Discriminative training for Automatic Speech Recognition using the Minimum Classification Error framework

Erik McDermott

NTT Communication Science Labs

NTT Corporation

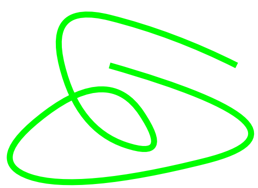
mcd@cslab.kecl.ntt.co.jp

Motivation & Overview

- Need for discriminative training
  - Overcome modeling limitations
  - Focus on *directly* improving recognition performance
- Overview:
  - MCE fundamentals
    - Smoothed error rate
      $\rightarrow$ parallel with large margin training
  - MCE vs. Maximum Likelihood results for large-scale speech recognition tasks

MCE training for generic models

Training
pattern x from
category k



Recognition system
(parameters: Λ)

score(x, Λ, category 1)
.
.
score(x, Λ, category i)
.
score(x, Λ, category M)

max

decision = best category

error: 0 if best = k,
1 otherwise

new Λ = Λ + $\Delta\Lambda$

d_k(x, Λ) =
best incorrect score - correct score

loss = sigmoid(d_k(x, Λ))

use d_loss/d_Λ to define $\Delta\Lambda$

Searching for the Bayes classifier

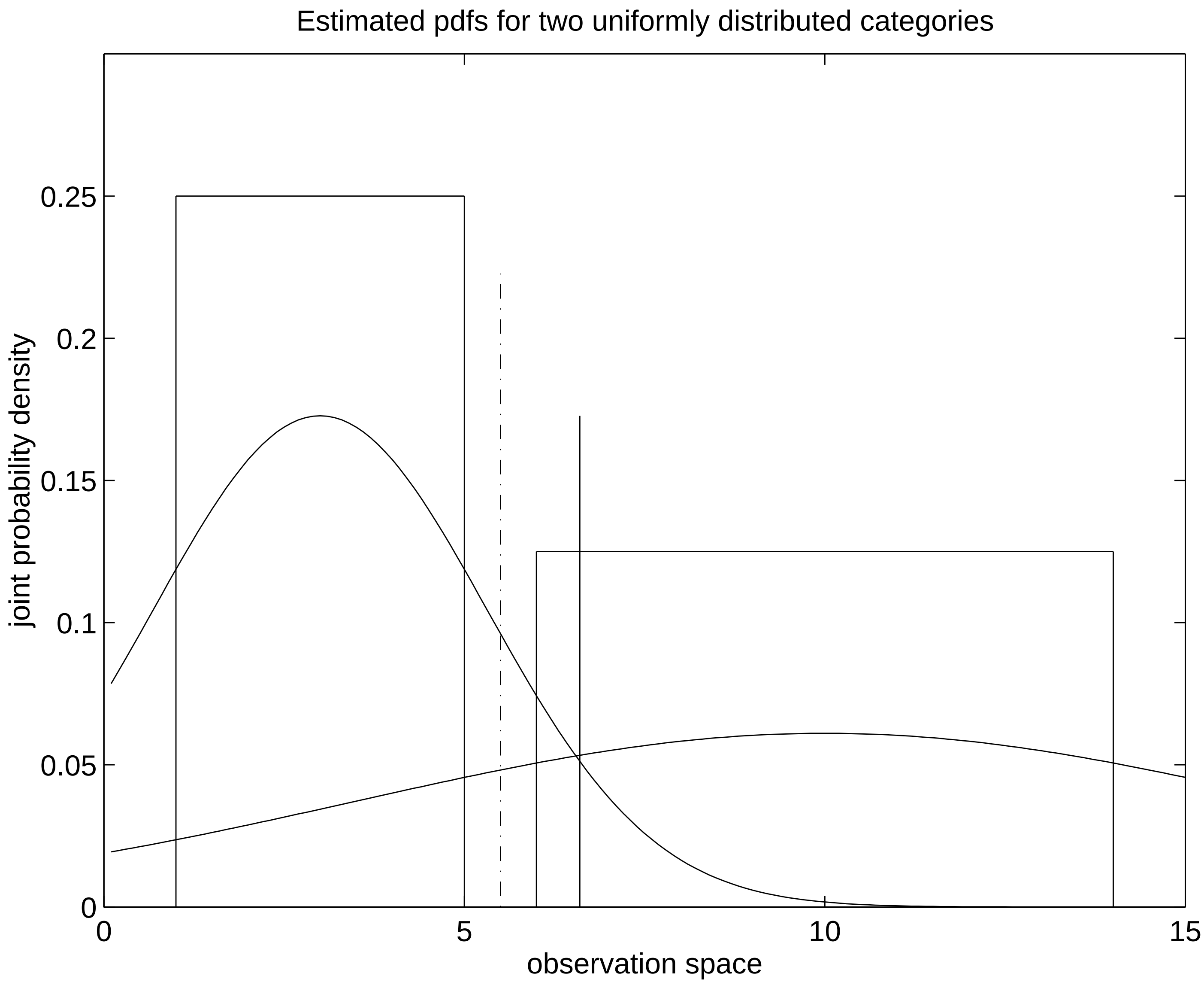
- Bayes decision rule:

$$\text{decide } C_i \text{ if } P(C_i|\mathbf{x}) > P(C_j|\mathbf{x}) \text{ for all } j \neq i$$

- In principle, the same optimal error can be attained using **discriminant functions**:

$$\text{decide } C_i \text{ if } g_i(\mathbf{x}, \mathbf{\Lambda}) > g_j(\mathbf{x}, \mathbf{\Lambda}) \text{ for all } j \neq i$$

Maximum Likelihood fails to separate!



MCE Misclassification Measure

- The m.m. compares **correct** and **best incorrect** categories:

$$d_k(\mathbf{x}_1^T, \mathbf{\Lambda}) = -g_k(\mathbf{x}_1^T, \mathbf{\Lambda}) + \max_{j \neq k} g_j(\mathbf{x}_1^T, \mathbf{\Lambda})$$

$d_k(\mathbf{x}_1^T, \mathbf{\Lambda}) < 0 \rightarrow$ correct classification, and
$d_k(\mathbf{x}_1^T, \mathbf{\Lambda}) \geq 0 \rightarrow$ incorrect classification.

- Special case of continuous definition (Chou, 1992):

$$d_k(\mathbf{x}_1^T, \mathbf{\Lambda}) = -g_k(\mathbf{x}_1^T, \mathbf{\Lambda}) + \log \left[\frac{1}{M-1} \sum_{j \neq k} e^{g_j(\mathbf{x}_1^T, \mathbf{\Lambda})\psi} \right]^{\frac{1}{\psi}}$$

