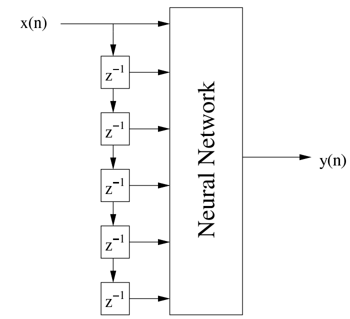
Time-Delay Neural Networks and NN/HMM Hybrids: *A Family of Connectionst Continuous-Speech Recognition Systems*

(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

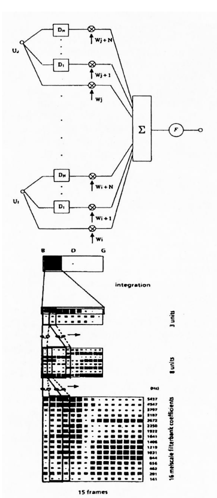
Introduction



Methods:
- Time-Delay Neural Networks (TDNN
- Multi-state Time-Delay Neural Ne works (MS-TDNN)
- Neural Networks with Hidden Marko Model (NN/HMM)

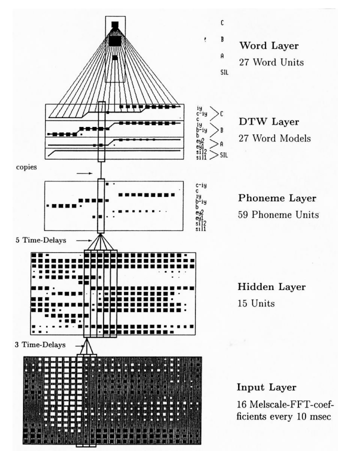
Applications:
- Continuous spelling recognition
- Lipreading (audio-visual hybrids)
- Large-vocabulary conversation ov telephone

The first Time-Delay NN's

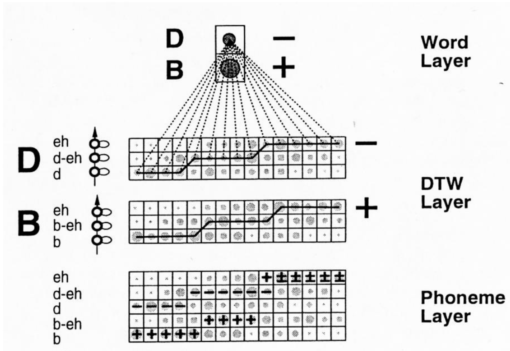


- Developed by Weibel and Lang in 1987.
- Calculate phoneme scores directly from speech segments.
- Feed time-delayed segments into NN with equal weights.
- Trained with phonemes /b/, /d/ and /g/.
- Shift-invariant model, with not too large set of parameters.

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.

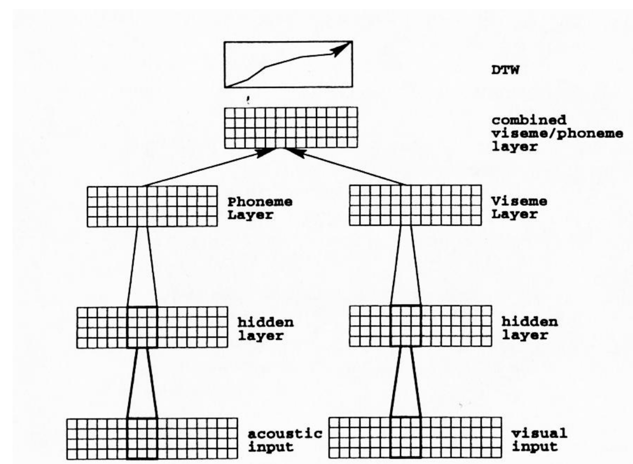


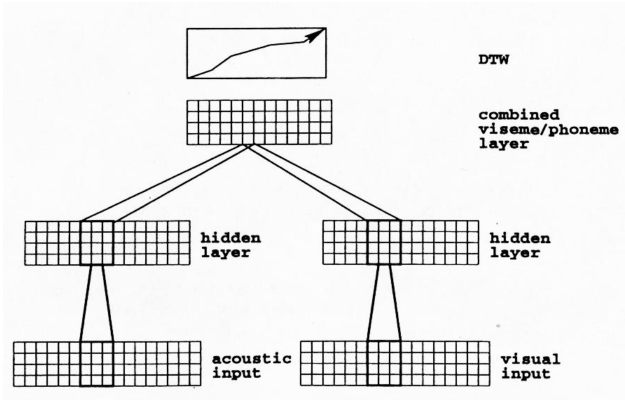
- In word-level training an alignment path is searched which maximise the phoneme sum scores.
- Path searching similarly as in HMM.
- Phonemes on the correct path receive positive training and on other paths negative training.

