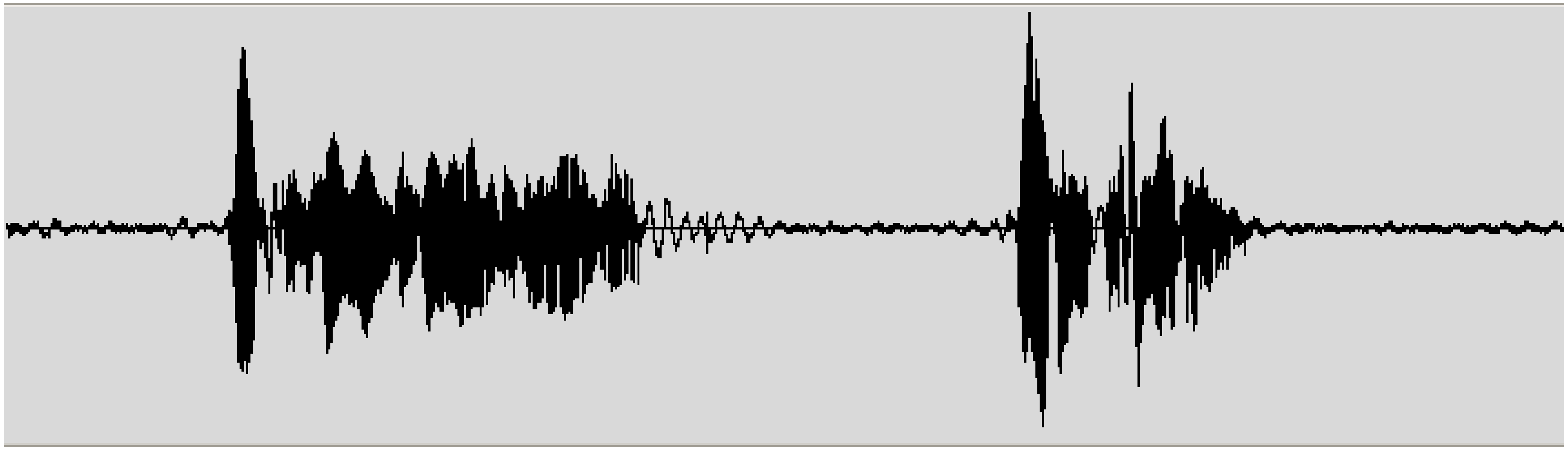
Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

Paul Hensch



Complications

Speaker

Accents
Dialect
Style
Emotion
Coarticulation
Reduction
Pronunciation
Hesitation
...

Environment

Noise
Side talk
Reverberation
...

Device

Head phone
Land phone
Speaker phone
Cell phone
...

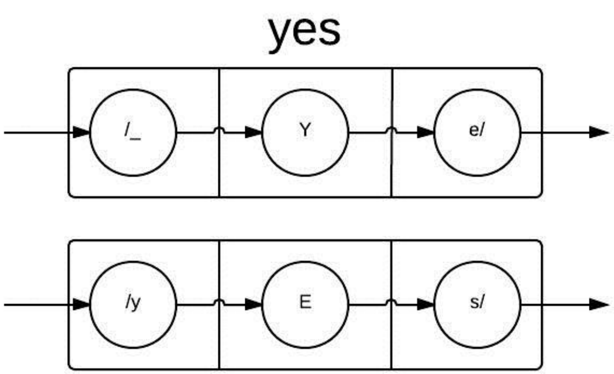
Large-Vocabulary **Speech Recognition**

- **Hidden Markov Models**
  - Find Phones, which matches input
- **Neural networks**
  - Look on frequency changes over time to recognize
- **Hybrid**
  - CD-GMM-HMM
  - CD-DNN-HMM

CD-DNN-HMM

- **Similar classes of sounds, or phones**
  - Diphone
  - Triphone
  - Quinphone

