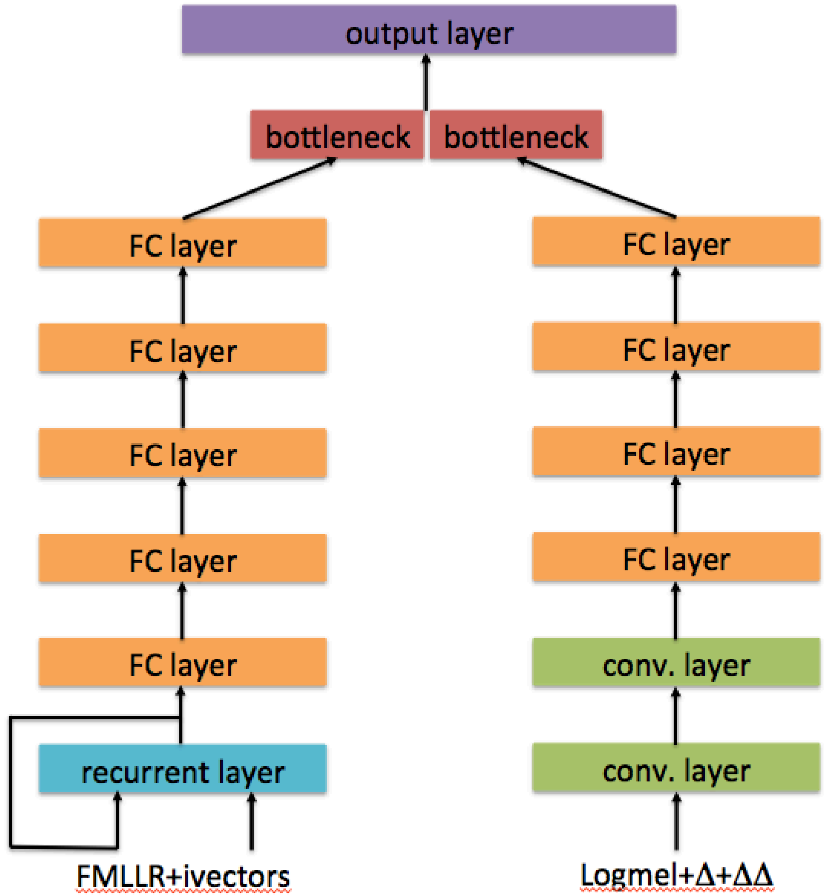
- **Key ingredients:**
  - AM: joint RNN/CNN
  - LM: model "M" + NN

- **Results:**

Model	Hub5'00 SWB	Hub5'00 CH
CNN	10.4	17.9
RNN	9.9	16.3
Joint RNN/CNN	9.3	15.6
+ LM rescoring	8.0	14.1

G. Saon, H. Kuo, S. Rennie, M. Picheny, "The IBM 2015 English conversational telephone speech recognition system", Interspeech 2015.

Joint RNN/CNN



H. Soltau, G. Saon, T. Sainath, "Joint training of convolutional and non-convolutional neural networks", ICASSP 2014.
T. N. Sainath, A.-r. Mohamed, B. Kingsbury, B. Ramabhadran, "Deep convolutional neural networks for LVCSR", ICASSP 2013.

2016 system

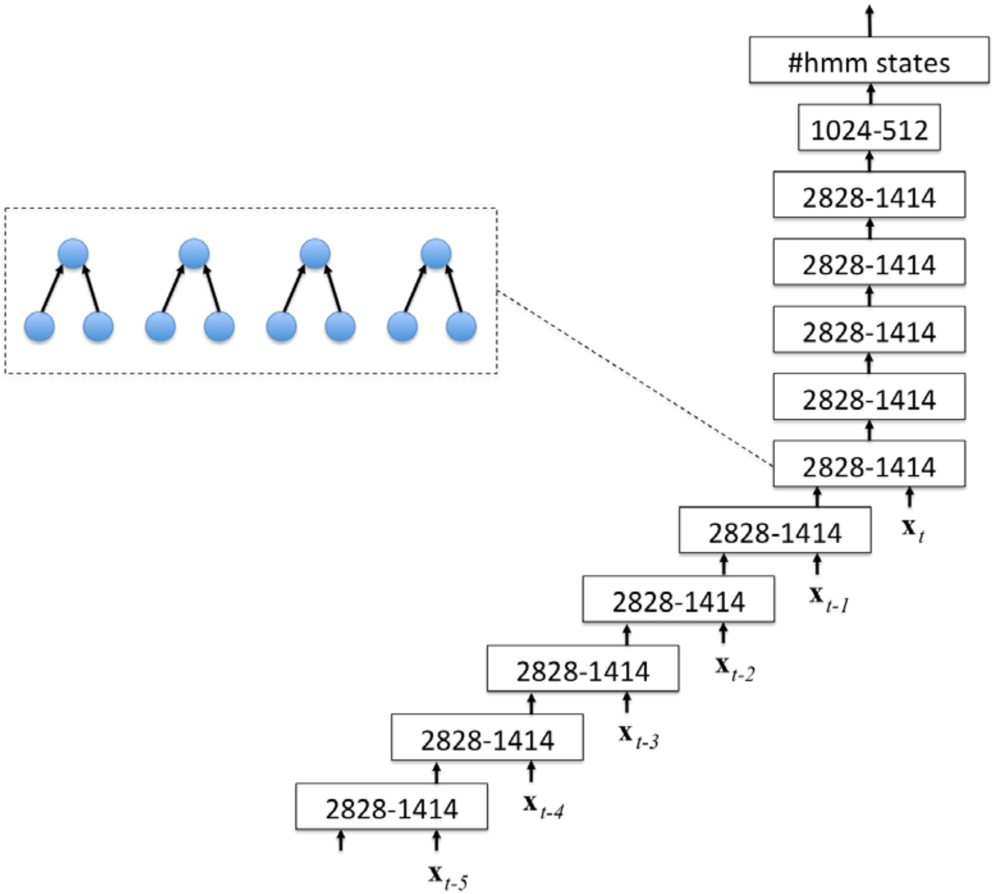
- **Key ingredients:**
  - AM: RNN Maxout + LSTM + VGG
  - LM: same as 2015 (vocab. increase)

- **Results:**

Model	Hub5'00 SWB	Hub5'00 CH
RNN	9.3	15.4
VGG	9.4	15.7
LSTM	9.0	15.1
RNN+VGG+LSTM	8.6	14.4
+ LM rescoring	6.6	12.2

