Chapter 1

Introduction

1.1 Modeling Dynamic Events

Understanding dynamic situations and performing appropriate behaviors in the situations is indispensable mechanisms for biological systems to survive in the real world. The mechanisms are also crucial for artificial information systems to realize intellectual functions, such as recognizing dynamic scenes, predicting events in the scene, controlling themselves to be desirable states, and providing useful information to users.

To understand dynamic situations in the real world, systems measure multiple time-varying signals from multiple sensors. The sensors can be a set of cameras, microphones, and tactile sensors. From the acquired multimodal signals, the system first recognizes "where" and "what objects" exist in the scene (object recognition), and then recognizes "when" and "what kind of / how" dynamic events have occurred or are occurring (event recognition).

In general case, these two processes of object and event recognition can be done in parallel. Biological motion in human perception analyzed by moving light displays [Joh73], or its interpretation systems [Ras80, CS94] are typical examples where motion itself plays an important role to determine objects (e.g., arms and legs). However, to concentrate on temporal aspect of event recognition, we here assume that object recognition is done beforehand.

Similar to event recognition, systems determine "when" and "what kind of / how" dynamic events should be generated in response to recognized events from dynamic situations to perform appropriate behaviors. For instance, one of the most important issues in robotics is the method to determine the activation patterns of actuators in the situations in which the robot body contacts with envi-

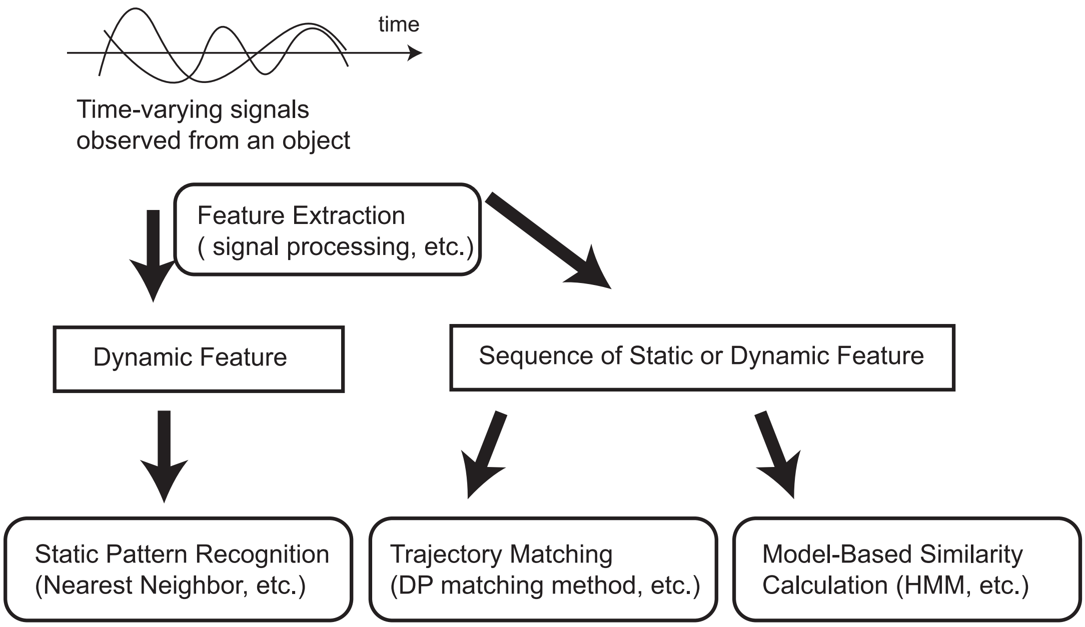
time
Time-varying signals observed from an object
Feature Extraction (signal processing, etc.)
Dynamic Feature
Sequence of Static or Dynamic Feature
Static Pattern Recognition (Nearest Neighbor, etc.)
Trajectory Matching (DP matching method, etc.)
Model-Based Similarity Calculation (HMM, etc.)

Figure 1.1: Flow of Event Recognition.

ronment and interacts with human.

As for the concrete methods of event recognition and generation, there exist many studies in wide variety of areas that cope with real-world problems; for example, computer vision, speech, and robotics [AC99, Gav99, Nak00]. Especially, the methods of event recognition, which we mainly handle in this thesis, can be categorized into several cases based on the flow of information processing shown in Figure 1.1:

- The use of static pattern recognition methods with dynamic features.
- The use of trajectory-based temporal pattern matching.
- The use of state-space models for similarity calculation.

In the following, we briefly describe each of the cases above. As we will introduce there, the model proposed in this thesis can be categorized as one of the state models (or state-transition models). Since we will discuss state models in details as it becomes relevant (in Section 1.2), in this section, we give only a concise comparison among the event recognition techniques.

Static Patter Recognition Methods with Dynamic Features

The event recognition methods that utilize static patterns of dynamic features have been introduced since the dawn of gesture and activity recognition studies. Once a set of time-varying signals is given as an observation of the object (e.g., a set of pixels from each of captured images), these methods extract dynamic features based on signal processing; for example, spectrum of signals, optical flow fields of visual features, and temporal integral of signals (e.g., motion-energy or motion-history images) can be dynamic features. The methods then exploit extracted dynamic features with classical static pattern recognition techniques, such as nearest neighbor methods and template matching, and classify the category of the input signals [PN94, NA94, EP97, BD97, ZMI01].

Since these approaches utilize dynamic features that become interfaces between time-varying signals and static pattern recognition techniques, the user can select variety of pattern recognition methods once the dynamic features are extracted. However, the approaches are sometimes sensitive to spatio-temporal fluctuation of input signals because most of the methods are difficult to represent the variance of signals (e.g., time warping). In addition, these methods often require problem-specific signals, such as periodic signals in gait motion, and restricted to use in a narrow domain.

Trajectory-Based Temporal Pattern Matching

Straightforward methods to recognize events from the input time-varying signals or temporal sequences of features, which we refer to as observations or observed sequences, are the use of dynamic programming (DP) matching methods. In these methods, a prototypical (or reference) temporal pattern (trajectory) is selected for each of the event categories in advance of recognition. In the recognition phase, similarity between an observed sequence and each of the prototypical patterns in every category is calculated based on DP matching. The DP matching methods have been used in speech recognition and gesture recognition systems [DP93, TSKO94, NMN97, KP98].

Although these methods enable us to use sequences of static features on behalf of dynamic features, and to search simple temporal patterns with inexpensive computational cost, we often fail to model the inner-class variation of patterns. This is because the prototypical pattern does not have expressive power to represent distribution of patterns. The expression of inner-class variation is often essential to realize speaker independent speech or gesture recognition systems. As a result, state-space models described in successive paragraphs are widely used in the current recognition systems that require expressing inner-class variation. In fact, some of DP matching methods can be deduced as a special case of the state-space models.

Model-Based Dynamic Event Recognition and Generation

State models represent temporal structure of events based on the change of the states apart from observation space (i.e., feature space or original signal space). Hidden Markov models (HMMs) and differential equation systems are well-known models widely used in speech and visual motion recognition fields. Thanks to the states, the models can represent variation of patterns in the observation space by defining mapping functions between states and observations. In the recognition phase, one of the states is activated by the combination of observed data and the previous state. Since state-space models maintain memory as the activation of the states, they can successfully track an observation sequence, and classify the sequence into one of the event categories. We can also use the model to generate time-varying signals if the model is a generative model.

The model we propose in this thesis is categorized as one of the state models. The state and its transition are however designed by the integration of two different system concepts: *dynamical systems*, which is suitable for describing physical phenomena (consider time as physical metric entity), and *discrete-event systems*, which is suitable for describing human subjective or intellectual activities (consider time as ordinal state transition).

In the remaining of this chapter, we first describe two different "concepts of time" in Section 1.2, and show the two kinds of systems, each of which represents "events" based on different concept of time. In Section 1.3, we describe an intuitive concept of integrating the two systems, and introduce some existing studies about the methods of the integration. The main idea of our proposed system is described in Section 1.4, where we introduce the use of "temporal intervals" to integrate the two different concepts of time, and show the target situation of the proposed system where we focus in this thesis. Finally, we show the overview of this thesis in Section 1.5.

