HMM-Based Speech Synthesis

David Kappel

Advanced Signal Processing Seminar SS 2010

June 25, 2010

Outline

Outline

Introduction

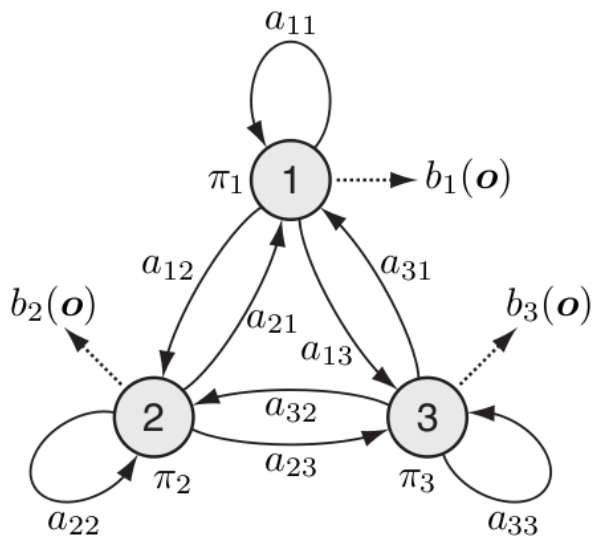
- Speech signal very quickly within milliseconds.
- large fluctuations even within a single speaker.
- humans are able to recover the underlying textual information.

- Speech signal assumed to be produced under some intention.
- Not observed directly.
- Has to be recovered from the speech signal.
- The hidden Markov model transfers this idea into a statistical model

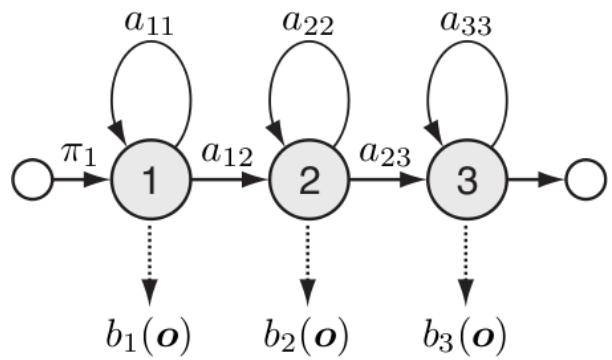
Outline

The Hidden Markov Model

- Statistical time series model.
- Describes a signal $\mathbf{O} = \{\mathbf{o}_1 \dots \mathbf{o}_T\}$.
- Underlying process: probabilistic state machine with N states



(a) Ergodic model



(b) Left-to-right model

State Sequence

- Sequence of latent state variables $\mathbf{q} = \{q_1 \dots q_T\}$.
- Temporal dependence between time instances captured in state variables.
- If state variables given observations become independent.
- $\rightarrow$ Markov property.

Parameter

HMM characterised by its set of parameters $\theta = \{\mathbf{A}, \mathbf{B}, \mathbf{\Pi}\}$

- $\mathbf{A}$: *transition probabilities* $a_{ij} = \mathrm{p}(q_t = i \mid q_{t-1} = j)$.
- $\mathbf{\Pi}$: *prior probabilities* $\pi_i = \mathrm{p}(q_1 = i)$ and $\mathbf{B}$.
- $\mathbf{B}$: table of *observation probabilities*.

Observation Probabilities

Mixture of multivariate Gaussian distributions

$$b_i(\mathbf{o}) = \sum_{m=1}^{M} w_{im} \mathcal{N}(\mathbf{o}; \mu_{im}, \mathbf{\Sigma}_{im}),$$

- $\mathcal{N}$: Multivariate Gaussian distribution.
- μ_{im}: Mean.
- $\mathbf{\Sigma}_{im}$: Covariance matrix.
- w_{im}: Mixture weights.

