Reverberant speech recognition combining deep neural networks and deep autoencoders

Masato Mimura

Shinsuke Sakai

Tatsuya Kawahara

ACCMS, Kyoto University

The REVERB challenge 2014 workshop

Introduction

- Use **deep learning** in both **frontend** and **backend** of the speech recognizer to handle reverberant speech.
  - Frontend: speech feature enhancement (dereverberation) w/ **deep autoencoder**
  - Backend: acoustic modeling w/ **deep neural networks**

Our submitted results for the challenge and final results on paper

Our submitted results

	Room1		Room2		Room3		Ave	room1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
Real-time	12.9	13.4	15.9	26.4	18.5	30.5	19.6	52.2	52.3	52.3
Full batch	13.0	13.3	15.4	24.9	17.9	28.6	28.8	50.6	50.5	50.6

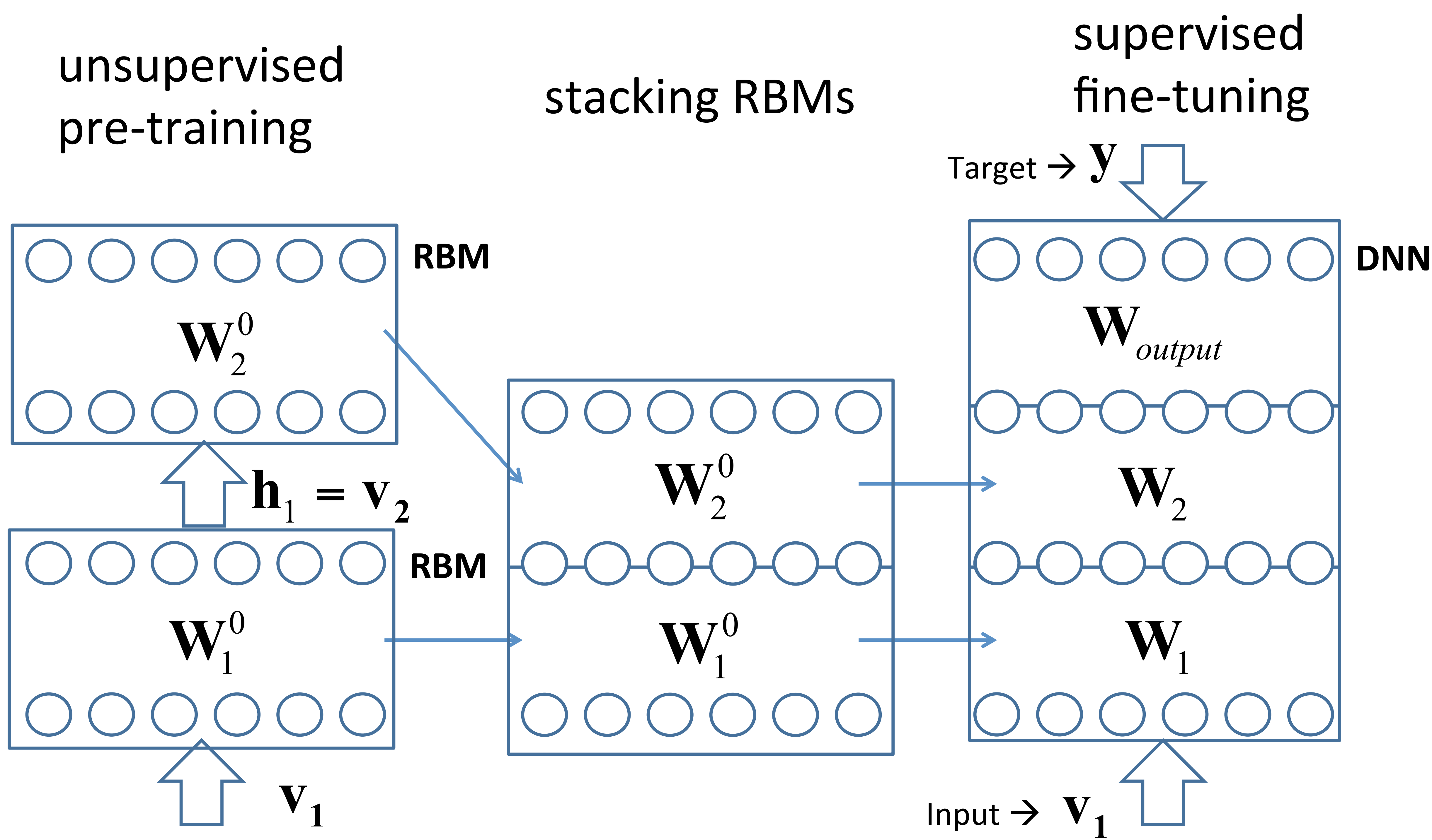
Forgot to include results by full batch adaptation in the paper. Sorry!

