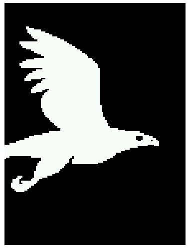
Interval-Based Hybrid Dynamical System for Modeling Dynamic Events and Structures

Hiroaki Kawashima

Event Recognition

Real world



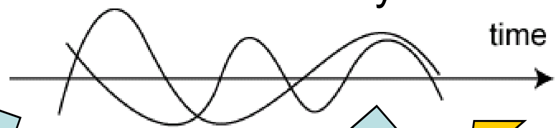
Sensor (ex. camera, microphone)

Time-varying signals

Feature Extraction
(signal processing)

Dynamic Feature

Sequence of Static or Dynamic Feature

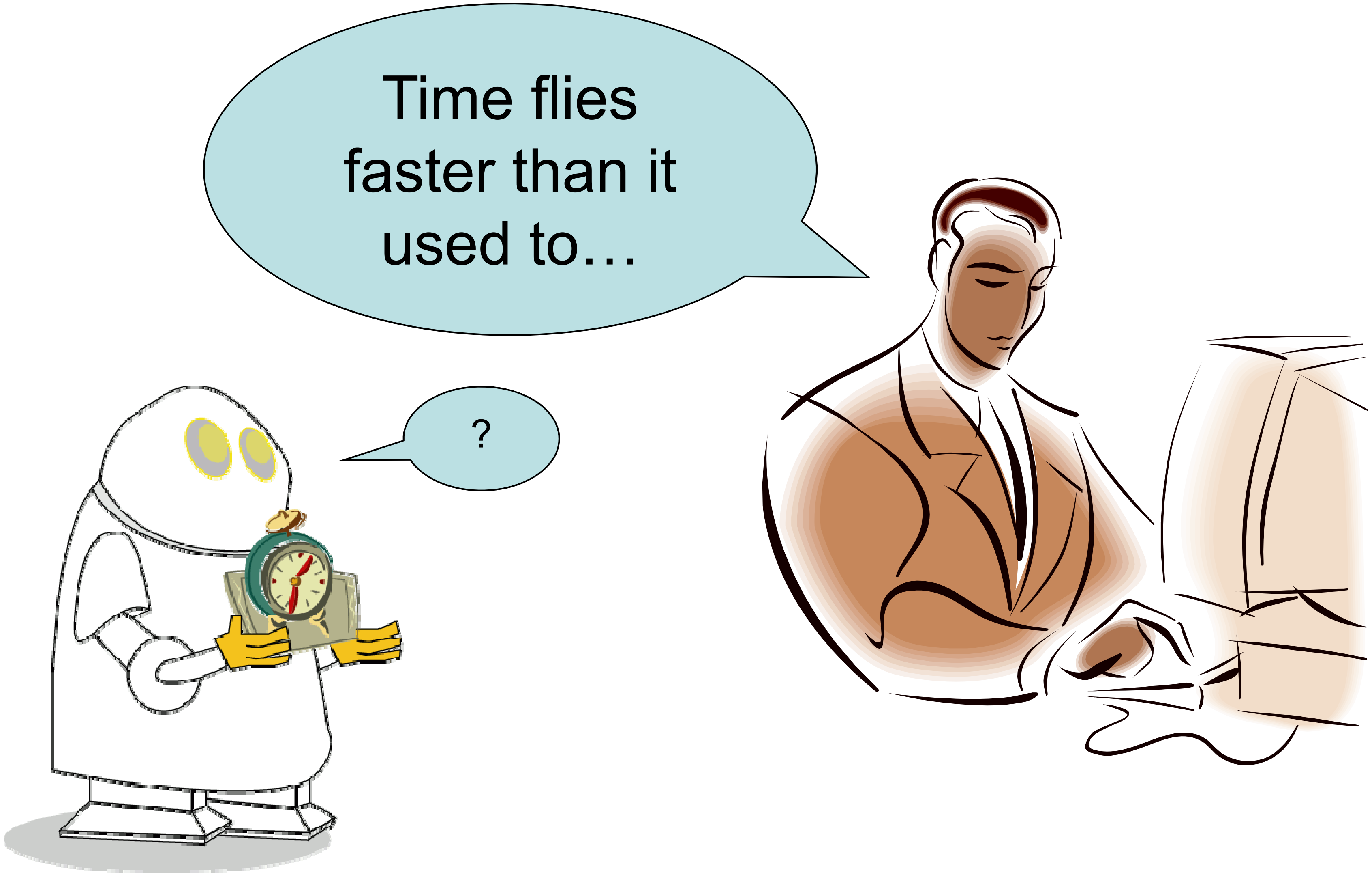


Static Pattern Recognition
(ex. Nearest Neighbor)

Trajectory Matching
(ex. DP matching method)

State-Transition Model
(ex. HMM, Dynamical System, Automaton)

Two Concepts of Time



Subjective time (*Kairos*)

Time has no metric: (N, ≦)

Event sequence

s_i : Countable set N ⟶ Finite set A

A = {up, down}

up→down→up→down ･･･

0 1 2 3 ･･･

Order of discrete events

Q: state set, E: input event set

Discrete-Event Systems

$$M(\text{State}_{now}, \text{input event}) = \text{State}_{next}$$

$M: Q \times A \to Q$

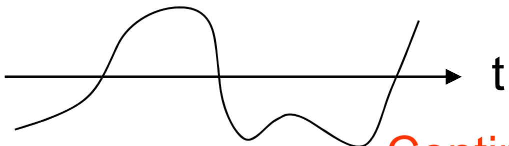
Automata, Petri nets
(Turing machine 1936)

Objective (Physical) time (*Chronos*)

Time has metric : (R, ≦, dist)

Signal data

$y(t)$: Real num. R ⟶ Continuous space R^k



Continuous change

R^n: state space

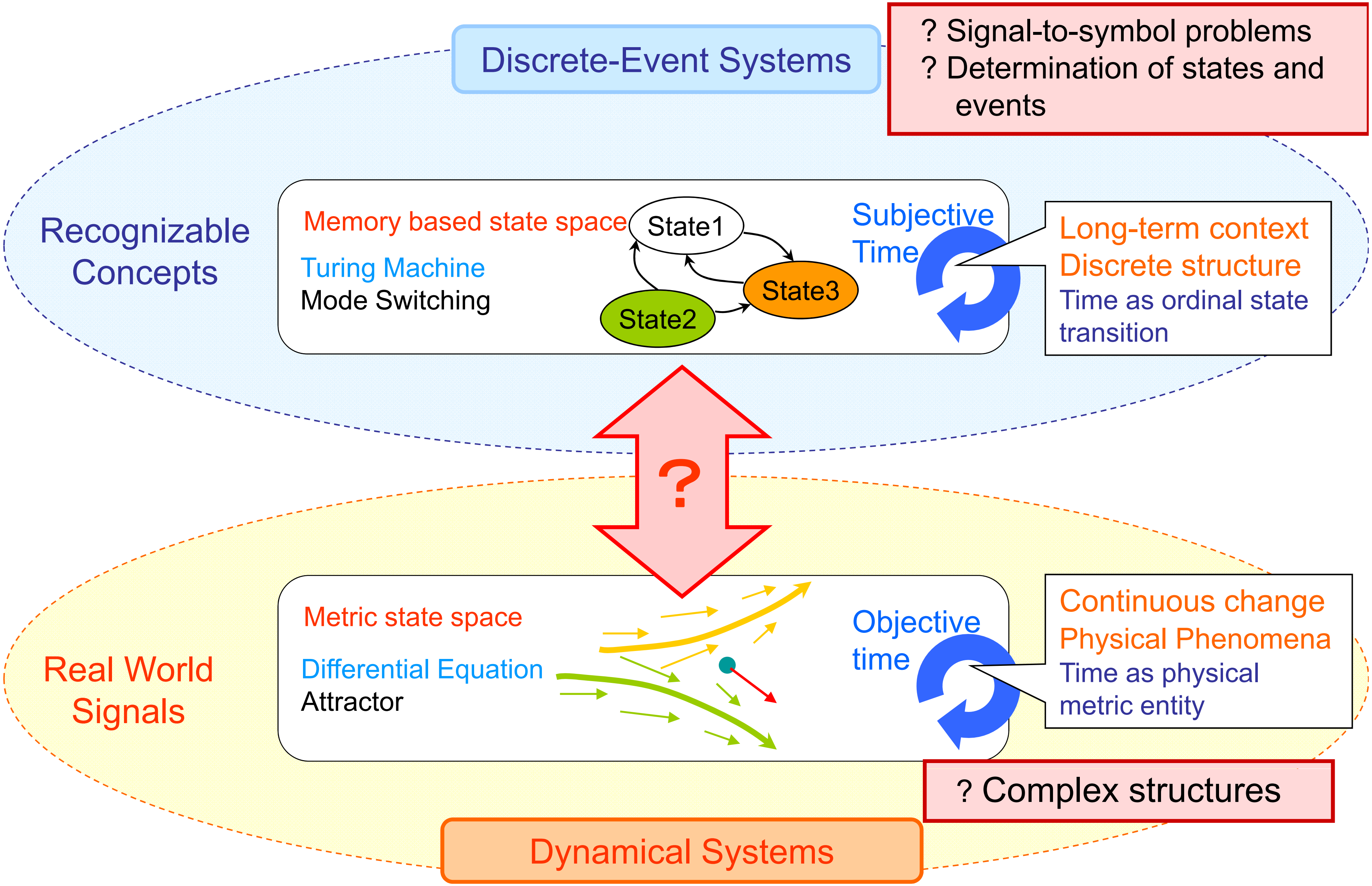
Dynamical Systems

$$\frac{d\ \text{State}(t)}{dt} = F(\text{state}(t))$$

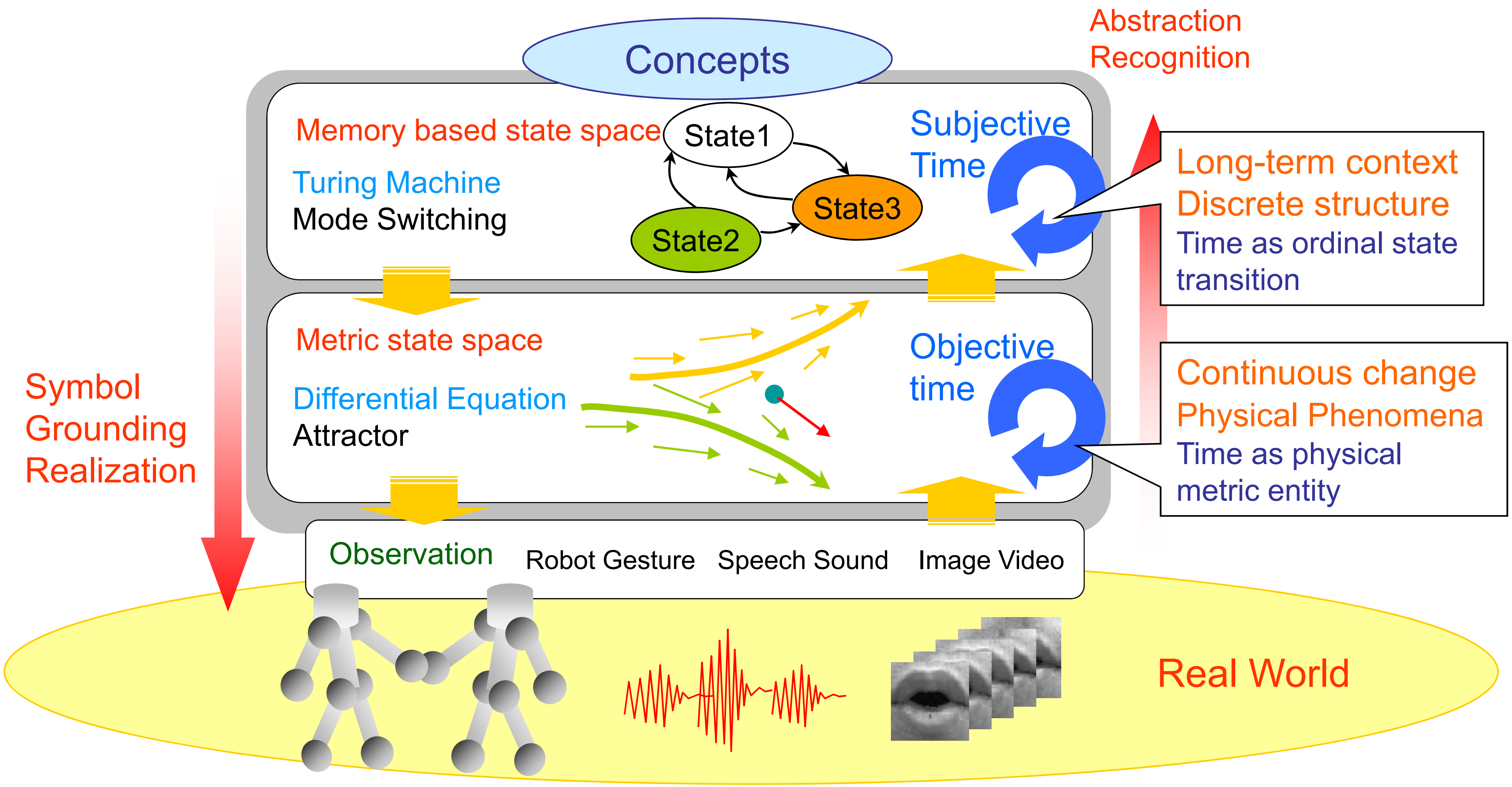
$F: R^n \to R^n$

Control systems, Neural networks
(Cybernetics 1947)

Advantages and Disadvantages



Hybrid Dynamical Systems



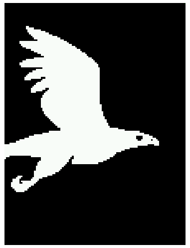
Existing Studies

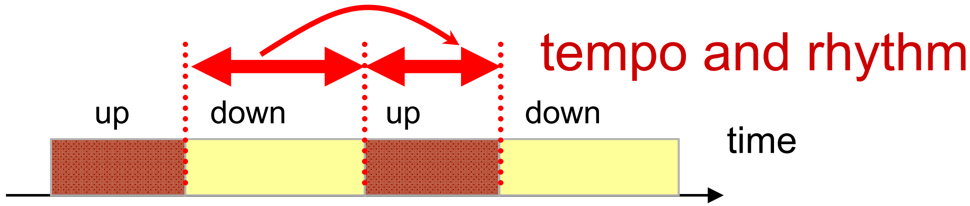
- Computer vision
  - Hybrid dynamical models [C. Bregrler 1997]
  - Multi-class condensation [B. North, A. Blake, M. Isard and J. Rittscher, 2000]
  - Switching linear dynamical systems [K.P. Murphy 1998, V. Pavlovic 1999]
- Speech recognition
  - Segment models [M. Ostendorf 1996]
- Computer graphics
  - Motion textures [Y. Li, T. Wang, H.Y. Shum 2002]
- Neural networks, Control theory, etc.
  - Piecewise linear models [R. Batruni 1991]
  - Switching space models [Z. Ghahramani 1996]
  - Piecewise affine maps [L. Breiman 1993]
  - Hybridautomata [R. Alur 1993]

Integration of “subjective time” and “objective time”?

