★NN

Neural Network Acoustic Modelling Across the Decades

Steve Renals
University of Edinburgh
s.renals@ed.ac.uk

ICSI, 14 March 2015

BDNN, CDNN, ...

Neural Network Acoustic Modelling Across the Decades

Steve Renals
University of Edinburgh
s.renals@ed.ac.uk

ICSI, 14 March 2015

DNN

Steve Renals
University of Edinburgh
s.renals@ed.ac.uk

ICSI, 14 March 2015

DNN

Decades of Neural Networks

Steve Renals
University of Edinburgh
s.renals@ed.ac.uk

ICSI, 14 March 2015

So how was it in the early 90s?

So how was it in the early 90s?

- Big neural networks trained as acoustic classifiers

So how was it in the early 90s?

- Big neural networks trained as acoustic classifiers
- Use an HMM for (limited) sequence processing

So how was it in the early 90s?

- Big neural networks trained as acoustic classifiers
- Use an HMM for (limited) sequence processing
- RNNs starting to look attractive as large-scale acoustic models

So how was it in the early 90s?

- Big neural networks trained as acoustic classifiers
- Use an HMM for (limited) sequence processing
- RNNs starting to look attractive as large-scale acoustic models
- Worrying about speaker adaptation, modelling acoustic context, robustness, …

So how was it in the early 90s?

- Big neural networks trained as acoustic classifiers
- Use an HMM for (limited) sequence processing
- RNNs starting to look attractive as large-scale acoustic models
- Worrying about speaker adaptation, modelling acoustic context, robustness, …
- Use vector processors to train networks quickly

So how was it in the early 90s?

Big Dumb Neural Nets: A Working Brute Force Approach to Speech Recognition

Nelson Morgan

Abstract— Neural networks with over a million connections have been trained using online backpropagation and a speech corpus with over 6 million training examples (speech frames). These networks provide phonetic probabilistic estimates that are used in a continuous speech recognizer. Issues of the network training and application are discussed in this paper.

So how was it in the early 90s?

Big Dumb Neural Nets: A Working Brute Force Approach to Speech Recognition

Nelson Morgan

Abstract— Neural networks with over a million connections have been trained using online backpropagation and a speech corpus with over 6 million training examples (speech frames). These networks provide phonetic probabilistic estimates that are used in a continuous speech recognizer. Issues of the network training and application are discussed in this paper.

In other words, they ain't so big, and they ain't so dumb.

So how was it in the early 90s?

Big Dumb Neural Nets: A Working Brute Force Approach to Speech Recognition

Nelson Morgan

If Moore's Law had applied to this, then by 2015:
16 billion weights
96 billion examples

Abs
millio
online
with over 6 million training examples (speech frames). These networks provide phonetic probabilistic estimates that are used in a continuous speech recognizer. Issues of the network training and application are discussed in this paper.

In other words, they ain't so big, and they ain't so dumb.

So how was it in the early 90s?

Not much has changed?

So how was it in the early 90s?

Context-dependent acoustic modelling

- Then
  - largely context-independent NNs
  - some context-dependence
- Now
  - large scale context-dependent NNs – CD classes typically (not always) derived from HMM/GMM decision tree

Context-dependent acoustic modelling (then)

Context-dependent acoustic modelling (then)

Factoring Networks by a Statistical Method

Nelson Morgan
International Computer Science Institute, Berkeley, CA 94704 USA

Hervé Bourlard
International Computer Science Institute, Berkeley, CA 94704 USA
and
Lernout & Hauspie Speechproducts, Ieper, B-8900, Belgium

We show that it is possible to factor a multilayered classification network with a large output layer into a number of smaller networks, where the product of the sizes of the output layers equals the size of the original output layer. No assumptions of statistical independence are required.

CDNN: A CONTEXT DEPENDENT NEURAL NETWORK FOR CONTINUOUS SPEECH RECOGNITION

Hervé Bourlard [†,‡], *Nelson Morgan* [‡], *Chuck Wooters* [‡], *and Steve Renals* [‡]

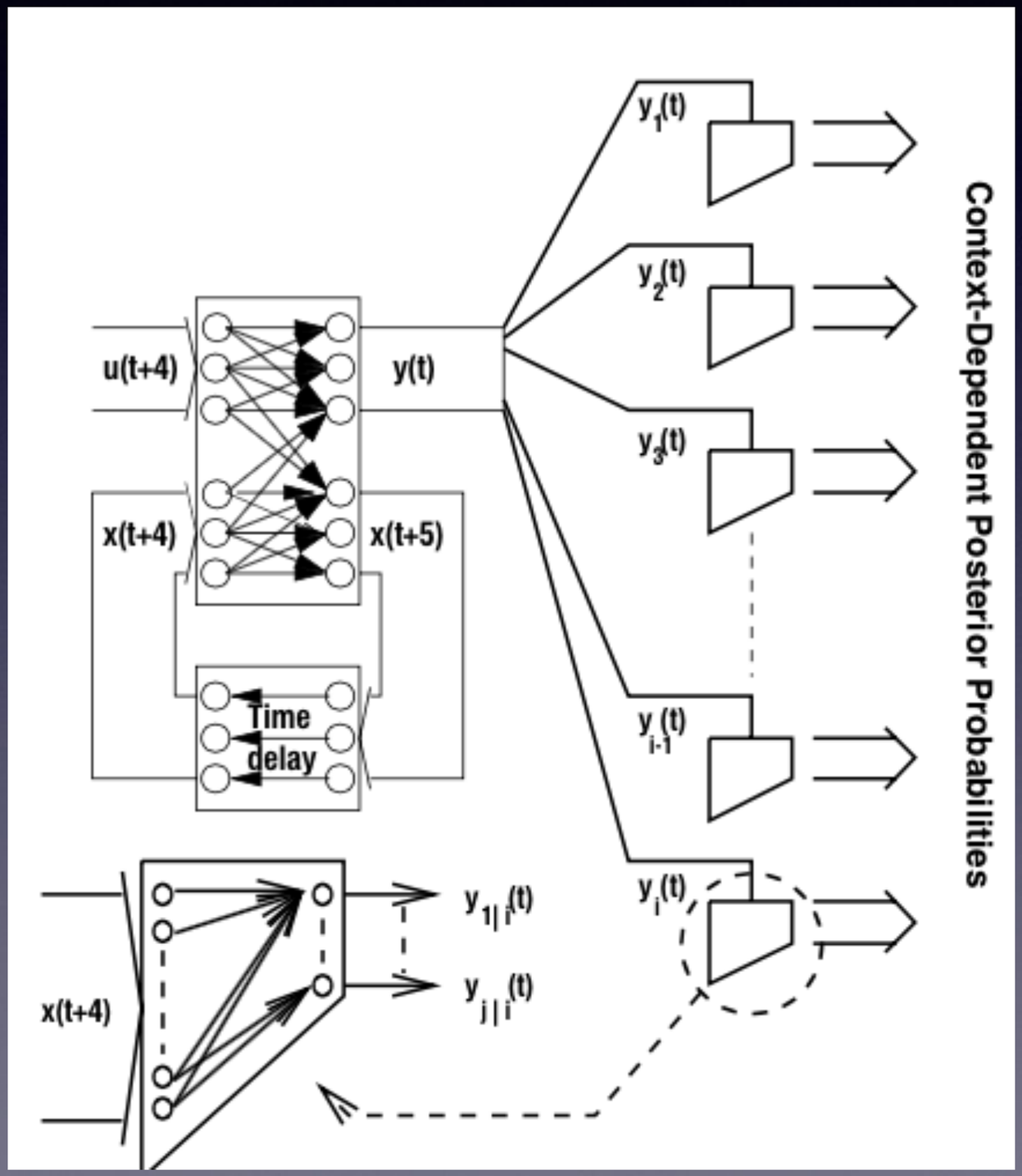
[†] L&H SpeechProducts, Ieper, B-8900 BELGIUM
[‡] International Computer Science Institute, 1947 Center St., Berkeley CA 94704, USA

