Communicative Rhythm in Gesture and Speech

Ipke Wachsmuth

Faculty of Technology, University of Bielefeld
D-33594 Bielefeld, Germany
ipke@techfak.uni-bielefeld.de

Abstract. Led by the fundamental role that rhythms apparently play in speech and gestural communication among humans, this study was undertaken to substantiate a biologically motivated model for synchronizing speech and gesture input in human computer interaction. Our approach presents a novel method which conceptualizes a multimodal user interface on the basis of timed agent systems. We use multiple agents for the purpose of polling presemantic information from different sensory channels (speech and hand gestures) and integrating them to multimodal data structures that can be processed by an application system which is again based on agent systems. This article motivates and presents technical work which exploits rhythmic patterns in the development of biologically and cognitively motivated mediator systems between humans and machines.

1 Introduction

Gesture and speech are the corner stones in natural human communication. Not surprisingly, they are each paid considerable attention in human-machine communication. It is apparent that advanced multimedia applications could greatly benefit from multimodal user interfaces integrating gesture and speech. Nevertheless, their realization faces obstacles for which research solutions to date have barely been proposed. The multimodal utterings of a user have to be registered via separate channels, as concurrent speech and gesture percepts. These channels have different time delays, that is, information from signal preprocessing is distributed in time. In order to process gesture and speech in their semantic connection, their temporal correspondence must first be reconstructed.

Observations in diverse research areas suggest that human communicational behavior is significantly rhythmic[1] in nature, for instance, in the way how spoken syllables and words are grouped together in time (speech rhythm) or how they are accompanied by body movements, i.e. gestures[2]. In theoretic and practical approaches attempting to mimic natural communication patterns in human-computer interaction, rhythmic organization has so far played a non-existent role. This paper takes a stance

[1] **Rhythm**: Following Martin [12] we define "rhythm" to mean relative timing between adjacent and nonadjacent elements in a behavior sequence, i.e., the locus of each element along the time line is determined relative to the locus of all other elements in the sequence.

[2] **Gesture**: For the purpose of this paper it is sufficient to understand "gestures" as body movements which convey information that is in some way meaningful to a recipient.

that rhythmic patterns[3] provide a useful mechanism in the establishment of intra-individual and inter-individual coordination of multimodal utterances. Based on a notion of timed agent systems, an operational model is proposed which is stimulated by findings from empirical research and which was explored in multimodal perception and integration of concurrent modalities, in particular, speech and hand gestures.

In the next section, we discuss representative findings from empirical research that substantiate the function and role of rhythm as it pertains to human communication. We then argue, in Section 3, that the idea of rhythmic organization should be a good starting point to deal with some problems of multimodal interfaces for accepting open input. The original contribution of the article lies in conceptualizing an agent-based model, described in Part 4, that accounts for some of the empirical findings and makes them available for technical solutions. A multimodal input agency is described which builds on rhythmic patterns and which served as a framework for conceptualizing a human-computer interface. Results and further prospects are discussed in Part 5. In the age of information society, rhythms might also be a more general paradigm for human machine communication, and we conclude with a brief vision of this aspect.

2 Rhythm in Human-Human Communication

Various findings from psychological and phonetics research have revealed forms of rhythmic synchronization in human communicational behavior, with respect to both the production and the perception of utterances. Like the coordination of rhythmic limb movement (for a review, cf. [21]), speech production and gesturing requires the coordination of a huge number of disparate biological components. When a person speaks, her arms, fingers, and head move in a structured temporal organization (self-synchrony), which was found to be synchronized across multiple levels [4]. The so-called gesture stroke is often marked by a sudden stop which is closely coupled to spoken words. Particularly for stress-timed languages[4], when spoken fluently, temporal regularities are observed between stressed syllables and accompanying gesture strokes. They are more clear for pointing gestures/deictics [17], whereas gestural beats and verbal stress are not synchronized in a strict rhythmic sense [16]. Furthermore, it was found that the rhythm in a speaker's utterances is readily picked up by the hearer (interactional synchrony), in that the body of a listener, within short latency following

[3] **Rhythmic patterns** are event sequences in which some elements are marked from others (accented); the accents recur with some regularity, regardless of tempo (fast, slow) or tempo changes (accelerate, retard) within the pattern. Since rhythmic patterns have a time trajectory that can be tracked without continuous monitoring, perception of initial elements in a pattern allows later elements to be anticipated in real time; cf. [12], [13].