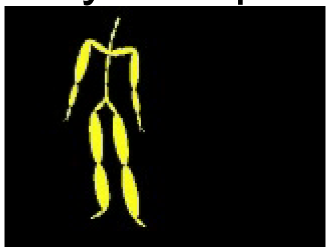
Dynamic Structures of Events

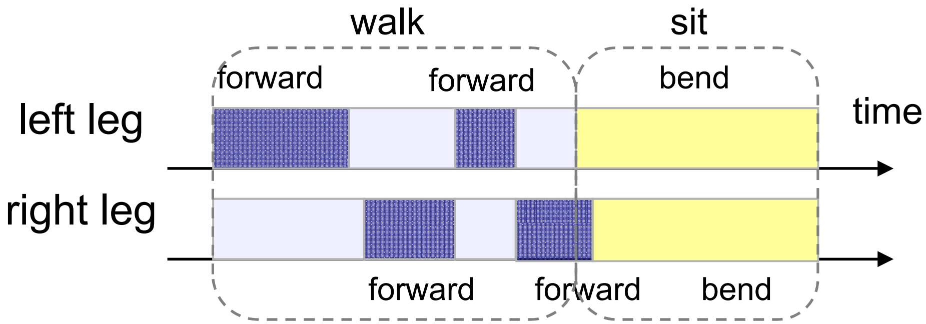
Sequential structures



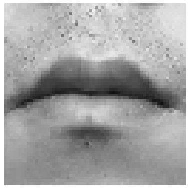


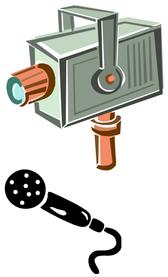
Constituted by multiple objects

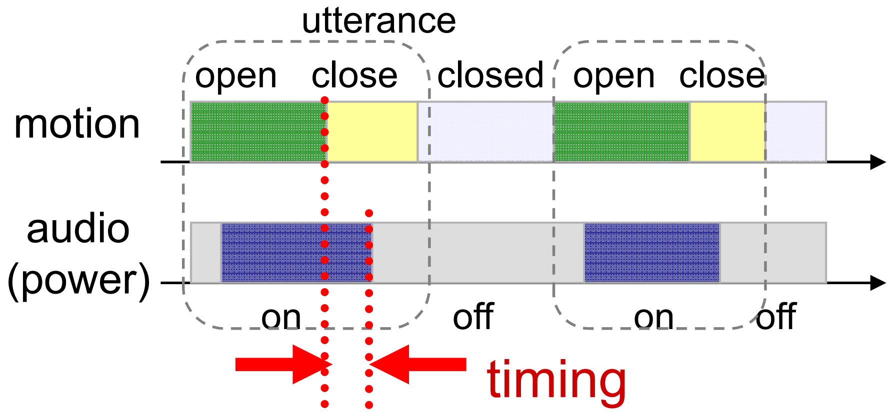




Observed as different media signals

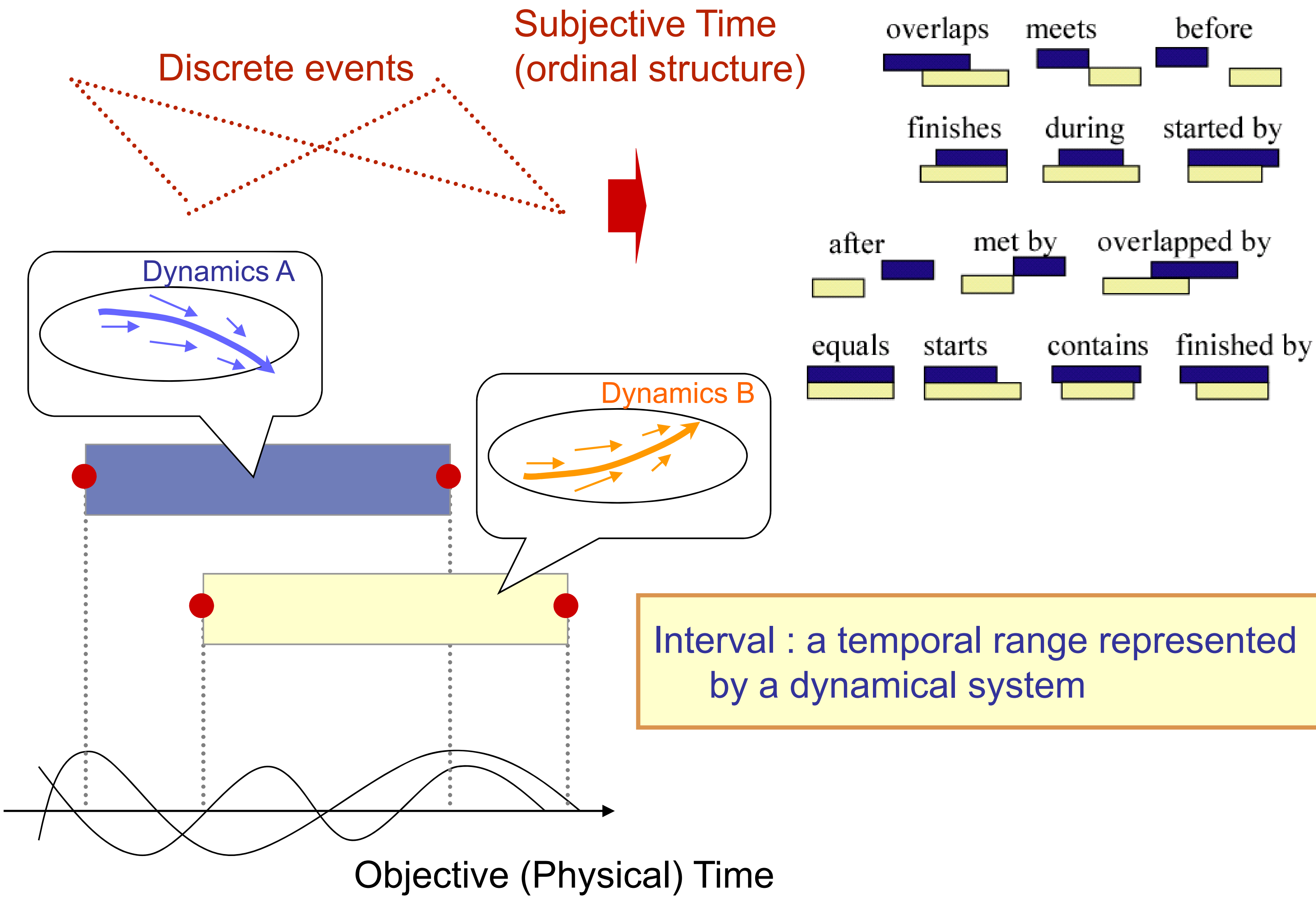






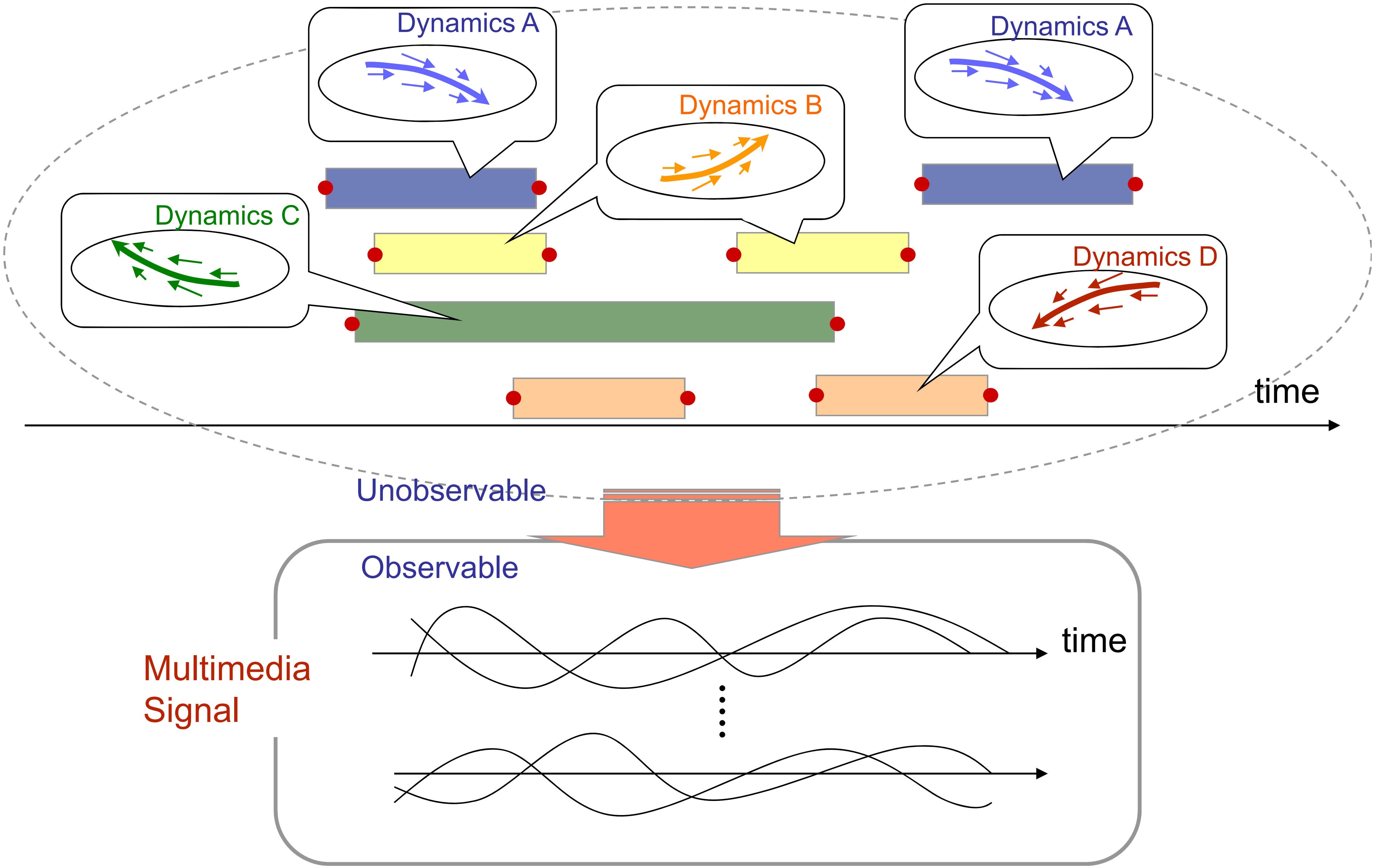
How to represent “a sense of time” of human?

Temporal Intervals



Interval : a temporal range represented by a dynamical system

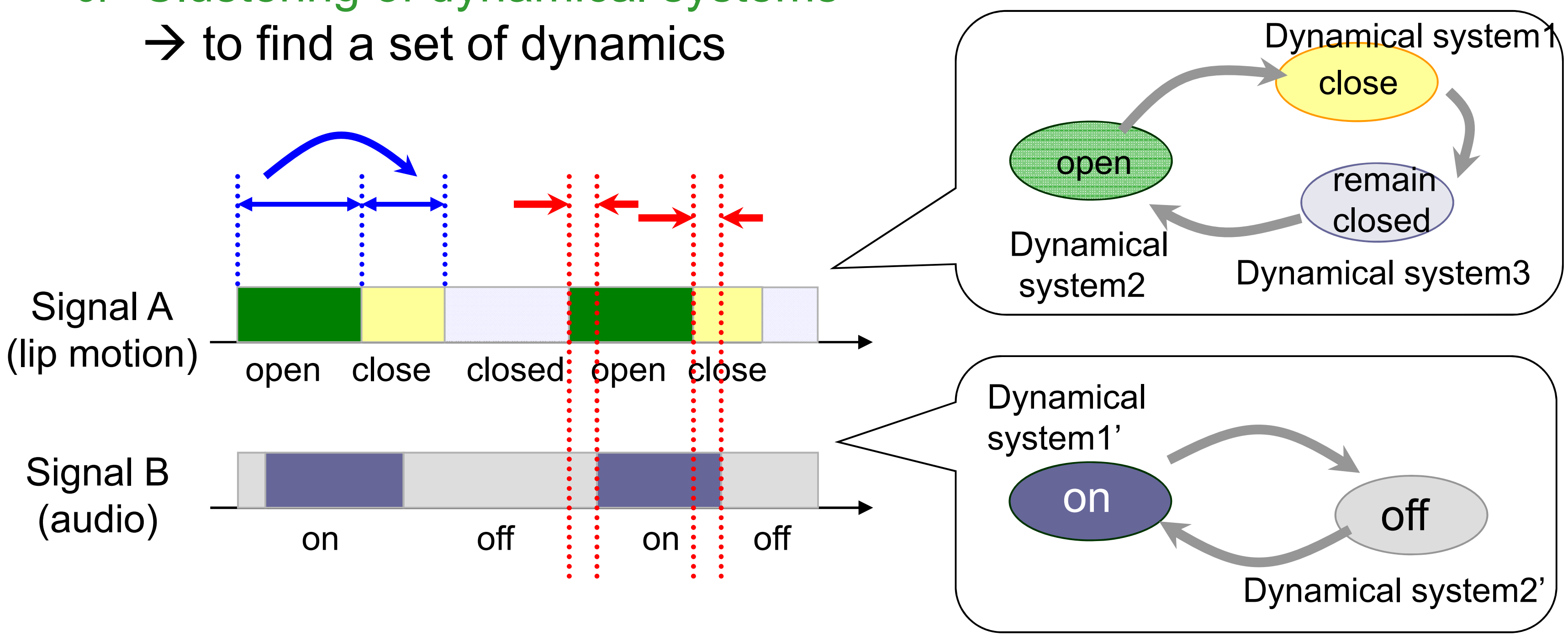
Orchestration of Dynamics



Interval-Based Hybrid Dynamical System

represents complex temporal structures with physical-time grounding

1. Interval-based state transition
   → to model rhythm and tempo of a single signal patterns
2. Timing structure model
   → to model timing structure in multiple signal patterns
3. Clustering of dynamical systems
   → to find a set of dynamics



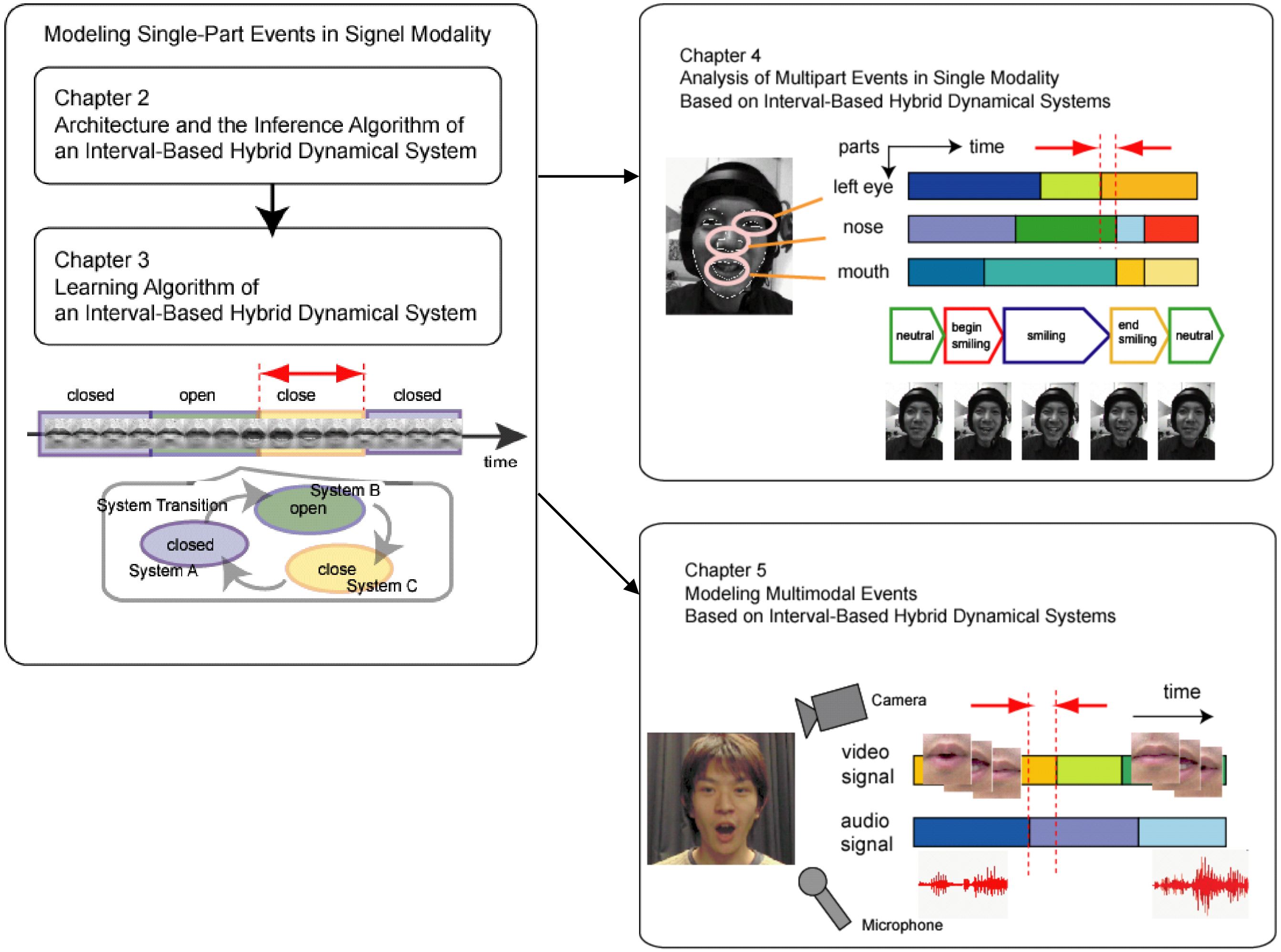
Overview of the Thesis

Modeling Single-Channel Signals
(Segmentation, Tempo, Rhythm)

Modeling Multi-Channel Signals
(Timing Structure)

Single-Media Signal

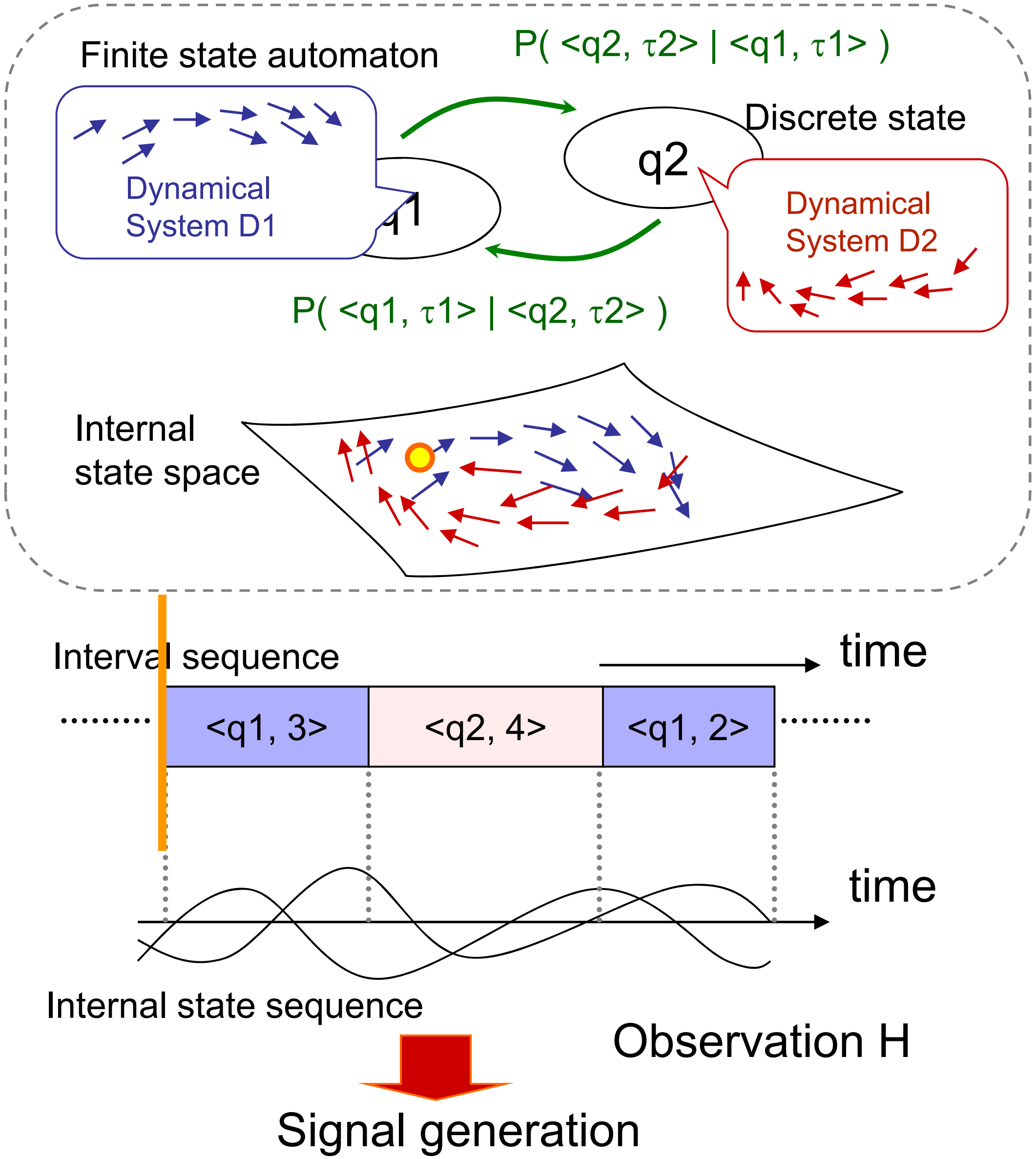
Multimedia Signal



Chapter 2

Interval-Based Hybrid Dynamical System

Interval-Based Hybrid Dynamical System



Interval-based transition

$$P(I_k =< q_j, \tau >| I_{k-1} =< q_i, \tau_p >)$$

Intervals

$$< q_j, \tau >$$

duration

Linear dynamical systems

State transition

$$x_t = F^{(i)} x_{t-1} + g^{(i)} + w_t^{(i)}$$

Observation

$$y_t = H x_t + v_t$$

Linear Dynamical Systems

- Internal state $x \in \mathrm{R}^{\mathrm{n}}$
- State transition

transition matrix

$$x_t = F^{(i)} x_{t-1} + g^{(i)} + w_t^{(i)}$$

bias process noise

- Observation

observation matrix

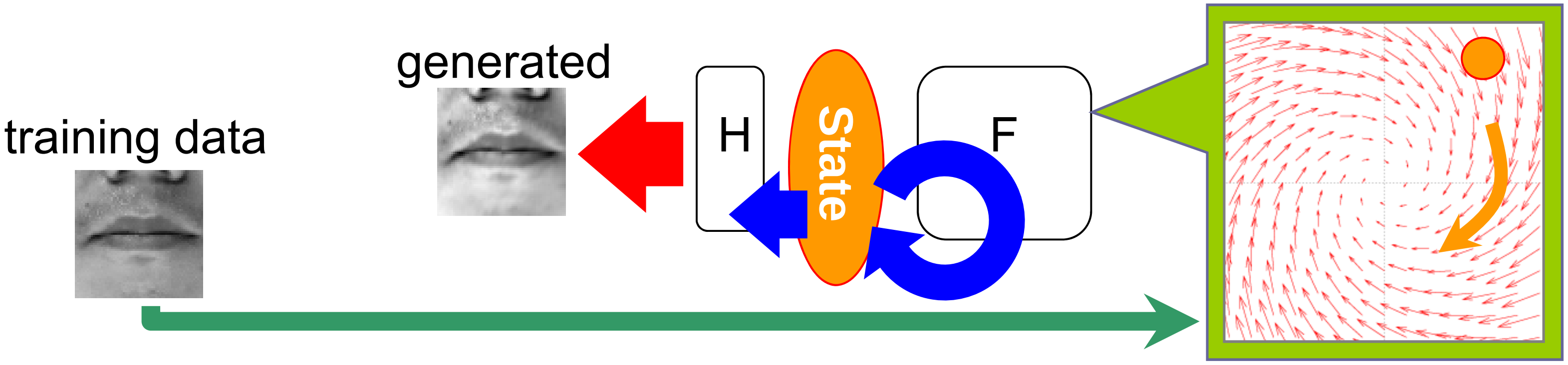
$$y_t = H x_t + v_t$$

observation noise

(w_t, v_t ~ Gaussian)

- Parameters: { F, g, H, noise covariance matrices, initial state}

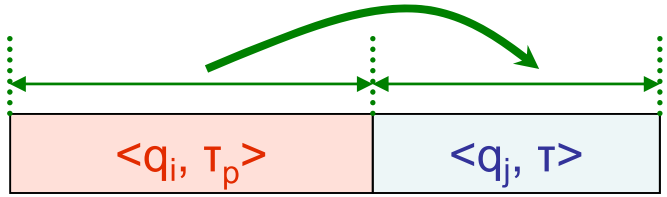
- Generation of continuous change



Interval-Based State Transition

- Modeling relation of duration lengths between adjacent intervals
  - to represent tempo and rhythm
  - to enhance robustness of segmentation (top-down constraints)

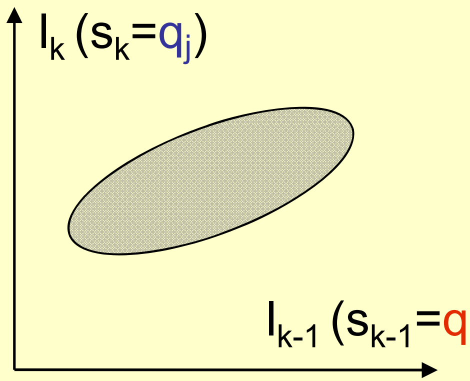
$$\boxed{P(\, I_k=\langle q_j,\ \tau\rangle \mid I_{k-1}=\langle q_i,\ \tau_p\rangle \,)} = P(\, s_k=q_j \mid s_{k-1}=q_i \,)\ \boxed{P(l_k=\tau \mid s_k=q_j, s_{k-1}=q_i, l_{k-1}=\tau_p)}$$

Interval sequence .... 

