Clustering of Bootstrapped Acoustic Model with Full Covariance

Xin Chen[1], Xiaodong Cui[2], Jian Xue[2], Peder Olsen[2], John Hershey[3], Bowen Zhou[2] and Yunxin Zhao[1]

Prague, Czech Republic

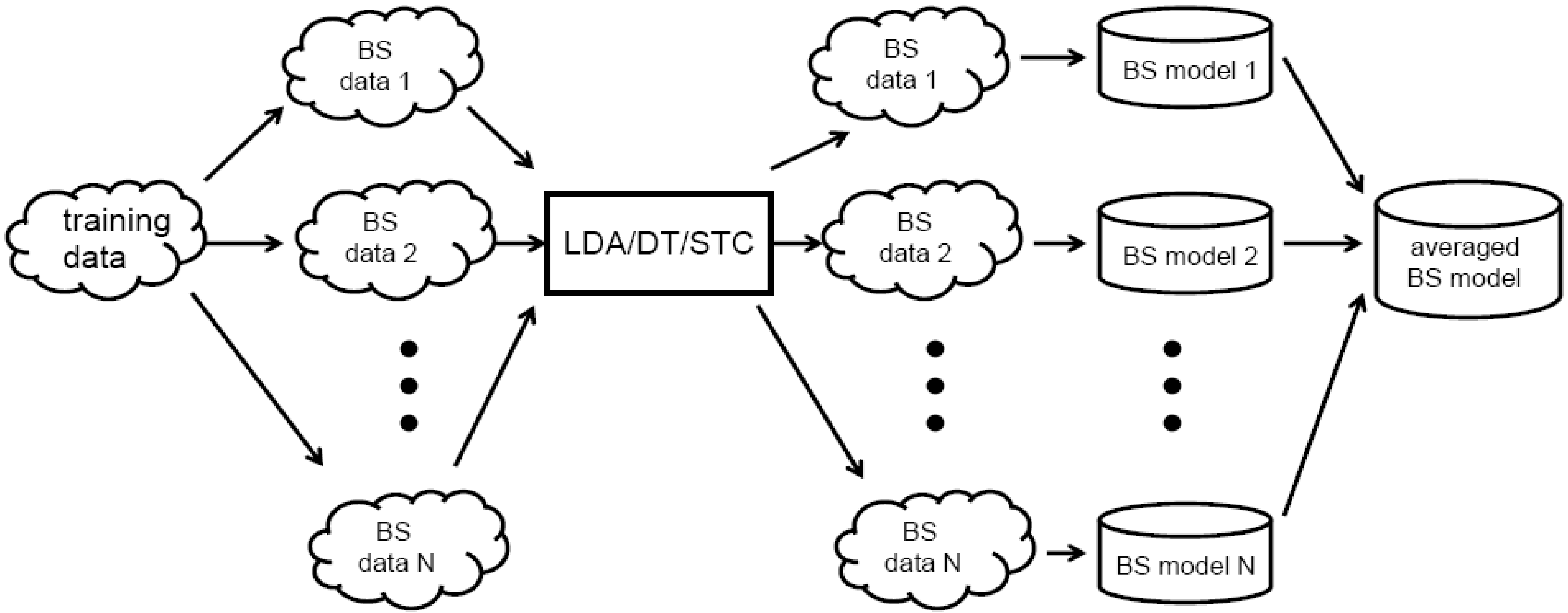
5.26.2011

SLIPL Lab, University of Missouri[1]
IBM T. J. Watson Research Center[2]
Mitsubishi Electric Research Laboratories[3]

Outline

- Overview of Bootstrap and Restructuring (BSRS) acousticmodeling
- Motivation
  - Why clustering?
  - Why full covariance?
- How to do the clustering?
  - Distance (similarity) measurements Investigated
    - Entropy, KL, Bhattacharyya, Bayes error, Chernoff
  - Clustering Algorithms proposed and Investigated
    - N-Best distance Refinement (NBR)
    - Global optimization
    - Model structure optimization
- Experimental results on proposed clustering methods
- Experimental results on BSRS with full covariance
- Future extensions

