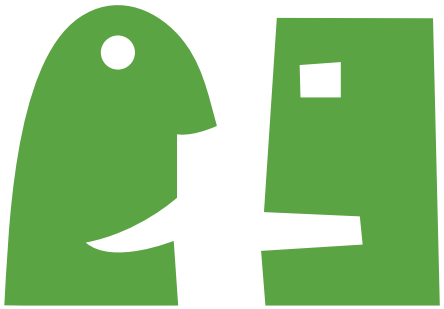
A lecture transcription system combining neural network acoustic and language models



Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Peter Bell

THE UNIVERSITY of EDINBURGH

23 May 2013

This talk

Natural Speech Technology

- We describe a system for lecture transcription built collaboratively by the University of Edinburgh and the National Institute of Information and Communications Technology in Japan (NICT)
- Joint work with Pawel Swietojanski, Fergus McInnes, Hitoshi Yamamoto and Youzheng Wu
- The system combines acoustic models from Edinburgh with language models from NICT
- Key novel features of the system:
  - Deep neural networks incorporating out-of-domain features
  - Factored recurrent neural network language models
- We achieve very competitive results on the speech recognition task from IWSLT

THE UNIVERSITY of EDINBURGH

TED lectures

Natural Speech Technology

The system is designed for automatic transcription of TED talks. (We also perform machine translation on the output).

TED is a nonprofit devoted to Ideas Worth Spreading. It started out (in 1984) as a conference bringing together people from three worlds: **Technology, Entertainment, Design.** Since then its scope has become ever broader. Along with two annual conferences -- the TED Conference on the West Coast each spring, and the TEDGlobal conference in Edinburgh UK each summer -- TED includes the award-winning TED Talks video site, the Open Translation Project and TED Conversations, the inspiring TED Fellows and TEDx programs, and the annual TED Prize.

The IWSLT evaluation

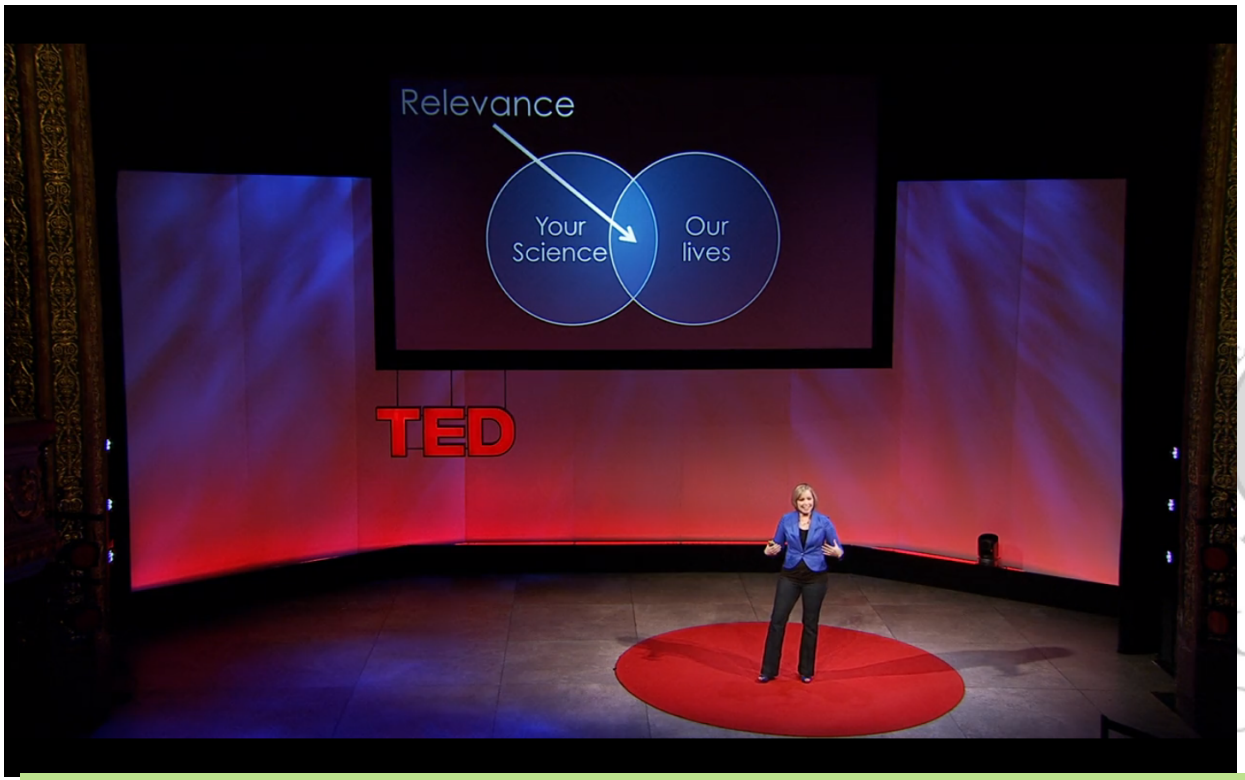
Natural Speech Technology

- The annual IWSLT brings together researchers in speech recognition (ASR) and machine translation
- In recent years, it has set a series of challenges for researchers, centred around the transcription and translation of TED talks.
- NICT had the best performing ASR system at IWSLT 2012

THE UNIVERSITY of EDINBURGH

The TED lecture task

The TED lecture "ASR task" is defined for the IWSLT evaluation campaign. The task consists of several development/test sets, each containing 8-11 single-speaker talks of around 10 minutes' duration. All talks are pre-segmented into utterances.



Characteristics of the task

- Generally clear, planned speech directed at an audience
- Single speaker per talk
- Training data is readily available on the web
- A wide vocabulary is used

THE UNIVERSITY
of EDINBURGH

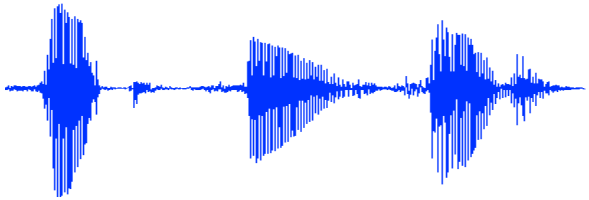
The key components of the system

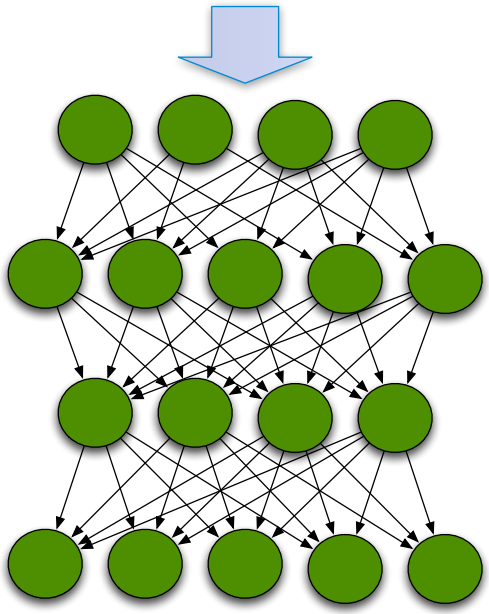
Natural Speech Technology

- Acoustic models based on **deep neural networks**
- Domain adaptation with **multi-level** networks
- **Recurrent neural network** language models

