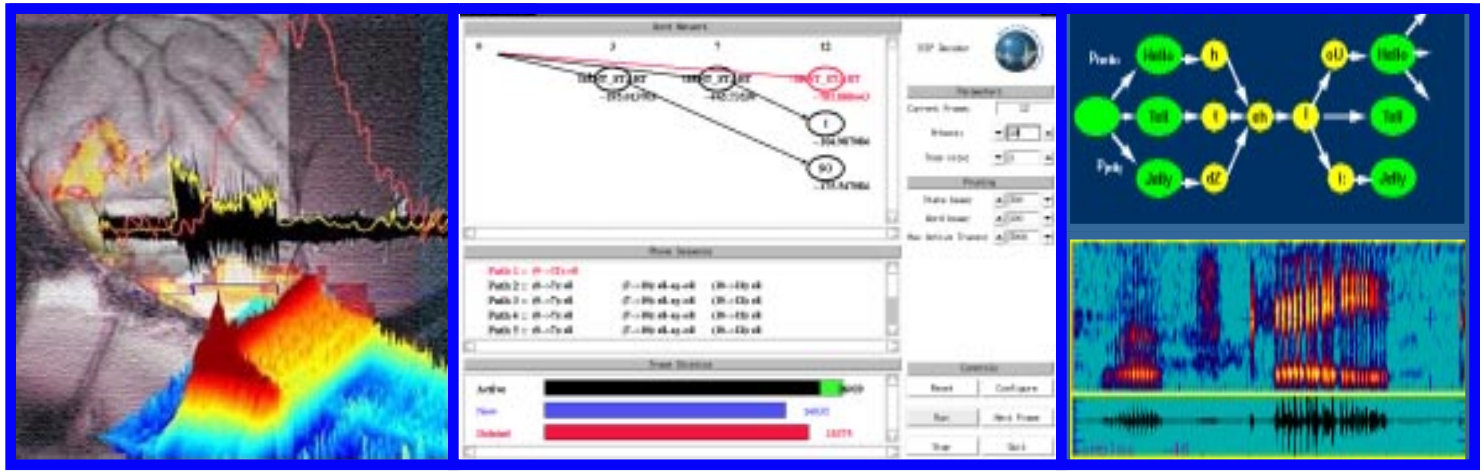
SPEECH RECOGNITION: REASONING UNDER UNCERTAINTY

Joseph Picone and Jon Hamaker
Institute for Signal and Information Processing
Mississippi State University

Aravind Ganapathiraju
Speech Scientist
Conversay Computing Corporation



- Contact Information:
  Box 9571
  Mississippi State University
  Mississippi State. Mississippi 39762
  Tel: 662-325-3149, Fax: 662-325-2298
  Email: picone@isip.msstate.edu

- This material is based upon work supported by the National Science Foundation under Grant No. IIS0095940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

This talk is available at:
http://www.isip.msstate.edu/publications/seminars/external/2002/utd_hlt

INTRODUCTION

ABSTRACT:

Modern speech recognition systems, heavily based in statistical methods, are testaments to how we can exploit complex and powerful computers to provide useful engineering solutions to real world problems. Yet, these systems are extremely primitive compared to how humans learn to recognize and understand speech. In this talk, we will review state of the art in speech recognition, and describe a new generation of technology based on principles of discrimination and risk minimization. We will show that an implementation of this approach based on Support Vector Machines resulted in a 10% reduction in word error rate on a small vocabulary task.We will describe on-going research focused on making such techniques more feasible for large scale tasks and developing more robust parameter estimation techniques.

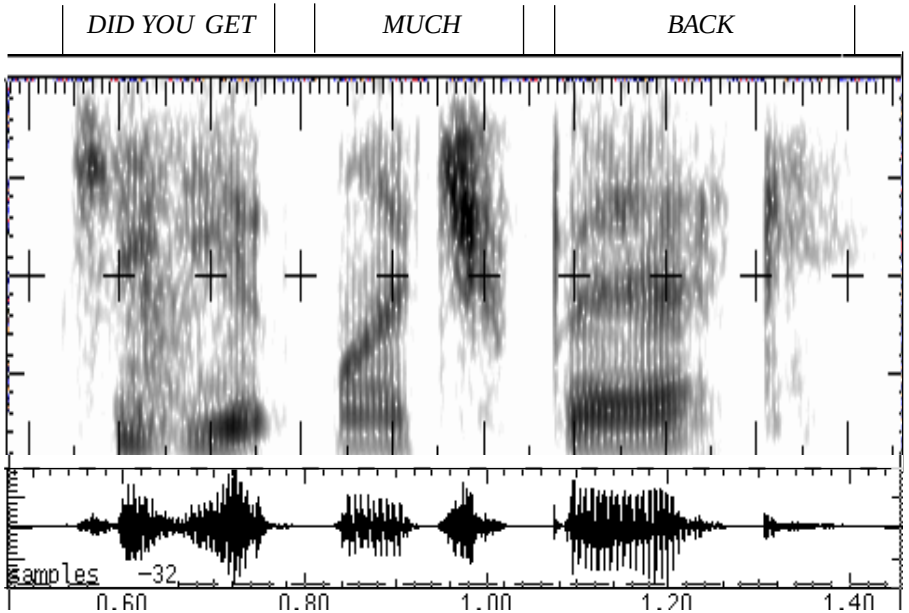
BIOGRAPHY:

Joseph Picone is currently a Professor in the Department of Electrical and Computer Engineering at Mississippi State University, where he also directs the Institute for Signal and Information Processing. He has previously been employed by Texas Instruments and AT&T Bell Laboratories. Dr. Picone received his Ph.D. in Electrical Engineering from Illinois Institute of Technology in 1983. He is a Senior Member of the IEEE and a registered Professional Engineer.