....

$P(l_k=\tau, l_{k-1}=\tau_p \mid s_k=q_j, s_{k-1}=q_i)$

Assume Gaussian distribution

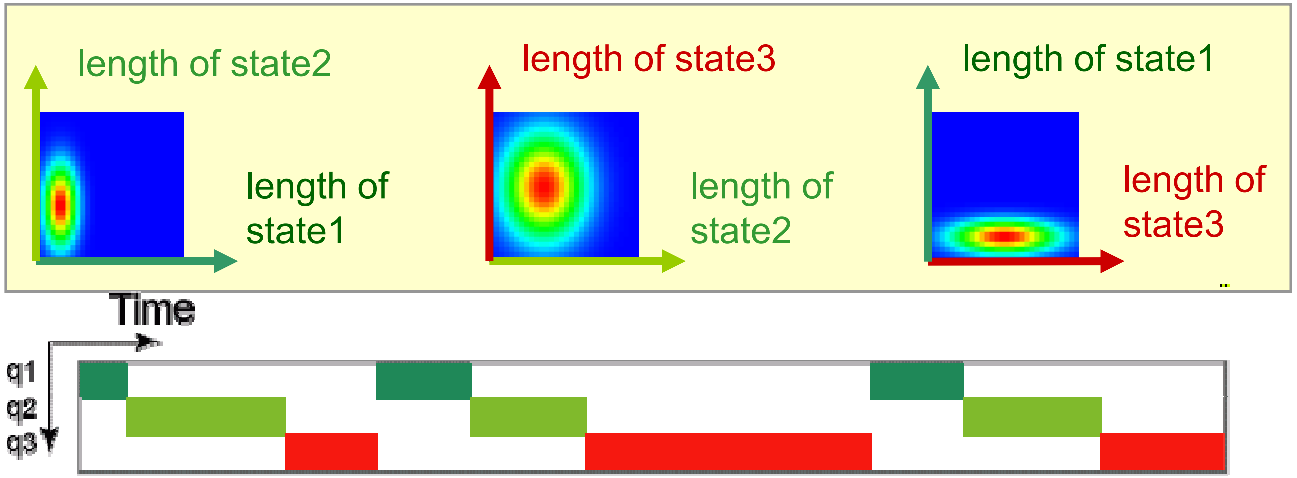


Example of Interval Sequence Generation

- Randomly generated interval sequences using manually given distributions

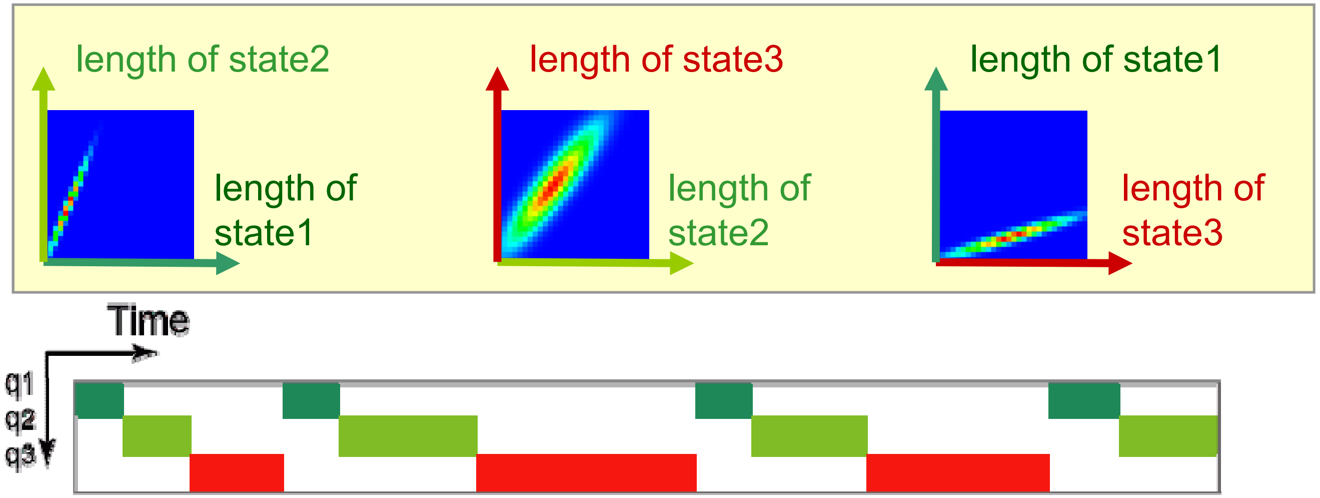
Three discrete states

Correlation = 0



Generated Sequence

Correlation ≒ 0.9



Generated Sequence

Chapter 3

Learning Method for the Interval-Based Hybrid Dynamical System

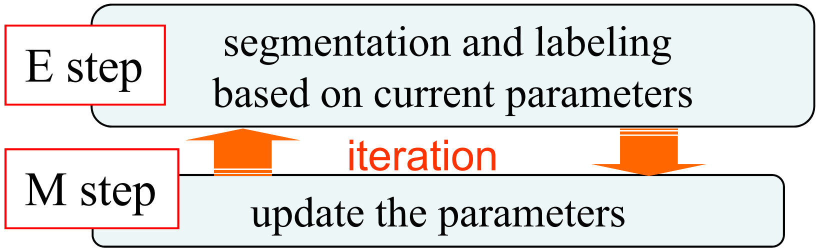
Difficulty of the Parameter Estimation

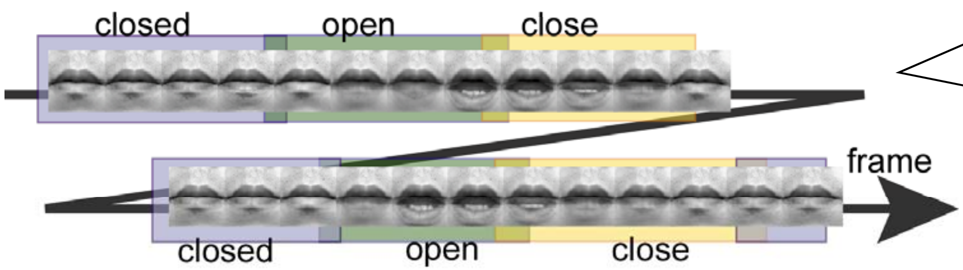
Assume that only a set of vector sequences is given

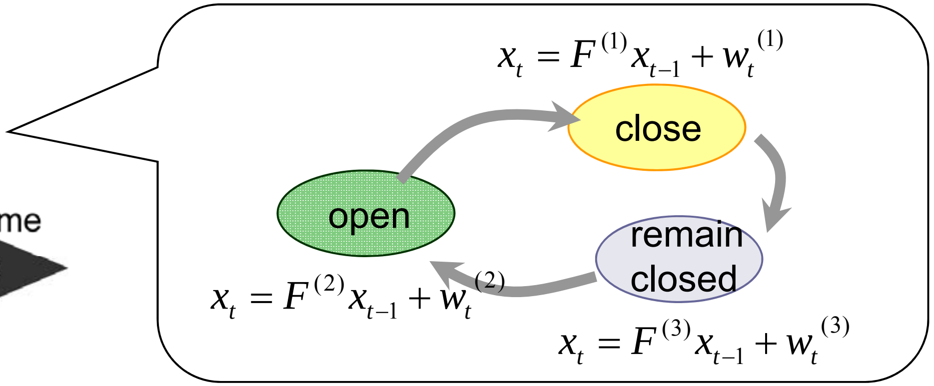
- Need to define a set of dynamics (primitives)
  - Defined manually in existing work

- Need to solve paradoxical nature of parameter estimation
  - Segmentation requires identified dynamical systems
  - Identification of dynamical systems requires segmentation

- Need to solve initialization problem of the EM algorithm
  - Strongly depends on its initial parameters

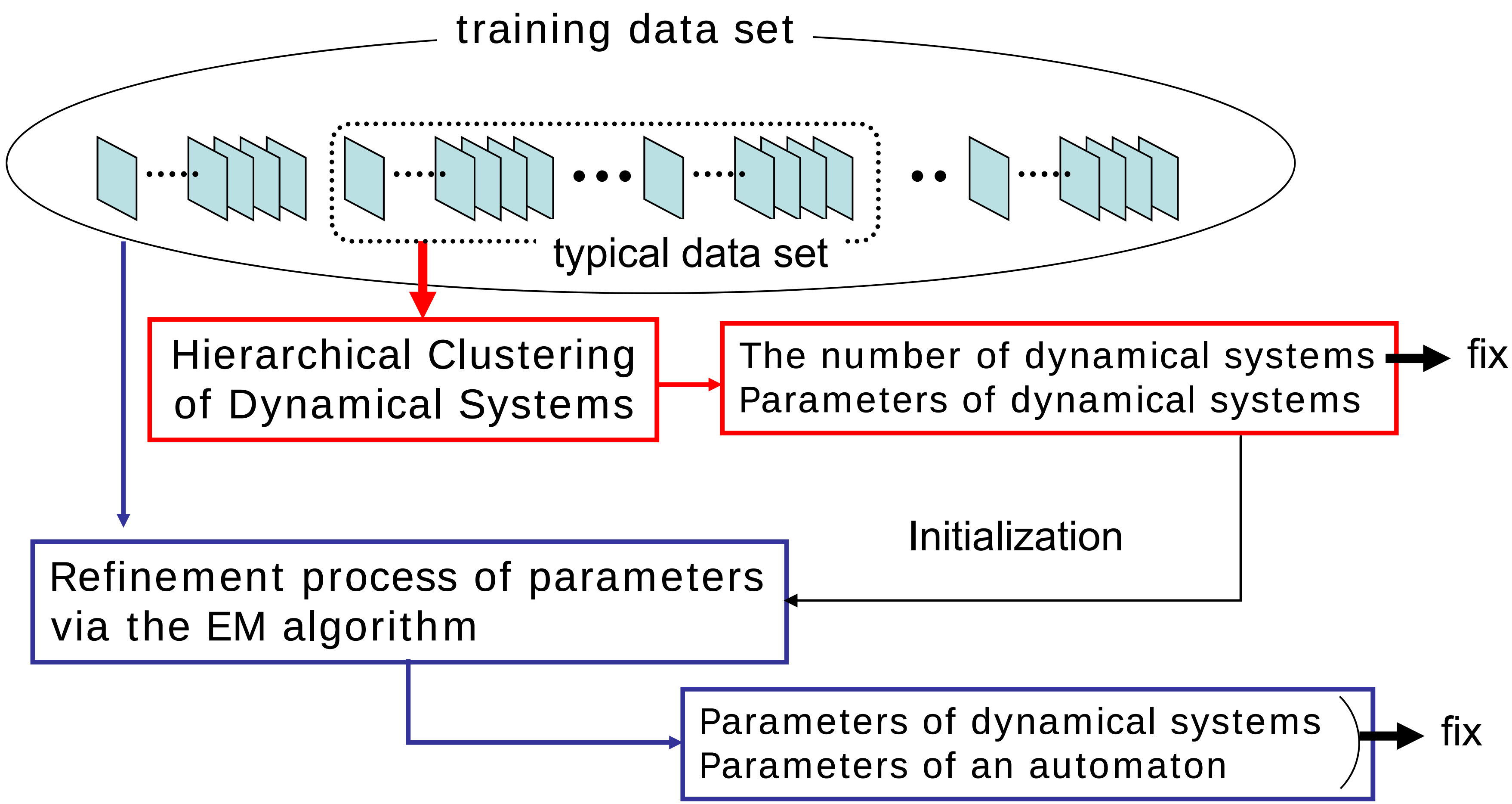






Overview of the Training

- Two-step learning method

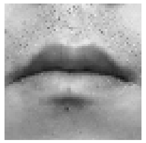


Example of Two-Step Learning Method

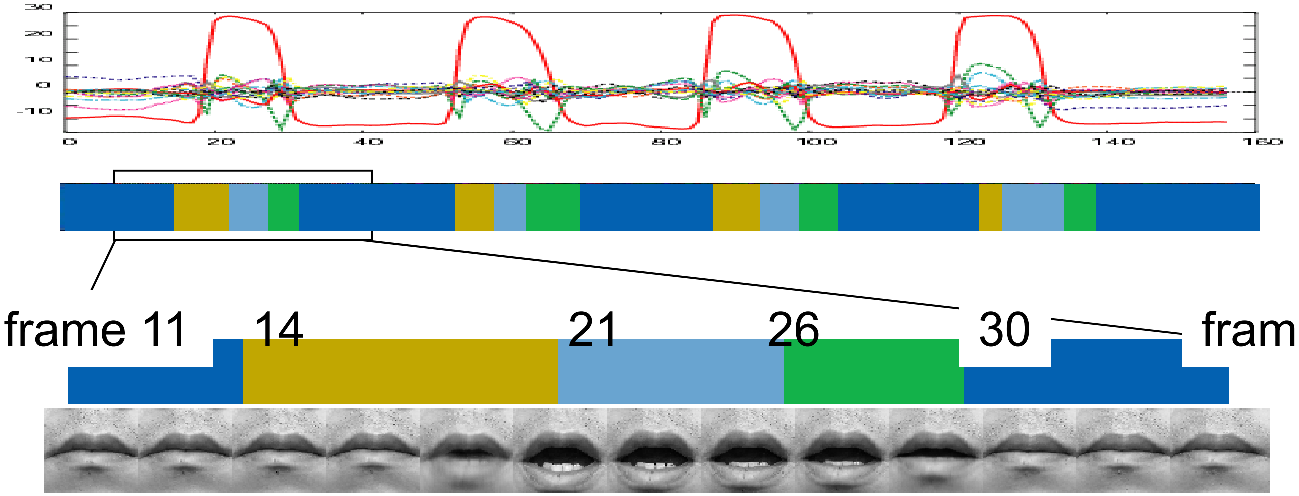
Hierarchical Clustering

feature vector sequence (PCA coef.)

time

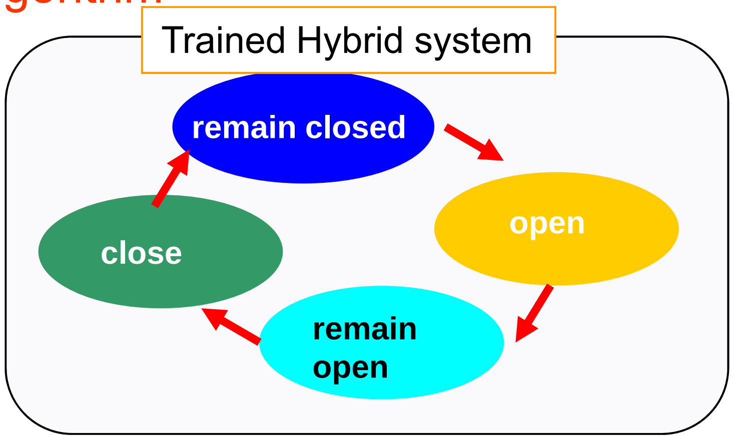


input lip image:
/mamamama/

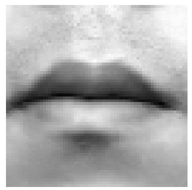


remain closed open remain open close

Training via the EM algorithm

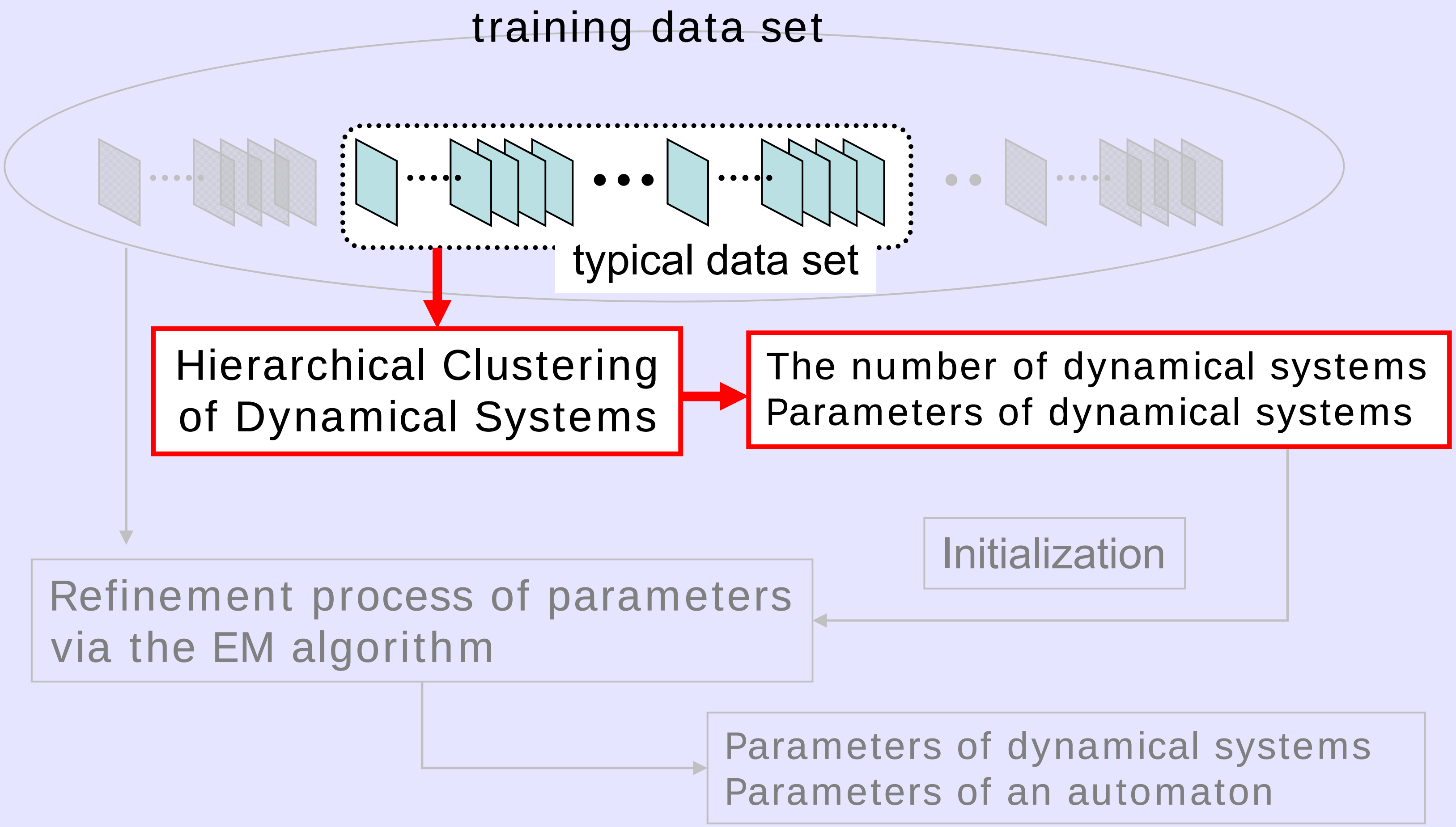


initialized by a closed lip



generation

Hierarchical Clustering of Dynamical Systems



Constrained Linear System Identification

- State transition

transition matrix

$$x_t = F^{(i)} x_{t-1} + g^{(i)} + w_t^{(i)}$$

bias process noise

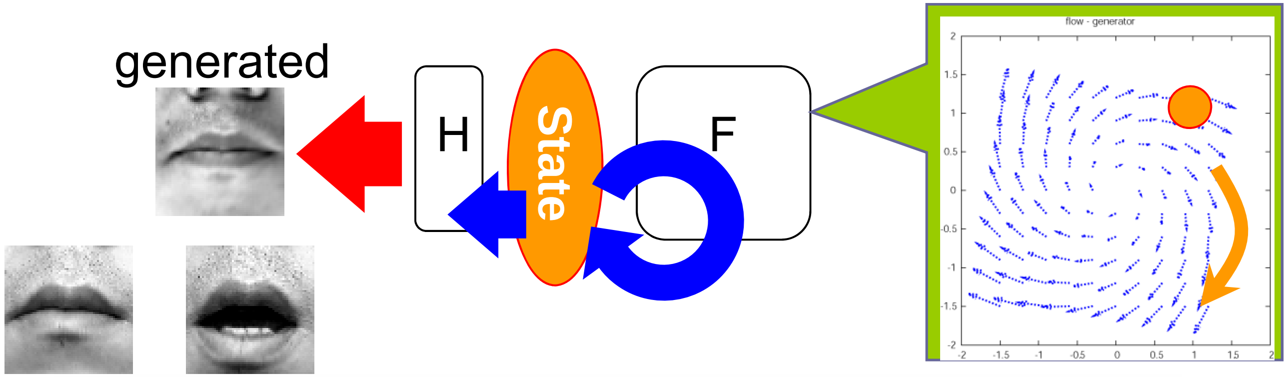
- Observation

observation matrix

$$y_t = H x_t + v_t$$

observation noise

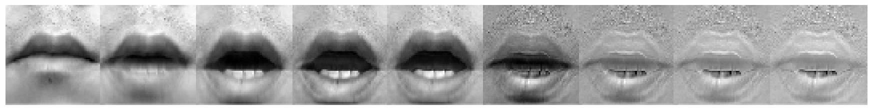
(w_t, v_t ~ Gaussian)



Overfitting problem

t = 1 5 9

no constraint



System behavior is determined by transition matrix F

Need constraints on the transition matrix F

Class of Linear Dynamical Systems

- Temporal evolution of the state

$$x_t = F^t x_0 = c_1 \lambda_1^t e_1 + c_2 \lambda_2^t e_2 + \cdots + c_n \lambda_n^t e_n$$

$$F = E \Lambda E^{-1} = [e_1, \ldots, e_n] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} [e_1, \ldots, e_n]^{-1}$$

(ex.) n = 2

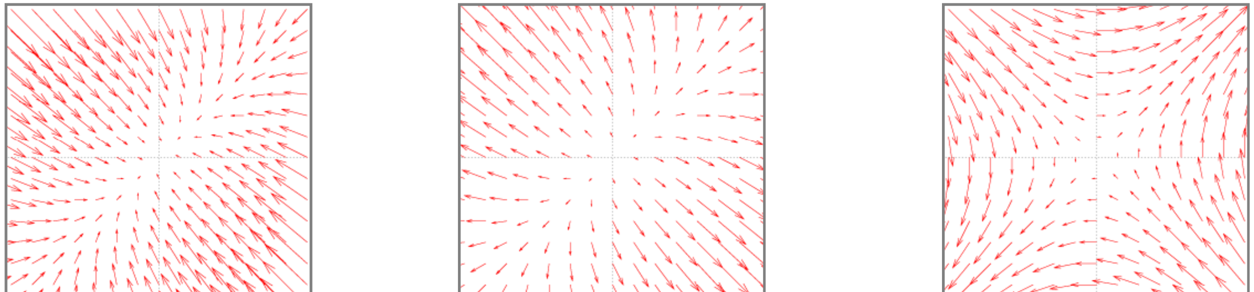
|λ1| < 1 and |λ2| < 1

|λ1| > 1 and |λ2| > 1

|λ1| > 1 and |λ2| < 1

λ1 and λ2 is positive real

λ1 and λ2 is complex num.



