Reformulating the HMM as a Trajectory Model

Keiich Tokuda   Heiga Zen   Tadashi Kitamura   (Nagoya Institute of Technology, Japan)

Background

Typical LVCSR framework

Feature vector: MFCC + ΔMFCC

Acoustic model: context-dependent HMM

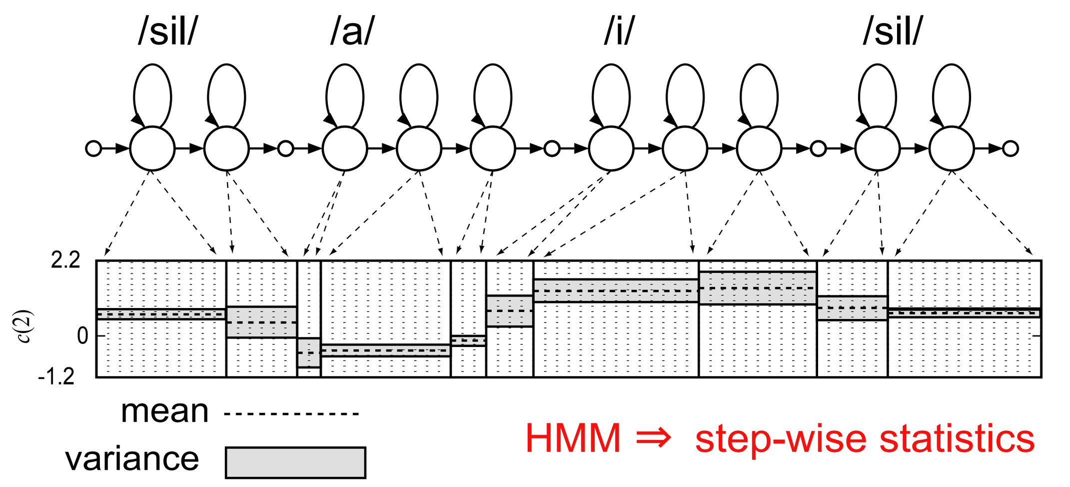
Language model: word N-gram

Limitations of the HMM for modeling speech

(1) Constant statistics within an state

(2) Conditional independence assumption

(3) Weak duration modeling



HMM ⇒ step-wise statistics

Alternative approach

- Alternative models
  ex.) HDM, SSM, SLDS, DBN, HSMM, etc.
- Dynamic feature
  (dynamic and accelaration coefficients)

⇒ Improper in the sense of statistical modeling

Trajectory-HMM

- Impose explicit relationship between static and dynamic features on the HMM
  ⇒ Reformulate the HMM as a trajectory model
- Overcome some limitations of the HMM
- Derive a training algorithm
- Improvements both in speech recognition and synthesis

Objective

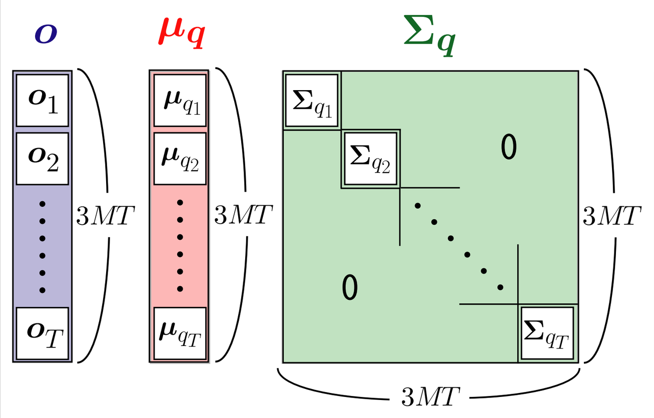
- Summarize the definition and algorithms
- Sampling algorithm for the model
  (Omitted from this poster m(._.)m)
- Relation to other statistical models
- Consistency between training and recognition

Reformulating HMM as trajectory-HMM

Output probability of an observation $\boldsymbol{o}$ for HMM λ

$$P(\boldsymbol{o} \mid \lambda) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda)$$

$$P(\boldsymbol{o} | \boldsymbol{q}, \lambda) = \mathcal{N}(\boldsymbol{o} | \boldsymbol{\mu_q}, \boldsymbol{\Sigma_q}) \text{ (single Gaussian, } 3MT)$$