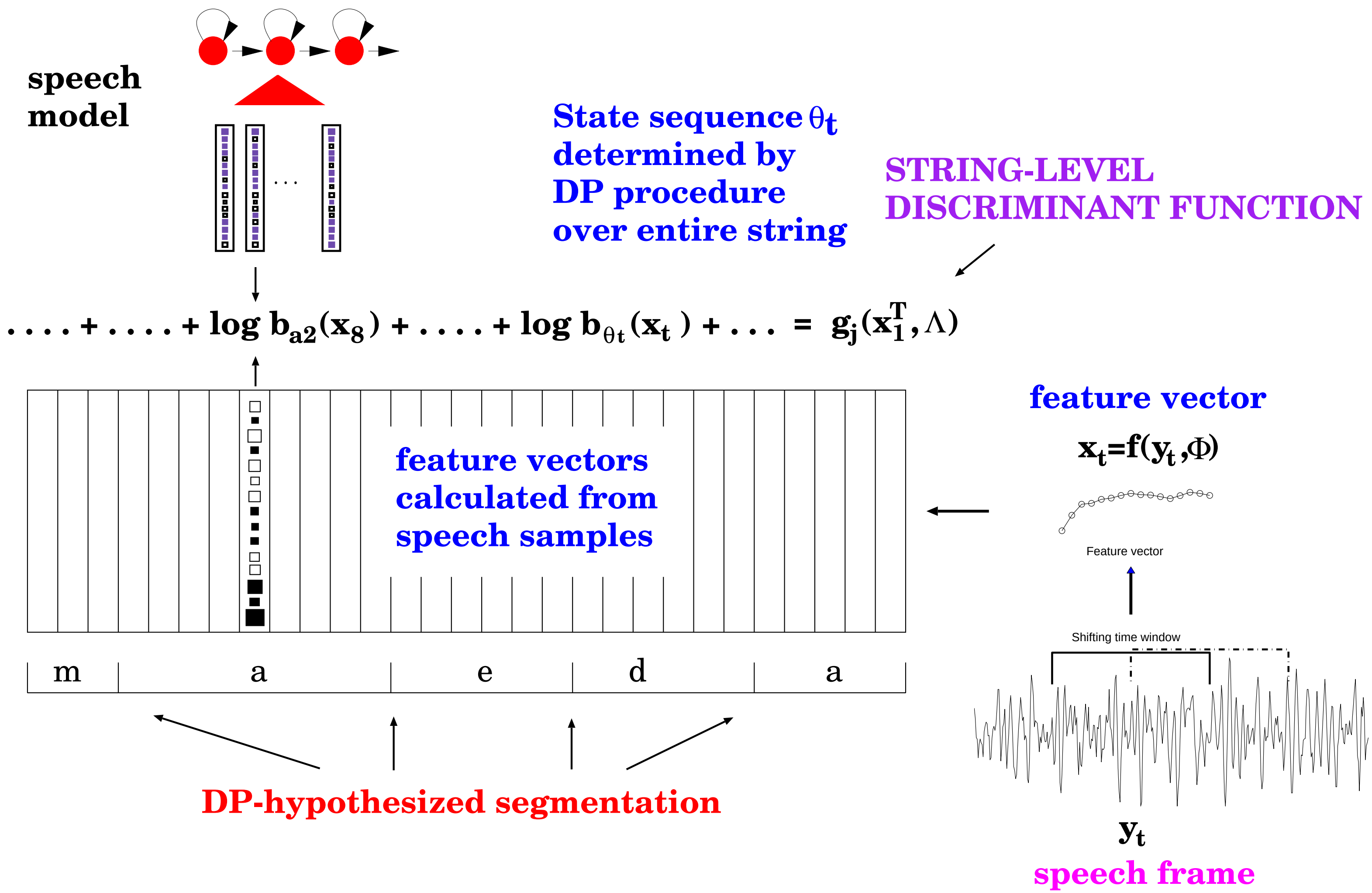
Discriminant function for HMMs

- Defined using best Viterbi path Θ^j:

$$g_j(\mathbf{x}_1^T, \mathbf{\Lambda}) = \log P(S_j) + \sum_{t=1}^{T} \log a_{\theta_{t-1}^j \theta_t^j} + \sum_{t=1}^{T} \log b_{\theta_t^j}(\mathbf{x}_t)$$

- Input: sequence of feature vectors, $\mathbf{x}_1^T = (\mathbf{x}_1, ..., \mathbf{x}_T)$

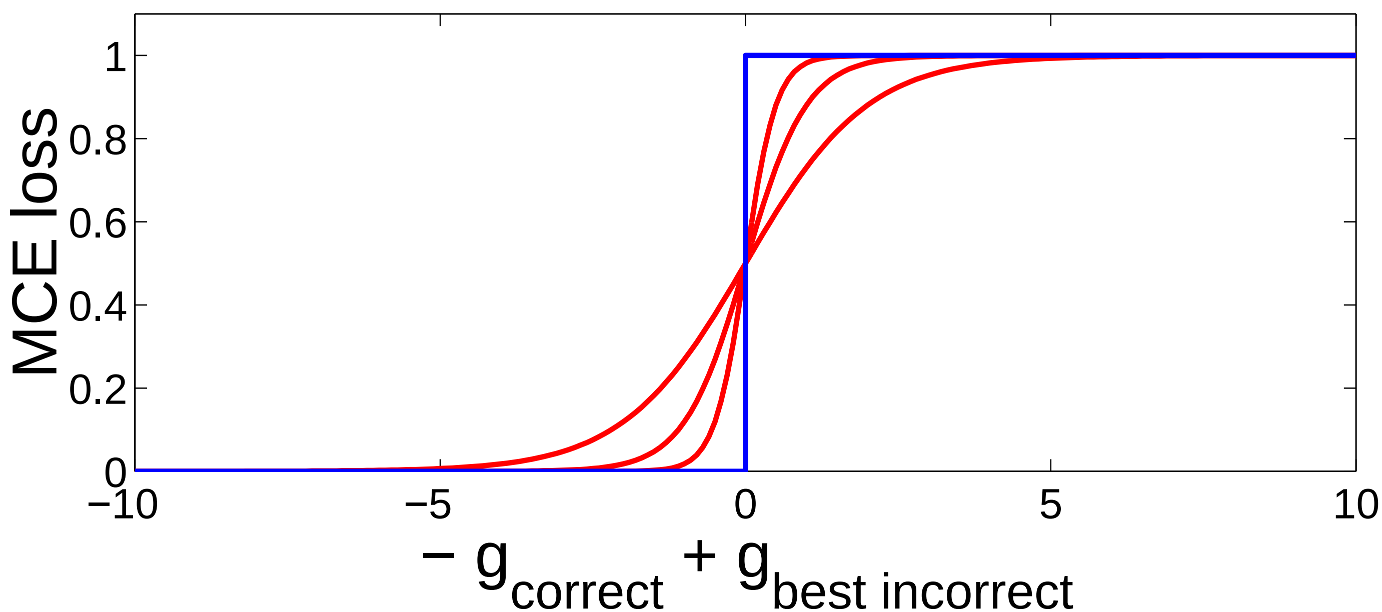
String-level HMM discriminant function



MCE loss function

Reflects classification success/failure:

$$\ell(d_k(\mathbf{X}_1^T, \mathbf{\Lambda})) = \frac{1}{1 + e^{-\alpha d_k(\mathbf{X}_1^T, \mathbf{\Lambda})}}$$



Overall loss and optimization

- A practical definition of overall loss is the **Average Empirical Cost**:

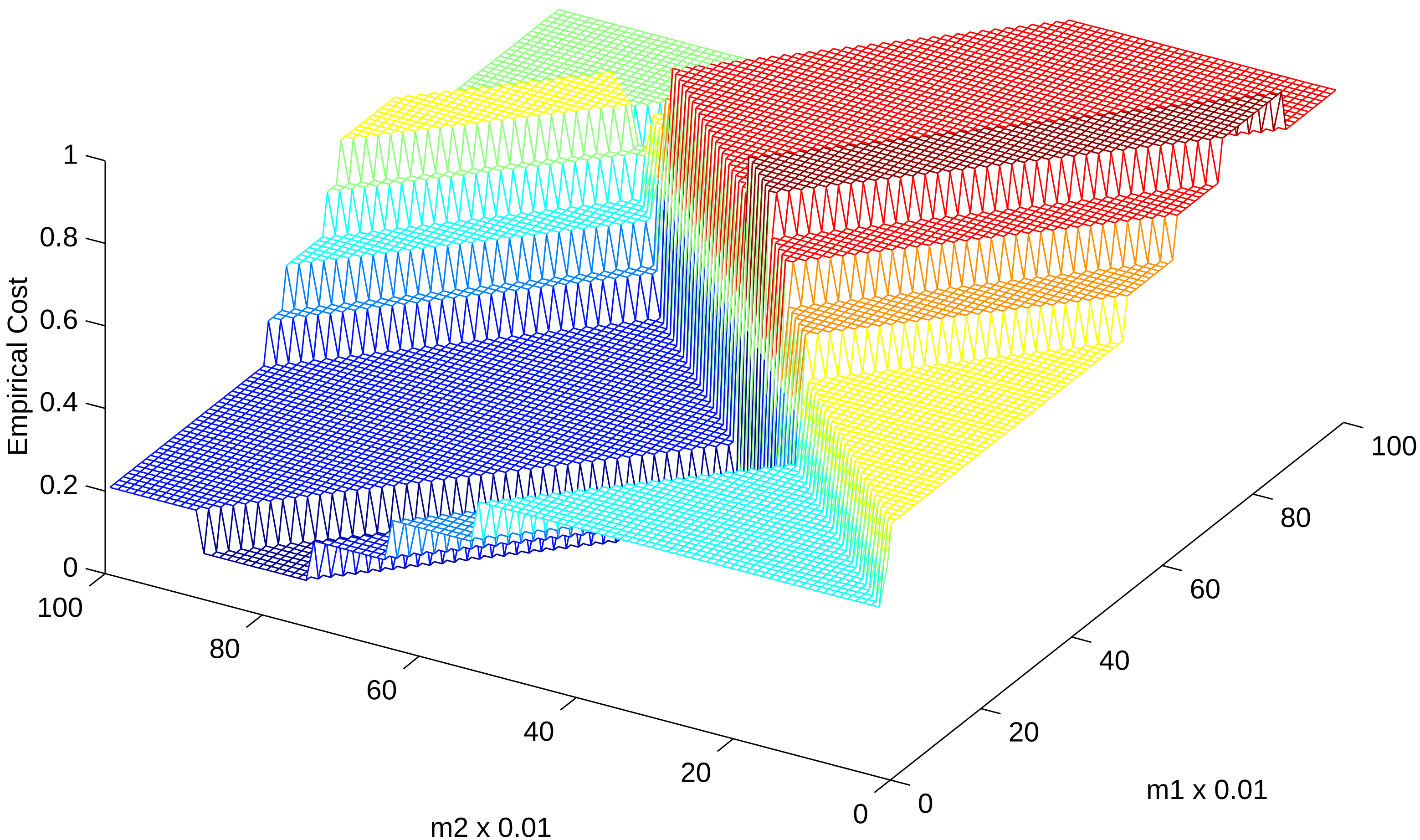
$$L(\mathbf{\Lambda}) = \frac{1}{N} \sum_{k}^{M} \sum_{i=1}^{N_k} \ell_k(\mathbf{x}_{ik}, \mathbf{\Lambda})$$

