Ersch. in Kognitionswissenschaft (2000) 8: 151-159

Kommunikative Rhythmen in Gestik und Sprache*

Ipke Wachsmuth

Technische Fakultät, Universität Bielefeld, 33594 Bielefeld
(e-mail: ipke@techfak.uni-bielefeld.de)

Communicative Rhythm in Gesture and Speech

Summary. 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 pre-semantic 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.

Zusammenfassung. Als Eckpfeiler der natürlichen Verständigung zwischen Menschen sind Gestik und Sprache in der Mensch-Maschine-Kommunikation von großem Interesse. Jedoch gibt es bislang kaum Lösungsvorschläge dafür, wie die multimodalen Äußerungen eines Systemnutzers – als zeitlich gestreute Perzepte auf getrennten Kanälen registriert – in ihrem zeitlichen Zusammenhang zu rekonstruieren sind. In diesem Beitrag wird anhand der Beobachtung, daß menschliches Kommunikationsverhalten von signifikant rhythmischer Natur ist, eine neuartige Methode zur Konzeption eines multimodalen Eingabesystems entworfen. Es basiert auf einem zeitgetakteten Multiagentensystem, mit dem eine präsemantische Integration der Sensordaten von Sprach- und Gesteneingaben in einer multimodalen Eingabedatenstruktur vorgenommen wird. Hiermit werden erste technische Arbeiten beschrieben, die rhythmische Muster für biologisch und kognitiv motivierte Mittlersysteme zwischen Mensch und Maschine ausnutzen.

---

* Dieser Beitrag ist eine leicht überarbeitete, ergänzte deutsche Fassung des unter o.g. englischem Titel erschienenen Beitrags (Wachsmuth, 1999), mit freundlicher Genehmigung von Springer.

1 Einleitung

Mit dem immer stärkeren Eintritt des Menschen in multimediale "virtuelle" Umgebungen finden Formen der nichtverbalen körperlichen Äußerung, insbesondere Gesten, als Mittel der Informationsübermittlung an maschinelle Systeme erhebliches Interesse. Untersucht werden in jüngerer Zeit auch 'koverbale' Gesten, also Gesten, die sprachliche Äußerungen mehr oder weniger spontan begleiten, z.B. wenn man auf einen Gegenstand zeigt ("dieses Rohr") oder eine Drehrichtung ("so herum") signalisiert. Es ist leicht erkennbar, daß eine derartige Eingabeform für Anwendungssysteme, wie sie heute schon im virtuellen Entwurf eingesetzt werden, erheblichen Komfortgewinn erbringen könnte.

Als eine Herausforderung stellt sich dabei die *multimodale Integration*, insbesondere die zeitliche Kopplung der beiden komplementären Modalitäten gesprochener Sprache und Gestik: Die von der Natur her multimodalen Äußerungen eines Systemnutzers werden als nebenläufige Sprach- und Gestenperzepte auf getrennten Kanälen technisch registriert und müssen für die Steuerung von Anwendungen zusammengeführt und interpretiert werden. Bei der Vorverarbeitung der in der Signalerfassung aufgenommenen Meßdaten kommt es zu spezifischen Verzögerungen, und die Zeitkonstanten dieser Prozesse sind verschieden, das heißt, die zentrale Verfügbarkeit von Informationen aus der Signalvorverarbeitung ist zeitlich gestreut. Um für die Interpretation der Meßergebnisse den inhaltlichen Zusammenhang im System herzustellen, ist also zunächst ihr zeitlicher Zusammenhang zu ermitteln. Technische Verfahren müssen den Zeitverlauf schon deshalb rekonstruieren, damit die Integration des Zeichenhaften (z.B. Zeigegeste) mit dem Signalgehalt (z.B. Zeigevektor im Moment des Zeigens) gelingen kann. Für das Problem der zeitlichen Integration multimodaler Eingaben haben sich bislang keine befriedigenden Lösungen finden lassen.

Beobachtungen in verschiedenen Forschungsbereichen zeigen nun, daß das menschliche Kommunikationsverhalten von signifikant rhythmischer[1] Natur ist, zum Beispiel in der Weise, wie gesprochene Silben und Wörter im zeitlichen Ablauf gruppiert sind (Sprechrhythmus) oder wie sie von sinnfälligen Körperbewegungen, d.h. Gesten[2] begleitet sind. In theoretischen wie praktischen Ansätzen der Nachahmung natürlicher Kommunikationsmuster in der Mensch-Maschine-Kommunikation hat der Gedanke einer rhythmischen Organisation bislang keine Rolle gespielt. Dieser Beitrag verfolgt als Kernbotschaft den Gedanken, daß

[1] *Rhythmus*: Nach Martin (1972) definieren wir "Rhythmus" hier als übergeordnetes zeitliches Muster zwischen einzelnen Elementen einer Verhaltensfolge, d.h. der Ort jeden Elements auf der Zeitachse ist relativ zu allen anderen Elementen der Folge determiniert.

[2] *Geste*: Für den Zweck dieses Artikels ist es ausreichend, "Gesten" als Körperbewegungen zu verstehen, welche Information übermitteln, die in irgendeiner Weise bedeutsam für einen Empfänger ist. Im Hinblick auf das Kernthema dieses Beitrags sei erwähnt, daß ein harmonischer Wechsel zwischen körperlicher Spannung und Entspannung – im Sport zuweilen als 'Bewegungsrhythmus' bezeichnet –, sich gerade auch bei gestischen Körperbewegungen beobachten läßt.

rhythmische Muster[3] einen nützlichen Mechanismus zur Koordination intra- und inter-individueller multimodaler Äußerungen bereitstellen. Auf der Basis eines Konzepts zeit-kontrollierter Agentensysteme wird eine operative Methode rhythmischer Verarbeitung entworfen, die durch empirische Befunde angeregt ist und die in einem Verfahren der multimodalen Äußerungsrezeption und -integration (Sprache und Handgestik) erprobt wurde.

Im folgenden Abschnitt diskutieren wir repräsentative Befunde aus empirischer Forschung, die das Phänomen und die vermutete Rolle von Rhythmen in der menschlichen Kommunikation verdeutlichen. In Abschnitt 3 wird argumentiert, daß rhythmische Organisation einen guten Ausgangspunkt zum Umgang mit einigen offenen Problemen bei multimodalen Schnittstellen darstellt. Der originäre Beitrag des Artikels liegt in der Konzeption eines agentenbasierten Verarbeitungsmodells, das einigen der empirischen Befunde Rechnung trägt und sie für technische Lösungen erschließt. Die in Abschnitt beschriebene multimodale Eingabeagentur gründet sich auf rhythmische Verarbeitungsmuster und dient als Rahmen für ein interaktives Grafiksystem, das sprachlich-gestische Eingaben verarbeitet. Eine Diskussion der Resultate und ein Ausblick auf weitere Arbeiten folgen in Abschnitt 5. Wir schließen mit einer kurzen Vision darüber, daß rhythmische Systeme ein allgemeineres Prinzip für die Mensch-Maschine-Kommunikation darstellen und möglicherweise zur Gestaltung 'angenehm' empfundener Mensch-Maschine-Systeme beitragen könnten.

