State of the art in Speech Recognition

Steve Renals
Automatic Speech Recognition – ASR Lecture 17
23 March 2017

G Saon et al, "English Conversational Telephone Speech Recognition by Humans and Machines", arXiv:1703:02136

Human Transcription Experiments

	WER SWB	WER 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 WER from [1]	5.9	11.3

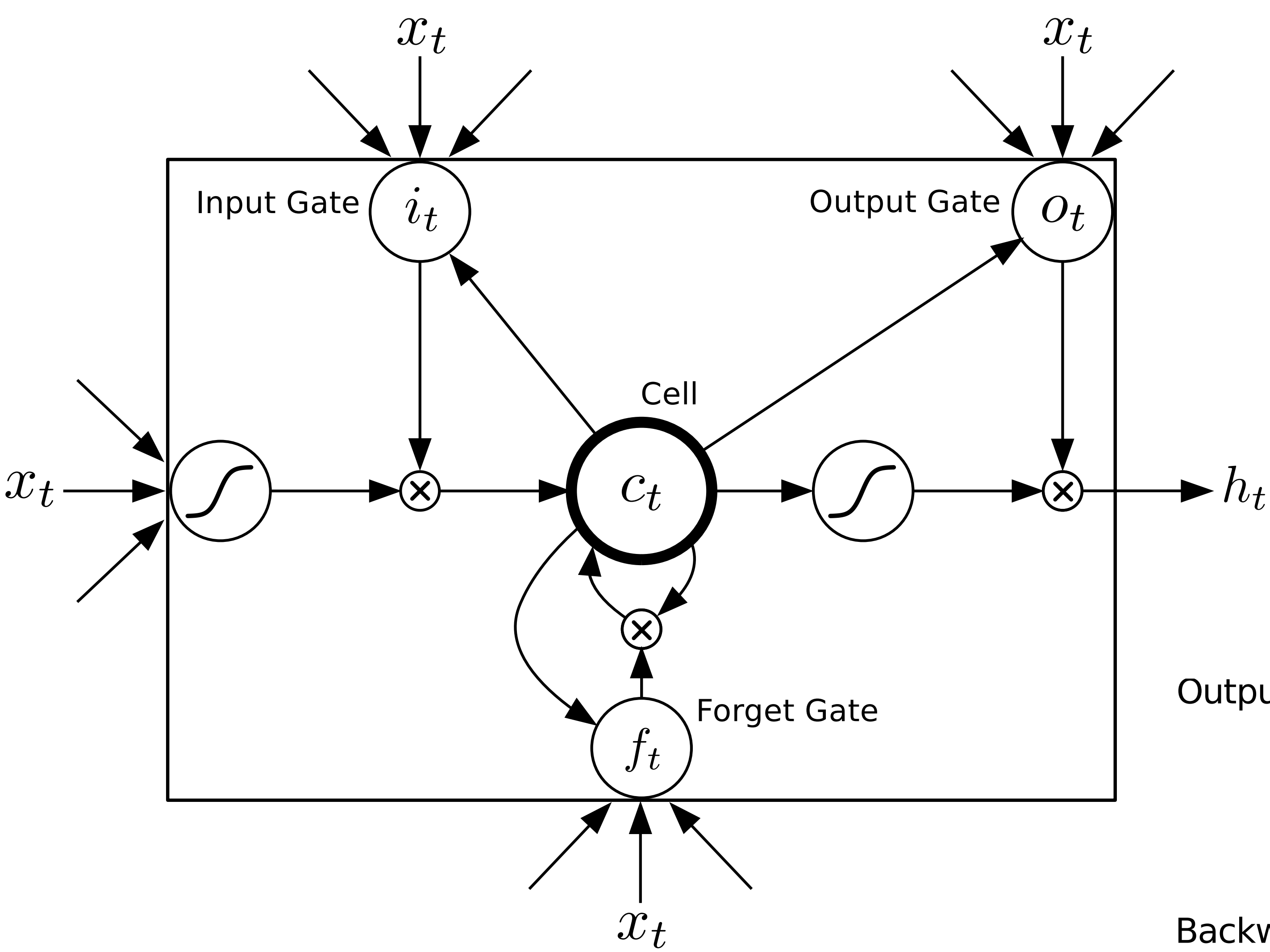
Task

- Conversational telephone speech
- Total 1975h training data
- 5 test sets, totalling 24h

