# Motivation

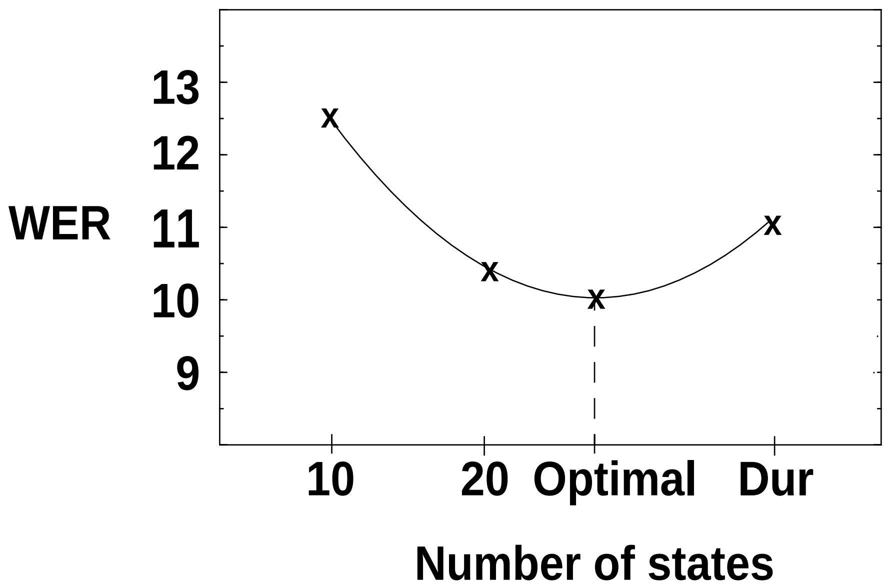
- ☞ Optimization of model topology not a common feature of most contemporary acoustic modeling systems
- ☞ Model topology/size decided heuristically and often uniformly (eg. 3-state triphones)
- ☞ Analysis indicates existence of optimal model size for larger acoustic units like syllables
- ☞ Information theoretic measures can be used to learn model size (number of states for HMMs)

# Current Approach

- Most triphone systems have 3 states/model
- Assume transition probabilities encode duration information
- Syllable models with number of states proportional to average model duration
- Syllable models with equal number of states
- Bayesian model merging used for Markov models (eg. pronunciation modeling)

# Previous Experience

- **Sys1**
  - **No. States: median duration / 20 msec**
  - **WER: 11.1% on OGI Alphadigits**
- **Sys2:**
  - **Max. states/syllable: 20 states**
  - **WER: 10.4% on OGI Alphadigits**



# Bhattacharyya Distance

- A separability measure between two Gaussian distributions

$$D = (M_2 - M_1)^T \cdot \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} \cdot (M_2 - M_1) + \frac{1}{2}\log\frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1| \cdot |\Sigma_2|}}$$

- Two terms represent separability due to class means and variances

- Used in phone clustering (Mak et. al., 1996)

# Kullback-Leibler Distance

- Divergence between two Gaussian distributions

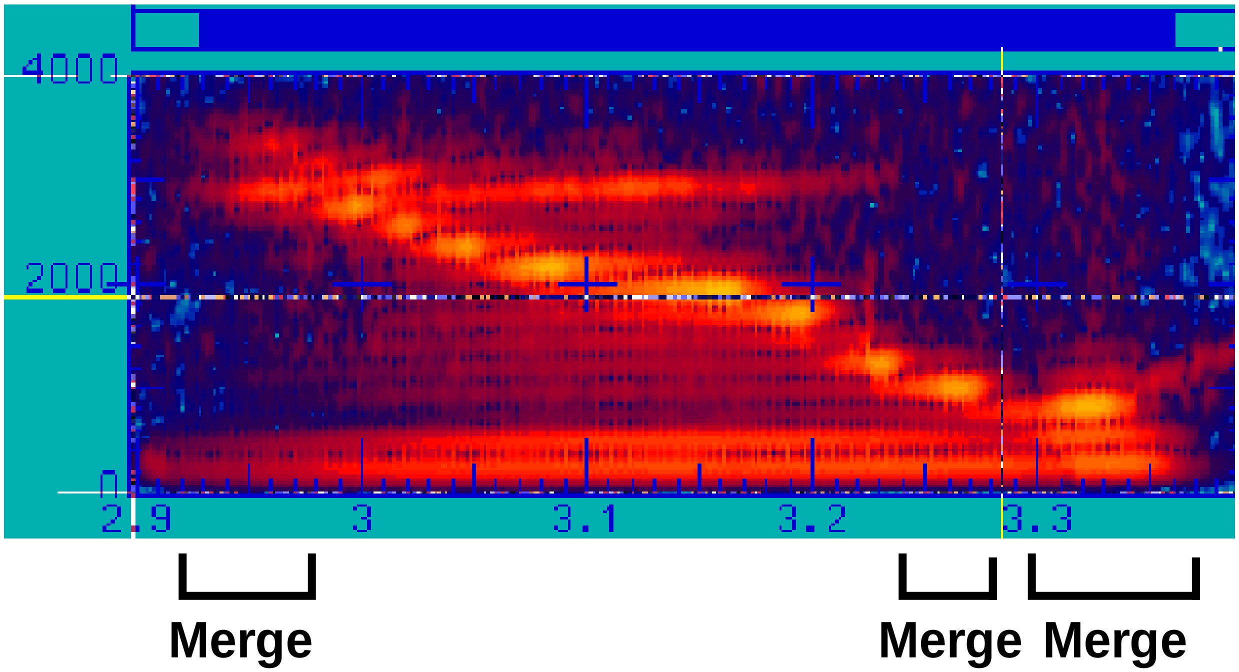
$$KL2(A;B) = \frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2} + (\mu_A - \mu_B)^2 \cdot \left(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right)$$

