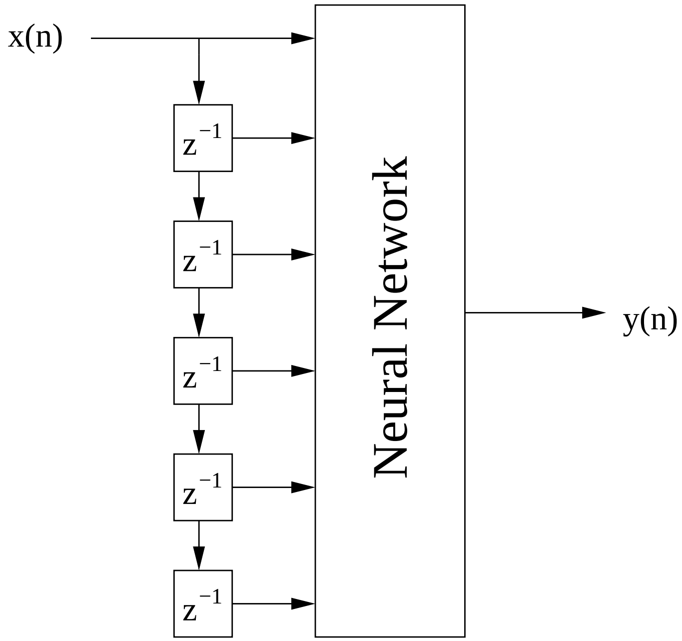
Time-Delay Neural Networks and NN/HMM Hybrids: *A Family of Connectionst Continuous-Speech Recognition Systems*

(Jürgen Fritsch, Hermann Hild, Uwe Meier and Alex Waibel)

Tom Bäckström
Laboratory of Acoustics and Audio Signal Processing
Helsinki University of Technology
http://www.acoustics.hut.fi/~tbackstr
mailto:tom.backstrom@hut.fi

21.2.2002

Introduction



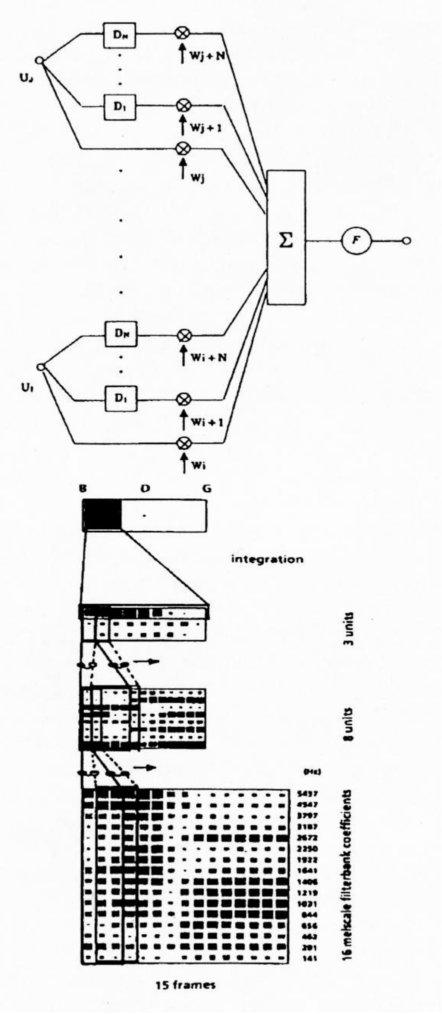
Methods:

- Time-Delay Neural Networks (TDNN)
- Multi-state Time-Delay Neural Networks (MS-TDNN)
- Neural Networks with Hidden Markov Model (NN/HMM)

Applications:

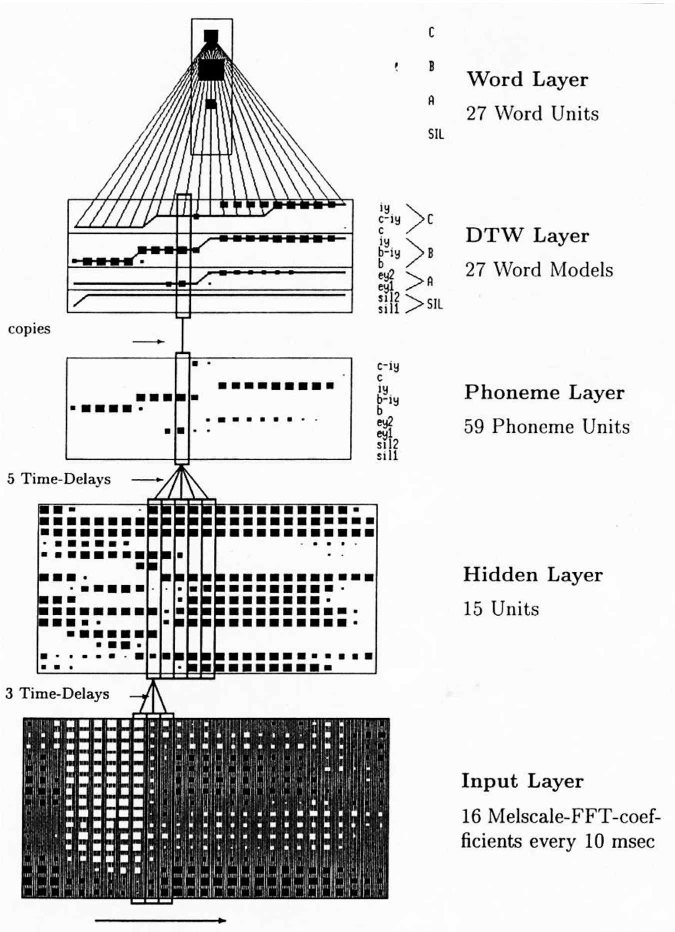
- Continuous spelling recognition
- Lipreading (audio-visual hybrids)
- Large-vocabulary conversation over telephone

