Double Articulation Analyzer for Unsegmented Human Motion using Pitman-Yor Language Model and Infinite Hidden Markov Model

Tadahiro Taniguchi, Shogo Nagasaka

***Abstract*— We propose an unsupervised double articulation analyzer for human motion data. Double articulation is a two-layered hierarchical structure underlying in natural language, human motion and other natural data produced by human. A double articulation analyzer estimates the hidden structure of observed data by segmenting and chunking target data. We develop a double articulation analyzer by using a sticky hierarchical Dirichlet process HMM (sticky HDP-HMM), a nonparametric Bayesian model, and an unsupervised morphological analysis based on nested Pitman-Yor language model which can chunk given documents without any dictionaries. We conducted an experiment to evaluate this method. The proposed method could extract unit motions from unsegmented human motion data by analyzing hidden double articulation structure.**

I. INTRODUCTION

A. Motion Segmentation for imitation learning

In the context of imitation learning in robotics, ‘when to imitate’ is an important problem to be solved[1]. When a robot tries to imitate a person’s behaviors, the learner (robot) has to decide what segment of behavior to imitate from the demonstrator (human). For example, suppose a person approaches a robot and performs several motions (e.g., raising his/her hands, nodding several times, turning around and waving good-bye, and leaving). The displayed motion is unsegmented. The learner (robot) does not know which segment is worth to imitate. Therefore, it is important that the learner segments the demonstrated behavior and extracts unit motions from the exhibited continuous motion. Motion segmentation method has been intensively studied by many researchers[2], [3], [4], [5], [6], [7], [8]. The segmentation methods can be categorized into four classes[9]. The first and most classical type of these methods segments a target time series by focusing on local features in continuous motion time series data[2]. The second type focuses on local dynamics and the predictability of motion data[3], [4], [5]. The third type uses more complex nonlinear predictors that also use short-term context information[6]. The fourth type finds repeated segments from a continuous time series[7], [8]. Roughly speaking, The first and second method tends to segment a target motion data into too fine short-term motion sequences. The third and fourth method tends to extract long-term motion sequences which can be regarded as a meaningful segment by human observer from a target motion data. Taniguchi pointed out that each approach has each problem in [9]. It is important to distinguish the short-term segment and long-term segment. To distinguish the two types of segments explicitly in a learning model, a structure of double articulation should be taken into consideration.

B. Double Articulation

In the context of motion segmentation, Barbic distinguished between *high-level behavior* and *low-level behavior* [3]. Low-level behavior is a simple short motion segment which can be modeled by linear dynamics in contrast that Barbic and we are interested in semantically meaningful segment. Barbic call such semantically meaningful segments high-level behavior, such as walking, running, sitting, throwing a ball, and swinging a stick. A high-level behavior is more complex than a low-level behavior. The third and fourth types of methods discussed in the previous subsection can extract higher-level behaviors better than the first and second types.

In this paper, we propose a method for extracting high-level behavior by connecting several low-level behaviors by analyzing hidden double articulation structure of human motion. The method is based on the concept of double articulation, which is well known in semiotics.

Fig. 1 shows the basic concept of double articulation in our unsupervised double articulation analyzer. We, humans, have a double articulation structure in our spoken language and other many semiotic data. Most theories of speech recognition have following assumptions. First, a spoken auditory signal is segmented into phonemes (letters). Second, the phonemes are chunked into words. In most cases, we do not give any meanings to phonemes, but give certain meanings to words. We assume that human motion also has double articulation structure. This is our basic assumption of our proposed motion segmentation method. We assume that unsegmented motion is segmented to low-level behaviors based on its linearity or its locality of distribution in its state space. We call low-level behavior *elemental motion* in this paper. Elemental motions are chunked into a *unit motion*, which corresponds to a word in spoken language. We propose unsupervised motion segmentation method which analyzes double articulation structure in observed unsegmented data and extracts unit motions. We describe the algorithm of the double articulation analyzer based on nonparametric Bayesian theory.

