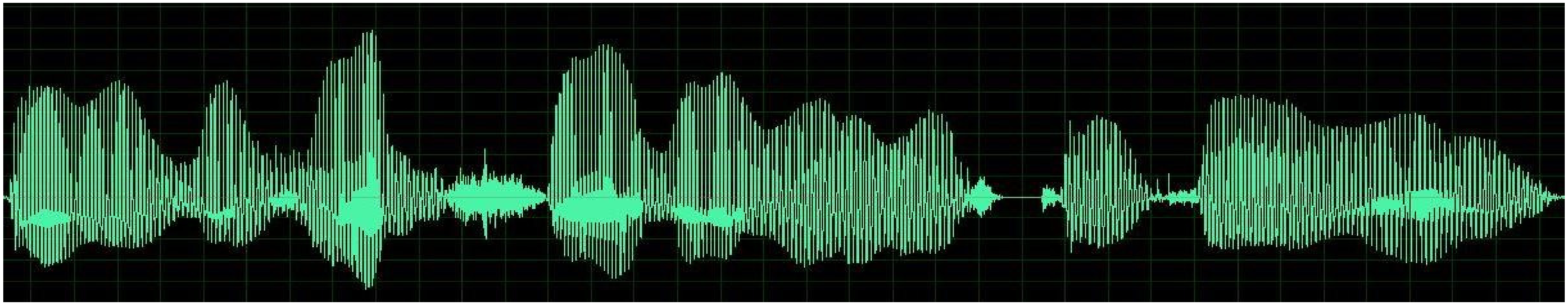
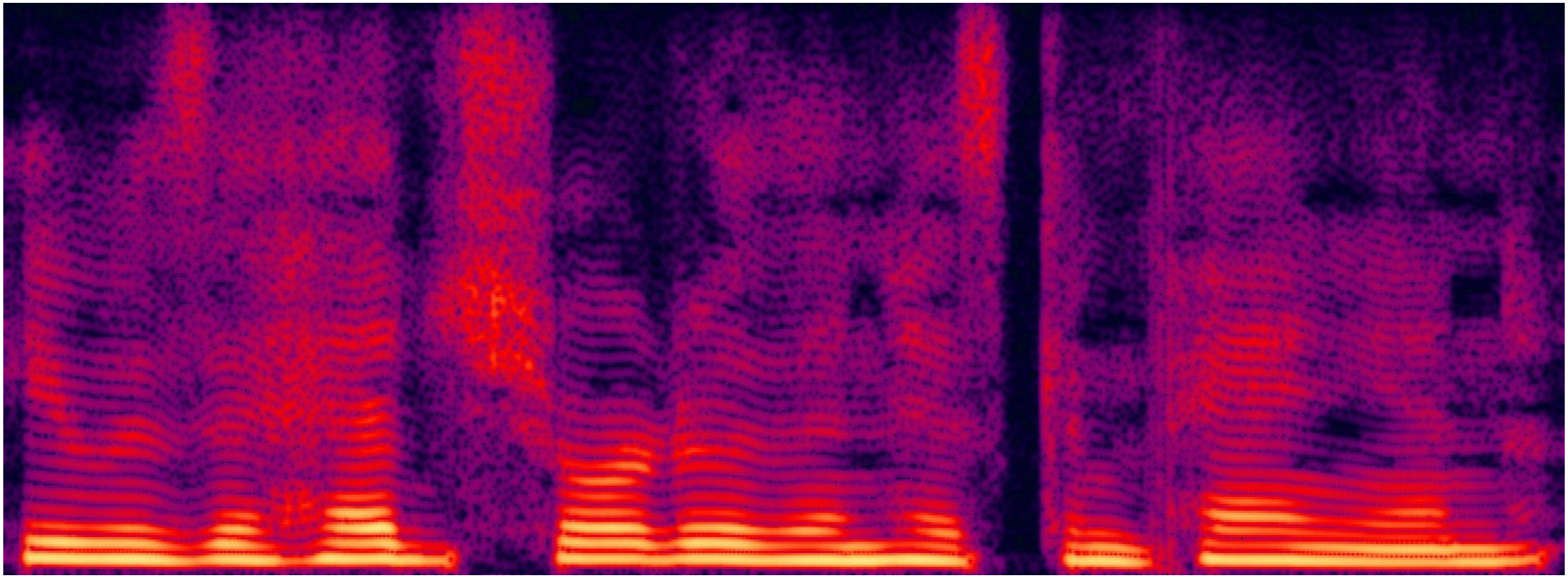
Speech Recognition Using Deep Believe Networks