[4] **Stress-timed language:** In general phonetics, it is assumed that "stress-timed" languages like English, German, and Danish tend to have a relatively constant duration of stress groups, independent of the actual number of phones or syllables involved in these groups. Thus, the time duration between the capitalized syllables in e.g. (a) "the BUS to GIF" and (b) "the BUSes to VerSAILLES" may be expected to be approximately the same when spoken by the same speaker under the same external conditions; cf. [3].

sound onset, entrains to the articulatory structure of a speaker's speech [4]; there may even be interpersonal gestural rhythm [16].

Under constrained conditions, Cummins and Port [6] found a metrical 'foot' to be a salient unit in the production of speech for native English speakers. Quasi-rhythmical timing phenomena in unconstrained speech production (text reading, mostly Swedish) are reported by Fant and Kruckenberg [7]: An average of interstress intervals[5] of the order of 500 ms (milliseconds) appears to function as a basic metrical reference quantum for the timing of speaking pause duration, and quantal rhythmic sub-units of the metrical foot are suggested by average durations of stressed syllables, unstressed syllables and phoneme segments of the order of 250 ms, 125 ms and 62.5 ms. The tempo and coherence of rhythmic patterns is speaker-specific; and average segment durations within a phrase are influenced by the density of content words and thus are not entirely "on foot". Similarly, Broensted and Madsen [3] have found intra-speaker variabilities in speech rates of English and Danish speakers due to time equalization of stress groups and utterances.

As for perception, Martin [12]; [13] observed that rhythmic and segmental aspects of speech are not perceived independently in that segmentation is guided by rhythmic expectancy. Temporal phenomena were identified by Pöppel [20] on two significant time scales. Indication was found for a high-frequency processing system that generates discrete time quanta of 30 ms duration, and a low-frequency processing system that sets up functional states of ~3s. Evidence for the high-frequency processing systems comes, in part, from studies on temporal order thresholds: Independent of sensory modality, distinct events require a minimum of 30 ms to be perceived as sucessive. The low-frequency mechanism binds successive events of up to 3s into perceptual units. Support for such a binding operation comes from studies on the temporal reproduction of stimuli with different duration; temporal integration for intervals up to 2-3s has also been observed with movement control and with the temporal segmentation of spontaneous speech. This integration is viewed to be automatic and presemantic in that the temporal limit is not determined by what is being processed.

Explanations found by the above-mentioned researchers agree in the observation that communicative rhythm may be seen as a coordinative strategy which enhances the effectiveness of speaker-listener entrainment. By expectable periodicities, rhythm seems to provide anticipations which help listeners perform segmentation of the acoustic signal and synchronize parts of speech with accompanying gesture. That is, the listener is apparently able to impose a temporal, 'time window'-like structure in the perception of utterances which aids in the grouping and integration of the information transmitted. A specific universal integration mechanism is suggested by the Pöppel [20] studies: Intervals of up to 3s can be mentally preserved, or grasped as a unit. This is particularly true for cross-connections among the different sensory modalities, and this temporal integration is viewed as a general principle of the neuro-cognitive machinery.

[5] **Interstress interval:** the time measured from the onset of the vowel in a stressed syllable to the onset of a vowel in the next stressed syllable, excluding those interrupted by a syntactic boundary.

3 Rhythm in Human-Machine Communication

As was argued above, there is evidence that communication among humans is strikingly rhythmic in nature. When this is true, then this observation should also be relevant in human-machine communication. For instance, Martin [13] has suggested that computational models of speech perception by humans should incorporate a *rhythmic expectancy component* which, starting from utterance onset, extrapolates ahead within the constraints supplied by the current information. In human-machine communication such approaches to mimic biological communication patterns have yet to be attempted.

