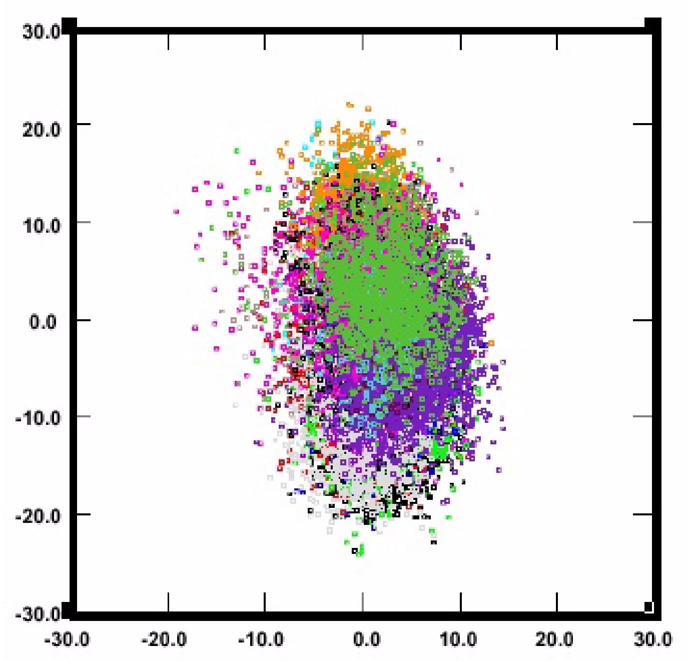
Motivation

- ☞ Mislabeled data in speech corpora — ~5% inherent WER in Switchboard
- ☞ Learning in such noisy environments crucial for robust classifier design
- ☞ Learning in SVMs can be made efficient by identifying mislabeled data
- ☞ Differs from other techniques — mislabeled data identified within the estimation loop
- ☞ Need for an explicit data cleanup stage eliminated

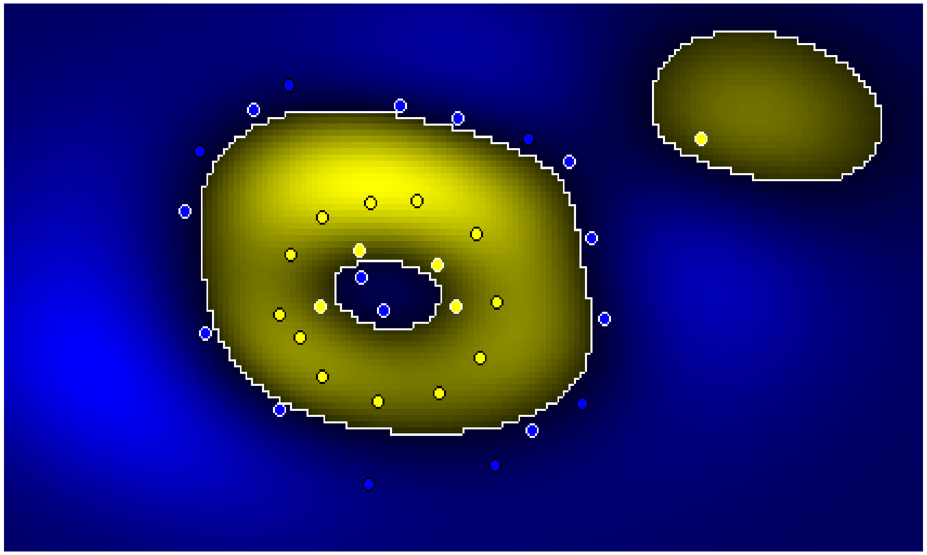
Data Overlap



first two cepstral coefficients
for vowels in Switchboard

- Significant overlap in real speech data
- Use the non-overlap region to learn a decision surface
- Good open-loop performance — possibly worse closed-loop performance