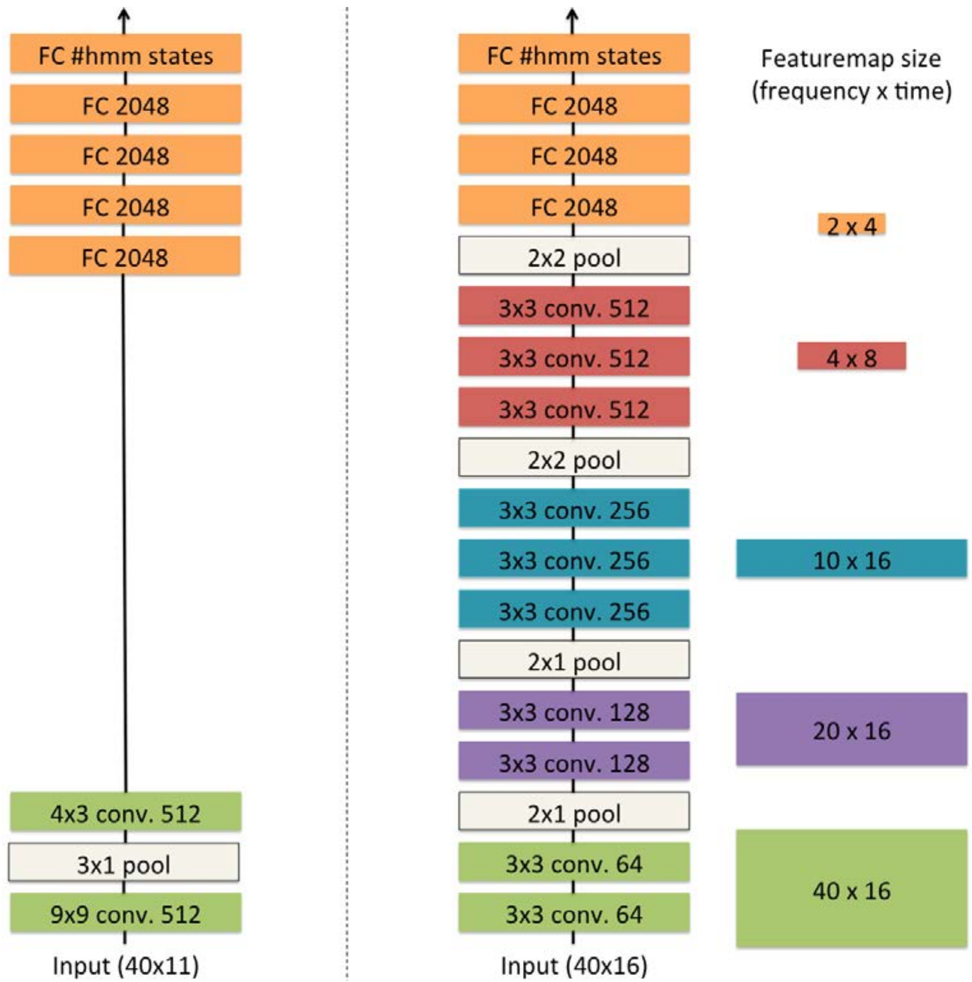
G. Saon, H. Kuo, S. Rennie, M. Picheny, "The IBM 2016 English conversational telephone speech recognition system", Interspeech 2016.

Maxout RNN with annealed dropout



I. Goodfellow, D. Ward-Farley, M. Mirza, A. Courville, Y. Bengio, "Maxout networks", arXiv 2013.
S. Rennie, V. Goel, S. Thomas, "Annealed dropout training of deep networks", SLT 2014.

Very deep CNNs (VGG nets)



K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv 2014.
T. Sercu, V. Goel, "Advances in very deep convolutional networks for LVCSR", arXiv 2016.

2017 system (as of Interspeech)

- **Key ingredients:**
  - AM: LSTM + ResNet
  - LM: model “M” + LSTM + WaveNet

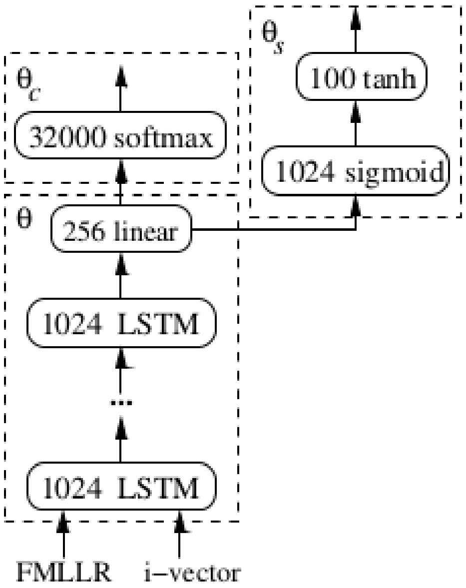
- **Results:**

Model	Hub5’00 SWB	Hub5’00 CH
LSTM	7.2	12.7
ResNet	7.6	14.5
LSTM+ResNet	6.7	12.1
+ LM rescoring	5.5	10.3

G. Saon et al., “English conversational telephone speech recognition by humans and machines”, Interspeech 2017

Speaker-adversarial training for LSTMs

- **Predict i-vectors and subtract gradient component**



$$
\begin{aligned}
\hat{\theta}_c &= \theta_c - \epsilon \frac{\partial \mathcal{L}_{CE}(\mathbf{x})}{\partial \theta_c} \\
\hat{\theta}_s &= \theta_s - \epsilon \frac{\partial \mathcal{L}_{MSE}(\mathbf{x})}{\partial \theta_s} \\
\hat{\theta} &= \theta - \epsilon \left(\frac{\partial \mathcal{L}_{CE}(\mathbf{x})}{\partial \theta} - \lambda \frac{\partial \mathcal{L}_{MSE}(\mathbf{x})}{\partial \theta} \right)
\end{aligned}
$$

- **Results:**

Model	Hub5'00 SWB	Hub5'00 CH
Baseline	7.7	13.8
SA-MTL	7.6	13.6

Y. Ganin et al., "Domain-adversarial training of neural networks", arXiv 2015.

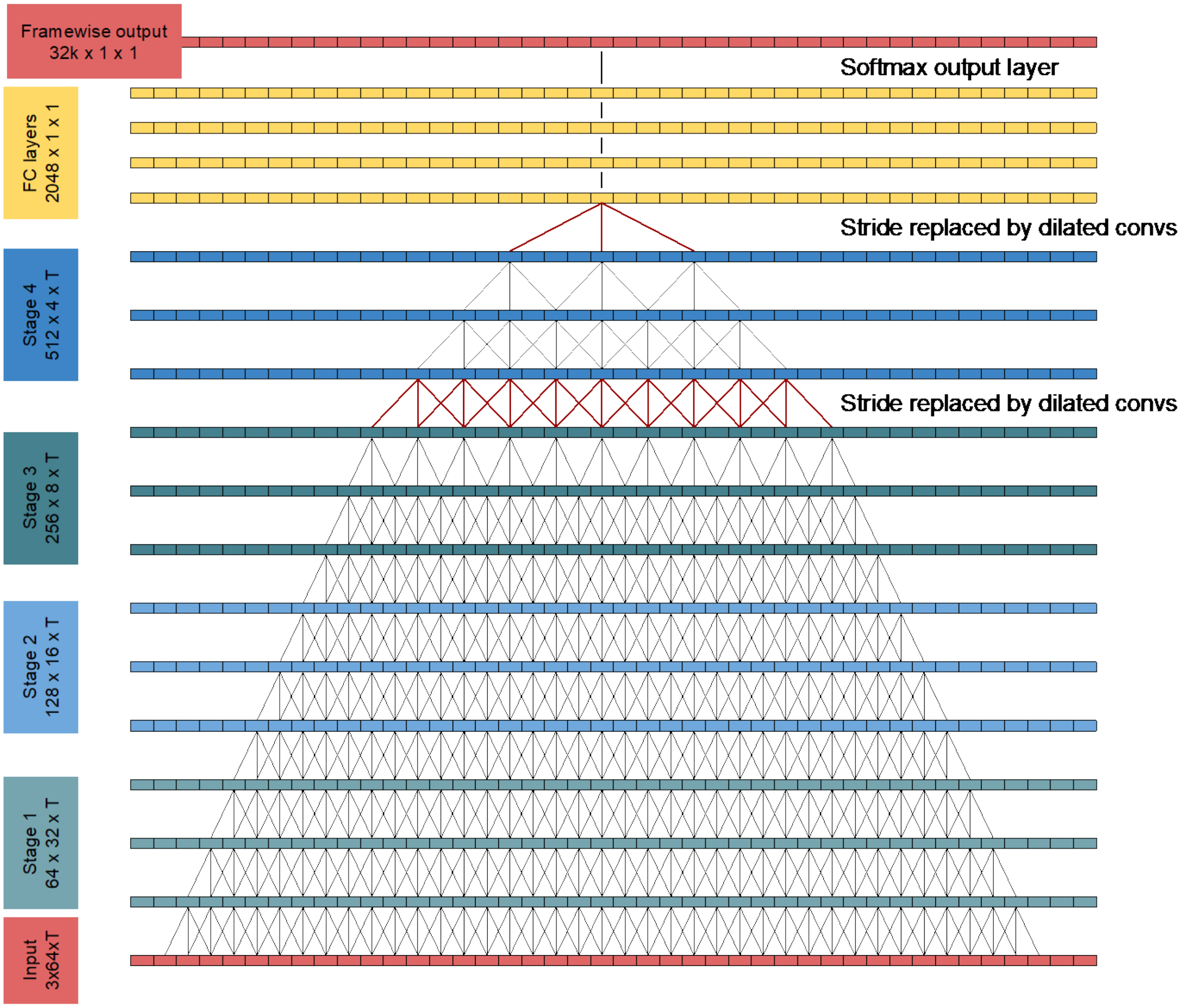
Feature fusion for LSTMs

- **Train bidirectional LSTMs on 3 feature streams:**
  - 40-dimensional FMLLR
  - 100-dimensional i-vectors
  - 120-dimensional Logmel + Δ + $\Delta\Delta$

