Continuous Stochastic Feature Mapping Based on Trajectory HMMs

Heiga Zen[1,2] Yoshihiko Nankaku[1] Keiichi Tokuda[1]

[1] Nagoya Institute of Technology, Japan

[2] Presently, Toshiba Cambridge Research Lab., UK

2009/03/30@Univ. Birmingham

Background

GMM / HMM-based feature mapping

- **Applications**
  - Speaker conversion [Stylianou;'98, Kain;'98]
  - Acoustic-to-articulatory inversion mapping [Shiga;'04, Toda';08]
  - Noise compensation [Droppo;'02, Cui;'07]
- **Mapping process**
  - Model joint PDF between src. & tgt. feats. by a GMM or HMMs
  - Convert src. feats. into tgt. ones using conditional PDFs
- **Problems**
  - Discontinuities caused by frame-by-frame mapping
  - Mapping with dynamic-feature constraints [Toda';07]

    ⇒ Inconsistencies between training & mapping

Background

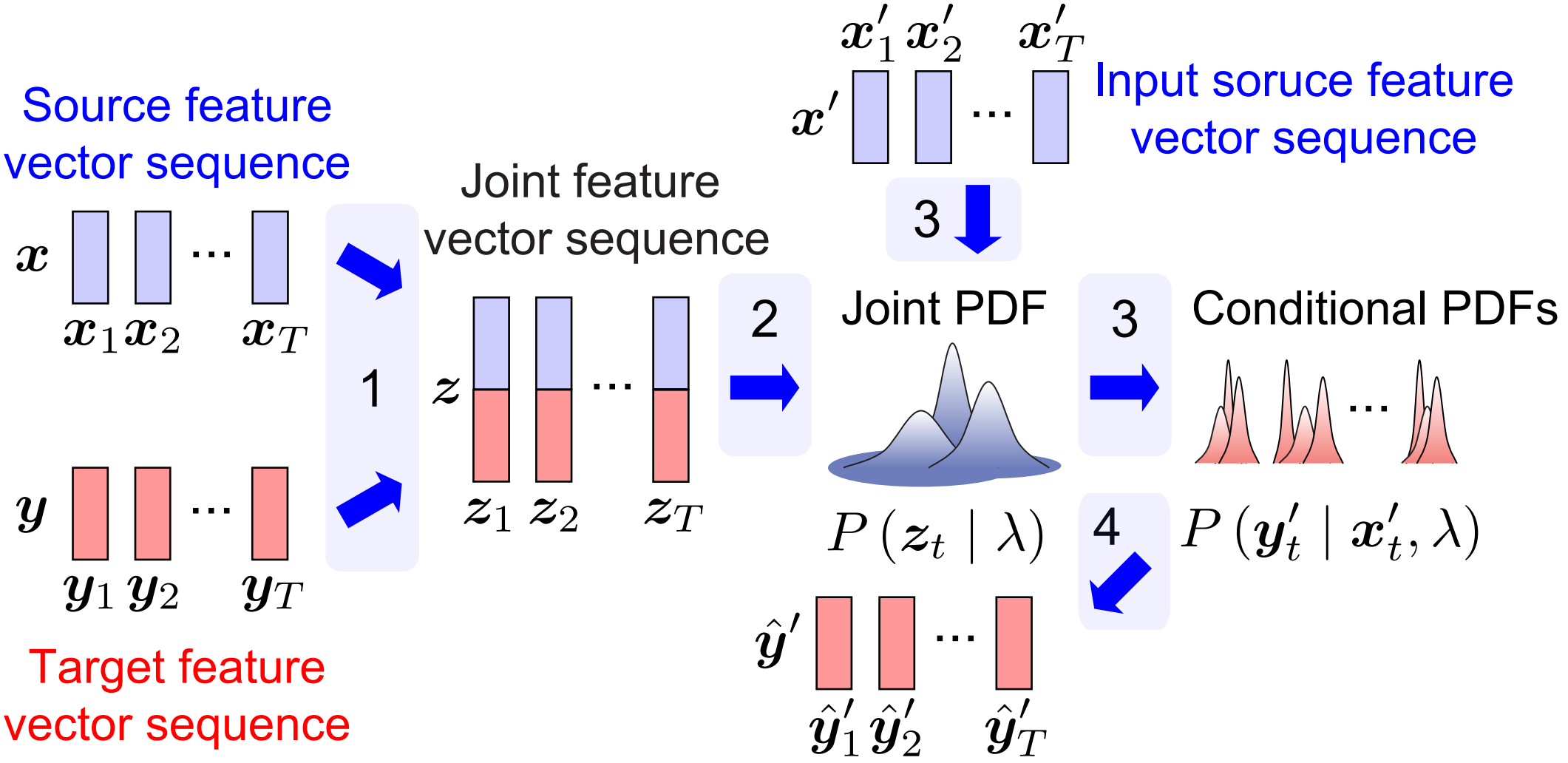
Trajectory GMM / HMM-based feature mapping

- **Trajectory HMM** [Zen;'07]
  - Impose explicit relationships between static & dynamic feats.
    ⇒ HMM is reformulated as a trajectory model
  - Avoid underlying assumptions of HMMs
    * Frame-wise conditional independence of state-output probs.
    * Piece-wise constant statistics within an HMM state
- **Feature mapping based on trajectory HMMs / GMMs**
  - Using dynamic feature constraints in both training & mapping
    ⇒ Make training & mapping consistent
  - Entire utterance-level transformation
    ⇒ Appropriate static & dynamic characteristics

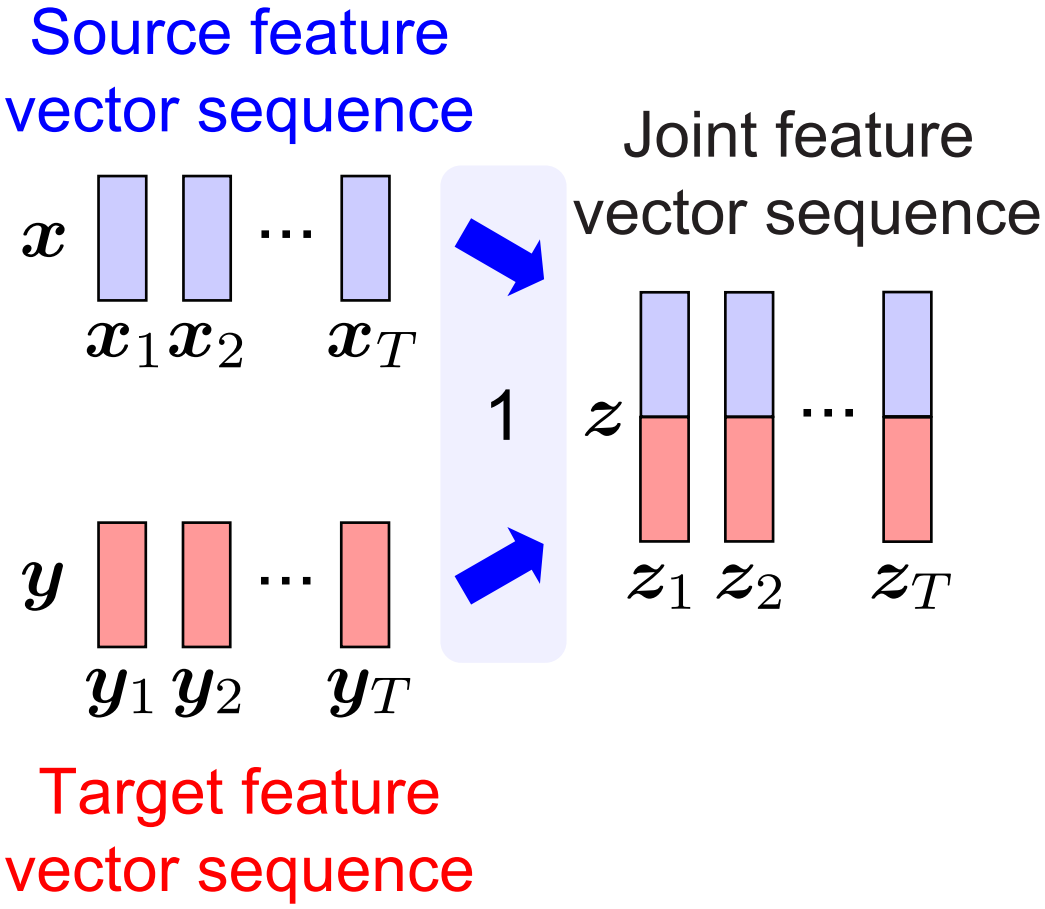
Outline

- **Background**
- **GMM-based feature mapping**
- **GMM-based feature mapping with dynamic features**
- **Trajectory GMM / Trajectory HMM-based mapping**
- **Experiments**
  - Speaker conversion
  - Acoustic-to-articulatory inversion mapping
  - Noise compensation
- **Conclusions**

GMM-based feature mapping [Kain;'98]