1.2 Subjective Time and Physical Time

1.2.1 Definition of Events and Time

Before describing two different system concepts in state models, we first define the terms and notions of event and time, which are significant to differentiate between the two system concepts.

Discrete Events and Dynamic Events

The term *event*, which we used in the previous section, has two different notions, each of which has been used in different ways. One is considered to occur at an instantaneous point in time (e.g., "switch the coffee maker on") [TKT$^+$00, KCB95], and the other is considered to occur in a temporal region and have a duration length (e.g., "travel from A to B") [All83]. To be precise, the former event does not necessarily occur at a single instant in time, however, the duration of the occurrence is negligible or nonessential. On the other hand, the continually-changing patterns of objects are essential in the latter event.

In this thesis, we therefore use the following terms to distinguish the notions of event:

Discrete events. Discrete events are the occurrence of something that are independent of temporal metric; the discrete events take discrete values (i.e., elements in a finite set). In this thesis, we assume that the discrete events occur instantaneously in time similar to delta functions.

Dynamic events. Dynamic events are the occurrence of something that is described by time-varying signals and has physical energy in the real world; thus, dynamic events have duration lengths. The sensors can convert dynamic events as trajectories of observable signals in the space that have spatio-temporal metric based on the exchange of the energy (e.g., optical-electrical conversion).

Comparing the above definitions with the existing studies in artificial intelligence, discrete events correspond to the actions used in *situation calculus* [MMM69], which represents a temporal (causal) relation between an action of an agent at one situation and the result of the action at the next situation in time. Whereas the situation calculus works well when there exists a single agent performing instantaneous or discrete actions, it fails to represent duration of actions or temporal structures among actions of multiple agents, which can be overlapped with each other.

Alternative formalisms to represent these temporal structures are known as *event calculus* [KS86], Allen's *interval-based temporal logic* [All83, All84], and so on. These approaches assume that events have duration lengths in time; therefore, those events used in the approaches correspond to dynamic events defined here.

Subjective Time and Physical Time

In ancient Greek, there were two different concepts of time described by two words: *kairos* and *chronos*, which are the names of Greek gods originally. Kairos is the moment or occasion of making meaning; kairotic time is measured by discrete events. For example, a marriage and childbirth is a kairotic time (moment). Chronos, on the other hand, is the time that flows linearly; chronological time is a concept of time that describes the continual change of dynamic events, which are measurable by clocks (e.g., Newton's laws of motion).

Generalizing the notions of kairos and chronos, we define the following two kinds of time.

Subjective Time (Kairos): temporal order among recognized discrete events. Let $\mathbf{T}_s$ be a countable set that has no mathematical structures, such as distances and relations, among the elements. The set $(\mathbf{T}_s, \leq)$ becomes a subjective time axis, where $\leq$ denotes an ordered relation between elements in the underlying set $\mathbf{T}_s$. We often use a set of linearly ordered natural numbers $\mathbf{N}$ for $\mathbf{T}_s$.

Physical or Objective Time (Chronos): metric entity that linearly progresses. Let $\mathbf{T}_p$ be a set that has no mathematical structures. The set $(\mathbf{T}_p, \leq, d)$ becomes a physical time axis, where d denotes a distance function between two elements in the underlying set $\mathbf{T}_p$ (e.g., $d(t_1, t_2) = |t_1 - t_2|$, where $t_1, t_2 \in \mathbf{T}_p$). We often use set $\mathbf{R}^+ = \{t | t \geq 0, t \in \mathbf{R}\}$ for $\mathbf{T}_p$, where $\mathbf{R}$ is a set of real numbers.

Note that the most significant difference of physical time from subjective time lies in the existence of metric property.

1.2.2 Discrete-Event Systems and Dynamical Systems

Based on the two concepts of time described in the previous subsection, we can categorize existing state models, which are used for event recognition and generation, into following systems:

- Discrete-event systems
- Dynamical systems

As we will describe in this subsection, each of discrete-event systems and dynamical systems defines its state transition based on subjective time and physical time, respectively.

The integrated system of these two systems is referred to as *hybrid dynamical systems*, as we will describe in Section 1.3.

Discrete-Event Systems

A discrete-event system has a set of *discrete states*, which is represented by a finite set, and it does not change the discrete state before an input discrete event occurred. Therefore, the state transition is described based on the subjective time. The simple case of the state transition becomes the following function, which is used in finite state automata:

$$M(\text{state}_{\text{now}}, \text{event}_{\text{input}}) = \text{state}_{\text{next}} \qquad (M : Q \times A \rightarrow Q), \tag{1.1}$$

where A and Q is a set of discrete events and discrete states, respectively.

The representative model of discrete-event systems is the Turing machine, which initially models "a man in the process of computing a real number" based on the finite number of discrete states (configurations) [Tur36, Tur50]. Finite state automata [KCB95, BW97, WBC97, MM98b, WM00], HMMs [Rab89, HAJ90, Nak00, YOI92, SP95, BOP97], and Petri nets [DAJ95] are examples of discrete-event systems that are widely used for modeling structure of discrete events.

Discrete-event systems have the advantage of being able to model long-term contexts, relations of temporal order, overlaps, and inclusion among events [PMB97], and other discrete structures such as coupling between action and perception [KCB95]. However, discrete-event systems have signal-to-symbol problems; that is, it requires us to define an event set in advance. Moreover, human-designed states depend on our recognizable event scales. Therefore, the continual changes produced by smooth dynamics are difficult for discrete-event systems to describe.

Dynamical Systems

In contrast to discrete-event systems, dynamical systems define the state transition based on the physical time using the formulation of differential or difference equations. The simple case of the differential equation becomes:

$$\frac{\mathrm{d}\,\mathrm{state}(t)}{\mathrm{d}t} = F(\mathrm{state}(t)) \qquad (F : \mathbf{R}^n \to \mathbf{R}^n), \tag{1.2}$$

where $t \in \mathbf{R}^+$ and n is the dimensionality of the internal (continuous) state space, where *internal states* are defined. Note that the system changes the state even if there are no input signals.

Cybernetics, which had been advanced by Norbert Wiener since 1940s, is the study of "teleological mechanisms" involving regulatory feedback in animals (living organisms) and machines [Wie61]. Cybernetics influenced wide area of automatic systems including dynamical systems in control theory [AM79]. Gaussian and non-Gaussian linear dynamical systems [Rao97, IB98, RB99, BCMS01, DCWS03, DD05] are often used for modeling dynamic events that have physical dynamics. As for nonlinear dynamics, recurrent neural networks [FH88, Rob94, Dor96, Mor96, UT00, HWK01] and other nonlinear dynamical systems [dFNG98, GR99, OTN02] are used for modeling complex events such as robot motion.

Dynamical systems have the advantage of modeling continually-changing dynamic events such as human motion and utterance. However, the systems are not suitable to represent complex structures of signals, such as duration lengths of dynamic events, patterns of the duration lengths, and other temporal relations exist in multiple dynamic events that occur concurrently.

1.3 Hybrid Dynamical Systems

Hybrid dynamical systems (hybrid systems) that integrate dynamical systems and discrete-event systems are introduced to overcome the disadvantages of the two systems in a complementary manner.