- E.g.: **Quickprop** (Fahlman, 1988) can be seen as a modified Newton's method:
  - use a second order Taylor expansion to model $L(\mathbf{\Lambda})$:

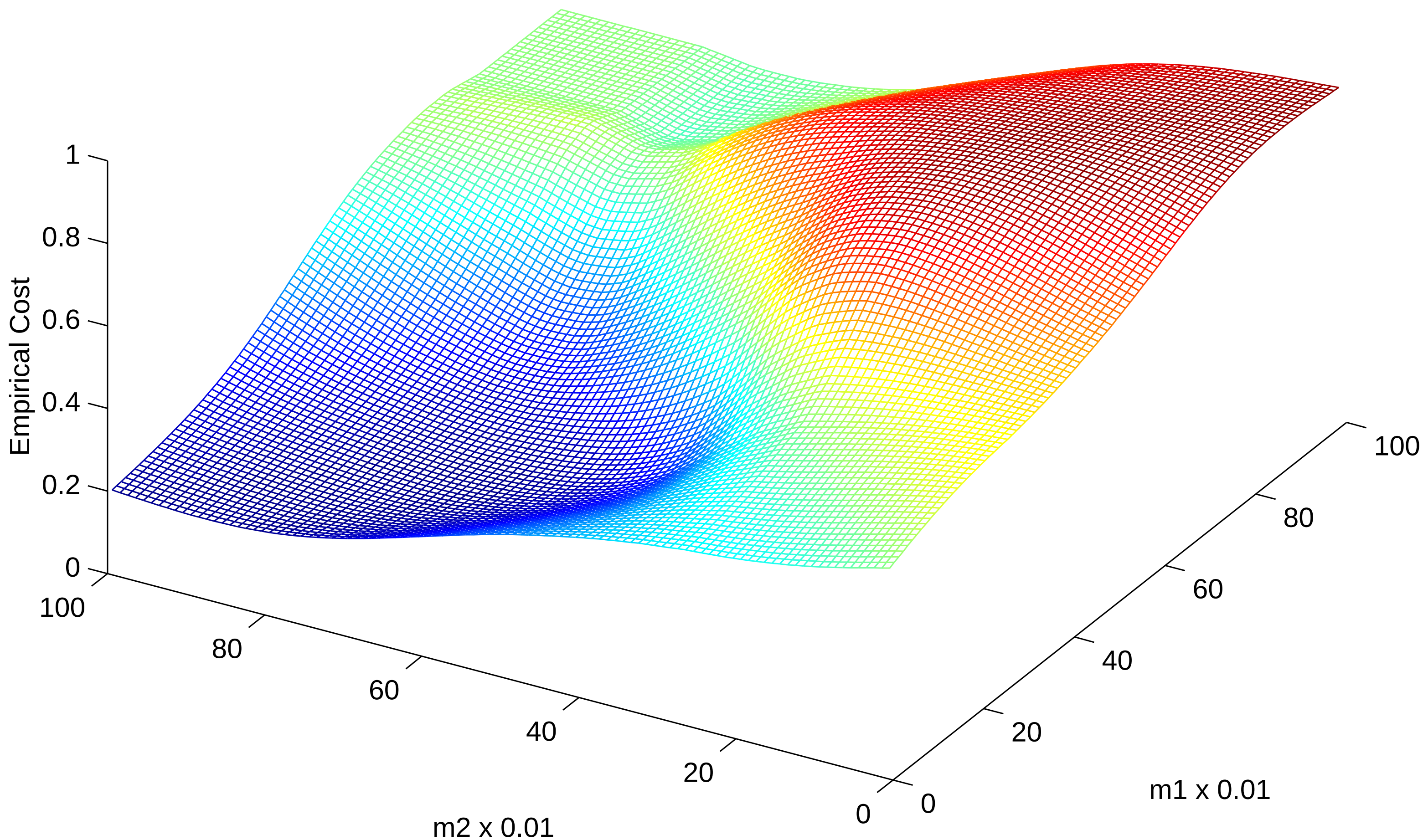
$$L(\mathbf{\Lambda} + s) \approx M(\mathbf{\Lambda} + s) = L(\mathbf{\Lambda}) + \nabla L(\mathbf{\Lambda})^t s + \frac{1}{2} s^t \nabla^2 L(\mathbf{\Lambda}) s$$

  - calculate the step size that moves to the minimum of the model

Actual classification error



Smoothed MCE loss function



MCE & MMI

- Unified framework for MCE & MMI (Schlueter, 1998):

$$\mathcal{F}(\mathbf{\Lambda}) = \frac{1}{R} \sum_{r=1}^{R} f \left(\log \frac{p_\Lambda(X_r|S_r)^\psi P(S_r)^\psi}{\sum_{S \in \mathcal{M}_r} p_\lambda(X_r|S)^\psi P(S)^\psi} \right)$$

- Optimization :
  - Gradient-based methods
  - Extended Baum-Welch algorithm; see Kanevsky, 1995
    $\rightarrow$ see Macherey et al., Eurospeech 2005 for application to MCE
- MMI (= Cross-entropy) fails to separate: Gopalakrishnan et al. ICASSP 1988: “Decoder selection based on cross-entropies”

Minimum Phone/Word Error

- MPE, MWE (Povey, 2002):