The first Time-Delay NN's



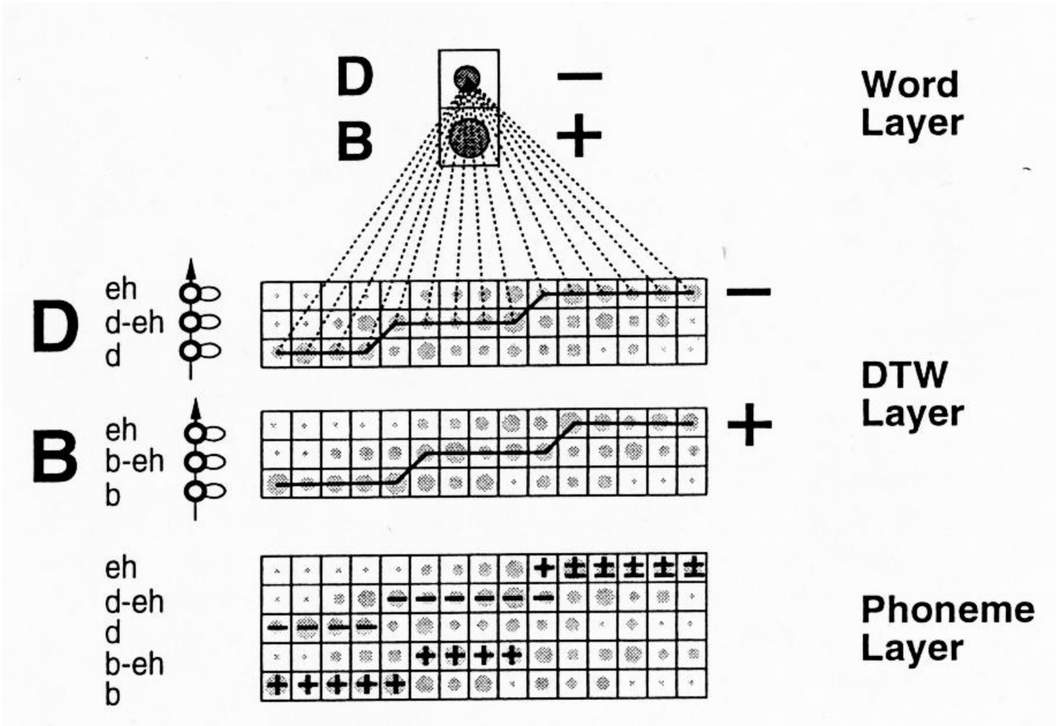
- Developed by Weibel and Lang in 1987.
- Calculate phoneme scores directly from speech segments.
- Feed time-delayed segments into NN with equal weights.
- Trained with phonemes /b/, /d/ and /g/.
- Shift-invariant model, with not too large set of parameters.

Multi-state TDNN



- Extension of TDNN to phoneme sequences.

- Five layers; input, hidden, phoneme, dynamic time warping (DTW) and word layer.
- Phonemes are trained frame-by-frame using the three first layers (input, hidden and phoneme) with standard back-propagation.



- In word-level training an alignment path is searched which maximise the phoneme sum scores.
- Path searching similarly as in HMM.
- Phonemes on the correct path receive positive training and on other paths negative training.

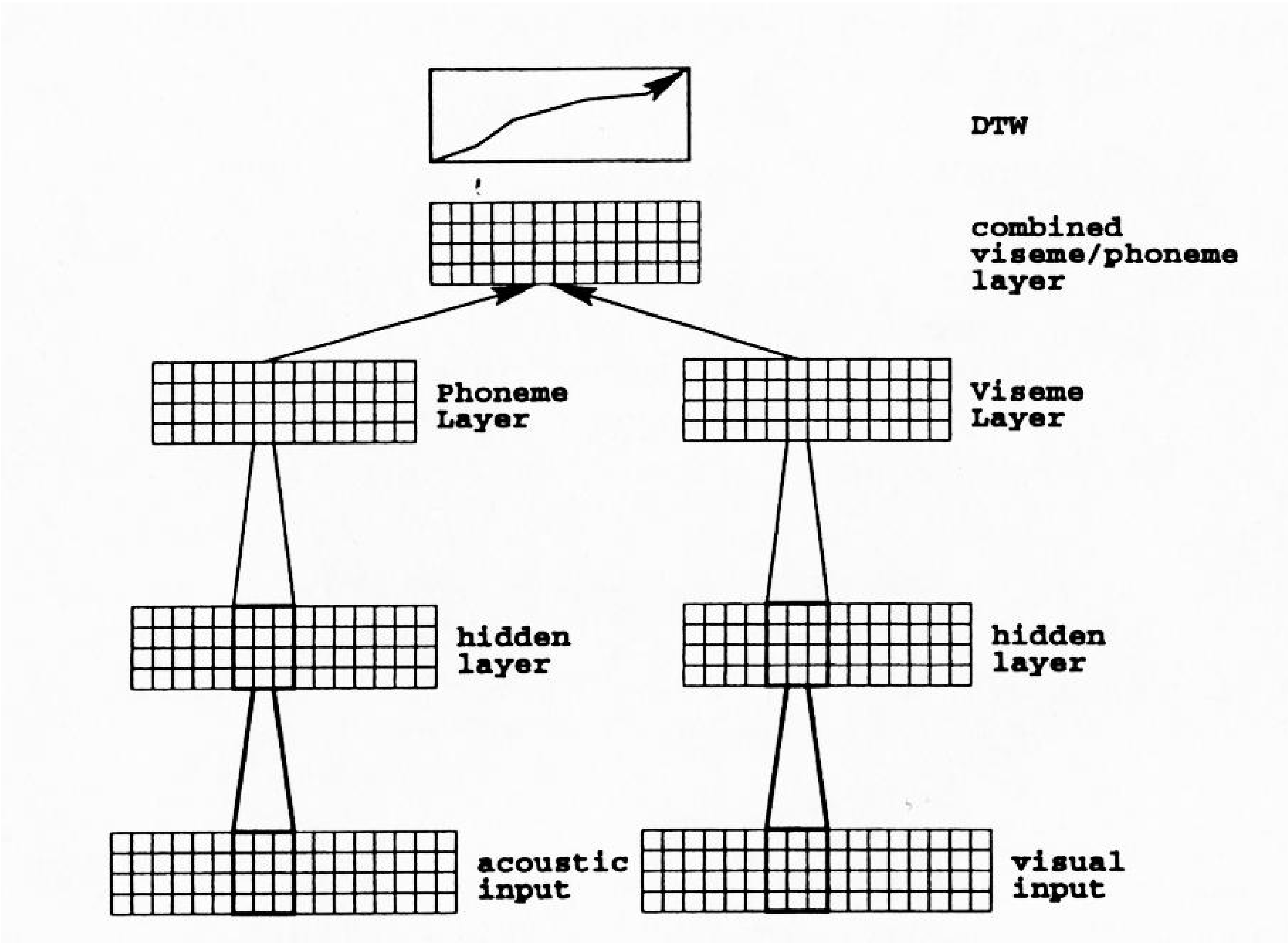
- Sentence-level training is achieved similarly.

