Interval-based Modeling of Human Communication Dynamics via Hybrid Dynamical Systems

Hiroaki Kawashima *
Kyoto University
kawashima@i.kyoto-u.ac.jp

Takashi Matsuyama
Kyoto University
tm@i.kyoto-u.ac.jp

Abstract

Temporal structures such as utterance pauses/overlaps, and duration of facial expression play a crucial roles in realizing smooth and natural human communication. We introduce a novel method for modeling the human communication dynamics based on interval-based representations via hybrid dynamical systems and show several examples to demonstrate how this method captures the characteristics of temporal structures in facial expressions and multimodal signals.

1 Introduction

The primary objectives of human-machine interaction systems are understanding the meaning of user commands and presenting the appropriate information. Therefore, most of the existing researches have aimed to realize interaction systems that can understand the semantic information specified by a user and generate attractive presentations through multimedia data (e.g., texts, pictures, videos, and sounds). Furthermore, currently advanced systems that can understand spoken words and gestures are being developed. While the multimedia interactions are important, users sometimes feel frustrated when the systems do not adequately consider human interaction protocols. For example, systems often ignore dynamic features, such as acceleration patterns, pause lengths, tempo speeds, and rhythms, which convey rich nonverbal information in human communication.

We have been attempting to model such dynamic features or temporal structures in human communication based on a hybrid dynamical system (HDS, or, in short, hybrid system) [1]. Essentially, an HDS is basically an integrated model of dynamical systems and discrete-event systems. In HDSs, dynamical systems modeled by using differential equations are suitable for describing smooth and continuous physical phenomena (considering time as a physical metric entity), while discrete-event systems are suitable for describing not only discontinuous changes in physical phenomena but also subjective or intellectual activities (considering time as ordinal state transition).

The rationale we use HDSs in the context of modeling human communication dynamics is the followings. Firstly, we assume that a complex human behavior consists of dynamic primitives, which are often referred to as motion elements, movemes, visemes, etc. For example, a periodic lip motion can be described by a periodic sequence of symbols that represent simple lip movements such as “open,” “close,” and “remain closed.” Once the set of dynamic primitives is determined, a complex behavior can be partitioned into “temporal intervals,” each of which is characterized by a label of the dynamic primitive and its temporal duration length.

Secondly, we assume that not only temporal orders of the primitives but also their duration lengths or temporal relations (gaps/overlaps) of the beginning and end time points of those temporal intervals convey rich information in human communication. For example, some psychological experiments have suggested that the duration lengths of facial actions influence the human judgments of basic facial-expression categories [2].

*Currently, a JSPS Postdoctoral Fellow for Research Abroad.

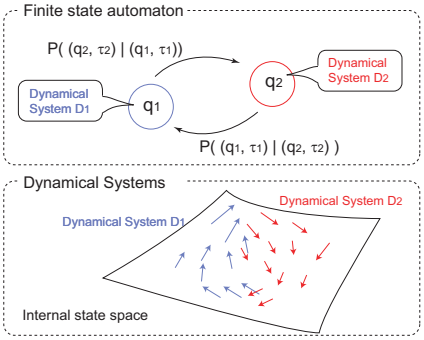

Figure 1: Interval-based HDS

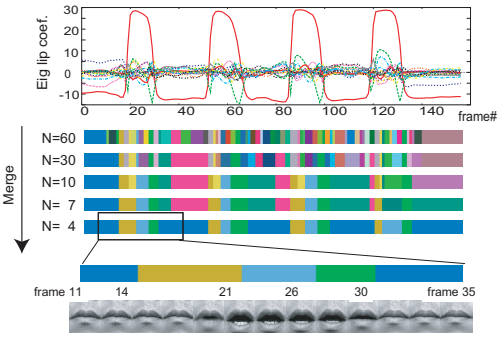

Figure 2: Agglomerative clustering of LDSs

Based on the abovementioned assumptions, we propose a specific class of HDS [1], which can be regarded as an extension of segmental models [3]. The system has a two-layer architecture consisting of a finite state automaton (Fig. 1 top) and a set of linear dynamical systems (LDSs) $\{D_i\}_{i=1}^{N}$ (Fig. 1 bottom). In this architecture, each LDS represents the dynamics of each primitive (e.g., motion element) and has a one-to-one correspondence with a discrete state of the automaton. In order to connect the physical temporal axis used in the LDSs with the automaton, we introduce *intervals* described by (m, τ), where $m \in \{q_1, ..., q_N\}$ and τ denote an automaton state and its temporal duration length, respectively. We therefore refer to our model as an interval-based HDS (IHDS).

Once the IHDS is identified (learned) from given training data in the manner described in the next section, the trained model can generate a multivariate sequence; that is, the automaton first generates a sequence of intervals, while the LDS corresponding to each interval is subsequently activated and generates a multivariate sequence. The activation timing and period of the constituent LDSs are thus controlled by the automaton. Similar to other generative models such as HMMs, the trained system can also be used to segment newly observed input data via the Viterbi algorithm.

