Learning representations of speech in neural network acoustic models

Steve Renals

Centre for Speech Technology Research
University of Edinburgh

s.renals@ed.ac.uk

3 December 2014

Learning representations of speech in neural network acoustic models

Pawel Świętojański, Peter Bell, and Steve Renals

Centre for Speech Technology Research
University of Edinburgh

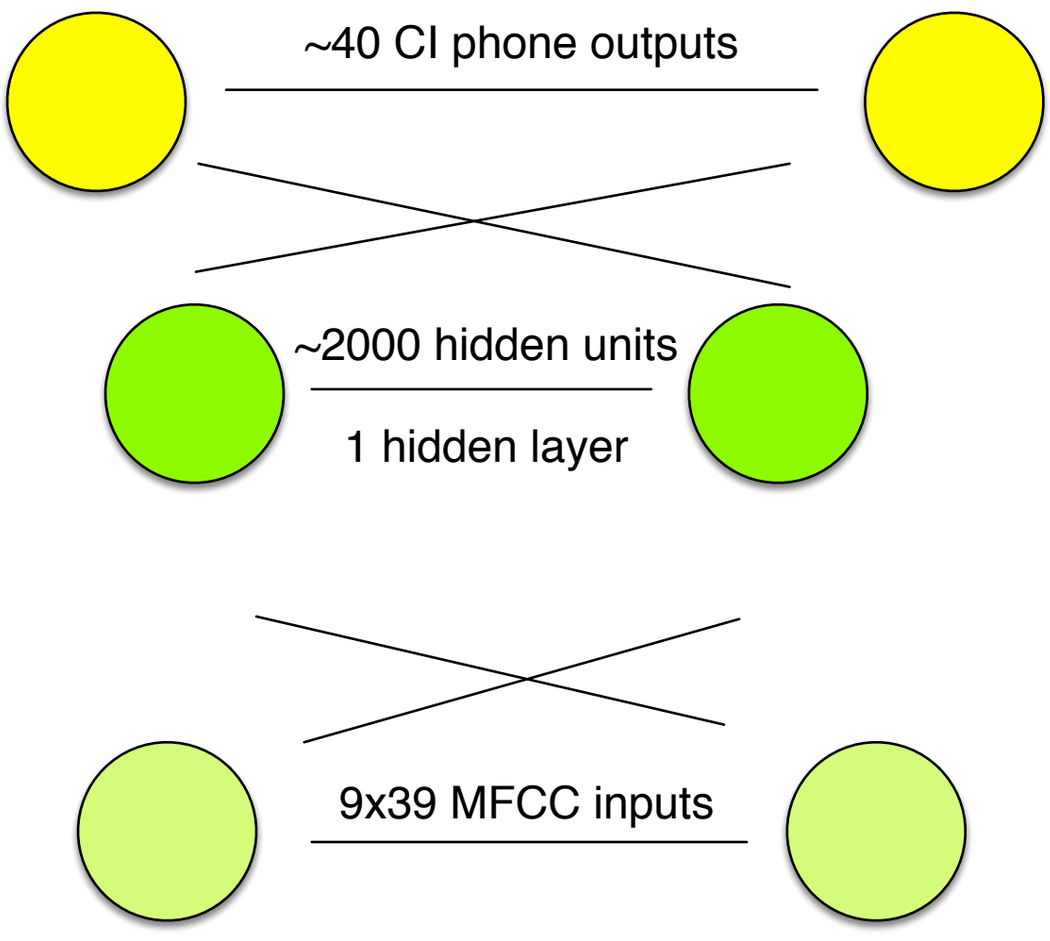
s.renals@ed.ac.uk

3 December 2014

Introduction

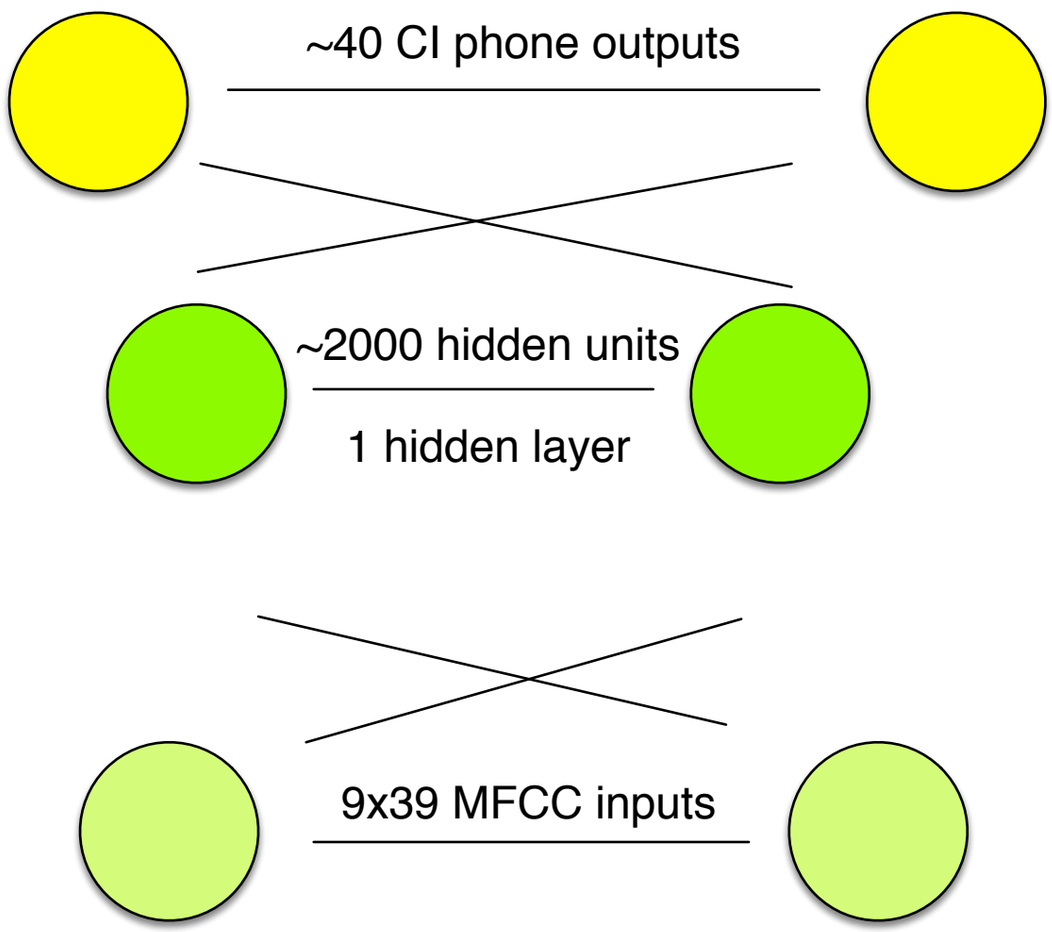
Neural network acoustic models (1990s)

Neural network acoustic models (1990s)

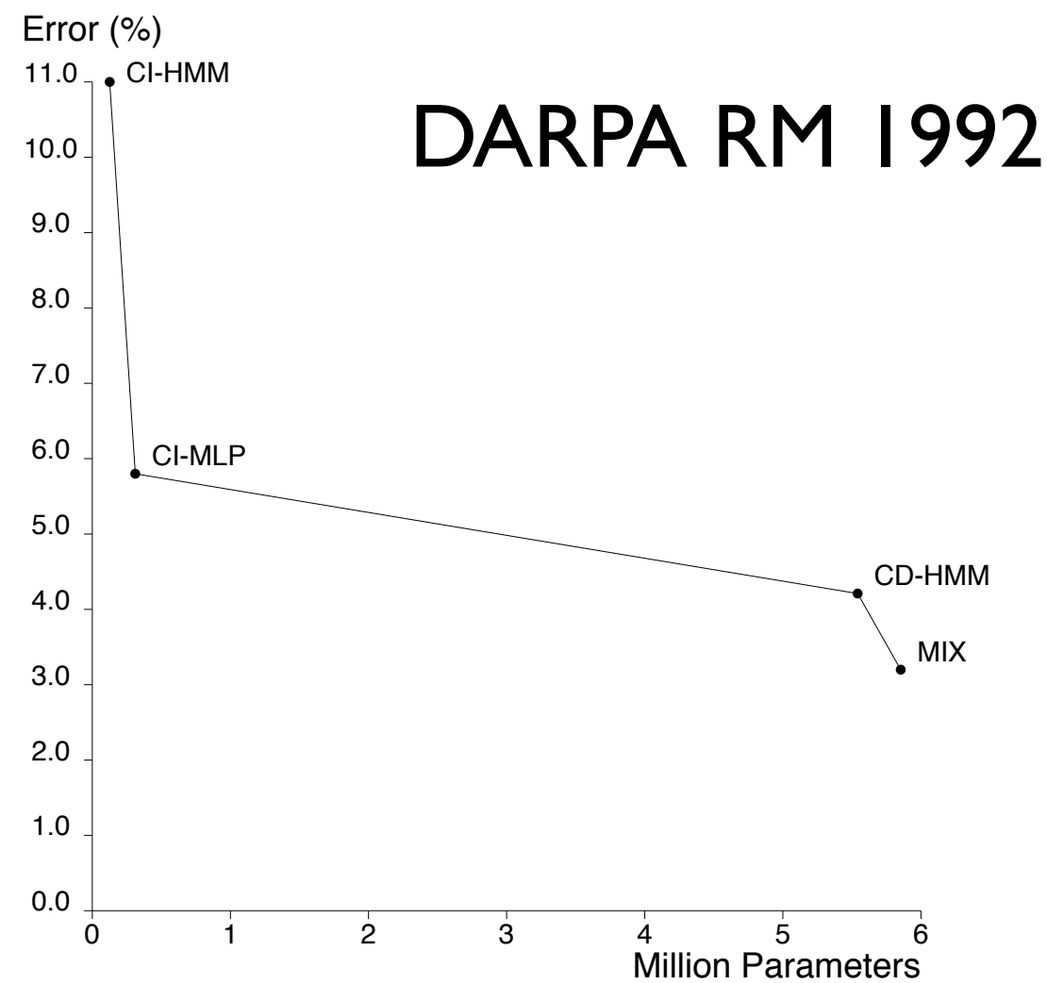


Bourlard & Morgan, 1994

Neural network acoustic models (1990s)

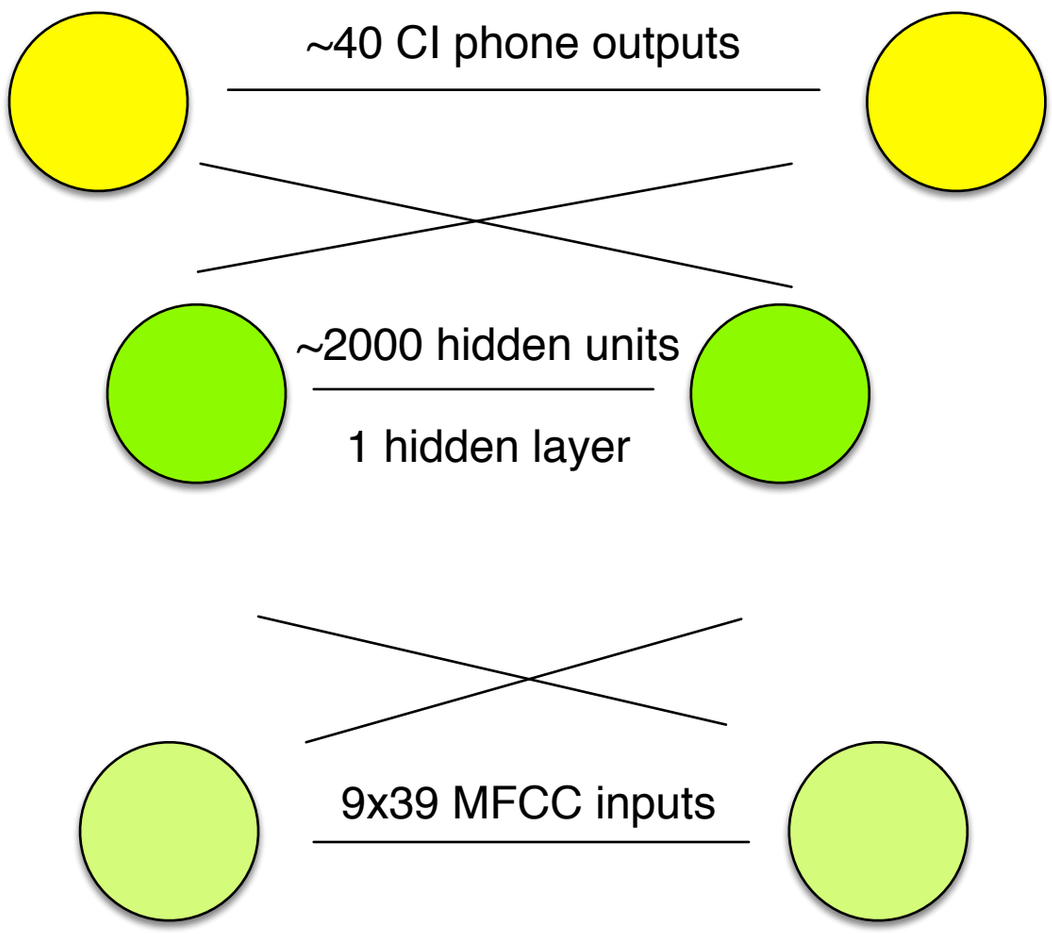


Bourlard & Morgan, 1994

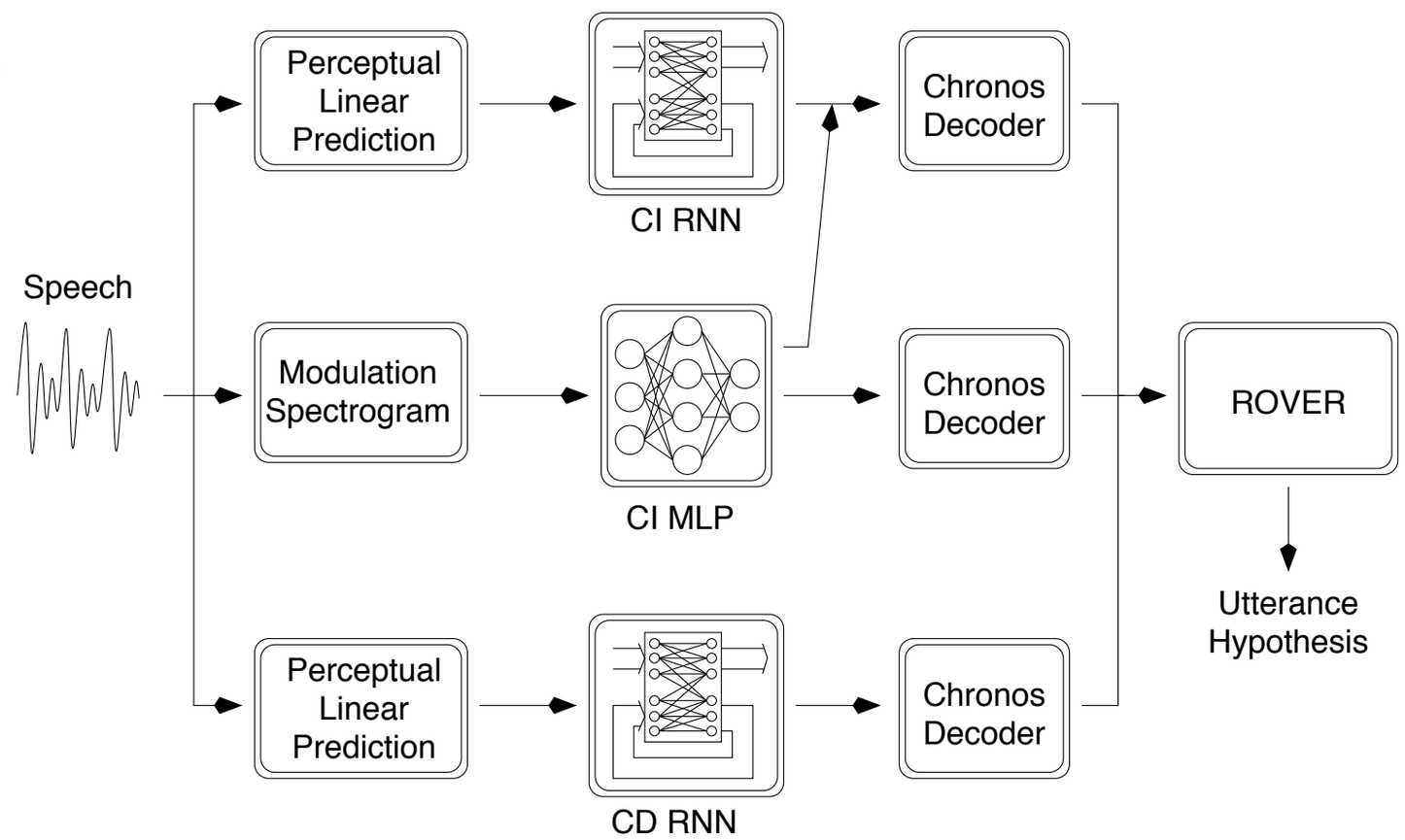


Renals, Morgan, Cohen & Franco, ICASSP 1992

Neural network acoustic models (1990s)



Bourlard & Morgan, 1994



Broadcast news 1998

20.8% WER

(best GMM-based system, 13.5%)

Cook, Christie, Ellis, Fosler-Lussier, Gotoh, Kingsbury, Morgan, Renals, Robinson, & Williams, DARPA, 1999

NN acoustic models Limitations vs GMMs

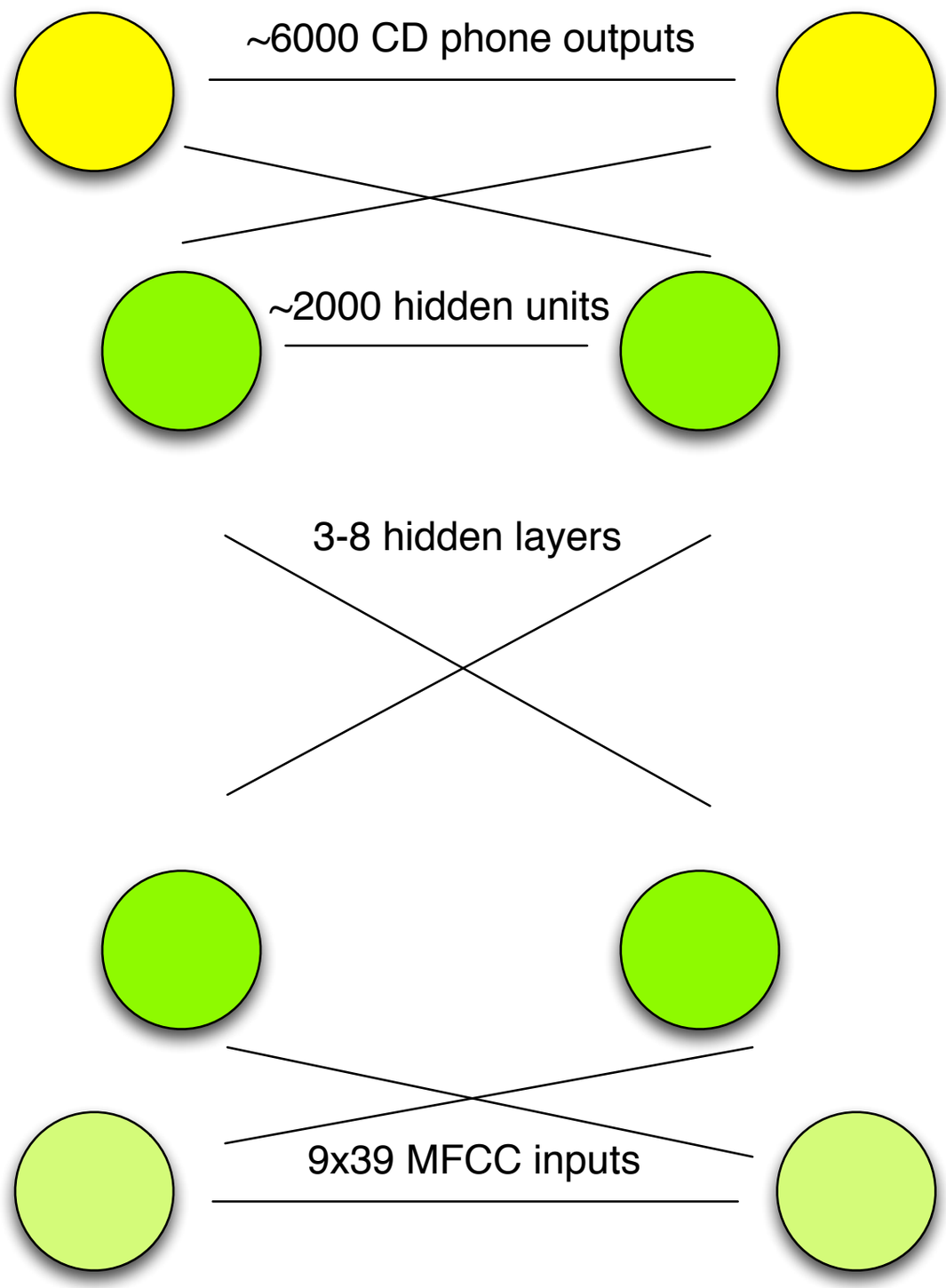
- Computationally restricted to monophone outputs
  - CD-RNN factored over multiple networks – limited within-word context
- Training not easily parallelisable
  - experimental turnaround slower
  - systems less complex (fewer parameters)
    - RNN – <100k parameters
    - MLP – ~1M parameters
- Rapid adaptation hard (cf MLLR)

NN acoustic models
Benefits

- Fewer limitations on inputs
  - Correlated features
  - Multi-frame windows
- Discriminative training criteria (frame level and sequence level)
- Can be used to generate ‘higher-level’ features
  - tandem, posteriorgrams
  - bottleneck features

(Deep) neural network acoustic models (2010s)

