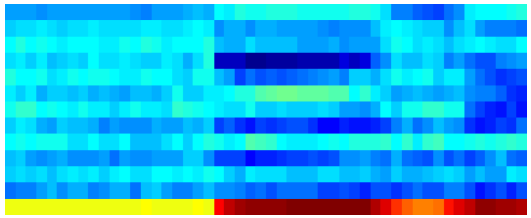
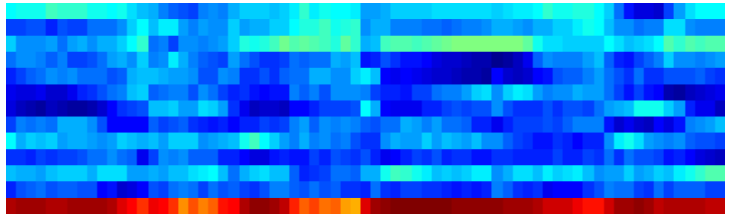
Overview of Building a Speech Recognizer Using HMMs

Recording and Front-End

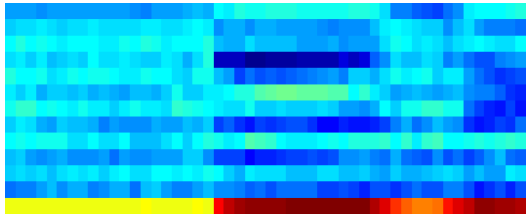


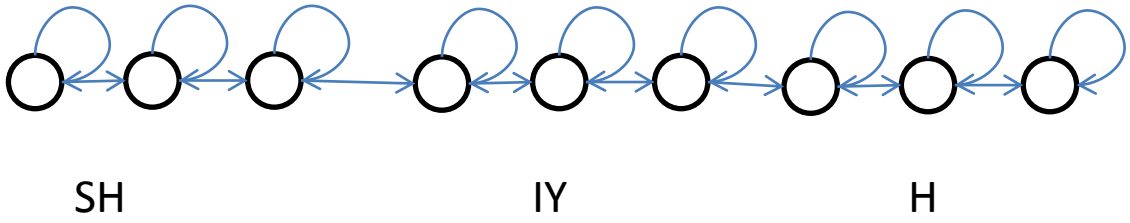
SH IY H AE D Y AO R D AA R K



AO TH ER AH V DH AH D EY N JH ER

Acoustic modeling





-Concatenate models for each utterance

-usually triphones are used (e.g. *SH*-***IY***-*H*)

-Initialize models e.g. randomly

-Go through the whole data set

-Calculate forward/backward probabilities

-state occupation probability for each model

-Re-estimate

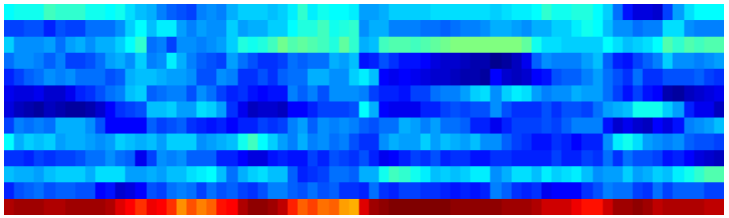
-calculate new transition probabilities, means and variances based on data and state occupation probabilities

-Iterate

-In practice, models have to be pooled

-decision-tree state clustering

Recognition



= ?

Construct language model from text data
-N-grams, for example
Expand word into acoustic models
-a gigantic network
Find the most likely state sequence given the observations with Viterbi
This state sequence will give us the recognized speech + timing + confidence