Context-dependent acoustic modelling (then)

$$p(r_j . s_\ell \mid x_n) = p(s_\ell \mid x_n) \times p(r_j \mid s_\ell . x_n) \tag{2.2}$$

Thus, the desired probability is the product of one coarse category posterior probability and a second conditional probability. The former can be realized with a standard MLP probability estimator, using the same

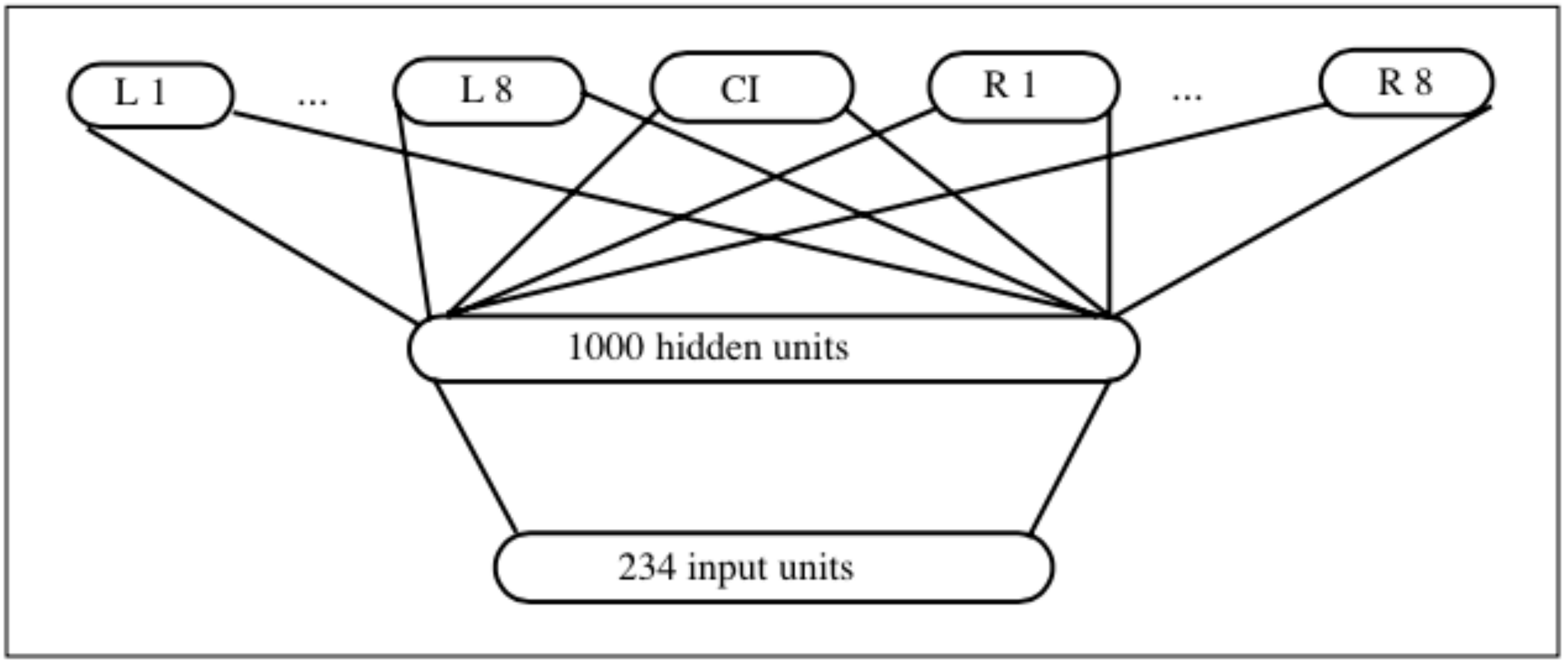
Context-dependent acoustic modelling (then)



Context-dependent acoustic modelling (then)

Test Set	ABBOT '94	ABBOT '95 (CI)	ABBOT '95 (CD)	$\text{Red}^{n}_{\text{WER}}$
H1:C0	14.1†	15.3	13.1	14.4
H1:P0	12.4‡	13.3	11.2	15.6
H3:P0-DT		24.6	21.9	11.0
H3:C1B		15.3	13.6	11.1

Context-dependent acoustic modelling (then)