At the same time the call for multimodal user interfaces, like interfaces that combine the input modalities of speech and gesture in a computer application, requires a more explicit understanding of how these modalities are perceived and integrated. Multimodal input facilities are crucial for a more natural and effective human-computer interaction where information of one modality can serve to disambiguate information conveyed by another modality [14]. Building multimodal input systems requires, on the one hand, the processing of single modalities and, on the other hand, the integration of multiple modalities [5]. To enable a technical system to coordinate and integrate perceived speech and gestures in their natural flow, two problems have to be solved [23]:

The segmentation problem: Given that the system is to process open input, how is the right chunk of information determined that the system takes in for processing at a time? How are consecutive chunks linked together?

The correspondence problem: Given that the system is to integrate information from multiple modalities, how does it determine cross-references, i.e., which information from one modality complements information from another modality?

To date, research solutions have barely been proposed how to reconstruct a user's multimodal utterings, which are registered on separate channels and distributed in time, in their natural temporal connection. Early attempts to realize a multimodal input system are the PUT-THAT-THERE system [1] and CUBRICON [18]. These systems are restricted to analyze speech and gestural input sequentially, and they do not allow gestural input in a natural form but, rather, as static pointing direction. More recent systems, e.g. [9]; [2]; [19], allow the parallel processing of two or more modalities. Nevertheless these approaches do not support what is called open input, i.e. instructing a system without defining where an instruction starts or ends, as well as the resolution of redundancies or inconsistencies between pieces of information of different modalities.

The observations in the previous section suggest that the analysis of communicative rhythm could be used to improve technical mediator systems between humans and machines. By exploiting segmentation cues, such as gesture stroke and stress beat in speech, the communicative rhythm could be reproduced, and possibly anticipated on, by the system. It could help to impose time windows for signal segmentation and determine correspondence of temporally distributed speech and gesture percepts which precede semantic analysis of multimodal information.

4 A Multimodal Interface Based on Timed Agents

In a first technical approach we have employed the idea of communicative rhythm to determine how spoken words and hand pointing gestures belong together. For a preview, the multimodal input stream is segmented in time windows of equal duration, starting from utterance onset in one modality. Input data from multiple modalities registered within one time cycle are considered as belonging to the same instruction segment, and cross-references are resolved by establishing correspondence between gesture percepts and linguistic units registered within a time cycle. As this will not always work, time-cycle-overspanning integration needs also be considered. These ideas are in the first place motivated by the above-mentioned findings on temporal perception in humans [20] and earlier ideas about rhythmic expectancy in speech perception [13].

4.1 Materials and Methods

The setting of our work is communicating with virtual environments, i.e., computer-graphics-based three-dimensional scenes which can be changed interactively by user interventions. The study reported here was carried out in the VIENA project [24] where the prototypical application example is the design of a virtual office environment. The VIENA system can process instructions from a user to execute alterations of the scene by means of an agent-based interface. Instructions can be transmitted by spoken natural language and by pointing gestures which are issued via a simple Nintendo data glove. In this study we have used a Dragon Dictate Version 1.2b speech recognizer which processes (speaker-dependent) isolated words. An instruction is spoken as a sequence of words:

put | <gesture> this | computer | on | <gesture> that | table

where the sound onsets of consecutive words follow each other by approx. 600 ms. Pointing gestures are issued, at about the time of the spoken "this" or "that", by glove-pointing at one of the displayed objects. A glimpse of the environment that was used in this study can be obtained from Figure 1.

As the principal method to register and process information perceived from different sensory channels, we use a processing model that realizes distributed functionalities by the interplay of multiple software agents. The single agent is an autonomous computational process that communicates and cooperates with other agents based on a variant of the contract-net protocol [25]. A system of such agents, termed "agency", realizes a decentral processing of information. The core of the VIENA agency (cf. Figure 2) consists of a number of agents that take part in mediating a user's instruction to change the scene in color and spatial layout. Typically, the functionality of each single agent is achieved in a sense-compute-act cycle, i.e., `sense` input message data, `compute` function, `act` by sending resulting messages to other agents, or to effectors like the graphics system.