Our final results with DAE feature enhancement (and some bug fixes)

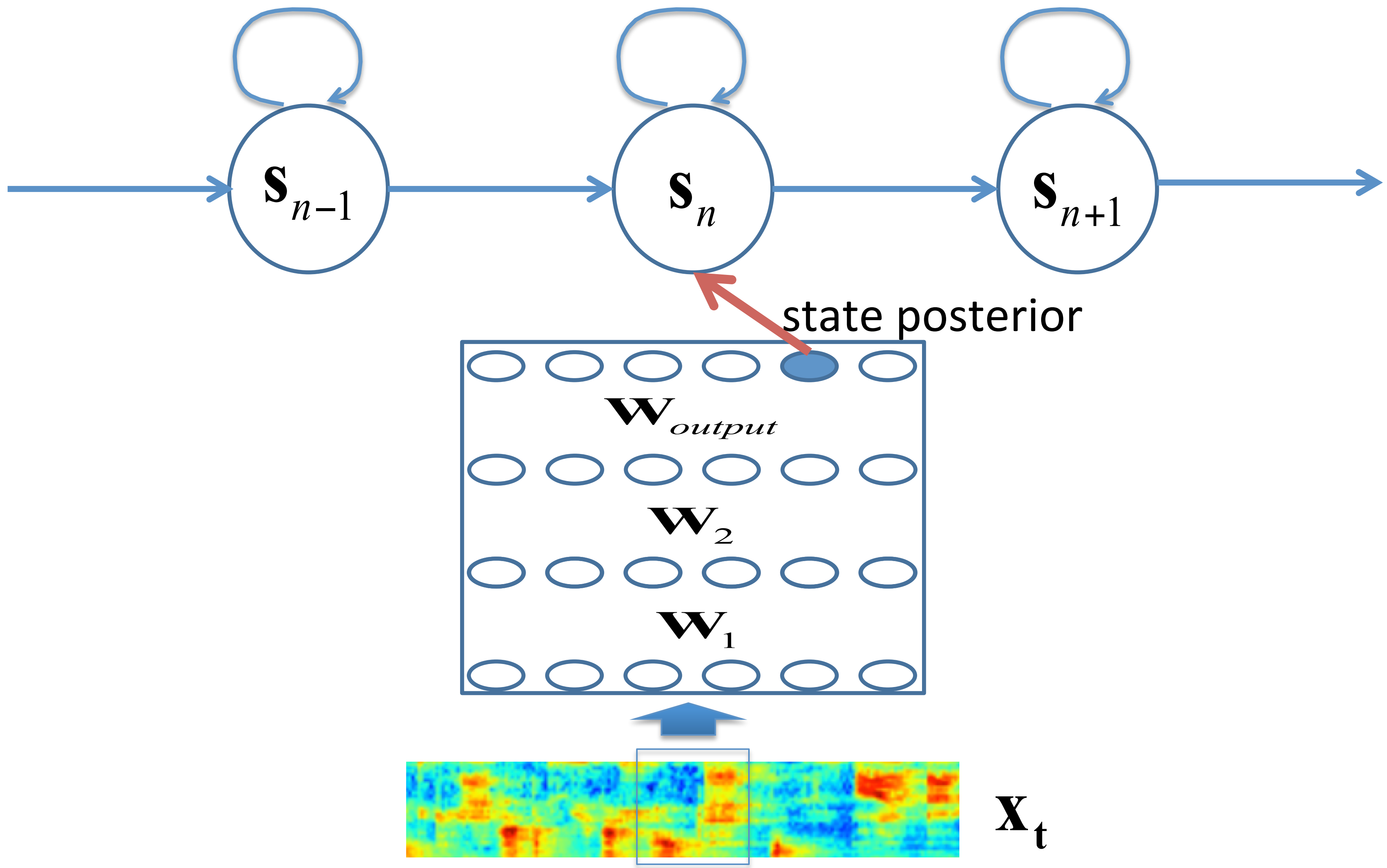
	Room1		Room2		Room3		Ave	room1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
Real-time(a)	10.3	10.6	12.9	21.4	14.1	23.3	15.5	49.3	48.1	48.7
Real-time(b)	14.2	14.2	13.3	19.5	14.0	18.8	15.7	45.5	45.2	45.4