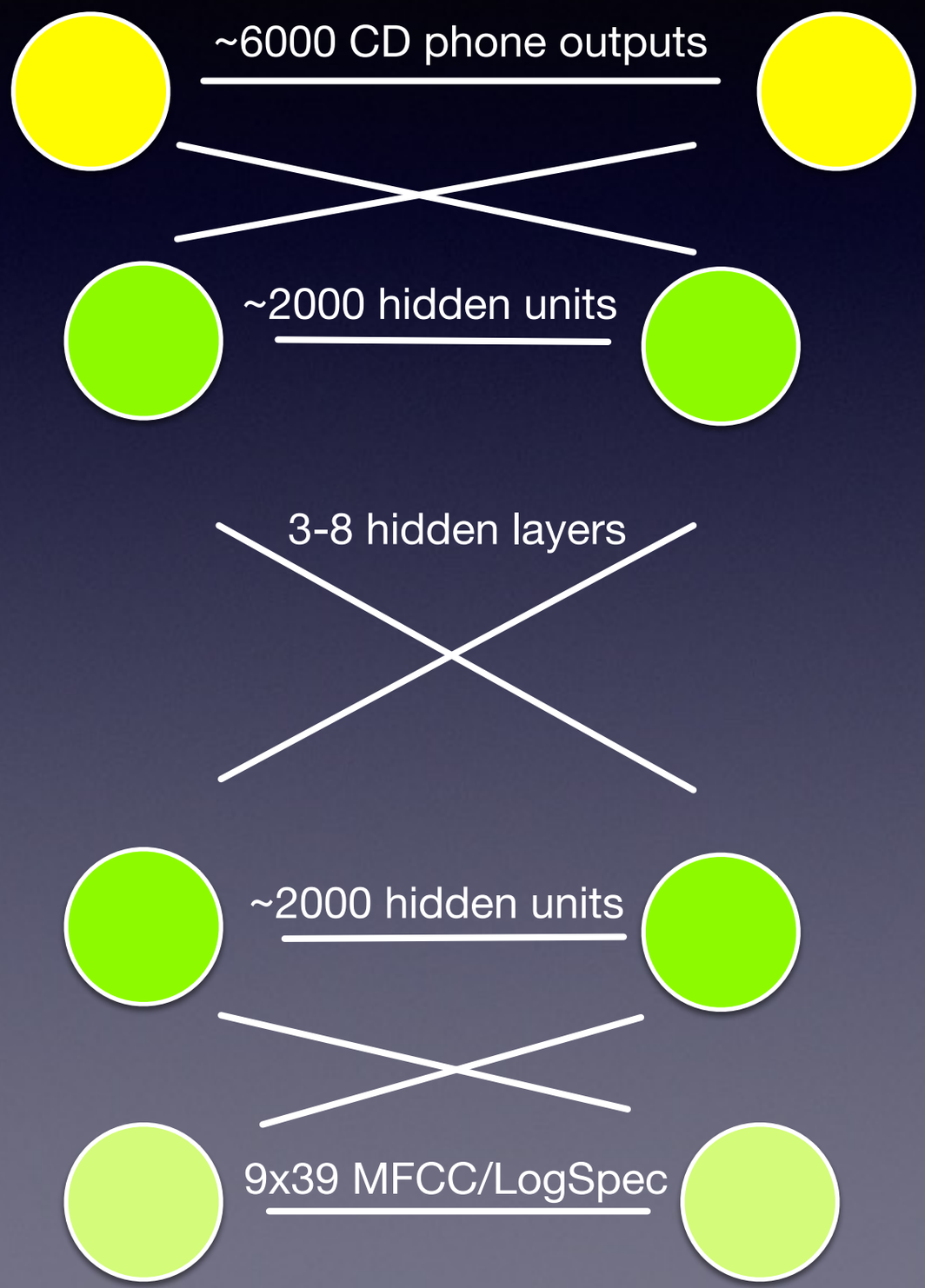


Context-dependent acoustic modelling (then)

Table I: WORD PAIR GRAMMAR

	CI MLP	CD MLP	CD HMM
Feb 91	5.8	4.7	3.8
Sept 92 ind	10.9	7.6	10.1
Sept 92 dep	9.5	6.6	7.0
Overall test	8.77	6.3	7.0

Context-dependent acoustic modelling (now)



Why model context-dependence?

- Divide and conquer the acoustic modelling space, if enough training data

- Context-sensitive adaptation of phone models to local acoustic/phonetic context – enhances discrimination of acoustically confusable phones

Why model context-dependence?

- Divide and conquer the acoustic modelling space, if enough training data

  More important for GMM-based generative models than neural networks?

  Neural networks focus on class boundaries rather than within-class structure

- Context-sensitive adaptation of phone models to local acoustic/phonetic context – enhances discrimination of acoustically confusable phones

Drawbacks of context-dependent phone models in neural networks

Drawbacks of context-dependent phone models in neural networks

- No distinction between different phones and the same phone in different contexts
  - leads to hidden units learning discriminations that are not useful
  - discriminations depend on particular state clustering

Drawbacks of context-dependent phone models in neural networks

- No distinction between different phones and the same phone in different contexts
  - leads to hidden units learning discriminations that are not useful
    - discriminations depend on particular state clustering
- Data sparsity

Drawbacks of context-dependent phone models in neural networks

- No distinction between different phones and the same phone in different contexts
  - leads to hidden units learning discriminations that are not useful
  - discriminations depend on particular state clustering
- Data sparsity
- How to avoid these problems?
  - factorised CDNN
  - sequence-level discriminative training

Another way? Multitask CD and CI

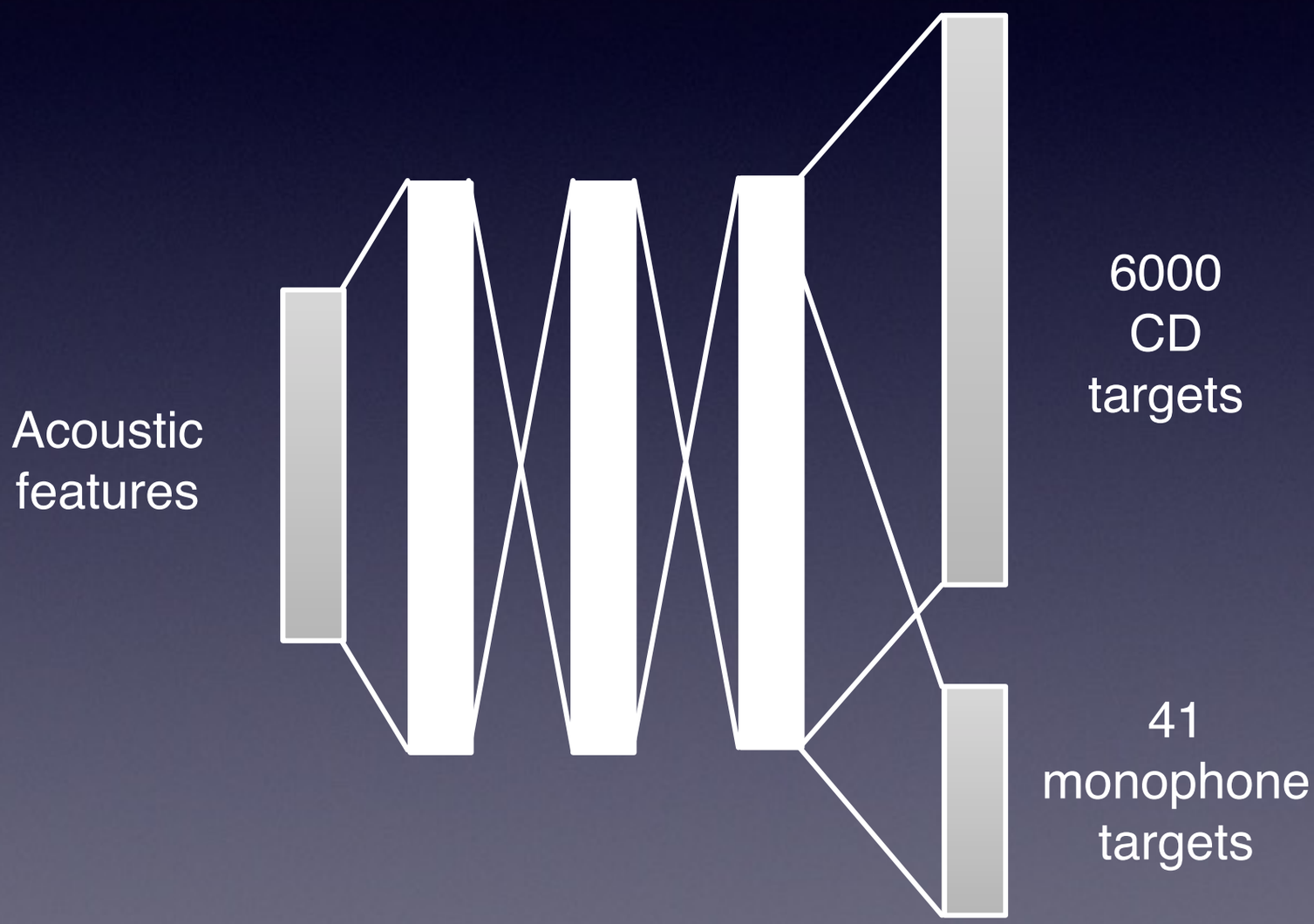
- Multitask learning – learning CD and CI phones together

