

Beyond Symbolic AI

Stephen M. Omohundro

International Computer Science Institute

Berkeley, California

- Symbolic AI and connectionism: toward a synthesis.
- Learning and Recognition.
- The Bayesian framework.
- Model merging and bumptrees.
- Surface learning and lip recognition.
- Stochastic grammar learning and speech recognition.

Symbolic AI

Strengths:

- Coherent underlying semantics
- Represents dynamic structures: Higher-order logic
- Representation of knowledge is understandable

Weaknesses:

- Tend to be brittle, (eg. expert systems)
- Poor for representing uncertainty
- Poor for quantitative knowledge, eg. speech, vision
Need to ground symbols in perception.
- Learning is fairly weak
- High development costs (need for large amounts of hand-entered knowledge makes construction time-consuming and expensive).

Neural Networks

Strengths:

- Evidential representation
- Naturally quantitative, eg. vision and speech
- Quantitative learning
- Naturally parallel

Weaknesses:

- Fixed network structure (eg. variable binding, relational knowledge)
- Opaque knowledge representation
- Poor semantics
- Need structural learning

Toward a Synthesis

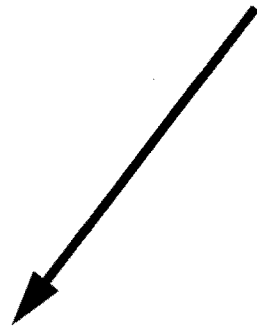
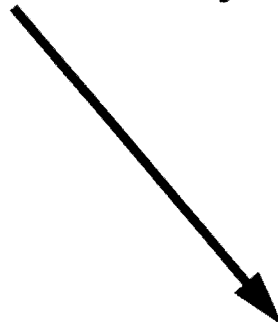
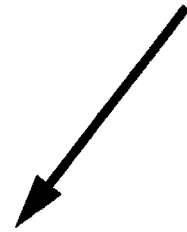
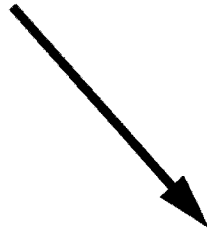
Symbolic AI

Neural Networks

Symbols + Uncertainty

Structured Connectionism

Synthesis



Desirable characteristics

Large amounts of knowledge are necessary.

-> Must use learning to help build knowledge bases.

Symbols must be grounded in the real world.

-> Must be integrated with perception.

- Coherent probabilistic semantics
- Representation of geometric and physical information
- Powerful learning and generalization
- Dynamic structure and relational representation
- Computationally efficient

Induction in Recognition and Learning

- Both are inductive model building processes.
A *model* is an explanation that accounts for the data.

- ***Recognition:*** sensation -> perception

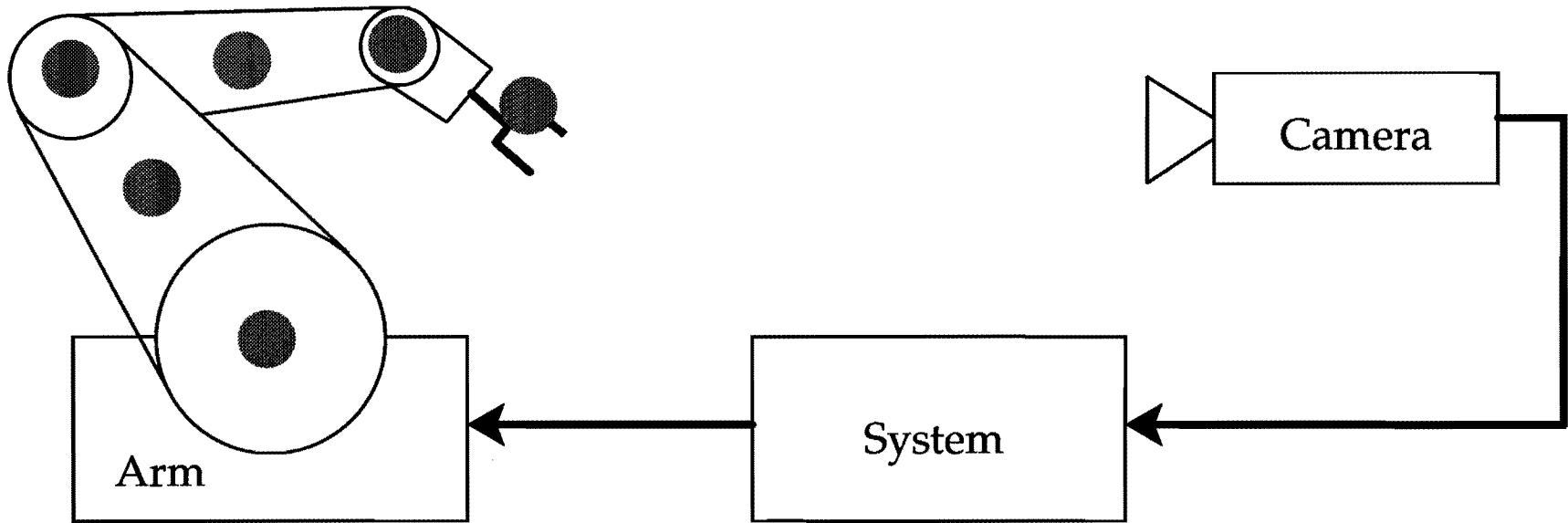
- ***Learning:*** past experience -> knowledge base

- Learning provides the model construction materials for recognition.

Simple Neural Net Learning

- Fixed space of possible models.
(eg. a neural network)
- Simple parameterization of model space.
(eg. network weights)
- Start at random parameter value.
(eg. random weights)
- Gradually modify parameters to improve performance on data.
(eg. backpropagation for stochastic error gradient descent)
- Possibly include terms to help prevent overfitting .
(eg. weight decay or cross-validation)

A Visually Guided Robot Arm.



