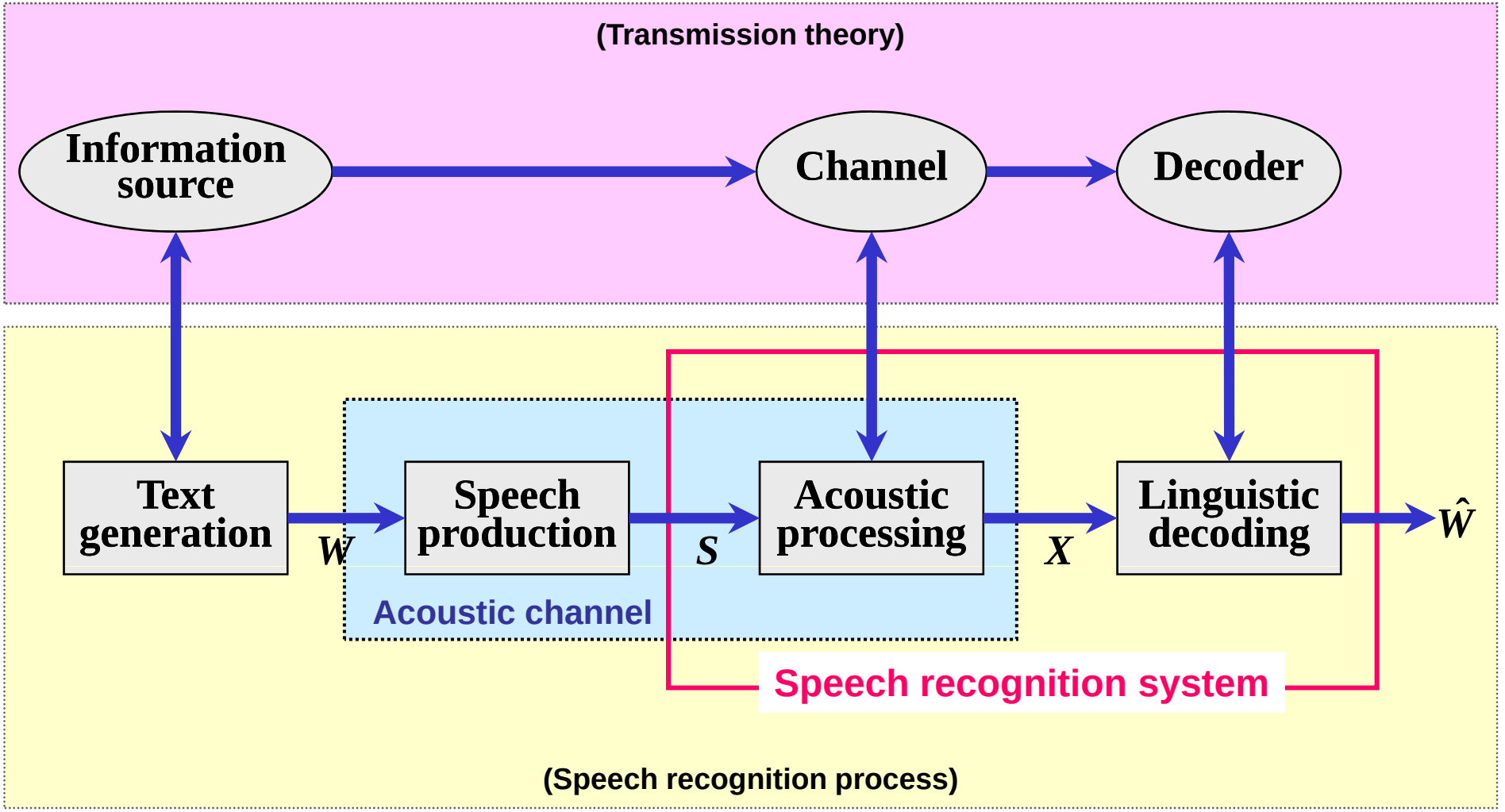
Speech Recognition (Acoustic Modeling)

Sadaoki Furui
Tokyo Institute of Technology
Department of Computer Science
furui@cs.titech.ac.jp

Structure of speech production and recognition system based on information transmission theory

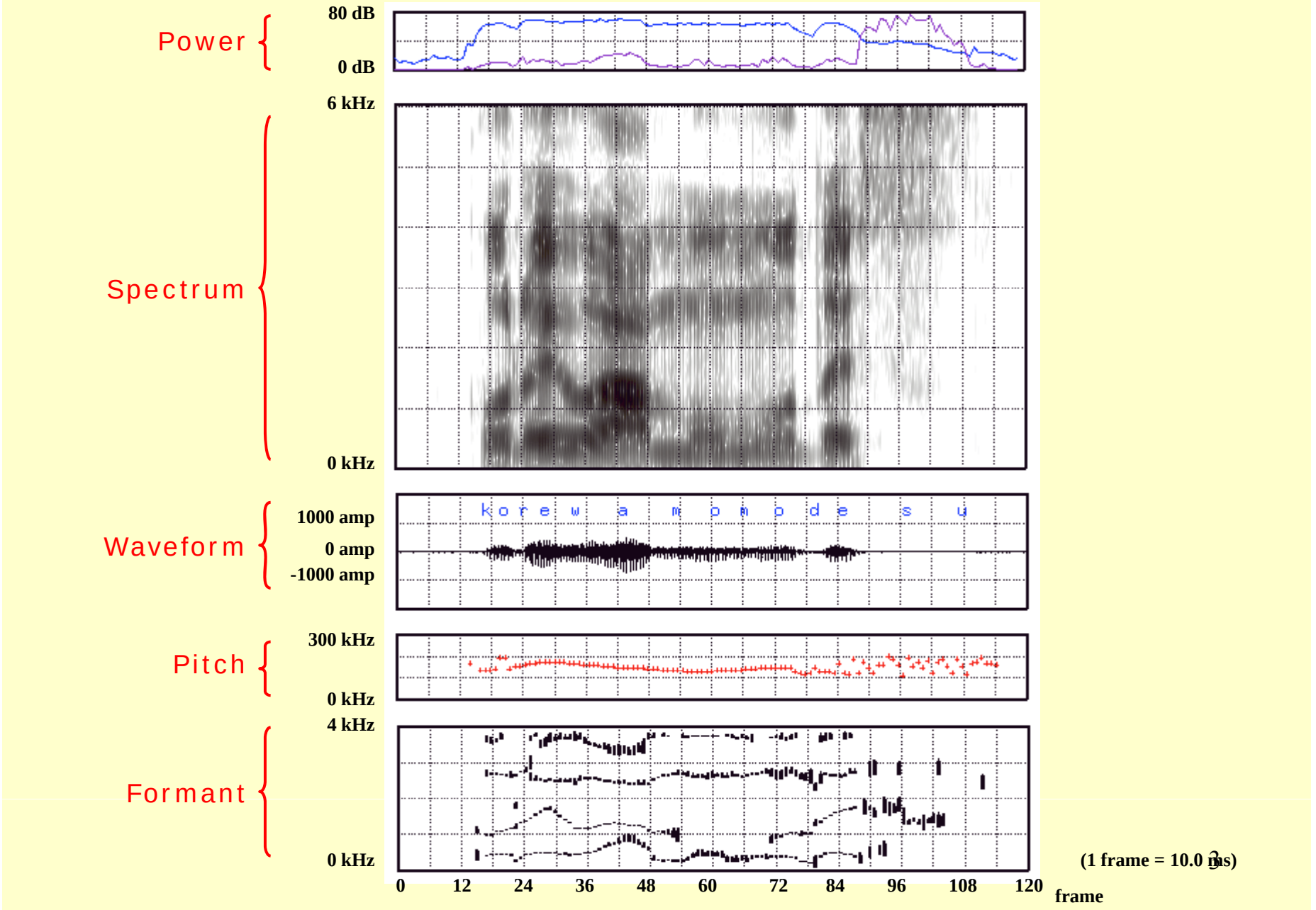


$$\hat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)}$$

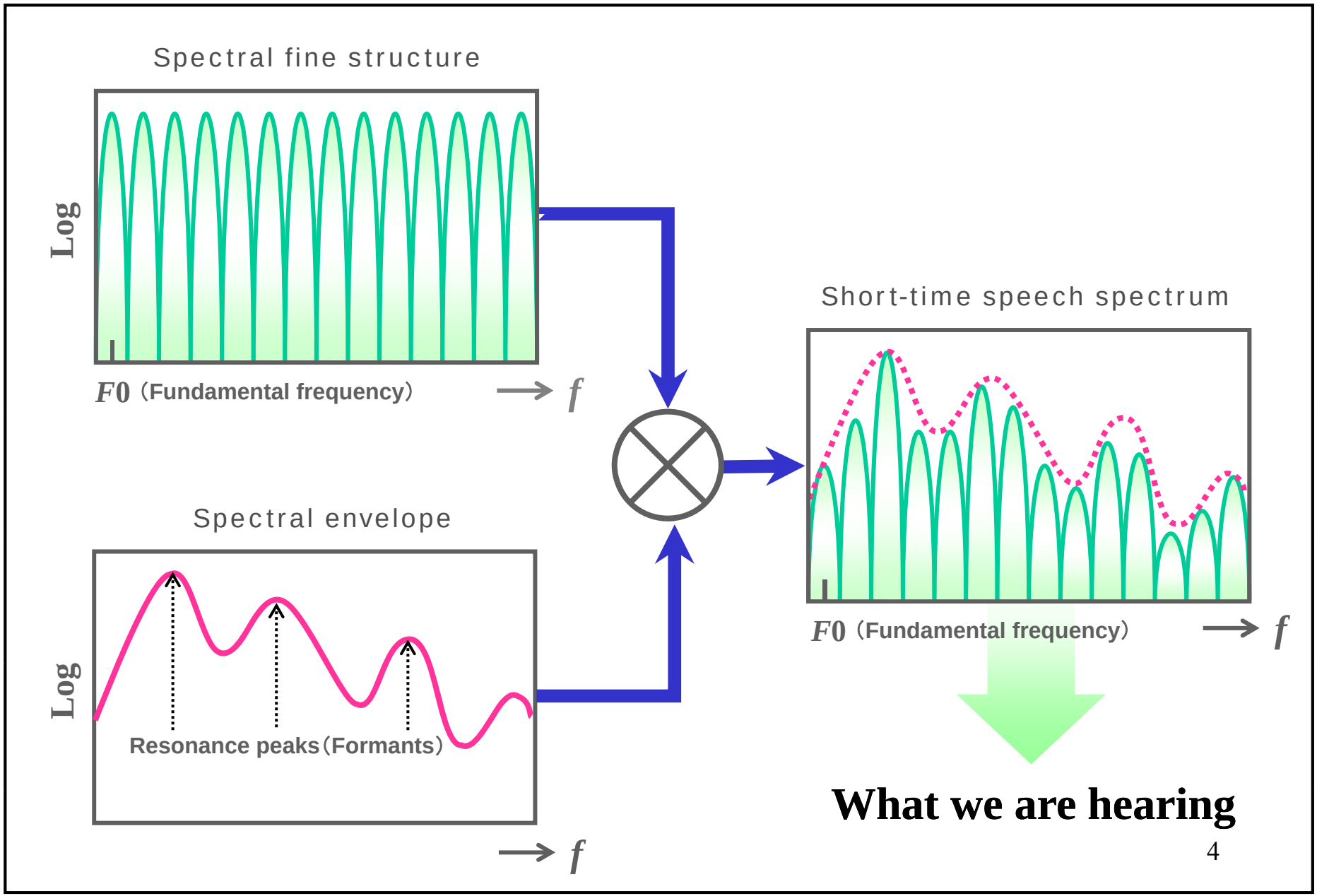
Choosing the right observation

1. Statistical characteristics
2. Associated distortion measures
3. Physical significance
4. Processing convenience
5. Effectiveness for the task