In the following subsections, we first give a basic idea of hybrid dynamical systems to see how the disadvantages can be solved by the interaction between

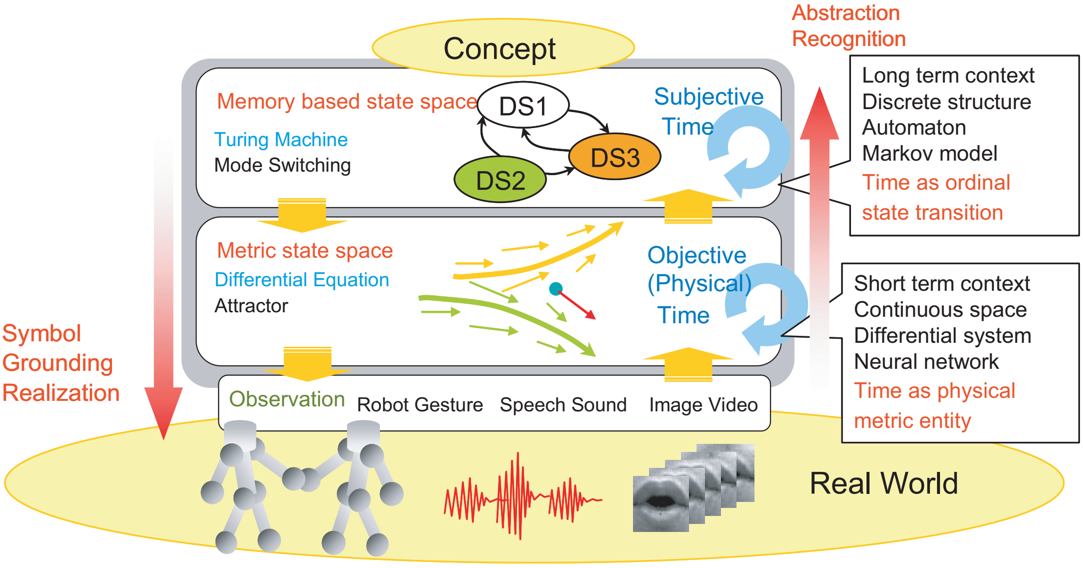

Figure 1.2: The concept of hybrid dynamical systems.

discrete-event systems and dynamical systems that constitute the overall system (Subsection 1.3.1). We then introduce some existing approaches that integrate discrete-event systems and dynamical systems, and discuss the idea of the integration in most of the approaches are different from integrating the two different concepts of time (Subsection 1.3.2).

1.3.1 Interaction in a Hybrid Dynamical System

Figure 1.2 shows the concept of hybrid dynamical systems. In a hybrid system, a discrete-event system decides the activation timing of multiple dynamical systems. The discrete-event system provides a solution to represent a discrete structure of primitives; meanwhile, the dynamical systems represent detailed dynamics in each primitive and also provide metric properties among the primitives. A typical interaction among the discrete-event system and the dynamical systems in a hybrid dynamical system is as follows:

1. Dynamic events in the real world, such as speech utterance, lip motion, and human (robot) motion, are measured as multimedia signals.

2. At the boundary of the hybrid dynamical systems and the real world, dynamic primitives are represented by various attractors in the internal state space based on differential equations. Each dynamical system, therefore,

changes the internal state in the physical-time domain based on observed time-varying signals. This process can be regarded as the “resonance” of dynamical systems with observed signals, and the signals are partitioned into temporal intervals, where each partitioned interval corresponds to the dynamical system that resonated the most with the observed signal.

3. The segmentation result of the observed signal determines the macro-transition of the discrete states, which can be considered as the flow of subjective time in the discrete-event system. If the probability of each state transition is given in advance, the discrete-event system affects the activation order and timing of constituting dynamical systems.

4. Hence, the interaction between top-down and bottom-up information occurs simultaneously at the two boundaries; that is, the boundary between the real world and the constituent dynamical systems (see 2), and the boundary between the dynamical systems and the discrete-event system (see 3).

Due to the interaction above, each of disadvantages in discrete-event systems and dynamical systems can be solved as follows:

- The dynamical systems provide interfaces between signals in the real world and discrete states (symbolic entities) in the discrete-event system; this architecture resolves the signal-to-symbol problem of discrete-event systems.
- The discrete-event system represents the structures of discrete events that are produced by the constituent dynamical systems; thus, complex temporal relation among dynamics are described in this architecture.

1.3.2 Existing Approaches

Because of the high capability of modeling nonlinear and complicated events, hybrid dynamical systems are currently attracting great attention in various fields including controls, robotics, computer vision, graphics, neural networks, and other computer science fields. In the following paragraphs, we introduce some of the existing hybrid dynamical systems in these fields. Mathematical analysis of hybrid dynamical systems can be found in [GV89, MMdB$^+$91, ACH$^+$95, ZM95].

As we will see in the remaining of this subsection, the notion of “hybrid” differs among the studies. Note that most of the existing hybrid dynamical system focus on integrating discrete states, represented by symbols, and internal states, represented by continuous values, rather than integrating the two different concepts of time (i.e., subjective and physical time) described in Subsection 1.2.1.

Piecewise ARX Models

Piecewise AutoRegressive eXogenous (ARX) models are the ARX models that use piecewise linear (PWL) or piecewise affine (PWA) maps as the regression function [FMLM03, RBL04, KHS+04]. A PWL map constructs a nonlinear function $f(x)$ by partitioning the domain $\mathcal{X} \subset \mathbf{R}^n$ into several regions $\mathcal{X}_1, ..., \mathcal{X}_N$ with polyhedral boundaries, which are referred to as *guardlines*. In each region, a linear mapping function is defined individually, and they are switched by the condition of $x \in \mathcal{X}$. As a result, the mapping function becomes nonlinear as a whole.

Piecewise ARX models are a class of hybrid systems for which the switching law between the affine submodels is specified by the shape of the guardlines; thus, the model represents nonlinear signals due to the switching. The conditions of discrete-state transition in the model, however, can be regarded as static because they are determined beforehand based on the design of guardlines.

Switching Dynamical Systems

Bregler et al. [Bre97] proposed a multilevel modeling method of human gate motion based on an architecture of a hybrid dynamical system. The model comprises multiple linear dynamical systems as its subsystems, and an HMM that switches the constituent subsystems. As a similar approach to the Bregler's model, a switching linear dynamical system (SLDS), which switches linear dynamical systems based on the state transition of the HMM, have become a common approach for modeling complex dynamics such as human motion [GH96, PRCM99, PRM00] (see [Mur98] for the survey of similar models). The stochastic linear hybrid systems [LWS02, BHJT04] are also the extension of Breglar's model.

In these approaches, the discrete and internal states are integrated. However, the macro-transition between constituent subsystems (i.e., linear dynamical systems) is modeled in the same time axis as the internal-state transition of each subsystem (i.e., physical time axis). Assuming that the system consists of a set of subsystems $\mathcal{Q} = \{q_1, ..., q_N\}$, then the SLDS models the transition from subsystem q_i to q_j as a conditional probability $P(s_t = q_j | s_{t-1} = q_i)$, where t is synchronized to the internal-state transition in each subsystem. Some other method use particle filters on behalf of linear dynamical systems (Kalman filters) [BIR00], however the method also use the physical time to model the state transition in the HMM.

Segment Models

Segment models [ODK96] have been proposed in speech recognition fields as the unified model of segmental HMMs [Lev86, HAJ90] and other segment-based models [Mur02]. In contrast to SLDSs, segment models use *segments* as descriptors. Each of the segments represents a temporal region in which one of the discrete states is activated. Since a discrete state corresponds to a dynamic event such as phonemes and subwords, each of which is represented by a subsystem, the discrete-state transition of the segment model represents temporal order of dynamic events apart from the physical-time domain. Thus, the conditional probability of the state transition becomes $P(s_k = q_j | s_{k-1} = q_i)$, where k represents the temporal order of the subsystem activation. Motion texture [LWS02], which is proposed for motion generation purpose, can be also categorized as one of the segment models.

Since the transition between the subsystems is modeled independently from the physical-time domain, the model handles one aspect of integrating the concepts of physical time and subjective time. However, because this model is proposed as a unified framework of segmental HMMs, it focuses on modeling only state duration rather than complex temporal structures among discrete events. We will discuss the details of this point in the next section.

1.4 Interval-Based Hybrid Dynamical System

In this thesis, we propose a novel hybrid dynamical system that integrates the concepts of subjective and physical time by exploiting *temporal intervals* (intervals, in short) defined in this section. We refer to the system as an *interval-based hybrid dynamical system* (interval system, in short). Interval systems are similar to segment models in respect that the both models are able to describe the temporal order of dynamic events, each of which is represented by a subsystem, apart from the physical-time domain. However, the concept of interval systems are different from the segment models because we concentrate on modeling temporal structure among multiple discrete events extracted by constituent subsystems (i.e., temporally dividing points of complex dynamic events) rather than only modeling the duration lengths of dynamic events (i.e., temporally divided parts of complex dynamic events).