This research was partially supported by a Grant-in-Aid Creative Scientific Research 2007-2011 (19GS0208) funded by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

T. Taniguchi and S. Nagasaka are in Department of Human & Computer Intelligence, Ritsumeikan University,1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan taniguchi@ci.ritsumei.ac.jp

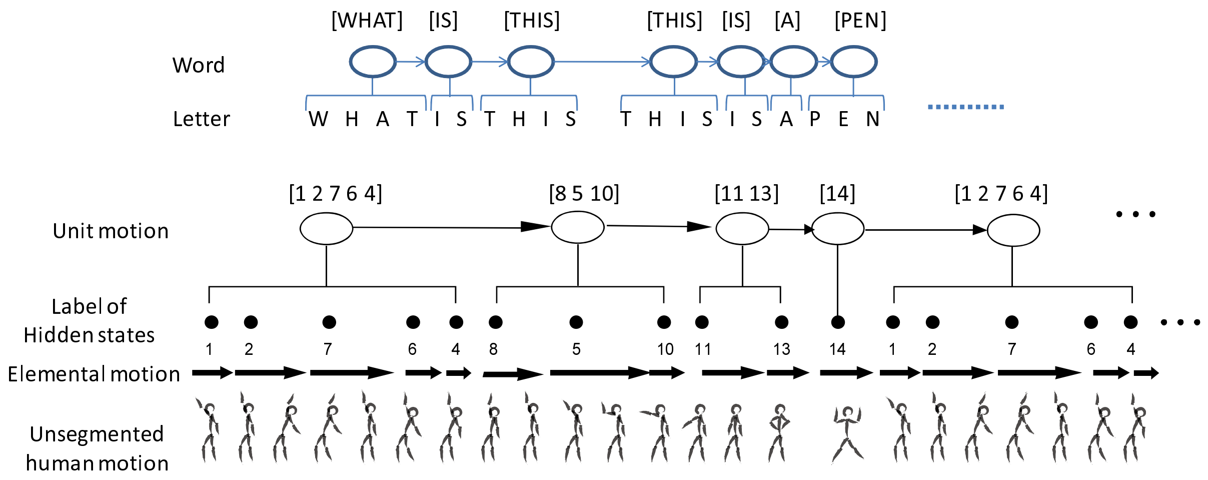


Fig. 1. Assumption of double articulation in motion segmentation

II. Algorithm

A. Overview

We give an overview of our proposed double articulation analyzer in this subsection. Fig. 2 shows a schematic overview of the overall learning architecture.

First, a large amount of high-dimensional motion data are observed by a robot and recorded. Singular value decomposition or other method reduces their dimensionality as preprocess. This reduces successive computational costs and extracts low-dimensional features, which mainly relate to unit motions embedded in unsegmented motions.

A sticky hierarchical Dirichlet process HMM (sticky HDP-HMM) [10] is used to segment and model the target preprocessed unsegmented human motion data. By using a sticky HDP-HMM, a robot can obtain elemental motions and sequences of labels of hidden states without fixing the number of types of elemental motions. We call a sequence of labels of hidden states that corresponds to observed unsegmented motion data a *document* (see Fig. 3). After obtaining a document, it is chunked into a sequence of words (sequence of letters). Taniguchi et al. used chunking method based on minimal description length (MDL) principle to solve the same problem. However, this method requires much computational time. In addition, the chunking method is heuristic probabilistic model, not a pure generative model. Mochihashi proposed an unsupervised morphological analysis method based on nested Pitman-Yor language model. Nested Pitman-Yor language model (NPYLM) is a nonparametric Bayesian language model[11]. That has two hierarchical Pitman-Yor (HPY) process. One is a language model which is an N-gram model of words and the other is a word model which is an N-gram model of letters. The language model is named NPYLM because an HPY word model is nested by an HPY language model. The language model enables unsupervised morphological analysis, unsupervised chunking of letters in other words. In this paper, we proposed to use NPYLM to chunk elemental motions to unit motions.