THE UNIVERSITY of EDINBURGH

Deep neural networks acoustic models





Tandem vs hybrid neural network systems

Tandem:

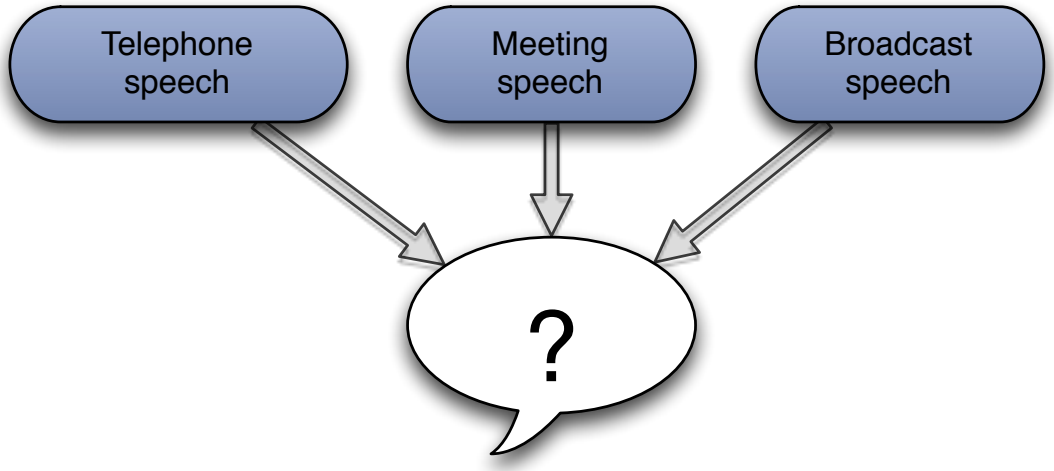
- Neural networks are used to derive features from training data, which are augmented with the standard acoustic features and used to train gaussian mixture models (GMMs).
- Can use decorrelated posterior features (eg over monophones), or bottleneck features

Hybrid:

- Neural networks used to generated posterior probabilities over states, used as likelihoods in decoder, scaled by state priors.
- In modern systems, we model the probabilities of tied triphone states.

We always use deep neural networks with RBM pre-training.

Domain adaptation



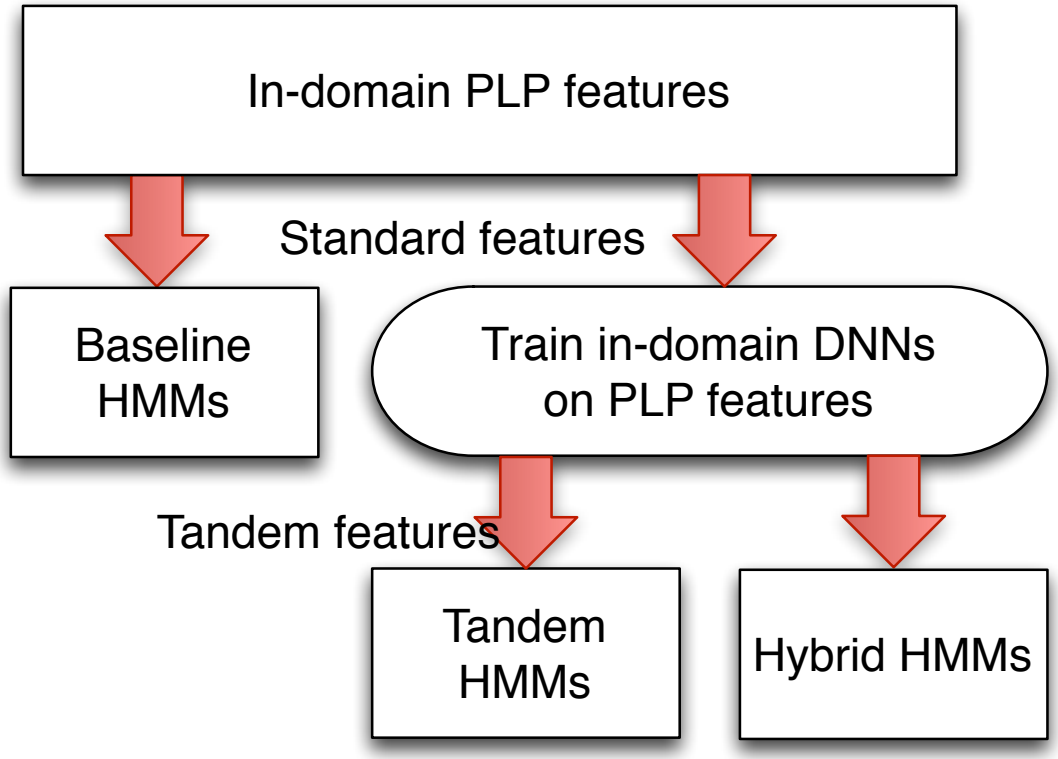
- If we are starting out on a new speech recognition task, it may be helpful to use data that we already have from other domains.
- But speech varies a lot in style, accent, and environmental conditions – how can we use out-of-domain data without harming performance?