2 Rhythmus in der menschlichen Kommunikation

Phänomene des Rhythmus' in der Kommunikation – hier verstanden als wechselseitige, weitgehend beabsichtigte Informationsübertragung zwischen Partnern – sind in einer Vielzahl von Publikationen beschrieben worden, die hier nur im Ansatz gewürdigt werden können. Verschiedene Befunde aus der psychologischen und phonetischen Forschung haben Hinweise auf eine rhythmische Organisation des menschlichen Kommunikationsverhaltens erbracht, und das sowohl im Hinblick auf die Produktion als auch die Rezeption von Äußerungen. Wie die rhythmisch koordinierte Bewegung der Gliedmaßen in der Lokomotion – im Abriß beschrieben in (Schöner & Kelso, 1988) – erfordert die Produktion sprachlicher und gestischer Äußerungen die Koordination einer großen Zahl disparater biologischer Komponenten. Wenn eine Person spricht, bewegen sich oft viele Teile des Körpers: Arme, Finger, der Kopf etc. zur gleichen Zeit und in präziser hierarchisch-rhythmischer Organisation mit der sprachlichen Artikulationsstruktur. Dieses bei Condon (1986) mit "Selbstsynchronität" beschriebene Phänomen beinhaltet unter anderem die Beobachtung, daß eine sprechende Person während der Äußerung die Spannungspostur der oberen Gliedmaßen

[3] *Rhythmische Muster* sind Ereignisfolgen, in denen bestimmte Elemente gegenüber anderen hervorgehoben (akzentuiert) sind; die Akzente wiederholen sich innerhalb des Musters ansatzweise regelhaft, unabhängig von Tempo (schnell, langsam) oder Tempoänderungen (Beschleunigung, Verlangsamung). Da rhythmische Muster eine zeitliche Trajektorie haben, die ohne stetiges Überwachen verfolgt werden kann, erlaubt die Perzeption anfänglicher Musterelemente die Antizipation späterer Elemente in Realzeit; cf. (Martin 1972; 1979).

für verlängerte Dauer aufrecht erhält (zur Illustration mag Abbildung 1 dienen), mitunter in einem Einsekunden-Rhythmus, wobei zuweilen ein diese Zeit halbierendes "Wippen" der Gliedmaßen festzustellen ist.

Abb. 1. Marvin Minsky; Momentbild aus einem Videointerview

Die Ausführung einer Geste läßt sich nach (McNeill, 1992); (Kendon, 1972) in mehrere Phasen unterteilen, von denen die expressive Phase (*stroke*) die wichtigste ist. Der *stroke* ist häufig durch einen abrupten Halt gekennzeichnet, der mit den gesprochenen Wörtern zeitlich in enger Beziehung steht. Sprache und Körperbewegungen zeigen dabei charakteristische Periodizitäten; z.B. finden sich in allen germanischen Sprachen, die man den sog. "stress-timed languages" zurechnet[4], bei flüssigem Sprechen Korrelationen zwischen den (durch zeitliche Dehnung) betonten Silben und einhergehenden Gesten-*strokes*. Experimente haben ergeben, daß ein betontes Wort in der Regel nicht vor dem *stroke* der koverbalen Geste geäußert wird, sondern der *stroke* tritt kurz zuvor oder spätestens mit dem betonten Wort auf (McNeill, 1992). Dieses Phänomen zeigt sich deutlicher bei Zeigegesten (*deictics*; McNeill, 1992), während betonungsunterstützende Gesten-"Schläge" (*beats*; McNeill, 1992) und Sprechbetonung nicht in einem strikt rhythmischen Sinn synchronisiert sind (McClave, 1994).

Auch in der sprachlichen Äußerung allein lassen sich rhythmische Akzentuierungen beobachten, die sich im *timing* des Sprechens äußern (Kien & Kemp, 1994); (Fant & Kruckenberg, 1996). Ebenfalls ist verschiedentlich beobachtet worden (Condon, 1986); (McClave, 1994), daß die Äußerungsrhythmik eines Sprechers – mit sehr kurzer Latenz nach

[4]*Stress-timed language*: In der allgemeinen Phonetik wird verschiedentlich angenommen, daß "stress-timed" Sprachen wie Deutsch, Englisch und Dänisch in der Tendenz eine relativ konstante Dauer von Stress-Gruppen haben, unabhängig von der tatsächlichen Zahl von Phonen oder Silben einer Gruppe. Somit läßt sich der zeitliche Abstand zwischen den hervorgehobenen Wörtern bzw. Silben z.B. in (a) "der ZUG nach KÖLN" und (b) "die ZÜge nach BerLIN" als ungefähr gleich erwarten, wenn vom selben Sprecher unter gleichen äußeren Bedingungen gesprochen; cf. (Broensted & Madsen, 1997).

Sprechbeginn – in körperlichen Reaktionen des Hörers übernommen wird; dieses Phänomen wird in den Arbeiten Condons mit "Interaktionssynchronität" bezeichnet.

Unter fixierten Randbedingungen konnten Cummins und Port (1998) rhythmische Phänomene in der Sprachproduktion englischer Sprecher nachweisen: In den 'speech cycling'-Experimenten, bei denen Versuchspersonen einen englischen Satz wiederholt zum Takt eines Metronoms sprechen mußten, fanden sie Evidenz für ein rhythmisches Einstimmen des Sprechtakts zu dem vorgegebenen, fixen Metronomtakt. Der Sprechanfang (*sound onset*) betonter Silben verlegte sich nach kurzer Dauer auf Zeitpunkte, die den durch das Metronom getakteten Satzwiederholungszyklus in ganzzahligem Verhältnis derart unterteilen, daß Akzentuierungen (*stress beats*) zu vorhersagbaren Phasen des Metronomtaktes auftreten. Diese Beobachtung erklären die Autoren – in einer Theorie gekoppelter Oszillatoren – als harmonische Einstimmung zwischen Satzrhythmus und Satzwiederholungszyklus.

Quasi-rhythmische Phänomene in freierer Sprachproduktion, nämlich beim Vorlesen von Texten (hauptsächlich Schwedisch) sind von Fant und Kruckenberg (1996) beobachtet worden. Wie dabei festgestellt wurde, läßt die durchschnittliche Länge des Intervalls zwischen aufeinander folgenden (*stress*-)betonten Silben[5] in der Größenordnung von 500 ms (Millisekunden) Vorhersagen für die Dauer der nächstauftretenden Sprechpausen zu; genauer variieren Sprechpausen in quantalen Teilungen von ca. 500 ms, wobei die präzisen Werte vom lokalen Durchschnitt der letzten 8 Zwischenbetonungsintervalle (bzw. rund 4 Sek.) abhängen. Die Messung der Durchschnittslängen von betonten Silben, unbetonten Silben und Phonemsegmenten in Größenordnungen von 250 ms, 125 ms bzw. 62.5 ms legt des weiteren nahe, daß der "Grundtakt" von 500 ms quantal unterteilt wird, also in 1/2, 1/4 und 1/8-Teilungen eines metrischen Referenzquantums, das sich in den Zwischenbetonungsintervallen äußert. Das tatsächliche Tempo und die Kohärenz des rhythmischen Musters ist sprecherspezifisch und wird überdies von der Dichte der Inhaltswörter beeinflußt, ist also nicht von völlig musikalischer Strenge. Ähnlich beobachten Broensted und Madsen (1997) Intra-Sprecher-Variabilitäten in der Sprechrate von englischen und dänischen Sprechern aufgrund des Zeitausgleichs von Betonungsgruppen in den einzelnen Äußerungen.