Acoustic Models

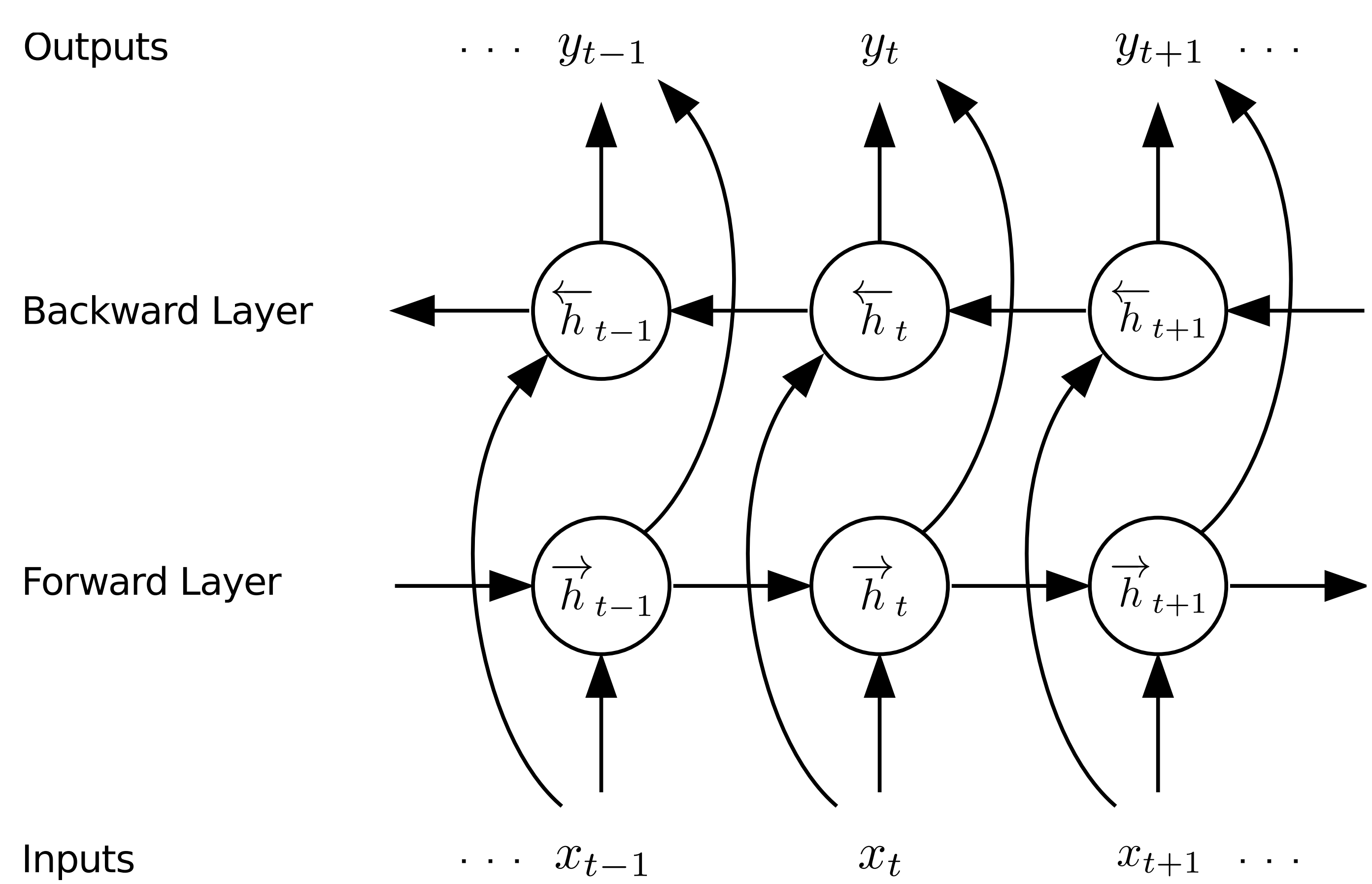
- LSTM recurrent neural networks
- Speaker adversarial multi-task learning networks (SA-MTL)
- Very deep convolutional networks – ResNet Acoustic Models
- Model Combination (frame-level)

