# Chapter 6

# Conclusion

## 6.1 Summary

In this thesis, we proposed a novel computational model, named an interval-based hybrid dynamical system, to model dynamic events and structures. As we described in Chapter 2, we exploited temporal intervals as an interface between 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).

To overcome the paradoxical nature of the learning process, which requires to solve temporal segmentation and system identification problems simultaneously, we proposed a two-step learning method in Chapter 3. Due to the proposed method, we can extract linear dynamical systems that model primitive dynamics of the event from the given temporal signals.

In Chapter 4, we applied the proposed model to describe structured dynamic events that consists of multipart primitives. We showed that the systems can analyze dynamic features based on the timing structures extracted from temporal intervals. We examined the effectiveness of the timing structure analysis to discriminate fine-grained facial expression categories such as intentional and spontaneous smiles of which existing methods had difficulty to represent the difference.

In Chapter 5 we proposed a "timing structure model" that directly represents timing structures in multimedia signals, such as synchronization and mutual dependency with organized temporal differences among temporal patterns of media signals. Experiments on simultaneously captured audio and video data showed that time-varying signals of one media signal can be generated from an-

other related media signal using the trained timing structure model.

In the next section, we show some open issues that we were not able to cope with in this thesis, which can be divided into two aspects: (1) the extension of the proposed computational model, and (2) the situations that the model can be applicable including multiparty interaction.

## 6.2 Future Work

### 6.2.1 Extension of the Interval-Based Hybrid Dynamical System

In this subsection, we show some directions that the proposed interval-based hybrid dynamical system should be extended for future work.

#### (a) Non-linear Dynamical Systems

The selection of appropriate dynamical models depends on the nature of signals and the design policies of users. We chose to use linear dynamical systems because most of the continuously changing human motions can be approximated by linear dynamics. This is because the motions are generated by the expansion and contraction of muscles, and are controlled to be stable (e.g. no oscillation). However, nonlinear dynamical systems sometimes can be more reasonable choice for modeling time-varying patterns such as consonants in human speech. One of the straight forward methods to extend our model is the use of "kernel methods", which are major approach for the nonlinear data analysis. The kernel methods convert nonlinear algorithms in the original data space into linear algorithms in higher (or infinite) dimensional feature space. For example, kernel principal component analyses utilize inner products in the higher dimensional space [HTF01]. We need further discussion to give a guideline to select models.

Some motion generation researches in robotics design the overall system as a nonlinear dynamical system rather than a hybrid dynamical system. For example, Okada represented the motion of robots as a cyclic attractor in the configuration space [Oka95], and modeled the switching process between cyclic attractors in the configuration space based on continuous dynamics [ONN03]. Recurrent neural networks (RNN) also utilize its nonlinear dynamics to represent complex motion. Ogata et al. use the RNN with parametric bias (RNNPB) to extracting [OOK+05], which change its internal dynamics based on the additional input to the network (parametric bias). Morita et al. proposed RNN with non-monotonous function

for modeling temporal pattern recognition [MMM02] and extended the model to represent symbolic contexts of patterns using selective desensitization of some of elements [MMM02, MMMS04].

**(b) Modeling Transition Process between Dynamical Systems in adjacent Intervals**

The state in the internal state space often changes discontinuously when the automaton changes the dynamical systems. To model co-articulated dynamics such as phonemes in speech data and to generate smooth motion, we need transitional process modeling between two dynamical systems (e.g., the modulation of dynamics by the preceding dynamics). A straightforward method is to model the interpolation of two dynamics in adjacent intervals. Although the interpolation provides low-cost method to smoothing two dynamics, it sometimes generates unnatural motion at the joint of the two intervals. Li et al. proposed to set the end constraints of a synthesized segments [LWS02]. They deduced a block-banded system of linear equation from the constrains, and realized smooth motion texton synthesis by solving the equation.

**(c) Modeling Temporal Structures among More Than Three Signals**

While we concentrated on a timing structure model in two media signals in Chapter 5, we can apply the model to represent the structures among more than three media signals by defining pairs of signals and constructing timing structure models for each of the pairs similar to coupled HMMs [BO97]. On the other hand, we will be required to introduce other timing structure models if we consider a problem specific causality between signals. For example, we can introduce a unobservable interval sequence that controls a generation timing of observable media patterns. This model might be applied to a large area of human (animal) behavior because many of muscular motions are controlled by unobservable spike signals from a brain with physical delay [Pop85].

