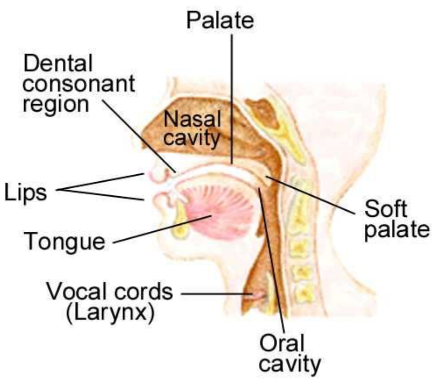
Deep Learning for Speech and Language

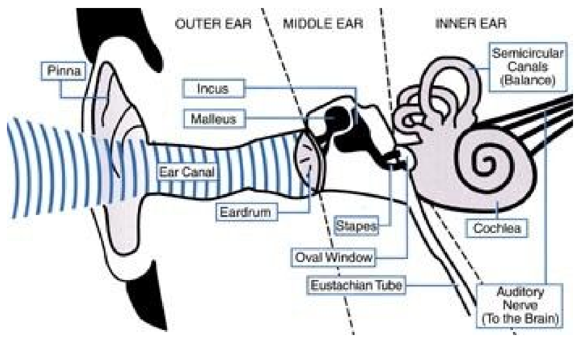
End-To-End Speech Recognition with Recurrent Neural Networks

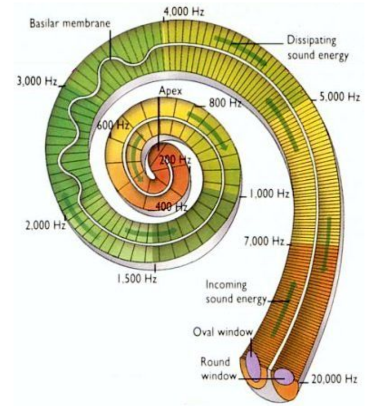
José A. R. Fonollosa
Universitat Politècnica de Catalunya

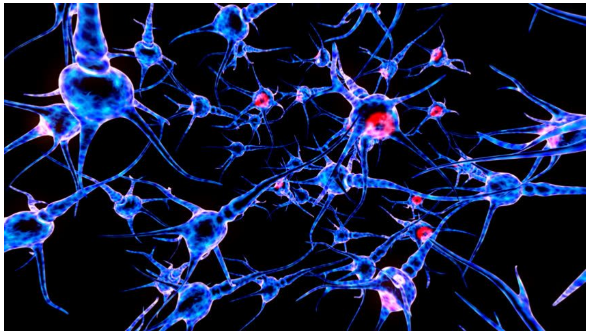
Barcelona, January 26, 2017

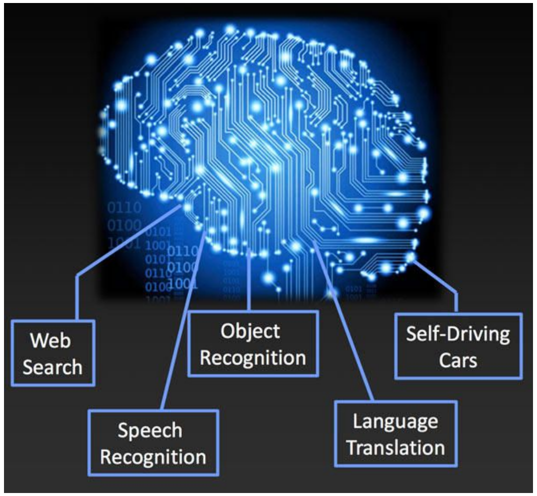
From speech processing to machine learning







