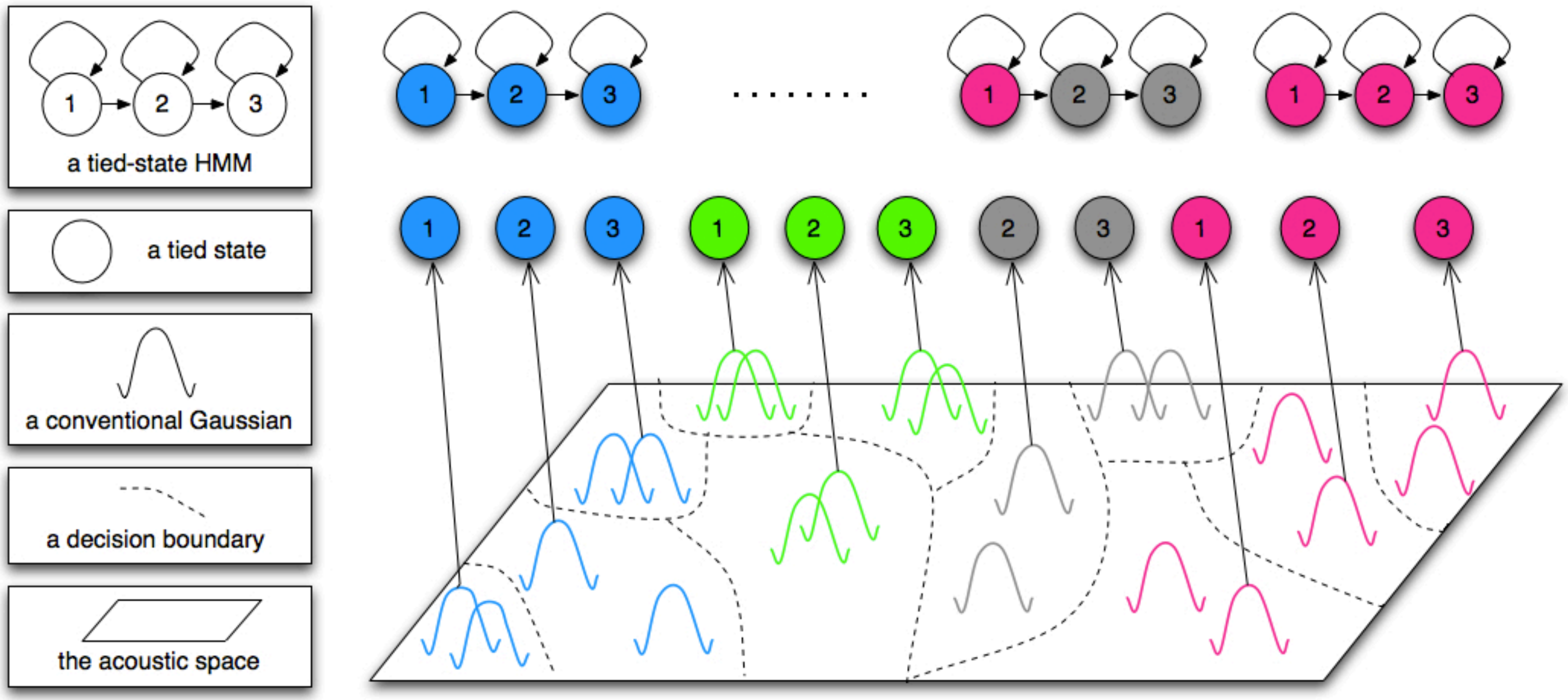
Discriminative Dynamic Gaussian Mixture Selection for Accented Speech Recognition