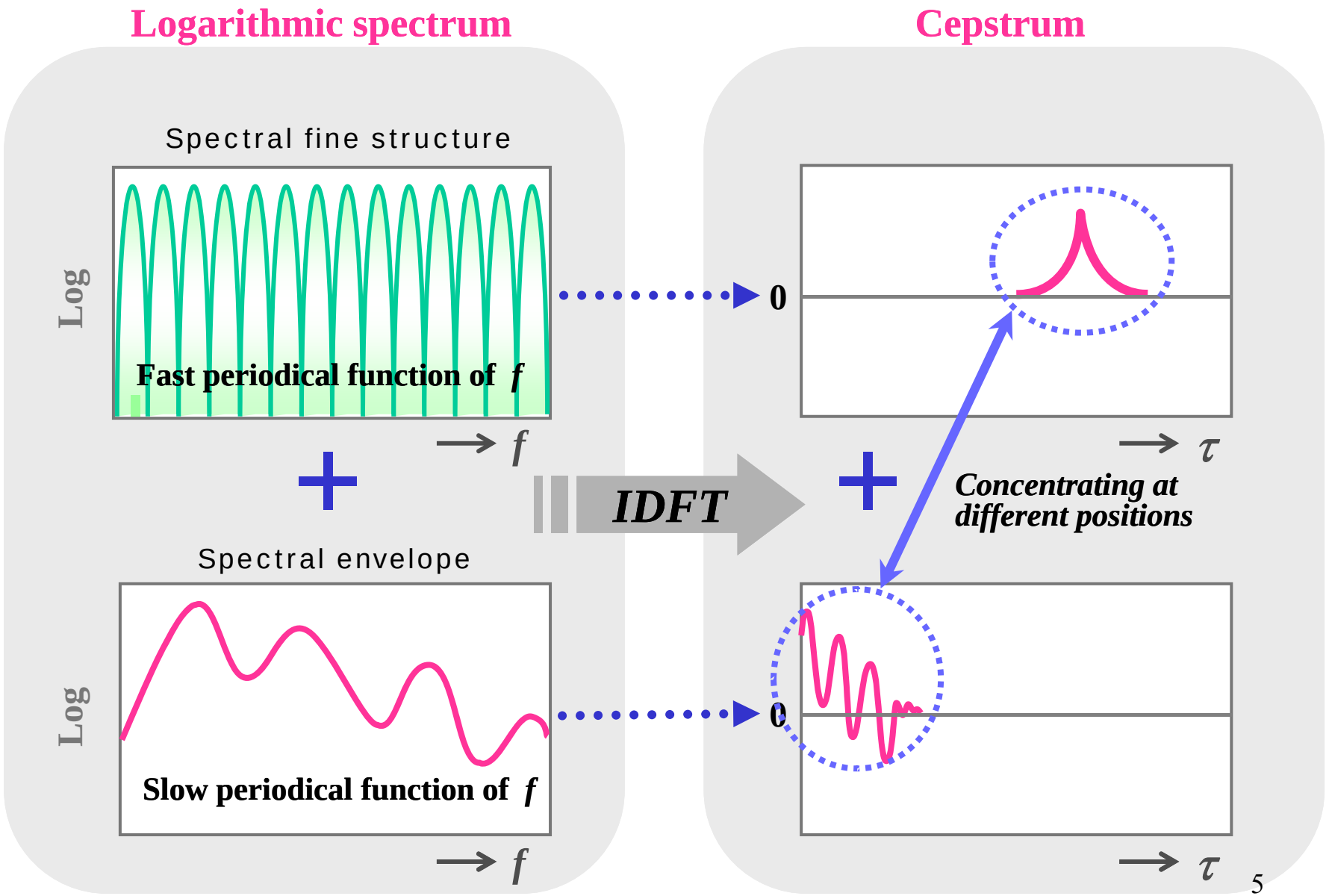
Digital sound spectrogram



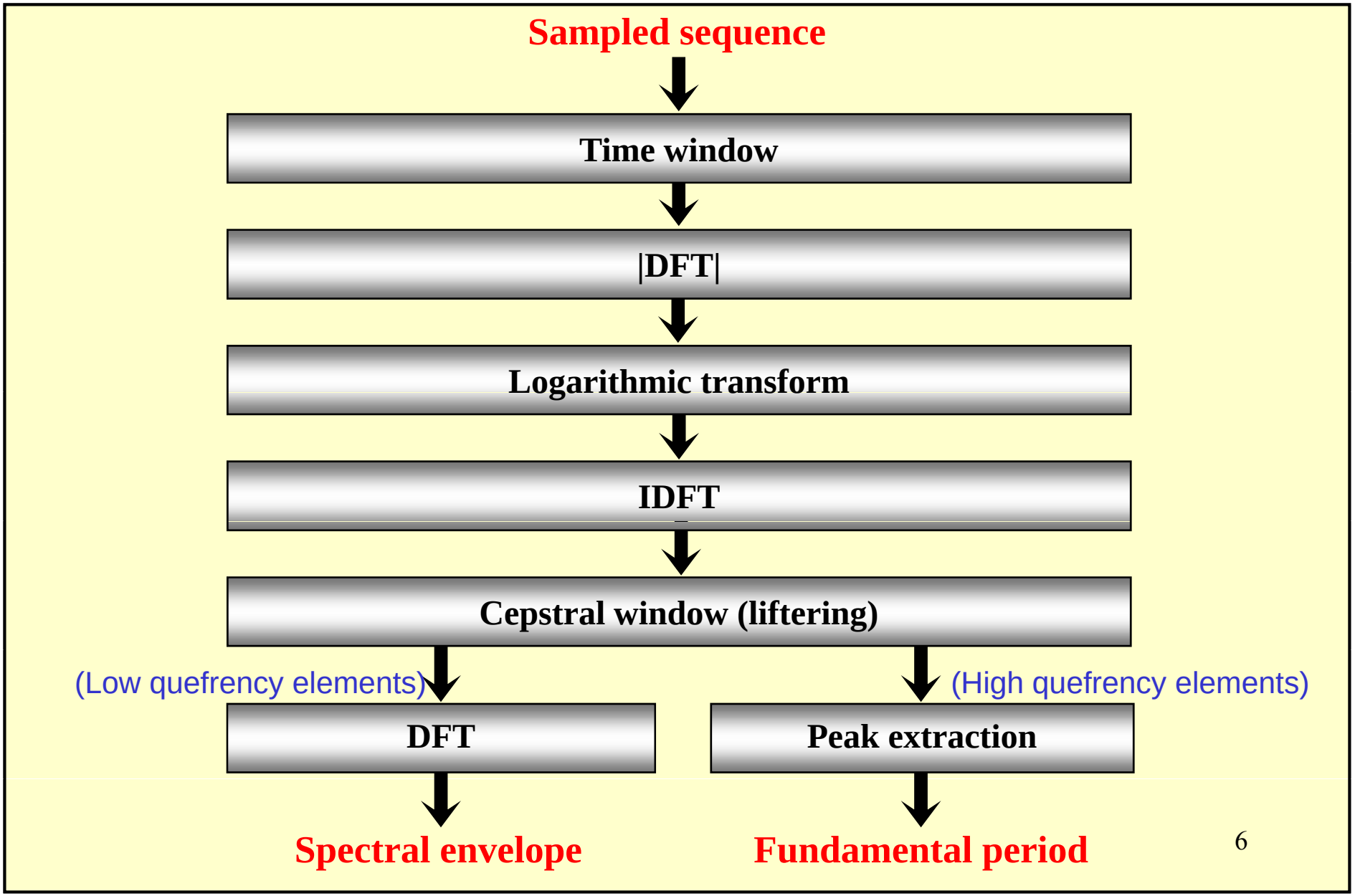
Spectral structure of speech



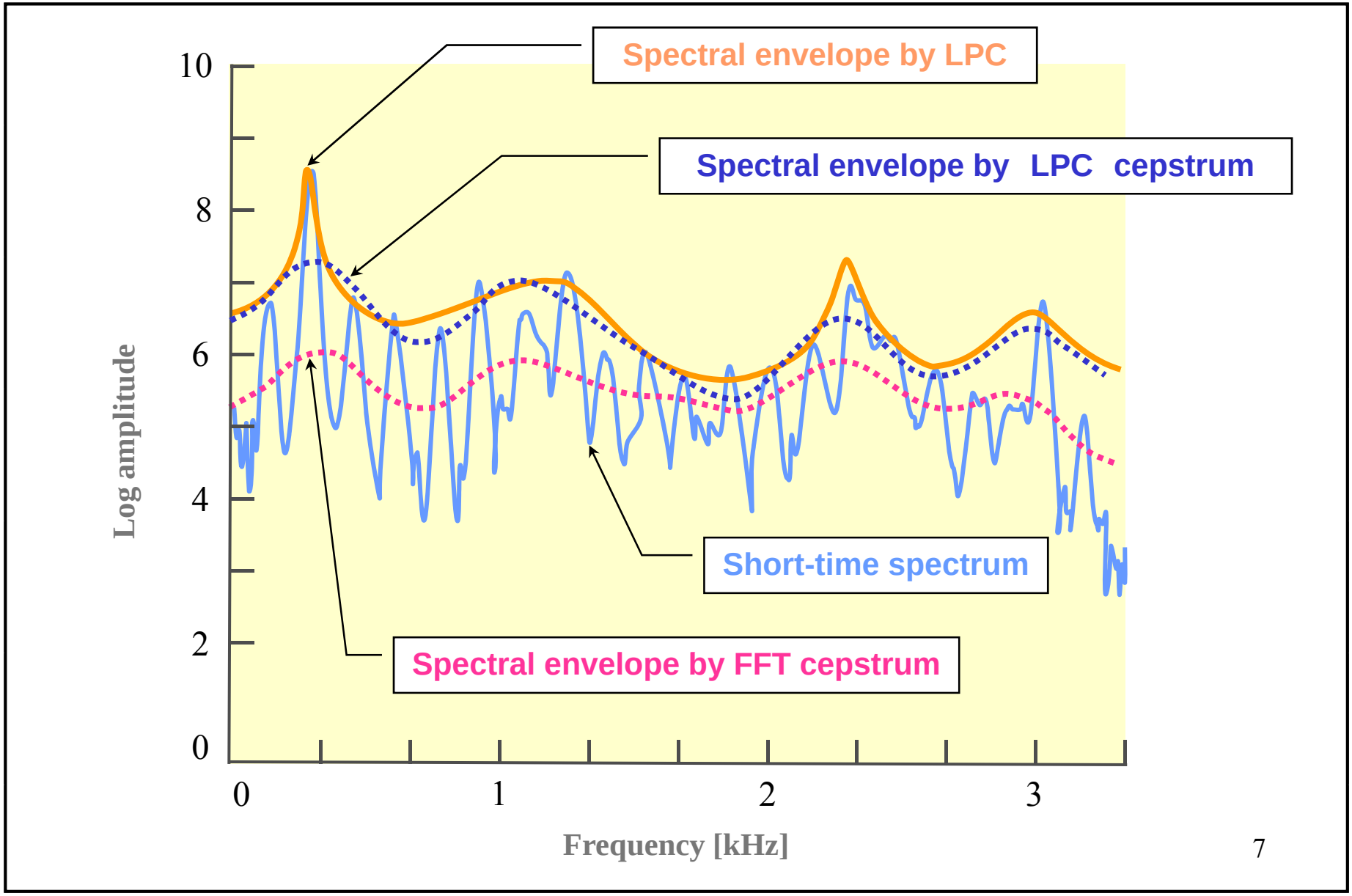
Relationship between logarithmic spectrum and cepstrum



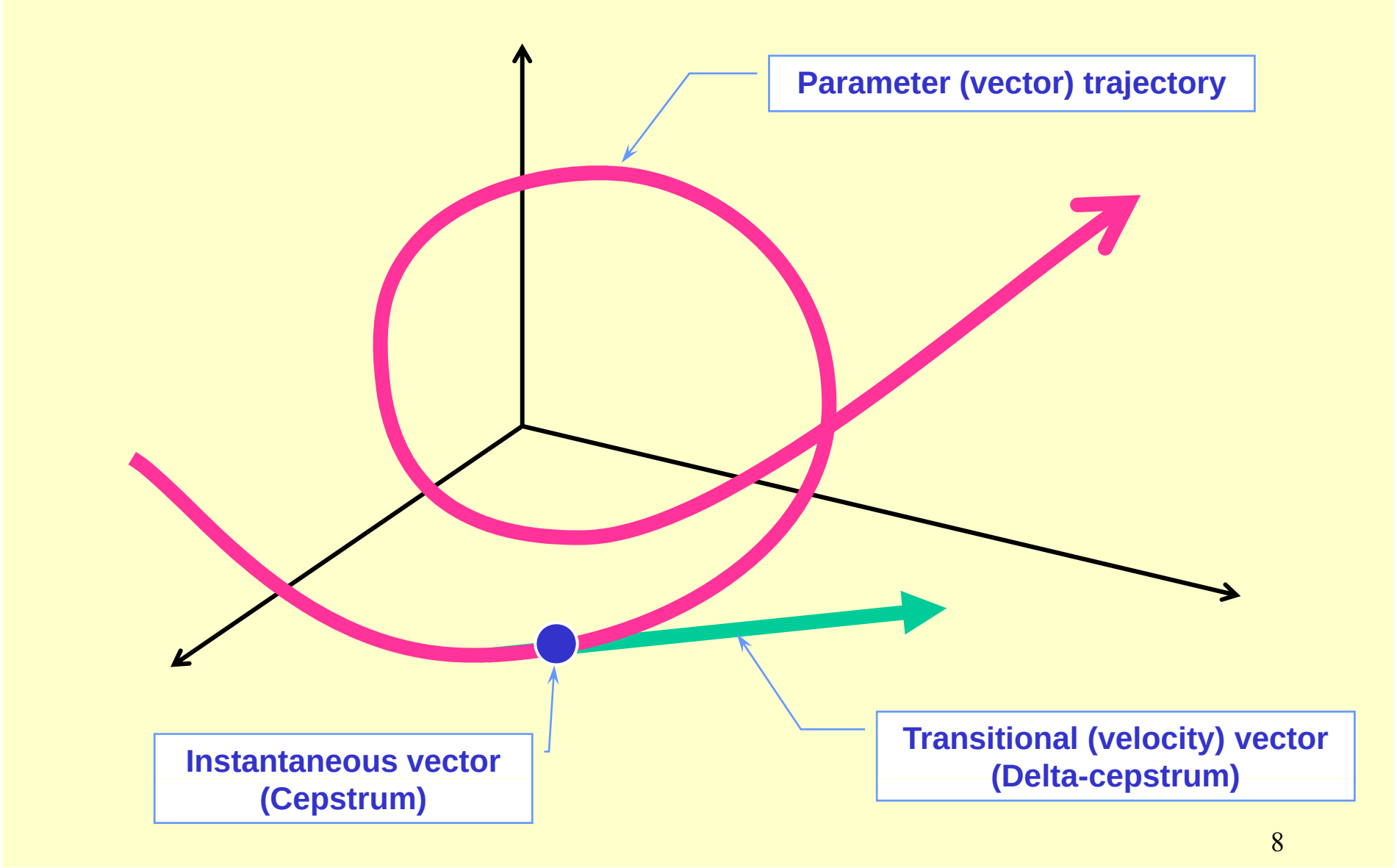
Block diagram of cepstrum analysis for extracting spectral envelope and fundamental period



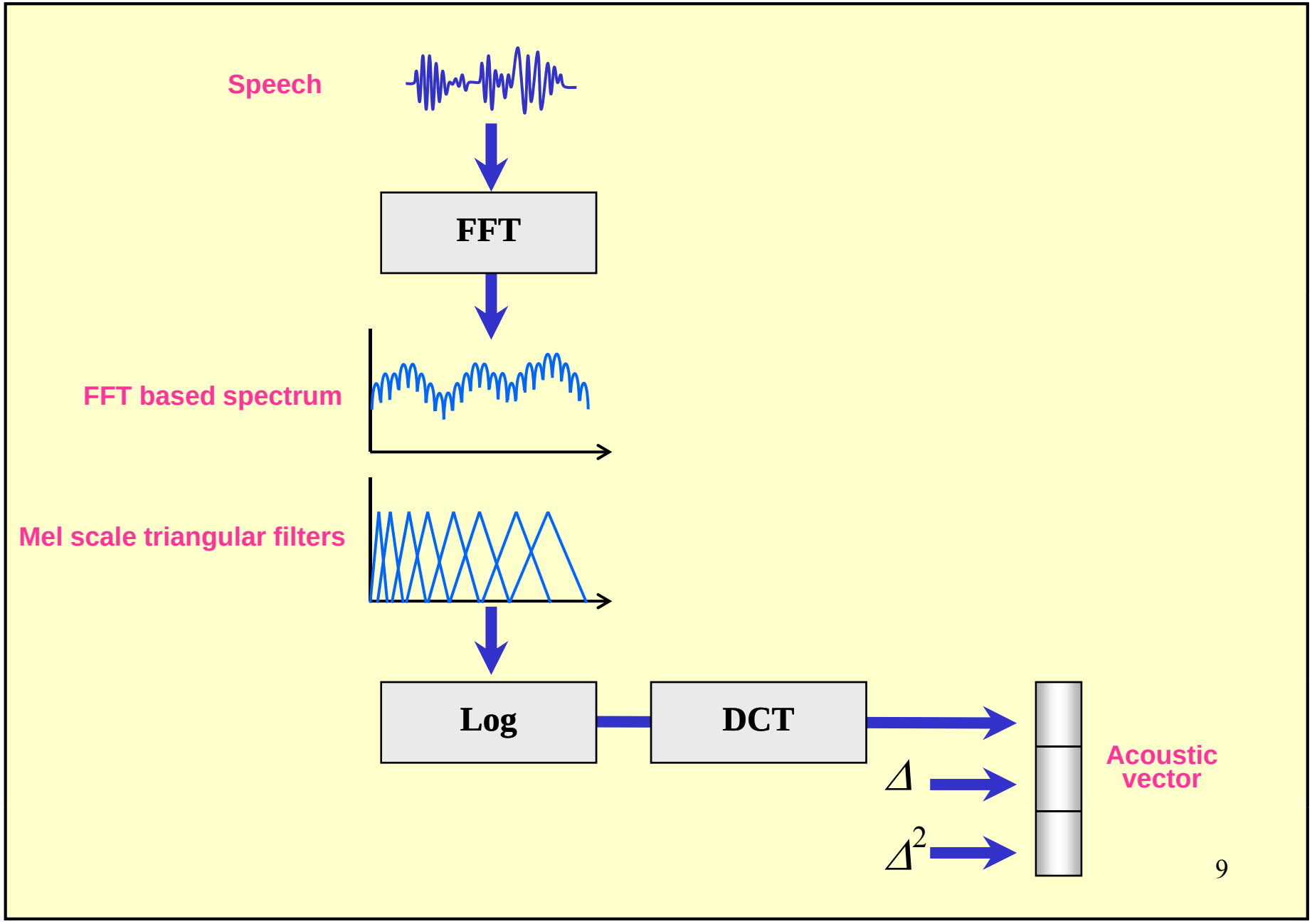
Comparison of spectral envelopes by LPC, LPC cepstrum, and FFT cepstrum methods