INTRODUCTION

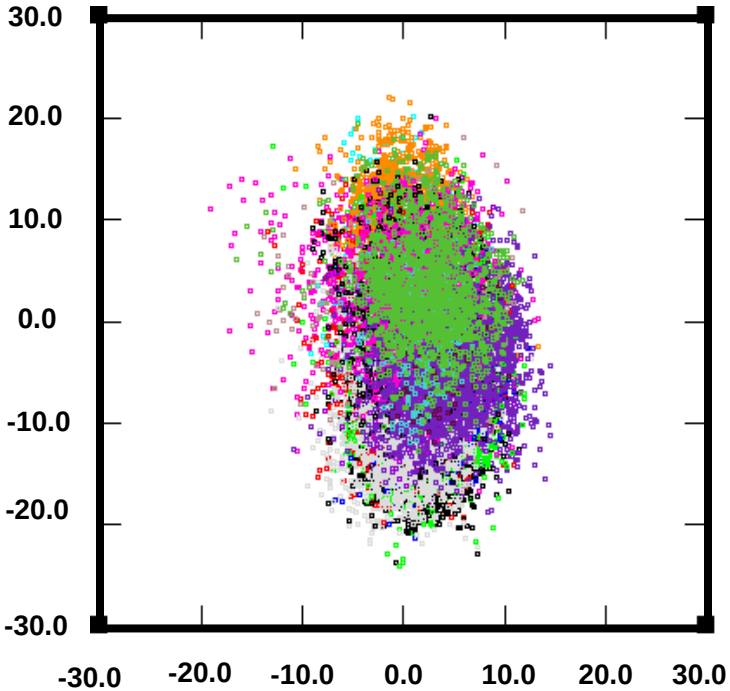
Two fundamental challenges for modern speech recognition systems:

Pronunciation Modeling:



- “Did” is reduced and merged into “you get” such that the resulting word is pronounced “jyuge.”
- Deletion rate for phonemes: ~12%
  Deletion rate for syllables: ~1%
- Syllables are a promising acoustic unit.

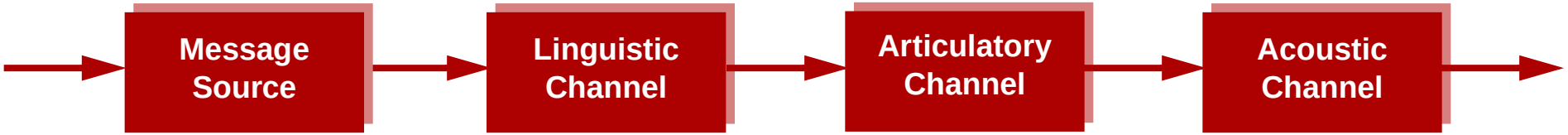
Acoustic Confusability:



- First two cepstral coefficients for all vowels (based on a conversational speech corpus — SWITCHBOARD).
- Overlap represents a fundamental barrier for good classification.

STATE OF THE ART

A noisy communication channel model for speech production and perception:



Observable: Message | Words | Phones | Features

Bayesian formulation for speech recognition:

$$P(W|A) = P(A|W)P(W)/P(A)$$

Objective: minimize the word error rate by maximizing $P(W|A)$

Approach: maximize $P(A|W)$ (training)

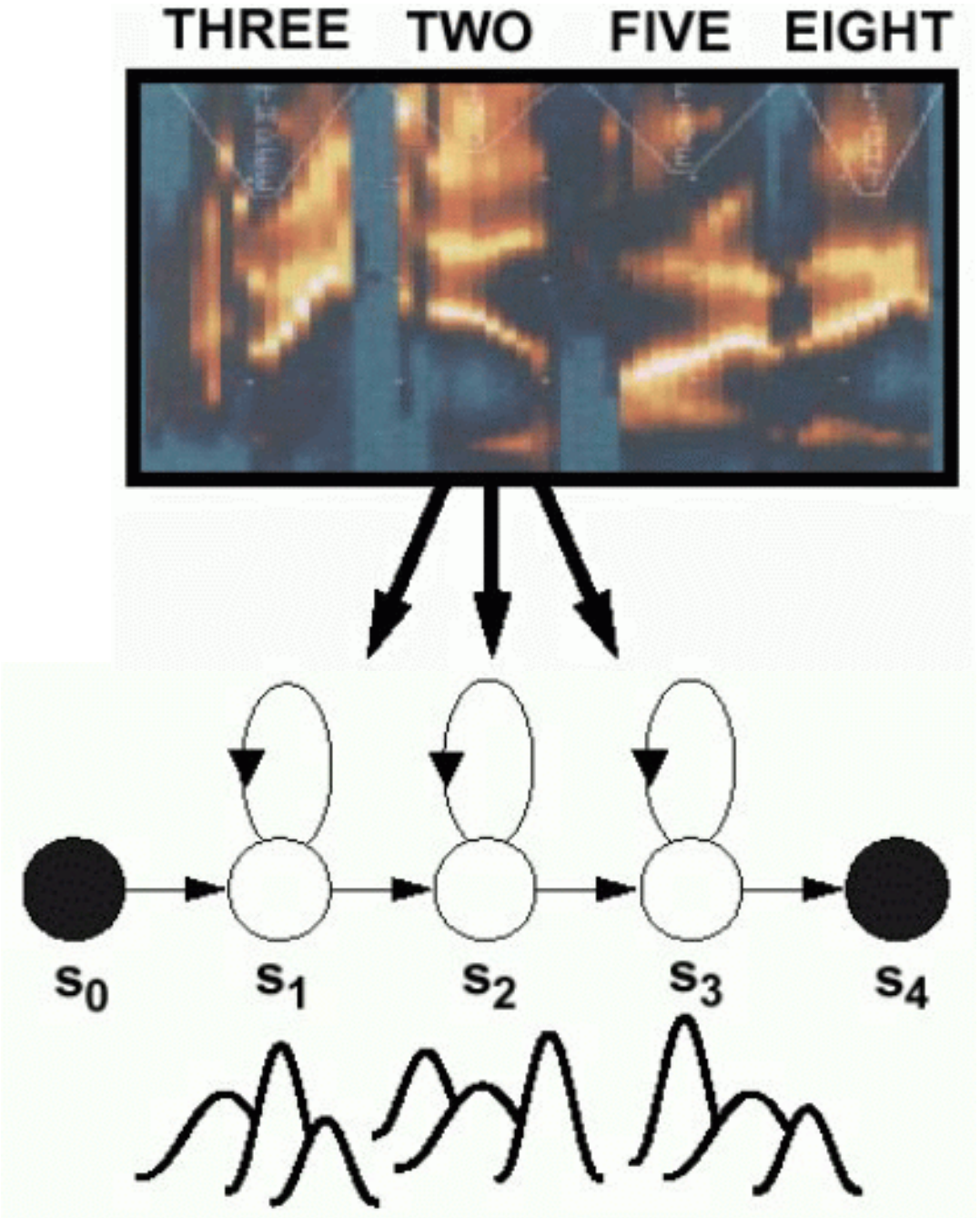
Components:

- $P(A|W)$: acoustic model (hidden Markov models, mixture of Gaussians)
- $P(W)$: language model (statistical, N-grams, finite state networks)
- $P(A)$: acoustics (ignore during maximization)

The language model typically predicts a small set of next words based on knowledge of a finite number of previous words (N-grams) — leads to search space reduction.

STATE OF THE ART

- Acoustic models encode the temporal evolution of the features (spectrum).
- Gaussian mixture distributions are used to account for variations in speaker, accent, and pronunciation.
- Sharing model parameters is a common strategy to reduce complexity.
- The goal of our research is to replace the Gaussian likelihood computation at each state with a machine that incorporates notions of:
  - **discrimination ("one vs. all")**
  - **Bayesian statistics (priors)**
  - **confidence**
  - **sparsity**
- Maintain computational efficiency?



STATE OF THE ART

- Data-driven modeling supervised only from a word-level transcription.
- The expectation/maximization (EM) algorithm is used to improve our estimates:

$$\log P(Data|\bar{\lambda}) \geq \log P(Data|\lambda)$$

if:

$$Q(\lambda, \bar{\lambda}) \geq Q(\lambda, \lambda)$$

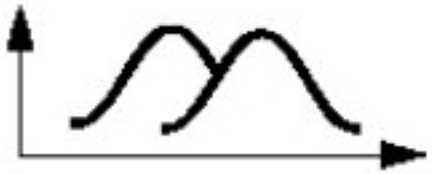
Approach: maximum likelihood estimation

- Computationally efficient training algorithms (Forward-Backward) have been crucial.
- Batch mode parameter updates are typically preferred.
- Decision trees are used to optimize sharing parameters, minimize system complexity, and integrate additional linguistic knowledge.

- Initialization

- Single Gaussian Estimation

- 2-Way Split

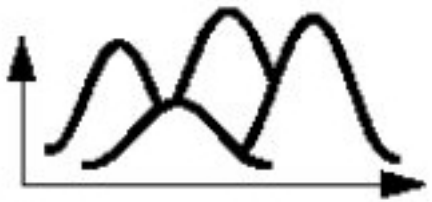


- Mixture Distribution Reestimation

- 4-Way Split

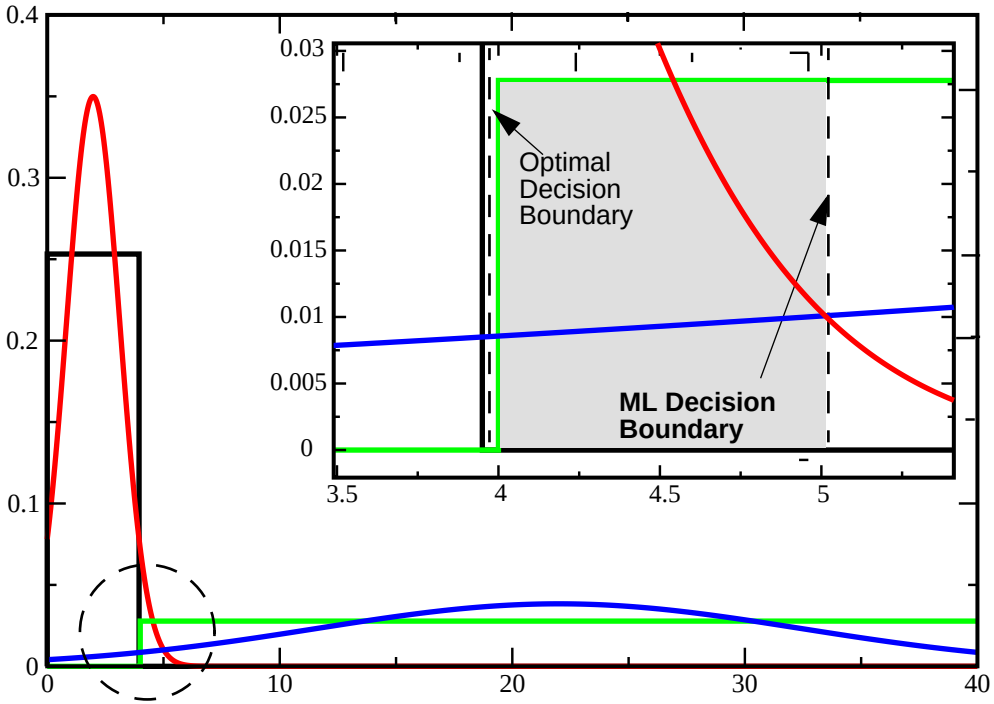
- Reestimation

...