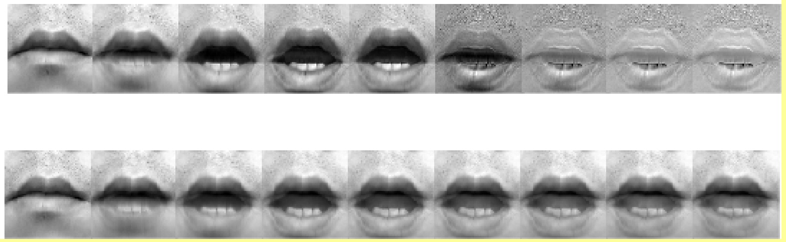
$\lambda_{max} = 3.0$

$\lambda_{max} < 1.0$

t = 1 5 9

no constraint

with constraint



Constrained Linear System Identification

- Eigenvalue constraints
  - Upper bound of eigenvalues is determined by

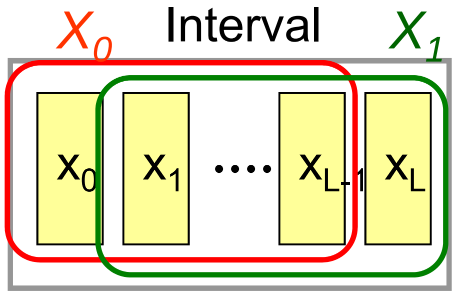
$$F^{(i)} = \begin{array}{cc} & v \\ & \left[\begin{matrix} f_{11}^{(i)} & \cdots & f_{1n}^{(i)} \\ \vdots & f_{rc}^{(i)} & \vdots \\ f_{n1}^{(i)} & \cdots & f_{nn}^{(i)} \end{matrix}\right] & u \end{array}$$

$$UpperBound = \boxed{\max_r \sum_{c=1}^{n} \left| f_{rc}^{(i)} \right| = 1}$$

Deduced from
Gershgorin's theorem

- Estimation of transition matrix

Least squares problem



$$F^{(i)*} = \arg\min_{F^{(i)}} \| F^{(i)} X_0^{(i)} - X_1^{(i)} \|^2$$

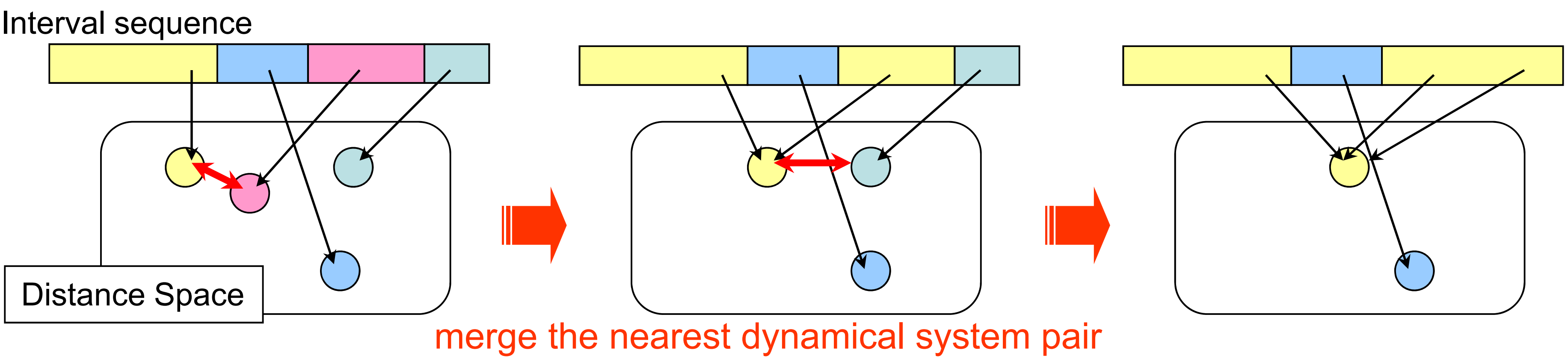
$$= \lim_{\delta^2 \to 0} X_1^{(i)} \underline{X_0^{(i)\mathrm{T}} (X_0^{(i)} X_0^{(i)\mathrm{T}} + \delta^2 I)^{-1}} = X_1^{(i)} \underline{X_0^{(i)+}}$$ (pseudo-inverse)

Stop the limit before δ^2 converges to 0

δ^2 controls the scale of matrix elements

Algorithm of Hierarchical Clustering

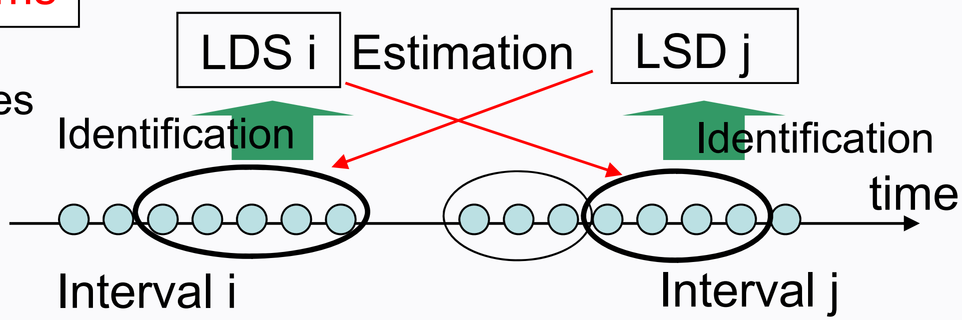
1. Divide the training data into short intervals (Initialization)
2. Identify the parameters of the dynamical systems from each interval
3. Calculate distances of all the dynamical system pairs
4. Merge the **nearest pair** of dynamical systems (intervals are also merged)
5. Identify the new dynamical system from the merged intervals
6. Repeat 3 to 5



Distance between dynamical systems

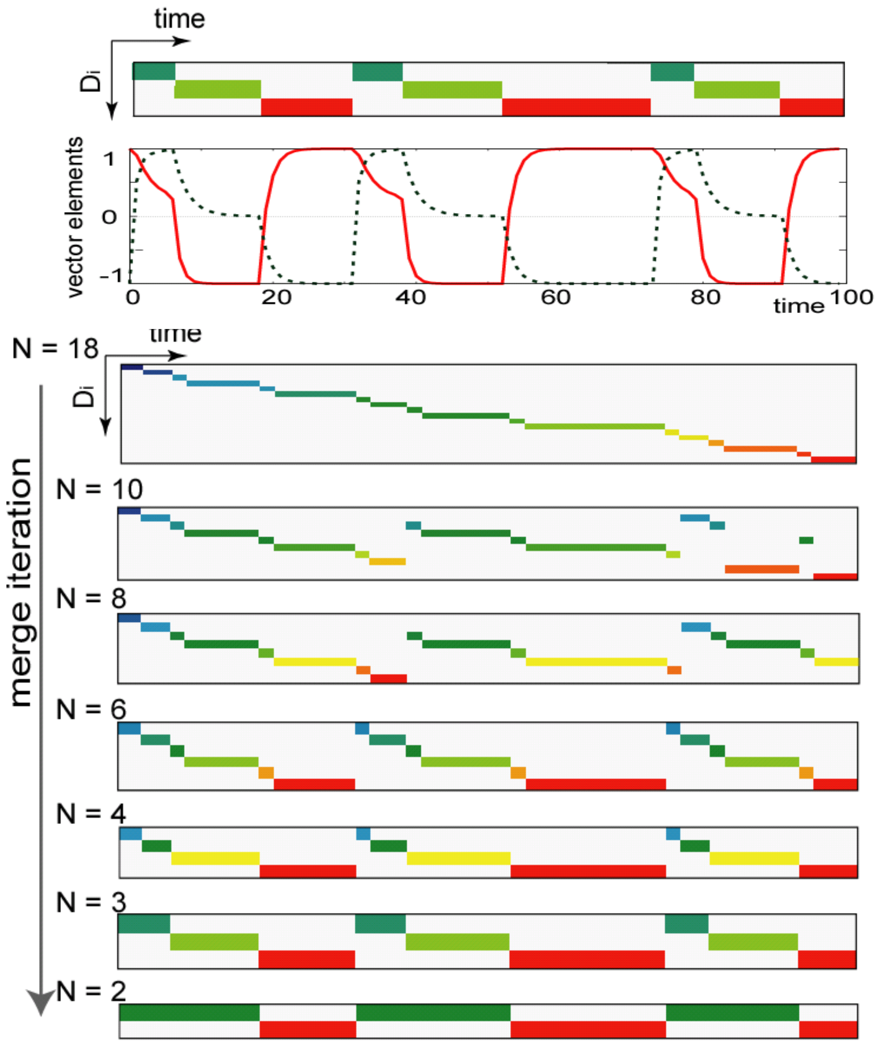
Average of Kullback-Leibler divergences

$$Dist(D_i, D_j) = \frac{KL(D_i \| D_j) + KL(D_j \| D_i)}{2}$$

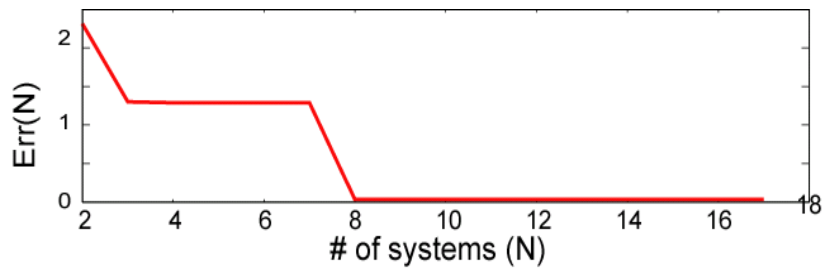


Simulation Result

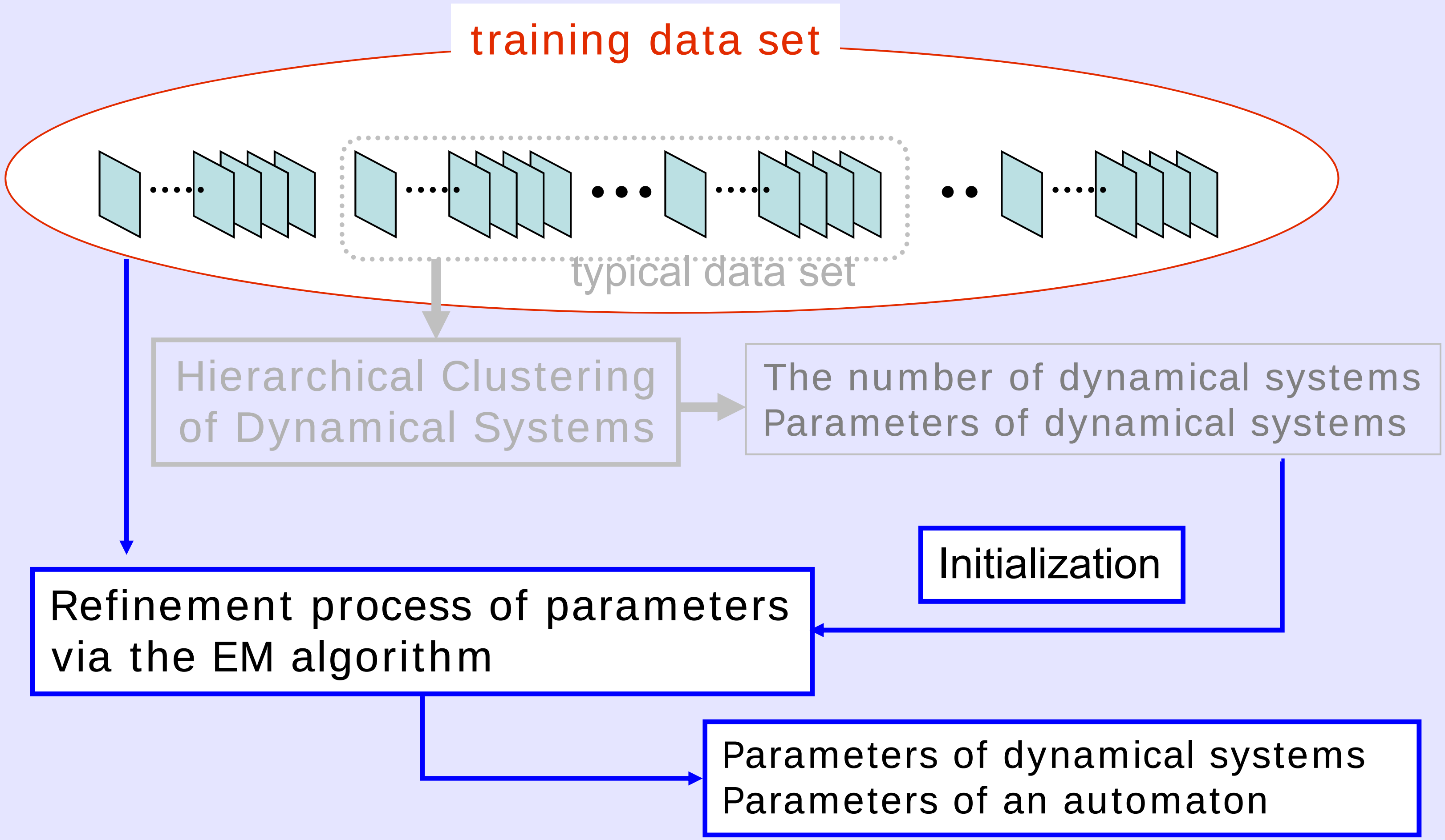
- Input data: generated from three dynamical systems



Prediction error of overall systems at each iteration step



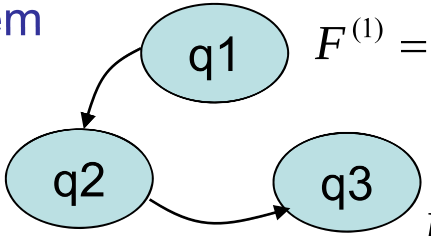
Refinement Process of All the Parameters



Evaluation based on Simulated Data

- Interval-Based HDS with known parameters

Original system

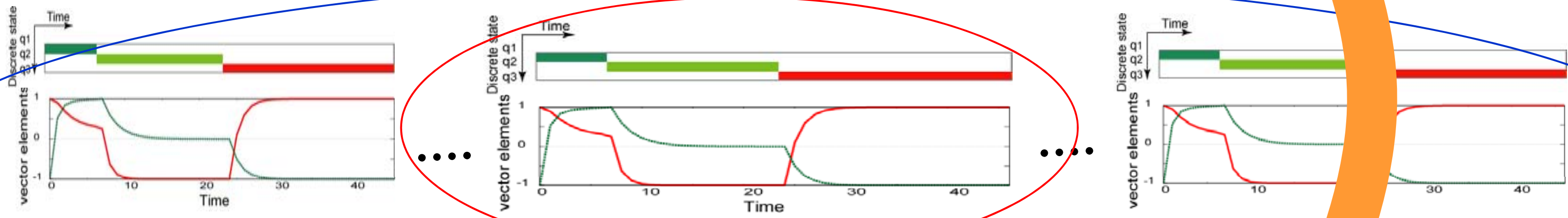


$$F^{(1)} = \begin{bmatrix} 0.6 & -0.1 \\ -0.1 & 0.2 \end{bmatrix}$$

$$F^{(2)} = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.6 \end{bmatrix}$$

$$F^{(3)} = \begin{bmatrix} 0.5 & 0.1 \\ -0.1 & 0.3 \end{bmatrix}$$

Generation

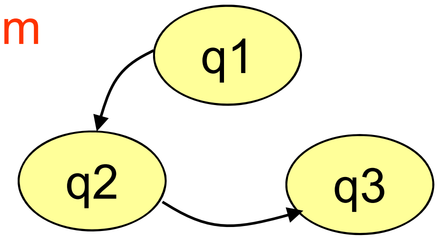


Used for clustering (one sequence)

Used for EM algorithm (10 sequences)

Compare

Trained system



Estimated parameters

Simulation Result

- Comparison between given and estimated parameters

Original (ground truth)

$$F^{(1)} = \begin{bmatrix} 0.60 & -0.10 \\ -0.10 & 0.20 \end{bmatrix} \quad F^{(2)} = \begin{bmatrix} 0.30 & 0.00 \\ 0.00 & 0.60 \end{bmatrix} \quad F^{(3)} = \begin{bmatrix} 0.50 & 0.10 \\ -0.10 & 0.30 \end{bmatrix}$$

Estimated parameters via clustering

$$F^{(1)} = \begin{bmatrix} 0.01 & -0.14 \\ -0.01 & 0.21 \end{bmatrix} \quad F^{(2)} = \begin{bmatrix} 0.86 & 0.30 \\ -0.21 & -0.06 \end{bmatrix} \quad F^{(3)} = \begin{bmatrix} 0.74 & -0.44 \\ -0.11 & 0.75 \end{bmatrix}$$

Estimated parameters via EM algorithm

$$F^{(1)} = \begin{bmatrix} 0.60 & -0.10 \\ -0.10 & 0.20 \end{bmatrix} \quad F^{(2)} = \begin{bmatrix} 0.32 & 0.02 \\ 0.06 & 0.52 \end{bmatrix} \quad F^{(3)} = \begin{bmatrix} 0.49 & 0.09 \\ -0.10 & 0.29 \end{bmatrix}$$

Discussion

- Interval-based hybrid dynamical system
  - Interval-based state transition to model tempo and rhythm
  - Linear dynamics to model continuously changing patterns

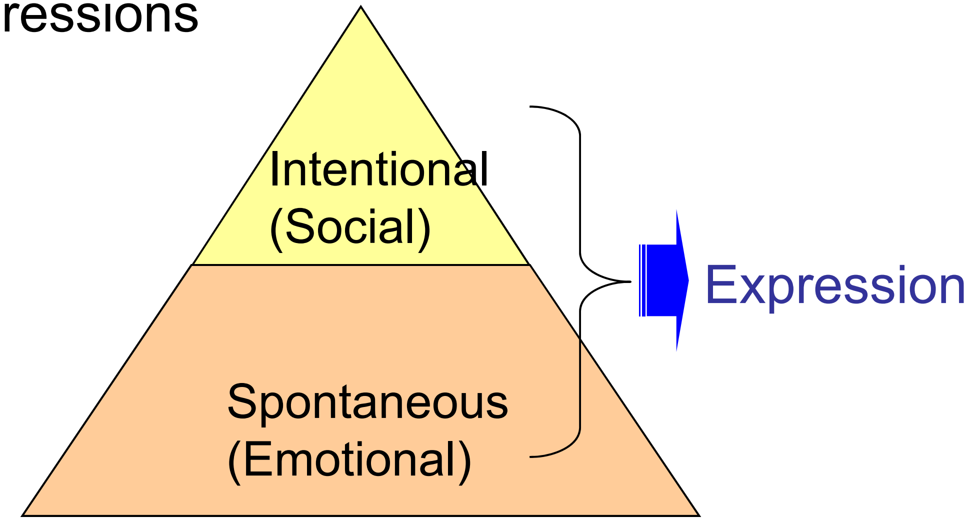
- Two-step learning method for the interval-based
  - Clustering of dynamical systems + EM algorithm
  - Constrained system identification based on eigenvalues

Chapter 4

Analysis of Timing Structures in Multipart Motion of Facial Expression

Facial Expression as Communication Protocol

- Communication via facial expressions
  - Generation
    - Express internal state
  - Recognition
    - Estimate internal state
- Acquisition of expressions
  - Intrinsic
    - Smile, cry, surprise
  - Learned from experience (parents)
    - Social contexts



Facial expression = Emotional category

Facial expression = Communication protocol

Related work

- FACS (Facial Action Coding System) (Ekman, et al.)
  - AU (Action Unit) : motion primitives in faces
  - Describe facial expressions based on combination of AU

    (ex.) Surprise = AU1+2+5+26
  - Describes only emotional categories
    - {happiness, surprise, fear, anger, disgust, sadness}

Problem: cannot describe dynamic structures
(synchronous/asynchronous motions, duration of motions, etc.)