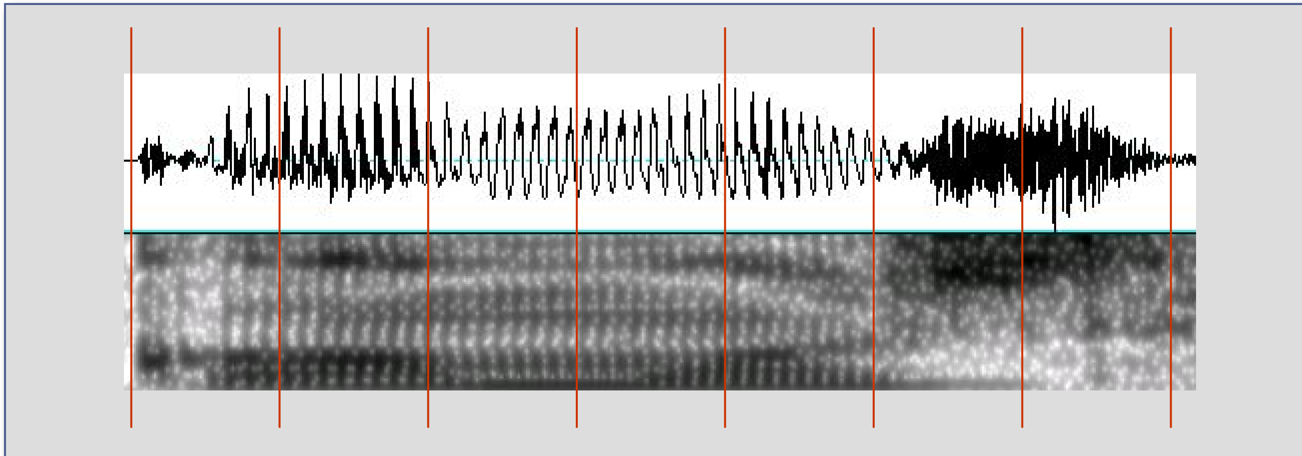




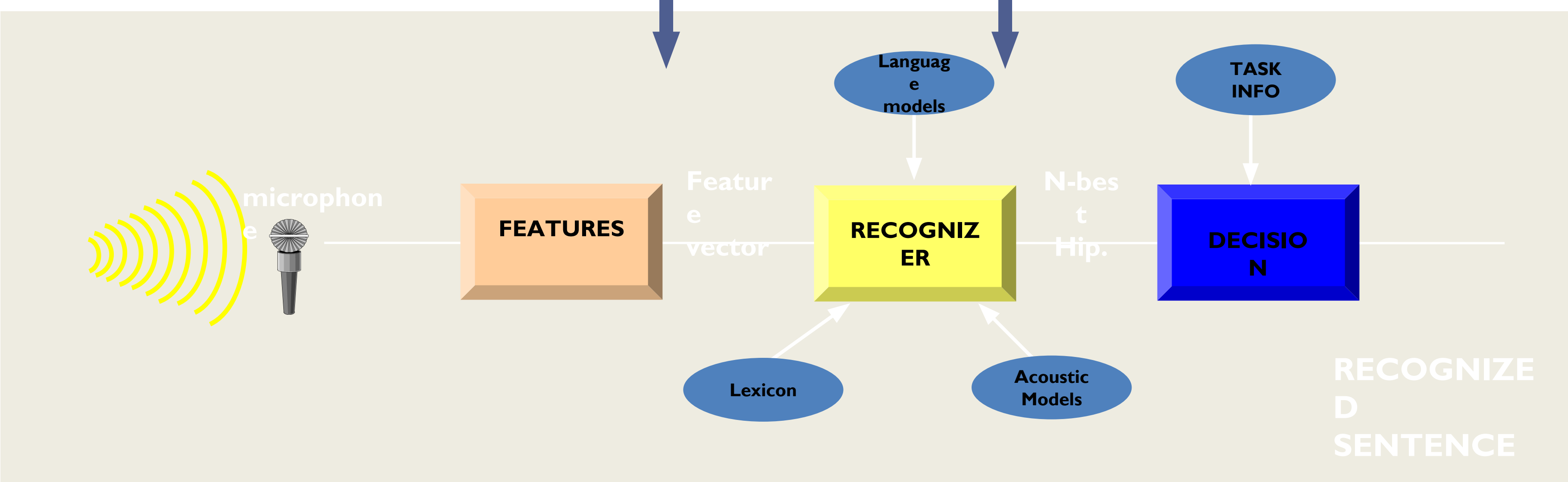
Towards end-to-end RNN Speech Recognition?

- Architectures
  - GMM-HMM: 30 years of feature engineering
  - DNN-GMM-HMM: Trained features
  - DNN-HMM: TDNN, LSTM, RNN, MS
  - DNN for language modeling (RNN)
  - End-to-end DNN?
- Examples
  - Alex Graves (Google)
  - Deep Speech (Baidu)

Recognition system



$\mathbf{x} = x_1 \ldots x_{|\mathbf{x}|}$ $\mathbf{w} = w_1 \ldots w_{|\mathbf{w}|}$



GMM-HMM

Perceptual Feature Extraction (MFCC, PLP, FF, VTLN,GammaTone, ..)

↓

Feature Transformation (Derivative, LDA, MLLT, fMLLR, ..)

↓

GMM (Training: ML, MMI, MPE, MWE, SAT, ..)

↓

HMM

↓

N-GRAM

Acoustic Model

Phonetic inventory

Pronunciation Lexicon

Language Model

DNN-GMM-HMM

Feature Extraction (MFCC, PLP, FF)

DNN (Tandem, Bottleneck, DNN-Derived)

Feature Transformation (LDA, MLLT, fMLLR)