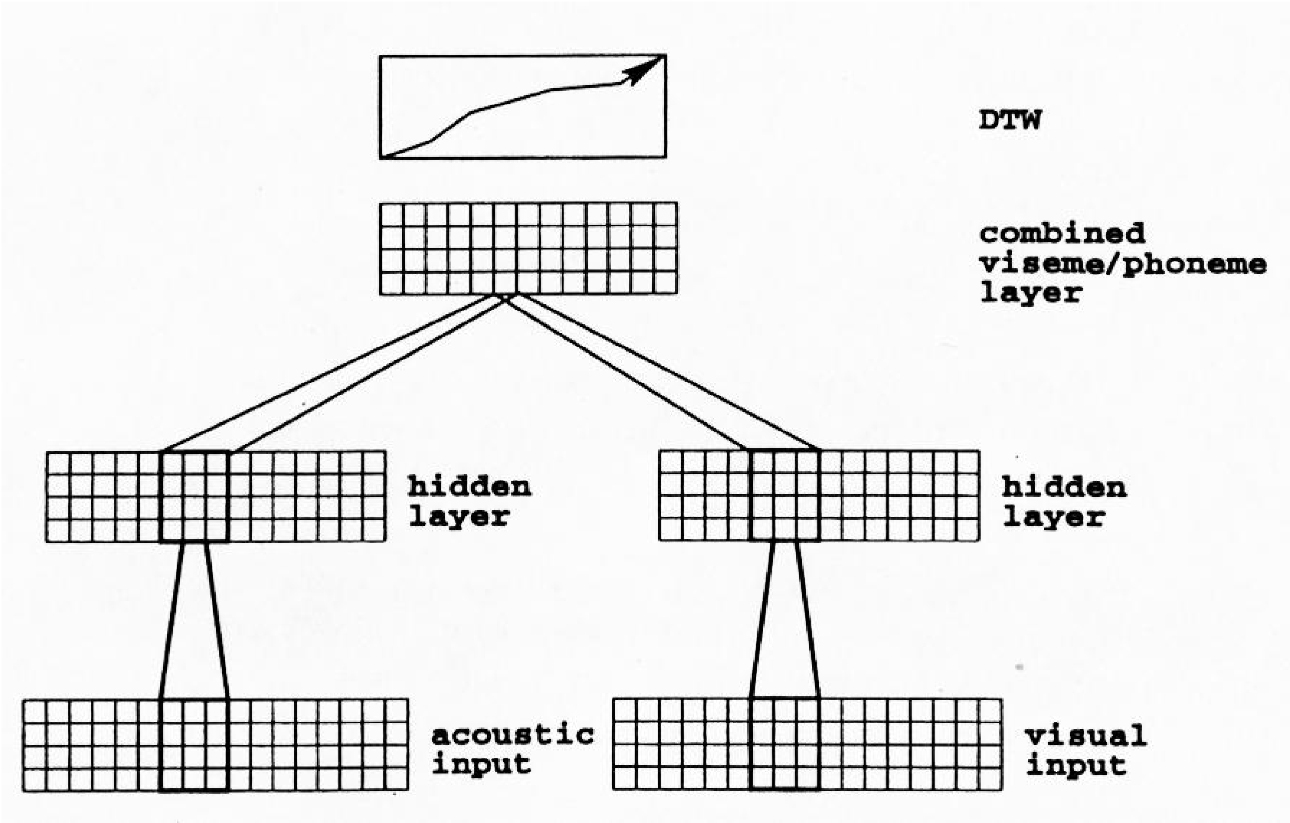
- Special rules to avoid insertion errors; e.g. T is recognised as TE.
  $\rightarrow$ Use word-entrance penalties.