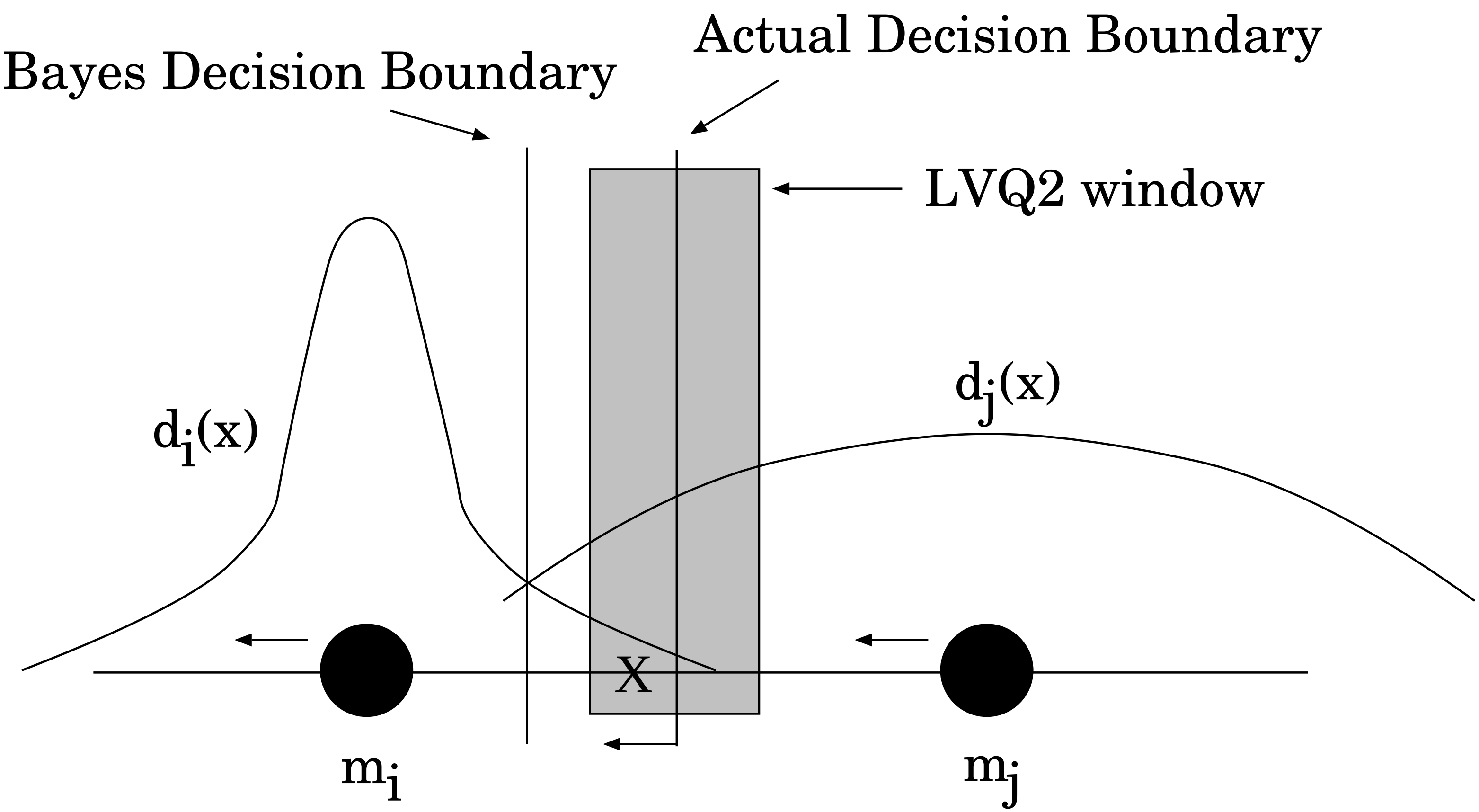
$$\mathcal{F}(\mathbf{\Lambda}) = \frac{1}{R}\sum_{r=1}^{R}\log\frac{\sum_S p_\Lambda(X_r|S)^\psi P(S)^\psi \mathcal{G}(S,S_r)}{\sum_S p_\lambda(X_r|S)^\psi P(S)^\psi}$$

- cf. unified framework for MCE & MMI (Schlueter, 1998):

$$\mathcal{F}(\mathbf{\Lambda}) = \frac{1}{R}\sum_{r=1}^{R} f\left(\log\frac{p_\Lambda(X_r|S_r)^\psi P(S_r)^\psi}{\sum_{S\in\mathcal{M}_r} p_\lambda(X_r|S)^\psi P(S)^\psi}\right)$$

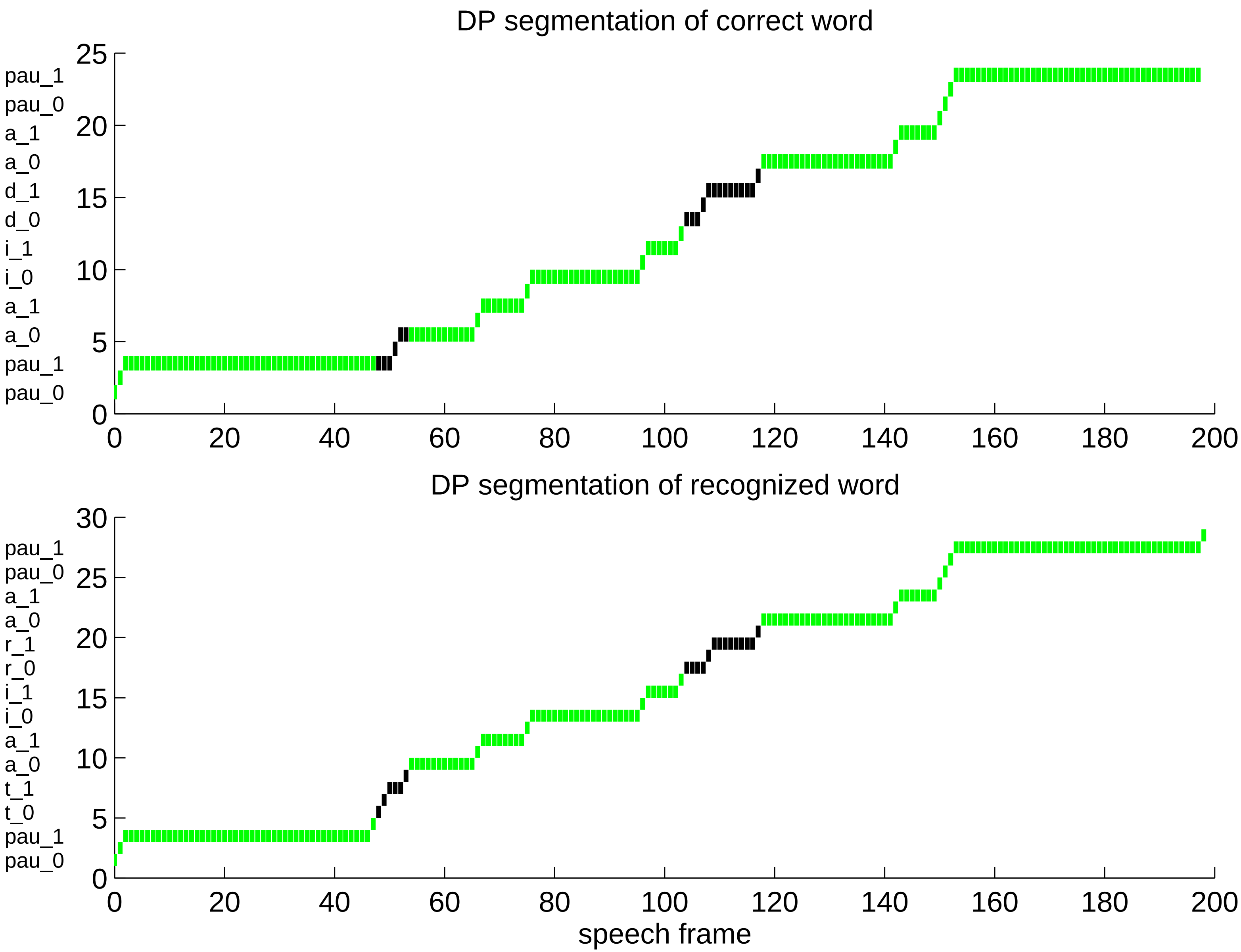
- See Macherey et al., Eurospeech 2005 for comparison between MCE, MPE and MMI on Wall Street Journal task.

LVQ = an application of MCE!