GMM

HMM

N-GRAM

Acoustic Model

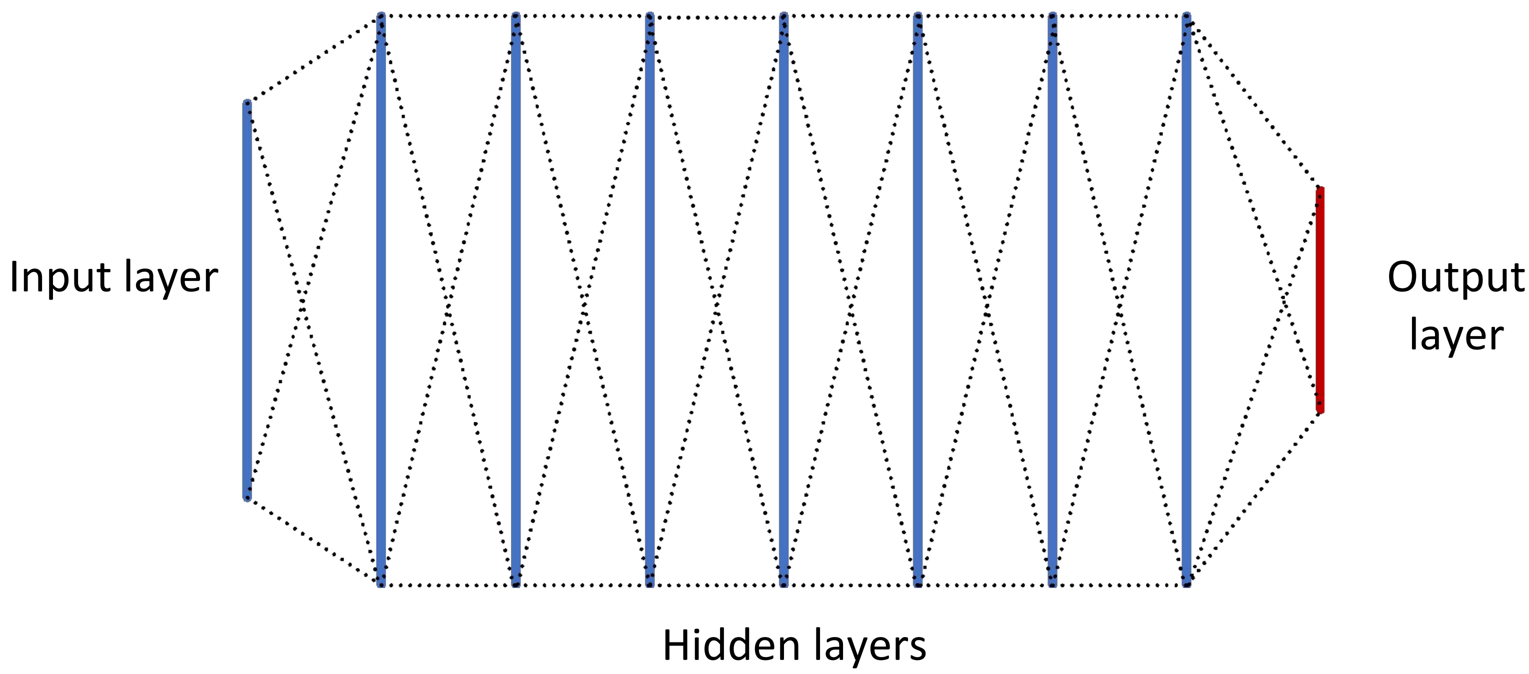
Phonetic inventory

Pronunciation Lexicon

Language Model

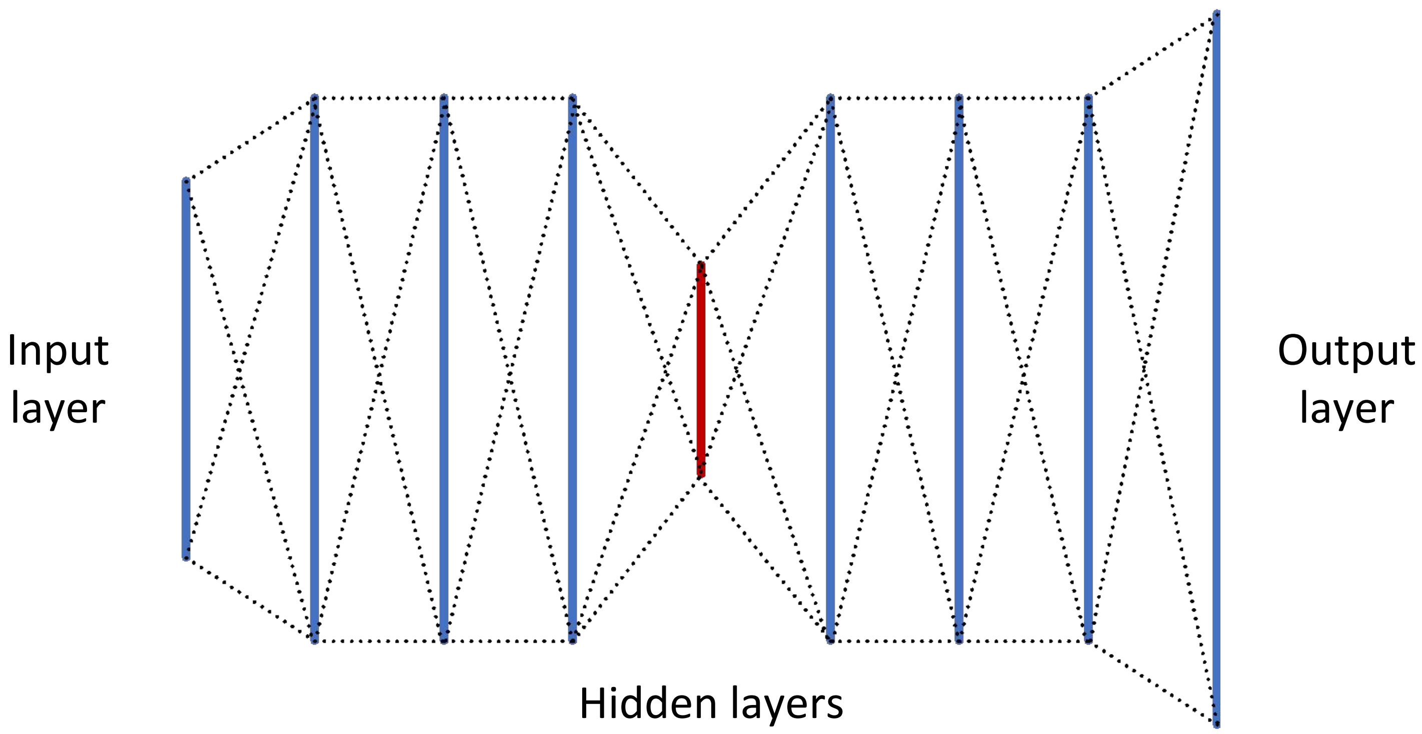
Tandem

- **MLP outputs as input to GMM**



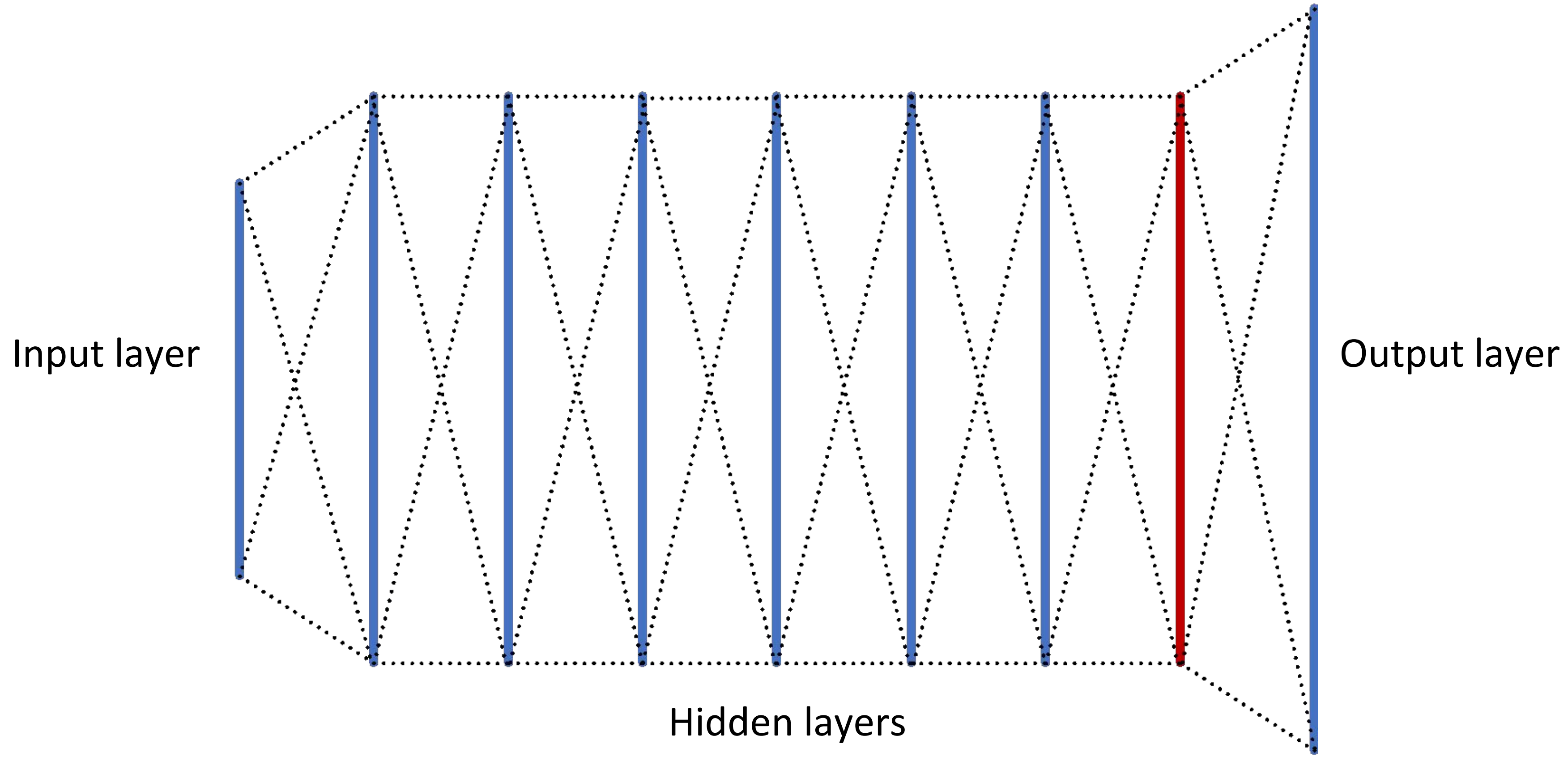
Bottleneck Features

- **Use one narrow hidden layer. Supervised or unsupervised training (autoencoder)**



DNN-Derived Features

- Zhi-Jie Yan, Qiang Huo, Jian Xu: A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR. INTERSPEECH 2013: 104-108



DNN-GMM-HMM

DNN (Tandem, Bottleneck, DNN-Derived)

↓

Feature Transformation (LDA, MLLT, fMLLR, …)

↓

GMM

↓

HMM

↓

N-GRAM

Acoustic Model

Phonetic inventory

Pronunciation Lexicon

Language Model

DNN-HMM

DNN

↓

HMM

↓

N-GRAM

Acoustic Model

Phonetic inventory

Pronunciation Lexicon

Language Model

DNN-HMM+RNNLM

DNN

HMM

(N-GRAM +) RNN

Acoustic Model

Phonetic inventory

Pronunciation Lexicon

Language Model

RNN-RNNLM

RNN

Acoustic Model

(N-GRAM +) RNN

Language Model

End-to-End RNN

- Alex Graves et al. (2006) “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks” Proceedings of the International Conference on Machine Learning, ICML
- Florian Eyben, Martin Wöllmer, Björn Schuller, Alex Graves (2009) “From Speech to Letters - Using a Novel Neural Network Architecture for Grapheme Based ASR”, ASRU