RISK MINIMIZATION

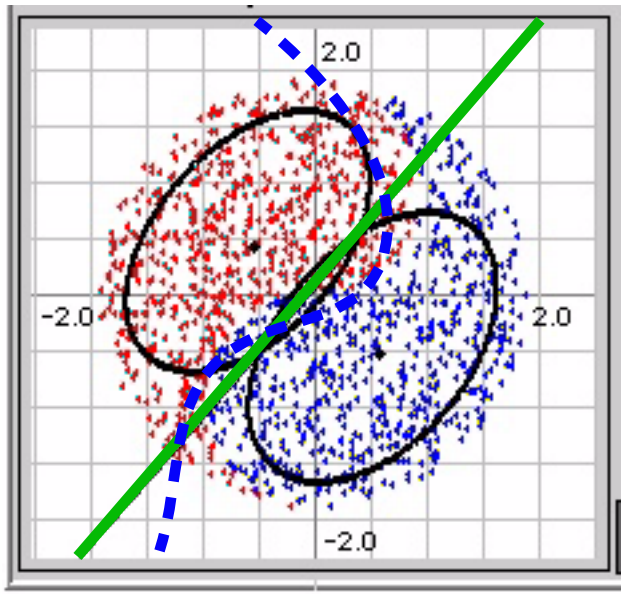
Convergence in maximum likelihood does not translate to optimal classification:



- Error results from fitting uniform distributions with Gaussians (and using an ML boundary).
- Since the classes are separable, finding the optimal decision surface is trivial.

Discrimination

Class-Dependent PCA (MLE Gaussian)



- Data not separable by a hyperplane (a nonlinear classifier is needed).
- Gaussian MLE models tend towards the center of mass (overtraining).

Solution: must balance representation and discrimination in a common framework (rather than alternating between the two in a bootstrapped training procedure).

RISK MINIMIZATION

- Expected Risk:

$$R(\alpha) = \int \frac{1}{2} |y - f(x, \alpha)| \, dP(x, y)$$

Not possible to estimate $P(x, y)$.

- Empirical Risk Minimization:

$$R_{emp} = \frac{1}{2l} \sum_{i=1}^{l} |y_i - f(x_i, \alpha)|$$

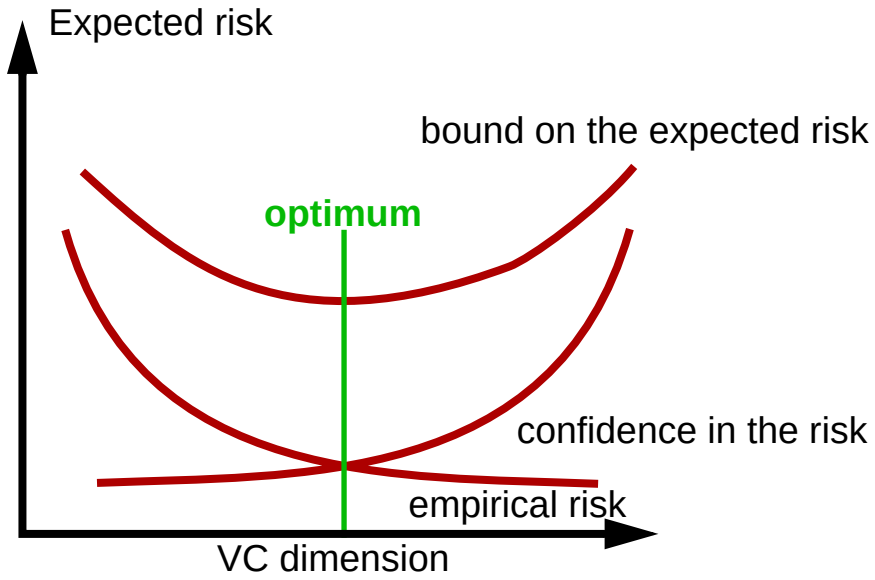
- Related by VC (Vapnik-Chervonenkis) dimension:

$$\boldsymbol{R(\alpha) \leq R_{emp}(\alpha) + f(h)}$$

$$f(h) = \sqrt{\frac{h(\log((2l/h) + 1)) - \log(\eta/4)}{l}}$$

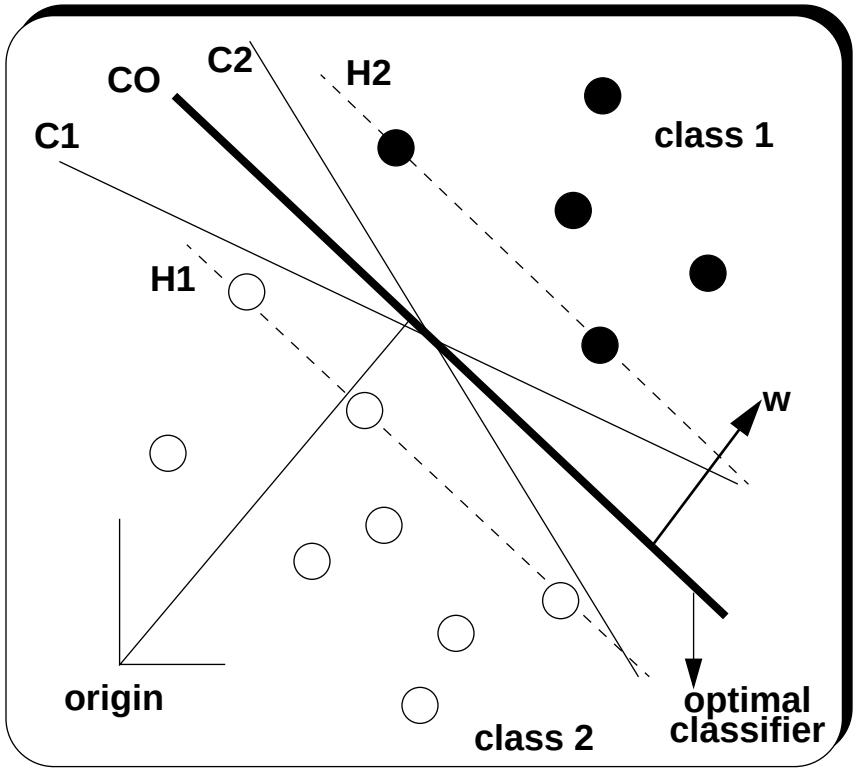
$f(h)$ is referred to as the VC confidence, η is a confidence measure $(0 \leq \eta \leq 1)$.