Psychological experiments

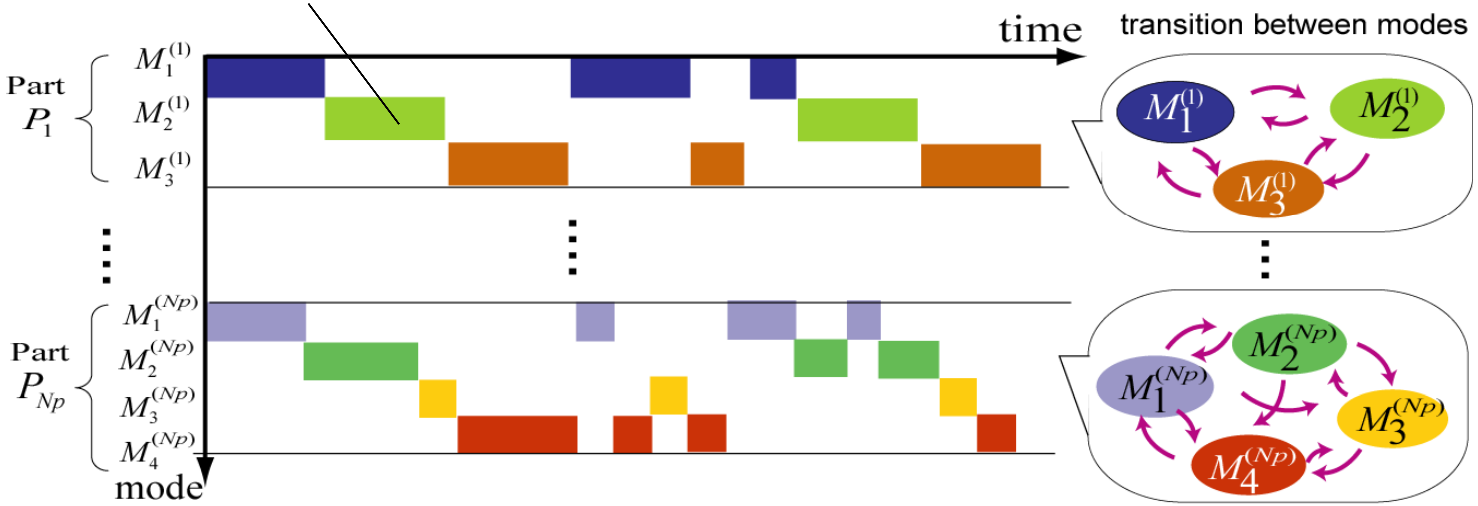
・Temporal difference of beginning time between eyes and mouth is important to discriminate social, pleasant, and unpleasant smiles（Nishio&Koyama1997）

・Human recognition of facial expressions depends on duration of motion（Ekman&Friesen1982, Kamachi2001, Krumhuber2005）

Facial Score: Interval-Based Facial Action Description

- Define facial parts – move independently
- Define modes - motion primitives (dynamics)

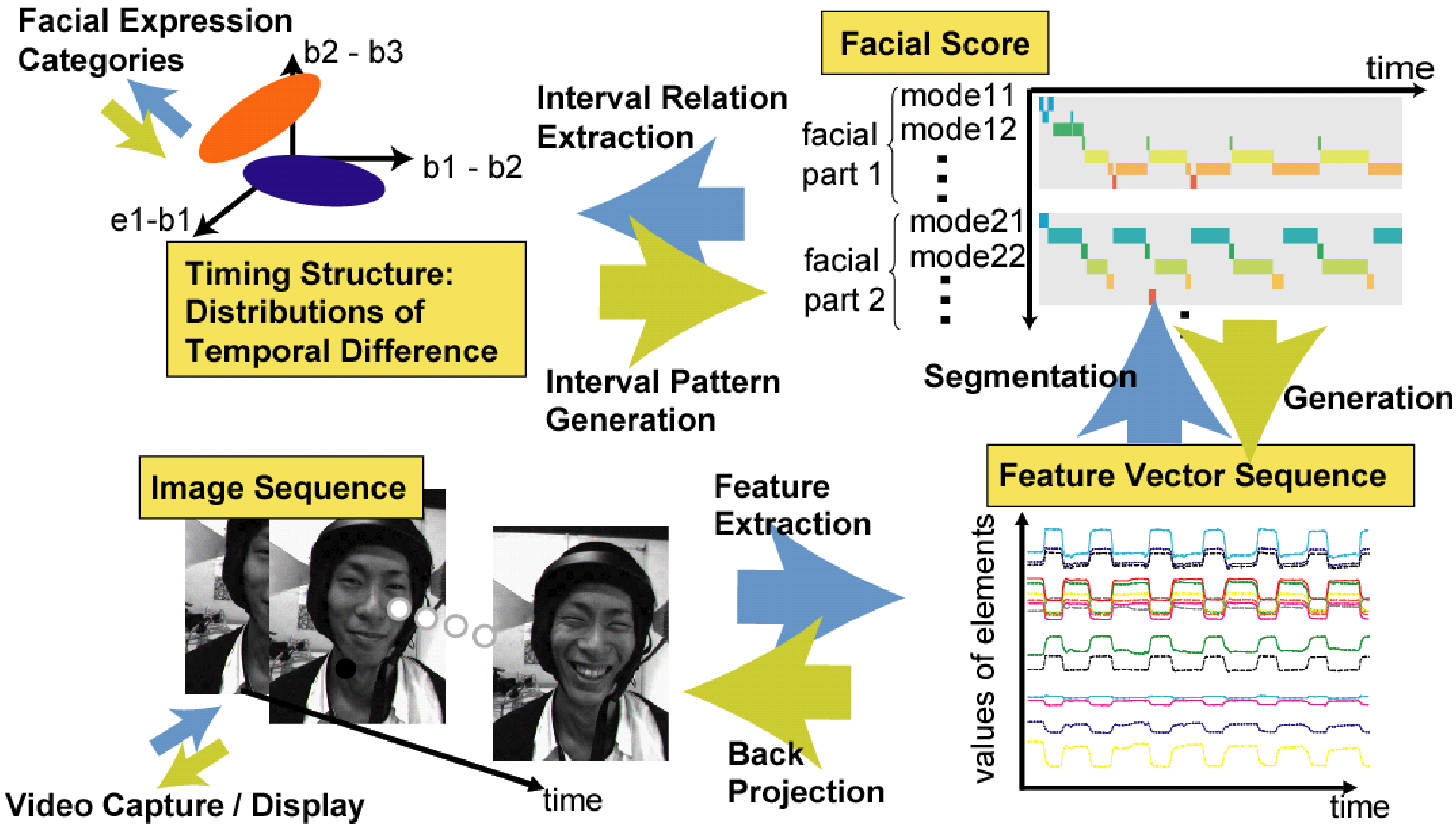
Interval: { beginning point, ending point , mode label}



Facial Score: {Interval set of part1, … , Interval set of part N}

Represent timing structure among modes (dynamics)

Facial Expression Generation and Recognition



1. Definition of Facial Scores
2. Automatic Acquisition of Facial Scores
3. Evaluation

- Definition of parts
- Definition of modes

Facial Parts in Facial Scores

- Follows Ekman's definition
  - Parts = {left/right eyebrow, left/right eye, nose, mouth}
- Feature vector of each part
  - x,y coordinates of feature points

$$z^{(a)} = (x_1^{(a)}, y_1^{(a)}, \ldots, x_{Na}^{(a)}, y_{Na}^{(a)})^{\mathrm{T}}$$

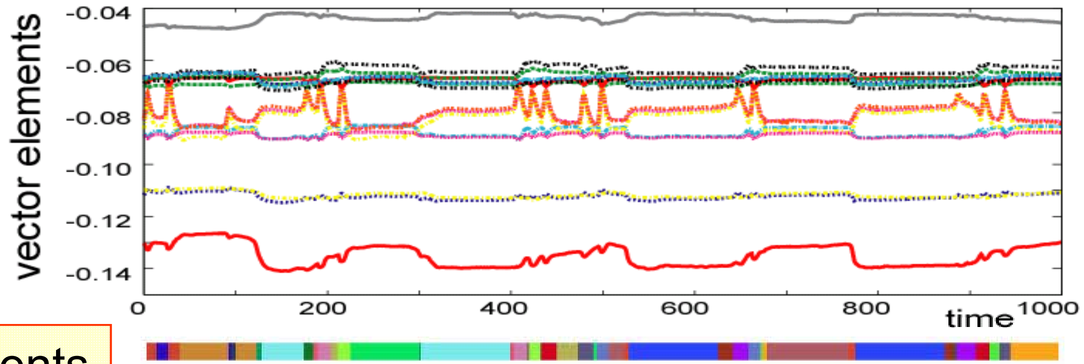
$$a \in \mathrm{Parts}$$

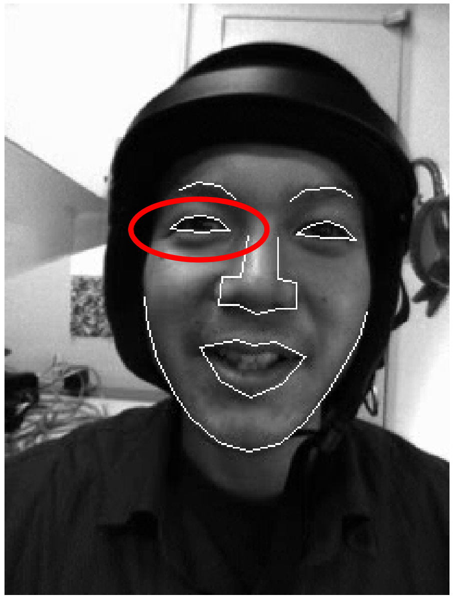
(dimensionality: each eyebrow : 10, each eye: 16, nose: 22, mouth: 16)

Active Appearance Model(AAM) (Cootes 1998)

Facial Modes in Facial Scores

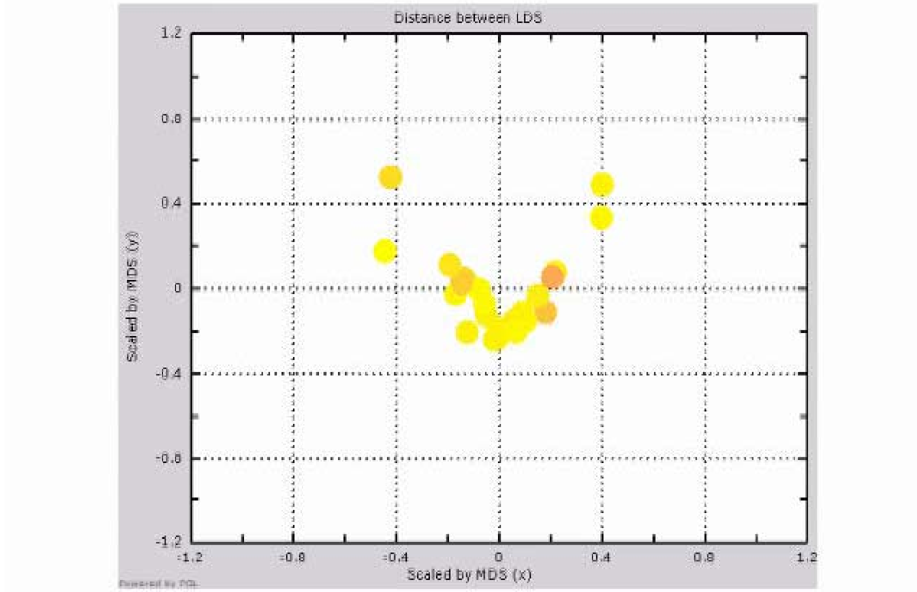
- Smiled four times
- Feature vector: x,y coordinates of feature points around right eye (eight points)
- Length: 1000 frames





segments

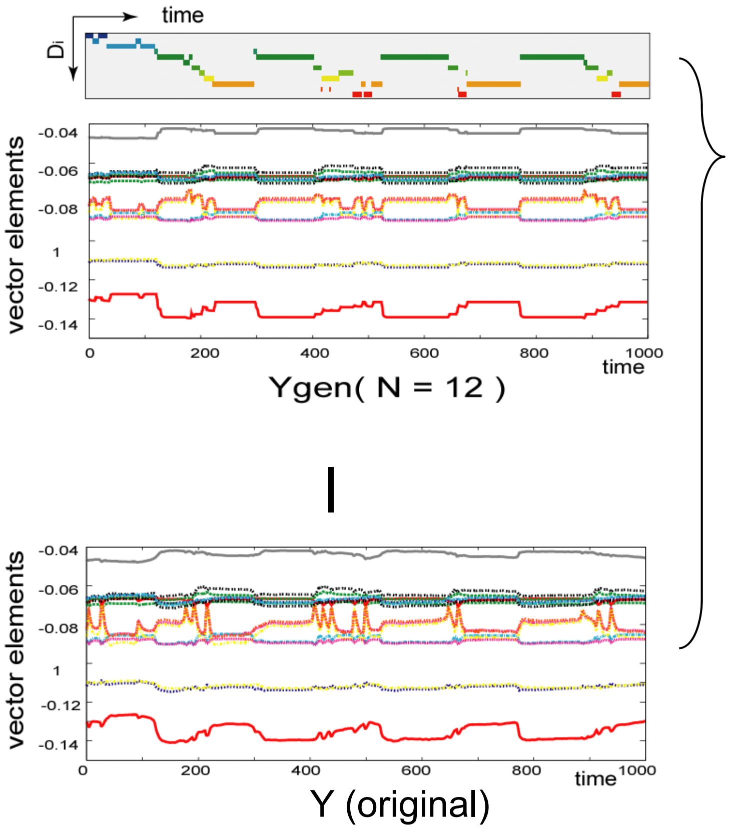
pseudo distance space



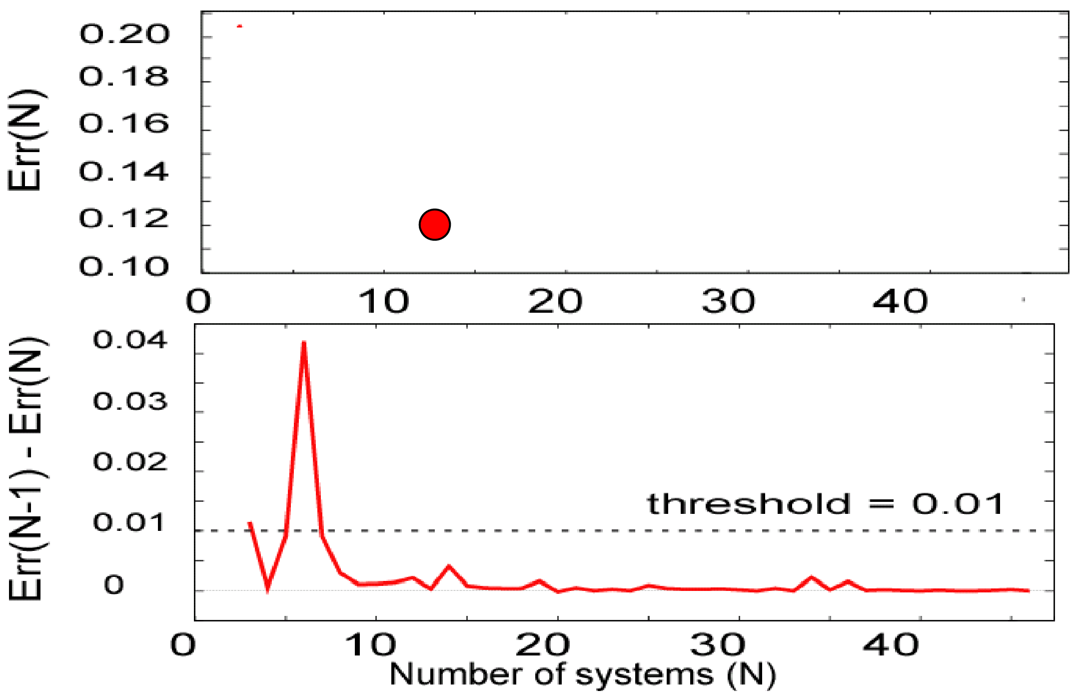
Tracked by Active Appearance Model
(Cootes 1998)
Thanks to Stegmann's AAM-API

Determine the number of modes (dynamical systems)

- Find a rapid change of model fitting error curve



$$Err(N) = \sqrt{\sum_{t=1}^{L} \| \mathrm{orig}_{(t)} - \mathrm{gen}^{(\mathrm{N})}{}_{(t)} \|^2}$$

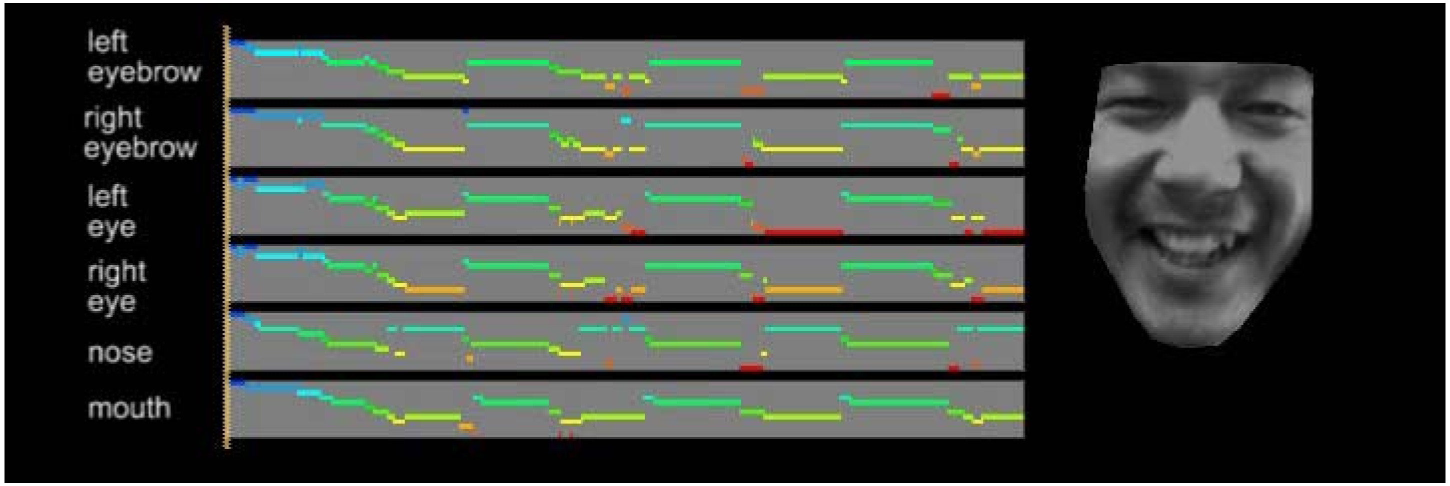


1. Definition of Facial Scores
2. Automatic Acquisition of Facial Scores
3. Evaluation

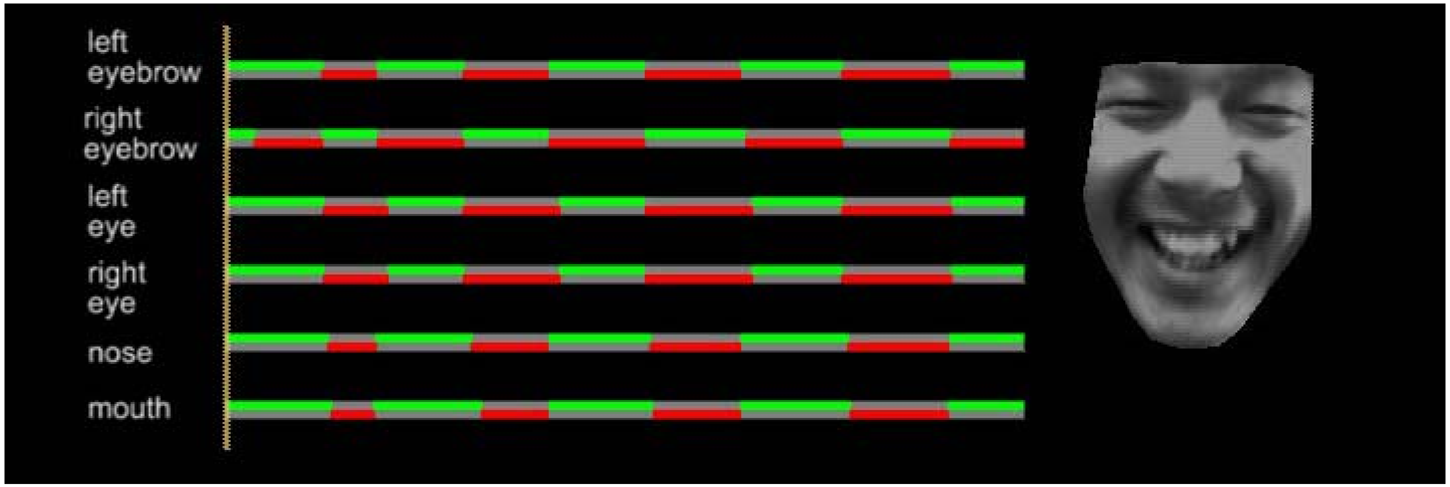
- Generation of expressions
- Discrimination of expressions