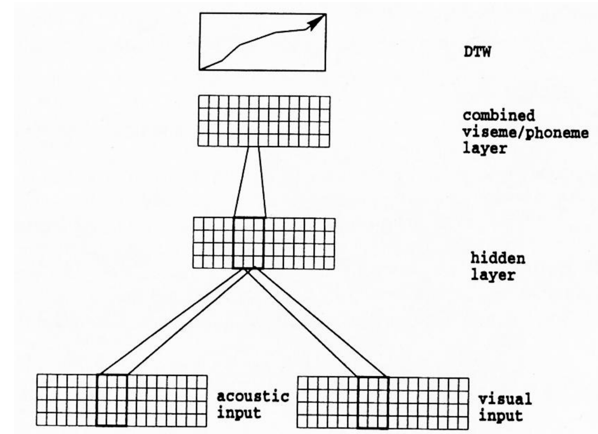
- Sentence-level training is achieved similarly.
- Special rules to avoid insertion errors; e.g. T is recognised as TE.
  → Use word-entrance penalties.
- In experiments, MS-TDNN performs better then HMM, mixed TDNN/H
  and linear predictive NN in letter recognition tasks.

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





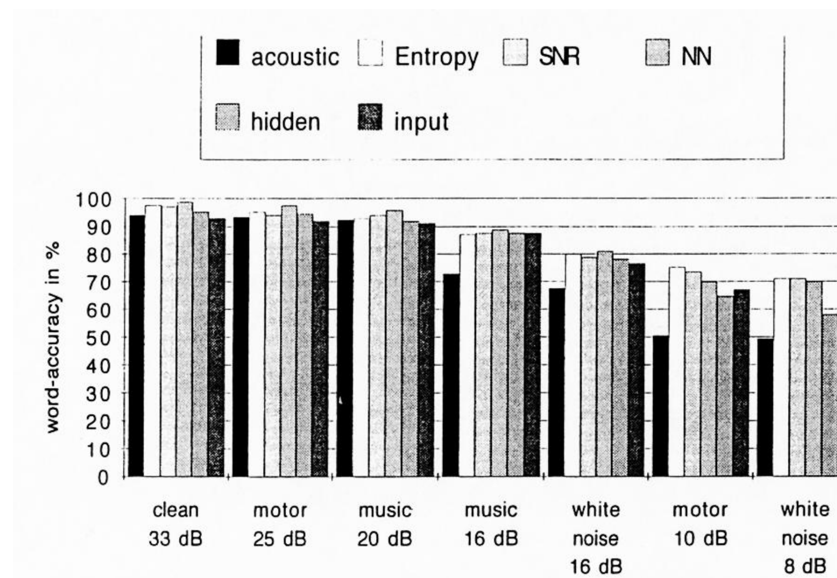


- Auditory and visual information importance can be weighted such that the audiovisual activation h_{AV} for a phoneme is

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

  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 → 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.



Hierarchical mixtures of experts (HME)

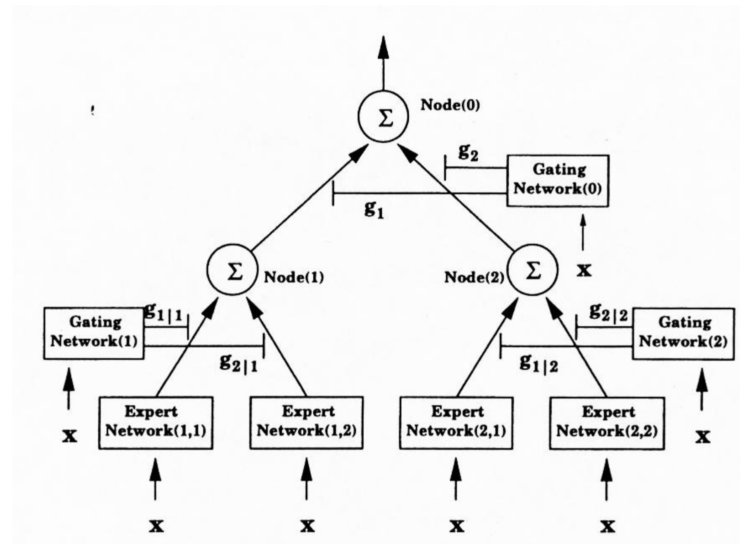
- Divide-and-conquer strategy: Learning task is divided into overlapping regions which are trained separately with experts.