Was die Perzeption betrifft, hat Martin (1972, 1979) beobachtet, daß beim Menschen ein Zusammenhang zwischen dem Sprechrhythmus und der Segmentierung gesprochener Sprache beobachtet werden kann, den er als "rhythmische Erwartung" bezeichnet. Pöppel (1997) beschreibt Zeitphänomene auf zwei signifikanten Zeitskalen. Seine Beobachtungen lassen einerseits ein hochfrequentes neuro-kognitives Verarbeitungssystem annehmen, das diskrete Zeitquanten von 30 ms Dauer generiert, und auf der anderen Seite ein niederfrequentes Verarbeitungssystem, das funktionale Zustände von ~3 s Dauer im Bewußtsein etabliert. Evidenz für das hochfrequente Verarbeitungssystem ergibt sich etwa durch Untersuchungen der zeitlichen Ordnungsschwelle: Unabhängig von der betrachteten Sinnesmodalität erfordern unterscheidbare Ereignisse demnach einen Mindestabstand von 30 ms, um als aufeinander

[5]*Interstress interval* ("Zwischenbetonungsintervall"): die gemessene Zeit zwischen dem Vokalanfang einer betonten Silbe zum Vokalanfang der nächsten betonten Silbe, außer solchen, die von einer syntaktischen Grenze unterbrochen sind.

folgend wahrgenommen zu werden. Der niederfrequente Mechanismus integriert aufeinander folgende Ereignisse in bis zu 3 s langen Einheiten der bewußten Wahrnehmung und betrifft insbesondere die Verbindungen zwischen den verschiedenen Sinnesmodalitäten. Eine Basis für diese Annahme bilden Untersuchungen über die Reproduktion von Stimulus-Mustern mit unterschiedlicher Dauer; die zeitliche Integration in Intervallen von 2-3 s läßt sich nach Pöppel (1997) auch in der Bewegungskontrolle und in der zeitlichen Segmentierung gesprochener Sprache beobachten. Diese zeitliche Integration wird insofern als automatisch und präsemantisch angesehen, als die Zeitgrenze nicht davon abzuhängen scheint, was inhaltlich verarbeitet wird.

Kommunikative Rhythmen in Gestik und Sprache könnten auf Basis solcher Befunde als koordinative Strategie des menschlichen Äußerungs- und Wahrnehmungsapparats gedeutet werden, die der rhythmischen Koordination bei Lokomotion und manuellen Aufgaben gleicht (Cummins & Port, 1997). Rhythmen scheinen hier eine Art "Pulse" bereitzustellen, die hierarchisch synchronisierte Strukturen im Kommunikationsverhalten emergieren lassen und Sprecher-Hörer-Einstimmung bewirken (Condon, 1986). Erklärungsversuche, die in den o.g. Arbeiten vorgelegt werden, lassen weiter annehmen, daß Rhythmus in der Sprache vereinzelbare Prozeßeinheiten ("Zeitfenster") hervorbringt, die dem Rezipienten durch erwartbare Periodizitäten das Segmentieren des übertragenen Signals erleichtern und das Abwechseln im Dialog unterstützen (Martin, 1979); (Fant & Kruckenberg, 1996). Die Beobachtungen Pöppels (1997) legen schließlich nahe, daß einerseits sehr kleine und davon eher entkoppelte größere Prozeßeinheiten im Mehrsekundenbereich eine Erklärungsgrundlage für eine zeitlich kontrollierte Modalitätenintegration des Menschen bieten könnten.

3 Rhythmus in der Mensch-Maschine-Kommunikation

Im vorangehenden Abschnitt wurde argumentiert, daß eine rhythmische Organisation im Kommunikationsverhalten des Menschen evident scheint. Wenn dies so ist, dann sollte diese Beobachtung auch für die Mensch-Maschine-Kommunikation von Bedeutung sein. Zum Beispiel wurde bereits von Martin (1979) vorgeschlagen, daß Computermodelle der Sprachperzeption eine Komponente "rhythmischer Erwartung" beinhalten sollten, die ausgehend vom ersten Ansatz einer Äußerung (*utterance onset*) auf die zeitliche Struktur der nachfolgend übermittelten Information extrapoliert. In der Mensch-Maschine-Kommunikation steht ein solcher Versuch der Imitation biologischer Kommunikationsmuster bislang noch aus.

Auf der anderen Seite sollte aufgrund des Forschungsziels der Verwirklichung multimodaler Interfaces, die beispielsweise die Eingabemodalitäten Sprache und Gestik kombinieren, erhebliches Interesse an den kognitiven Prinzipien der Modalitätenperzeption und -integration bestehen. Die Entwicklung multimodaler Eingabesysteme erfordert neben der Verarbeitungsmöglichkeit einzelner Modalitäten Methoden für die Integration multipler Modalitäten (Coutaz, Nigay & Salber, 1995). Die Möglichkeit multimodaler Eingabe gilt als

zentral für eine natürlichere und effektivere Mensch-Maschine-Interaktion, bei der die Information einer Modalität dazu beiträgt, die Information einer anderen zu disambiguieren (Maybury, 1995). Gesprochene Sprache und Gestik sind jedoch zunächst essentiell kontinuierliche Prozesse. Um ein technisches System zu befähigen, die perzipierten Sprach- und Gesteneingaben in ihrem natürlichen Fluß zu koordinieren und zu integrieren, sind vor einer semantischen Analyse übermittelter Information folgende zwei "logistischen" Probleme zu lösen (Srihari, 1995):

(1) Das Segmentierungsproblem: Wenn ein System offene Eingaben verarbeiten soll, wie sind die Prozeßeinheiten zu determinieren, die das System in einem Zyklus verarbeitet? Wie sind aufeinander folgende Prozeßeinheiten zu verbinden?

(2) Das Korrespondenzproblem: Wenn ein System Information multipler Modalitäten integrieren soll, wie sind die Querbezüge zwischen den Modalitäten zu determinieren, das heißt genauer, welche Information einer Modalität komplementiert Information einer anderen?

Mit bisherigen Ansätzen liegen erst in geringem Umfang Lösungsvorschläge dafür vor, wie die multimodalen Äußerungen eines Benutzers, die auf getrennten Kanälen und zudem zeitlich gestreut registriert werden, in ihrem natürlichen Zusammenhang rekonstruiert werden können. Frühe Versuche, ein multimodales Eingabesystem zu realisieren, sind das PUT-THAT-THERE-System (Bolt, 1980) und CUBRICON (Neal & Shapiro, 1991). Jedoch beschränken sich diese Systeme auf eine sequentielle Analyse und darauf folgende Zusammenführung von Sprach- und Gesteneingaben; zudem erlauben sie keine gestischen Eingaben in natürlicher Bewegungsform, sondern nur als statische Zeigerichtungen. Jüngere Ansätze, zum Beispiel in (Koons, Sparrell, & Thórisson, 1993); (Bos, Huls, & Claasen, 1994); (Nigay & Coutaz, 1995), erlauben die parallele Verarbeitung von zwei oder mehr Modalitäten. Allerdings unterstützen diese Systeme noch keine offene Eingabe – das heißt, Eingaben ohne definite Festlegung des Anfangs und Endes – und ebensowenig die Auflösung von Redundanzen und Inkonsistenzen zwischen Anteilen der unterschiedlichen Modalitäten.