Observation Probabilities

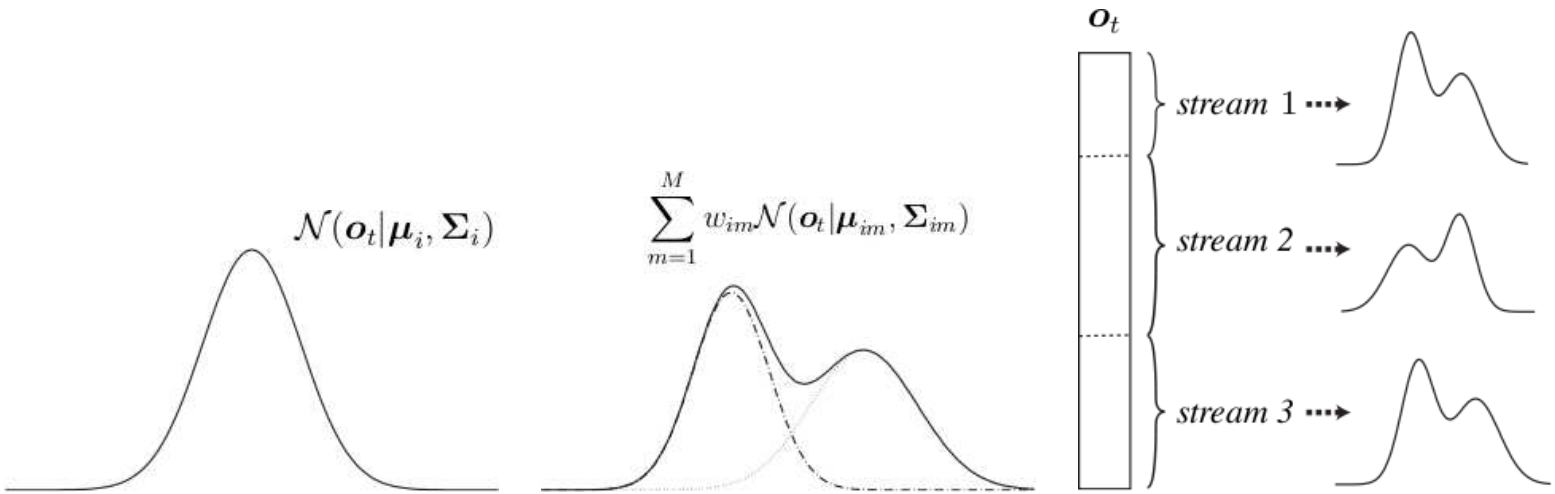
Observation sequence consists of S stochastic-independent data streams

$$\mathbf{o} = \{\mathbf{o}_1^\top \dots \mathbf{o}_T^\top\}$$

observation probabilities formulated by *products* of Gaussian mixture densities

$$b_i(\mathbf{o}) = \prod_{s=1}^{S} \left(\sum_{m=1}^{M_s} w_{ism} \mathcal{N}(\mathbf{o}_s; \mu_{ism}, \mathbf{\Sigma}_{ism}) \right) .$$

Observation Probabilities



(a) Gaussian PDF (b) Multi Mixture PDF (c) Multi Stream PDF

3 basic problems

- Calculate probability of observation sequence $\mathbf{O}$ for given parameter $p(\mathbf{O}|\theta)$.

  $\rightarrow$ *forward-backward* algorithm

- Given observation sequence $\mathbf{O}$ find most probable path.

  $\rightarrow$ *Viterby* algorithm

- Adjust model parameters θ to given observation sequence $\mathbf{O}$.

  $\rightarrow$ *Baum-Welch* algorithm

The Baum-Welch algorithm

Ccalculate optimal model parameter $\rightarrow$ maximise likelihood function

$$\mathcal{L}(\lambda|\mathbf{O}) = p(\mathbf{O}|\lambda) = \sum_{\mathbf{q}} p(\mathbf{O}, \mathbf{q}|\lambda).$$

- no closed form solution to solve this problem.
- *EM* algorithm $\rightarrow$ *local* maximum.
- incremental update rule:
  Expectation (E) - and *Maximisation* (M) steps.

Q-Function

Auxiliary function Q of old and new parameter set:

$$Q(\lambda^{old}, \lambda^{new}) = \sum_{\mathbf{q}} \mathrm{p}(\mathbf{q} \mid \mathbf{O}, \lambda^{old}) \log \mathrm{p}(\mathbf{q}, \mathbf{O} \mid \lambda^{new}).$$

- E-step: Evaluated the Q-function based on λ^{old}.
- M-step: Find λ^{new} that maximises Q and replace λ^{old} by λ^{new}.