The basic model of agent performance is event-driven, that is, there are no temporal constraints as to when a cycle is completed. However, in the context of integrating

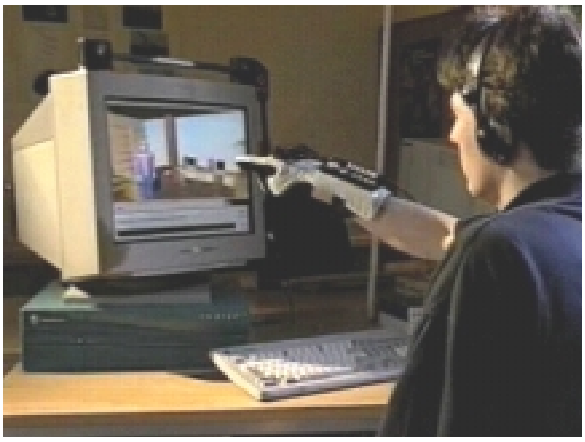

Fig. 1. Instructing the VIENA system by combined speech and gesture input

modalities from different sensors, temporal processing patterns become also relevant and especially so when taking into account a close coupling of speech and gesture input. Led by this observation, we have extended the basic agent model to be *timed*. To this end, we have provided for a temporal buffer for sensed information and, besides event-driven control, temporal constraints by way of time-cycle-driven patterns of processing, supporting a low-frequency "rhythmic" segmentation procedure.

In our first approach, time cycles spanning a `sense-buffer-compute-act` sequence executed by the single agents have a fixed duration which can be varied for experiments. The multimodal input agency described below is comprised by a number of agents dedicated to (1) sensory and linguistic input analysis and (2) the coordination and processing of multimodal input information.

4.2 Multimodal Input Agency

To address the aspects of open input and correspondence in multimodal instructions, we have developed a multimodal input agency, as shown in the right part of Figure 2. It is comprised by a set of timed listener agents which record, analyze, and elaborate input information from different sensory channels, and a coordinator mechanism, also realized as a timed agent system, which integrates analyzed sensory information. This information is then passed on to the application system (mediating agency) shown in the left part of Figure 2.

The input agency consists of a set of modality-specific input listeners, a parser for linguistic analysis, and a coordinator. Three listener agents, i.e., a speech listener, a type listener, and a gesture listener, track and analyze sensor data from the microphone, the keyboard, and the data glove, respectively. Assisted by the parser, the coordinator analyzes and integrates the inputs received from the listeners and generates an internal task description that is posted to mediator agents. The mediating agency determines the according changes in the virtual environment and updates the scene visualization.

Multimodal instructions are issued by speaking to the microphone and using the glove for pointing. Typewritten input can be used in unimodal (verbal) instructions.

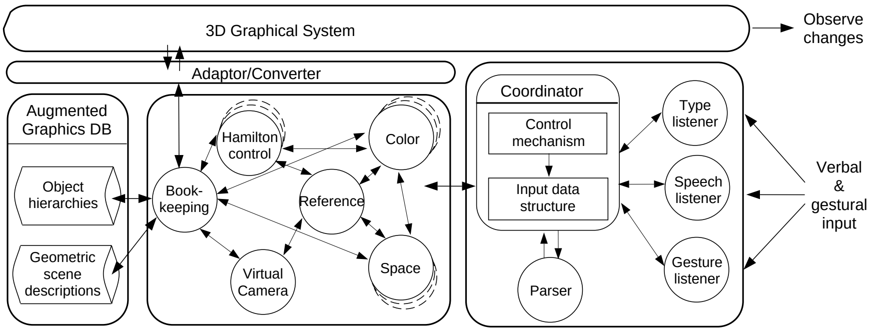

Fig. 2. VIENA agent interface with mediators (left) and multimodal input agency (right)

The input agency performs a time- and event-driven routine to integrate multiple (speech and gesture) modalities. Whereas input agents are "listening" for input events in short polling cycles of 100 ms, the coordinator agent processes information in fixed time cycles of a longer periodicity of 2 seconds. The actual values were found by experiments with the VIENA system which have shown that time cycles with durations of 100 ms and 2 seconds, resp., work best for the single-word recognition system and glove-based gesture recognizer used in the study. The 100 ms rhythm was determined by the fact that the glove sends a maximum of 10 data packets per second; thus a higher-frequency polling would cause unnecessary communication overhead.