Results with DAE enhancement not in time for result submission deadline

Standard procedure for training DNN

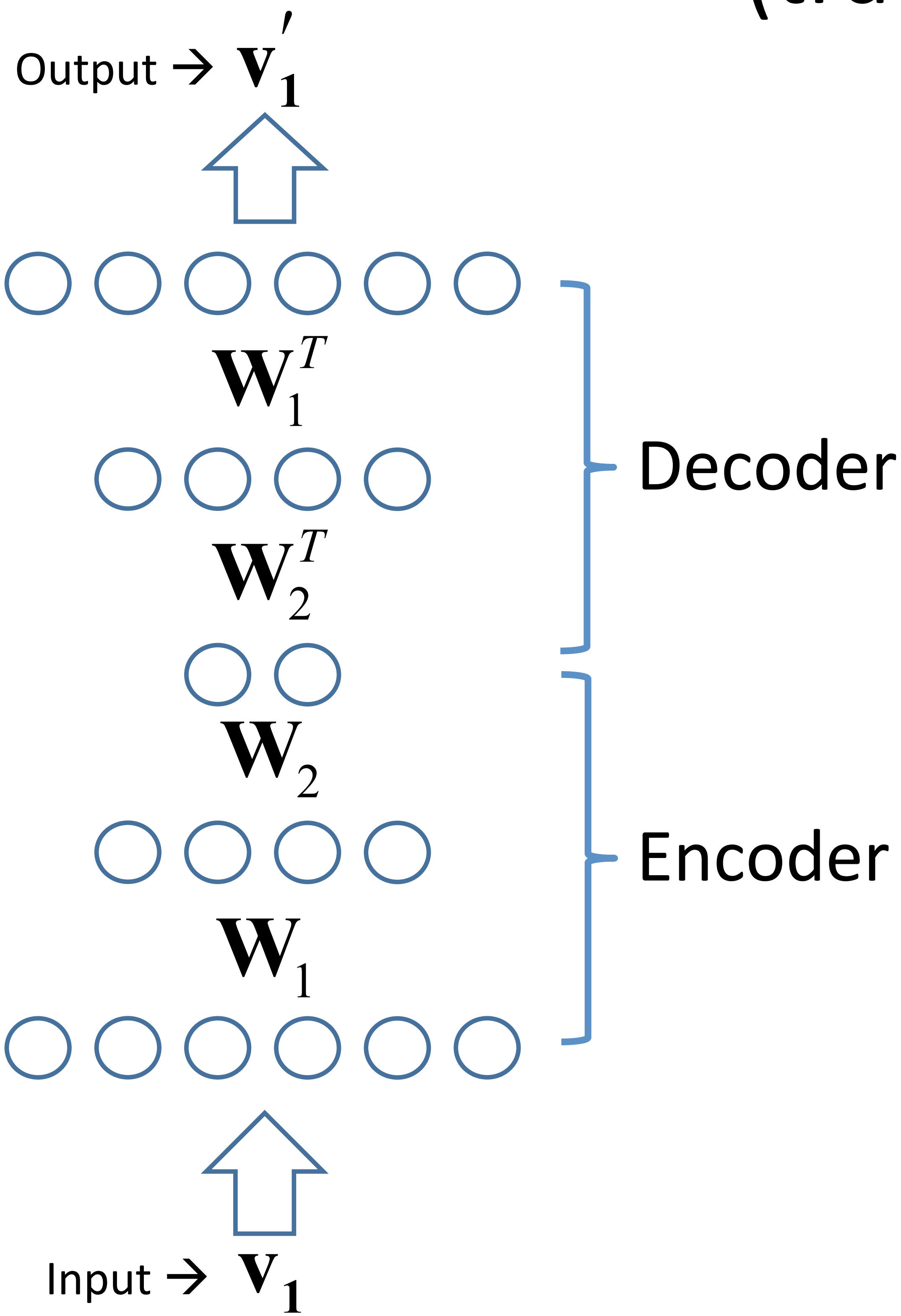


Hybrid model (DNN-HMM) [Mohamed 12][Dahl 12]



