

Speaker Segmentation and Clustering

J. Ajmera^{†‡}, C. Wooters[‡], B. Peskin[‡], and C. Oei[‡]

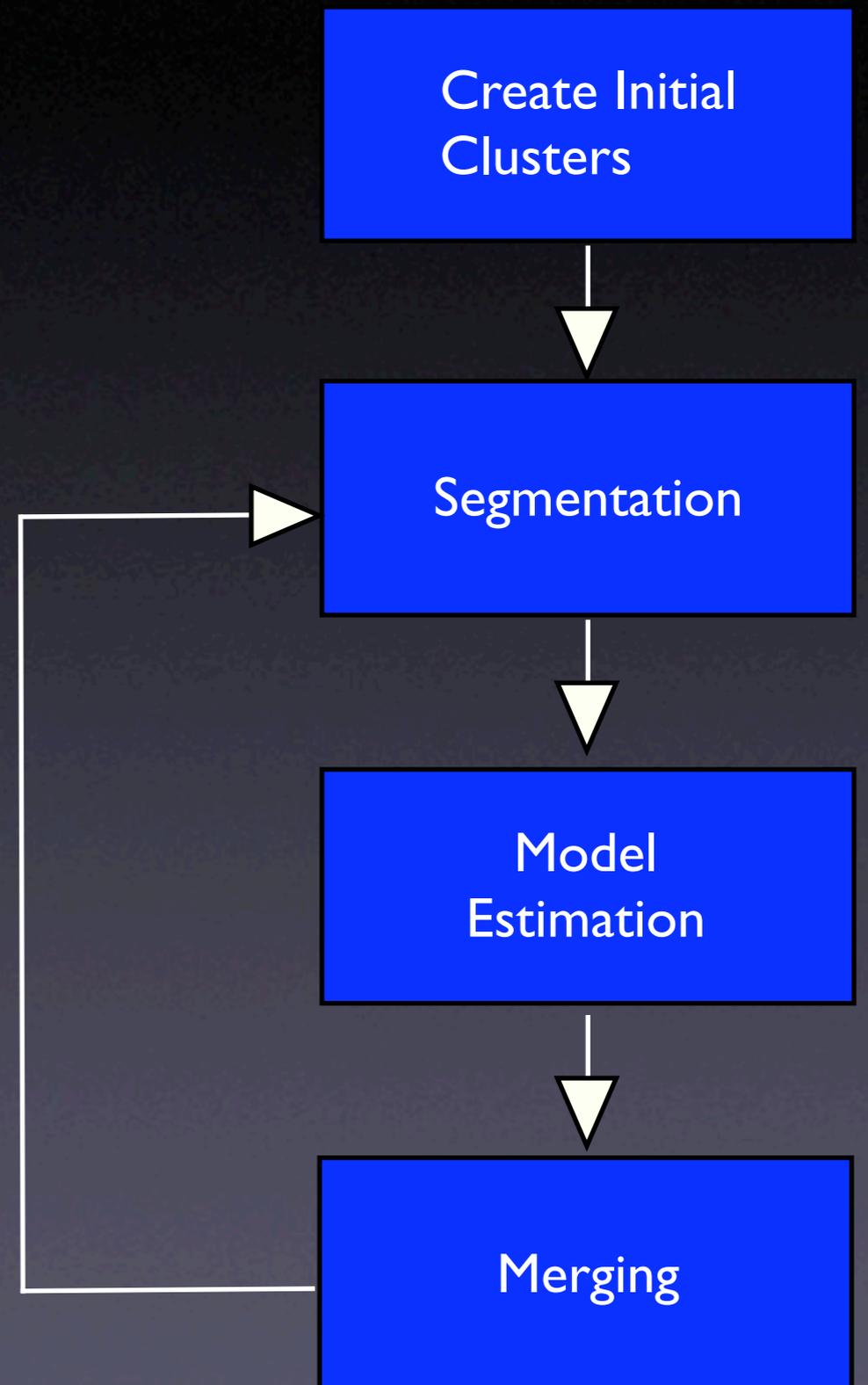
[†]IDIAP [‡]ICSI

Overview

- The clustering algorithm
- Speech/Music Classifier
- Results

The Algorithm

No training data
Few tunable parameters
No prior segmentation assumed
Num. of params remains constant

