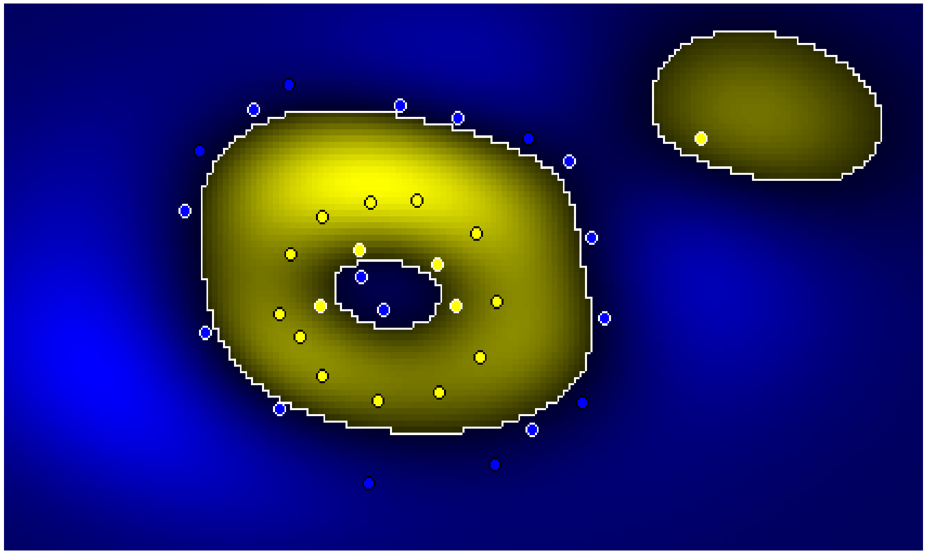
Motivation

- Support Vector Machine (SVM) is a machine learning technique applied to a variety of tasks like classification and function estimation
- Successfully applied to image and text classification problems
- This is the first attempt at applying SVMs to large vocabulary continuous speech recognition
- Encouraging results on Switchboard (SWB) rescoring experiments

SVM Preliminaries



- Based on Structural Risk Minimization
- Discriminative learning technique
- Models non-linear decision regions by transformation to higher dimension

SVM Fundamentals

- **Hyperplane:** $$\sum_{i=1}^{l} y_i \alpha_i \cdot K(x_i \bullet x) + b = 0, \alpha_i \ge 0$$
- **Constraints:** $$\xi_i \ge 0, y_i\left(\sum_{j=1}^{l} y_j \alpha_j \cdot K(x_i \bullet x_j) + b\right) \ge 1 - \xi_i$$
- **Optimize:** $$\phi = \frac{1}{2}(w \cdot w) + C\sum\xi_i \quad , \qquad w = \sum_{i=1}^{l} y_i \alpha_i \cdot x_i + b$$
- **Training vectors with non-zero α are called support vectors**
- **K is the non-linear kernel**
- C **controls the penalty for errors**
- $\sum\xi_i$ **is an upper bound on errors**