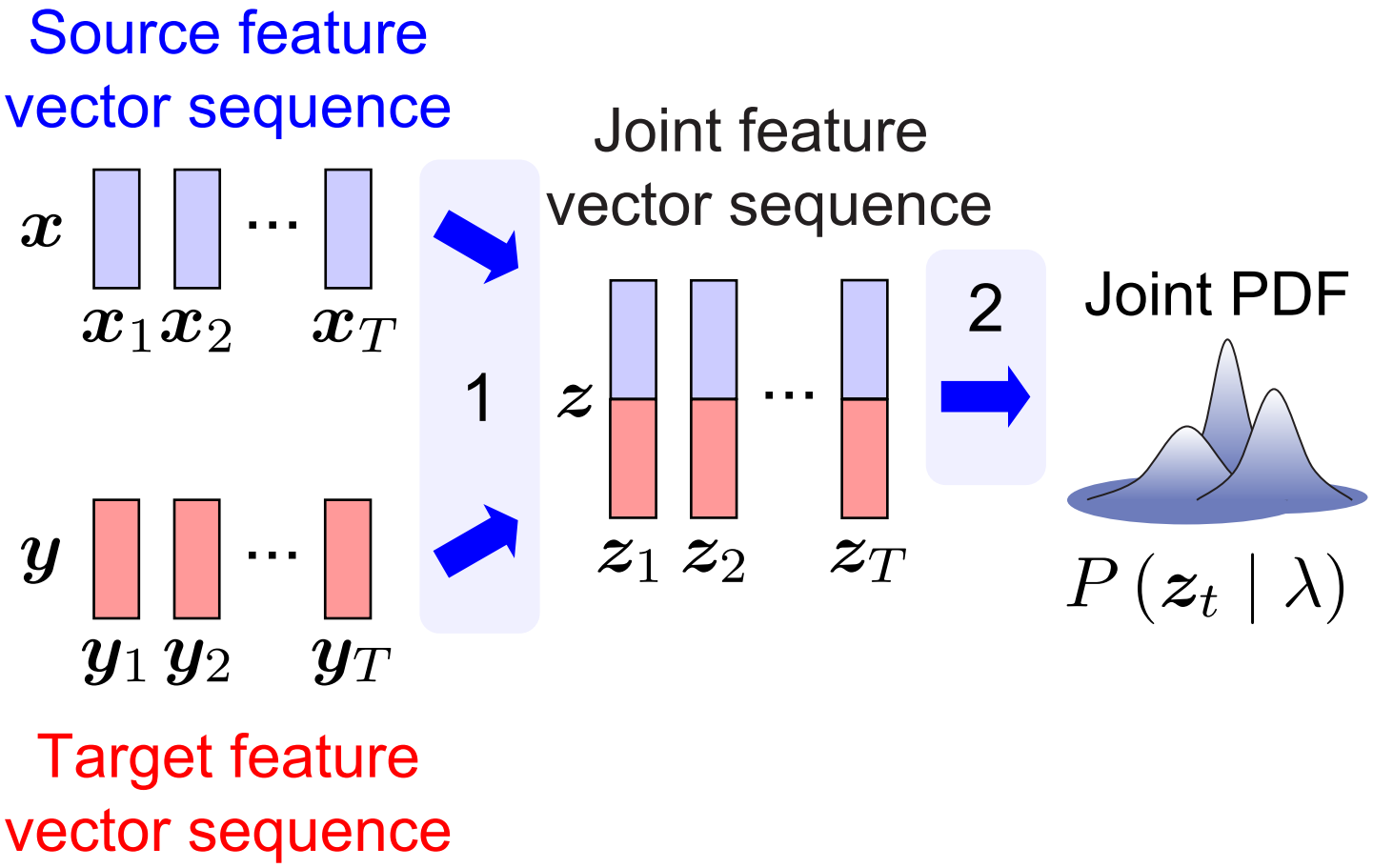


GMM-based feature mapping [Kain;'98]



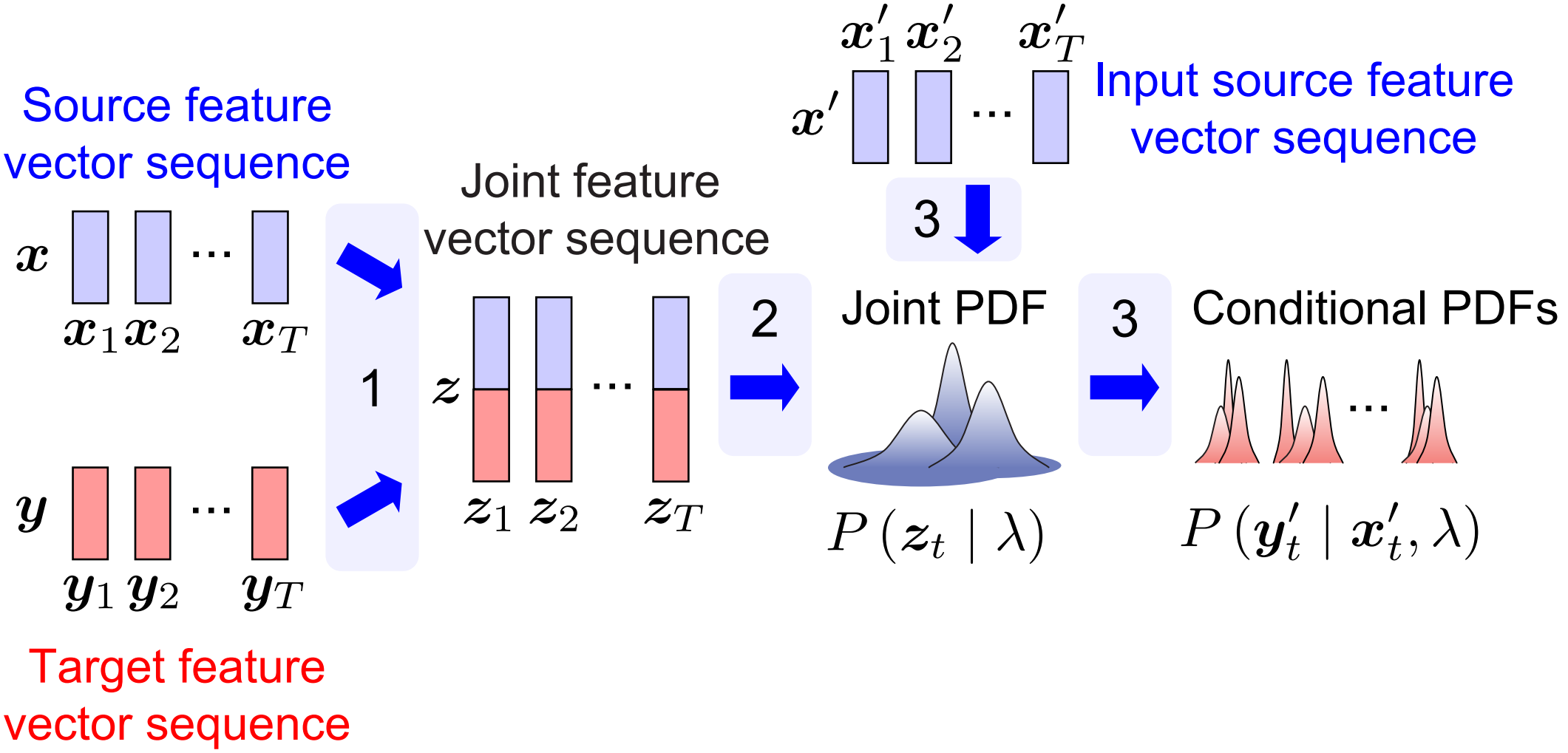
1. Make joint feat. vector $\boldsymbol{z}_t$ from source & target vectors $\boldsymbol{x}_t$ & $\boldsymbol{y}_t$

GMM-based feature mapping [Kain;'98]



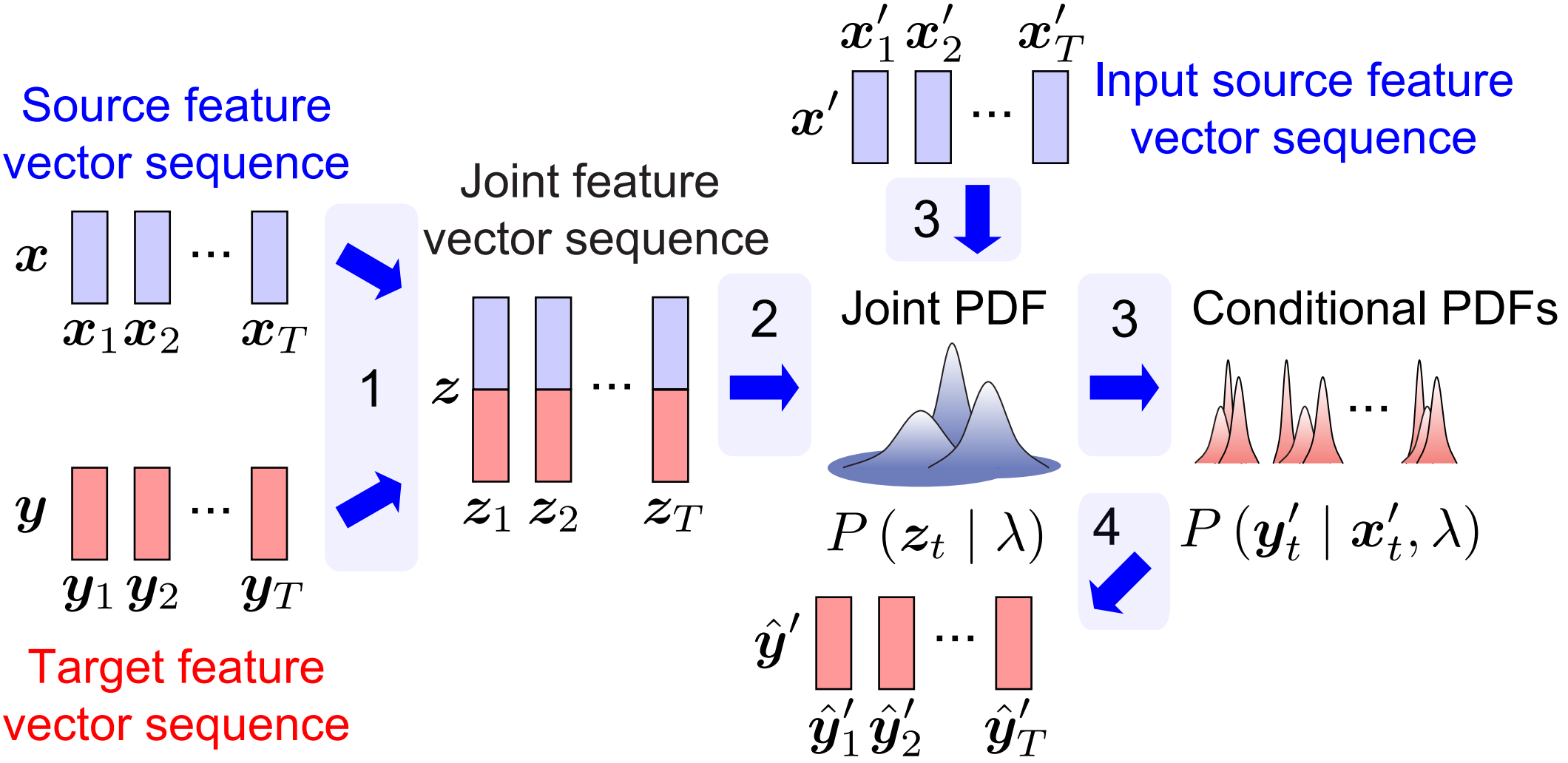
2. Model frame-level joint PDF $P(z_t \mid \lambda)$ by GMM