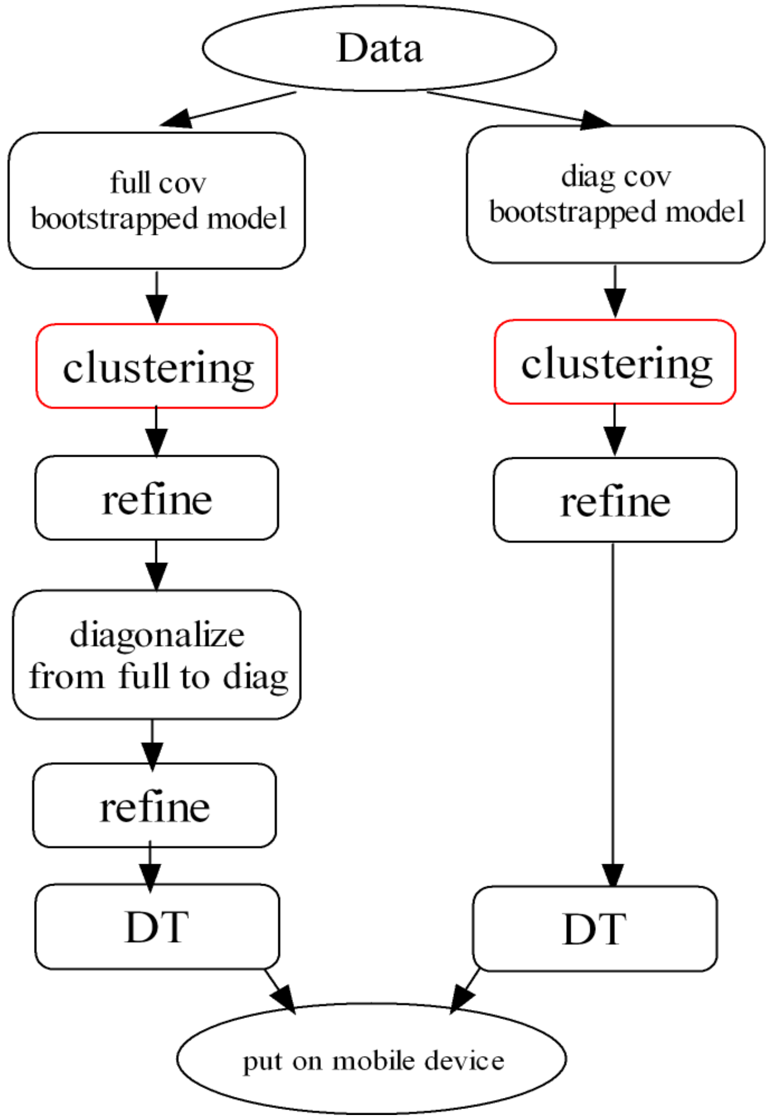
Bootstrap Based Acoustic Modeling



- Bootstrap the original training data S into N subsets $\{S_1, S_2, \cdots, S_N\}$ without replacement.
- Each subset covers a fraction of the original data $S_i = r \cdot |S|$.
- Combine all the subsets for training of LDA, decision tree and STC (therefore shared LDA/DT/STC and single graph in decoding).
- Perform EM training in parallel on N subsets for N HMMs.

Bootstrap and Restructuring (BSRS) with full covariance (1)

- Aggregated N BS Acoustic model
  - Performs very well
  - Too Large and restructuring is needed
- 1. BS+Diag strategy
  - Train diagonal covariance model in all steps
- 2. BS+Full→Diag strategy
  - Keep all the info until the last step
  - Train full covariance up to the last steps
    - Full covariance clustering needed



Bootstrap and Restructuring (BSRS) with full covariance(2)

- Clustering is a critical step
  - Remove the redundancy
  - Scale down the model (able to put on mobile device)
  - Flexible
    - Train large model and scale down to desirable size
  - Full covariance clustering
    - Needed for BS+Full→Diag strategy

Distance Measurements for Clustering (1)

- Entropy
  - measures the change of entropy after two distributions are merged
- KL divergence
  - KL divergence

$$D_{\mathrm{kl}}(f_1 \parallel f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx$$

  - Symmetric KL divergence

$$D_{\mathrm{kls}}(f_1 \parallel f_2) = \int \left[f_1(x) \log \frac{f_1(x)}{f_2(x)} + f_2(x) \log \frac{f_2(x)}{f_1(x)} \right] dx$$

- Bhattacharyya

$$D_{\mathrm{bhat}}(f_1 \parallel f_2) = \int \sqrt{f_1(x), f_2(x)} dx$$

Distance Measurements for Clustering (2)

- Bayes error $D_{\text{bayes}}(f_1 \parallel f_2) = \int \min(f_1(x), f_2(x))dx$
  - measures the overlap of two distributions.
  - No closed-form even for multivariate Gaussians.
  - A variational approach is applied based on the Chernoff distance.