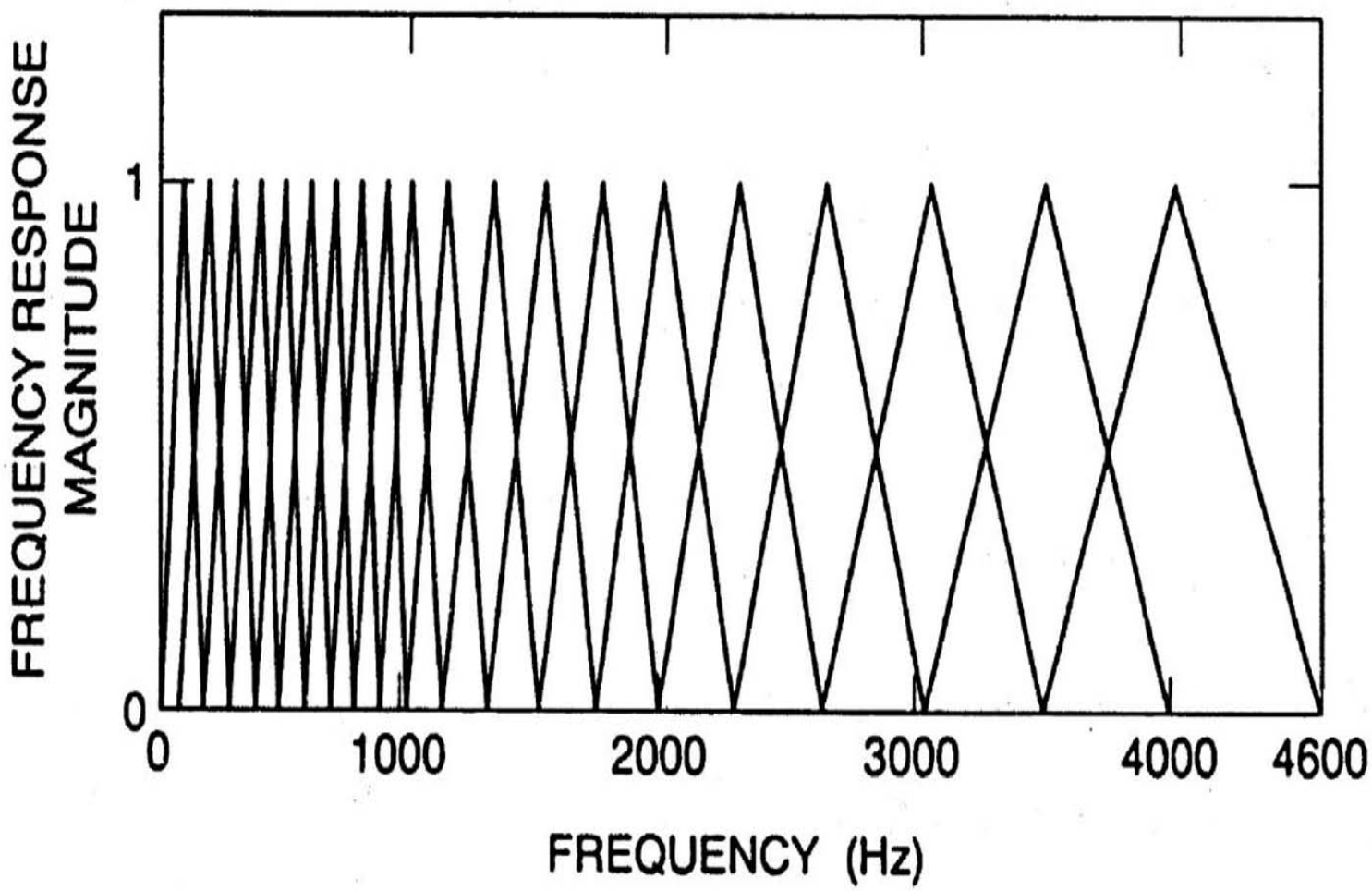
Cepstrum and delta-cepstrum coefficients



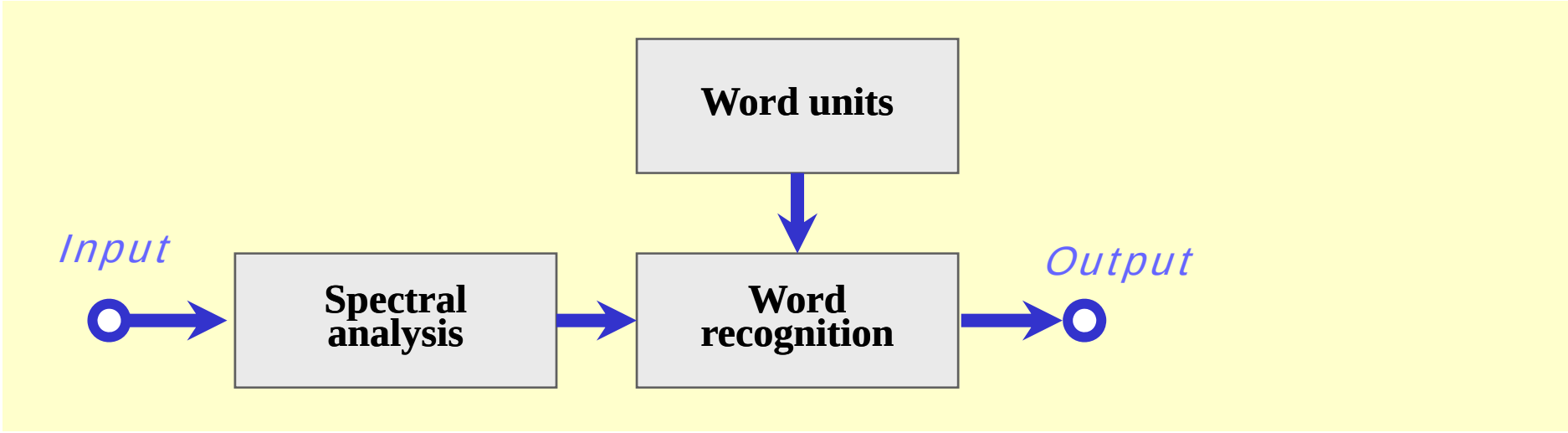
MFCC-based front-end processor



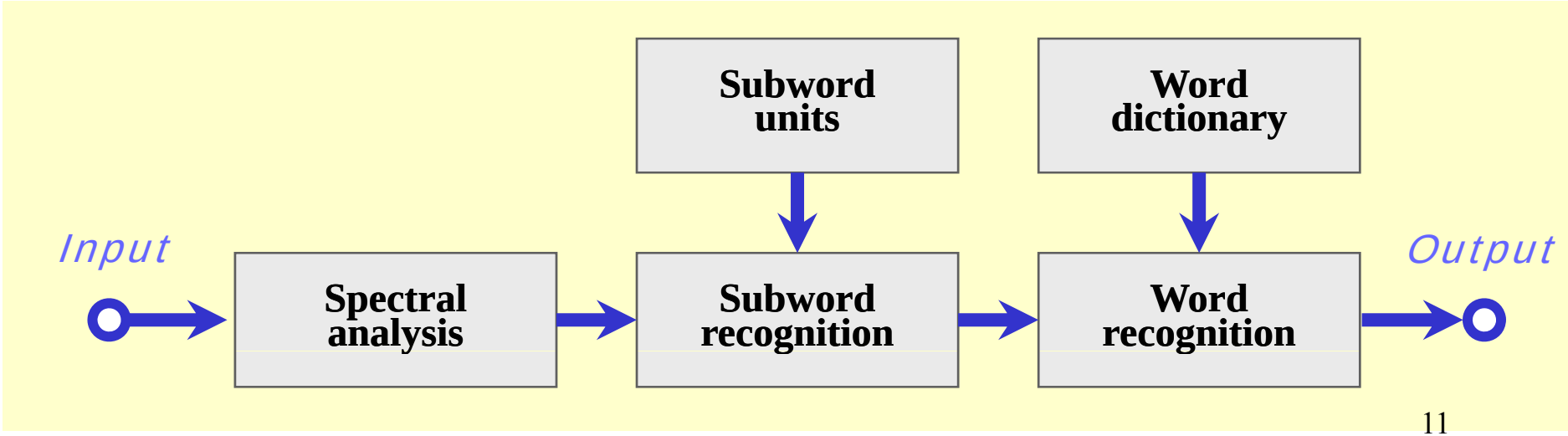
Filter bank for generating mel-based cepstral coefficients



Structure of word recognition systems

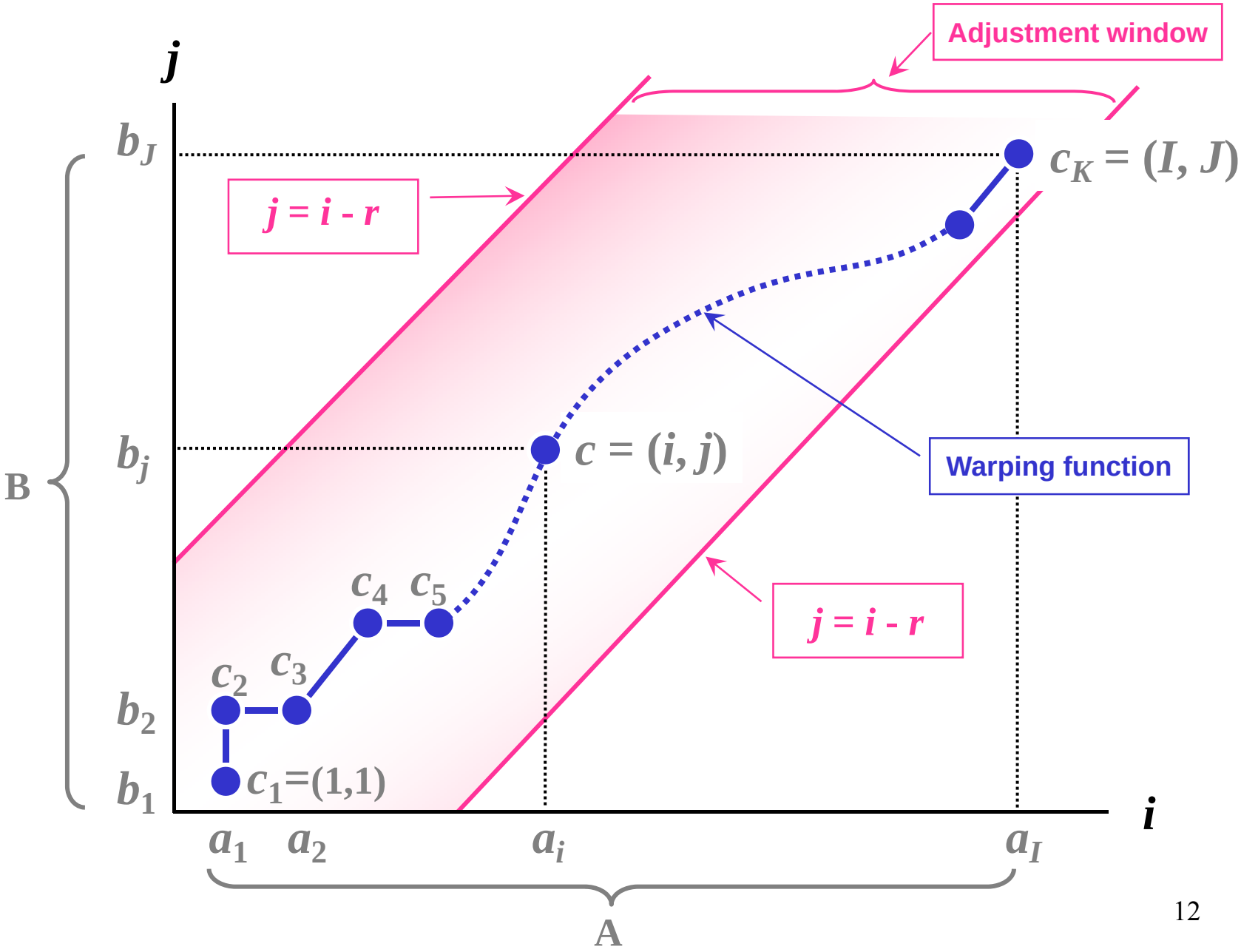


(a) word-based recognition



(b) subword-based recognition

DTW between two time sequences, A and B



HMM for speech recognition

1. **Isolated word recognition**
2. **Connected word recognition**
3. **Speech recognition using subword units**
4. **Language models**

Isolated word recognition with HMM

Vocabulary: M words

Training data: K utterances each of the M words.

