A Sparse Modeling Approach to Speech Recognition Based on Relevance Vector Machines

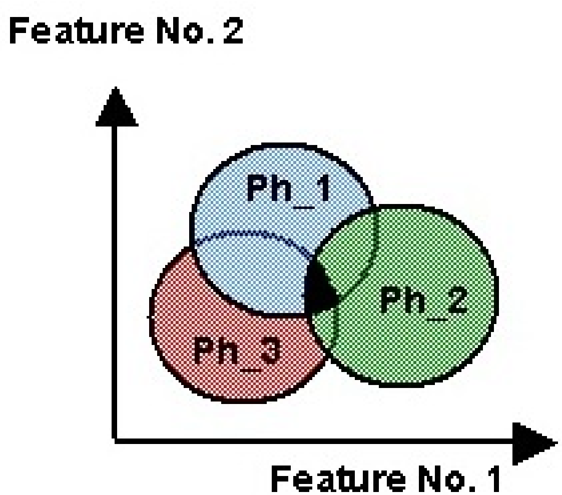
Jon Hamaker and Joseph Picone
hamaker@isip.msstate.edu
Institute for Signal and Information Processing
Mississippi State University

Aravind Ganapathiraju
aganapathiraju@conversay.com
Speech Scientist
Conversay Computing Corporation

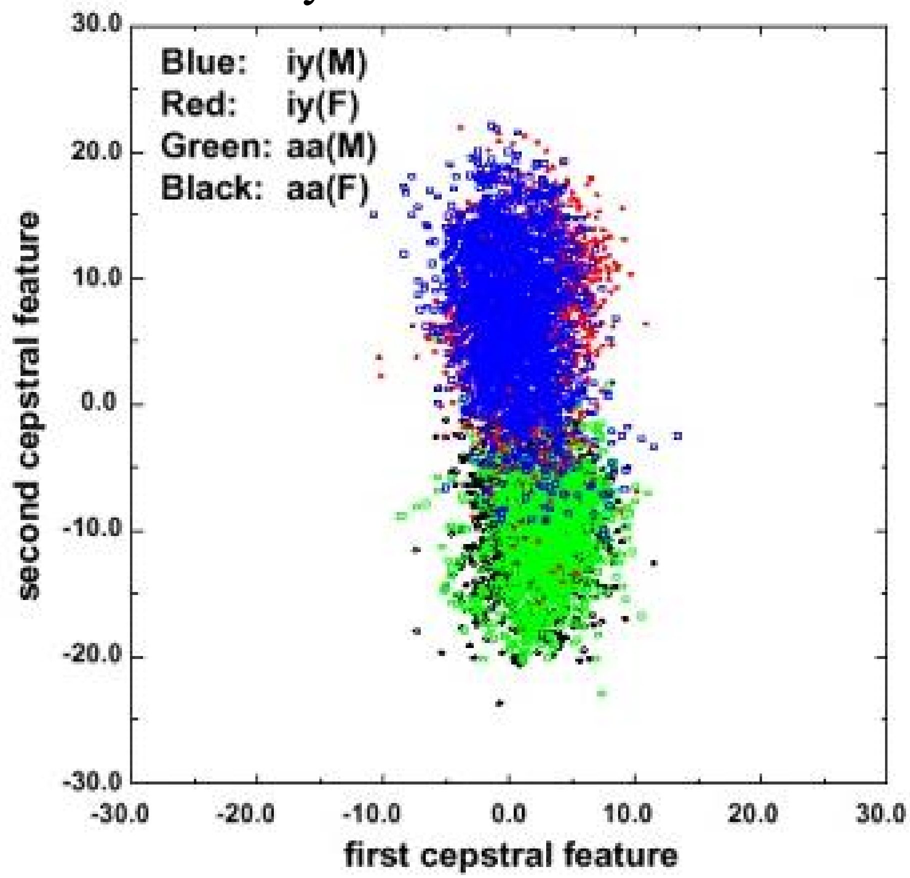
This material is based upon work supported by the National Science Foundation under Grant No.IIS0095940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

MOTIVATION

Acoustic Confusability: Requires reasoning under uncertainty!



Comparison of “aa” in “lOck” and “iy” in “bEAt” for SWB



- Regions of overlap represent classification error
- Reduce overlap by introducing acoustic and linguistic context.

