The JanusRTk Switchboard/CallHome System

Hybrid Acoustic Modeling

Jürgen Fritsch
Michael Finke

Interactive Systems Laboratories
Carnegie Mellon University
University of Karlsruhe

May 15, 1997

Carnegie Mellon

Interactive Systems Laboratories

Overview

Hybrid NN/HMM acoustic modeling:

- Context-dependent multi-state HMMs
- Context classes adopted from HMM polyphone cluster tree
- Hierarchical modular classification scheme
- Using tree structured hierarchy of neural networks (HNN)

Stream Interpolation:

- Interpolation of multiple (heterogenous) acoustic modeling streams
- Linear opinion pooling
- Mixture of experts model – adaptive gating networks

Hybrid NN/HMM Acoustic Modeling

Approach:

- Focus on modularity through factorization of posterior state emission probabilities
- Use neural networks to estimate these posteriors
- Convert to scaled likelihoods according to Bayes' rule

Why hybrid NN/HMM models?

- Improved discrimination (MAP instead of ML)
- Compact acoustic models, faster decoding
- Diversification of acoustic modeling streams
- Use posteriors for phone deactivation pruning in decoder

Factorization of Posteriors

Consider n-state left-right HMM topology, modeling

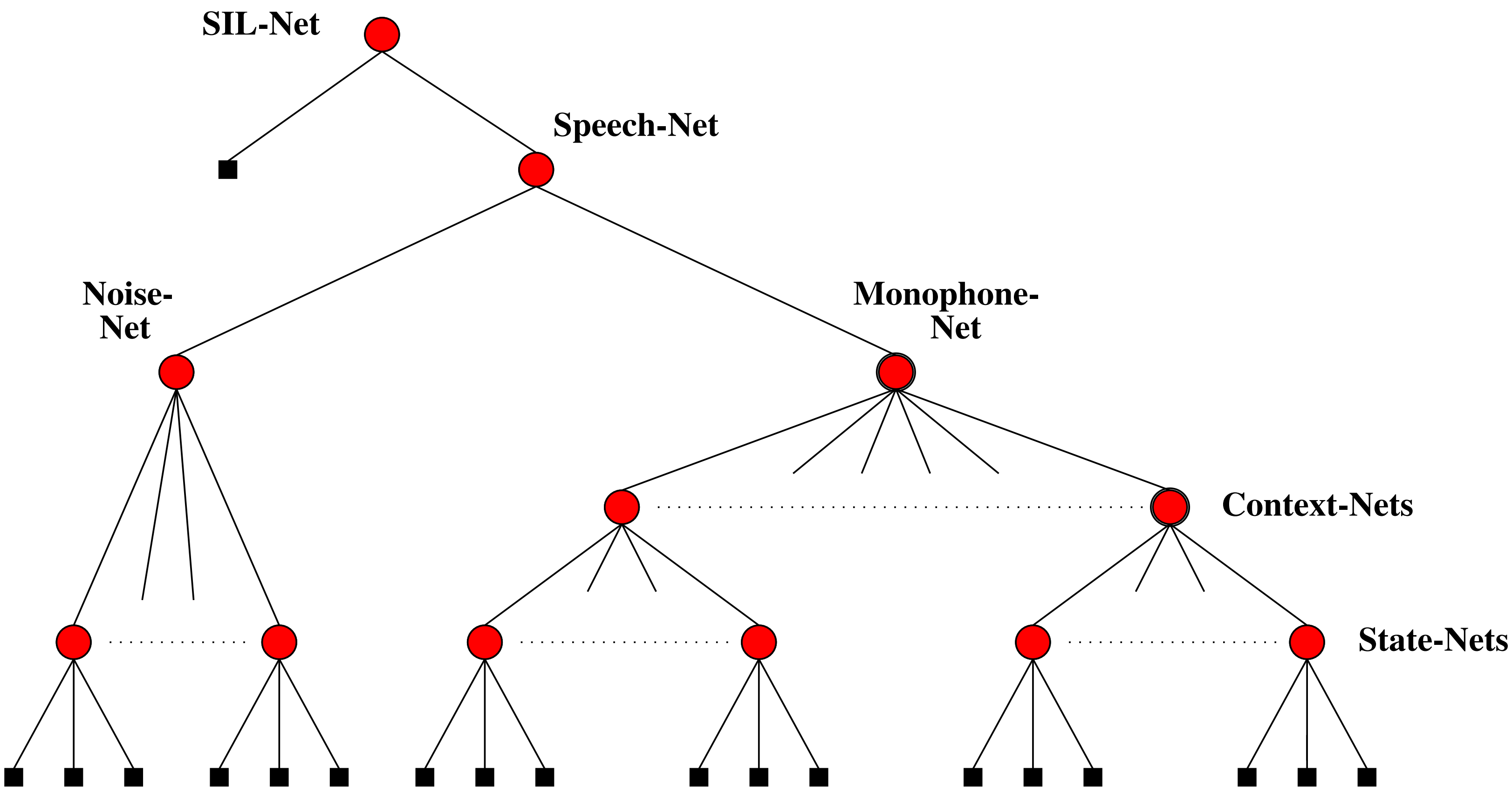
- monophones ω_i
- in contexts c_{ij}
- with states s_{ijk}