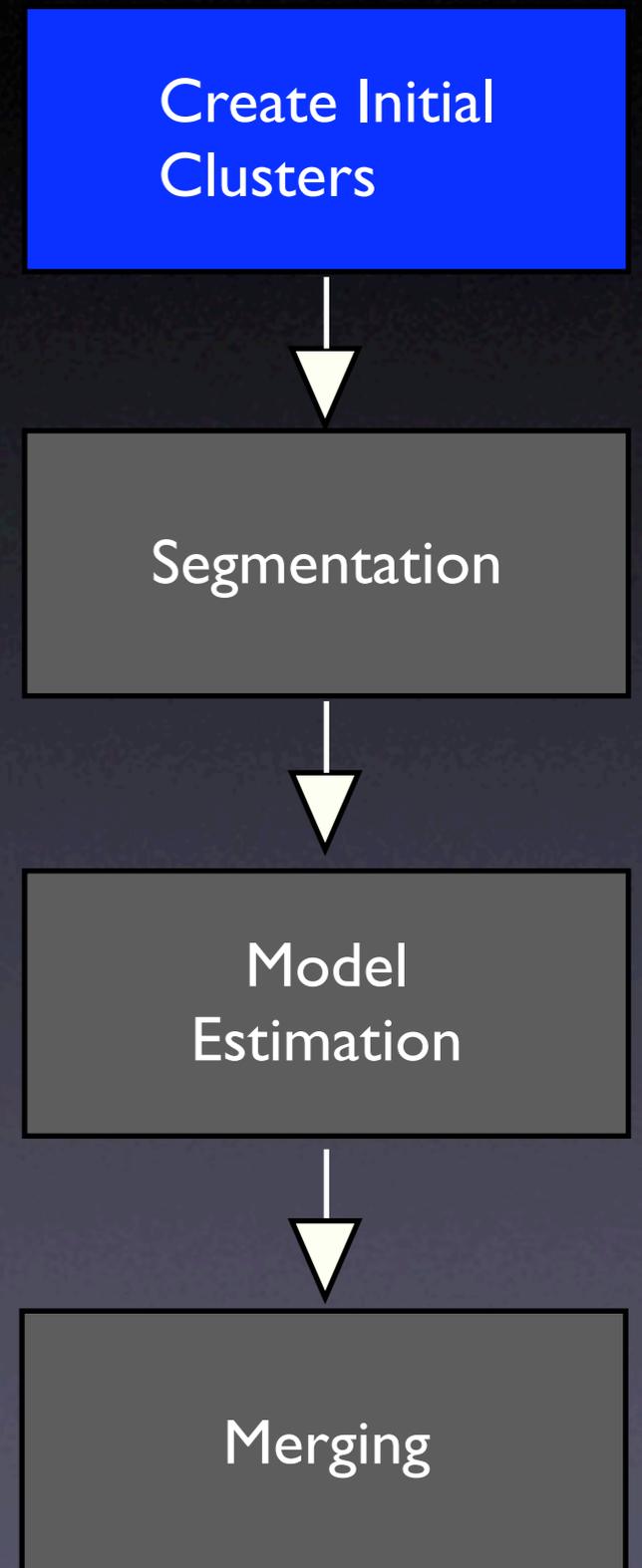


Create Initial Clusters

A cluster is a set of frames and we model it using a parametric model (GMM).

Choose a number of clusters (K) that is larger than the expected number of speakers.