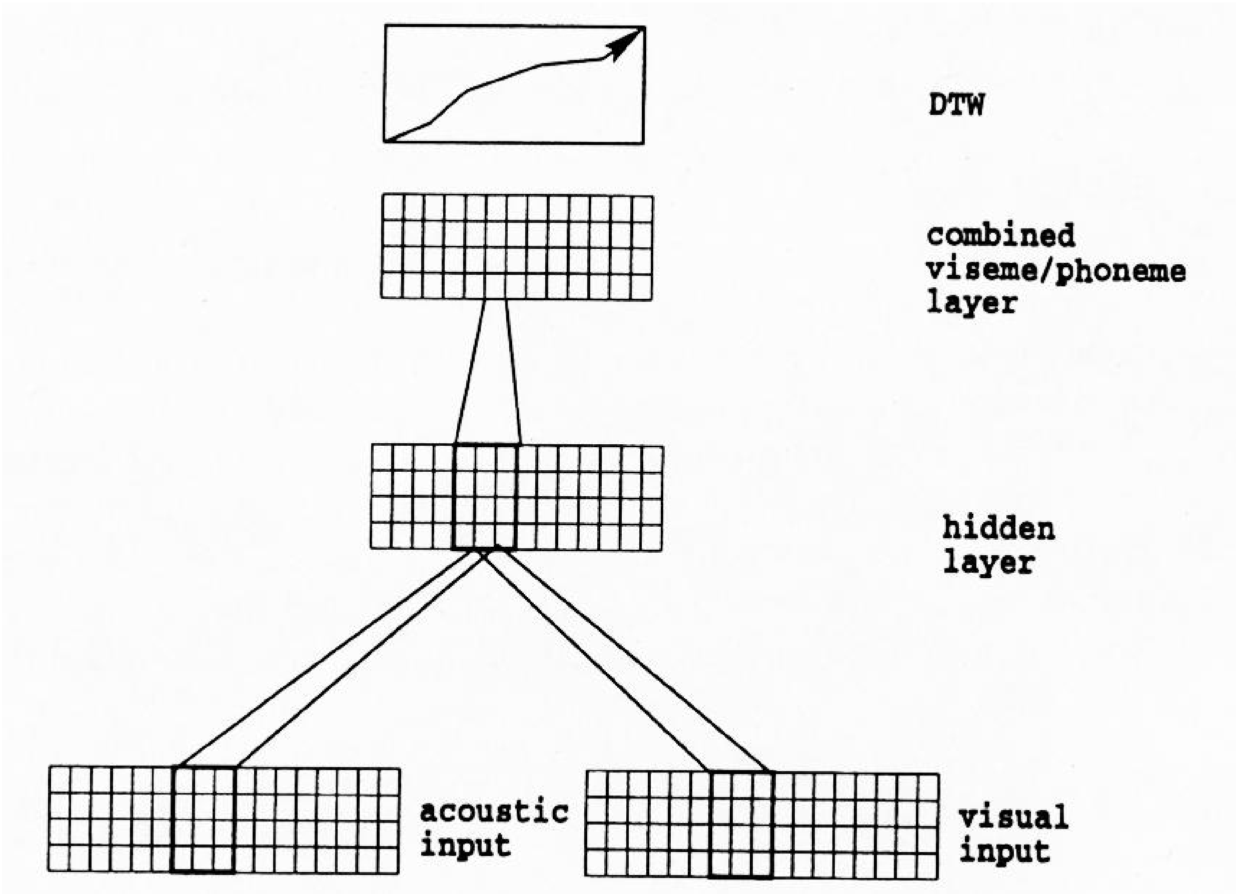
- In experiments, MS-TDNN performs better then HMM, mixed TDNN/HMM and linear predictive NN in letter recognition tasks.

Combining lipreading with acoustic signal

- The idea is to include visual information into the acoustic model.
- Fundamental visual unit corresponding to a phoneme is a *viseme*.
- Again, this is an imitation of human speech recognition, e.g. at a cocktail-party, looking at the speaker helps separate signal from others.
- The combination of visual and acoustic data can be done
  - after the phoneme and viseme layers on an additional layer
  - by combining the phoneme and viseme layers
  - by feeding both phoneme and viseme data into hidden layer


DTW
combined viseme/phoneme layer
Phoneme Layer
Viseme Layer
hidden layer
hidden layer
acoustic input
visual input


DTW
combined
viseme/phoneme
layer
hidden
layer
hidden
layer
acoustic
input
visual
input