SVM Classification



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

SVM Theory

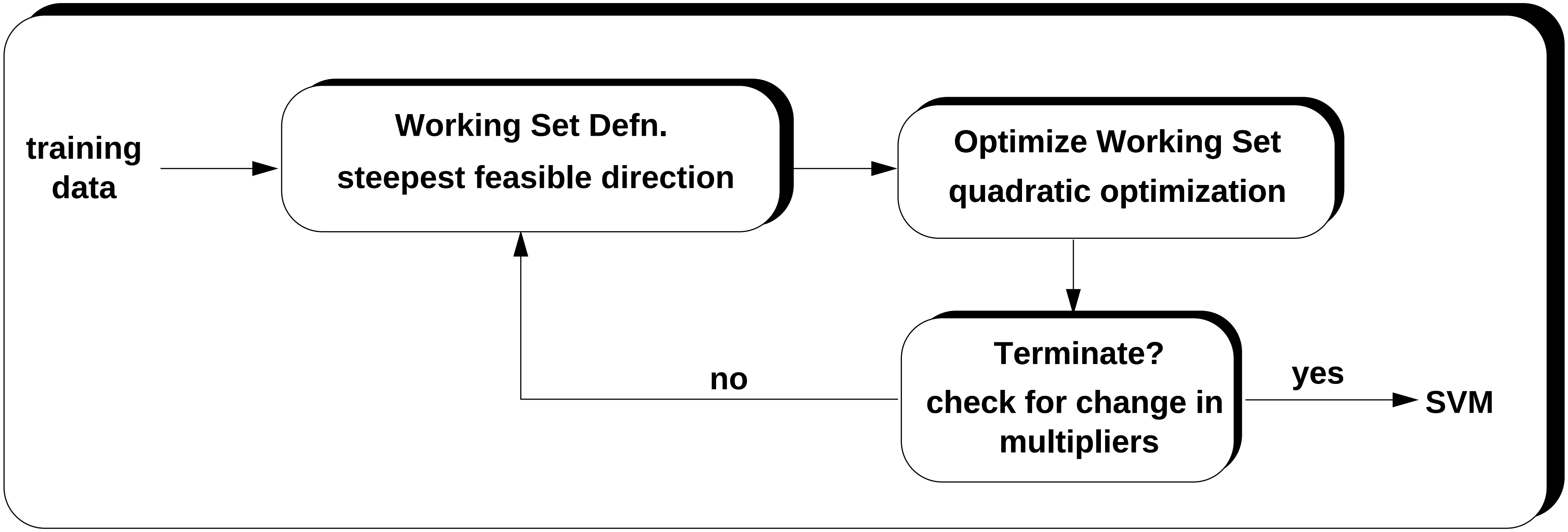
- **Hyperplane:** $\sum_{i=1}^{l} y_i\alpha_i \cdot K(x_i \bullet x) + b = 0, \alpha_i \ge 0$

- **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:** $\phi = \frac{1}{2}(w \cdot w) + C\sum\xi_i$ **,** $w = \sum_{i=1}^{l} y_i\alpha_i \cdot x_i + b$

- **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 approximation for the number of errors allowed for the training set**

Chunking Algorithm



training data

Working Set Defn.

steepest feasible direction

Optimize Working Set

quadratic optimization

Terminate?