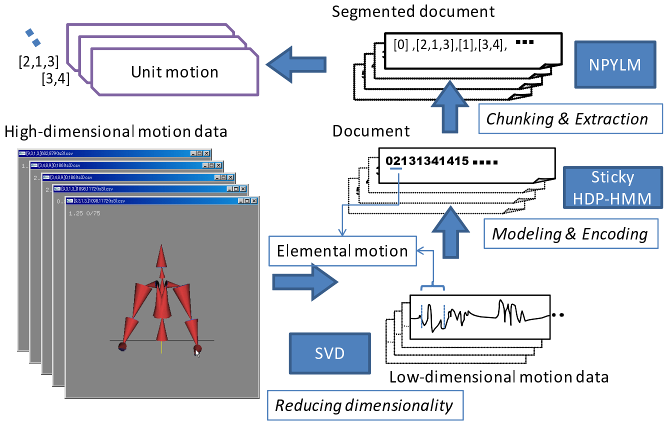
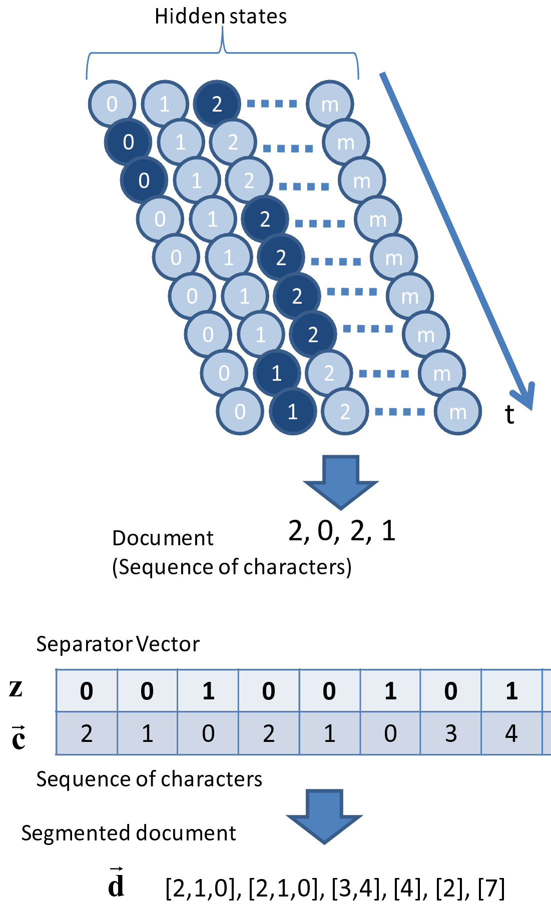


Fig. 2. Overview of proposed imitation learning architecture

B. Infinite Hidden Markov Model

The infinite hidden Markov model (iHMM), proposed by Beal [12], is a first nonparametric Bayesian statistical model which can be substituted for an HMM. HMM's selection probability of hidden states is temporally related in a Markovian manner. A potentially infinite number of hidden states are assumed with the iHMM. Through its inference process, the iHMM can flexibly estimate the number of hidden states. In a conventional HMM, the number of hidden states is fixed. The iHMM is a flexible statistical model whose number of hidden states is determined adaptively depending on given training data. However, it did not have an adequate generative model and an efficient inference algorithm.

The [13] extends the HDPM into the hierarchical Dirichlet process-hidden Markov model (HDP-HMM), which is an adequate generative model for iHMM. In the HDP-HMM, the SBPs GEM(γ) having the concentration parameter γ produces β, which produces π_k for all hidden states. π_k is a multinomial distribution corresponding to each hidden state. In a generative process, the next state is selected using a



$\mathbf{z}$	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**1**
$\vec{\mathbf{c}}$	2	1	0	2	1	0	3	4	2	7

Fig. 3. (Left) sequence of hidden states is transformed into document. (Right) separator vector determines segmentation of given document.

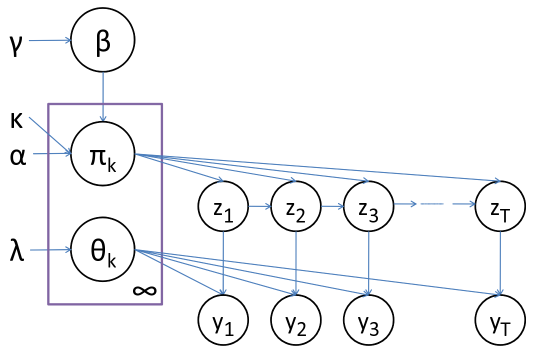


Fig. 4. Graphical model of sticky HDP-HMM

multinomial distribution corresponding to the hidden states. This corresponds to transition matrix in classical HMM. This means that the HDP-HMM has transition matrices having potentially infinite dimensions.However, the HDP-HMM has a problem that hidden states transit to other hidden states too frequently. This comes from the fact that π_k does not have any self transition bias. In contrast, a hidden state is expected to be sustained for a certain number of time steps from practical use of HMMs in continuous dynamical systems, e.g., modeling and segmenting, spoken language, human motion, and data from sensory networks. To overcome this problem, stick HDP-HMM was proposed.

