(Deep) Neural Networks

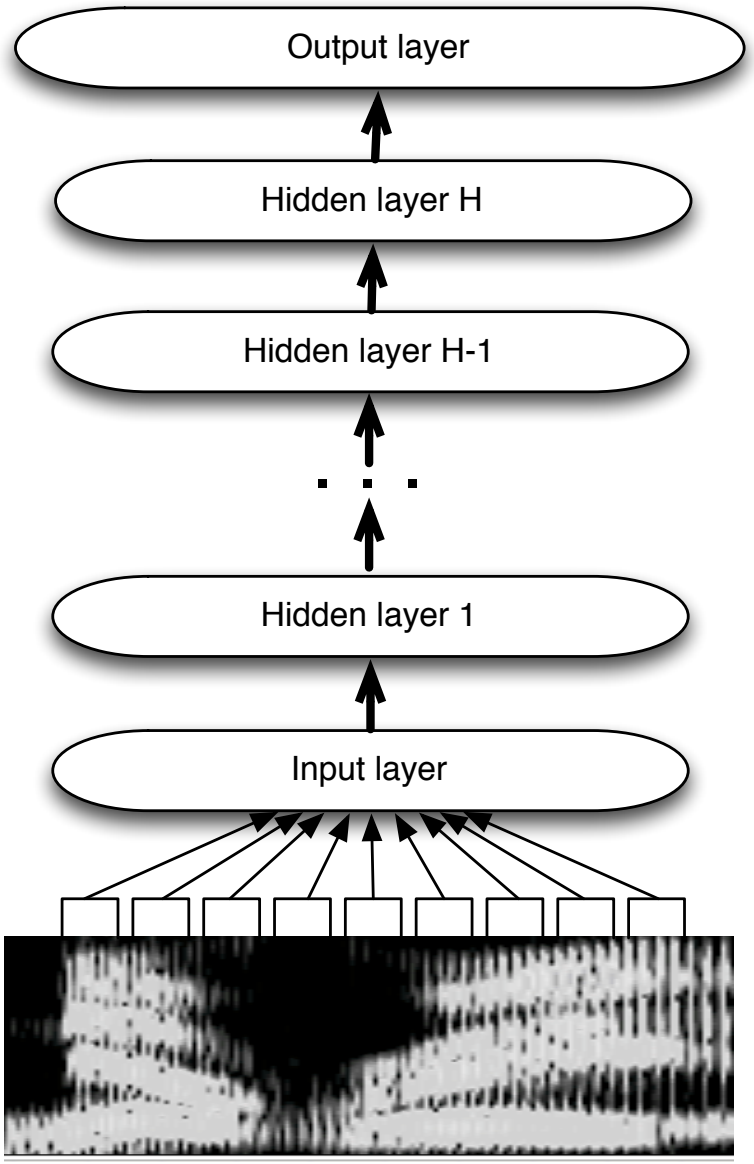
Steve Renals

Automatic Speech Recognition— ASR Lecture 11
26 February 2015

Neural networks for speech recognition

- Introduction to Neural Networks
- Training feed-forward networks
- **Hybrid neural network / HMM acoustic models**
- **Neural network features – Tandem, posteriorgrams**
- **Deep neural network acoustic models**
- **Neural network language models**