- **<u>Senons</u> are used as the DNN output unit**



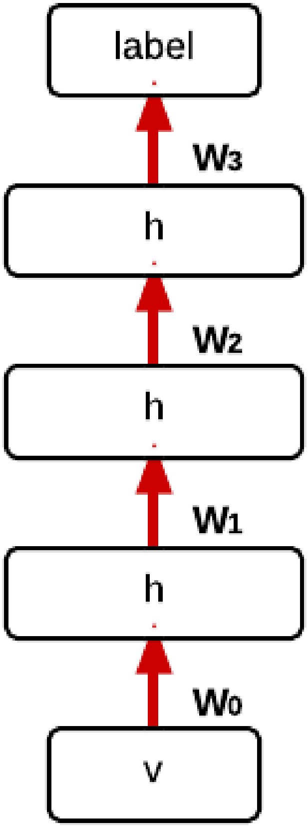
CD-**DNN**-HMM

Deep Belief Network

(not dynamic Bayes net)

2 Stages:

- Pre-training
- Fine-tuning

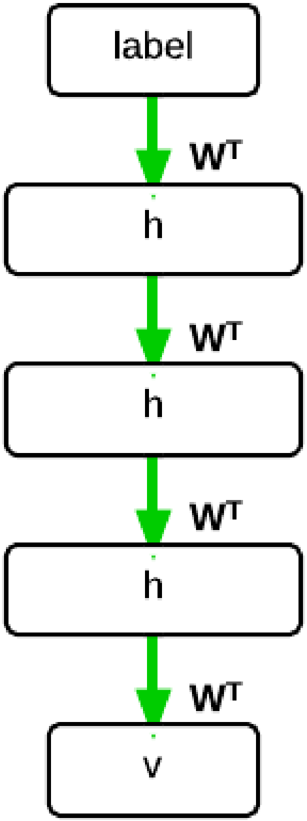


CD-**DNN**-HMM

TECHNISCHE UNIVERSITÄT DARMSTADT

Pre-training advantages:

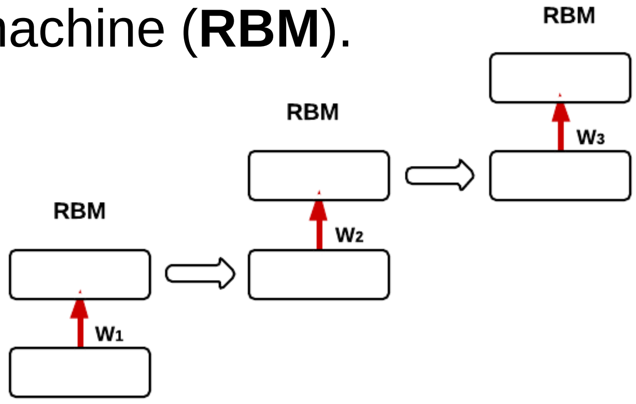
- Often also achieve lower training error
- Sort of data-dependent regularization