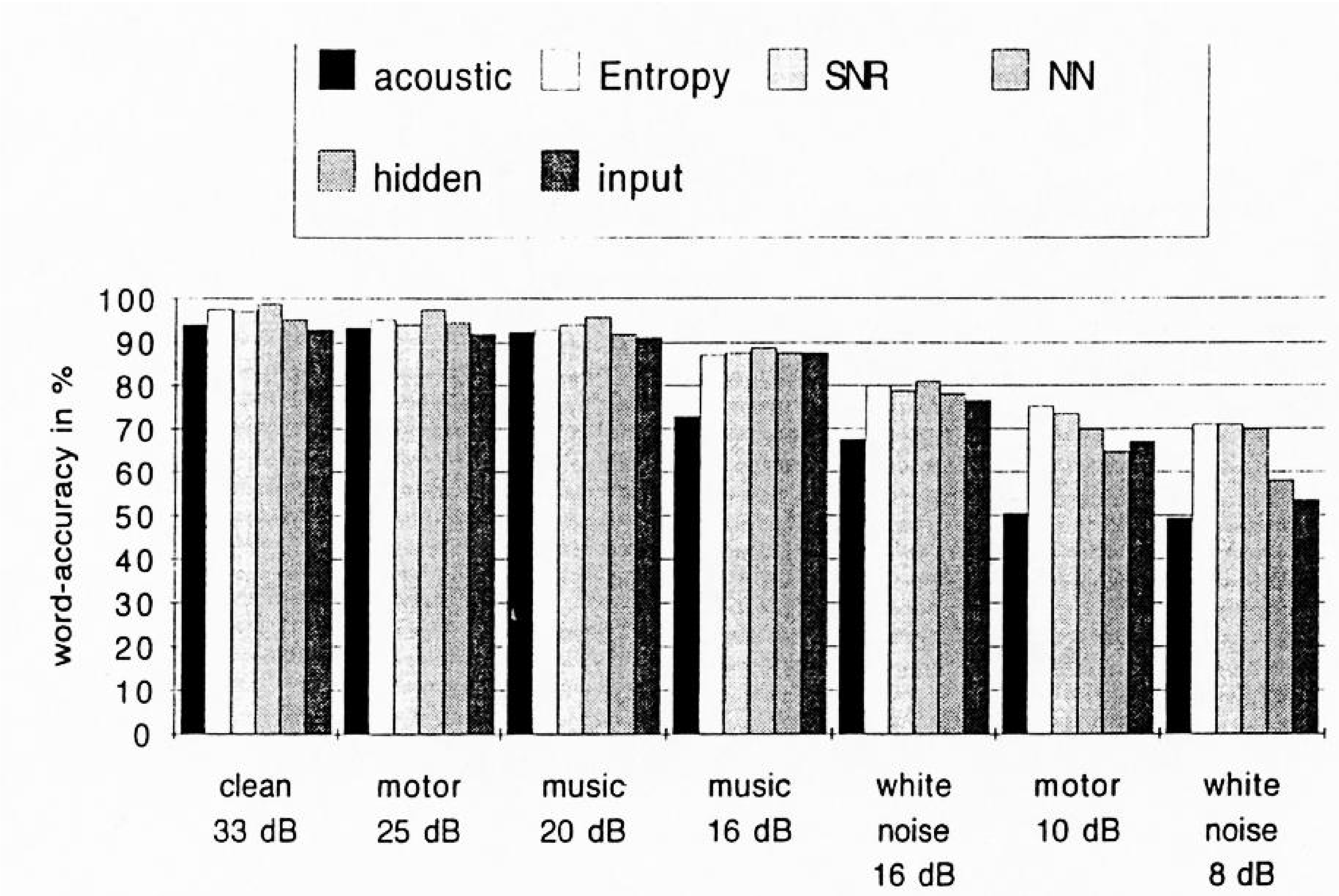

DTW
combined
viseme/phoneme
layer
hidden
layer
acoustic
input
visual
input

- Auditory and visual information importance can be weighted such that the audiovisual activation h_{AV} for a phoneme is

$$h_{AV} = \lambda_A h_A + \lambda_V h_V, \qquad \text{where } \lambda_A + \lambda_V = 1 \tag{1}$$

  Useful mostly when acoustic and visual phoneme activations are calculated separately.

- E.g., if the acoustic signal is noisy then rely more on visual information
- Choice of λ_A and λ_V can be done with:
  - *Entropy weight* – If phoneme and viseme activations are evenly spread then the respective phoneme and viseme entropies are high. High entropy means high ambiguity $\rightarrow$ lower weight.