Zhang Chao

CSLT, RIIT, Tsinghua University

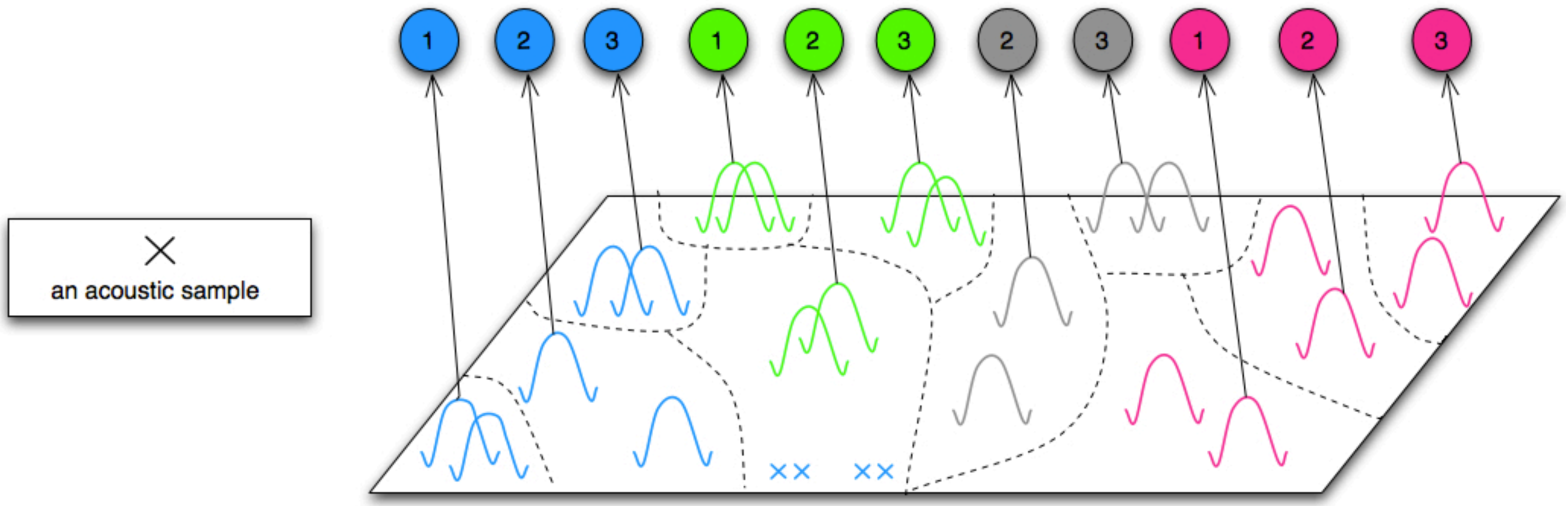
Principle of Tied-State HMMs

- Tied-sate HMMs are usually used in context-dependent acoustic modeling



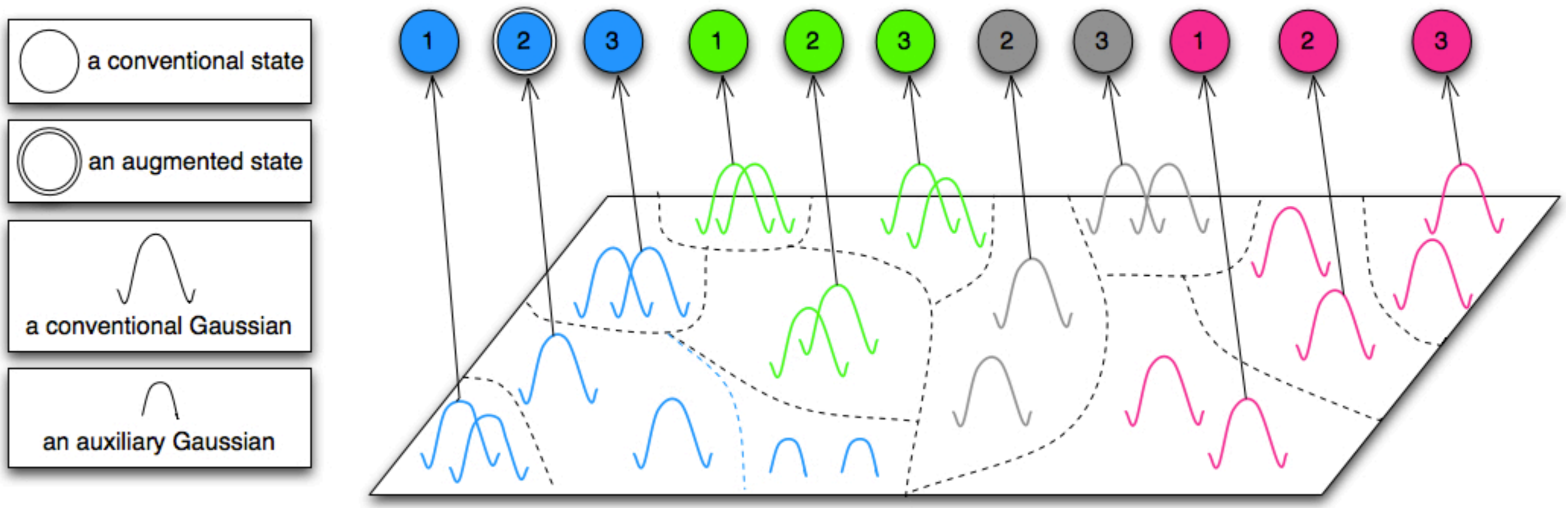
How accents impact on pronunciations?