C. sticky HDP-HMM

Fox et al. [10] proposed a sticky HDP-HMM with a self-transition bias [10]. This model is an extension of the HDP-HMM. By biasing the self transition probability, this sticky HDP-HMM can reduce the frequency of transition among hidden states. Therefore, this model is more effectively used to model and segment a continuous observed real data stream, e.g., speaker diarization and speech recognition. If the segmentation process, outputting elemental motions, produces too many fragments, i.e., too many state transitions, the posterior word extraction process does not work well and cannot extract unit motions. For our purpose, the stickiness the sticky HDP-HMM provides is important. A graphical model of sticky HDP-HMM is shown in Fig.4.

Fox et al. also describes a numerical computation algorithm using a blocked Gibbs sampler. Straight-forward application of the forward filtering-backward sampling algorithm for an HMM [14] to the iHMM is not feasible because it is impossible to accumulate forward messages for an infinite number of hidden states. Therefore, halting an SBP and truncating the number of hidden states are unavoidable. Fox et al. proposed a blocked Gibbs sampler by adopting weak-limit approximation. This accelerates the inference sampling process in the HDP-HMM. Practically, the approximation is not so problematic for the purpose of motion learning. Therefore, we adopted the blocked Gibbs sampler proposed by Fox et al. [10]. The precise formulation, derivation, and discussion of the sticky HDP-HMM and its blocked Gibbs sampler is omitted in this paper [1]. In this paper, we use the weak-limit approximation by Fox et al. for practical use.

D. Nested Pitman-Yor Language Model

We assume that a unit motion consists of a chunk of elemental motions. This corresponds to the relationship between a spoken word and phonemes. Taniguchi et al. [15], [16] proposed an imitation learning architecture that enables a robot to extract characteristic unit motions from unsegmented hand movement by using heuristic keyword extraction method. However, this keyword extraction method highly depends on several hand-coded parameters and initial conditions. In contrast, Tanaka et al. and Taniguchi et al. developed a motion segmentation method [17] based on the MDL principle. However, that requires high computational cost.

To chunk sequential letters into several words corresponds to morphological analysis in linguistics. Several researchers have already proposed unsupervised morphological analysis methods for segmenting documents by using nonparametric Bayesian language models [11], [18]. In the field of motion segmentation, unit motions are usually unknown in contrast with words in spoken language recognition. Therefore, when we analyze unsegmented motion data the analyzer has to

[1] For more information, see Fox et al.'s research [10].

treat with unknown words (sequence of letters corresponding to hidden states in HMM). Unsupervised morphological analysis does not assume preexisting dictionary. Therefore, it is suitable for motion analysis. Mochihashi [11] proposed an unsupervised morphological analysis method based on Nested Pitman-Yor language model (NPYLM). It uses letter N-gram in addition to word N-gram model. The both of them use Pitman-Yor process to smooth their probability. Mochihashi uses NPYLM and probabilistic dynamic programming to chunk sentences written in natural language.

1) Pitman-Yor process: HPYLM is an N-gram language model using hierarchical Pitman-Yor process. Pitman-Yor process is a stochastic process whose base measure is itself Pitman-Yor process which is a generalization of Dirichlet process.

In HPYLM, probability of word w after a context $h = w_{t-n} \dots w_{t-1}$ is calculated as follows.

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_h}{\theta + c(h)} p(w|h') \qquad (1)$$

h' is a context whose order is one less than h, i.e., $h' = w_{t-n-1} \dots w_{t-1}$. Therefore, $p(w|h')$ becomes a prior probability of w after h, and their probability is calculated recursively. $c(w|h)$ is a count of w after in a context h, and $c(h)$ is summation of all words' counts in a context of h.

t_{hw} is a count that w is estimated to be generated from the context of h', and t_h is summation of t_{hw} in a context of h. Discount parameter d and concentration parameter θ is hyper parameters of HPYLM.