LSTM Acoustic Model



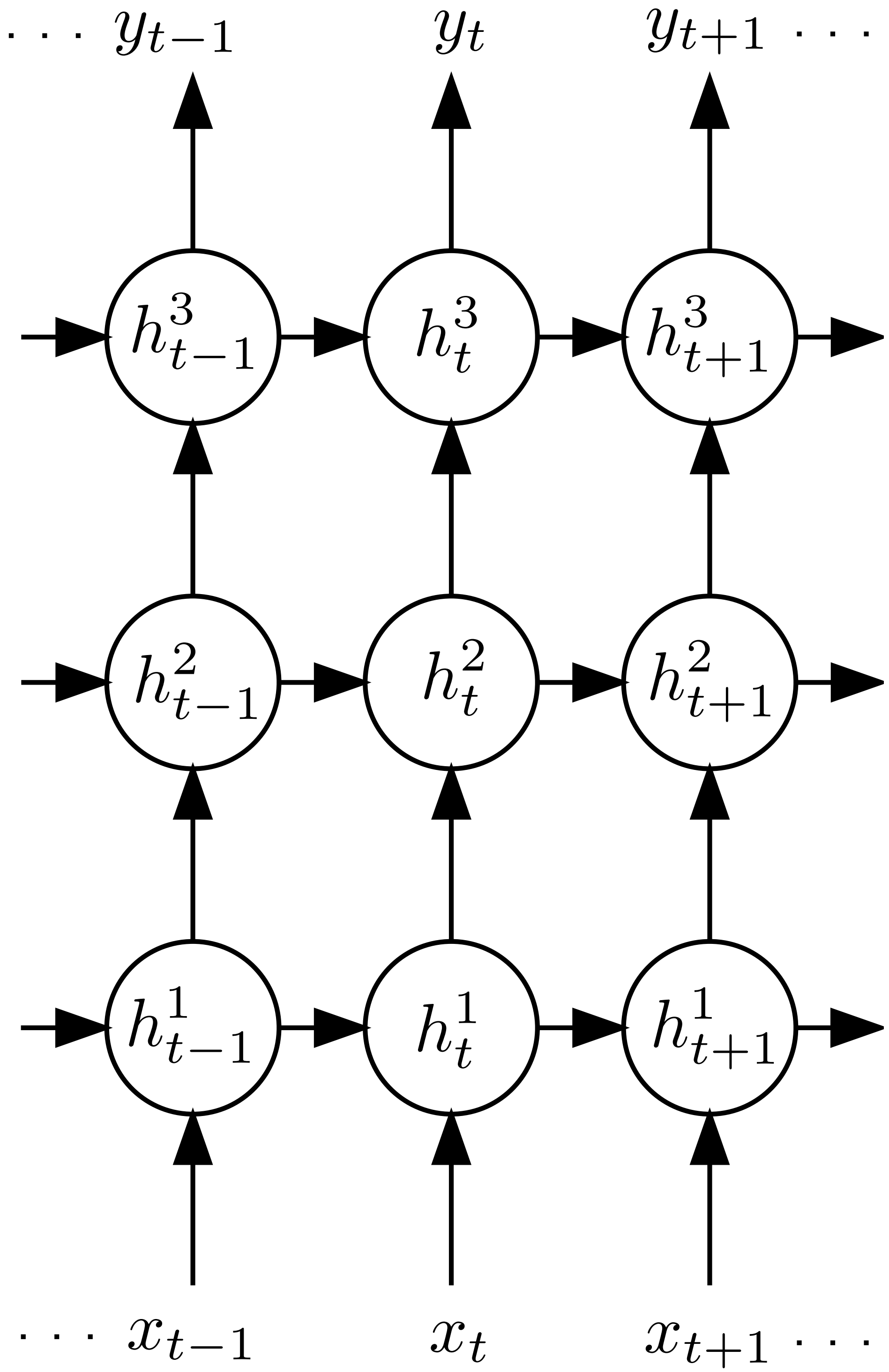
LSTM Cell

Bidirectional RNN

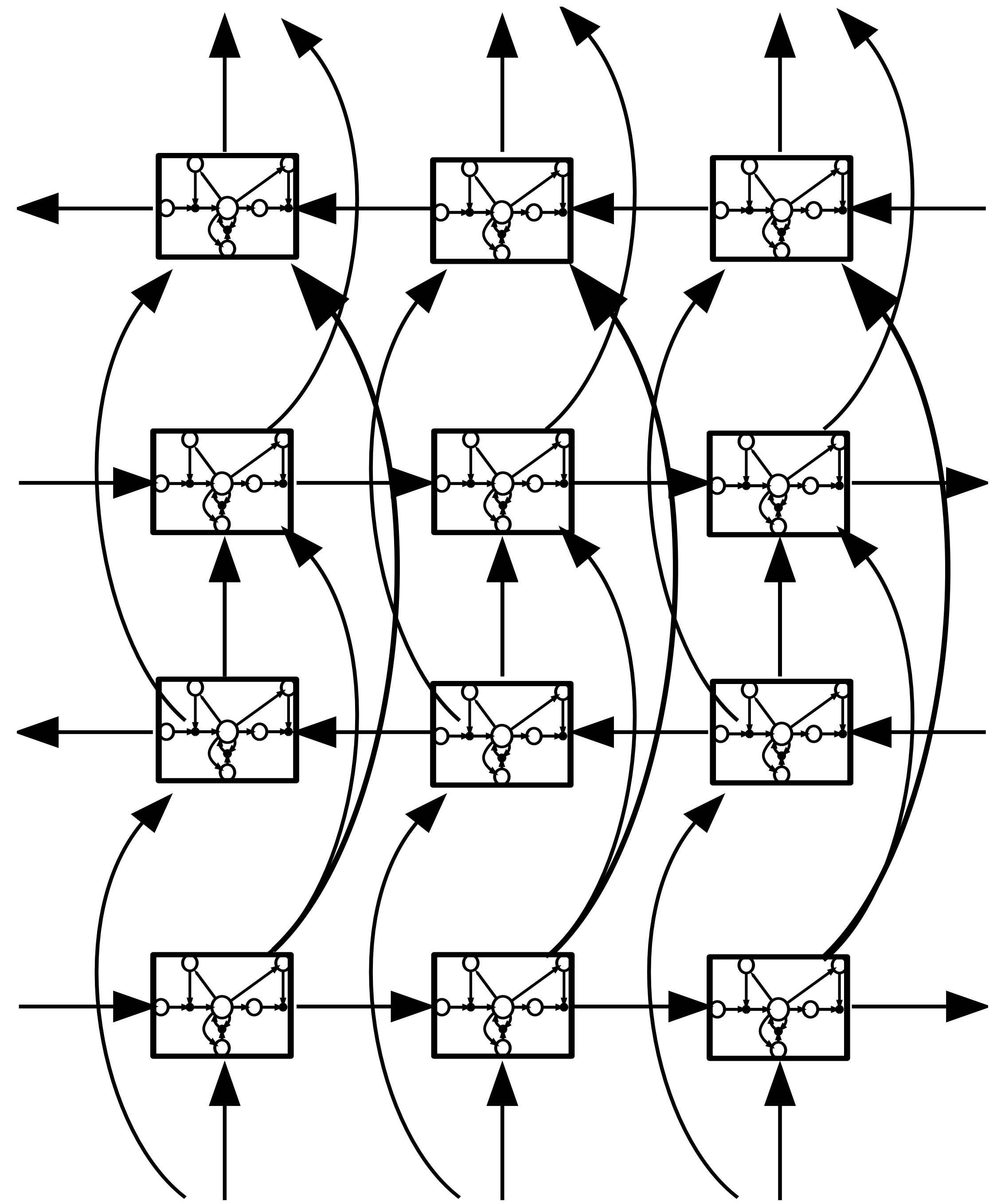


Graves et al, Hybrid speech recognition with deep bidirectional LSTM, ICASSP-2013.

LSTM Acoustic Model



Deep RNN



Deep Bidirectional LSTM

LSTM Architecture and Setup

- LSTM has 4-6 32k bidirectional layers with 1024 cells/layer (512 each direction)

