

# Continuous Stochastic Feature Mapping Based on Trajectory HMMs

Heiga Zen,<sup>1,2</sup> Yoshihiko Nankaku,<sup>1</sup> Keiichi Tokuda<sup>1</sup>

<sup>1</sup> Nagoya Institute of Technology, Japan

<sup>2</sup> Presently, Toshiba Cambridge Research Lab., UK

# 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

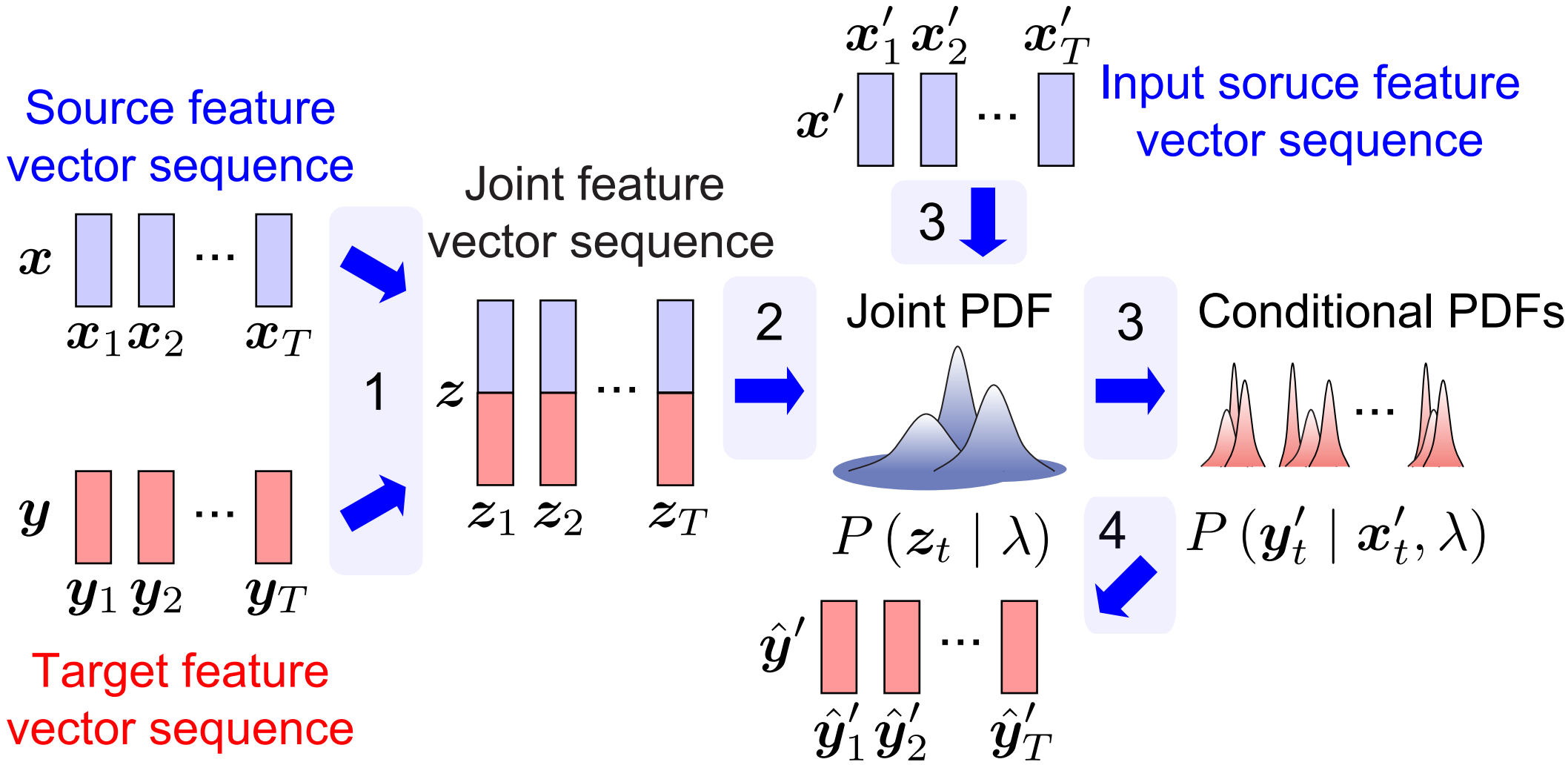
## Trajectory GMM / HMM-based feature mapping

- **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

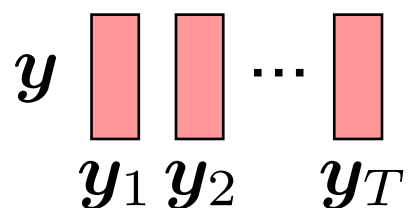
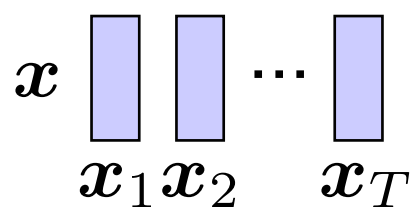
- **Background**
- **GMM-based feature mapping**
- **GMM-based feature mapping with dynamic features**
- **Trajectory GMM / Trajectory HMM-based mapping**
- **Experiments**
  - Speaker conversion
  - Acoustic-to-articulatory inversion mapping
  - Noise compensation
- **Conclusions**

# GMM-based feature mapping [Kain;'98]



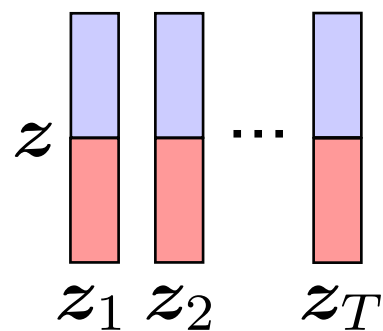
# GMM-based feature mapping [Kain;'98]

Source feature  
vector sequence



Target feature  
vector sequence

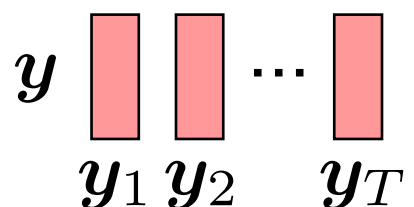
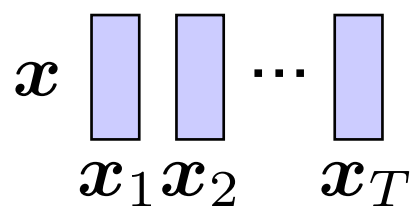
Joint feature  
vector sequence



1. Make joint feat. vector  $z_t$  from source & target vectors  $x_t$  &  $y_t$

# GMM-based feature mapping [Kain;'98]

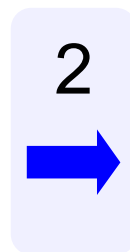
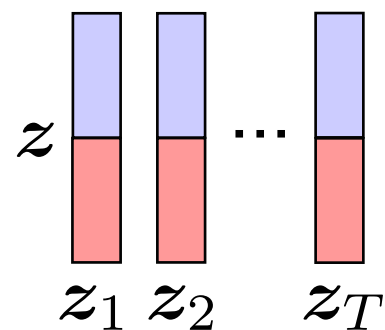
Source feature  
vector sequence