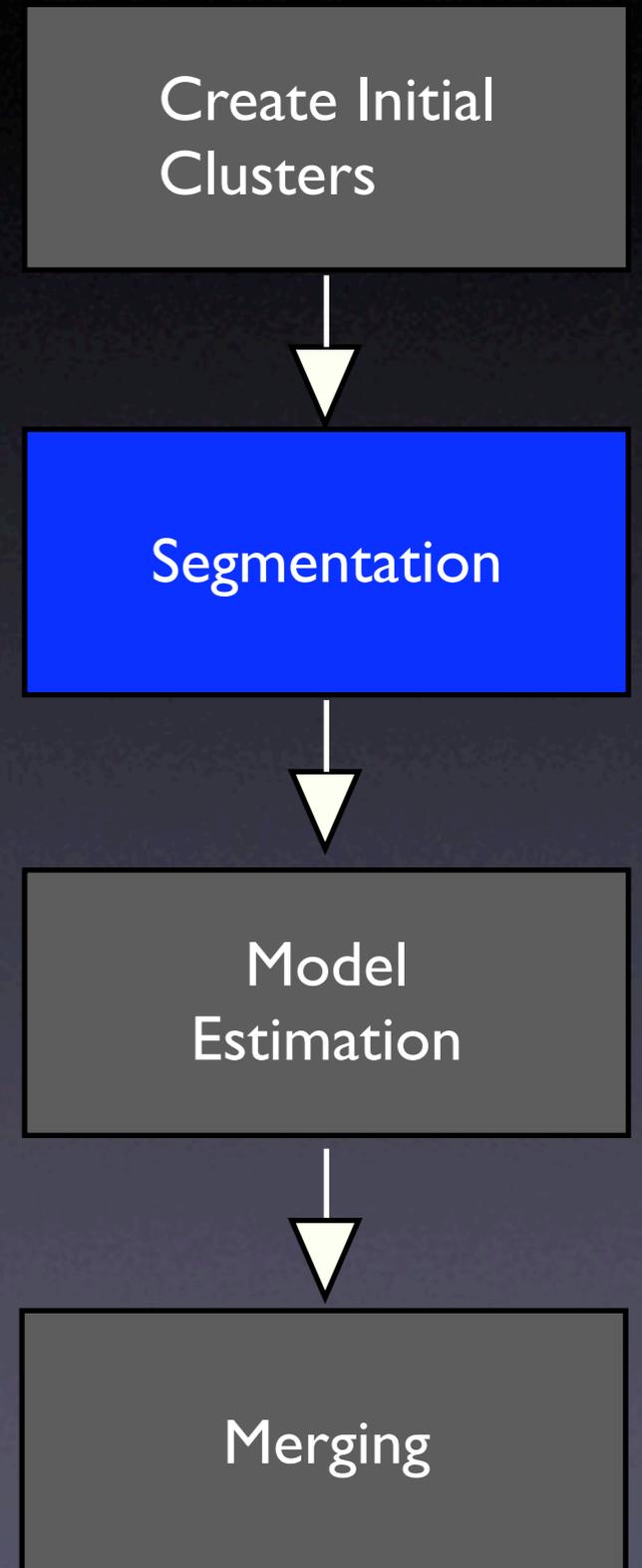
Uniformly segment data to create initial set of K clusters.



Segmentation

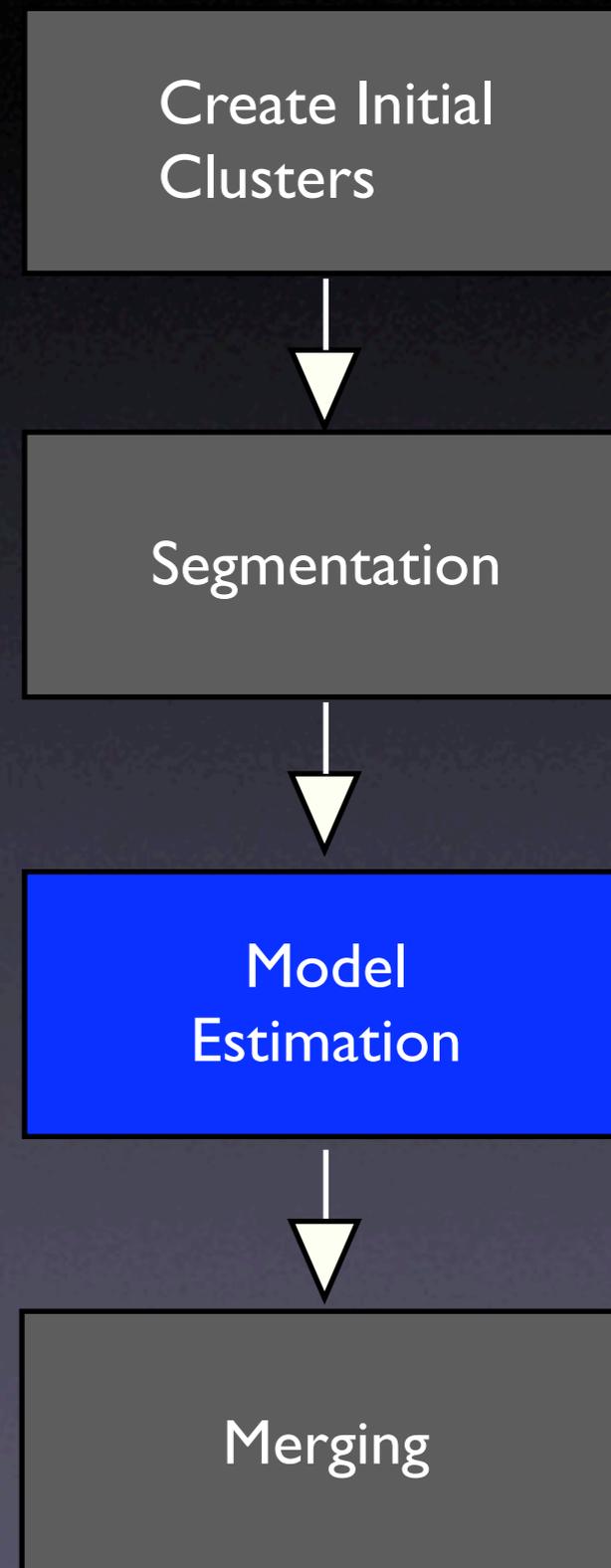
Create a multi-state HMM for each cluster; HMMs have a specified minimum duration (e.g. 2 secs.)