Generation of Facial Expression

Num. of modes: 12 in each parts



Num. of modes: 2 in each parts

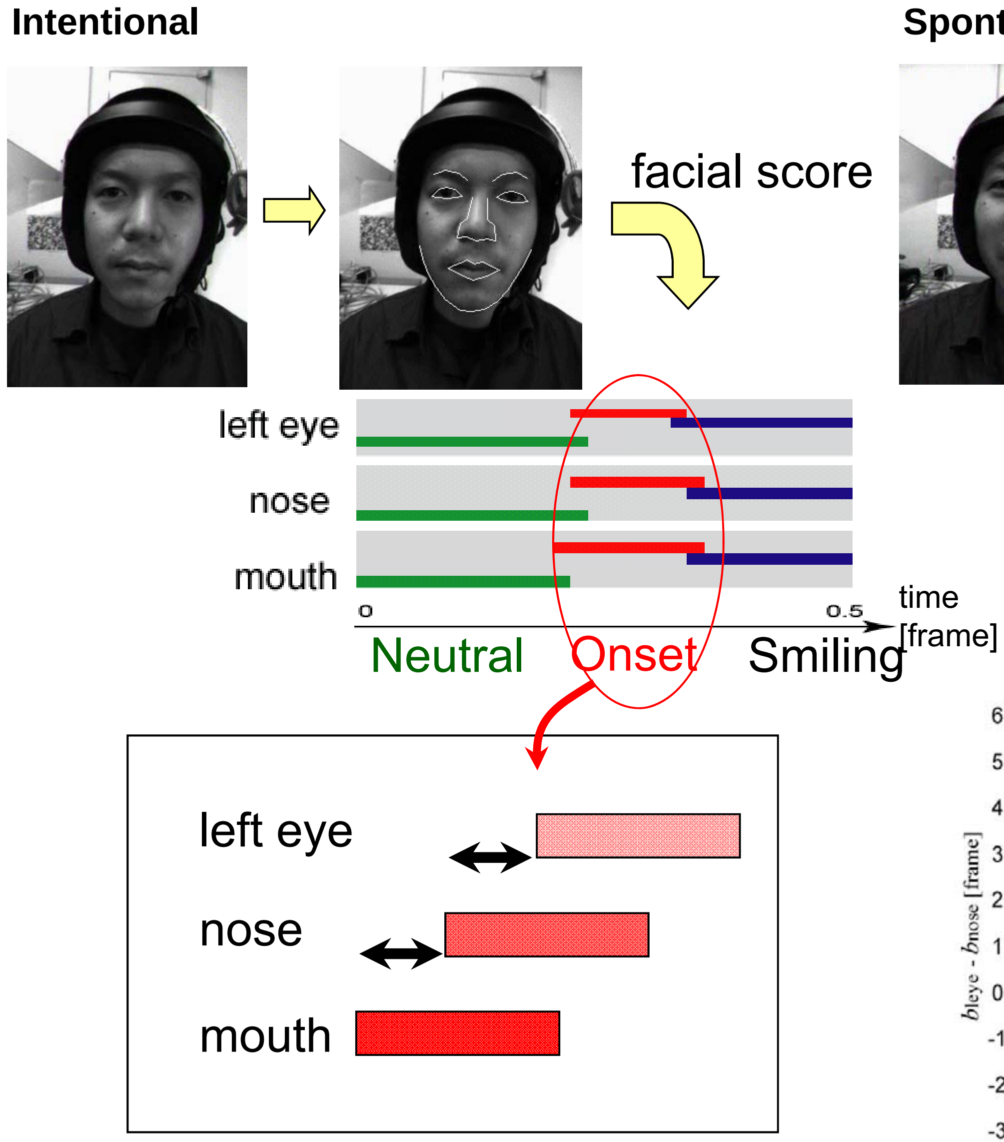


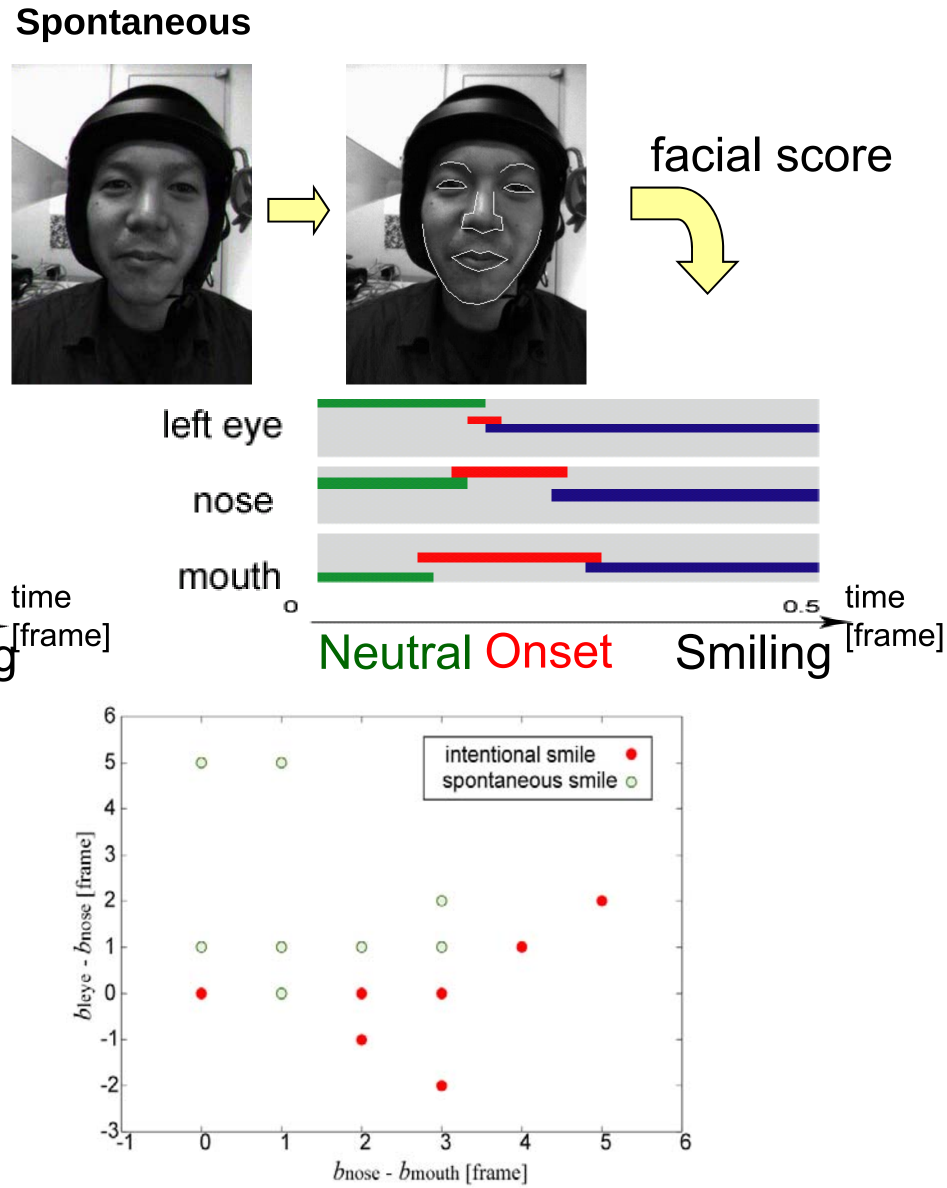
Discrimination of facial expressions

- Subjects
  - Intentional smile
  - Spontaneous smile

  separable?
  - Six (male)
  - about 30-50 times for each smile category
- Method
  - Video data
    - VGA480x640(down sampling to 240x360) , 60fps
  - Instruction of expressions
    - Start from neutral face
    - Intentional: make smile during watching a disgust movie
    - Spontaneous: watch Japanese stand-up comedy movies

Timing Structure in Facial Scores



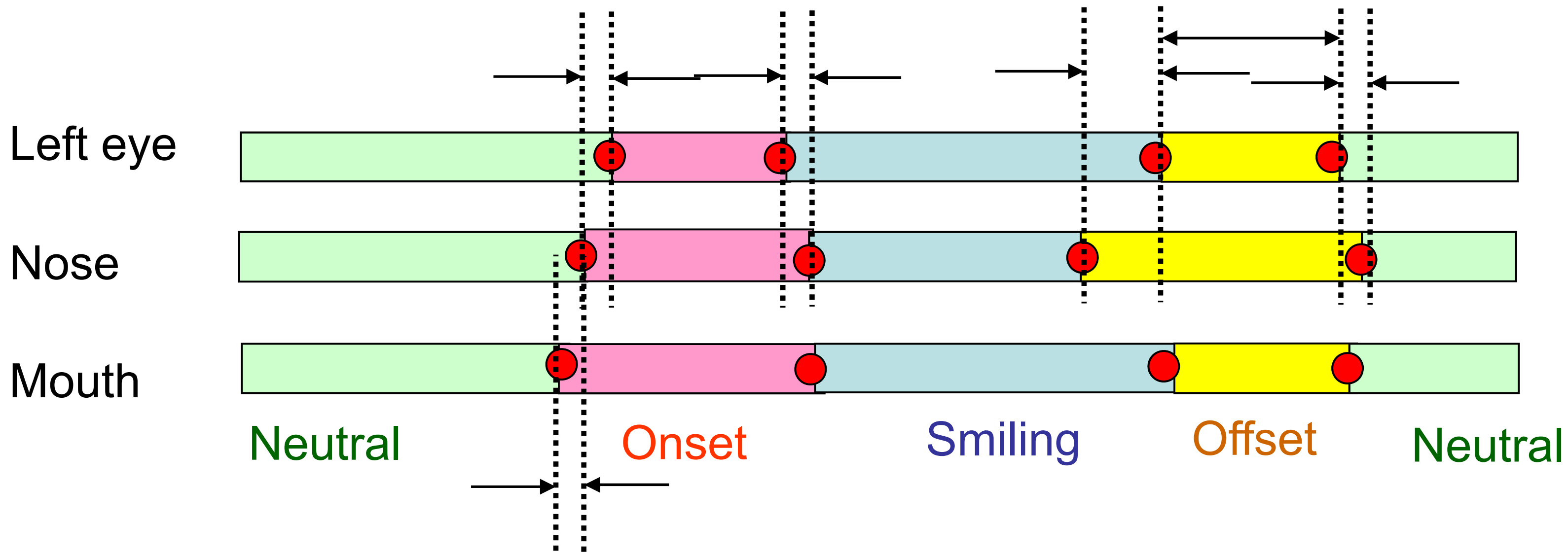


Extract Timing Structure from Facial Scores

1. Use temporal differences among the beginning and ending points of "onset" and "offset" motion

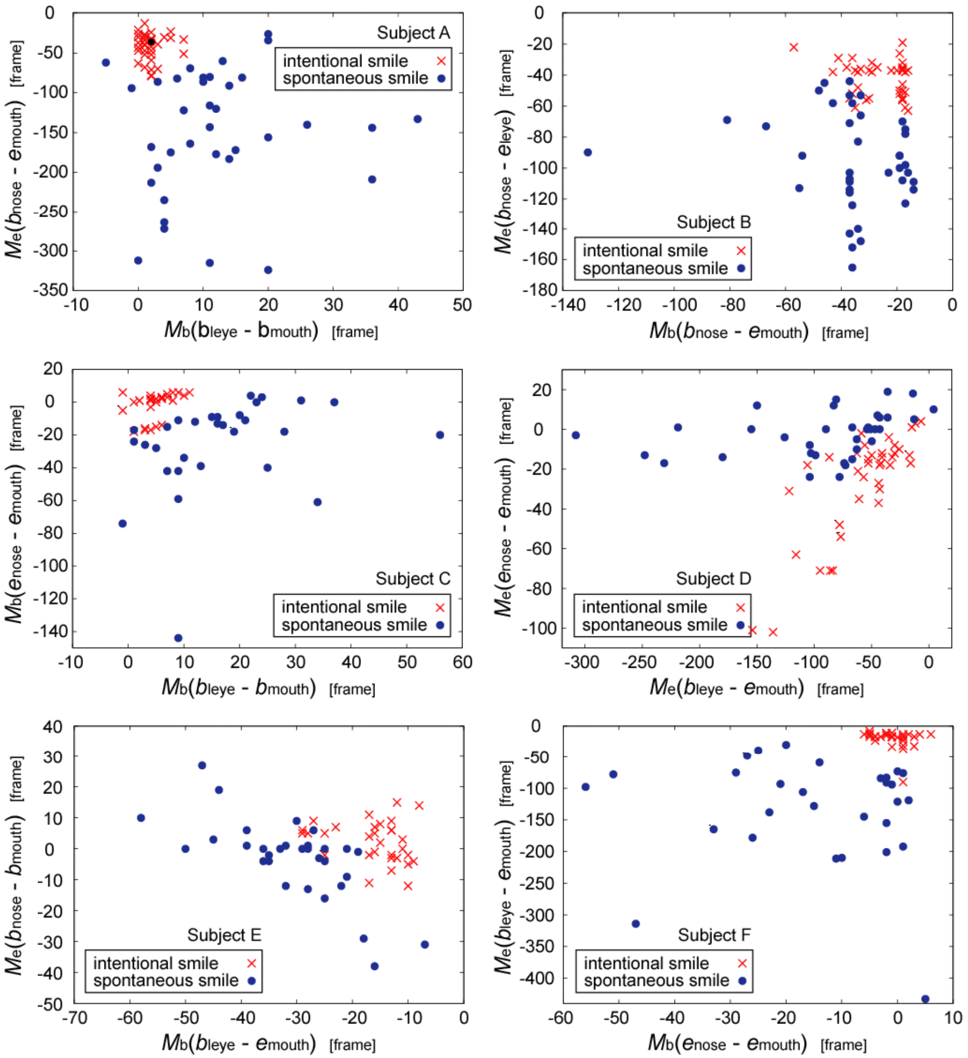
2. Calculate two-dimensional distributions using a combination of two temporal differences as the axes

3. Calculate distance between the distributions of two smiles for all the combinations in 2



Result of Discriminating Two Smiles

Chose two axes that provides the maximum distance between two distributions



Six subjects
intentional smile ×
spontaneous smile ●

Recognition rate of each subject
based on the support vector machine
(leave-one-out method)

subject	intentional (%)	spontaneous (%)
A	100	83.8
B	100	79.4
C	82.4	96.4
D	85.1	79.7
E	85.3	90.3
F	96.6	93.1

Discussion

- Analysis of timing structure in multipart motion of facial expression
  - Successfully discriminated and recognized intentional and spontaneous smiles

Future Work

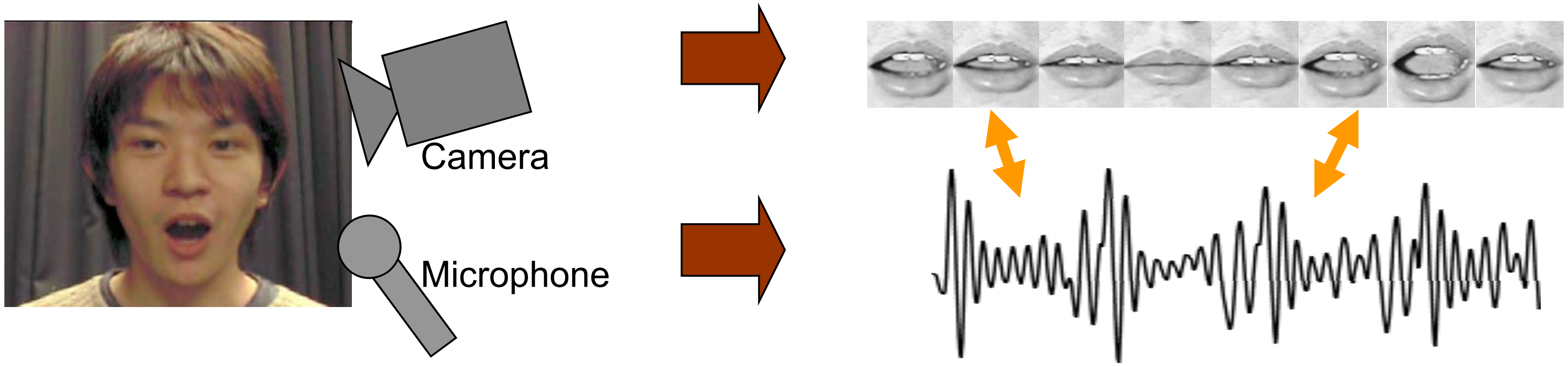
1. Long term observation (video capturing)
    - Find expression categories in a bottom-up manner
2. Expression in a context
    - conversation, singing, watching movies
    - relation among multiple subjects
3. Personality
    - Common structure and modes
    - Specific structure and modes

Chapter 5

Modeling Timing Structures in Multimedia Signals

Temporal Relation in Multimedia Signals

Multimedia signal



Recognition by multimedia integration (ex.) Audio-visual speech recognition

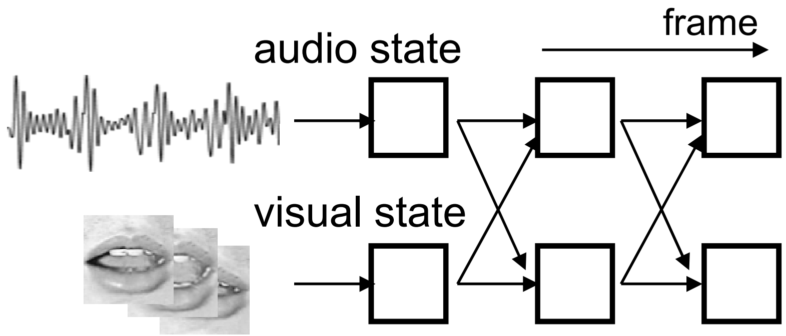
Speech → Audio, Video, Motion → Integration → Classification

Media signal generation from another related signal (ex.) Lip sync.