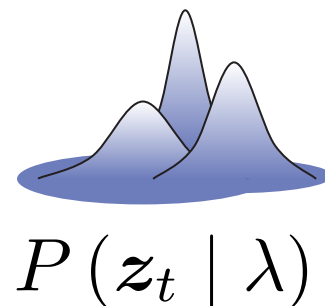
Target feature  
vector sequence



Joint feature  
vector sequence

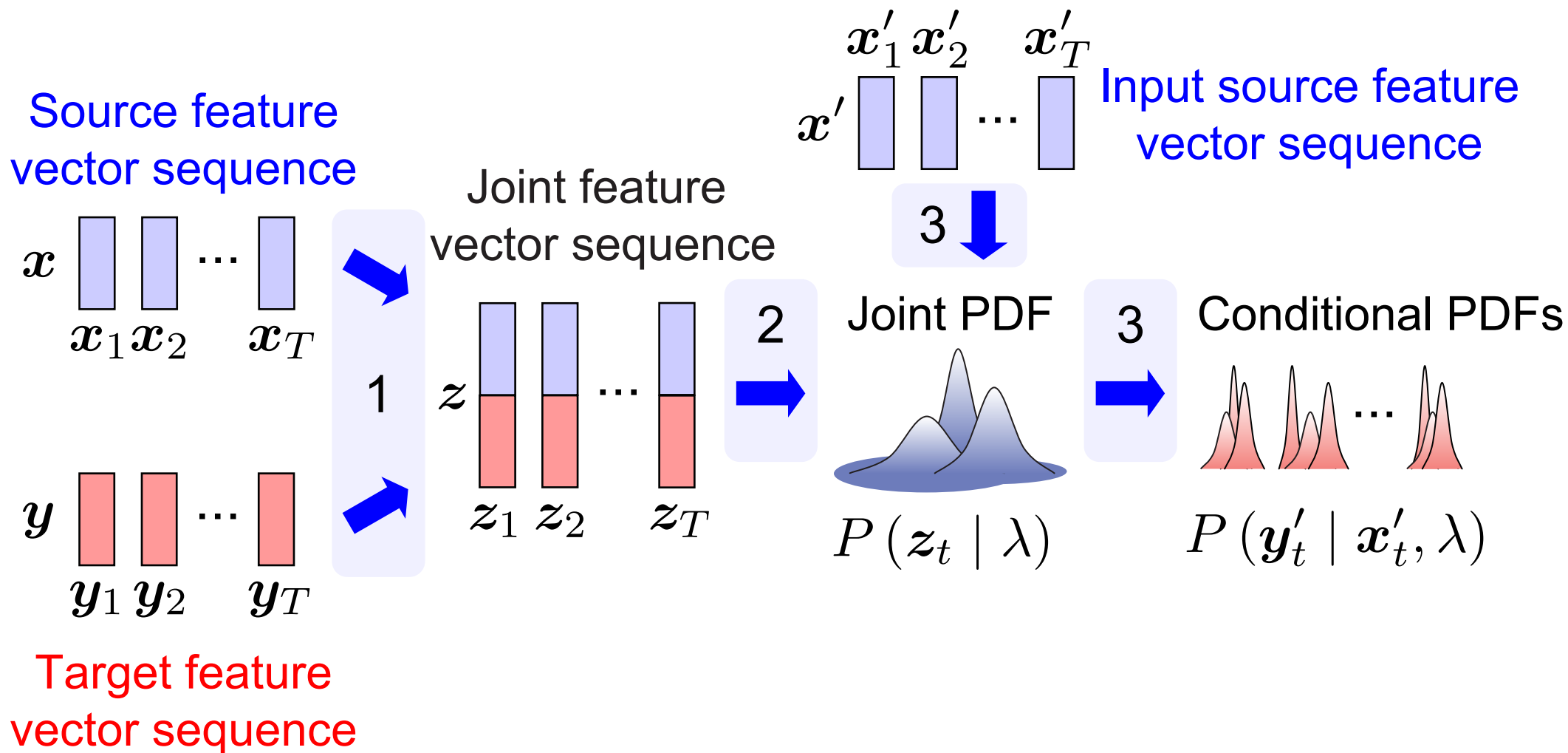


Joint PDF



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

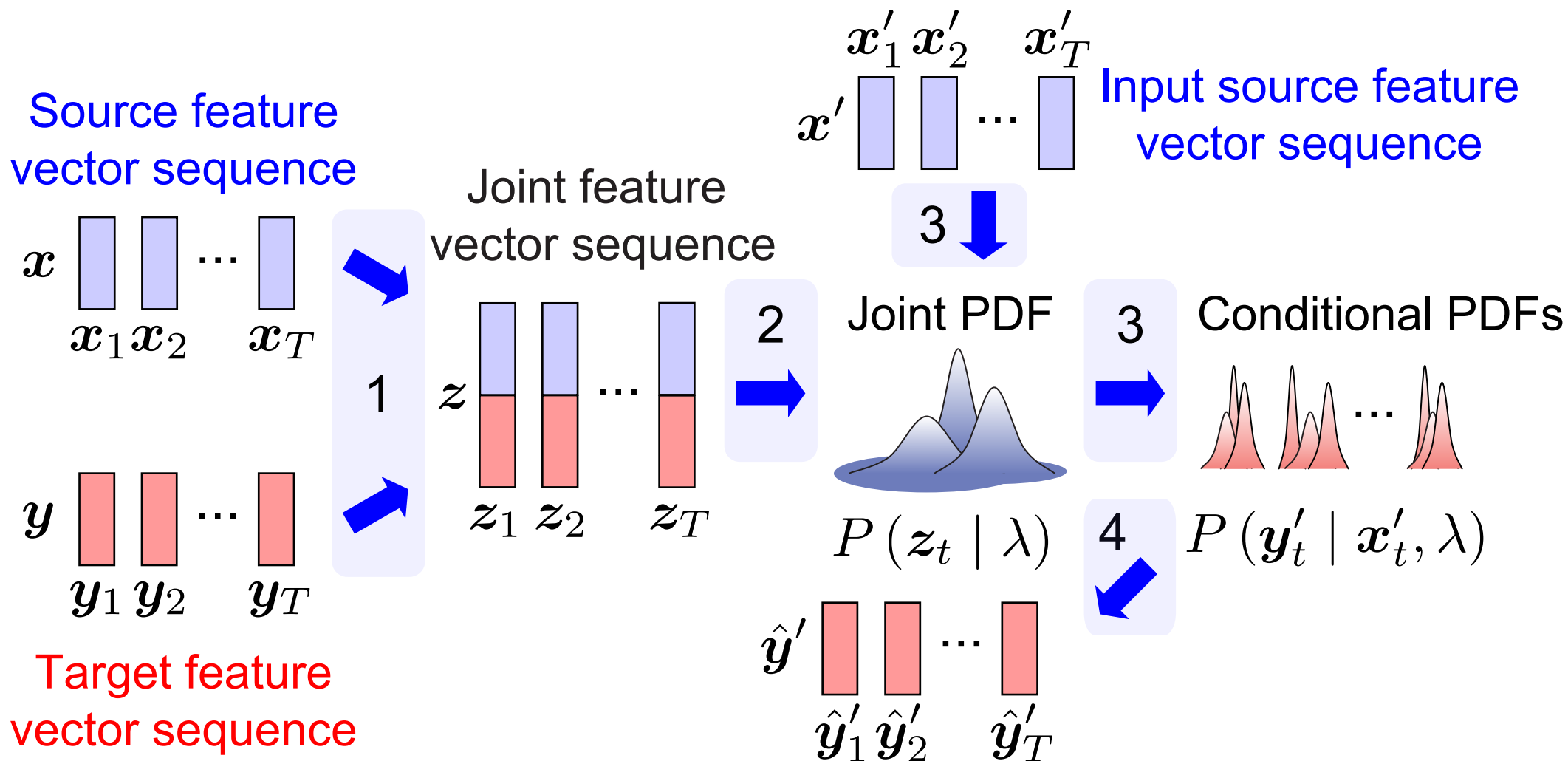
# GMM-based feature mapping [Kain;'98]



