连续语音的HMM训练

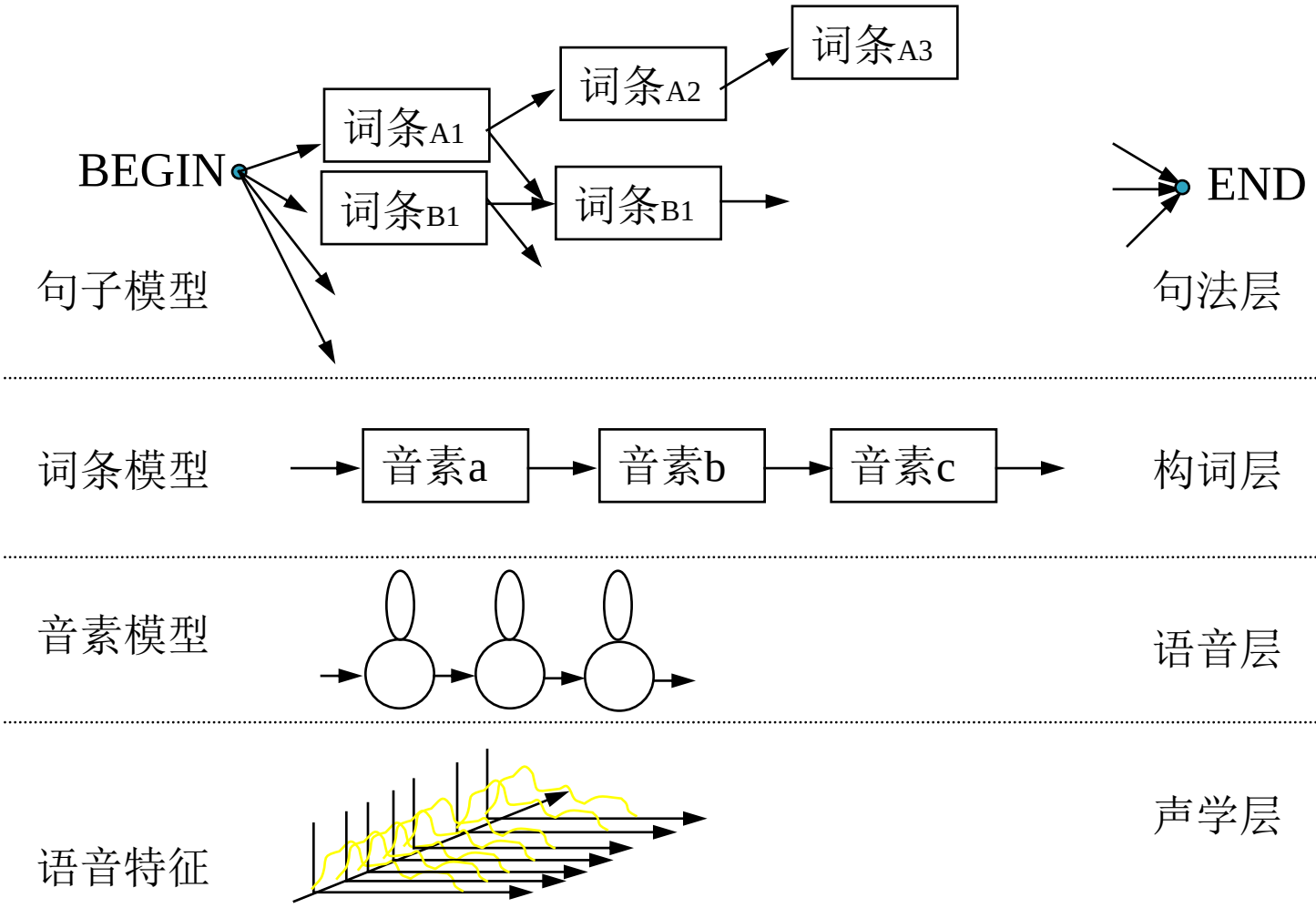
洪青阳 副教授

厦门大学信息科学与技术学院
qyhong@xmu.edu.cn

要点

- 连续语音的HMM训练
- 音素的上下文建模
- 决策树

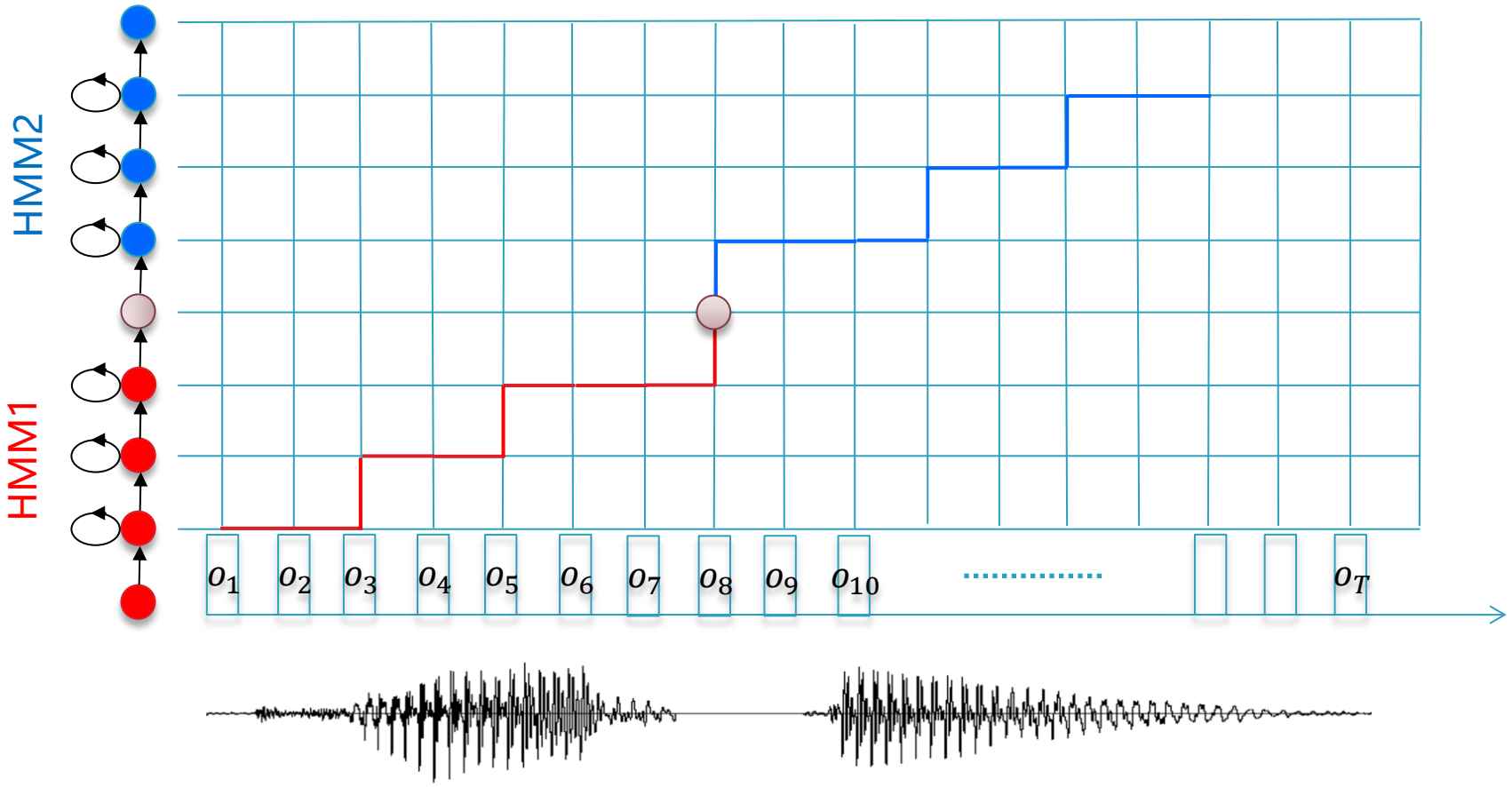