**(d) Modeling Complex Structures of Dynamic Events**

We exploited a simple finite state automaton as a discrete-event model in order to concentrate on modeling human body actions and motions, which have relatively simple grammatical structures (represented by a regular grammar) compared to natural languages. To model languages (e.g., human speech and signs) and other

complex situations (e.g., human communication and strategies), more complex grammatical structures such as N-gram models and context free grammars will be required together with its parameter estimation method.

There are several aspects of structures to extend our model. In the following, we show some of the existing approaches in machine learning and computer vision to represent complex structures.

**Context-free grammars.** A stochastic context-free grammar (SCFG), which defines probabilities of each of productions in a context-free grammar, is used for modeling grammatical structures among primitive events. Ivanov and Bobick use the SCFG model to recognize dynamic situations of parking area and human gestures [IB00]. Moore and Essa extended the model to detect errors and to recover the detected errors, and applied for modeling behavior of players involved in card game situation [ME02].

**Layered structures.** A hierarchical HMM (HHMM) was proposed by Fine, Singer and Tishby in machine learning community, and some computer vision applications were realized based on the model [NBVW03, BPV04]. The model is the extension of HMMs that each of states is capable to have not only output probability but a child HMM. As a result, the model can represent layered structure of events based on the recursive definition of HMMs. The HHMM is the simplified model of SCFG, therefore, the computation cost of probabilistic inference in HHMM is lower than that of SCFG.

**Higher-order Markov models.** Variable-length N-gram model was also proposed by Ron and Tishby [RST96]. They used a prediction suffix tree to represent and construct a variable-length N-gram model from an input symbol sequence. The model can be converted to a finite state automaton in which each state corresponds to a sequence of symbols. Galata et al. applied the model to represent long-term context of human behavior and provided some preliminary results [GJH01].

A key issue when we introduce layered structures of discrete states into the interval-based hybrid dynamical system is how to determine the layer in which temporal intervals and temporal relations among the intervals are defined. As we assumed in this thesis, a set of modes (dynamic primitives) corresponds directly to a set of discrete states, and the modes mapped one-to-one to the discrete-states. An intuitive extension is to consider a sequence of modes (linear dynamical systems) as a single discrete state based on the variable-length N-gram model. In a

sense, the sequence of modes constitute a "phrase" of dynamic primitives, and the discrete-state transition determines the sequence of the phrases. In this case, we can regard duration of phrase as an interval, and can introduce duration lengths of phrases. We can use a prediction suffix tree to train this model after the set of modes are determined by the clustering method proposed in Chapter 3.

### 6.2.2 Modeling Multiparty Interaction

In this thesis, we concentrated on applying the interval-based hybrid dynamical system to model a single human behavior rather than multiparty interaction, because our first concern is to see the effectiveness of the proposed model for modeling and learning dynamic events and structures from multimodal signals (see Figure 6.1 left). Extending the proposed scheme to model multiparty interaction and to realize human-machine interaction systems, we have to aim at finding key features of interaction dynamics or protocols in human-human communication, and exploiting the found features for natural and smooth human-machine communication (Figure 6.1 right).

We discuss how the proposed framework of the system can be applicable for modeling multiparty interaction and what are insufficient for our current system in the following paragraphs.

#### Timing Structures in Speech Conversation

In human conversation, there exist many lexical, prosodic and syntactic elements that help create the dialog structure [Shi05]. Especially, utterance timing (such as transition interval in Figure 6.1 upper-right) and speaking tempo can help to add a smooth, tense, lively or relaxing tone to the dialog. It is with this tone that dialog can evoke feelings of pride, sorrow, fear, and enjoyment. Since we humans are exposed to a large amount of timing structures in speech dialog during their development, in other words, we are professional to recognize and generate timings; the users are sensitive to unnatural utterance timing and speaking tempo of speech dialog systems.

As for the analysis of the timing structures and their effects in human speech dialogs, Ichikawa and Sato showed that many of backchannel utterances occur within about 0.4 second after the keyword appears in the speech of the others [IS94]. Nagaoka et al. analyzed dialog of operators and customers in a telephone shopping situation, and showed that utterance timings are one of the es-