GMM-based feature mapping [Kain;'98]



3. Convert joint PDF $P(z_t \mid \lambda)$ to conditional PDF $P(y'_t \mid x'_t, \lambda)$

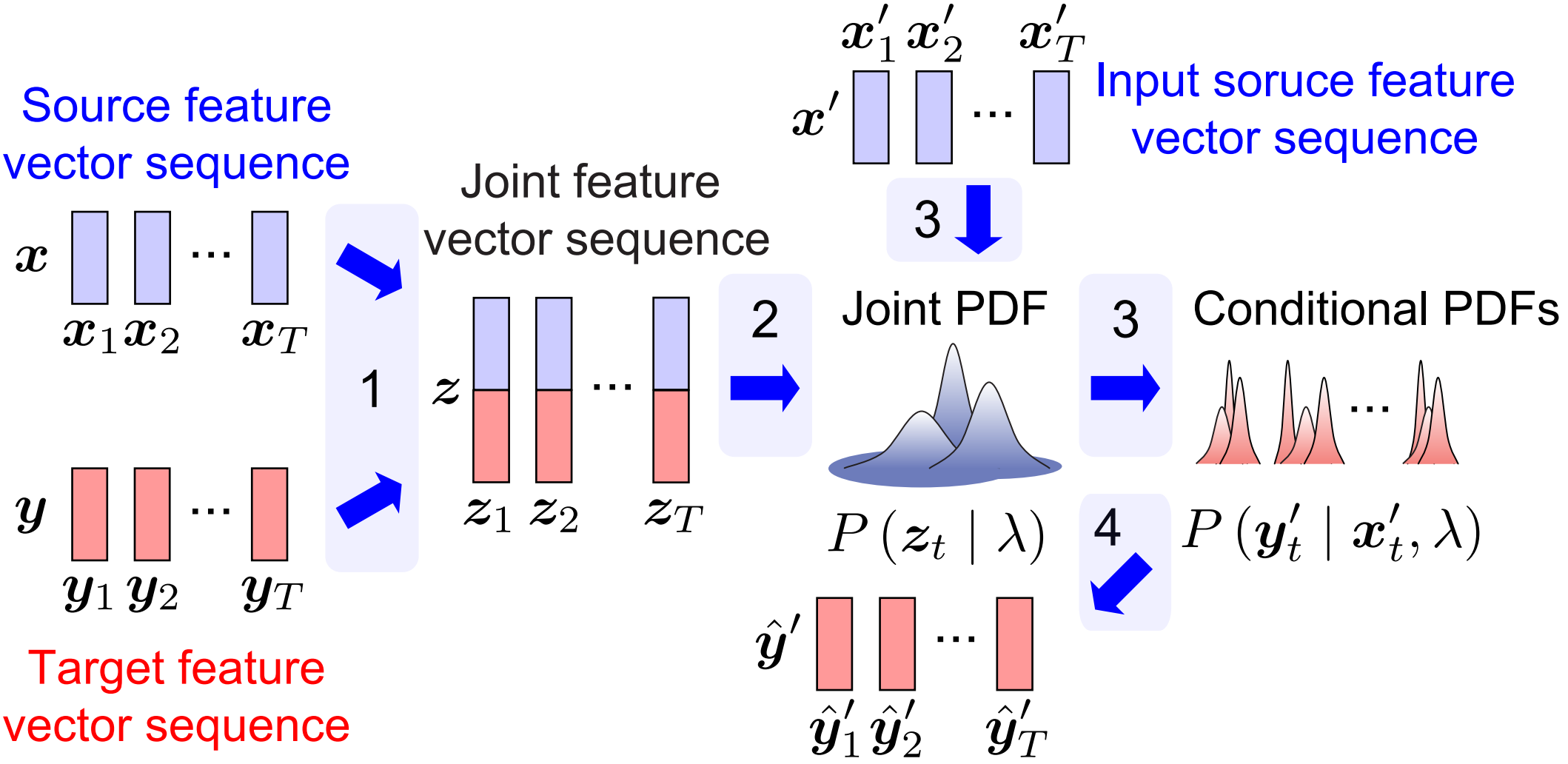
GMM-based feature mapping [Kain;'98]



4. Estimate $\hat{\boldsymbol{y}}'_t$ from conditional PDF by MMSE

$$\hat{\boldsymbol{y}}'_t = \sum_{i=1}^{N} \gamma_i \left[\boldsymbol{\mu}_i^{(\boldsymbol{y})} + \boldsymbol{\Sigma}_i^{(\boldsymbol{yx})} \boldsymbol{\Sigma}_i^{(\boldsymbol{xx})^{-1}} \left(\boldsymbol{x}'_t - \boldsymbol{\mu}_i^{(\boldsymbol{x})} \right) \right]$$

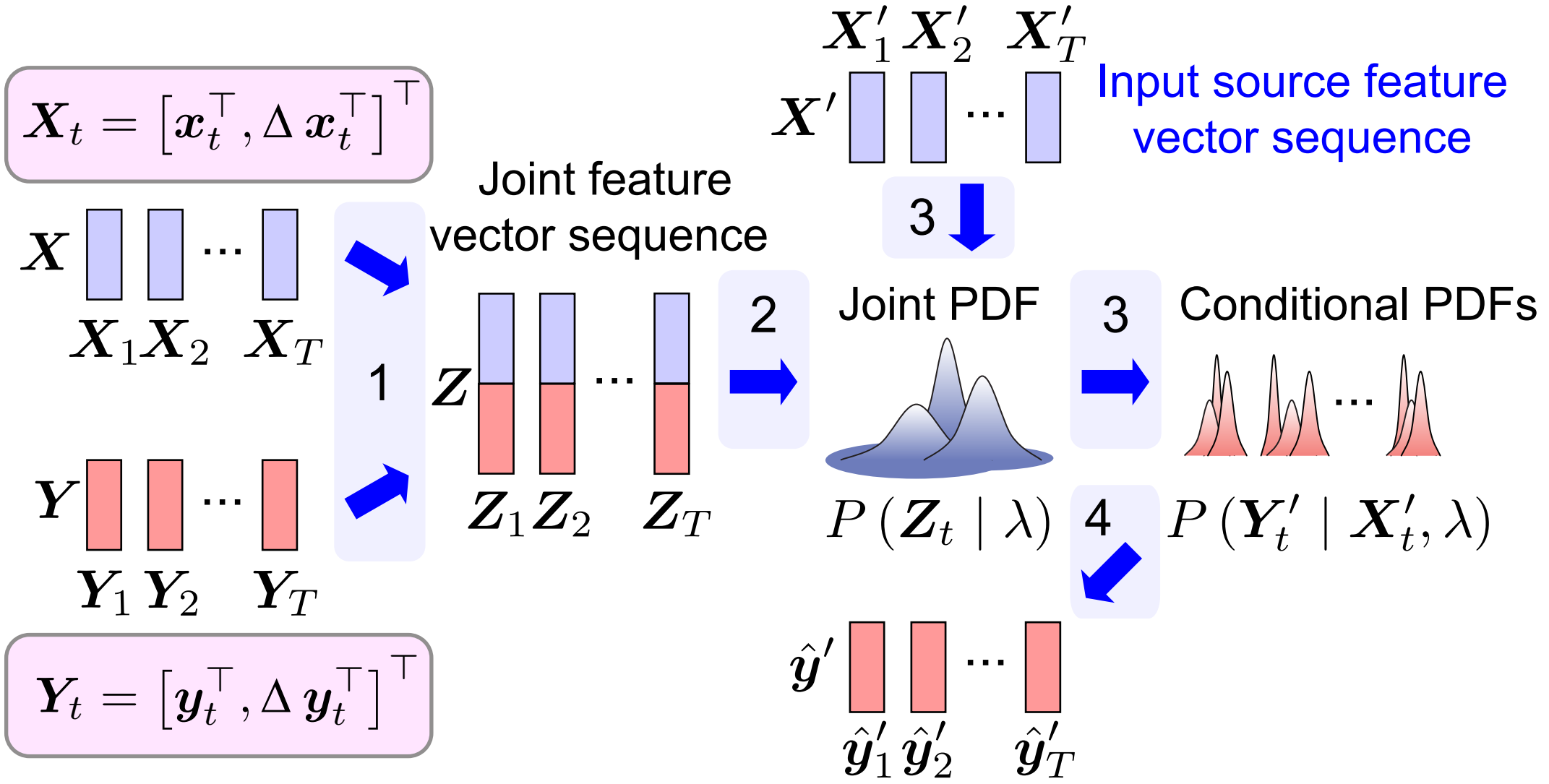
GMM-based feature mapping [Kain;'98]



Frame-by-frame mapping

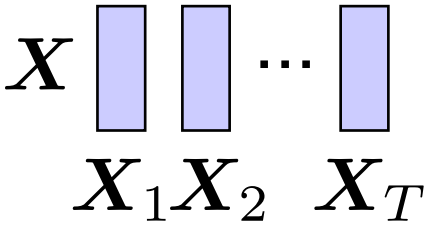
⇒ Mapped features are sometimes discontinuous

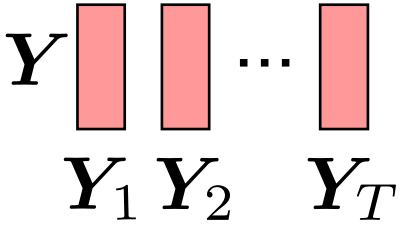
GMM-based mapping with dyn. feats. [Toda;'07]



