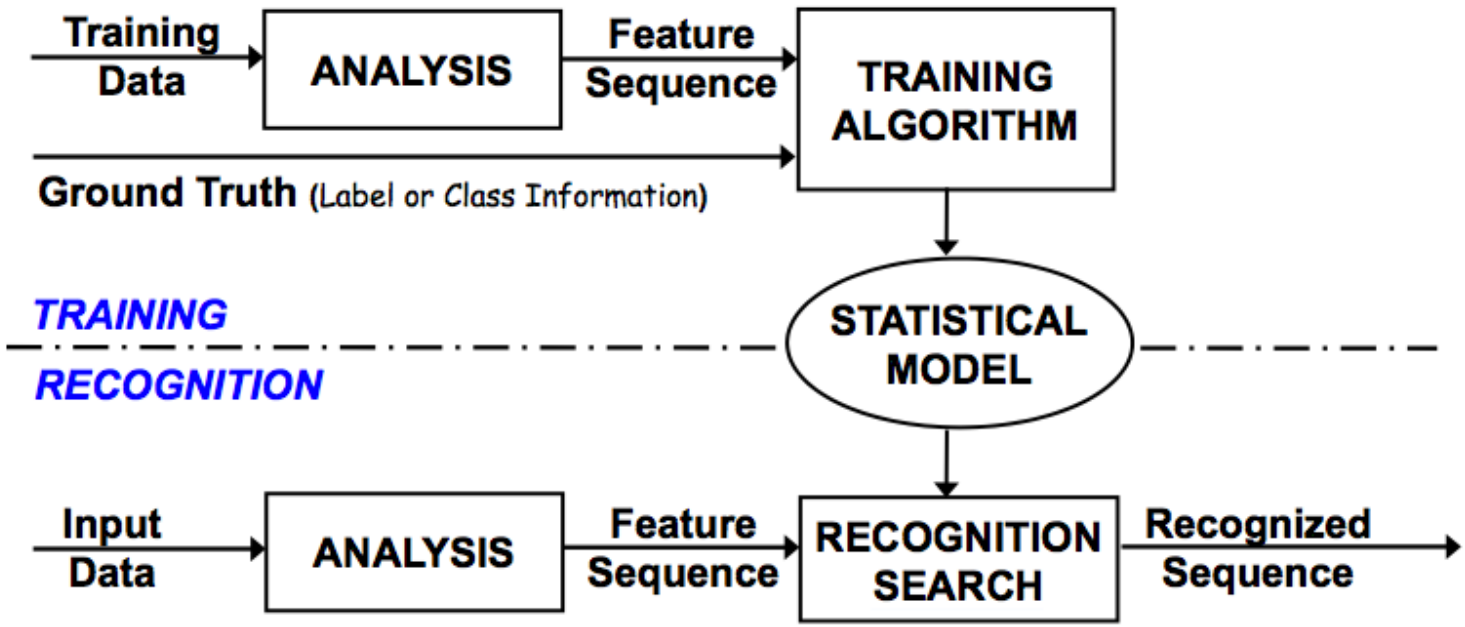
Speech Recognition

Leonid E. Zhukov

School of Applied Mathematics and Information Science
National Research University Higher School of Economics