- Interpret as regulariser? Compare with pretraining

  - unsupervised RBM

  - discriminative monophone pretraining (cf curriculum learning)

- Can combine pretraining with multitask learning



Joint work with Peter Bell

Another way? Multitask CD and CI

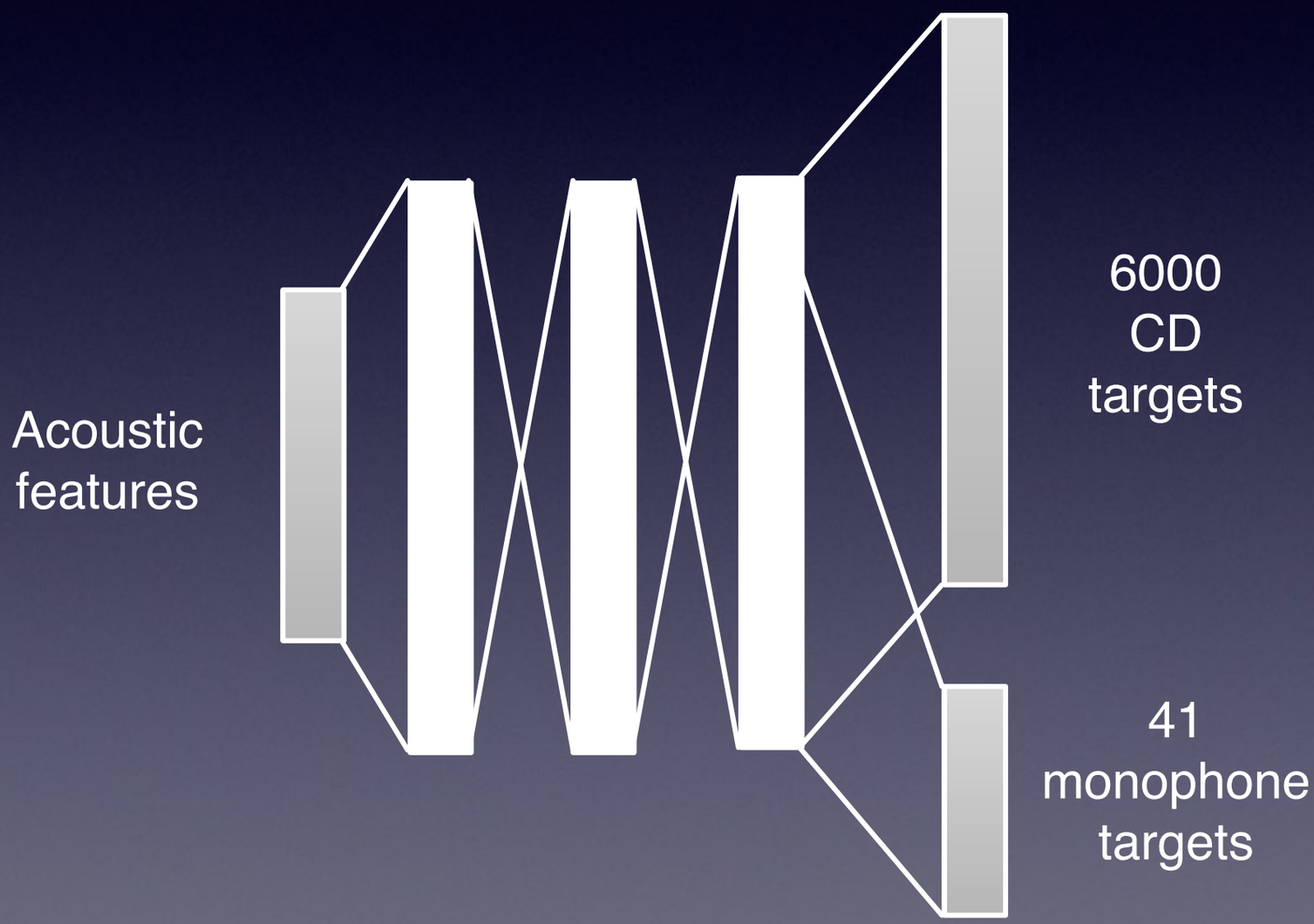
- Multitask learning – learning CD and CI phones together