$\boldsymbol{o}_t$: Observation ($3M \times 1$)

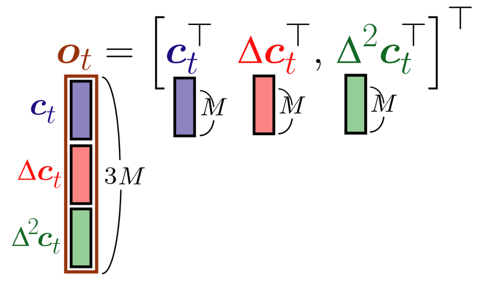
$\boldsymbol{q}$: State sequence

$\boldsymbol{\mu}_{q_t}$: Mean of state q_t

$\boldsymbol{\Sigma}_{q_t}$: Covariance of state q_t

Observation vector ⇒ static and dynami features

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top\right]^\top$$



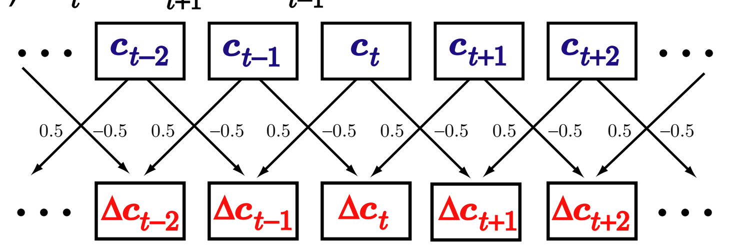
$\boldsymbol{c}_t$: static feature ($M \times 1$)

$\Delta \boldsymbol{c}_t$: 1st-order dynamic feature

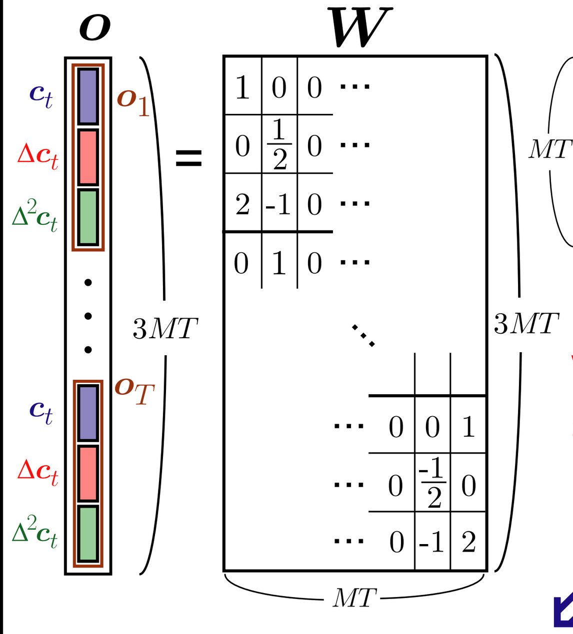
$\Delta^2 \boldsymbol{c}_t$: 2nd-order dynamic feature

Dynamic feature ⇒ computed from static features

Ex.) $\Delta \boldsymbol{c}_t = 0.5\boldsymbol{c}_{t+1} - 0.5\boldsymbol{c}_{t-1}$