- Accents cause pronunciation changes (‘*n*’ changes to ‘*l*’) that yield relevant acoustic samples shift to unexpected subspaces
- e.g. Samples belong to ‘blue’ locat at a subspace of ‘green’



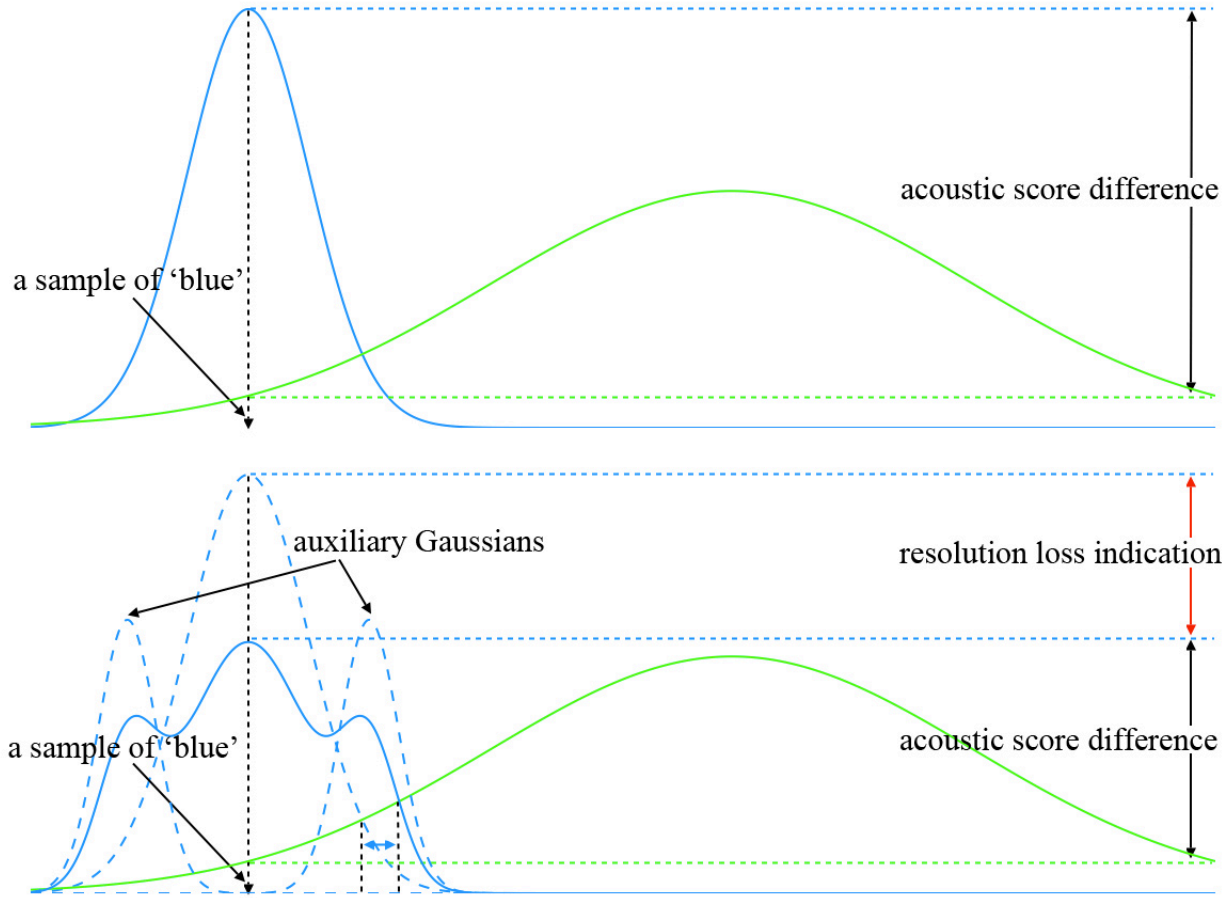
Handle Accent Changes by Model Augmentation