Die Beobachtungen des vorangehenden Abschnitts sind hier nun der Ausgangspunkt für eine in den folgenden Abschnitten beschriebene Studie, in der ein multimodales Eingabesystem mit obigen Ansprüchen auf der Basis rhythmischer Organisation realisiert wurde. Die Grundannahme dabei war, daß eine Art grundlegender "Takt" im Äußerungsverhalten des Menschen besteht und daß durch Verwertung von Segmentierungshinweisen – wie Gesten-*stroke* und Sprechtakt – der kommunikative Rhythmus systemseitig reproduziert und damit antizipiert werden könnte. Dies sollte dabei helfen, durch Einführung entsprechender "Zeitfenster" die Korrespondenzen der zeitlich gestreuten Sprach- und Gestenperzepte wieder herzustellen und dadurch die semantische Analyse multimodaler Information erleichtern.

4 Ein multimodales Interface auf Basis zeitgetakteter Agenten

An der Universität Bielefeld ist seit mehreren Jahren das Thema der Gestenerkennung für Mensch-Maschine-Schnittstellen ein Forschungsfokus (Wachsmuth & Fröhlich, 1998). Die im Einleitungsabschnitt geschilderten Beobachtungen führten uns zu dem Gedanken, die Analyse kommunikativer Rhythmen zur Verbesserung der Leistungsfähigkeit technischer Mittlersysteme zwischen Mensch und Maschine und insbesondere im Hinblick auf die multimodale Integration auszunutzen.

In einem ersten technischen Ansatz haben wir diesen Gedanken für ein Verfahren der Synchronisation gesprochener Wörter und Handzeigegesten umgesetzt. Es beruht kurz gesagt darauf, daß der multimodale Eingabestrom aus Signalen durch Sprach- und Gestenerkenner registriert und in gleichlangen Zeitfenstern, die mit dem Äußerungsansatz einer Modalität beginnen, segmentiert wird. Eingabedaten der unterschiedlichen Modalitäten, die in einem Zeitzyklus registriert werden, werden demselben Instruktionsabschnitt zugeschrieben; modale Querbezüge werden dadurch aufgelöst, daß Korrespondenzen zwischen Gestenperzepten und linguistischen Einheiten innerhalb je eines Zeitzyklus' ermittelt werden. Da dies nicht immer ausreicht, ist des weiteren eine Zeitzyklus-überspannende Integrationsmethode zu betrachten. Diese Ansätze sind in erster Linie durch die oben erwähnten Beobachtungen über die zeitliche Verarbeitung im menschlichen Wahrnehmungssystem (Pöppel, 1997) und das Postulat einer rhythmischen Erwartung bei der Sprachperzeption (Martin, 1979) motiviert.

4.1 Gegenstand und Methoden

Der Kontext unserer Arbeiten ist die Kommunikation mit virtuellen Umgebungen, das sind Computergrafik-basierte dreidimensionale Szenen, die durch Benutzerinstruktionen interaktiv verändert werden können. Die hier beschriebene Studie wurde im VIENA-Projekt durchgeführt (Wachsmuth & Cao, 1995), wo das Design einer virtuellen Büroumgebung als prototypisches Anwendungsbeispiel diente. Das VIENA-System verarbeitet Benutzerinstruktionen, um Änderungen der visualisierten Büroszene mit Hilfe eines agentenbasierten Interface-Systems auszuführen. Die Instruktionen werden mit gesprochener Sprache und eingegeben und durch Zeigegesten ergänzt, die über einen einfachen Nintendo-Datenhandschuh erfaßt werden. In dieser Studie wurde ein Dragon Dictate (Version 1.2b) Spracherkenner benutzt, der sprecherabhängige Einzelworte verarbeitet. Instruktion werden als Folge von (englischen) Wörtern gesprochen wie folgt:

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

wobei der Sprechbeginn (*sound onset*) aufeinanderfolgender Wörter etwa 600 ms auseinander liegt. Zeigegesten werden ungefähr zur Zeit des gesprochenen "this" oder "that" durch

Handschuhzeigen auf dargestellte Objekte oder Positionen eingegeben; zur Illustration siehe Abb. 2.

Abb. 2. Instruktion des VIENA-Systems durch sprachlich-gestische Eingaben

Zur Aufnahme und Verarbeitung von Eingabeinformation der verschiedenen sensorischen Kanäle benutzen wir ein Basisverarbeitungsmodell, das verteilte Funktionen durch die Wechselwirkung multipler Software-Agenten realisiert. Der einzelne Agent ist ein autonomer Berechnungsprozeß, der mit anderen solchen Agenten auf Basis einer Variante des Kontraktnetz-Protokolls (Wooldridge & Jennings, 1995) kommuniziert und kooperiert. Ein System solcher Agenten, im folgenden "Agentur" genannt, realisiert eine dezentrale Informationsverarbeitung. Der Kern der VIENA-Agentur (siehe Abb. 3) besteht aus einer Anzahl von Agenten, die an der Vermittlung von Benutzereingaben beteiligt sind, um die Szene in Farbgebung und räumlicher Anordnung zu ändern. Typischerweise wird die Funktionalität der einzelnen Agenten in einem sense-compute-act-Zyklus erreicht, betreffend die Aufnahme von Nachrichtendaten (sense), Berechnung der jeweiligen Funktion (compute) und schließlich das Senden entsprechender technischer Kommandos (act) an andere Agenten oder ein Effektorsystem wie hier das Grafiksystem.

Das Basisverarbeitungsmodell von Agentensystemen ist ereignisgetrieben, das heißt, es gibt keine zeitliche Beschränkungen dafür, wann ein solcher sense-compute-act-Zyklus abschließt. Im Kontext der Modalitätenintegration verschiedener sensorischer Kanäle sind jedoch auch *zeitliche* Verarbeitungsmuster von Bedeutung, vor allem, wenn es auf eine enge Kopplung von sprachlichen und gestischen Eingaben ankommt. Aufgrund dieser Tatsache haben wir das Basisverarbeitungsmodell von Agenten im Hinblick auf eine zeitliche Taktung erweitert. Zu diesem Zweck wurden zeitliche Puffer für sensorisch erfaßte Information eingeführt und es wurden – neben ereignisgetriebener Ablaufsteuerung – zeitgetriebene Verarbeitungsmuster in den Agenten etabliert, die auf zwei verschiedenen Ebenen eine zeitlich getaktete, "rhythmische" Abarbeitung multimodaler Eingaben unterstützen.

Im ersten Anlauf haben die nun für eine sense-buffer-compute-act-Sequenz ausgelegten Zeitzyklen der Agenten eine fixe Dauer, die sich für Experimente variieren läßt. Die

nachfolgend beschriebene Multimodale Eingabe-Agentur (MEA) besteht aus einer Anzahl dedizierter Agenten für (1) sensorische und linguistische Eingabeanalyse und (2) die Koordination und Integration multimodaler Information.