Abdel-rahman Mohamed

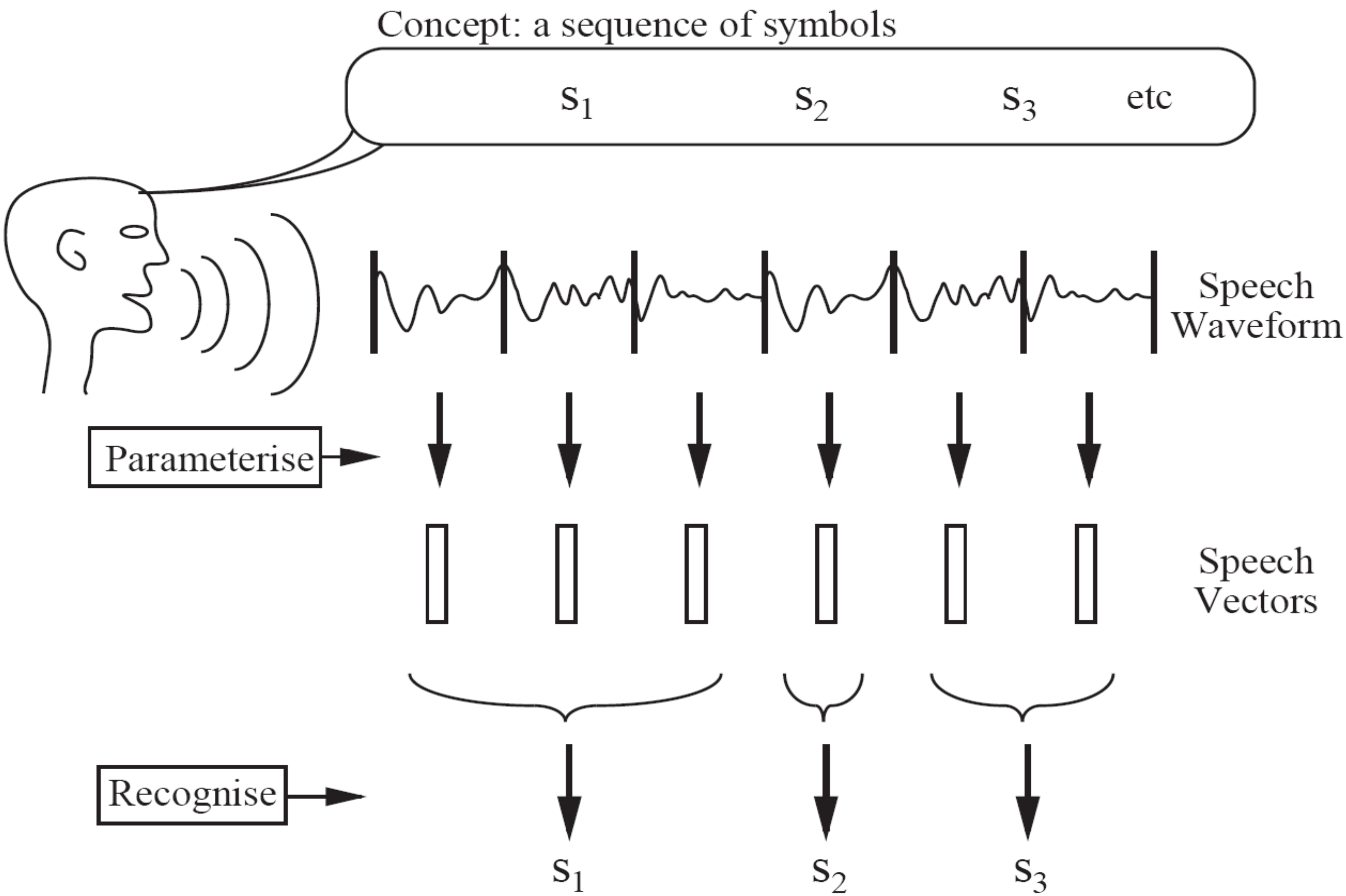
Department of Computer Science
University of Toronto