基于HMM的连续语音识别系统



基于子词单元的HMM训练

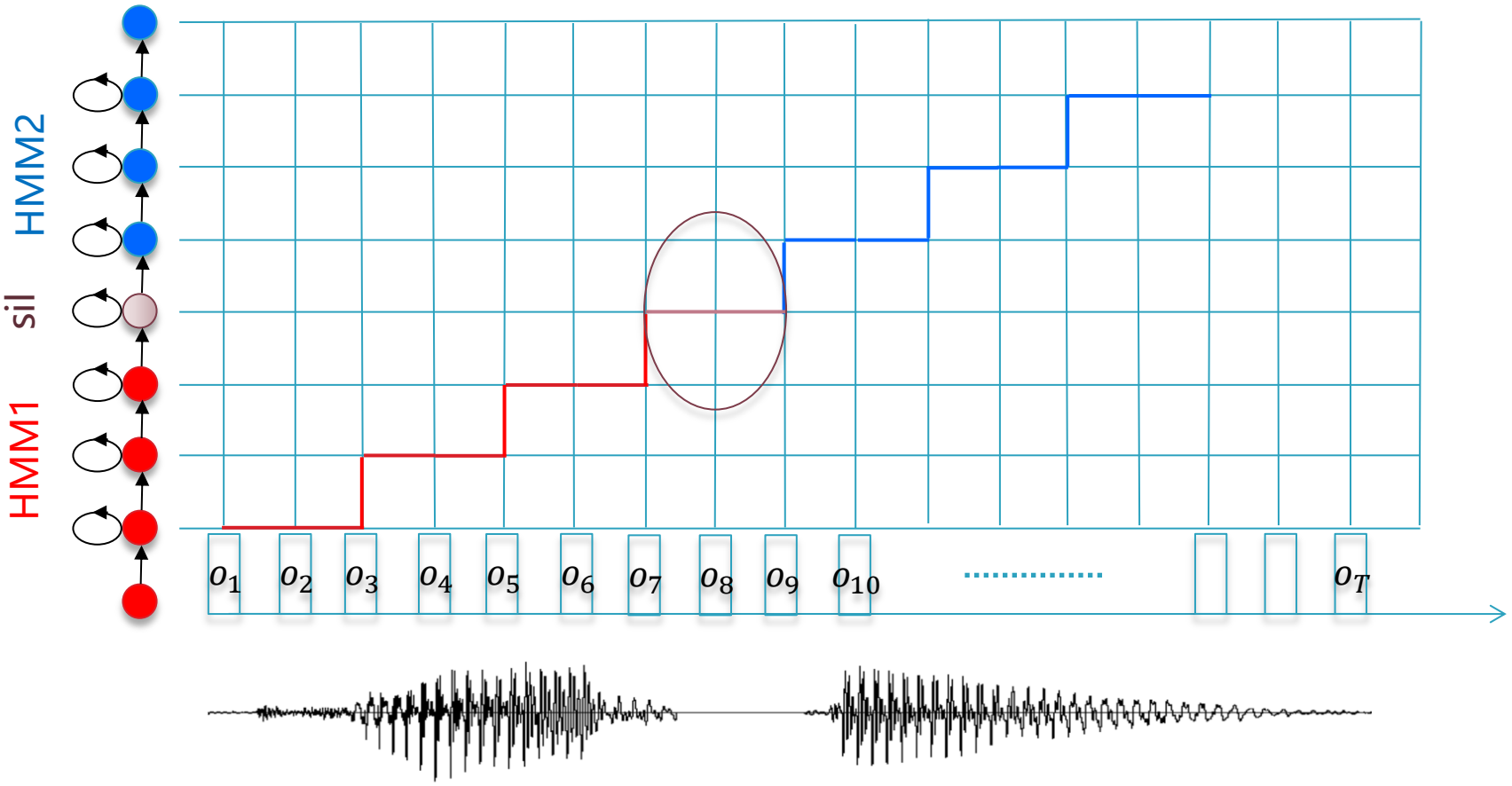
- 子词单元的HMM一般采用从左到右的结构，状态数固定为3到5个。
- 在语音段中，子词太短，无法精确标出语音边界。
- 训练时，用一种很粗糙的方法进行初始分段，例如等长分段，形成初始模型 。