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



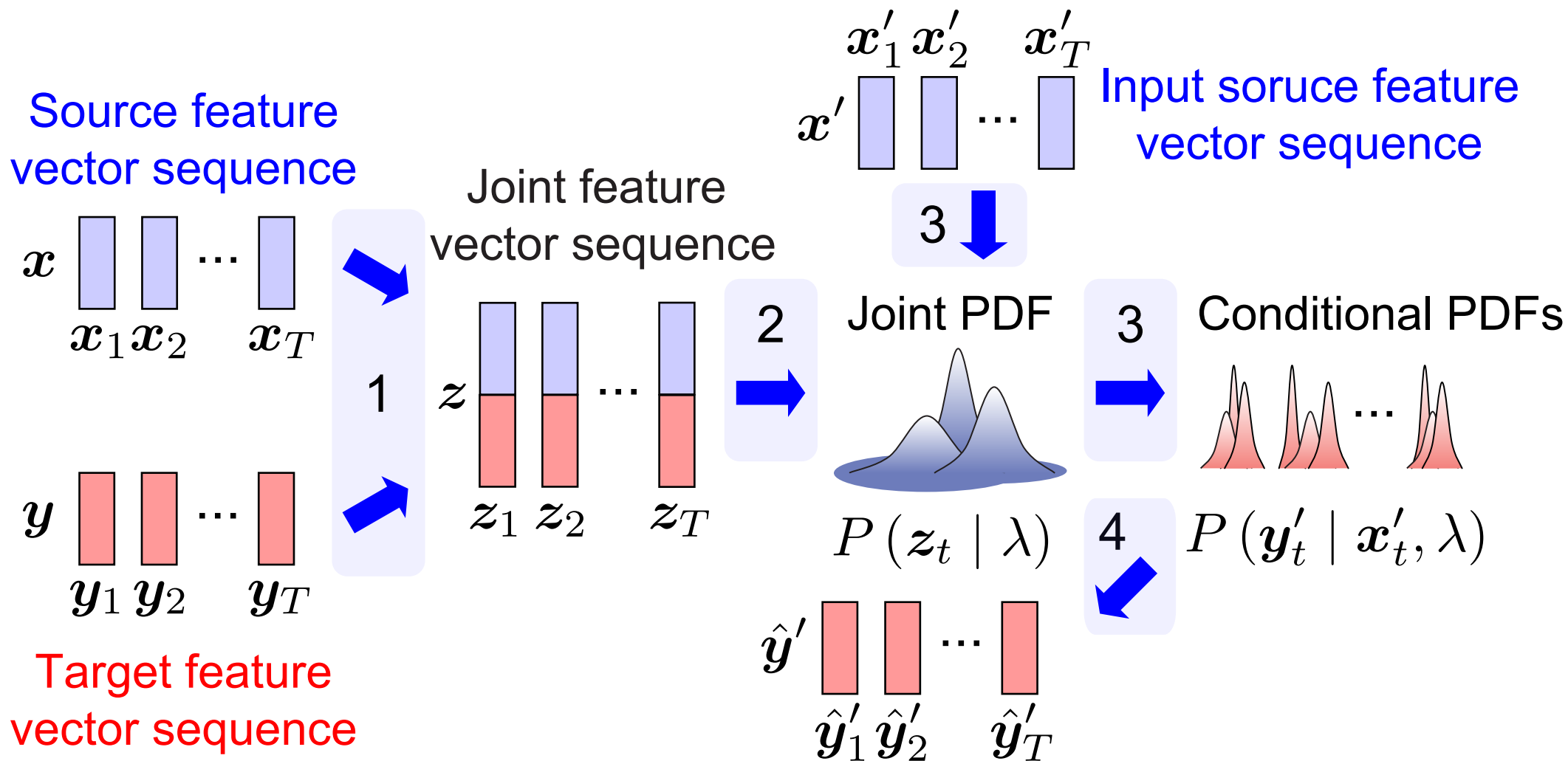
# GMM-based feature mapping [Kain;'98]



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

$$\hat{\mathbf{y}}'_t = \sum_{i=1}^N \gamma_i \left[ \mu_i^{(y)} + \Sigma_i^{(yx)} \Sigma_i^{(xx)^{-1}} \left( \mathbf{x}'_t - \mu_i^{(x)} \right) \right]$$

# GMM-based feature mapping [Kain;'98]

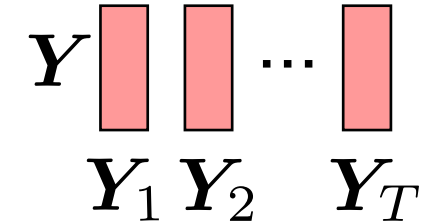
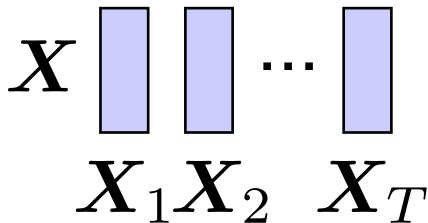


Frame-by-frame mapping

⇒ Mapped features are sometimes discontinuous

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

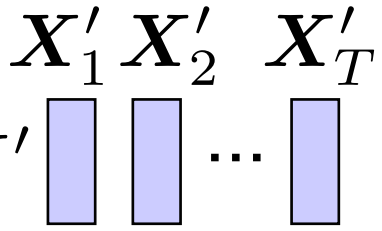
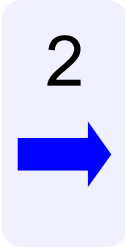
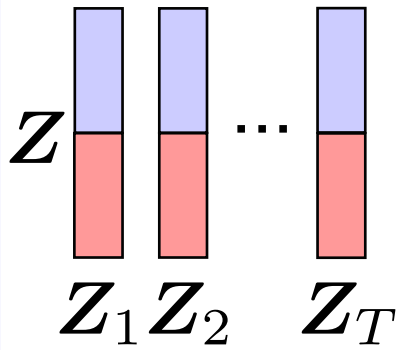
$$\mathbf{X}_t = [\mathbf{x}_t^\top, \Delta \mathbf{x}_t^\top]^\top$$



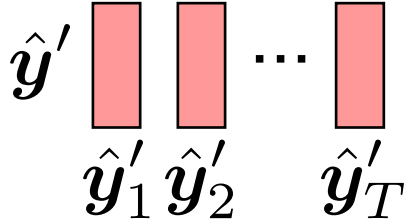
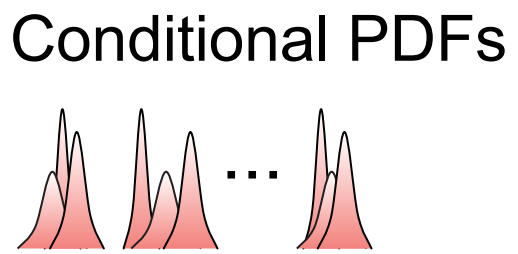
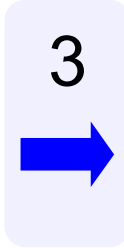
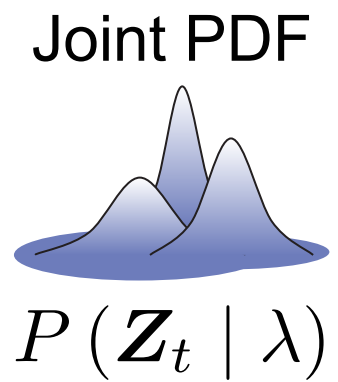
$$\mathbf{Y}_t = [\mathbf{y}_t^\top, \Delta \mathbf{y}_t^\top]^\top$$



Joint feature vector sequence



Input source feature vector sequence



$$P(\mathbf{Y}'_t | \mathbf{X}'_t, \lambda)$$

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

$$\mathbf{X}_t = [\mathbf{x}_t^\top, \Delta \mathbf{x}_t^\top]^\top$$

$$\mathbf{X} \begin{array}{c} \color{blue}{\boxed{\phantom{X_1}}} \\ \color{blue}{\boxed{\phantom{X_2}}} \\ \dots \\ \color{blue}{\boxed{\phantom{X_T}}} \end{array}$$

$\mathbf{X}_1 \mathbf{X}_2 \dots \mathbf{X}_T$

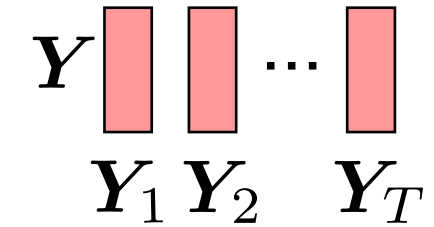
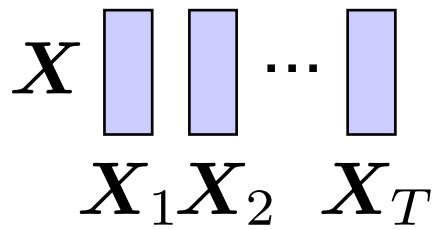
$$\mathbf{Y} \begin{array}{c} \color{red}{\boxed{\phantom{Y_1}}} \\ \color{red}{\boxed{\phantom{Y_2}}} \\ \dots \\ \color{red}{\boxed{\phantom{Y_T}}} \end{array}$$