The speech signal

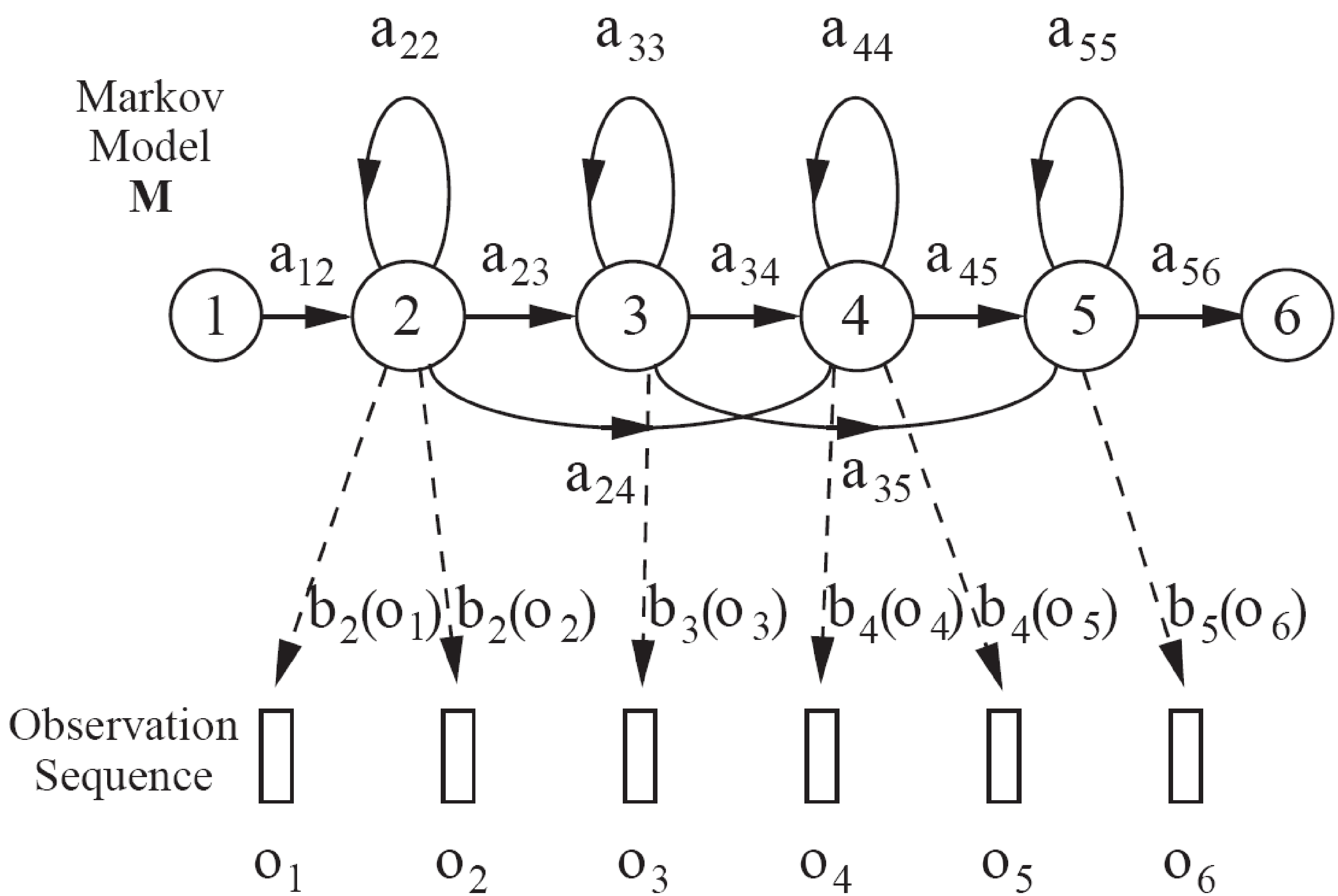




Automatic Speech Recognition (ASR)



ASR existing models



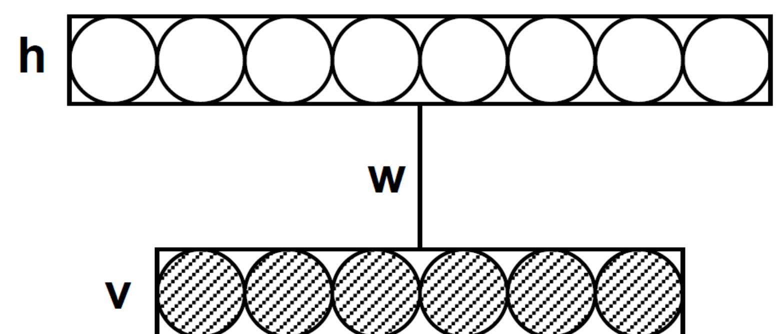
Motivation

- The state-of-the-art techniques for acoustic modeling suffers from unrealistic independence assumptions.