By applying IHDSs to various human communication behaviors, we successfully extract dynamic features of the behaviors on the basis of the relations of temporal intervals. In this paper, we give two examples: the classification of fine-grained facial expressions (Sec. 3) [4] and the modeling of the synchronization/delay mechanisms between mouth movements and speech sounds (Sec. 4) [5].

2 Identification of Interval-based HDS

In the learning stage, we assume that only a set of multivariate sequences is given. Under this assumption, the identification process of the IHDS has a paradoxical nature. That is, since the system comprises a set of subsystems ($\mathcal{D}$), the parameter estimation of each subsystem requires partitioned training data to be modeled by the subsystem; meanwhile, the segmentation process of the given training data requires a set of identified subsystems as well as the number of subsystems. Moreover, iterative methods such as the expectation-maximization (EM) algorithm do not work properly in many cases because it depends strongly on the initial parameters. Therefore, we propose the following two-step learning method [1].

[Step 1] Clustering of Dynamical Systems: The first step is an agglomerative clustering that finds a set of LDSs required to describe the training data, i.e., the estimated number of LDSs (N) and their parameters. This is a model-based hierarchical clustering, where the models are LDSs and the distances between LDSs are measures such as the Kullback-Leibler divergence. Figure 2 shows an example in which the algorithm was applied to the periodic lip motion during the utterance of /mamamama/. The sequence of four LDSs appears repeatedly (here the colors represent the labels of the LDSs), and the periodic property is successfully extracted when $N = 4$.

[Step 2] Refinement of Parameters: The second step is a refinement process of the system parameters by using the EM-based Viterbi-approximated algorithm. This iterative algorithm is initialized by the result of the clustering step, which provides a rough estimation of the parameters of $\mathcal{D}$. The state-transition probability and the duration-length distributions are estimated along with the refinement of the parameters of LDSs at each iteration step.

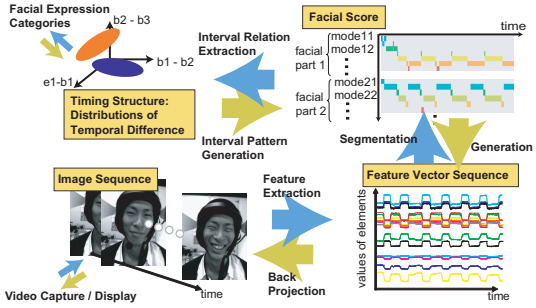

Figure 3: Overall flow of facial expression recognition/generation.

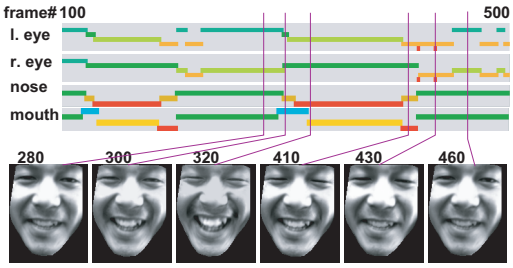

Figure 4: Generated image sequence from the obtained facial score.

3 Analysis of Timing Structures in Facial Expressions

Human facial expressions can be considered as being generated based on two mechanisms: (1) emotional expressions produced by spontaneous muscular action and (2) intentional displays to convey certain intentions to others. Therefore, an analysis of their dynamic structures is required to recognize human emotion and intention from facial expressions. However, most of the existing approaches depend on the facial action coding system (FACS) [6], whose objective is to describe basic emotional expressions (anger, happiness, sadness, disgust, etc.) using the combination of "action units" rather than determining the dynamic characteristics of facial expressions such as duration lengths and temporal differences among the movements.

To extract the dynamic characteristics of facial action, we propose a novel facial motion description that we refer to as a *facial score*; it can be acquired by applying multiple IHDSs to the movements of multiple parts in a human face (see also Fig. 3 bottom): (1) Track each facial part; in this study, we use six parts—left/right eyes, left/right eyebrows, mouth, and nose. (2) Extract the movements of each part as a time-varying sequence of shape vectors; we obtain six corresponding sequences. (3) Use an IHDS for each sequence, and segment the signal into an interval sequence. (4) Align the six interval sequences obtained in (3) along the common temporal axis. In this manner, we can construct a facial score (Fig. 3 top right) as a set of intervals; the facial score is similar to a music score in the sense that both representations describe the timing of elements (music notes).

Generation of facial expressions: By using this score, we can describe facial expressions as a spatio-temporal combination of the intervals. Figure 4 shows an example of the full set of the facial score that describes the dynamic characteristics of all facial parts during intentional smiles. Figure 4 (bottom) shows the facial motion generated by activating IHDS; in other words, each constituent dynamical system, which represents a simple motion, in the IHDS is activated in accordance with the timing described in the facial score.