For the above reason, we use the term "intervals" instead of "segments". In other words, our motivation is bringing Allen's interval-based temporal logic [All83, All84], which exploits 13 topological relations between two intervals (e.g., meets, during, starts with, etc.), into the class of hybrid systems. Once the intervals are explicitly defined, we can fabricate flexible models to represent complex structures among multiple types of dynamics, which can be appeared concurrently in human behavior and interaction (e.g., tempo and rhythms of utterance, synchronization/delay mechanism of speech and lip motion, and action timing generation in response to input events in interactive systems).

In the following subsections, we first define the notion of an "interval" (Subsection 1.4.1), and show how the temporal structures that have vital information for human can be described by the intervals (Subsection 1.4.2). We then give a concept of an interval-based hybrid dynamical systems (Subsection 1.4.3), and finally we discuss the expressive power and the limitations (Subsection 1.4.4).

1.4.1 Definition of Intervals

The definition of "dynamic events" in Subsection 1.2.1 is independent of cognitive processes, however, significant dynamic events are perceptible units for some "cognitive subjects". We therefore define an "interval" as a temporal difference between the beginning and ending points of the dynamic event that is perceived by some cognitive processes of humans or artificial systems. The length of the interval corresponds to the duration of the perceived dynamic event in the physical-time domain (see Figure 1.3).

Regarding human cognition, those perceptible units are not restricted to the dynamic events recognized consciously. While humans are not able to be aware of some events, the unconsciously perceived events are often processed without awareness, and exploited to provide appropriate decision and action. For instance, as we learn techniques in football, the learner can be aware of primitive motions (dynamic events) constituting an overall kicking action, such as pulling the leg back and moving it forward. However, once the learner has acquired the skill of the action, he or she can provide the action without awareness of each primitive motions.

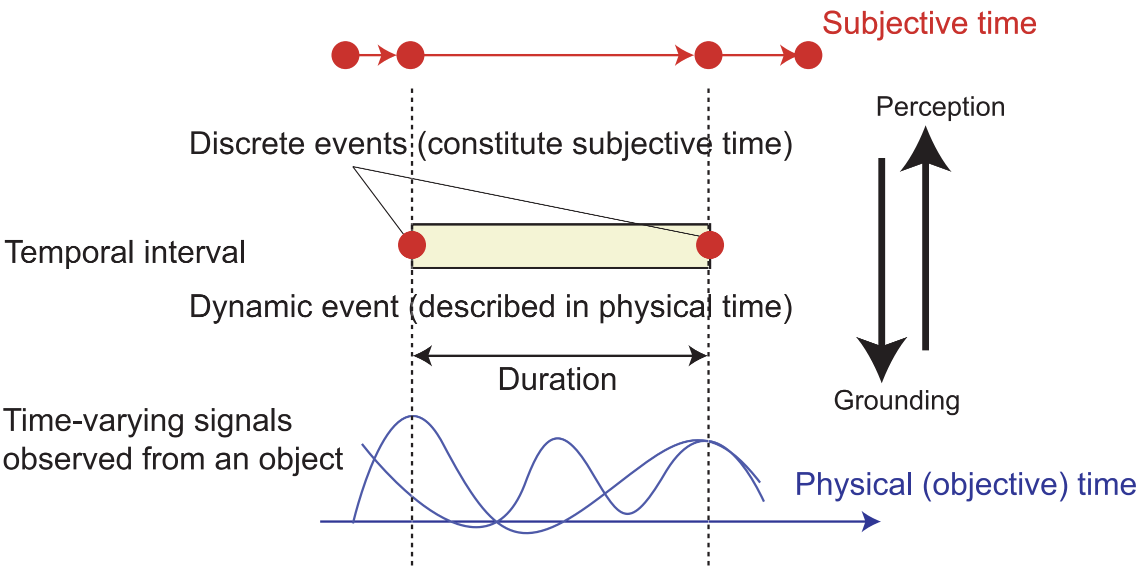
Subjective time
Perception
Discrete events (constitute subjective time)
Temporal interval
Dynamic event (described in physical time)
Duration
Grounding
Time-varying signals observed from an object
Physical (objective) time

Figure 1.3: Definition of a temporal interval.

Whether the humans are aware of the units or not, the beginning and ending points of perceptible dynamic events are essential to understand the temporal property of situations. We consider these instantaneous points in time as discrete events as shown in Figure 1.3; thus, a set of the discrete events constitutes subjective time, and temporal ordering relations among the discrete events become important for the situations incorporated by discrete-event systems as we described in Subsection 1.2.1. In the next subsection, we see how the significant temporal structures are described by the intervals.

1.4.2 Dynamic Structures Exploited by Humans

Temporal relations among discrete events (i.e., beginning and ending time points of perceptible dynamic events) have significant information for humans and artificial information systems to describe the situations of environments, to understand the meaning of object behaviors, and to generate actions in appropriate occasion.

Allen proposed an interval-based temporal logic to describe temporal relation among multiple actions that occur simultaneously and have many interact with each other [All83, All84]. The logic represents the relationships between temporal intervals based on temporal ordering relations of discrete events (beginning and ending time points) obtained from two intervals (Figure 1.4). As a result, it

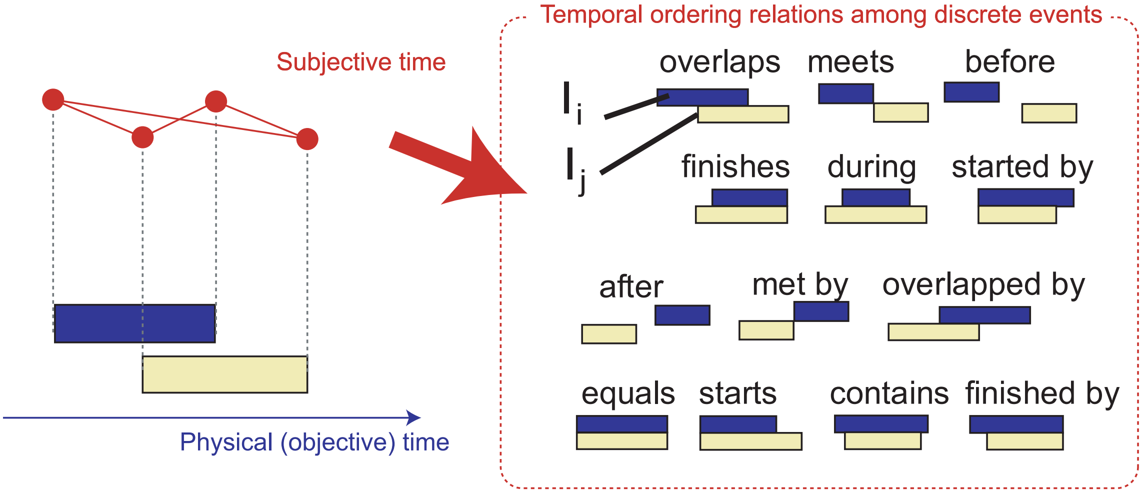

Figure 1.4: 13 temporal relations between two intervals I_j and I_j used in Allen's interval-based temporal logic, which exploits temporal ordering relations among discrete events (beginning and ending points of intervals) in the subjective time.

successfully represents temporal relations between multiple intervals in a hierarchical manner using constraint propagation techniques.

While temporal ordering relations among discrete events are indispensable to realize intelligent functions, we humans exploit not only temporal orders but also metric properties such as temporal differences among discrete events. In particular, these metric properties have crucial information for understanding temporal features appeared in the real world such as in verbal and nonverbal human communication, and for performing appropriate behaviors in complex environment.

In subsequent paragraphs, we see some examples of metric properties that we humans exploit.

Duration lengths of dynamic events. As we described in the previous subsection, each interval has a duration length as its metric property. Some psychological experiments suggest that duration lengths of facial actions play important roles for human judgments of basic facial-expression categories [KBM+01, KK05]. In addition, the duration lengths of stationary gaze often used to estimate his or her interest to the objects [WYH05].