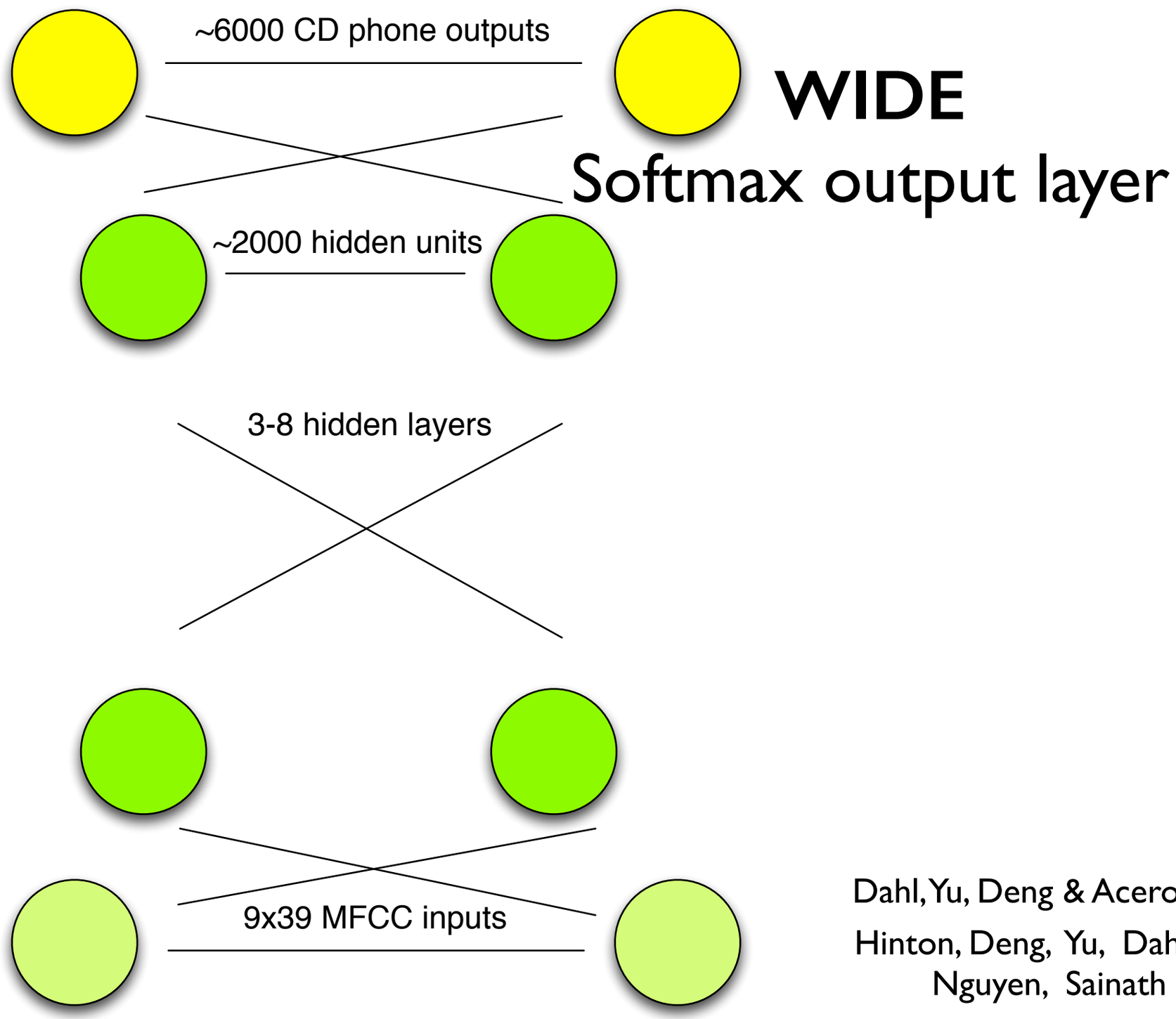
(Deep) neural network acoustic models (2010s)



Dahl, Yu, Deng & Acero, IEEE TASLP 2012

Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, IEEE SP Mag 2012

(Deep) neural network acoustic models (2010s)



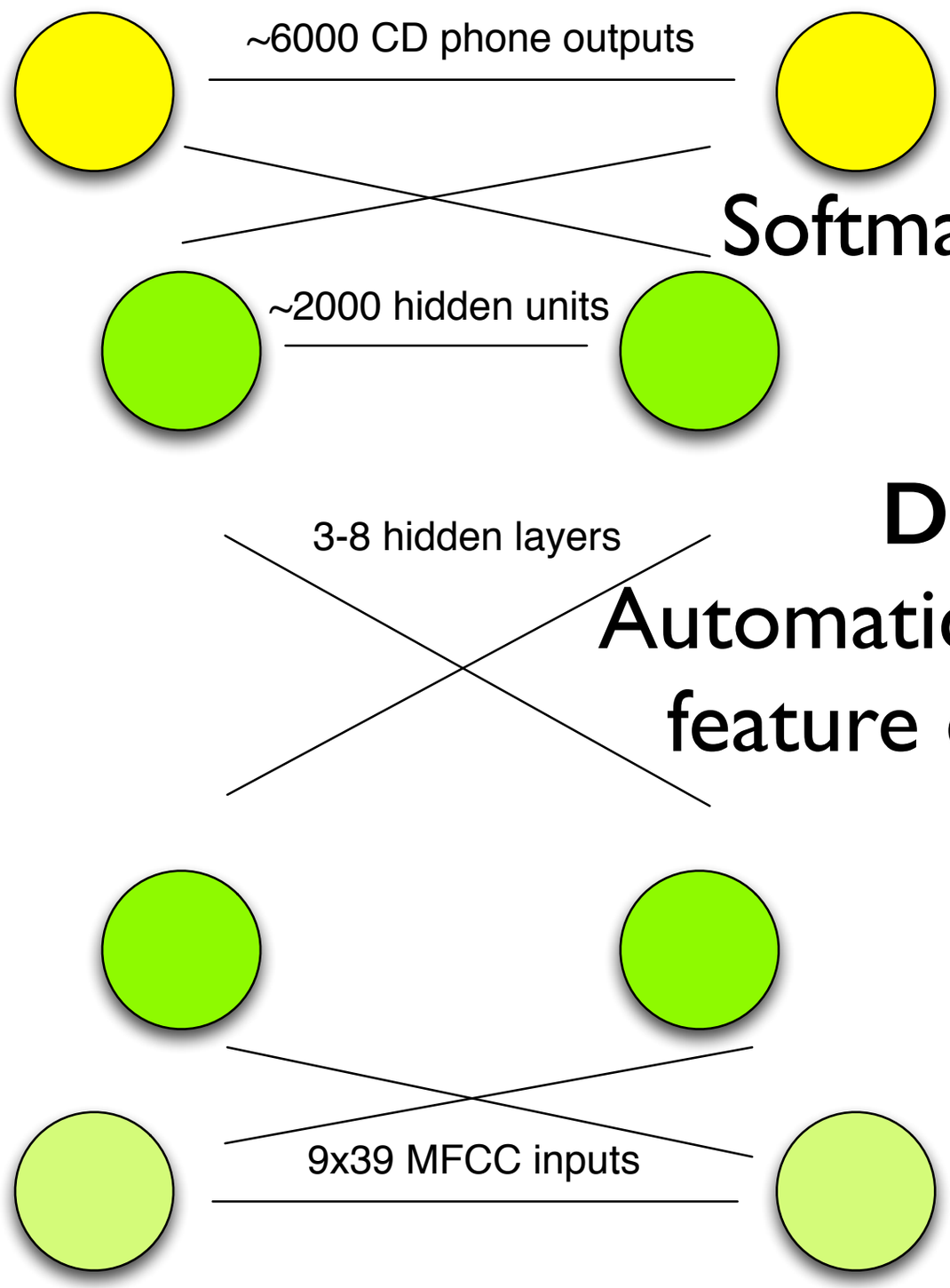
WIDE

Softmax output layer

Dahl, Yu, Deng & Acero, IEEE TASLP 2012

Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, IEEE SP Mag 2012

(Deep) neural network acoustic models (2010s)



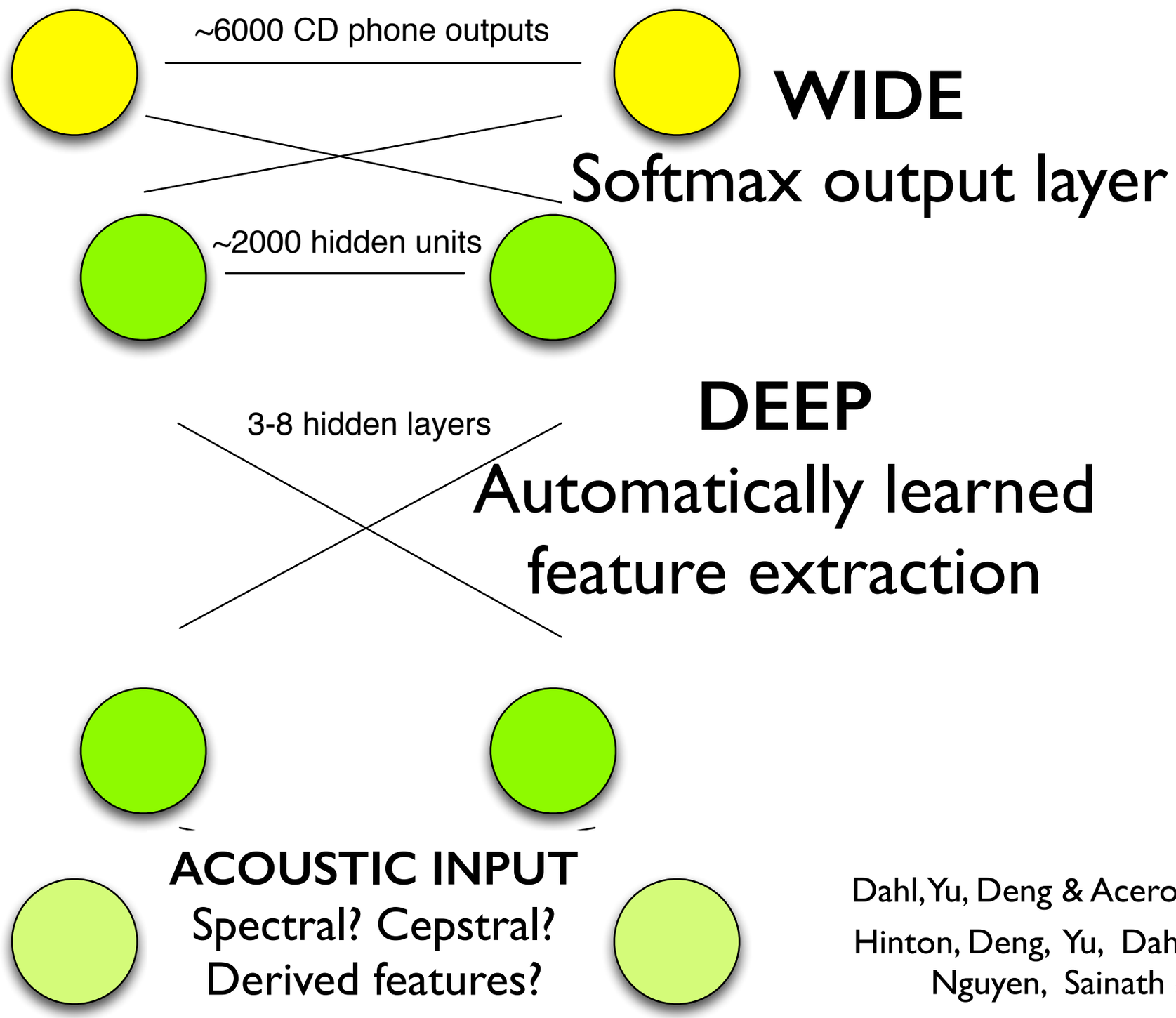
WIDE
Softmax output layer

DEEP
Automatically learned feature extraction

Dahl, Yu, Deng & Acero, IEEE TASLP 2012

Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, IEEE SP Mag 2012

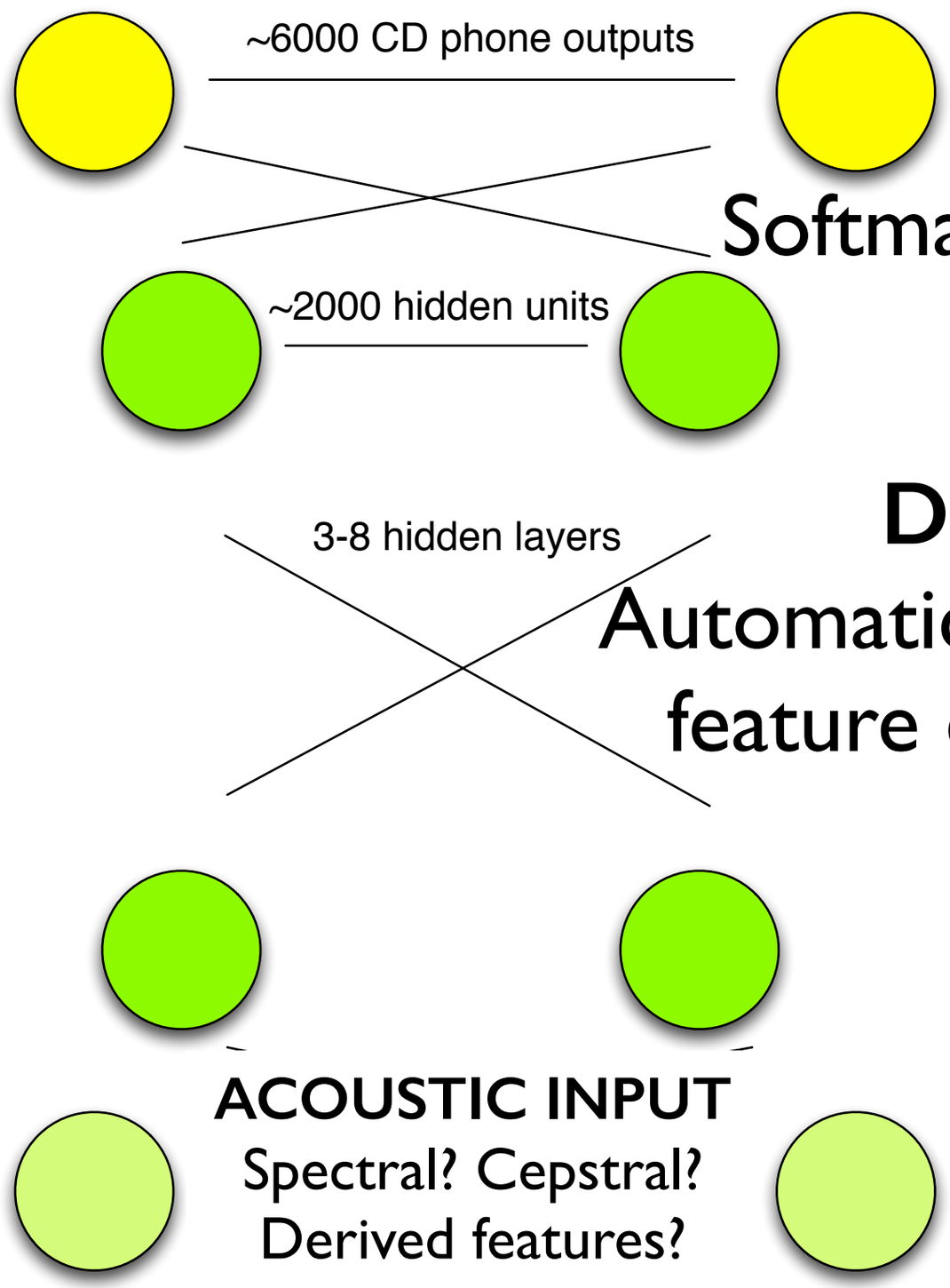
(Deep) neural network acoustic models (2010s)



Dahl, Yu, Deng & Acero, IEEE TASLP 2012

Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, IEEE SP Mag 2012

(Deep) neural network acoustic models (2010s)

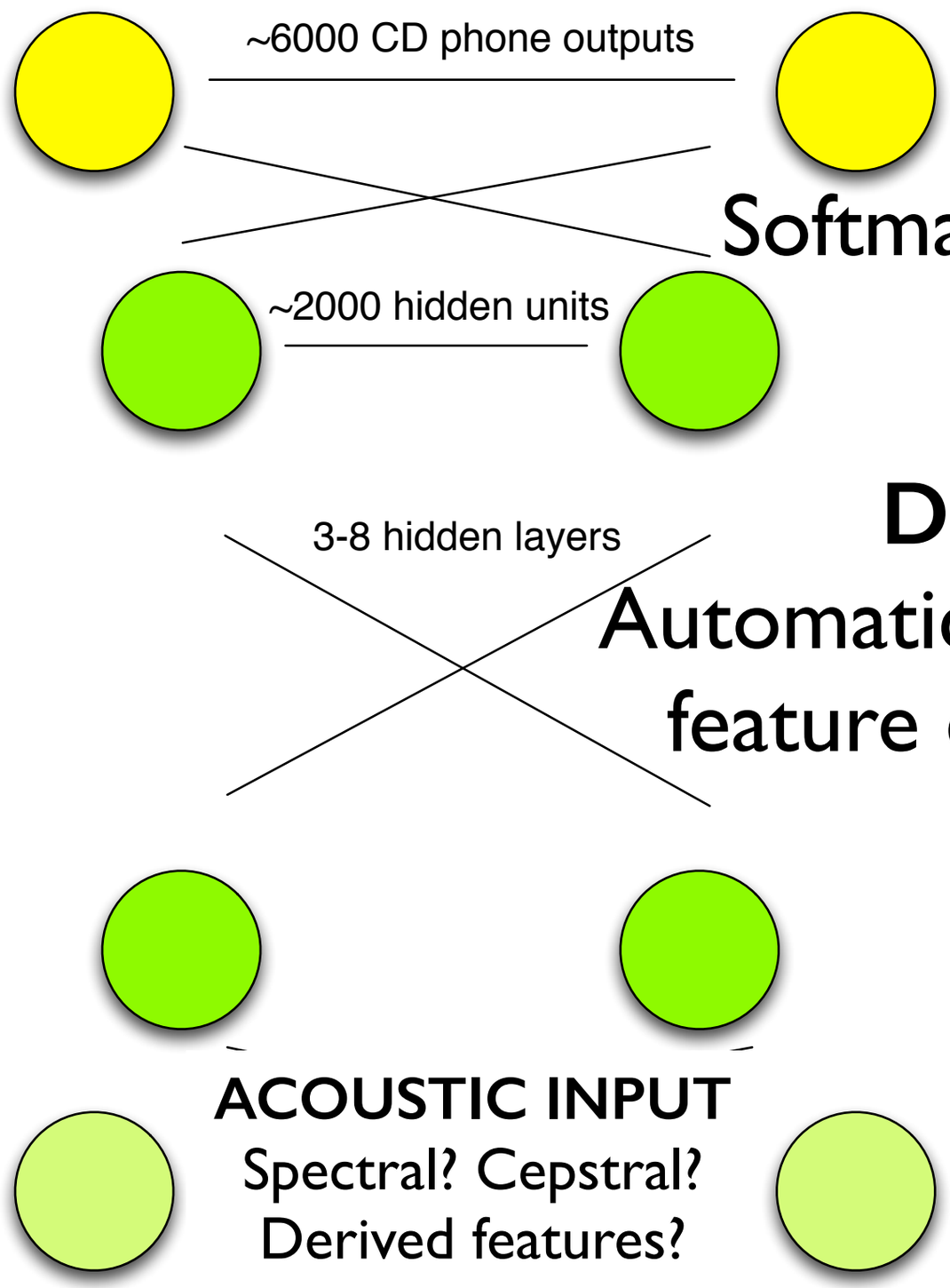


WIDE
Softmax output layer

DEEP
Automatically learned feature extraction

ACTIVATION FUNCTIONS
– pooling, RELU, gated units

(Deep) neural network acoustic models (2010s)



WIDE

Softmax output layer

DEEP

Automatically learned feature extraction

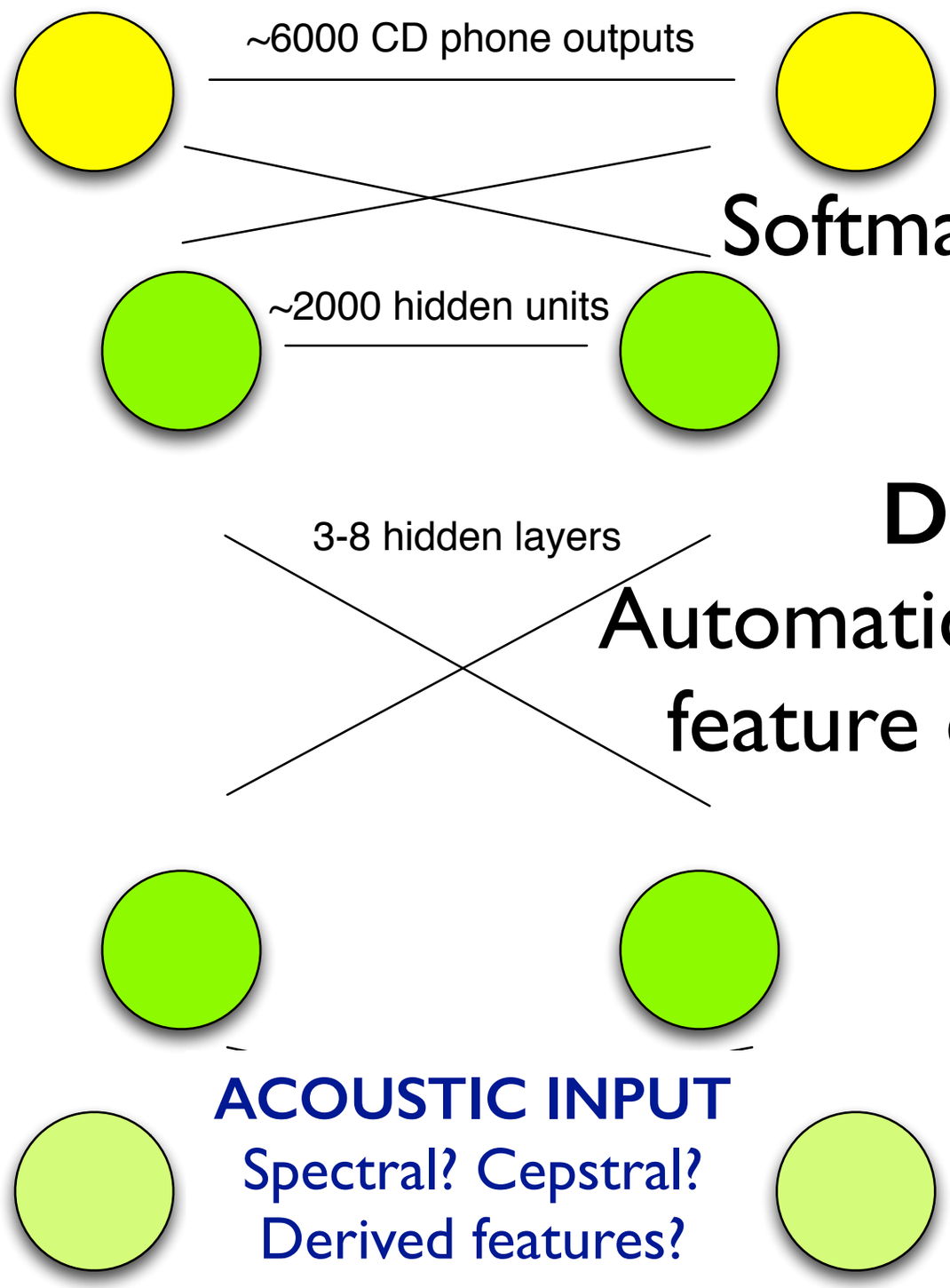
WEIGHT SHARING

– adaptation, CNNs

ACTIVATION FUNCTIONS

– pooling, RELU, gated units

(Deep) neural network acoustic models (2010s)