Use Viterbi alignment to assign cluster labels to the speech data.



Cluster Model Reestimation

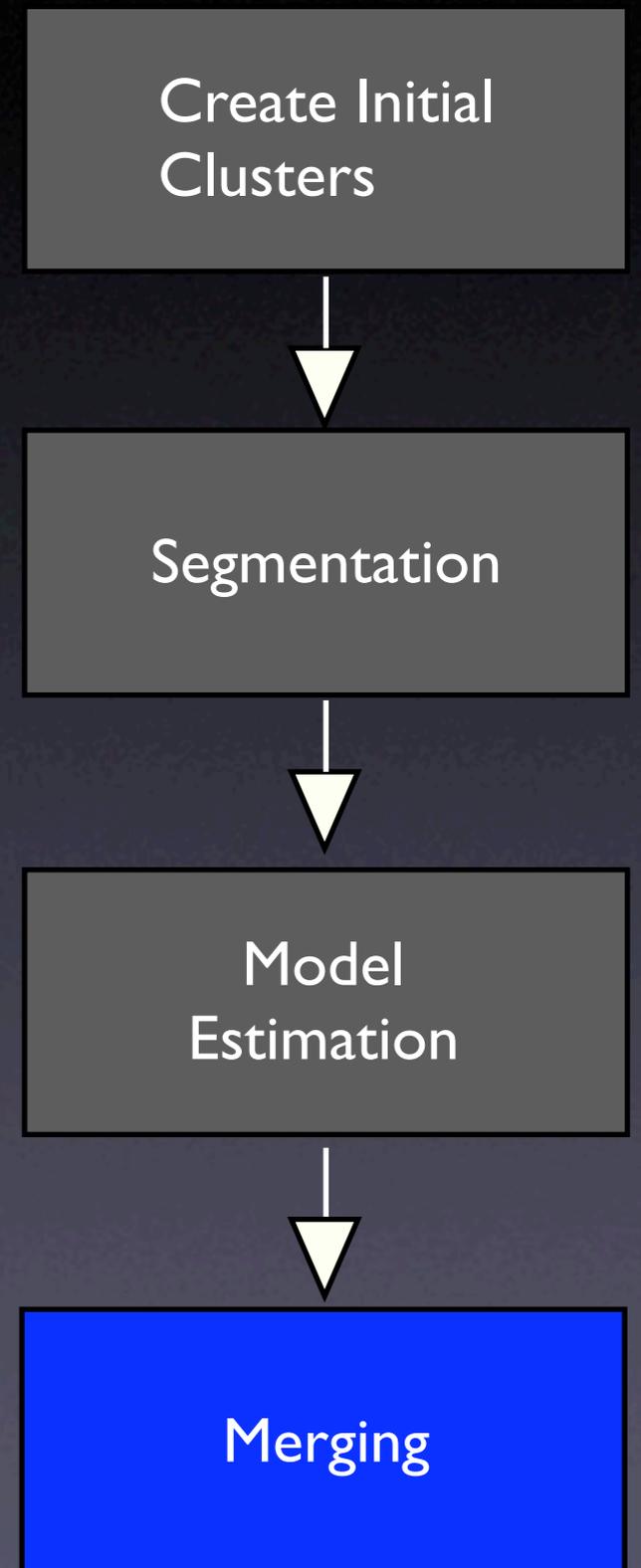
Retrain the parameters of the cluster models using the new segmentation.