- Alex Graves, Navdeep Jaitly, (Jun 2014) “Towards End-To-End Speech Recognition with Recurrent Neural Networks”, International Conference on Machine Learning, ICML
- Jan Chorowski et al. (Dec 2014) “End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results”, Deep Learning and Representation Learning Workshop: NIPS

End-to-End RNN

- **Awni Hannun et al (Dec 2014), "Deep Speech: Scaling up end-to-end speech recognition", arXiv:1412.5567 [cs.CL]**
- D. Bahdanau et al. (Dec 2014) "End-to-End Attention-based Large Vocabulary Speech Recognition", arXiv:1508.04395 [cs.CL]
- Miao, Y., Gowayyed, M., and Metze, F. (2015). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. arXiv:1507.08240 [cs]
- **Baidu Research – 34 authors- (Dec 2015), "Deep Speech 2: End-to-end Speech Recognition in English and Mandarin", arXiv:1512.02595 [cs.CL]**

End-to-End RNN

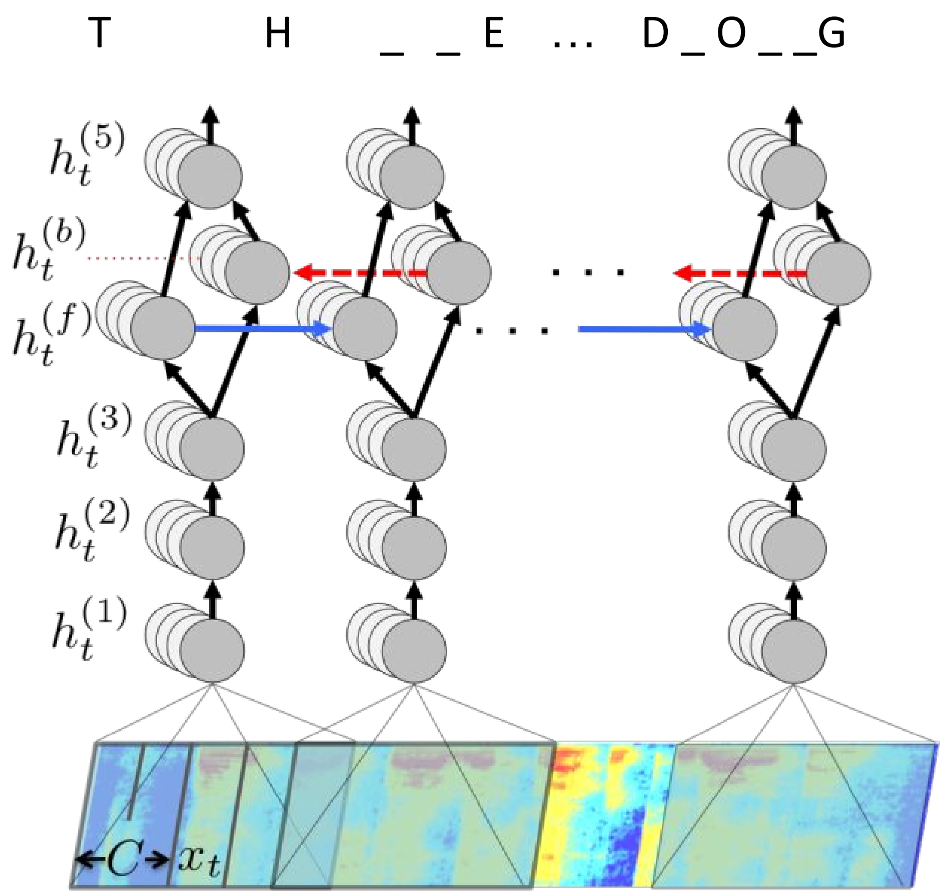
- No perceptual features (MFCC). No feature transformation. No phonetic inventory. No transcription dictionary. No HMM.
- The output of the RNN are characters including space, apostrophe, (not CD phones)
- Connectionist Temporal Classification (No fixed aligment speech/character)
- Data augmentation. 5,000 hours (9600 speakers) + noise = 100,000 hours. Optimizations: data parallelism, model parallelism
- Good results in noisy conditions