- **Results:**

Model	Hub5'00 SWB	Hub5'00 CH
Baseline (FMLLR+ivecs)	7.7	13.8
Fusion	7.2	12.7

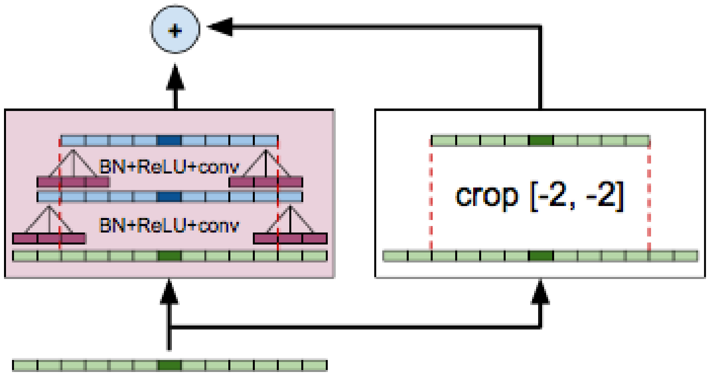
ResNets



K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition”, arXiv 2015.
T. Sercu, V. Goel, “Dense prediction on sequences with time-dilated convolutions for speech recognition”, arXiv 2016.

ResNets

- **Residual blocks with identity shortcut connections**



- **Results:**

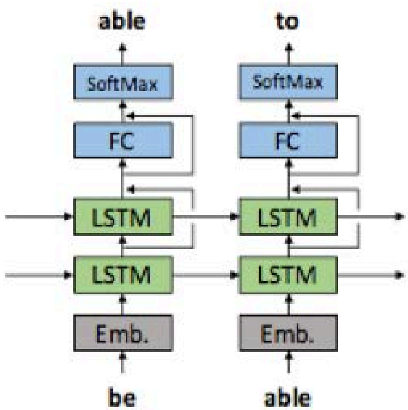
Model	Hub5'00 SWB	Hub5'00 CH
LSTM	7.2	12.7
ResNet	7.6	14.5
LSTM+ResNet	6.7	12.1

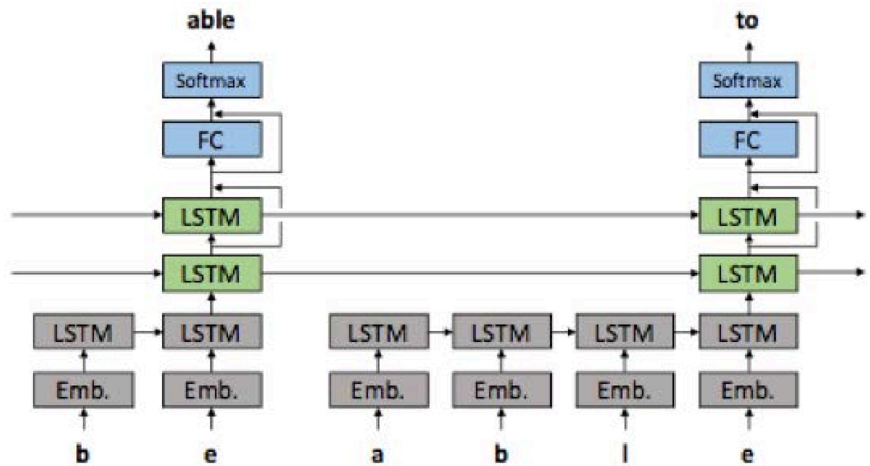
Other AM techniques

- **Speaker adaptation:**
  - Feature normalization: per-speaker CMVN, VTLN [Lee'96], FMLLR [Gales'97]
  - I-vectors [Dehak'11] as auxiliary inputs [Saon'13]
- **Architecture:**
  - Large output layer (32000 CD HMM states)
  - Bottleneck layer [Sainath'13]
- **CE training:**
  - Minibatch SGD with frame randomization [Seide'11]
  - Balanced sampling training [Sercu'16]
  - LSTM training for hybrid models [Sak'15, Mohamed'15]
- **Sequence discriminative training:**
  - Objective: sMBR [Gibson'06] or boosted MMI [Povey'08]
  - Optimization: Hessian-free [Kingsbury'12] or SGD with CE smoothing [Su'13]

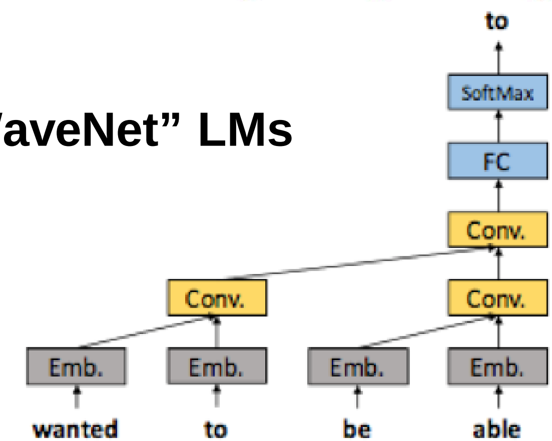
Language modeling (Interspeech'17)

- **Word and character LSTMs**





- **Convolutional "WaveNet" LMs**



G. Kurata et al., "Empirical exploration of LSTM and CNN language models for speech recognition", Interspeech 2017.

Language modeling (ASRU'17)

- **Highway LSTMs: add carry and transform gates to the memory cells and hidden states**

$$
\begin{aligned}
g_T &= \mathrm{sigm}(W_T x + b_T) \\
g_C &= \mathrm{sigm}(W_C x + b_C) \\
y &= x \odot g_C + \tanh(W x + b) \odot g_T
\end{aligned}
$$

- **Unsupervised LM adaptation:**
  - Reestimate interpolation weights between component LMs based on rescored output
  - Use each testset as a heldout set

R. Srivastava, K. Greff, J. Schmidhuber, "Highway networks", arXiv 2015.
G. Kurata, B. Ramabhadran, G. Saon, A. Sethy, "Language modeling with highway LSTM", ASRU 2017.