$\mathbf{Y}_1 \mathbf{Y}_2 \dots \mathbf{Y}_T$

$$\mathbf{Y}_t = [\mathbf{y}_t^\top, \Delta \mathbf{y}_t^\top]^\top$$

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

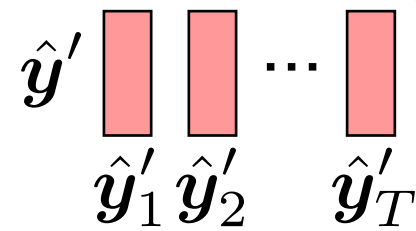
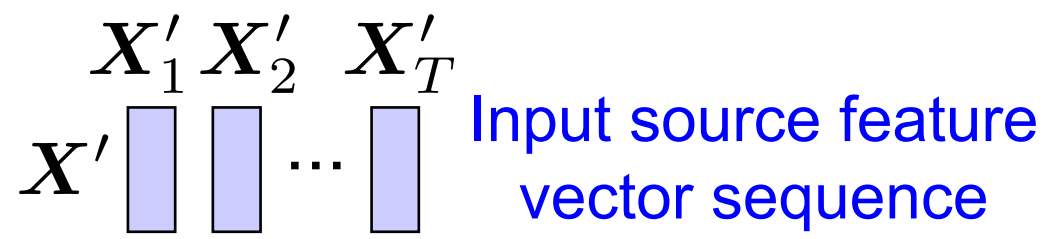
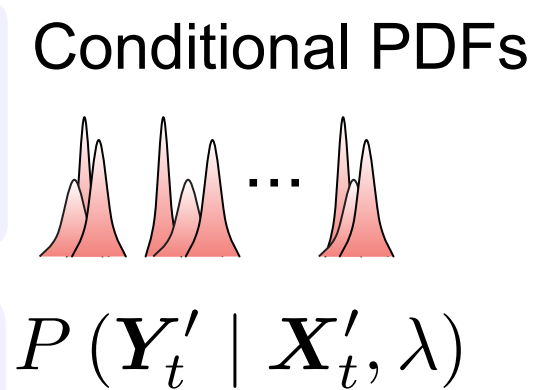
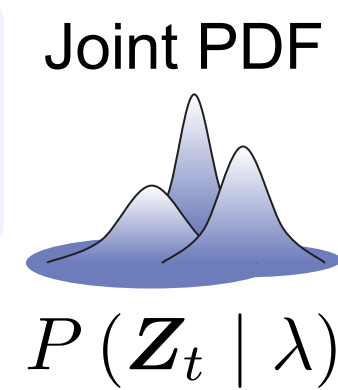
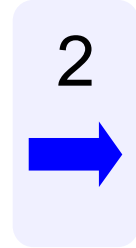
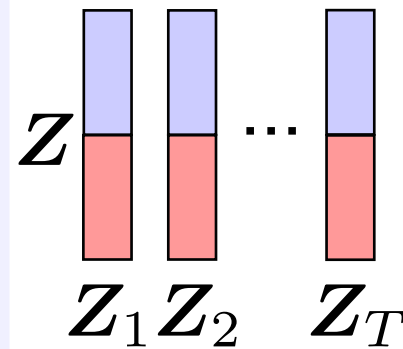
$$\mathbf{X}_t = [\mathbf{x}_t^\top, \Delta \mathbf{x}_t^\top]^\top$$



$$\mathbf{Y}_t = [\mathbf{y}_t^\top, \Delta \mathbf{y}_t^\top]^\top$$



Joint feature vector sequence



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

# Mapping with dynamic-feature constraints

$$\hat{\mathbf{y}}' = \arg \max_{\mathbf{Y}'} P(\mathbf{Y}' | \mathbf{X}', \lambda) = \arg \max_{\mathbf{y}'} P(\mathbf{W}\mathbf{y}' | \mathbf{X}', \lambda)$$

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

$$\mathbf{E}_q^{(\mathbf{y})} = \left[ \mathbf{E}_{tq_1}^{(\mathbf{y})\top}, \mathbf{E}_{tq_2}^{(\mathbf{y})\top}, \dots, \mathbf{E}_{tq_T}^{(\mathbf{y})\top} \right]^\top \quad \mathbf{q} = \{q_1, q_2, \dots, q_T\}$$

$$\mathbf{D}_q^{(\mathbf{y})} = \text{diag} \left[ \mathbf{D}_{q_1}^{(\mathbf{y})}, \mathbf{D}_{q_2}^{(\mathbf{y})}, \dots, \mathbf{D}_{q_T}^{(\mathbf{y})} \right]$$

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

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

- **Entire utterance-level transformation**

  - ⇒ Mapped feats. have proper static & dynamic characteristics

- **Dynamic-feature constraints are used at mapping only**

  - ⇒ Inconsistencies between training & mapping

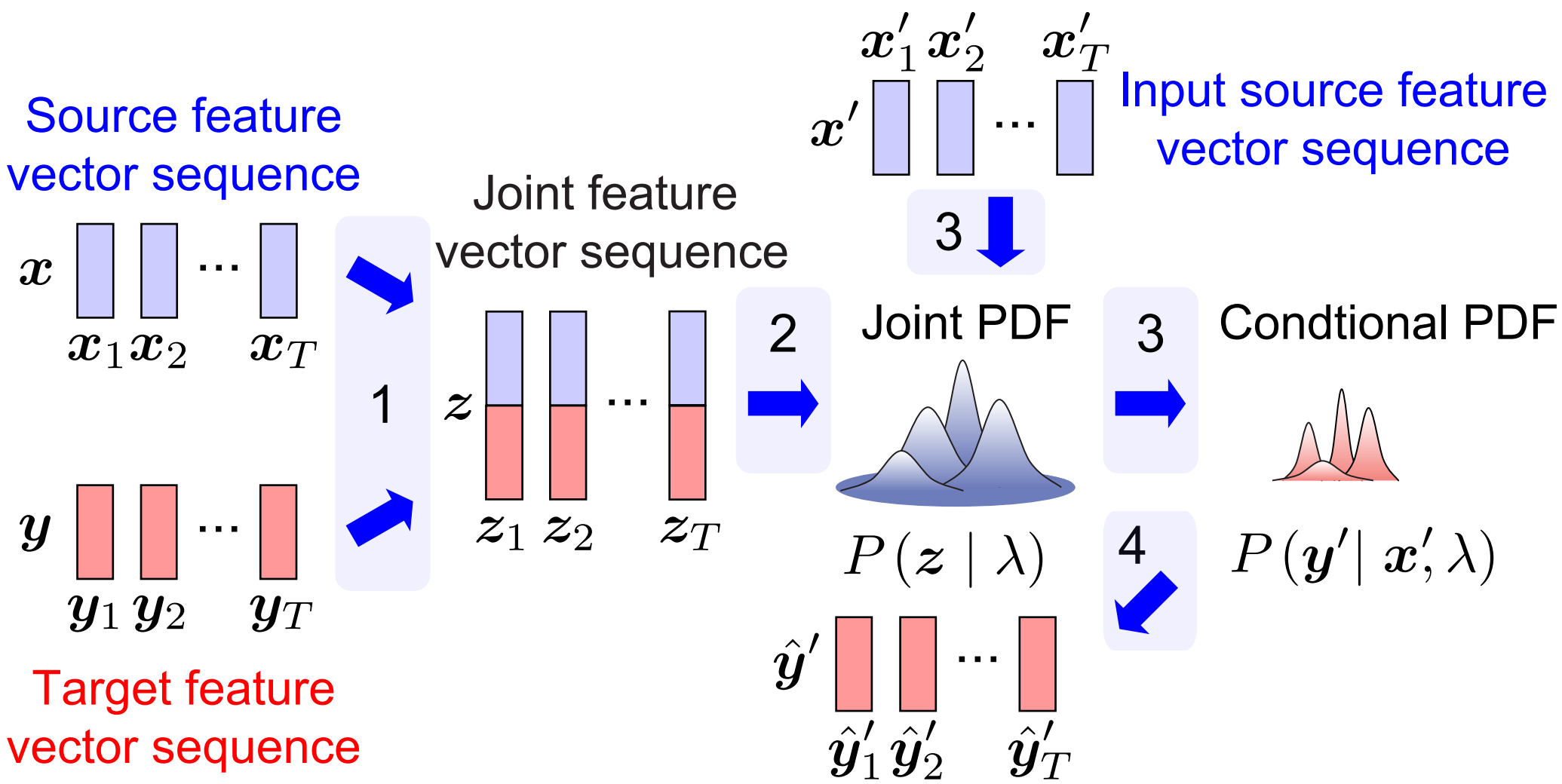
# Trajectory HMM [Zen;'07]