- Chernoff distance
  - Chernoff function can be viewed as variational way to measure the Bayes error, the Chernoff distance is defined as

$$D_{\text{chern}}(f_1 \parallel f_2) = \underset{0 \le s \le 1}{\operatorname{argmin}} \int f_1(x)^s f_2(x)^{1-s} dx$$

  - Note that the Bhattacharyya is Chernoff function with s =0.5

Distance Measurements for Clustering (3)

- Chernoff distance (Details elaborated in [2])

Let $c(s) = \log C(s)$, which can be computed as

$$c(s) = \log Z(s\theta_1 + (1-s)\theta_2) - s \log Z(\theta_1) - (1-s)\log Z(\theta_2)$$

$c(s)$ is a convex function of s. Apply Newton-Raphson algorithm

$$s_{k+1} = s_k - \frac{c'(s)}{c''(s)}$$

where $$c'(s) = \log \frac{Z(\theta_2)}{Z(\theta_1)} + \sum_{i=1}^{n} \left[\frac{u_i v_i + s u_i^2 - \frac{1}{2}\xi_i}{1 + s\xi_i} - \frac{\frac{1}{2}\xi_i (v_i + s u_i)^2}{(1 + s\xi_i)^2} \right]$$

$$c''(s) = \sum_{i=1}^{n} \left[\frac{u_i^2}{1 + s\xi_i} - \frac{2\xi_i u_i v_i + 2s\xi_i u_i^2 - \frac{1}{2}\xi_i^2}{(1 + s\xi_i)^2} + \frac{\xi_i^2 (v_i + s u_i)^2}{(1 + s\xi_i)^3} \right]$$

also has an analytical form for a derivative free approach.

$$c(s) = -\frac{1}{4} s(1-s)(\mu_1 - \mu_2)^{\mathsf{T}} \left[(1-s)\Sigma_1^{-1} + s\Sigma_2^{-1} \right] (\mu_1 - \mu_2) - \frac{1}{2} \log \left[\frac{|(1-s)\Sigma_1 + s\Sigma_2|}{|\Sigma_1|^{(1-s)} + |\Sigma_2|^{(s)}} \right]$$

Outline of Investigated Algorithms

- Investigated Algorithms

- Bottom-up
  - Greedy
    - N-Best distance Refinement
      - To improve the speed
  - Non-Greedy
    - K-step look ahead
    - Search the best path
      - For global optimization

- 2-Pass strategy to improve model structure

Bottom-up Approaches

$$f(x) = \sum_{i=1}^{M} w_i \mathcal{N}(x; \mu_i, \Sigma_i) \quad \text{where} \quad M = \sum_{i=1}^{T} K_i$$