- Model Augmentation uses extra (auxiliary) Gaussians trained by accented samples to cover accent changes
- e.g. the subspace of ‘green’ with accented samples belong to ‘blue’ is covered by auxiliary Gaussians of ‘blue’



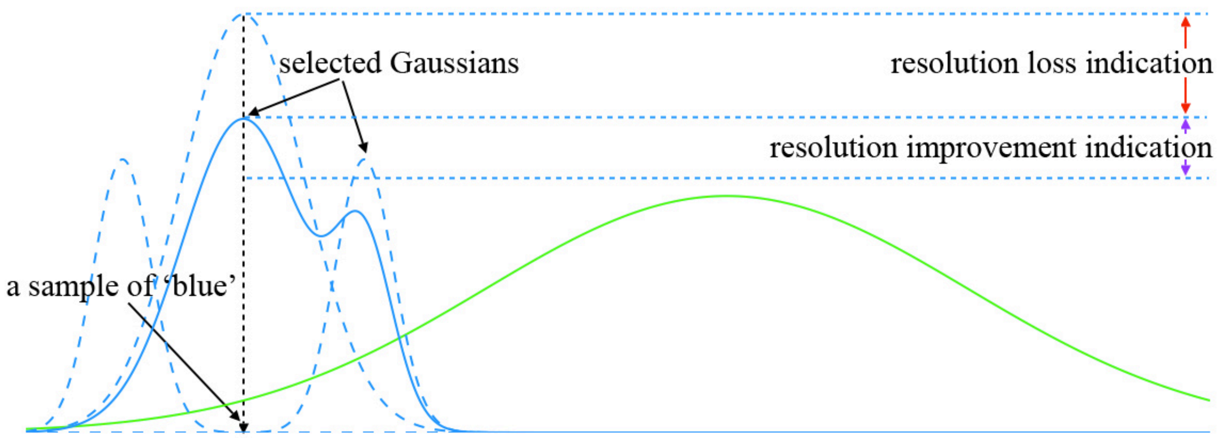
Side-Effect of Model Augmentation

- Model Augmentation reconstructs tied-states statically by borrowing auxiliary Gaussians and degrades model resolution



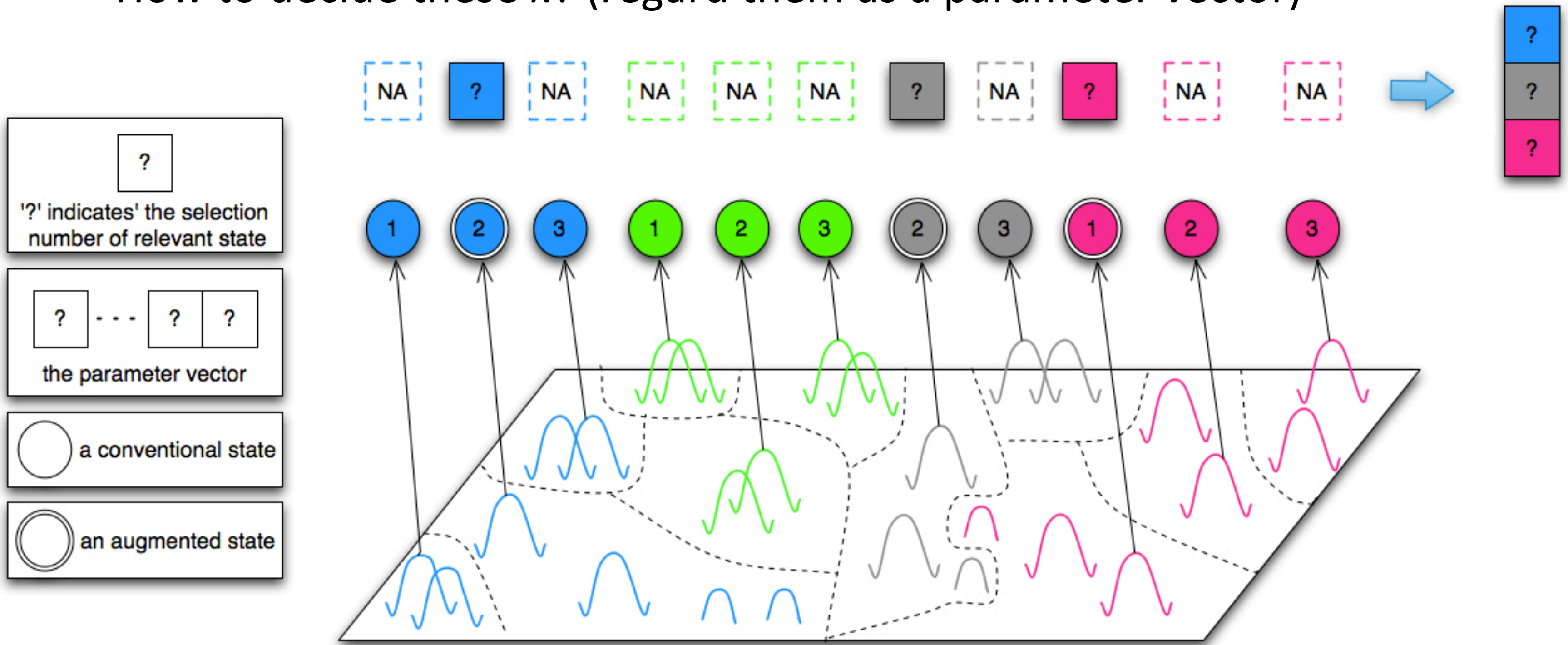
Dynamic Gaussian Mixture Selection (DGMS)