Cluster Merging

Since the algorithm is essentially an agglomerative clustering technique, we need a method to calculate the “distance” (or degree of similarity) between two clusters.

Check to see if we would be better off (i.e. obtain a higher likelihood) by modeling the two data sets using a single model or two separate models, while holding the number of parameters constant.



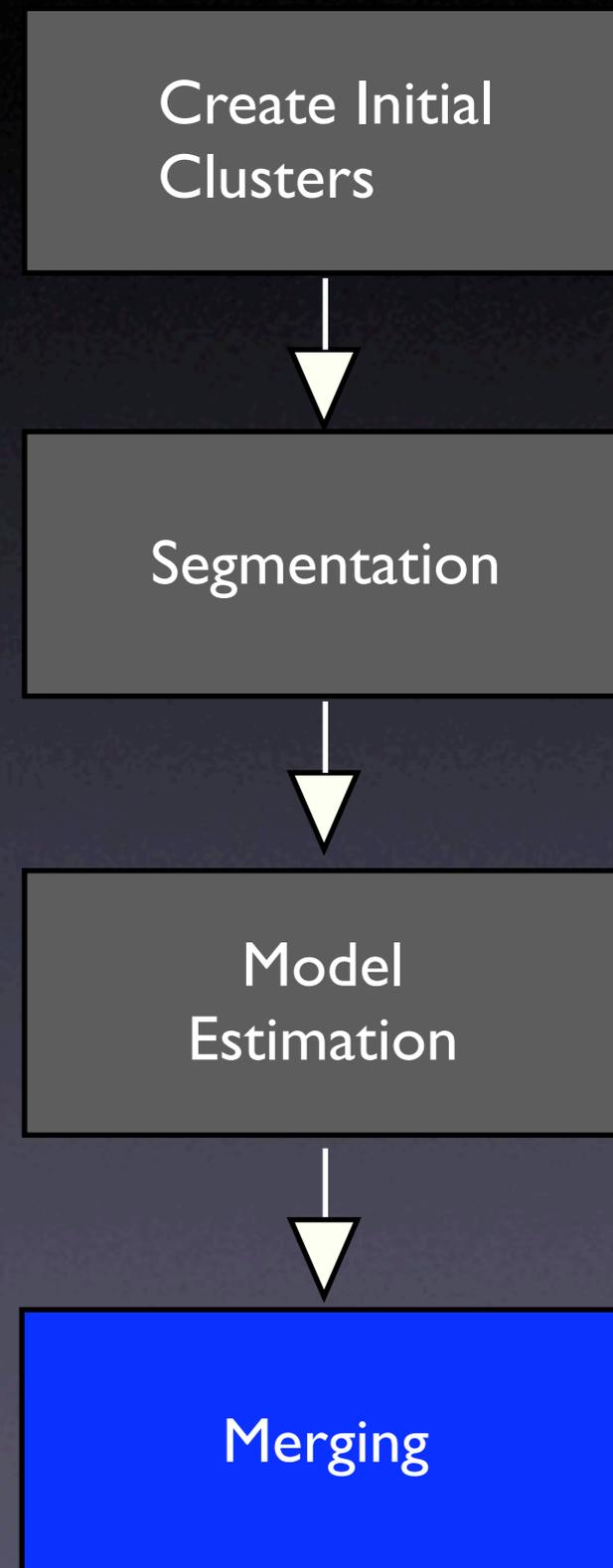
Cluster Merging: con't

Find the pair of clusters with the smallest distance and merge them.