- 256 unit linear bottleneck layer

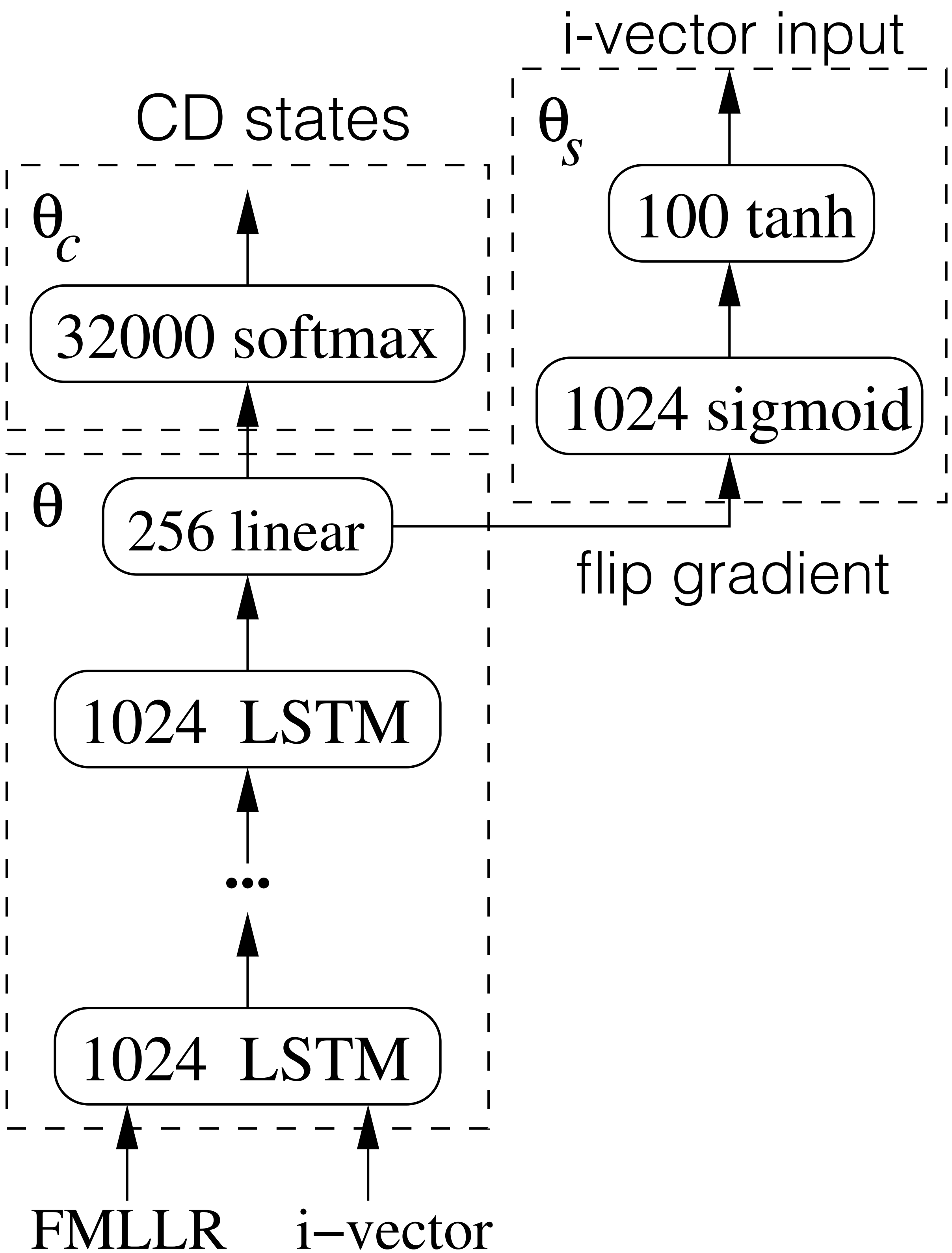
- 32k context-dependent state outputs

- 40-dimension FMLLR input features + 100-dimension i-vector

- 14 passes CE (frame-level) training, 1 pass sequence training

- Training took 2 weeks on a GPU

Speaker-adversarial multi-task learning (SA-MTL)