基于子词单元的HMM训练



基于子词单元的HMM训练

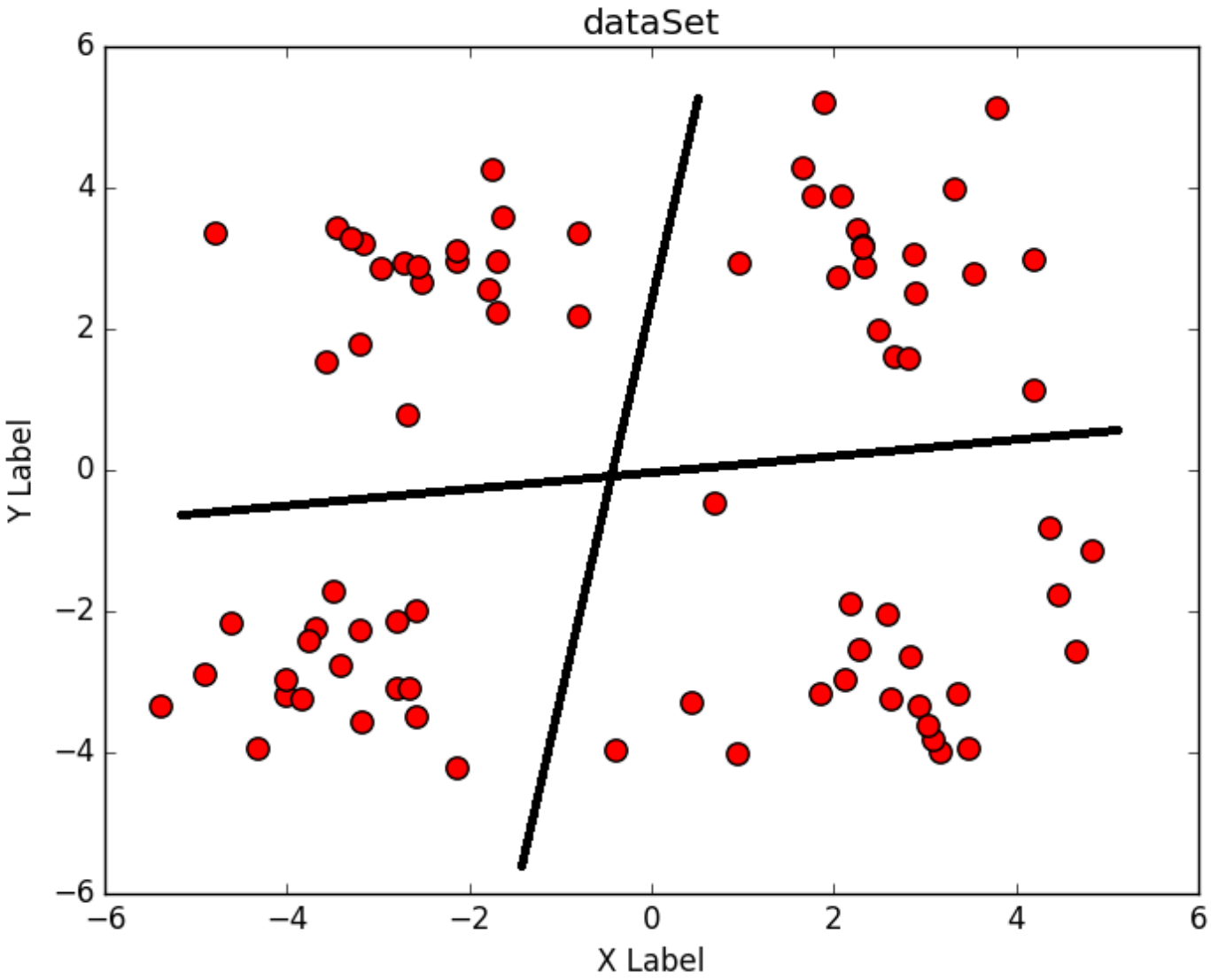
考虑静音



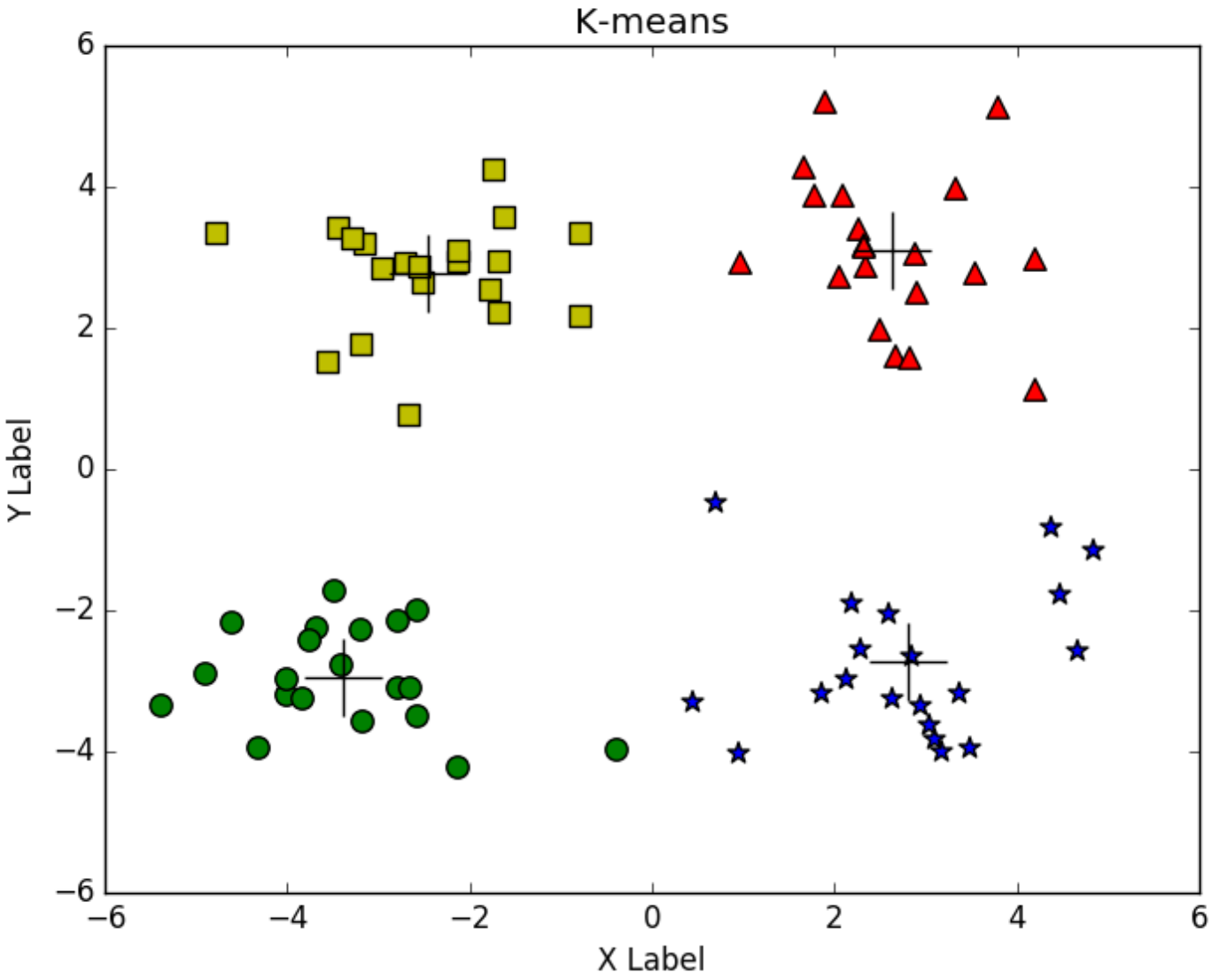
基于子词单元的HMM训练

- 分段K-Means算法
  - 初始化：将每个训练语句线性分割成子词单元，将每个子词单元线性分割成状态，即假定在一个语句中，子词单元及其内部的状态驻留时间是均匀的；
  - 聚类：对每个给定子词单元的每一个状态，其在所有训练语句段中特征矢量用K-Means算法聚类；