ACOUSTIC MODELS

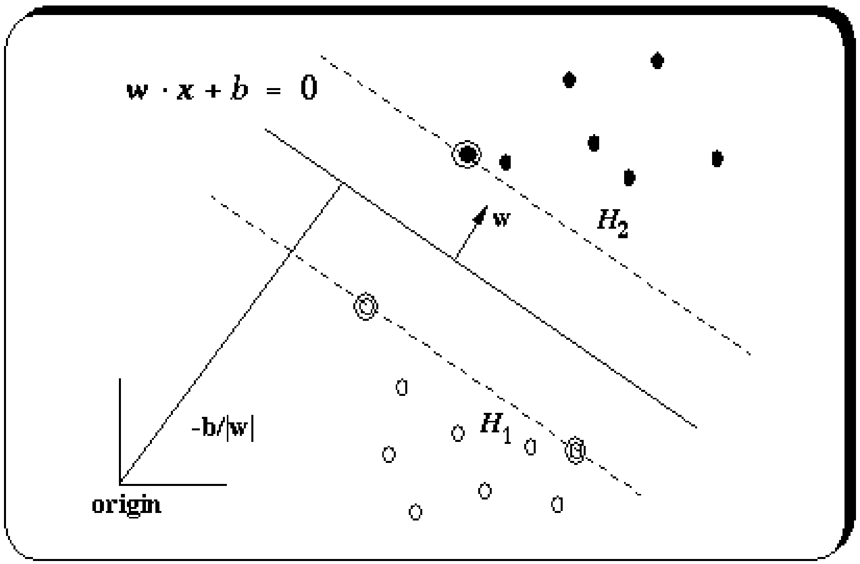
Acoustic Models Must:

- Model the temporal progression of the speech
- Model the characteristics of the sub-word units

We would also like our models to:

- Optimally trade-off discrimination and representation
- Incorporate Bayesian statistics (priors)
- Make efficient use of parameters (sparsity)
- Produce confidence measures of their predictions for higher-level decision processes

SUPPORT VECTOR MACHINES



$$f(x) = \sum_i \alpha_i y_i K(x_i, x) + b$$

$$y_i = \pm 1$$

$$K(x_i, x) = \Phi(x_i) \bullet \Phi(x)$$

- Maximizes the margin between classes to satisfy SRM.
- Balances empirical risk and generalization.
- Training is carried out via quadratic optimization.
- Kernels provide the means for nonlinear classification.
- Many of the multipliers go to zero – yields sparse models.

DRAWBACKS OF SVMS

- Uses a binary decision rule
  - Can generate a distance, but on unseen data, this measure can be misleading
  - Can produce a “probability” using sigmoid fits, etc. but they are inadequate
- Number of support vectors grows linearly with the size of the data set
- Requires the estimation of trade-off parameters via held-out sets

RELEVANCE VECTOR MACHINES

- A kernel-based learning machine

$$y(x;w) = w_0 + \sum_{i=1}^{N} w_i K(x_i, x)$$

$$P(t=1 \mid x_i; w) = \frac{1}{1+e^{-y(x_i;w)}}$$

- Incorporates an automatic relevance determination (ARD) prior over each weight (MacKay)

$$P(w \mid \alpha) = \prod_{i=0}^{N} N(w_i \mid (\mu_i = 0), \frac{1}{\alpha_i})$$

- A flat (non-informative) prior over α completes the Bayesian specification.

RELEVANCE VECTOR MACHINES

- The goal in training becomes finding:

$$\hat{w},\hat{\alpha} = \underset{w,\alpha}{\arg\max}\; p(w,\alpha \mid t, X) \quad \textit{where}$$

$$p(w,\alpha) = \frac{p(t \mid w,\alpha, X)\, p(w,\alpha \mid X)}{p(t \mid X)}$$

- Estimation of the "sparsity" parameters is inherent in the optimization – no need for a held-out set!
- A closed-form solution to this maximization problem is not available. Rather, we iteratively reestimate $\hat{w}$ *and* $\hat{\alpha}$.