- Looking for new models that offer more representational capacity.

Restricted Boltzmann Machines (RBMs) (1)

- An RBM is a bipartite graph in which visible units are connected to binary stochastic hidden units using undirected weighted connections.
- RBMs have an efficient generative training procedure as well as discriminative fine tuning mechanisms.



RBMs (2)

- The energy of the joint configuration (v, h) is given by:

$$E(\mathbf{v},\mathbf{h};\theta) = \sum_{i=1}^{\mathcal{V}} \frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{\mathcal{V}} \sum_{j=1}^{\mathcal{H}} w_{ij} v_i h_j - \sum_{j=1}^{\mathcal{H}} a_j h_j$$

- The probability that the model assigns to a visible vector v is:

$$p(\mathbf{v};\theta) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{u}} \sum_{\mathbf{h}} e^{-E(\mathbf{u},\mathbf{h})}}$$

- Conditional distributions $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$ are factorial and given by:

$$\begin{aligned} p(h_j = 1|\mathbf{v};\theta) &= \sigma(\sum_{i=1}^{\mathcal{V}} w_{ij} v_i + a_j) \\ p(v_i = 1|\mathbf{h};\theta) &= \mathcal{N}(\sum_{j=1}^{\mathcal{H}} w_{ij} h_j + b_i, 1) \end{aligned}$$

Using RBMs for phone recognition

(Mohamed, Hinton, ICASSP 2010)

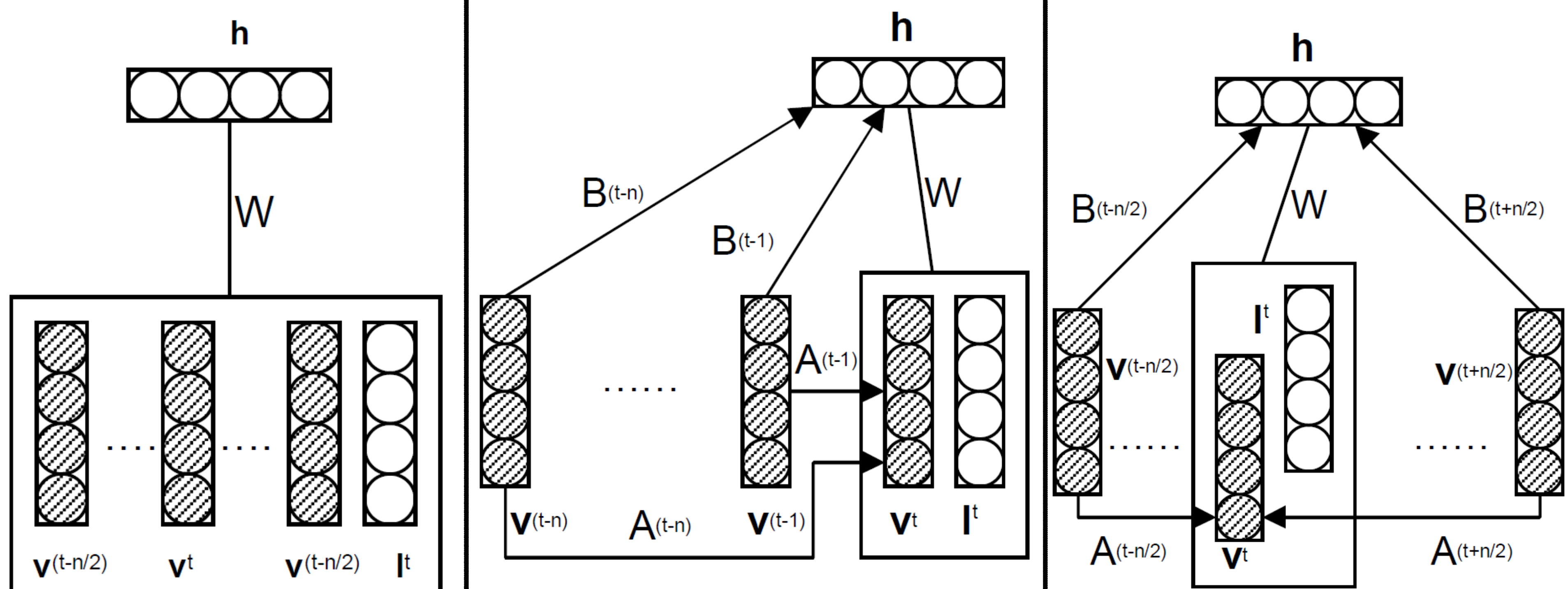
- A context window of successive feature vectors is used to set the states of the visible units.
- To train an RBM to model the joint distribution of data and labels, the visible vector is concatenated with a binary vector of class labels.