4.2 Multimodale Eingabe-Agentur (MEA)

Um die in Abschnitt 3 genannten Problemstellungen der Segmentierung offener Eingaben und Korrespondenzherstellung für sprachlich-gestische Instruktionen anzugehen, haben wir eine multimodale Eingabe-Agentur (MEA) entwickelt, wie in der rechten Hälfte von Abb. 3 gezeigt. Sie besteht aus drei modalitätsspezifischen Eingabeagenten – den *listeners* –, einem Parser für linguistische Analyse und schließlich einem Koordinator, der die zu symbolischen Tokens vorverarbeitete Sensorinformation integriert. Die drei Eingabeagenten (speech listener, type listener und gesture listener) registrieren Sensordaten von Mikrofon, Tastatur bzw. Datenhandschuh und verarbeiten sie zu Sprach- bzw. Gestenperzepten. Unterstützt durch den Parser analysiert und integriert der Koordinator die von den *listener*-Agenten erhaltene Information und generiert daraus eine interne Aufgabenbeschreibung, die an eine Mediator-Agentur übergeben wird (in der linken Hälfte von Abb. 3 gezeigt). Die Mediator-Agentur berechnet die resultierenden Änderungen der Szenenbeschreibung und veranlaßt eine entsprechende Aktualisierung der Szenenvisualisierung durch das 3D-Grafiksystem. Multimodale Instruktionen werden durch Einsprechen ins Mikrofon und Zeigen mit dem Datenhandschuh eingegeben. Tastatureingaben werden nur für sprachliche (unimodale) Instruktionen verwendet.

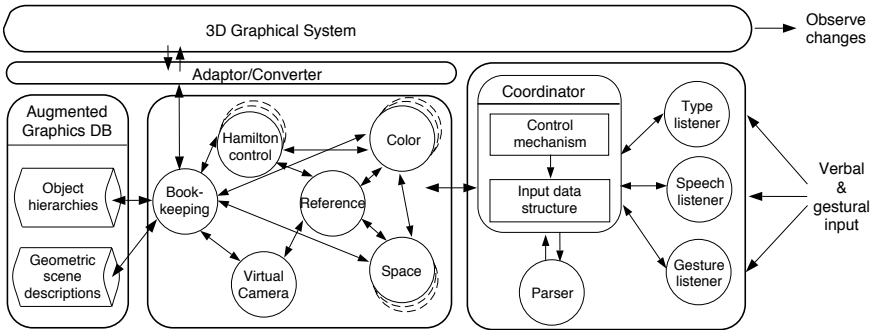

Abb. 3. VIENA-System mit multimodaler Eingabe-Agentur (rechts) und Mediator-Agentur (links)

Zur Integration der multiplen Modalitäten (Sprache und Gestik) führt die MEA eine zeit- und ereignisgetriebene Routine aus. Während die Eingabeagenten in kurzen Zyklen von 100 ms auf Eingabeereignisse “horchen”, verarbeitet der Koordinatoragent die aufgelaufene Information in fixen Zeittakten von 2 Sekunden Dauer. Diese Werte wurden in Experimenten mit dem VIENA-System ermittelt, bei denen sich zeigte, daß Zeitzyklen von 100 ms bzw. 2 Sekunden Dauer für die in der Studie verwendete Eingabetechnik (Einzelworterkenner und handschuhbasierter Gestenerkenner) die besten Resultate liefern. Der 100 ms-Zyklus wurde

auf der Basis festgelegt, daß der verwendete Datenhandschuh maximal 10 Datenpakete pro Sekunde liefert und somit eine höhere Abtastrate unnötigen Mehraufwand erbracht hätte.

Der 2-sekündige Integrations-Rhythmus beruht auf Experimenten, in denen der Gesamtdurchsatz im VIENA-System ermittelt wurde, gemessen vom Anfang einer gesprochenen Instruktion bis zur Ausgabe einer neuen Szenenvisualisierung, während die Integrationszeitzyklen in 1-Sekunden-Inkrementen variiert wurden. In diesen Experimenten wurden (unimodale) Sprachinstruktionen unterschiedlicher Länge verwendet: eine 4-Wort, eine 7-Wort und eine 10-Wort-Eingabe. Der Sprechanfang (*sound onset*) der einzelnen Wörter erfolgte durch Computersteuerung im 600 ms-Abstand, unabhängig davon, ob ein- oder mehrsilbige Wörter gesprochen wurden. Das heißt, die Eingabedauer für die 4-Wort, 7-Wort und 10-Wort-Instruktionen betrug etwas mehr als 1800, 3600 bzw. 5400 ms. Die folgenden Sprachinstruktionen wurden benutzt ("saturn" und "andromeda" sind Namen, die sich auf die beiden in Abb. 2 zu sehenden Computer beziehen):

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

Der durch die MEA realisierte Integrationsprozeß ist eine Kombination von zeit- und ereignisgetriebenen Berechnungen. In den folgenden Abschnitten wird detaillierter erläutert, wie das Segmentierungs- und das Korrespondenzproblem (vgl. Abschnitt 3) in der MEA des VIENA-Systems bewältigt wird. Ausführlich ist das Verfahren in (Lenzmann, 1998) beschrieben.

4.3 Segmentierung offener Eingaben: Das 3-Zustands-Rhythmusmodell

Der Basisansatz zur Segmentierung des multimodalen Eingabedaten"stroms" beruht auf der Idee, Eingabeereignisse von den verschiedenen Modalitäten in Zeitzyklen zu registrieren, mit denen der Koordinatoragent der Sequenz von Eingabeereignissen Zeitfenster aufprägt. Dies wird durch ein 3-Zustands-Rhythmusmodell bewerkstelligt, das in Abb. 4 veranschaulicht ist. Eingabeereignisse innerhalb eines Zeitfensters werden dem gleichen Eingabesegment zugeschrieben. Entsprechend puffert der Koordinatoragent Information, die von den Sprach- und Gesten-*listeners* kommt, und integriert sie, wenn ein Zyklus abgeschlossen ist (siehe Abschnitt 4.4).

Der erste Zeitzyklus (z1) beginnt mit dem Einsatz des Signals, wenn der Benutzer eine (verbale oder gestische) Äußerung übermittelt, woraus ein erstes Eingabeereignis resultiert (e1 zur Zeit t1). Dies versetzt den Koordinator in einen Zustand "swing", der der registrierten Ereignissequenz getaktete Zeitfenster aufprägt und solange anhält, wie Signale auf einem der *listener*-Kanäle registriert werden, wodurch eine rhythmische Erwartung modelliert wird. Der Koordinator geht in einen Abkling-Zustand ("subside") über, wenn in einem vollen Zyklus

keine weiteren Eingabeereignisse registriert werden; dabei wird die Zeitfenstertaktung aufrecht erhalten. Der Abkling-Zustand wechselt in den Wartezustand ("wait"), wenn *k* (in unserem System: 2) ereignisfreie Zyklen erkannt werden, oder kehrt zurück nach "swing", falls im Zustand "subside" ein erneutes Eingabeereignis auftritt. Der Wartezustand hält für unbestimmte Dauer an; er wechselt erneut nach "swing", wenn ein neues Eingabeereignis registriert wird.

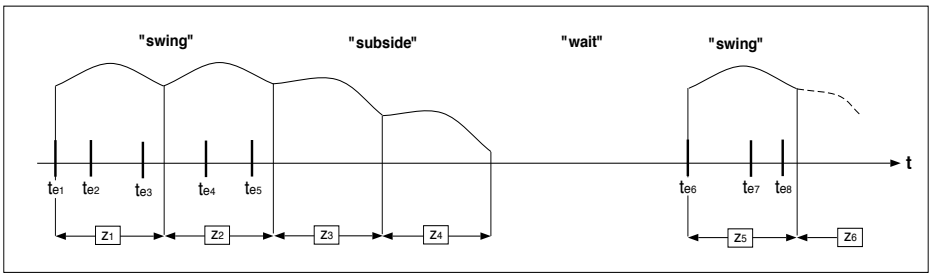

Abb. 4. 3-Zustands-Rhythmusmodell (swing–subside–wait); jeder Zyklus im Zustand "swing" und "subside" hat gleiche zeitliche Länge.