28.11.2012

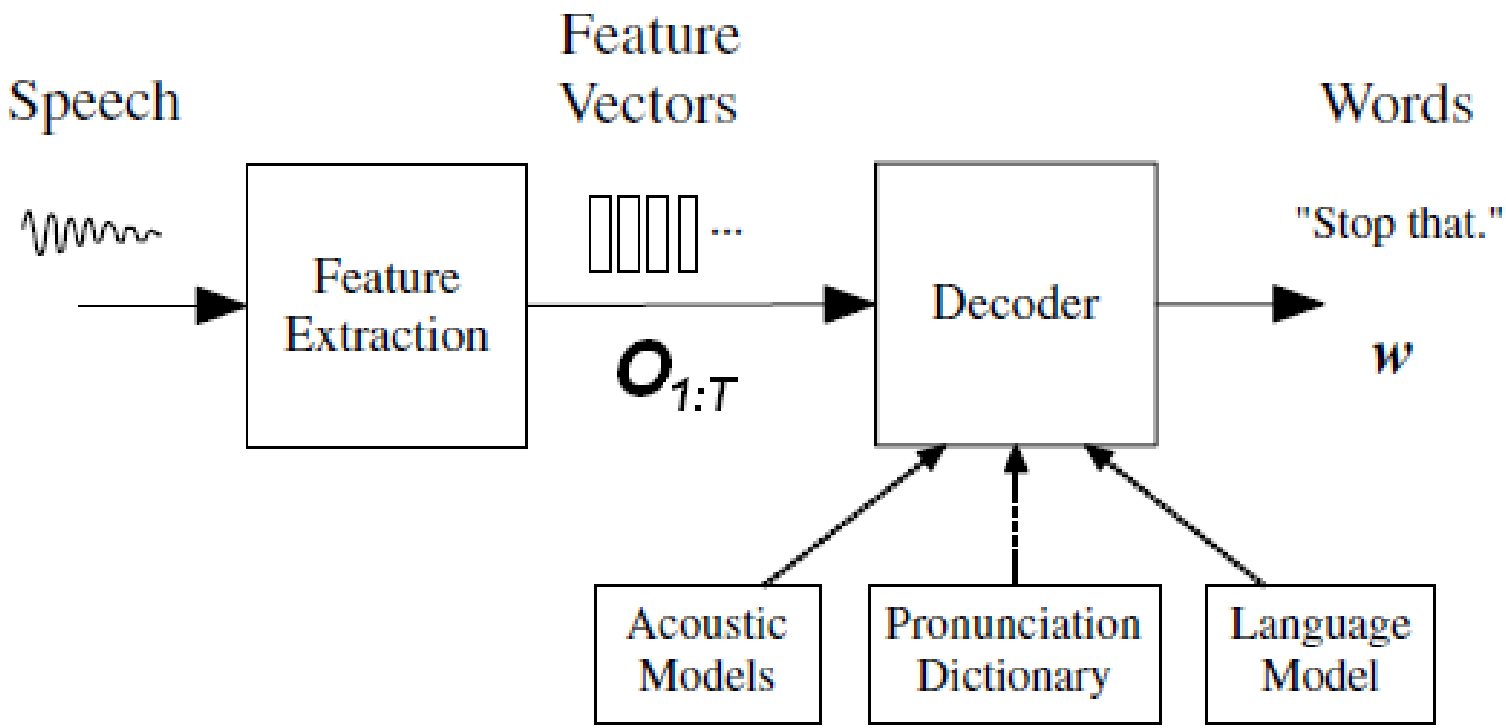
Statistical modeling



Speech recognition

- Speaker dependend / independent
- Isolated word / connected speech
- Language / lexicon
- Noisy channel
- HMMs

Architecture of HMM speech recognizer



Architecture

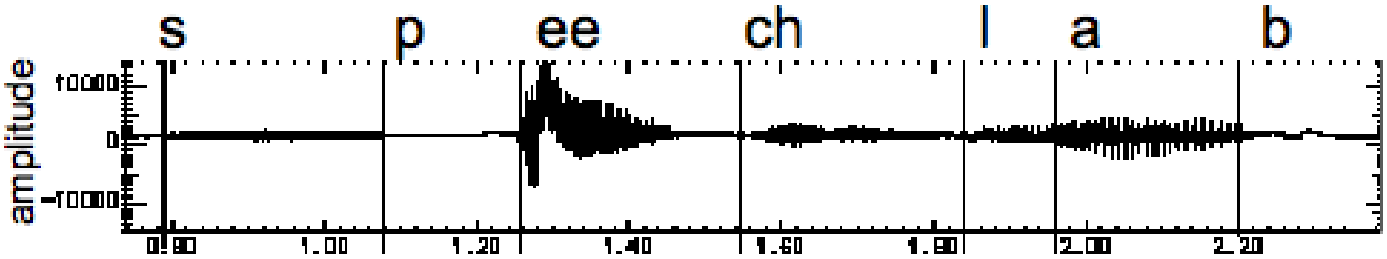
Acoustic input: $O = O_1 O_2 O_3 .... O_t$
Sentence output: $W = W_1 W_2 W_3 ... W_n$
What is the most probable sentence from language L?