Relationship between $\boldsymbol{o}$ and $\boldsymbol{c}$ in a matrix form

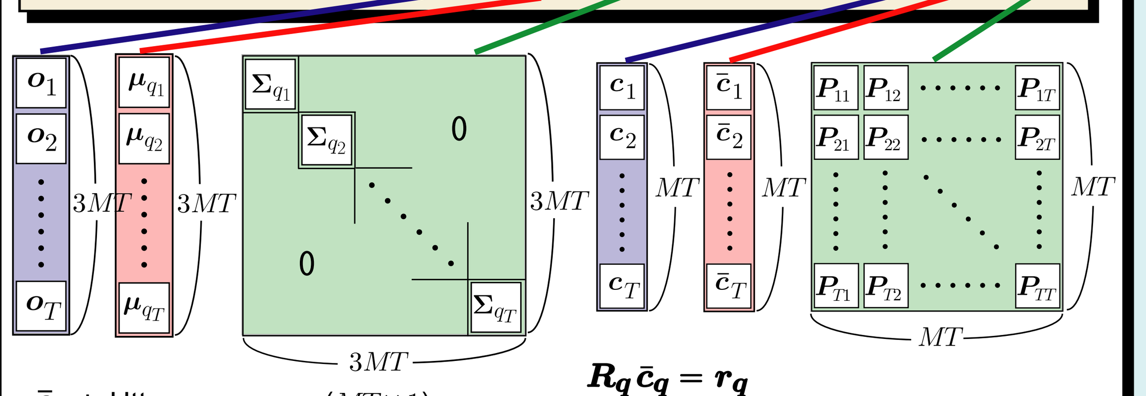


$\boldsymbol{W}$: Window matrix ($3MT \times MT$)

Under this relationship, we should regard only $\boldsymbol{c}$ as the random variable

⇒ Rewrite HMM likelihood as a function of $\boldsymbol{c}$

$$P(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda) = \mathcal{N}(\boldsymbol{Wc} \mid \boldsymbol{\mu_q}, \boldsymbol{\Sigma_q}) = K_{\boldsymbol{q}} \cdot \mathcal{N}(\boldsymbol{c} \mid \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P_q})$$



$\bar{\boldsymbol{c}}_{\boldsymbol{q}}$: Utterance mean ($MT \times 1$)

$\boldsymbol{P_q}$: Utterance covariance ($MT \times MT$)

$K_{\boldsymbol{q}}$: Constant term independent of $\boldsymbol{c}$

$$\boldsymbol{R_q}\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{r_q}$$

$$\boldsymbol{R_q} = \boldsymbol{W}^\top \boldsymbol{\Sigma_q}^{-1} \boldsymbol{W} = \boldsymbol{P_q}^{-1}$$

$$\boldsymbol{r_q} = \boldsymbol{W}^\top \boldsymbol{\Sigma_q}^{-1} \boldsymbol{\mu_q}$$

Definition of trajectory-HMM

$P(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda)$ shoud be normalized by $K_{\boldsymbol{q}}$:

$$P(\boldsymbol{c} \mid \lambda) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda)$$

where

$$P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = K_{\boldsymbol{q}}^{-1} \cdot P(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda) = K_{\boldsymbol{q}}^{-1} \cdot \mathcal{N}(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda) = \mathcal{N}(\boldsymbol{c} \mid \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P_q})$$

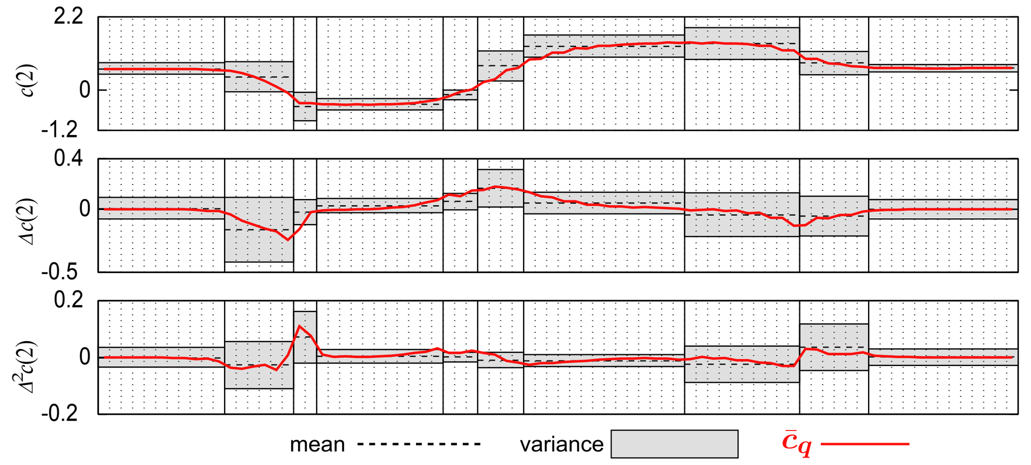
⇒ referred as "trajectory-HMM"