Requires estimation of conditional likelihoods:

$$
\begin{aligned}
p(\mathbf{x}|\omega_i, c_{ij}, s_{ijk}) &= \frac{p(\omega_i, c_{ij}, s_{ijk}|\mathbf{x})}{P(\omega_i, c_{ij}, s_{ijk})} \quad p(\mathbf{x}) \\
&= \frac{p(c_{ij}, s_{ijk}|\omega_i, \mathbf{x})}{P(c_{ij}, s_{ijk}|\omega_i)} \quad \frac{p(\omega_i|\mathbf{x})}{P(\omega_i)} \quad p(\mathbf{x}) \\
&= \frac{p(s_{ijk}|\omega_i, c_{ij}, \mathbf{x})}{P(s_{ijk}|\omega_i, c_{ij})} \quad \frac{p(c_{ij}|\omega_i, \mathbf{x})}{P(c_{ij}|\omega_i)} \quad \frac{p(\omega_i|\mathbf{x})}{P(\omega_i)} \quad p(\mathbf{x})
\end{aligned}
$$

state nets context nets monophone net

Hierarchy of Neural Networks (HNN)



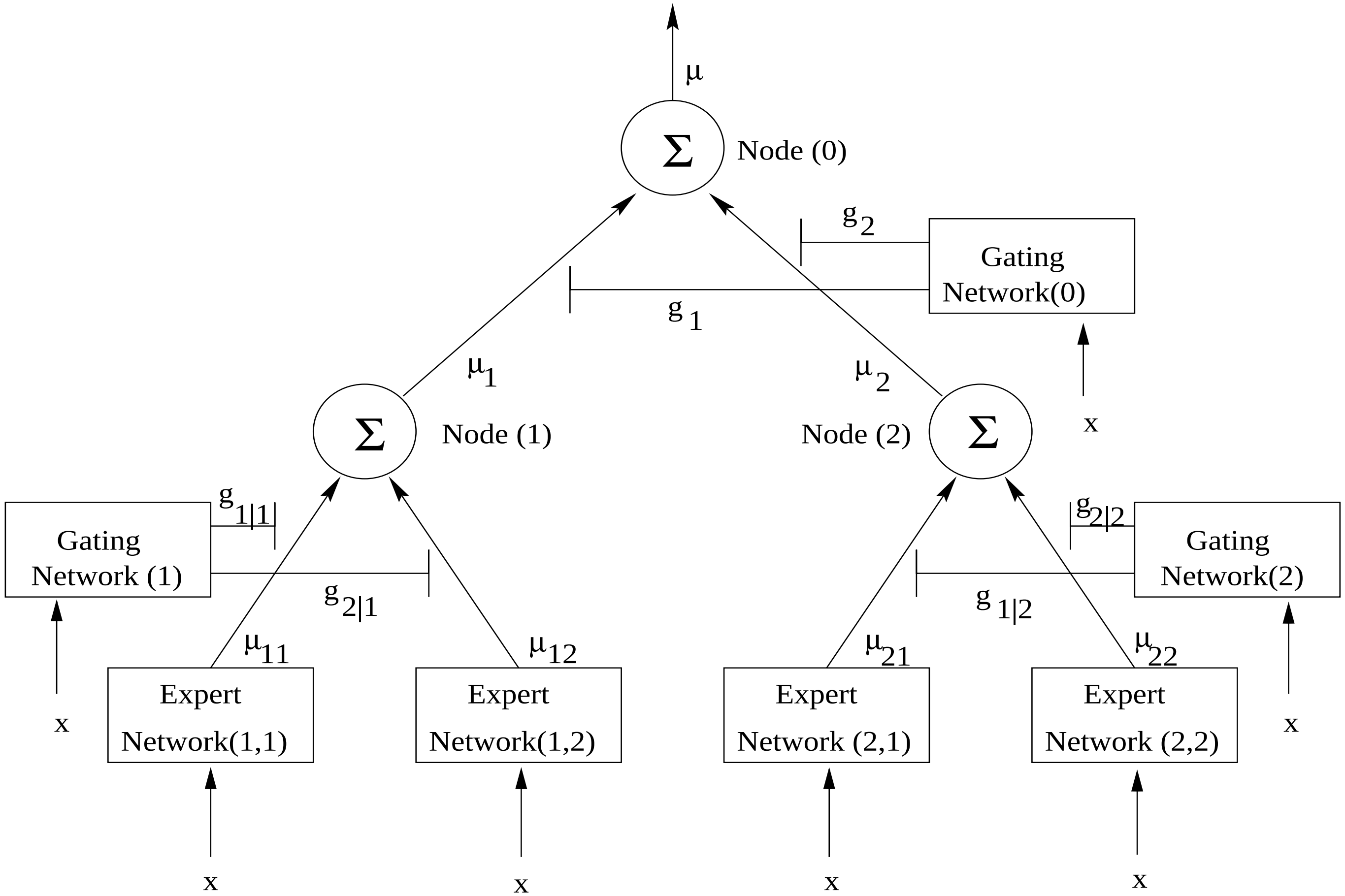
Neural Networks for HNN

Applied variety of NN models:

- **Generalized Linear Models (GLIM)**