GMM-based mapping with dyn. feats. [Toda;'07]

$$\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta \boldsymbol{x}_t^\top\right]^\top$$





$$\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta \boldsymbol{y}_t^\top\right]^\top$$

GMM-based mapping with dyn. feats. [Toda;'07]

$$\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta \boldsymbol{x}_t^\top\right]^\top$$

$\boldsymbol{X}$ $\boldsymbol{X}_1 \boldsymbol{X}_2 \cdots \boldsymbol{X}_T$

$\boldsymbol{Y}$ $\boldsymbol{Y}_1 \boldsymbol{Y}_2 \cdots \boldsymbol{Y}_T$

$$\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta \boldsymbol{y}_t^\top\right]^\top$$

1

Joint feature vector sequence

$\boldsymbol{Z}$ $\boldsymbol{Z}_1 \boldsymbol{Z}_2 \cdots \boldsymbol{Z}_T$

2

Joint PDF

$P(\boldsymbol{Z}_t \mid \lambda)$

3

Conditional PDFs

$P(\boldsymbol{Y}'_t \mid \boldsymbol{X}'_t, \lambda)$

$\boldsymbol{X}'$ $\boldsymbol{X}'_1 \boldsymbol{X}'_2 \cdots \boldsymbol{X}'_T$

Input source feature vector sequence

3

4

$\hat{\boldsymbol{y}}'$ $\hat{\boldsymbol{y}}'_1 \hat{\boldsymbol{y}}'_2 \cdots \hat{\boldsymbol{y}}'_T$

4. Estimate $\hat{\boldsymbol{y}}'$ from conditional PDF by ML under the constraints between static & dynamic features ($\boldsymbol{Y}' = \boldsymbol{W}\boldsymbol{y}'$)

Mapping with dynamic-feature constraints

$$\hat{\boldsymbol{y}}' = \arg\max_{\boldsymbol{Y}'} P\left(\boldsymbol{Y}' \mid \boldsymbol{X}', \lambda\right) = \arg\max_{\boldsymbol{y}'} P\left(\boldsymbol{W}\boldsymbol{y}' \mid \boldsymbol{X}', \lambda\right)$$

$$\approx \left(\boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{q}}^{(\boldsymbol{y})^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\boldsymbol{q}}^{(\boldsymbol{y})^{-1}} \boldsymbol{E}_{\boldsymbol{q}}^{(\boldsymbol{y})} \quad \text{(Viterbi approx.)}$$

$$\boldsymbol{E}_{\boldsymbol{q}}^{(\boldsymbol{y})} = \left[\boldsymbol{E}_{tq_1}^{(\boldsymbol{y})^\top}, \boldsymbol{E}_{tq_2}^{(\boldsymbol{y})^\top}, \ldots, \boldsymbol{E}_{tq_T}^{(\boldsymbol{y})^\top}\right]^\top \qquad \boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$$

$$\boldsymbol{D}_{\boldsymbol{q}}^{(\boldsymbol{y})} = \mathrm{diag}\left[\boldsymbol{D}_{q_1}^{(\boldsymbol{y})}, \boldsymbol{D}_{q_2}^{(\boldsymbol{y})}, \ldots, \boldsymbol{D}_{q_T}^{(\boldsymbol{y})}\right]$$

$$\boldsymbol{E}_{ti}^{(\boldsymbol{y})} = \boldsymbol{\mu}_i^{(\boldsymbol{y})} + \boldsymbol{\Sigma}_i^{(\boldsymbol{y}\boldsymbol{x})} \boldsymbol{\Sigma}_i^{(\boldsymbol{x}\boldsymbol{x})^{-1}} \left(\boldsymbol{X}_t' - \boldsymbol{\mu}_i^{(\boldsymbol{x})}\right)$$

$$\boldsymbol{D}_i^{(\boldsymbol{y})} = \boldsymbol{\Sigma}_i^{(\boldsymbol{y}\boldsymbol{y})} - \boldsymbol{\Sigma}_i^{(\boldsymbol{y}\boldsymbol{x})} \boldsymbol{\Sigma}_i^{(\boldsymbol{x}\boldsymbol{x})^{-1}} \boldsymbol{\Sigma}_i^{(\boldsymbol{x}\boldsymbol{y})}$$

- **Entire utterance-level transformation**
  - ⇒ Mapped feats. have proper static & dynamic characteristics
- **Dynamic-feature constraints are used at mapping only**
  - ⇒ Inconsistencies between training & mapping

Trajectory HMM [Zen;'07]

- **Derived from HMMs with explicit dyn. feat. constraints**
- **Underlying generative model of HMM-based synthesis**

$$P(\boldsymbol{z} \mid \lambda) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{q} \mid \lambda) P(\boldsymbol{z} \mid \boldsymbol{q}, \lambda)$$

$$P(\boldsymbol{z} \mid \boldsymbol{q}, \lambda) = \mathcal{N}\left(\boldsymbol{z};\ \bar{\boldsymbol{z}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right)$$

$$\boldsymbol{R}_{\boldsymbol{q}} \bar{\boldsymbol{z}}_{\boldsymbol{q}} = \boldsymbol{r}_{\boldsymbol{q}}$$

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^{\top}, \ldots, \boldsymbol{\mu}_{q_T}^{\top}\right]^{\top}$$

$$\boldsymbol{\mu}_i = \left[\boldsymbol{\mu}_i^{(\boldsymbol{x})\top}, \boldsymbol{\mu}_i^{(\boldsymbol{y})\top}\right]^{\top}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \operatorname{diag}\left[\boldsymbol{\Sigma}_{q_1}^{-1}, \ldots, \boldsymbol{\Sigma}_{q_T}^{-1}\right]$$

$$\boldsymbol{\Sigma}_i^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_i^{(\boldsymbol{xx})} & \boldsymbol{\Omega}_i^{(\boldsymbol{xy})} \\ \boldsymbol{\Omega}_i^{(\boldsymbol{yx})} & \boldsymbol{\Omega}_i^{(\boldsymbol{yy})} \end{bmatrix}$$

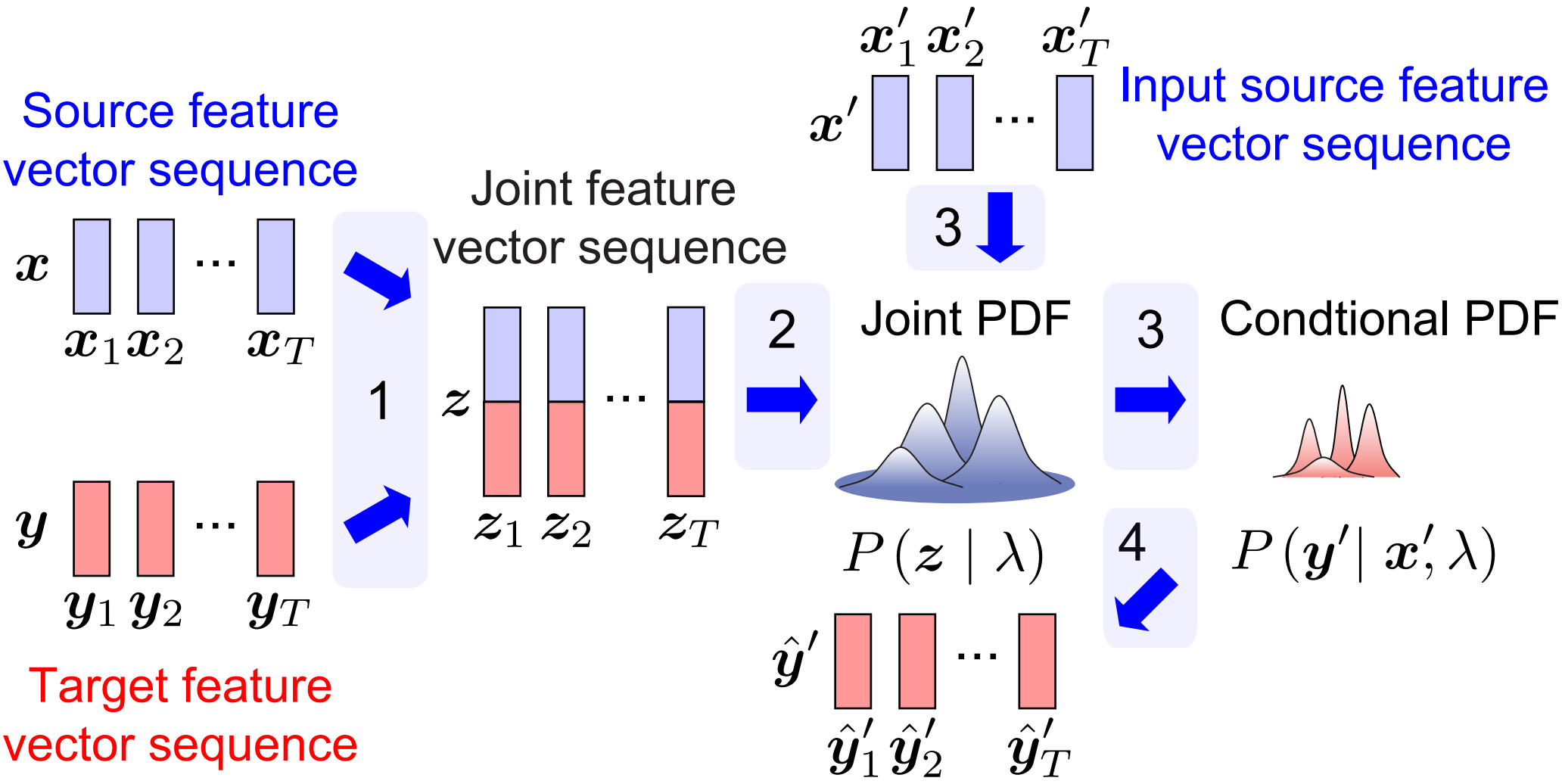
- $\bar{\boldsymbol{z}}_{\boldsymbol{q}}$ **is determined as a smooth trajectory**