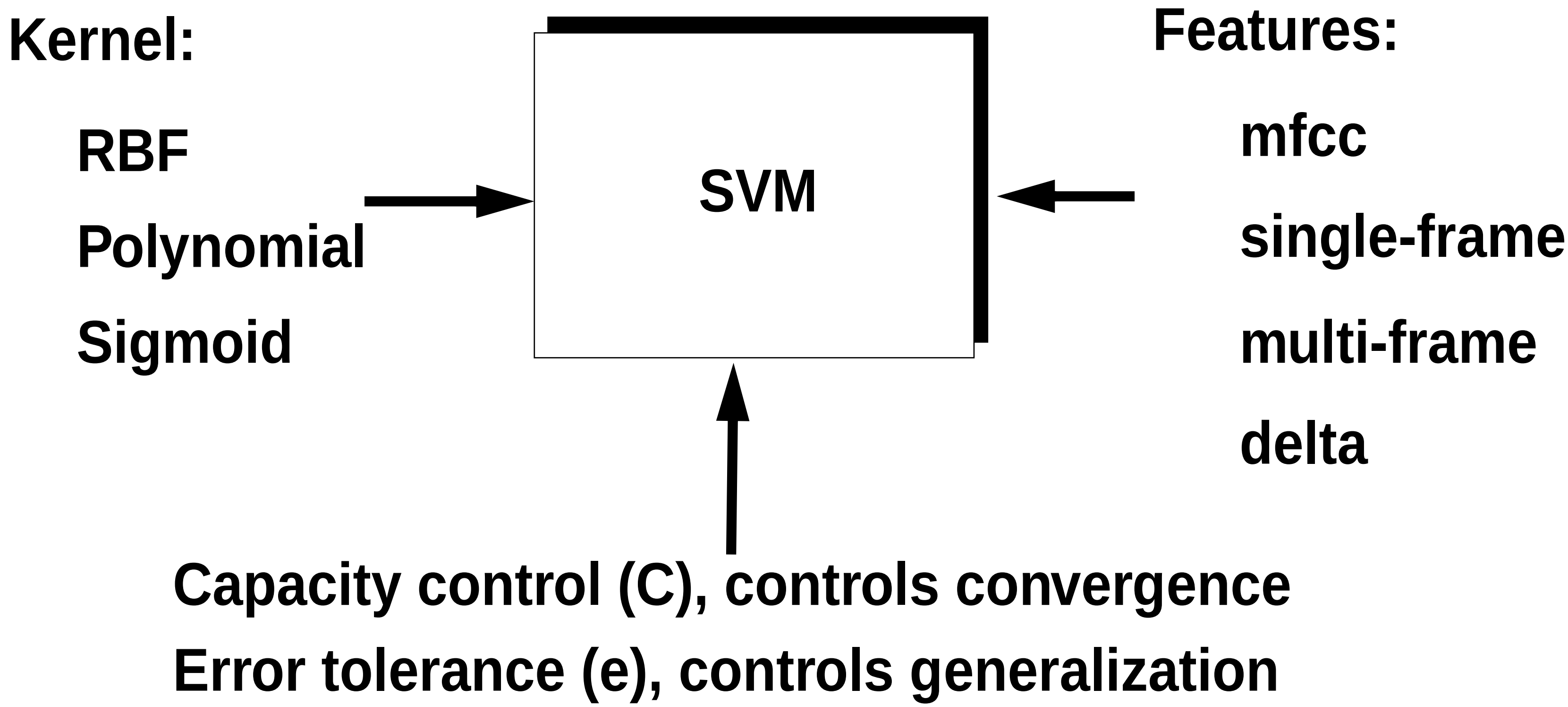
Previous Applications of SVM

- Mostly applied to static pattern recognition problems
- Digit recognition (Vapnik et. al. 1995)
- Text characterization (Joachims et. al., 1998)
- Speaker identification (Gish et. al., 1996)
- Phonetic classification including TIMIT (Clarkson, et. al., 1998)

SVMs in LVCSR

- Temporal modeling is crucial in LVCSR
- SVMs when used as classifiers can be used to compute likelihoods at the frame level
- Avoid temporal modeling/search problems by using a phone alignment rescoring paradigm (Picone et. al., 1998)
- Need to convert SVM distances to likelihoods:
  - ❑ Simple linear regression between distances and probabilities
  - ❑ ANN that maps distances to probabilities

SVMs in Practice



- Careful choice of input features can compensate for lack of temporal modeling
- RBF kernels have been more successful