- Since the augmented tied-states suffer from resolution ability loss while the left ones do not, model augmentation leads to serious performance degradation in pruned beam search
- We dynamically reconstruct the statically reconstructed tied-states by selecting *k* Gaussians nearest to current input frame to compute its acoustic likelihood



A Key Issue of DGMS

- In DGMS method, each statically reconstructed tied-state has a selection number *k* used to dynamically reconstructed it
- Hundreds of *k* exist in a set of acoustic models
  - How to decide these *k*? (regard them as a parameter vector)



Discrete Variable Optimization

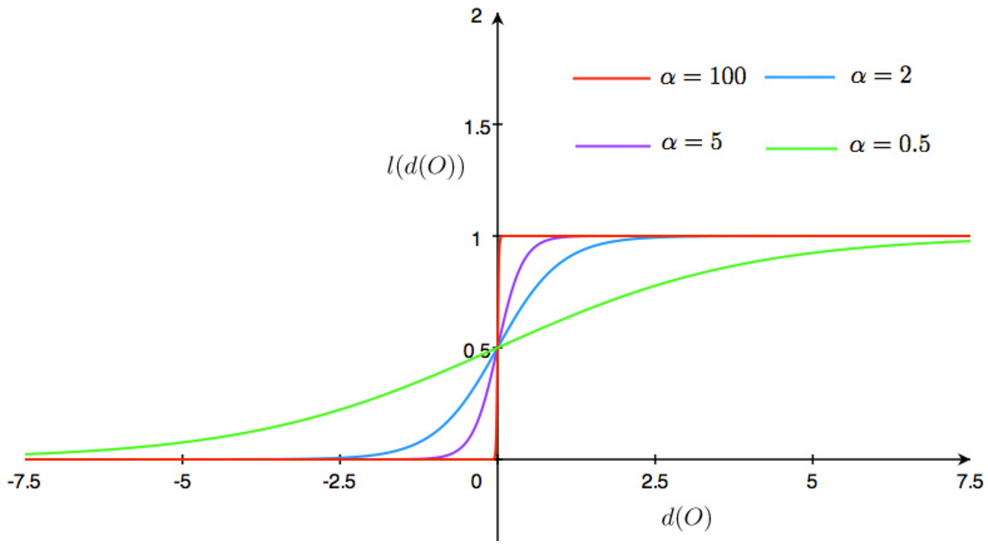
- Deciding the parameter vector is actually a discrete variable optimization problem
- The optimization criterion
  - Discriminative criteria (MCE, MPE, MMIE *etc.*) always outperforms generative criteria (ML, MAP *etc.*)
  - We choose MCE criterion, which leads to discriminative DGMS
- The optimization algorithm
  - The selection numbers are discrete variables with no derivatives that causes traditional optimization methods unsuitable (EM, BFGS *etc.*)
  - There are exponentially possible vectors and are hard to be exhausted
  - We hire GA to heuristically accelerate the optimization