K-Means算法



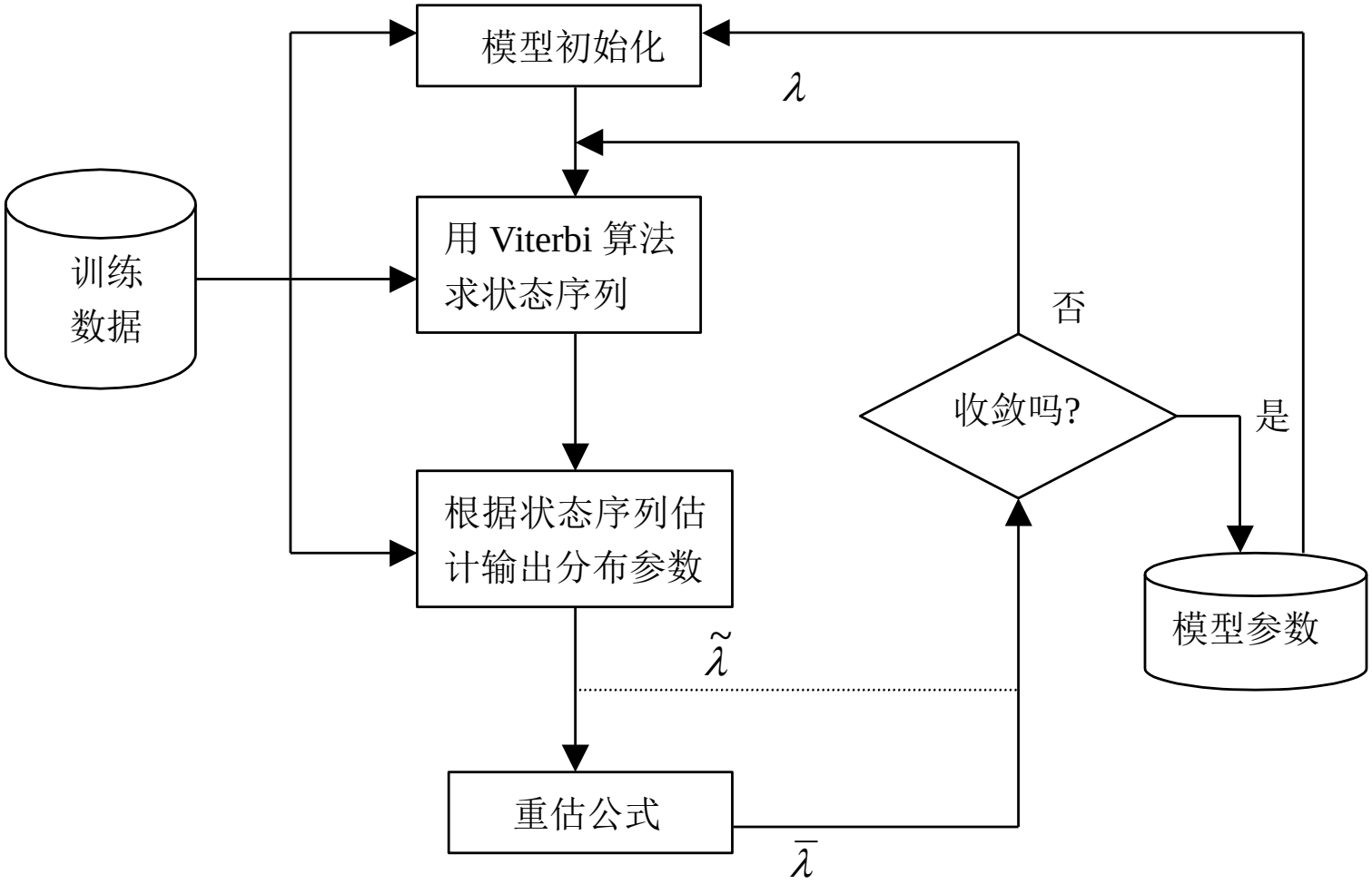
K-Means算法



分段K-Means算法

- 参数估计：根据聚类的结果计算均值、各维方差和混合权值系数；
- 分段：根据上一步得到的新的子词单元模型，通过Viterbi算法对所有训练语句再分成子词单元和状态，重新迭代聚类和参数估计，直到收敛。

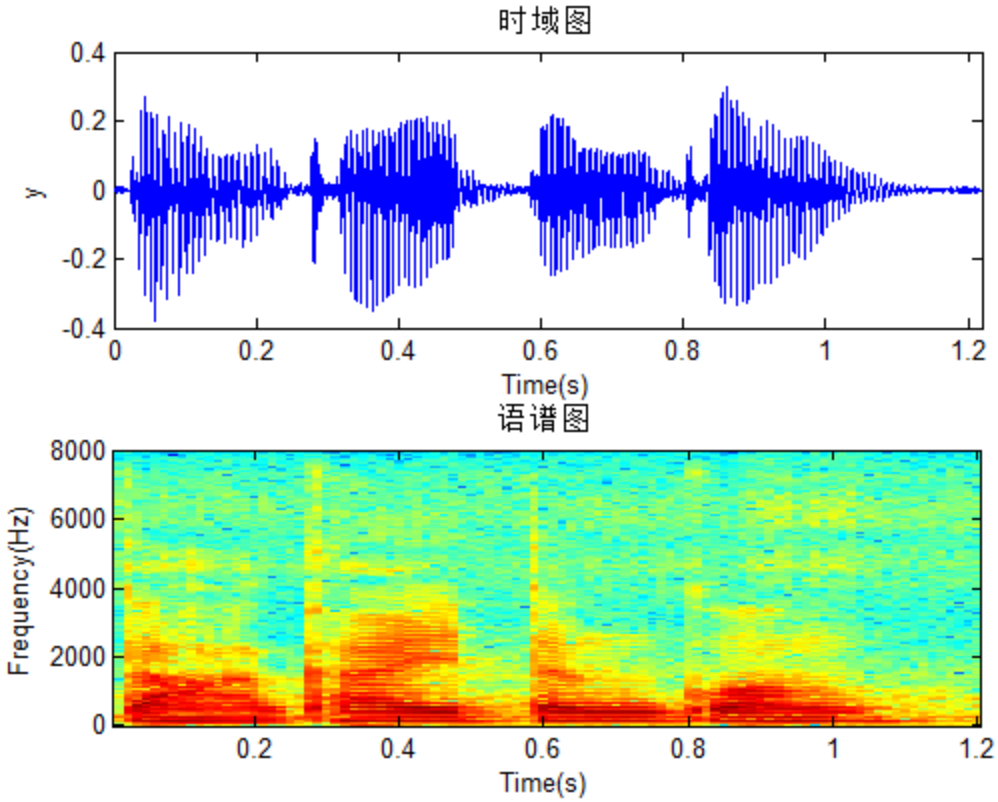
基于子词单元的HMM训练



音素的上下文建模

- 上下文关联的发音
- 上下文建模
  - 三音子(triphone)