Domain adaptation with DNNs

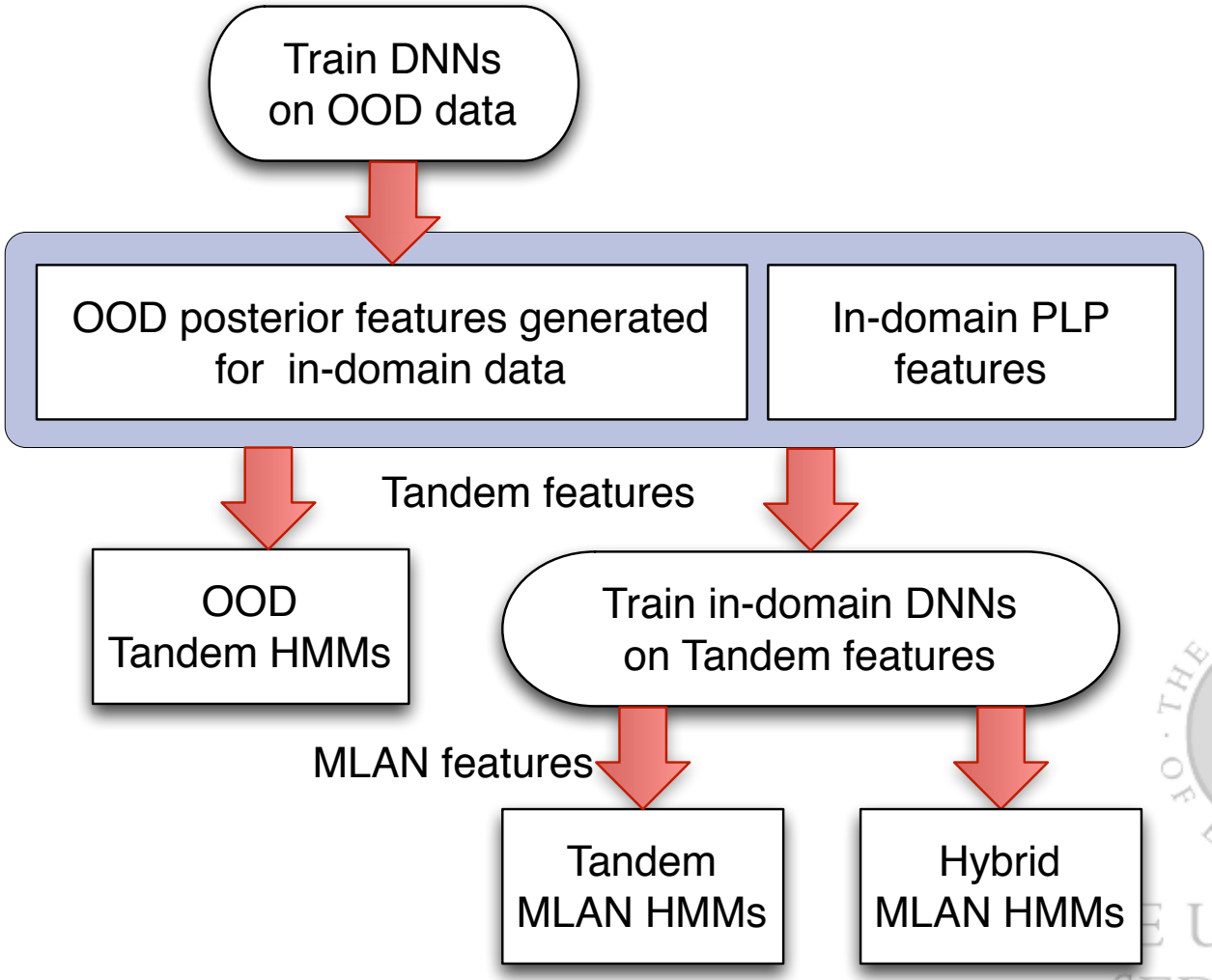
- Features derived from neural networks are known to provide a degree of domain-independence.
- Features trained on one domain may be used add discriminative ability to another domain (perhaps one where data is limited)
- Multi-level adaptive networks (MLAN): use a second DNN to discriminatively select which OOD features are most effective in the new domain.

THE UNIVERSITY
of EDINBURGH

Standard scheme



MLAN scheme


Train DNNs
on OOD data
OOD posterior features generated
for in-domain data
In-domain PLP
features
Tandem features
OOD
Tandem HMMs
Train in-domain DNNs
on Tandem features
MLAN features
Tandem
MLAN HMMs
Hybrid
MLAN HMMs

Feature space adaptation

Natural Speech Technology

How to enable the hybrid system to benefit from speaker adaptation? (Particularly important in the when there is a lot of data for each speaker)

- The simplest method is to perform feature-space adaptation on the input acoustic features
- Estimate a single linear transform for each speaker using the baseline GMMs.
- Retrain the hybrid DNNs on the speaker-normalised feature space.

THE UNIVERSITY of EDINBURGH

Language models

Natural Speech Technology

In summary:

- Decoding is performed with a **trigram** model
- Lattices are rescored with a **4-gram** model
- We later score complete sentences with a **factored recurrent neural network** model
- Models are trained on the transcriptions of TED talks, with selected out-of-domain data
- All LMs used here were trained at NICT

THE UNIVERSITY of EDINBURGH

Language model adaptation

- There is a relatively small amount of in-domain data. We need to select out-of-domain text that matches the TED domain.
- Use **cross-entropy difference** metric which biases towards sentences that are both like the in-domain corpus D_I and unlike the average of the out-of-domain:

$$D_s = \{s \mid H_I(s) - H_O(s) < \tau\},\ s \in D_O$$

- The threshold τ is empirically set to minimise the perplexity of the development set