When we calculate unigram probability $p(w|h)$, $p(w|h')$ does not exist. To overcome this problem, we use letter N-gram smoothed by HPYLM as a base measure of word unigram. This gives word HPYLM reasonable base measure without preparing word dictionary.

E. Morphological Analysis using blocked Gibbs sampler

NPYLM enables us to calculate N-gram probability without prepared dictionary. Our proposed method analyzes sequence of letters by using NPYLM. Blocked Gibbs sampler and probabilistic dynamic programming enables NPYLM to chunk given letter sequences without heavy computational time.

Blocked Gibbs sampler eliminates words included in a sequence (document) from the language model and sample new chunking by using language model. After chunking, it adds the sampled words to the language model and updates the language model. NPYLM becomes optimized by sampling segmentation repeatedly. Forward filtering-Backward sampling algorithm is used to segment target sentences in the unsupervised morphological analysis in [11].

III. EXPERIMENT

We conducted an experiment to evaluate our double articulation analyzer, a sticky HDP-HMM for modeling unsegmented human motion, and unsupervised morphological analysis method using NPYLM. In this experiment, recorded high dimensional time series data representing unsegmented human upper body motions were input into the learning architecture as learning data samples[2].

A. Experimental Conditions

Human upper body motion, which includes a DOF of 36, was recorded using the motion capture system Gypsy 5 Torso (Meta motion). Each joint angle of a human's upper body. A participant was required to manipulate three target objects. When the participant manipulated an object, he/she was required to exhibit a corresponding unique unit motion.

In these experiments, we used captured human motion data as a learning data set. We asked a participant to move for 20 seconds while being recorded. The frame rate was 60 Hz. During the recording session, we asked the participant to manipulate three types of target objects, a toy, ball, and stick. The participant was asked to arbitrarily switch target objects, and the continuous motion was recorded as an unsegmented motion. The participant was allowed to insert small arbitrary motions between two specific unit motions. Therefore, the time series contained the three types of bodily motions without any explicit segmentation.The dimensionality of the recorded data was reduced to 6 by using singular value decomposition[9] .

We set the sticky HDP-HMM parameters $\alpha = 0.1, \gamma = 0.1, and \kappa = 0.9$ as default values. Before the sticky HDP-HMM learning phase, the Gaussian prior distribution's mean value and variance were calculated and set to $\mu_0 = 0$, and $\Sigma_0 = d_s I$, respectively, which are the hyperparameters of the prior distribution of the mean vectors of each emission distribution. We set the DOF as 9. $\Delta = d_f I$ for the inverse-Wishert distribution, which is a prior distribution of the variance-covariance matrices of the Gaussian distribution. We set the hyperparameter of emission distribution $d_f = 0.5 \times 10^{-4}$ by referring to the variance of data in state space. We iterated Gibbs sampling 10 times in sticky HDP-HMM. In NPYLM, we used discount parameter $d = 0.5$ and concentration parameter $\theta = 0.1$. Blocked Gibbs sampler repeated 200 times.

B. Result

The output sequences corresponding to three data from sticky HDP-HMM are shown in Fig. 5. This sequence was given to unsupervised morphological analyzer using NPYLM as input data set. The output from NPYLM is shown in 6. The parenthetic subsequences are chunked letters corresponding to words in documents. The chunked letters correspond to a unit motion.

Fig .6 shows that (10 4), (16 4 6 11 6), (16 17), (10), (4 16 4), (10 4 10 4), (16), (17) were extracted as unit motions. By reviewing the recorded movie, we found that (16 4 6 11 6) , (4 16 4) and (10 4) are corresponding to playing with a toy (Fig 7), rolling a ball (Fig. 8) and swinging a stick (Fig. 9), respectively. (10 4) was a subsequence of (10 4 10 4). Unit motions corresponding to (16 17), (16), (17) were found to be shorter than 1 second. They are motions

[2] This experiment uses the same data set as [9]