- Approach: **choose the machine that gives the least upper bound on actual risk**



- The VC dimension, h is a measure of the capacity of the learning machine.
- Principle of structural risk minimization (SRM) (Vapnik, 1979) involves finding the subset of functions that minimizes the bound on the actual risk.
- Optimal hyperplane classifiers achieve zero empirical risk for linearly separable data.

SUPPORT VECTOR MACHINES



- Hyperplanes C0-C2 achieve perfect classification — zero empirical risk.
- C0 is optimal in terms of generalization.
- The data points that define the boundary are called support vectors.

Optimization (Separable Data)

- Hyperplane: $\boldsymbol{x} \cdot \boldsymbol{w} + b$
- Constraints:

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1 \geq 0 \qquad \forall i$$

The data points that satisfy the equality are called support vectors.

- Optimize:

$$L_P = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) + \sum_{i=1}^{N} \alpha_i$$

- Minimization of this Lagrange functional minimizes risk criterion (maximizes margin).
- Final classifier:

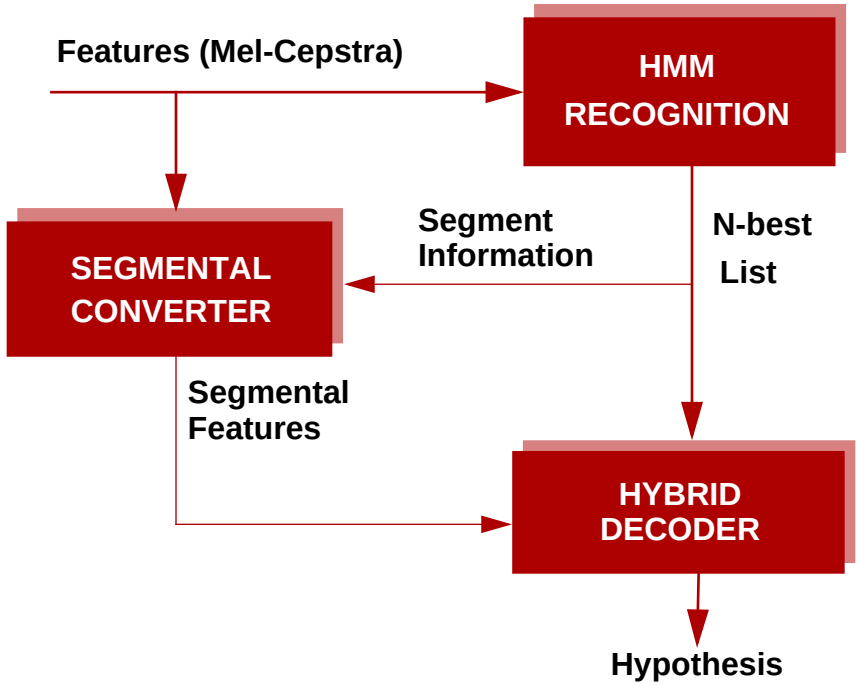
$$f(\boldsymbol{x}) = \sum_{i=1}^{numSVs} \alpha_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b$$

SUPPORT VECTOR MACHINES

- Experimental Results: Deterding Vowel (11 vowels spoken in “h*d” context)

Approach	Error Rate
K-Nearest Neighbor	44%
Gaussian Node Network	44%
SVM: Polynomial Kernels	49%
SVM: RBF Kernels	35%
Separable Mixture Models	30%
RVM: RBF Kernels	30%

- A Hybrid Speech Recognition Framework



- Experimental Results: Continuous Speech

Information Source		HMM		Hybrid	
Transcription	Segmentation	AD	SWB	AD	SWB
N-best	Hypothesis	11.9	41.6	11.0	40.6
N-best	N-best	12.0	42.3	11.8	42.1
N-best + Ref.	Reference	—	—	3.3	5.8
N-best + Ref.	N-best + Ref.	11.9	38.6	9.1	38.1

- Rescore N-best lists using phone classifiers
- Use a segmental modeling approach for phone classifiers
- 10.6% on AD task using hybrid system that combines HMM and SVM scores

RELEVANCE VECTOR MACHINES

Drawbacks of SVMs:

- Complexity scales linearly with the training data for nontrivial problems (prohibitive for large speech recognition tasks).
- Sparsity of the model should be explicit in the optimization of the model.
- Need a posterior probability, not distance.
- The sigmoid approximation tends to overestimate confidence (Tipping).

Relevance Vector Machines:

- A kernel-based learning technique.
- A Bayesian approach (MacKay) that incorporates an automatic relevance determination (ARD) prior over each model parameter.
- RVMs typically require an order of magnitude less parameters that SVMs, but require significantly more training time.

- As with SVMs, the RVMs are formed by defining a vector-to-scalar mapping:

$$y(\boldsymbol{o};\boldsymbol{w}) = w_o + \sum_{i=1}^{M} w_i \phi_i(\boldsymbol{o}) = \boldsymbol{w}^T \phi(\boldsymbol{o})$$

- RVMs take a Bayesian approach and explicitly define an ARD prior distribution over the weights:

$$p(\boldsymbol{w}|\alpha) = \prod_{i=0}^{N} N\left(w_i|0, \frac{1}{\alpha_i}\right) = \frac{1}{\sqrt{(2\pi)^{N+1}|\boldsymbol{A}^{-1}|}} e^{-\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w}}$$

- To complete the Bayesian specification of the model, we use a non-informative (flat) prior for α_i.
- The likelihood of the training data set can be written as:

$$P(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{O}) = \prod_{n=1}^{N} \sigma_n^{t_n} (1-\sigma_n)^{1-t_n}$$

where $\sigma_n = \sigma\{y(\boldsymbol{o}_n;\boldsymbol{w})\}$.