Some Notation

Probability of making transition $j \to k$ at time t:

$$\xi_t(i,j) = \mathrm{p}(q_t = i, q_{t-1} = j \mid \mathbf{O}, \lambda)$$

Probability of being in state i:

$$\gamma_t(i) = \mathrm{p}(q_t = i \mid \mathbf{O}, \lambda)$$

M-step

Set of parameters that maximises Q-function:

$$\begin{aligned} \pi_i^{new} &= \gamma_1(i), \\ a_{ij}^{new} &= \frac{\sum_{\tau=1}^{T} \xi_t(i,j)}{\sum_{\tau=1}^{T} \gamma_t(i)}. \end{aligned} \tag{1}$$

For the case of a Gaussian observation probability:

$$\begin{aligned} \mu_i^{new} &= \frac{\sum_{\tau=1}^{T} \gamma_t(i) \cdot \mathbf{o}_t}{\sum_{\tau=1}^{T} \gamma_t(i)}, \\ \mathbf{\Sigma}_i^{new} &= \frac{\sum_{\tau=1}^{T} \gamma_t(i) \cdot (\mathbf{o}_t - \mu_i)(\mathbf{o}_t - \mu_i)^\top}{\sum_{\tau=1}^{T} \gamma_t(i)}, \end{aligned} \tag{2}$$

The forward-backward algorithm

Find an efficient evaluation of the M-step:

$$
\begin{aligned}
\xi_t(i,j) &= \mathrm{p}(q_t = i, q_{t-1} = j \mid \mathbf{O}) \\
&= \frac{\mathrm{p}(\mathbf{O} \mid q_t = i, q_{t-1} = j)\mathrm{p}(q_t = i, q_{t-1} = j)}{\mathrm{p}(\mathbf{O})} \\
&= \frac{\mathrm{p}(\mathbf{o}_1 \dots \mathbf{o}_{t-1}, q_{t-1} = j) a_{ij} b_i(\mathbf{o}_t) \mathrm{p}(\mathbf{o}_{t+1} \dots \mathbf{o}_T \mid z_t = k)}{\mathrm{p}(\mathbf{O})}
\end{aligned}
$$

Two recursive equation

- one running forward:

$$\alpha_t(i) = \mathrm{p}(\mathbf{o}_1 \dots \mathbf{o}_t, q_t = i) = \sum_{j=1}^{N} a_{ij} \cdot b_i(\mathbf{o}_t) \cdot \alpha_{t-1}(j)$$

- one running backward

$$\beta_t(i) = \mathrm{p}(\mathbf{o}_{t+1} \dots \mathbf{o}_T \mid q_t = i) = \sum_{j=1}^{N} a_{ji} \cdot b_j(\mathbf{o}_{t+1}) \cdot \beta_{t+1}(j).$$

Initialisation:

$$\alpha_1(i) = \pi_i b_i(\mathbf{o}_1), \qquad \beta_T(i) = 1.$$

Efficient M-step

$$\xi_t(i,j) = \frac{\alpha_{t-1}(j)a_{ij}b_i(\mathbf{o}_t)\beta_t(i)}{\mathrm{p}(\mathbf{O})}$$