- Interpret as regulariser? Compare with pretraining

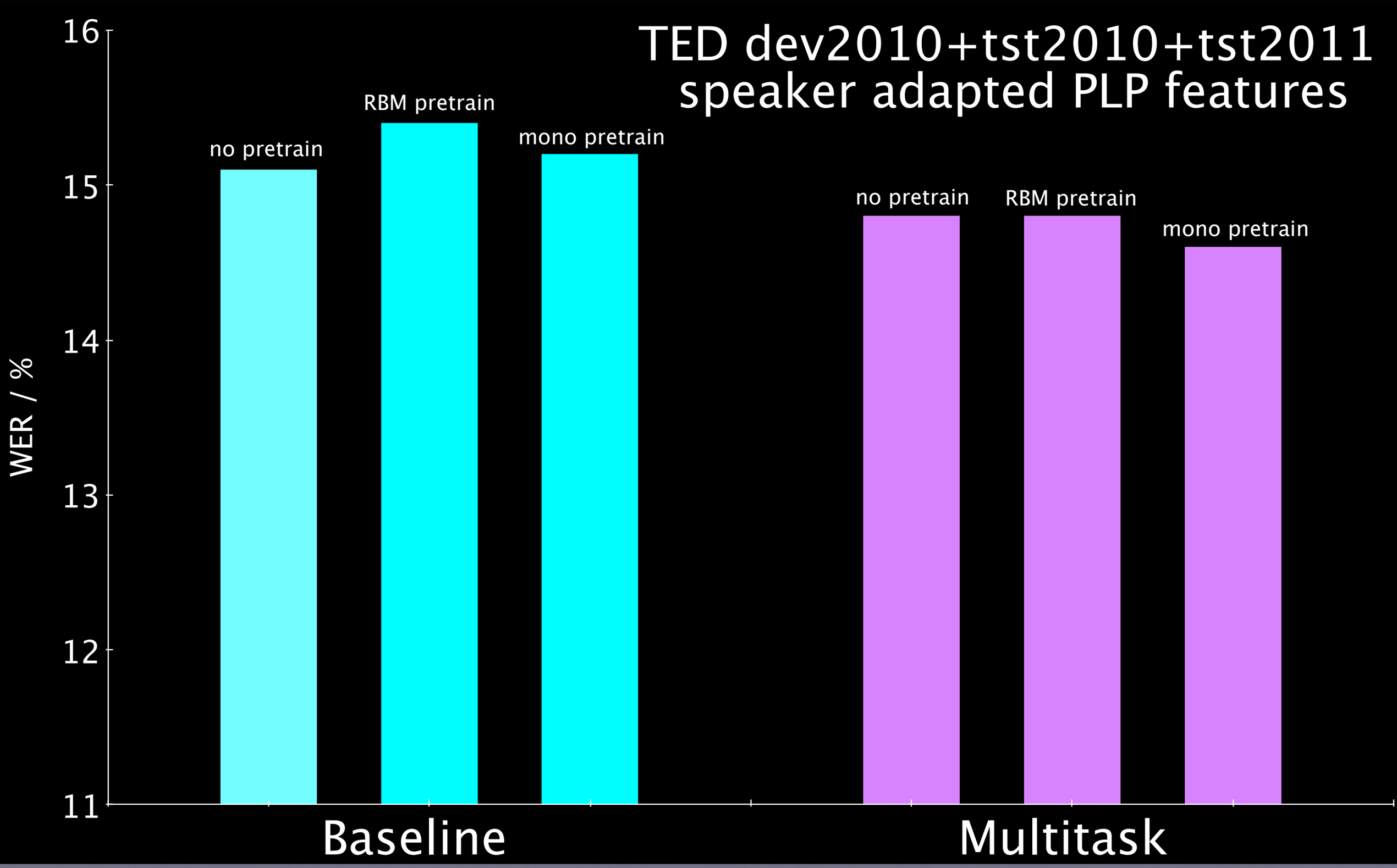
  - unsupervised RBM

  - discriminative monophone pretraining (cf curriculum learning)

- Can combine pretraining with multitask learning



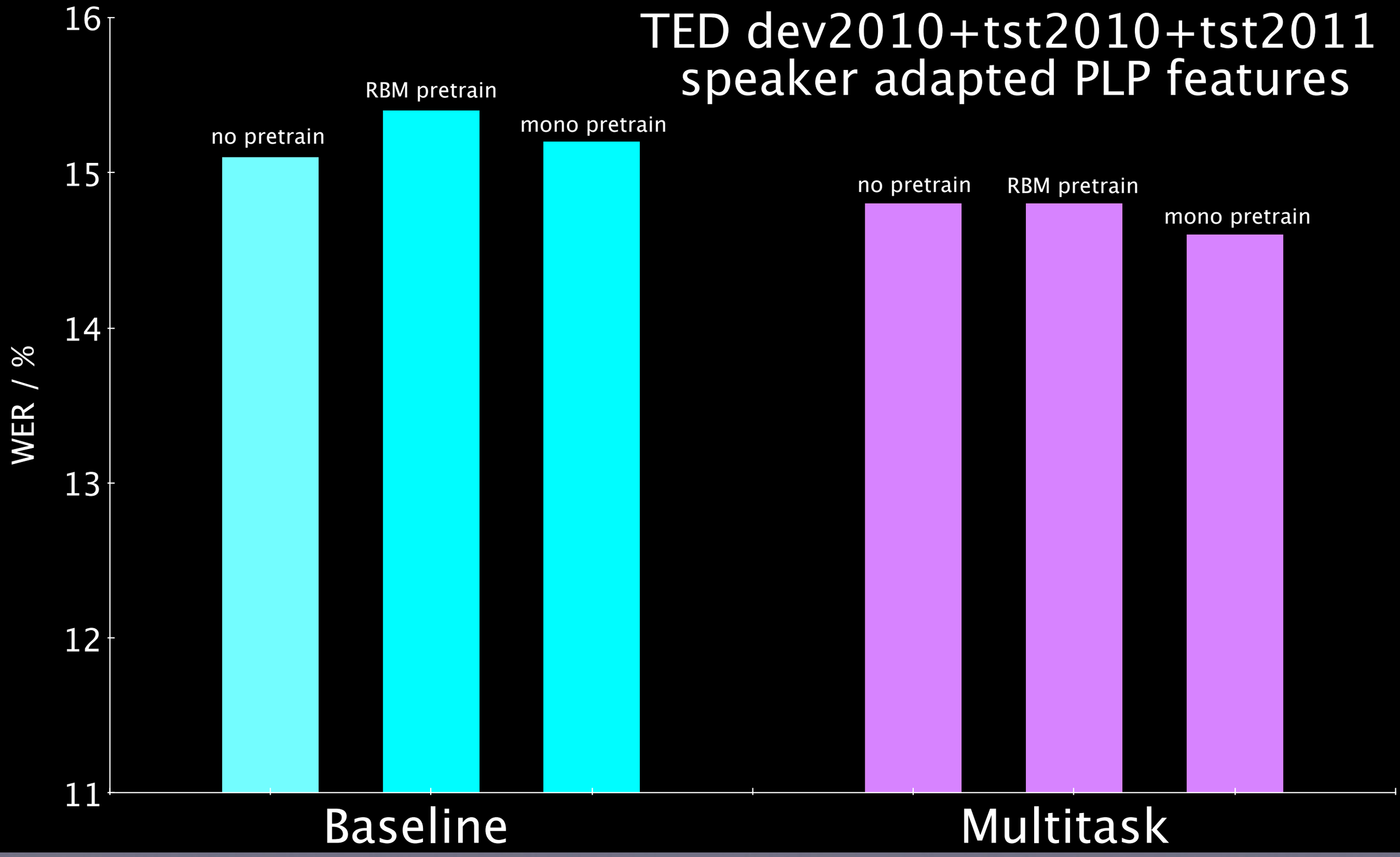
Experiment


TED dev2010+tst2010+tst2011
speaker adapted PLP features
WER / %
16
15
14
13
12
11
no pretrain
RBM pretrain
mono pretrain
no pretrain
RBM pretrain
mono pretrain
Baseline
Multitask