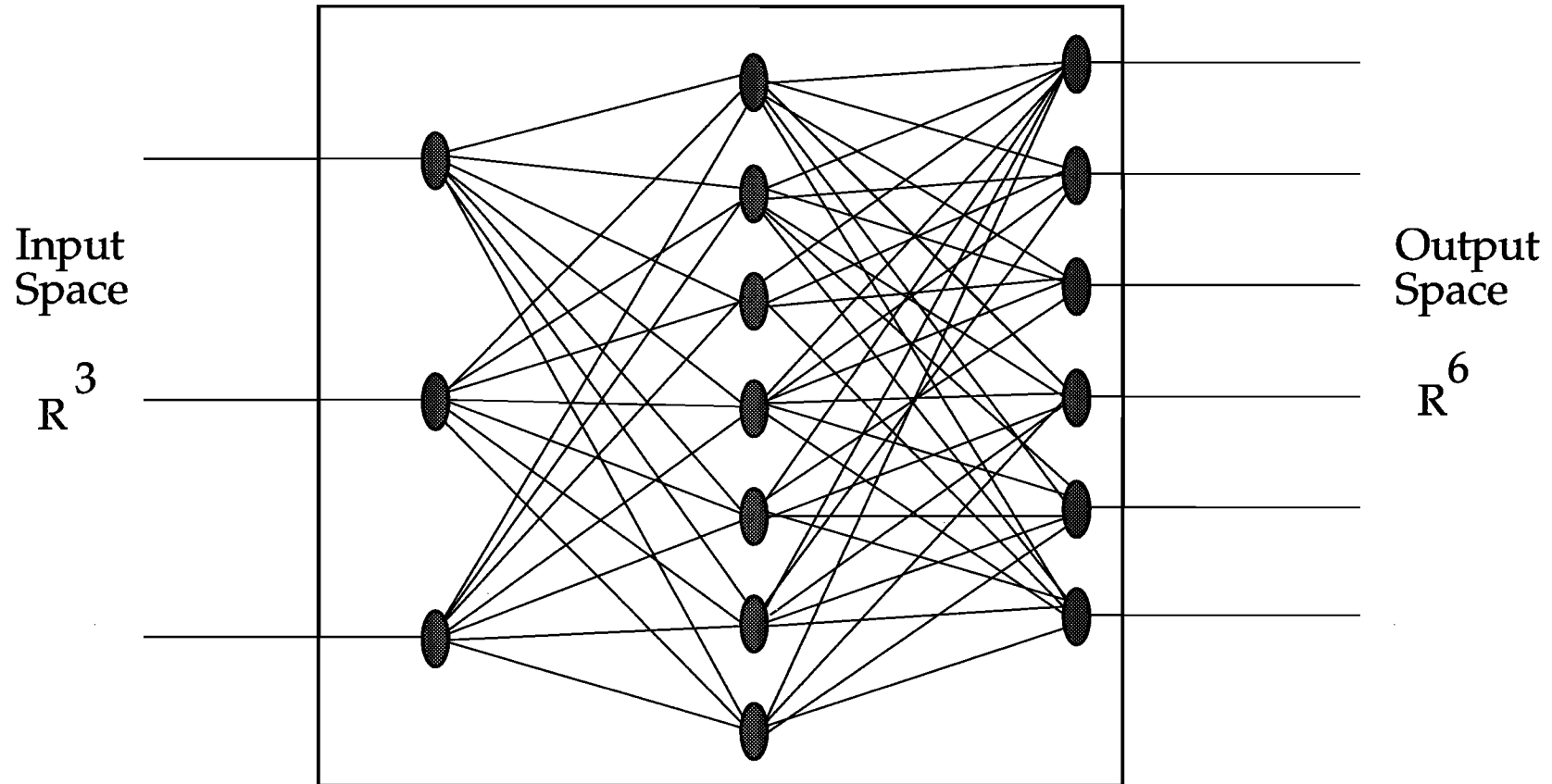
Kinematic space

Visual space

R^3

R^{12}

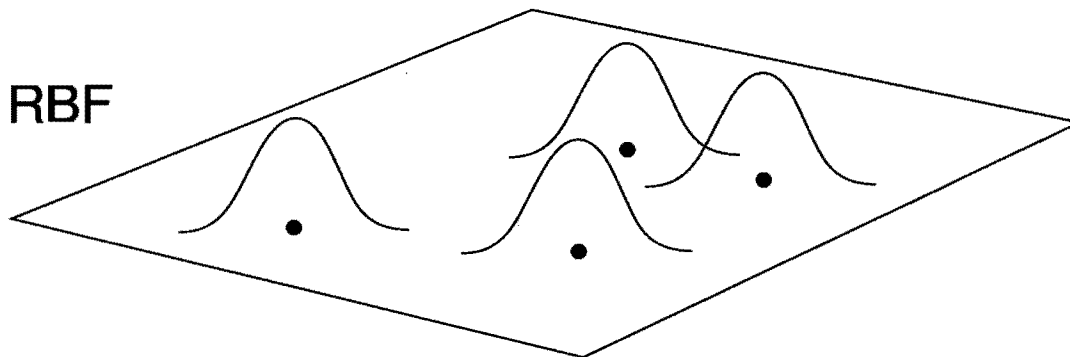
A Backpropagation Neural Network



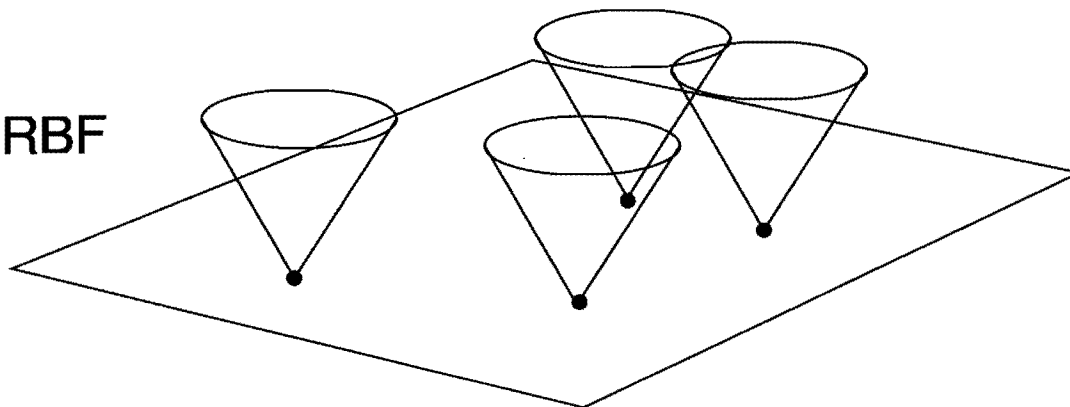
Radial Basis Functions

$$f(x) = \sum_i w_i g_i(x - x_i)$$

Gaussian RBF



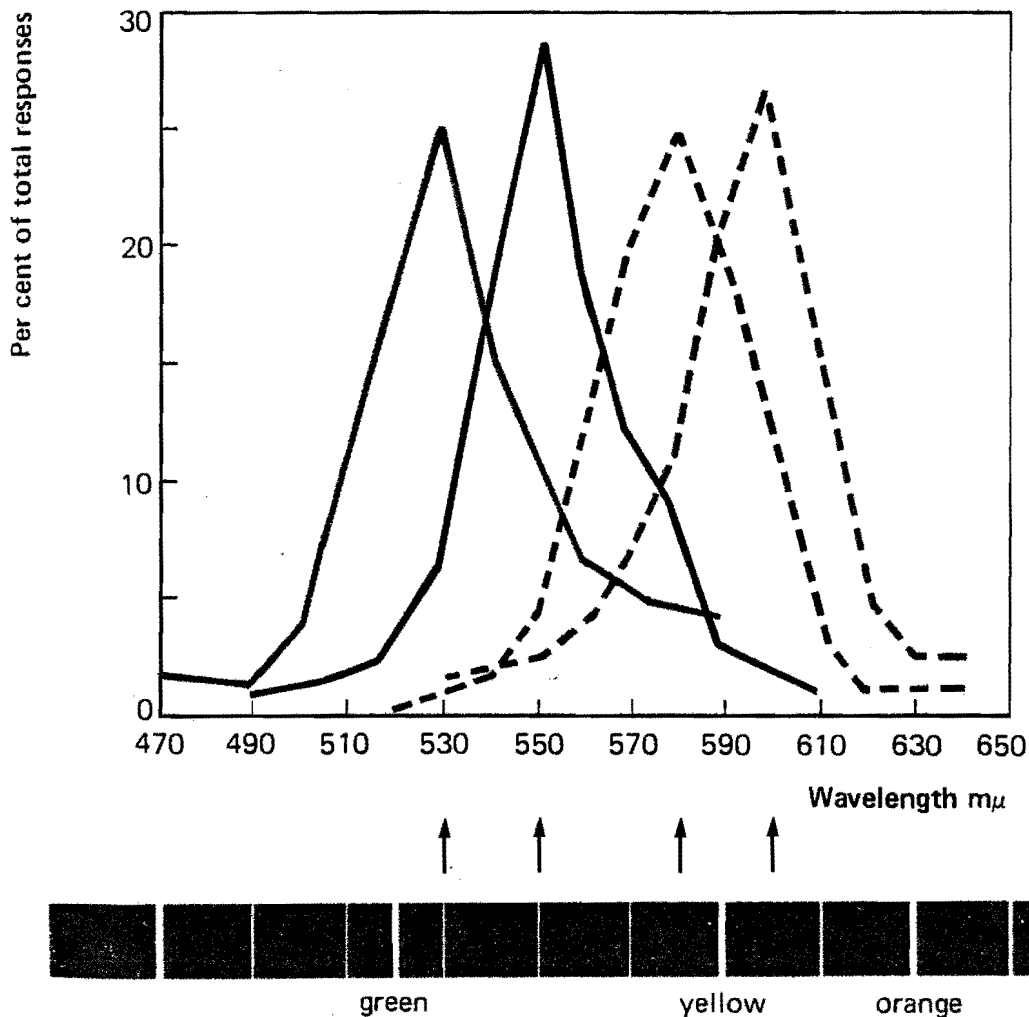
Linear RBF



Problems with Simple Learning Approaches

- Cognitively implausible: doesn't know what it knows.
(always thinks it has the full model)
- Incapable of one-shot learning. *(single examples rarely have a big impact)*
- Subject to inappropriate choice of model space.
(eg. too small to represent or approximate the true model)
- Subject to overfitting when the model space is too large.
(too many parameters for the amount of data)
- Can't allocate resources.
(may be large enough, but can't put model components where needed)
- Slow to learn. *(gradually varies parameters)*
- Liable to get stuck in local minima.
(model parts interfere with one another)
- Computationally expensive. *(full model is evaluated for each example)*
- No coherent underlying semantics. *(compositionality?)*

2.5 The phenomenon of stimulus generalisation appears to be a universal property of learnt responses. In the particular case illustrated, four groups of pigeons were rewarded for pecking at a disc on to which light of the four different wavelengths indicated by the arrows was projected. When the wavelength of light was changed, there was an orderly decline in the pigeons' rates of response: the greater the stimulus change, the greater was the decline.