The baseline language model

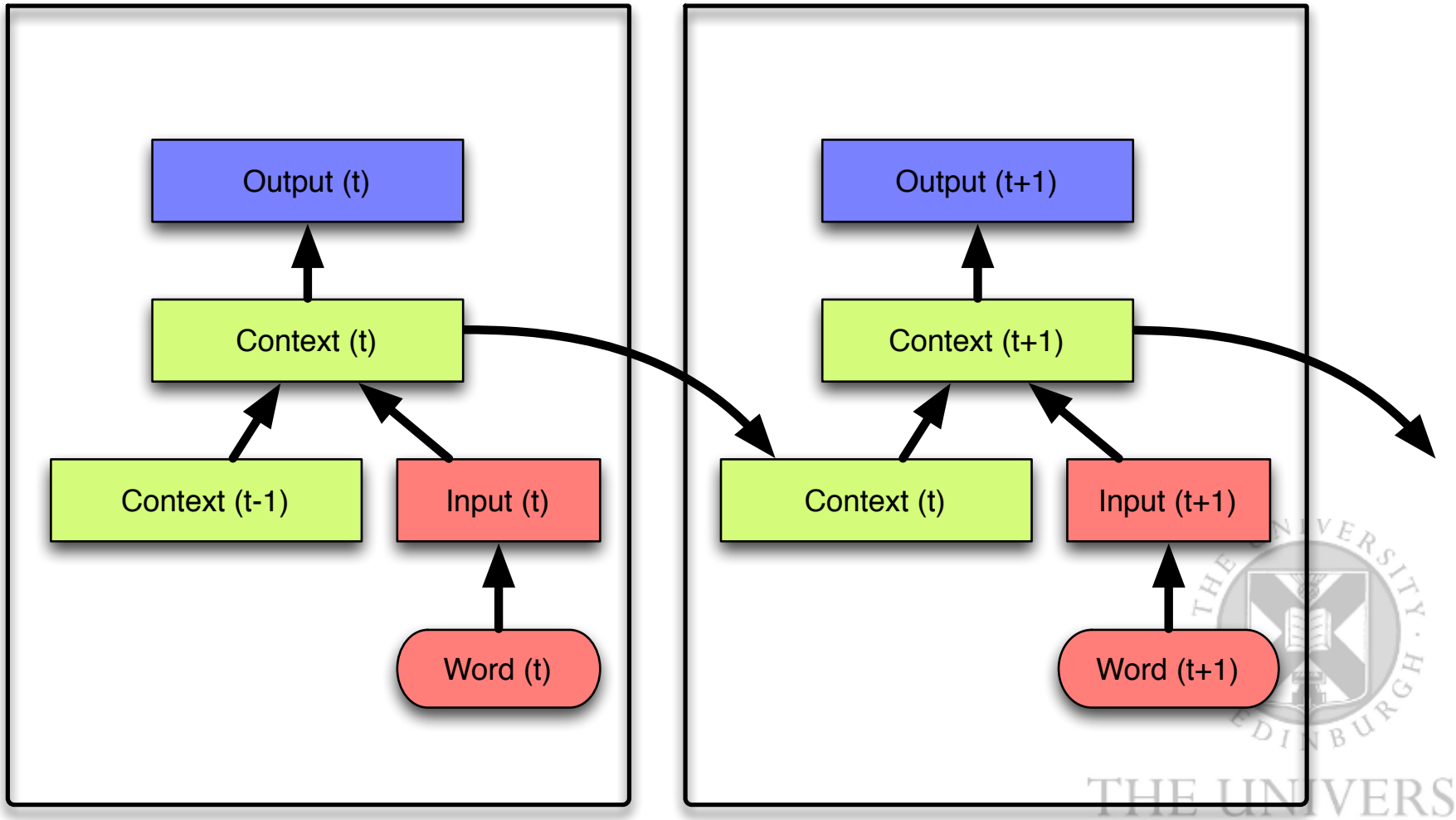
Natural Speech Technology

- Standard n-gram model using modified Kneser-Ney smoothing, trained on a mixture of in-domain and out-of-domain text data
- Some statistics:

Unigrams	**55,000**
Bigrams	**24,000,000**
Trigrams	**129,000,000**
4-grams	**43,000,000**

THE UNIVERSITY of EDINBURGH

Recurrent neural network language models



Rescoring *n*-best lists

Natural Speech Technology

- As the RNN LMs are not finite state, we rescore *n*-best lists, generated from the output lattices of the recogniser
- we can compute LM probabilities over whole sentences with the new model

```
-6796.31 -20.6985 8 <s> A BIT TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6787.48 -20.9777 8 <s> BUT WE TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6850.2  -19.0512 8 <s> A BIT TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6841.38 -19.3304 8 <s> BUT WE TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6787.89 -21.4919 8 <s> THEY'LL BE TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6841.79 -19.8447 8 <s> THEY'LL BE TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6729.71 -23.5939 8 <s> A BE TOO MUCH RAIN NOT ENOUGH FRAME </s>
```

Rescoring *n*-best lists

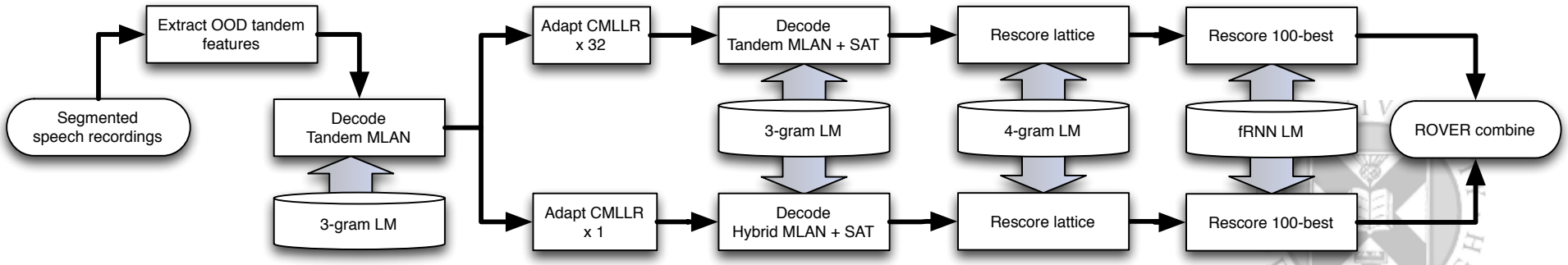
Natural Speech Technology

- As the RNN LMs are not finite state, we rescore *n*-best lists, generated from the output lattices of the recogniser
- we can compute LM probabilities over whole sentences with the new model