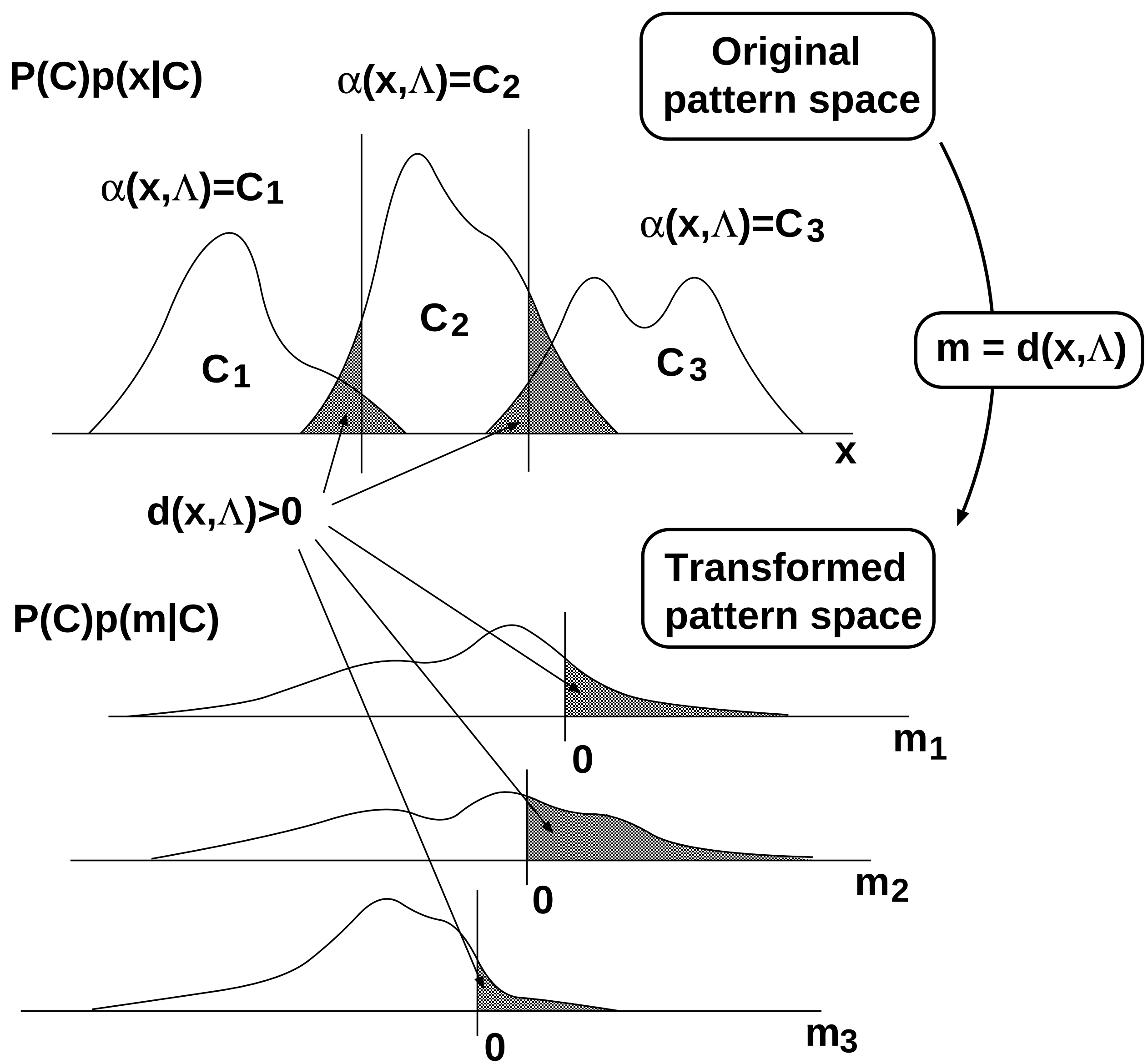


$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) - \alpha(t)(\mathbf{x}(t) - \mathbf{m}_i(t))$$
$$\mathbf{m}_j(t+1) = \mathbf{m}_j(t) + \alpha(t)(\mathbf{x}(t) - \mathbf{m}_j(t))$$

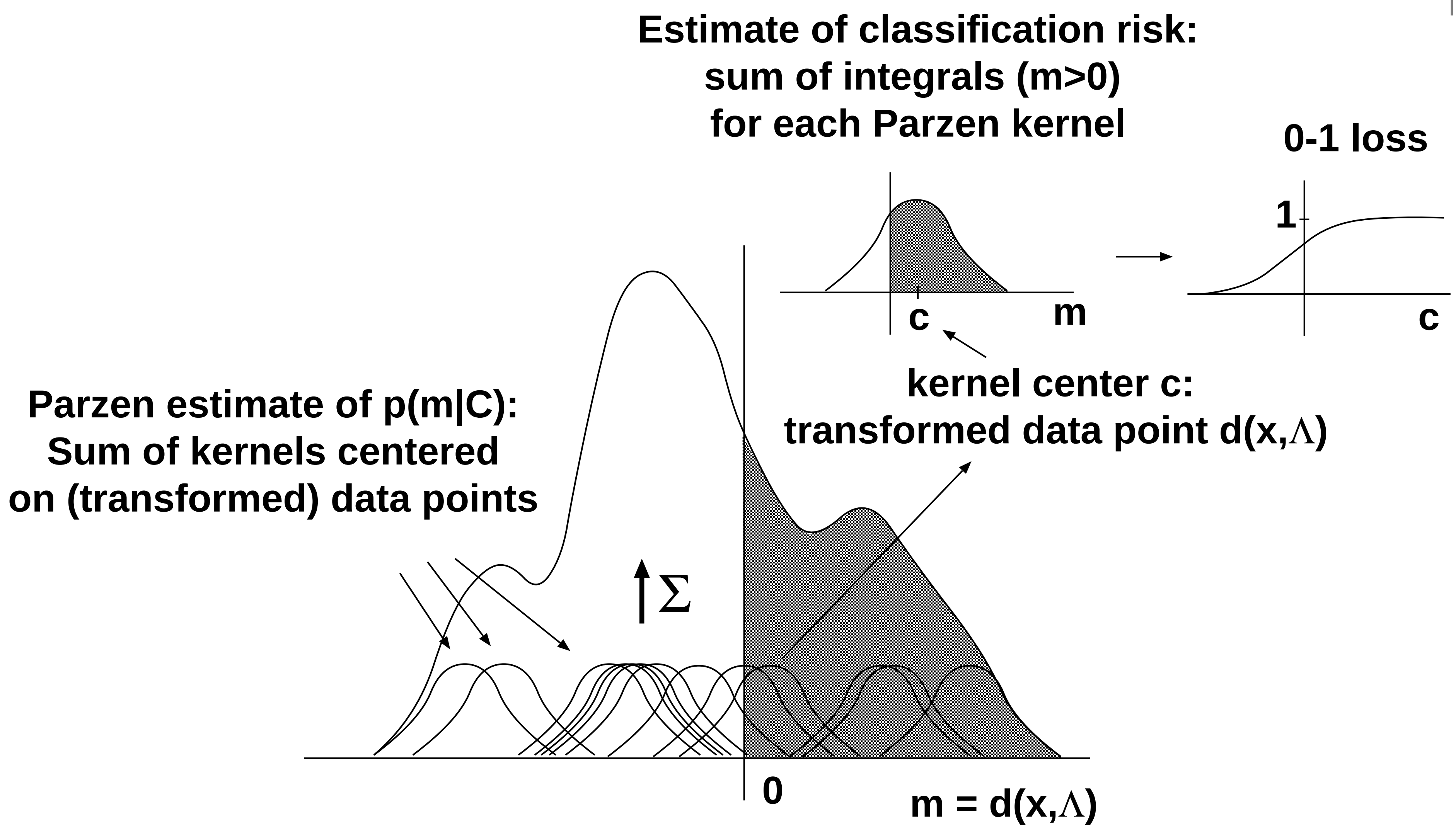
Gradient along correct & incorrect paths