- Train a speaker classifier in parallel with main classifier
- Subtract the gradient component from the speaker classifier when training
- Speaker classifier trained to predict input i-vector

LSTM Results

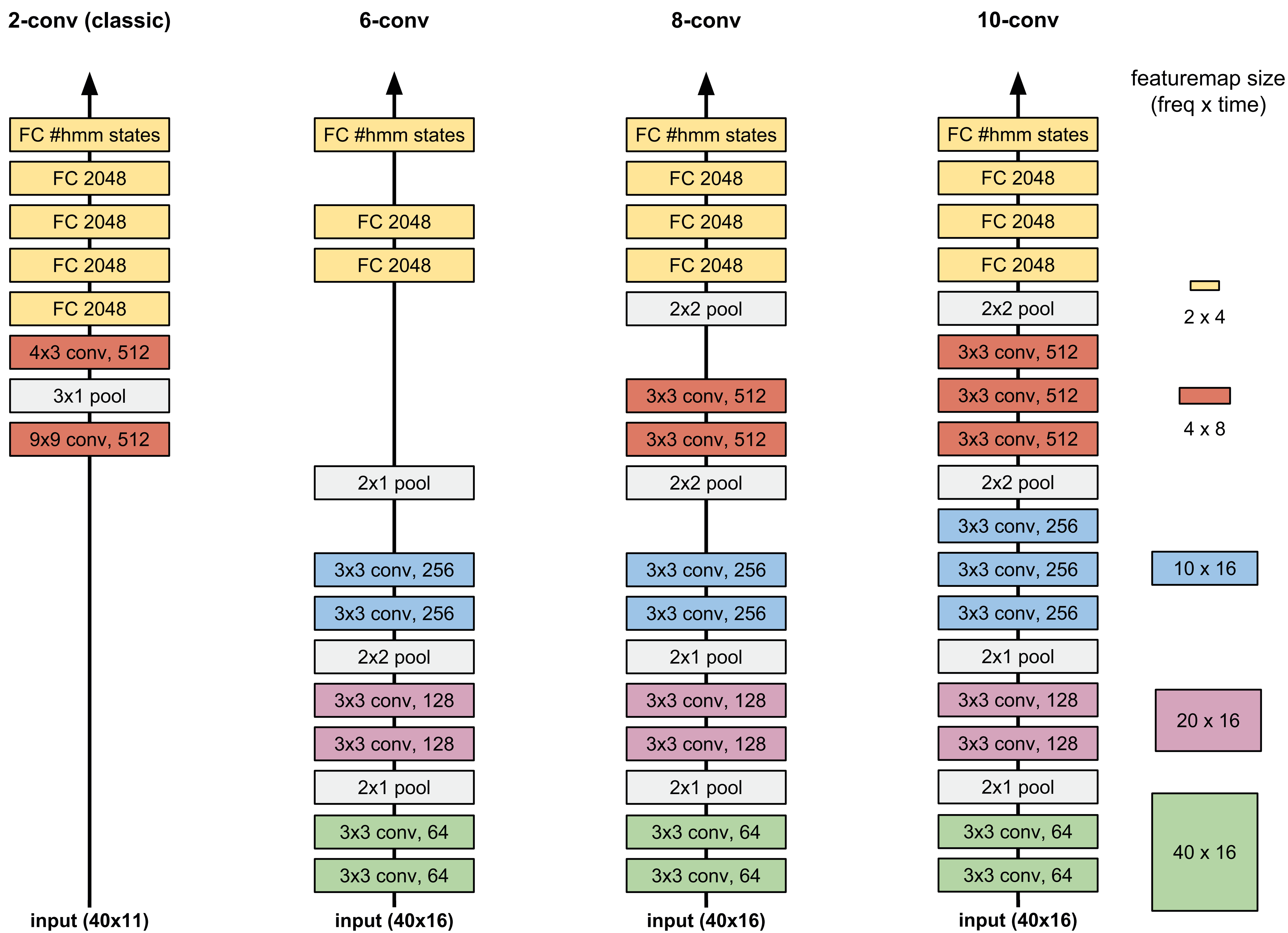
LSTM	SWB	CH	RT'02	RT'03	RT'04	DEV'04f
4-layer	8.0	14.3	12.2	11.6	11.0	10.8
6-layer	7.7	14.0	11.8	11.4	10.8	10.4
Realigned	7.7	13.8	11.7	11.2	10.8	10.2
SA-MTL	7.6	13.6	11.5	11.0	10.7	10.1
Feat. fusion	7.2	12.7	10.7	10.2	10.1	9.6