WIDE

Softmax output layer

DEEP

Automatically learned feature extraction

WEIGHT SHARING

– adaptation, CNNs

ACTIVATION FUNCTIONS

– pooling, RELU, gated units

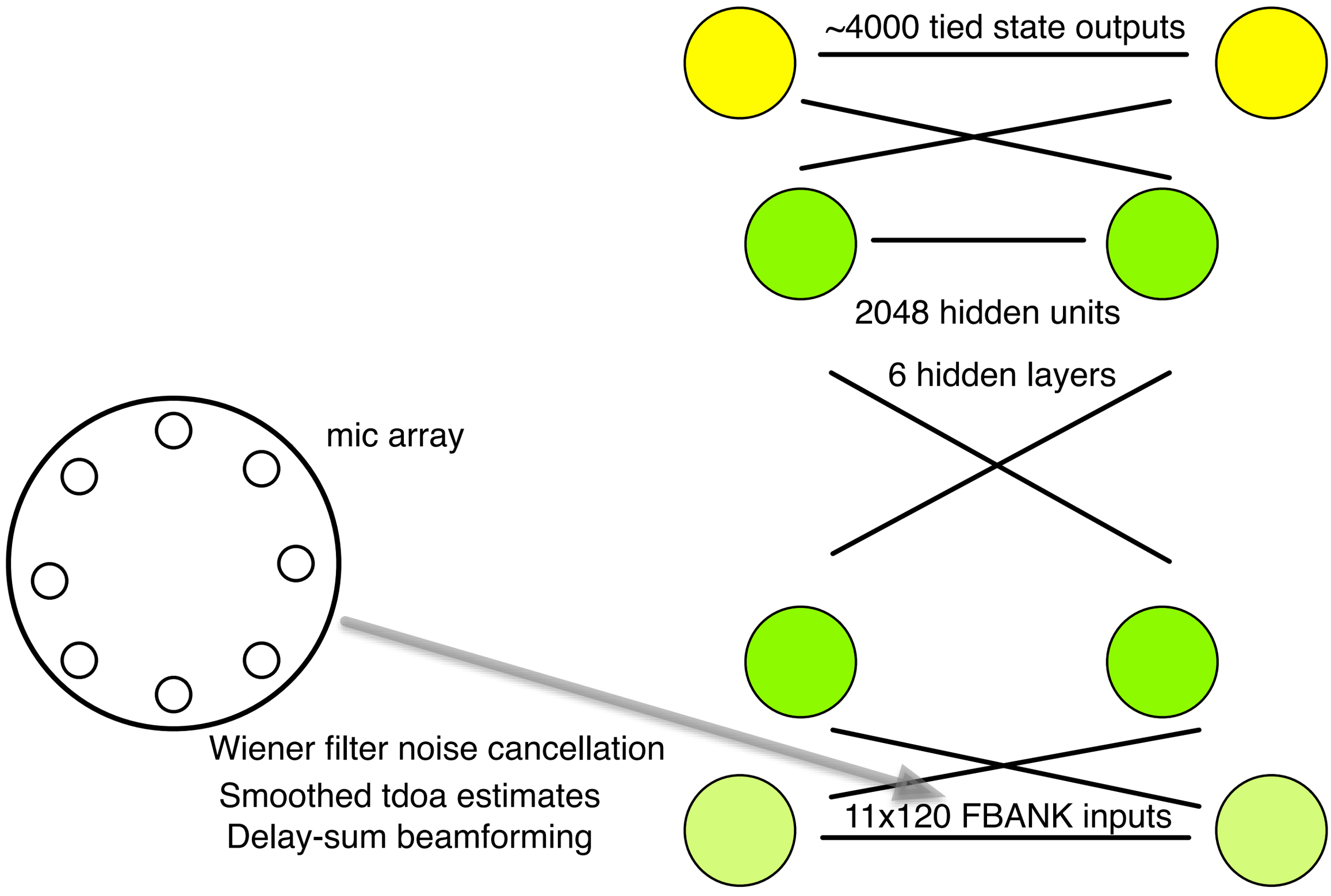
Representation learning for acoustic models

- DISTANT SPEECH RECOGNITION
  - Combining input channels from multiple microphones
  - Convolutional Neural Networks
- SPEAKER ADAPTATION
  - Model-based adaptation of NN acoustic model – unsupervised, compact, efficient, accurate
  - Learning Hidden Unit Contributions (LHUC)
- DOMAIN ADAPTATION
  - Limited in-domain data
  - Multi-Level Adaptive Networks (MLAN)