Testsets

Testset	Duration	Nb. speakers	Nb. words
Hub5'00 SWB	2.1h	40	21.4K
Hub5'00 CH	1.6h	40	21.6K
RT'02	6.4h	120	64.0K
RT'03	7.2h	144	76.0K
RT'04	3.4h	72	36.7K

LM rescoring results (full and simplified system)

- **Full system:**

	Hub5'00 SWB	Hub5'00 CH	RT'02	RT'03	RT'04
n-gram	6.7	12.1	10.1	10.0	9.7
+ model M	6.1	11.2	9.4	9.4	9.0
+ LSTM+DCC	5.5	10.3	8.3	8.3	8.0
+ Highway LSTM	5.2	10.0	8.1	8.1	7.8
+ Unsup. adaptation	5.1	9.9	8.2	8.1	7.7

- **Simplified system 1 AM + 1 rescoring LM:**

n-gram	7.2	12.7	10.7	10.2	10.1
+ LSTM	6.1	11.1	9.0	8.8	8.5

Human speech recognition experiments

Issues in measuring human speech recognition performance

- **References are created by humans**
  - No absolute gold standard, inherent ambiguity
  - Measure inter-annotator agreement
- **No “world champions” for speech transcription**
  - Verbatim transcription is not a natural task for humans
  - Use experts who do this for a living
- **Multiple estimates of human WER for the same testset**
  - Depends on transcriber selection and transcription procedure

Transcription of Switchboard testsets (done by Appen)

- 3 independent transcribers quality checked by a 4th senior transcriber
- Native US speakers selected based on quality of previous work
- Transcribers familiarized with LDC transcription protocol
- Utterances are processed in sequence, just like ASR system
- Transcription time: 12-13xRT for first pass, 1.7-2xRT for second pass

Human WERs on Hub5'00 SWB and CH

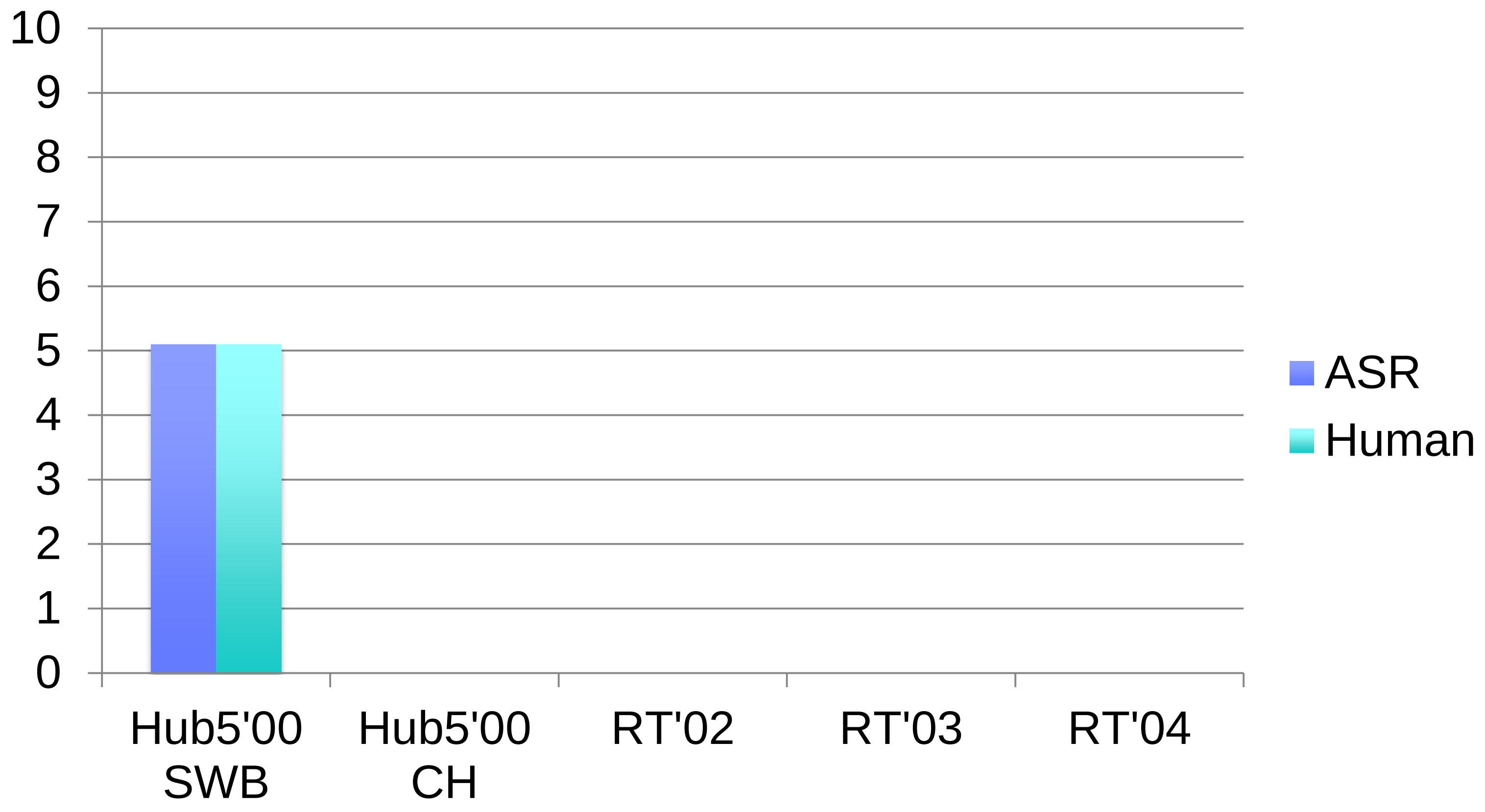
	Hub5'00 SWB	Hub5'00 CH
Transcriber 1 raw	6.1	8.7
Transcriber 1 QC	5.6	7.8
Transcriber 2 raw	5.3	6.9
Transcriber 2 QC	**5.1**	**6.8**
Transcriber 3 raw	5.7	8.0
Transcriber 3 QC	5.2	7.6
Human estimate by MSR*	5.9	11.3

*Xiong et al. "Achieving Human Parity in Conversational Speech Recognition", arXiv 2016.