LAPLACE'S METHOD

- Fix α and estimate **w** (e.g. gradient descent)

$$\hat{w} = \arg\max_{w} \; p(t \mid w)\, p(w \mid \alpha)$$

- Use the Hessian to approximate the covariance of a Gaussian posterior of the weights centered at $\hat{w}$

$$\Sigma = -\left\{ \nabla_w \nabla_w \left[p(t \mid w)\, p(w \mid \alpha) \right] \right\}^{-1}$$

- With $\hat{w}$ and Σ as the mean and covariance, respectively, of the Gaussian approximation, we find $\hat{\alpha}$ by finding

$$\hat{\alpha}_i = \frac{\gamma_i}{\hat{w}_i^2} \quad where \;\; \gamma_i = 1 - \alpha_i \Sigma_{ii}$$

CONSTRUCTIVE TRAINING

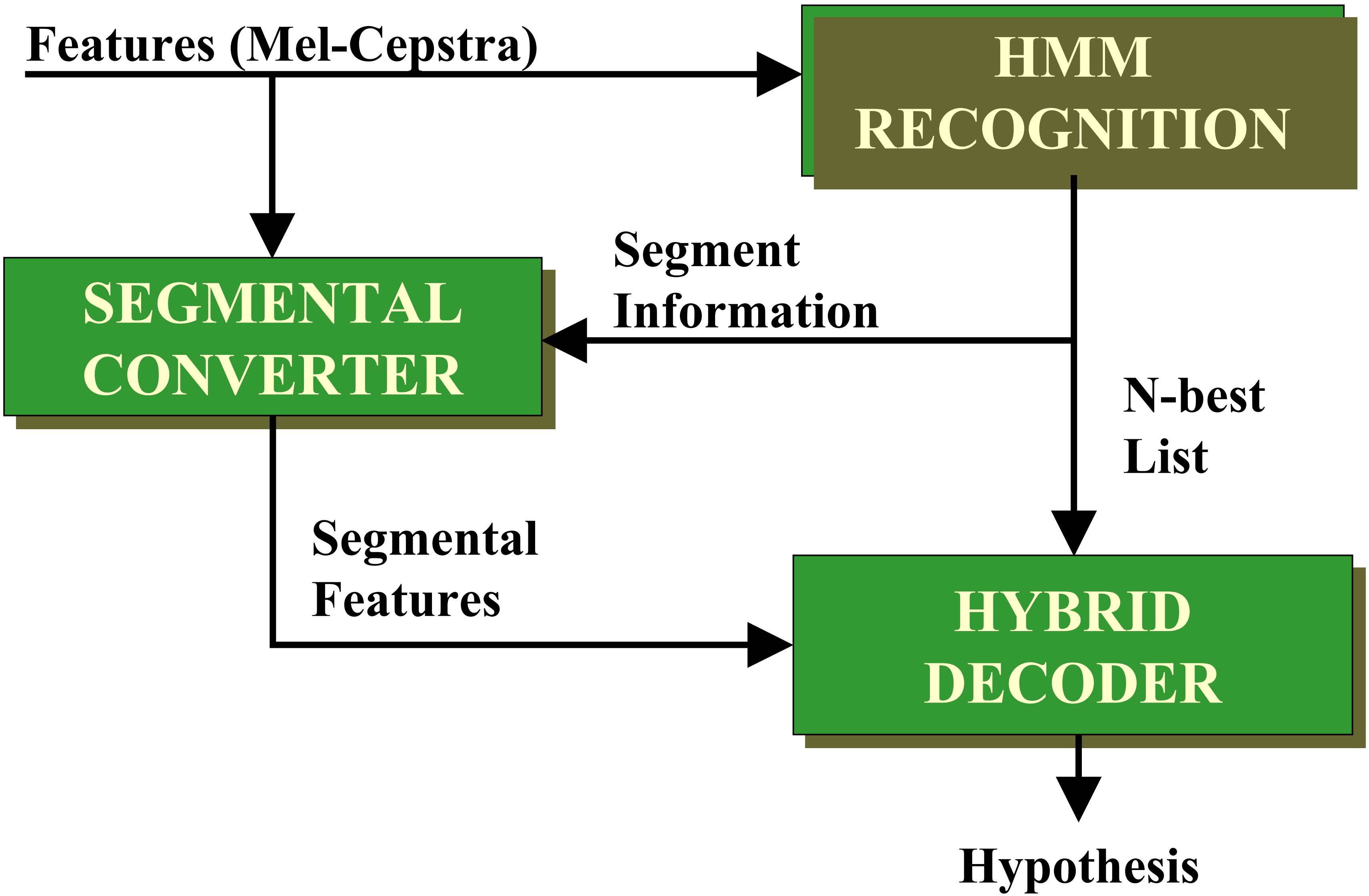
- Central to this method is the inversion of an MxM hessian matrix: an $O(N^3)$ operation initially
- Initial experiments could use only 2-3 thousand vectors
- Tipping and Faul have defined a constructive approach
  - Define $L(\alpha) = L(\alpha_{-i}) + l(\alpha_i)$
  - $L(\alpha)$ has a unique solution with respect to α_i
  - The results give a set of rules for adding vectors to the model, removing vectors from the model or updating parameters in the model
  - Begin with all weights set to zero and iteratively construct an optimal model without evaluating the full NxN matrix.