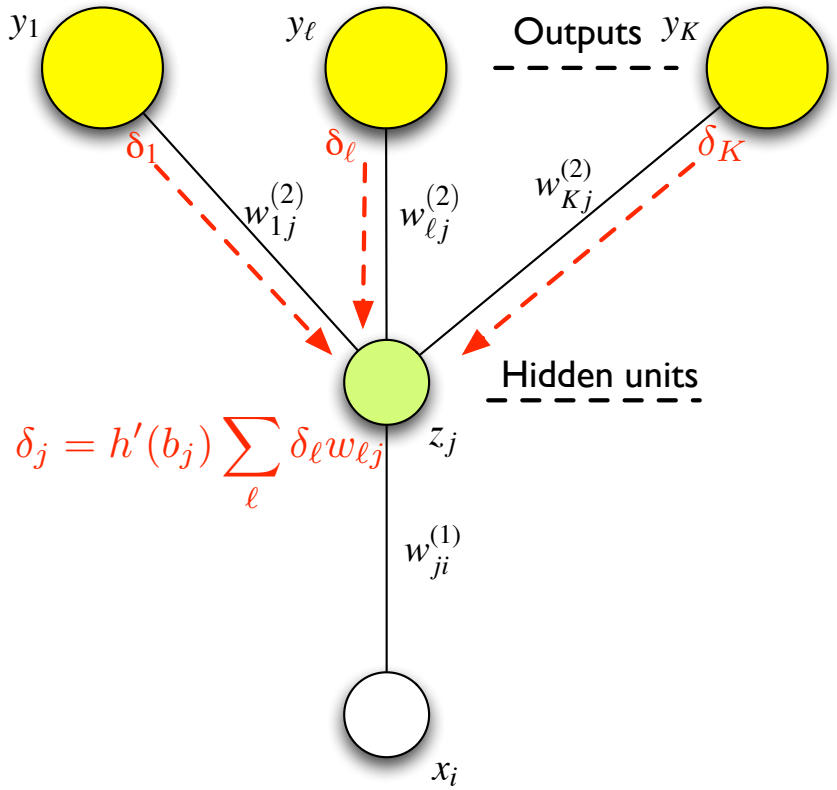
Neural network acoustic models



- Input layer takes several consecutive frames of acoustic features
- Output layer corresponds to classes (e.g. phones, HMM states)
- Multiple non-linear hidden layers between input and output
- Neural networks also called multi-layer perceptrons