$$g(x) = \sum_{i=1}^{N} w_i \mathcal{N}(x; \mu_i, \Sigma_i)$$

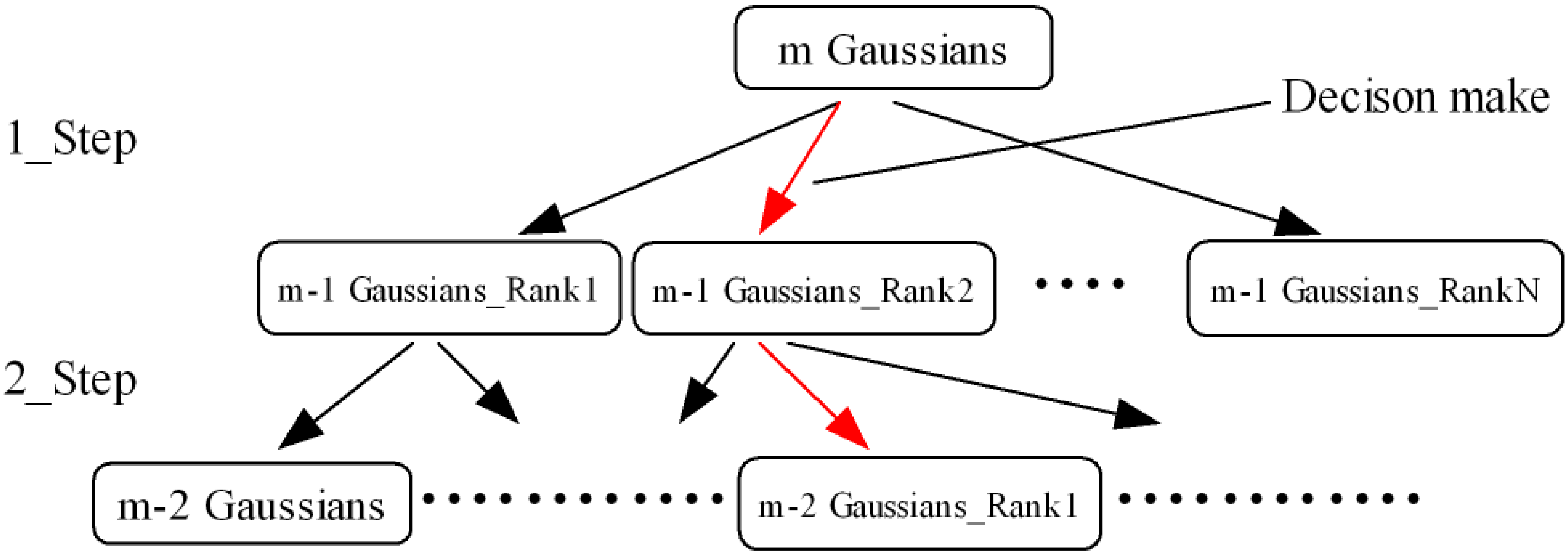
- bottom-up strategy
  - every time the two most similar Gaussians [Gaussian f_a and Gaussian f_b] are combined to one under certain criterion.

$$D(f,g) = \sum_{i=1}^{M-N} Distance_i(f_a, f_b)$$

  - Minimize $Distance_i(f_a, f_b)$ (Greedy)
  - Minimize D(f,g) (Global optimization) [**Our Target**]

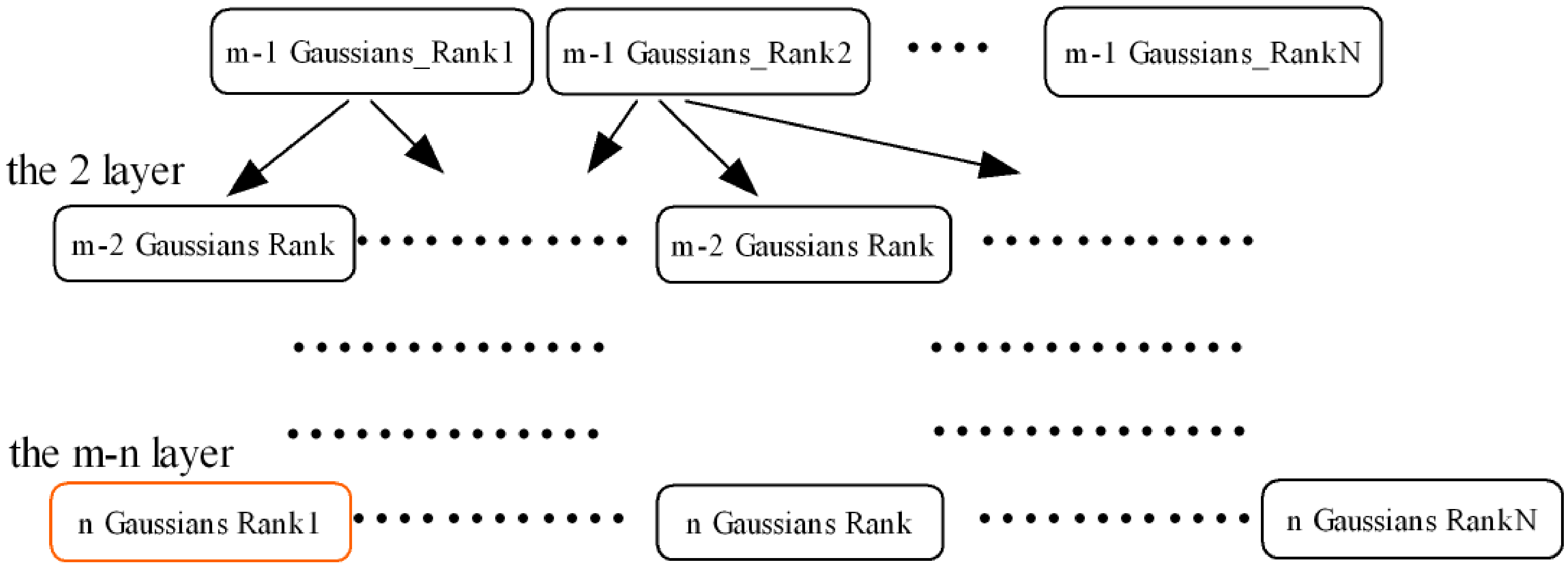
Global Optimization (1)