Das zeit- und ereignisgesteuerte Integrationsverfahren ist mit dem Segmentierungsprozeß verzahnt. Es besteht aus einem vierschrittigen zyklischen Prozeß, der aus den Funktionen "sense", "buffer", "compute" und "act" besteht und vom Koordinator-Agent ausgeführt wird. Während "sense" und "buffer" im Wechsel fortgesetzt werden, bis der laufende 2-Sekunden-Zeitzyklus abgeschlossen ist, werden "compute" und "act" am Ende eines jeden Zeitzyklus' ausgeführt. Die Funktion "sense" tut dabei nichts weiter, als die Nachrichten von den *listener*-Agenten als Eingabeereignisse zu registrieren. Die Funktion "buffer" extrahiert relevante Nachrichteninformation und akkumuliert sie in einer multimodalen Eingabedatenstruktur (EDS; siehe unten). Am Ende eines Zeitzyklus' interpretiert die Funktion "compute" die in der Eingabedatenstruktur des Koordinators akkumulierte multimodale Information. Anschließend bestimmt die Funktion "act" geeignete Agenten in der Mediator-Agentur und übergibt die korrespondierenden Aufgaben an sie, um eine Berechnung der Szenenänderung zu veranlassen.

4.4 Korrespondenz in der multimodalen Integration

Die Aufgabe der Interpretationsfunktion "compute" ist es, Korrespondenzen zwischen verbaler und gestischer Information in der Eingabedatenstruktur (EDS) aufzulösen und eine Gesamtaufgabenbeschreibung zu berechnen, die der multimodalen Benutzereingabe entspricht. Dabei sind zwei Fälle zu unterscheiden: (1) In der Zeitzyklus-internen Interpretation wird allein Information des aktuellen Zeitzyklus' verwertet; (2) in der Zeitzyklus-überspannenden Interpretation wird Information auch zurückliegender Zeitzyklen ausgewertet.

Nachdem der Koordinator bestimmt hat, welcher der beiden Fälle vorliegt – im einzelnen siehe bei (Lenzmann, 1998) – werden Sprach- und Gestenmodalität separat analysiert und die Resultate in einer mehrschrittigen Auswertung verarbeitet, die sowohl zeitliche wie auch linguistische Attribute verwertet, um die wahrscheinlichsten Korrespondenzen zu berechnen. Die unter Umständen mehrdeutigen Referenzen werden mit Hilfe spezieller Agenten in der Mediator-Agentur aufgelöst und die resultierende Repräsentation der Instruktion auf Vollständigkeit im Hinblick auf ihre Verarbeitbarkeit geprüft. Ein Beispiel, wo bei einer Zeitzyklus-überspannenden Interpretation eine bereits als vollständig bewertete EDS korrigiert würde, ist die Instruktion "move that chair to the left wall". Fiele nämlich "wall" in einen neuen Zeitzyklus (und würde also im Zustand "swing" registriert), so müßte die bereits vollständige EDS im Hinblick auf die Ziellokation revidiert werden. Die möglicherweise bereits an die Mediator-Agentur übergebene Aufgabenbeschreibung wird in einem solchen Fall abgefangen bzw. durch eine erneute Szenenänderung korrigiert. Bei unvollständiger EDS-Repräsentation wartet der Koordinator auf weitere Eingabeereignisse, die der Erwartung nach im nächsten Zeitzyklus anfallen müßten, oder es wird – wenn der Schwingvorgang ohne weitere Eingabeereignisse abgeklungen ist – der Benutzer mit seiner unvollständigen Eingabe in einem Editor-Fenster konfrontiert, um sie zu vervollständigen.

Die tatsächliche multimodale Integration wird in der "compute"-Phase dadurch vorgenommen, daß Korrespondenzen zwischen Gestenperzepten und sog. Gestenplätzen innerhalb eines (zweisekündigen) Integrationsintervalls hergestellt werden. *Gestenplätze* sind dabei zeitgestempelte Informationsplatzhalter, die die zur Sprachanalyse verwendete NL-Grammatik um Erwartungen anreichern, die Spracheingabe an diesen Stellen um Gestikinformation – im Hinblick auf zusätzliche Objekt- oder Richtungsspezifikationen – zu ergänzen. Die Gestenplätze bilden somit Ankerpunkte für den Aufbau von Querreferenzen zwischen Sprach- und Gestikereignissen. Dabei wird ein heuristisches Qualitätsmaß maximiert, das den Zeitversatz und den Ambiguitätsgrad der den Gestenplatz tragenden Sprachinformation berücksichtigt.

Die Bewertung potentieller Gestenplätze geschieht auf Basis der Heuristik "Je ambiger die Referenzobjekte oder Lokationen in der Sprachanweisung beschrieben werden, desto höher ist die Bewertung des entsprechenden Gestenplatzes". Wenn zwei Gestenplätze für nur ein Gestenperzept vorliegen, wird die Auflösung der Korrespondenz durch die zeitliche Nähe geleitet und durch den Vergleich von Ambiguitätswerten, die den Sprachabschnitten bei der linguistischen Analyse zugeordnet wurden; zum Beispiel ist in der Eingabe "put the chair there" die Phrase "the chair" im Hinblick auf eine Szenenreferenz weniger ambig als das deiktische "there". Ein Beispiel, bei dem die zeitliche Nähe den Ausschlag gibt, ist die Instruktion

put | this | computer | on | that | table

sofern nur ein Gestenperzept vorliegt (unter der Präsupposition, daß einer der deiktischen Ausdrücke vom vorangehenden Kontext geklärt ist). In diesem Fall ist die zeitliche Nähe von

Gestenplatz und Gestenperzept ausschlaggebend dafür, daß eines der Paare “<Geste> this” oder “<Geste> that” ein höheres Qualitätsmaß hätte. Einige weitere Beispiele für mögliche Kombinationen sprachlich-gestischer Eingaben zur Disambiguierung von Objekt- oder Lokationsinformation folgen:

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

Insgesamt kann die Segmentierung multimodaler Eingaben mit den beschriebenen Verfahren derart realisiert werden, daß insbesondere die Verarbeitung zeitlich offener Eingaben möglich ist, bei denen Anfang und Ende einer Instruktion nicht explizit mitgeteilt werden müssen; die Segmentierung erfolgt allein durch den im Koordinator evozierten Rhythmus. Ergänzt durch den mehrschrittigen Integrationsmechanismus lassen sich auch (hier nicht mehr betrachtet) Redundanzen und Inkonsistenzen in Eingaben komfortabel handhaben, um Korrespondenzen bei der multimodalen Integration zu etablieren.