opera glasses). Given what we know (but Watson did not know) about the maturation of certain fears, especially of animals, it is easy to accept Valentine's conclusion that certain stimuli may elicit a 'lurking fear in the background' (owing to innate factors) so that an added disturbance can bring out a full-blown fear reaction. In spite of the undoubted importance of classical conditioning, then, these observations suggest that we should be cautious in attributing all the fear-arousing properties of a particular stimulus to this form of learning.

From -Jeffrey Gray "The Psychology of Fear and Stress"²⁵

Human and Animal Induction

- Very different from the popular models.
- One shot learning: individual experiences can have a big impact.
- Episodic memory for specific events, especially when first learning about a domain. (each experience is precious)
- Initially generalize by similarity. (eg. Shepard's universal exponential law of generalization after one training example).
- As experience accumulates, build more complex models as warranted.
- Adaptive structure: complex models in one portion, simple models in others.
- Mostly avoid overfitting and getting stuck in local minima.
- Focus mechanisms to only access relevant information.
- Organism knows what it knows. (confidence in model).
- Humans are adaptive to a wide variety of domains.
- Some (eg. Anderson) claim Bayesian underlying semantics.

Newton's Euclidean space with an even more abstract four-dimensional Riemannian manifold (14).

Analogously in psychology, a law that is invariant across perceptual dimensions, modalities, individuals, and species may be attain-