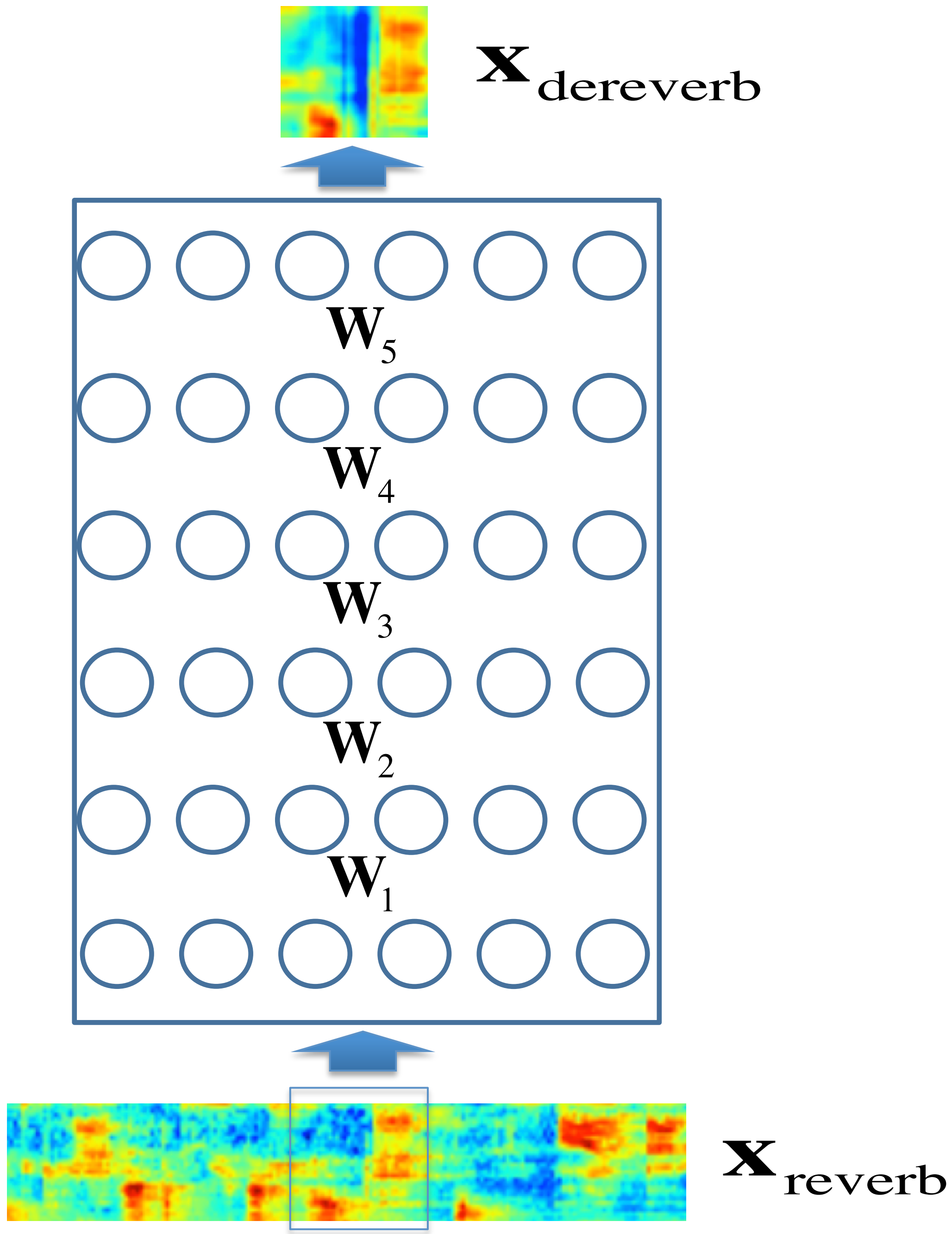
- GMMs for calculating state probabilities replaced by a single DNN
- Other parameters like transition probabilities copied from a well-trained GMM-HMM