- K-step Look Ahead(KLA)



Global Optimization (2)

- Search the optimized path
  - Breadth First Search (BFS), when beam is set to N
    - Keep N candidates at each layer
    - Extend to next layer from N candidates



2-Pass Model Structural Refinement

- Original approach $S_i^{new} = S_i * \frac{N}{M}$
  - Every state has the same compression rate
- Every state can have a variable compression rate.
  - 2-Pass $(S_i * \frac{N}{M}) - K, .., S_i * \frac{N}{M}, .., (S_i * \frac{N}{M}) + K$
  - A Criteria is used to decide the compression rate from the candidates.
    - Bayesian Information Criteria [3]
    - Fixed BIC for all states, different compression rate.

Let $S = s_1, s_2, ...s_k$ be the current k cluster GMM, suppose we combine s_1, s_2 to s_1'. then we will have $S' = s_1', ...s_k$. The change from S to S' if measured with BIC

$$= -(w_1 + w_2) \cdot \log|\Sigma| + w_1 \cdot \log|\Sigma_1| + w_2 \cdot \log|\Sigma_2| + N(d + 0.5 * d(d+1))$$

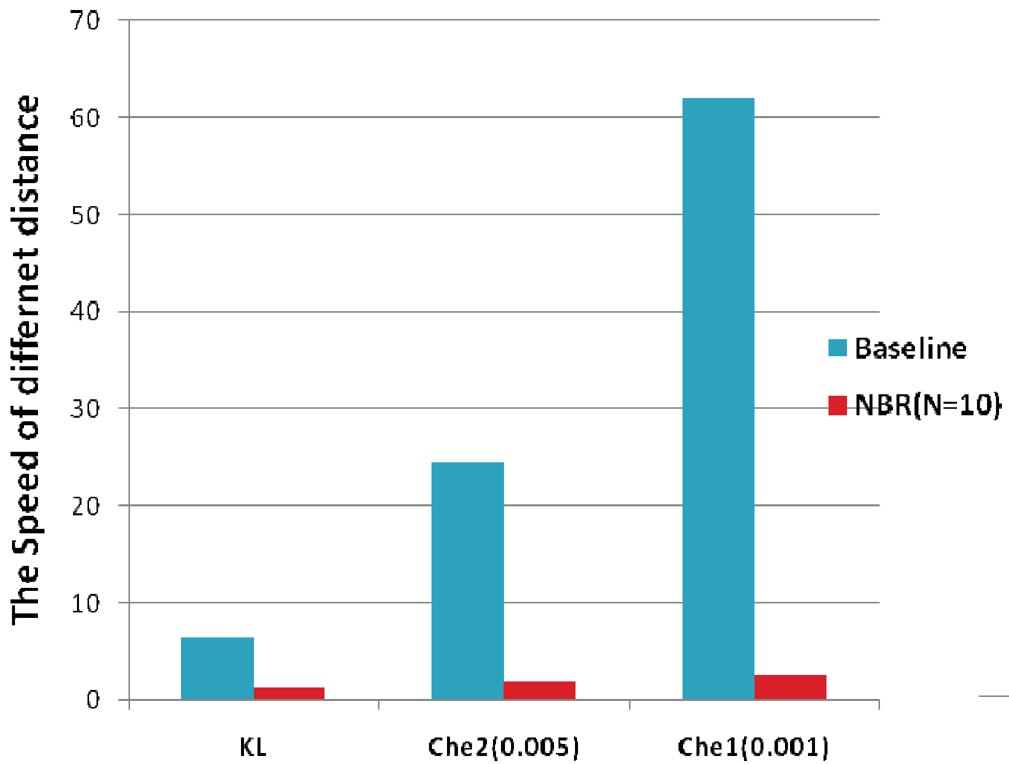
    - Entropy

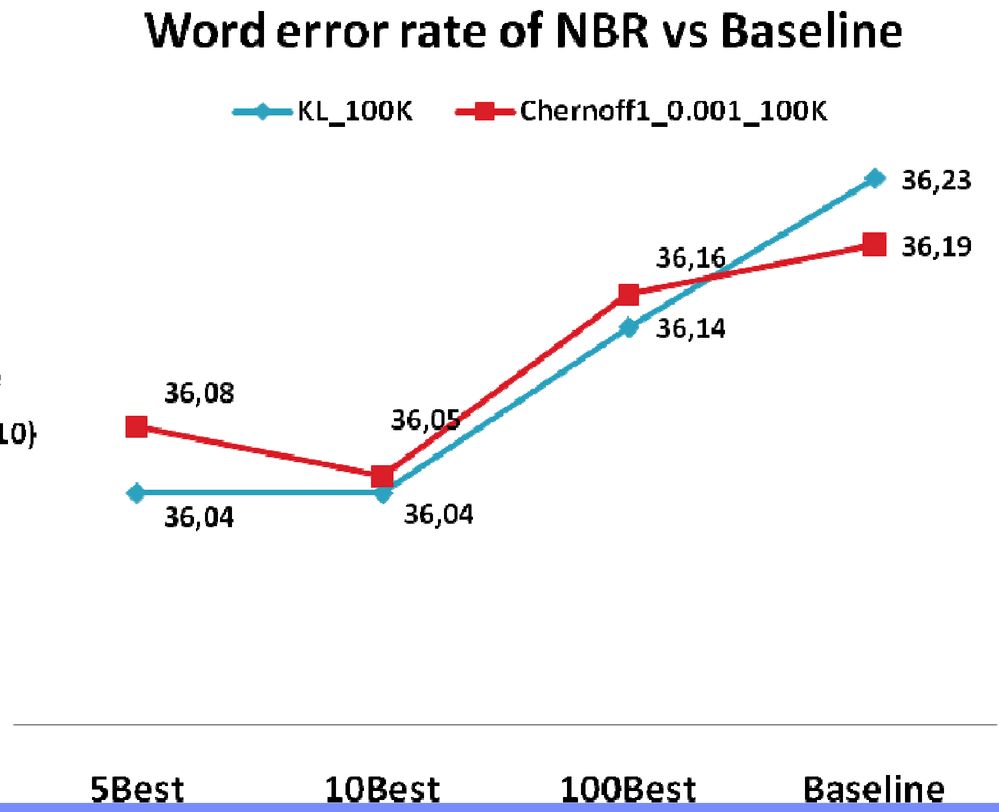
Experiments: ASR Setup

- Pashto data set(from TRANSTAC)
  - 135 hours of training data
  - 24 dimension PLP features
  - Speaker independent
  - Test set: 6896 sentences (10 hours)
  - Both training and testing data are spontaneous speech
  - 15 Bootstrapped model has 6K states and total 1.8M Gaussians
    - This big model has a WER of 35.46%

N-Best Distance Refinement (NBR)

- Chernoff and KL distance measures are slow to obtain
- Entropy (ENT) is fast and effective
  - Using ENT to find the N best candidate pairs
  - Using Chernoff/KL to recalculate the distances