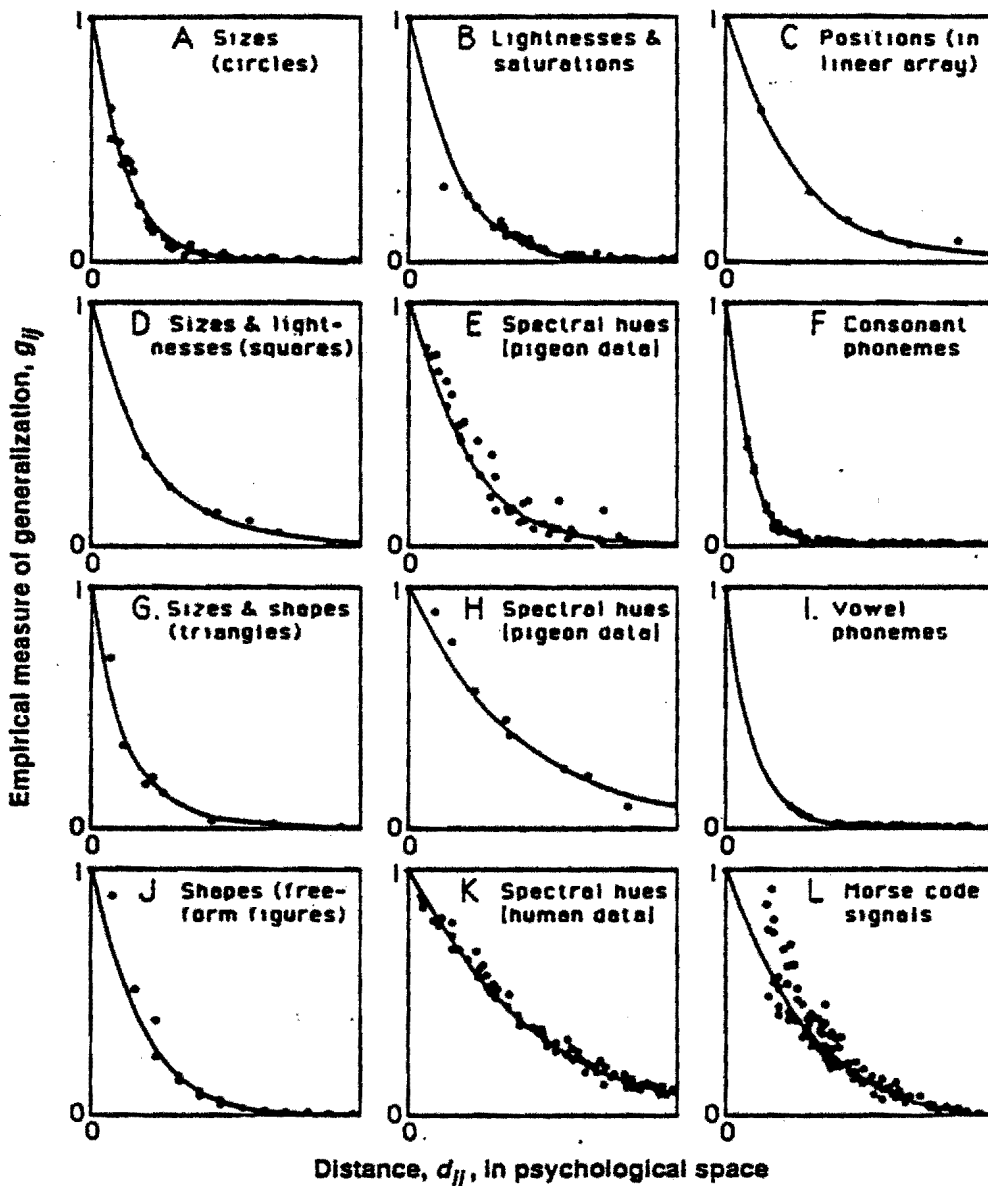


Fig. 1. Twelve gradients of generalization. Measures of generalization between stimuli are plotted against distances between corresponding points in the psychological space that renders the relation most nearly monotonic. Sources of the generalization data (g) and the distances (d) are as follows. (A) g , McGuire (33); d , Shepard (7, 18). (B) g , Shepard (7, 17); d , Shepard (7, 18). (C) g , Shepard (17); d , Shepard (8). (D) g , Attneave (25); d , Shepard (8). (E) g , Guttman and Kalish (4); d , Shepard (11). (F) g , Miller and Nicely (34); d , Shepard (35). (G) g , Attneave (25); d , Shepard (8). (H) g , Blough (36); d , Shepard (11). (I) g , Peterson and Barney (37); d , Shepard (35). (J) g and d , Shepard and Cermak (38). (K) g , Ekman (39); d , Shepard (18). (L) g , Rothkopf (40); d , Cunningham and Shepard (41). The generalization data in the bottom row are of a somewhat different type. [See (39) and the section "Limitations and Proposed Extensions."]

distances must satisfy (9, 15, 17): For each set of two most widely separated points, the distance between those two points to the

The uniqueness of the law is implicit in the following constraints (18, 19): Provided not too small relative to the rank order of the $n(n-1)/2$ a close approximation to the relation by an arbitrary set of points I found that for n dimensional space, distribution average correlation with the points, the correlation

The actual determination of the associated distances is achieved by multidimensional scaling and Kruskal (20) and scaling. In a specified representing the n stationary configurations defined measure of distance the generalization measures. Configurations can be defined in n dimensions, and even a previous representation is monotonicity is accepted measures g_{ij} against the configuration is interpreted psychologically rather than to be determined in the stimuli.

Intimations of

For a given set of experiment yields, the empirical estimate of stimulus i is made by the method is usually applied to generalization measures, g_{ij} normalization such a

Bayesian Approach

- Assume agents that must take actions based on uncertain data.
- Assume they can compare their preferences for states of the world.
- Under very mild assumptions: There exists a “prior” probability distribution and a “utility” function such that any rational agent behaves according to Bayesian decision making.
- If not, agent is subject to “dutch books” (*loses however things come out*).
- Procedure: Using prior and data compute posterior.
- Act to maximize posterior expected utility.
(*The expectation requires an expensive integral over the model space.*)

$$\operatorname{argmax}_a \int U(a, m) p(m) p(d|m) dm$$

- Need approximations to do the average.

Choice of Prior

- The prior defines the modelling language. The data overwhelms any non-zero prior, but a good choice speeds learning enormously.
- Prior for recognition is the posterior for learning.
- Prior for learning should be based on universal properties of physics (and maybe biology, sociology, psychology, etc).
- Bertrand Russell's "On Human Knowledge".
- **Time**
- **Continuity**
- **Sparseness**
- **Locality**
- **Natural kinds**
-

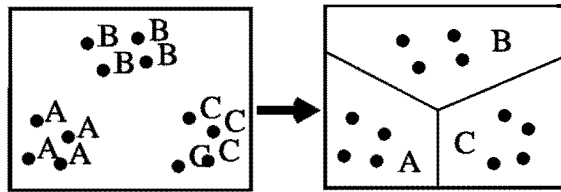
Time and Continuity

- **Time**: Expect the future to be like the past.

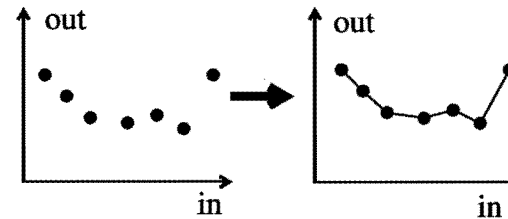
Without out this prior, induction is impossible.

- **Continuity**: Unless known to be otherwise, assume that nearby perceptions correspond to nearby states of the world.

Reflects the *geometric* nature of space.



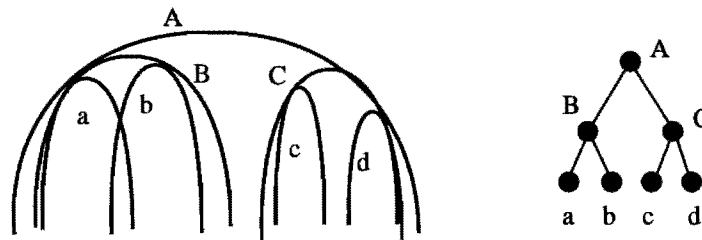
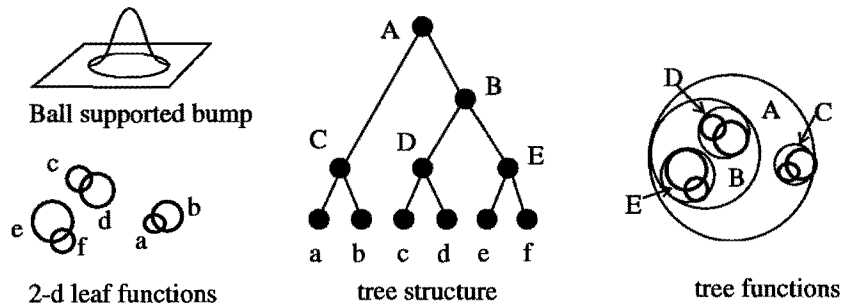
Classification Learning



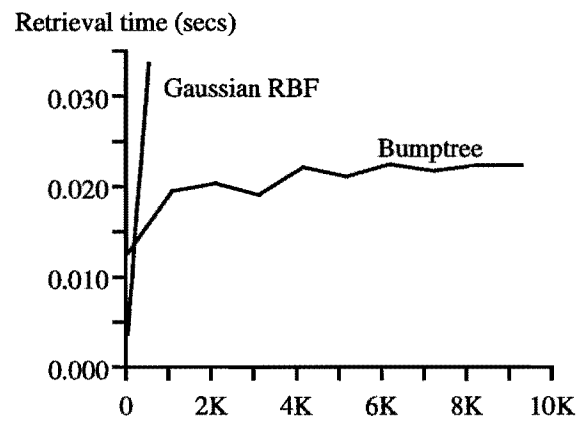
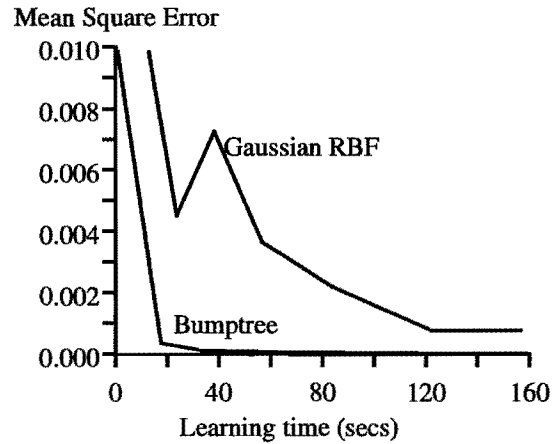
Mapping Learning

Bumptrees.

A *bumptree* is a complete binary tree with a function associated to each node such that an interior node's function bounds all leaf functions beneath it.