The 2s integration rhythm was determined in experiments probing the overall computational cost of the VIENA system, as measured from the onset of a speech instruction to the output of a new scene visualization while varying the length of the integration cycle time by 1-second increments. In these experiments we used instructions of different lengths, i.e. a 4-word, a 7-word, and a 10-word instruction. The sound onsets of consecutive words were computer-controlled to follow each other by 600 ms, independent of whether one-, two-, or four-syllable words were spoken in. That is, speech input for the 4, 7, 10-word sentences took a bit more than 1800, 3600, and 5400 ms, respectively. The following, unimodal, spoken instructions were used ("saturn" and "andromeda" are names that refer to the two computers shown on the screen in Figure 1):

move | the | chair | left
put | the | palmtree | between | saturn | and | andromeda
put | the | palmtree | between | the | back | desk | and | the | bowl

The integration process realized in the input agency is a combination of time and event-driven computations. In the following sections we explain in more detail how the segmentation and the correspondence problem (cf. Section 3) are treated in the VIENA multimodal input agency. In full detail the method is described in [11].

4.3 Open Input Segmentation: The Tri-State Rhythm Model

The basic approach to segment the multimodal input stream is to register input events from the different modalities in time cycles imposed by the coordinator agent, resulting in a tri-state rhythm model which is illustrated in Figure 3. As input data within one time cycle is considered as belonging to the same instruction segment, the coordinator agent, accordingly, buffers information received from the speech and gesture listeners, to integrate them when a cycle is completed (cf. Section 4.4).

The first time cycle (z1) starts at signal onset when the user inputs a (verbal or gestural) instruction, resulting in a first input event (e1 at time te1). This causes the coordinator to reach a state "swing" which continues as long as signals are received on one of the listener channels, modeling a rhythmic expectancy. The coordinator subsides swinging when no further input event occurs within a full cycle. The "subside" state changes to "wait" once that k, e.g. 2, event-free cycles are recognized or, when triggered by a new event, returns to "swing". The "wait" state is of indefinite time; it will change to the "swing" state again upon receiving a new input event.

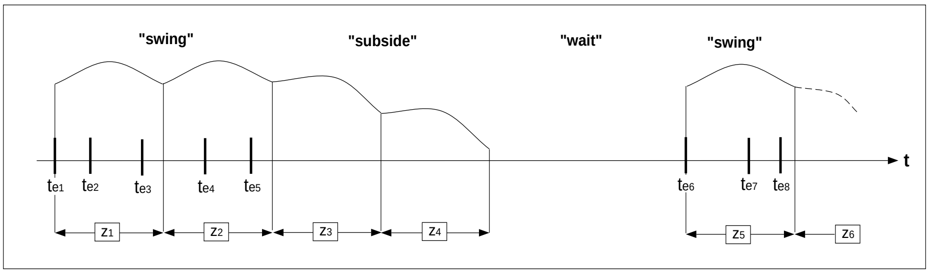

Fig. 3. Tri-state rhythm model (swing–subside–wait); each cycle in state "swing" or "subside" timed equally.

The time- and event-driven integration method is interwoven with the segmentation process. It consists of a cyclical four-step process comprised by functions `sense`, `buffer`, `compute`, and `act`. Whereas `sense` and `buffer` are continued until the current time cycle is completed, `compute` and `act` are executed at the end of each time cycle. The function `sense` allows that input events sent by the listeners are received as messages, whereas the function `buffer` extracts relevant message information and collects them in an input data structure which is organized in time cycles. The coordinator agent performs these two steps as long as the current time cycle has not elapsed. At the end of a time cycle, the function `compute` interprets the multimodal information stored in the input data structure. Afterwards, the function `act` determines appropriate agents in the mediator agency and posts the corresponding tasks to them.