STATIC CLASSIFICATION

- Deterding Vowel Data: 11 vowels spoken in "h*d" context.

Approach	Error Rate	# Parameters
K-Nearest Neighbor	44%	
Gaussian Node Network	44%	
SVM: Polynomial Kernels	49%	
SVM: RBF Kernels	35%	83 SVs
Separable Mixture Models	30%	
RVM: RBF Kernels	30%	13 RVs

FROM STATIC CLASSIFICATION TO RECOGNITION



ALPHADIGIT RECOGNITION

- OGI Alphadigits: continuous, telephone bandwidth letters and numbers
- Reduced training set size for comparison: 10000 training vectors per phone model.
  - Results hold for sets of smaller size as well.
  - Can not yet run larger sets efficiently.
- 3329 utterances using 10-best lists generated by the HMM decoder.
- SVM and RVM system architecture are nearly identical: RBF kernels with gamma = 0.5.
  - SVM requires the sigmoid posterior estimate to produce likelihoods.

ALPHADIGIT RECOGNITION

Approach	Error Rate	Avg. # Parameters	Training Time	Testing Time
SVM	15.5%	994	3 hours	1.5 hours
RVM	14.8%	72	5 days	5 mins

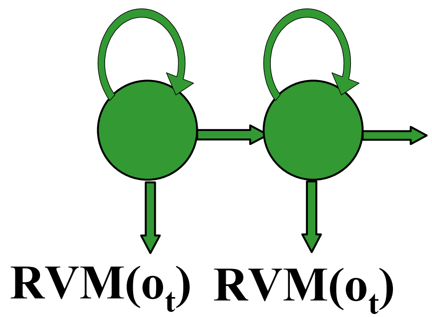
- RVMs yield a large reduction in the parameter count while attaining superior performance.
- Computational costs mainly in training for RVMs but is still prohibitive for larger sets.
- SVM performance on full training set is 11.0%.

CONCLUSIONS

- Application of sparse Bayesian methods to speech recognition.
  - Uses automatic relevance determination to eliminate irrelevant input vectors: Applications in maximum likelihood feature extraction?
- State-of-the-art performance in extremely sparse models.
  - Uses an order of magnitude fewer parameters than SVMs: Decreased evaluation time.
  - Requires several orders of magnitude longer to train: Need more efficient training routines that can handle continuous speech corpora.

CURRENT WORK

HMMs with RVM Emission Distributions



Iterative Parameter Estimation

- Frame-level classification
- Convergence properties and efficient training methods are critical
- A "chunking" approach is in development
  - Apply the algorithm to small subsets of the basis functions
  - Combine results from each subset to reach a full solution
  - Optimality?

REFERENCES

- M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning*, vol. 1, pp. 211-244, June 2001.
- A. Faul and M. Tipping, "Analysis of Sparse Bayesian Learning," in T.G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14,* pp. 383-389, MIT Press, 2002.
- M. Tipping and A. Faul, "Fast Marginal Likelihood Maximization for Sparse Bayesian Models," *Artificial Intelligence and Statistics '03,* preprint, August, 2002.
- D.J.C. MacKay, "Probable Networks and Plausible Predictions --- A Review of Practical Bayesian Methods for Supervised Neural Networks," *Network: Computation in Neural Systems,* vol. 6, pp. 469-505, 1995.
- A. Ganapathiraju, *Support Vector Machines for Speech Recognition*, Ph.D. Dissertation, Mississippi State University, Mississippi State, Mississippi, USA, 2002.
- J. Hamaker, *Sparse Bayesian Methods for Continuous Speech Recognition,* Ph.D. Dissertation (preprint), Mississippi State University, Mississippi State, Mississippi, USA 2003.