5 Diskussion und Ausblick

Diese explorative Untersuchung wurde im Kontext von Forschungsarbeiten ausgeführt, deren Gesamtziel die Entwicklung natürlicherer – insbesondere multimodaler – Mensch-Maschine-Schnittstellen ist. Ein Verfahren wurde beschrieben, das verschiedene Eingabemodalitäten mit Hilfe rhythmischer Zeitzyklen koordiniert und integriert. Auf Basis der neuartigen Konzeption zeitgetakteter Agentensysteme, die rhythmische Muster zur zeitlichen Segmentierung und Korrespondenzherstellung nebenläufiger Sensor-Modalitäten realisieren, gelang es im ersten Ansatz

- ein technisiertes Modell der zeitlichen Wahrnehmung und Integration multipler Eingabe-modalitäten zu entwickeln
- das Modell in einer prototypischen Anwendung zu implementieren und damit als operabel nachzuweisen
- Einsichten über die Vorteile des “richtigen” Rhythmus’ durch Exploration des laufenden Modells zu gewinnen.

In unseren ersten Experimenten haben wir, vor allem wegen des dabei noch verwendeten einfachen Einzelwort-Spracherkenners, zunächst mit nur sehr groben Sprechrhythmen arbeiten können. Dennoch war gerade die Tatsache, daß sowohl die Produktion wie auch die

technische Rezeption multimodaler Benutzereingaben rhythmischen Mustern genügte, entscheidend für den vergleichsweise einfachen Lösungsansatz für die multimodale Integration. Eine rhythmische Erwartung realisierend, hält das 3-Zustands-Rhythmusmodell eine zeitliche Rasterung über das momentan übermittelte Signal hinaus aufrecht und unterstützt damit die schritthaltende Verarbeitung eines kontinuierlichen Eingabestroms. Selbst wenn unsere Methode noch weit von einem feiner nuancierten kommunikativem Rhythmus entfernt ist, konnten doch Erfolge im Hinblick auf die Segmentierung offener Eingaben und das Korrespondenzproblem erlangt werden. Es besteht durchaus die Aussicht, daß diese Ideen weiter führen, selbst wenn noch viele Fragen offen blieben.

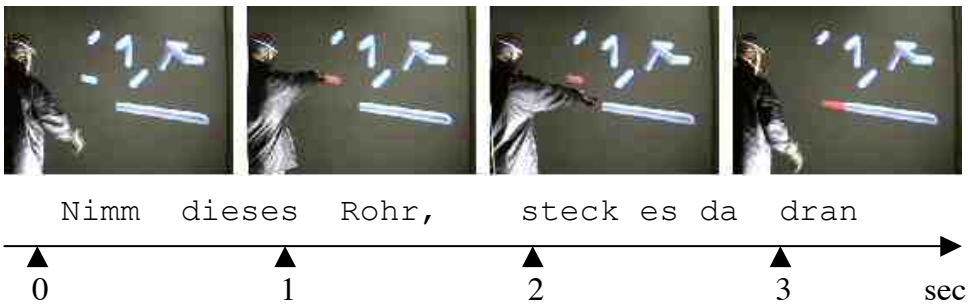

Abb. 5. Natürliche Sprach- und Gesteneingabe beim virtuellen Konstruieren

Die Realisierung eines entwickelteren Systemprototyps, der von der Erkennung komplexerer Körpergesten über die Integration auch kontinuierlicher Sprache mit Gesten bis zur Anbindung an eine Zielapplikation des virtuellen Konstruierens reicht, ist ein Kernziel des SGIM-Projekts ("Sprach- und Gesten-Interfaces für Multimedia") in Bielefeld. Unser Basisansatz wurde im Hinblick auf eine natürlichere multimodale Interaktion verfeinert, wie die Illustration in Abb. 5, aus dem SGIM-Interaktionsszenario, es andeuten soll. Bestimmte Methoden, mit denen wir derzeit experimentieren, erwarten als Parameter den Zeitpunkt eines Taktschlags, der die Grenze eines semantischen Segments andeutet; er liegt je nach Signaltyp innerhalb oder direkt am Anfang einer neuen semantischen Einheit und ist für die Ablaufsteuerung der Integration von Bedeutung. Bei den komplexeren, in schneller Folge zu verarbeitenden Gesten im SGIM-Szenario ist die Gewinnung von Segmentierungshinweisen von entscheidender Wichtigkeit. Dabei machen wir uns unter anderem zunutze, daß zwischen je zwei ausgeprägten Gesten sich die Hand kurzfristig entspannt, was über die Meßsignale eines Datenhandschuhs feststellbar ist. Ein regelbasiertes Rahmensystem, in dem die zeitliche Integration symbolischer Information aus unterschiedlichen Modalitäten realisiert wird, ist in (Sowa, Fröhlich & Latoschik, 1999) beschrieben.

Des weiteren wurden Arbeiten begonnen, mit denen das natürliche Timing bei der Gestengenerierung dadurch untersucht werden soll, daß eine artikulierte synthetische Figur in die Lage versetzt wird, sie (in Realzeit) zu produzieren; siehe Abb. 6. Dies ist im Ansatz bereits gelungen (Kopp & Wachsmuth, 1999).

Abb. 6. Sequenz von Armposturen einer artikulierten Figur: Aufwärtsbewegung des Arms in Vorbereitung des Gesten-*strokes*, der abwärts gerichtet erfolgt.

Eine mögliche Fortsetzung unserer Arbeiten betrifft die Frage, wie sich ein rhythmisch gesteuertes Eingabesystem automatisch auf den individuellen kommunikativen Rhythmus unterschiedlicher Benutzer einstimmen läßt. Dazu haben wir erste Experimente unternommen, die zeigen, daß adaptive Oszillatoren (McAuley, 1994) eine Grundlage dafür bieten könnten, die bislang fixen systemseitigen Zeitfenster in verhältnismäßig kurzer Zeit (1-2 s) automatisch an einen vorgegebenen Rhythmus anzupassen. Hiermit könnte in Reaktion auf ein individuelles Äußerungstempo eine Verlangsamung oder Beschleunigung der Zeitfenster-Rasterung erzielt werden (wie ein musikalisches Ritardando oder Accelerando), unter Beibehaltung der hierarchischen zeitlichen Struktur der Integrationsintervalle. Von weiterem Interesse in solchen Untersuchungen wird voraussichtlich ein Halbsekundentakt sein, der ein Raster zu markieren scheint, auf dem Akzentuierungen (wie *stress*-betonte Silben) mit Wahrscheinlichkeit auftreten (Kien & Kemp, 1994). Schließlich sind Erkenntnisse anzustreben, wie sich ein niederfrequentes Segmentierungsverfahren – wie in der VIENA-Studie verwendet – mit rhythmischen Mustern auf einer feiner gekörnten Zeitskala in Zusammenhang bringen läßt.

Abschließend sei hier die Chance genutzt, im Anschluß an obige Ausführungen eine Vision darzulegen, die die Annehmlichkeit zukünftiger Mensch-Maschine-Systeme verbessern könnte, nämlich "rhythmische" Systeme. Während Informatik- und Ingenieur-ansätze sich in erster Linie um die Beschleunigung der Durchsatzzeiten interaktiver Anwendungssysteme bemühen, wird kaum darüber nachgedacht, ob Geschwindigkeit der einzige oder wichtigste Gesichtspunkt sein sollte. Könnte man sich aussuchen, ob eine Systemantwort in schnellstmöglicher, jedoch unbestimmter, oder aber zu *antizipierbarer* Zeit eintrifft, mögen etliche Benutzer die zweite Option bevorzugen. Es scheint deshalb sinnvoll, Systeme zu konzipieren, die in dem Sinne "rhythmisch" sind, daß Antworten auf Benutzereingaben nach erwartbarer (dennoch akzeptabel kurzer) Zeit erfolgen, so daß Benutzer nicht mehr so stark "am Bildschirm kleben" müßten, während sie die Antwort abwarten. Es braucht nicht extra gesagt zu werden, daß ein solches Vorhaben ein noch tiefergehendes Verständnis davon voraussetzen müßte, welcher kommunikative Rhythmus

dem Menschen natürlich und angenehm ist. Technisch scheint es jedoch nicht völlig abwegig, Lösungen zu entwickeln, die zu stetigen Durchsatzzeiten führen und damit weder Ungeduld noch unangenehme Hast verursachen.