4.4 Correspondence in Multimodal Integration

The interpretation function `compute` resolves cross-references between verbal and gestural information in the input data structure and produces an overall task description that corresponds to the multimodal input of the user. Two cases are distinguished: (1) in the *time-cycle-internal interpretation*, information of just the most recent time cycle is used; (2) in the *time-cycle-overspanning interpretation*, data of the last n time cycles is used. Having determined what kind of interpretation has to be performed, the coordinator analyzes the speech and gesture modality separately and merges information of the different modalities in a multi-step evaluation procedure that considers both temporal and linguistic features to compute the most appropriate cross-references. Then it disambiguates all kinds of references with the help of specific agents in the mediator agency, and checks whether or not the resulting instruction is complete with respect to domain-dependent requirements. If incomplete, the coordinator waits for information that expectedly would occur within the next time cycles or, when cycling has subsided, it presents the user with his/her incomplete instruction for editing.

The actual integration in the `compute` phase is done by establishing correspondence between gesture percepts and so-called gesture places within integration intervals. *Gesture places* are time-stamped information slots, determined in spoken-language analysis, which formalize expectations about events that provide missing object or direction specifications from the gesture channel. Potential gesture places are specifications of reference objects or locations derived from speech input. The valuation of gesture places is calculated by the heuristics "the more ambiguous a reference described in the verbal instruction, the higher the valuation of a gesture place." If there are two gesture places for only one gesture percept, resolution of correspondence between cross-modal events is led by their closeness in time and by comparing ambiguity values associated with speech sections; e.g., "the chair" is less ambiguous (with respect to reference) than the deictical "there" in the sentence "put the chair there." An example where closeness in time is relevant is the instruction "put this computer on that table" if only one gesture percept is available (presupposing that one indexical is clear from previous context). In this case closeness in time would be indicative in that one of the pairs "<gesture> this" or "<gesture> that" would have higher weight. Further examples for possible combinations of speech and gesture inputs to disambiguate objects and locations are following:

put | <gesture> this | computer | on | the | blue | table
move | <gesture> that | to | the | left
make | <gesture> this | chair | green
put | <gesture> this | thing | <gesture> there
put | the | bowl | between | <gesture> this | and | <gesture> that | computer

Segmentation of multimodal input streams is thus realized in a way that open input is possible where the start and end of instructions need not be defined. Augmented by a multi-step fusion mechanism, redundancies and inconsistencies in the input stream can be handled comfortably to establish correspondence in multimodal integration.

5 Discussion and further prospects

This exploratory study was carried out in the context of research toward advanced human-computer interfaces and with the rationale to establish more natural forms of multimodal human machine communication. In detail we have desribed a method that is based on processing patterns which coordinate different input modalities in rhythmic time cycles. Based on the novel notion of timed agents realizing rhythmic mechanisms in temporal perception, we were able to

- develop a theoretical model of temporal integration of multiple input modalities
- implement the model in a prototype application and show that it is operational
- gain further insights into advantages of the 'right' rhythm by exploring the running model in experiments

In our first experiments we have used data-glove pointing and a simple word-by-word speech recognizer, allowing only very crude speech rhythm. Nevertheless, the very fact that the production as well as the technical perception of multimodal user utterings was rhythmically constrained in time was decisive for the comparably simple solution of multimodal integration. Realizing rhythmic expectancy, the tri-state segmentation model sustains equal temporal constraints beyond the current portion of signal transmitted and aids in the processing of a steady input stream. Even when our method is still far from mimicking communicative rhythm more succintly, we feel that some progress was made with respect to open input segmentation and the correspondence problem. There is reason to believe that these ideas carry further even when more obstacles have to be overcome.