Deep autoencorders (DAEs) [06 Hinton] (traditional)



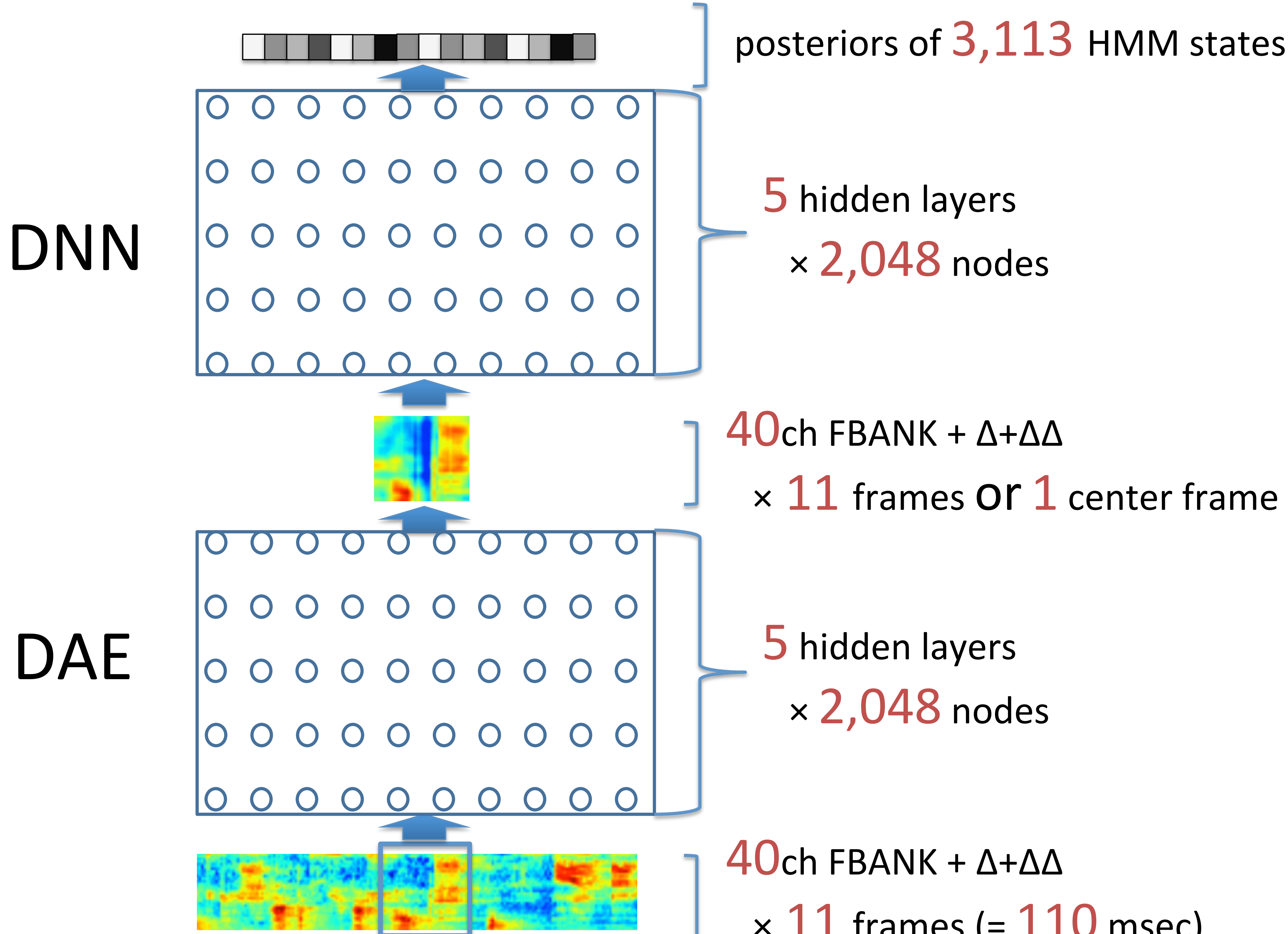
- Deep neural networks used for regression tasks
- Encoder layers generate compact representation for Decoder to recover the input data
- DAE trained as **denoising autoencoder**:
  - Input = corrupted data
  - Target = clean data