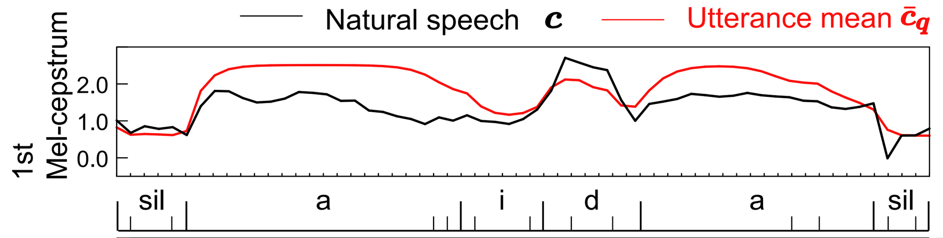
The mean vector $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ is given as a trajectory

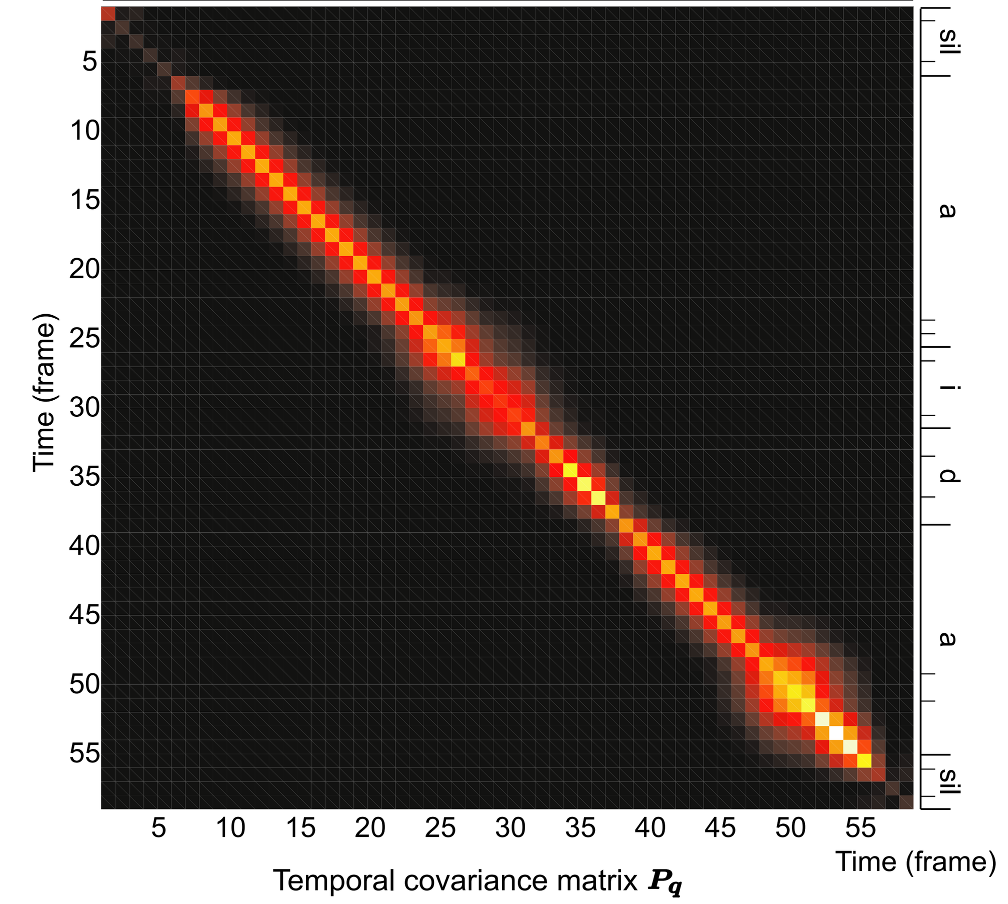
→ Variable statics in a state

The covariance matrix $\boldsymbol{P_q}$ is full

→ Dependency of state output probabilities







Temporal covariance matrix $\boldsymbol{P_q}$

Mean ⇒ varies in a state

Temporal correlation ⇒ modeled by $\boldsymbol{P_q}$

HMM-based speech synthesis

Speech parameter generation from HMM [Tokuda;'00]

⇒ Generate obs. vec. seq. maximizing its output prob.