Minimum Classification Error Criterion (MCE)

- MCE minimizes an estimation of train-set sentence errors

$$l(d(O)) = 1 \Big/ \left(1 + e^{-\alpha \cdot d(O)}\right)$$

- MCE embeds acoustic score difference between competing and target models $d(O)$ into a sigmoid function to
  - Count sentence errors by different degrees
  - Guarantee derivatives of the parameters exist (unnecessary for us)

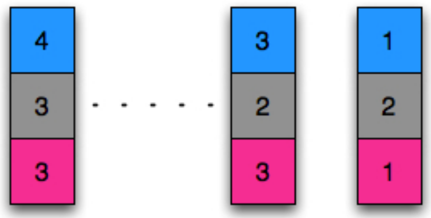


Genetic Algorithm (GA)

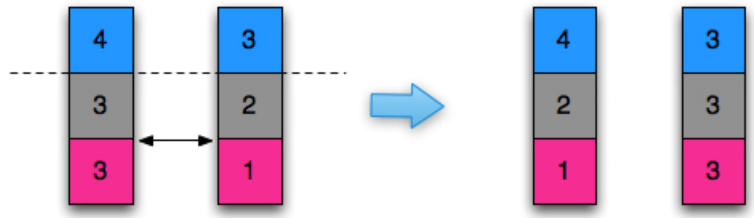
- GA is a random “search for solution” algorithm
  - Mimics the survival of the fittest process of natural evolution
  - To find the optimum by examining over only a small fraction of the possible candidates
- Employ GA to find the optimal selection numbers
  - See each parameter vector as a possible candidate (a chromosome)
  - We use $l(d(O))$ as the fittest-function to evaluate the candidates: the fitter the candidate, the smaller its fittest-function
  - $l(d(O))$ is computed on the acoustic models with DGMS enabled using current parameter vector; the competing string is approximated with the best one among N-Best hypotheses

Essential Steps of GA

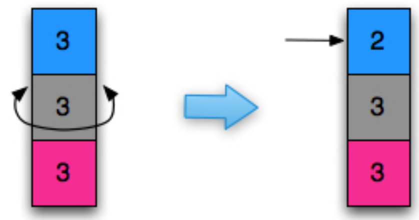
- Maintaining a population of candidates



- Generating new candidates randomly from the population by
  - Mating: fitter candidates have the priority to mate