Bumptree learning and retrieval times



Sparseness and Locality

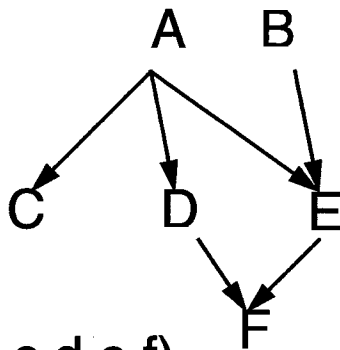
- **Sparseness**: The world is composed from component models which typically interact only sparsely with one another.
- **Locality**: Sensory data is composed of components which typically interact only sparsely with the components of the world.

Independence: X Y $p(x,y)=p(x)p(y)$

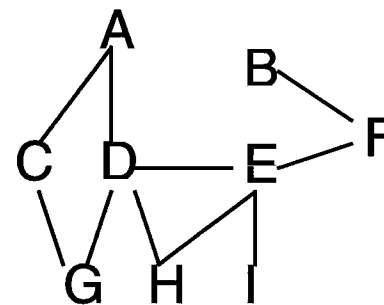
Conditional Independence: X—Y—Z $p(x,y,z)=p(x|y)p(y)p(z|y)$

Bayesian Networks and

Markov Networks:



$$p(a,b,c,d,e,f) = p(a)p(b)p(c|a)p(d|a)p(e|a,b)p(f|d,e)$$



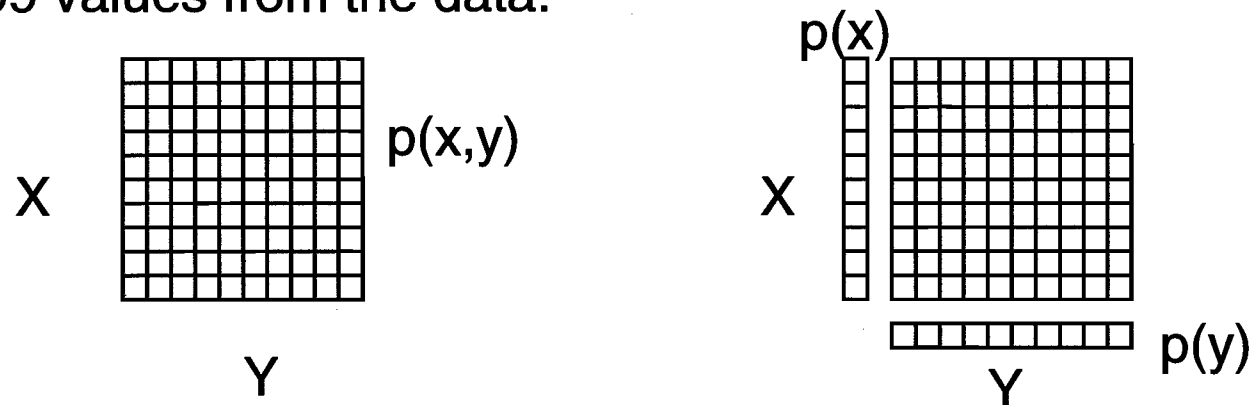
Fixing A and G makes C independent of B,D,E,F,H,I

Practical Success for Bayesian Networks

- Medical diagnosis: eg. anesthesia complications, heart failure, abdominal pain, etc.
- Diagnosis of plant disease
- Pharmaceutical approval
- Financial modeling
- Scheduling
- ...

The Power of Independence for Learning

Eg. to estimate $p(X,Y)$ where X and Y can each take 10 values, must estimate 99 values from the data:



If X and Y are independent: $p(x,y)=p(x)p(y)$, need only estimate 18 values.

And can determine that X and Y are independent with many fewer examples than would be needed to estimate $p(x,y)$.

Herskovits and Cooper learn the structure of a 37 variable ALARM Bayes net modelling anesthesia problems with 46 args with only 10,000 examples (with 2 extra and 2 missing arcs).

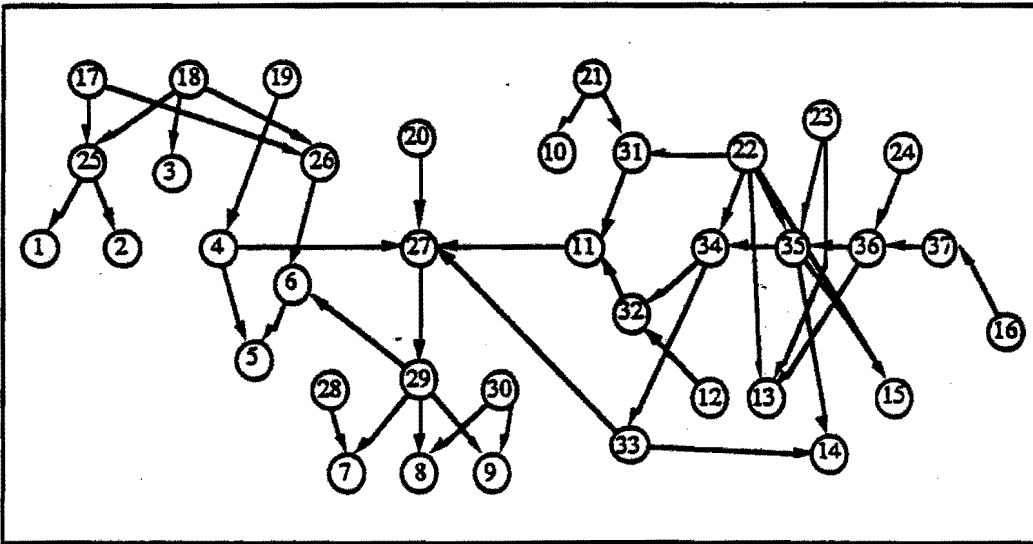


Figure 2 The ALARM belief network, with 37 nodes and 46 arcs.

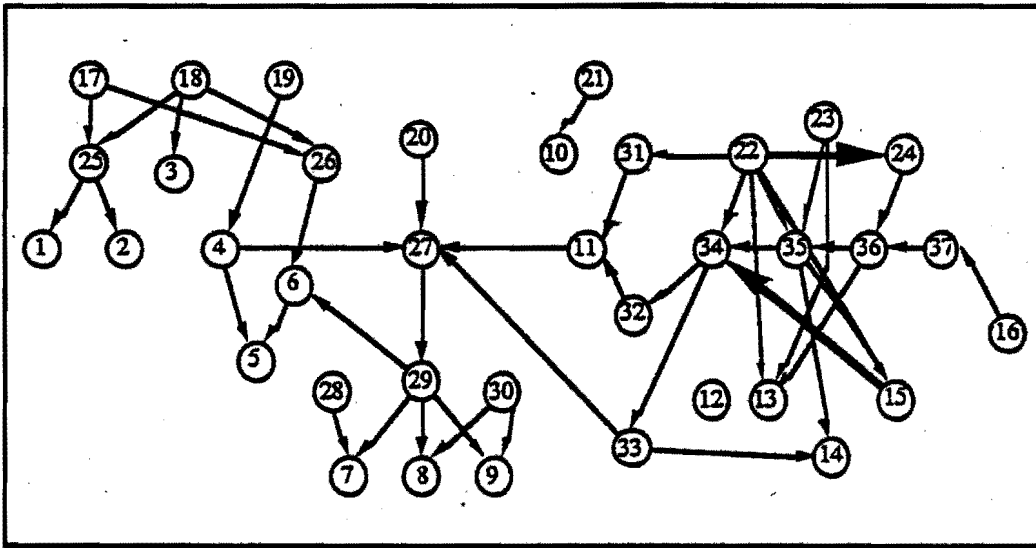


Figure 3 The ALARM network generated by Kutató from a 10,000-case database. The arcs from node 21 to node 31, and from node 12 to node 32 are missing, and extra arcs (bold) from node 15 to node 34 and from node 22 to node 24 have been added.