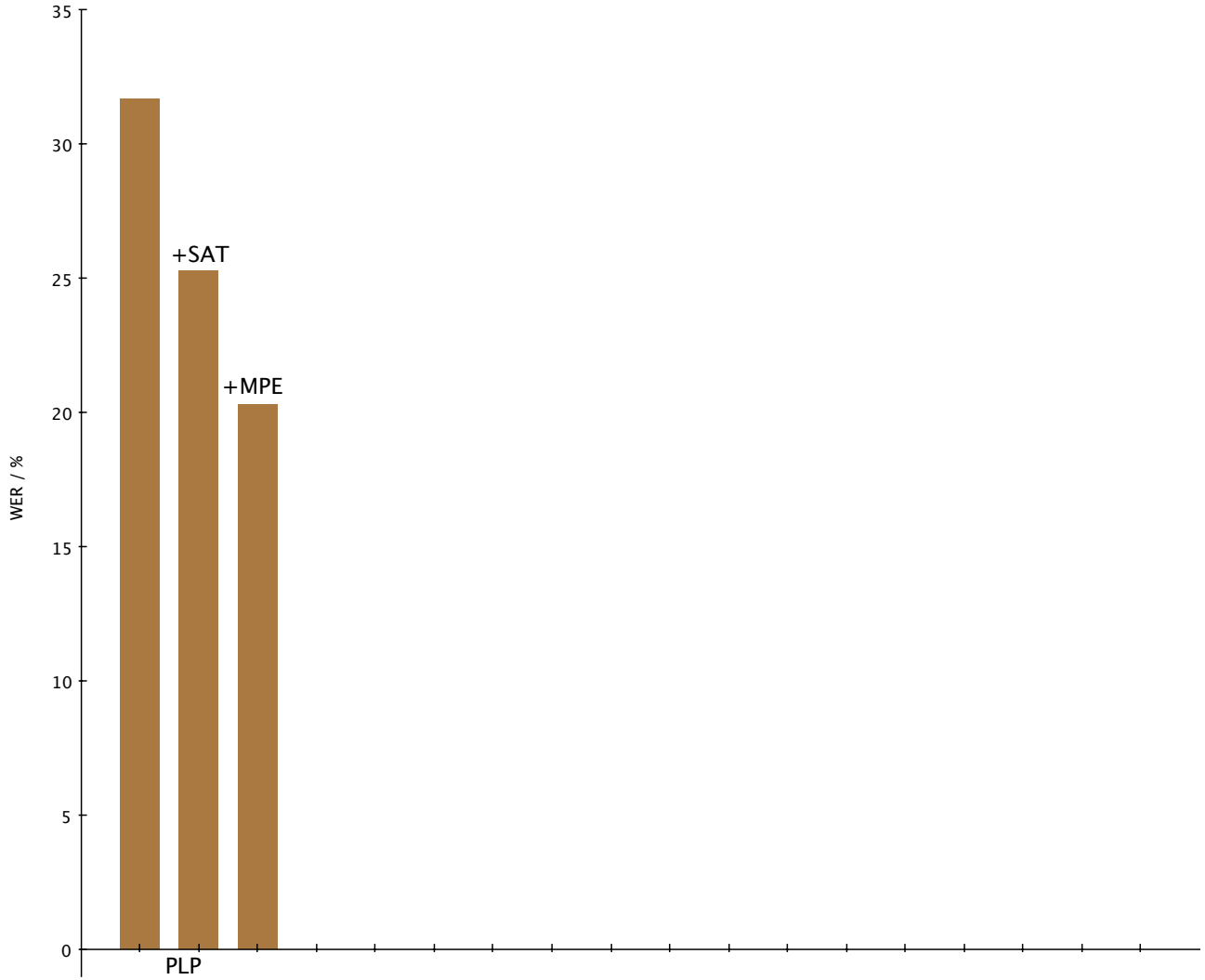
```
-6796.31 -20.6985 8 <s> A BIT TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6787.48 -20.9777 8 <s> BUT WE TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6850.2  -19.0512 8 <s> A BIT TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6841.38 -19.3304 8 <s> BUT WE TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6787.89 -21.4919 8 <s> THEY'LL BE TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6841.79 -19.8447 8 <s> THEY'LL BE TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6729.71 -23.5939 8 <s> A BE TOO MUCH RAIN NOT ENOUGH FRAME </s>
```

The complete system

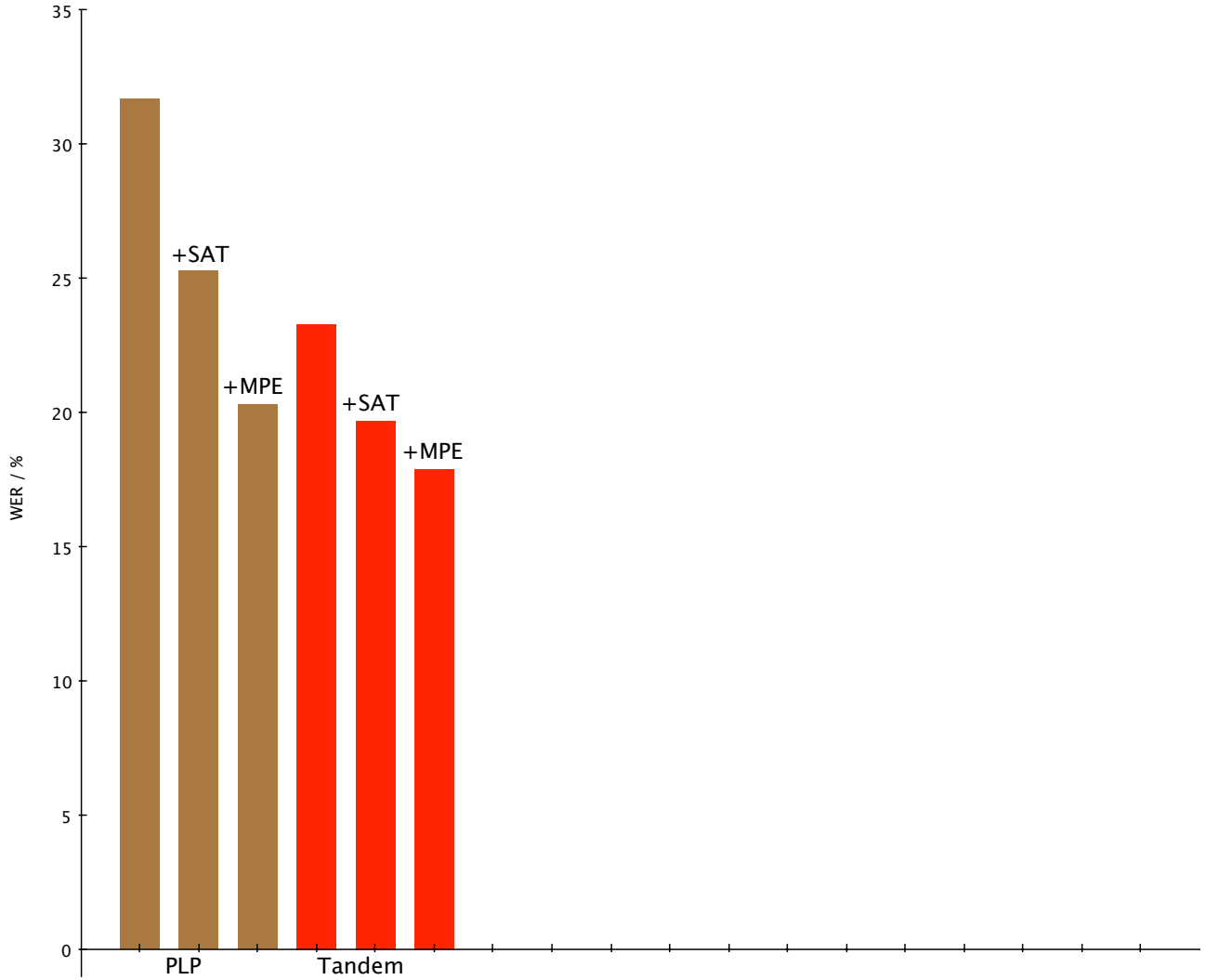
- We run a first-pass recognition which is used to estimate speaker adaptation transforms
- Then we run the complete decoding process with both Tandem and Hybrid models using MLAN features
- Finally, the systems are combined using ROVER