$$E(\mathbf{v},\mathbf{l},\mathbf{h};\theta) = -\sum_{i=1}^{\mathcal{V}}\sum_{j=1}^{\mathcal{H}} w_{ij}h_jv_i - \sum_{k=1}^{\mathcal{L}}\sum_{j=1}^{\mathcal{H}} w_{kj}h_jl_k - \sum_{j=1}^{\mathcal{H}} a_jh_j - \sum_{k=1}^{\mathcal{L}} c_kl_k + \sum_{i=1}^{\mathcal{V}} \frac{(v_i-b_i)^2}{2}$$

$$p(l_k=1|\mathbf{h};\theta) = \mathrm{softmax}(\sum_{j=1}^{\mathcal{H}} w_{kj}h_j + c_k)$$

- The DBN produces a probability distribution over the possible labels of the central frame. Then probabilities are fed to a standard Viterbi decoder.

RBM and its variants



(a) RBM (b) CRBM (c) ICRBM

RBM training: Generative training

- By maximizing the likelihood function of the visible data, we get:

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$$

- The Contrastive Divergence (CD) approximation is used:

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_1$$

- For the CRBM, the directed connection updates are:

$$\Delta A_{ij}^{(t-q)} = v_i^{(t-q)} (\langle v_j^t \rangle_{data} - \langle v_j^t \rangle_1)$$
$$\Delta B_{ij}^{(t-q)} = v_i^{(t-q)} (\langle h_j^t \rangle_{data} - \langle h_j^t \rangle_1)$$

RBM training: Discriminative training

- $p(\mathbf{l}|\mathbf{v})$ can be computed exactly by:

$$p(\mathbf{l}|\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{l},\mathbf{h})}}{\sum_{\mathbf{l}} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{l},\mathbf{h})}}$$

- The gradient of $log\ p(\mathbf{l}|\mathbf{v})$ can also be computed exactly. The update rule for the vis-hid weights is:

$$\Delta w_{ij} = v_i \sigma \left(a_j + w_{jm} + \sum_{i=1}^{\mathcal{V}} w_{ij} v_i \right) \ - \ v_i \sum_{k=1}^{\mathcal{L}} p(l_k = 1|\mathbf{v}) \sigma \left(a_j + w_{jk} + \sum_{i=1}^{\mathcal{V}} w_{ij} v_i \right)$$

- To avoid model overfitting, we follow the gradient of: $f(\mathbf{v}, \mathbf{l}) = \alpha\ log\ p(\mathbf{l}|\mathbf{v}) + log\ p(\mathbf{v}|\mathbf{l})$

Evaluation Setup

- The core test set of the TIMIT database is used. The MIT development set (50 speakers) was used for model tuning.
- 12^{th} order MFCC and energy along with 1^{st} and 2^{nd} derivatives were used as features.
- A context window of 11 feature frames was used.
- All architectures contain 2000 hidden units.
- We used 183 target class labels (3 states*61 phones).
- We used a bigram language model over phones, estimated from the training set of TIMIT.