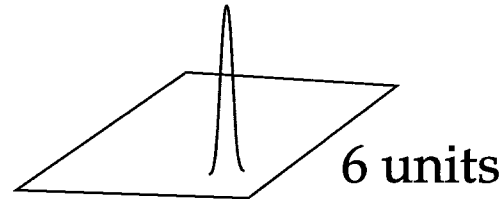
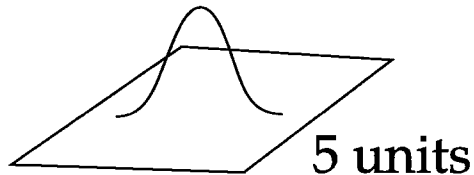
From Herskovits & Cooper "Kutató"
 ALARM net from Beinlich to model anesthesia problems.
 37 nodes, 46 arcs.

Stochastic Grammars: Dynamic Independence

- Bayesian and Markov networks are based on a *fixed* underlying graph.
- The conditional independence relationships are not data dependent.
- This is sufficient for domains like medical diagnosis but is not rich enough for vision or natural language. Also not rich enough for learning, eg. distribution over Bayes nets is not described by a Bayes net.
- Stochastic grammars are the simplest form of probability distribution over dynamic structures. Conditional independence is data dependent.
- Stochastic regular grammars: Hidden markov models
- Stochastic context free grammars: Each production has an associated probability. Probability of a parse tree is the product of the probabilities of the productions.

Overfitting and Occam's Razor

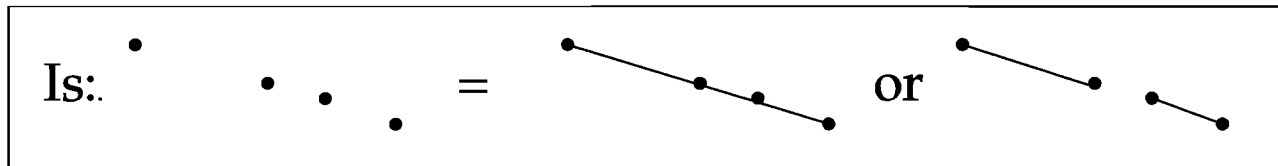
Poor approximations to the posterior average lead to overfitting. Eg. Choosing the maximum a posteriori probability model (MAP):



Choose between a net with 5 hidden units versus one with 6.

6 might give a better fit and so have higher posterior, but because the space is larger, the integral of the posterior might be larger over 5. So MAP picks the wrong dimension. "Overfitting"

Same issues apply to perception as learning:



Avoiding Overfitting

- Bayesian Occam Factor (*Approximate posterior integral by Gaussian approximation at peaks, automatically prefers lower dimensions*)
- Vapnik-Chervonenkis Dimension
- Akaike Information Criterion
- Rissanen's Minimum Description Length Criterion
- Weight decay and complexity terms in error functions
- Cross validation, bootstrap, jackknife methods
- ...

Vapnik and Nested Families

- Vapnik computes number of samples needed to get bounds on error in selecting distributions from families.
- The larger the family, the more samples are needed in general.
- If we choose too small a family, we may not be able to fit the data, if too large, we may need too much data to validate the choice.

- Vapnik suggests that we work with a nested family of models, trying to fit in the small families first, working our way up.
- The number of samples needed to validate a model depends only on the size of the space it is from, not on the size we potentially could have gone to.

Best First Model Merging

- Small amount of data -> Remember and use similarity
Large amount of data -> Fit models and find regularities
- Build complex model from simple model components.
- Start with one component per example.
- Successively merge models always doing the best first.
- Stop when error is not made up for in reduction of complexity.
- The merged models may be from a more complex class than the merged components because there is more data available.
- eg. Density estimation: Gaussian mixture kernel estimate, successively more complex components.
- eg. Mapping learning: Start with local affine fits, move to quadratic, etc.
- eg. Grammar learning: Start with products, move to regular, context free, etc.

Approximating Curves and Surfaces

- Use local models to approximate curves and surfaces.
- Eg. find best set of segments approximating a given curve.

Merge the pair of segments which increases the error the least repeatedly until an error criterion is reached.

- This best first approach does a good job of “homing” in on the “corners” and makes ambiguous decisions only after all the clear decisions have been made.
- The class of the merged model may be from a more complex family since there is more data to provide evidence.
- eg. A pair of affine segments may be fit by a quadratic.