The realization of a more elaborated system prototype, reaching from recognition of complex gestures over (continuous) speech-and-gesture integration to linkage with a target application of virtual prototyping, is now the goal of the SGIM project (Speech and Gesture Interfaces for Multimedia) in Bielefeld. We have taken steps to refine our basic approach to the demands of a more natural multimodal interaction. The illustrations in Figure 5, taken from the SGIM interaction scenario, convey that work is underway to realize more fluent speaking and gesturing in multimodal input. Segmentation cues are available from speech as well as gestural rhythm; we were able to make

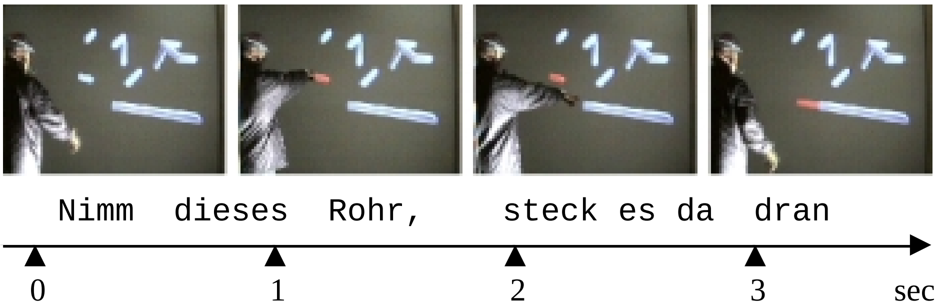

Fig. 5. Natural speech and gesture input in a virtual construction scenario ("Take this pipe, mount it there-to")

use of some of them in first instances. Work is underway to further build on these ideas [22]. We have also begun to research the issue of natural timing of generative gesture by making an articulated figure able to produce it in real time [10].

An issue for future work is how the system could be enabled to entrain to the communicative rhythm exhibited by the individual user. We have successfully completed first experiments which support the idea that adaptive oscillators [15] could provide a method to adjust the so far equal-sized integration time windows in reasonably short latency, i.e., within about 1-2s. This adjustment might allow to mimic a stretching or shrinking of segmentation time windows (like musical ritardando or accelerando, resp.) by responding to the tempo of user utterances while preserving the hierarchical temporal structure of integration intervals. Of further interest in our research will be the .5s beat that seems to mark a grid on which accented elements (e.g., stressed syllables) are likely to occur [8]. We hope to get insights as to how a low frequency segmentation mechanism, as used in the VIENA study, goes together with rhythm patterns on a finer-grained time scale.

Finally, I would like to take the chance to express my vision of an idea that I feel could be beneficial for future information society, namely, "rhythmic" systems. Whereas computer scientists and engineers have been mainly concerned with making throughput cycles of interactive applications faster, little thought was given to the question if speed is the only or most important issue. Given a choice of awaiting a system response as fast as possible, but at indeterminate time, or at *anticipatory* time, many users might prefer the second over the first option. Thus it seems worthy to conceive systems that are 'rhythmic' in the sense that they produce their response to a user's query in expectable time, so the user is not as much 'soaked' in waiting for a system output. Needless to say, such a conception would require a still more profound understanding of the communicative rhythm that is natural and comfortable to a human. It does not seem totally off hand to pursue technical solutions achieving steady throughput cycles which neither stress patience nor impose uncomfortable haste on users, by meeting rhythmic expectancy as experienced natural by humans.

Acknowledgments

This work profits greatly from contributions of the members of the AI Group at the University of Bielefeld. In particular, Section 4 builds on the dissertation by Britta Lenzmann. Her assistance, as that of my research assistants and doctoral students Timo Sowa, Martin Fröhlich, Marc Latoschik, and Ulrich Nerlich are gratefully acknowledged. The VIENA project was in part supported by the Ministry of Science and Research of the Federal State North-Rhine-Westphalia under grant no. IVA3-107 007 93.

References

1. R.A. Bolt. "Put-That-There": Voice and gesture at the graphics interface. *Computer Graphics, 14*(3): 262-270, 1980.
2. E. Bos, C. Huls, & W. Claasen. EDWARD: Full integration of language and action in a multimodal user interface. *Int. Journal Human-Computer Studies, 40:* 473-495, 1994.