Adam Coates

Bidirectional Recursive DNN

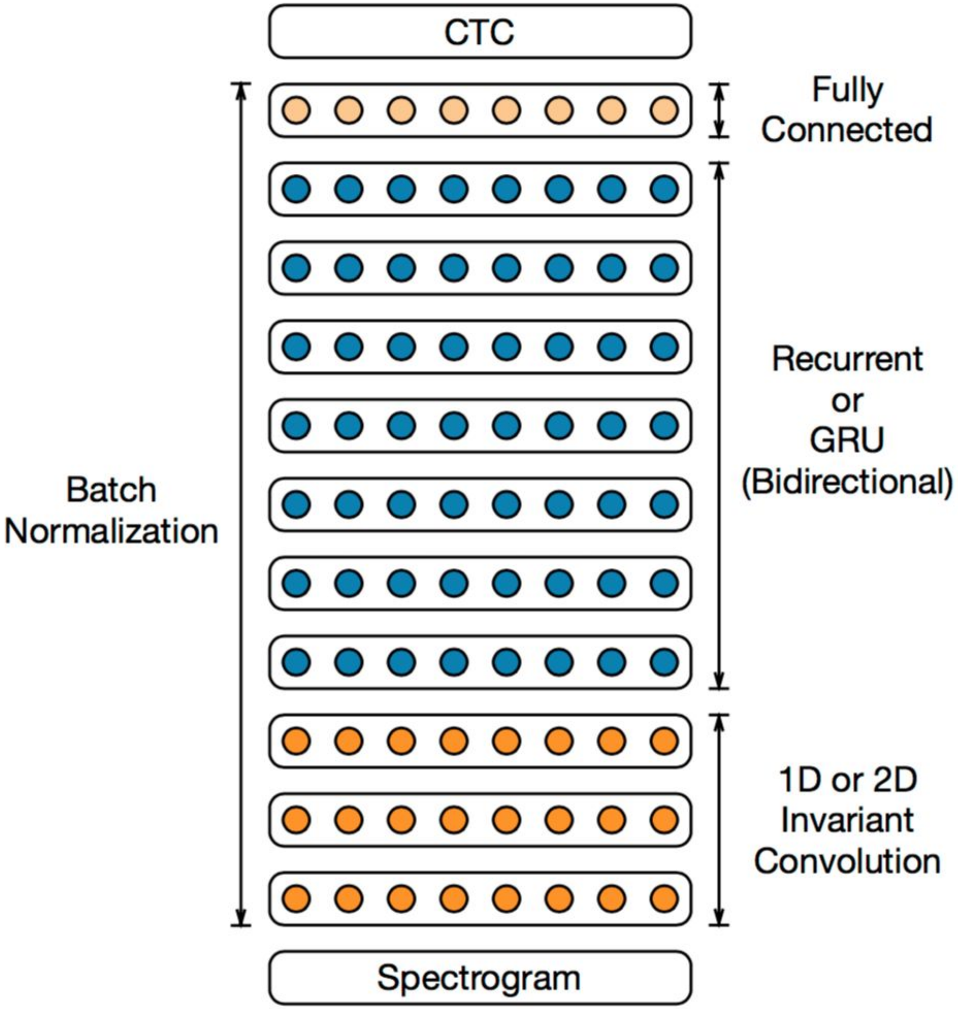
Unrolled RNN

Spectogram
Clipped ReLu
Accelerated gradient method
GPU friendly



$$h_t^{(f)} = g(W^{(4)} h_t^{(3)} + W_r^{(f)} h_{t-1}^{(f)} + b^{(4)})$$
$$h_t^{(b)} = g(W^{(4)} h_t^{(3)} + W_r^{(b)} h_{t+1}^{(b)} + b^{(4)})$$

Deep Speech II (Baidu)



Language Model

English: Kneser-Ney smoothed 5-gram model with pruning.

Vocabulary: 400,000 words from 250 million lines of text

Language model with 850 million n-grams.

Mandarin: Kneser-Ney smoothed character level 5-gram model with pruning

Training data: 8 billion lines of text.

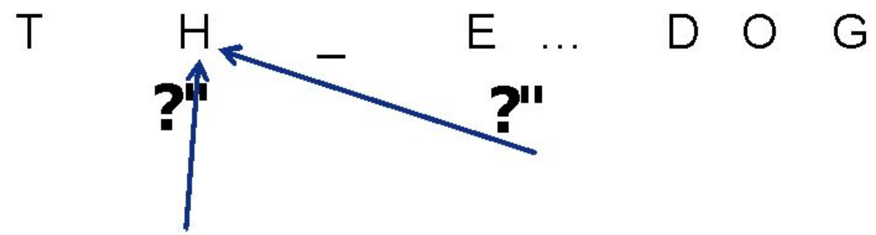
Language model with about 2 billion n-grams.

Maximize $Q(y) = \log(p_{ctc}(y|x)) + \alpha \log(p_{lm}(y)) + \beta \, word\_count(y)$

Language	Architecture	Dev no LM	Dev LM
English	5-layer, 1 RNN	27.79	14.39
English	9-layer, 7 RNN	14.93	9.52
Mandarin	5-layer, 1 RNN	9.80	7.13
Mandarin	9-layer, 7 RNN	7.55	5.81

Table 6: Comparison of WER for English and CER for Mandarin with and without a language model. These are simple RNN models with only one layer of 1D invariant convolution.