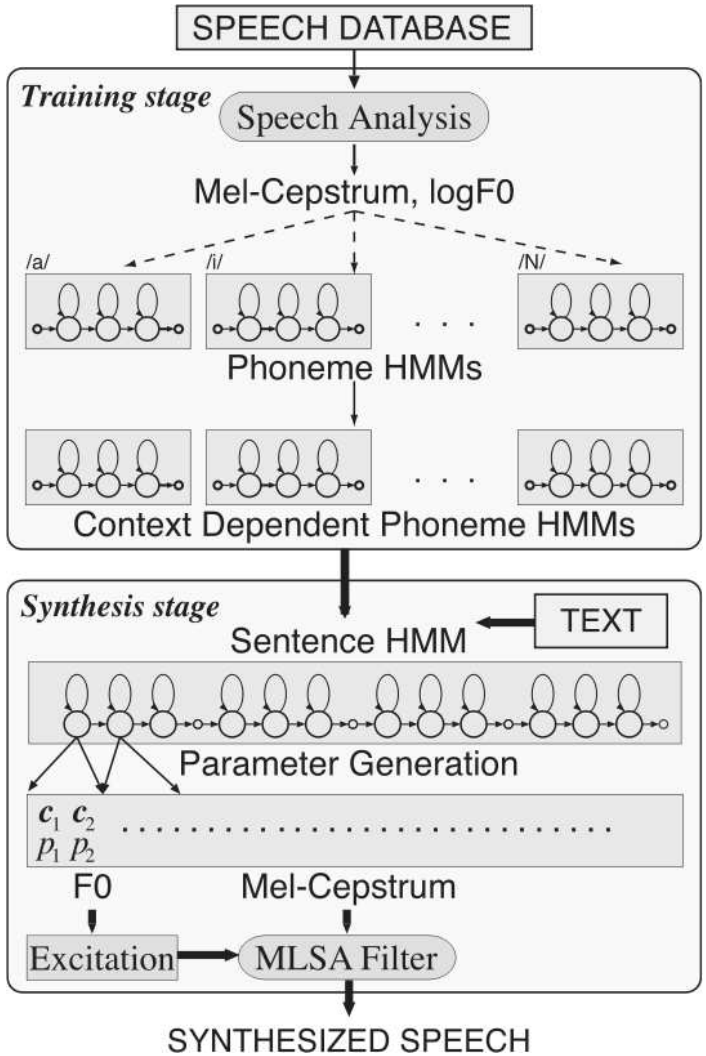
$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathrm{p}(\mathbf{O})}$$

- Each message needs to be computed once $\leftarrow$ efficient.
- Computing $\mathrm{p}(\mathbf{O} \mid \lambda)$:

$$\mathrm{p}(\mathbf{O} \mid \theta) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

Outline

A HMM-based TTS System



Speech Parameter Generation

- Again optimisation problem:

$$\mathbf{O}^* = \arg\max_{\mathbf{O}} \mathrm{p}(\mathbf{O} \mid \lambda, T).$$

- Again no closed form solution.

Approximation: Split expression up into 2 separate optimisation problems.

$$\begin{aligned}
\mathbf{O}^* &= \operatorname*{argmax}_{\mathbf{O}} \mathrm{p}(\mathbf{O} \mid \lambda, T) \\
&= \operatorname*{argmax}_{\mathbf{O}} \sum_{\mathbf{q}} \mathrm{p}(\mathbf{O}, \mathbf{q} \mid \lambda, T) \\
&\approx \operatorname*{argmax}_{\mathbf{O}} \max_{\mathbf{q}} \mathrm{p}(\mathbf{O}, \mathbf{q} \mid \lambda, T),
\end{aligned}$$

Optimising O

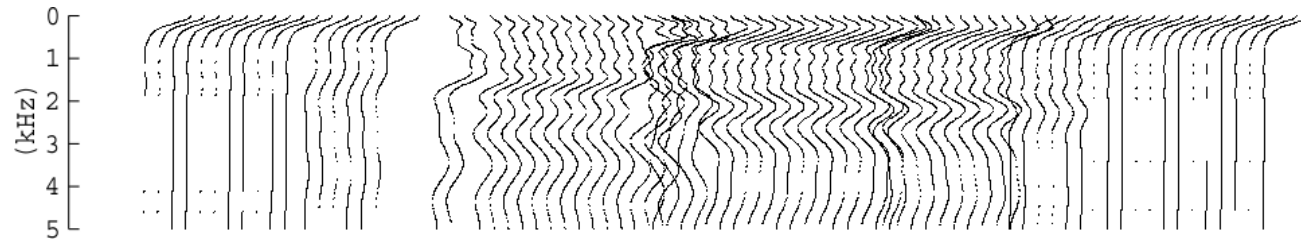
- Assume optimal state vector $\mathbf{q}^*$ given.
- Drawing speech parameter independently causes discontinuities $\rightarrow$ clicks
- *dynamic features*: $\mathbf{o}_t = [\mathbf{c}_t, \Delta\mathbf{c}_t, \Delta^2\mathbf{c}_t]$.