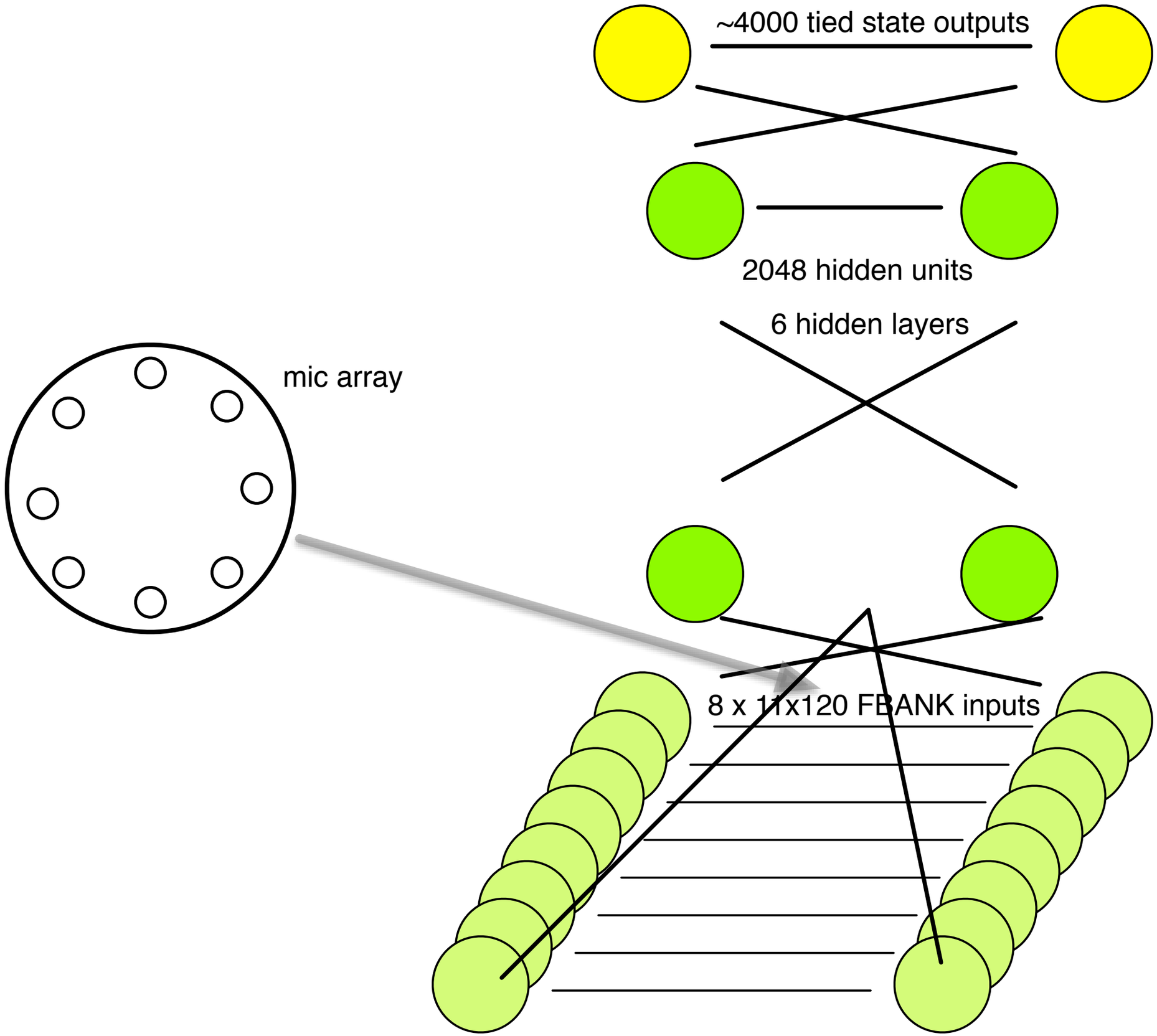
Part One

Distant Speech Recognition

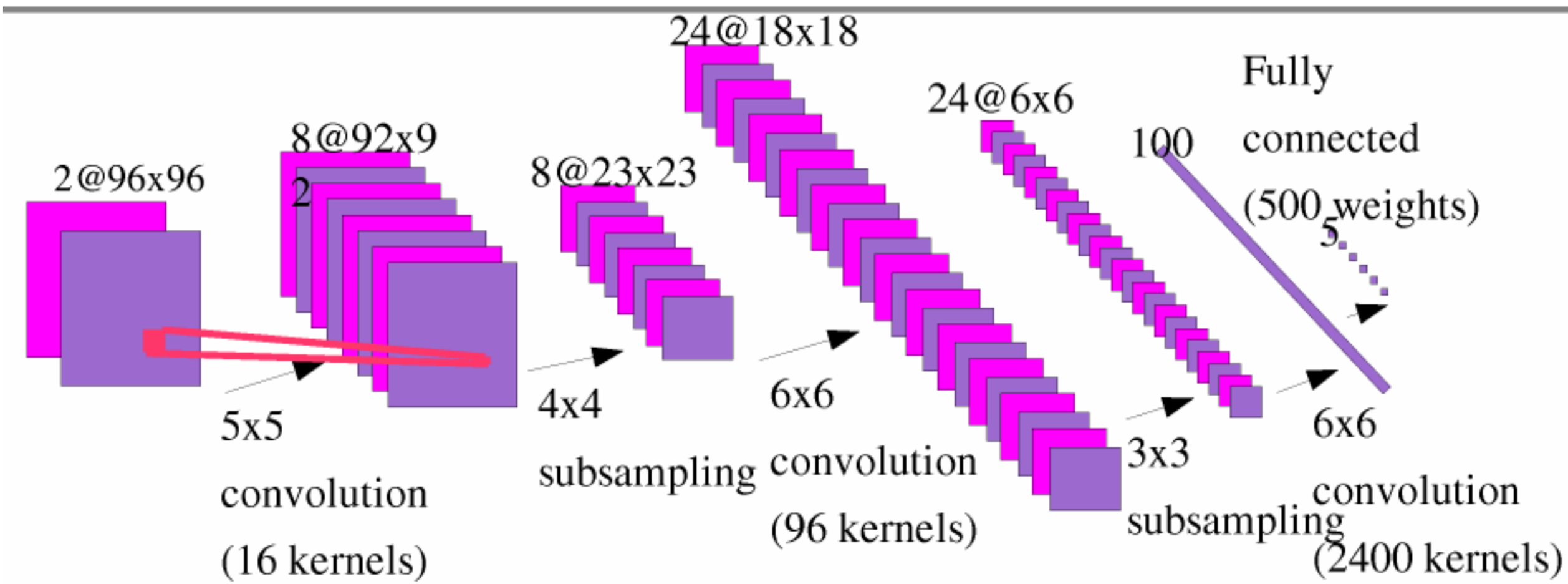
Distant speech recognition



Distant speech recognition

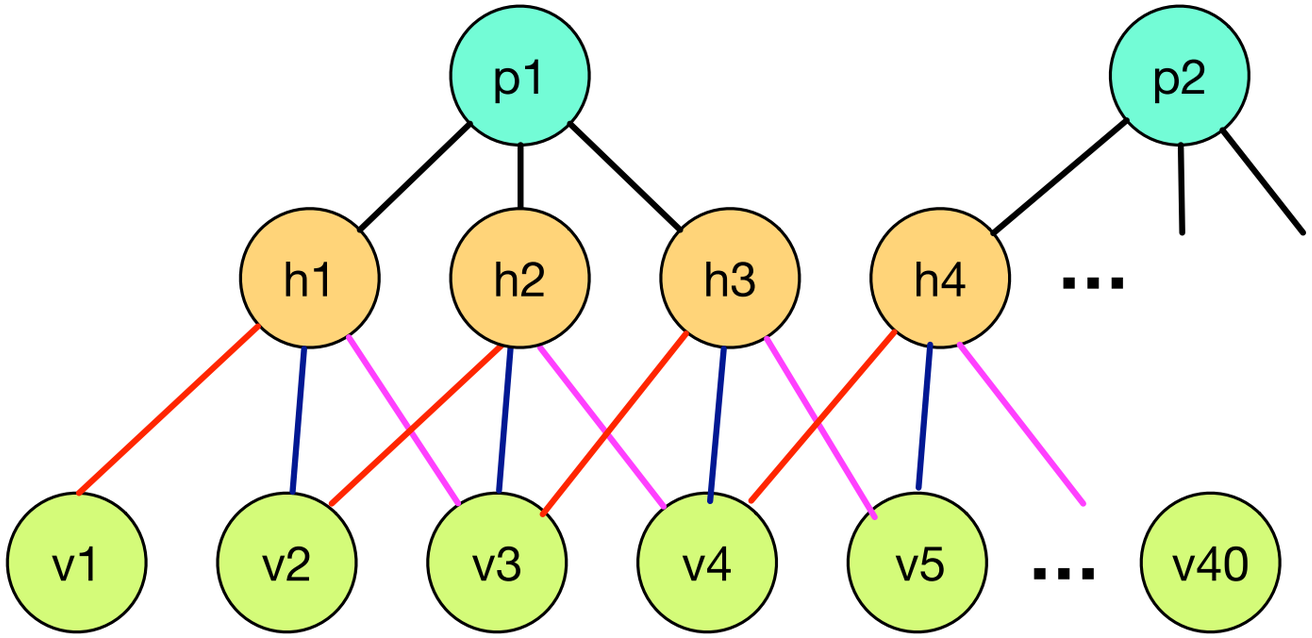


Convolutional Neural Networks

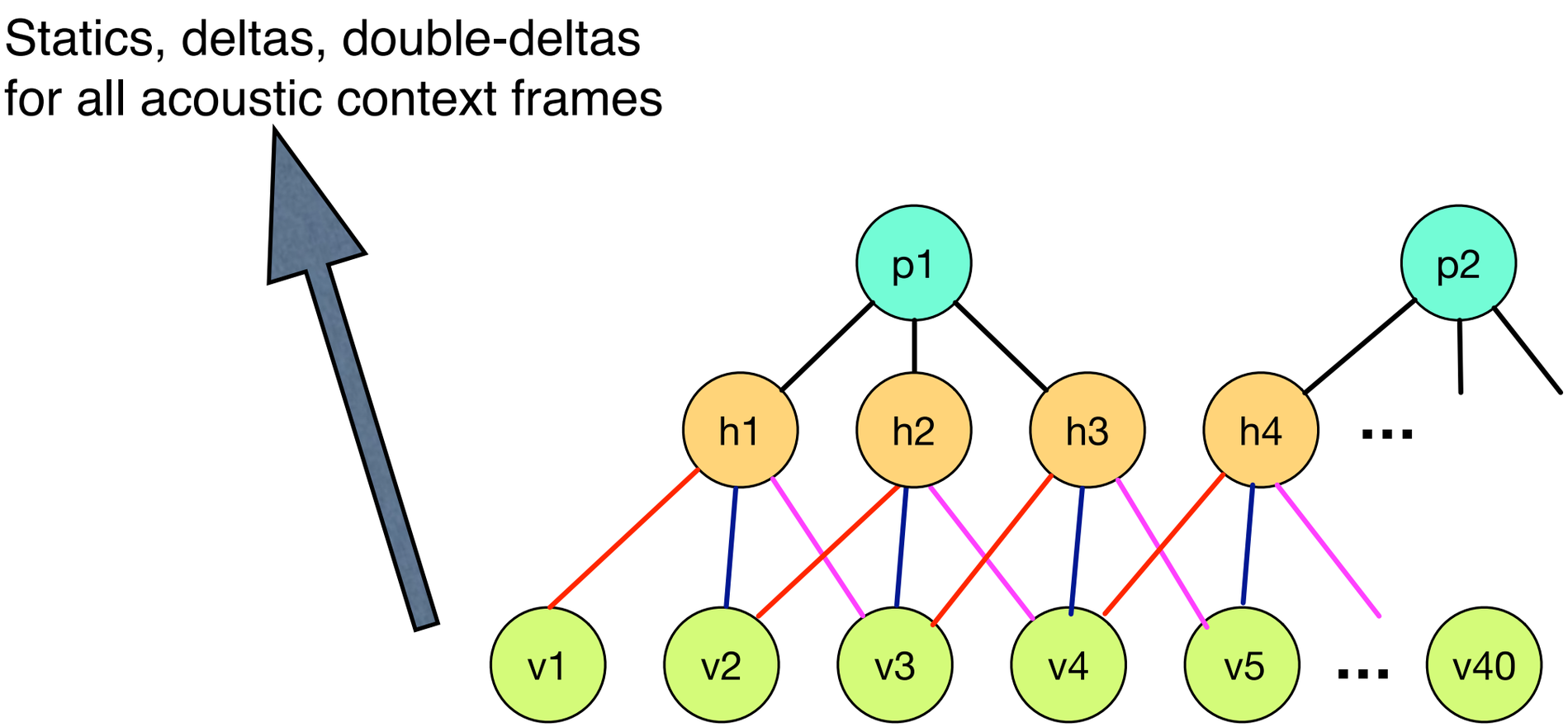


Yann Le Cun, 1989 onwards

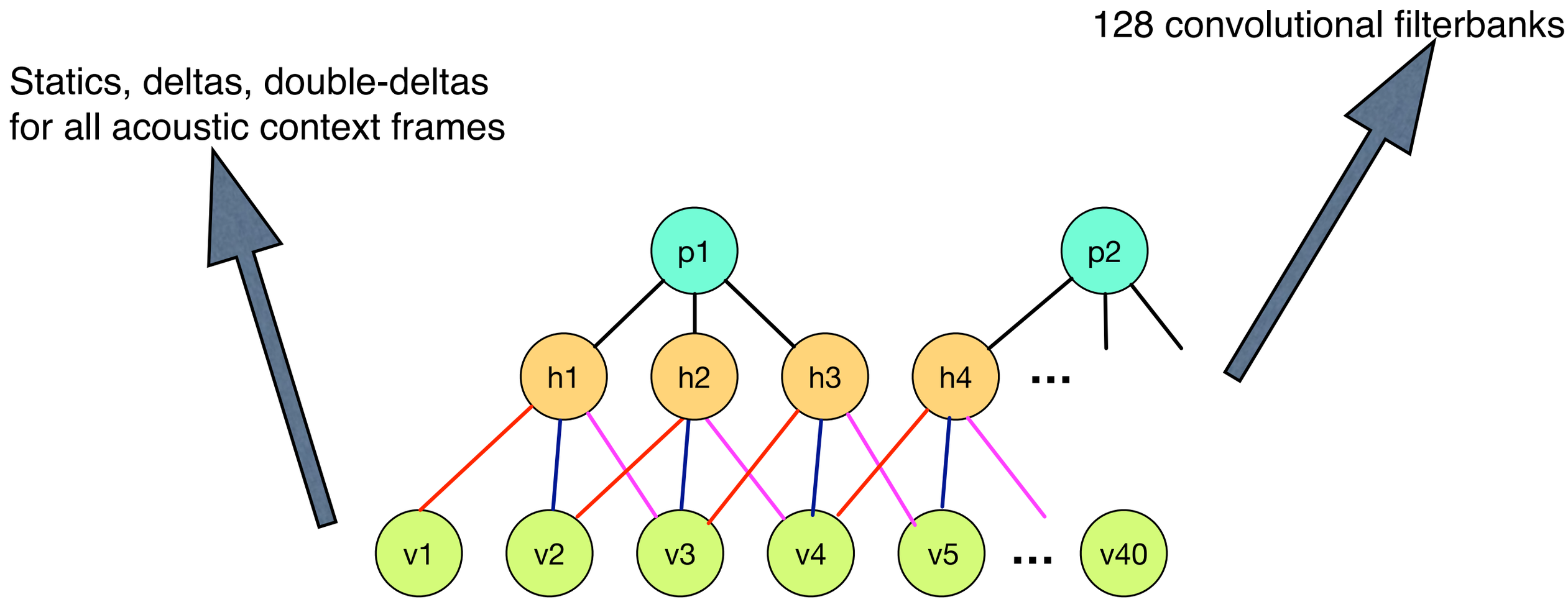
CNN acoustic model



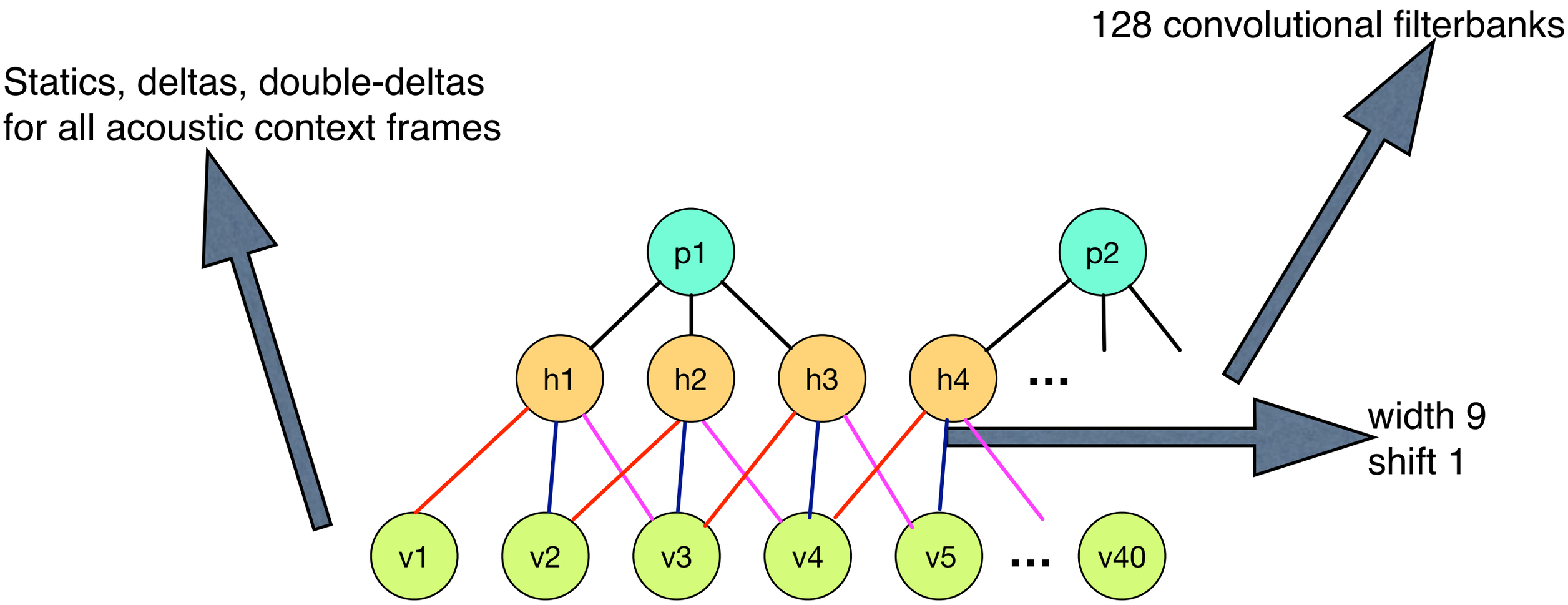
CNN acoustic model



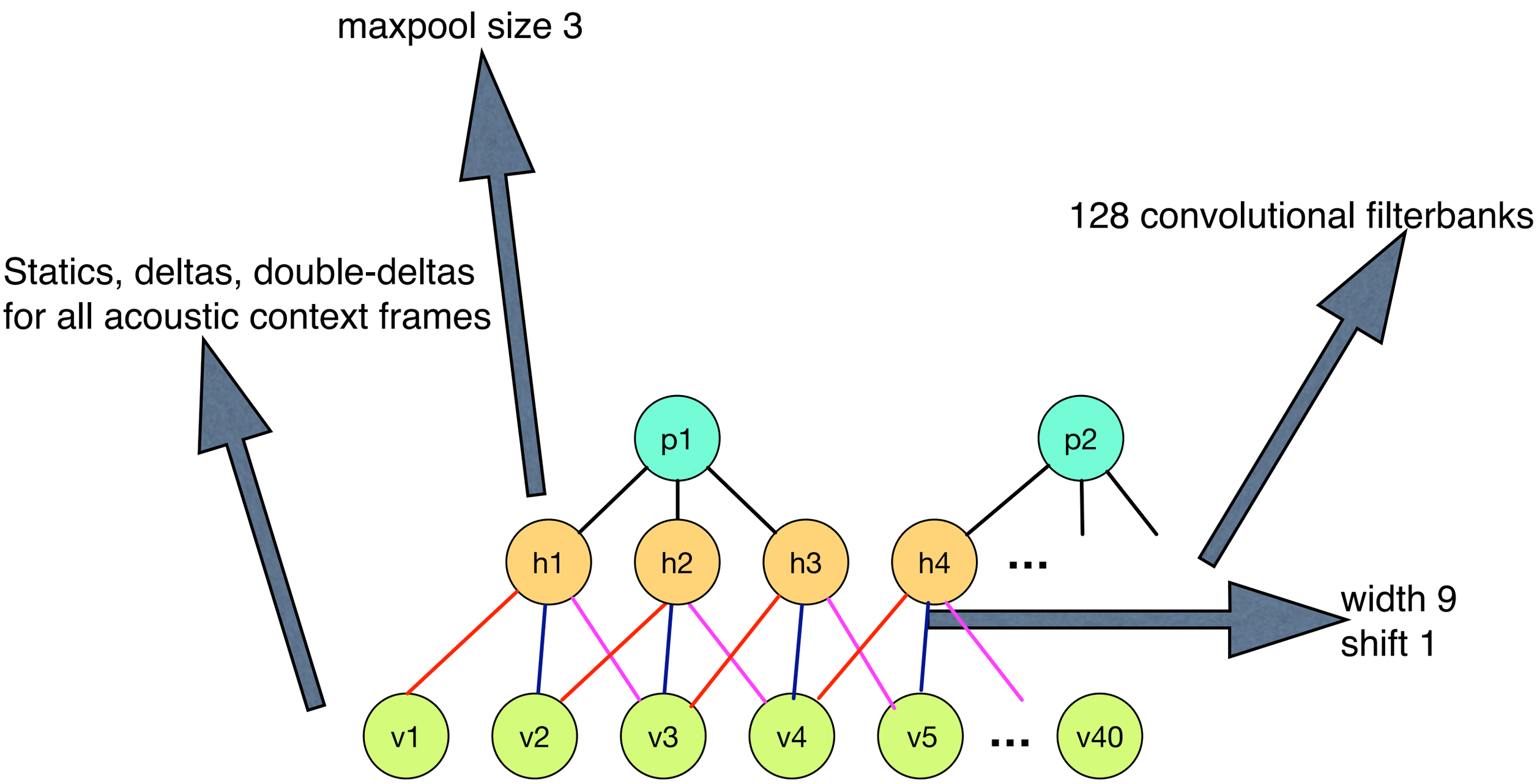
CNN acoustic model



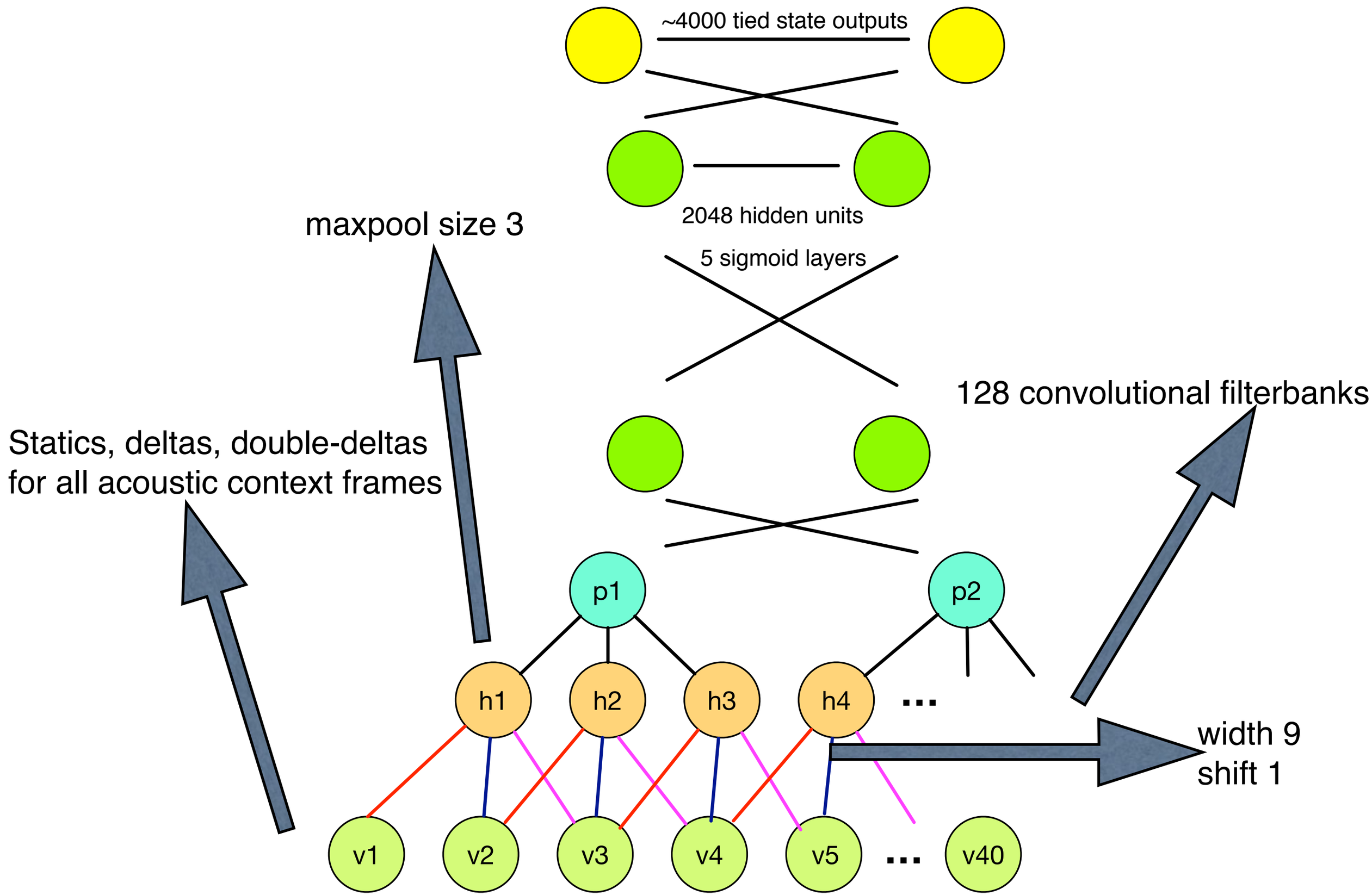
CNN acoustic model



CNN acoustic model

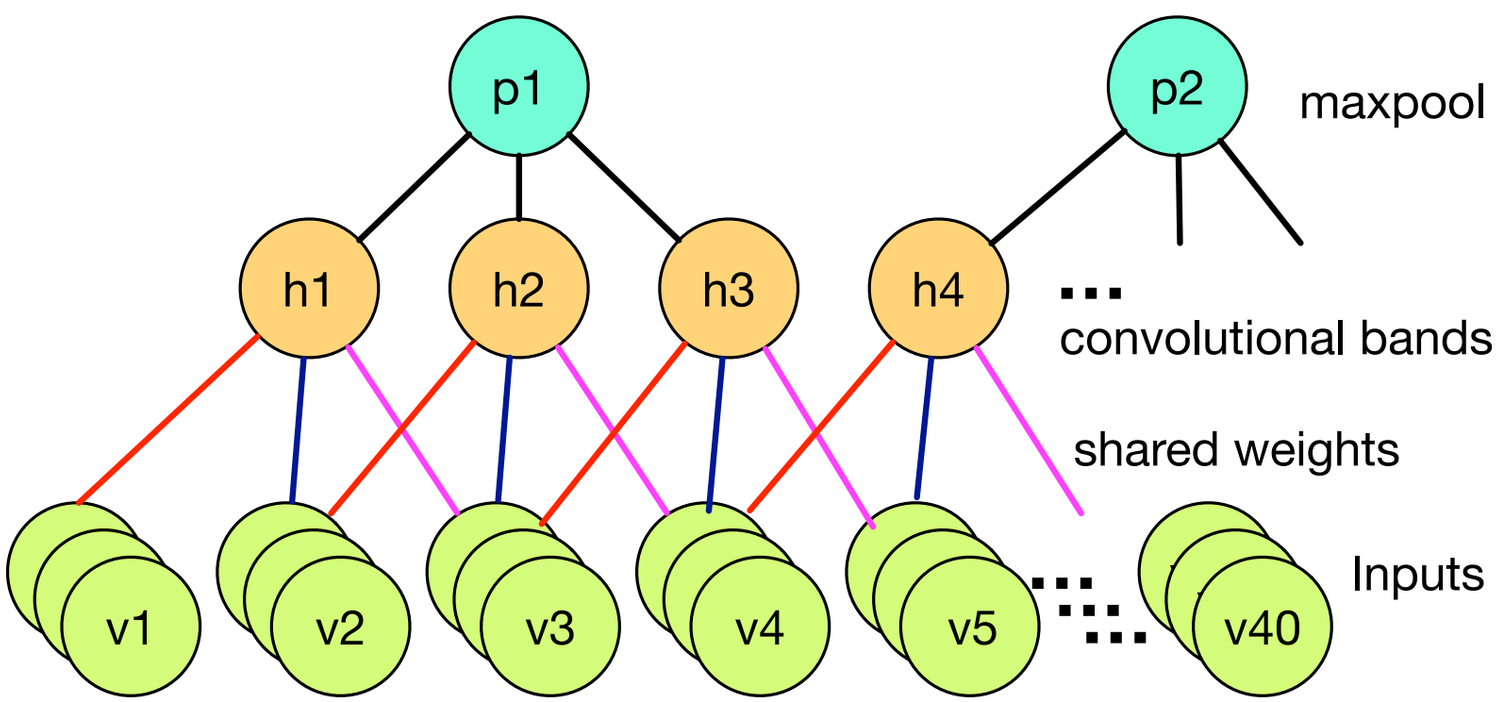


CNN acoustic model

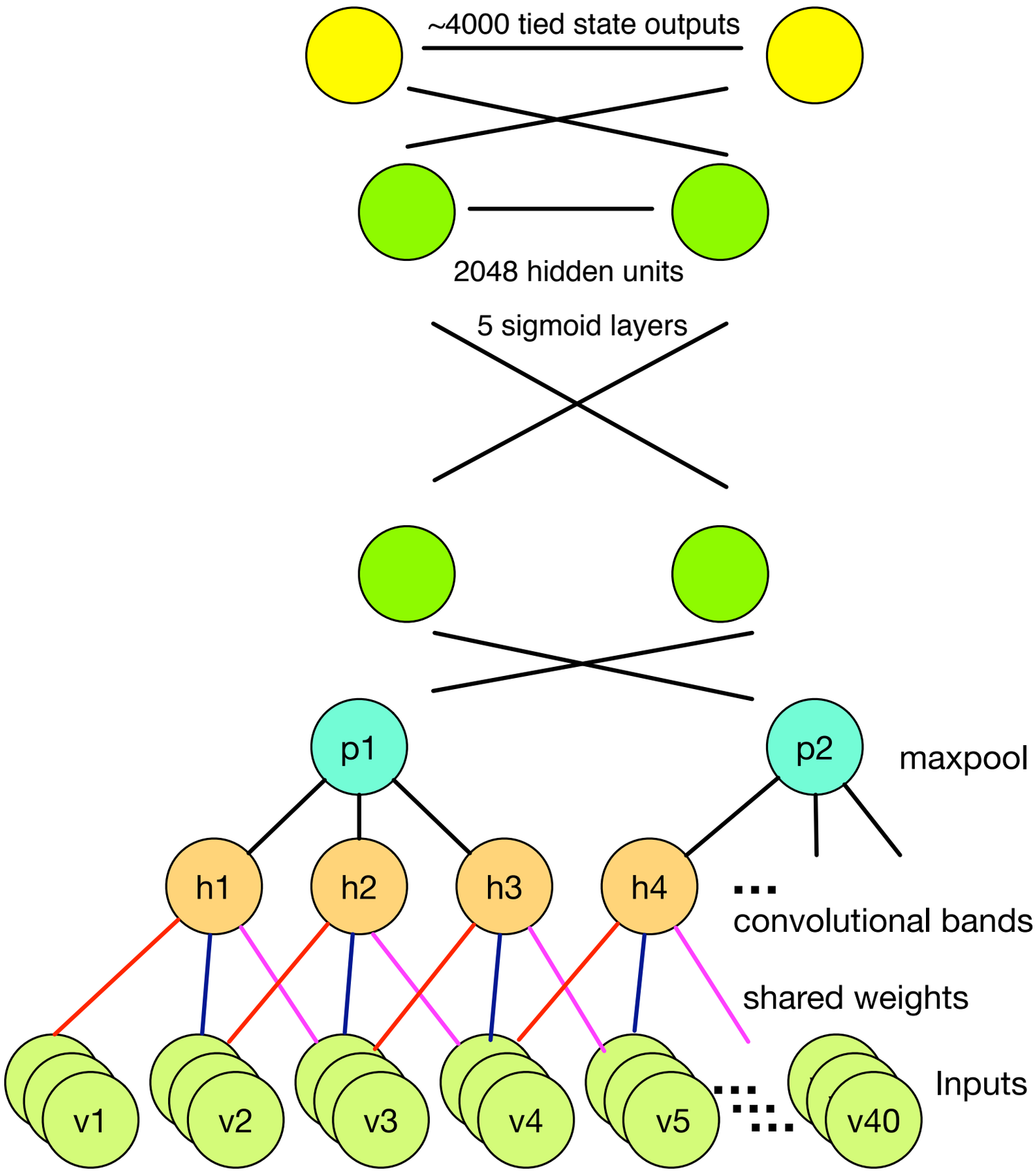


CNN – Multi-channel

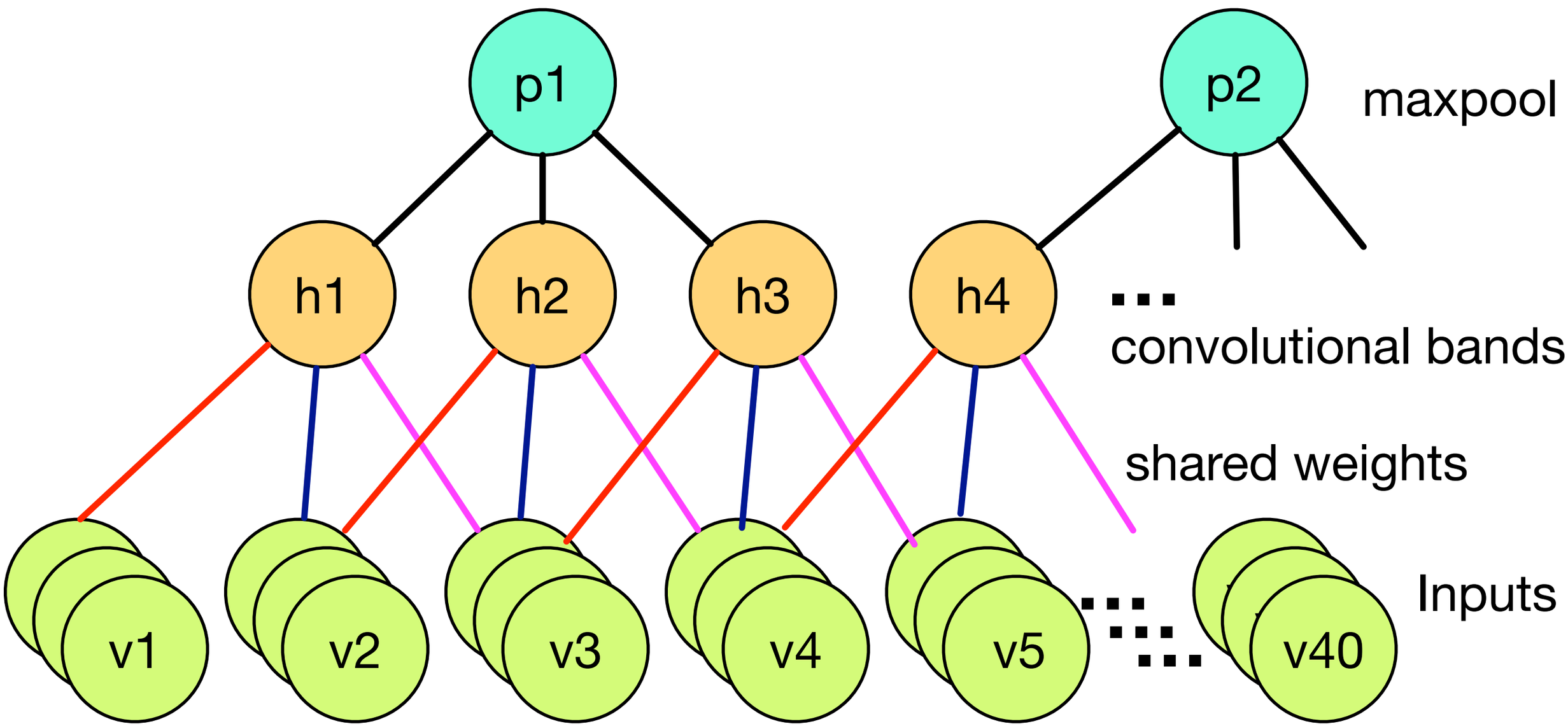
CNN – Multi-channel



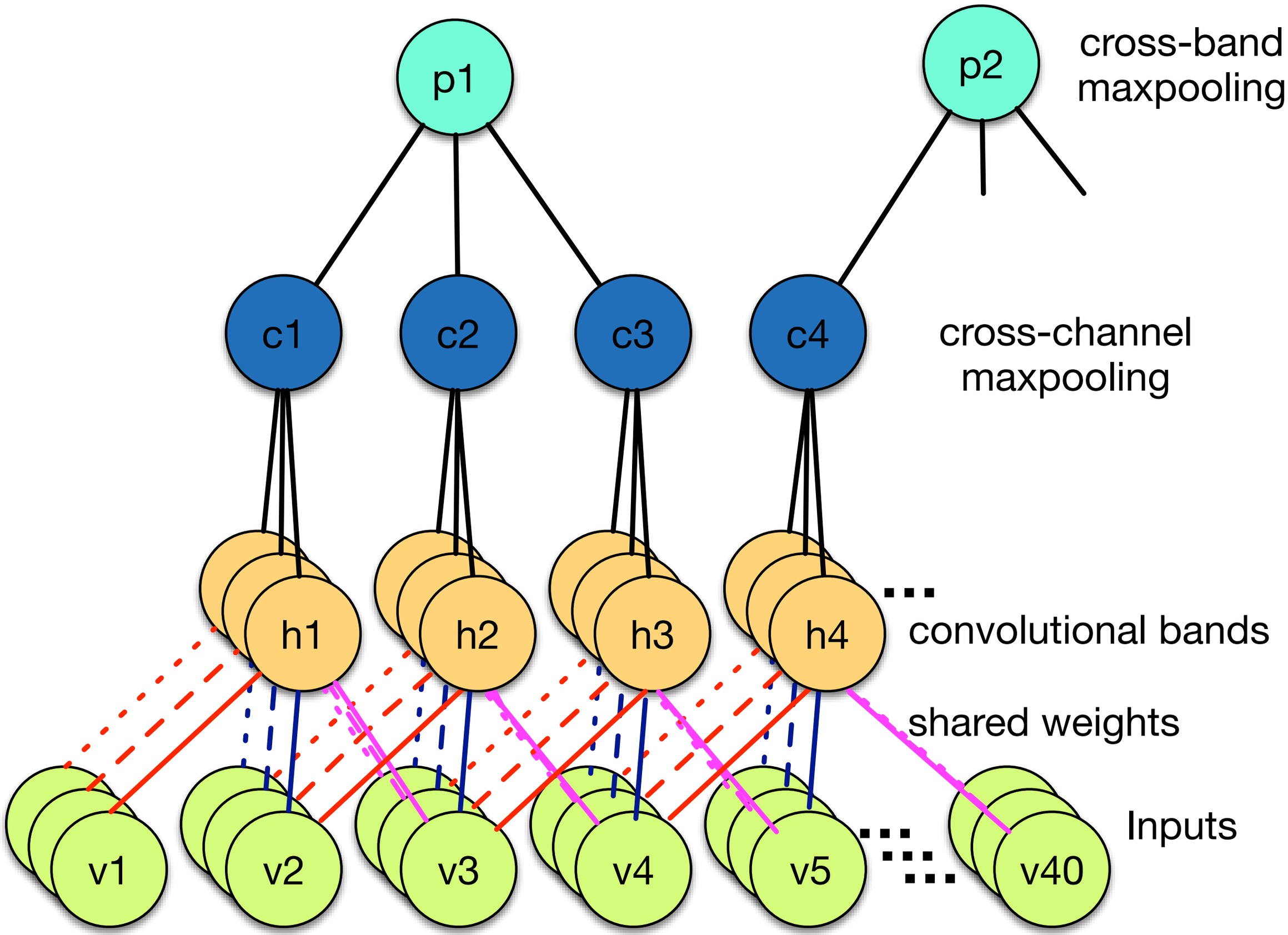
CNN – Multi-channel



CNN – Multi-channel



CNN – Cross-channel



Details?
Experimental Results?
Some Conclusions?