Evaluation

- Using the generative objective, PER percentages are:

RBM	CRBM	ICRBM
36.9 %	42.7 %	39.3 %

- Using the hybrid objective function:

RBM	ICRBM
27.5 %	26.7 %

Evaluation

- Comparison with feedforward neural networks

NN (random weights)	NN (RBM weights)	ICRBM
28.7 %	28.3 %	26.7 %

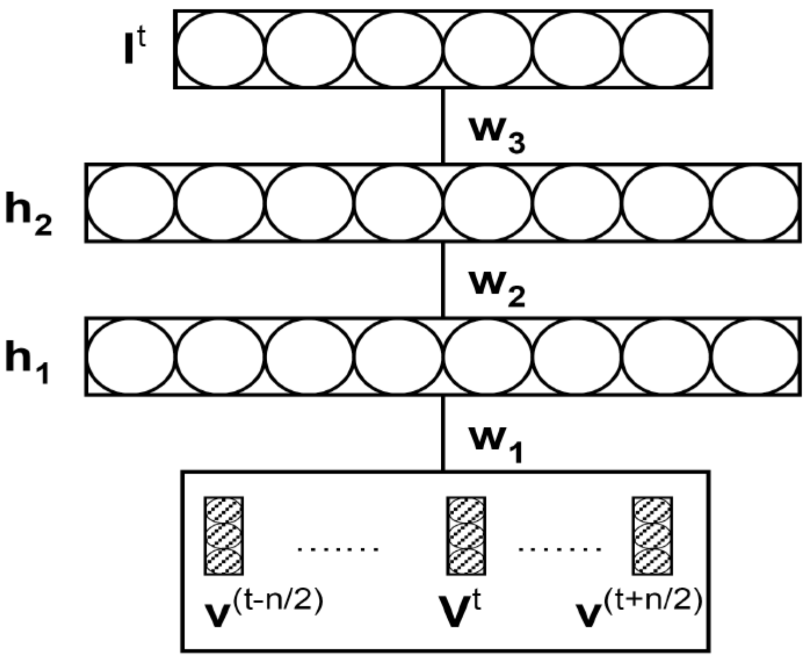
- A two-tailed Matched Pairs Sentence-Segment Word Error (MAPSSWE) significance test showed that ICRBM is significantly better.

Using DBNs for phone recognition

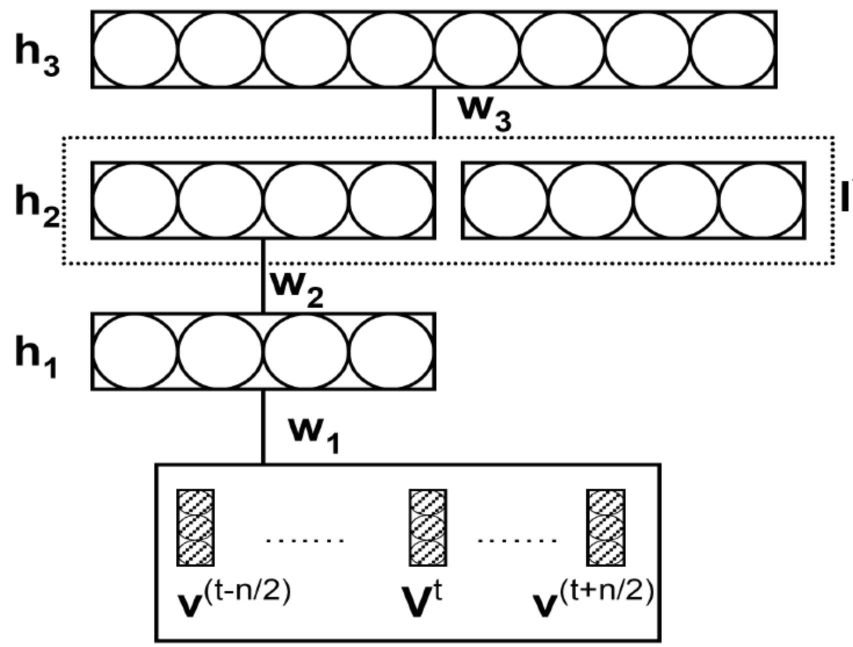
(Mohamed, Dahl, Hinton, NIPS workshop 2009)

We employed two types of DBN architectures:

- The BP-DBN: It performs a purely discriminative fine-tuning phase using backpropagation.
- The AM-DBN: It has an RBM associative memory for the final layer to model joint density of labels and inputs. The hybrid objective function is used for fine-tuning.

