Probabilistic Feature Mapping Based on Trajectory HMMs

Heiga Zen Yoshihiko Nankaku Keiichi Tokuda (Nagoya Institute of Technology)

1. Introduction

GMM / HMM-based feature mapping

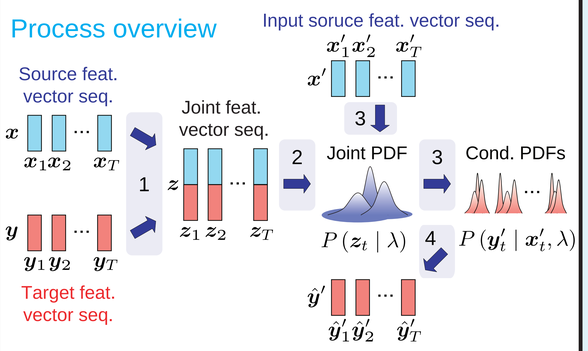
- Continuously transform src. features into tgt. ones
  - Model joint PDF betw. src. & tgt. by GMM or HMMs
  - Map src. feats. into tgt. ones via conditional PDFs
- Applications
  - Voice conversion [Stylianou;'98, Kain;'98]
  - Acoustic-articulatory inv. mapping [Shiga;'04, Toda;'08]
  - Noise compensation [Droppo;'02, Cui;'07]
- Problem
  - Frame-by-frame mapping
    ⇒ Inappropriate dynamic characteristics
  - Mapping with dynamic feature constraints [Toda;'07]
    ⇒ Improper in the sense of statistical modeling

Trajectory GMM / HMM-based feature mapping

- Trajectory HMM [Zen;'07]
  - Impose explicit relationships betw. static & dyn. feats.
    ⇒ HMM is reformulated as a trajectory model
  - Avoid limitations of HMMs w/o additional parameters
    * Frame-wise conditional independence assumption
    * Piece-wise constatnt statistics within an HMM state
- Feature mapping based on trajectory HMMs / GMMs
  - Entire utterance-level mapping
    ⇒ Appropriate static & dynamic characteristics
  - Using dyn. feat. constraints in both training & mapping
    ⇒ Make training & mapping consistent

2. GMM-based mapping [Kain;'98]

Process overview



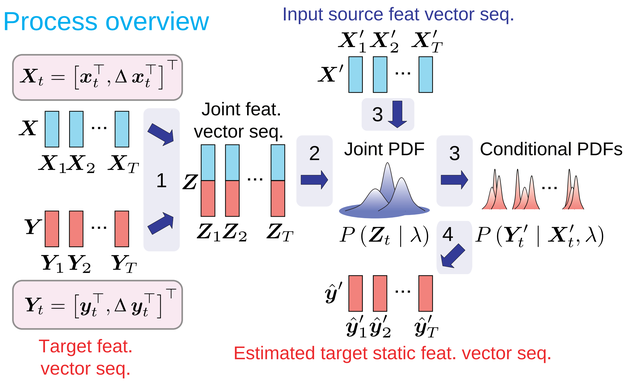
1. Make joint feat. vector $\boldsymbol{z}_t$ from src. & tgt. vectors $\boldsymbol{x}_t$ & $\boldsymbol{y}_t$
2. Model frame-level joint PDF $P(\boldsymbol{z}_t \mid \lambda)$ by GMM
3. Convert joint PDF $P(\boldsymbol{z}_t \mid \lambda)$ to cond. PDF $P(\boldsymbol{y}'_t \mid \boldsymbol{x}'_t, \lambda)$
4. Estimate $\hat{\boldsymbol{y}}'_t$ from conditional PDF by MMSE

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

Each frame is transformed independently
⇒ Mapped features are sometimes discontinuous

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

Process overview



1〜3 are the same as those of mapping w/o dynamic features

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

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

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

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

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

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

$$\boldsymbol{D}_i^{(\boldsymbol{y})} = \boldsymbol{\Sigma}_i^{(\boldsymbol{yy})} - \boldsymbol{\Sigma}_i^{(\boldsymbol{yx})} \boldsymbol{\Sigma}_i^{(\boldsymbol{xx})^{-1}} \boldsymbol{\Sigma}_i^{(\boldsymbol{xy})}$$

- Entire utterance-level mapping
  ⇒ Mapped feats. has proper static & dyn. characteristics
- Dynamic feature constraints are used only on mapping
  ⇒ Inconsistency between training & mapping

4. Trajectory HMM / Trajectory GMM

Trajectory HMM [Zen;'07]

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

$$P(\boldsymbol{z} \mid \lambda) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{q} \mid \lambda) P(\boldsymbol{z} \mid \boldsymbol{q}, \lambda) \qquad \boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_T}^\top \right]^\top$$

$$P(\boldsymbol{z} \mid \boldsymbol{q}, \lambda) = \mathcal{N}\left(\boldsymbol{z} ; \bar{\boldsymbol{z}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}} \right) \qquad \boldsymbol{\mu}_i = \left[\boldsymbol{\mu}_i^{(\boldsymbol{x})^\top}, \boldsymbol{\mu}_i^{(\boldsymbol{y})^\top} \right]^\top$$

$$\boldsymbol{R}_{\boldsymbol{q}} \bar{\boldsymbol{z}}_{\boldsymbol{q}} = \boldsymbol{r}_{\boldsymbol{q}} \qquad \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{q_1}^{-1}, \ldots, \boldsymbol{\Sigma}_{q_T}^{-1} \right]$$