$$W^* = \arg\max_{W \in L} P(W|O) = \arg\max_{W \in L} P(O|W)P(W)$$

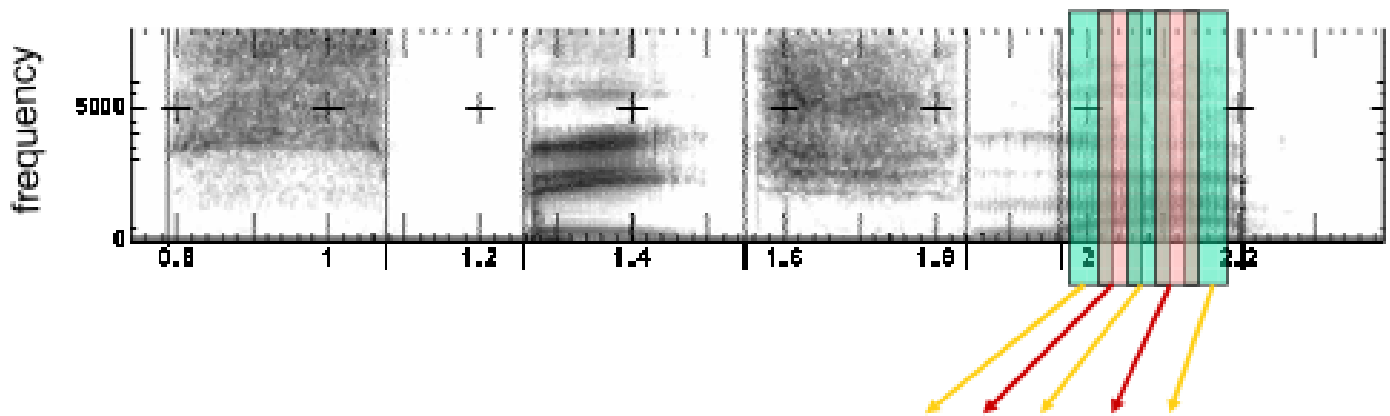
$P(O|W)$ - acoustic model
$P(W)$ - language model

Feature extraction

- Discrete time sampling computed every 10 ms
- Window 25ms
- Spectral transform DCT (FFT)
- spectral vectors (40 dim)