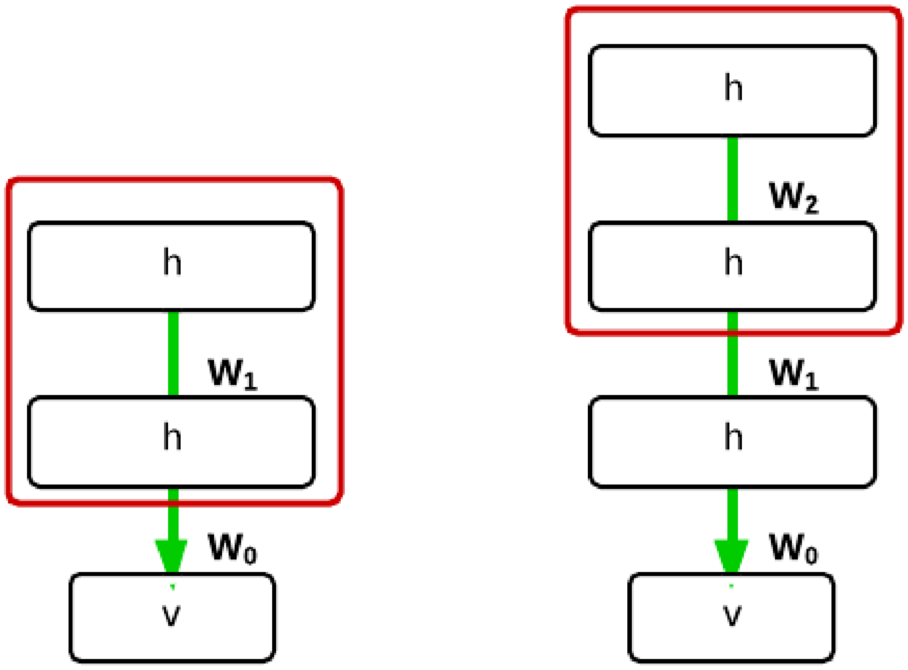


CD-**DNN**-HMM

Pre-training

learning the connection weights in a DBN that is equivalent to training each adjacent pair of layers as an restricted Boltzmann machine (**RBM**).





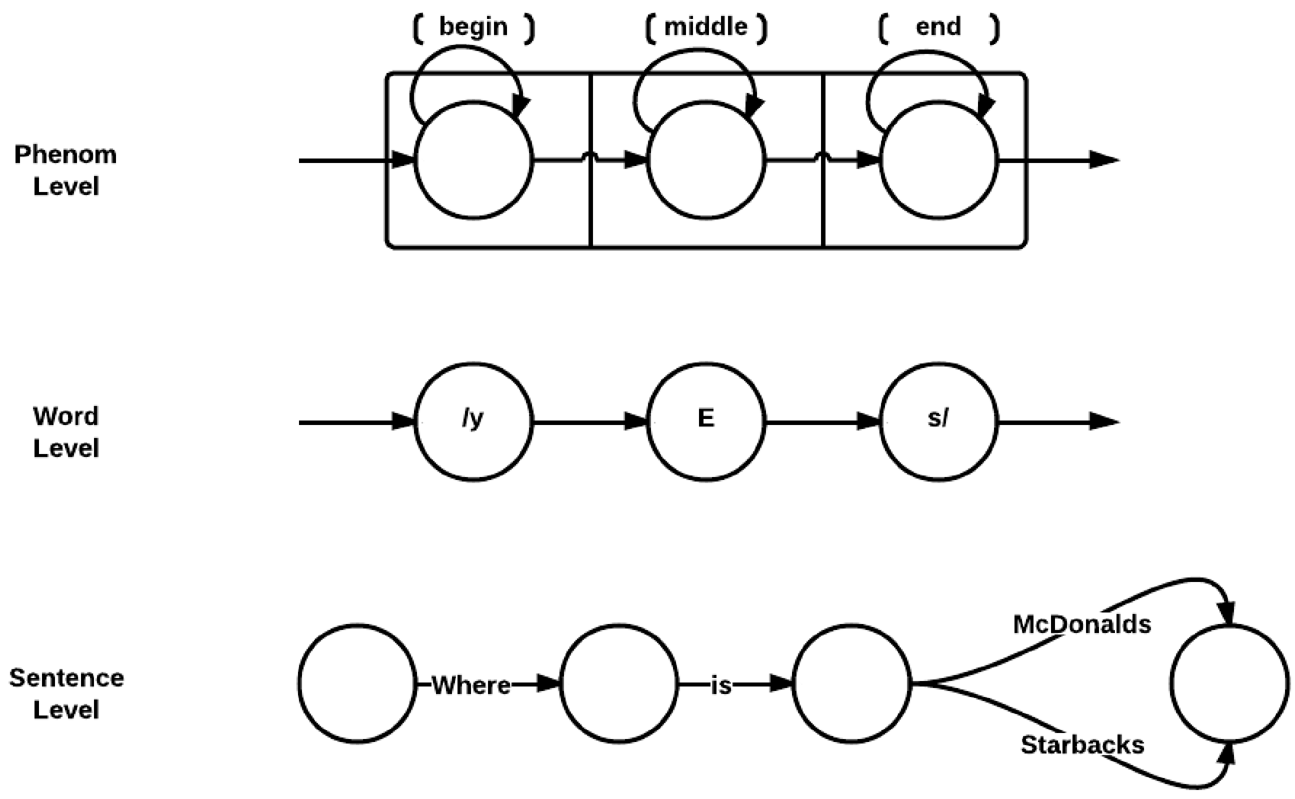
CD-**DNN**-HMM

Last steps of training:

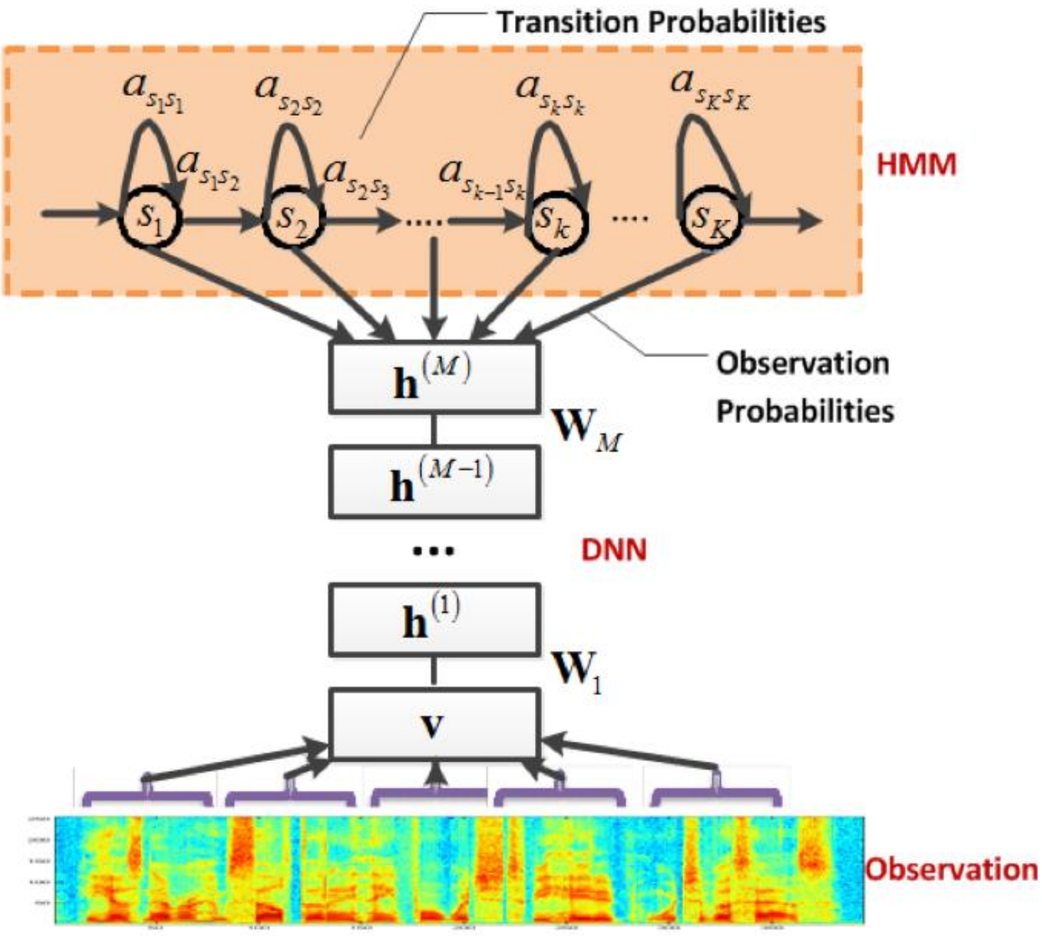
With pre-training complete, add a randomly initialized softmax output layer and use backpropagation to fine-tune all the weights.