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

$$\begin{aligned} P(\mathbf{z} | \lambda) &= \sum_{\forall \mathbf{q}} P(\mathbf{q} | \lambda) P(\mathbf{z} | \mathbf{q}, \lambda) & \boldsymbol{\mu}_{\mathbf{q}} &= [\boldsymbol{\mu}_{q_1}^\top, \dots, \boldsymbol{\mu}_{q_T}^\top]^\top \\ P(\mathbf{z} | \mathbf{q}, \lambda) &= \mathcal{N}(\mathbf{z}; \bar{\mathbf{z}}_{\mathbf{q}}, \mathbf{P}_{\mathbf{q}}) & \boldsymbol{\mu}_i &= [\boldsymbol{\mu}_i^{(x)\top}, \boldsymbol{\mu}_i^{(y)\top}]^\top \\ \mathbf{R}_{\mathbf{q}} \bar{\mathbf{z}}_{\mathbf{q}} &= \mathbf{r}_{\mathbf{q}} & \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} &= \text{diag}[\boldsymbol{\Sigma}_{q_1}^{-1}, \dots, \boldsymbol{\Sigma}_{q_T}^{-1}] \\ \mathbf{R}_{\mathbf{q}} &= \mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \mathbf{W} = \mathbf{P}_{\mathbf{q}}^{-1} & \boldsymbol{\Sigma}_i^{-1} &= \begin{bmatrix} \boldsymbol{\Omega}_i^{(xx)} & \boldsymbol{\Omega}_i^{(xy)} \\ \boldsymbol{\Omega}_i^{(yx)} & \boldsymbol{\Omega}_i^{(yy)} \end{bmatrix} \\ \mathbf{r}_{\mathbf{q}} &= \mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \boldsymbol{\mu}_{\mathbf{q}} \end{aligned}$$

- $\bar{\mathbf{z}}_{\mathbf{q}}$  is determined as a smooth trajectory  
⇒ Speech dynamics can be modeled explicitly
- Intra- & inter-frame covariance matrix  $\mathbf{P}_{\mathbf{q}}$  becomes full  
⇒ Capture both intra- & inter-frame dependencies

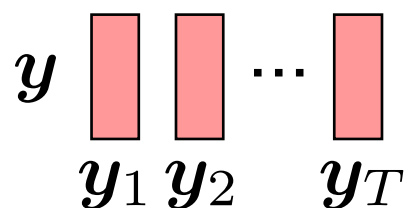
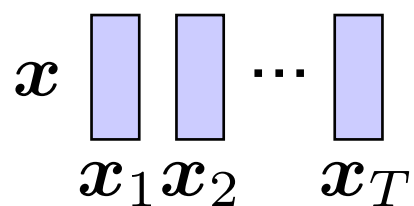
# Trajectory GMM / Trajectory HMM-based mapping





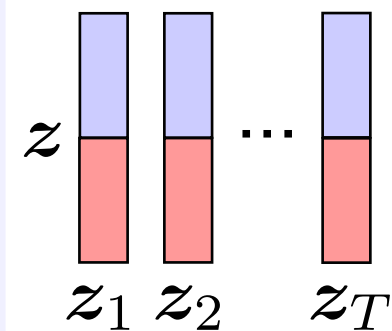
# Trajectory GMM / Trajectory HMM-based mapping

Source feature  
vector sequence



Target feature  
vector sequence

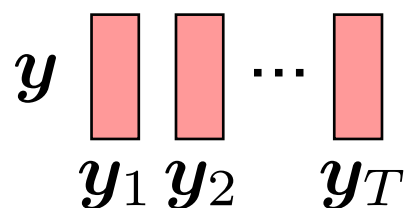
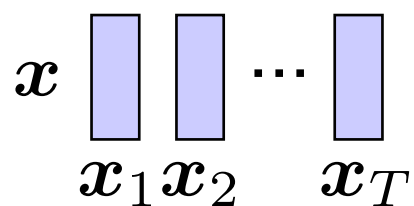
Joint feature  
vector sequence



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

# Trajectory GMM / Trajectory HMM-based mapping

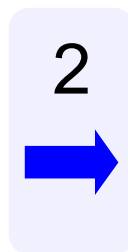
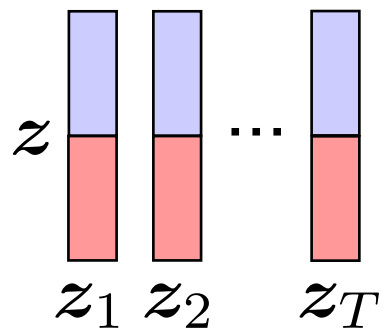
Source feature  
vector sequence