Vowel Classification

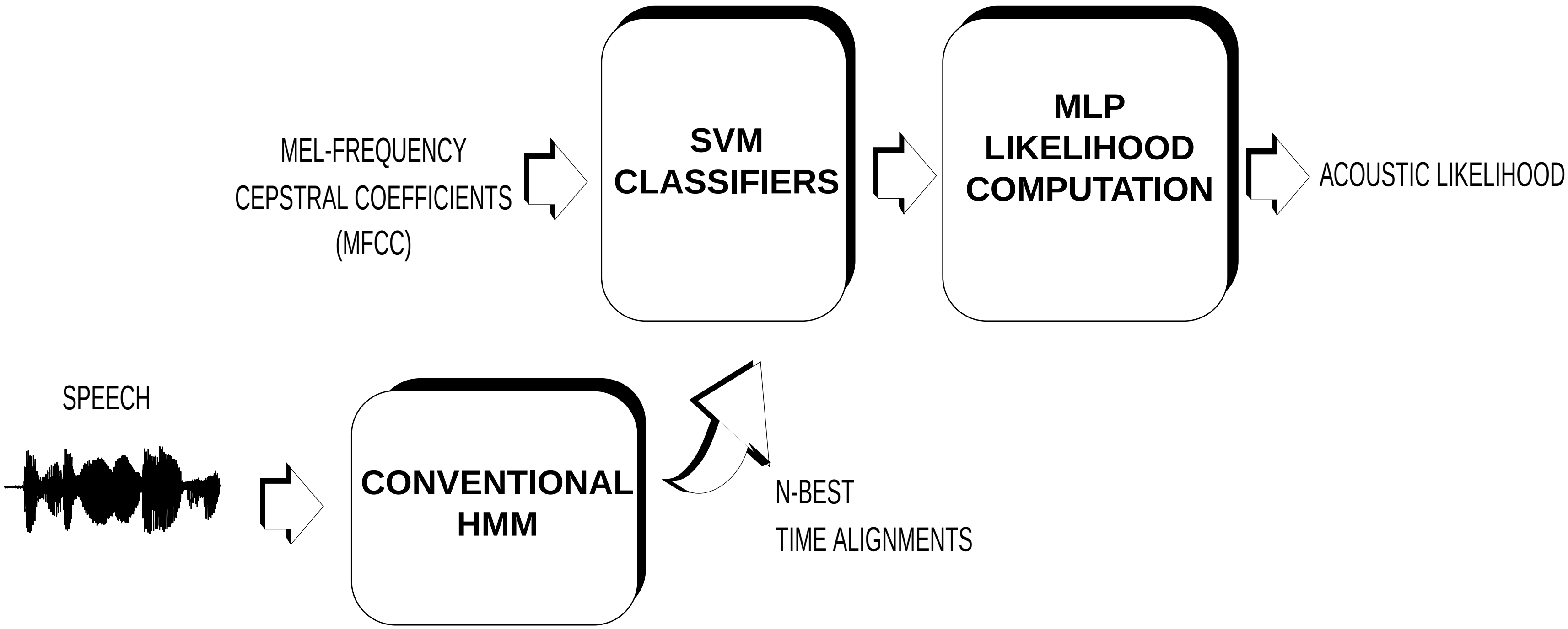
order/gamma/ hidden-units	RBF	Polynomial	Gaussian Node Network
2/0.025/22	32	**42**	46
3/0.05/88	**31**	44	47
4/0.1/528	32	45	**45**

- ☛ **Deterding Vowel data used — a standard data set to benchmark non-linear classifiers**
- ☛ **11 classes → 11 two-class classifiers**
- ☛ **Results comparable to state-of-the-art performance on this data set: 29% using separable mixture models (Tenenbaum et. al. 1997)**

SWB Phone Classification

- A natural pilot experiment before evaluating on continuous speech
- 16 phones, including vowels, fricatives and nasals extracted from SWB
- RBF kernel achieves best performance of 52% classification error
- Polynomial kernel of order 10 achieves a performance of 62% classification error: in general RBF kernel outperforms the polynomial kernel

SWB Rescoring Methodology



- **Cross-word triphone system used for alignments**
- **44 two-class classifiers trained using RBF kernel — in-class vs. all other classes**
- **MLP used for likelihood computation**

SWB Rescoring Results

N-Best	SVM Rescore	N-Best Error
1	—	43.8%
5	48.5%	50.6%
10	49.8%	52.8%
15	52.4%	51.5%
20	55.8%	55.8%
25	55.4%	57.9%
30	54.5%	53.8%

☛ **For 5-best**

- ❑ **word-internal triphone system — 52.0%**
- ❑ **syllable system — 50.9%**

Analysis

- Number of support vectors (SV) proportional to complexity of the problem
- High degree of support vector sharing — over 75% of SVs shared by at least 2 classifiers
- Issues:
  - Number of classifiers in context dependent systems - not practical
  - Training time and resources: ~24 hours per classifier for 10 hours of speech data on a 300MHz Pentium PII machine with 0.5Gb RAM

Summary

- First successful application of SVMs to large vocabulary speech recognition - vowel and phone classification, SWB rescoring
- SWB rescoring results (48.5% WER for 5-best) comparable to triphone systems
- Need a faster optimization process — Sequential Minimal Optimization (Platt, 1998)
- Explore incorporation of context into feature vectors (Clarkson et. al., 1998)