1. For each word m, construct HMM λ_m using the training utterances for that word.

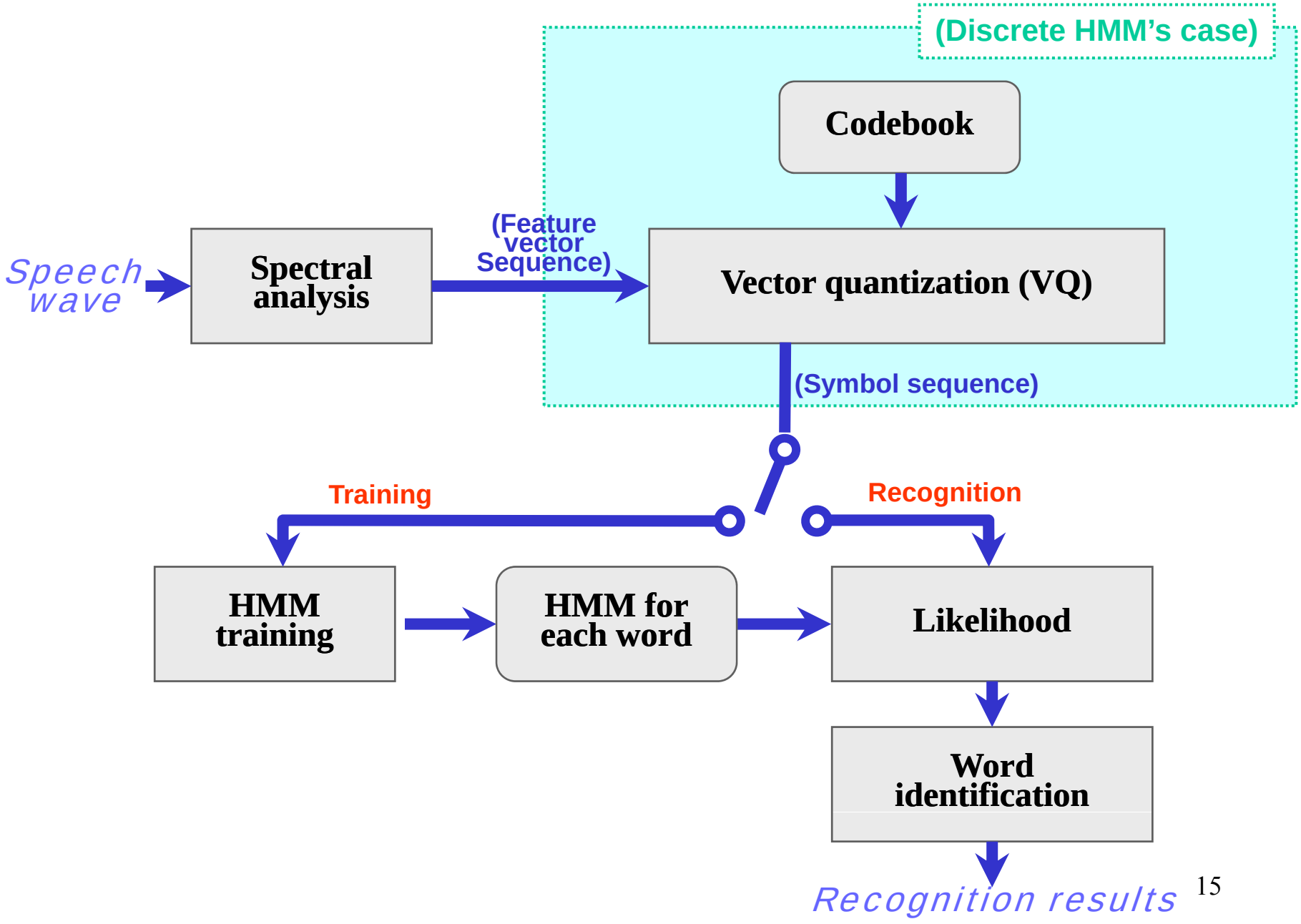
2. For each unknown word x to be recognized:

--- evaluate $P(x|\lambda_m)$, $m = 1, 2, \ldots, M$

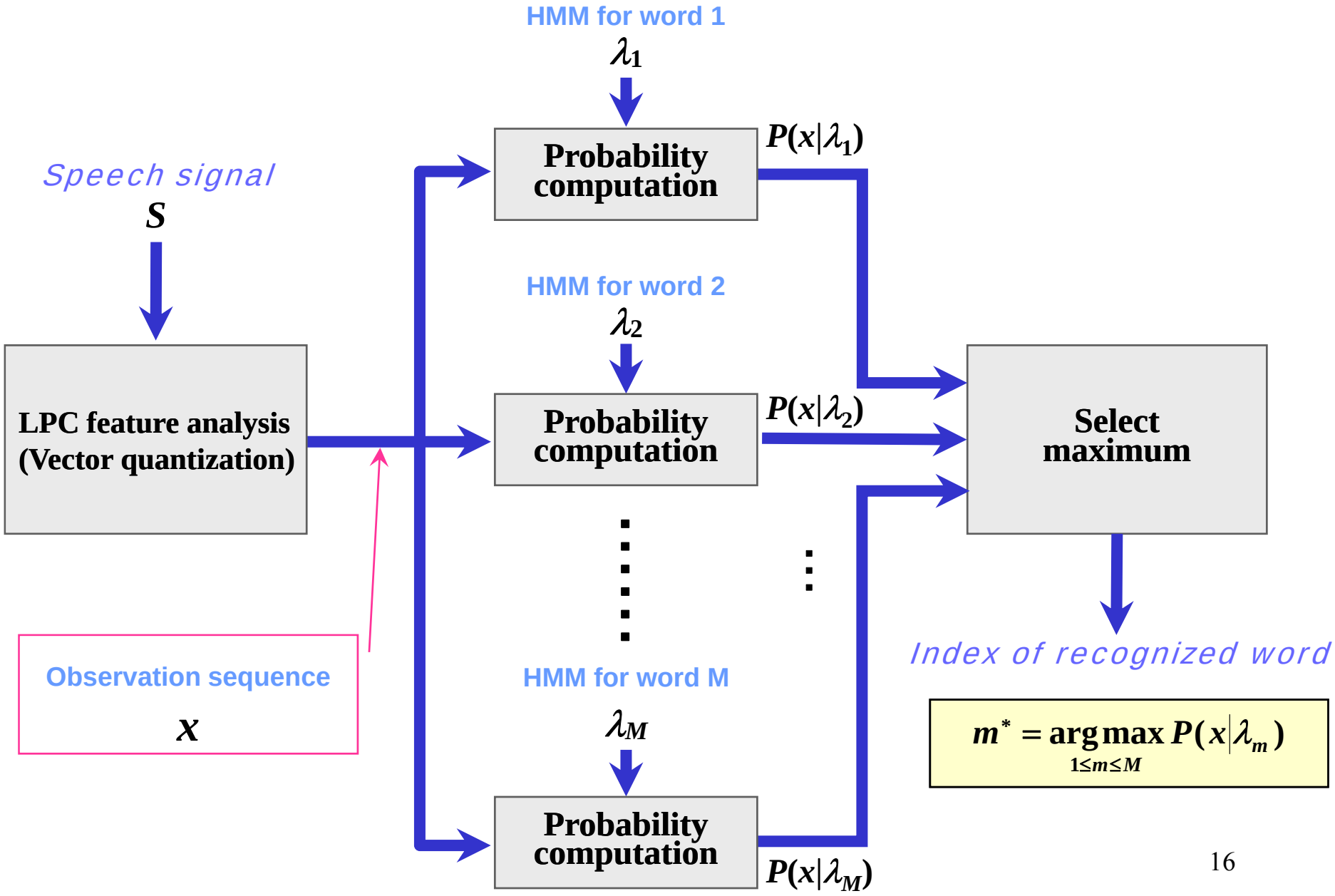
--- select word whose model likelihood is highest, i.e.

$$m^* = \arg\max_{1, 2, \ldots, M} P(x|\lambda_m)$$

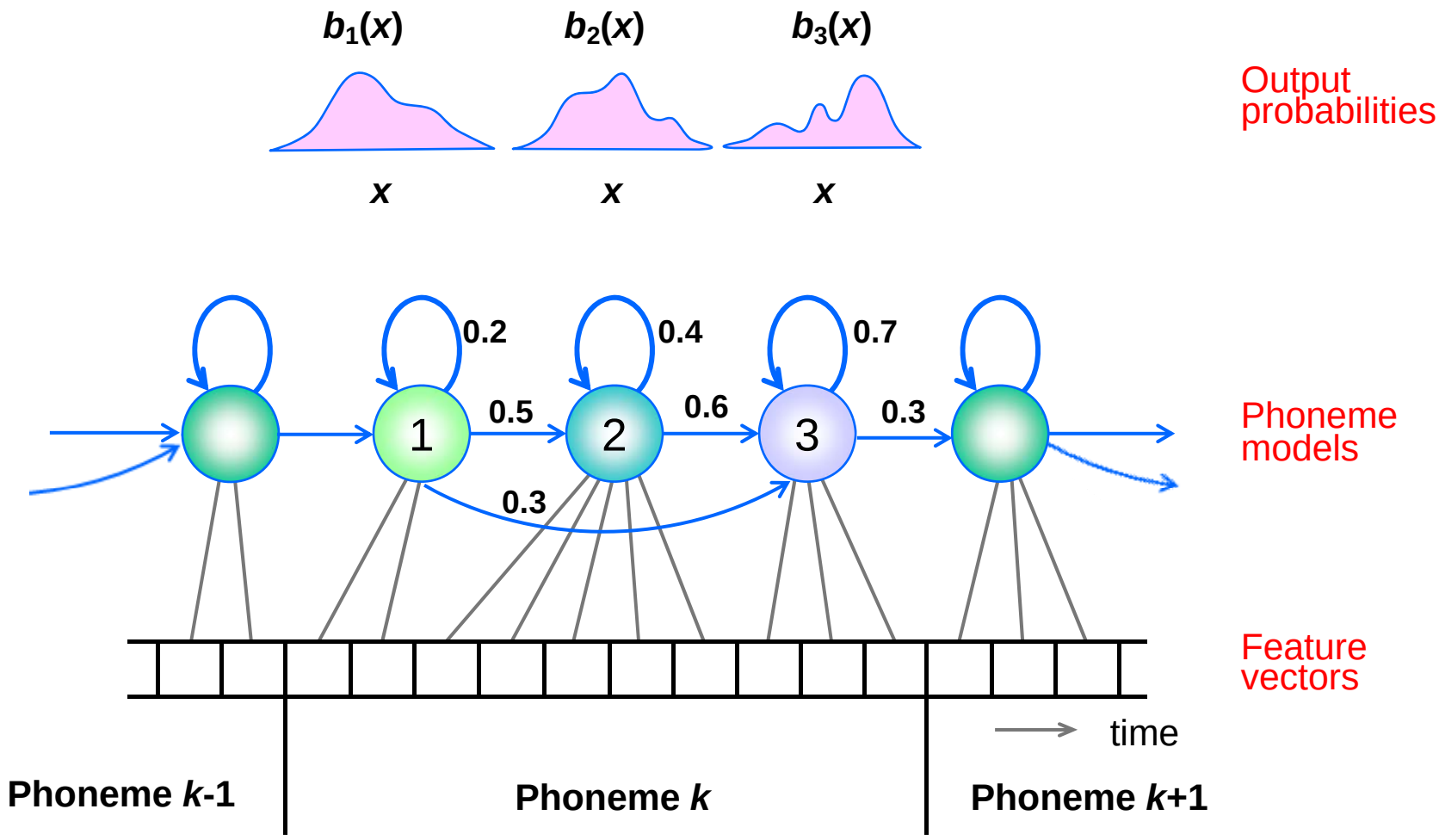
Structure of word recognizer based on HMM



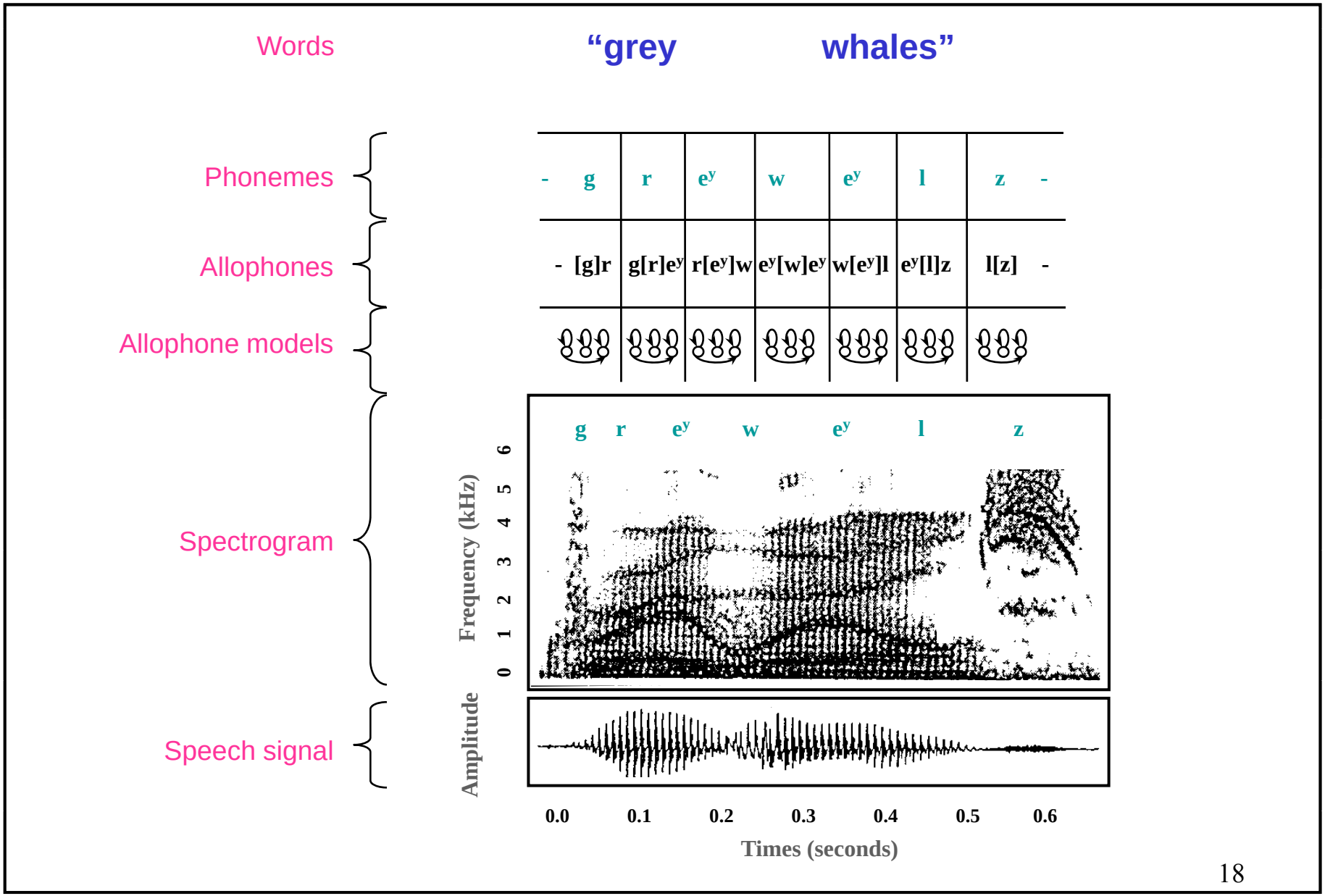
Block diagram of an isolated word HMM recognizer