check for change in multipliers

no

yes

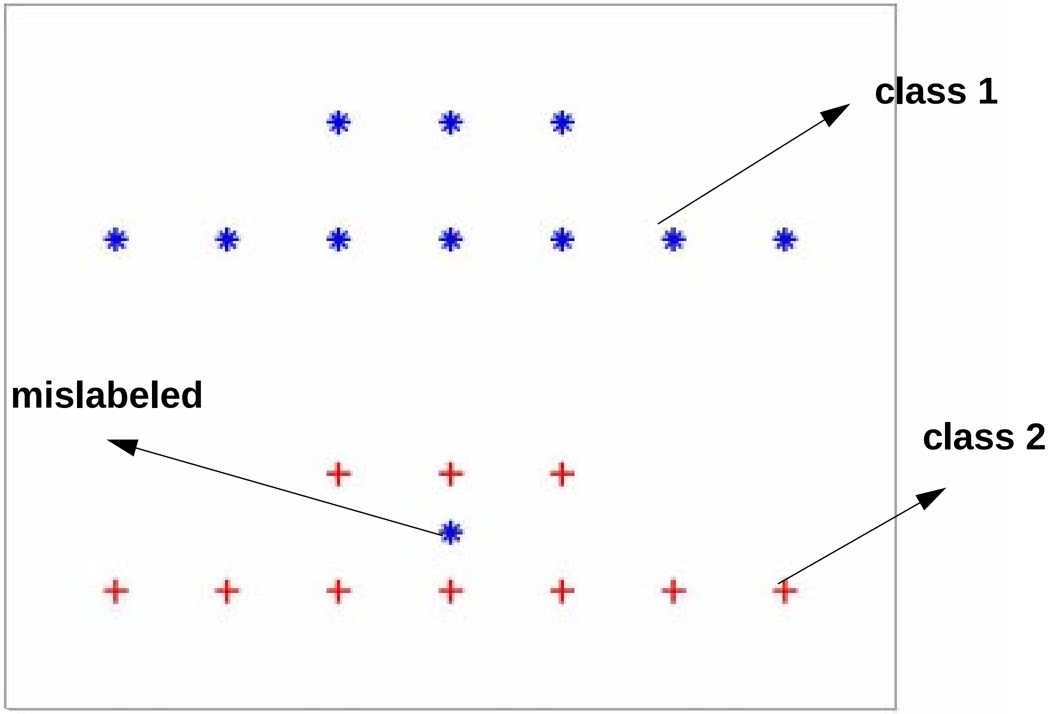
SVM

- Proposed by Osuna et al.
- Guarantees convergence to global optimum
- Working set definition is crucial

Bounded Support Vectors (BSV)

- **Chunking converges faster when the working set is composed of examples that violate the Karush-Kuhn-Tucker optimality conditions**
- **Several support vectors with multipliers at the upper bound (C) — they form the BSVs**
- **If an example is identified as a BSV for several iterations, the example is probably mislabeled (or noise)**
- **Elimination of such examples from further estimation gives faster convergence and better classifiers**