- 10 4 16 4 6 11 6 16 4 6 11 6 16 17 4 16 4 10 4 10 4 10
- 10 4 10 4 10 4 16 4 6 11 6 16 4 6 11 6 16 4 6 11 6 16 4 16 4 10
- 10 4 16 4 6 11 6 16 4 6 11 6 16 17 10 4 10 4 16 4 6 11 6 16 17 4 16 4 10

Fig. 5. Transition of hidden states of Sticky HDP-HMM

- (10 4) (16 4 6 11 6) (16 4 6 11 6) (16 17) (4 16 4) (10 4) (10 4) (10)
- (10 4 10 4) (10 4) (16 4 6 11 6) (16 4 6 11 6) (16 4 6 11 6) (16) (4 16 4) (10)
- (10 4) (16 4 6 11 6) (16 4 6 11 6) (16 17) (10 4 10 4) (16 4 6 11 6) (16) (17) (4 16 4) (10)

Fig. 6. A sample of chunked label sequence obtained by using NPYLM

switching between the main unit motions. This result shows that our double articulation analyzer could extract three unit motions corresponding to three objects from unsegmented motion data.

We compared the computational time of proposed method with the MDL-based chunking method Taniguchi et al. used[9]. We prepared larger data sets by duplicating measured motion data. Fig. 10 shows the relative computational time compared with the time chunking 3 data sets required. This shows the increase in computational time in NPYLM is smaller than that in MDL. MDL approach requires calculation of description length for each repetition. This increases computational time when the size of data set increases. On the other hand, NPYLM does not require such recalculation requiring big computational cost. Mochihashi [11] reported that blocked Gibbs sampler reduced computational time greatly. The same result was obtained in this experiment.

IV. Conclusions and Future Works

In this paper, we proposed a new double articulation analyzer using sticky HDP-HMM and NPYLM and evaluate its effectiveness through a simple experiment. The analyzer could extract unit motion by using NPYLM without preexisting dictionary. It also shown that NPYLM requires less computational time than MDL-based chunking method[9].

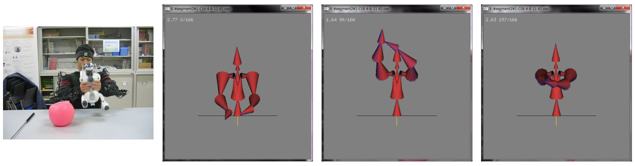


Fig. 7. Holding up a dog-like robot (16 4 6 11 6)

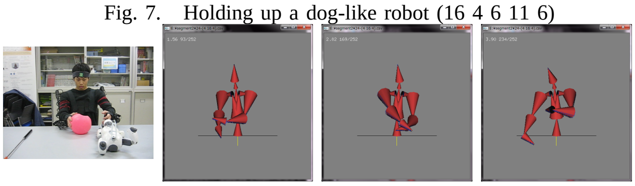


Fig. 8. Rotating a ball (4 16 4)

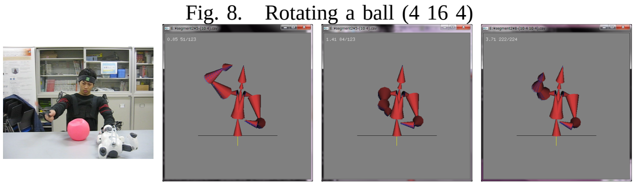


Fig. 9. Swinging a stick (10 4 10 4)

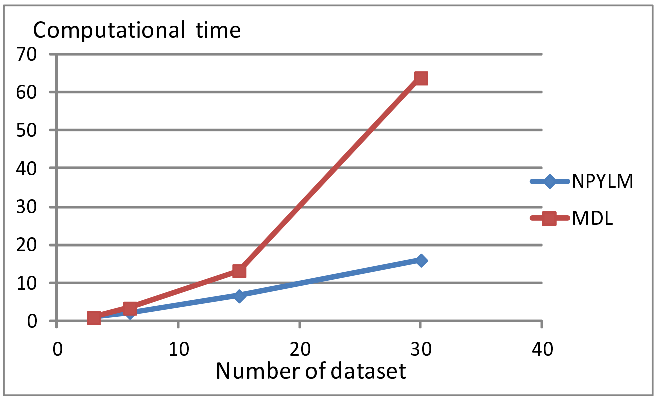


Fig. 10. Comparison of computation time of each chunking method

However, the proposed method is not a theoretically unified model. Sticky HDP-HMM and NPYLM are both based on nonparametric Bayesian theory, but work separately. Ideally, the double articulation analyzer works by using a language model and a dynamical model (HDP-HMM) interactively. If a double articulation analyzer uses the information of language model when it segment unsegmented original time series, we can unify the segmentation process and chunking process into a generative model. This is our future work. In addition, reducing computational time of sticky HDP-HMM which requires much computational time is necessary to apply this method to a large data set.