上下文关联的发音



打 开 灯 光

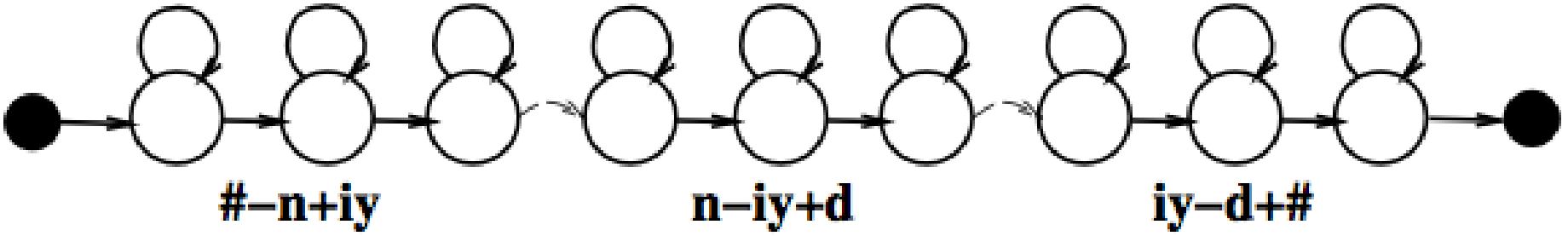
da kai deng guang

上下文建模

- 单音子(monophone)
  - 声母：d, k,g
  - 韵母：a,ai,eng,uang

- 三音子(triphone)
  - d+a+k
  - k+ai+d
  - d+eng+g
  - g+uang+sil

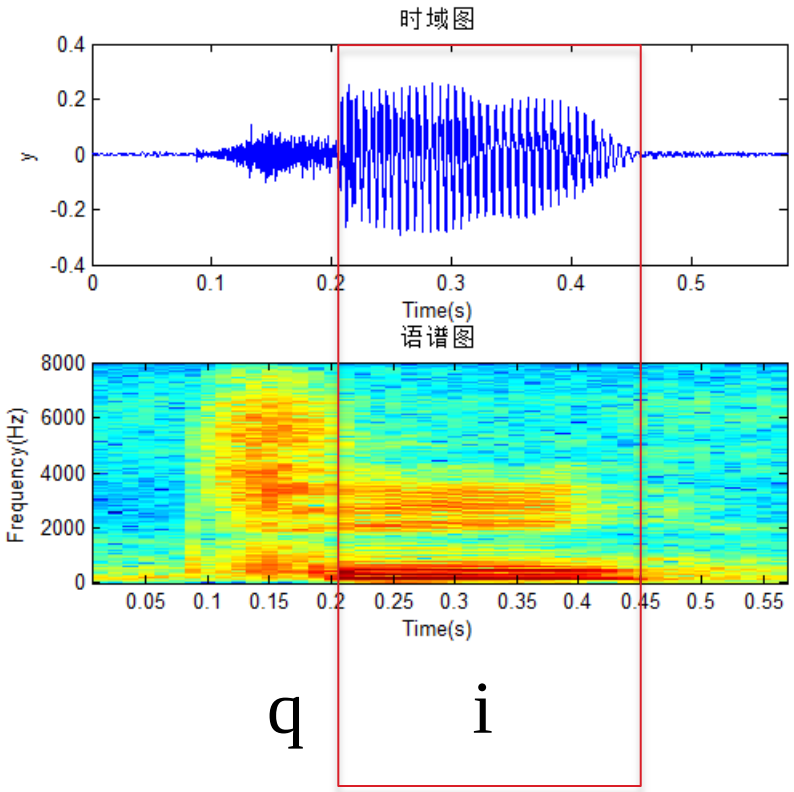
英文“need”的三音子模型

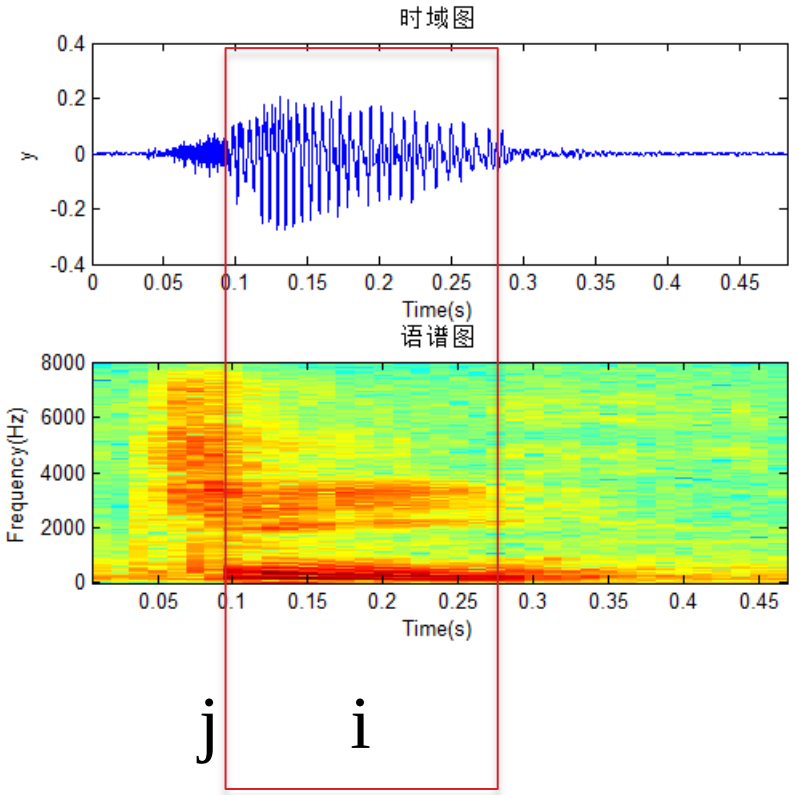


Triphone

- 对音素的上下文更精细的建模
- 数量巨大！例如
  - 普通话：27(声母)*36(韵母)*29(声母+sil/sp)=28188
  - 英语：28(辅音)*20(元音)*30(元音+sil/sp)=16800
- 如果加上声调，数目更多！而训练数据有限！
- 解决办法：
  - 采用双音子
  - 状态绑定(state tying)