Rhythms of dynamic events. Because the term "rhythm" is often used ambiguously, we need a definition of the term. Once an action is partitioned into

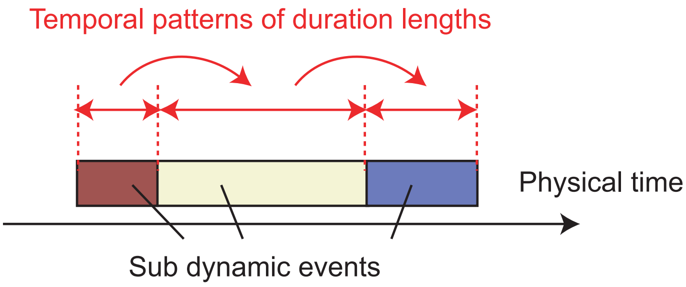

Figure 1.5: Rhythm of dynamic events observed in a single signal.

subactions, we obtain a sequence of sub dynamic events. For a simple definition, we here refer to patterns of duration lengths appeared in the sequence of subactions as rhythms (Figure 1.5). We humans are sensitive to the rhythms of not only music performances but also general events in various situations; for example, human gait motion (easy to detect an injury of others), swinging arm motion in communication (phasic gestures [WBC97]), and sports (feint motion in ball games to foil the others rhythmic prediction).

Synchronization among dynamic events. Intervals can be obtained from not only a single media signal but multiple media signals captured from multipart motion, multiple sensor modalities, and other situations; the intervals of concurrent dynamic events therefore can be overlapped each other. In these case, temporal differences between beginning points or between ending points among dynamic events often become significant in some situations (e.g,. synchronization/delay mechanisms among dynamic events) (Figure 1.6). For example, it is well-known fact that the simultaneity between auditory and visual patterns influences human perception (e.g., the McGurk effect [MM76]). Synchronized motion or sound generation among performers are also indispensable for music and dance performances [Mat96].

Action timing generation in response to perceived events. Timing generation mechanisms of actions exploits the metric properties as well. One can control the beginning timing of utterances based on the other's speech signals (Figure 1.7). This pause and overlap lengths often convey rich information of one's intention or affective states [OKYI96, NDKN02]. Timing generation is also essential to perform articulated motions; humans and animals optimize control timing (e.g., insertion of torque power) of each different parts to realize effective body

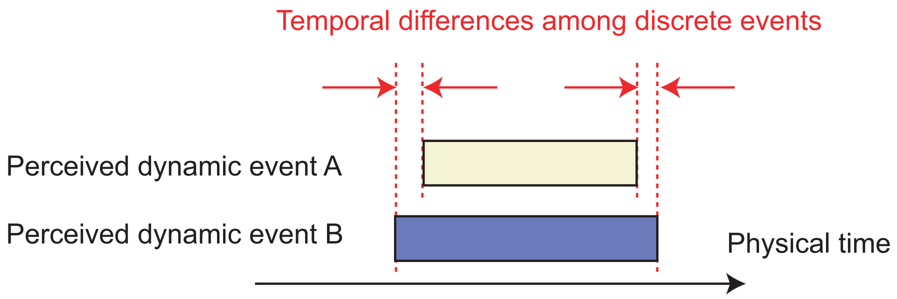
Temporal differences among discrete events
Perceived dynamic event A
Perceived dynamic event B
Physical time

Figure 1.6: Temporal relation between dynamic events appeared in multiple objects (parts) or different media signals (e.g., synchronization).

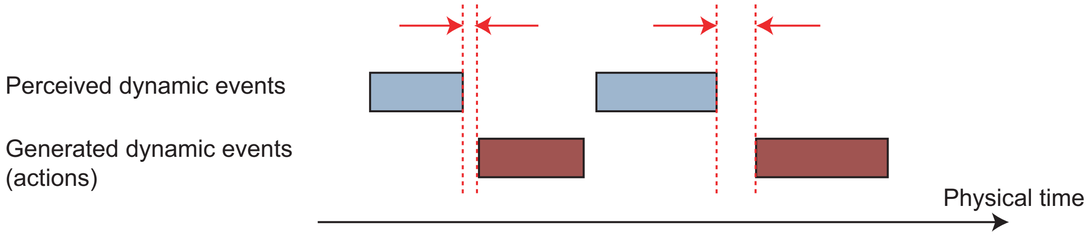

Figure 1.7: Generation of action timing in response to perceived events.

action [KOT$^+$04, YKM06]. The timing between body parts are also essential; for example, multi-part motion appeared in human body often described by timing of motions as we see in dance notation (e.g., Labanotation [Nak01]).

The objective of this thesis is to provide a computational model that represents dynamic structures described in the preceding paragraphs:

- Duration lengths and their patterns of dynamic events in a single signal
- Temporal metric relation of multiple dynamic events observed in multiple parts, objects, and different media signals, which can be overlapped each other
- Temporal metric relations of perceived and generating events

In this thesis, we use the term *timing structures* for these metric relations described by the distances among discrete events. In the next subsection, we exploit relations of temporal intervals and introduce a concrete system for modeling the timing structures.

1.4.3 Interval-Based Hybrid Dynamical System

An interval-based hybrid dynamical system (interval system) is the integration of a discrete-event system and multiple dynamical systems similar to existing hybrid dynamical systems described in Section 1.3. However, the rationale of the integration in an interval system is different from the existing studies. That is, we use a set of dynamical systems as cognitive processes, which we described in Subsection 1.4.1, for determining "temporal intervals".

Due to the intervals, the discrete-event system is able to describe a complex structure of dynamic events based on the relations of discrete events that are obtainable from the intervals as their beginning and ending time points. As a result, the intervals work as interfaces between the discrete-event systems, which represent the structure of discrete events in the subjective-time domain, and the dynamical systems, which represent continually-changing dynamic events in the physical-time domain.

Let us consider a human kicking motion as a dynamic event for example. If we observe the motion, we recognize that the leg moves based on several types of dynamics, such as two types of dynamics in bending backward and kicking forward if we assume bi-phasic motion, and we see that each of dynamics appears as temporal intervals in the physical time. Because the motions of other parts (e.g., arms) are essential for the kicking motion to take balance of the overall body, we also observe that several arm motions are closely related to the leg motions. Thus, primitive motions appear in multiple parts during a single motion execution can be represented by a structured multiple dynamics. This process can be generalized as the following conception; the observable signals (Figure 1.8 bottom) are produced by the *orchestration of dynamics* (Figure 1.8 top) that determines the activation timing of dynamics in the internal state spaces (Figure 1.8 middle).

What we want to do here is describing these structures that have significant temporal relations among multiple intervals each of which represented by a dynamical system. In addition, we demand the model to be learned from training data observed as input multivariate signals or extracted feature sequences from the signals. These objectives require two essential issues to be considered:

1. How to determine the concrete model of the dynamical system that represents each of dynamic events (primitives); the type of dynamic events should be considered because the model of dynamics affects the discrete events and their structures represented by the overall system.

2. How to model the temporal relations among discrete events (i.e., beginning and ending points of intervals) with their metric properties; the complexity of the model is too high to be trained if we take all the relations among discrete events into account, some simplification is therefore required.

1. Types of the Dynamics for Modeling Dynamic Events

In this thesis, we focus on modeling human behaviors observed in communication. Therefore, it is plausible to use linear dynamical systems as a type of dynamics for modeling dynamic events, such as visual motion, because most of dynamics produced by humans is the effect of muscular action.

Another option for modeling dynamics is the use of nonlinear dynamical systems; for example, recurrent neural networks [MM98a], polynomial systems [OTN02], and other systems that use nonlinear mapping functions in their state transition or observation (e.g., extended Kalman filtering [SHST00]). Nonlinear dynamical systems might be important for modeling such as consonant sounds in speech; we however assume that most types of dynamics are represented by linear dynamical systems because of the following reasons.

- Nonlinearity in the signal can be reduced to some degree (1) if we assume enough order for the Markov process, and (2) if we select appropriate static or dynamic features in the feature extraction phase.

- Nonlinear signals or nonlinear feature sequences can be represented by piecewise-linear systems if we choose appropriate units of primitives.

In particular, we exploit the second points; that is, we assume a complex dynamic event comprising a set of sub dynamic events. Those sub dynamic events are often referred to as motion primitives [NNYI04], movemes [Bre97], visemes [NLP$^+$02], motion textons [LWS02], modes [NKHM05], and so on.

Then, we assume the observed signals or feature sequences that are describing each of the temporal regions of sub dynamic events is represented by a linear dynamical system. For example, a cyclic lip motion can be described by a set of simple lip motions such as “open”, “close”, and “remain closed” (Figure 1.9). Once the set of sub dynamic events is determined, a complex action can be partitioned by temporal intervals that have labels of the linear dynamical systems and their duration lengths.