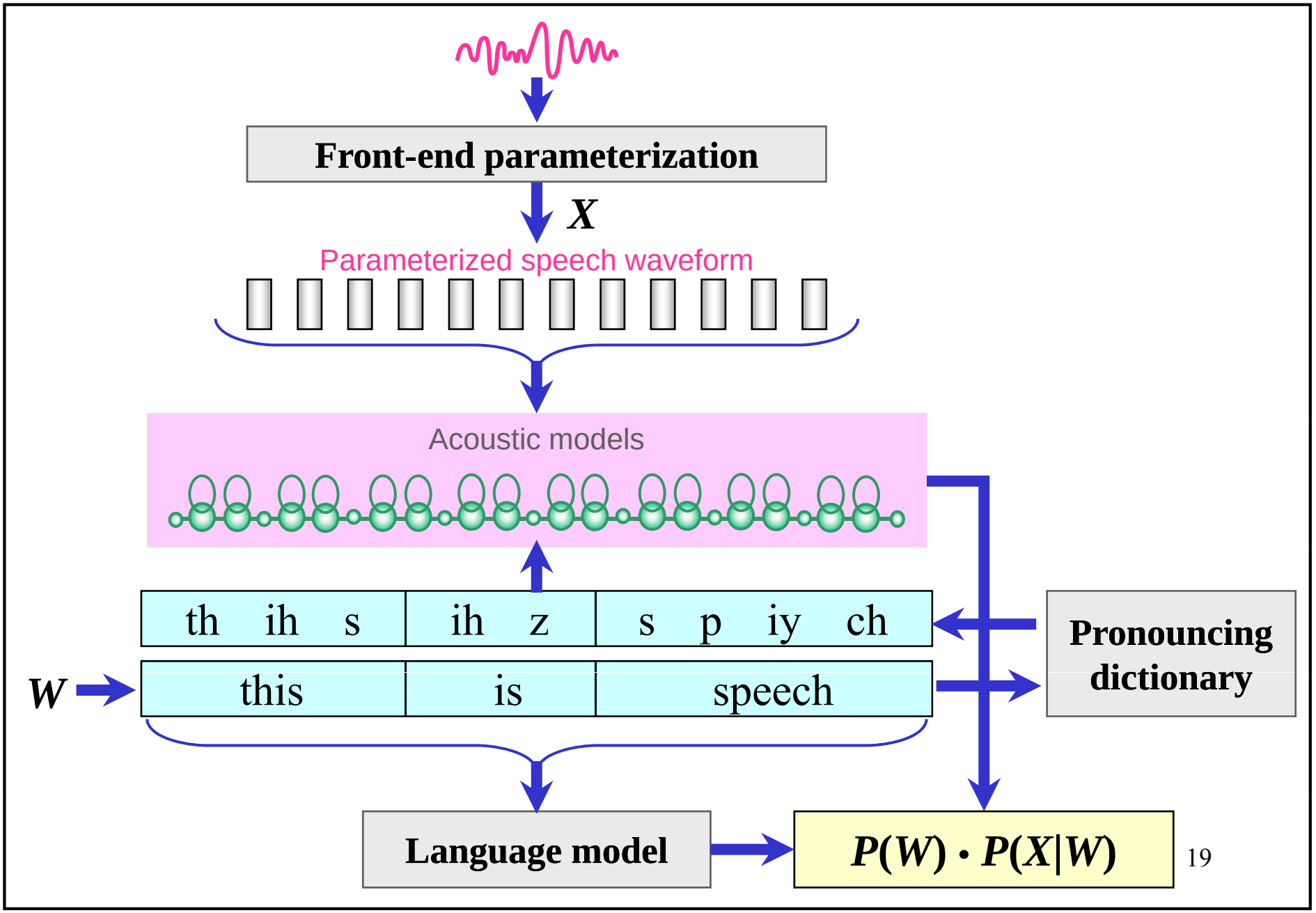
Structure of phoneme HMMs



Units of speech



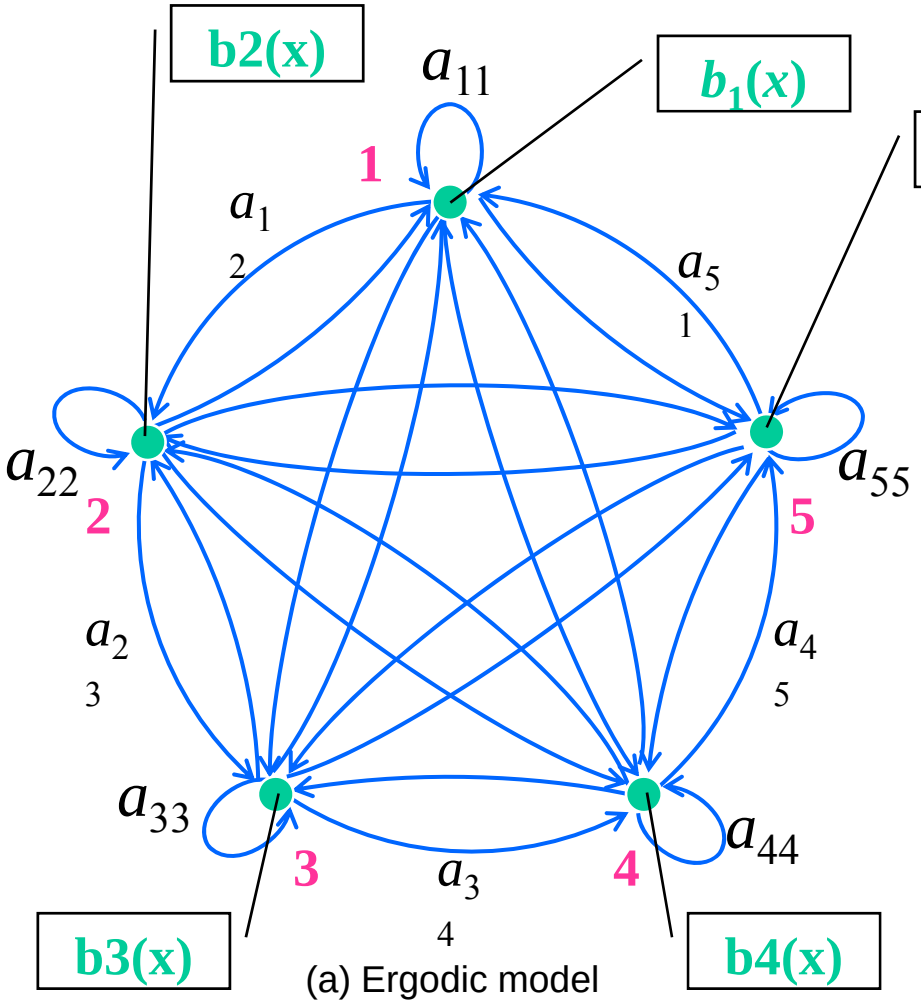
Overview of statistical speech recognition



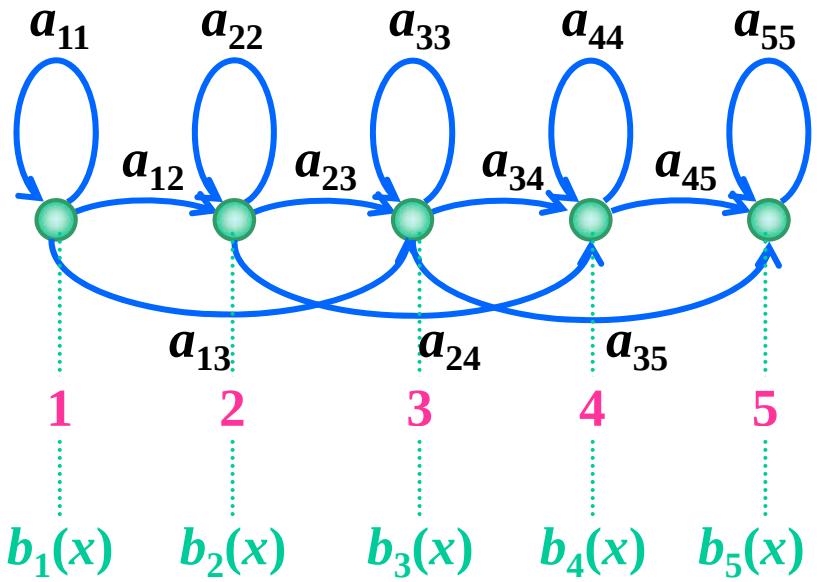
Choice of model structure

1. Left-to-right models vs. ergodic models
2. Number of states
   - number of distinct sounds in word
   - number of observations in word
   - relation with other competing word
3. Observation
   - continuous vectors
   - discretized(VQ) indices
4. Number of mixtures
   - statistical characteristics of observation vectors (single speaker, multiple speaker, speaker independent case, etc.)
   - robustness of the model.