5% relative improvement

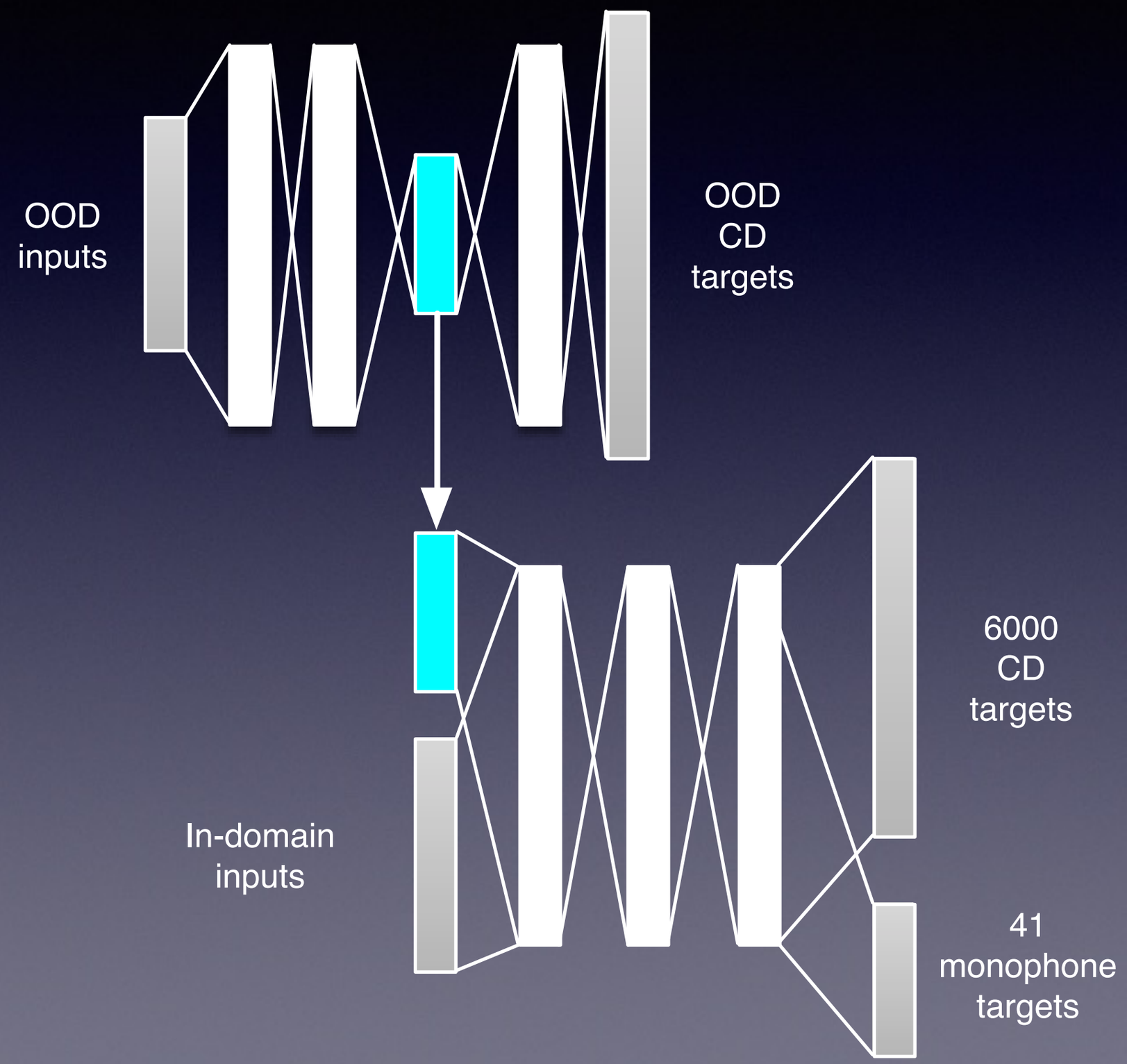
Experiment

TED dev2010+tst2010+tst2011
speaker adapted PLP features



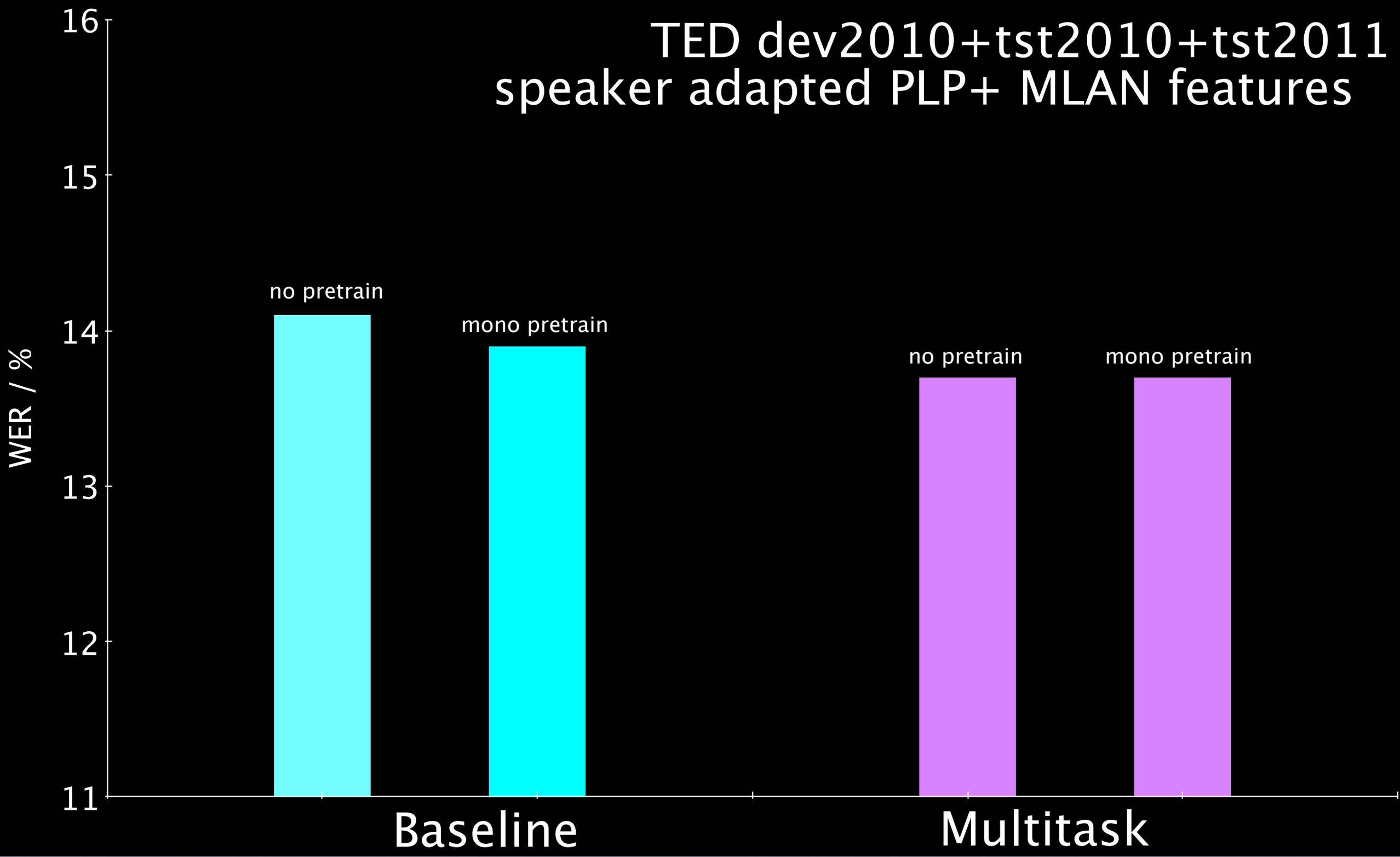
MLAN features