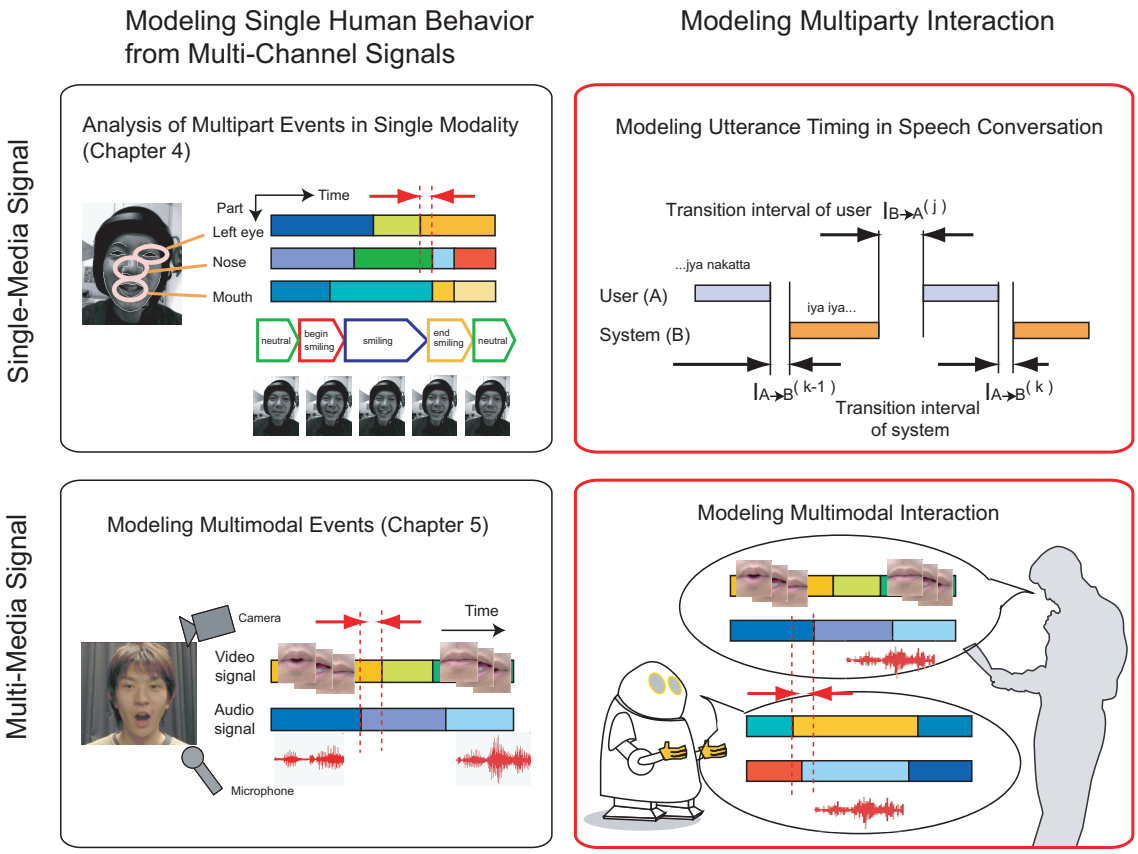


Figure 6.1: Extension to Multiparty Interaction. We investigated left half of the figure in this thesis. (See also Figure 1.11.)

sential cues for determining an impression about the speaker [NDKN02]. As for dialog systems that control utterance timing, Okato proposed the use of pitch patterns in utterances to estimate the timing of backchannel responses. Kitaoka et al. developed a response timing generator for speech dialog systems based on the use of pitch and power patterns, power, utterance lengths, and some lexical information [KTNN05]. Fujie developed a robot that makes backchannel feedbacks with overlaps based on the use of pitch and power patterns [FFK05].

Although the existing approaches generate timing of backchannel utterances, non-backchannel utterances also have significant features in its generation timings. We have to use more fine categories of utterances based on the functions or purpose of the speech in order to realize dialog systems that communicate purpose, attitude and feeling of the spoken word. The use of dialog acts (e.g. "question", "statement", "opinion", etc.), which correspond to illocutionary acts in speech acts [Sea86], will help to analyze and categorize backchannel and non-backchannel utterances based on the functions of them [JSFC98, SRC+00, DBCS04]. Once the categories of dialog acts are defined, we can learn the timing structure model for each of the dialog acts, and can apply the timing generation algorithm described in Chapter 5.

**Timing Structures in Multimedia Interaction**

Another disadvantage of existing timing generation applications is that the systems have no unified framework to integrate multimodal information captured as different media signals. For example, human utterance timings are defined by not only audio information but visual features such as facial expressions and lip motions of others. We can also see that the facial expression of one person affects the others expressions in our daily communication.

As we described in this thesis, the interval-based hybrid dynamical system has a capability of modeling the mutual dependency among multiple signals. We can therefore exploit this system to realize human-machine interaction systems by extending the system to model timing structures among more than three signals (see Subsection 6.2.1 (c)). The use of nonlinear dynamical systems and more complex structures should be considered to represent these general interaction patterns (see Subsection 6.2.1 (a) and (d)).