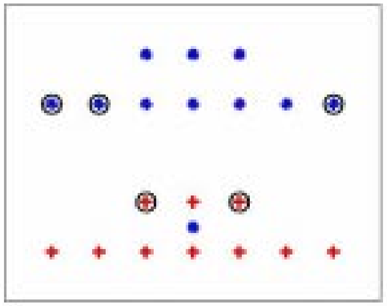
Synthetic Data Example - I



☛ 2-D data - simple classifier sufficient

☛ Noisy data generated by intentionally mislabeling some negative examples

Synthetic Data Example - II



Mislabeled data identified

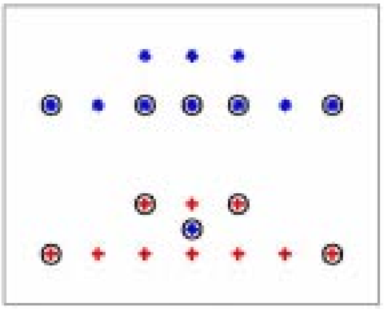
Fewer SVs

Simple classifier

Mislabeled data not identified

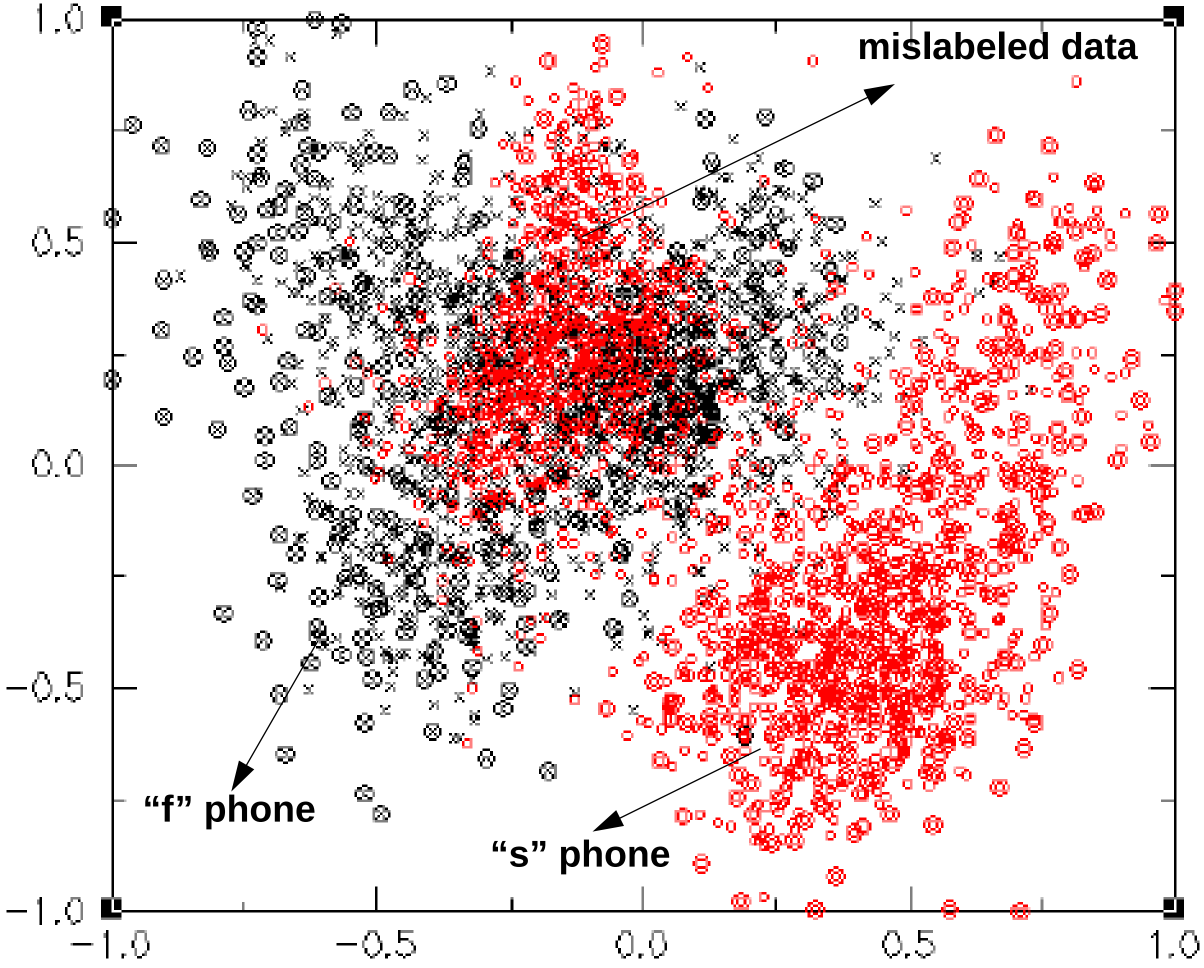
Increased number of SVs

Complex classifier



- Consistent identification of BSVs leads to effective culling of mislabeled data
- Identifying BSVs results in simpler and more effective classifiers