GMM/ML	21.2	36.4
GMM/BMMI	18.6	33.0
DNN/CE	14.2	25.7
DNN/MMI	12.9	24.6

Vesely et al (2013)

Feature fusion: append log mel filter bank features (+ first and second derivatives) to FMLLR and i-vector features

Deep CNN Acoustic Models



ResNet
Deep Residual Networks

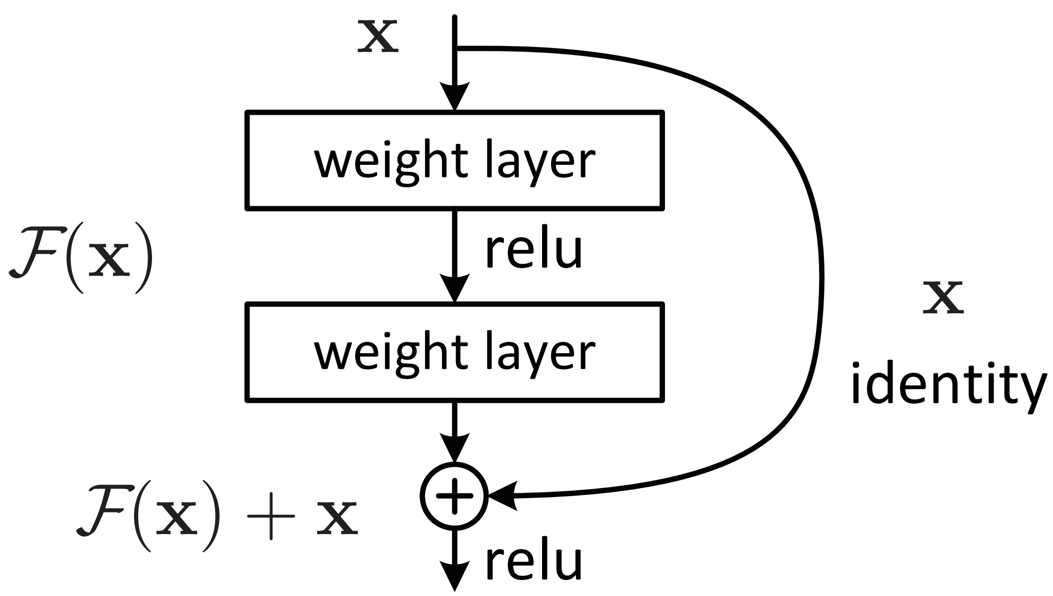
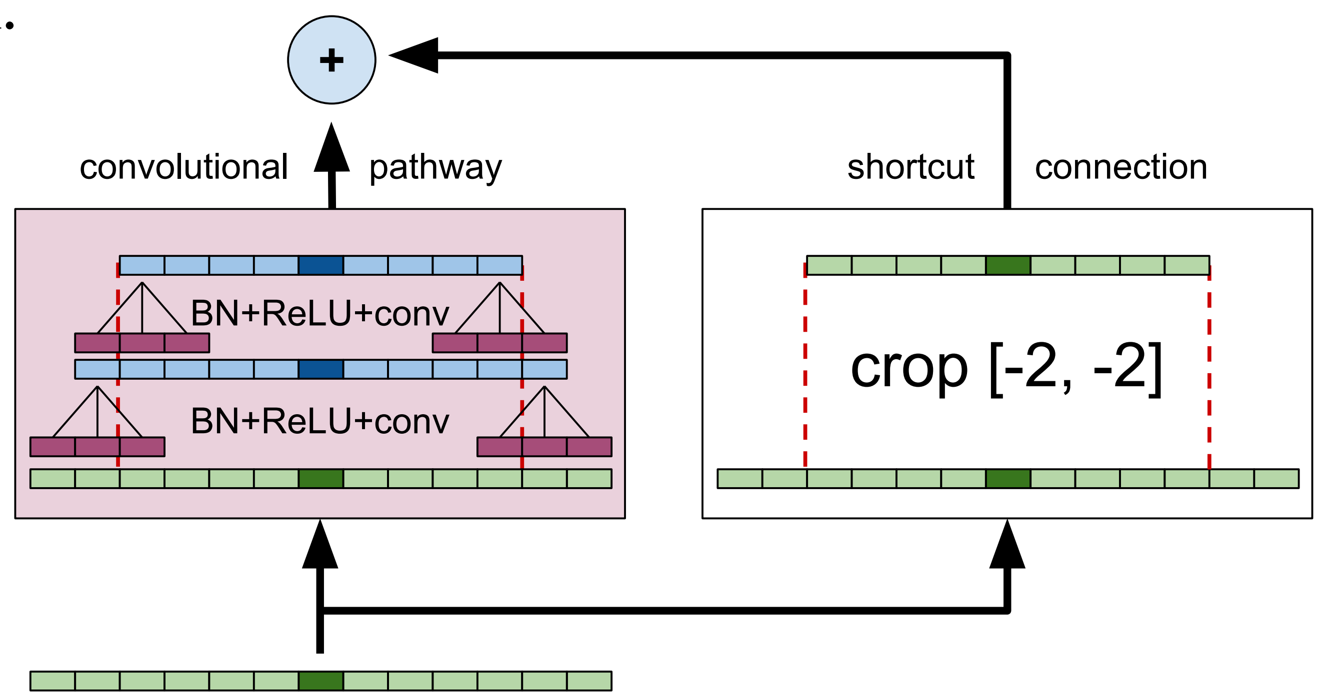


Figure 2. Residual learning: a building block.