Details?
Experimental Results?
Some Conclusions?

Swietojanski & Renals,
Convolutional Neural Networks for
Distant Speech Recognition

Poster later today,

IEEE SPL, 21:1120-1124, 2014.

Part Two

Speaker Adaptation

Speaker/Channel Adaptation

Speaker/Channel Adaptation

- GMM-based systems
  - Maximum likelihood linear transform (MLLT) family
    - MLLR, CMLLR, …
  - MAP, Eigenvoices, CAT, …

Speaker/Channel Adaptation

- GMM-based systems
  - Maximum likelihood linear transform (MLLT) family
    - MLLR, CMLLR, …
  - MAP, Eigenvoices, CAT, …
- Adapting DNN acoustic models
  - Feature space transforms (e.g. CMLLR, LIN)
  - Auxiliary features (e.g. iVectors)
  - Model-based

Speaker/Channel Adaptation

- GMM-based systems
  - Maximum likelihood linear transform (MLLT) family
    - MLLR, CMLLR, …
  - MAP, Eigenvoices, CAT, …
- Adapting DNN acoustic models
  - Feature space transforms (e.g. CMLLR, LIN)
  - Auxiliary features (e.g. iVectors)
  - Model-based
- Ideally: COMPACT / UNSUPERVISED / EFFICIENT

Model-based NN adaptation

Model-based NN adaptation

- Adaptation of different **weight subsets** (Liao 2013)
  - 5% relative decrease in WER when all 60M weights adapted

Model-based NN adaptation

- Adaptation of different **weight subsets** (Liao 2013)
  - 5% relative decrease in WER when all 60M weights adapted
- Adaptation cost based on **KL divergence** between SI and SA output distributions (Yu et al 2013)
  - 3% relative decrease in WER on Switchboard

Model-based NN adaptation

- Adaptation of different **weight subsets** (Liao 2013)
  - 5% relative decrease in WER when all 60M weights adapted
- Adaptation cost based on **KL divergence** between SI and SA output distributions (Yu et al 2013)
  - 3% relative decrease in WER on Switchboard
- Increase compactness by **SVD factorisation** of weight matrix (Xue et al 2014)

Model-based NN adaptation

- Adaptation of different **weight subsets** (Liao 2013)
  - 5% relative decrease in WER when all 60M weights adapted
- Adaptation cost based on **KL divergence** between SI and SA output distributions (Yu et al 2013)
  - 3% relative decrease in WER on Switchboard
- Increase compactness by **SVD factorisation** of weight matrix (Xue et al 2014)
- Automatically adapt specific parameter subsets – **output biases** (Yao et al 2012), **slope and bias** of hidden units (Siniscalchi et al 2013)

Model-based NN adaptation

- Adaptation of different **weight subsets** (Liao 2013)
  - 5% relative decrease in WER when all 60M weights adapted
- Adaptation cost based on **KL divergence** between SI and SA output distributions (Yu et al 2013)
  - 3% relative decrease in WER on Switchboard
- Increase compactness by **SVD factorisation** of weight matrix (Xue et al 2014)
- Automatically adapt specific parameter subsets – **output biases** (Yao et al 2012), **slope and bias** of hidden units (Siniscalchi et al 2013)
- **Speaker codes** (Bridle & Cox 1990; Abdel-Hamid & Jiang 2013) – model-based adaptation using auxiliary features

LHUC

Learning Hidden Unit Contributions

LHUC

Learning Hidden Unit Contributions

Key idea: add a learnable
speaker-dependent amplitude
to each hidden unit

LHUC

Learning Hidden Unit Contributions

Key idea: add a learnable ***speaker-dependent amplitude*** to each hidden unit

$$\mathbf{h}_m^l = a(\mathbf{r}_m^l) \circ \sigma^l(\mathbf{W}^{l\top}\mathbf{h}_m^{l-1})$$

LHUC

Learning Hidden Unit Contributions

Key idea: add a learnable
speaker-dependent amplitude
to each hidden unit

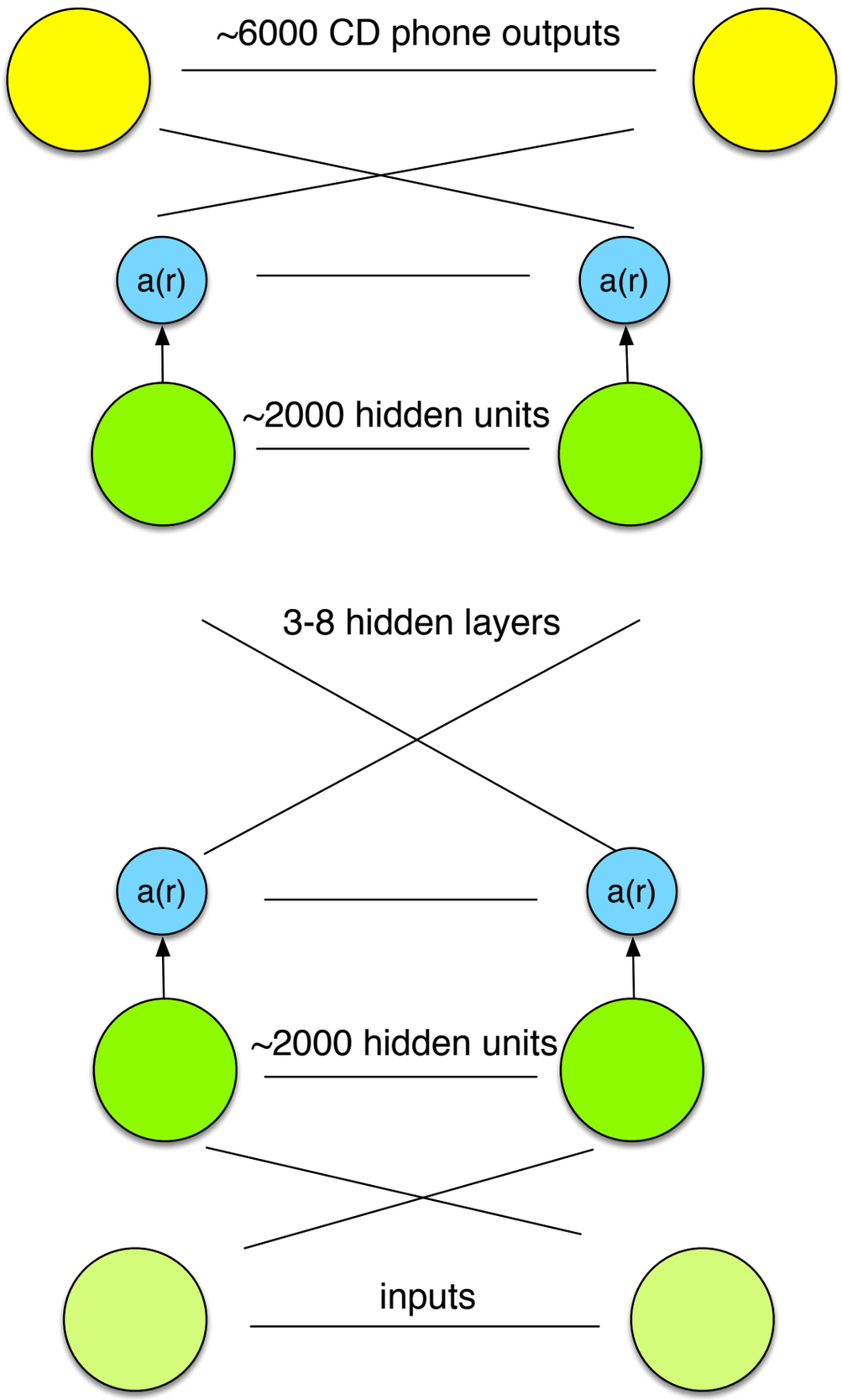
$$\mathbf{h}_m^l = a(\mathbf{r}_m^l) \circ \sigma^l(\mathbf{W}^{l\top}\mathbf{h}_m^{l-1})$$

Amplitude is a sigmoid in range 0–2,
parameterised by **r**

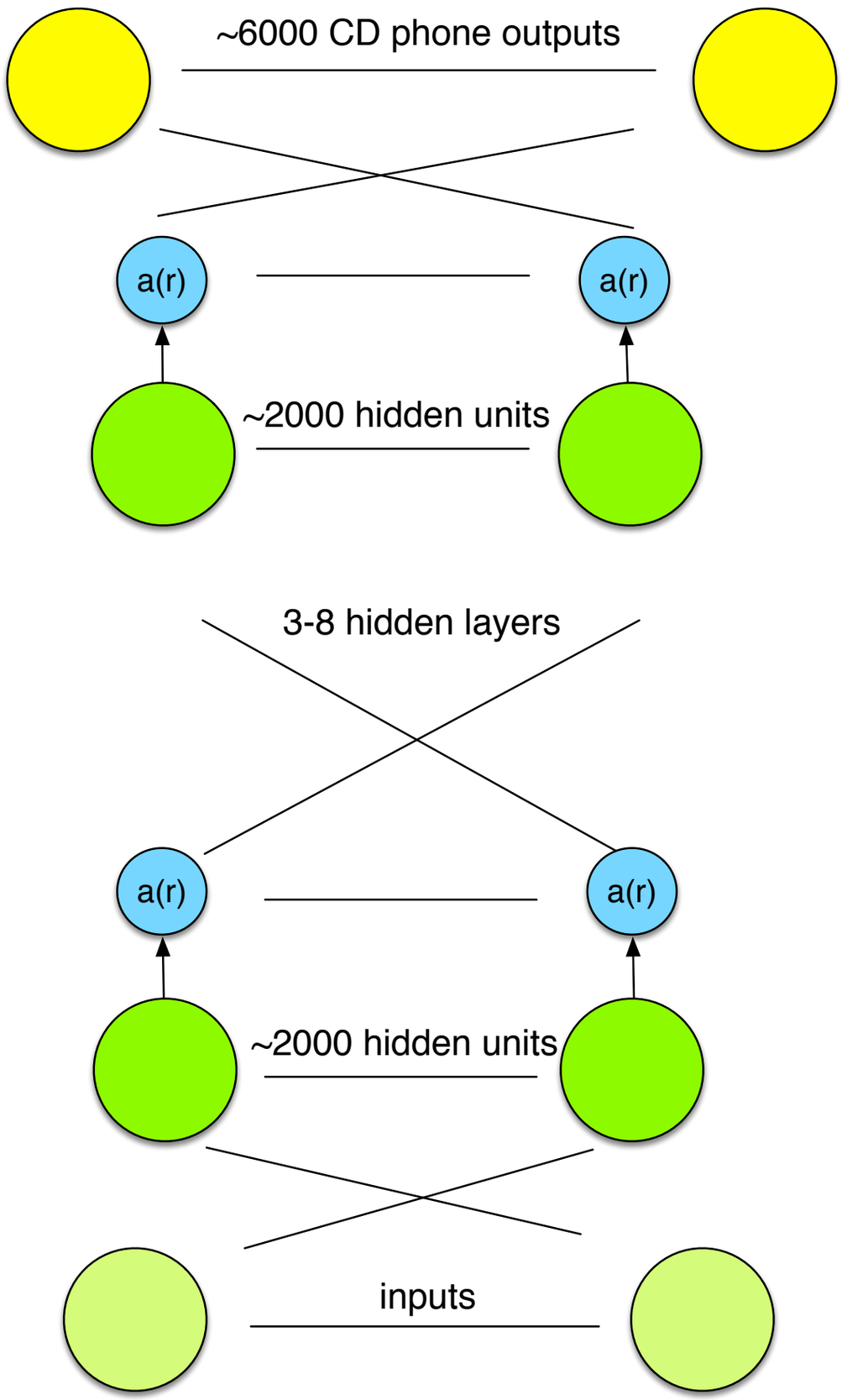
SI Model: set amplitudes to 1
SD Model: learn amplitudes from data, per speaker

Use first pass decoding to obtain adaptation targets

LHUC

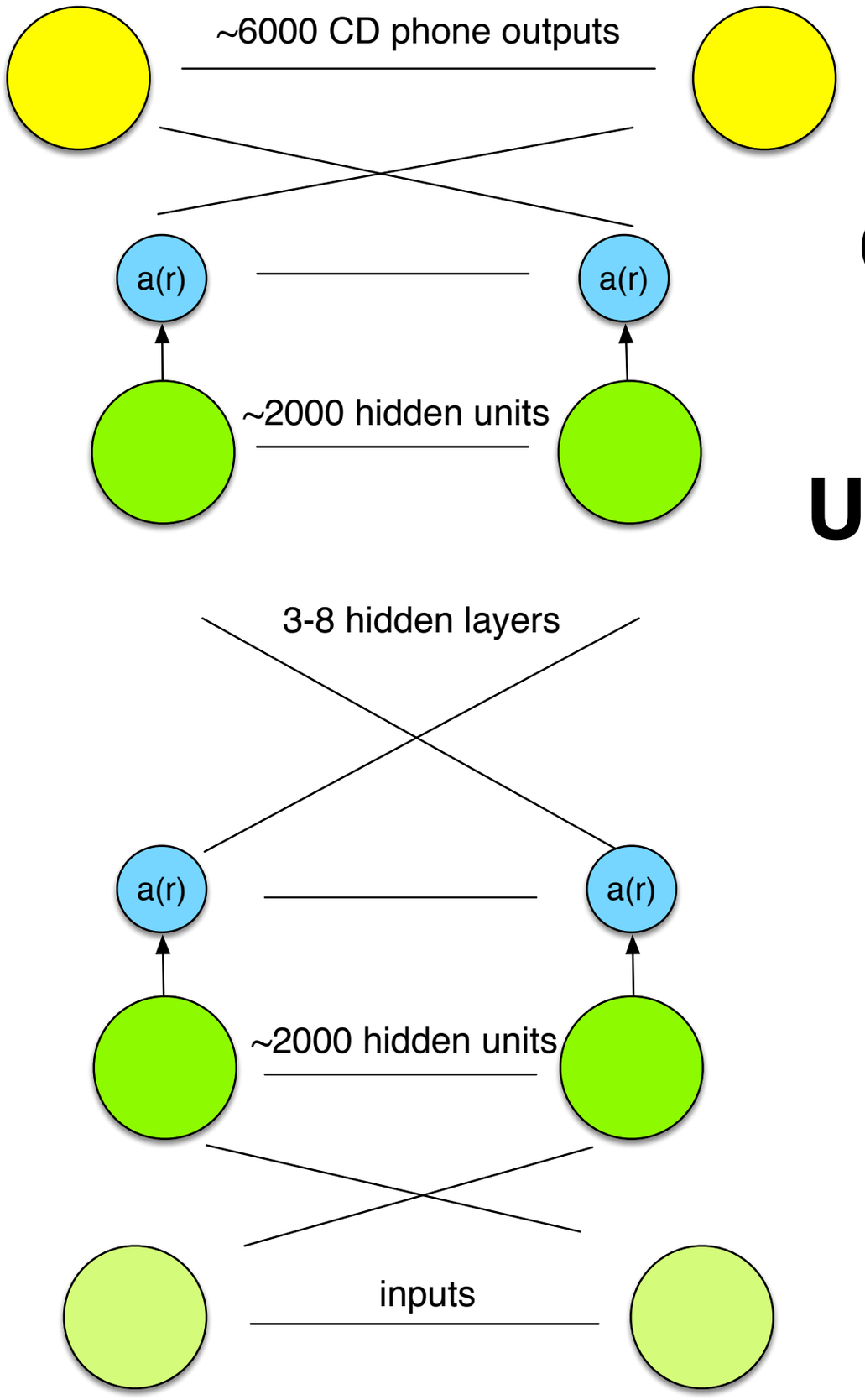

~6000 CD phone outputs
a(r)
a(r)
~2000 hidden units
3-8 hidden layers
a(r)
a(r)
~2000 hidden units
inputs