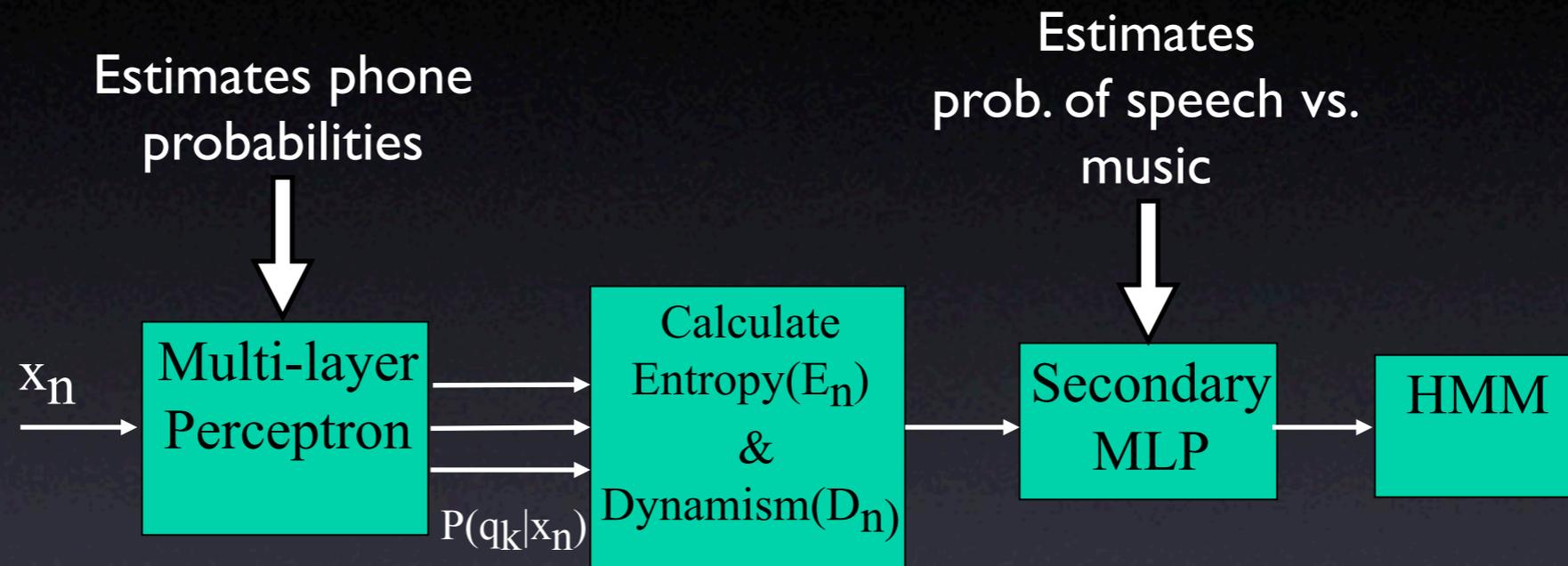
The merged cluster contains the same number of parameters as the two separate clusters.

If no pair satisfies the merging criterion, stop.

Else, return to segmentation step.



Speech/Music Classifier



$$E_n = -\sum_k P(q_k|x_n) \log P(q_k|x_n)$$

$$D_n = \sum_k (P(q_k|x_n) - P(q_k|x_{n+1}))^2$$

Where, q_k = k 'th phoneme, and x_n = feature vector at time n

SMC Analysis

	lpcc+smc			lpcc		
	MS	FA	SE	MS	FA	SE
VOA	3.4	6.8	12.0	0.1	9.0	14.0
PRI	4.2	5.7	11.0	0.1	10.1	11.5
MNB	0.5	8.8	2.1	0.0	16.7	2.1

Speech/Music Classifier (SMC) increases Missed Speech (MS), decreases False Alarms (FA). Helps in all the shows, esp. in MNB.

Our MS and FA errors are very high. A better SMC can help us improve on this.

Our Speaker error times (SE) are reasonably low, showing the effectiveness of the clustering algorithm on speech regions.

Results

Data	lpcc+smc (primary)	Only lpcc	mfcc+smc
VOA	22.17	23.09	17.39
PRI	20.94	21.68	23.73
MNB	11.32	18.82	11.90
Overall	18.70%	21.40%	18.09%

lpcc: Linear Predictive Cepstral Coeff (12)

mfcc: Mel Frequency Cepstral Coeff (19)

smc: Speech/Music Classifier

References

- J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, “Unknown-multiple speaker clustering using HMM”, ICSLP, 2002.
- J. Ajmera, H. Bourlard, I. Lapidot, “Improved Unknown-Multiple Speaker Clustering Using HMM”, IDIAP Research Report- 02-23. <http://www.idiap.ch>
- J. Ajmera, I. McCowan, and H. Bourlard, “Speech/Music Discrimination using Entropy and Dynamism Features in a HMM Classification Framework”, Speech Communication, vol 40/3 pp 351 - 363.