ResNet Architectures and Results

	(a)	(b)	(c)	(d)
Summary	Bottleneck 1-3333	1-3333 NoTimestride	1-2222 Timestride	1-3333 Timestride
# param	64.3 M	67.1 M	60.8 M	67.1 M
Input	$3 \times 64 \times 31$	$3 \times 64 \times 55$	$3 \times 64 \times 56$	$3 \times 64 \times 76$
Stage 0 64x32xT	conv5x5, 64 maxpool (2x1)	conv5x5, 64 maxpool (2x1)	conv5x5, 64 maxpool (2x1)	conv5x5, 64 maxpool (2x1)
Stage 1 (64x32xT)	*initStride 1x1* 3x [conv 1x1, 64 conv 3x3, 64 conv 1x1, 256]	*initStride 1x1* 3x [conv 3x3, 64 conv 3x3, 64]	*initStride 1x1* 2x [conv 3x3, 64 conv 3x3, 64]	*initStride 1x1* 3x [conv 3x3, 64 conv 3x3, 64]
Stage 2 (128x16xT)	*initStride 2x1* 3x [conv 1x1, 128 conv 3x3, 128 conv 1x1, 512]	*initStride 2x1* 3x [conv 3x3, 128 conv 3x3, 128]	*initStride 2x1* 2x [conv 3x3, 128 conv 3x3, 128]	*initStride 2x1* 3x [conv 3x3, 128 conv 3x3, 128]
Stage 3 (256x8xT)	*initStride 2x1* 3x [conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024]	*initStride 2x1* 3x [conv 3x3, 256 conv 3x3, 256]	*initStride 2x1* 2x [conv 3x3, 256 conv 3x3, 256]	*initStride 2x1* 3x [conv 3x3, 256 conv 3x3, 256]
Stage 4 (512x4xT)	*initStride 2x1* 3x [conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048] maxpool (2x1)	*initStride 2x1* 3x [conv 3x3, 512 conv 3x3, 512] maxpool (2x1)	*initStride **2x2*** 2x [conv 3x3, 512 conv 3x3, 512] maxpool (**2x2**)	*initStride **2x2*** 3x [conv 3x3, 512 conv 3x3, 512] maxpool (**2x2**)
Output	3x FC 2084 FC 1024 FC 32k	3x FC 2084 FC 1024 FC 32k	3x FC 2084 FC 1024 FC 32k	3x FC 2084 FC 1024 FC 32k
(XE-300) SWB	11.8	11.2	11.3	11.4
(XE) SWB		9.7	9.5	9.2
(ST) SWB		8.6	8.7	8.3
(ST) CH		15.5	15.0	14.9
(ST) RT'02		13.4	13.3	13.1
(ST) RT'03		13.1	12.7	12.7
(ST) RT'04		12.1	12.0	11.9
(ST) DEV'04f		11.3	11.1	11.2

Model combination

Model	SWB	CH	RT'02	RT'03	RT'04	DEV'04f
LSTM1 (SA-MTL)	7.6	13.6	11.5	11.0	10.7	10.1
LSTM2 (Feat. fusion)	7.2	12.7	10.7	10.2	10.1	9.6
ResNet	7.6	14.5	12.2	12.2	11.5	11.1
ResNet+LSTM2	6.8	12.2	10.2	10.0	9.7	9.4
ResNet+LSTM1+LSTM2	6.7	12.1	10.1	10.0	9.7	9.2