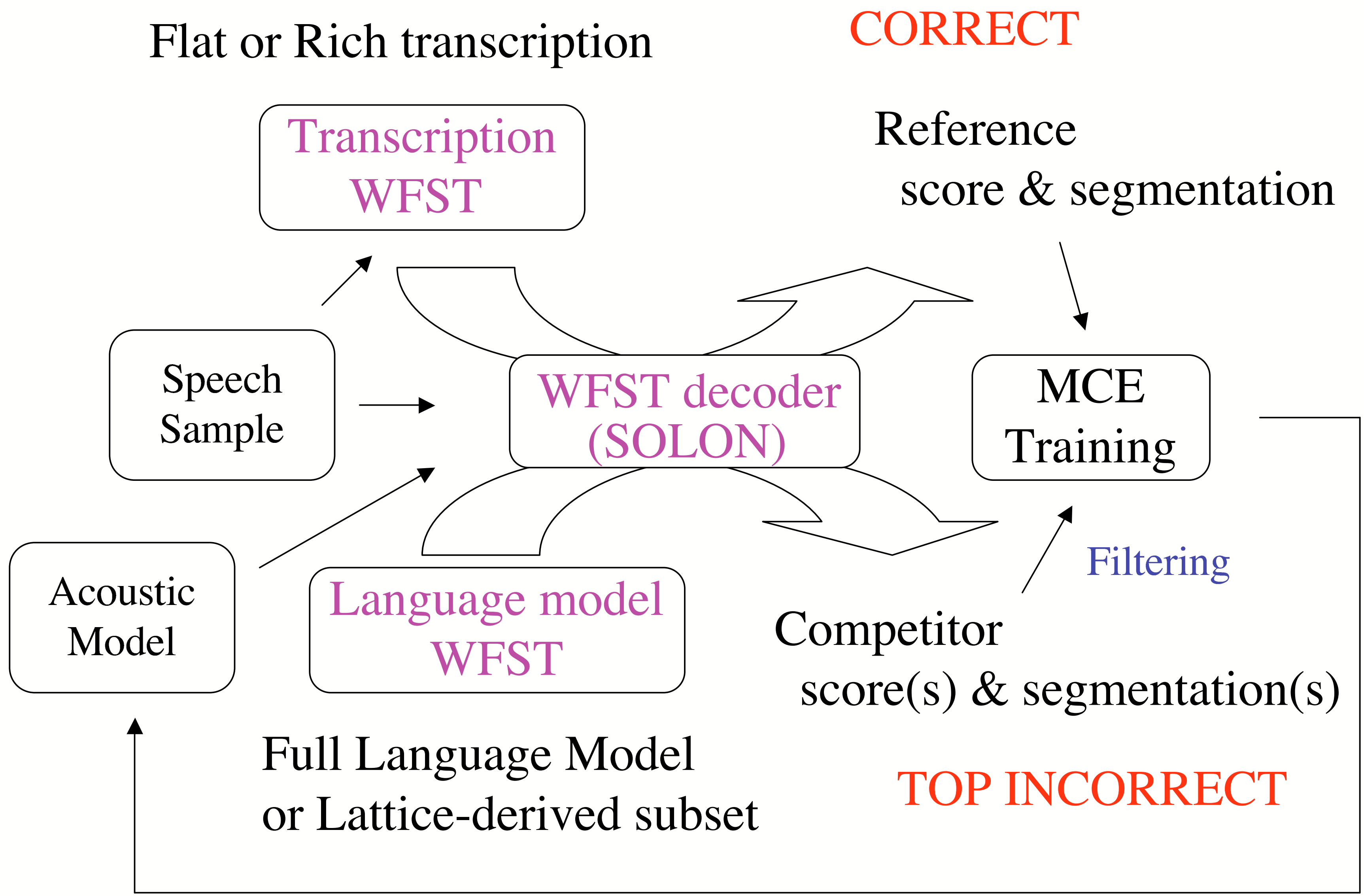
Defining risk in a new domain



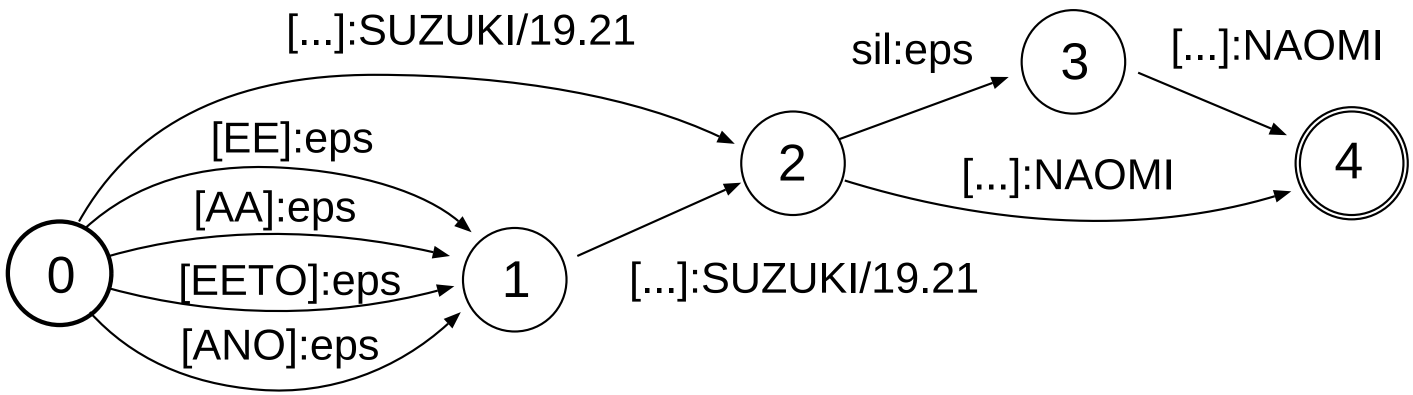
Parzen estimation of risk



WFST-based MCE Training



Flexible transcription model



- Define desired output as a *set* of strings, rather than a single string.
- Regular grammar model, represented as WFST.
- Use for MCE training:
  - Correct string $S_k \rightarrow$ correct string set, $\mathcal{K}$
  - Decoder finds best string within set $\mathcal{K}$ (with score and segmentation)

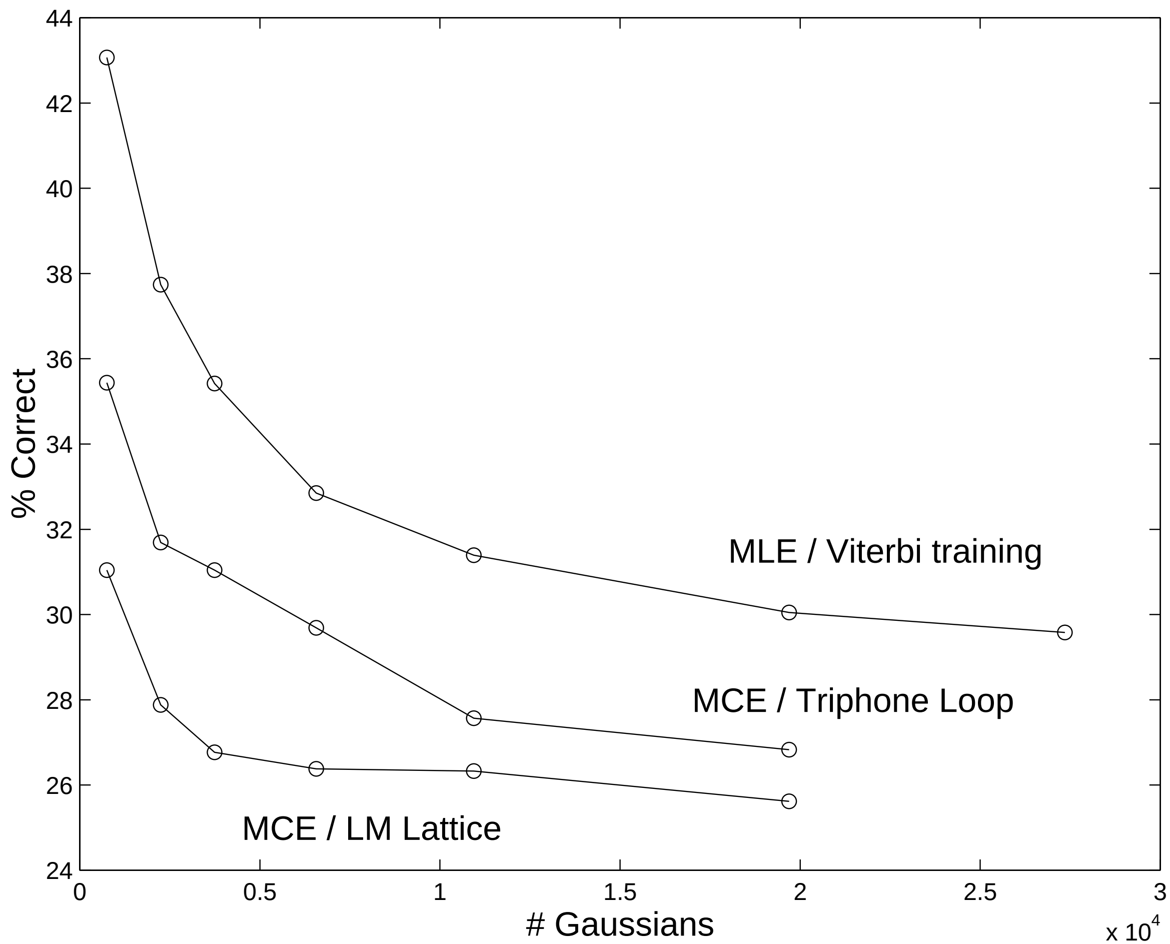
Telephone Based Name Recognition

(McDermott et al., ICASSP 2000, 2005)

- Task: telephone-based, speaker independent, open vocabulary **name recognition**
- Approx. 22,000 family & given names modeled
- Database:
  - > 35,000 training utterances (> 39 hours of audio)
- Evaluated:

1. ML / Viterbi Training vs. MCE training
2. Use of lattice-derived WFSTs to speed up training
3. “Strict” vs. “Flexible” transcription WFSTs