Recognition of spontaneous and intentional smiles: Figure 4 also suggests that the movement of each smile can be segmented into intervals based on the following four modes—two stationary modes ("neutral" and "smiling") and two dynamic modes ("onset" and "offset" of smiling). By comparing the onset and offset timings of the intervals observed in different facial parts, we can extract various temporal features that can be used to classify facial expressions (Fig. 3 top left). In particular, we examined person-dependent recognition of intentional and spontaneous smiles using the timing structures among facial parts. Support vector machines were used on the features of temporal differences. The performance evaluation showed that the rate of correct discrimination ranged from 79.4% to 100% depending on the subjects.

4 Modeling Cross-Media Timing Structures in Multimedia Signals

Multimedia data comprising media signals is obtained by measuring human communication with multiple sensors. Once we successfully model the mutual dependency among those signals, we can employ the model in a variety of applications such as human computer interfaces (e.g., audio-visual speech recognition systems [7]) as well as computer graphics techniques that generate one media signal from another (e.g., lip sync to input speech) [8].

While the existing methods enable us to represent frame-based cooccurrence or short-term cross-media relations, they are not well suited to describe systematic and long-term relationships. For example, an explosive sound /p/ is strongly synchronized with an opening lip movement, while a vowel sound loosely synchronized with the lip and its temporal gap has certain variance.

To represent such systematic synchronizations/delays of mutual dependencies among multimedia signals, we extend the multipart-relation modeling described in the previous section. Let us assume that two media signals S and S' are captured from different sensors and that each signal is modeled by a different IHDS. The temporal relationship between overlapped interval pairs can then be modeled by the following probabilistic distribution:

$$P(b(I) - b(I'), e(I) - e(I') \,|\, m(I), m(I'), I \cap I' \neq \emptyset),$$

where $b(I)$, $e(I)$ are the beginning and end points of interval I, respectively; I, I', intervals appearing in the segmentation results of signal S, S', respectively; and $m(I)$, $m(I')$, the labels of LDSs in the IHDS models of S, S', respectively. We estimate the above distribution for all the possible LDS pairs in signals S, S' and use the cross-media relation model together with other distributions such as $P(m(I), m(I') \,|\, I \cap I' \neq \emptyset)$. This enables us to construct various of applications such as mechanisms to realize sound-to-motion (e.g., lip sync [5]) and motion-to-sound generations (speech estimation in noisy environments via visual cues [9]). The basic idea behind these applications is that, once IHDSs are learned from both media signals, the media conversion can be executed in the interval representation by solving the optimization $\mathcal{I}_{\rm gen} = \arg\max_{\mathcal{I}} P(\mathcal{I}|\mathcal{I}_{\rm ref}, \Phi)$ via the dynamic programming, where $\mathcal{I}_{\rm gen}$ and $\mathcal{I}_{\rm ref}$ are generated and reference interval sequences, respectively, and Φ is the parameter set of the learned timing model (i.e., probability distributions).

5 Conclusion

We have proposed an interval-based representation for modeling human communication dynamics via multiple hybrid dynamical systems (HDSs). Since each media signal can be described by the switches in submodels using the identified HDS, the overall model successfully extracts the concurrence of and temporal relation between multipart and multimodal signals as the temporal differences between switching times. Although we have primarily concentrated on modeling the dynamics in individual human behaviors rather than multiparty in this study, we are currently extending this scheme in order to model mutual interactions including turn-taking based on the coupling of HDSs.

Acknowledgments

This study is supported by the Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contracts of 18049046 and 21680016.

References

[1] H. Kawashima and T. Matsuyama, "Multiphase learning for an interval-based hybrid dynamical system," *IEICE Trans. Fundamentals*, vol. E88-A, no. 11, pp. 3022–3035, 2005.

[2] M. Kamachi, V. Bruce, S. Mukaida, J. Gyoba, S. Yoshikawa, and S. Akamatsu, "Dynamic properties influence the perception of facial expressions," *Perception*, vol. 30, pp. 875–887, 2001.

[3] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech and Audio Process*, vol. 4, no. 5, pp. 360–378, 1996.

[4] M. Nishiyama, H. Kawashima, T. Hirayama, and T. Matsuyama, "Facial expression representation based on timing structures in faces," *IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures (W. Zhao et al. (Eds.): AMFG 2005, LNCS 3723)*, pp. 140–154, 2005.

[5] H. Kawashima and T. Matsuyama, "Interval-based linear hybrid dynamical system for modeling cross-media timing structures in multimedia signals," *Int. Conference on Image Analysis and Processing*, pp. 789–794, 2007.

[6] P. Ekman and W. V. Friesen, *Unmasking the Face*. Prentice Hall, 1975.

[7] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1–15, 2002.

[8] M. Brand, "Voice puppetry," *SIGGRAPH*, pp. 21–28, 1999.

[9] H. Kawashima, Y. Horii, and T. Matsuyama, "Speech estimation in non-stationary noise environments using timing structure between mouth movements and sound signals," *Interspeech*, pp. 442–445, 2010.