- Deviates from traditional definition of divergence to make it symmetric

- Successfully used for speaker change detection (Seigler et. al. 1997)

# Merging Phenomenon

**Syllable _y_uw**

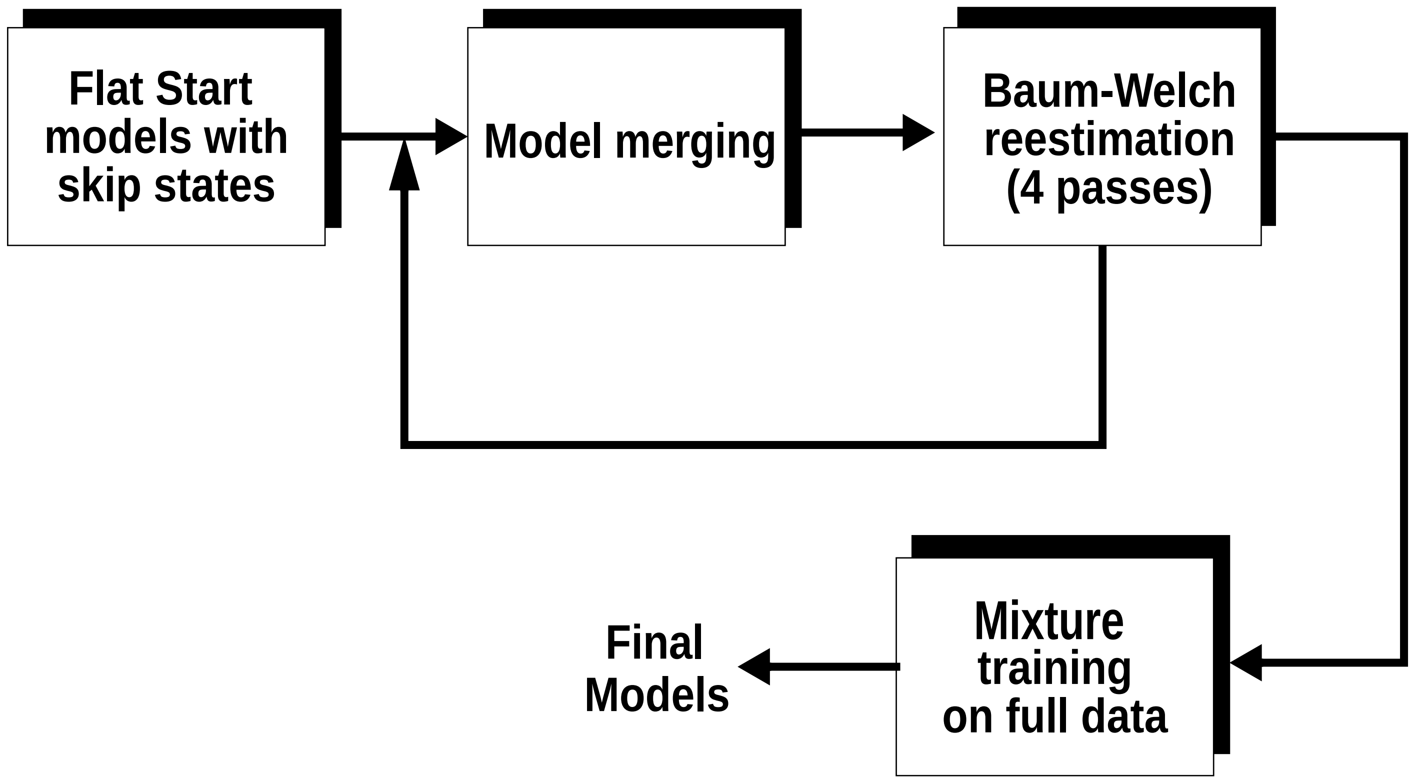



- **Ideally states best used to model transitional phenomenon**
- **Merge states around stable spectral regions**

# OGI Alphadigits

- Telephone database collected digitally using a T1 interface to the telephone network
- 3000 subjects in the corpus
- 19 or 29 alphanumeric strings per speaker
- Each utterance averages about six words in length (“8 H A 8 B H”, “8 W R W 8 E”)
- 1102 unique prompting strings
- Balanced phonetic context of bigrams

# Methodology




- Minimum probability of error used for Bhattacharyya distance is 0.2
- Minimum distance used for KL measure is 0.1