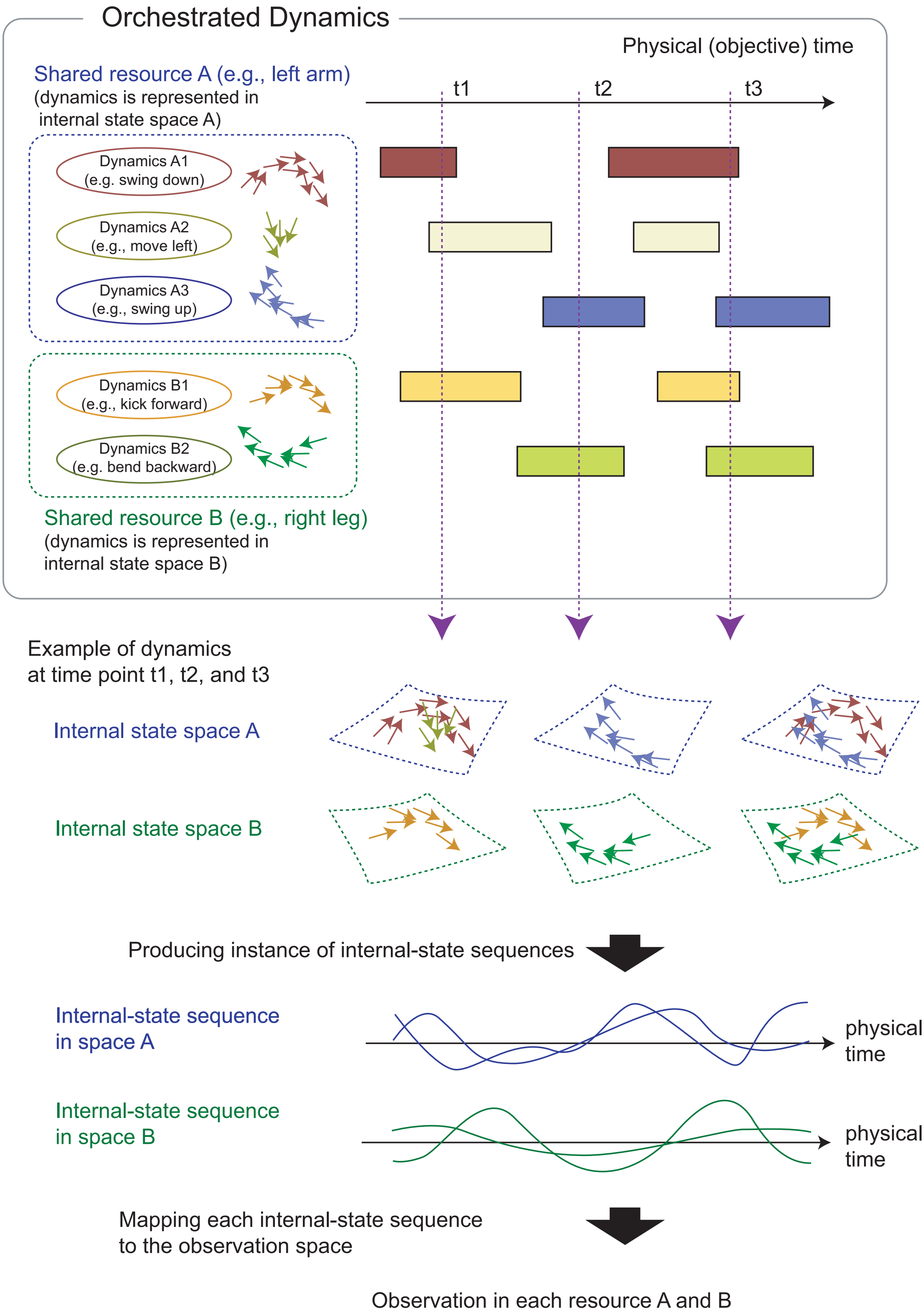

Figure 1.8: Orchestration of dynamics.

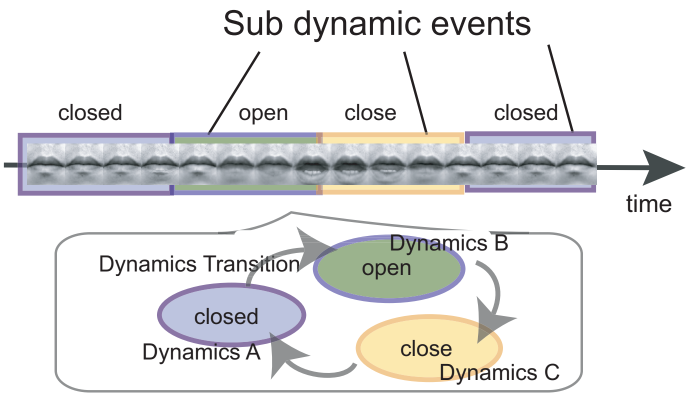

Figure 1.9: An example of dynamic events that consist of sub dynamic events.

2. Simplification for Modeling Dynamic Structures

As we see in Figure 1.8, the orchestrated dynamics have complex structures among multiple intervals. Whereas our final goal is to describe this kind of general structures, the complexity of the model describing the structure is not negligible because we require the model to be learned from training data; the stability and computational cost of the learning depend on the complexity of the model.

To reduce the complexity as simple as the model is trainable from real data, we set some assumptions for constraints of the model to describe a subset of general structures:

- A set of features or parameters that represent configurations of a single resource (e.g., a body part, an object, and a type of media) form a single multivariate signal.

- A signals of a single resource can be partitioned into intervals by multiple linear dynamical systems that share a single internal state space; each of the intervals is represented by a linear dynamical system.

- The intervals represented by linear dynamical systems of a single resource have no gaps or overlaps each other in the physical time; thus, the dynamics in a single resource switches from one to another, and the beginning points of one interval corresponds to the ending points of the next interval.

- The metric relation between intervals in different resources can be described by the temporal differences of beginning and ending points of the intervals (another assumption is introduced in Chapter 5 to specify the interval pairs).