LHUC



Compact: 1 SD parameter per hidden unit

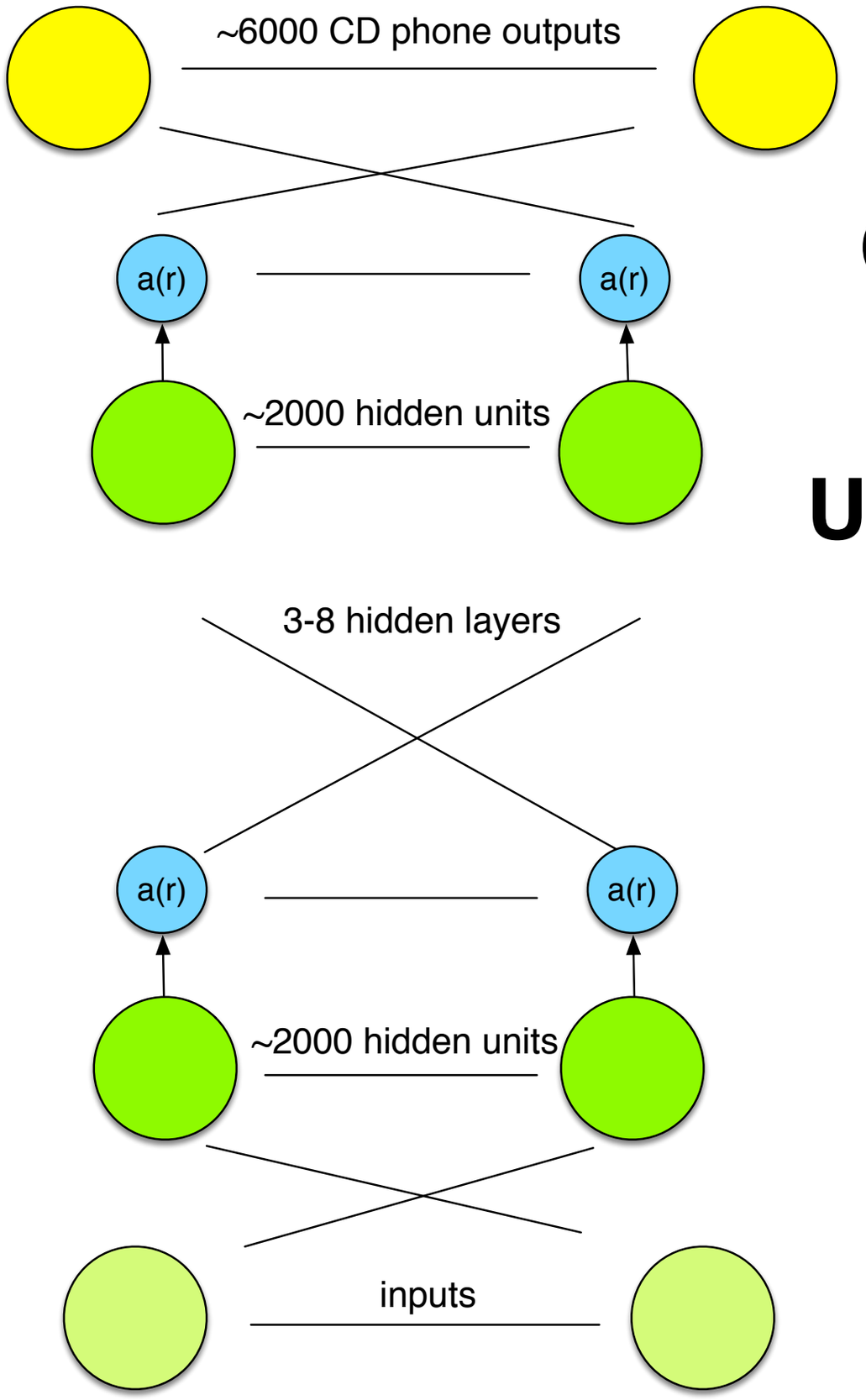
LHUC



Compact: 1 SD parameter per hidden unit

Unsupervised: No speaker-adaptive training

LHUC

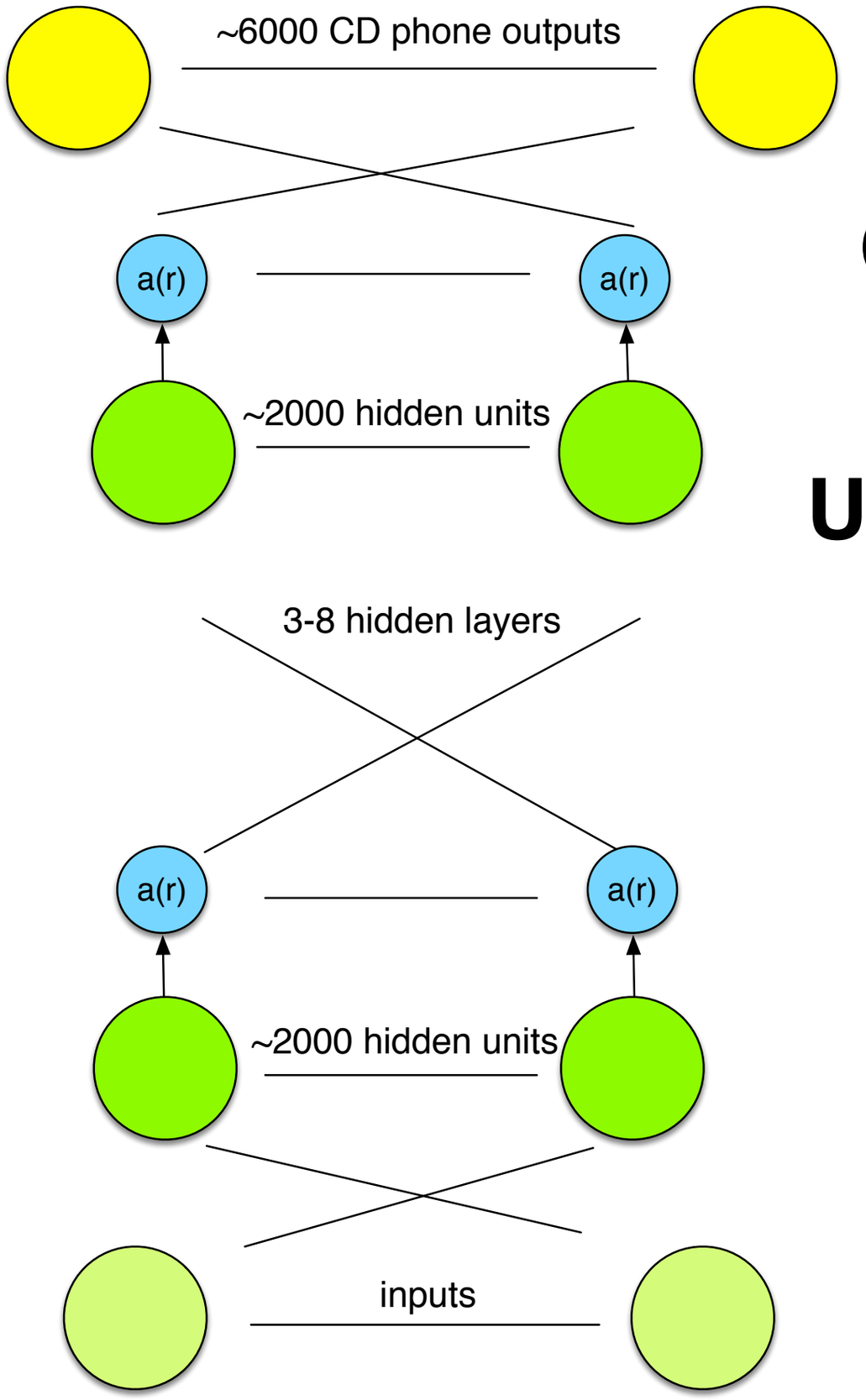


Compact: 1 SD parameter per hidden unit

Unsupervised: No speaker-adaptive training

Efficient: Few iterations of gradient descent on test/adapt data

LHUC



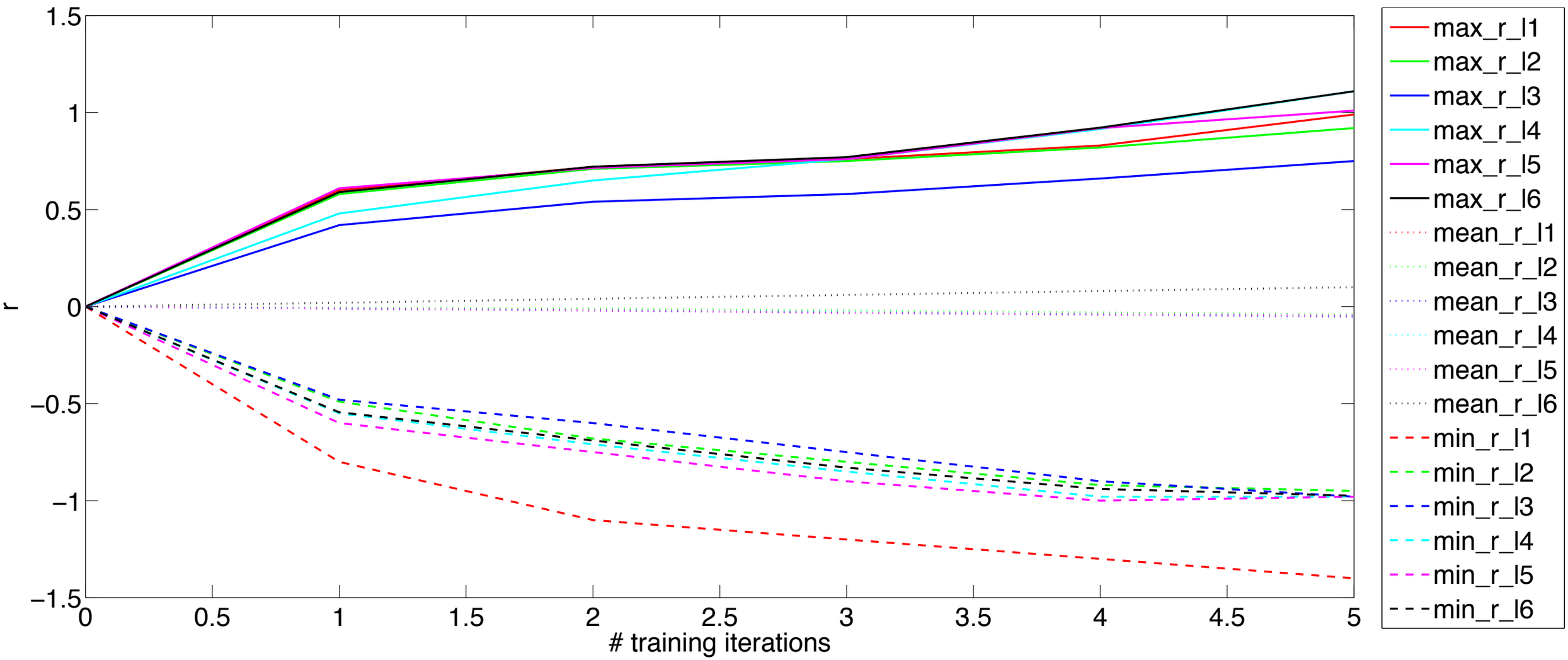
Compact: 1 SD parameter per hidden unit

Unsupervised: No speaker-adaptive training

Efficient: Few iterations of gradient descent on test/adapt data

Flexible: Applicable to different hidden unit transfer functions

Speaker-dependent amplitudes



Mean/Max/Min of adapted amplitudes in each layer

LHUC experiment 1
TED Talks (IWSLT)

Training: 143 hrs – 50M frames

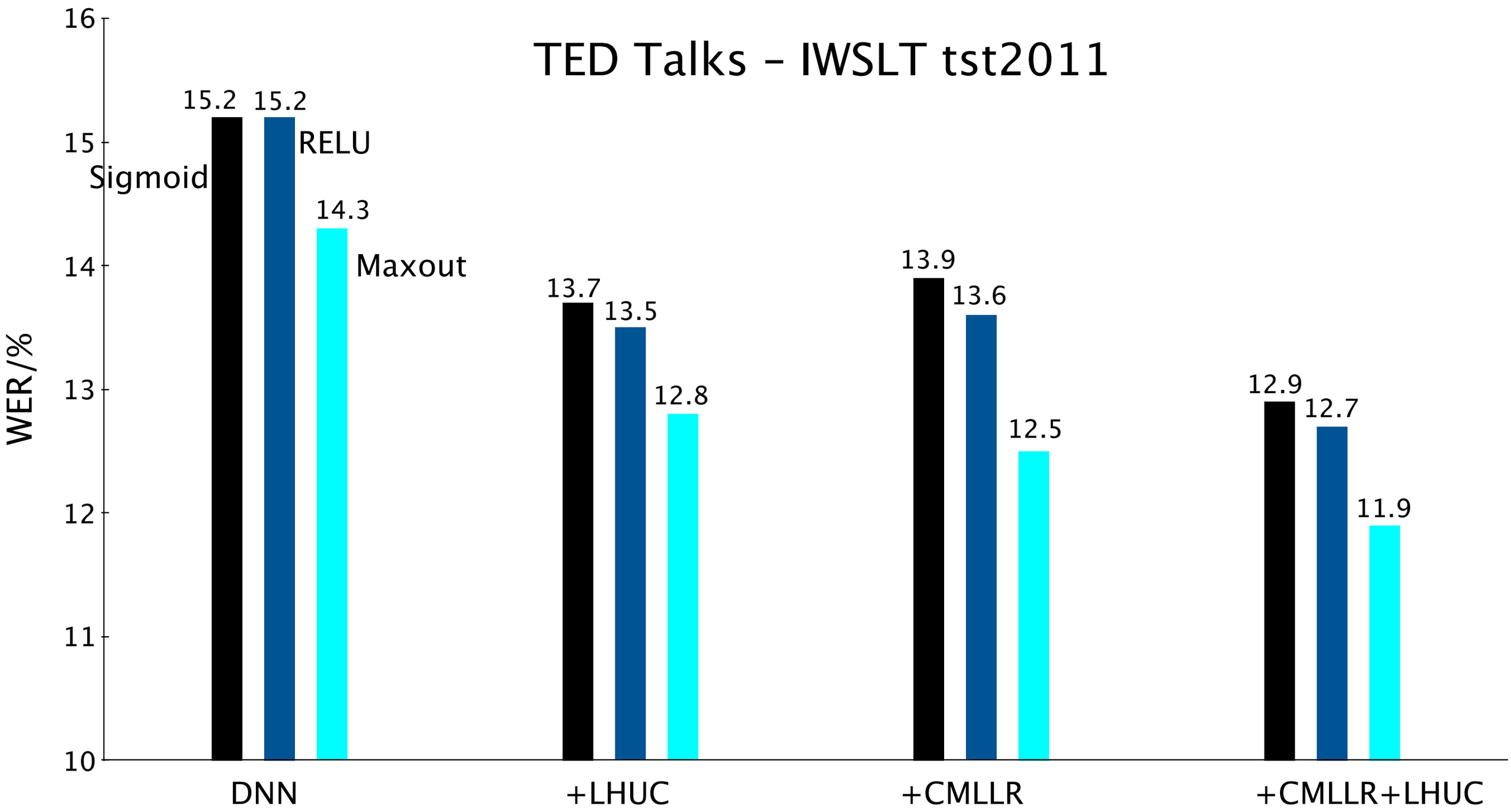
12,000 tied states

(sigmoid) 6 x 2000 unit hidden layers

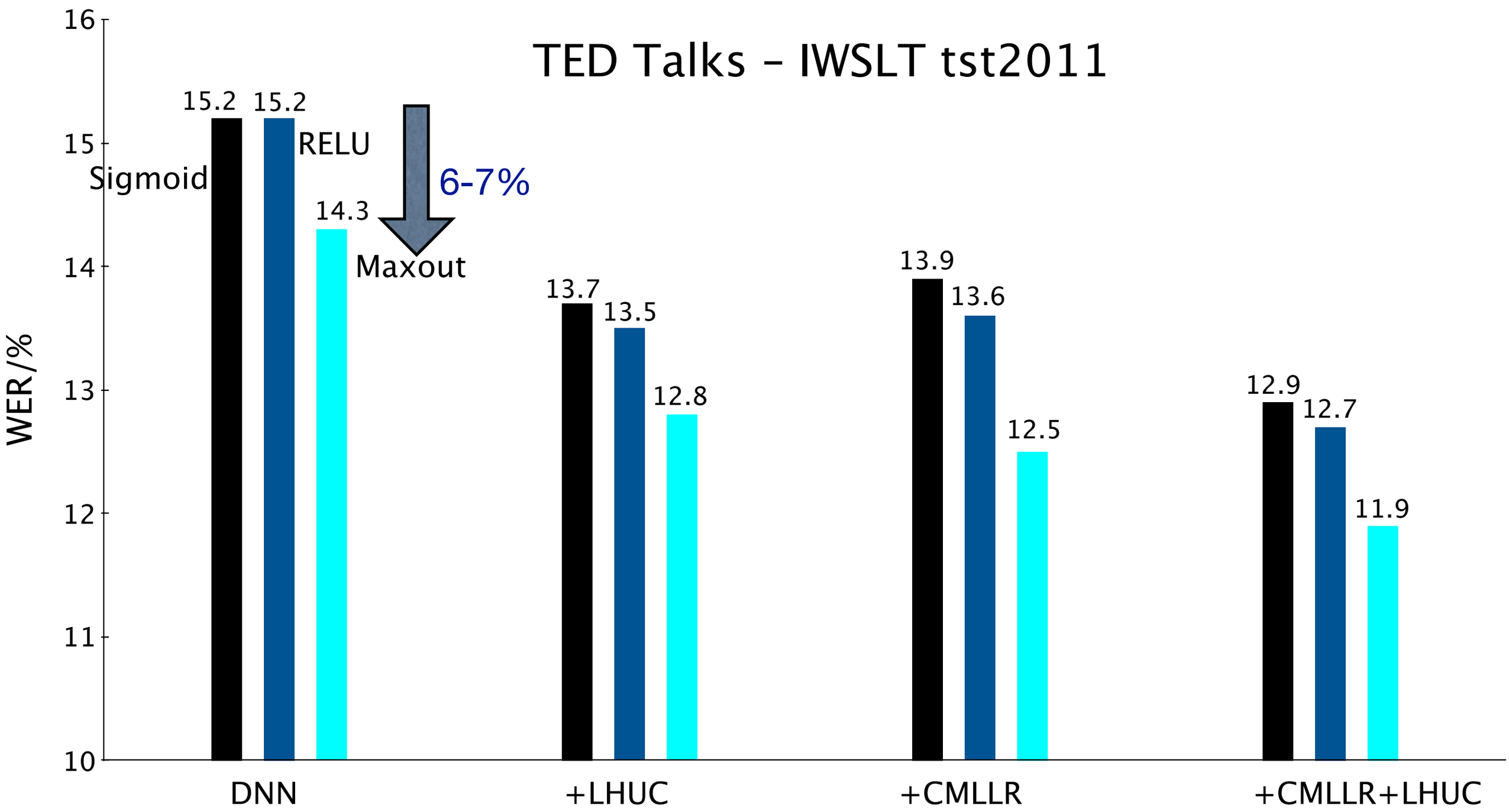
(maxout) 6 x 1500 maxpool[2] hidden layers