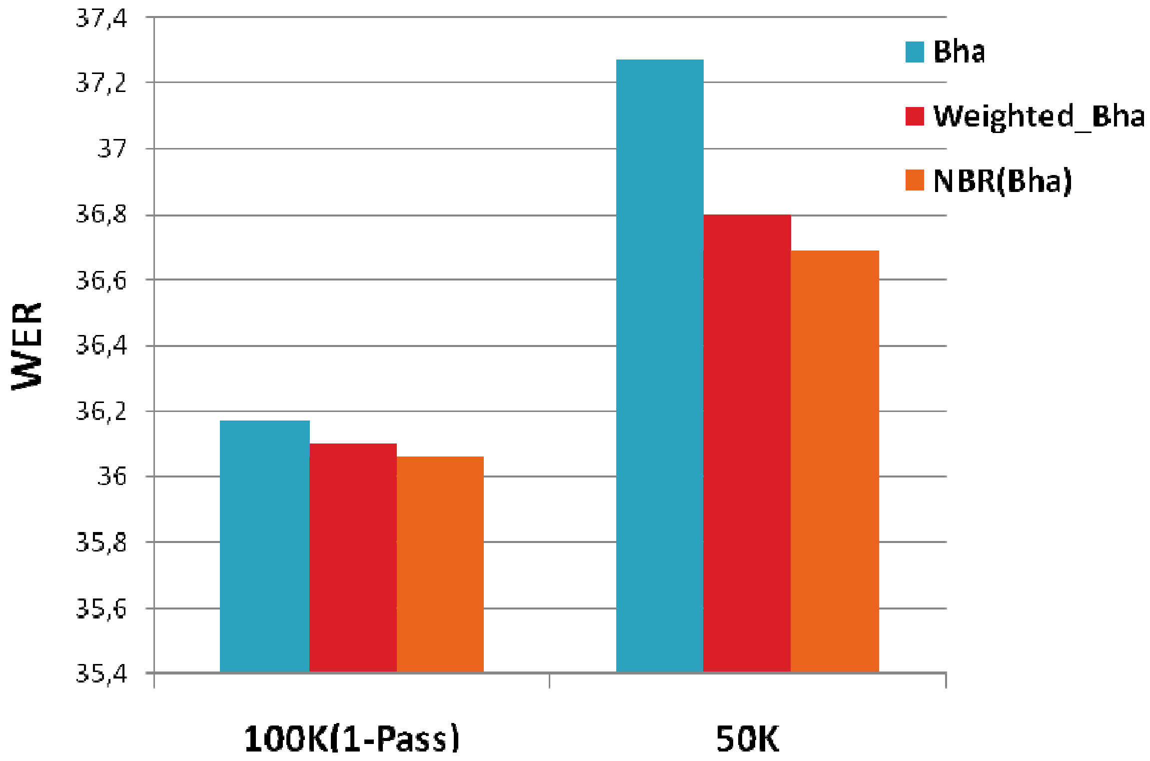




Weighted Distances

- The improvement in NBR suggest a weighted distance can be a potential improvement, as proposed in [4].
  - Evaluated on Weighted Bhatharraya distance

$$D_{\text{bhat}}(f_1 \parallel f_2) = \int \sqrt{w_1 f_1(x) w_2 f_2(x)} dx$$



Results for Global Optimization

100K	Baseline	2-step LA(10)	Search(2_4_8)
ENT Speed	1X	6.7X	23.8X
ENT_State0_D(f,g)	3336.8	3336.76	**3299.04**

WER 100K test	Baseline	NBR	2-step LA	Search
KL	36.23	**36.04**	36.11	36.14
ENT	36.11	N/A	36.08	**36.08**
Chernoff	36.19	36.05	N/A	N/A

WER 50K test	Baseline	NBR	2-step LA	Search
KL	37.27	**36.68**	N/A	37.27
ENT	36.77	N/A	36.81	**36.6**
Chernoff	37.33	36.79	N/A	N/A