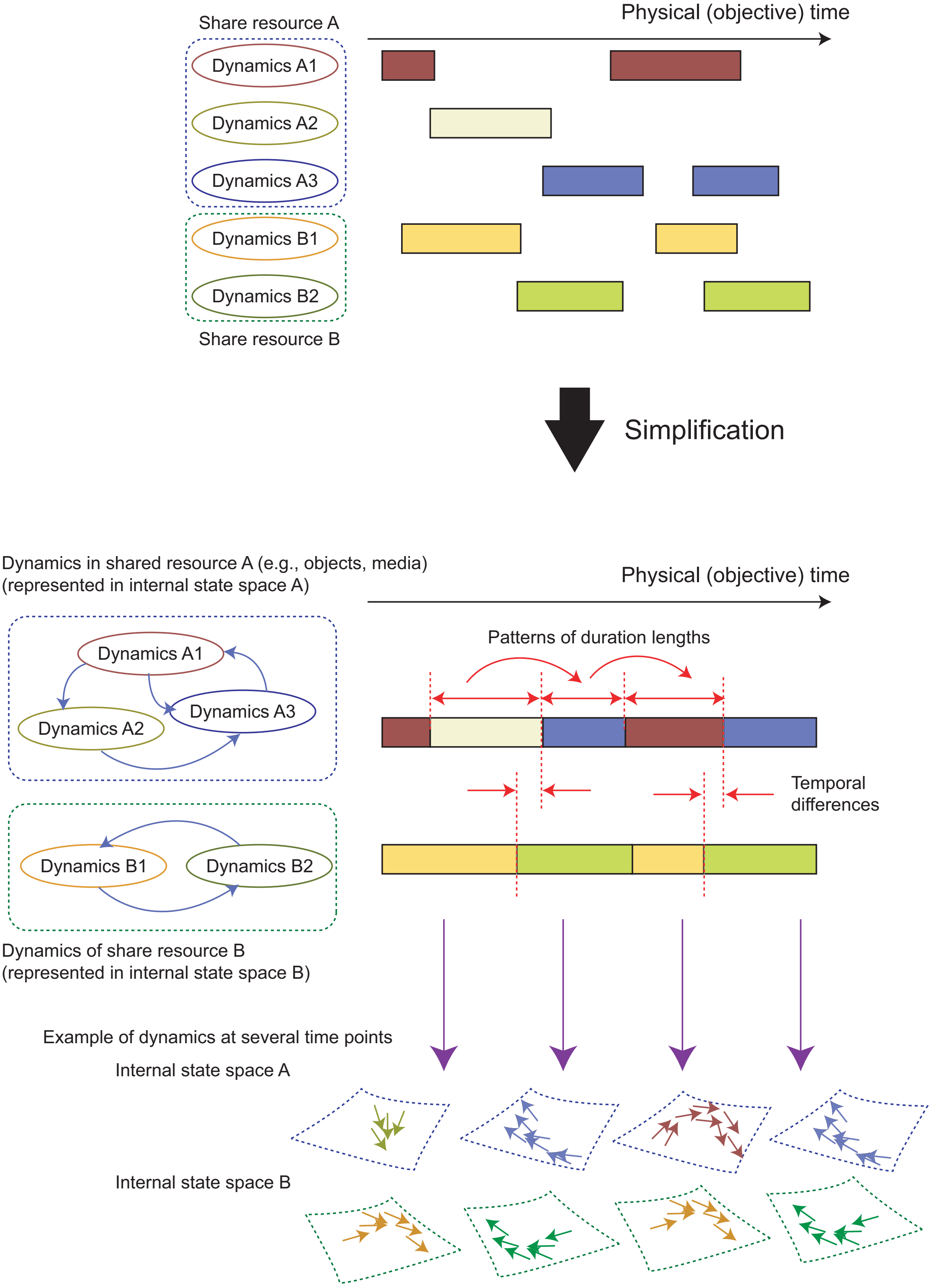
Share resource A
Physical (objective) time
Dynamics A1
Dynamics A2
Dynamics A3
Dynamics B1
Dynamics B2
Share resource B
Simplification
Dynamics in shared resource A (e.g., objects, media)
(represented in internal state space A)
Physical (objective) time
Patterns of duration lengths
Dynamics A1
Dynamics A3
Dynamics A2
Temporal
differences
Dynamics B1
Dynamics B2
Dynamics of share resource B
(represented in internal state space B)
Example of dynamics at several time points
Internal state space A
Internal state space B

Figure 1.10: Integrataion of two interval-based hybrid dynamical systems.

Based on the third assumption, the discrete events of one resource are linearly (totally) ordered. Note that each discrete event has a type of dynamics that is finished by the event. Therefore, we use simple automaton to model the order of the types of dynamics finished by each of discrete events in a single resource. Especially, we consider one-to-one correspondence between discrete states of the automaton and the type of linear dynamics (see Chapter 2 for details). In this thesis, we use the term "an interval-based hybrid dynamical system" (or an interval system) to refer the system that represents a single resource.

Consequently, a general structure of orchestrated dynamics, such as shown in the top of Figure 1.10, is simplified to the architecture comprises two interval systems as shown in the bottom of the figure. We see that only one type of dynamics can be activated at a single time point in one resource, and each automaton models the activation order of the multiple dynamics of the resource. The transition of dynamics provides complex dynamics as a whole, and determines the behavior of a produced instance of an internal-state sequence.

We view the interval system proposed in this thesis as an initial step toward understanding human-human interaction and realizing human-machine interaction systems. Therefore, as evaluating the interval system, we concentrate on verifying how the interval systems are suitable for modeling dynamic events produced by humans, such as facial motion, body motion, utterances, and behaviors, based on the assumptions above.

1.4.4 Expressive Power and Limitations

Expressive Power

Modeling duration length of a dynamic events. It is well known fact that the HMMs represent only the exponential distributions of state duration if we use the model with observations that occur in fix-length intervals [ODK96]. This situation is quite common in speech and gesture recognition systems that use sampled signals as input data. Let us consider a HMM that has more than two discrete states, and assume that the observations occur based on a fixed-length sampling rate. Let a_{ii} be a transition probability that the HMM preserves state q_i at an occurrence of an observation. Then the probability that the HMM sustains state q_i during time length t — and changes the state after the duration — becomes $a_{ii}^t(1 - a_{ii})$. Thus, the HMMs are restricted to represent the exponential distributions of discrete-state duration. On the other hand, the interval system is able to utilize general functions for the interval length distribution similar to segment models. Moreover, the interval system explicitly models the relation among interval lengths of multiple dynamic events, and provides the expressive power that we describe in the next paragraph.

Modeling patterns of duration lengths of dynamic events in a single signal. As for the expressive power of the interval-based hybrid dynamical system for modeling patterns of duration lengths, we compare the system with several existing models in the language theory. Let us consider the situation of modeling a dynamic event E that comprises two types of dynamics a and b, each of which represents a sub dynamic event constitutes E. Let us assume that each of dynamics a and b appears only once in this order, and that each dynamics continues the same length. To describe this situation, we can use a sequence $\{a^t b^t \mid t \in \mathbf{N}\}$ to denote the change of dynamics in the physical time, where we assume t represents the length of the dynamics described in a discrete time, which sampled by a fixed-length rate. Whereas an automaton can not represent these patterns because it is in the class of context-free grammar (CFG), the interval system can describe these relations; for example, we can explicitly set the adjacent interval to be the same duration lengths.

This result can be extended to the relation among three types of dynamics. Let us consider the situation that a dynamic event comprises three types of dynamics a, b, and c, and assume they appear in this order with the same length. We can use a sequence $\{a^t b^t c^t \mid t \in \mathbf{N}\}$ similar to the previous example. Although the sequence is represented by a context-sensitive grammar rather than a CFG in this case, we are able to describe this situation if we explicitly model the relation of the duration between dynamics pairs (a, b) and (b, c). Consequently, the interval system is more expressive than CFG in some aspect of modeling the patterns of duration lengths.

Modeling temporal differences among discrete events in multiple signals. Early integration [CR98] is one of common methods to model the relation between multimodal signals. This method combines two feature vectors observed from different modalities at a single frame, and forms a single vector. In other words, the early integration utilizes a frame-based integration. The method however have disadvantage of modeling metric structures of discrete events, such as lengths of temporal gaps (pauses) and overlaps in two intervals. To describe these structures, the method can not avoid increasing the Markov order of the model. For instance, if the maximum length of the temporal gaps between discrete events is $l_{\max}$, the Markov order in the model is required to be greater or equal to $l_{\max}$ because the relation between one discrete event at time t and the other event at $t + l_{\max}$ have to be preserved in the model. The size of $l_{\max}$ is however not small in general case. As a result, the computational cost and memory size of the frame-based models easily increases. In the proposed framework in this thesis, on the other hand, models these temporal gaps explicitly based on temporal differences among discrete events. Thus, the number of the model parameters becomes small enough to train and apply in real problems.

Limitations

Chaotic dynamics. Nonlinear systems often have important temporal or geometric features, and able to describe complex behavior without modeling stochastic processes [AIYK00]. For example, some dynamics have positive Lyapunov exponents, which determine how fast the system becomes unpredictable in time. These dynamical systems, which are referred to as *chaotic systems*, can represent a wide variety of signals in spite of using only small degree of freedom.

On the other hand, some dynamical systems have non-integer fractal dimension, and they generate strange attractors that have recursive structures in their internal state space. The systems therefore generate complex signals that have layered dynamics even if we use only a single dynamical system.

Despite of the capabilities of those nonlinear dynamics described above, in this thesis, we use only linear dynamical systems because nonlinear systems are sometimes hard to identify from real data, and difficult to predict their macro behaviors. The limitation that we use only linear dynamics becomes significant if the behavior of signals is inherently chaotic or have recursive strange attractors. We however anticipate that most of signals observed in human behaviors, such as motion and utterance, can be represented by a combination of linear dynamics, as we described in the previous subsection.