Neural network training

- Train multiple layers of hidden units nested nonlinear functions
  - Powerful feature detectors
  - Posterior probability estimation
  - Theorem: any function can be approximated with a single hidden layer



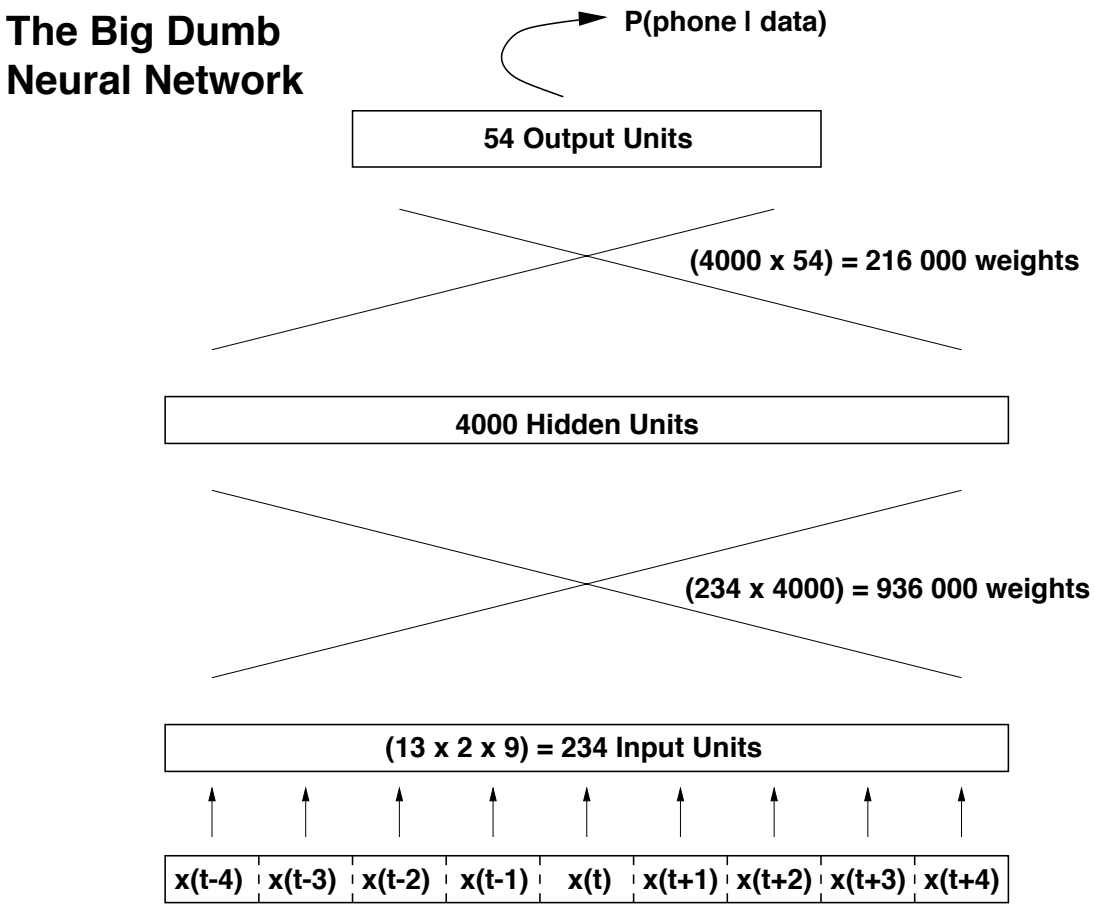
Using neural networks for acoustic modelling