LSTM Language Models

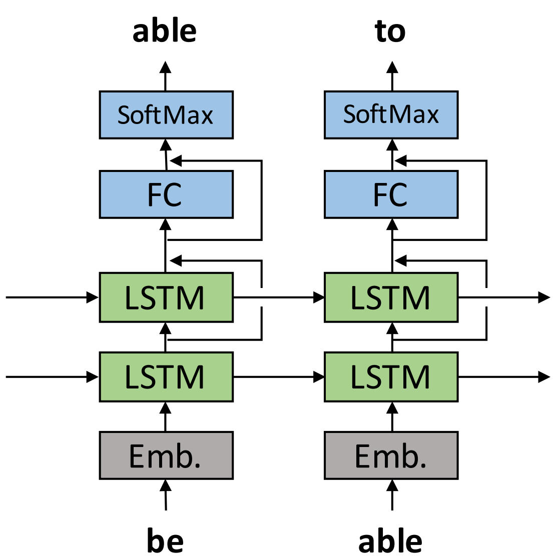


Figure 3: *Word-LSTM*

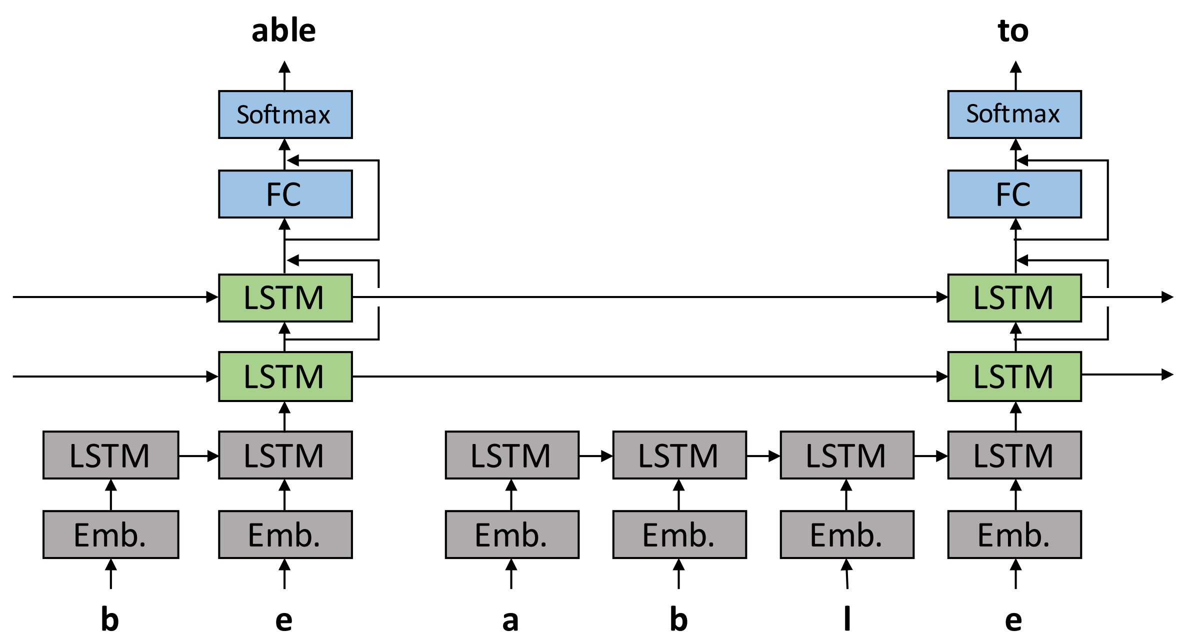


Figure 4: *Char-LSTM*

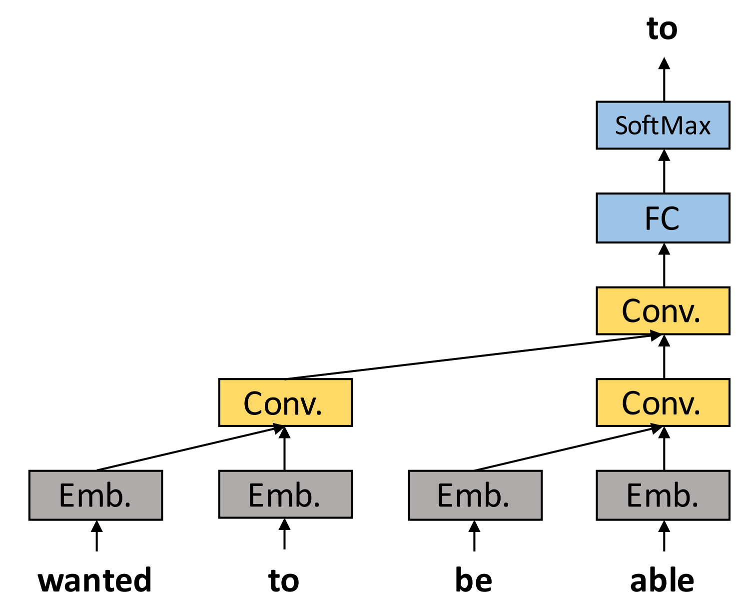


Figure 5: *Word-DCC*

Results with different LMs

	WER [%]	
	SWB	CH
n-gram	6.7	12.1
n-gram + model-M	6.1	11.2
n-gram + model-M + Word-LSTM	5.6	10.4
n-gram + model-M + Char-LSTM	5.7	10.6
n-gram + model-M + Word-LSTM-MTL	5.6	10.3
n-gram + model-M + Char-LSTM-MTL	5.6	10.4
n-gram + model-M + Word-DCC	5.8	10.8
n-gram + model-M + 4 LSTMs + DCC	**5.5**	**10.3**

Conclusions

- Acoustic model improvements
  - deep bidirectional LSTM, with feature fusion
  - deep residual networks
- Language modelling
  - recurrent and convolutional networks
  - word-based and character-based
- Parity with human performance not yet reached