- Tandem/bottleneck features from OOD NN
- OOD net trained on Switchboard



Experiment

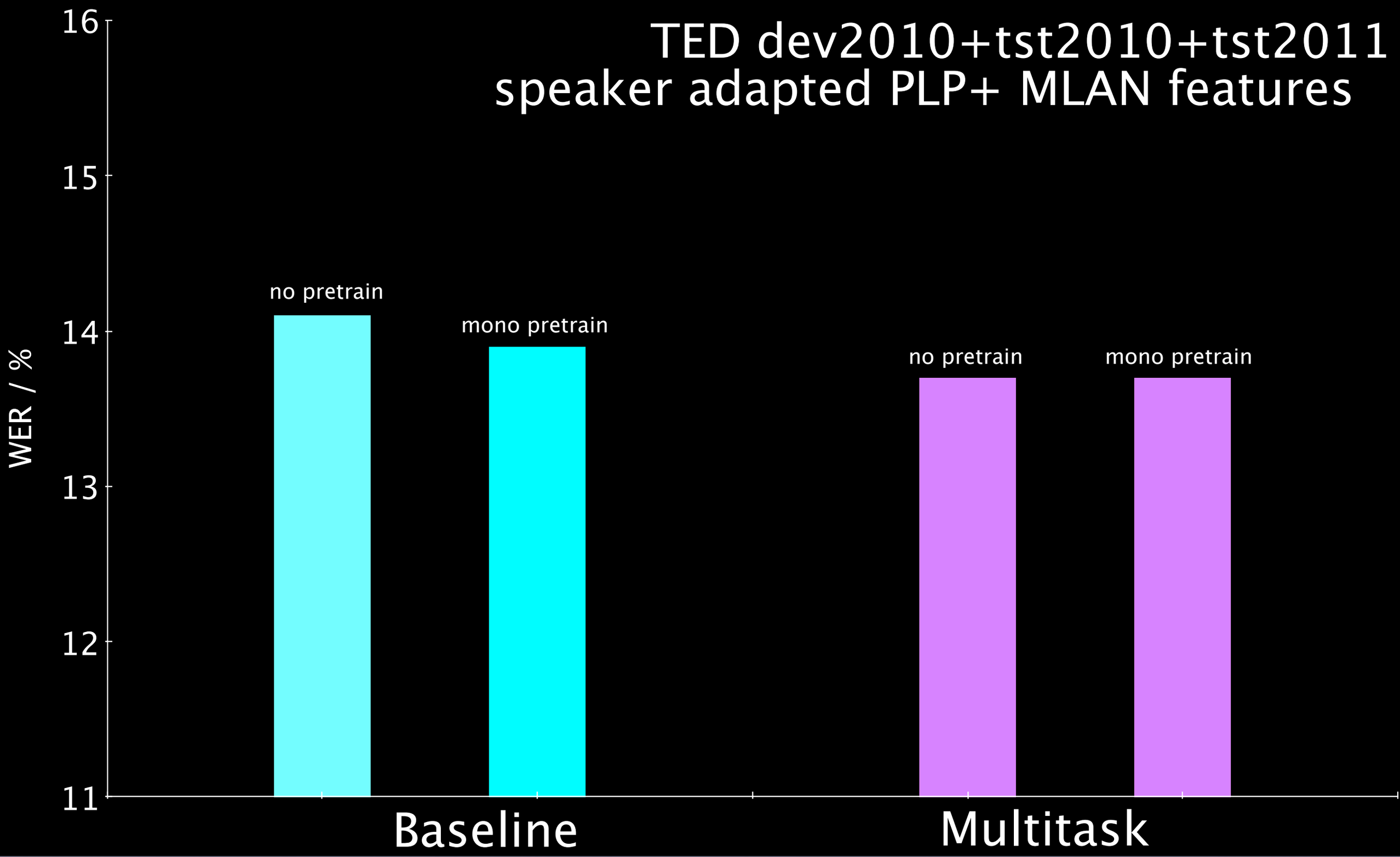
TED dev2010+tst2010+tst2011
speaker adapted PLP+ MLAN features