- **Hybrid** NN/HMM systems
  - Basic idea: in an HMM, replace the GMMs used to estimate output pdfs with the outputs of neural networks
  - Transform NN posterior probability estimates to *scaled likelihoods* by dividing by the relative frequencies in the training data of each class

$$P(\mathbf{x}_t \mid C_k) \propto \frac{P(C_k \mid \mathbf{x}_t)}{P_{\text{train}}(C_k)} = \frac{y_k}{P_{\text{train}}(C_k)}$$

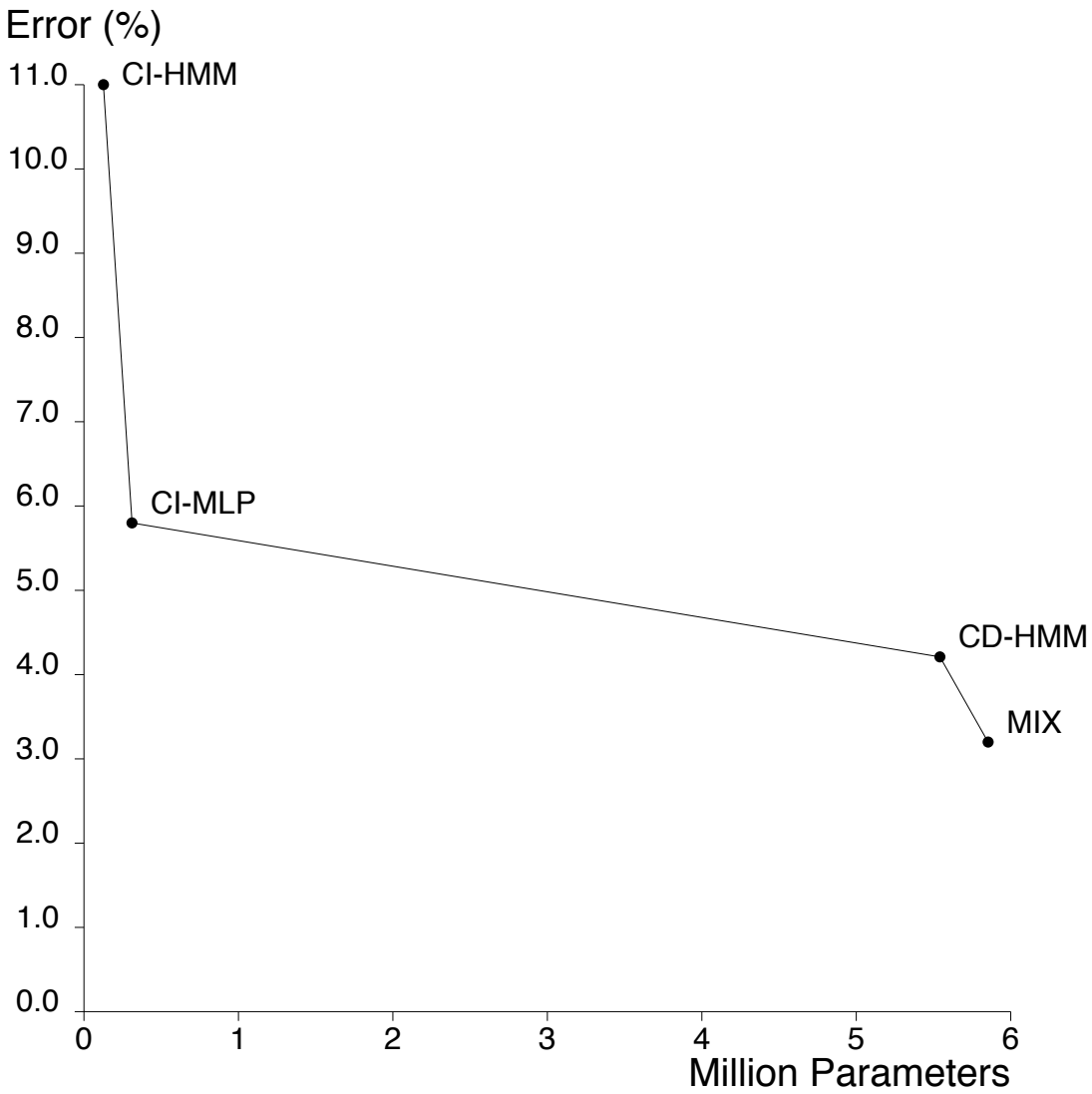
  - NN outputs correspond to phone classes or HMM states
- **Tandem** features
  - Use NN probability estimates as an additional input feature stream in an HMM/GMM system (posteriograms / bottleneck features)