COMPARISON

Support Vector Machines

Data:

Class labels: {-1,+1}; “one vs. all”

Goal:

Find decision surface that maximizes the margin between two classes

Training:

Adjust parameters under constraint:

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \geq 1 - \xi_i \qquad \forall i$$

Optimize:

$$L_P = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) + \sum_{i=1}^{N} \alpha_i$$

Training Complexity: $O(N^2)$

Classification: Threshold decoding (0.0)

Decoding:

- Rescoring N-best lists
- Segmental models

Relevance Vector Machines

Data:

Class labels: {0,1}; “one vs. all”

Goal: Learn posterior, $P(t|\boldsymbol{x})$.

Training:

$$y(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$

$$P(t|\boldsymbol{x}) = \frac{1}{1 + e^{-y(\boldsymbol{x})}}$$

$$p(\boldsymbol{w}|\alpha) = \prod_{i=0}^{N} N\left(w_i|(\mu_i = 0), \frac{1}{\alpha_i}\right)$$

find: $\underset{\bar{\boldsymbol{w}},\, \bar{\alpha}}{\operatorname{argmax}} P(\boldsymbol{w},\alpha|[t],[x])$

iteratively find $\hat{\boldsymbol{w}}|\alpha$ then $\hat{\alpha}|\hat{\boldsymbol{w}}$.

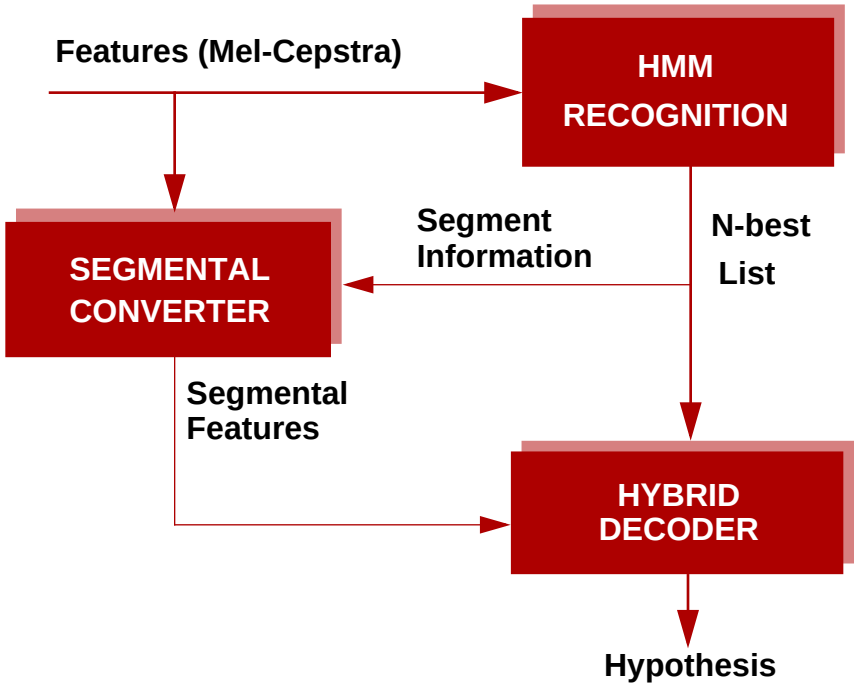
Training Complexity: $O(N^3)$

Classification: Threshold decoding (0.5)

Decoding: Integrated likelihood computation

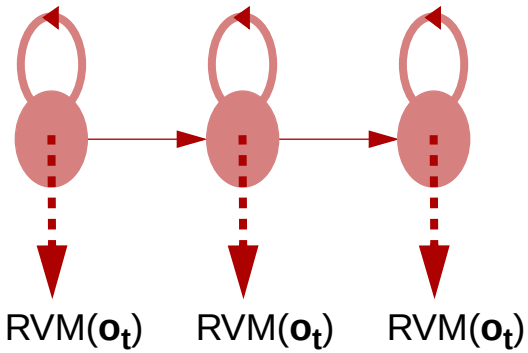
RELEVANCE VECTOR MACHINES

- First Attempt: A Hybrid Approach

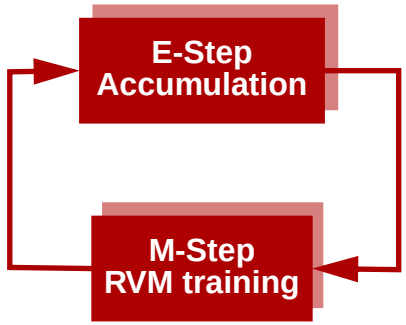


- Use HMM system to generate segmentations and phone hypotheses
- Use RVM to rescore phone hypotheses
- Problem: as with SVMs, RVMs not exposed to alternate segmentations

- Second Attempt (Under Development):



HMMs with RVM Emission Distributions



Iterative Parameter Estimation Similar to Baum-Welch Training of HMMs

- Convergence properties and efficient training methods are critical.
- Bootstrapping or incremental training
- Available as part of the ISIP speech recognition toolkit.

RELEVANCE VECTOR MACHINES

- Experimental Results: Deterding Vowel (11 vowels spoken in “h*d” context)

Approach	Error Rate
K-Nearest Neighbor	44%
Gaussian Node Network	44%
SVM: Polynomial Kernels	49%
SVM: RBF Kernels	35%
Separable Mixture Models	30%
RVM: RBF Kernels	30%

Approach	Avg. Parameter Count
SVM: RBF Kernels	83 SVs
RVM: RBF Kernels	13 RVs

- RVMs yield superior sparsity with comparable generalization.