Connectionist Temporal Classification

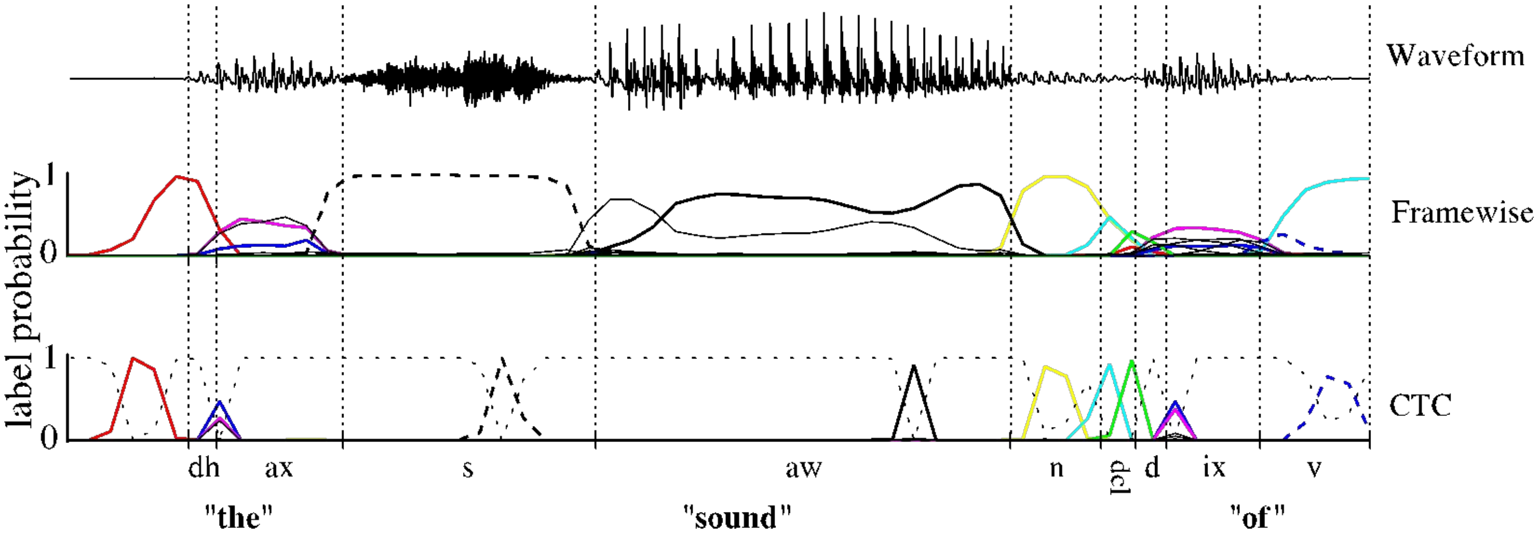


- How'to'connect'speech'data'with'transcription?'
  – Transcription'not'labeled'per'millisecond'
- Use'CTC,'from'[Graves'06]'
- Efficient'dynamic'programming'of'all'possible' alignments'to'compute'error'of'{audio,'transcription}'

Bryan'Catanzaro

Connectionist Temporal Classification

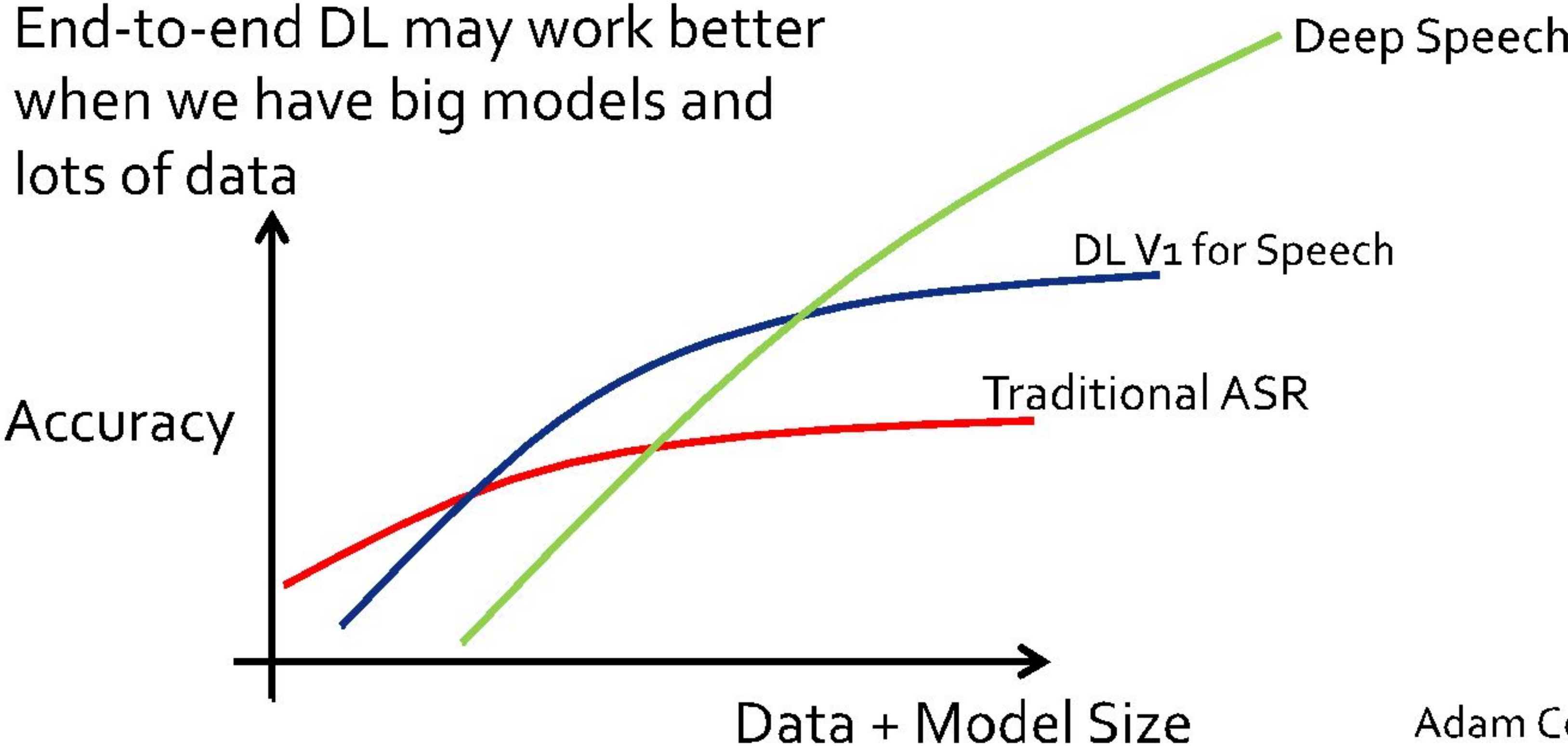
- The framewise network receives an error for misalignment
- The CTC network predicts the sequence of phonemes / characters (as spikes separated by 'blanks')
- No force alignment (initial model) required for training.