This method is unsupervised learning method. Therefore, obtained results fully depend on the dataset of motion data. In order to extract a unit motion from unsegmented motion data, the unit motion should be observed several times in the dataset because the NPYLM uses frequency of words to estimate probability of words. It's difficult to analyze how much data is necessary for proper segmentation. However, we showed that a big dataset is not necessary for motion segmentation in our experiment. To analyze our proposed method from this viewpoint is also our future work.

References

[1] A. Alissandrakis, C. Nehaniv, and K. Dautenhahn, “Imitation with alice: Learning to imitate corresponding actions across dissimilar embodiments,” *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 32, no. 4, pp. 482–496, 2002.

[2] A. Fod, M. Matarić, and O. Jenkins, “Automated derivation of primitives for movement classification,” *Autonomous robots*, vol. 12, no. 1, pp. 39–54, 2002.

[3] J. Barbič, A. Safonova, J. Pan, C. Faloutsos, J. Hodgins, and N. Pollard, “Segmenting motion capture data into distinct behaviors,” in *Proceedings of Graphics Interface 2004*, 2004, pp. 185–194.

[4] H. Kawashima and T. Matsuyama, “Multiphase learning for an interval-based hybrid dynamical system,” *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. E88-A, no. 11, pp. 3022–3035, 2005.

[5] Y. Li, T. Wang, and H. Shum, “Motion texture: a two-level statistical model for character motion synthesis,” in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 465–472.

[6] T. Ogata, S. Matsumoto, J. Tani, K. Komatani, and H. Okuno, “Human-robot cooperation using quasi-symbols generated by rnnpb model,” in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 2156–2161.

[7] H. Kadone and Y. Nakamura, “Segmentation, memorization, recognition and abstraction of humanoid motions based on correlations and associative memory,” in *IEEE-RAS International Conference on Humanoid Robotics*, 2006, pp. 1–6.

[8] S. Chiappa and J. Peters, “Movement extraction by detecting dynamics switches and repetitions,” in *Advances in Neural Information Processing Systems 23*, 2010, pp. 388–396.

[9] T. Taniguchi, K. Hamahata, and N. Iwahashi, “Unsupervised segmentation of human motion data using a sticky hierarchical dirichlet process-hidden markov model and minimal description length-based chunking method for imitation learning,” *Advanced Robotics*, vol. 25, pp. 2143–2172, 2011.

[10] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “The sticky hdp-hmm: Bayesian nonparametric hidden markov models with persistent states,” MIT Laboratory for Information and Decision Systems, Tech. Rep. 2777, 2007.

[11] D. Mochihashi, T. Yamada, and N. Ueda, “Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, 2009, pp. 100–108.

[12] M. Beal, Z. Ghahramani, and C. Rasmussen, “The infinite hidden Markov model,” *Advances in Neural Information Processing Systems*, vol. 1, pp. 577–584, 2002.

[13] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[14] S. Scott, “Bayesian methods for hidden markov models: Recursive computing in the 21st century,” *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 337–351, 2002.

[15] T. Taniguchi, N. Iwahashi, K. Sugiura, and T. Sawaragi, “Constructive approach to role-reversal imitation through unsegmented interactions,” *Journal ref: Journal of Robotics and Mechatronics*, vol. 20, no. 4, pp. 567–577, 2008.

[16] T. Taniguchi and N. Iwahashi, “Computational model of role reversal imitation through continuous human-robot interaction,” in *Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*. ACM New York, NY, USA, 2007, pp. 25–31.

[17] Y. Tanaka, K. Iwamoto, and K. Uehara, “Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle,” *Machine Learning*, vol. 58, no. 2, pp. 269–300, 2005.

[18] Y. Teh, “A hierarchical bayesian language model based on pitman-yor processes,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 985–992.