- Experimental Results: OGI Alphadigits (telephone bandwidth letters and numbers)

Approach	Error Rate	Avg. Parameter Count	Training Time	Testing Time
SVM	16.4%	257 SVs	1/2 hour	30 mins
RVM	16.2%	12 RVs	1 month	1 min

- Hybrid RVM system is mirror of hybrid SVM system (still has segmentation problem).
- Reduced training set size (2000 examples per phone class).
- RVM yields a large reduction in parameter count — translates to large efficiency boost for decoder.
- Computational cost mainly in training, but is still prohibitive for large data sets.

CONCLUSIONS

- Principles of structural risk minimization have been applied to speech recognition. Performance on pilot experiments such as the OGI Alphadigits corpus are promising. Oracle experiments indicate tremendous potential.

- Support Vector Machines (SVMs) can reduce system complexity by reducing parameter counts over traditional hidden Markov model-based systems. However, SVM complexity is still high for large speech recognition tasks.

- Relevance Vector Machines (RVMs) were introduced as a way of reducing overall system complexity, but maintaining high levels of performance. Preliminary results are promising.

- Though these techniques are conceptually attractive, there are many theoretical and computational issues to be explored before this technology will be as ubiquitous as HMMs. Perhaps the two most important issues are:

  $\Rightarrow$ Closed-loop training: Vector machines must be exposed to alternate segmentations during training (SVM lattice rescoring experiments have been disappointing).

  $\Rightarrow$ Computational complexity of RVMs is prohibitive — $O(N^3)$

- Future research will focus on integrating this technology with syllable modeling technology to circumvent the pronunciation modeling problem.

REFERENCES AND RESOURCES

[1] V.N. Vapnik, *Statistical Learning Theory*, John Wiley, New York, NY, USA, 1998.

[2] C. Cortes, V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 293-297, 1995.

[3] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *http://svm.research.bell-labs.com/SVMdoc.html*, AT&T Bell Labs, November 1999.

[4] M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning*, vol. 1, pp. 211-244, June 2001.

[5] M. Tipping, "The Relevance Vector Machine," in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen and K.-R. Muller, eds., pp. 652-658, MIT Press, 2000.

[6] D. J. C. MacKay, *"Bayesian Methods for Adaptive Models*," Ph. D. Dissertation, California Institute of Technology, Pasadena, California, USA, 1991.

[7] D. J. C. MacKay, "Probable networks and plausible predictions --- a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, 6, pp. 469-505, 1995.

[8] A. Ganapathiraju, *Support Vector Machines for Speech Recognition*, Ph.D. Dissertation, Department of Electrical and Computer Engineering, Mississippi State University, January 2002.

[9] J. Hamaker, *Sparse Bayesian Methods for Continuous Speech Recognition*, Ph.D. Dissertation, Department of Electrical and Computer Engineering, Mississippi State University, January 2002.

Speech Recognition Resources:

[10] "Internet-Accessible Speech Recognition Technology," *http://www.isip.msstate.edu/projects/speech/index.html,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

[11] "Speech Recognition System Training Workshop," *http://www.isip.msstate.edu/conferences/srstw/current/index.html,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 2002.

[12] "Speech Recognition Experiment Server," *http://www.isip.msstate.edu:8080/isip/jsa/index.jsp,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

Pattern Recognition:

[13] "Speech and Signal Processing Demonstrations," *http://www.isip.msstate.edu/projects/speech/software/demonstrations/index.html,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

[14] M. Sewell, "Support Vector Machine Links," http://www.support-vector-machine.org/links.htm, Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom.

APPENDIX: SUPPORT VECTOR MACHINES

- Data for practical applications typically not separable using a hyperplane in the original input feature space

- Transform data to higher dimension where hyperplane classifier is sufficient to model decision surface

$$\Phi : \Re^n \rightarrow \Re^N$$

- Kernels used for this transformation

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$$

- Final classifier:

$$f(\boldsymbol{x}) = \sum_{i=1}^{numSVs} \alpha_i y_i K(\boldsymbol{x}, x_i) + b$$

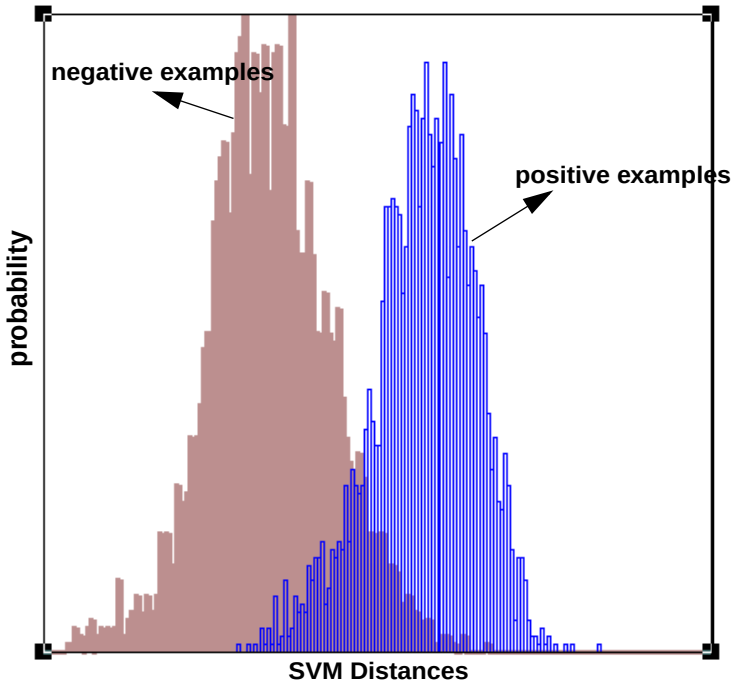
- Soft margin classifiers used in practice:

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \geq 1 - \xi_i \qquad \forall i$$

- SVMs do not generate likelihoods directly

- Posterior estimation required for speech

- Use a sigmoid function to map distances to posteriors:



$$p(y = 1 / f) = \frac{1}{1 + \exp(Af + B)}$$

APPENDIX: RELEVANCE VECTOR MACHINES

- First level of inference:

$$P(\boldsymbol{w}|D, H_i) = \frac{P(D|\boldsymbol{w}, H_i)P(\boldsymbol{w}|H_i)}{P(D|H_i)}$$

$\boldsymbol{w}$: the set of adjustable parameters

D: data from which we make inferences

H_i: overall model

- Second level of inference:

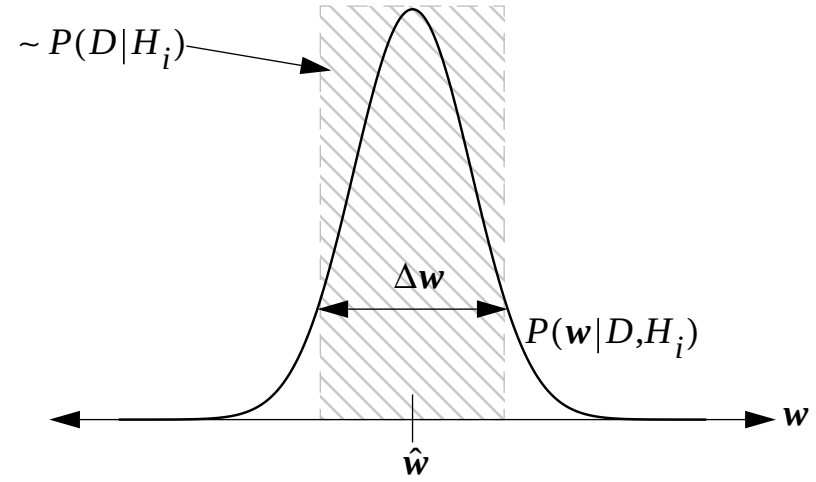
$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}$$

if $P(H_1) = P(H_2)$, best model chosen by evaluating evidence $P(D|H_i)$.

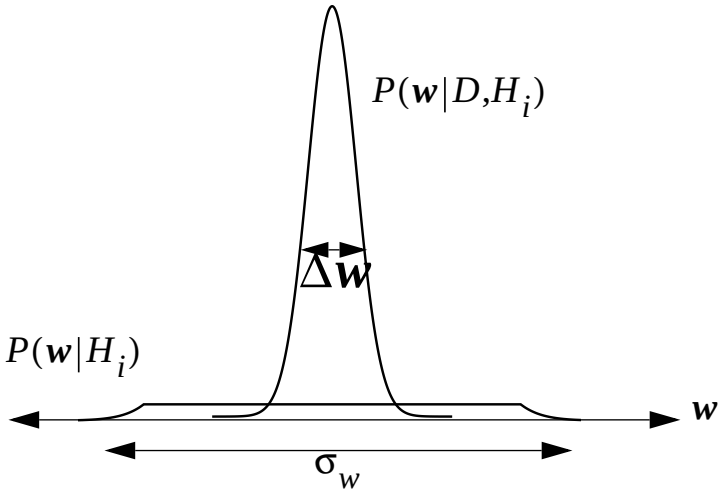
- Evidence marginalized across model parameters:

$$P(D|H_i) = \int P(D|\boldsymbol{w},H_i)P(\boldsymbol{w}|H_i)d\boldsymbol{w}$$

- It is impractical to compute this integral, so we need an approximation.



- Evidence approximation for a single model (Gaussian assumption)



- The parameter's prior distribution and the posterior distribution width determine the model complexity

APPENDIX: OVERFITTING

- Under the assumption that the posterior probability is Gaussian:

$$P(\boldsymbol{w}|D,H_i) \approx P(D|\boldsymbol{w},H_i)P(\boldsymbol{w}|H_i)$$

- The marginalization integral can be assumed to have a strong peak at the most probable value of the parameters, $\hat{\boldsymbol{w}}$.
- The evidence can then be approximated by multiplication of the height of the integrand and the width of the posterior, $\Delta \boldsymbol{w}$.
- The evidence is approximated by

$$P(D|H_i) \approx P(D|\hat{\boldsymbol{w}},H_i)P(\hat{\boldsymbol{w}}|H_i)\Delta \boldsymbol{w}$$

$P(D|\hat{\boldsymbol{w}},H_i)$ is the likelihood of the data given the best-fit parameter set

$P(\hat{\boldsymbol{w}}|H_i)\Delta \boldsymbol{w}$ is a penalty on the range of [0,1] which measures how well our posterior model fits our prior assumptions.

- The objective in training:

$$(\hat{\boldsymbol{w}}, \hat{\alpha}) = \underset{\boldsymbol{w},\, \alpha}{\operatorname{argmax}} \; p(\boldsymbol{w}, \alpha|\boldsymbol{t}, \boldsymbol{O})$$

- Using Bayes' rule:

$$p(\boldsymbol{w}, \alpha|\boldsymbol{t}, \boldsymbol{O}) = \frac{p(\boldsymbol{t}|\boldsymbol{w}, \alpha, \boldsymbol{O})\, p(\boldsymbol{w}, \alpha|\boldsymbol{O})}{p(\boldsymbol{t}|\boldsymbol{O})}$$

- A closed form solution to this maximization is not possible.
- An iterative approximation has been developed by MacKay that has complexity $O(N^3)$ and is based on Gaussian assumptions. Not feasible for large speech recognition tasks.
- This approach is similar to Minimum Description Length (MDL) and Bayesian Information Criterion (BIC).