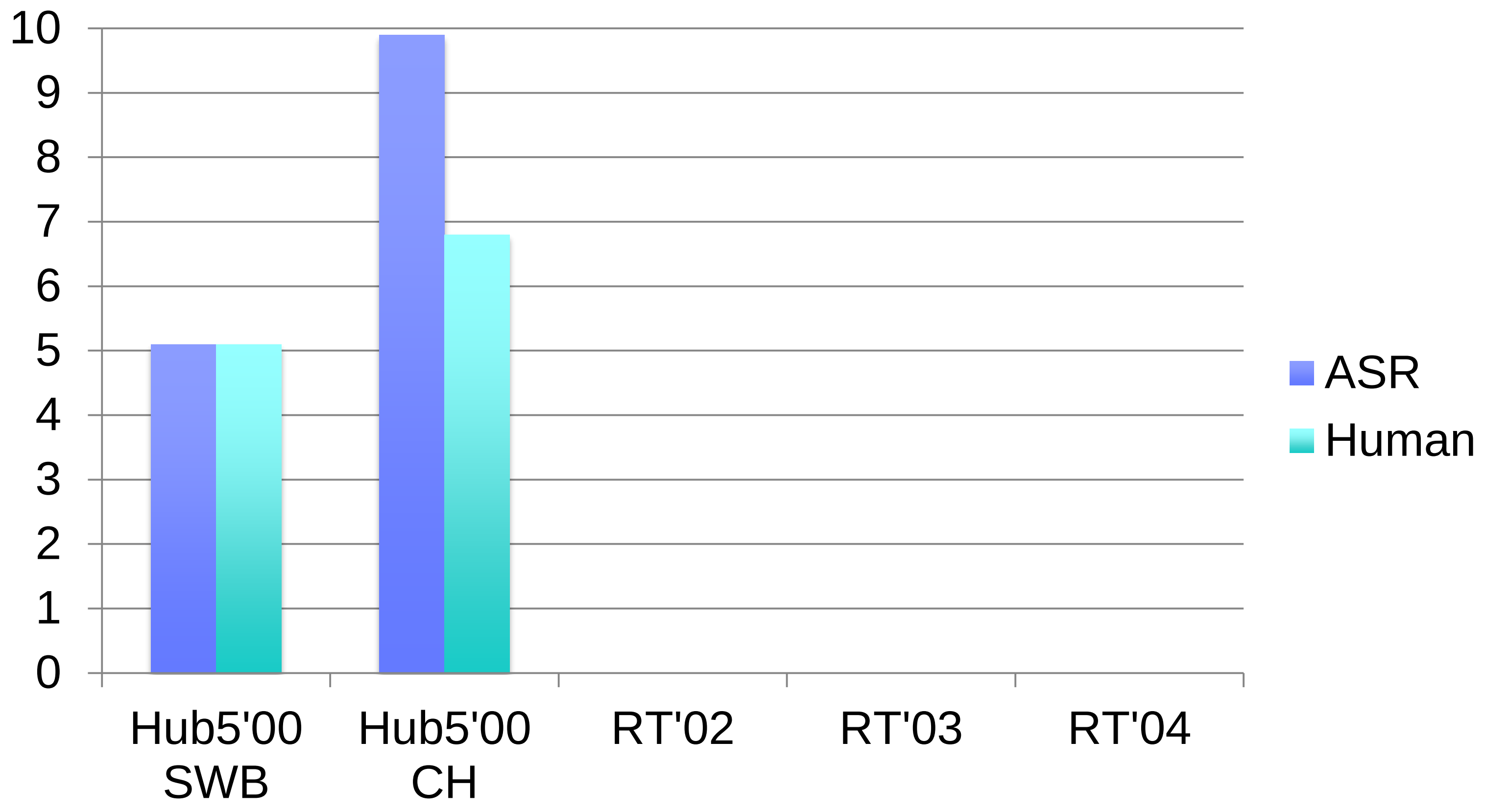
Inter-annotator agreement

Ref	Test	SWB	CH
T1	T2	6.8	9.2
T1	T3	7.0	9.4
T2	T3	6.3	8.3
T1QC	T2QC	6.0	8.1
T1QC	T3QC	6.0	8.1
T2QC	T3QC	5.6	7.8
LDC	T1QC	5.6	7.8
LDC	T2QC	5.1	6.8
LDC	T3QC	5.2	7.6

Man vs. machine: Hub5'00 SWB



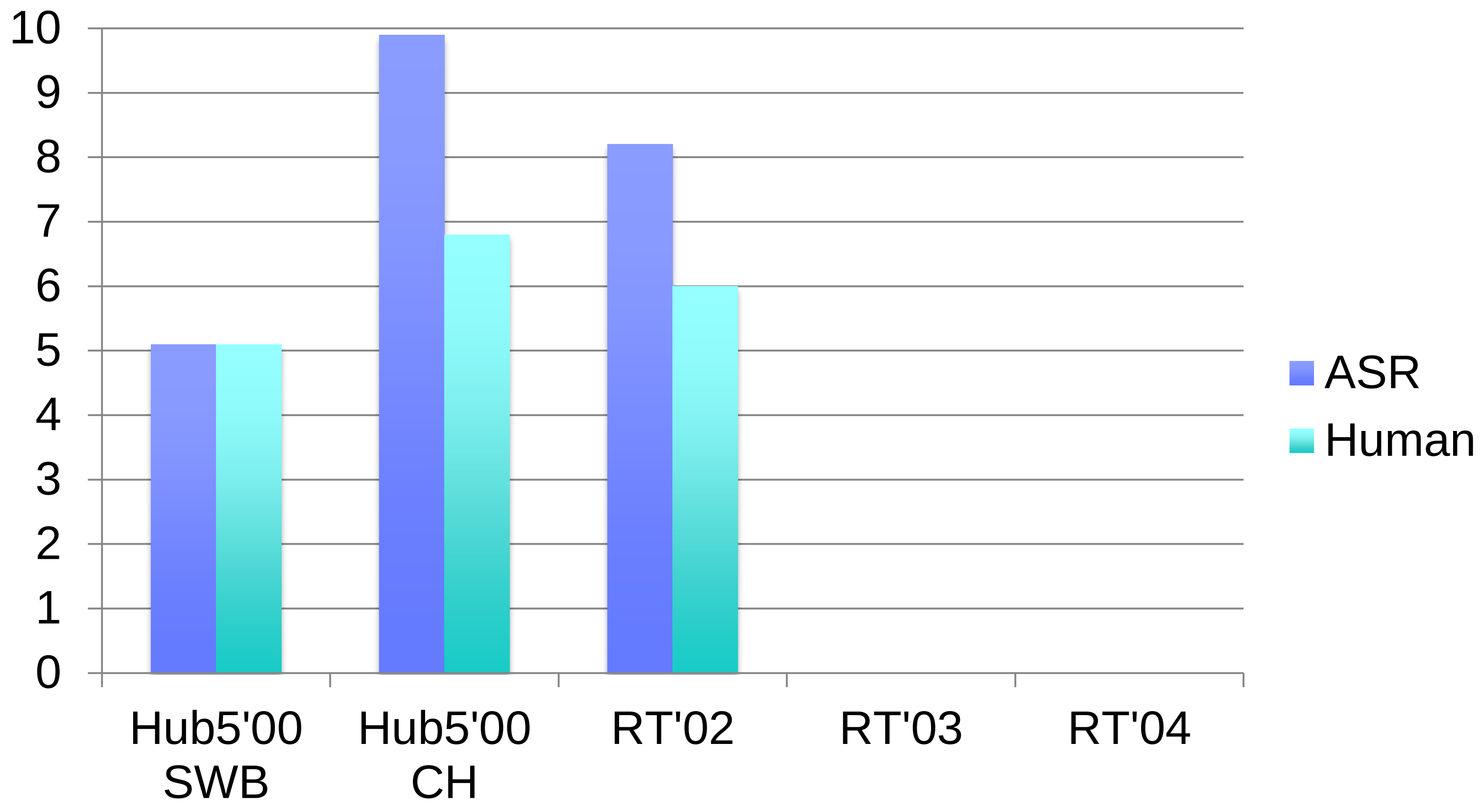
Man vs. machine: Hub5’00 CH



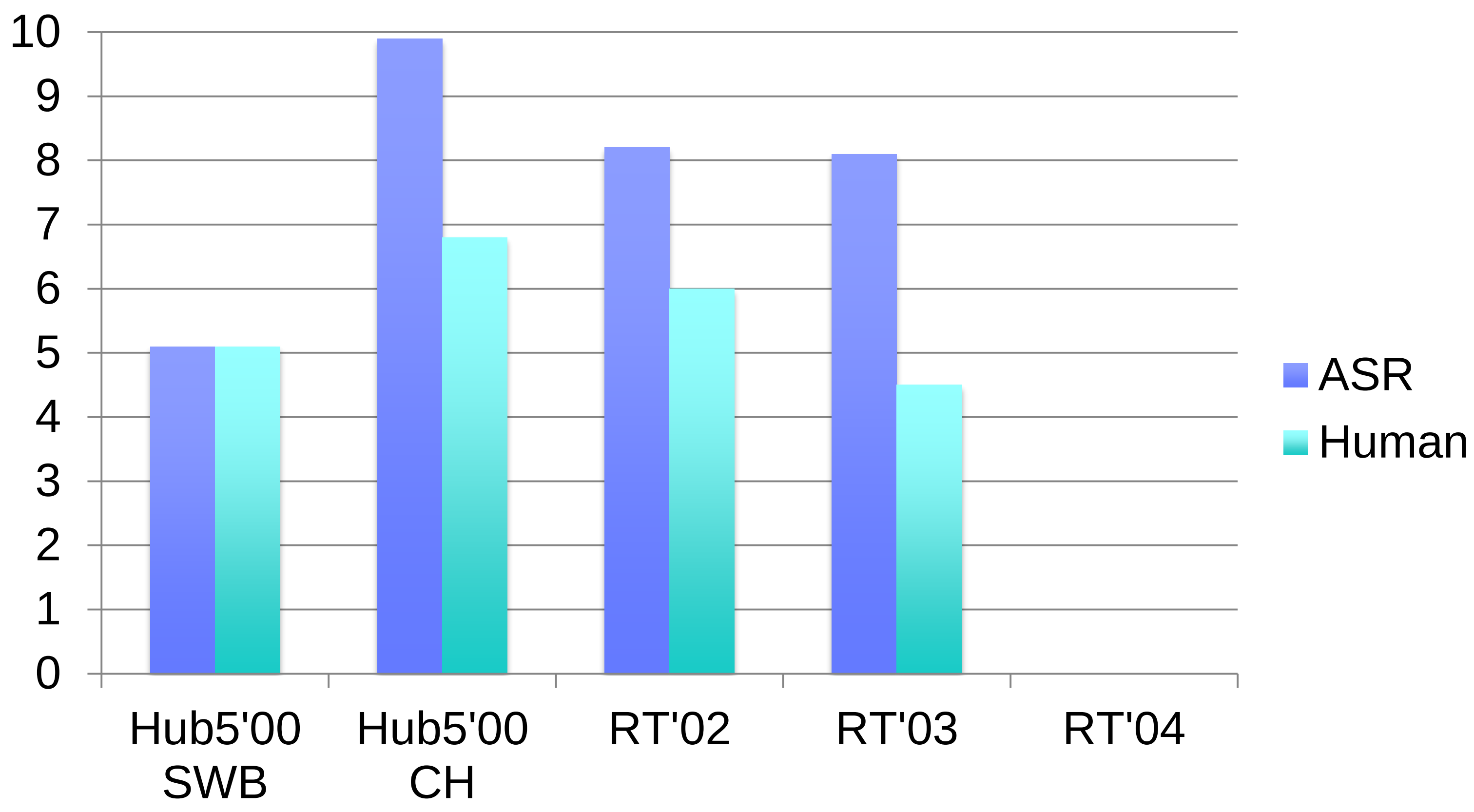
- Hub5’00 SWB: 36/40 test speakers appear in the training data (not an issue according to *)
- Hub5’00 CH: testset is mismatched (only 18 hours of training data)