Audio → Media Conversion → Video, Motion

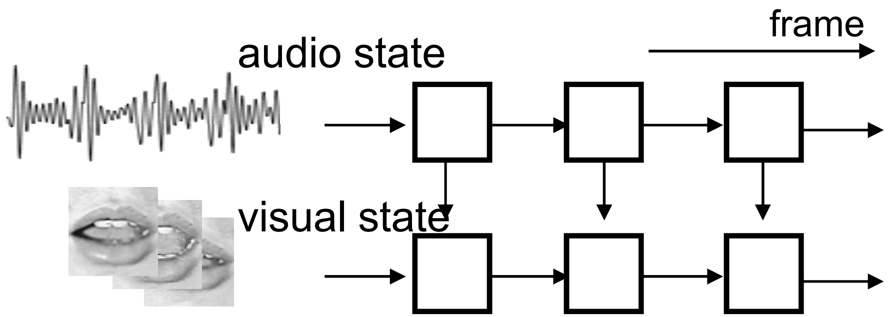
Related Work

State relation in adjacent frames



(ex.) Coupled HMM [Nefian, et al, ICASSP 2002]

Frame-wise state co-occurrence

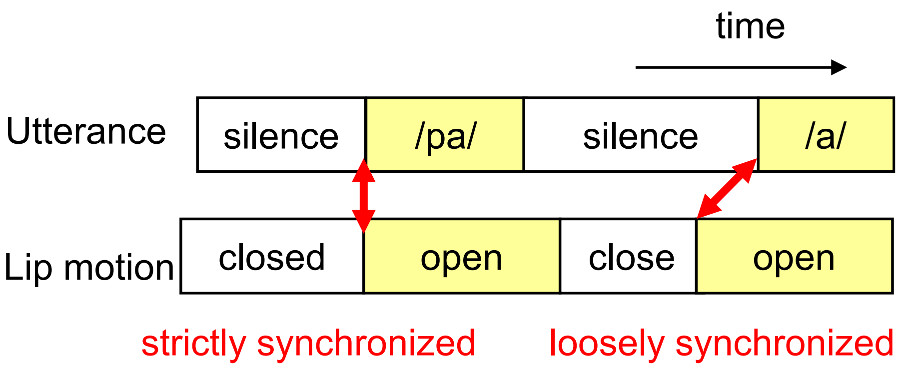


(ex.) Voice puppetry [Brand, SIGGRAPH 1999]

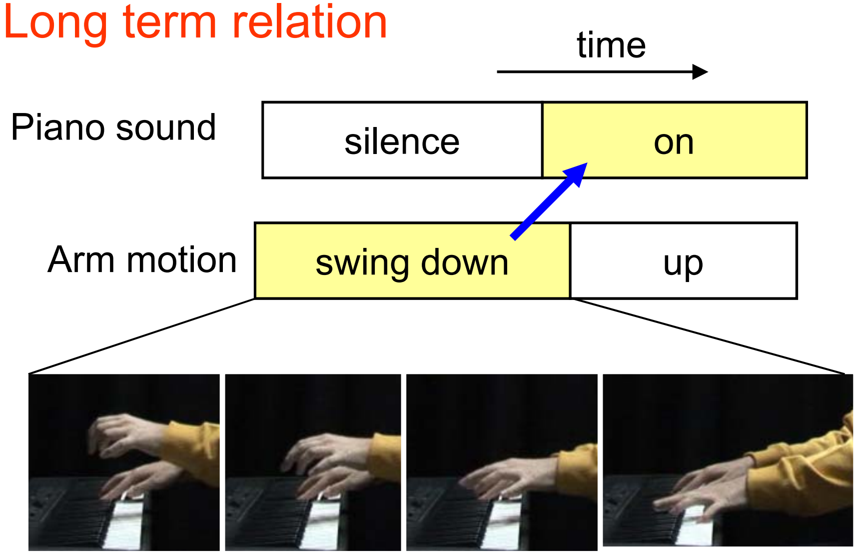
Frame-based

Open Issues

Synchronization mechanisms



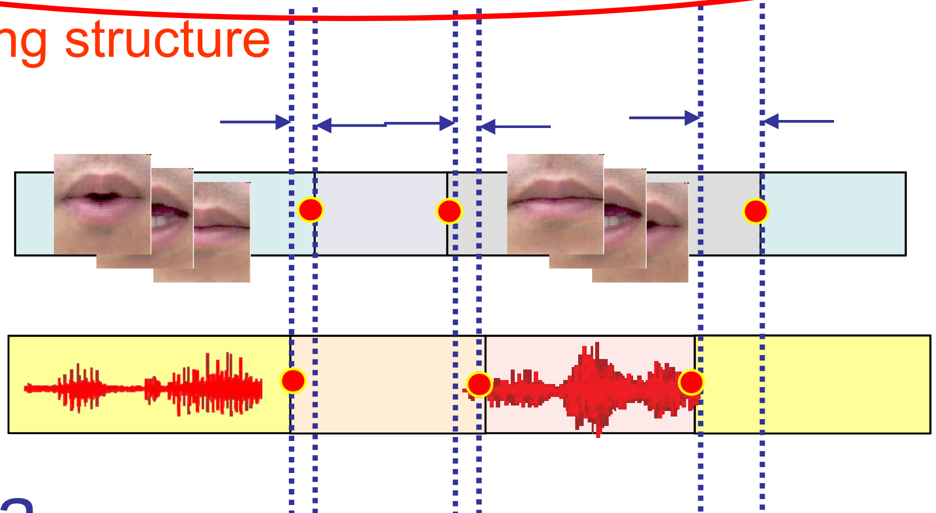
Long term relation



Timing Structure Model

There exists mutual dependency with organized temporal difference between signals

Timing structure



Key Idea

Directly model timing structure
using an interval based representation

1. How to divide signals into intervals?
2. How to model timing structure ?

Interval-based HDS

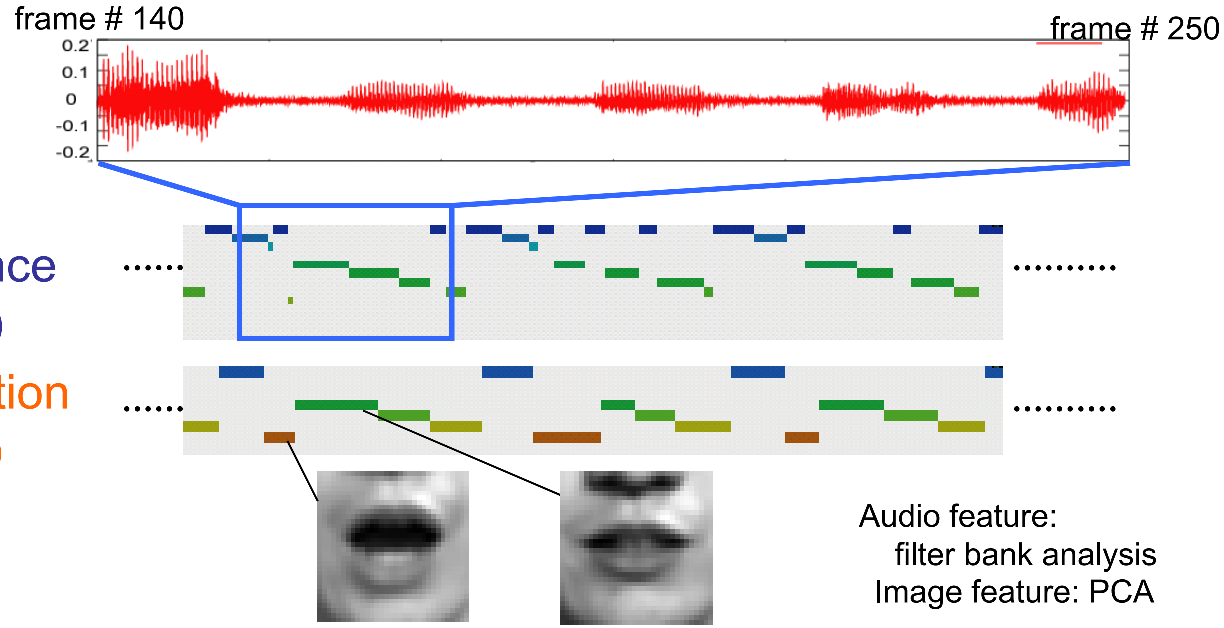
Temporal Relation of Intervals

Input speech:
/aiueo/
nine times
continuously

Segmentation result

Utterance (Audio)

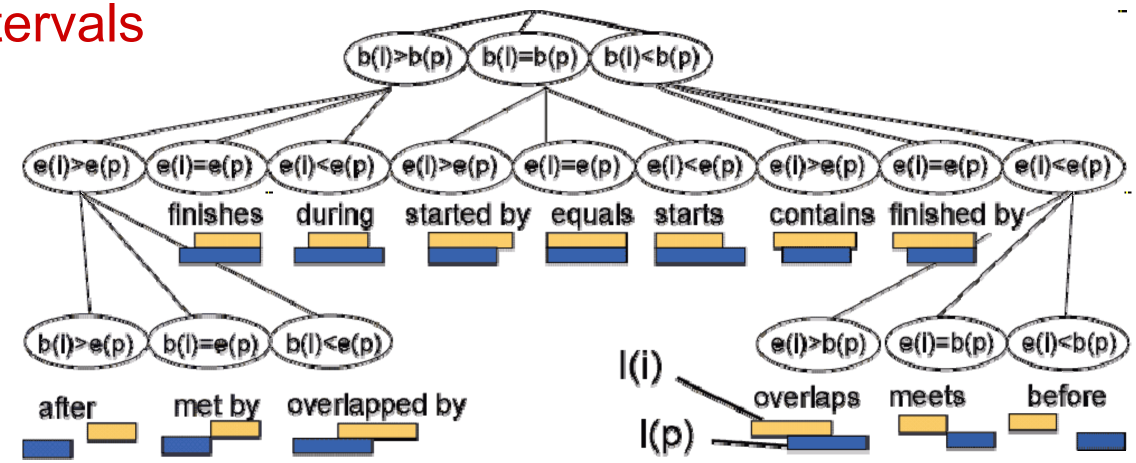
Lip motion (Video)



Audio feature:
filter bank analysis
Image feature: PCA

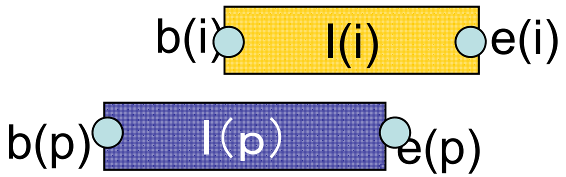
Temporal relation of two intervals
(without metric)

Focus on overlapped interval pairs

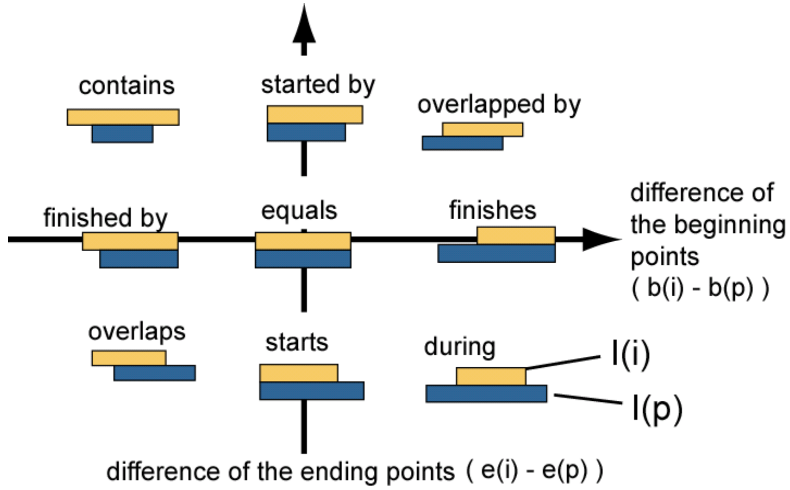


Metric Relation of Intervals

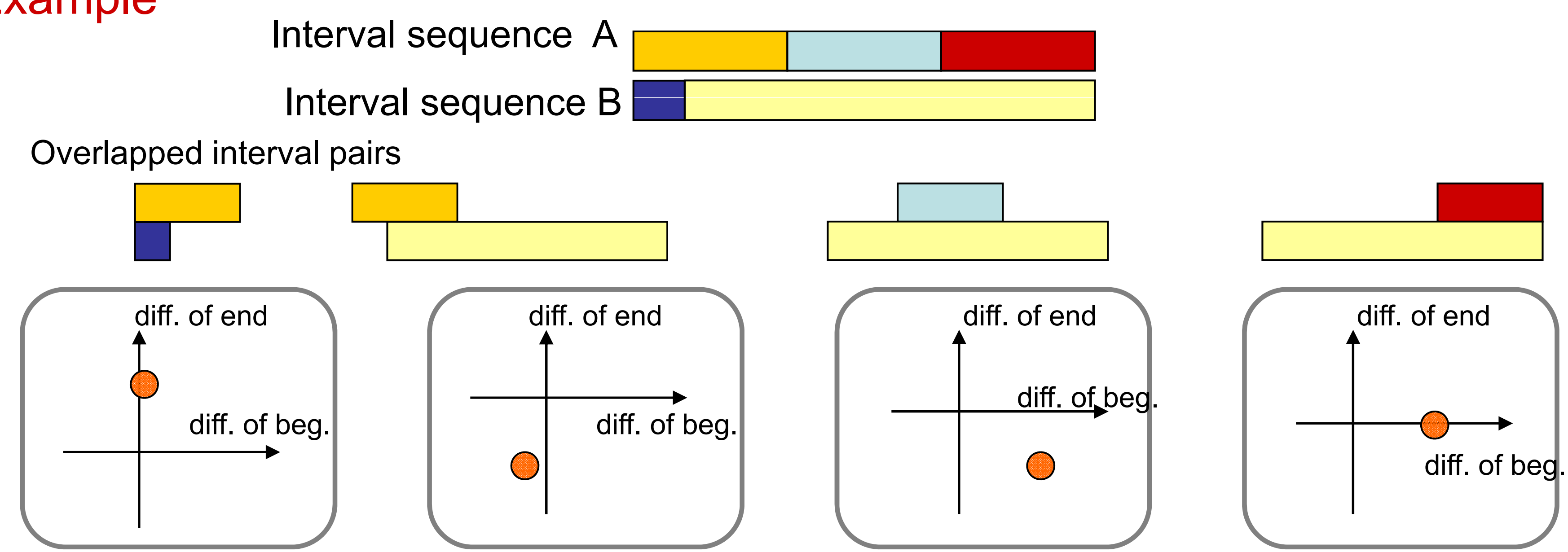
Overlapped interval pair



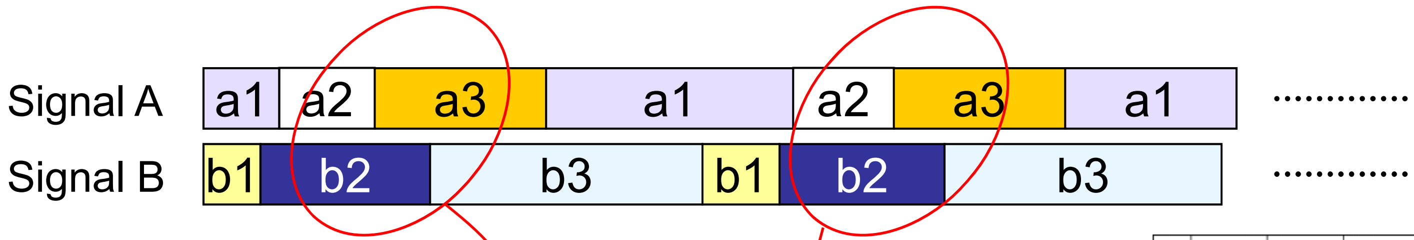
$$b(i) - b(p) \qquad e(i) - e(p)$$



Example

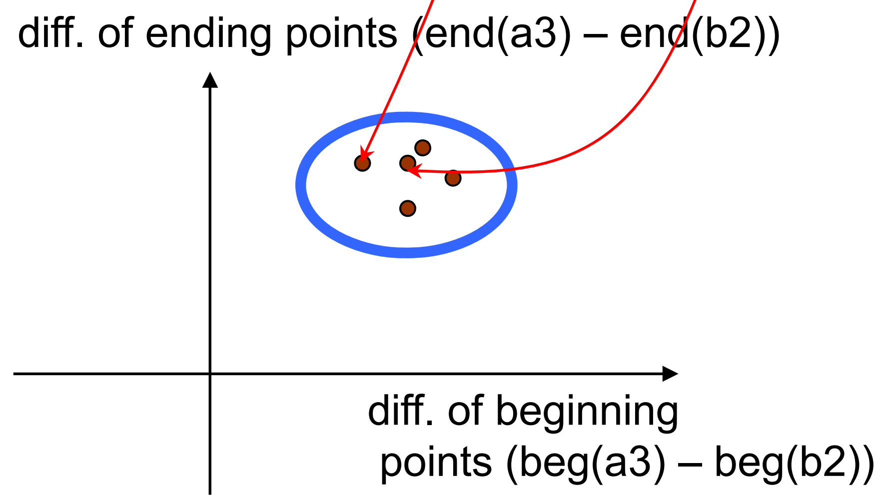


Timing Structure Model

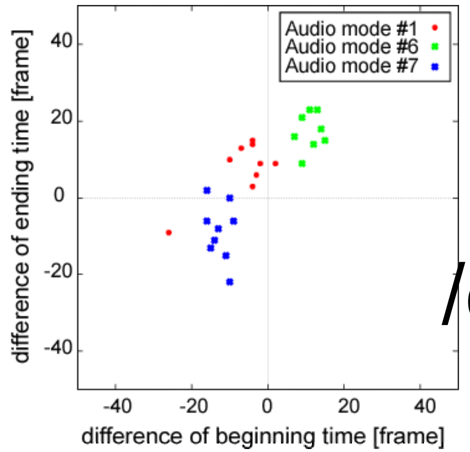


Learning: Prepare a distribution for each label pairs

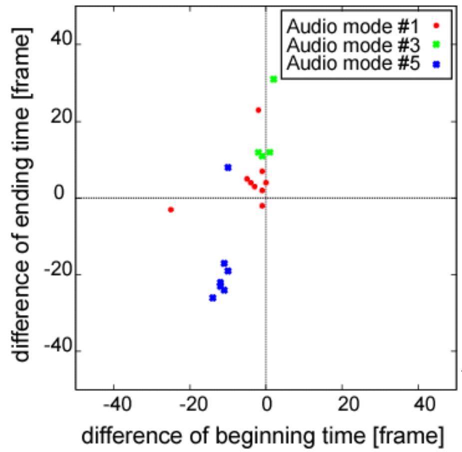
(ex.) **Temporal difference distribution** of pair (a3 b2)



Fit samples with a distribution function
(single Gaussian, mixture of Gaussian…)