  ⇒ Speech dynamics can be modeled explicitly

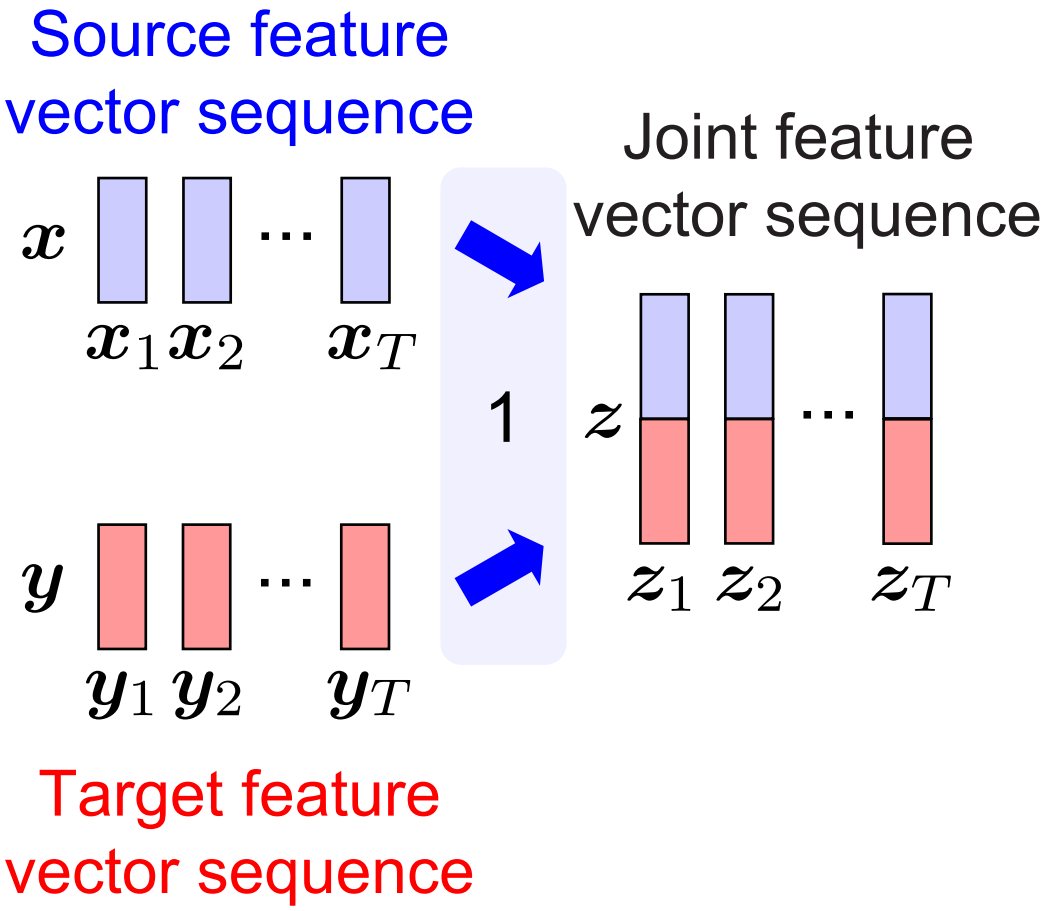
- **Intra- & inter-frame covariance matrix** $\boldsymbol{P}_{\boldsymbol{q}}$ **becomes full**

  ⇒ Capture both intra- & inter-frame dependencies

Trajectory GMM / Trajectory HMM-based mapping

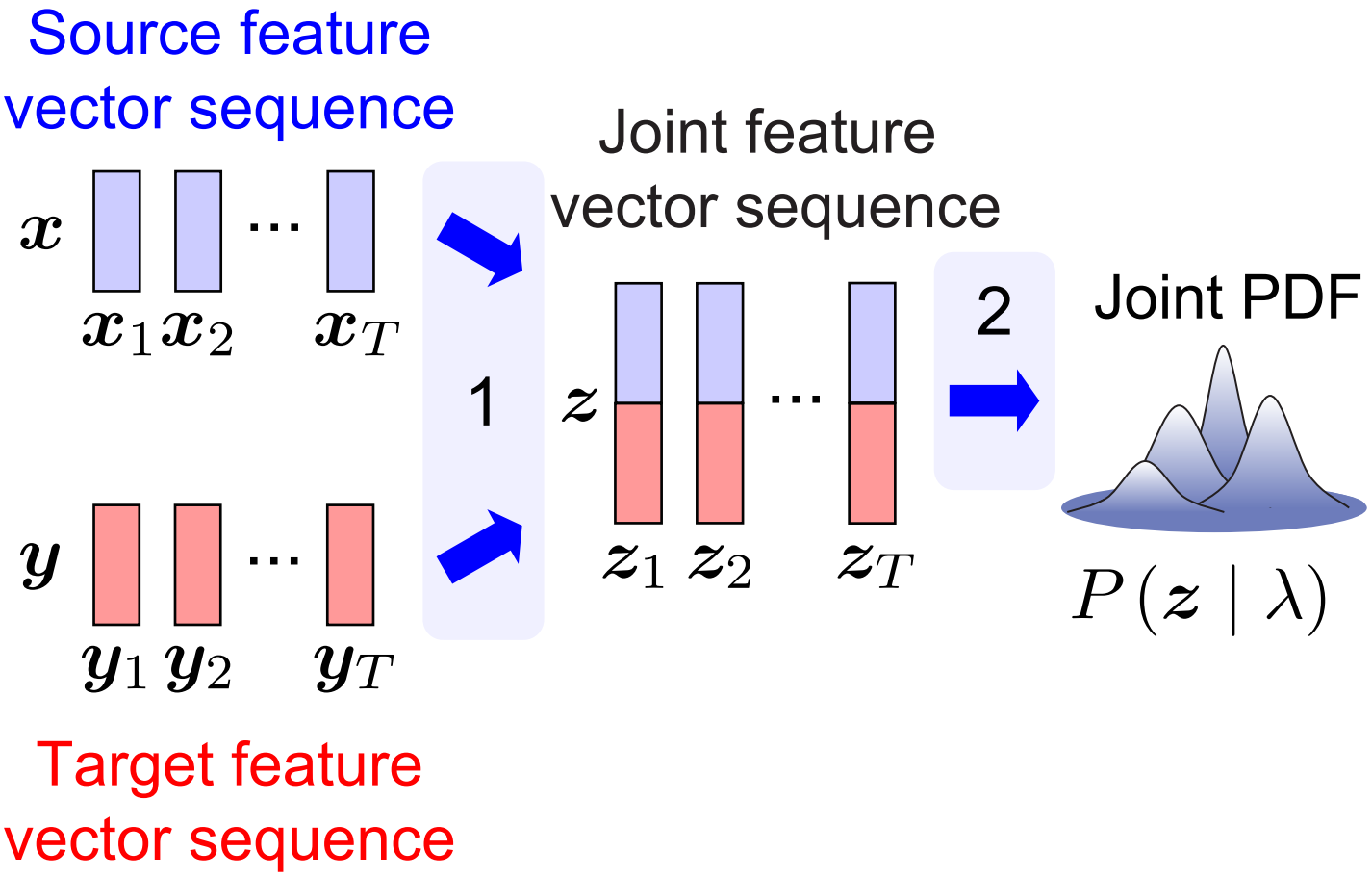


Trajectory GMM / Trajectory HMM-based mapping



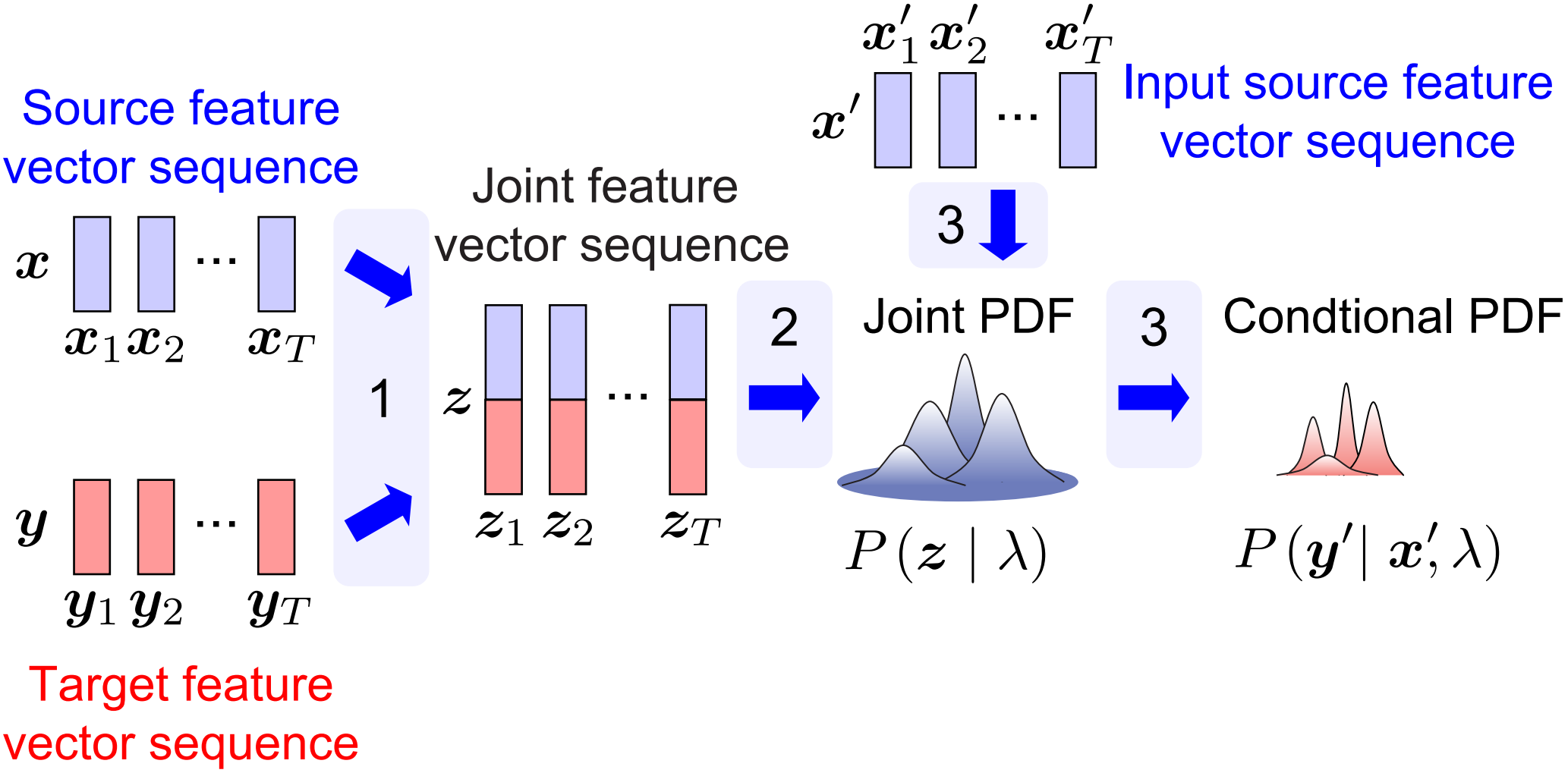
1. Make joint feat. vec. seq. z from src. & tgt. vec. seqs. x & y