普通话“i”

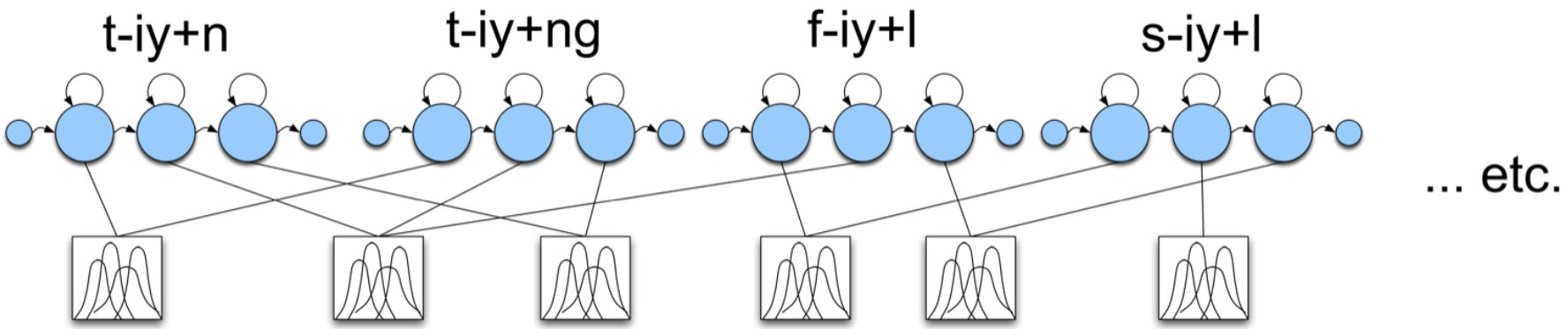

时域图
y
Time(s)
语谱图
Frequency(Hz)
Time(s)
q
i


时域图
y
Time(s)
语谱图
Frequency(Hz)
Time(s)
j
i

双音子(diphone)

- 例子：
  - qi: q+i, q-i
  - ji: j+i, j-I
- 除了sil/sp静音模型，最多只需
  - 普通话(不带声调)： 27(声母)*36(韵母)*2=1944
  - 英语： 28(辅音)*20(元音)*2=1120

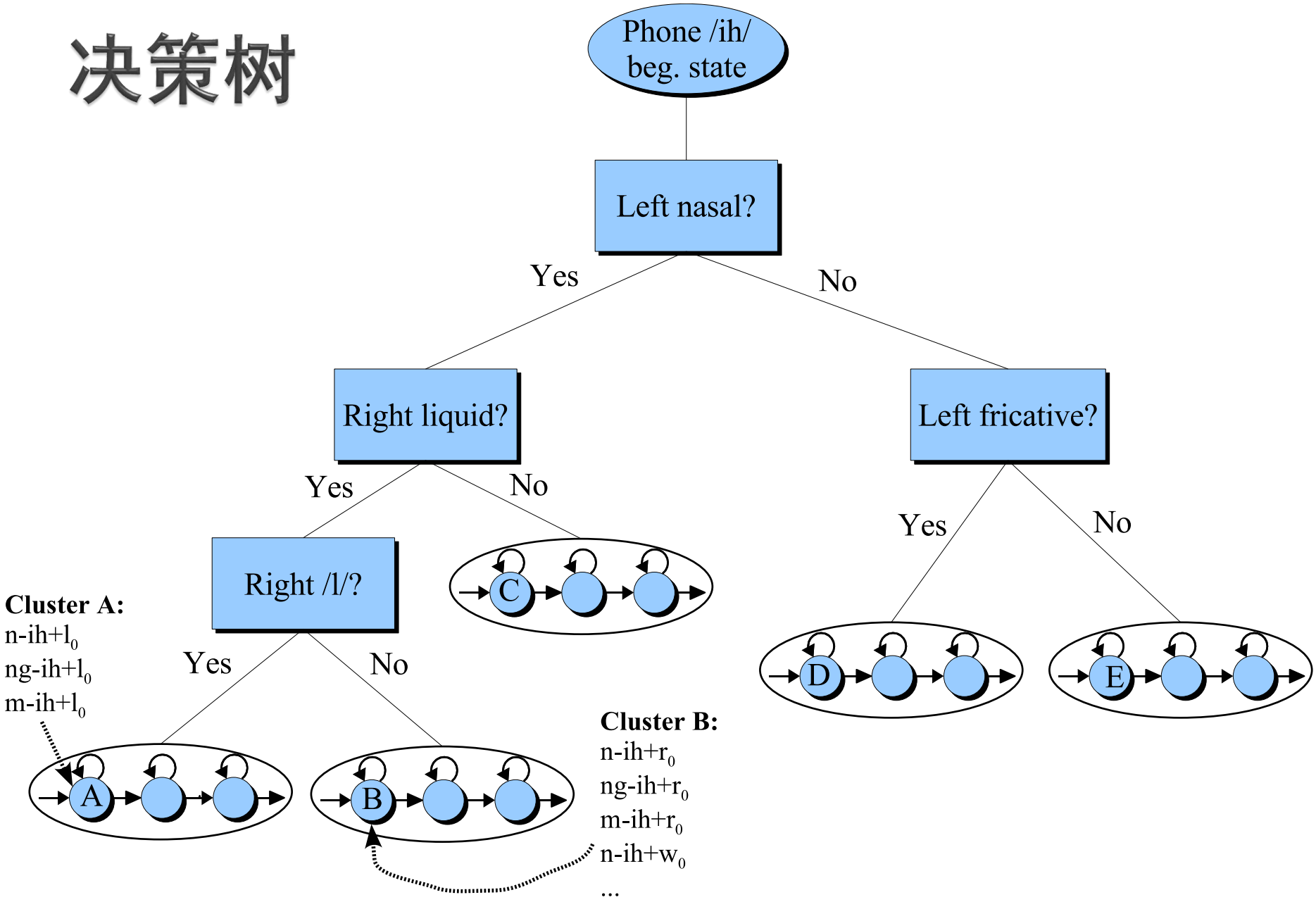
三音子的状态绑定(state tying)