Monophone HMM/NN hybrid system (1990s) (1)



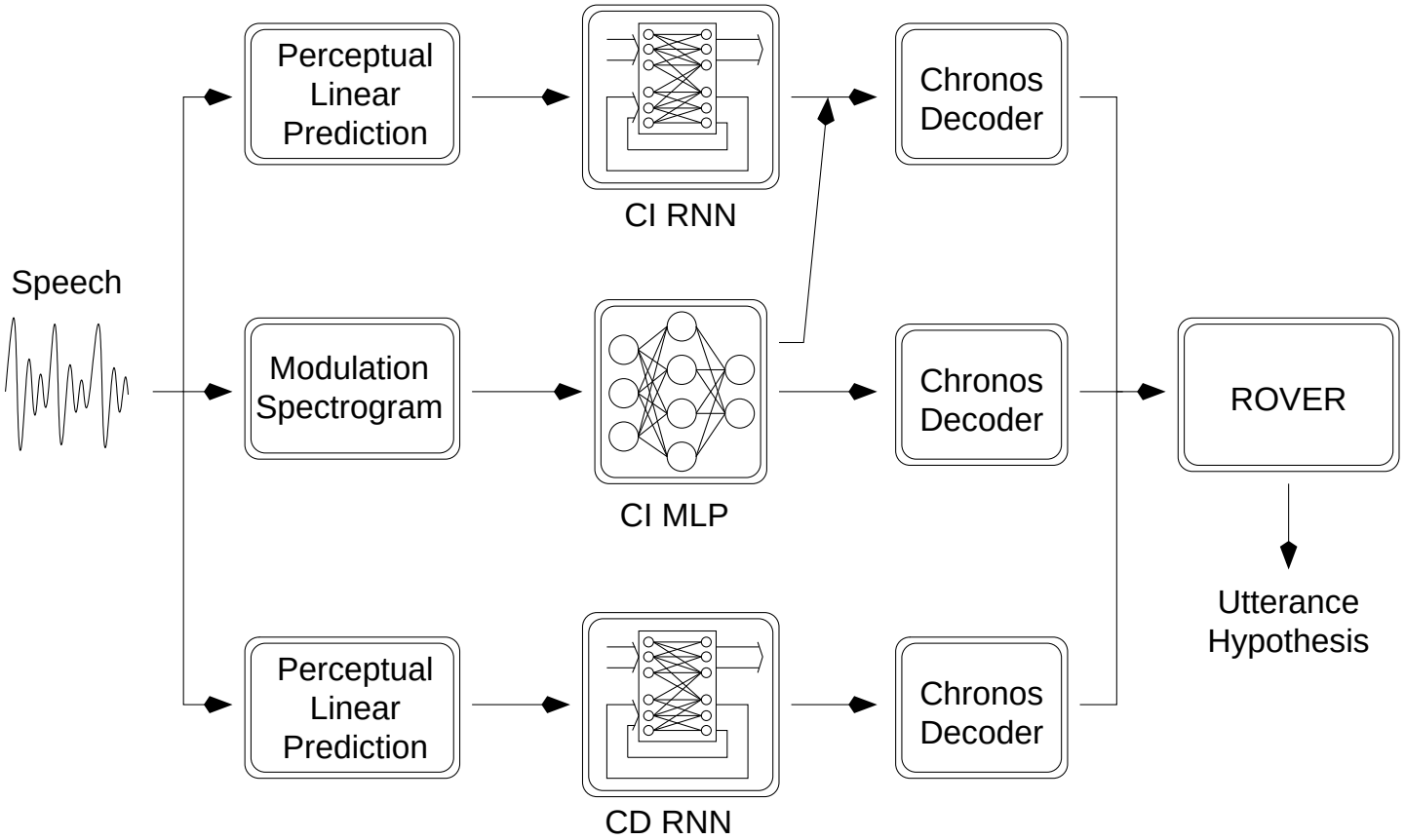
- Similar performance to context-dependent HMM/GMM systems on WSJ
- More errors on more complex tasks (broadcast news, conversational telephone speech)

Monophone HMM/NN hybrid system (1990s) (2)



Renals, Morgan, Cohen & Franco, ICASSP 1992

Monophone HMM/NN hybrid system (1990s) (3)



- Broadcast news transcription (1998) – 20.8% WER
- (best GMM-based system, 13.5%)
- Cook et al, DARPA, 1999