CD-DNN-**HMM**

Hidden Markov Models is the dominant technique for LVSR



CD-DNN-HMM



Transition Probabilities
$a_{s_1s_1}$
$a_{s_2s_2}$
$a_{s_ks_k}$
$a_{s_Ks_K}$
$a_{s_1s_2}$
$a_{s_2s_3}$
$a_{s_{k-1}s_k}$
s_1
s_2
s_k
s_K
HMM
Observation Probabilities
$\mathbf{h}^{(M)}$
$\mathbf{W}_M$
$\mathbf{h}^{(M-1)}$
DNN
$\mathbf{h}^{(1)}$
$\mathbf{W}_1$
$\mathbf{v}$
Observation

CD-DNN-HMM

More advantages:

- Implement a CD-DNN-HMM system with only minimal modifications to an existing CD-GMM-HMM system
- Any improvements in modeling units that are incorporated into the CD-GMM-HMM baseline system, such as cross-word triphone models, will be accessible to the DNN

EXPERIMENTAL RESULTS

Business search dataset collected from the Bing mobile voice search application

- Collected under real usage scenarios in 2008
- Sampled at 8 kHz
- Encoded with the GSM codec
- contains all kinds of variations: noise, music, side-speech, accents, sloppy pronunciation …

EXPERIMENTAL RESULTS

Information on the Business Search Dataset

	Hours	Number of Utterances
Training Set	24	32,057
Development Set	6.5	8,777
Test Set	9.5	12,758

- Dataset contains 65 K word unigrams, 3.2 million word bi-grams, and 1.5 million word tri-grams.
- Sentence length is 2.1 tokens

EXPERIMENTAL RESULTS

Compare sentences (sentence accuracy) instead of word accuracy.

G. Zweig and P. Nguyen, “A segmental CRF approach to large vocabulary continuous speech recognition”

Difficulties:

- “Mc-Donalds” “McDonalds”
- “Walmart” “Wal-mart”
- “7-eleven” “7 eleven” “seven-eleven.”

Maximum of 94% accuracy.

EXPERIMENTAL RESULTS

Baseline Systems

- trained clustered cross-word triphone GMM-HMM
- The performance of the best CD-GMM-HMM configuration is summarized in Table

maximum-likelihood

maximum mutual information

minimum phone error

Criterion	Dev Accuracy	Test Accuracy
ML	62.9%	60.4%
MMI	65.1%	62.8%
MPE	65.5%	63.8%

EXPERIMENTAL RESULTS

TECHNISCHE UNIVERSITÄT DARMSTADT

CD-DNN-HMM Results and Analysis

COMPARISON OF CONTEXT-INDEPENDENT MONOPHONE STATE LABELS AND CONTEXT-DEPENDENT TRIPHONE SENONE LABELS

# Hidden Layers	# Hidden Units	Label Type	Dev Accuracy
1	2K	Monophone States	59.3%
1	2K	Triphone Senones	68.1%
3	2K	Monophone States	64.2%
3	2K	Triphone Senones	69.6%