*A. Stolcke and J. Droppo, “Comparing human and machine errors in conversational speech transcription”, Interspeech 2017.

Man vs. machine: RT'02



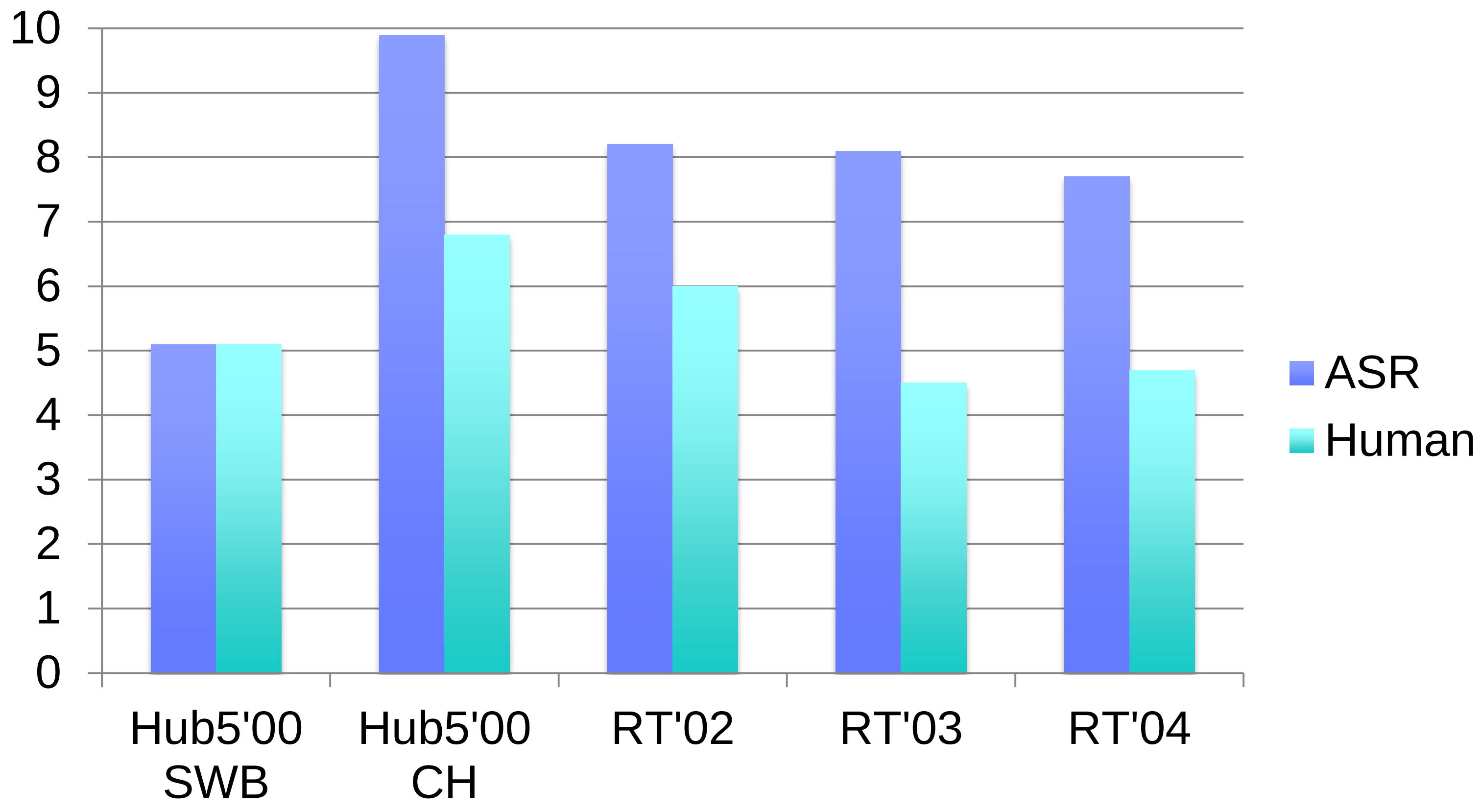
Man vs. machine: RT’03



- LDC reports inter-transcriber disagreement of 4.1 – 4.5% in *

*M. Glenn, S. Strassel, H. Lee, K. Maeda, R. Zakhary, X. Li, “Transcription methods for consistency, volume and efficiency”, LREC 2010.