ML vs. MCE performance

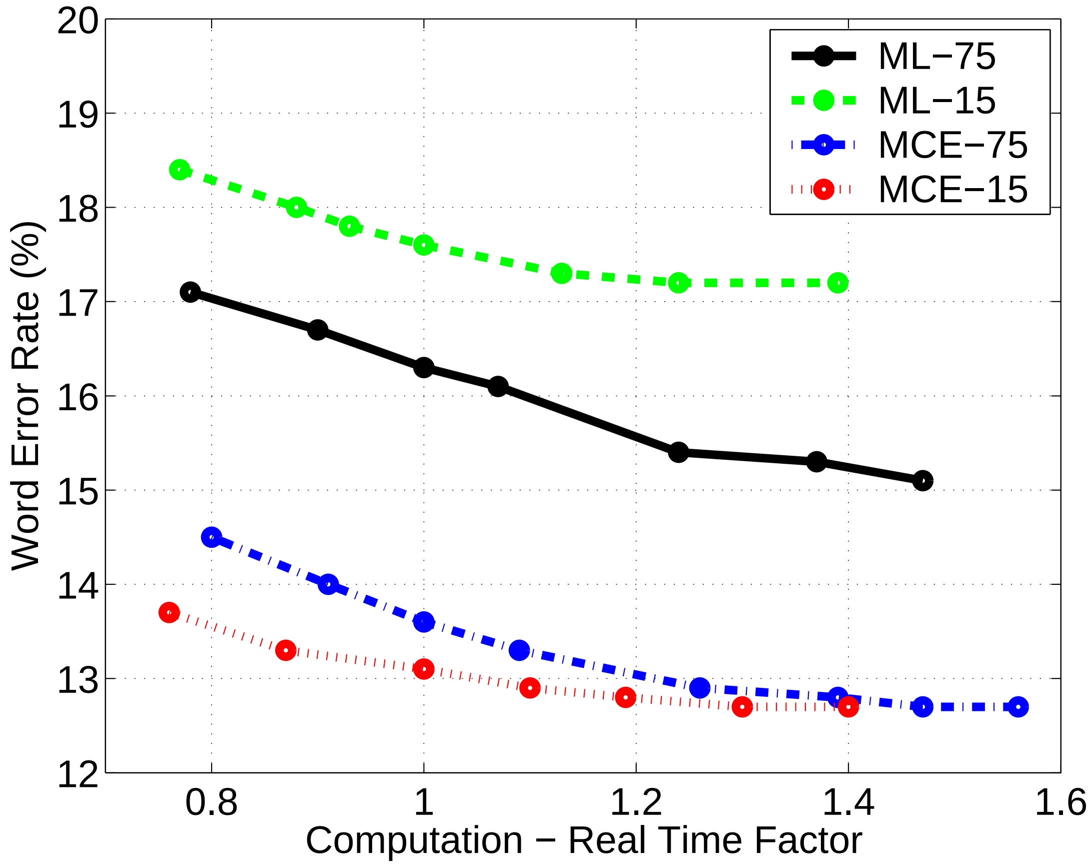


- Efficient use of parameters via MCE training
- Huge gains in system compactness

MCE for JUPITER/SUMMIT system

- Few MCE results for large, real-world tasks [McDermott, ICASSP 2000] (but MMI results for SWITCHBOARD [Woodland, 2002]).
- McDermott & Hazen, ICASSP 2004: evaluated application of MCE to MIT's online weather information system, **JUPITER**, based on **SUMMIT** recognition system
- Basic finding: for fixed real-time factor of 1.0, small models trained with MCE yielded a 20 % relative reduction in word error on in-vocab test set.

ML vs. MCE experiments on JUPITER

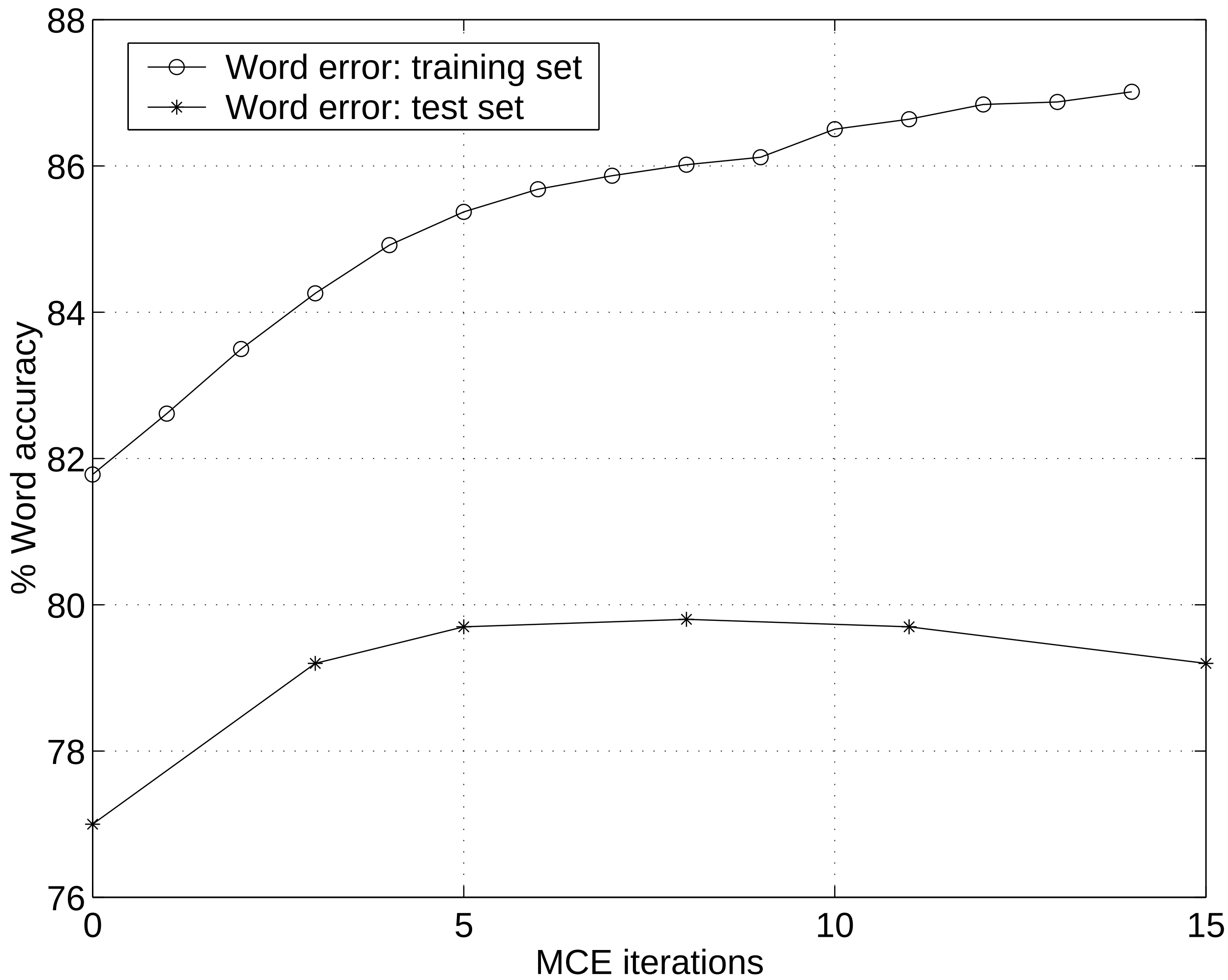


- MCE training of large/small models: 41777 gaussian pdfs (**MCE-75**) vs. 15245 gaussian pdfs (**MCE-15**); comparison with corresponding ML models.

Corpus of Spontaneous Japanese

- Task: lecture speech transcription (Kawahara, 2003)
  - $>$ 180,000 training utterances ($\approx$ 230 hours of audio)
    - $\approx$ 84,000,000 training vectors of 39 dimensions.
  - 10 test speeches ($\approx$ 2 hours of audio)
- Le Roux & McDermott, Eurospeech 2005
  - Evaluated different optimization methods (Quickprop, Rprop, BFGS, Probabilistic Descent)
- More Recent work:
  - MCE training with 68K unigram WFST (no lattices)
  - MCE training with 100K trigram WFST
  - Testing using 100K trigram LM
- Relative Word Error Rate reduction $\approx$ 9-12 %

Course of training - 100k words



- Optimization via Rprop

Recent CSJ results - 1

- MCE training with 68k unigram
- Testing with 30k word trigram
- Evaluate use of different HMM topologies

# States	# Gssns	ML-v30k	MCE-v30k
2000	16	23.4	22.3
2000	32	22.4	21.0
3000	8	24.1	22.5
3000	16	23.1	20.8
4000	16	22.8	20.8