EXPERIMENTAL RESULTS

TECHNISCHE UNIVERSITÄT DARMSTADT

CD-DNN-HMM Results and Analysis

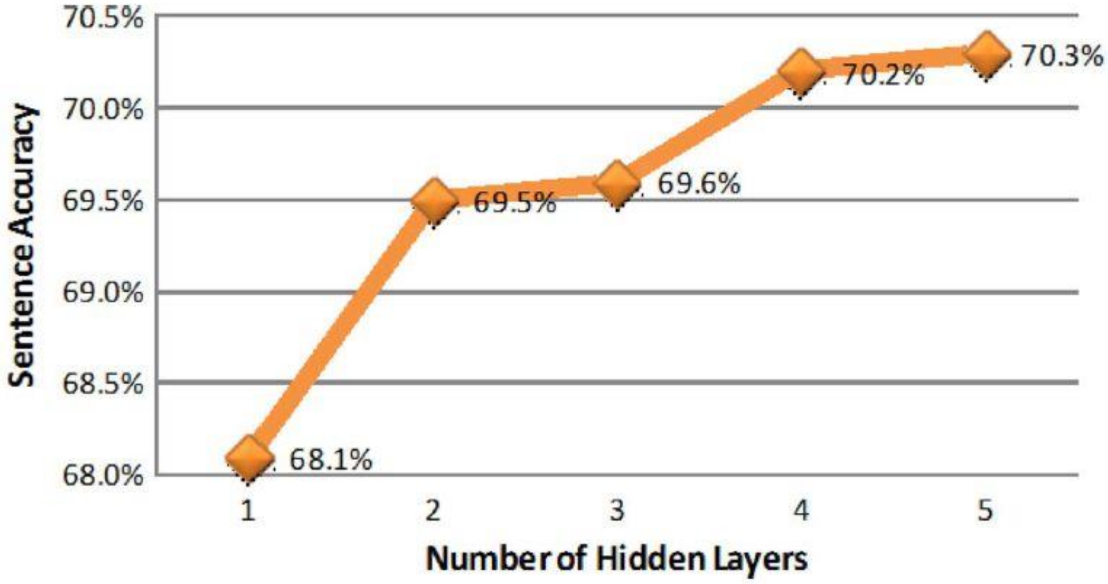
CONTEXT-DEPENDENT MODELS WITH AND WITHOUT PRE-TRAINING

Model Type	# Hidden Layers	# Hidden Units	Dev Accuracy
without pre-training	1	2K	68.0%
without pre-training	2	2K	68.2%
with pre-training	1	2K	68.1%
with pre-training	2	2K	69.5%

EXPERIMENTAL RESULTS

CD-DNN-HMM Results and Analysis

Relationship between the recognition accuracy and the number of layers. Context-dependent models with 2 K hidden units



EXPERIMENTAL RESULTS

Training and Decoding Time

- Trainer written in Python
- Carried out on Dell Precision T3500 workstation
  - Quad core computer
  - CPU clock speed of 2.66 GHz, 8 MB of L3 CPU cache
  - 12 GB of 1066 MHz DDR3 SDRAM.

  - NVIDIA Tesla C1060 (GPGPU), which contains 4 GB of GDDR3 RAM and 240 processing cores

EXPERIMENTAL RESULTS

TECHNISCHE UNIVERSITÄT DARMSTADT

Training and Decoding Time

SUMMARY OF TRAINING TIME USING 24 HOURS OF TRAINING DATA AND 2 K HIDDEN UNITS PER LAYER

Type	# of Layers	Time Per Epoch	# of Epochs
Pre-train	1	0.2 h	50
Pre-train	2	0.5 h	20
Pre-train	3	0.6 h	20
Pre-train	4	0.7 h	20
Pre-train	5	0.8 h	20
Fine-tune	4	1.2 h	12
Fine-tune	5	1.4 h	12

EXPERIMENTAL RESULTS

TECHNISCHE UNIVERSITÄT DARMSTADT

Training and Decoding Time

Processing Unit	# of Layers	DNN Time Per Frame	Search Time Per Frame	Real-time Factor
CPU	4	4.3 ms	1.5 ms	0.58
GPU	4	0.16 ms	1.5 ms	0.17
CPU	5	5.2 ms	1.5 ms	0.67
GPU	5	0.20 ms	1.5 ms	0.17

- five-layer CD-DNN-HMM, pre-training takes

0.2 x 50 + 0.5 x 20 + 0.6 x 20 + 0.7 x 20 + 0.8 x 20 = 62 hours

- Fine-tuning takes

1.4 x 12 = 16.8 hours

EXPERIMENTAL RESULTS

TECHNISCHE UNIVERSITÄT DARMSTADT

Training and Decoding Time

Observations:

The bottleneck in the training process is the mini-batch stochastic gradient descend (SGD) algorithm used to train the DNNs

It is extrapolated that using similar technics described in this paper, it should be possible to train an effective CD-DNN-HMM system that exploits **2000 hours** of training data in about **50 days**

CONCLUSION AND FUTURE WORK

- CD-DNN-HMM is more expensive than GMM
- CD-DNN-HMM performs better than GMM

- Finding new ways to parallelize training
- Finding highly effective speaker and environment adaptation algorithms for DNN-HMMs
- The training in this study used the embedded Viterbi algorithm, which is not optimal (MFCC)

Thank your for your attention

TECHNISCHE UNIVERSITÄT DARMSTADT

Questions?..