$\mathbf{c}_t$: cepstarl coefficients

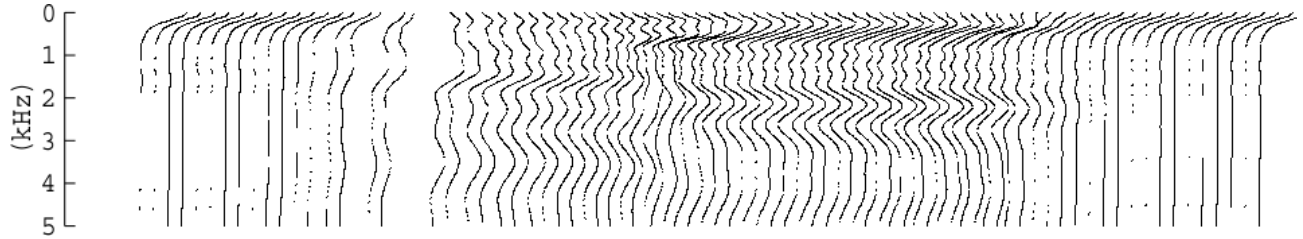
$$\Delta\mathbf{c}_t = \frac{1}{2}\left(\mathbf{c}_{t+1} - \mathbf{c}_{t-1}\right)$$

$$\begin{aligned}\Delta^2\mathbf{c}_t &= \frac{1}{2}\left(\Delta\mathbf{c}_{t+1} - \Delta\mathbf{c}_{t-1}\right) \\ &= \frac{1}{4}\left(\mathbf{c}_{t+2} + 2\mathbf{c}_t - \mathbf{c}_{t-2}\right)\end{aligned}$$

Dynamic vs. Static Features



(d) Without dynamic features



(e) With dynamic features

Optimising $\mathbf{q}$

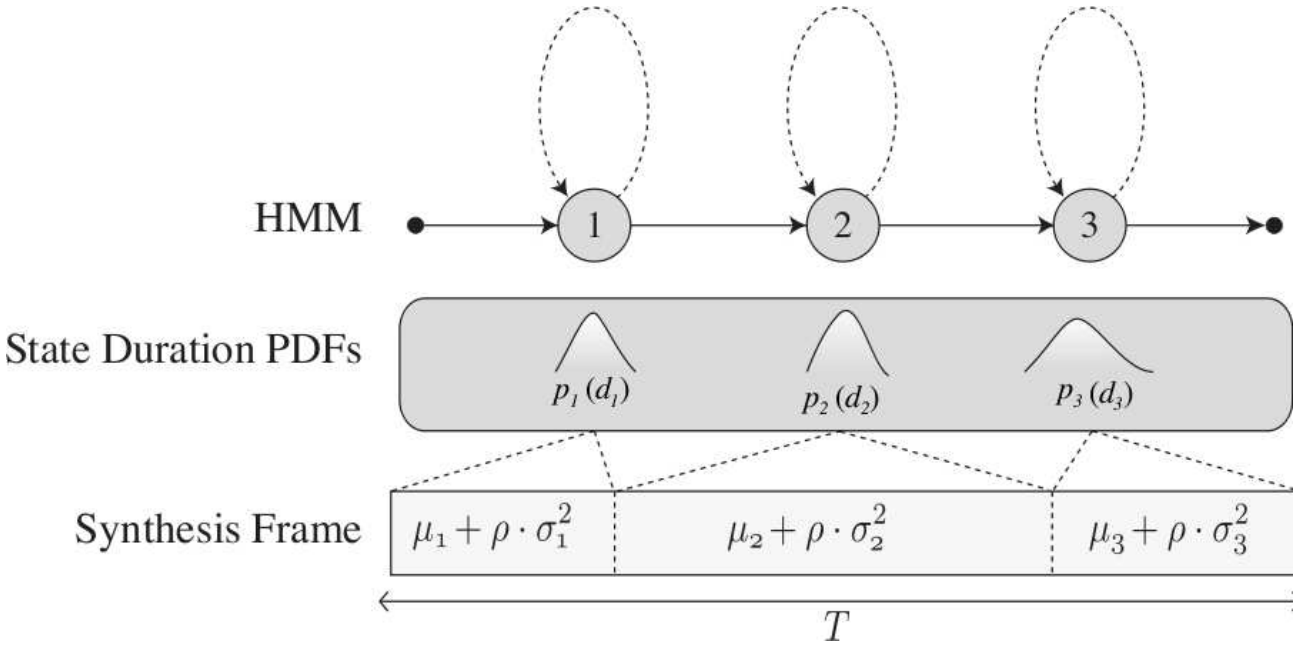
$$\mathbf{q}^* = \underset{\mathbf{q}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{q} \mid \lambda, T)$$