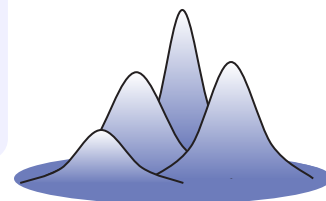
Target feature  
vector sequence



Joint feature  
vector sequence



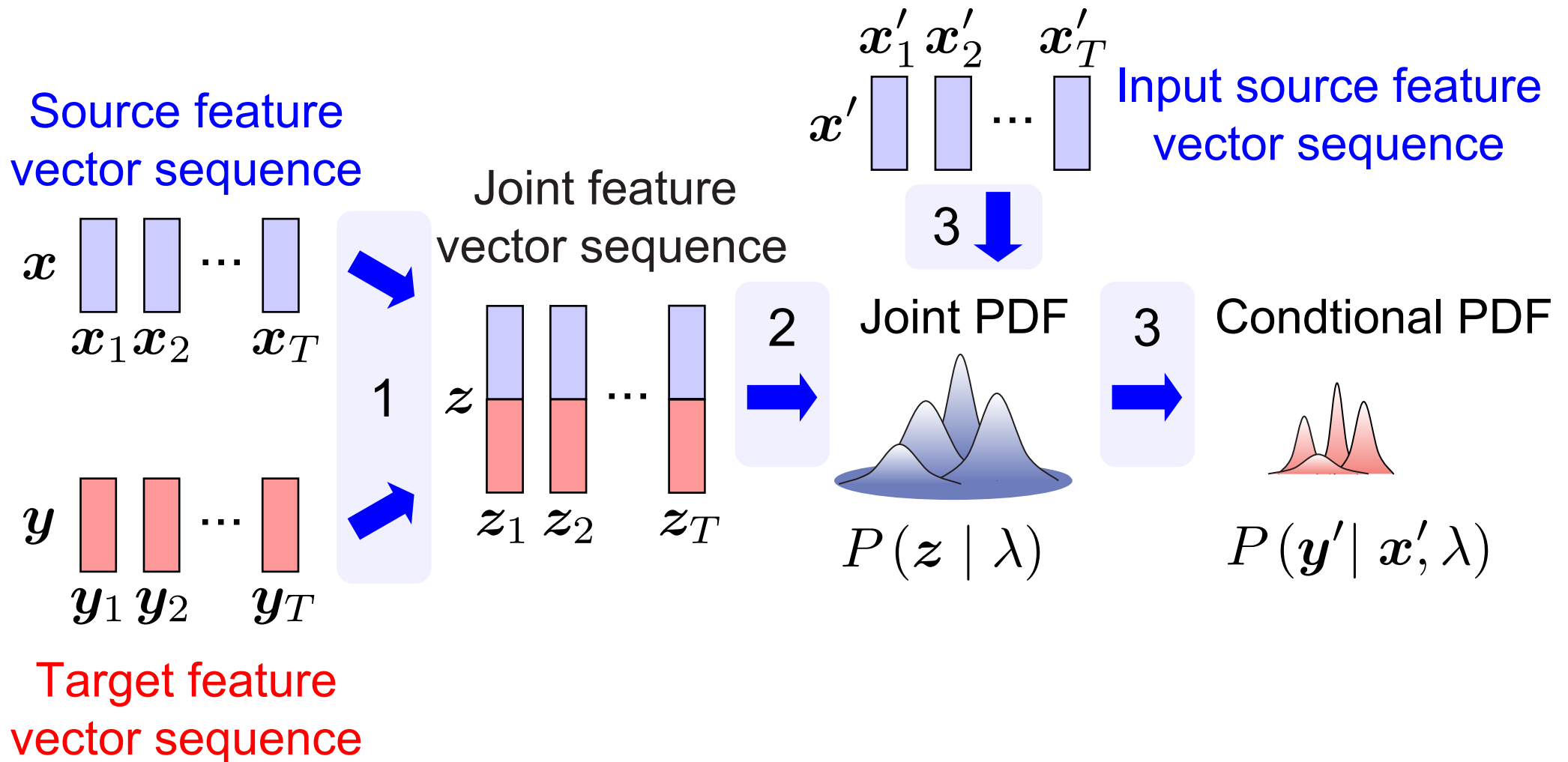
Joint PDF



$$P(z | \lambda)$$

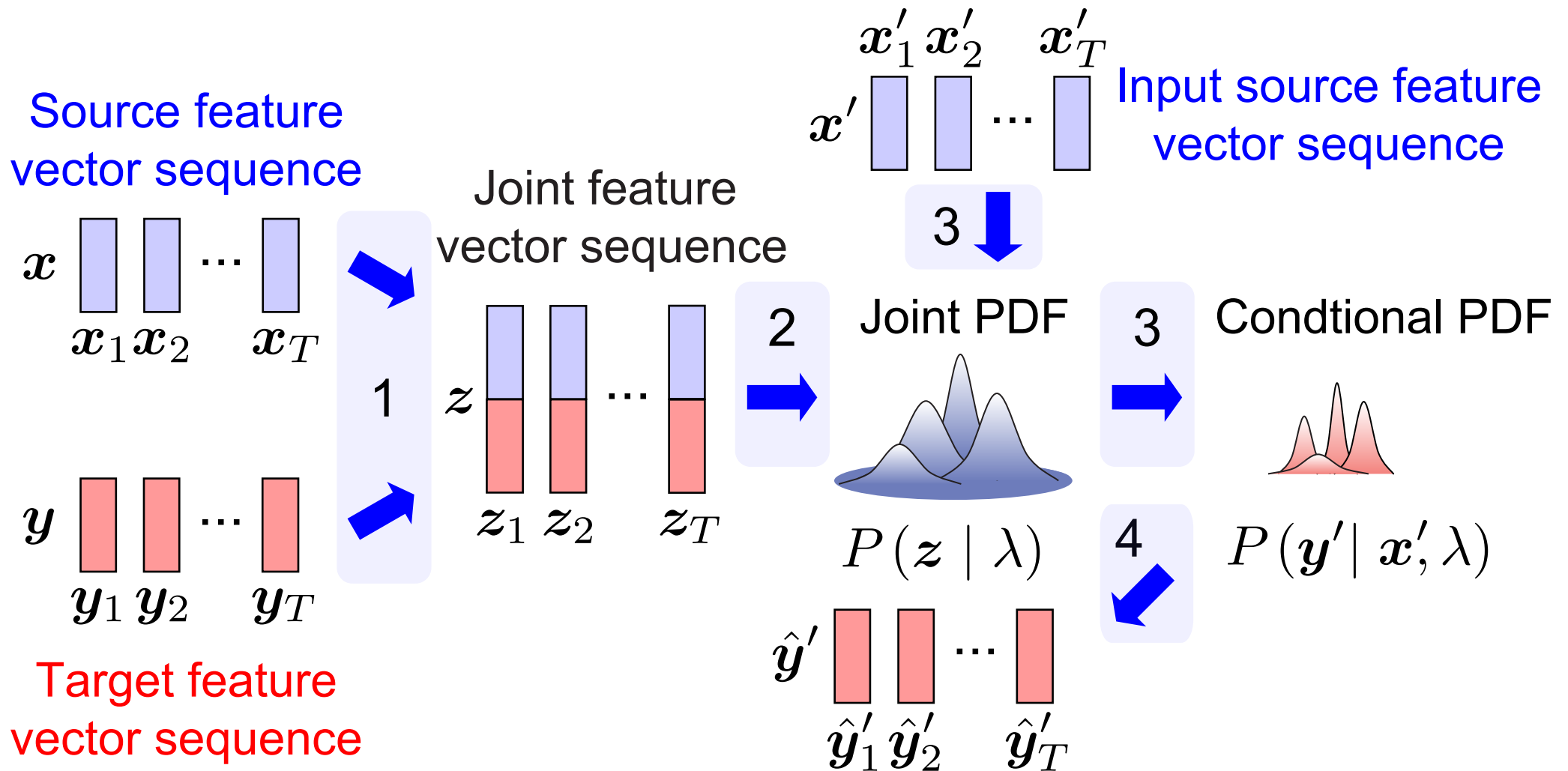
2. Model **utterance-level** joint PDF  $P(z | \lambda)$  by trajectory GMM

# Trajectory GMM / Trajectory HMM-based mapping



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

# Trajectory GMM / Trajectory HMM-based mapping



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

# Trajectory GMM / Trajectory HMM-based mapping

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

$$\hat{\mathbf{y}}'_{\text{MAP}} = \left( \sum_{\forall \mathbf{q}} \gamma'_{\mathbf{q}} \cdot \tilde{\mathbf{P}}_{\mathbf{q}}^{(yy)^{-1}} \right)^{-1}$$

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

- 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**

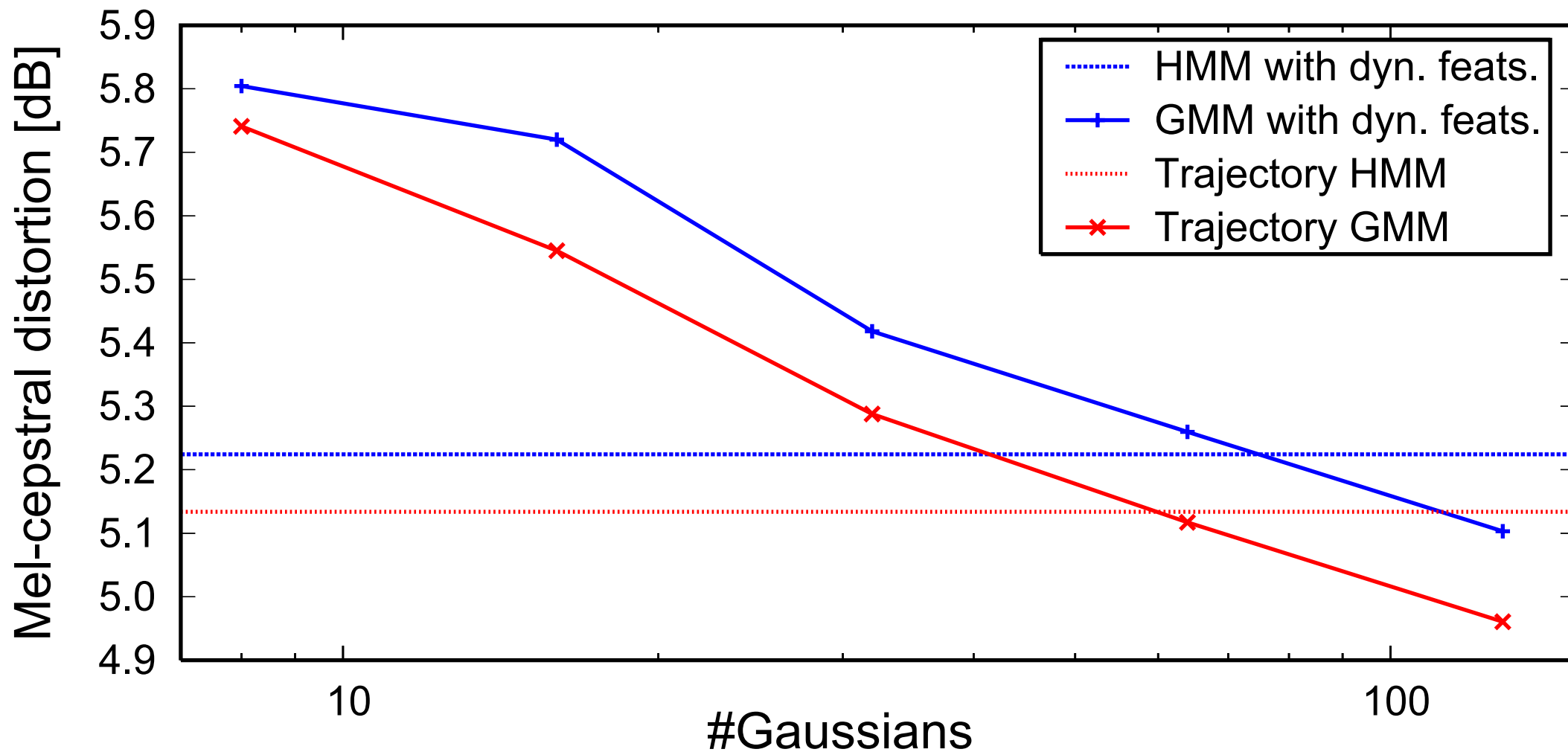
# Experiment (speaker conversion)