13-D PLP features + Δ + ΔΔ

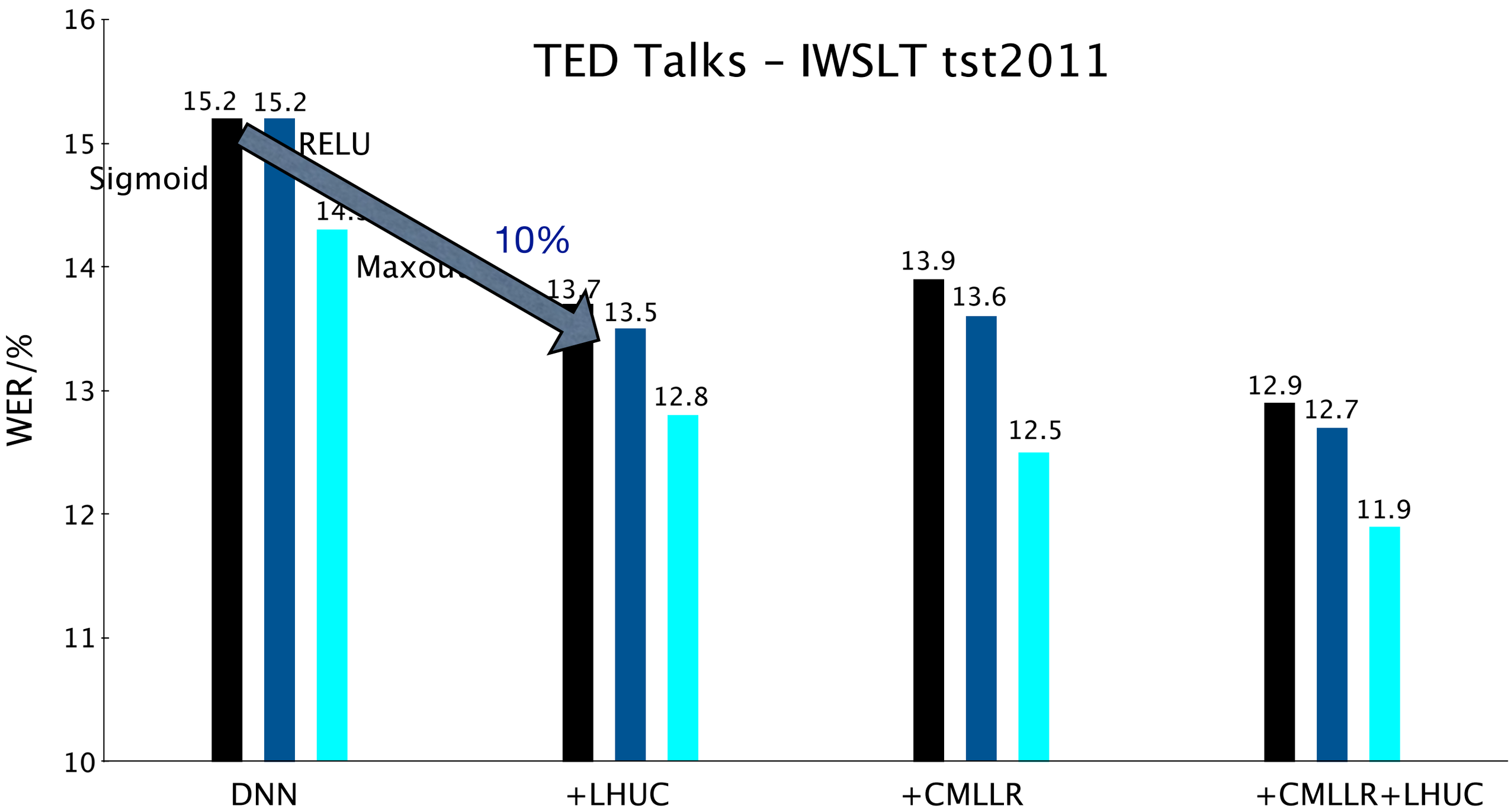
LHUC experiment I
TED Talks



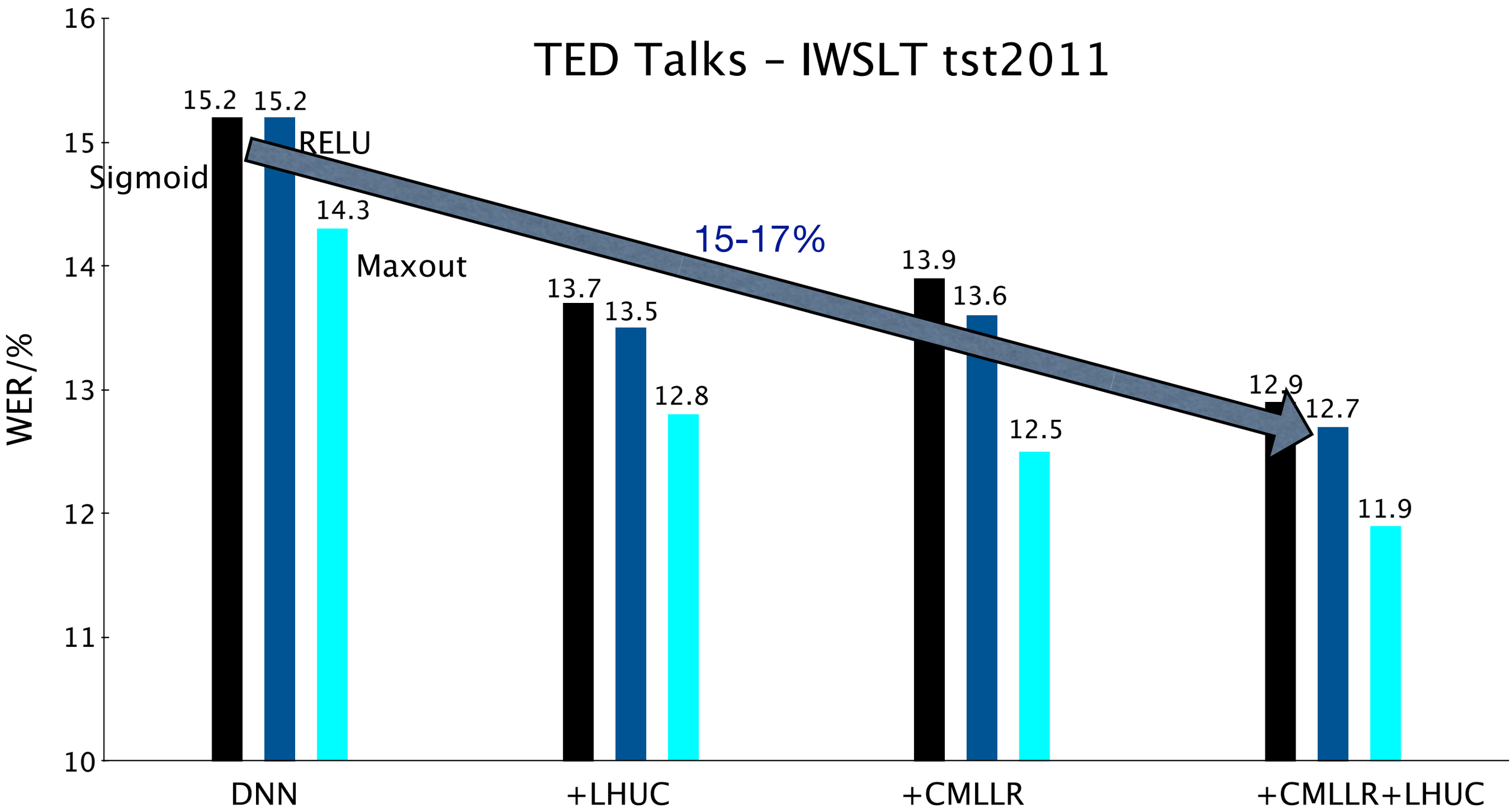
LHUC experiment I
TED Talks



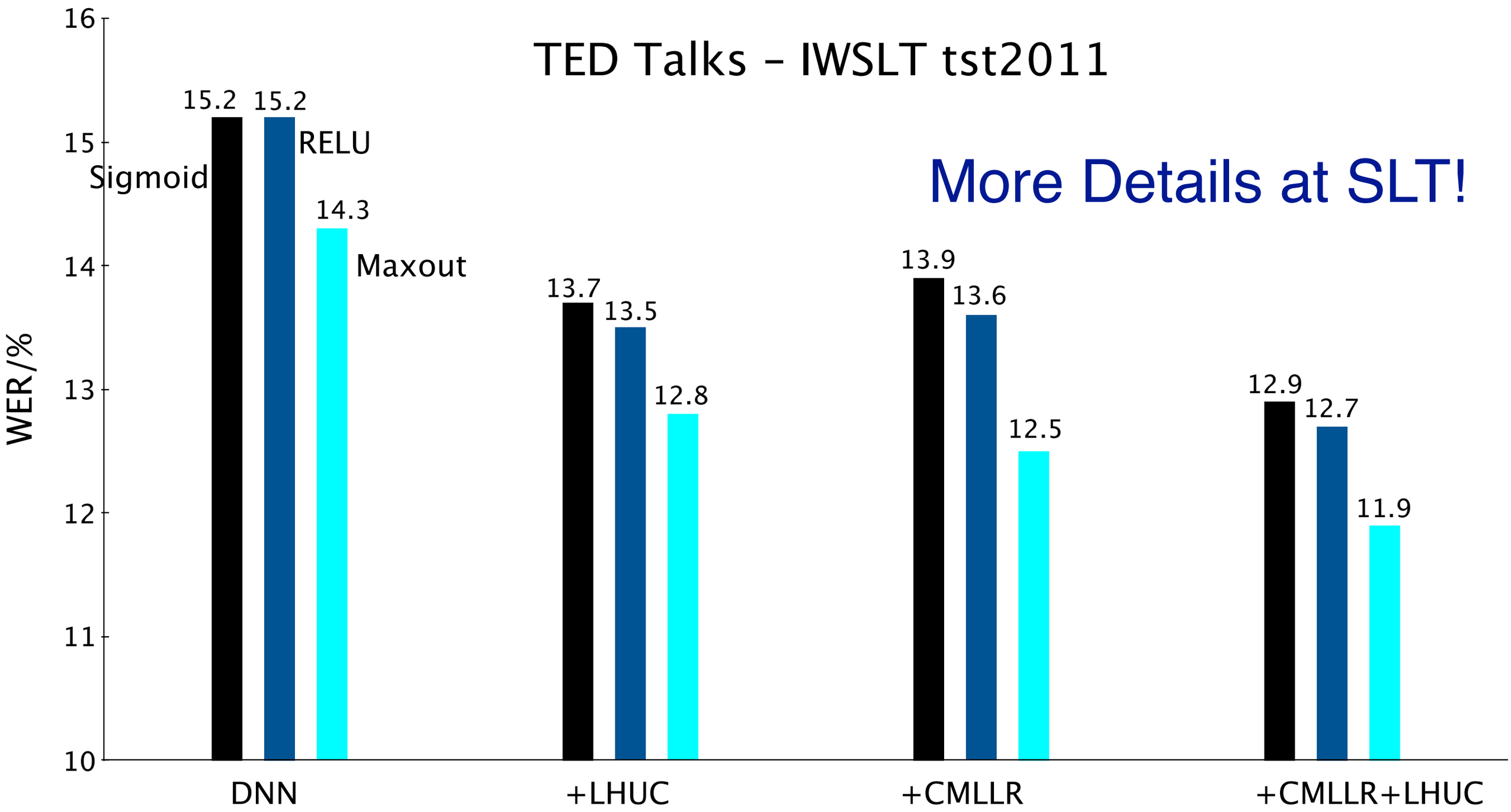
LHUC experiment I
TED Talks



LHUC experiment I
TED Talks



LHUC experiment I
TED Talks



More Details at SLT!

LHUC experiment 2
Aurora-4

LHUC experiment 2
Aurora-4

Standard Aurora-4 protocol

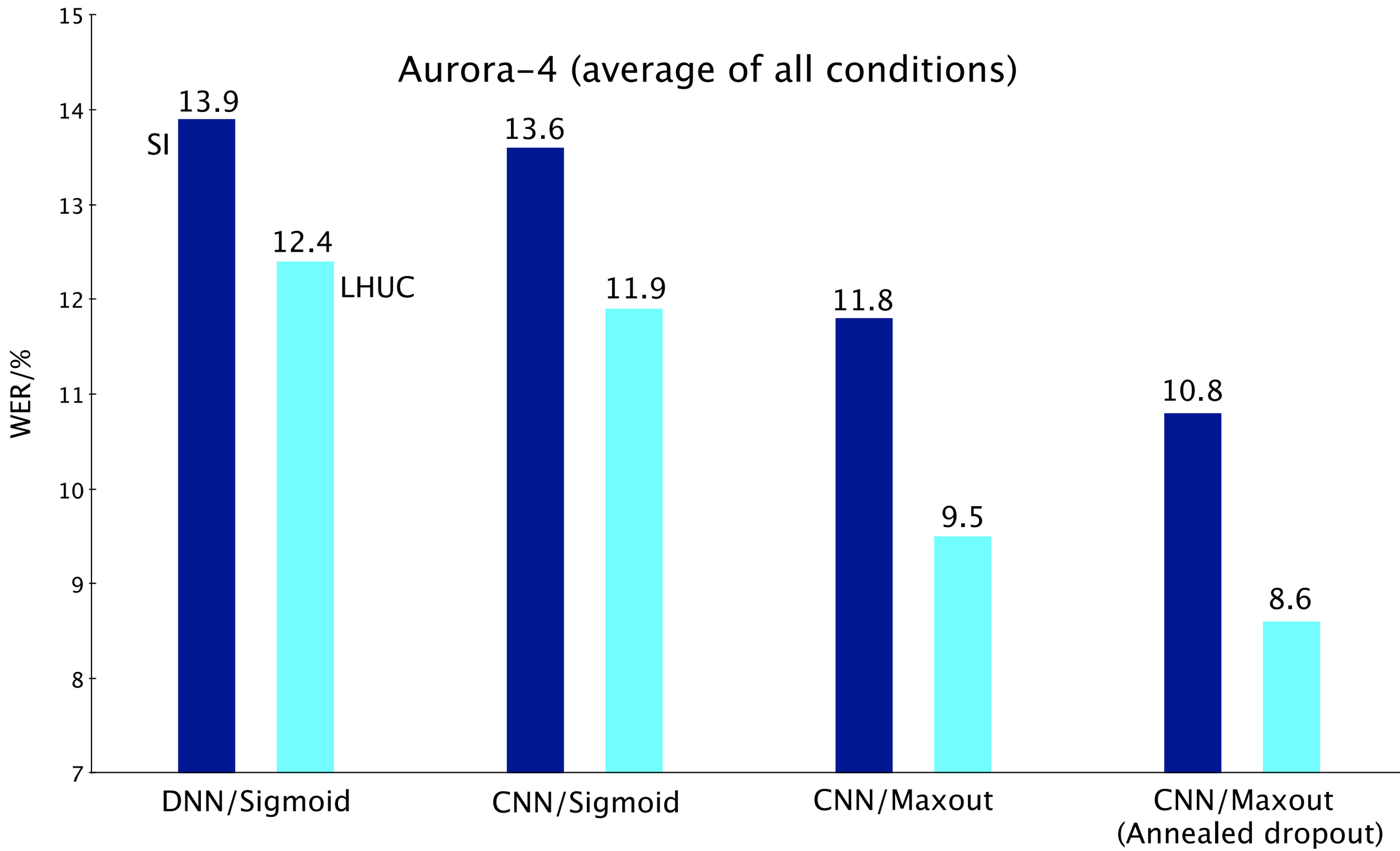
DNN/Sigmoid – 5x2000 unit hidden layers

CNN/Sigmoid – additional conv layer (300 filters)

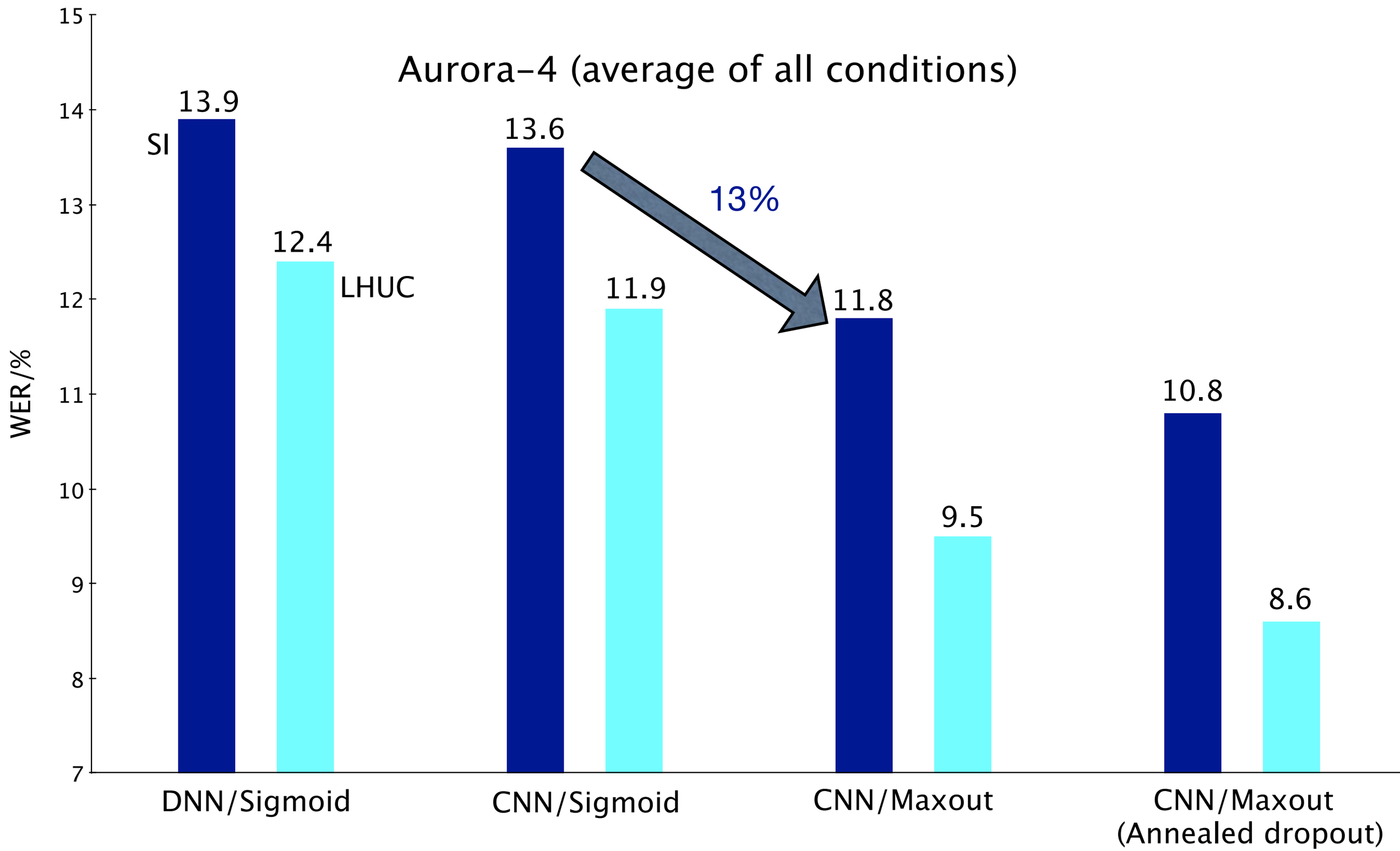
CNN/Maxout - 300 filters, 6x1000 maxpool[2] hidden layers

CNN/Maxout (Annealed Dropout) (Rennie et al, SLT 2014)

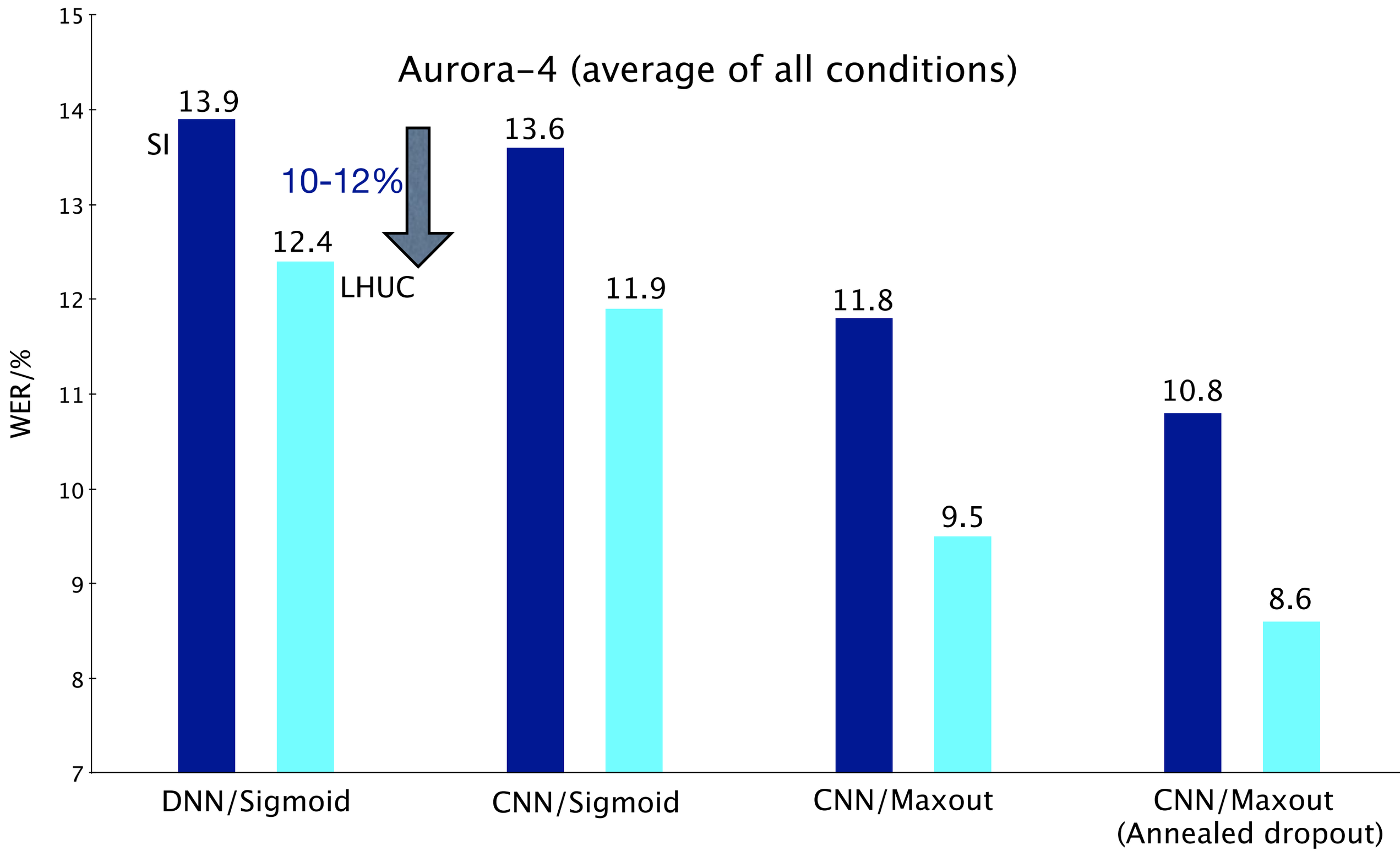
LHUC experiment 2
Aurora-4



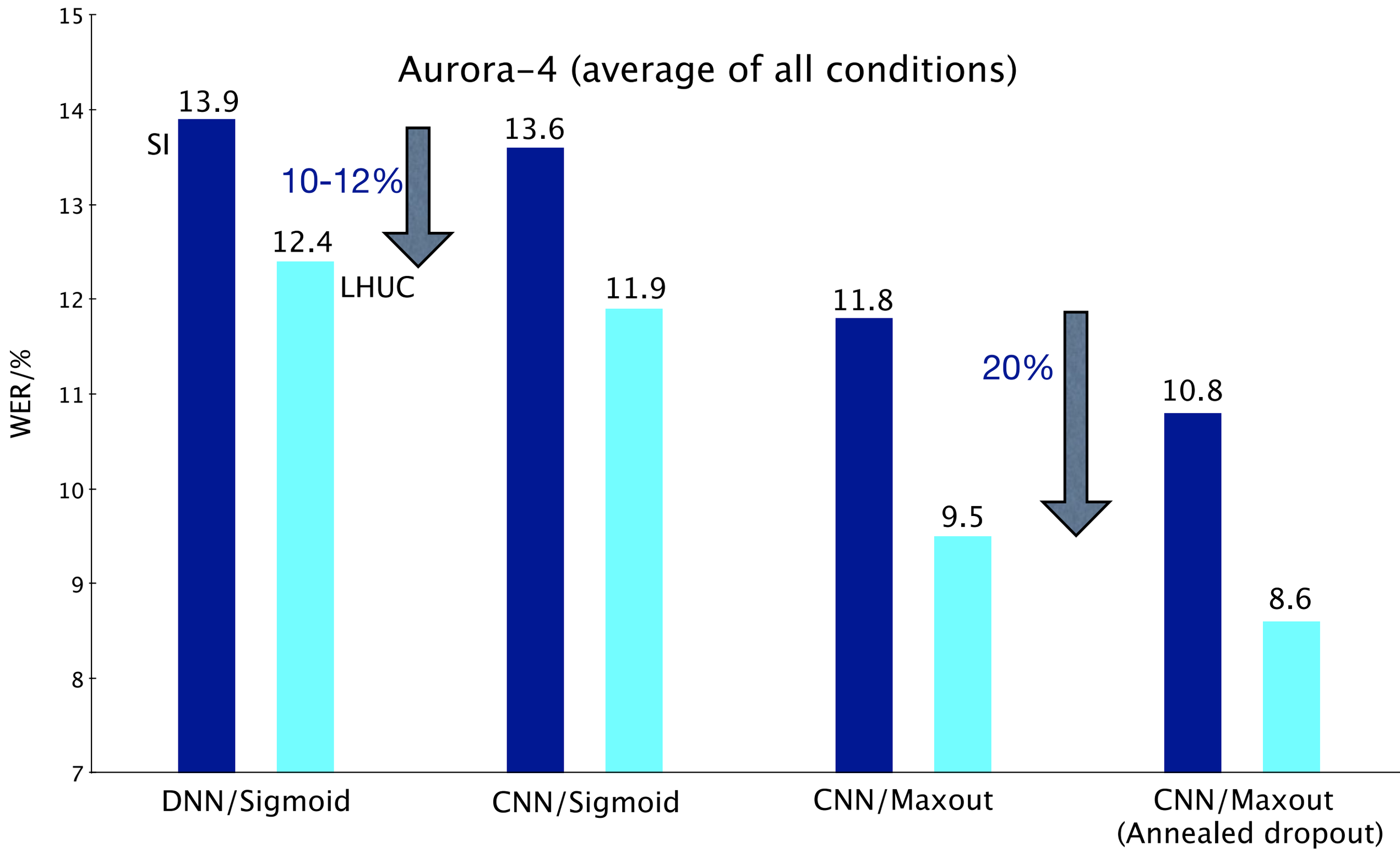
LHUC experiment 2
Aurora-4



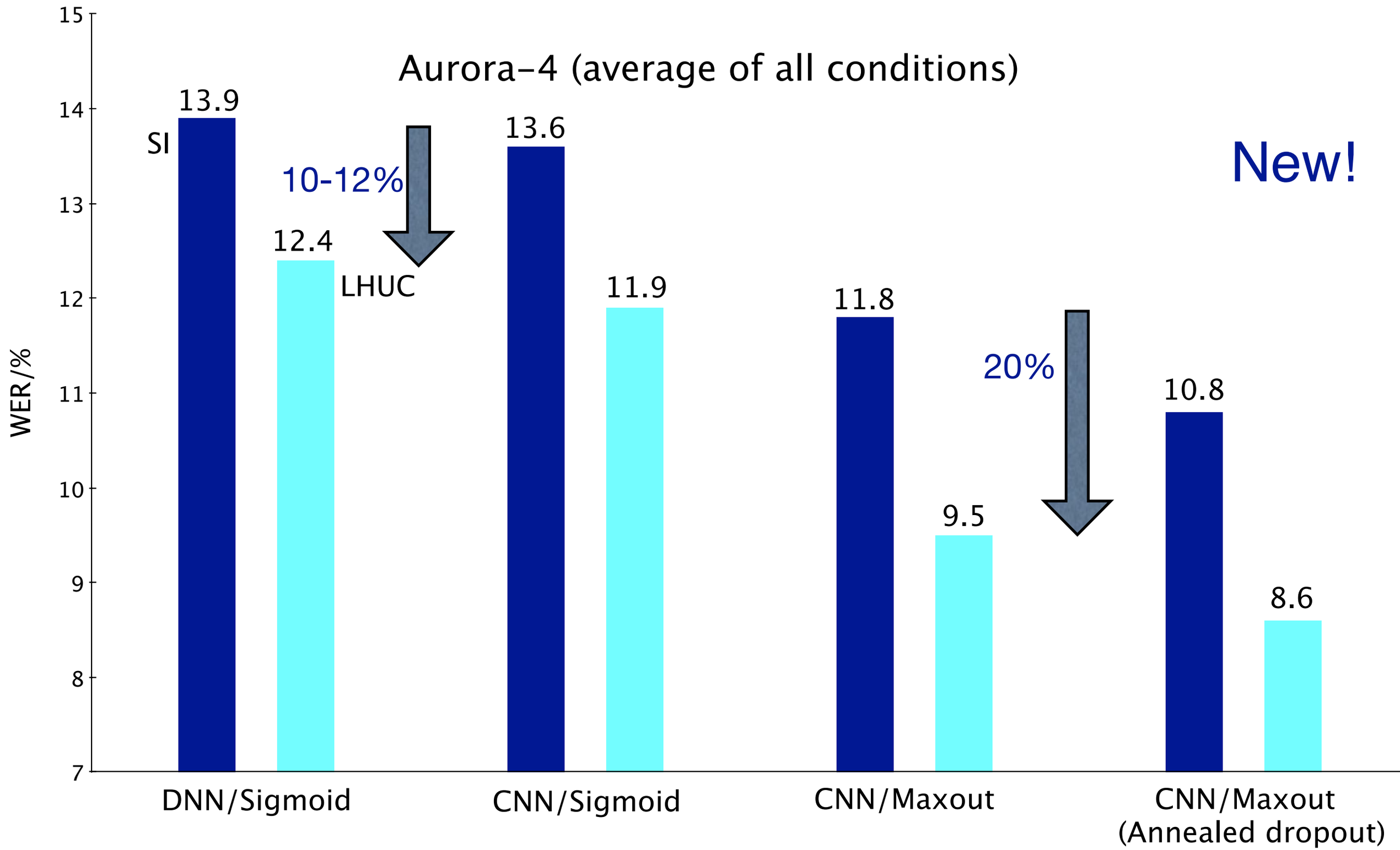
LHUC experiment 2
Aurora-4



LHUC experiment 2
Aurora-4



LHUC experiment 2
Aurora-4



New!