Without dynamic feature constraints

$$\boldsymbol{o}_{\max} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda)$$

$\boldsymbol{o}$: Obs. vec. sequence

$\boldsymbol{q}$: Gaussian sequence

⇒ $\boldsymbol{o}_{\max}$ becomes a sequence of mean vectors

With dynamic feature constraints

$$\boldsymbol{c}_{\max} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda)$$

By setting $\frac{\partial}{\partial \boldsymbol{c}} \log P(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda) = 0$

$$\boldsymbol{R_q}\boldsymbol{c}_{\max} = \boldsymbol{r_q}$$

$$\boldsymbol{R_q} = \boldsymbol{W}^\top \boldsymbol{\Sigma_q}^{-1} \boldsymbol{W} = \boldsymbol{P_q}^{-1} \qquad \boldsymbol{\mu_q} = \left[\boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_T}^\top\right]^\top$$

$$\boldsymbol{r_q} = \boldsymbol{W}^\top \boldsymbol{\Sigma_q}^{-1} \boldsymbol{\mu_q} \qquad \boldsymbol{\Sigma_q} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{q_1}, \ldots, \boldsymbol{\Sigma}_{q_T}\right]$$

Training algorithm

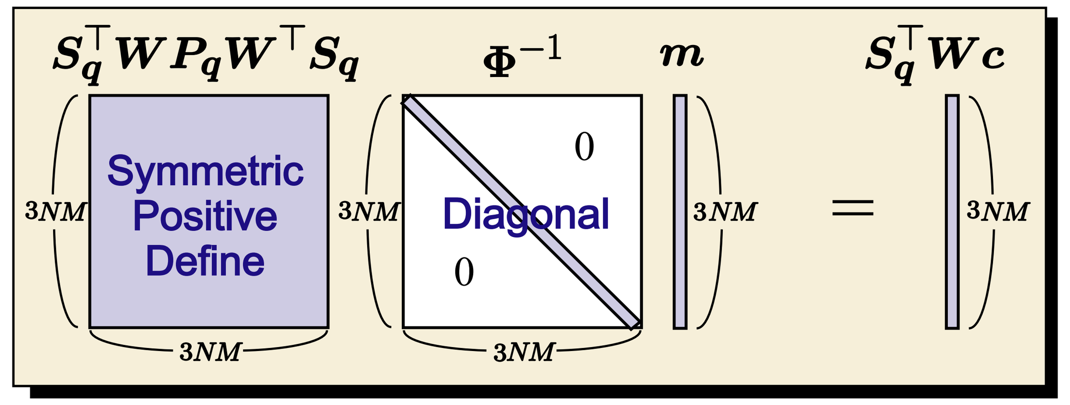
Vitebi-type training algorithm

- Determination of Gaussian sequence $\boldsymbol{q}$

⇒ A modified Viterbi algorithm based on recursive likelihood calculation and delayed-decision.

- Parameter update

By setting $\partial \log P(\boldsymbol{c}, \boldsymbol{q} \mid \lambda) \;/\; \partial \boldsymbol{m} = 0$,



$$\frac{\partial \log P(\boldsymbol{c}, \boldsymbol{q} \mid \lambda)}{\partial \boldsymbol{\phi}} = \frac{1}{2} \boldsymbol{S_q}^\top \mathrm{diag}^{-1}\Big(\boldsymbol{W P_q W}^\top - \boldsymbol{W c c}^\top \boldsymbol{W}^\top + 2\boldsymbol{\mu_q c}^\top \boldsymbol{W}^\top + \boldsymbol{W}\bar{\boldsymbol{c}}_{\boldsymbol{q}}\bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{W}^\top - 2\boldsymbol{\mu_q}\bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{W}^\top\Big)$$

⇒ Gradient method (steepest ascent)

where

$\boldsymbol{S_q}$: State sequence matrix $\begin{cases} \boldsymbol{\mu_q} = \boldsymbol{S_q m} \\ \boldsymbol{\Sigma_q}^{-1} = \mathrm{diag}(\boldsymbol{S_q \phi}) \end{cases}$