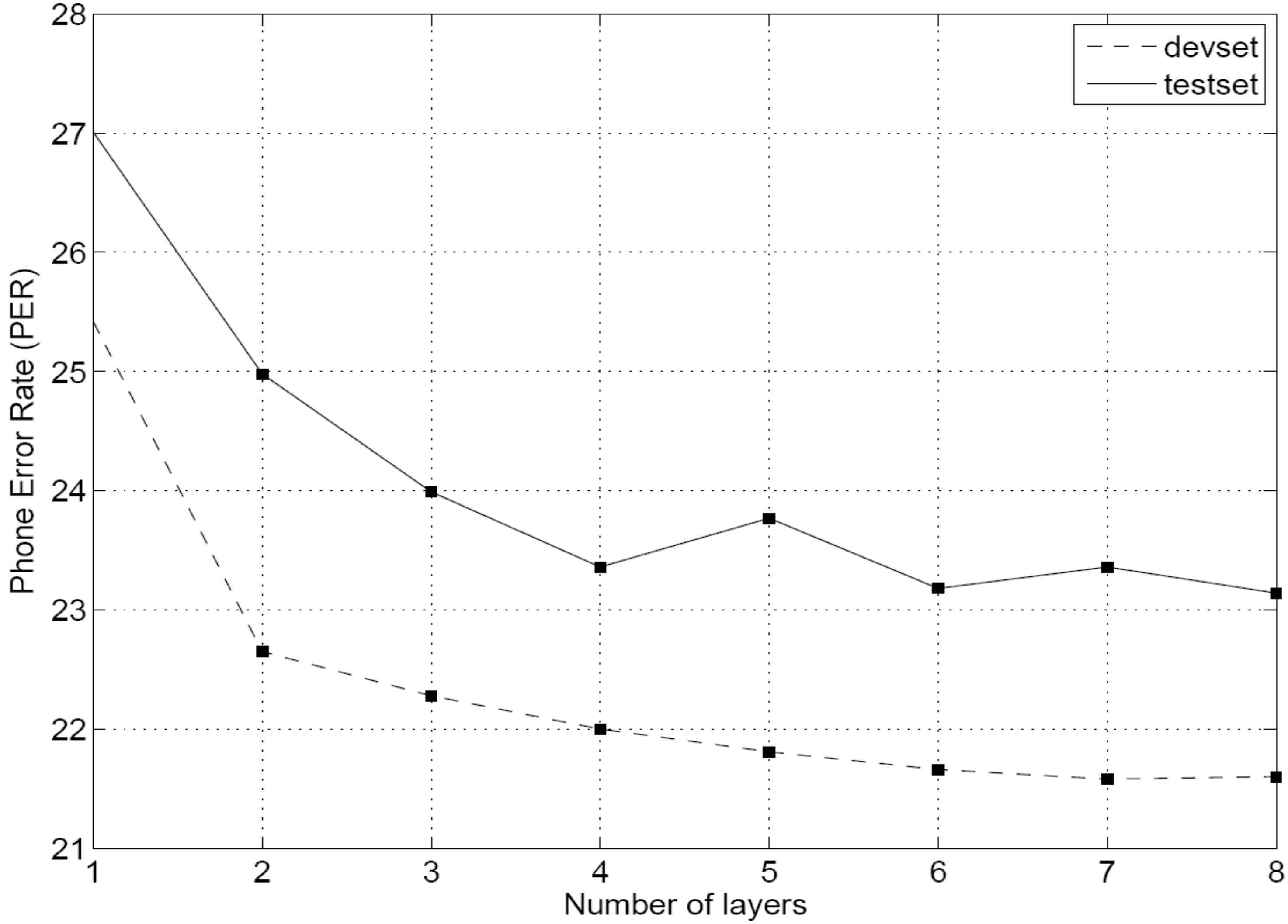


(a)A 2-layer BP-DBN



(b)A 3-layer AM-DBN

Evaluation: How deep should the model be?



Evaluation

Method	PER
Large Margin GMM	33.0 %
ML trained CD-HMM	27.3 %
ICRBM	**26.7 %**
Recurrent NN	26.1 %
Monophone HTMs	24.8 %
Heterogeneous Classifiers	24.4 %
Deep Belief Network (DBN)	**23 %**
CD-HMM trained with BMMI (IBM)	22.7%
DBN with mcRBM as the 1st layer	**20.5%**

Thank you