In addition, a timing structure model is useful to estimate internal state of humans considering the result of the facial expressions analysis in Chapter 4. Estimation of human internal states including intensions, interests, emotions, and

other unobservable mental events are essential for providing appropriate information to others. Because these mental events often affect our body via a neural system deliberately or involuntary, we can observe distinctive dynamic features of signals from multiple modalities that is sufficient for estimating internal states of others. Especially, we remark the following points for taking advantage of timing structure models for the internal state estimation:

- Internal states affects the timing structure (e.g., pause, tempo, and rhythms) of pitch and power in speech, gestures, eye gaze, gait motion, and other observable media signals from human activities.

- The timing of human reaction is affected by his or her internal states; for example, we can see some time lag of the response when the person concentrated on other things.

From the timing structures above, information systems will understand user's aims and situations, and will provide kind and timely guidance, which are the most important functions for realizing human-centered communication.

Consequently, the extension of the interval-based hybrid dynamical systems can be a fundamental basis of interaction systems that share a sense of time with humans based on the integration of physical and subjective time as we described in Chapter 1.

### 6.2.3 Hybrid Computing in Robotics

In this subsection, we show how the proposed concept of the interval-based hybrid dynamical system can be applicable to the area of robotics, especially the degree of freedom (DOF) of robots becomes very high (e.g., humanoids).

The existing methods to realize robots that control the body motion can be categorized into model-based approaches and behavior-based approaches. A model-based approach calculates and plans body motions and actions at the inside of computers based on the knowledge of the real world and robot bodies of robots (e.g., reasoning agents [RN02, PS99] and inverse-dynamics based controls [KYHH05]). A behavior-based approach exploits the interaction between robot bodies and the environment to emerge robot actions without modeling the real world or body (e.g., subsumption architectures [Bro86, Bro91] and a passive walk [CRTW05]). The integration methods of two approaches, which use different computing resources, are under the investigation [YK02]. This stream is

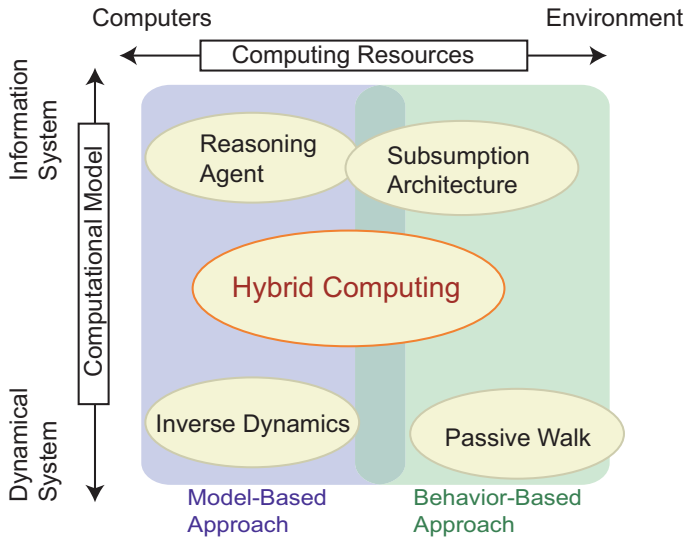


Figure 6.2: A Map of Robotics Approaches.

regarded as an integration of the horizontal direction in Figure 6.2.

On the other hand, these approaches can be divided based on different aspect depicted by the vertical direction in Figure 6.2: dynamical systems and information systems. For example, both the inverse-dynamics-based control and agent reasoning can be regarded as model-based approaches, which utilize the knowledge of the world including body of robots, the representation of the system is however completely different. That is, the inverse-dynamics-based control exploits dynamical systems and the agent reasoning uses information systems.

There exist some methods to integrate these two systems. The characteristics of these methods are that they first define some action primitives such as "move left hand up" and "move right leg forward", and then integrates these action primitives based on information systems such as finite state machines; meanwhile, each of primitives are realized by calculation of torque based on dynamical systems. The methods however have disadvantages that the primitives become too coarse and abstract to generate smooth motion. This is because the primitives are defined manually, and the temporal scale of each primitive cannot become smaller than human recognizable scales (Problem A). It is also difficult to define enough number of primitives when the DOF of robots becomes very high such as in humanoids (Problem B). Moreover, appropriate energy input timings are important to realize dynamic motion similar to human [KOT+04] (Problem C).

As we described in Chapter 3, the interval-based hybrid dynamics system pro-

posed in this thesis has a capability of learning a set of dynamic primitives from the given time-varying signals, it has therefore a potential to solve Problem A and B. The size of learned primitives can be controlled to be smaller than our intentional segmentation units. On the other hand, the timing generation method described in Chapter 5 can be the basis of modeling timing structure between the primitives (e.g., motion timing among body parts) and of determining activation timing of each primitives in response to the input or recognized events, which leads to the solution of the Problem C.