Layered structures among discrete events. There exist layered structures of dynamic events in space and time; for example, a running motion can be decomposed into several body motions such as arm motions, and the arm motion sometimes comprises different types of dynamics. However, as shown in the previous subsection, we focus on modeling two-layer structure; that is, relation between a dynamic events and its sub dynamic events. This limitation becomes significant if the dynamic events have grammatical structures (e.g., sign languages). In Chapter 6, we provide detailed discussion of extending the framework proposed in this thesis to deal with discrete events that have complex layered structures.

1.5 Overview of the Thesis

In this section, we present the organization of the subsequent chapters in this thesis. Figure 1.11 depicts the overview of this thesis.

Modeling Structures of a Single Signal (Chapter 2)

In this thesis, we first concentrate on modeling a single signal from a single source for the simplest case, and describe the relation of adjacent intervals based on the correlation of their duration lengths. Duration-length relation of the adjacent intervals corresponds directly to our cognitive sense of time such as tempo and rhythms, which are crucial information to represent features of the dynamic events.

Another advantage of explicitly modeling interval relations is that it enhances robustness against outliers during temporal partitioning process; in other words, the top-down knowledge works as a constraint to the lower-level process. For instance, if the duration distributions of the subsystems are biased toward a long length, the system will not change the subsystem before the activation of the subsystem sustains enough length. As a result, the system improves the robustness of representing temporal structures that can be partitioned into temporal intervals.

Chapter 2 describes a detailed model structure and an inference algorithm that searches the optimal interval sequence that provides the highest probability for the given observation. Then, we verify the inference algorithm using simulated data.

Learning Method of an Interval-Based Hybrid Dynamical System (Chapter 3)

In spite of the flexibility of hybrid dynamical systems, especially for modeling human behaviors such as gestures and facial expressions, few applications have exploited the system to handle real-world signals. The reason is largely due to the paradoxical nature of the learning process: temporal segmentation and system identification problems need to be solved simultaneously.

Chapter 3 proposes a two-step learning method to identify the interval system. In particular, we propose a novel clustering algorithm as the first step of the learning method; the algorithm extracts a set of dynamical systems from observed sequences, and is applicable to general hybrid dynamical systems. We evaluate the effectiveness of the proposed learning method using simulated and real data.

Analysis of Timing Structures in Multiple Signals (Chapter 4)

As we described in Subsection 1.4.2, temporal metric relations among multiple objects or multimodal signals often have significant structures to identify dynamic events. Applying the interval-based hybrid dynamical systems to describe structured dynamic events, we can analyze dynamic features based on the timing structures extracted from temporal intervals.

Chapter 4 shows how the interval system can be applied to describe and analyze temporal relation between multiple objects. We apply the system to represent complex motion appeared in each facial part independently, and examine the effectiveness of using the timing structures to analyze and discriminate fine-grained facial expression categories such as intentional and spontaneous smiles of which existing methods had difficulty to represent the difference.

Modeling Timing Structures in Multiple Signals for Timing Generation (Chapter 5)

Timing structures are also essential to provide appropriate behaviors at appropriate timing in response to the perception of dynamic events occurred in the situations. To realize the function of timing generation, we model temporal structures among different kind of media signals from multiple sensors by extending the analysis in Chapter 4.

Chapter 5 shows a general framework for modeling and utilizing mutual dependency among media signals based on the temporal relations among intervals. In this chapter, we provide a novel algorithm that generates timing of dynamic events in one media signals (e.g., lip motions in a visual signal) from another related input signal (e.g., an audio signal).

Finally, Chapter 6 summarizes the investigation in this thesis, and concludes with a discussion of open issues.

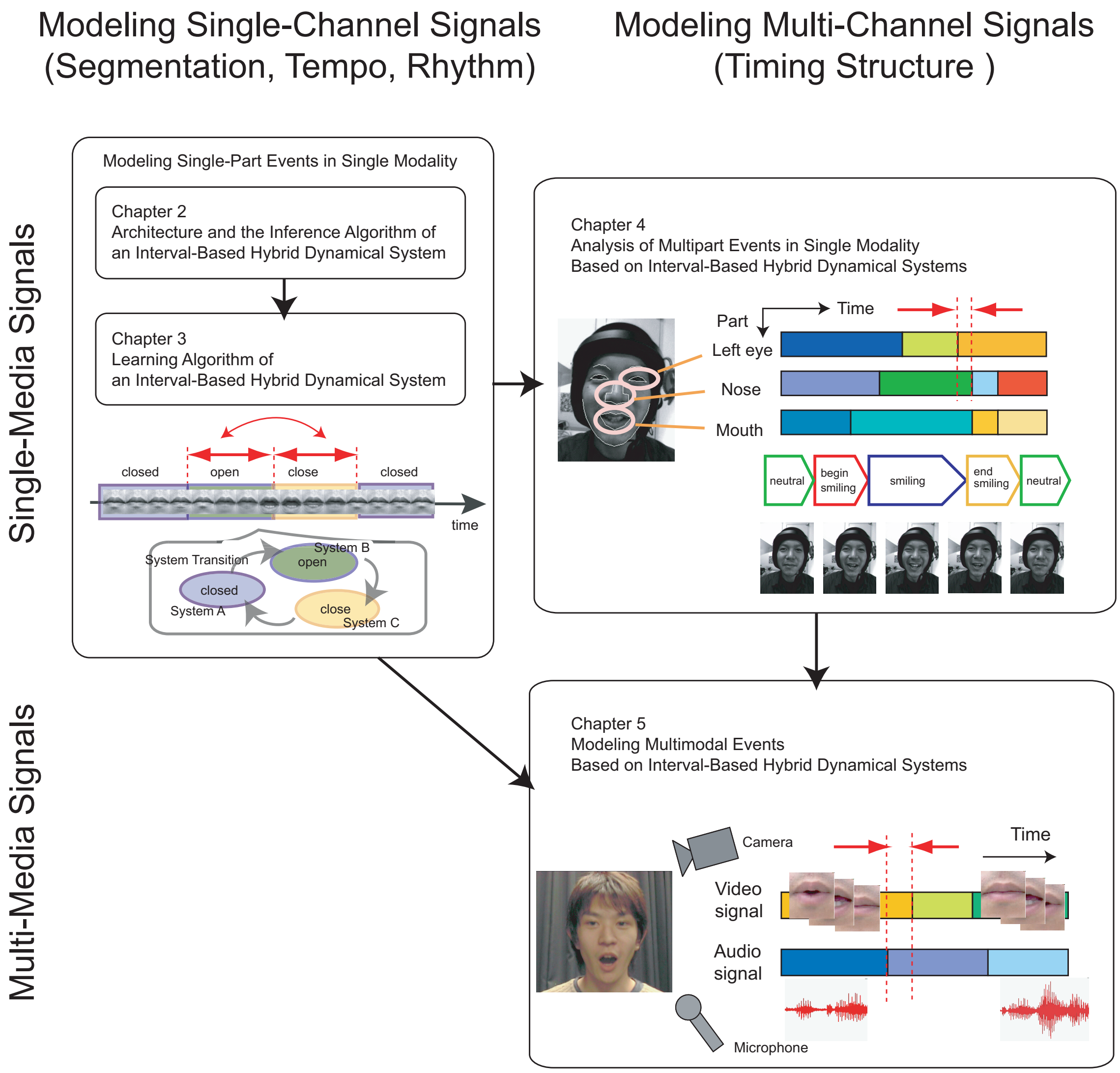
Modeling Single-Channel Signals
(Segmentation, Tempo, Rhythm)
Modeling Multi-Channel Signals
(Timing Structure)
Single-Media Signals
Multi-Media Signals
Modeling Single-Part Events in Single Modality
Chapter 2
Architecture and the Inference Algorithm of
an Interval-Based Hybrid Dynamical System
Chapter 3
Learning Algorithm of
an Interval-Based Hybrid Dynamical System
closed
open
close
closed
time
System Transition
System B
open
closed
System A
close
System C
Chapter 4
Analysis of Multipart Events in Single Modality
Based on Interval-Based Hybrid Dynamical Systems
Part
Time
Left eye
Nose
Mouth
neutral
begin smiling
smiling
end smiling
neutral
Chapter 5
Modeling Multimodal Events
Based on Interval-Based Hybrid Dynamical Systems
Camera
Time
Video signal
Audio signal
Microphone

Figure 1.11: Overview of the thesis.