## 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, $\Delta$ & $\Delta\Delta$
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)

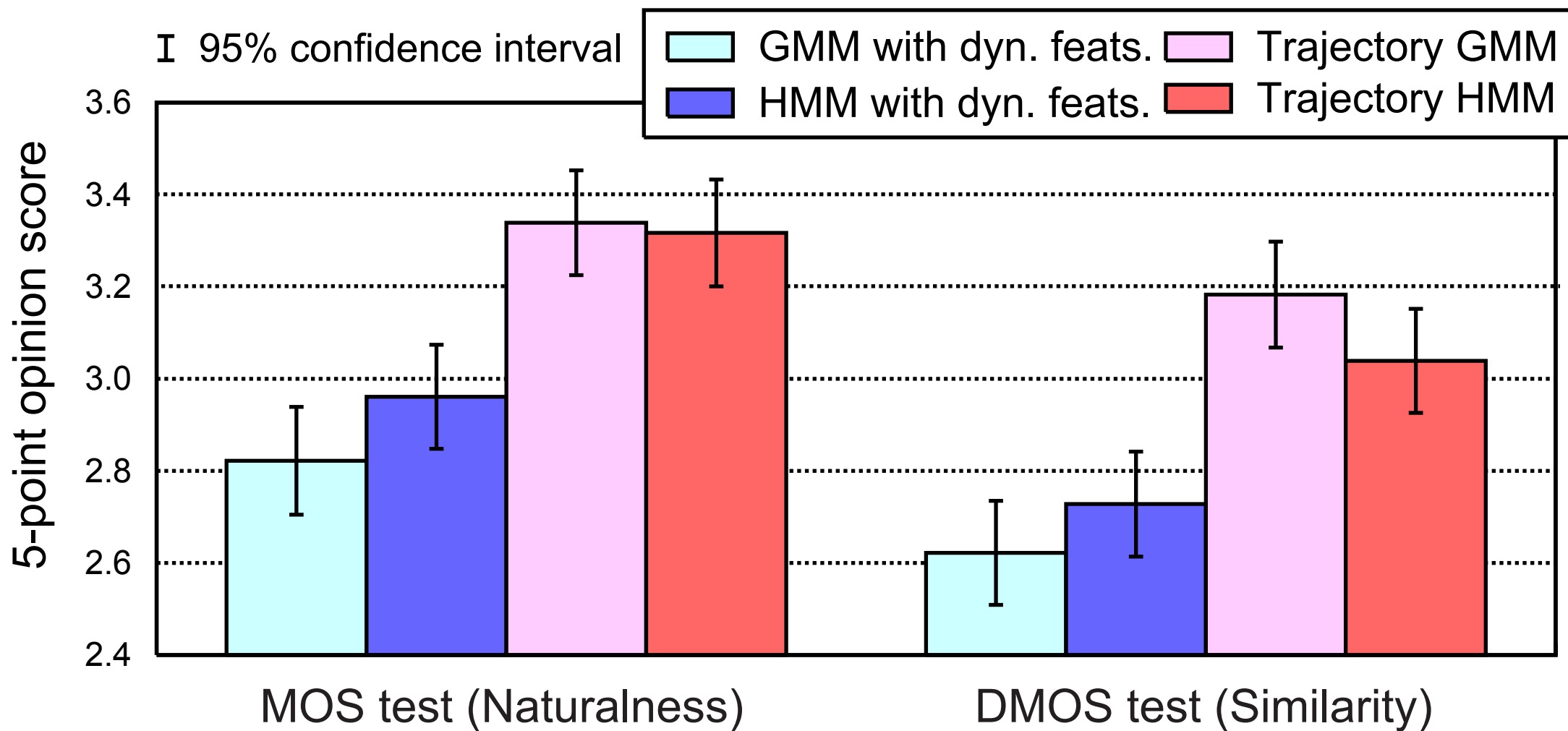
## Objective evaluation (mel-cepstral distortion)





# Experiment (speaker conversion)

## Subjective evaluation (MOS & DMOS)



(#subjects=12, 15 sentences/subjects)

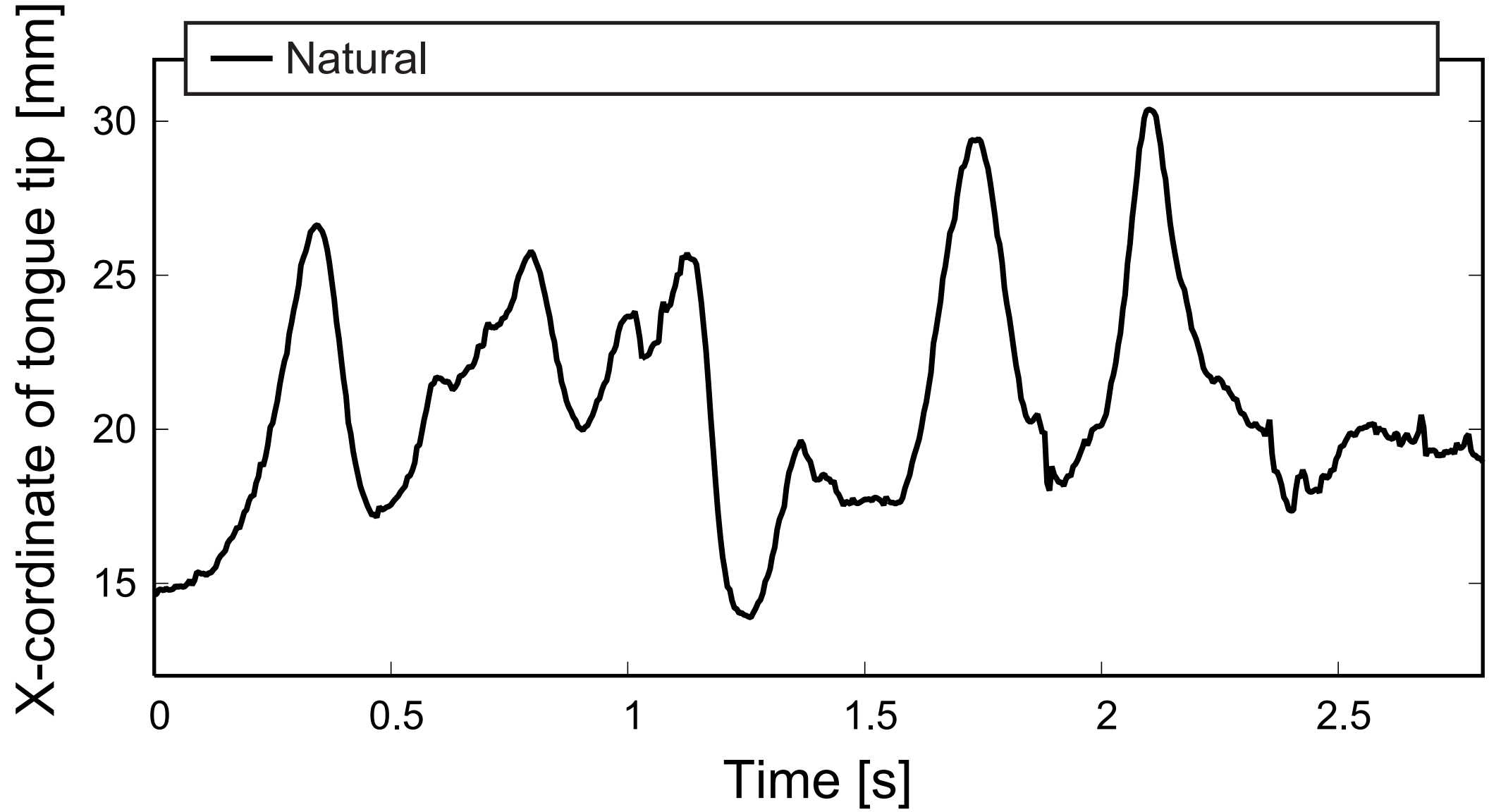
# Experiment (acoustic-articulatory inv. mapping)

## 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, $\Delta$ & $\Delta\Delta$ 7 articulators in x & y coordinates, $\Delta$ & $\Delta\Delta$
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