Product of all the state transitions along the path:

$$\mathrm{p}(\mathbf{q} \mid \lambda, T) = \prod_{t=1}^{T} a_{q_t, q_{t-1}}$$

- Calculation for all possible paths impracticable for full HMM $\rightarrow$ Restricting the number of paths.
- Assumed left-to-right HMM without skips.
- Separate sate duration distribution.

Duration Synthesis

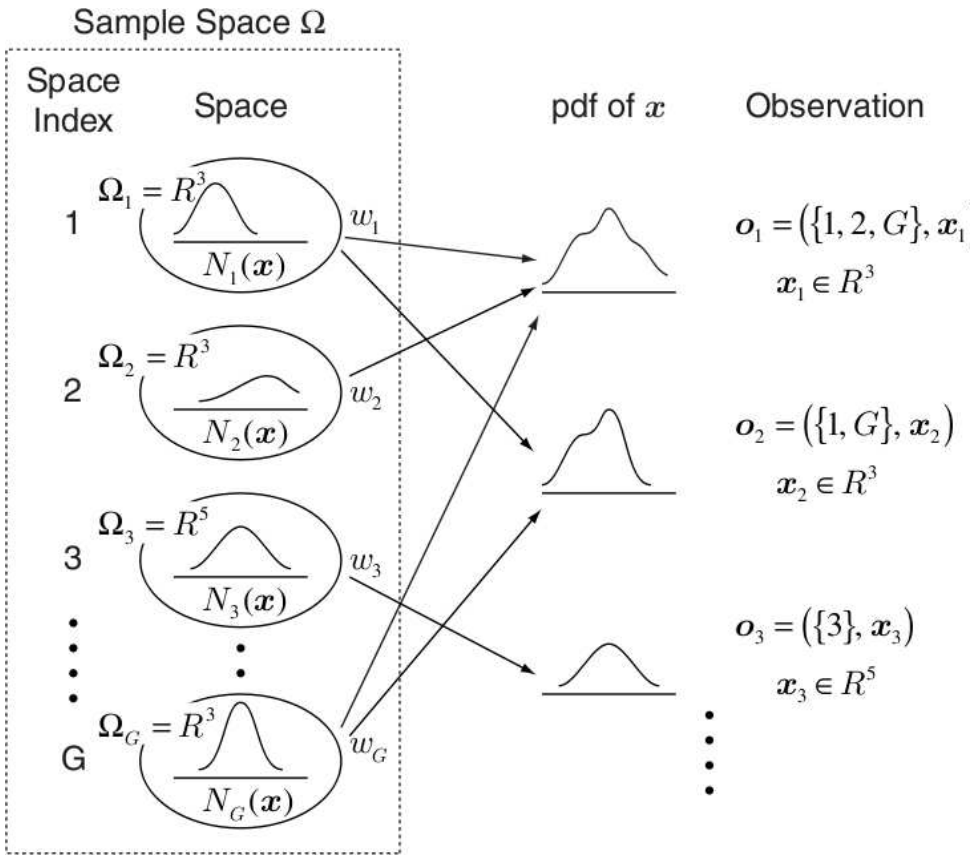


MSD-HMM

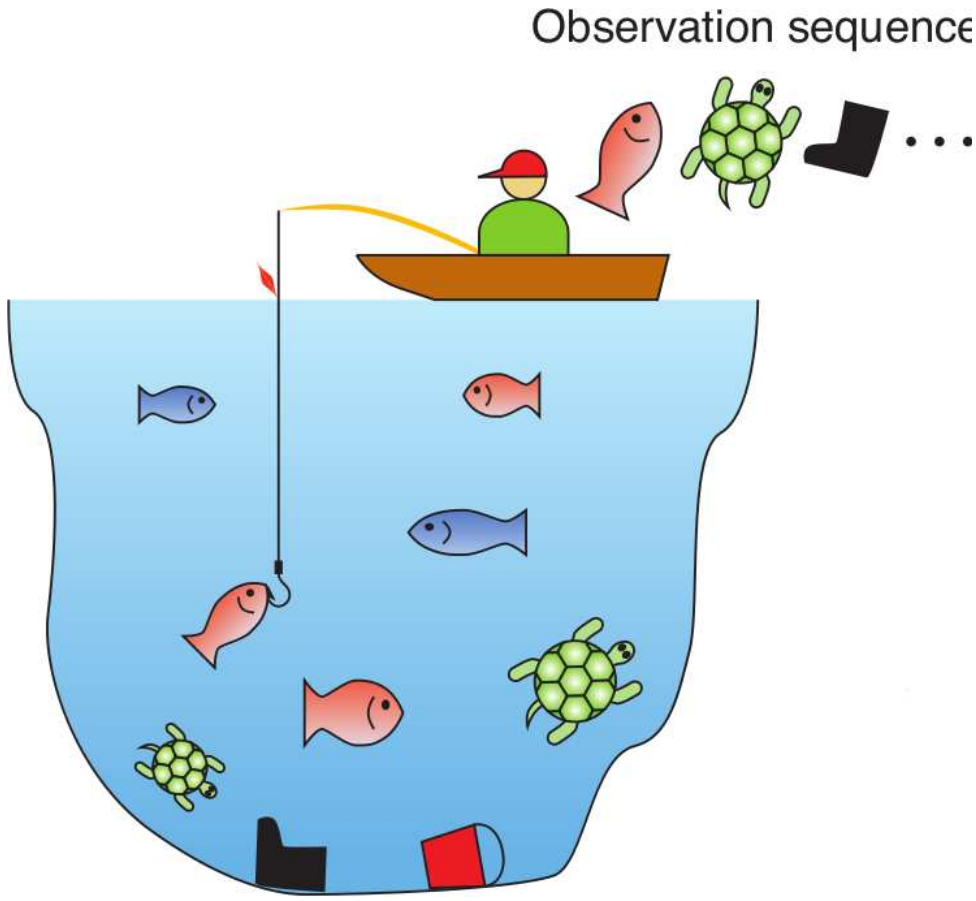
- extension of the HMM
- Observation probability: Multi-Space Probability Distribution (MSD).
- Multiple probability spaces Ω_g with n_g-dimensional probability distribution $\mathcal{N}_g(x)$.
- Observation consists of 2 random variables: $\mathbf{o} = (X, \mathbf{x})$

$$b_i(\mathbf{o}) = \sum_{g \in X} w_g \mathcal{N}_g(x).$$

MSD


Sample Space Ω
Space Index
Space
pdf of x
Observation
1
2
3
G
$\Omega_1 = R^3$
$\Omega_2 = R^3$
$\Omega_3 = R^5$
$\Omega_G = R^3$
$N_1(x)$
$N_2(x)$
$N_3(x)$
$N_G(x)$
w_1
w_2
w_3
w_G
$o_1 = (\{1, 2, G\}, x_1$
$x_1 \in R^3$
$o_2 = (\{1, G\}, x_2)$
$x_2 \in R^3$
$o_3 = (\{3\}, x_3)$
$x_3 \in R^5$

Like Fishing...



$F0$ Modelling

- Space with dimensionality 0 possible.
- Contain only one element with probability 1.
- Used for synthesis of $F0$ frequencies.
  - Spaces of dimensionality 1: frequencies
  - Spaces of dimensionality 0: unvoiced parts.

Outline

Questions?