Best first merging of segments to fit a curve



curve

err 0

err 1

err 2

err 3

err 4

err 5

err 10

err 20

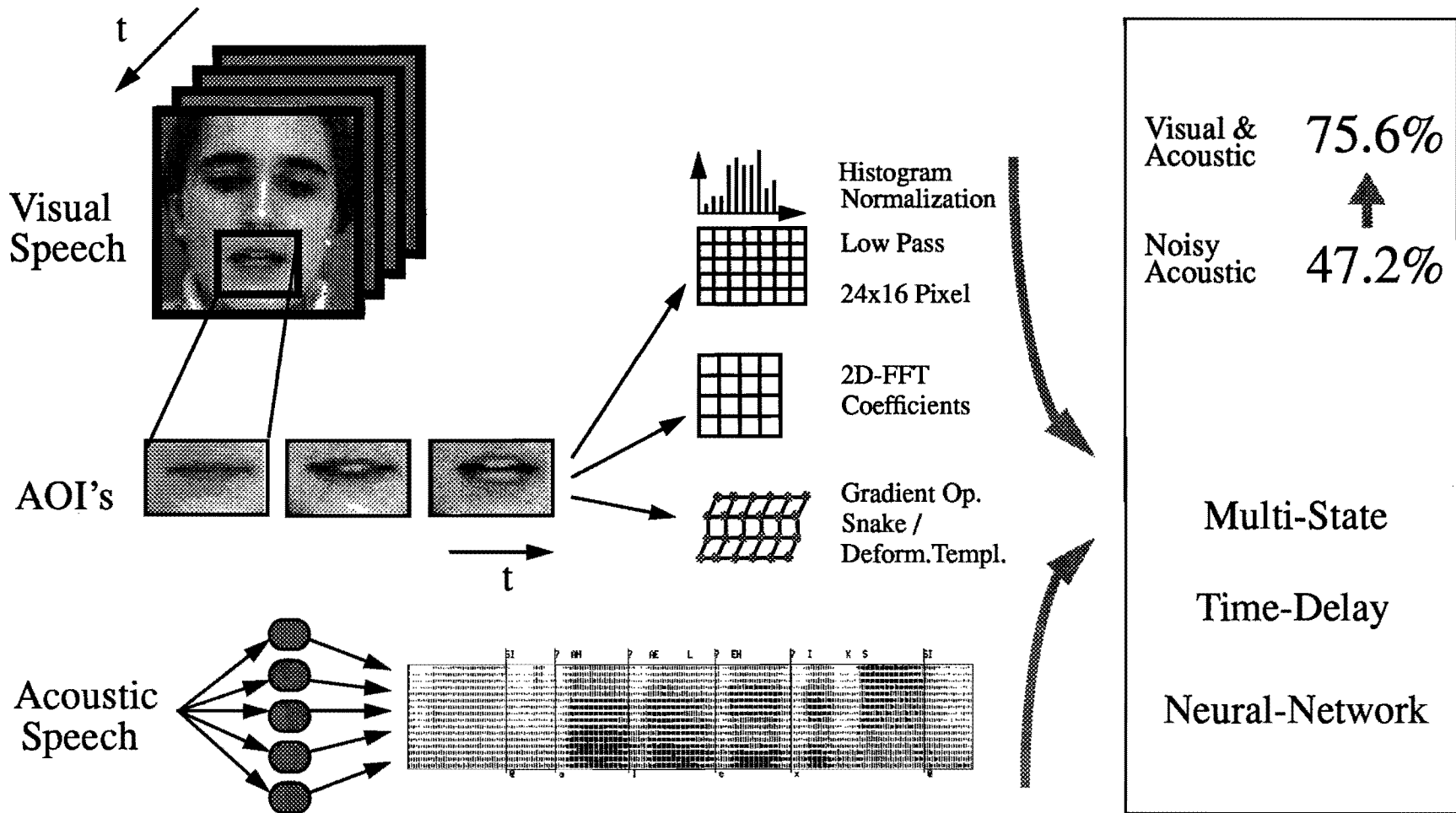
err 50

err 100

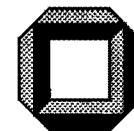


Improving Connected Letter Recognition by Lipreading

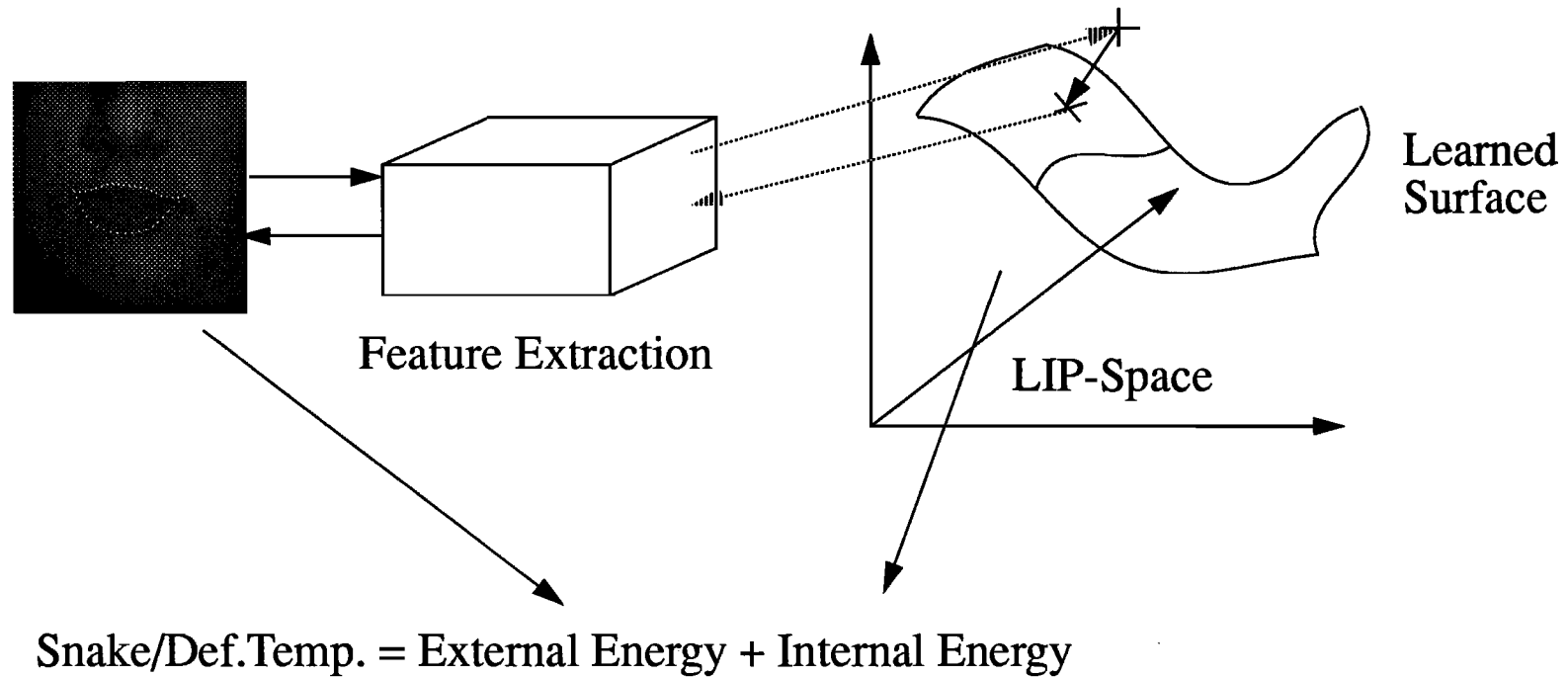
Christoph Bregler, Hermann Hild, Stefan Manke, and Alex Waibel



Multimodal Speech Recognition



Deformable Templates & Surface Learning¹



1. By Steve Omohundro, ICSI, Berkeley

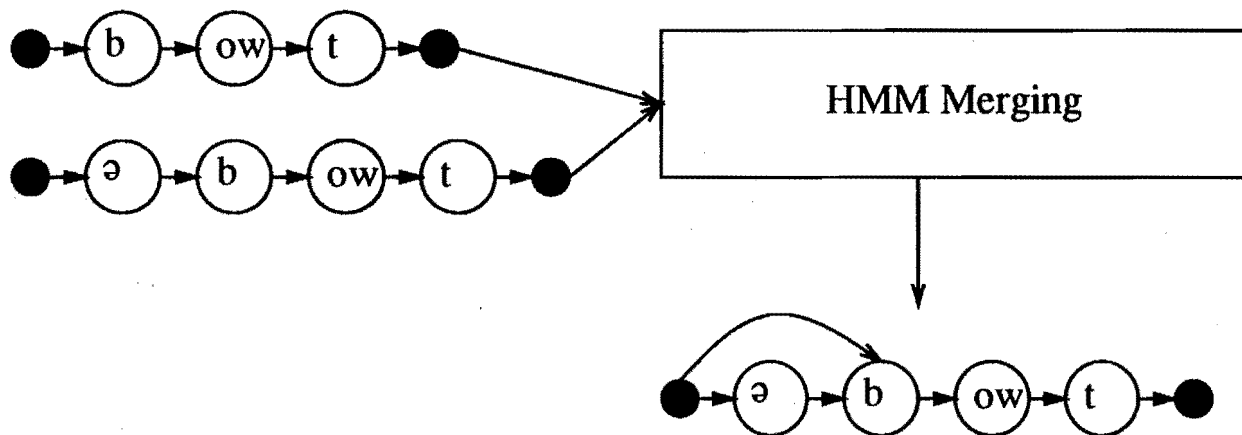
Learning Stochastic Grammars

- Stochastic grammar = Probability distribution on strings.
- Stochastic regular languages = Hidden Markov Models specified by states, transition probabilities and emission probabilities. *(Used extensively in speech recognition and cryptography)*
- MDL gives a natural prior. Maximum likelihood model has a state for each symbol in each example string. Merging states lowers likelihood but raises prior. Repeatedly merge best first until posterior peaks.
- Stochastic context free grammars *(Becoming important in natural language and vision)*
- Need both to merge nonterminals and “chunk” them into new non-terminals. Maximum likelihood model has a rule for each example. Merging and chunking lower likelihood but raise prior. Repeatedly merge best-first until posterior peaks.
- Beats the Baum-Welch E-M based approach.