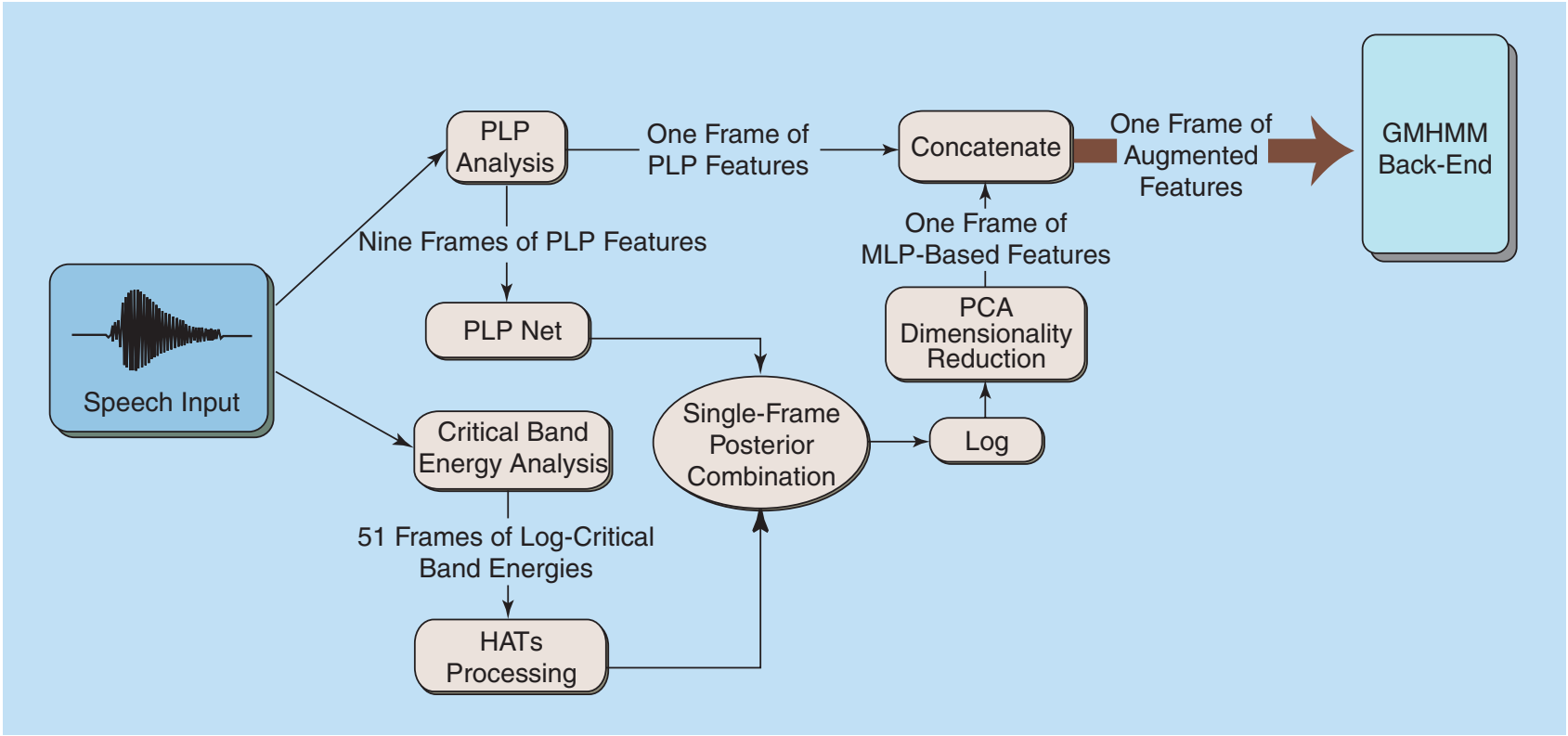
HMM/NN vs HMM/GMM

- Advantages of NN:
  - Incorporate multiple frames of data at input
  - More flexible than GMMs (i.e. not made of (nearly) local components) — GMMs inefficient for non-linear class boundaries
- Disadvantages of NN:
  - Context-independent (monophone) models
  - Weak speaker adaptation algorithms
  - Computationally expensive - more difficult to parallelise than GMM systems
  - Systems less complex than GMMSs (fewer parameters)
    - RNN – $<$ 100k parameters
    - MLP – $\sim$ 1M parameters
- Reading: Morgan and Bourlard (1995)

Tandem features (posteriorgrams)