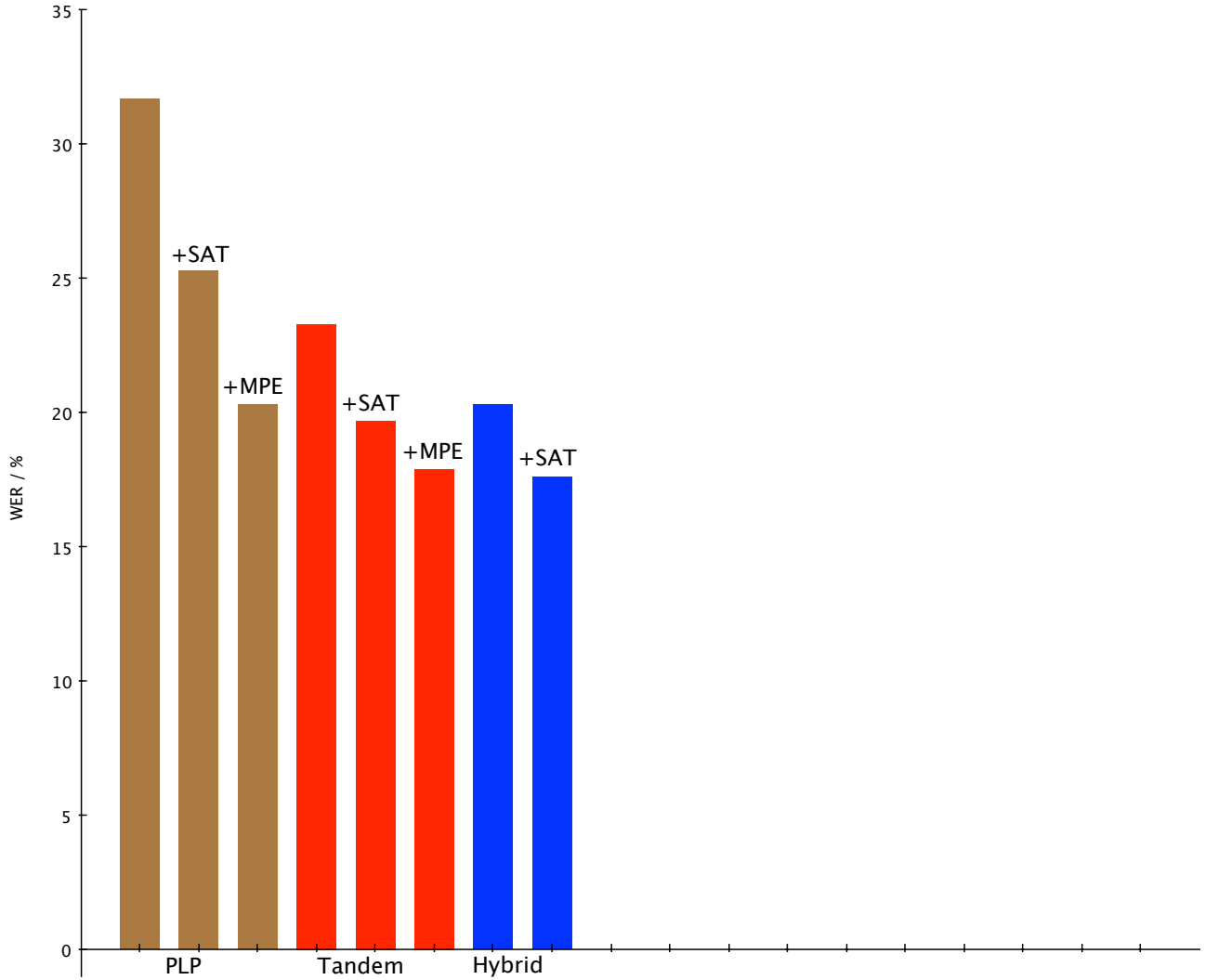
Results on the 2010 test set



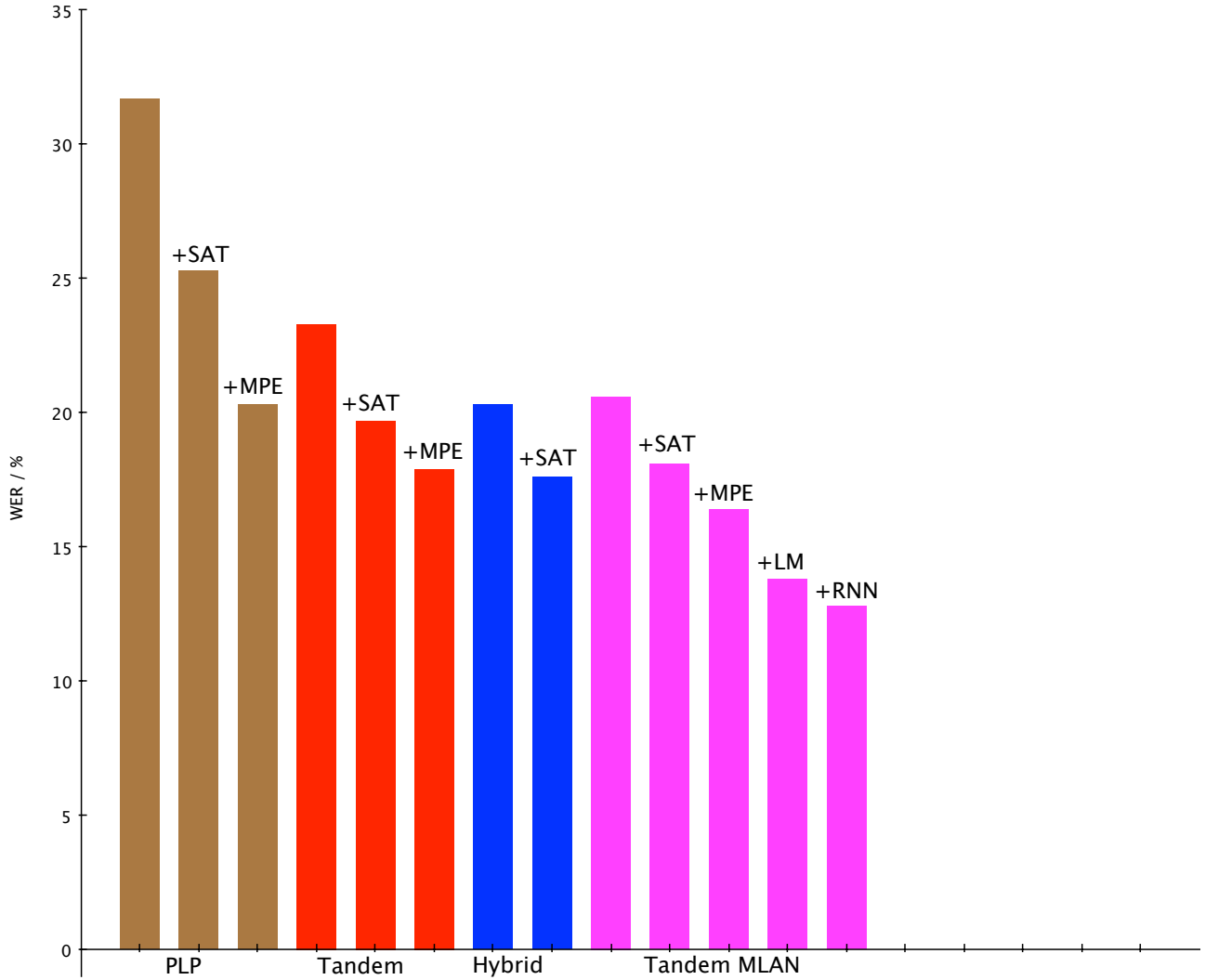
Results on the 2010 test set



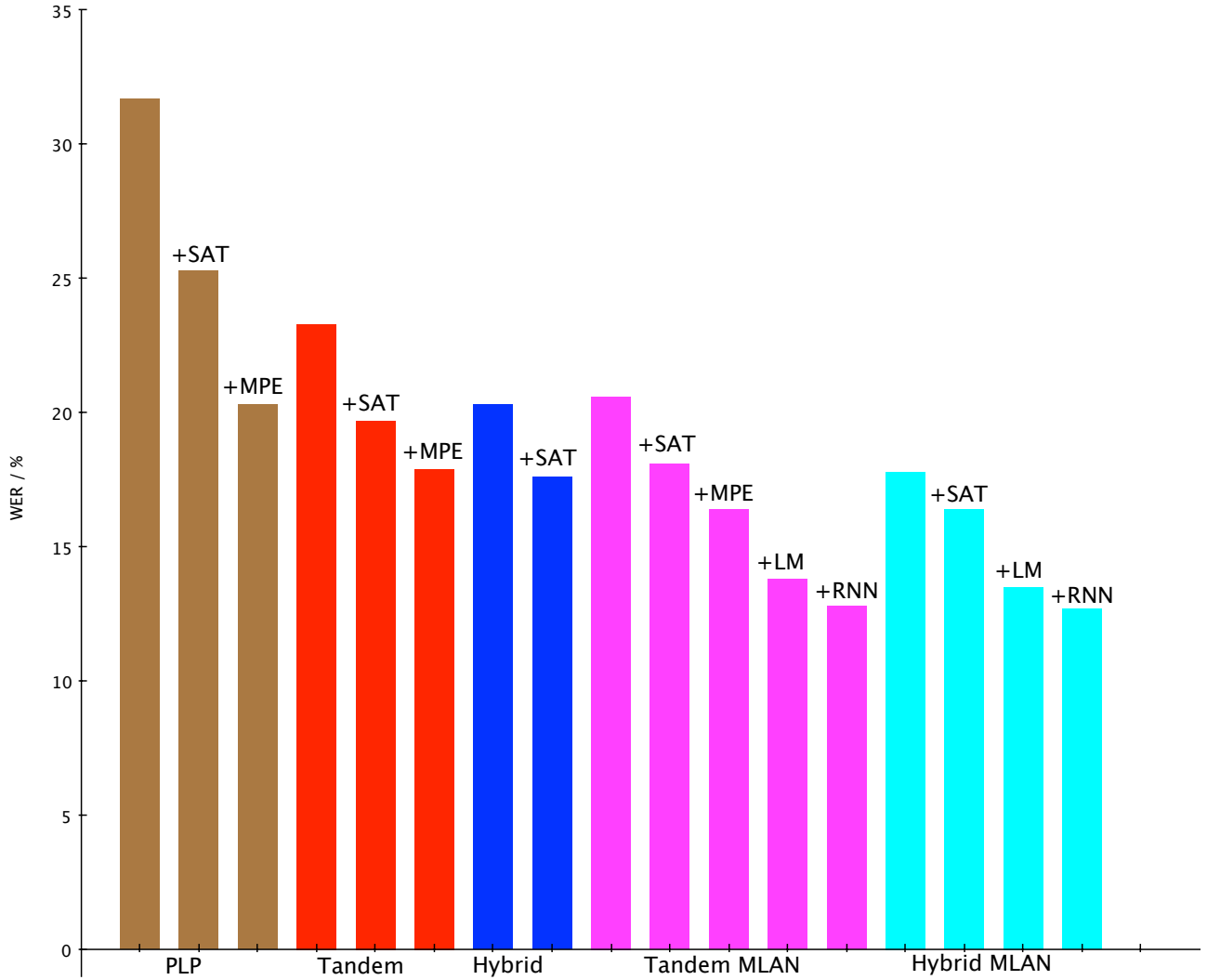
Results on the 2010 test set



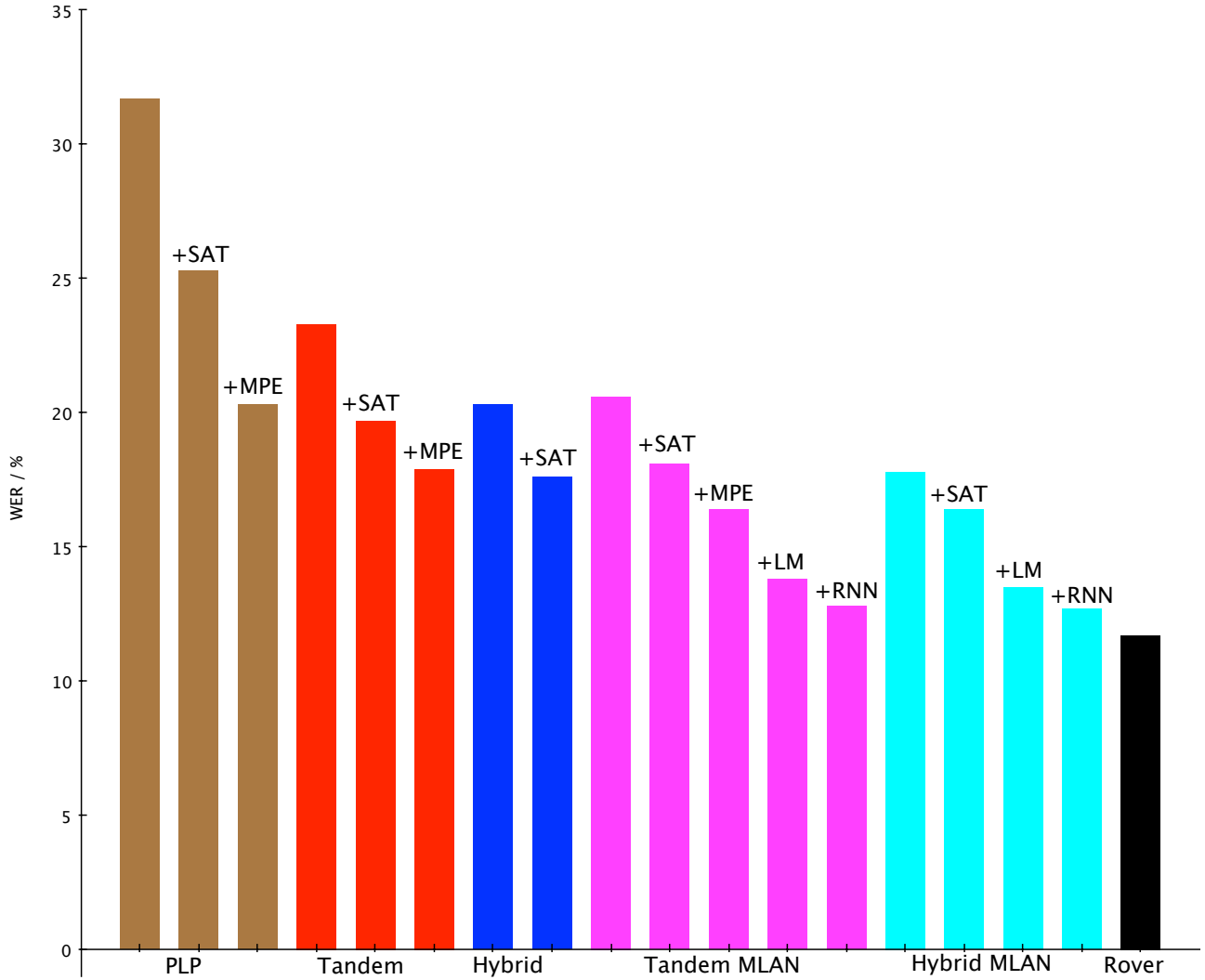
Results on the 2010 test set



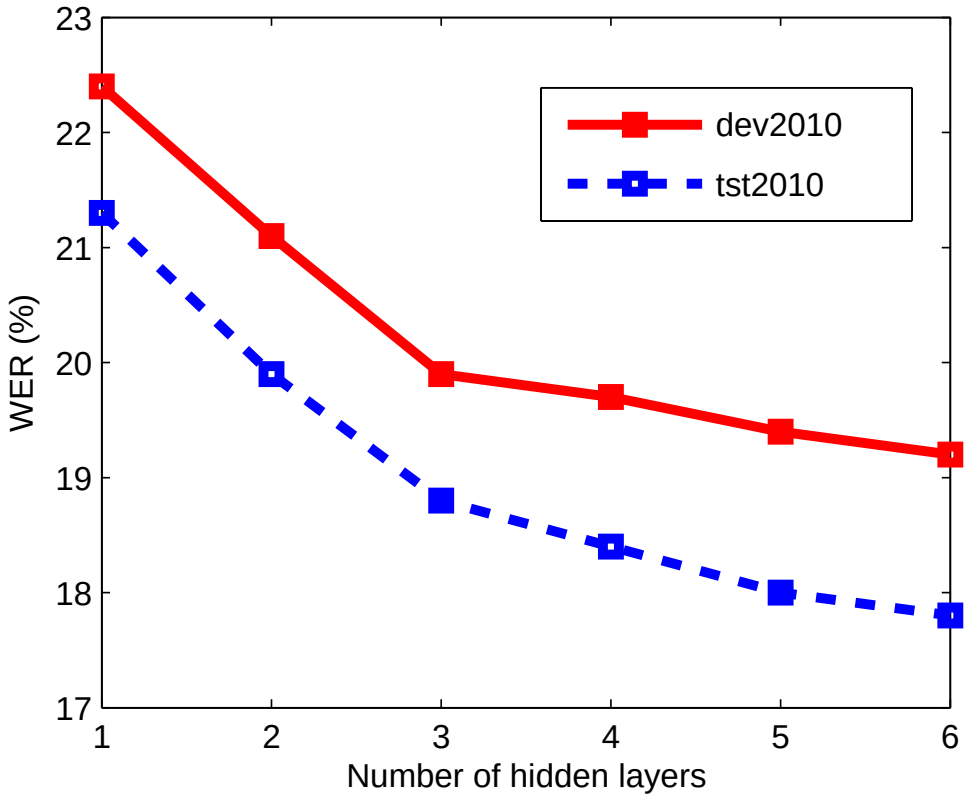
Results on the 2010 test set



Results on the 2010 test set



Increasing the number of layers in the DNN



The effect of increasing the number of DNN layers for the hybrid MLAN systems on TED lectures (systems without SAT)

Results of 2012 evaluation