3% relative improvement

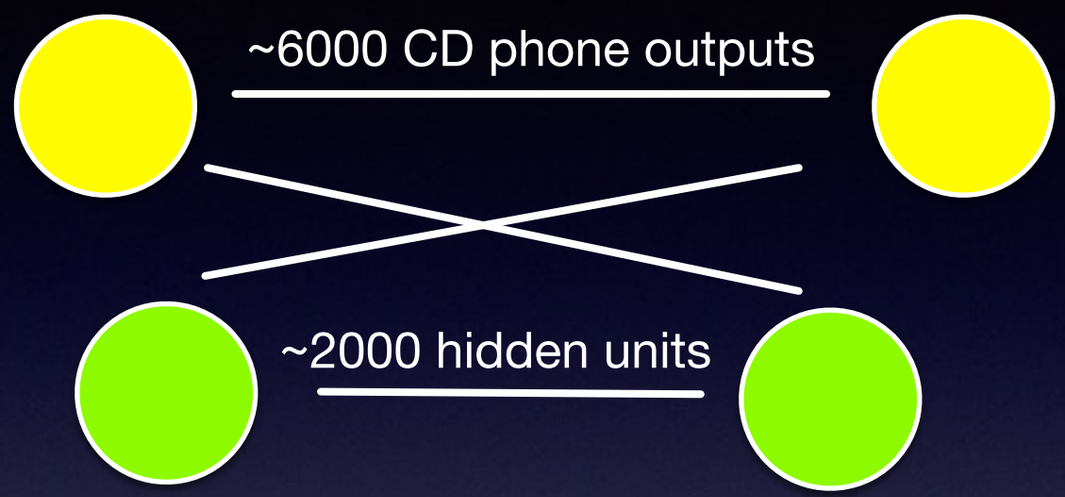
Experiment

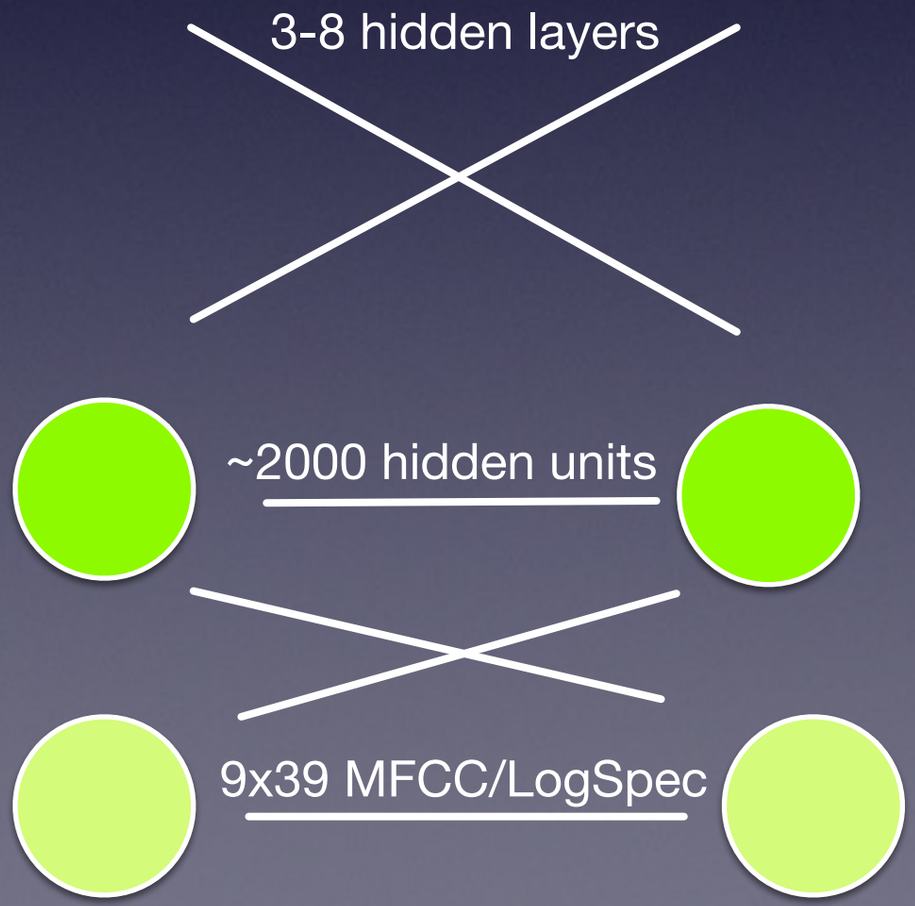
TED dev2010+tst2010+tst2011
speaker adapted PLP+ MLAN features



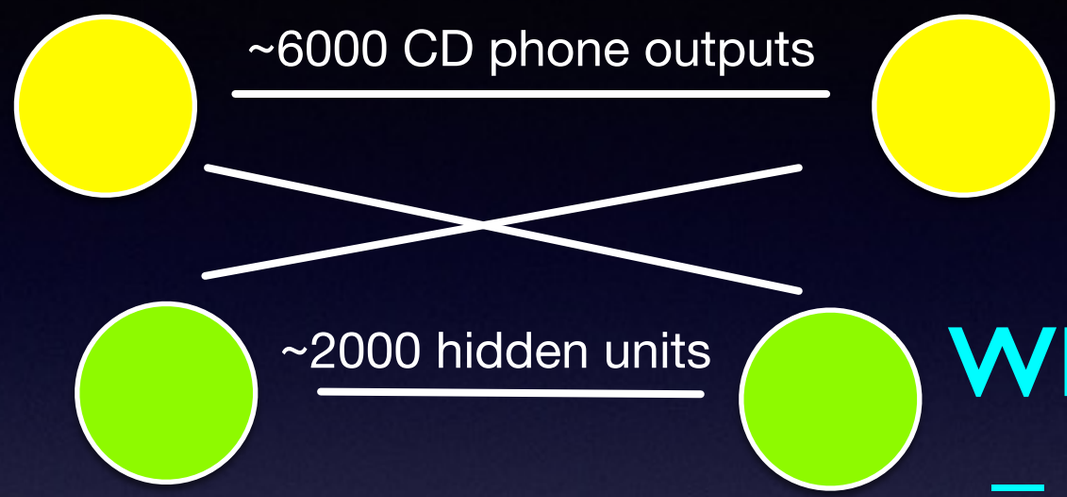
Practical Optimism

Practical Optimism



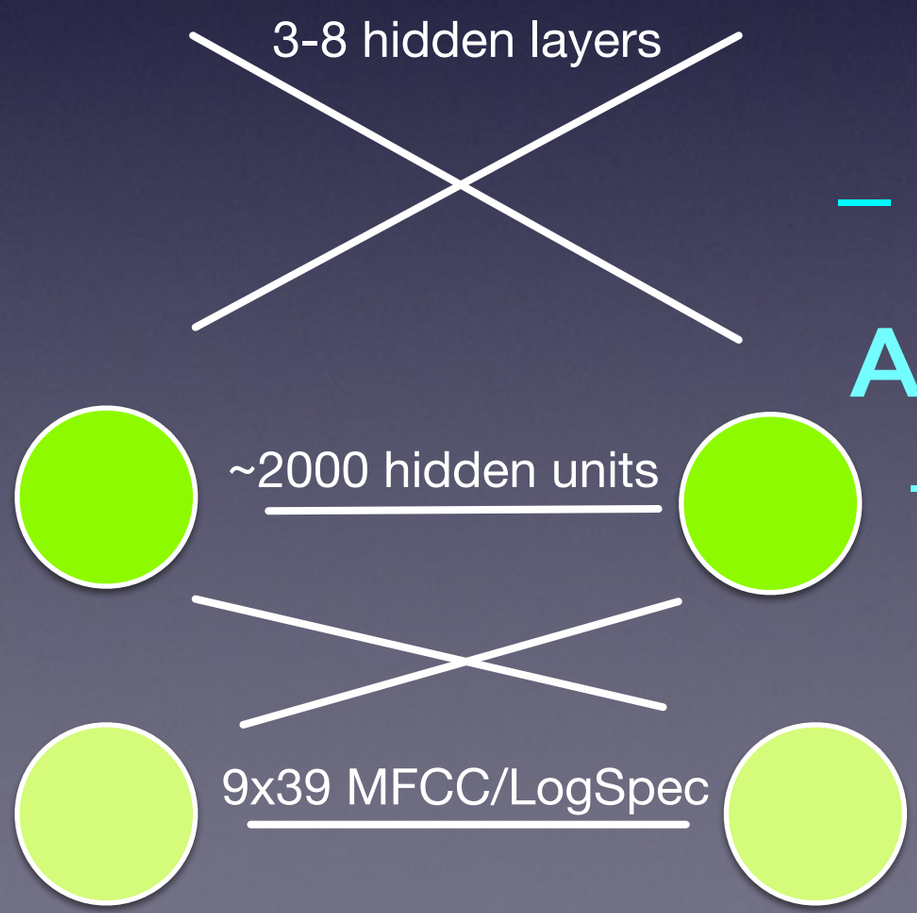


Practical Optimism



TRAINING
– objective function, optimisation

WEIGHT SHARING
– adaptation, CNNs



ARCHITECTURES
– convolutional, recurrent

ACTIVATION FUNCTIONS
– pooling, RELU, gated units

ACOUSTIC INPUT
– learning features?

Closing thoughts

- Morgan (and Hervé) set the framework within which all the current HMM/NN acoustic modelling stuff resides
- What a good idea to build a supercomputer to do NN training at scale 25 years ago
- And quite a few of things that seem to have been forgotten are worth remembering

Closing thoughts

- Morgan (and Hervé) set the framework within which all the current HMM/NN acoustic modelling stuff resides
- What a good idea to build a supercomputer to do NN training at scale 25 years ago
- And quite a few of things that seem to have been forgotten are worth remembering

Thanks!