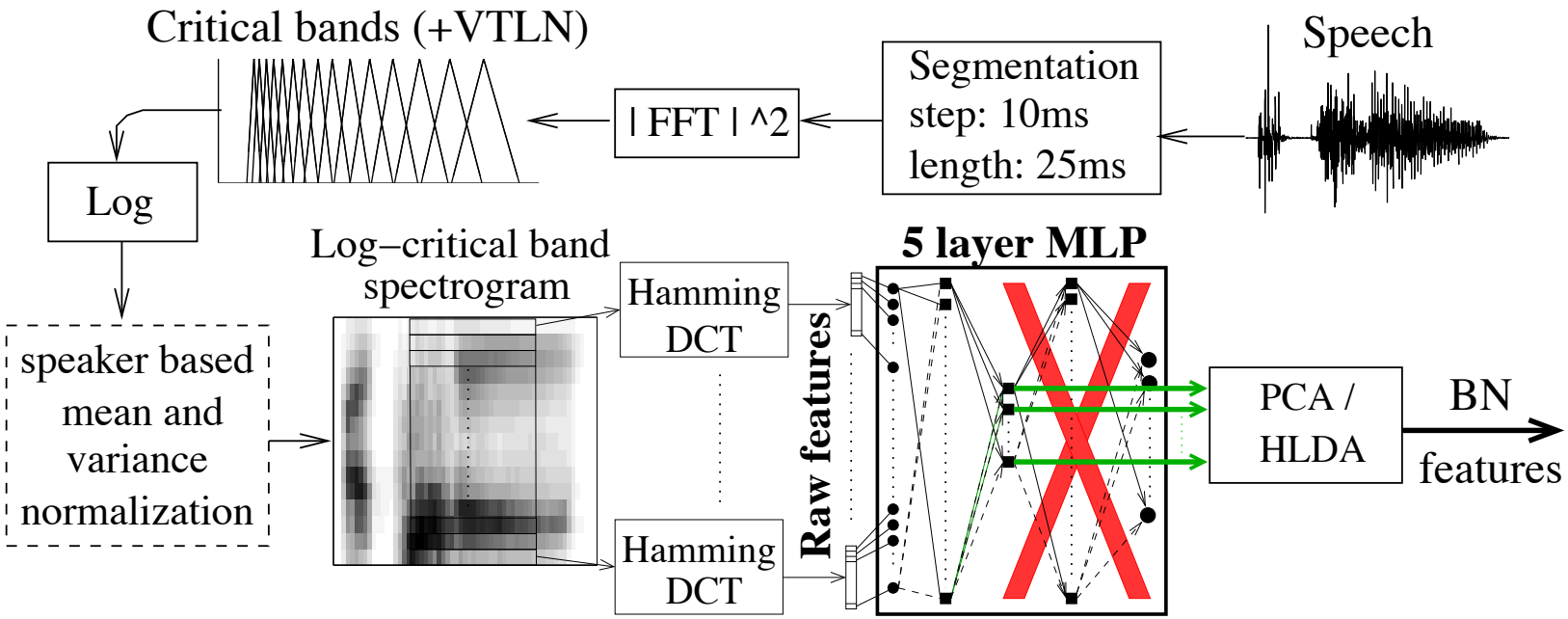
- Use NN probability estimates as an additional input feature stream in an HMM/GMM system —- (*Tandem* features (i.e. NN + acoustics), posteriorgrams)
- Advantages of tandem features
  - can be estimated using a large amount of temporal context (eg up to ± 25 frames)
  - encode phone discrimination information
  - only weakly correlated with PLP or MFCC features
- Tandem features: reduce dimensionality of NN outputs using PCA, then concatenate with acoustic features (e.g. MFCCs)

Tandem features



Morgan et al (2005)