  - *SNR weight* – Calculate estimates for SNR for phonetic and visual data and set weights accordingly.
  - *Neural net* – Make a simple back-prop network without a hidden layer to combine the visual and acoustic data.


acoustic
Entropy
SNR
NN
hidden
input
word-accuracy in %
100
90
80
70
60
50
40
30
20
10
0
clean 33 dB
motor 25 dB
music 20 dB
music 16 dB
white noise 16 dB
motor 10 dB
white noise 8 dB

Hierarchical mixtures of experts (HME)

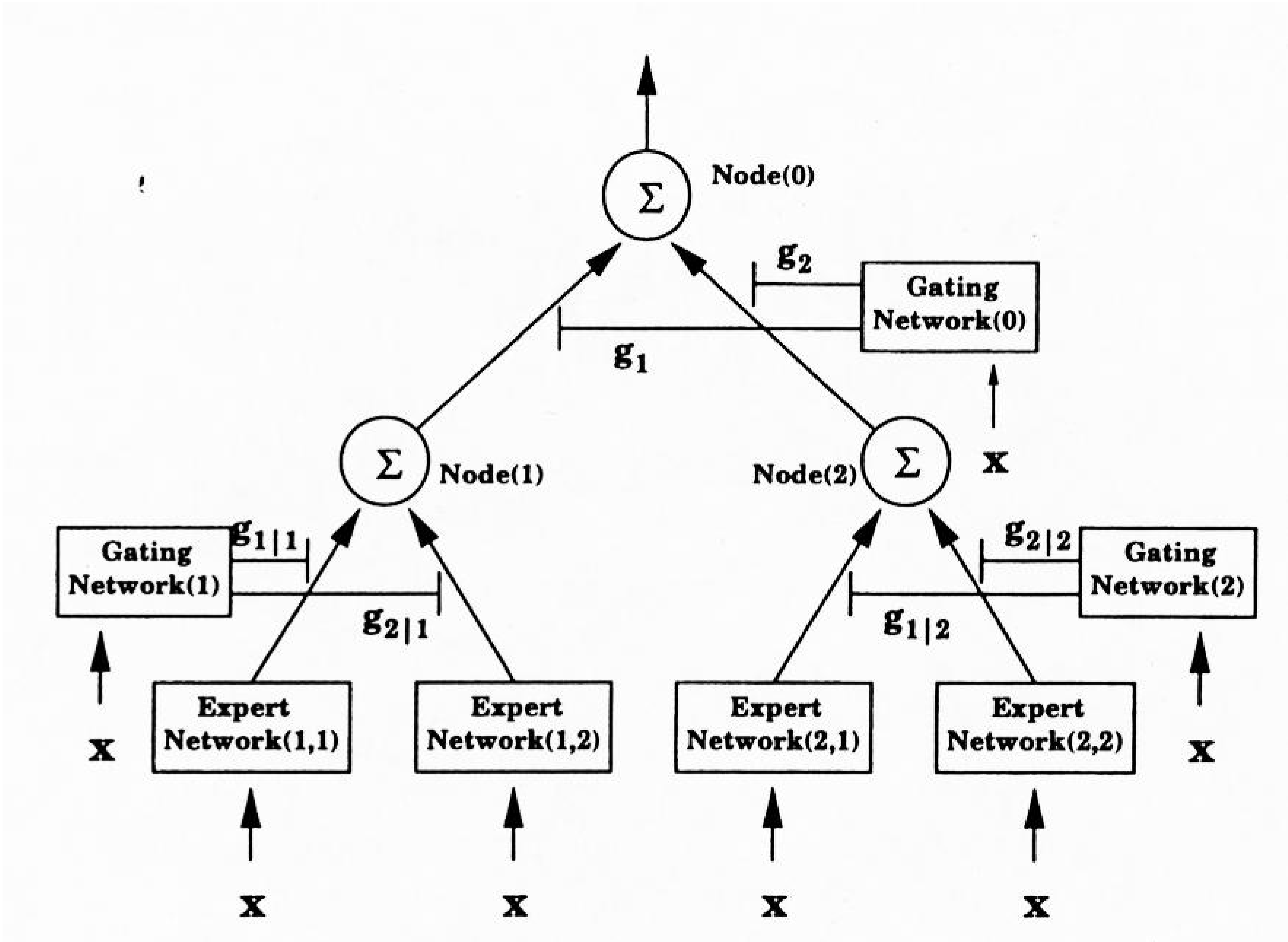
- Divide-and-conquer strategy: Learning task is divided into overlapping regions which are trained separately with experts.