3. T. Broendsted & J.P. Madsen. Analysis of speaking rate variations in stress-timed languages. *Proceedings 5th European Conference on Speech Communication and Technology (EuroSpeech),* pages 481-484, Rhodes 1997.
4. W.S. Condon, Communication: Rhythm and structure. In J. Evans & M. Clynes (Eds.): *Rhythm in Psychological, Linguistic and Musical Processes (pp. 55-77).* Springfield, Ill.: Thomas, 1986.
5. J. Coutaz, L. Nigay, & D. Salber. Multimodality from the user and systems perspectives. In *Proceedings of the ERCIM-95 Workshop on Multimedia Multimodal User Interfaces*, 1995.
6. F. Cummins & R.F. Port. Rhythmic constraints on stress timing in English. *Journal of Phonetics 26:* 145-171, 1998.
7. G. Fant. & A. Kruckenberg. On the quantal nature of speech timing. *Proc. ICSLP 1996*, pp. 2044-2047, 1996.
8. J. Kien & A. Kemp. Is speech temporally segmented? Comparison with temporal segmentation in behavior. *Brain and Language 46:* 662-682, 1994.
9. D.B. Koons, C.J. Sparrell, & K.R. Thórisson. Integrating simultaneous input from speech, gaze, and hand gestures. In M.T. Maybury (Ed.): *Intelligent Multimedia Interfaces (pp. 257-276).* AAAI Press/The MIT Press, Menlo Park, 1993.
10. S. Kopp & I. Wachsmuth. Natural timing in coverbal gesture of an articulated figure, Working notes, Workshop "Communicative Agents" at Autonomous Agents 1999, Seattle.
11. B. Lenzmann: *Benutzeradaptive und multimodale Interface-Agenten.* Dissertationen der Künstlichen Intelligenz, Bd. 184. Sankt Augustin: Infix, 1998.
12. J.G. Martin. Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review 79*(6): 487-509, 1972.
13. J.G. Martin. Rhythmic and segmental perception. *J. Acoust. Soc. Am. 65*(5): 1286-1297, 1979.
14. M.T. Maybury. Research in multimedia and multimodal parsing and generation. *Artificial Intelligence Review 9*(2-3): 103-127, 1995.
15. D. McAuley. Time as phase: A dynamical model of time perception. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, pages 607-612, Hillsdale NJ: Lawrence Erlbaum Associates, 1994.
16. E. McClave. Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research 23*(1), 45-66, 1994.
17. D. McNeill. *Hand and Mind: What Gestures Reveal About Thought.* Chicago: University of Chicago Press, 1992.
18. J.G. Neal & S.C. Shapiro. Intelligent multi-media interface technology. In J.W. Sullivan and S.W. Tyler, editors, *Intelligent User Interfaces, pages 11-43.* ACM Press, New York, 1991.
19. L. Nigay & J. Coutaz. A generic platform for addressing the multimodal challence. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI-95),* pages 98-105, Reading: Addison-Wesley, 1995.
20. E. Pöppel. A hierarchical model of temporal perception. *Trends in Cognitive Science 1*(2), 56-61, 1997.
21. G. Schöner & J.A.S. Kelso. Dynamic pattern generation in behavioral and neural systems. *Science, 239:* 1513-1520, 1988.
22. T. Sowa, M. Fröhlich, & M.E. Latoschik, Temporal symbolic integration applied to a multimodal system using gestures and speech, *this volume.*
23. R.K. Srihari. Computational models for integrating linguistic and visual information: a survey. *Artificial Intelligence Review 8:* 349-369, 1995.
24. I. Wachsmuth & Y. Cao: Interactive graphics design with situated agents. In W. Strasser & F. Wahl (eds.): *Graphics and Robotics (pp. 73-85)*, Springer, 1995.
25. M. Wooldridge & N.R. Jennings. Intelligent agents: Theory and practice. *Knowledge Engineering Review, 10*(2): 115-152, 1995.