Alex Graves 2006

GMM-HMM / DNN-HMM / RNN

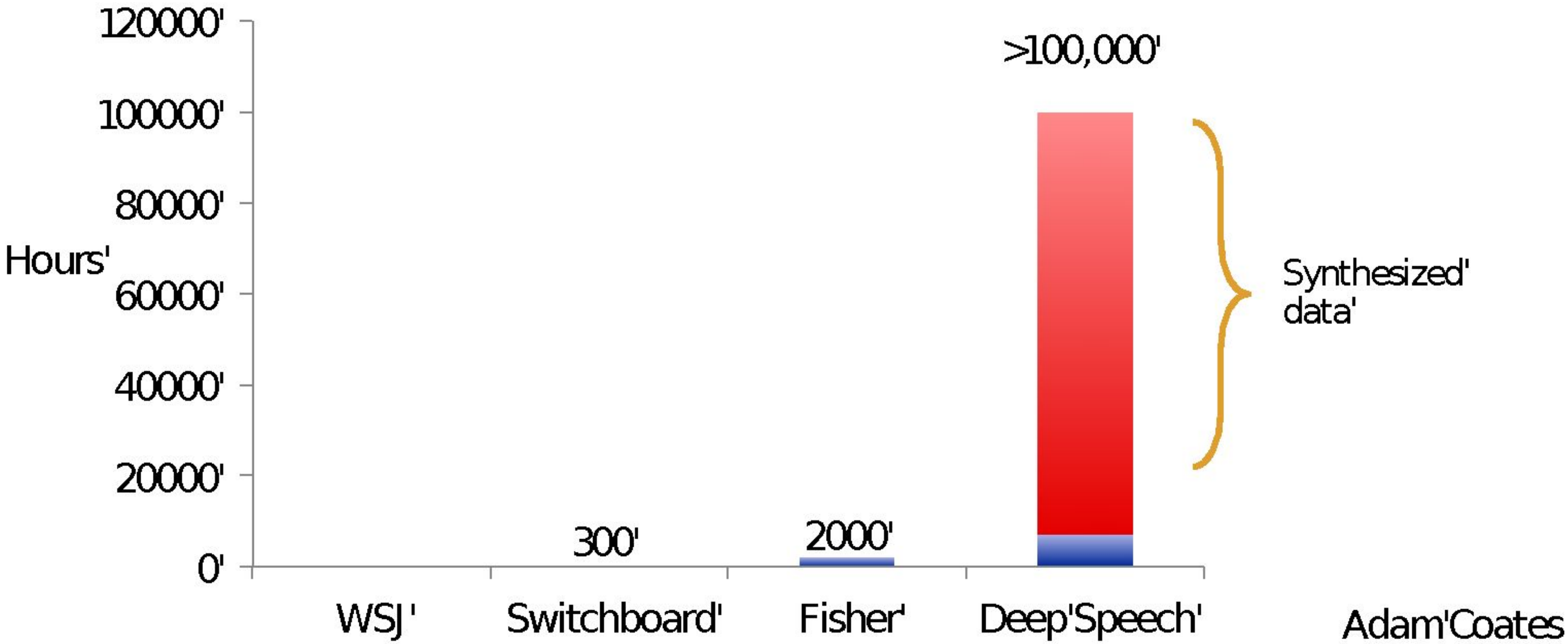
- End-to-end DL may work better when we have big models and lots of data



Adam Coates

Data Augmentation

This approach needs bigger models and bigger datasets.
Synthesis by superposition: reverberations, echoes, a large number of short noise recordings from public video sources, jitter
Lombart Effect.



Results
2000 HUB5 (LDC2002S09)

System	AM training data	SWB	CH
Vesely et al. (2013)	SWB	12.6	24.1
Seide et al. (2014)	SWB+Fisher+other	13.1	–
Hannun et al. (2014)	SWB+Fisher	12.6	19.3
Zhou et al. (2014)	SWB	14.2	–
Maas et al. (2014)	SWB	14.3	26.0
Maas et al. (2014)	SWB+Fisher	15.0	23.0
Soltau et al. (2014)	SWB	10.4	19.1
Saon et al (2015)	SWB+Fisher+CH	8.0	14.1

Hybrid(IBM) versus DS1(Baidu)

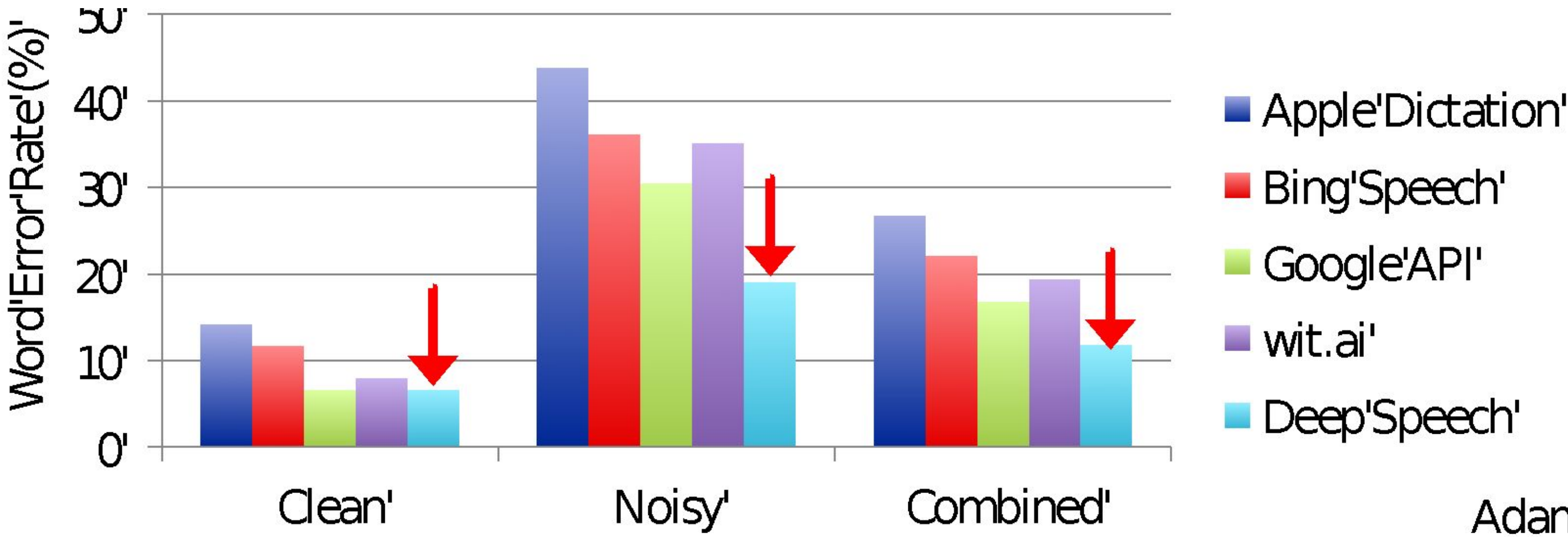
	IBM 2015	Baidu 2014
Features	VTL-PLP, MVN, LDA, STC, fMMLR, i-Vector	80 log filter banks
Alignment	GMM-HMM 300K Gaussians	-
DNN	DNN(5x2048) + CNN(128x9x9+5x2048) + +RNN 32000 outputs	4RNN (5 x 2304) 28 outputs
DNN Training	CE + MBR Discriminative Training (ST)	CTC
HMM	32K states (DNN outputs) pentaphone acoustic context	-
Language Model	37M 4-gram + model M (class based exponential model) + 2 NNLM	4-gram (Transcripts)