Typical structures of HMM used in speech recognition



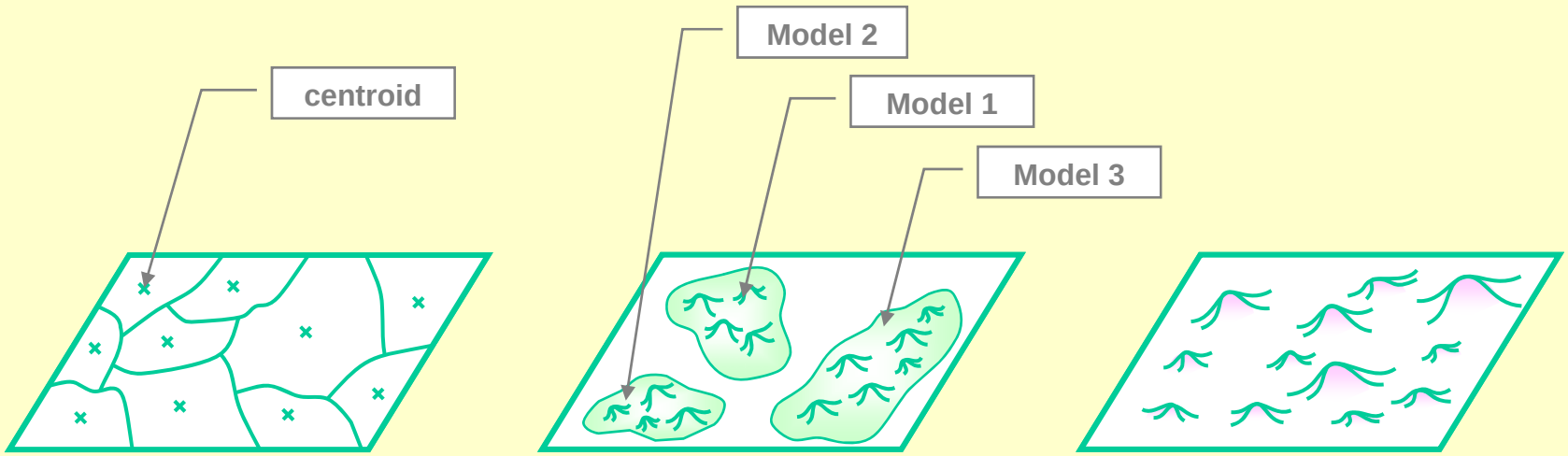
(a) Ergodic model



(b) Left-to-right model

a_{ij}: transition probability, $b_i(x)$: observation probability

Representations of the acoustic space of speech

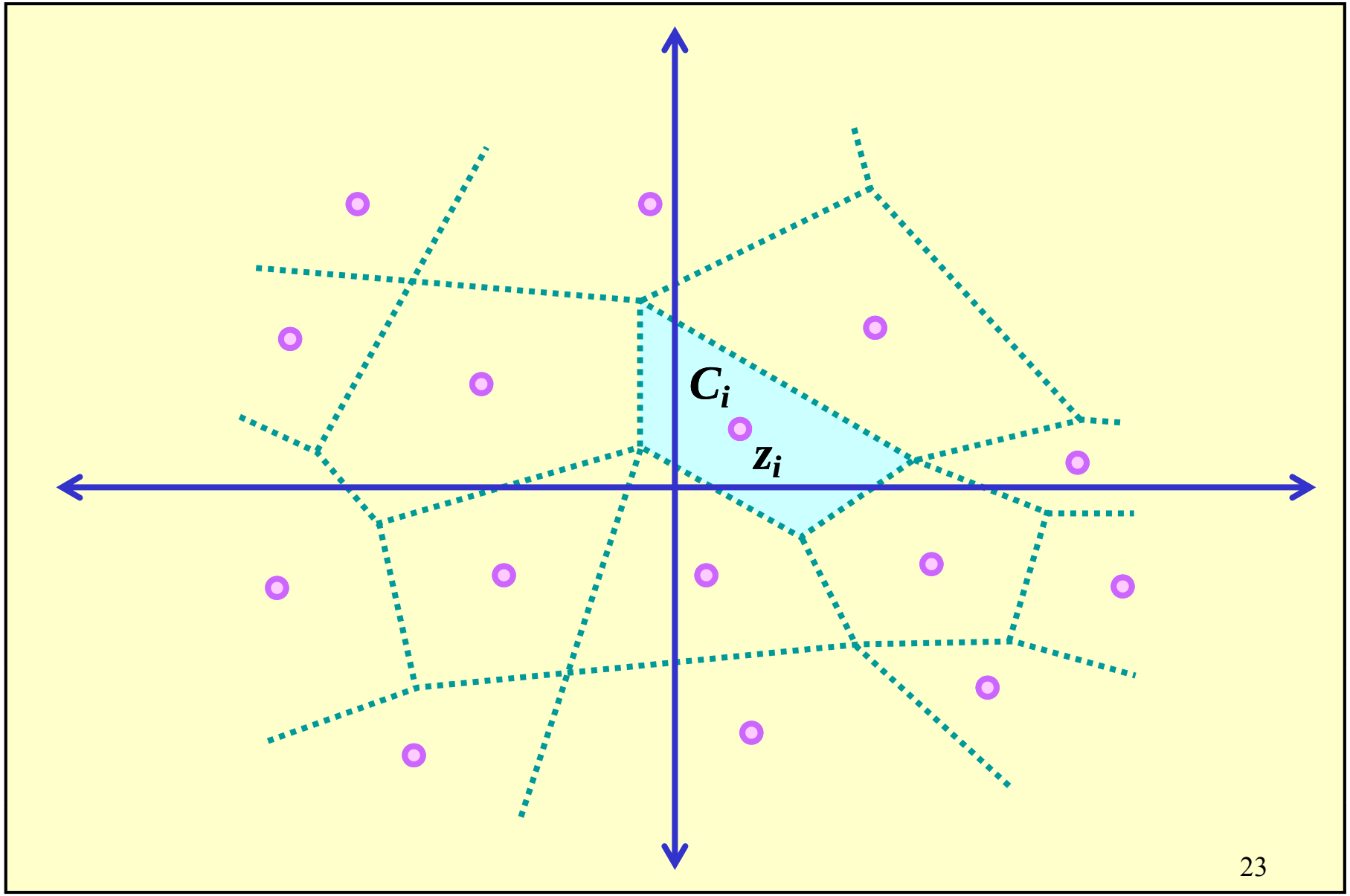


(a)
Discrete density VQ codebook

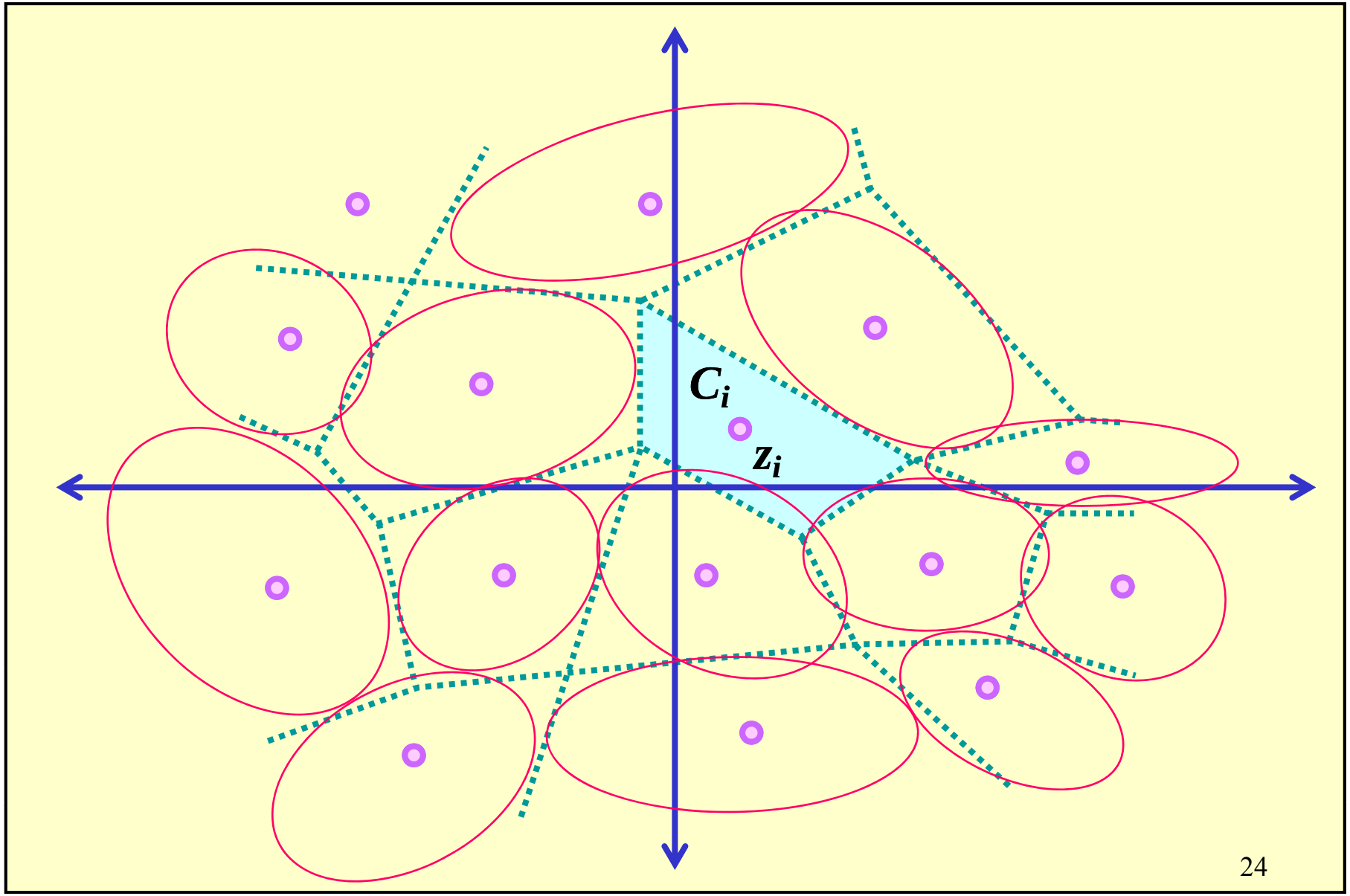
(b)
Continuous density mixture case

(c)
Continuous density codebook (semi-continuous case)

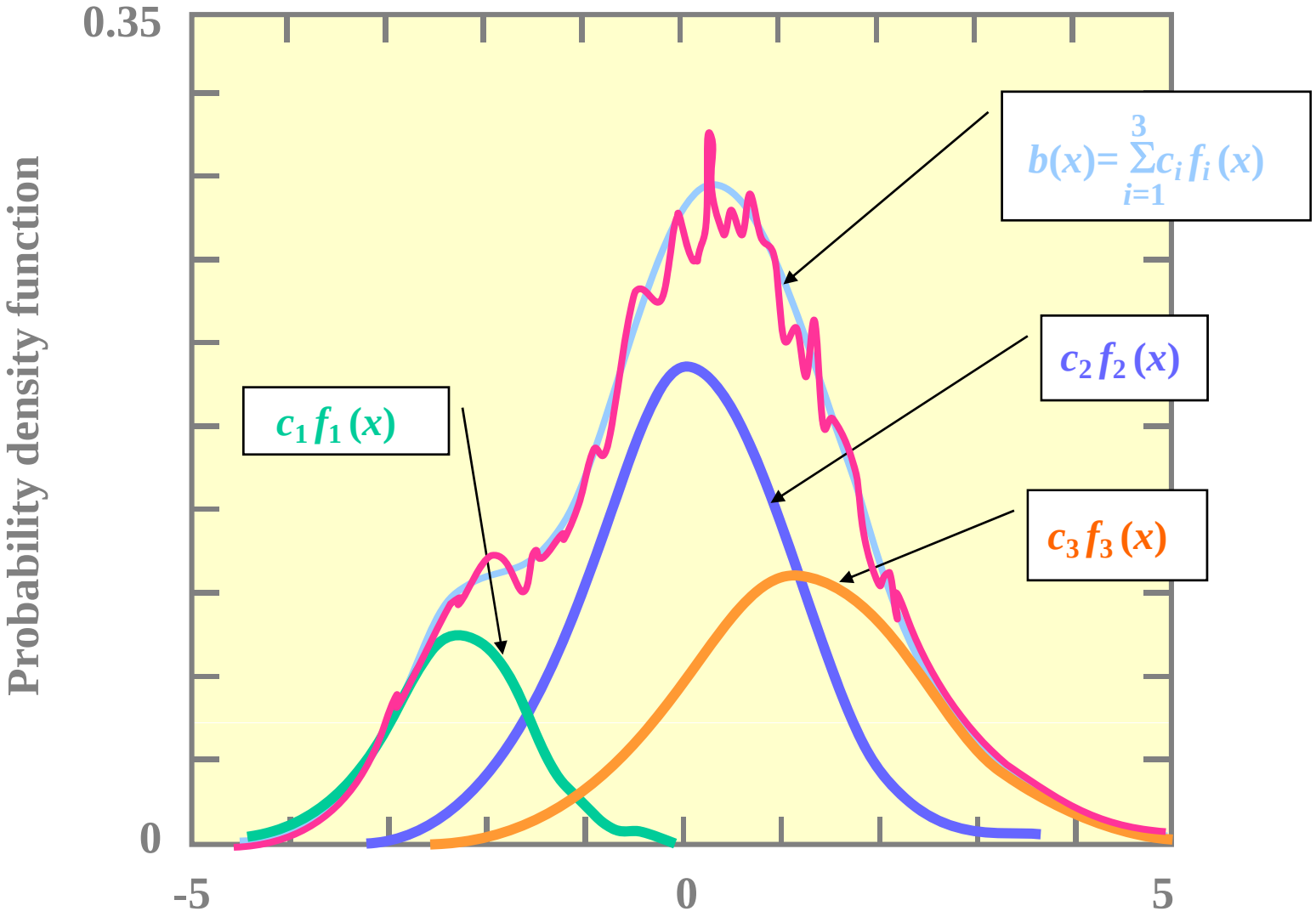
Partitioning of a two-dimensional space into 16 cells



Partitioning of a two-dimensional space into 16 cells



A mixture density function



Probabilistic functions of Markov chains hidden Markov models

1. N, number of states in the model
   - State index $i \in \{1, 2, \ldots, N\} = z_N$
   - State at time t, $s_t = i \in z_N$
2. State transition probabilities
   - $A = [a_{ij}]$
   - $a_{ij} = P\{s_t = j \mid s_{t-1} = i\}$
3. Observation probability distributions
   - $B = \{b_i(x_t)\}_{i=1}^{N}$
   - Observation at time t, x_t
   - $b_i(x_t) = P\{x_t \mid s_t = i\}$
4. Initial state distribution
   - $\pi' = [\pi_1, \pi_2, \ldots, \pi_N]$
   - $\pi_i = P[s_0 = i]$

Three basic problems of hidden Markov models

- Evaluation of the probability or likelihood, $P(x|\lambda)$.
- Determination of the most likely state sequence s that produces the observation x.
- Estimation of the Markov model parameters, $\lambda = (\pi, A, B)$.

HMM: $\lambda = (\pi, A, B)$

What is the probability of an observation sequence $x=\{x_t\}_{t=1}^{T}$ being generated by the hidden Markov source?

Consider a particular state sequence

$$s=\{s_t\}_{t=0}^{T}=\{s_0, s_1, s_2, \ldots, s_T\}$$

$$P(s|\lambda) = \pi_{s_0} a_{s_0 s_1} a_{s_1 s_2} \ldots a_{s_{T-1} s_T}$$

$$P(x|s, \lambda) = b_{s_1}(x_1)\, b_{s_2}(x_2) \ldots b_{s_T}(x_T)$$

$$P(x, s|\lambda) = P(x|s, \lambda)\, P(s|\lambda)$$

$$= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1} s_t}\, b_{s_t}(x_t)$$

Then,

$$P(x|\lambda) = \sum_{\text{all } s} P(x, s|\lambda)$$

$$= \sum_{\text{all } s} \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(x_t)$$

There are N^{T+1} possible s sequences. Direct calculation of $P(x|\lambda)$ requires $2TN^{T+1}$ multiplications, without counting the calculations for all the $b_{s_t}(x_t)$.

- $N = 5, T = 99 \rightarrow \sim 10^{72}$ multiply

- **Forward probabilities**

  $\alpha_t(i)$: joint probability of partial observation sequence $x_1, x_2, \ldots, x_t$ and state i at time t.

$$\alpha_t(i) = P(x_1, x_2, \ldots, x_t, s_t = i \mid \lambda)$$

- **Forward recursion**

  1. Initialization

$$\alpha_0(i) = \pi_i, \qquad i = 1, 2, \ldots, N$$

  2. Induction

$$\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji}\, b_i(x_t)$$

$$\text{for } t = 1, 2, \ldots, T, \quad i = 1, 2, \ldots, N$$

  3. Termination

$$P(x|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

- **Backward probabilities**

  $\beta_t(i)$: probability of partial observation sequence $x_{t+1}, x_{t+2}, \ldots, x_T$, given $s_t = i$ (i.e. at i^{th} state at time t).

  $\beta_t(i) = P(x_{t+1}, x_{t+2}, \ldots, x_T \mid s_t = i, \lambda)$

- **Backward recursion**

  1. Assumption of sure termination

     $\beta_T(i) = 1, \qquad i = 1, 2, ..., N$

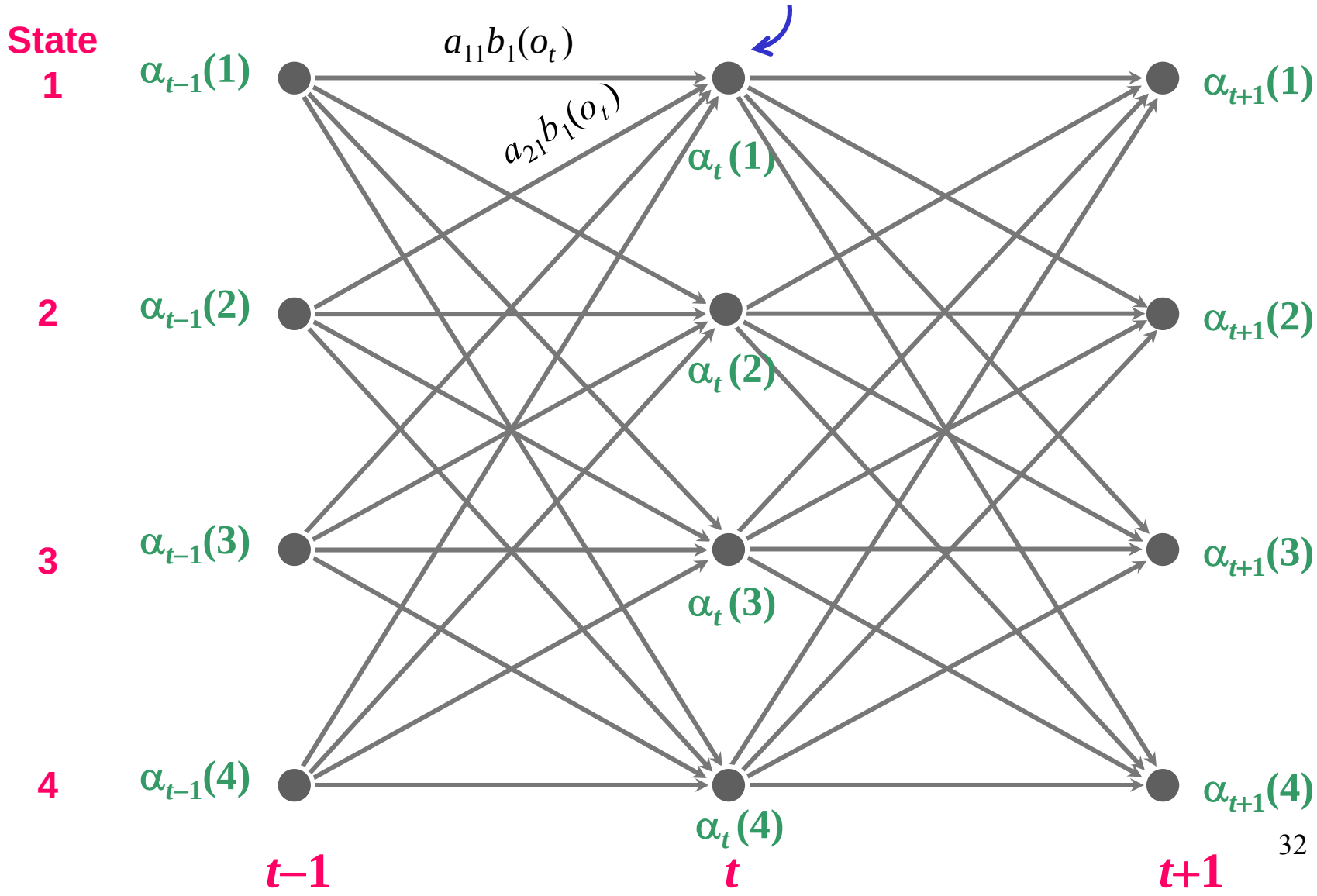
  2. Backward induction

     $$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j)\, a_{ji}\, b_i(x_{t+1})$$