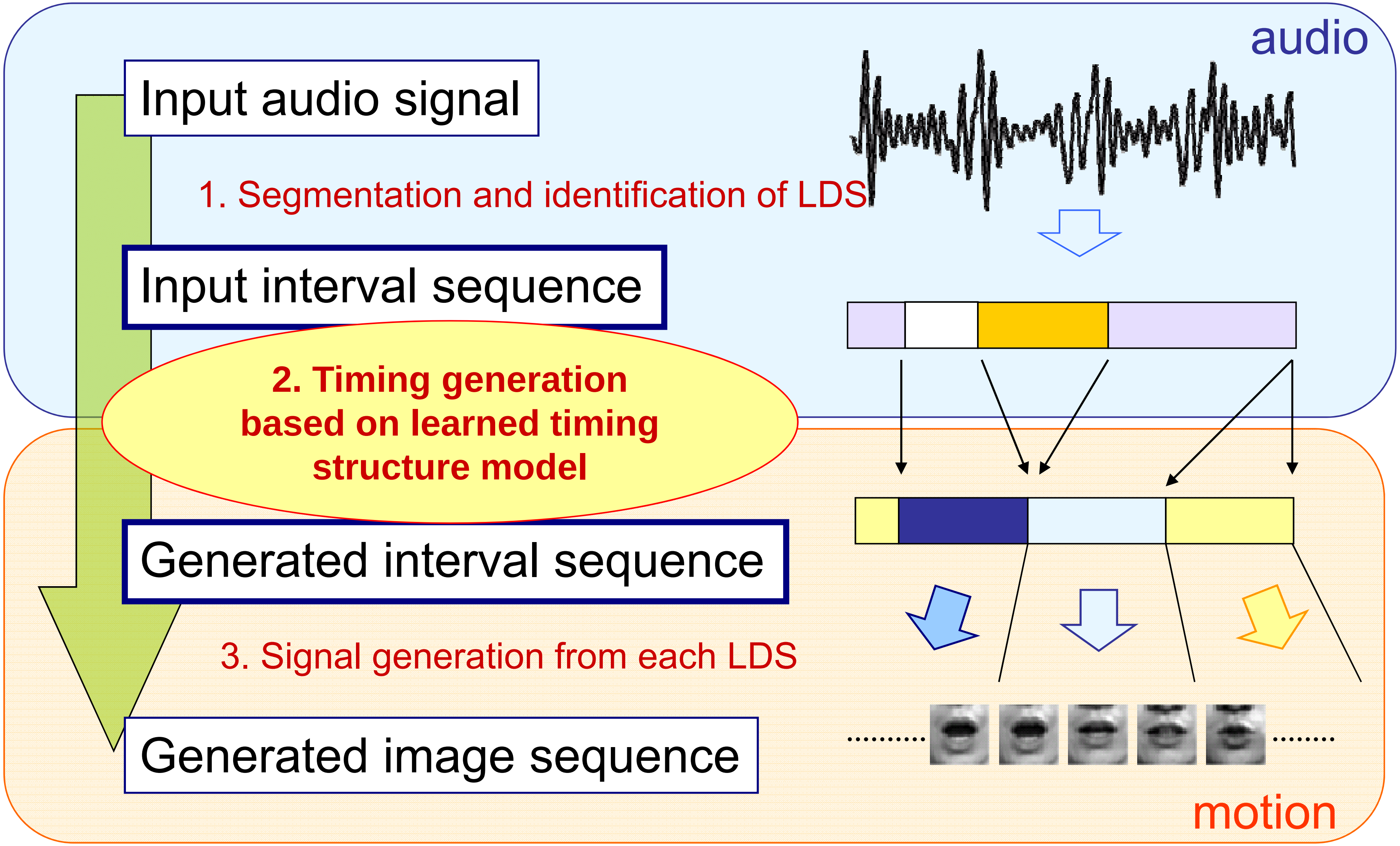


/o/

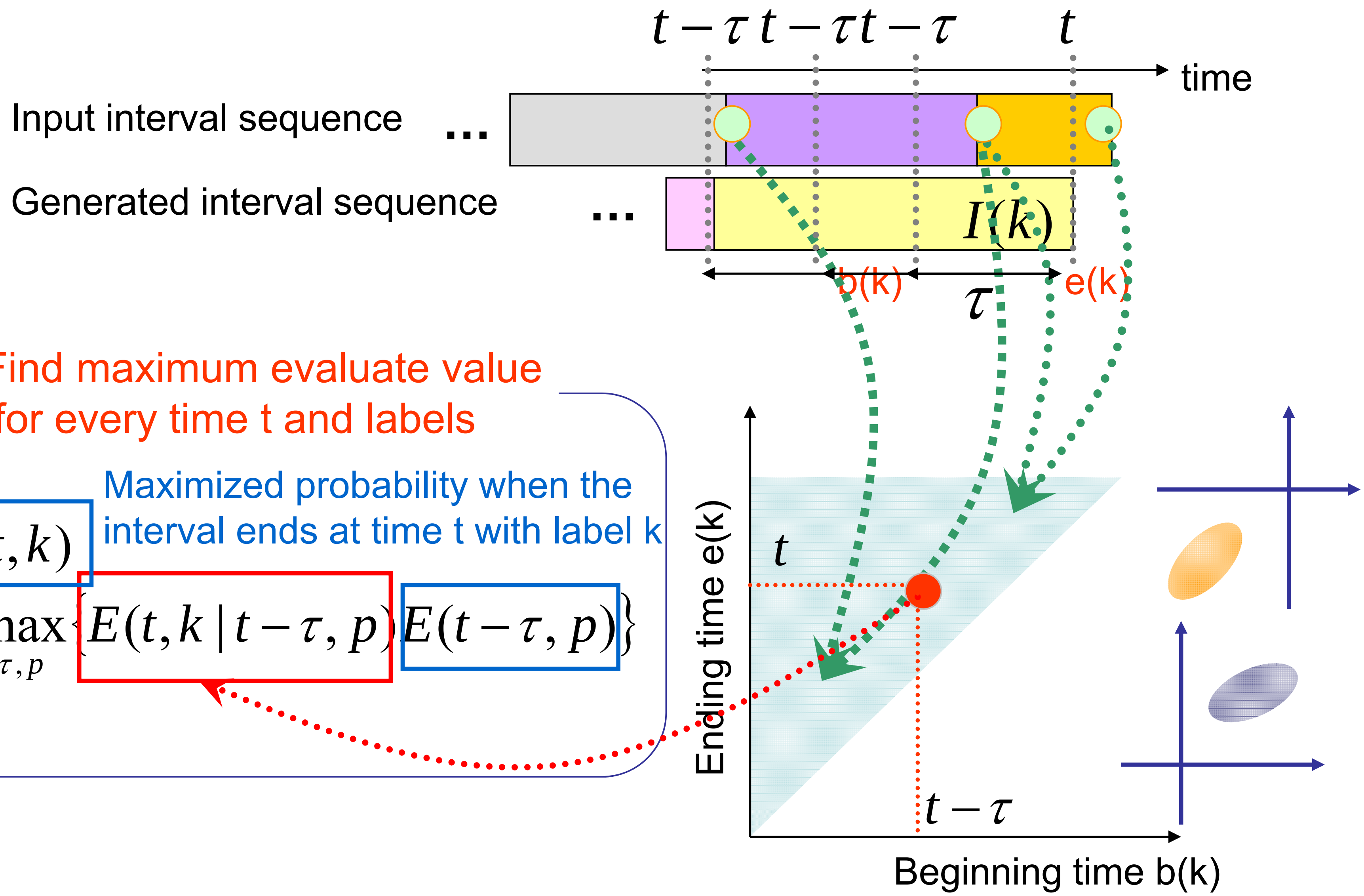


/i/

Media Signal Conversion



Timing Generation via DP Algorithm



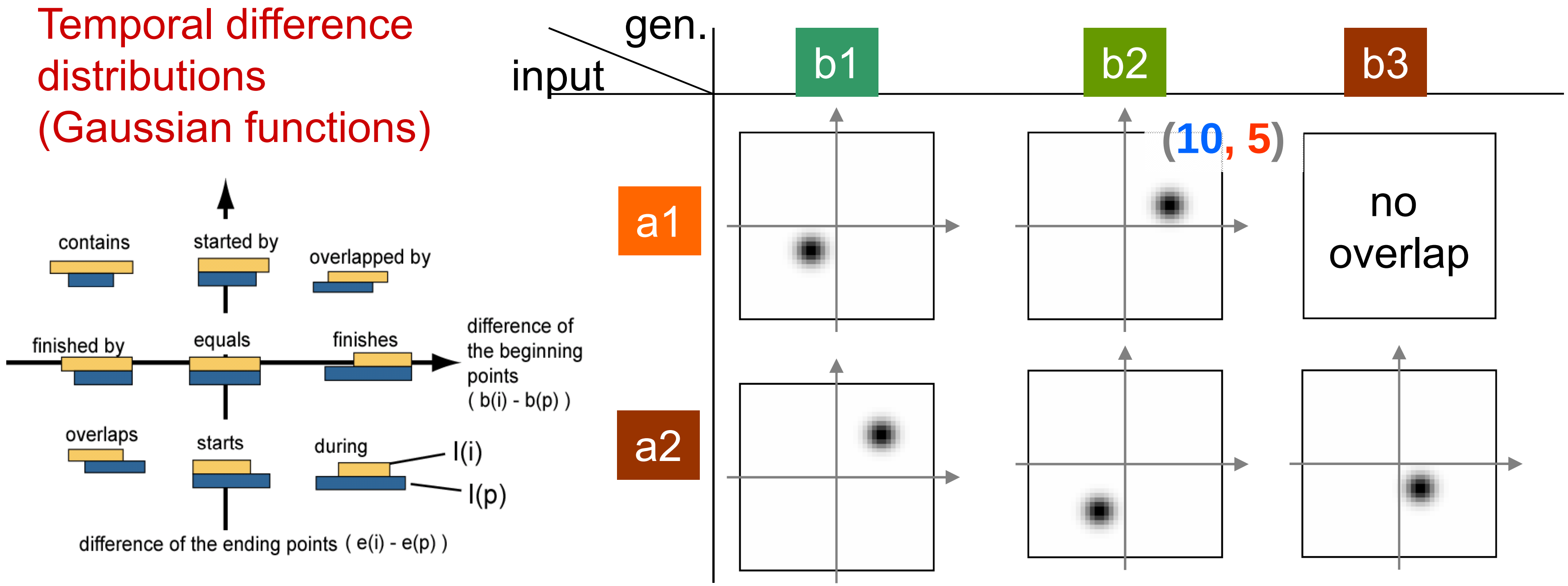
Find maximum evaluate value
for every time t and labels

Maximized probability when the interval ends at time t with label k

$$\boxed{E(t,k)} = \max_{\tau,p}\left\{\boxed{E(t,k \mid t-\tau, p)}\boxed{E(t-\tau, p)}\right\}$$

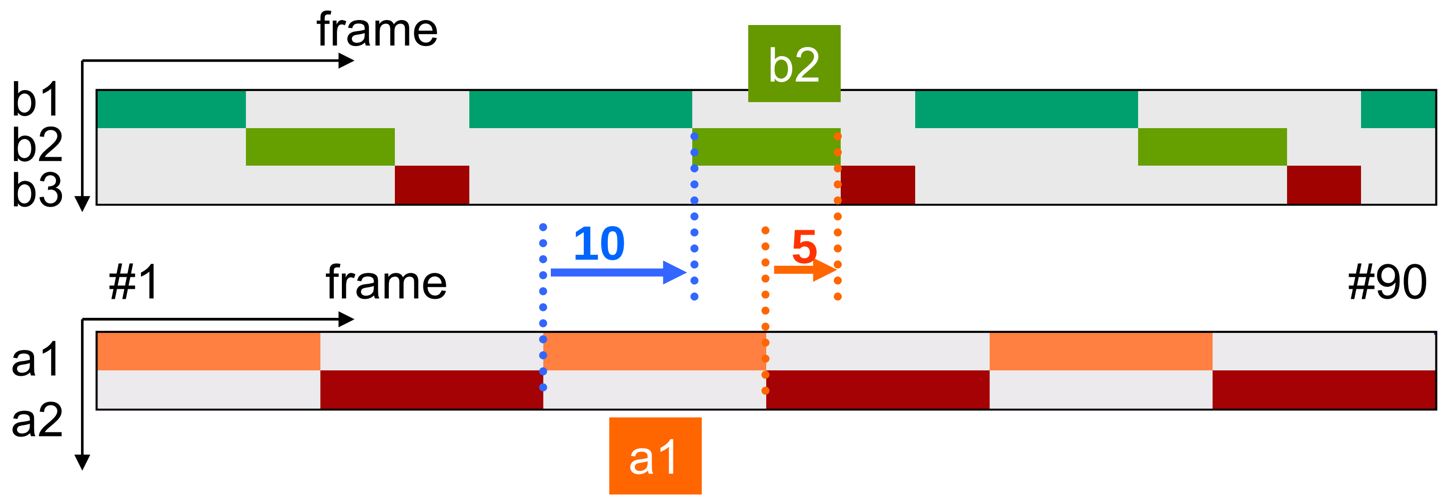
Verification of the Algorithm (Simulation)

Temporal difference distributions (Gaussian functions)



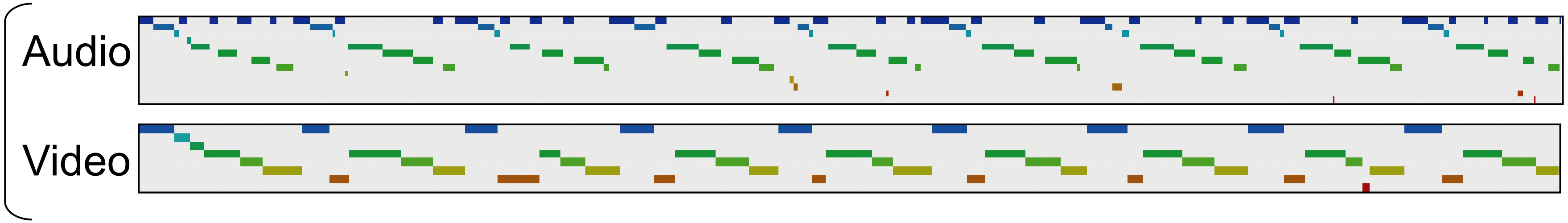
Generated interval seq.

Input interval seq.

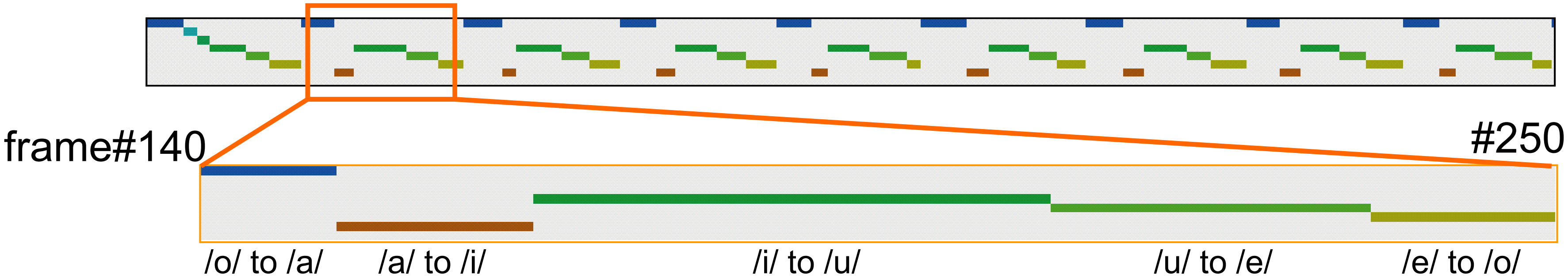


Lip Motion Generation from Audio Signal

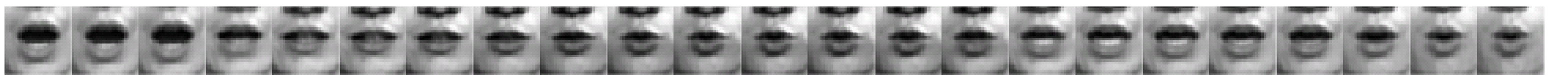
Training data



Generated interval sequence



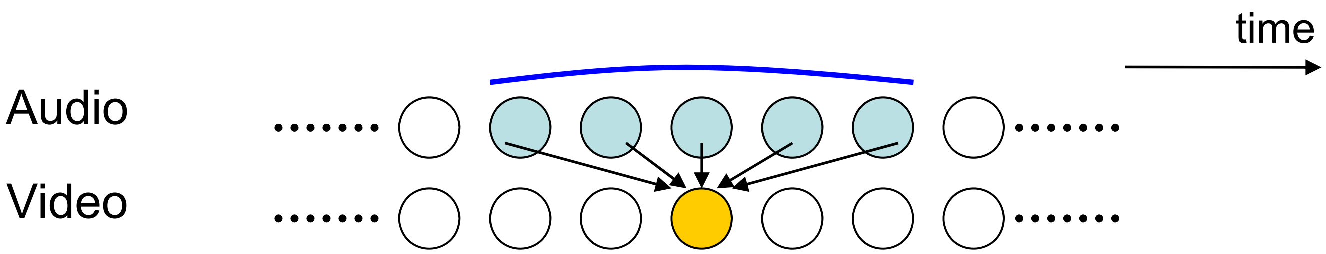
Generated image seq.



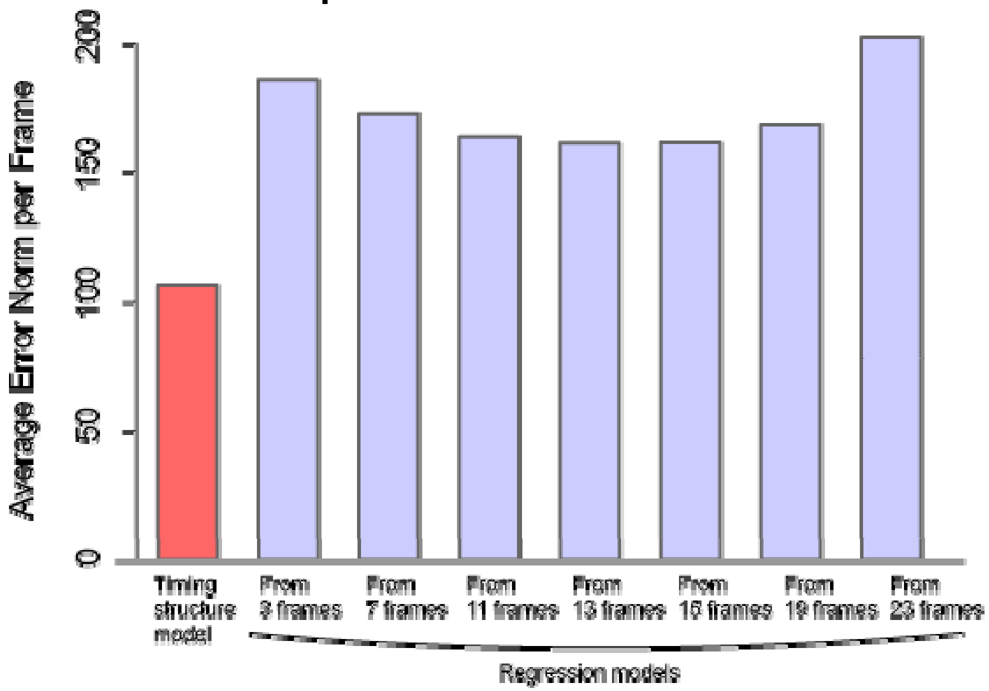
Original  Generated 

Comparison with Regression Models

- Linear regression models



- Average Error Norm per Frame



Pianist Motion Generation

Original

Generated

Discussion

- Timing structure model
  - Explicitly represents temporal metric relation between media signals
- Media conversion based on the timing structure model
  - Generates timing of one signal from other related media signals

Future work

- Apply to human-computer interaction
  (ex.)
  - audio-visual speech recognition
  - facial expression analysis
  - speaker detection in noisy environment
  - utterance timing generation for speech dialog system

Chapter 6

Conclusion

Summary

- Interval-based hybrid dynamical systems
  - integrate discrete-event systems (subjective time) and dynamical systems (objective (physical) time)
  - explicitly model temporal relations such as
    - tempo and rhythm in a signal
    - timing structure among different media signals

  based on *temporal intervals*

- Two-step learning method
  - Clustering of dynamical systems based on eigenvalue constraints
  - Refinement of parameters via the EM algorithm

Future Work

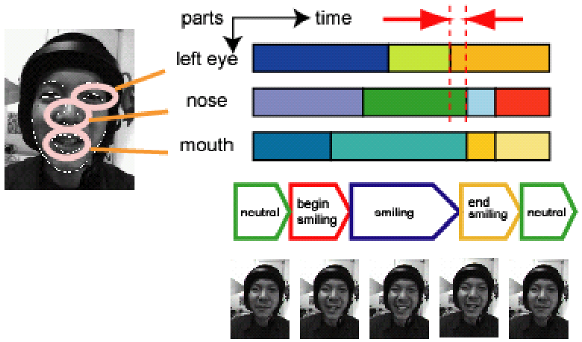
- Non-linear dynamical systems
  - Kernel method, neural networks
- Transition process between dynamics
  - Smooth signal generation
- Timing structure among more than three signals
  - Determination of causal relationship
  - Hidden interval sequence
- Hierarchical structures
  - Context-free grammar, hierarchical HMM
  - Variable length N-gram

Modeling Multiparty Interaction

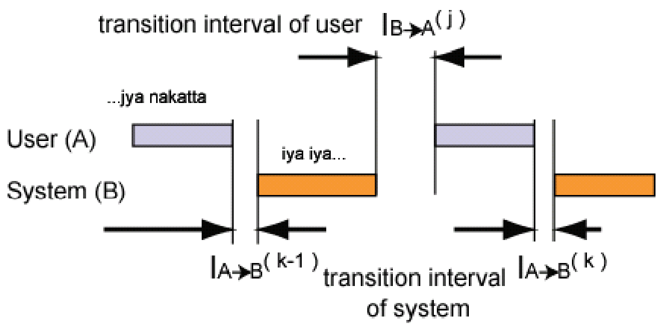
	Modeling Single Human Behavior from Multi-Channel Signals	Modeling Multiparty Interaction
Single-Media Signal	Analysis of Multipart Events in Single Modality (Chapter 4)	Modeling Utterance Timing in Speech Conversation
Multi-Media Signal	Modeling Multimodal Events (Chapter 5)	Modeling Multimodal Interaction

Single-Media Signal

Analysis of Multipart Events in Single Modality (Chapter 4)

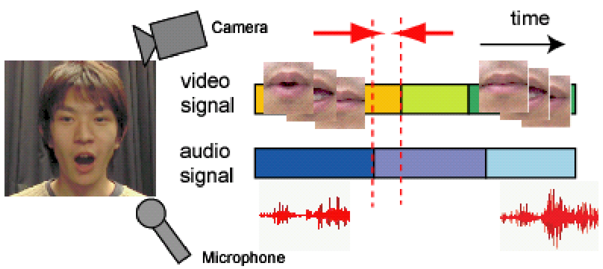


Modeling Utterance Timing in Speech Conversation



Multi-Media Signal

Modeling Multimodal Events (Chapter 5)



Modeling Multimodal Interaction