Bottleneck features

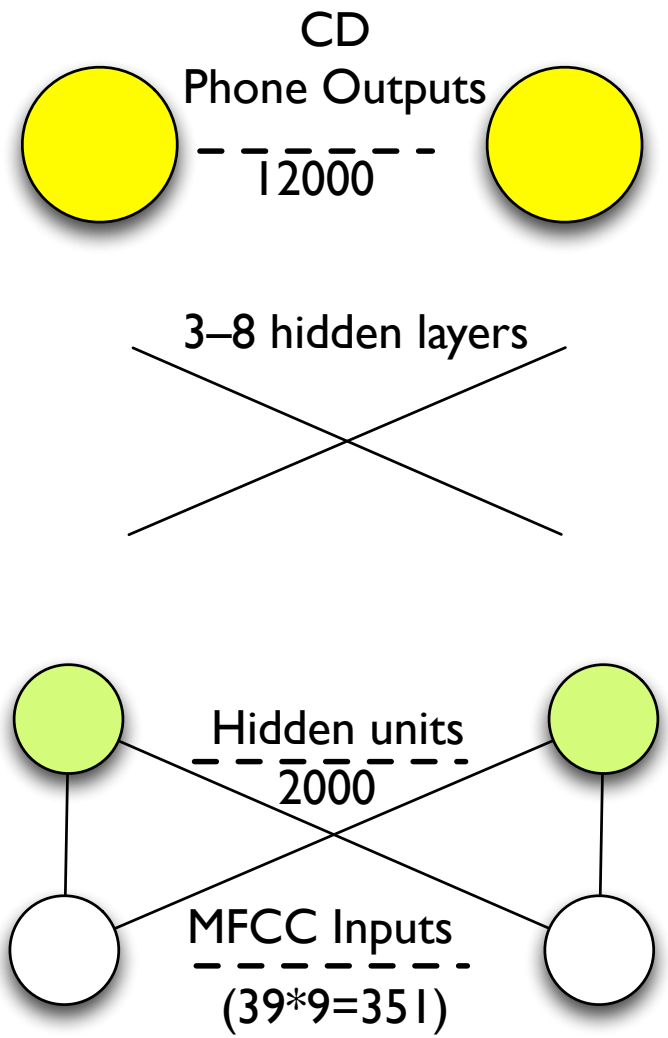


- Grezl and Fousek (2008)
- Use a "bottleneck" hidden layer to provide features for a HMM/GMM system

Interim summary

- Hybrid HMM/NN system can be powerful acoustic models — but unadapted monophone NN-based system have worse accuracies than state-of-the-art GMM systems on complex tasks
- Using NNs to provide tandem features (posteriorgrams, bottleneck features) for GMMs can significantly reduce word error rates (10-15%)

Deep neural networks (DNNs) — Hybrid system



DNNs — what's new?

- Training multi-hidden layers directly with gradient descent is difficult — sensitive to initialisation, gradients can be very small after propagating back through several layers.
  Unsupervised pretraining (see Hinton et al 2012)
  - Train a stacked restricted Boltzmann machine generative model (unsupervised), then finetune with backprop
  - Contrastive divergence training

  Layer-by-layer training
  - Successively train deeper networks, each time replacing output layer with hidden layer and new output layer
- Many hidden layers
  - GPUs provide the computational power
- Wide output layer (context dependent phone classes)
  - GPUs provide the computational power

Example: hybrid HMM/DNN phone recognition (TIMIT)

- Train a 'baseline' three state monophone HMM/GMM system (61 phones, 3 state HMMs) and Viterbi align to provide DNN training targets (time state alignment)
- The HMM/DNN system uses the same set of states as the HMM/GMM system — DNN has 183 (61*3) outputs
- Hidden layers — many experiments, exact sizes not highly critical
  - 3–8 hidden layers
  - 1024–3072 units per hidden layer
- Multiple hidden layers always work better than one hidden layer
- Pretraining always results in lower error rates
- Best systems have lower phone error rate than best HMM/GMM systems (using state-of-the-art techniques such as discriminative training, speaker adaptive training)

Example: hybrid HMM/DNN large vocabulary conversational speech recognition (Switchboard)

- Recognition of American English conversational telephone speech (Switchboard)
- Baseline context-dependent HMM/GMM system
  - 9,304 tied states
  - Discriminatively trained (BMMI — similar to MPE)
  - 39-dimension PLP (+ derivatives) features
  - Trained on 309 hours of speech
- Hybrid HMM/DNN system
  - Context-dependent — 9304 output units obtained from Viterbi alignment of HMM/GMM system
  - 7 hidden layers, 2048 units per layer
- DNN-based system results in significant word error rate reduction compared with GMM-based system
- Can also use DNNs in tandem configuration

Reading

- N Morgan and H Bourlard (May 1995). Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach, *IEEE Signal Processing Magazine*, **12**(3), 24–42.
- N Morgan et al (Sep 2005). Pushing the envelope — aside, *IEEE Signal Processing Magazine*, **22**(5), 81–88.
- F Grezl and P Fousek (2008). Optimizing bottleneck features for LVCSR, Proc ICASSP–2008.
- G Hinton et al (Nov 2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, **29**(6), 82–97.