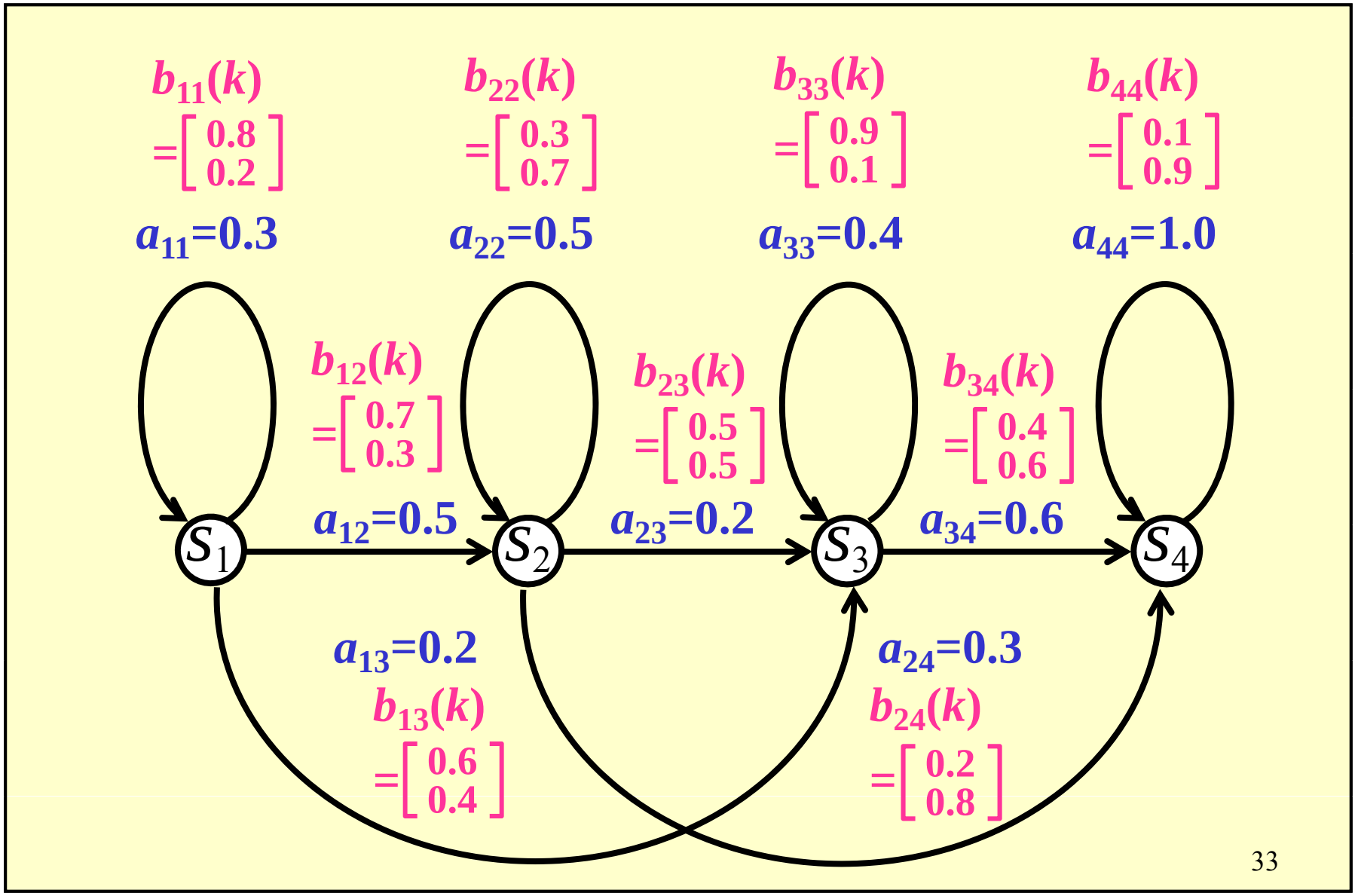
     for $t = T-1, T-2, \ldots, 1$, and $i = 1, 2, \ldots, N$

Forward probability induction

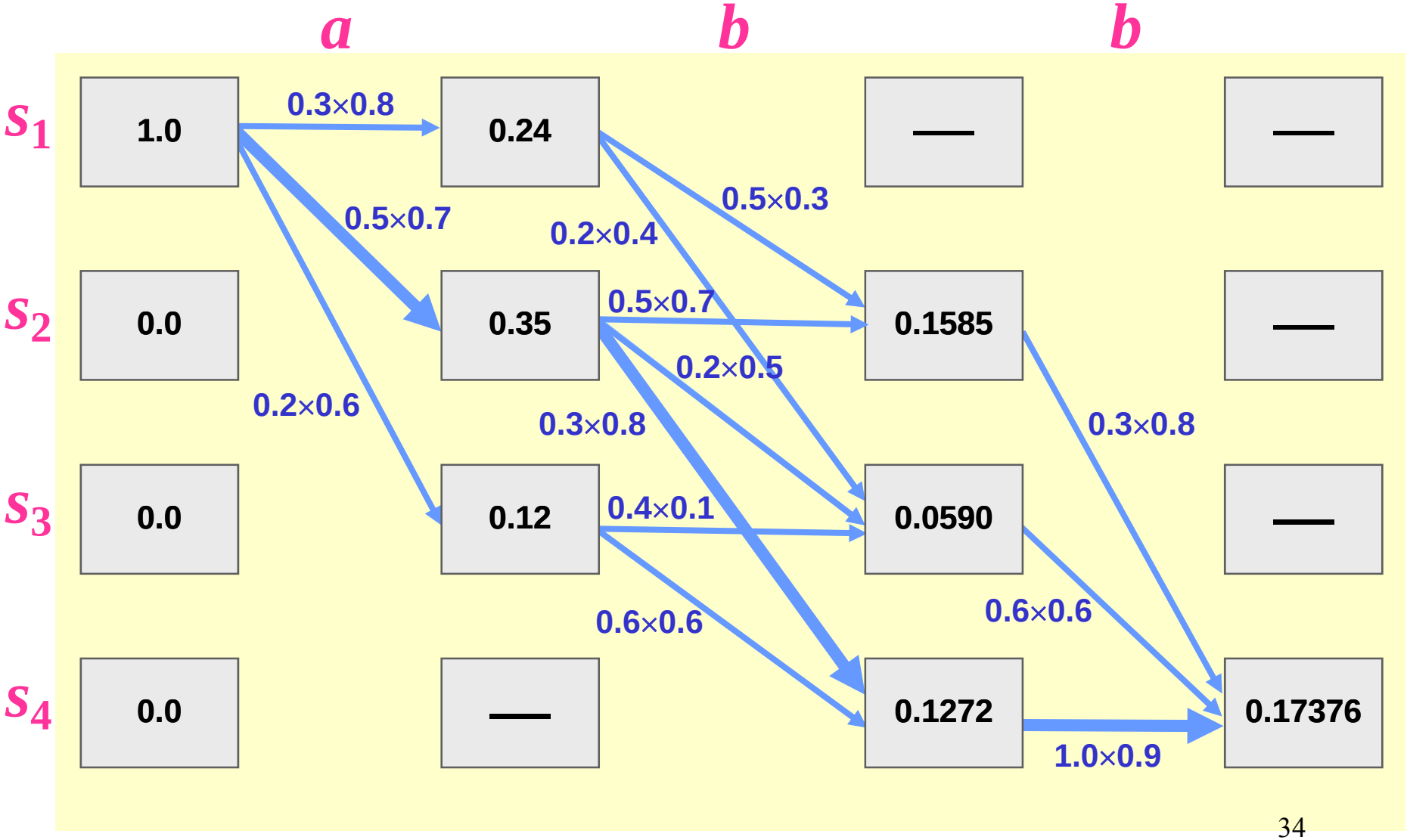
$$\alpha_t(1) = \alpha_{t-1}(1)\, a_{11} b_1(o_t) + \alpha_{t-1}(2)\, a_{21} b_1(o_t) + \alpha_{t-1}(3)\, a_{31} b_1(o_t) + \alpha_{t-1}(4)\, a_{41} b_1(o_t)$$



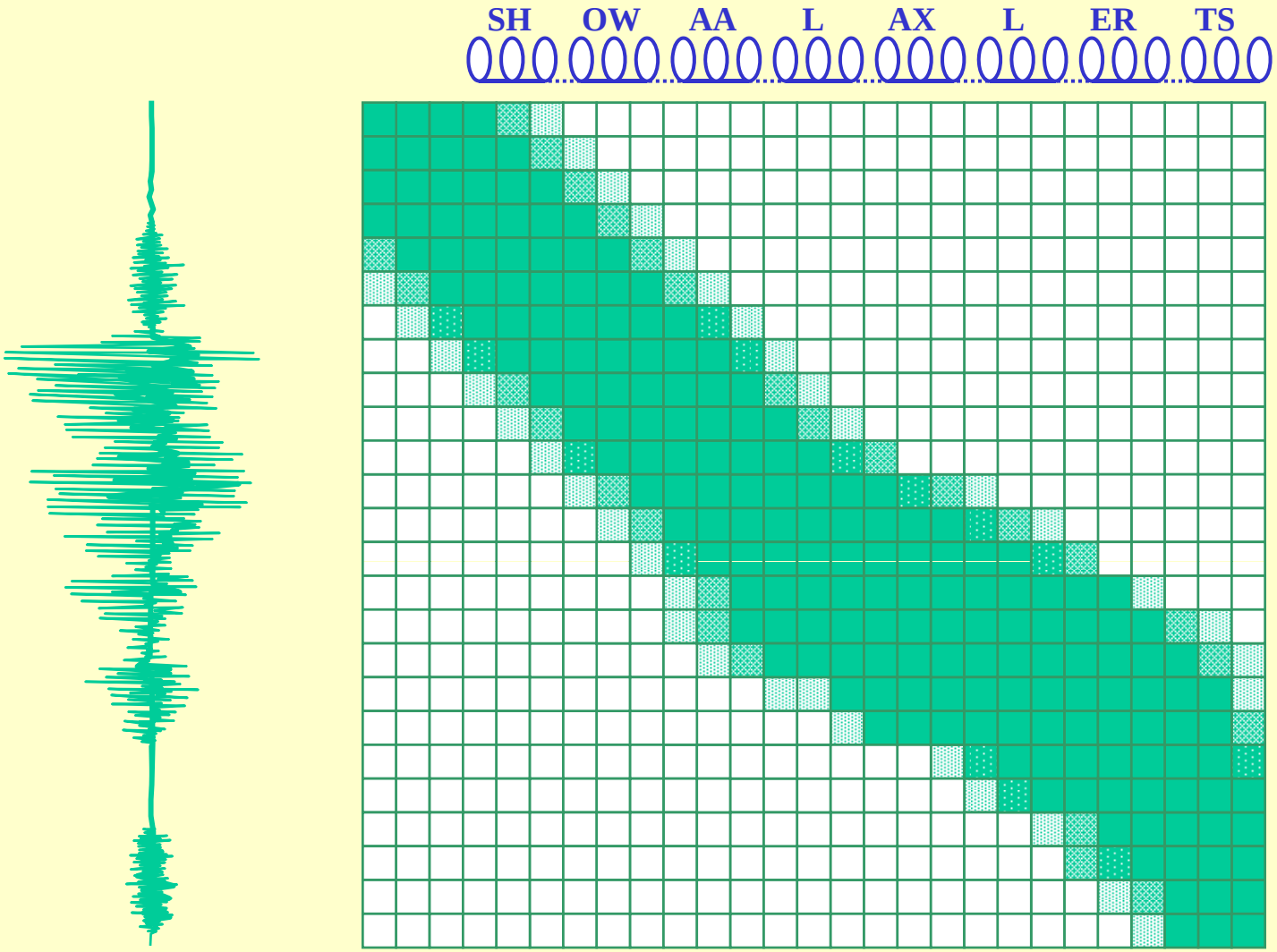
An example of HMM (Numbers in [ ] indicate the observation probabilities of symbols *a* and *b*



Calculation of P (abb) on the trellis for the previous example of HMM



The alignment of states vs. speech using the forward-backward algorithm. Darker squares correspond to large α and represent more plausible alignments.



Viterbi algorithm

1. Initialization

$$\delta_0(i) = \pi_i, \qquad i = 1, 2, ..., N$$

2. Recursion

$$\delta_t(i) = \max_{1 \le j \le N} [\delta_{t-1}(j)\, a_{ji}]\, b_i(x_t)$$

$$z_t(i) = \operatorname*{argmax}_{1 \le j \le N} [\delta_{t-1}(j)\, a_{ji}]$$

$$t = 1, 2, \ldots, T, \quad i = 1, 2, \ldots, N$$

3. Termination

$$P^* = \max_{1 \le j \le N} \delta_T(j)$$

$$s^*_T = \operatorname*{argmax}_{1 \le j \le N} \delta_T(j)$$

4. Backtracking

$$s^*_t = z_{t+1}(s^*_{t+1}), \quad t = T-1, T-2, \ldots, 0$$

Hidden Markov model parameter estimation
--- Training problem ---

Given a set of observation sequences, $\{x^{(i)}\}_{i=1}^{L}$,
how do we obtain a model $\lambda=(\pi, A, B)$ that best,
in the sense of some prescribed criteria, characterizes
the realization of the observation sequence set ?