$\boldsymbol{m} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \ldots, \boldsymbol{\mu}_N^\top\right]^\top$: Embedded mean vectors

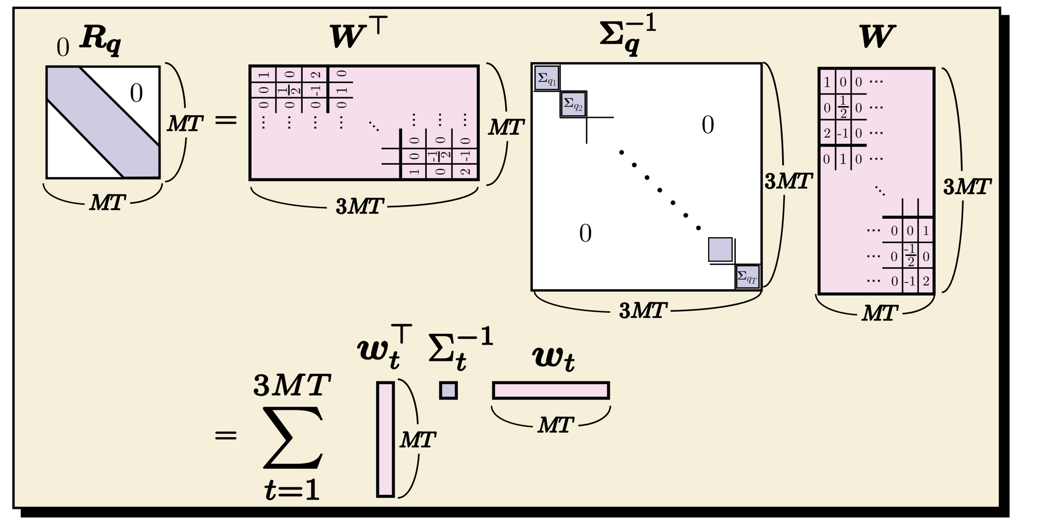
$\boldsymbol{\phi} = \left[\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \ldots, \boldsymbol{\Sigma}_N^{-1}\right]^\top$: Embedded covariance matrices

$\boldsymbol{\Phi}^{-1} = \mathrm{diag}(\boldsymbol{\phi})$ N : Number of state distributions

Relation to other statistical models

Relation to covariance modeling techniques

Temporal inverse covariance (precision) matrix $\boldsymbol{R_q}$



Weighted sum of rank-1 symmetric matrices

⇒ EMLLT model [Olsen et. al;'01] for temporal precision

Both mean and precision are constrained in a subspace

⇒ SPAM model [Axelrod et. al;'02]

Relation to Product of Experts (PoE)

Basis superposition (STC, EMLLT, SPAM, etc.) can be viewed as Product of Gaussians (PoG) [Sim et.al;'04]

Trajectory-HMM: EMLLT for temporal precision matrix

⇒ trajectory-HMM can also be viewed as PoG

PoG representation of the trajectory-HMM

$$P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = K_{\boldsymbol{q}}^{-1} \cdot P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) = K_{\boldsymbol{q}}^{-1} \cdot \mathcal{N}(\boldsymbol{o} \mid \boldsymbol{\mu_q}, \boldsymbol{\Sigma_q}) = \frac{1}{K_{\boldsymbol{q}}} \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{q_t}, \boldsymbol{\Sigma}_{q_t})$$

Consistency between training & decodnig

Database	ATR Japanese continuous speech database b-set speaker MHT
Training data	450 utterances
Test data	53 utterances

100-best rescoring

Training Decoding	HMM HMM	HMM trajectory	trajectory HMM	trajectory trajectory
PER (%) (rel. imp.)	19.7 ref.	19.5 1.0	72.1 -266.0	18.0 8.6

⇒ Importance of the consistency

(100-best + correct string) rescoring

Training Decoding	HMM HMM	HMM trajectory	trajectory HMM	trajectory trajectory
PER (%) (rel. imp.)	15.3 ref.			9.0 41.1

⇒ Necessity of a decoder specialized for the model