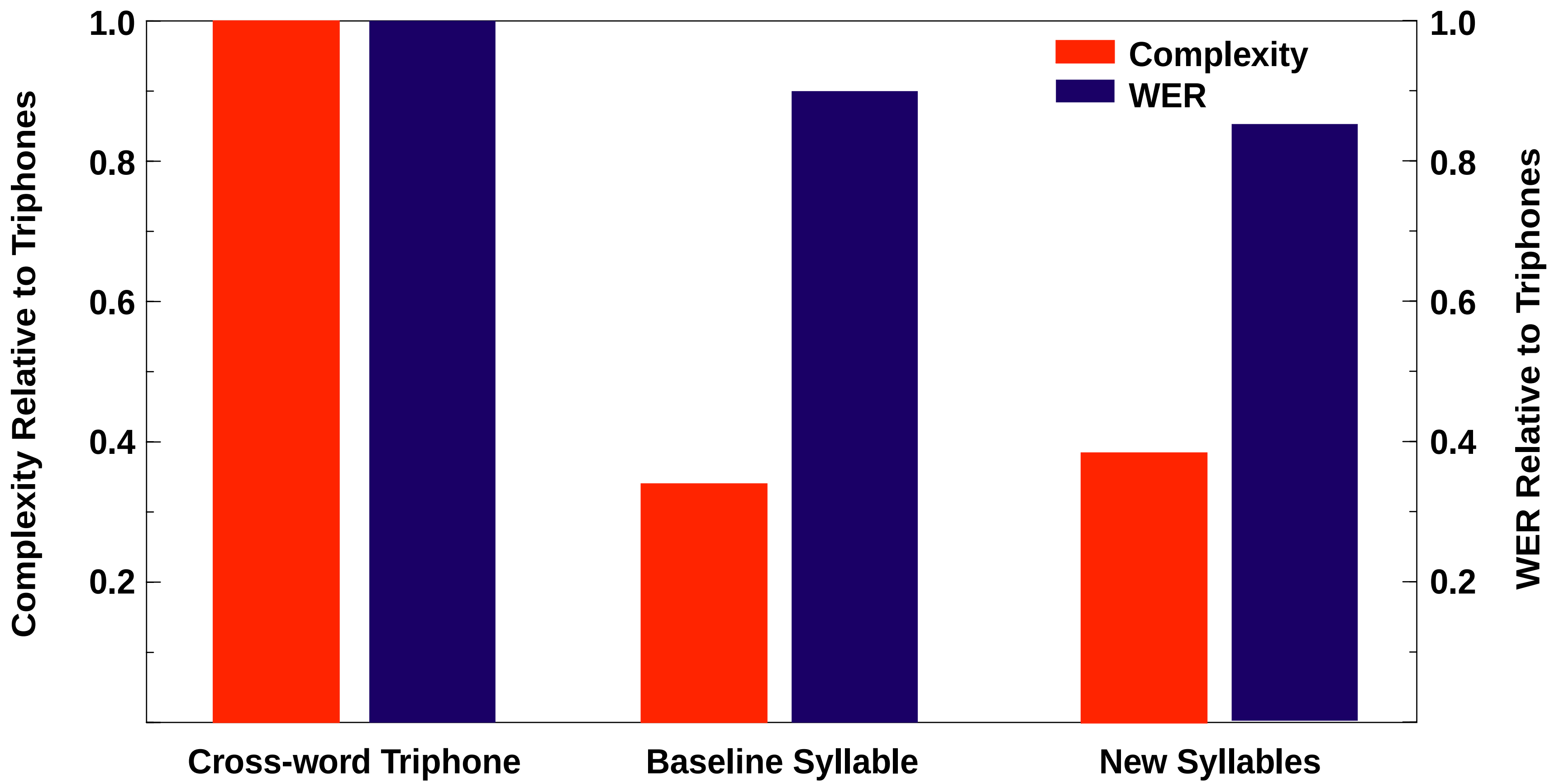
# Effect on Model Topology

- Initial models have number of states proportional to model duration (in frames)
- Before: models averaged 28 states/model
- After:

  Pass 1: average of 21 states/model

  Pass 2: average of 19 states/model
- Analysis shows that merges in the “stable” spectral portion of most models

# Results




- Using Bhattacharyya distance, WER reduced from 10.4% to 9.9% in a syllable system
- Models 13% longer on an average in the new system (compensated skip states)

# Analysis

- ☞ KL2 and Bhattacharyya distance metrics consistently give similar merges
- ☞ Convergence of some models slower than others, attributable to transitional formant structure
- ☞ Similar error modalities as previous syllable system
- ☞ Incorporation of skip states in models warrants higher word insertion penalty

# Summary

- ☞ Information theoretic measures for optimal model size — Bhattacharyya Distance and KL2 Distance
- ☞ Significant improvement in performance: reduced WER from 10.4% to 9.9% (alphadigits)
- ☞ Need to determine impact of model merging on syllable-based SWB systems
- ☞ Explore approaches like BIC and MDL for more general topology selection