Man vs. machine: RT’04

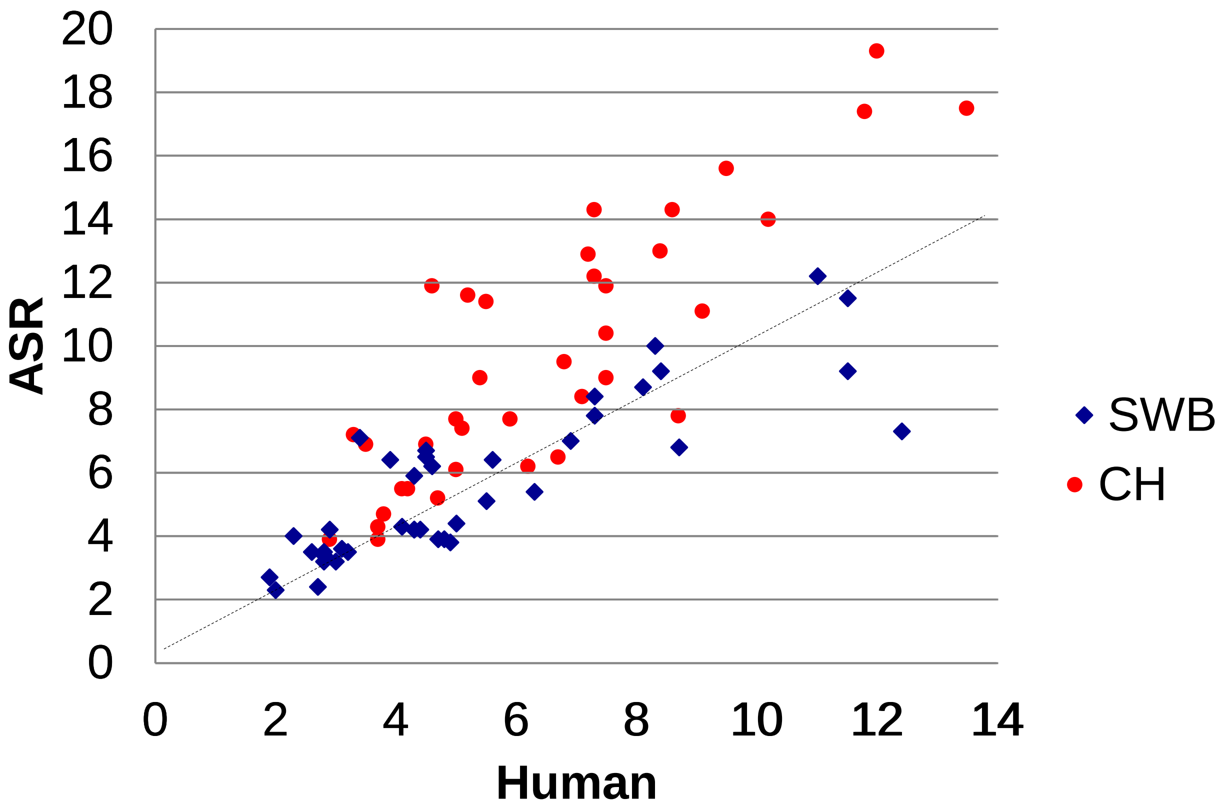


Most frequent errors for Hub5'00

SWB		CH	
ASR	Human	ASR	Human
11: and / in	16: (%hes) / oh	21: was / is	28: (%hes) / oh
9: was / is	12: was / is	16: him / them	22: was / is
7: it / that	7: (i-) / %hes	15: in / and	11: (%hes) / %bcack
6: (%hes) / oh	5: (%hes) / a	8: a / the	10: bentsy / benji
6: him / them	5: (%hes) / hmm	8: and / in	10: yeah / yep
6: too / to	5: (a-) / %hes	8: is / was	9: a / the
5: (%hes) / i	5: could / can	8: two / to	8: is / was
5: then / and	5: that / it	7: the / a	7: (%hes) / a
4: (%hes) / %bcack	4: %bcack / oh	7: too / to	7: the / a
4: (%hes) / am	4: and / in	6: (%hes) / a	7: well / oh

Deletions				Insertions			
SWB		CH		SWB		CH	
ASR	Human	ASR	Human	ASR	Human	ASR	Human
30: it	19: i	46: i	20: i	13: i	16: is	23: a	17: is
20: i	17: it	46: it	18: and	10: a	14: %hes	14: is	17: it
17: that	16: and	39: and	15: it	7: and	12: i	11: i	16: and
16: a	14: that	32: is	15: the	7: of	11: and	10: are	14: have
14: and	14: you	26: oh	14: is	6: you	9: it	10: you	13: a
14: oh	12: is	25: a	13: not	5: do	6: do	9: the	13: that
14: you	12: the	20: to	10: a	5: the	5: have	8: have	12: i
12: %bcack	11: a	19: that	10: in	5: yeah	5: yeah	8: that	11: %hes
12: the	10: of	19: the	10: that	4: air	5: you	7: and	10: not
11: to	9: have	18: %bcack	10: to	4: in	4: are	7: it	9: oh

Speaker error rates Hub5'00



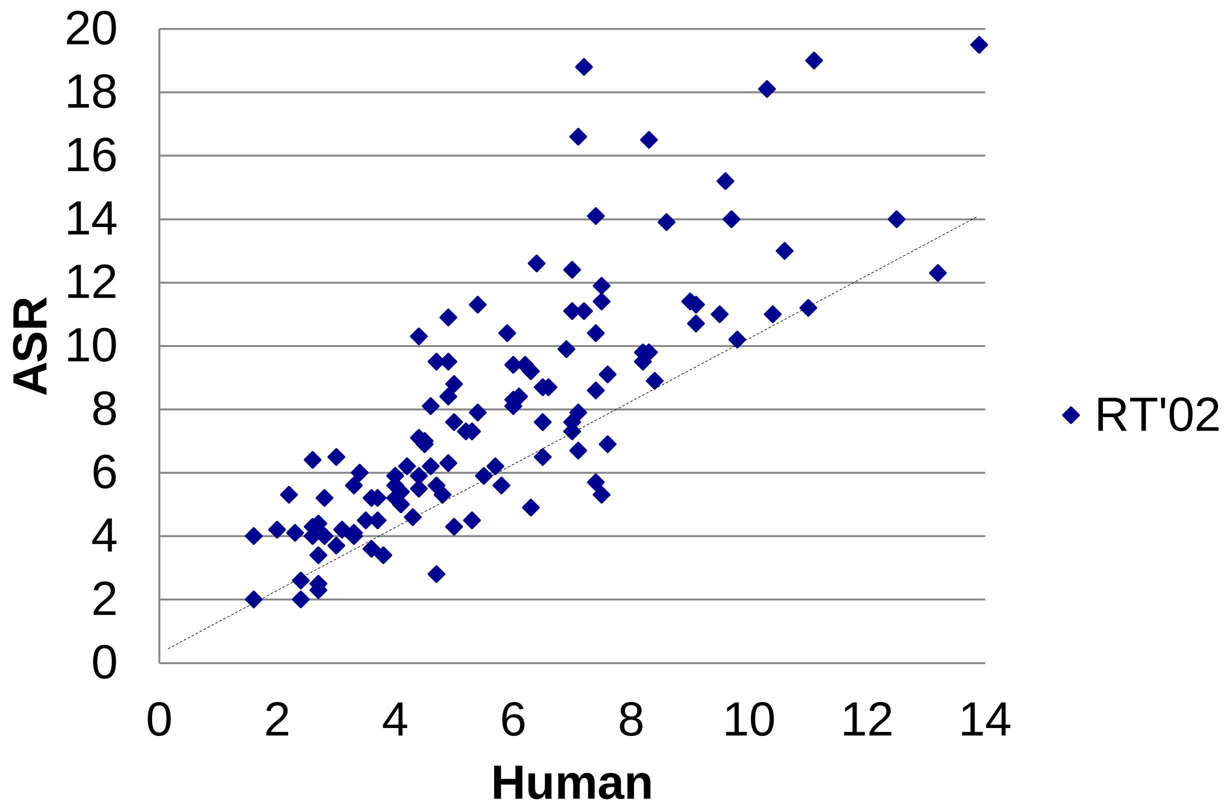
Speaker sw_4910-A

```
REF:    i do not know I i think (it-) ** a lot of it is just you know SEE    how other people live i mean you know I i tend to be you know in in my own circles of friends AND
HUMAN:  i do not know * i think it    IS a lot of it is just you know SEEING how other people live i mean you know * i tend to be you know in in my own circles of friends ***
Eval:                 D               I                               S                                         D                                                        D
ASR:    i do not know i i think it's     a lot of it is just you know SEEING how other people live i mean you know i i tend to be you know in in my own circles of friends and
Eval:                                                                 S
```