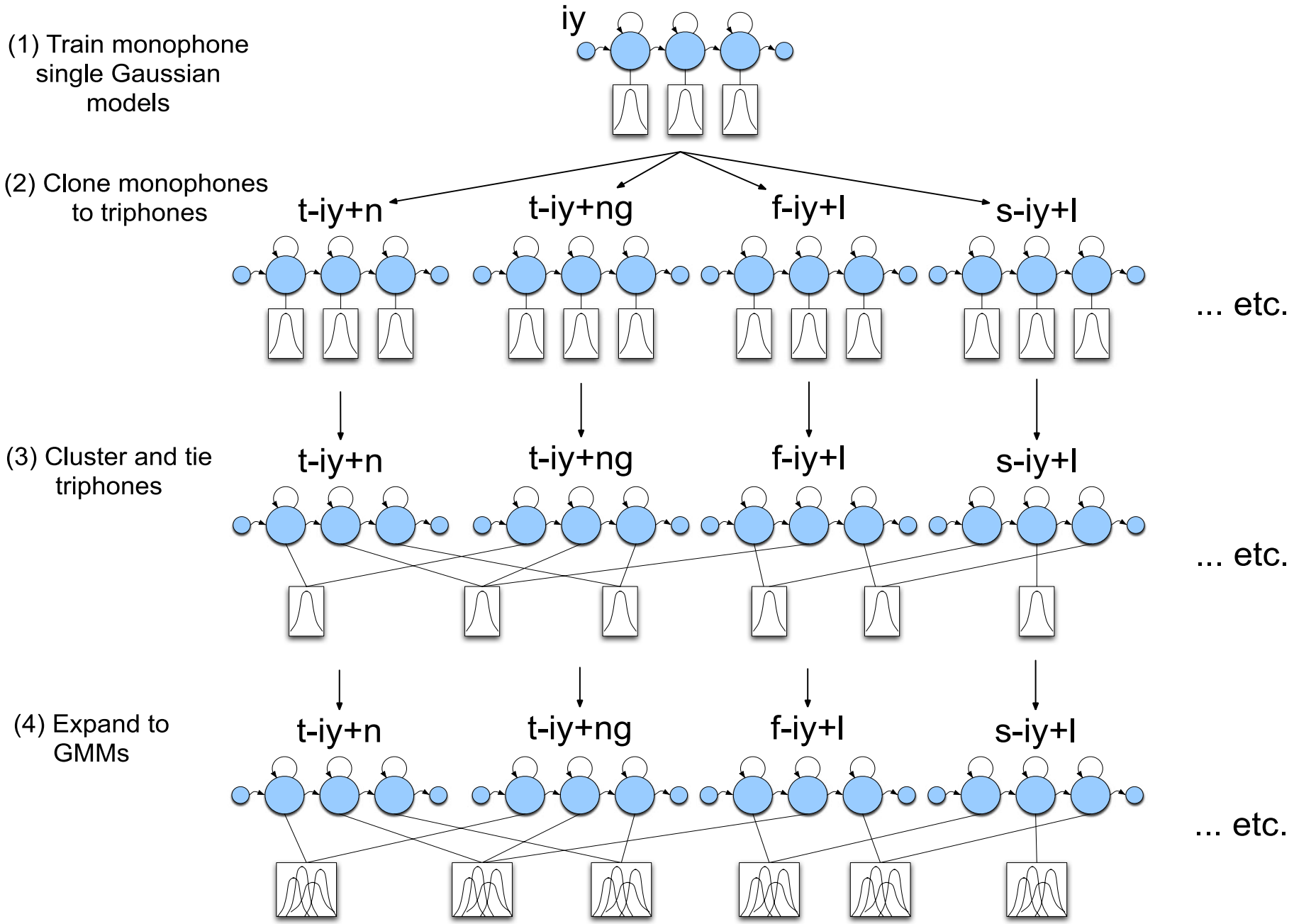
绑定依据

- 相似的辅音、元音
- 决策树

决策树


Phone /ih/ beg. state
Left nasal?
Yes
No
Right liquid?
Left fricative?
Yes
No
Yes
No
Right /l/?
C
Yes
No
D
E
Cluster A:
n-ih+l0
ng-ih+l0
m-ih+l0
A
B
Cluster B:
n-ih+r0
ng-ih+r0
m-ih+r0
n-ih+w0
...

State Tying: Young, Odell, Woodland 1994



Thank you!

Any questions?