Recent CSJ results - 2

- Same as before, but test with 100k word trigram

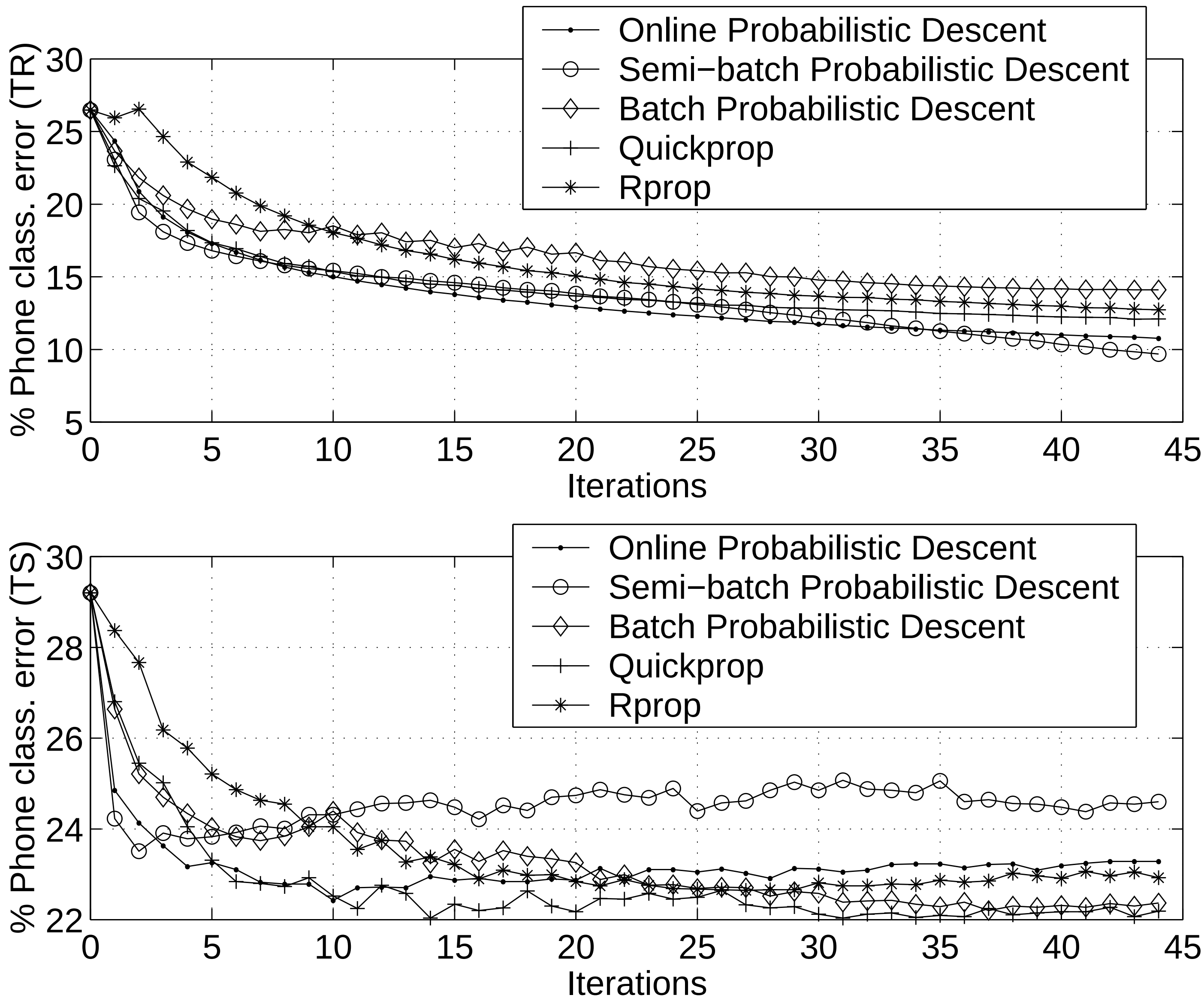
# States	# Gssns	ML-v100k	MCE-v100k
2000	16	23.0	21.1
2000	32	21.7	20.5
3000	8	-	21.1
3000	16	22.4	20.5
4000	16	22.1	20.1

Recent CSJ results - 3

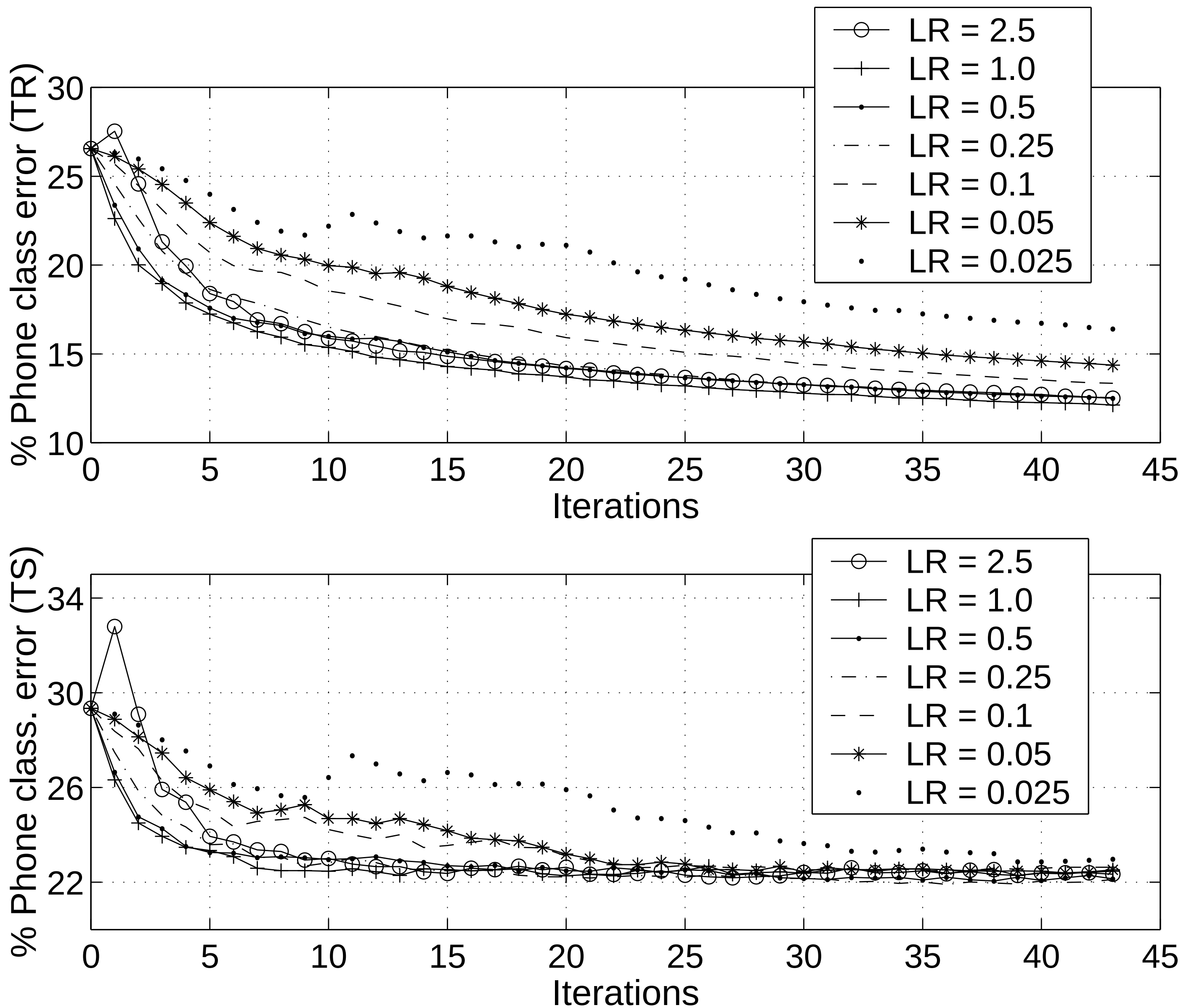
- Now train with 100k trigram LM
  Note: training and testing LMs are now matched

# States	# Gaussians	ML-v100k	MCE-v100k
2000	16	23.0	20.2
3000	16	22.4	20.5
4000	16	22.1	20.0

MCE for TIMIT phone classification



Sensitivity to Quickprop learning rate?



Summary

- MCE incorporates classification performance itself into a differentiable functional form.
  - By directly attacking the problem of interest, parameters are used efficiently.
- Large gains in performance and model compactness on challenging speech recognition tasks.
  - Telephone-based name recognition
  - MIT JUPITER weather information
  - Corpus of Spontaneous Japanese lecture speech transcription