Real Data Example



support vectors

indicated by circles

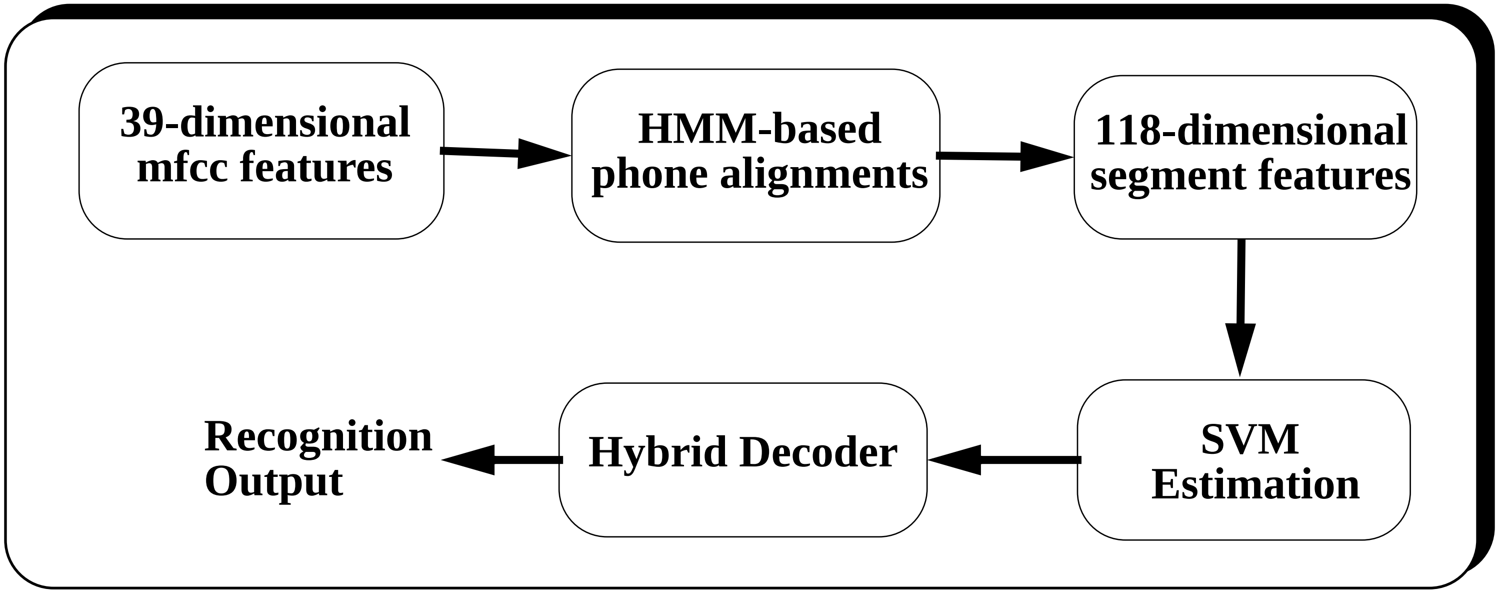
over the original data

radial basis function

used as kernel

☛ **First two cepstral coefficients of phones 's' and 'f' — example shows the need for identifying mislabeled data in real speech**

Hybrid ASR System


39-dimensional mfcc features
HMM-based phone alignments
118-dimensional segment features
SVM Estimation
Hybrid Decoder
Recognition Output

- 118-dimensional composite feature vectors used for SVM classifiers — log duration included
- Classifiers bootstrapped from a cross-word triphone system

System Performance

	Without Data Cleanup	With Data Cleanup
Substitutions	11.1	10.8
Deletions	0.5	0.5
Insertions	0.5	0.3
Total Error	12.1	11.6

- OGI Alphadigit data — 8500 training sentences and 1000 test sentences
- Mean relative improvement in classifier accuracy — 9%
- Decrease in the number of support vectors in new system — 41%
- Improvement in performance of the hybrid system — 7%

Conclusions

☞ Need for identification of mislabeled data — speech databases are not perfectly transcribed

☞ Identifying mislabeled data important for hybrid systems which use bootstrapping

☞ Improved hybrid system performance —
— 7% relative improvement in terms of WER
— 41% fewer support vectors in the new system

☞ Need for a data-driven methodology to estimate the training error penalty

☞ Need for formulation of a confidence measure