Deep autoencorders (DAEs)
(our network for dereverberation)



- Since our goal is not generating compact codes, we adopt network structure **without any bottleneck layer** for dereverberation

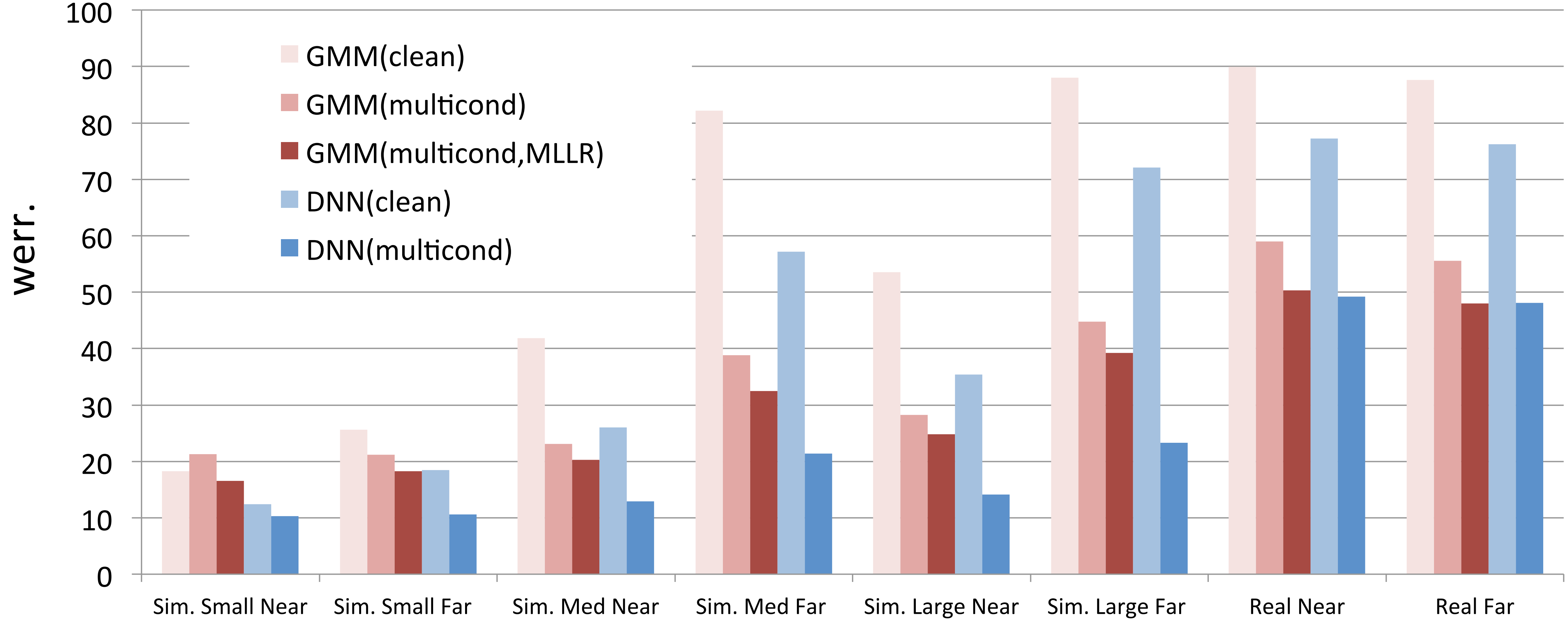
Our proposed network
(Combination of DNN-HMM and denoising DAE)



Speech recognition experiments

- DNN Training
  - **input:** Multi-condition data **target:** Frame-level state labels
- DAE Training
  - **input:** Multi-condition data **target:** Clean data
    - Reverberant speech frames and clean speech frames are **adjusted to be time-aligned**
- Test data
  - Simulated data: 3264 utts
    - Rooms: Small (T60 = 0.25s), Med (0.5s), Large (0.7s)
    - Mic. distances: Near (= 50cm), Far (= 200cm)
  - Real data: 372 utts:
    - Room: Large (T60 = 0.7s)
    - Mic. distances: Near (= 100cm), Far (= 250cm)