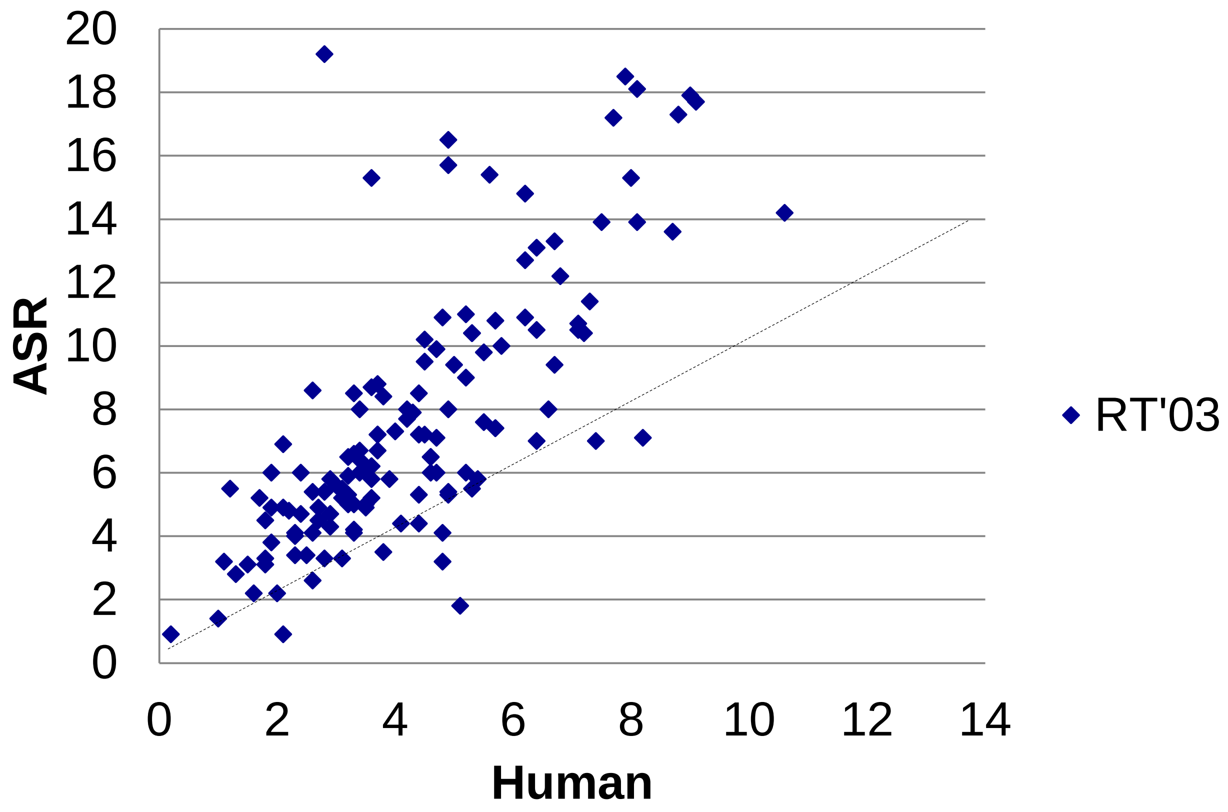
```
REF:    *********** but what about the people who do not make it i mean WHAT DO you what do you do with them
HUMAN:  %HESITATION but what about the people who do not make it i mean **** SO you what do you do with them
Eval:   I                                                           D    S
ASR:                but what about the people who do not make it i mean what do you what do you do with them
Eval:
```

```
REF:    WELL BY             JUST JU YOU KNOW you know living living in different ways than they do
HUMAN:  **** **             **** ** *** **** you know living living in different ways than they do
Eval:   D    D              D    D  D   D
ASR:    well by THE TIME I just ** you know you know living living in different ways than they do
Eval:           I   I    I      D
```

Speaker error rates RT'02



Speaker error rates RT'03



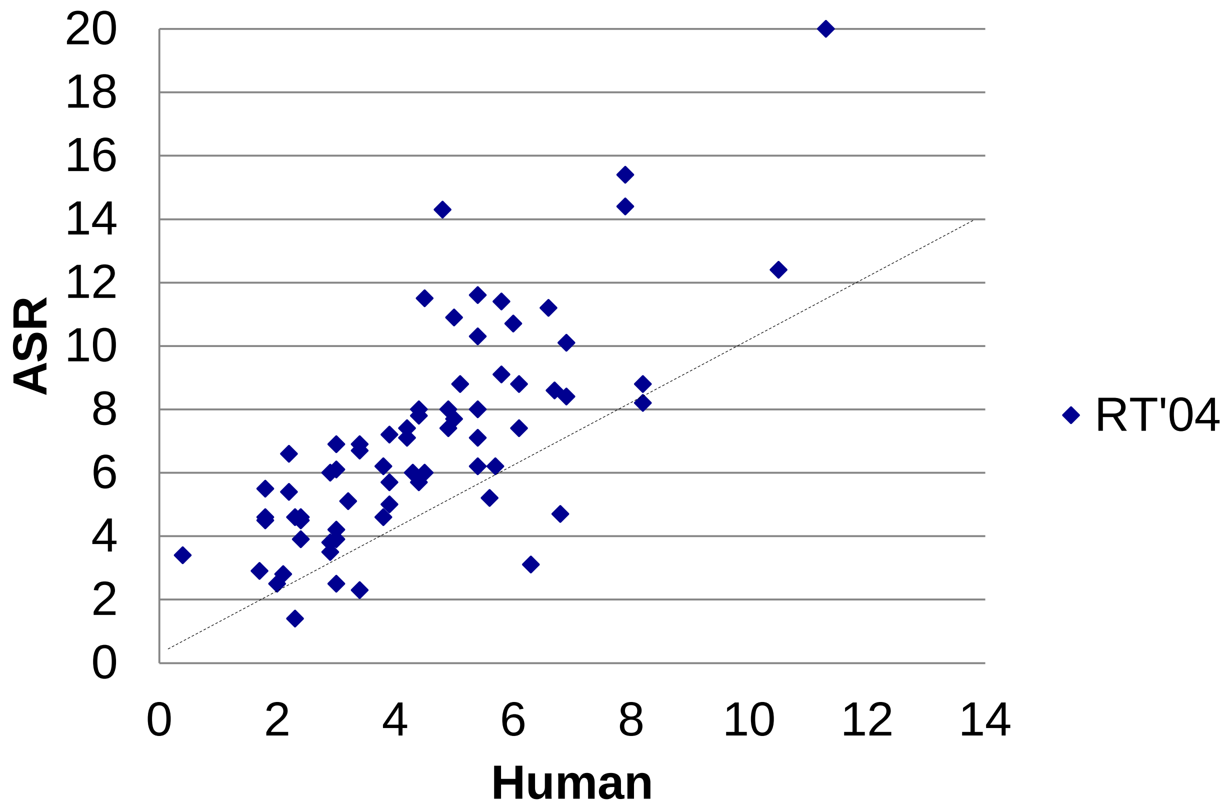
Speaker sw_46512-A

```
REF:    a can of pasta or something like that and you can not necessarily have IT     because it is not good
HUMAN:  a can of pasta or something like that and you can not necessarily have THAT   because it is not good
Eval:                                                                             S
ASR:    a can of pasta or something like that and you can not necessarily **** HAPPEN because ** ** MY  good
Eval:                                                                        D    S              D  D  S
```

```
REF:    what did you say  (i-) what did you say  i could not hear you you were breaking up
HUMAN:  what did you say       what did you say  i could not hear you you were breaking up
Eval:
ASR:    what *** SHE SAID      what *** SHE SAID i CAN   *** hear *** you were breaking up
Eval:        D   S   S              D   S   S      S     S        D
```

```
REF:    %bcack go to MICKEY D.'S and get  some fries   it is already done
HUMAN:  %bcack go to MICKY  D'S  and get  some fries   it is already done
Eval:                S      S
ASR:    %bcack go to MICKY  D'S  and GETS ON   PROCESS ** ** already done
Eval:                S      S        S    S    S       D  D
```

Speaker error rates RT'04



Summary

- **Ten-fold reduction in ASR WER in 25 years: 80% - 8%**
  - Data, speaker adaptation, discriminative training, deep learning in AM and LM
  - Competition drives the error rate down fast
- **Humans and machines make different errors**
  - Humans: low-volume speech, repetitions, short words
  - Machines: accented speech, mismatched training and test conditions
- **Humans have significantly lower WER on this task: ~5%**

- **Acknowledgment**
  - IBM: G. Kurata, T. Sercu, S. Rennie, H.-K. Kuo and M. Picheny
  - Appen: P. Hall, L.-L. Lim, B. Roomi, M. Levot and anonymous transcribers