  - state discriminators (3 state HMM's)
- **Multi-layer perceptrons (MLP)**
  - SIL-net, Speech-net, Noise-net (32 HU each)
  - Context-nets (128 HU each)
- **Hierarchical Mixtures of Experts (HME)**
  - depth 2, branching factor 4 HME
  - 2-layer MLP (128 HU) nodes (126k parameters)
  - to estimate monophone posteriors

All NN output layers need to be parameterized as multinomial probability models using softmax non-linearity

Hierarchical Mixtures of Experts (HME)



Training Issues

Hierarchy of neural networks (HNN):

- Order of factorization interchangeable
- Current topology reasonable, but design guided by heuristics
- Future Work will concentrate on agglomerative clustering of acoustic models $\rightarrow$ automatically derive HNN

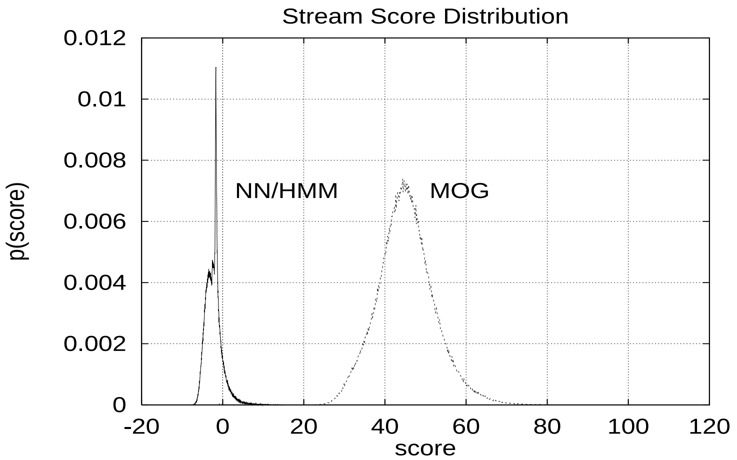
Training the neural networks in the HNN:

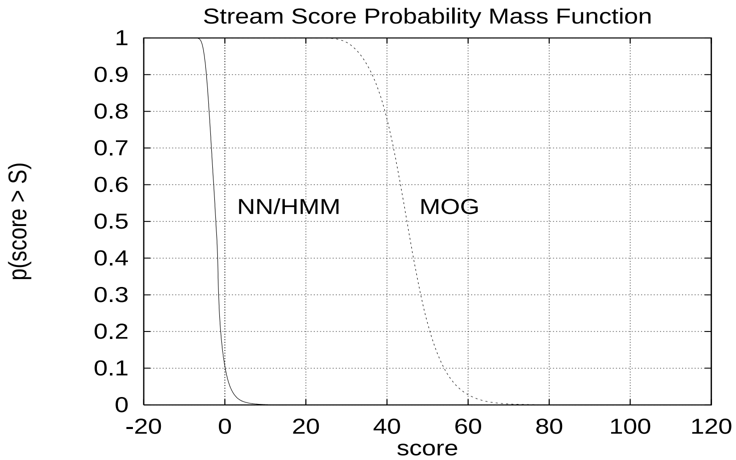
- training alignments using baseline mixture of Gaussians system
- stochastic/on-line gradient descent w/ momentum term
- 1-2 iterations on 161 hours SWB take 7-10 days
  on 2 UltraSparc 167Mhz
- training can easily be distributed/parallelized
  since individual networks in the HNN are trained independently

Stream Interpolation I

Stream score normalization:

- use histogram based approximations of pmf $p(s_i > S)$
- to map log-likelihoods to the range [0,1]
- acts as a probability model for an acoustic 'expert'
- allows to apply mixtures of experts paradigm





Stream Interpolation II

Interpolation techniques for normalized scores:

- **static linear interpolation:**

  $p(\mathbf{x}|\Theta) = \sum_i \gamma_i p_i(\mathbf{x}|\Theta_i) \quad with \quad \sum_i \gamma_i = 1$

- **adaptive gating / mixture of experts:**

  $p(\mathbf{x}|\Theta, v) = \sum_i g_i(\mathbf{x}|v_i) p_i(\mathbf{x}|\Theta_i) \quad with \quad \sum_i g_i(\mathbf{x}|v_i) = 1$

  - multinomial probability model for $g_i(\mathbf{x}|v_i)$'s using softmax NN
  - use cluster tree to assign acoustic models to multiple gating networks

- **confidence score based interpolation:**
  - assuming posterior word confidence measures c_i
  - compute posterior stream weights

$$\gamma_i = \alpha_i c_i / \sum_j \alpha_j c_j \quad \text{with priors } \alpha_i$$

Interactive Systems Laboratories

Results - Hybrid Acoustic Models

1997 devtest set (30 secs):

Hybrid NN/HMM	SWB WE	CH WE	SWB+CH WE
context-independent hybrid HME/HMM	56.5%	N/A	N/A
context-dependent modeling using polyphone cluster tree with 10k classes	42.1%	54.4%	47.1%
learning rate adaptation	40.6%	53.1%	45.9%
smoothing context posteriors	39.4%	51.7%	44.6%
using larger (GLIM $\rightarrow$ MLP) context networks	37.3%	49.3%	42.8%

Mixture of Gaussians	SWB WE	CH WE	SWB+CH WE
baseline system (w/o MLLR) used to align training data for hybrid NN/HMM	34.1%	47.3%	40.2%
evaluation system (MLLR,FTA,enhanced pronunciation modeling, multiple interpolated LMs)	27.1%	36.7%	31.7%

Results – Interpolated Acoustic Models

1997 devtest set:

	SWB WE	CH WE	SWB+CH WE	Relative Improvement
baseline mixture of Gaussians	34.1%	47.3%	40.2%	—
hybrid NN/HMM	37.3%	49.3%	42.8%	—
static linear interpolation	33.0%	45.7%	38.9%	3.2%
mixture of experts gating	33.1%	45.5%	38.8%	3.5%
CM based interpolation	33.9%	47.0%	40.0%	0.5%

1997 evalaluation set:

	SWB WE	CH WE	SWB+CH WE	Relative Improvement
Official Eval System	35.1%	54.4%	45.1%	—
Interpolated Hybrid+HMM	34.9%	53.7%	44.6%	1.1%