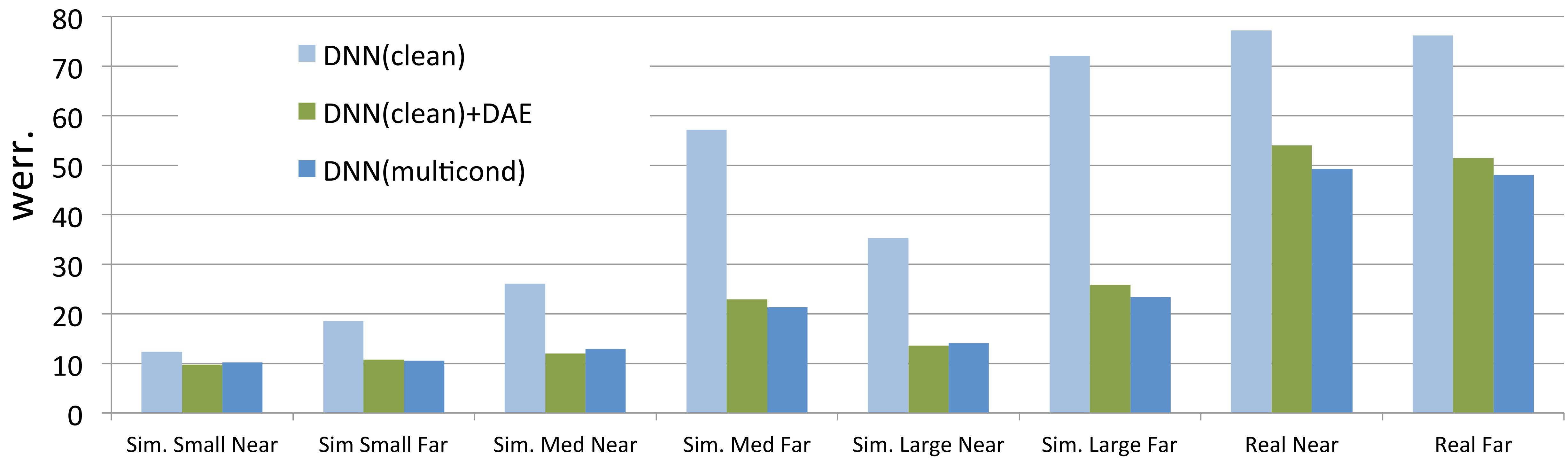
Performance of DNN-HMM for reverberant test data



vs. : **DNN-HMMs** achieves drastically higher accuracies than adapted **GMM-HMMs**

vs. : **multi condition training** effective for DNN-HMMs as well as GMM-HMMs (vs.)

Performance of DAE for reverberant test data

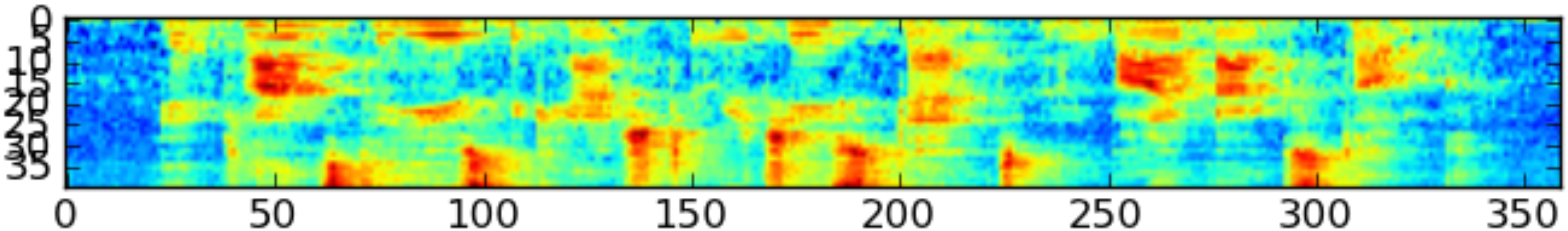


vs. : By **using DAE as frontend**, accuracies by clean DNN improved drastically

vs. : Interestingly, performance of **clean DNN combined with DAE** almost the same as **multicond. DNN without DAE**