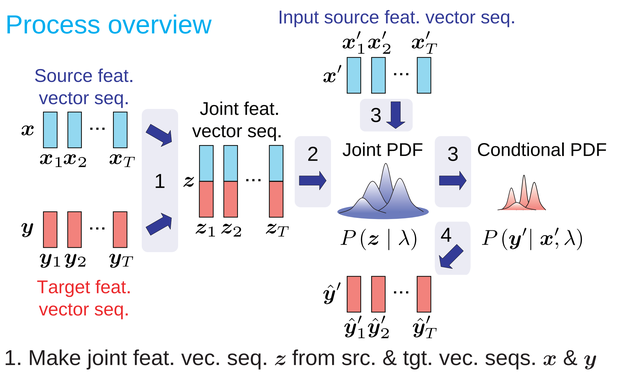
$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1} \qquad \boldsymbol{\Sigma}_i^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_i^{(\boldsymbol{xx})} & \boldsymbol{\Omega}_i^{(\boldsymbol{xy})} \\ \boldsymbol{\Omega}_i^{(\boldsymbol{yx})} & \boldsymbol{\Omega}_i^{(\boldsymbol{yy})} \end{bmatrix}$$

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

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

5. Trajectory GMM (HMM)-based mapping

Process overview



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

$$\boldsymbol{z} = \left[\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top, \ldots, \boldsymbol{z}_T^\top \right]^\top \quad \boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top \right]^\top$$

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

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

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

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

$$P(\boldsymbol{z} \mid \boldsymbol{q}, \lambda) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} ; \begin{bmatrix} \bar{\boldsymbol{x}}_{\boldsymbol{q}} \\ \bar{\boldsymbol{y}}_{\boldsymbol{q}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xx})} & \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \\ \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} & \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \end{bmatrix} \right)$$

$$P(\boldsymbol{y}' \mid \boldsymbol{x}', \lambda) = \sum_{\forall \boldsymbol{q}} \gamma_{\boldsymbol{q}} \cdot P(\boldsymbol{y}' \mid \boldsymbol{x}', \boldsymbol{q}, \lambda)$$

$$P(\boldsymbol{y}' \mid \boldsymbol{x}', \boldsymbol{q}, \lambda) = \mathcal{N}\left(\boldsymbol{y}' ; \tilde{\boldsymbol{y}}_{\boldsymbol{q}}, \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \right)$$

$$\tilde{\boldsymbol{y}}_{\boldsymbol{q}} = \bar{\boldsymbol{y}}_{\boldsymbol{q}} + \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \left(\boldsymbol{x}' - \bar{\boldsymbol{x}}_{\boldsymbol{q}} \right)$$

$$\tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} - \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xy})}$$

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

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

   - MAP

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

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

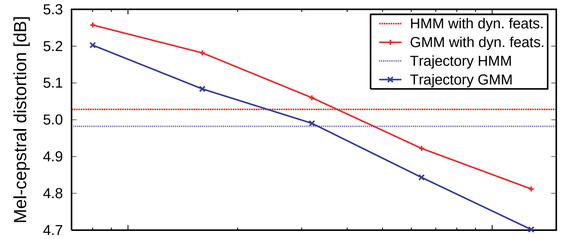
6. Voice conversion experiment

Conditions

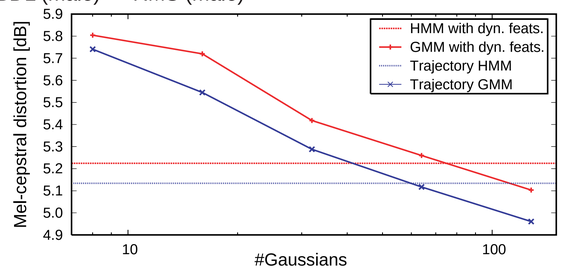
Database	CMU ARCTIC speech database
Training data	Speakers BDL, CLB, RMS, & SLT, first 593 utts. (Mapping: CLB→SLT, BDL→RMS)
Test data	Last 40 utterances
Sampling freq.	16 kHz
Analysis win.	25-ms Blackman window / 5-ms shift
Feature vec.	0〜24 order Mel-cepstral coefficients, Δ & ΔΔ
Topology	(Trajectory) 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

Objective evaluation (mel-cepstral distortions)

CLB (female) → SLT (female)

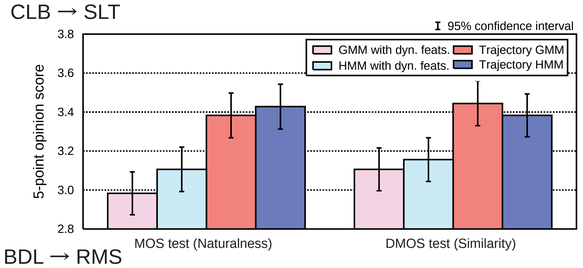


BDL (male) → RMS (male)



Subjective evaluation (MOS & DMOS)

CLB → SLT



BDL → RMS