Trajectory GMM / Trajectory HMM-based mapping



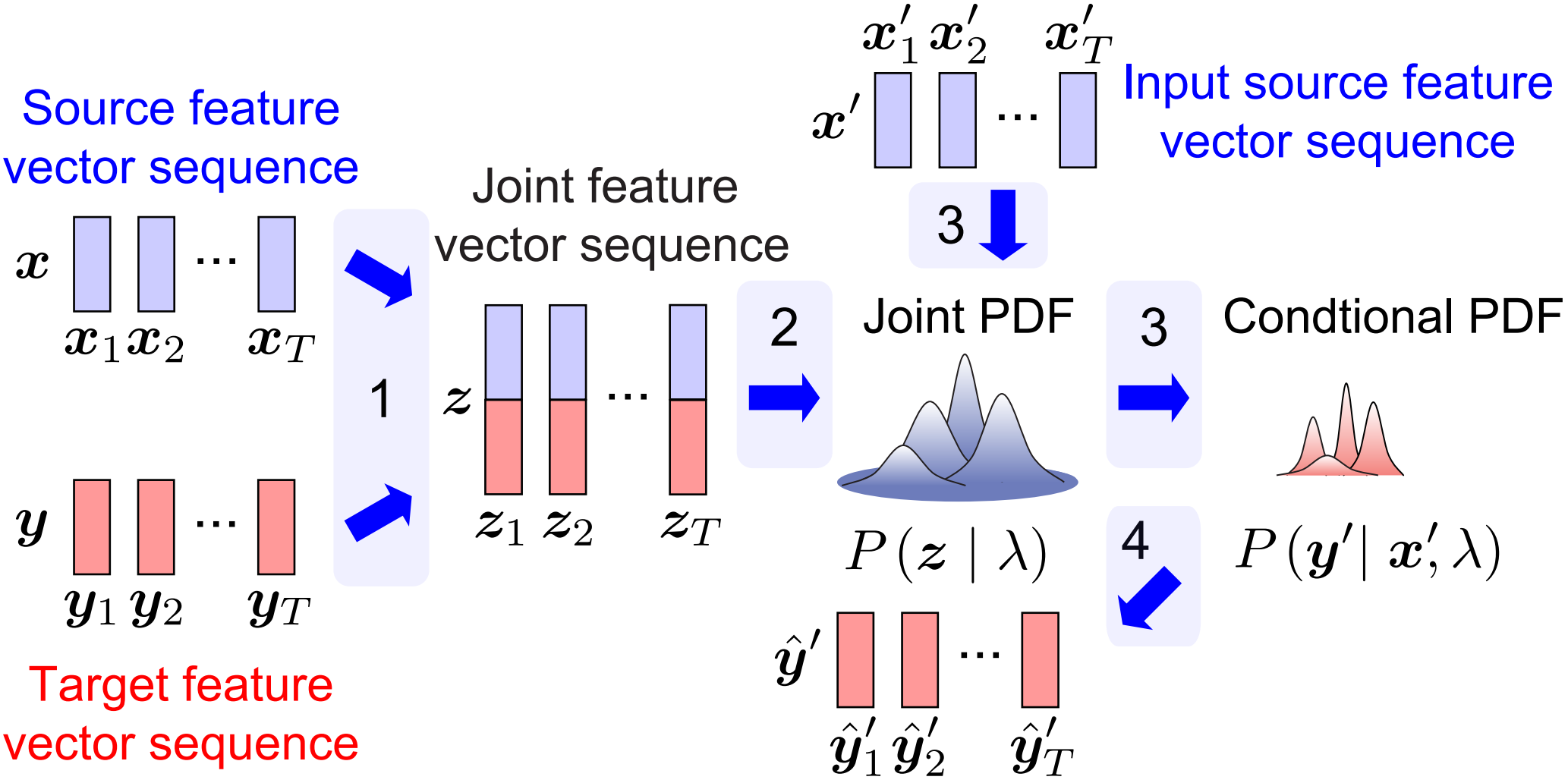
2. Model utterance-level joint PDF $P(\boldsymbol{z} \mid \lambda)$ by trajectory GMM

Trajectory GMM / Trajectory HMM-based mapping



3. Convert joint PDF $P(z \mid \lambda)$ to conditional PDF $P(y' \mid x', \lambda)$

Trajectory GMM / Trajectory HMM-based mapping



4. Estimate $\hat{y}'$ from conditional PDF by MMSE or MAP

Trajectory GMM / Trajectory HMM-based mapping

$$\hat{\boldsymbol{y}}'_{\mathrm{MMSE}} = \sum_{\forall \boldsymbol{q}} \gamma_{\boldsymbol{q}} \left[\bar{\boldsymbol{y}}_{\boldsymbol{q}} + \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \left(\boldsymbol{x}' - \bar{\boldsymbol{x}}_{\boldsymbol{q}} \right) \right]$$

$$\hat{\boldsymbol{y}}'_{\mathrm{MAP}} = \left(\sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \right)^{-1} \sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \left[\bar{\boldsymbol{y}}_{\boldsymbol{q}} + \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \left(\boldsymbol{x}' - \bar{\boldsymbol{x}}_{\boldsymbol{q}} \right) \right]$$

- Entire utterance-level transform
  ⇒ Mapped feats. have proper static & dyn. characteristics
- Dyn. feat. constraints are used in both training & mapping
  ⇒ Make training & mapping consistent

Outline

- **Background**
- **GMM-based feature mapping**
- **GMM-based feature mapping with dynamic features**
- **Trajectory GMM / Trajectory HMM-based mapping**
- **Experiments**
  - Speaker conversion
  - Acoustic-to-articulatory inversion mapping
  - Noise compensation
- **Conclusions**

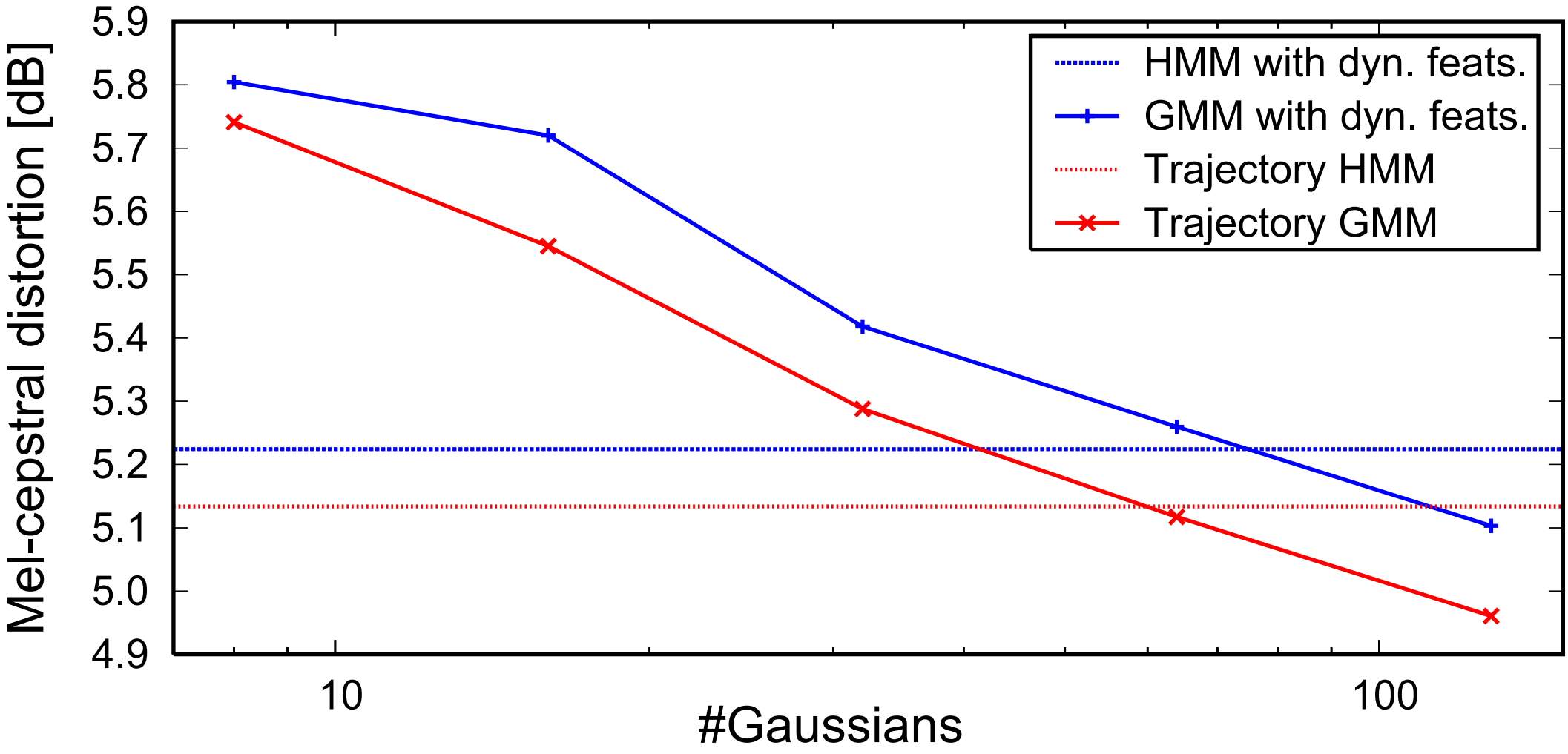
Experiment (speaker conversion)