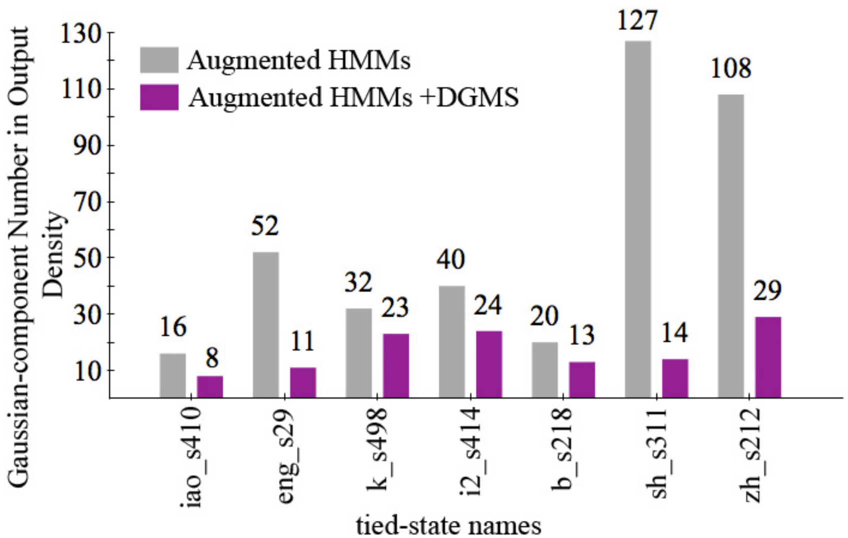
  - Mutation

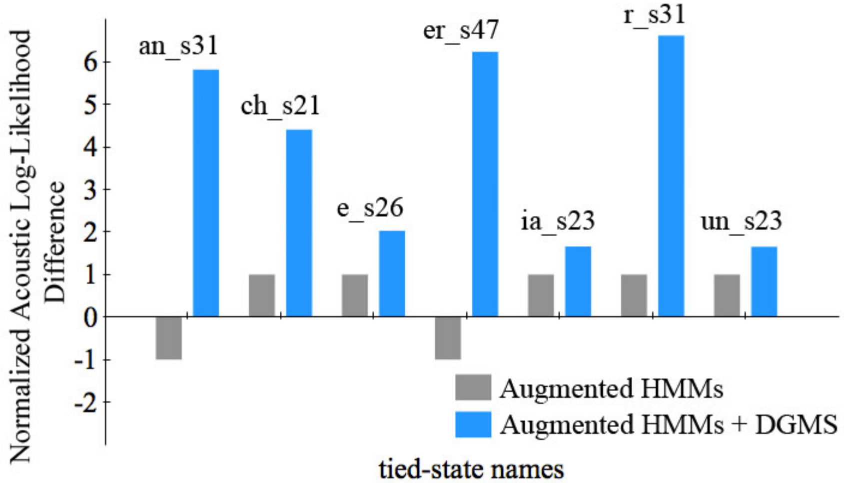


- Updating the population; repeating above steps; candidate with the smallest fittest-function value is the optimum

Experiments

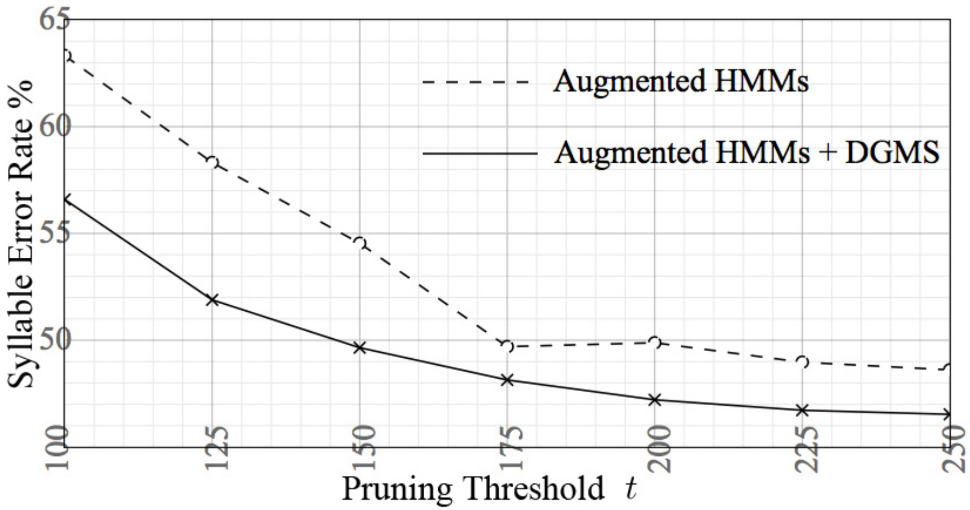
- We evaluate discriminative DGMS on multi-accent speech recognition task (with three accents: Chuan, Yue, and Wu)
  - Gaussian mixture model sizes in state observation densities for some representative tied-states with/without DGMS (the left figure)
  - Normalized relative model resolution for representative tied-states with/without DGMS evaluated on Yue accent (the right figure)





Experiments (Cont.)

– DGMS alleviates performance loss in pruned beam search



– Discriminative DGMS reduces free grammar syllable recognition error rate (without pruning search)

Data Type	Chuan	Yue	Wu	PTH
Relative Syllable Error Rate Reduction	3.63%	4.04%	3.96%	-0.32%

What's More …

- Many omitted details and algorithms are presented in our paper
  - "*Discriminative Dynamic Gaussian Mixture Selection with Enhanced Robustness and Performance for Multi-Accent Speech Recognition*"
- The performance on sentence level (not given here) are even significantly better than that of syllable level
  - 11% relative error rate reduction without pruning; 33% when t = 300
- Whether discriminative DGMS works on conventional HMMs and the relationship between continuous and this discrete variables discriminative HMMs are not yet clear
- Other criteria, more complex control strategy (other than the *k*-NN), and context-dependent selection may also work

Thank for your listening!