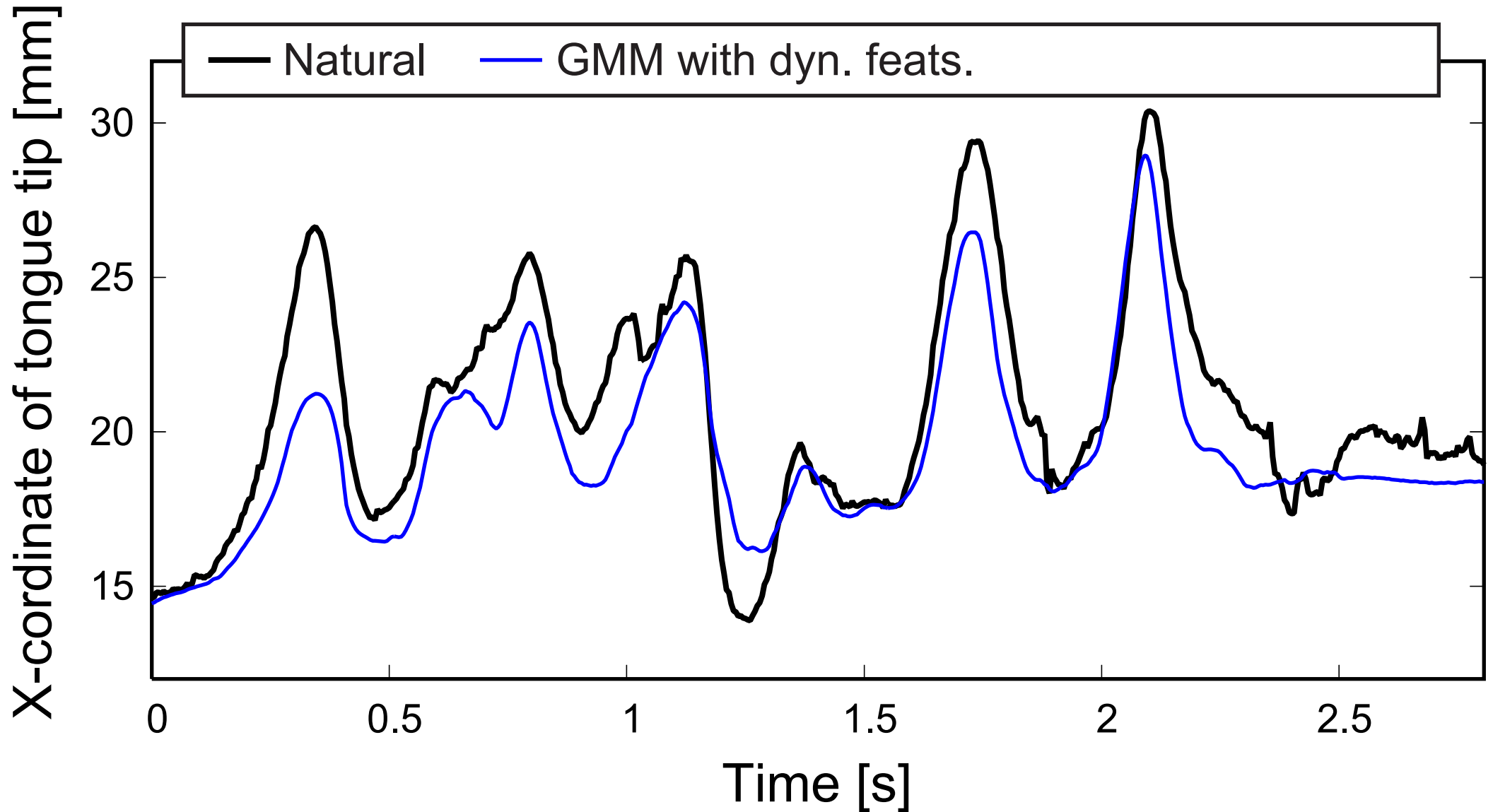
# Experiment (acoustic-articulatory inv. mapping)

## Examples of estimated articulatory movements



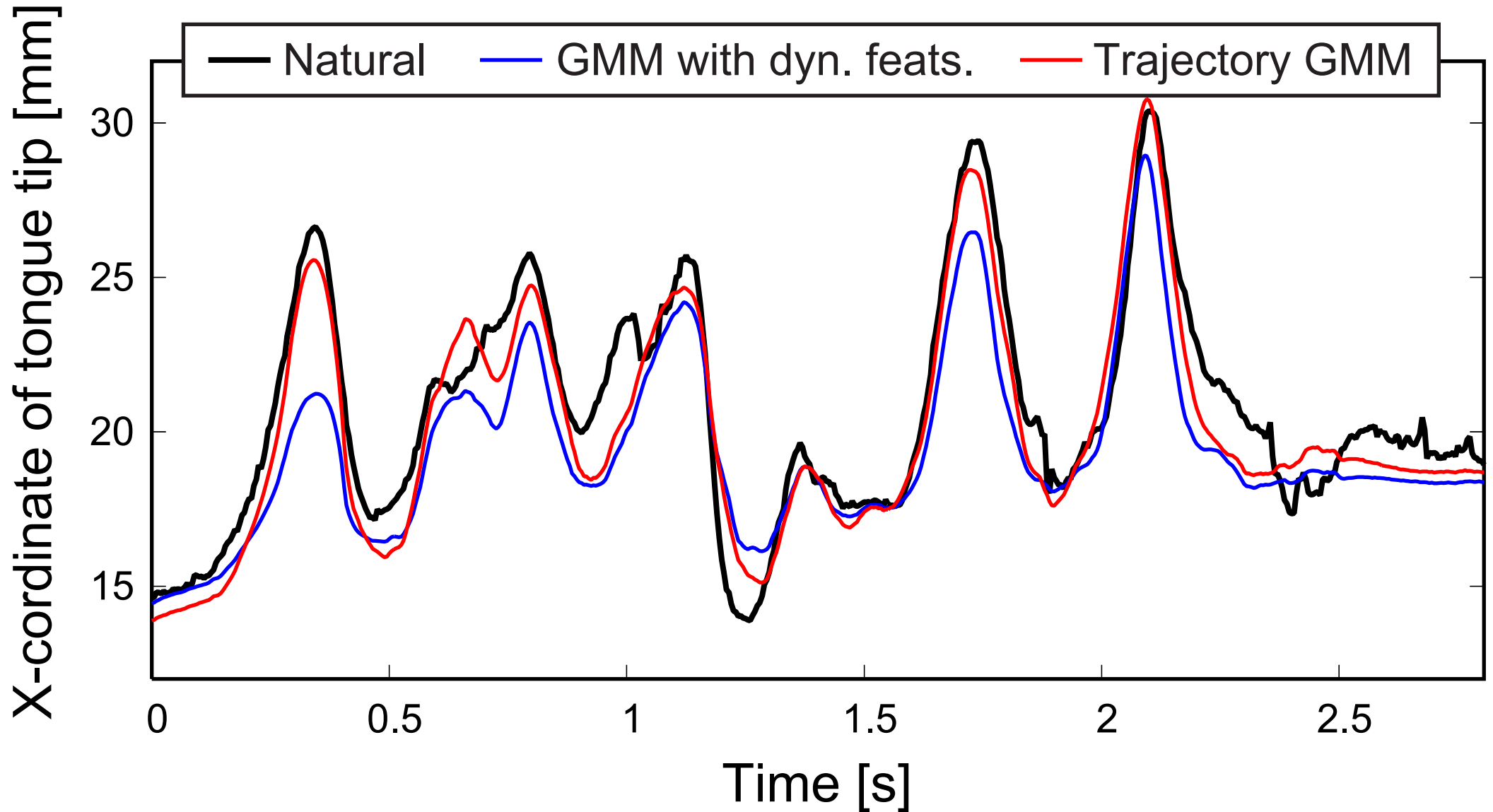
# Experiment (acoustic-articulatory inv. mapping)

## 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	<b>1.071</b>	<b>1.128</b>

# 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, $\Delta$ & $\Delta\Delta$
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, $\Delta$ & $\Delta\Delta$
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	<b>82.80</b>	77.92	83.92
GMM-dynamic	88.26	81.03	79.31	83.58
Trajectory GMM	<b>89.38</b>	81.54	<b>80.87</b>	<b>84.54</b>

# 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