Natural Speech Technology

System	tst2011	tst2012
FBK	15.4	16.8
RWTH	13.4	13.6
UEDIN	12.4	14.4
KIT-NAIST	12.0	12.4
MITLL	11.1	12.4
NICT	10.9	12.1

THE UNIVERSITY of EDINBURGH

Results of 2012 evaluation

Natural Speech Technology

System	tst2011	tst2012
FBK	15.4	16.8
RWTH	13.4	13.6
UEDIN	12.4	14.4
KIT-NAIST	12.0	12.4
MITLL	11.1	12.4
NICT	10.9	12.1

NICT-UEDIN systems	tst2011	tst2012
Tandem MLAN + fRNN	10.2	11.4
Hybrid MLAN + fRNN	10.3	11.3
ROVER combination	**9.3**	**10.3**

What next?

- Investigate the best choice of out-of-domain acoustic features
- Consider powerful methods for speaker adaptation of the hybrid DNN systems
- Further integrate the system with machine translation

References

- H. Yamamoto, Y. Wu, C.-L. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR system for the IWSLT2012," in *Proc. IWSLT*, 2012.
- E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN systems for the IWSLT 2012 evaluation," in *Proc. IWSLT*, 2012.
- P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc. ICASSP*, 2013, to appear.
- T. Mikolov, M. Karafiát, L. Burget, J. Černokcý, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010.