- Gating networks are trained to choose the right expert for each input.

- The overall output of the architecture is

$$\mu(\mathbf{x}, \mathbf{\Theta}) = \sum_{i=1}^{N} g_i(\mathbf{x}, \mathbf{v}_i) \sum_{i=1}^{N} g_{j|i}(\mathbf{x}, \mathbf{v}_{ij}) \mu_{ij}(\mathbf{x}, \theta_{ij}) \tag{2}$$

  where the g_i and $g_{j|i}$ are outputs of the gating networks and the μ_{ij} represent gating networks. The parameters of gating and expert networks are denoted by v_i, v_{ij} and θ_{ij}.


Node(0)
g_2
Gating Network(0)
g_1
x
Node(1)
Node(2)
$g_{1|1}$
Gating Network(1)
$g_{2|1}$
$g_{2|2}$
Gating Network(2)
$g_{1|2}$
Expert Network(1,1)
Expert Network(1,2)
Expert Network(2,1)
Expert Network(2,2)
x

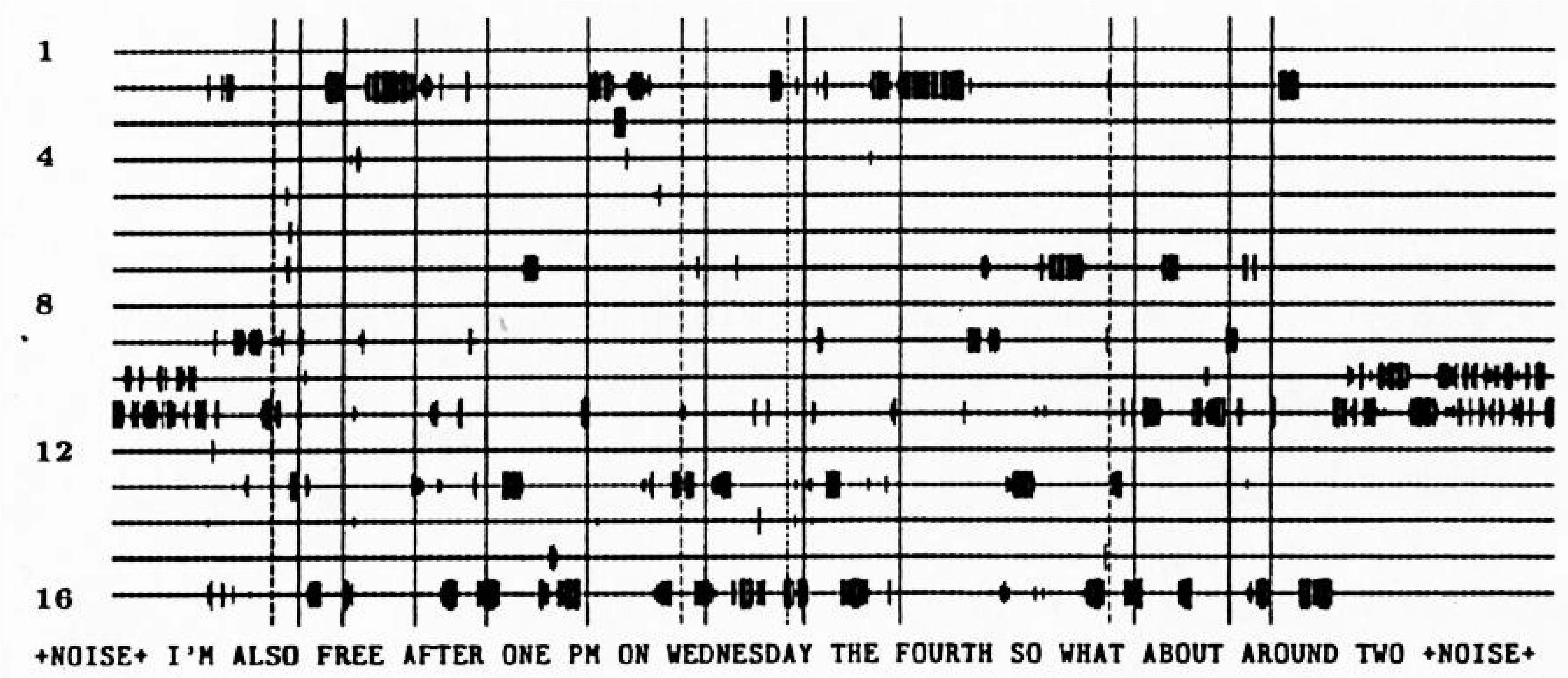
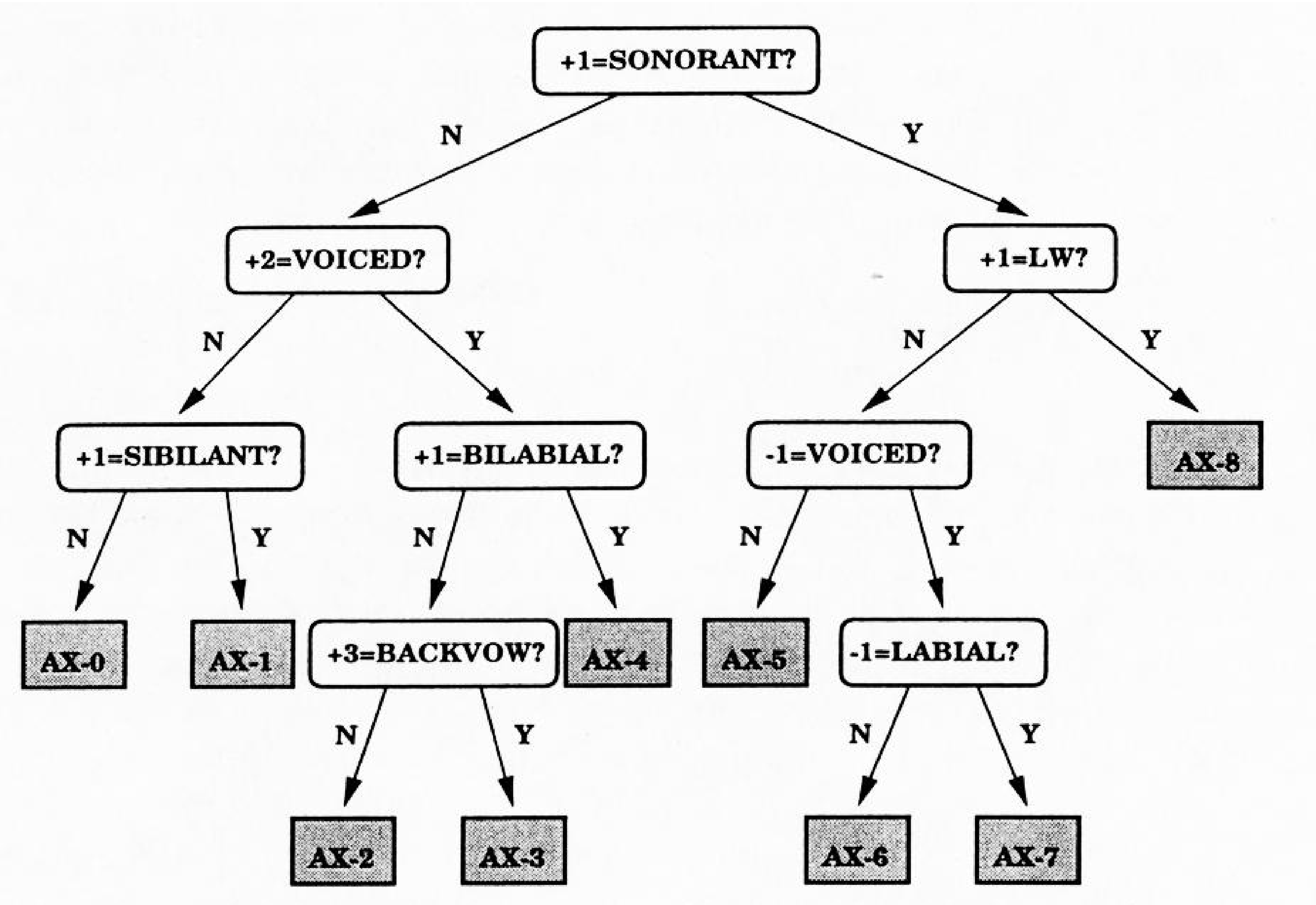