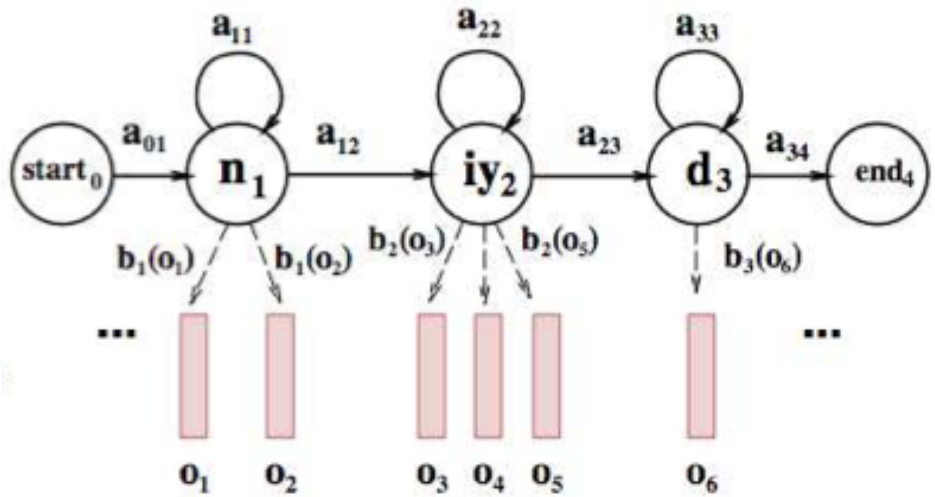


- Frequencies at each time slice processed into observation vectors



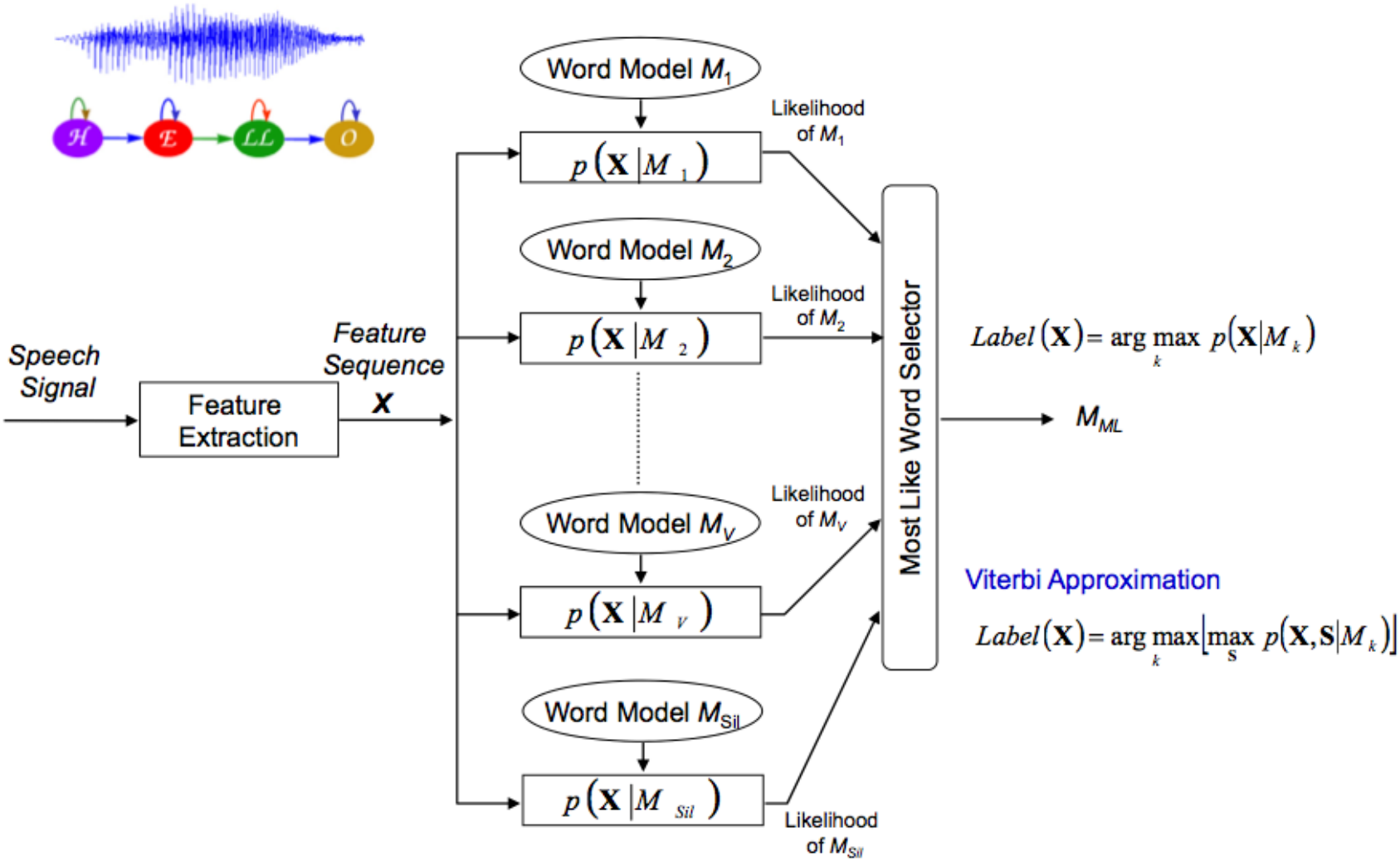
Acoustic model: training

- single word model: $P(O_i|W)$
- acoustic feature vector $O_i = \{o_1, o_2, o_3..\}$
- basic unit of sound: "phoneme" q
- monophone model, concatinating sequence of phones
- left-to-right HMM, $\lambda = (A, B, \pi)$
  $P(o_1, o_2..|\lambda) = \sum_{q_1,..q_T} \pi_{q_1} \prod_i b_{q_i}(o_i) a_{q_i q_{i+1}}$



Model λ parameters estimation, learning problem by Baum-Welch algorithm

Acoustic model: recognition



Probabilty evaluation problem, Forward-backward algorithm

Language model

Statistical language model: probability of a word sequence in a sentence

$$P(W) = P(W_1 W_2 .. W_N) = P(W_1)P(W_2|W_1)P(W_3|W_2, W_1)...$$
$$...P(W_n|W_1 .... W_{n-1}) = \prod_i^n P(W_i|W_1 .. W_{i-1})$$

unigram model: $P(W_1 W_2 .. W_N) = \prod_i^n P(W_i)$
bigram model: $P(W_1 W_2 .. W_N) = \prod_i^n P(W_i|W_{i-1})$

$$P(W_i|W_{i-1}) = \frac{count(W_{i-1}, W_i)}{count(W_{i-1})}$$

P(It is sunny today) =
= P(It|<beg>)P(is|it)P(sunny|is)P(today|sunny)P(<end>|today)

Algorithm

- Define set of words for modeling (lexicon)
- Collect labeled utterance of words (training set)
- Train HMMs on dictionary words, compute best model λ_i for each word
- For recognition for every acoustic word find $\arg\max_{\lambda_i} P(O|\lambda_i)$
- Compute most probable sequence of words in the sentence