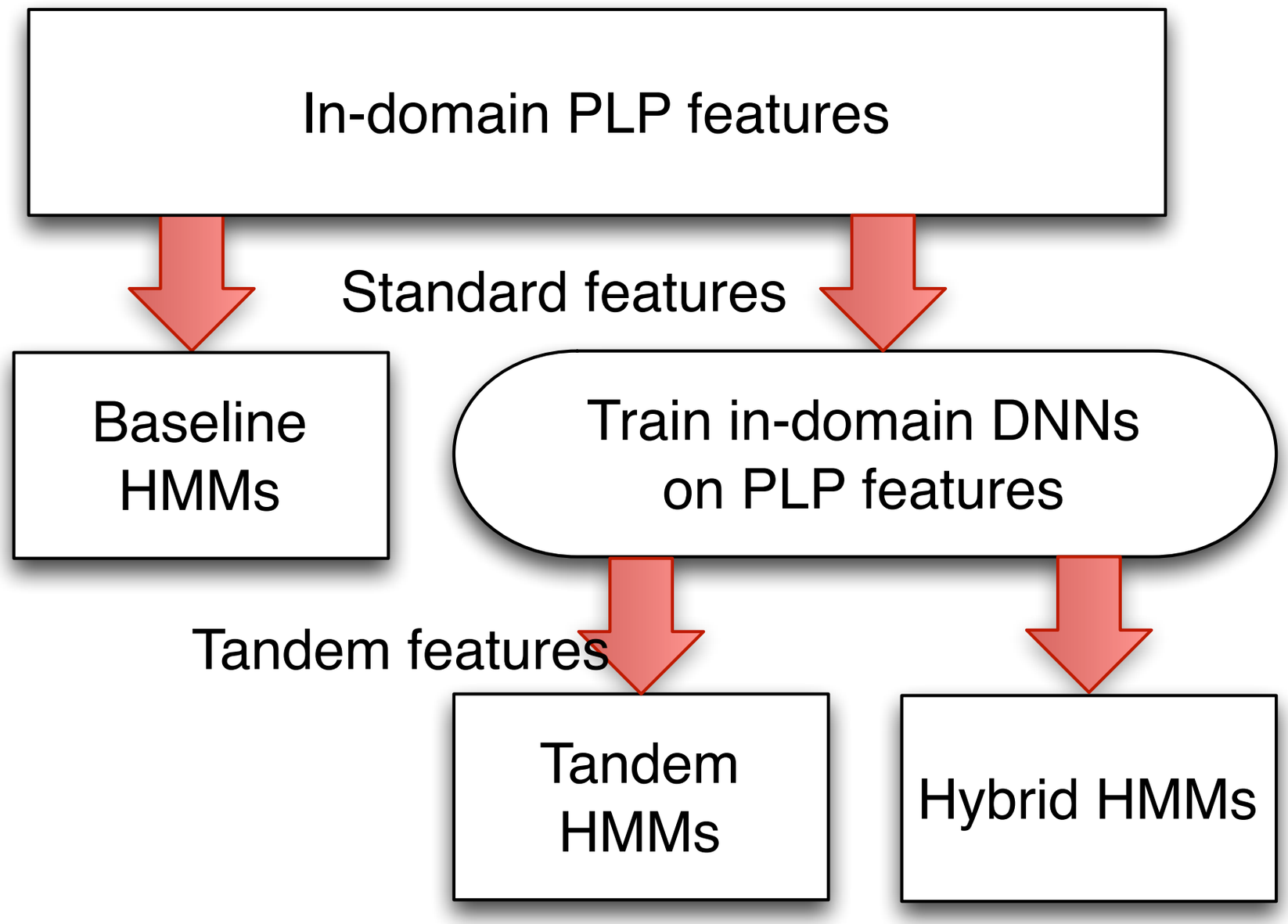
LHUC

- Speaker-dependent hidden unit amplitudes
- Compact, unsupervised adaptation scheme
- Applicable to different hidden unit functions
- Doesn't require adaptive training
- Experiments on TED/Aurora-4
  - 10-20% relative WER decrease (few minutes of speech)
  - 3% rel WER decrease from 10s speech
  - complementary to CMLLR transform
  - (need to combine with i-vector features)

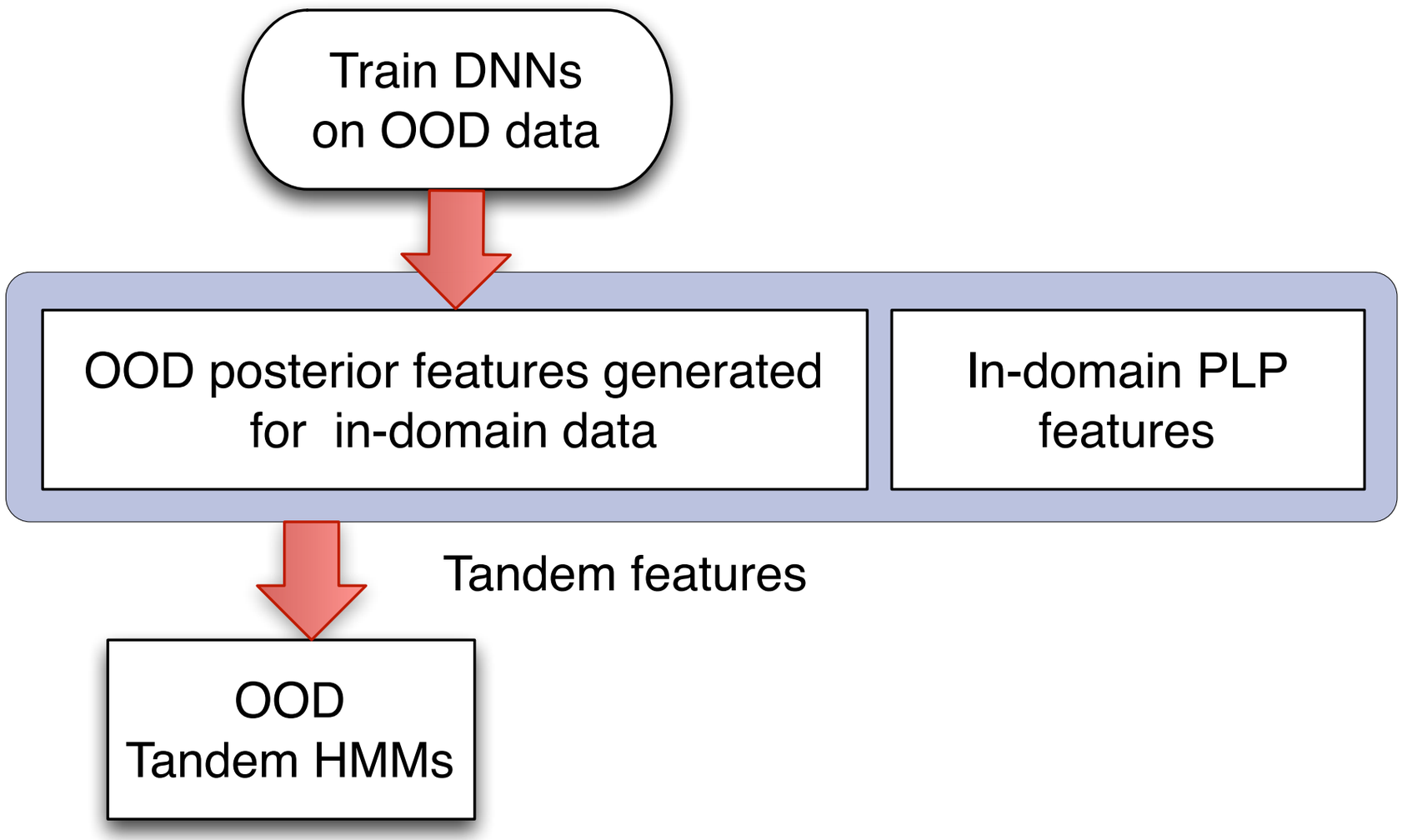
Part Three

Domain Adaptation

Conventional DNN

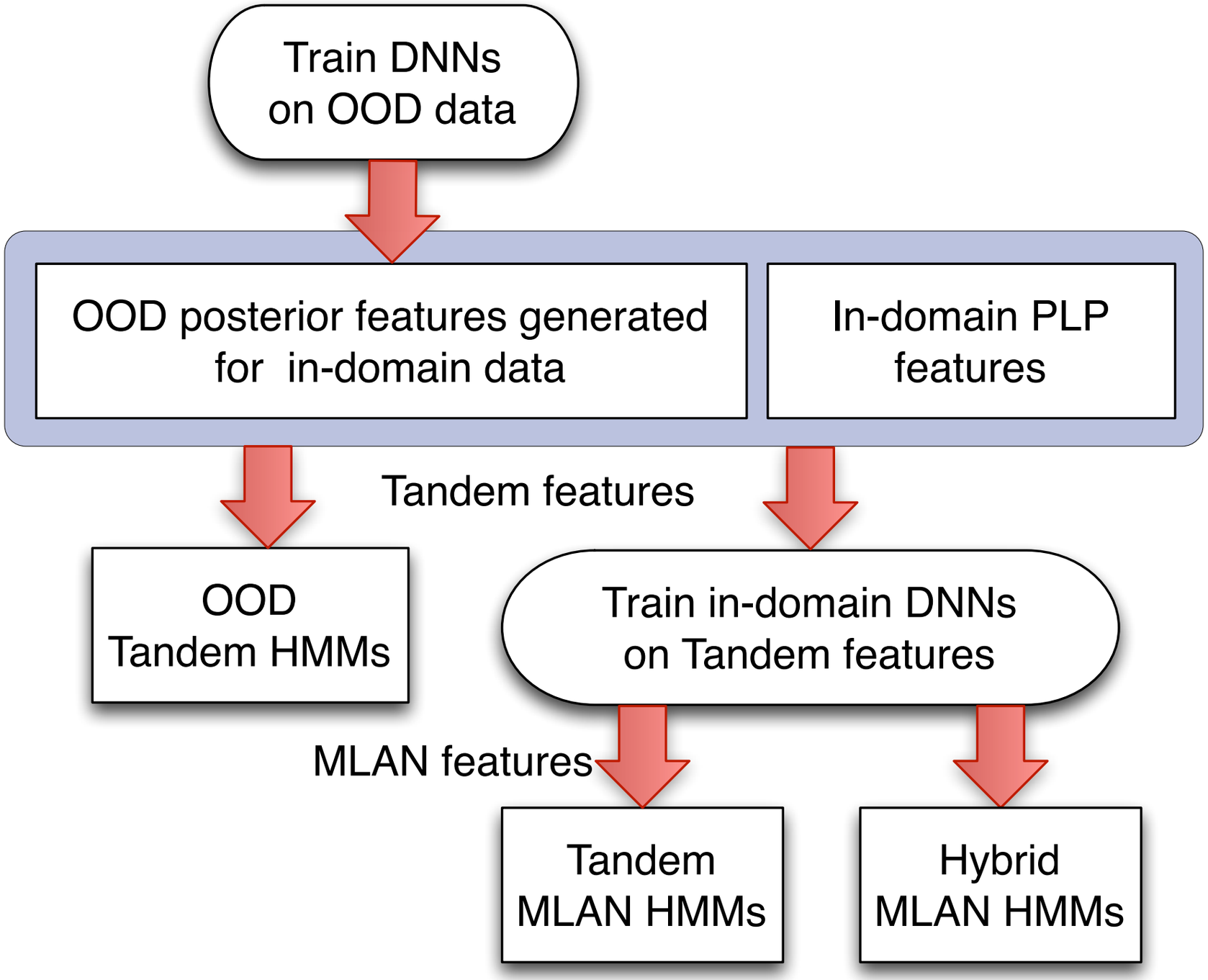


Out-of-domain NN features

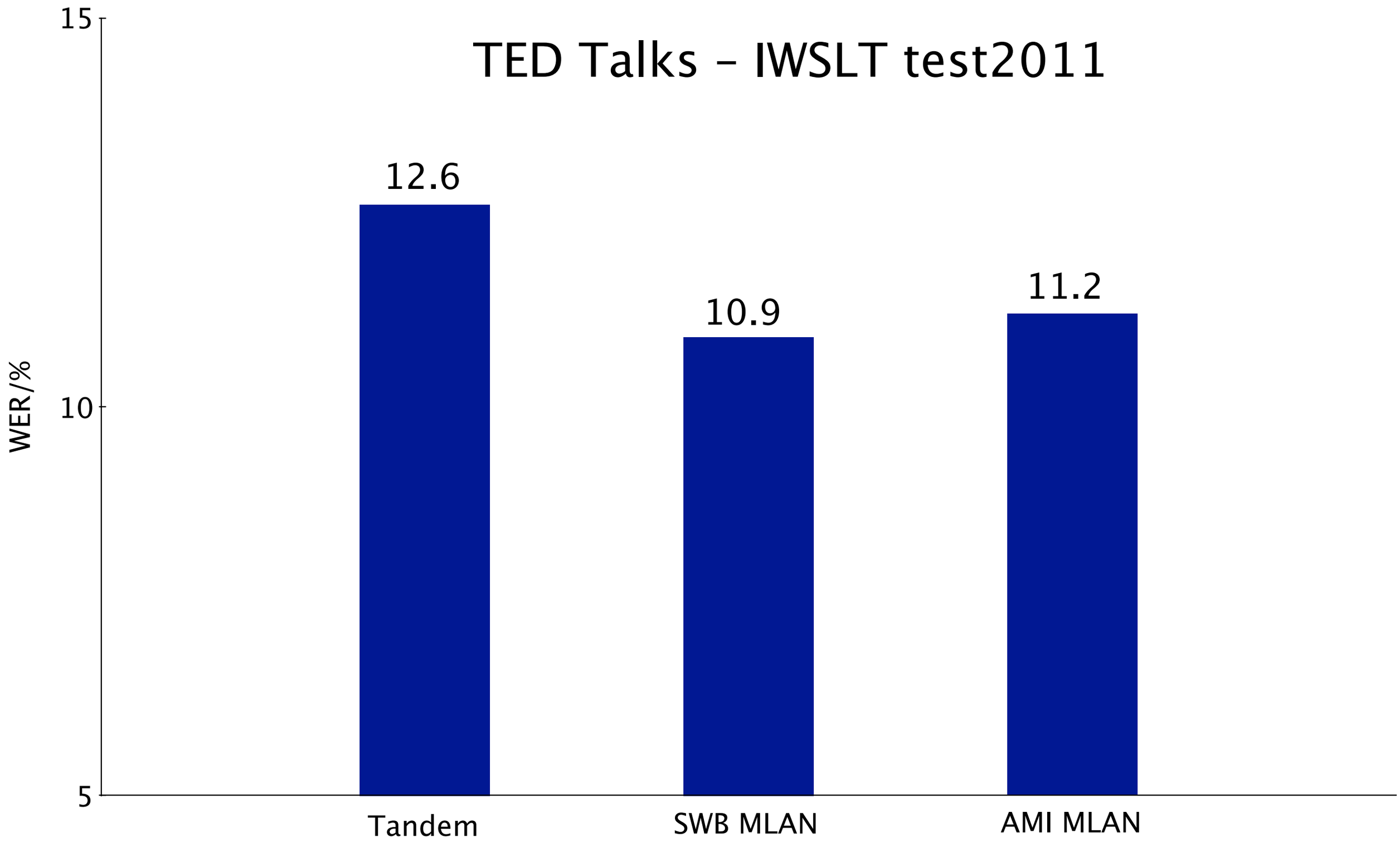


MLAN

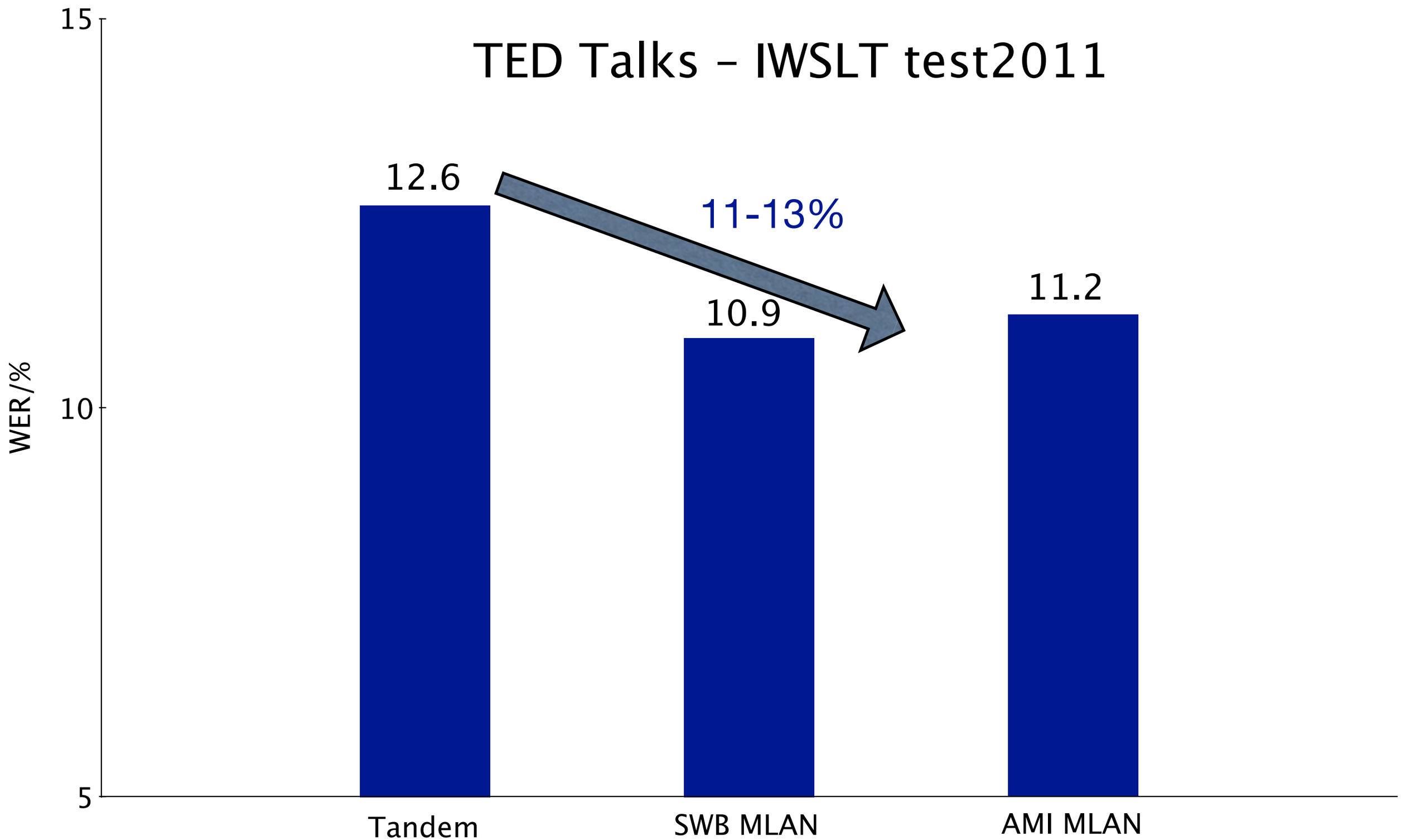
Multi-Layer Adaptive Networks



MLAN Experiment
TED Talks



MLAN Experiment
TED Talks



MLAN

- Use NNs trained on out-of-domain data to generate features with in-domain data
- Does not bias training to out-of-domain data
- Useful way to include tandem/bottleneck features in NN/HMM hybrids
- Can be used for cross-lingual ASR (SLT)

Conclusion

Conclusions

Conclusions

- Learning representations can give significant improvements in accuracy
- **CNNs for multichannel ASR**
  - beneficial for uncalibrated arrays
- **LHUC – speaker-dependent hidden unit amplitudes**
  - unsupervised, compact speaker adaptation
- **MLAN – out-of-domain NN acoustic features**
  - effective domain adaptation

Conclusions

- Learning representations can give significant improvements in accuracy
- **CNNs for multichannel ASR**
  - beneficial for uncalibrated arrays
- **LHUC – speaker-dependent hidden unit amplitudes**
  - unsupervised, compact speaker adaptation
- **MLAN – out-of-domain NN acoustic features**
  - effective domain adaptation

Thanks!

Conclusions

- Learning representations can give significant improvements in accuracy
- **CNNs for multichannel ASR**
  - beneficial for uncalibrated arrays
- **LHUC – speaker-dependent hidden unit amplitudes**
  - unsupervised, compact speaker adaptation
- **MLAN – out-of-domain NN acoustic features**
  - effective domain adaptation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield
EU★BRIDGE
inEvent
Accessing Dynamic Networked Multimedia Events
JST CREST uDialogue Project

Learning representations of speech in neural network acoustic models

Steve Renals
Centre for Speech Technology Research
University of Edinburgh

s.renals@ed.ac.uk

3 December 2014

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Aurora-4 results

	A	B	C	D	Avg
DNN/Sigmoid	5.1	9.3	9.3	20.8	13.9
DNN/Sigmoid +LHUC	4.5	8.6	7.4	18.3	12.4
CNN/Sigmoid	5.0	9.2	9.0	20.3	13.6
CNN/Sigmoid +LHUC	4.3	8.3	7.2	17.6	11.9
CNN/Maxout	4.2	7.7	8.2	17.7	11.8
CNN/Maxout +LHUC	3.8	6.5	5.5	14.3	9.5
CNN/Maxout (Annealed d/o)	4.0	5.9	7.0	15.8	10.8
CNN/Maxout +LHUC (Ann d/o)	3.6	6.7	5.2	13.5	8.6