- measure of goodness of fit
- solution mechanism
- characteristics and properties of the solution

HMM parameter estimation

Issues involved

- Estimation criterion
  - --- maximum likelihood
  - --- maximum joint state likelihood
  - --- maximum mutual information
  - --- maximum discrimination information
- Estimation/optimization procedure
- Convergence of the estimation procedure

Maximum likelihood

$\lambda=(\pi, A, B)$

$\bar{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, P(x|\lambda)$

→ Baum-Welch algorithm
Forward-backward algorithm
EM algorithm

Remarks:

1. No guarantee to reach globally optimal solution.
2. Only yield *fixed point* or locally optimal solutions.
3. Gradient techniques also work well although not as neat or convenient.

Iteration of the following 2 steps:

1. Determine the expectations:

$$Q(\lambda, \lambda') = E_{\lambda}\{\log P(x|\lambda')\}$$

2. Choose $\tilde{\lambda}$ which maximizes

$$E_{\lambda}\{\log P(x|\lambda')\}$$

as a function of λ'.

Note: $P(\tilde{x}|\lambda) \geq P(x|\lambda)$

As a result,

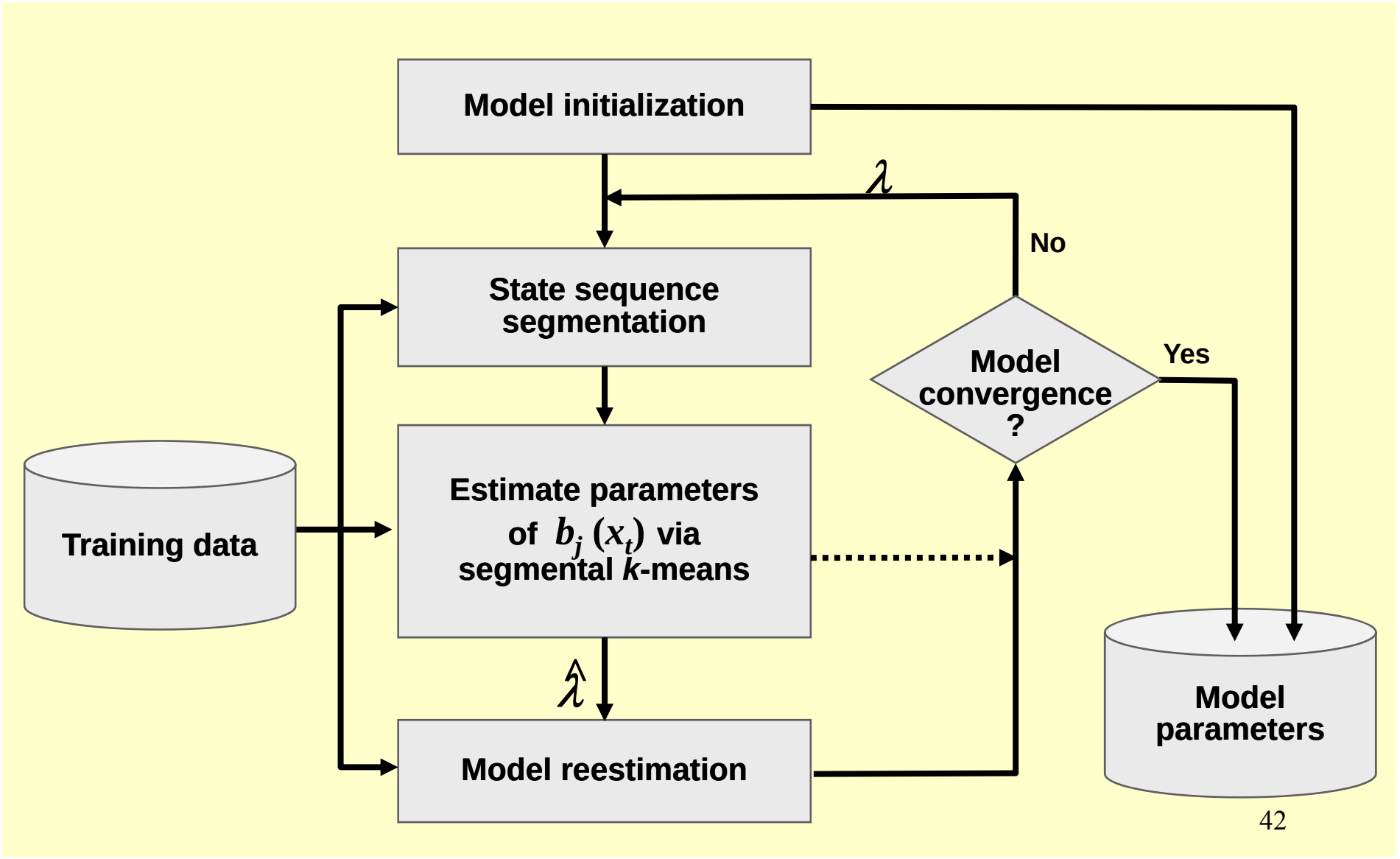
$$\tilde{\pi}_i = \frac{P(x, s_0 = i \mid \lambda)}{\sum_{j=1}^{N} P(x, s_0 = j \mid \lambda)}$$

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T} P(x, s_{t-1} = i, s_t = j \mid \lambda)}{\sum_{k=1}^{N} \sum_{t=1}^{T} P(x, s_{t-1} = i, s_t = k \mid \lambda)}$$

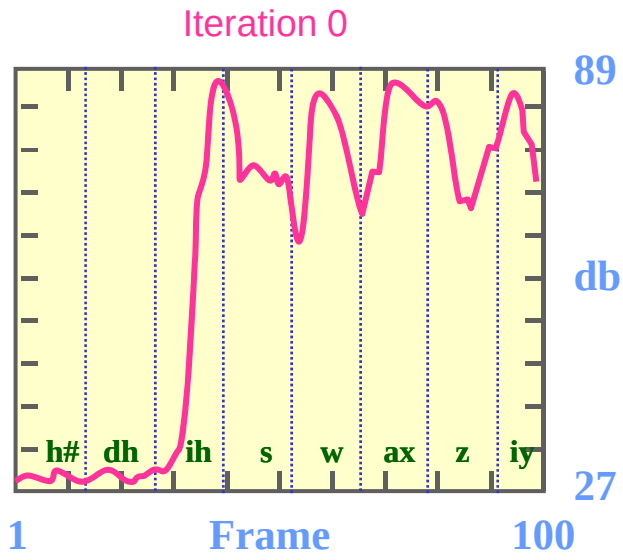
$$\tilde{b}_{ij} = \frac{\sum_{t=1}^{T} P(x, s_t = i, x_t = j \mid \lambda)}{\sum_{k=1}^{N} \sum_{t=1}^{T} P(x, s_t = i, x_t = k \mid \lambda)}$$

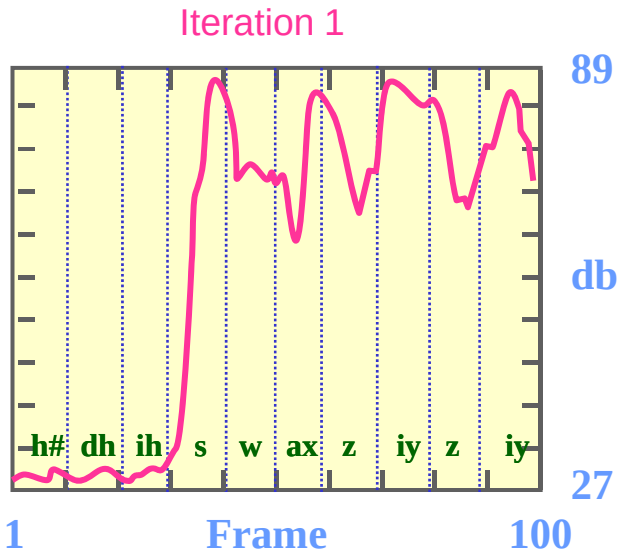
- Resembling conditional frequency of occurrences.
- Satisfying stochastic constraints.

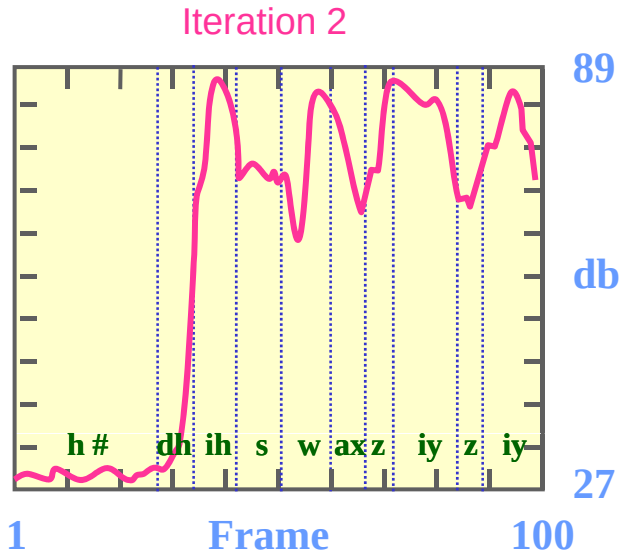
The segmental *k*-means training procedure used to estimate parameter values for the optimal continuous mixture density fit to a finite number of observation sequences

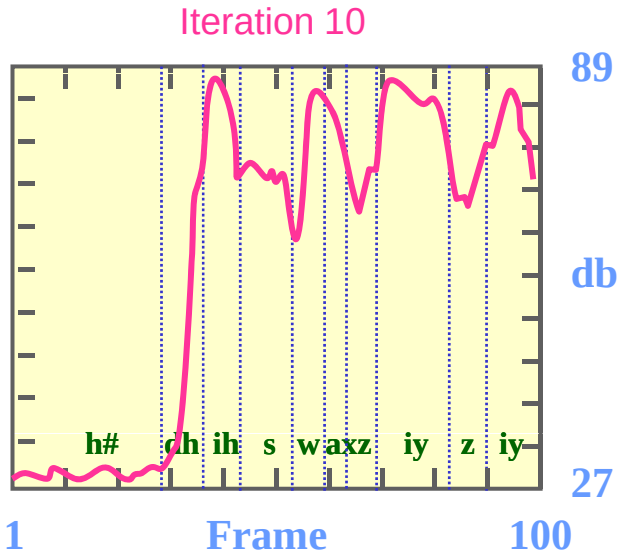


Segmental k-means training



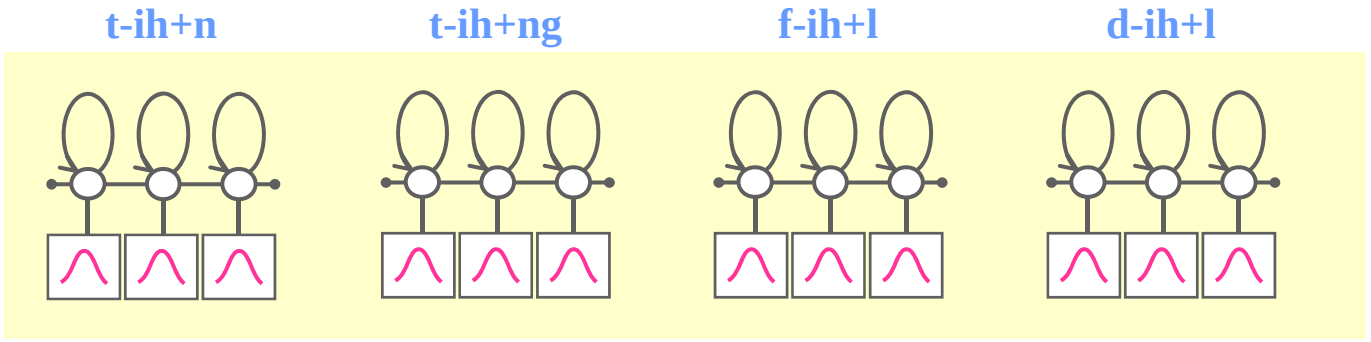




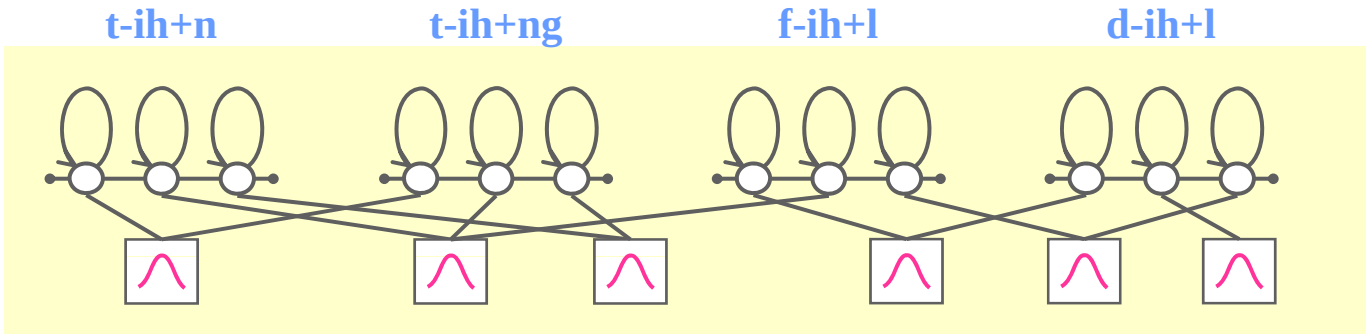


Tied-state triphone construction

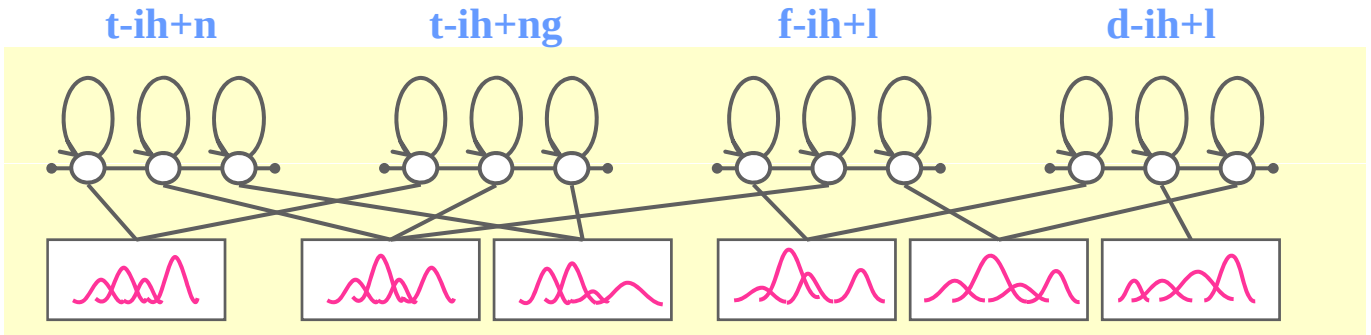
Single Gaussian triphones



State clustering



State clustered mixture Gaussian triphones



A decision tree for classifying the second state of *K*-triphone HMMs