Experimental conditions

Database	CMU ARCTIC speech database
Training data	Speakers BDL & RMS, first 593 utterances (Mapping: BDL→RMS)
Test data	Last 40 utterances
Sampling freq.	16 kHz
Analysis win.	25-ms Blackman window / 5-ms shift
Feature vec.	0～24 order Mel-cepstral coefficients, Δ & $\Delta\Delta$
Topology	(Trajectory) HMM: 3-state, left-to-right no-skip, monophone 1-mix (121 states) (Trajectory) GMM: 128-mix
Mapping	Spectrum: GMM (HMM) with dyn. or proposed F0: Linearly transformed in the log domain

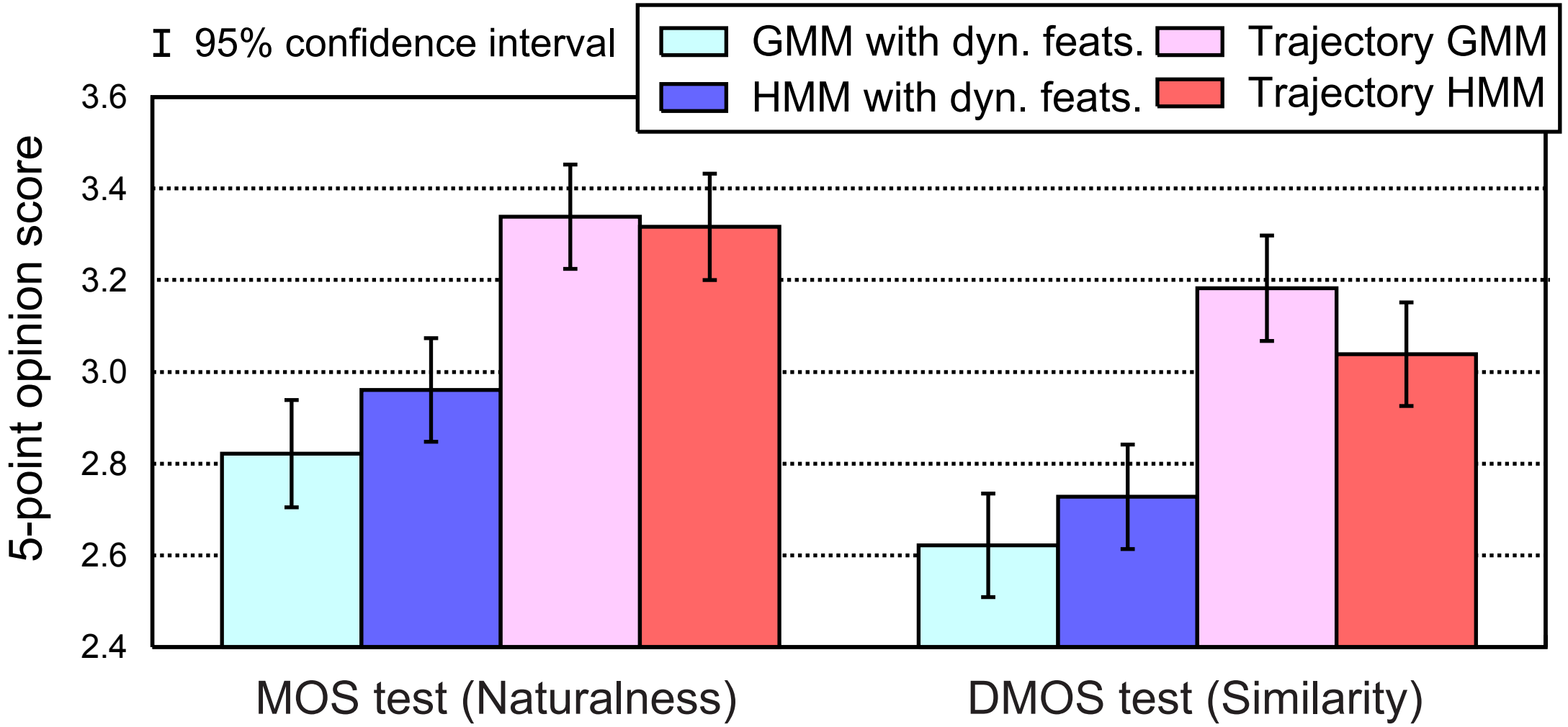
Experiment (speaker conversion)

Objective evaluation (mel-cepstral distortion)



Experiment (speaker conversion)

Subjective evaluation (MOS & DMOS)



(#subjects=12, 15 sentences/subjects)

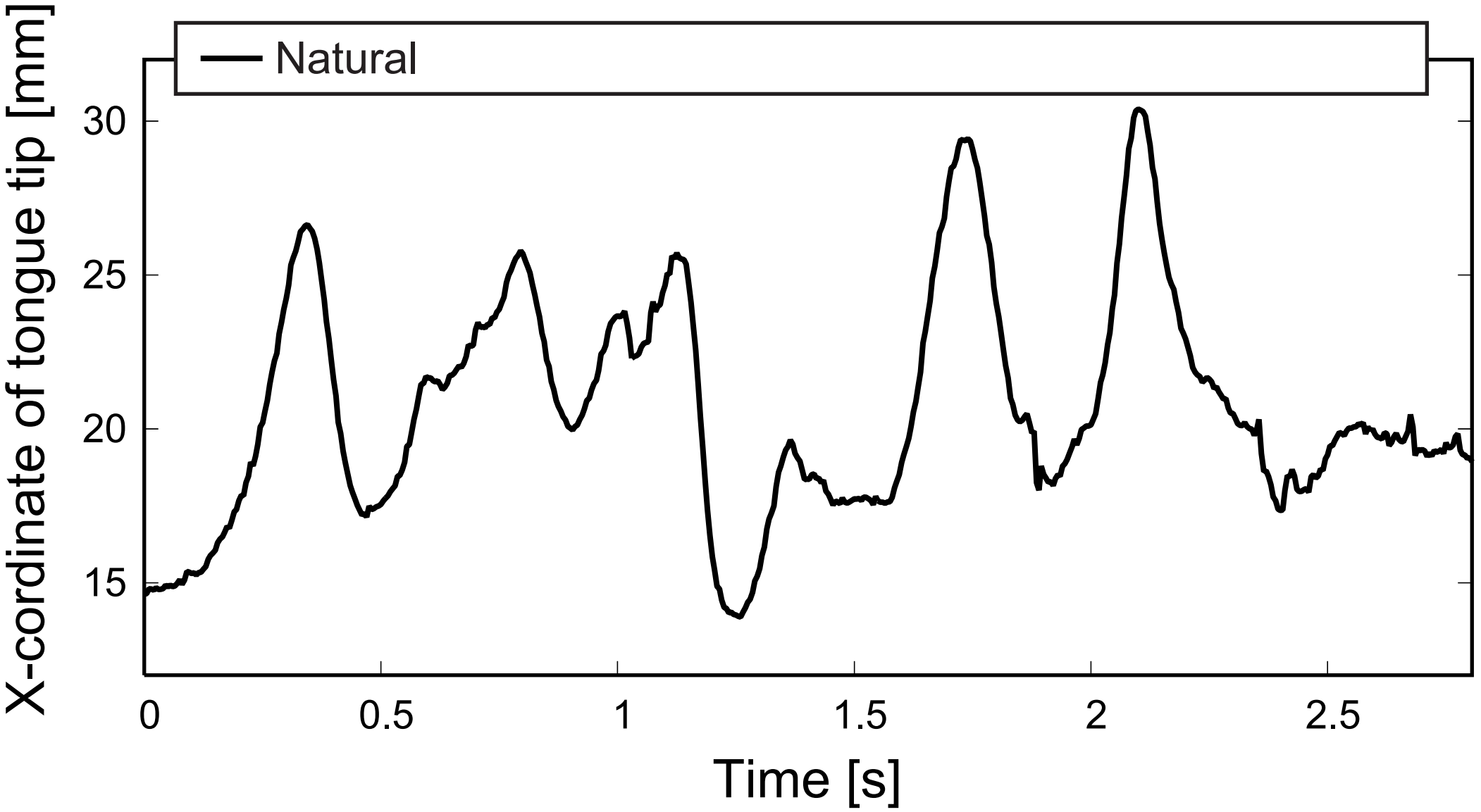
Experiment (acoustic-articulatory inv. mapping)

Experimental conditions

Database	MOCHA-TIMIT database
Training data	Speaker msak0, 368 utterances
Test data	92 utterances
Sampling freq.	16 kHz
Analysis win.	25-ms Blackman window / 5-ms shift
Feature vec.	0～13 order Mel-cepstral coefficients, Δ & $\Delta\Delta$ 7 articulators in x & y coordinates, Δ & $\Delta\Delta$
Topology	(Trajectory) HMM: 3-state, left-to-right no-skip, monophone 1-mix (135 states) (Trajectory) GMM: 128-mix
Mapping	GMM (HMM) with dynamic feature constraints or proposed trajectory GMM (HMM)-based one