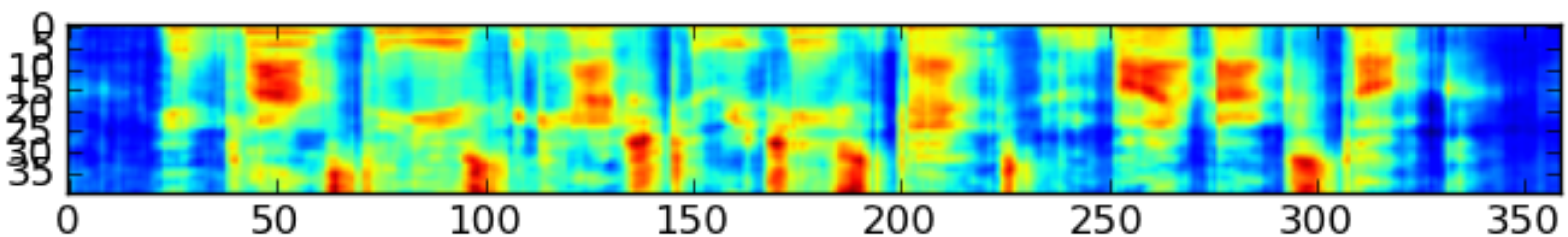
Example of DAE-enhanced speech feature

Reverberant FBANK feature



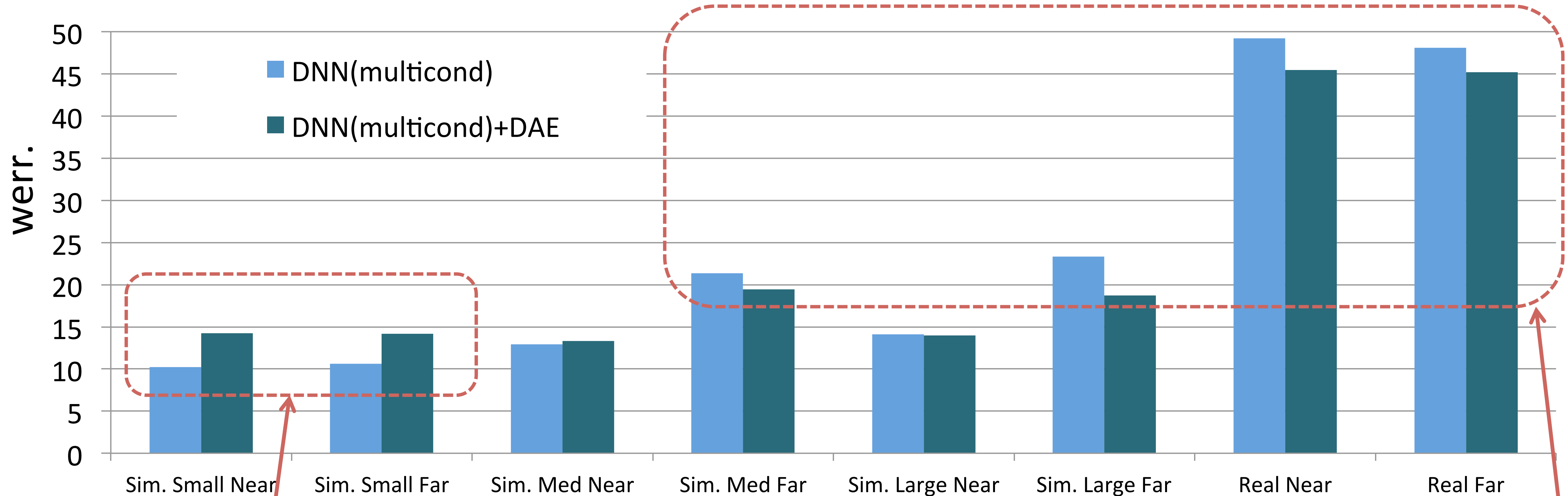
→ time

DAE-enhanced FBANK feature



→ time

Effectiveness of combination of multicond. DNN-HMM and DAE



In less adverse conditions, speech "enhancement" by DAE harmful

In very adverse conditions, significant improvements obtained by **combining DAE with multicond. DNN-HMM**

Conclusion

- **Deep learning effective** for reverberant speech recognition
  - Multi condition training of **DNN-HMMs**
  - Speech feature enhancement by **DAEs**
- Combined DAE and multicond. DNN-HMM achieves larger accuracy improvements **in more adverse reverberant conditions**.
- **Further error reduction by adapting** DNN-HMMs to the DAE-enhanced features