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

- The overall output of the architecture is

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

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}.



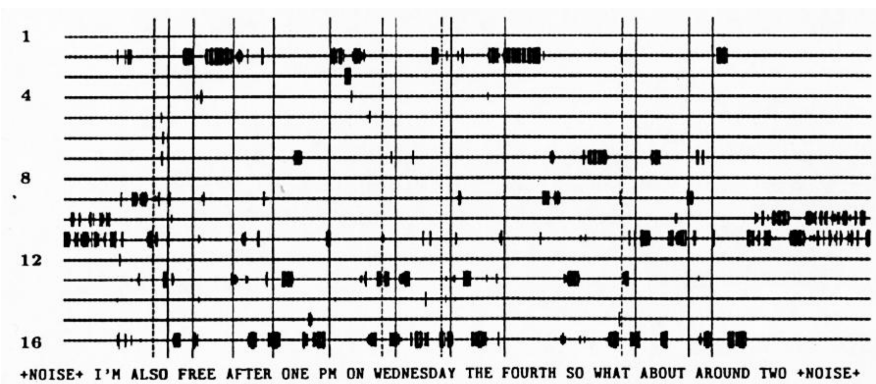
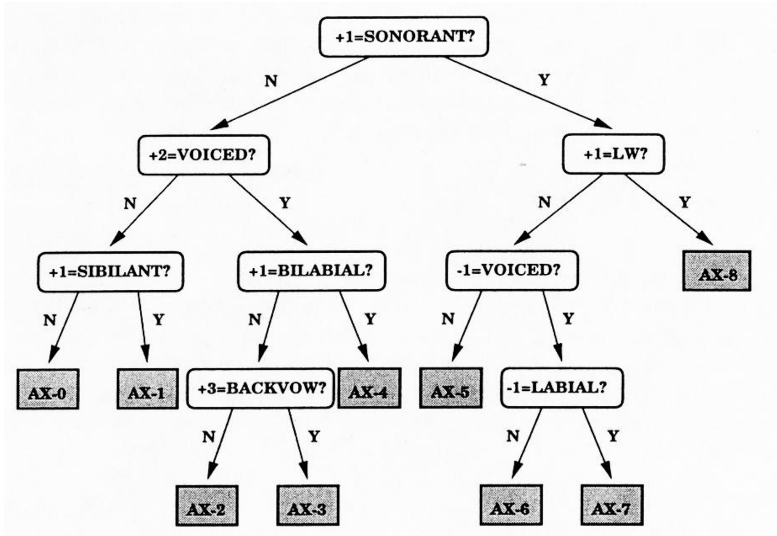


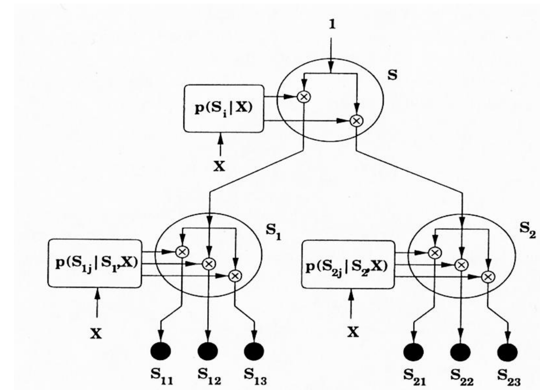
Figure 1: Expert activation diagram

Context modelling

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

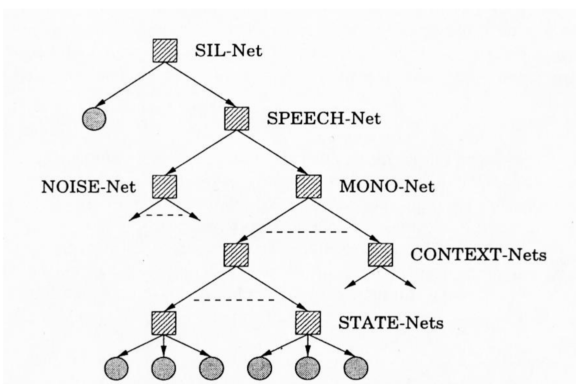


- The tree provides prior state probabilities.
- The neural network provides joint probability.
  → posteriori probabilities can be calculated.
- States can be calculated hierarchically.



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

Explain shortly which different approaches can be used to include visual information to speech recognition and how these approaches differ from each other.