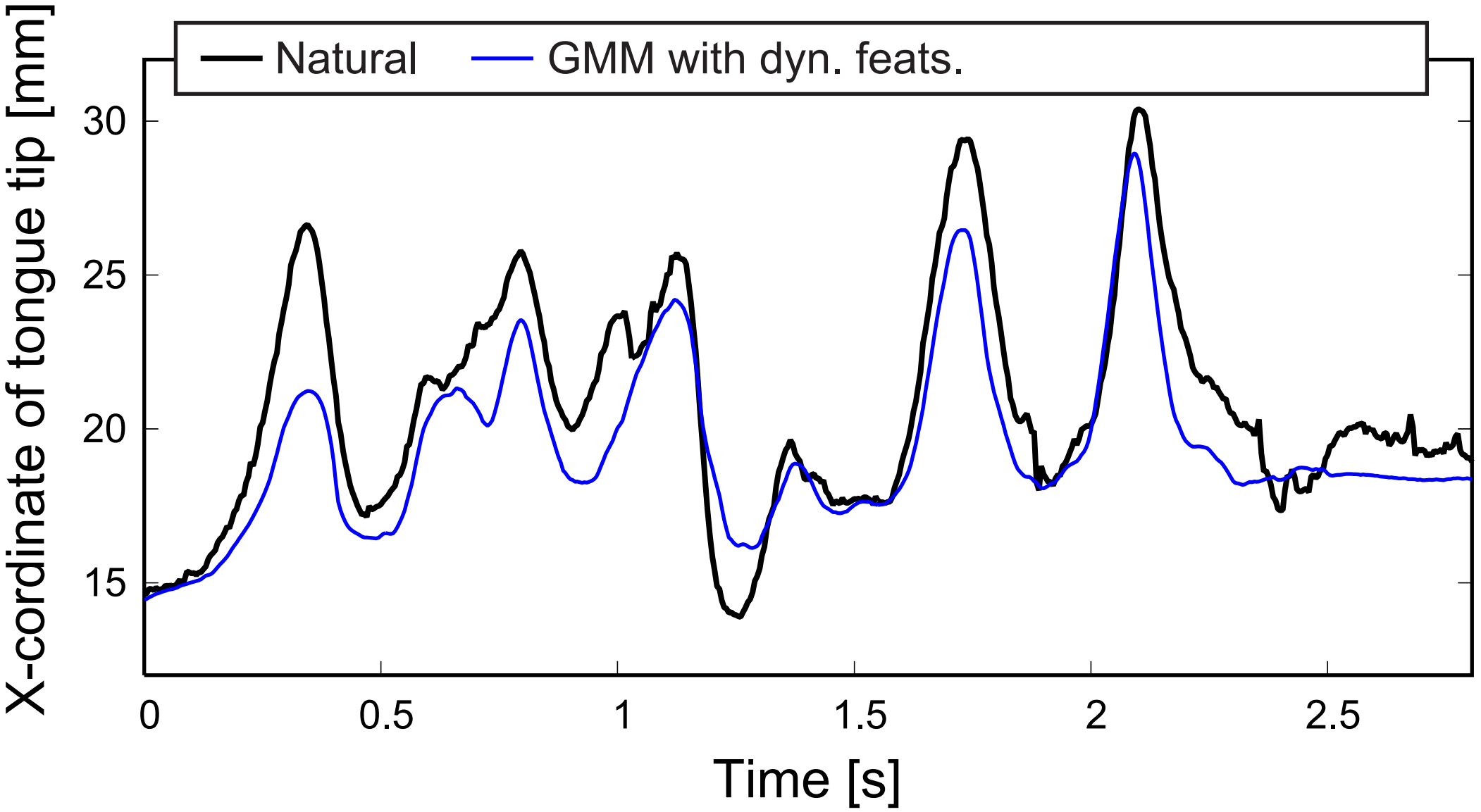
Experiment (acoustic-articulatory inv. mapping)

Examples of estimated articulatory movements



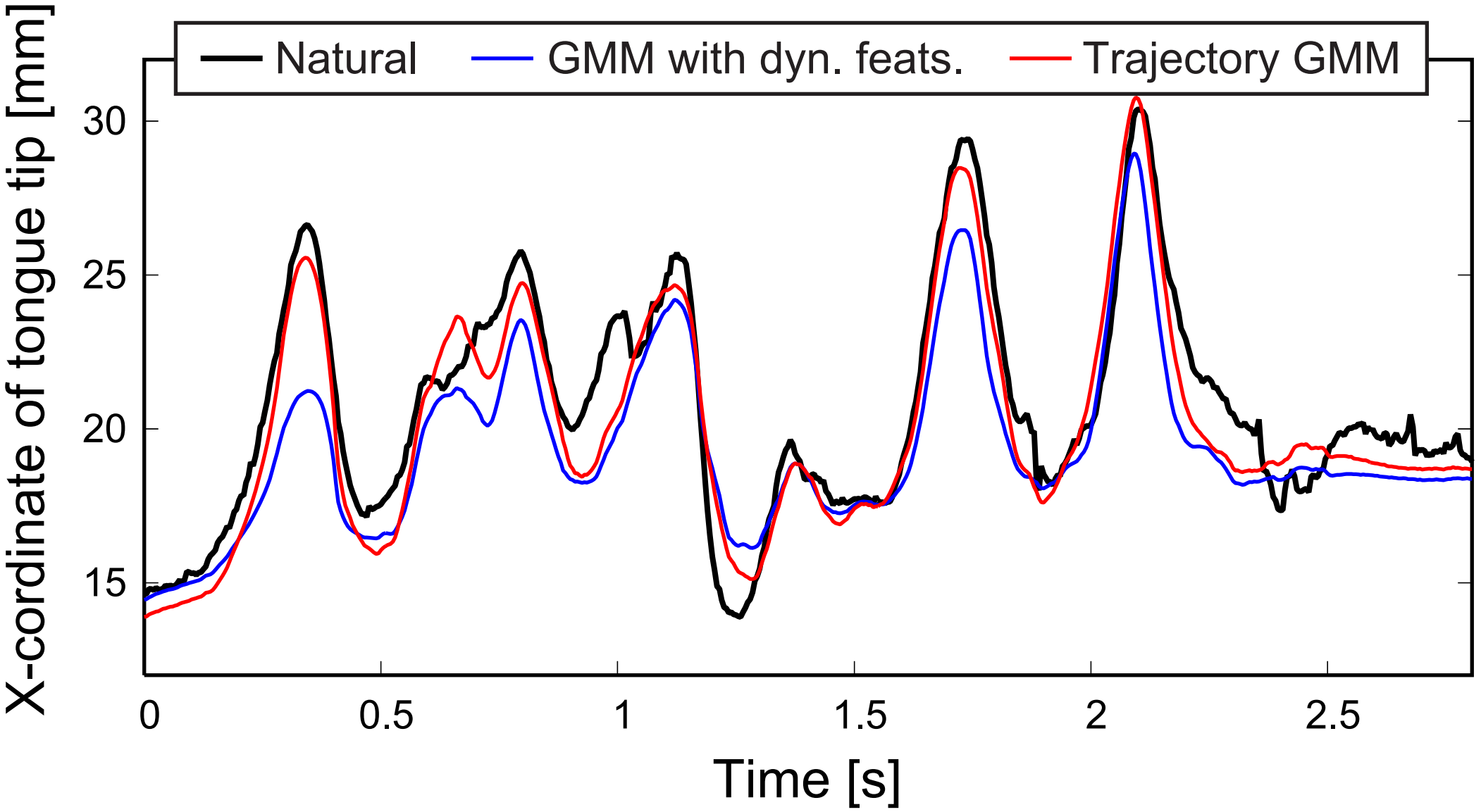
Experiment (acoustic-articulatory inv. mapping)

Examples of estimated articulatory movements



Experiment (acoustic-articulatory inv. mapping)

Examples of estimated articulatory movements



Experiment (acoustic-articulatory inv. mapping)

Average RMSE over the 14 channels (in mm)

Mapping	Dev. set	Eval. set
HMM with dyn. feats.	1.590	1.701
GMM with dyn. feats.	1.256	1.327
Trajectory HMM	1.412	1.523
Trajectory GMM	**1.071**	**1.128**

Experiment (noise compensation)

Experimental conditions

HMM training

Database	AURORA-2 database
Training data	Clean-condition training data used in the ETSI ref.
Feature vec.	12 MFCC & log energy by ETSI front-end 2.0, Δ & $\Delta\Delta$
Topology	16-state 3-mix left-to-right HMM for each digit

Noisy-clean mapping function training

Database	AURORA-2 database
Training data	Multi-condition training data used in the ETSI ref.
Feature vec.	12 MFCC & log energy by ETSI front-end 2.0, Δ & $\Delta\Delta$
Topology	(Trajectory) GMM: 256-mix for each noise condition

Experiment (noise compensation)

Evaluated techniques

- **GMM-static**
  - Train GMM using static feats. only
  - Map static feats. using GMM
- **GMM-complete**
  - Train GMM using both static & dynamic feats.
  - Map augmented feats. using GMM
- **GMM-dynamic**
  - Train GMM using both static & dynamic feats.
  - Map augmented feats. using GMM under dyn. feat. constraints
- **Trajectory GMM**
  - Map static feats. using trajectory GMM

Experiment (noise compensation)

Experimental results (average word accuracy (%))

Set A: seen noise, Set B: unseen noise, Set C: channel mismatch

Mapping	Set A	Set B	Set C	Average
w/o compensation	61.34	55.75	66.14	60.06
GMM-static	82.49	74.32	69.30	76.58
GMM-complete	88.03	**82.80**	77.92	83.92
GMM-dynamic	88.26	81.03	79.31	83.58
Trajectory GMM	**89.38**	81.54	**80.87**	**84.54**

Outline

- **Background**
- **GMM-based feature mapping**
- **GMM-based feature mapping with dynamic features**
- **Trajectory GMM / Trajectory HMM-based mapping**
- **Experiments**
  - Speaker conversion
  - Acoustic-to-articulatory inversion mapping
  - Noise compensation
- **Conclusions**

Conclusions

Trajectory GMM / Trajectory HMM-based mapping

- Entire utterance-level mapping
- Using dyn.-feat. constraints in both training & mapping
- Better performance in various tasks

Future works

- Real-time (or low-delay) mapping
- Integrate alignment process into model definition
- Applications to other tasks