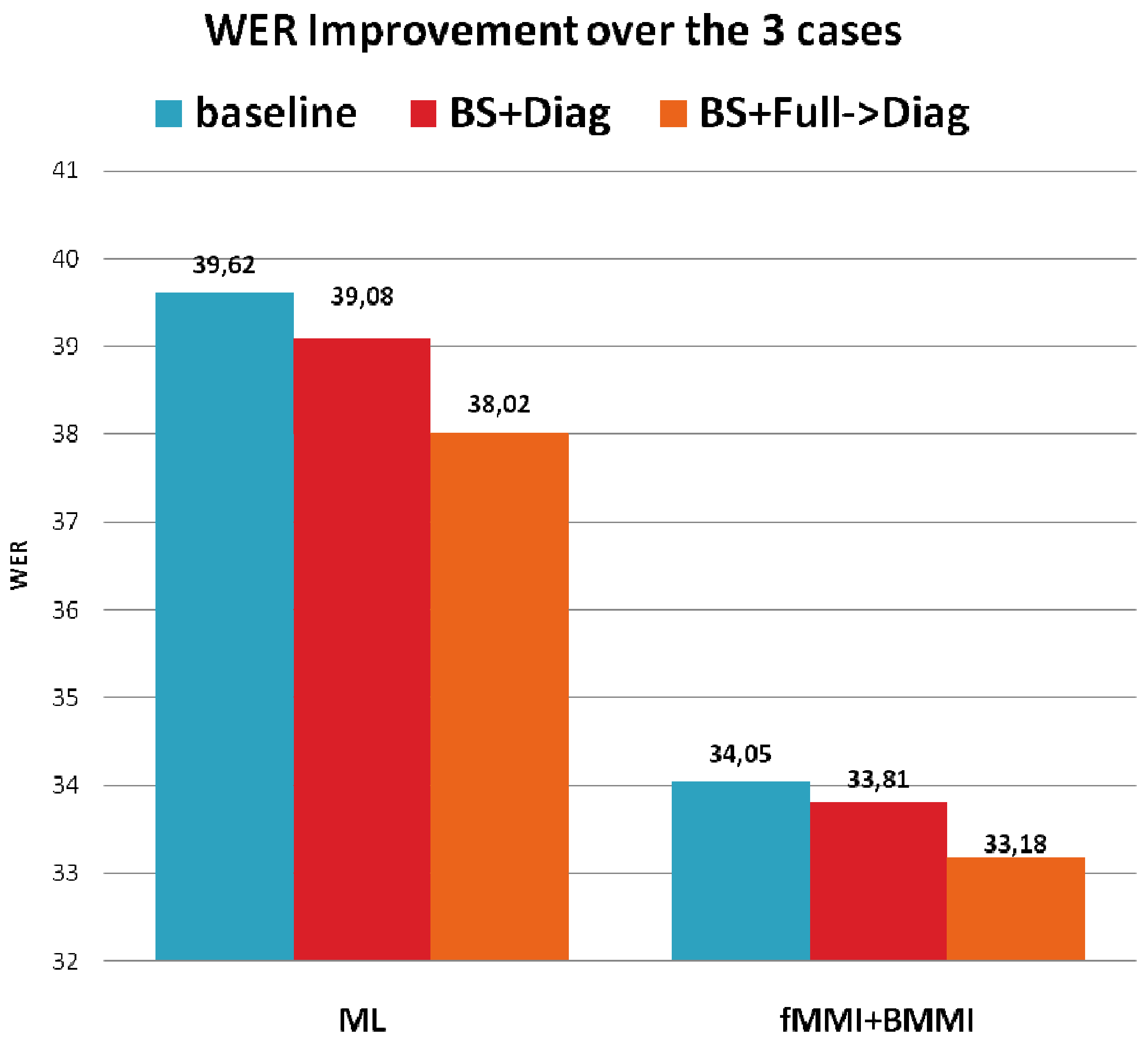
Results for 2-Pass Structural Optimization

Criteria for 2-Pass:
Find a threshold that Keep the clustered number of Gaussian is exactly the same as the 100K 1-pass model for a fair comparing

1-Pass (100K)	Baseline	NBR	2-step LA
KL	36.23	**36.04**	36.11
ENT	36.11	N/A	36.08
Chernoff	36.19	36.05	N/A

2-Pass (100K)	Baseline	NBR	2_step_LA
KL	36.18	36.02	36.04
ENT	36.04	N/A	36.04
Chernoff	36.12	**35.98**	N/A

From Full to Diagonal Comparison Results



Results are obtained by Xiaodong and Jian

Possible Future Extensions

- Search based
  - Auto adaptive beam
    - Beam can be based on a threshold
- K-step look ahead & Search optimize path
  - General approach can be extend to other similar tasks
    - Decision tree
- 2-Pass model structure optimization
  - Alternative criteria can be tried
    - MDL

References

[1] X. Cui, J. Xue, et. al.,“Acoustic modeling with bootstrap and restructuring for low resourced languages,” Proc.Interspeech, pp.291-294, 2010.
[2] X. Cui, et. al.,“Acoustic modeling with bootstrap and restructuring based on full covariance”, Submitted to Interspeech 2011.
[3] Scott Chen, Gopalakrishnan, P.S.,“Clustering via the Bayesian information criterion with applications in speech recognition,” Proc. ICASSP, pp.645-648, 1998.
[4] Ogawa, A. and Takahashi, S. ,“Weighted distance measures for efficient reduction of Gaussian mixture components in HMMbased acoustic model,” Proc. ICASSP, pp.4173-4176, 2008.

Thanks for your attention

Any questions?