DS1 versus DS2 (Baidu)

	Deep Speech 1 (Baidu 2014)	DS2 (Baidu 2015)
Features	80 log filter banks	?
Alignment	-	-
DNN	4RNN (5 x 2304) 28 outputs	9-layer, 7RNN, BatchNorm, Conv. Layers. (Time/Freq)
DNN Training	CTC	CTC
HMM	-	-
Language Model	4-gram	5-gram

Results II (DS1)

Training: Baidu database + data augmentation
Test: new dataset of 200 recording in both clean and noisy settings (no details)
Comparison against commercial systems in production



Adam'Coates

Deep Speech 2 Versus DS1

Test set	DS1	DS2
Baidu Test	24.01	13.59

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Deep Speech 2 Versus DS1

Accented Speech			
Test set	DS1	DS2	Human
VoxForge American-Canadian	15.01	7.55	4.85
VoxForge Commonwealth	28.46	13.56	8.15
VoxForge European	31.20	17.55	12.76
VoxForge Indian	45.35	22.44	22.15

Table 14: Comparing WER of the DS1 system to the DS2 system on accented speech.

Noisy Speech			
Test set	DS1	DS2	Human
CHiME eval clean	6.30	3.34	3.46
CHiME eval real	67.94	21.79	11.84
CHiME eval sim	80.27	45.05	31.33

DS2 Training data

Dataset	Speech Type	Hours
WSJ	read	80
Switchboard	conversational	300
Fisher	conversational	2000
LibriSpeech	read	960
Baidu	read	5000
Baidu	mixed	3600
Total		11940

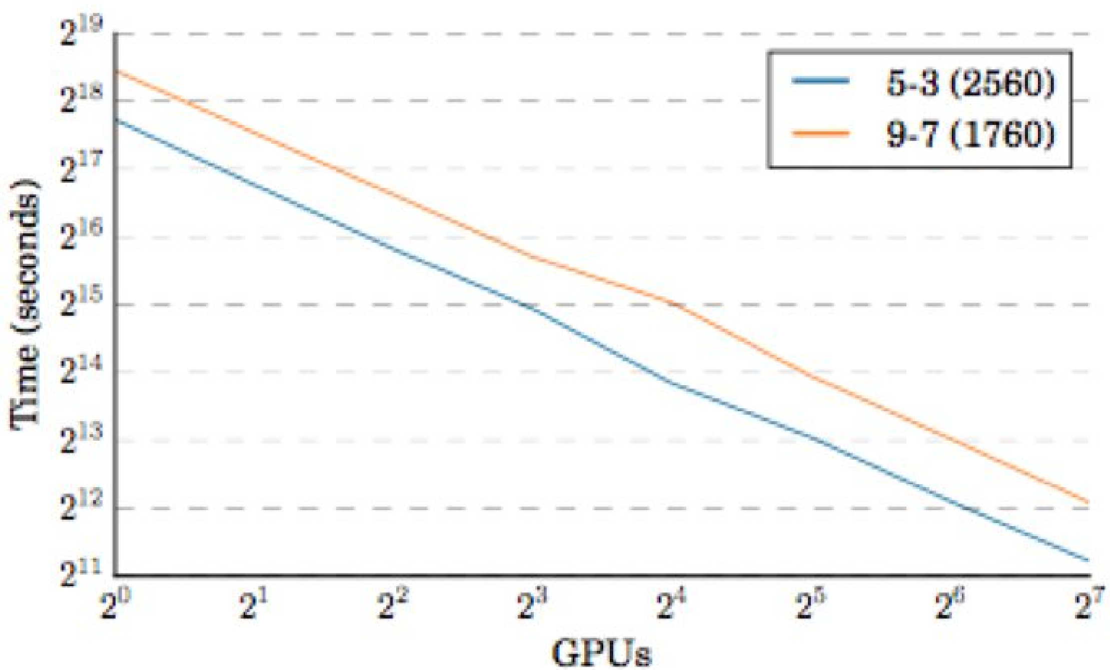
DS2 Training data

Dataset	Speech Type	Hours
WSJ	read	80
Switchboard	conversational	300
Fisher	conversational	2000
LibriSpeech	read	960
Baidu	read	5000
Baidu	mixed	3600
Total		11940

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

System optimization

- Scalability and data-parallelism
- GPU implementation of CTC loss function
- Memory allocation



Deep Speech Demo

- http://www.ustream.tv/recorded/60113824/highlight/631666

References

- Alex Graves et al. (2006) “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks” Proceedings of the International Conference on Machine Learning, ICML
- Florian Eyben, Martin Wöllmer, Björn Schuller, Alex Graves (2009) “From Speech to Letters - Using a Novel Neural Network Architecture for Grapheme Based ASR”, ASRU
- Alex Graves, Navdeep Jaitly, (Jun 2014) “Towards End-To-End Speech Recognition with Recurrent Neural Networks”, International Conference on Machine Learning, ICML
- Jan Chorowski et al. (Dec 2014) “End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results”, Deep Learning and Representation Learning Workshop: NIPS
- Awni Hannun et al (Dec 2014), “Deep Speech: Scaling up end-to-end speech recognition”, arXiv:1412.5567 [cs.CL]
- George Saon et al. “The IBM 2015 English Conversational Telephone Speech Recognition System”, arXiv:1505.05899 [cs.CL]