Sieht man sich die anfangs angerissenen, durchaus frappierenden Befunde zu den hier thematisierten "kommunikativen Rhythmen" näher an, so zeichnet sich das folgende Bild ab: In der Kommunikation ist zeitlich-strukturellen – und damit auch rhythmischen – Merkmalen offenbar ein ebenso großer Stellenwert einzuräumen wie der semantischen Informationsverarbeitung. In der Mensch-Maschine-Kommunikation haben solche Erkenntnisse bislang kaum Eingang gefunden. Unsere Untersuchungen lassen hoffen, daß Rhythmus der Schlüssel für einige schwierige Probleme, insbesondere in der multimodalen Kommunikation sein könnte. Auf jeden Fall eröffnet sich hiermit ein Diskursthema, das spannende Forschungsfragen für die kognitiven Disziplinen verspricht.

Dank und Hinweise. Dank gebührt den Mitgliedern meiner Arbeitsgruppe, deren Forschungsbeiträge die hier vorgestellten Arbeiten tragen, in besonderem Maße Britta Lenzmann, auf deren Dissertation der in Abschnitt 4 erläuterte Ansatz – bei weiterer Zuarbeit von Timo Sowa und Ulrich Nerlich – wesentlich beruht. Das VIENA-Projekt wurde von 1993 bis 1996 in dem Forschungsverbund "Anwendungen der Künstlichen Intelligenz in Nordrhein-Westfalen" (KI-NRW) vom Wissenschaftsministerium Nordrhein-Westfalen gefördert und mit dem Ende des Jahres 1997 abgeschlossen. Das SGIM-Projekt wurde von 1996 bis 1999 im Forschungsverbund "Multimedia NRW: Die Virtuelle Wissensfabrik" ebenfalls vom Wissenschaftsministerium Nordrhein-Westfalen gefördert.

Literatur

Bolt, R.A. (1980) "Put-That-There": Voice and gesture at the graphics interface. *Computer Graphics,* 14(3): 262-270

Bos, E., Huls, C., & Claasen, W. (1994) EDWARD: Full integration of language and action in a multimodal user interface. *Int. Journal Human-Computer Studies,* 40: 473-495

Broendsted, T. & Madsen, J.P. (1997) Analysis of speaking rate variations in stress-timed languages. *Proceedings 5th European Conference on Speech Communication and Technology (EuroSpeech)*, pp. 481-484, Rhodes

Condon, W.S. (1986) Communication: Rhythm and Structure. In J. Evans and M. Clynes (Eds.): *Rhythm in Psychological, Linguistic and Musical Processes*. Springfield, Ill.: Thomas, pp. 55-77

Coutaz, J., Nigay, L., & Salber, D. (1995) Multimodality from the user and systems perspectives. *Proceedings of the ERCIM-95 Workshop on Multimedia Multimodal User Interfaces*

Cummins, F. & Port, R.F. (1998) Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26: 145-171

Fant, G. & Kruckenberg, A. (1996) On the Quantal Nature of Speech Timing. *Proc. ICSLP-96*, pp. 2044-2047

Kendon, A. (1972) Some relationships between body motion and speech – An analysis of an example. In A.W. Siegman & B. Pope (eds) *Studies in Dyadic Communication*. New York: Pergamon Press

Kien, J. & Kemp, A. (1994) Is speech temporally segmented? Comparison with temporal segmentation in behavior. *Brain and Language* 46: 662-682

Koons, D.B., Sparrell, C.J., & Thórisson, K.R. (1993) Integrating simultaneous input from speech, gaze, and hand gestures. In M.T. Maybury (ed.) *Intelligent Multimedia Interfaces*. AAAI Press/The MIT Press, Menlo Park, pp 257-276

Kopp, S. & Wachsmuth, I. (1999) Natural timing in coverbal gesture of an articulated figure, Working notes, Workshop "Communicative Agents" at Autonomous Agents 1999, Seattle

Lenzmann, B. (1998) *Benutzeradaptive und multimodale Interface-Agenten*. Dissertationen der Künstlichen Intelligenz, Bd. 184, Sankt Augustin: Infix

Martin, J.G. (1972. Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review* 79(6): 487-509

Martin, J.G. (1979) Rhythmic and segmental perception. *J. Acoust. Soc. Am.* 65(5): 1286-1297

Maybury, M.T. (1995) Research in multimedia and multimodal parsing and generation. *Artificial Intelligence Review* 9(2-3): 103-127

McAuley, D. (1994) Time as phase: A dynamical model of time perception. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*. Hillsdale NJ: Lawrence Erlbaum Associates, pp 607-612

McClave, E. (1994) Gestural Beats: The Rhythm Hypothesis. *Journal of Psycholinguistic Research* 23(1), 45-66

McNeill, D. (1992) *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press

Neal, J.G. & Shapiro, S.C. (1991) Intelligent multi-media interface technology. In J.W. Sullivan and S.W. Tyler (eds): *Intelligent User Interfaces*. ACM Press, New York, pp 11-43

Nigay, L. & Coutaz, J. (1995) A generic platform for addressing the multimodal challence. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI-95)* Reading: Addison-Wesley, pp. 98-105

Pöppel, E. (1997) A hierarchical model of temporal perception. *Trends in Cognitive Science* 1(2), 56-61

Schöner, G. & Kelso, J.A.S. (1988) Dynamic pattern generation in behavioral and neural systems. *Science,* 239: 1513-1520

Sowa, T., Fröhlich, M. & Latoschik, M. (1999) Temporal symbolic integration applied to a multimodal system using gestures and speech. In A. Braffort et al. (eds). *Toward a Gesture-based Communication in Human-Computer Interaction (Proceedings Internat. Gesture Workshop, Gif-sur-Yvette, France, March 1999)*. Berlin: Springer (LNAI 1739). 291-302

Srihari, R.K. (1995) Computational models for integrating linguistic and visual information: a survey. *Artificial Intelligence Review* 8: 349-369

Wachsmuth, I. (1999) Communicative rhythm in gesture and speech. In A. Braffort et al. (eds). *Toward a Gesture-based Communication in Human-Computer Interaction (Proceedings Internat Gesture Workshop, Gif-sur-Yvette, France, March 1999)*. Berlin: Springer (LNAI 1739). 277-290

Wachsmuth, I. & Cao, Y. (1995) Interactive graphics design with situated agents. In W. Strasser & F. Wahl (eds). *Graphics and Robotics*. Berlin: Springer, pp. 73-85

Wachsmuth, I. & Fröhlich, M. (1998) *Gesture and Sign Language in Human-Computer Interaction (Proceedings International Gesture Workshop, Bielefeld, Germany, September 17-19, 1997)*. Berlin: Springer (LNAI 1371)

Wooldridge, M. & Jennings, N.R. (1995) Intelligent agents: Theory and practice. *Knowledge Engineering Review,* 10(2): 115-152