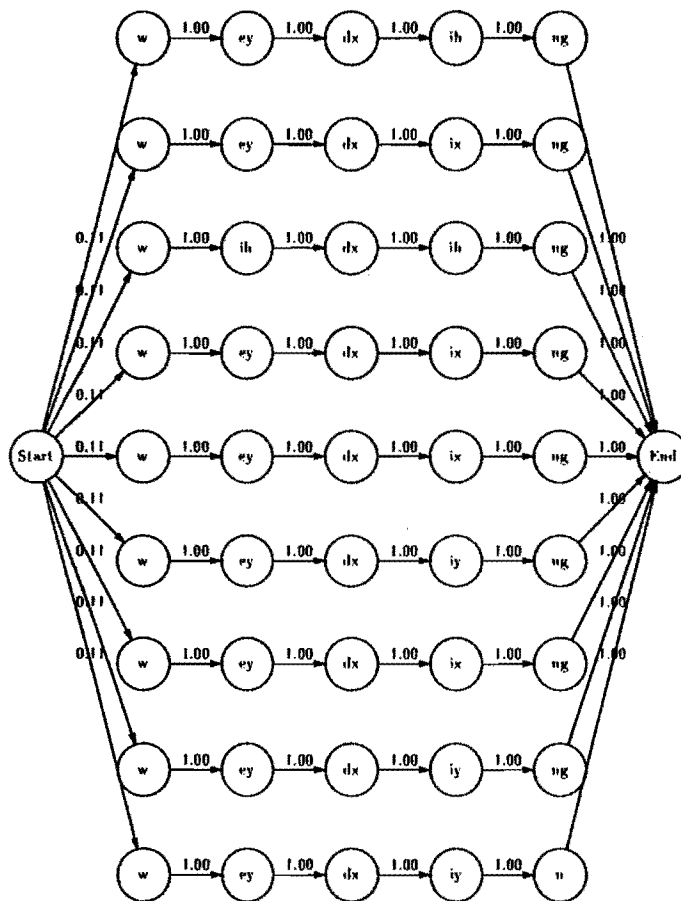
HMM Merging

(Work by Stolcke & Omohundro)

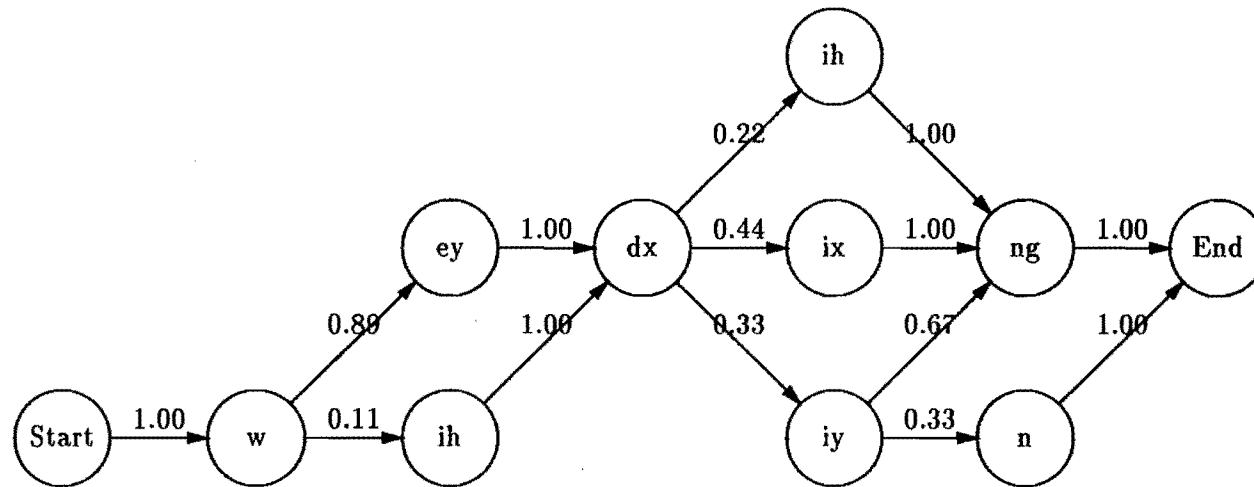
"about"



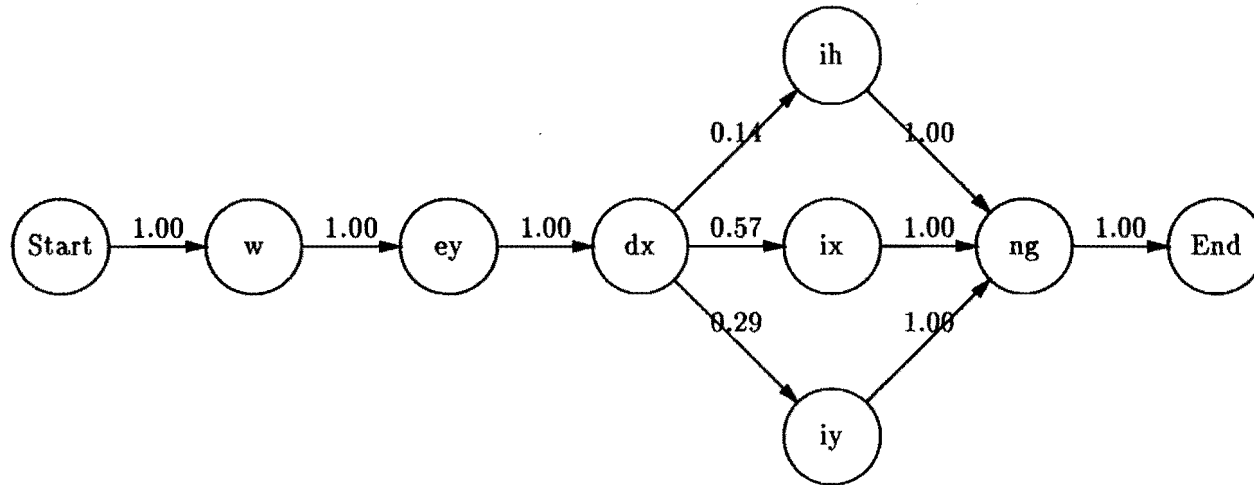
Unmerged HMM for "waiting"



Merged but unpruned HMM for "waiting"



Pruned HMM for "waiting"



Conclusions

A synthesis of the best of symbolic AI and connectionism has great potential. Such an approach will have:

- A coherent underlying semantics (*like logic and Bayesian statistics*)
- Support evidential reasoning (*like neural nets and probabilistic models*)
- Support geometric and perceptual knowledge (*like neural nets and geometric models*)
- Support dynamic structures (*like grammars and logic-based formalisms*)
- Support quantitative learning (*like neural nets using parameter optimization*)
- Support structural learning (*using operations like model merging*)
- Be computationally efficient (*using data structures like bumptrees*)