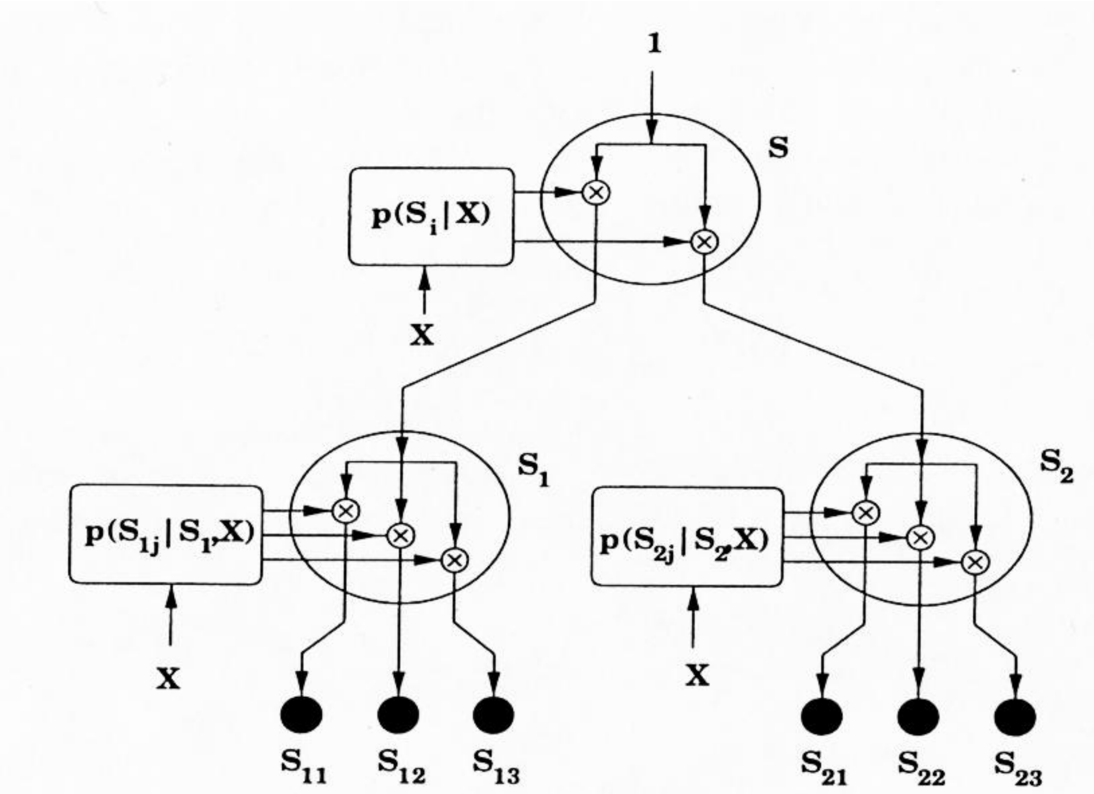
Figure 1: Expert activation diagram

Context modelling

- The context of the current phoneme is the combination of previous and following phonemes.
- Problem:
  - Context affects phonemes and a context model would therefore improve recognition.
  - Number of different phoneme combinations is large.
- One solution: Clustering context classes
  - Cluster phonemes in phoneme groups (e.g. labial phonemes) with relevant information for the context.
  - Represent phonemes classes in a binary decision tree — trees may end up with ~ 24.000 leaves.
  - Each leaf in the tree represent a state in the HMM.

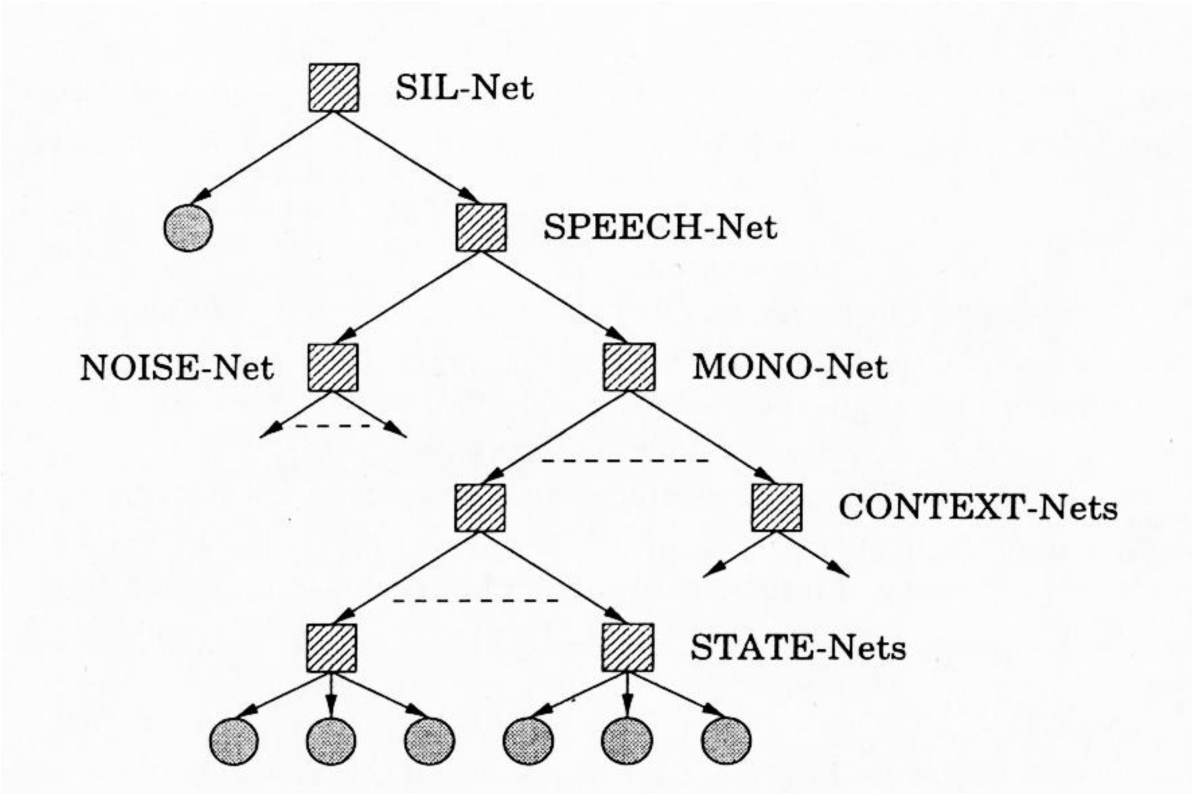

+1=SONORANT?
N
Y
+2=VOICED?
+1=LW?
N
Y
N
Y
+1=SIBILANT?
+1=BILABIAL?
-1=VOICED?
AX-8
N
Y
N
Y
N
Y
AX-0
AX-1
+3=BACKVOW?
AX-4
AX-5
-1=LABIAL?
N
Y
N
Y
AX-2
AX-3
AX-6
AX-7

- The tree provides prior state probabilities.
- The neural network provides joint probability.
  $\rightarrow$ posteriori probabilities can be calculated.
- States can be calculated hierarchically.



Structuring further

- NN's can be manually set to a hierarchy structure on higher level then context. I.e., noise-speech classification and similar.



- NN's can be automatically clustered through training.

Home assignment

On the basis of chapter 7, please explain briefly the differences and similarities between MS-TDNN and HMM-NN.