Consequently, the extension of the interval-based hybrid dynamical system is expected to realize a series of smooth behavior of robots, including humanoids, in dynamic situations where a robot contacts with objects and humans.

### 6.2.4 Relation to Human Consciousness

In this subsection, we consider the relation of the concepts in the interval-based hybrid dynamical system to "human consciousness"[1]. We do not intend to discuss what consciousness is and where the consciousness exists in our human body; consciousness comprises many aspects, and the definition of consciousness often differs from person to person. Our motivation here is focusing on one important aspect of human consciousness, the function of "temporal coordination", to bring some crucial issues as extending interval-based hybrid dynamical systems.

Why we human enjoy rhythms? The reason that the ability to enjoy rhythmic patterns has acquired during the process of evolution may not be only for playing music. We human are required to control and coordinate the timing of a series of actions in response to perceived events. Therefore, the sensitivity to dynamic structures among various events, which we described in Subsection 1.4.2, are indispensable for humans to survive in the real world. Especially, the temporal coordination within consciousness among dynamic events must be advantageous to humans under selection pressure compared to coordination in subconscious.

[1]The discussion in this subsection may also be applicable to animal consciousness. It is interesting to consider the difference of consciousness between humans and animals [Ecc89]; however, this topic is beyond the scope here.

### Essential Features of Temporal Coordination within Consciousness

We first consider the essential features that constitute temporal coordination within consciousness. A necessary condition for the function is to handle events (e.g., perceived input and generating actions) apart from the physical-time domain. However, this feature can also be used in subconscious temporal coordination. For example, human sometimes use a clutch of a manual-transmission car without awareness of the operation, which requires dynamic structure among multiple events. We here concentrate on more crucial features of the temporal coordination:

- A single time axis is used in mind for coordinating among multiple events (e.g., action and utterance). While multiple processes can be unconsciously activated in parallel [Lib04] (e.g., control of multiple body parts), the unified time in mind work as coordinator to maintain the consistency among the processes.

- Crucial time points that exist in various abstraction levels are dynamically selected, logically combined, and coordinated. While multiple time points of discrete events are recognizable, some points are crucial to achieve an overall action (e.g., "knacks" of robot action [KOT+04]). We pay attention to those crucial time points as the occasion demands. Once the crucial time points are coordinated, the dynamic structure among discrete events in lower abstraction levels is also coordinated unconsciously.

Hence, we can handle dynamic structures that have non-fixed patterns by exploiting these features.

### Learning of Structure among Discrete Events in Multiple Abstraction Levels

As for learning of a novel action such as a gymnastic exercise, crucial points are also variable. We human first find the temporal ordering relations among sensory information (e.g., visual input) and muscular activation. We then search the timing among perceived and generating events to realize the best performance of the action.

Once the action is acquired, we can orchestrate the control of multiple body parts in response to perceived input without awareness if the structure among events is fixed or simple enough; meanwhile, the learning phase requires awareness of fine-grained events that determine the performance of the action. In other

words, dynamic structures of acquired actions are push down to subconscious domain for concentrating on obtaining and realizing more complex or compositional actions that have structures in higher abstraction levels.

**Direction to Extend the Interval-Based Hybrid Dynamical System**

Compared to the temporal coordination within human consciousness described above, the interval-based hybrid dynamical system proposed in this thesis is quite restricted. As we discussed in Subsection 6.2.1, the interval system finds temporal points of discrete events based on linear dynamics and it controls only a single level of dynamic structure among those time points. As a more important issue, the timing generation method proposed in Chapter 5 is only able to control the temporal position of discrete events that have simple static distributions. In a sense, the system handles the subconscious coordination of events.

Considering temporal coordination in human, we anticipate the following features are essential to design information systems that fulfill the enough functions of human-machine interaction (Subsection 6.2.2) and robotics (Subsection 6.2.3):

- The mechanism that dynamically finds the crucial points in multiple abstraction levels based on the context and situation
- Temporal coordination function that maintains consistency of lower abstraction levels
- Learning method that reuses the dynamic structures obtained in the past learning to construct more complex structures

The function of temporal coordination in human consciousness also affects the number of events of which human is aware, and may influence the length of cognitive time in the experience, which attract many scientists' interest [Tsu87]. We believe the design of the computational model that has satisfactory functions to continue and survive in the real situations is necessary not only for engineering purpose but also for understanding the mechanisms of mind process, such as cognitive sense of time, in humans and animals. We hope the concept of the interval-based hybrid dynamical system serve as the first step of these objectives.