Rhythmus in der Mensch-Maschine-Kommunikation

Ipke Wachsmuth
Technische Fakultät
Universität Bielefeld

Kurzfassung

Mit dem verstärkten Eintritt des Menschen in multimediale "virtuelle" Umgebungen finden Formen der nichtverbalen körperlichen Äußerung, insbesondere Gesten, als Mittel der Informationsübermittlung an maschinelle Systeme starkes Interesse. Untersucht werden in jüngerer Zeit auch 'koverbale' Gesten, also Gesten, die sprachliche Äußerungen mehr oder weniger spontan begleiten. Als Herausforderung stellt sich dabei die multimodale Integration, insbesondere die zeitliche Kopplung der beiden komplementären Modalitäten gesprochener Sprache und Gestik. 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. Dieser Beitrag motiviert anhand kognitionswissenschaftlicher Befunde den Stellenwert 'kommunikativer Rhythmen' in Äußerungsformen des Menschen und gibt Einblick in erste technische Arbeiten, die rhythmische Muster für die Entwicklung kognitiv motivierter Mittlersysteme zwischen Mensch und Maschine ausnutzen.

1 Einleitung

Gestik und Sprache sind die Eckpfeiler in der natürlichen Verständigung zwischen Menschen. Nicht von ungefähr wird daher auch in der Forschung über Mensch-Maschine-Kommunikation Gesten- und Sprachschnittstellen erhebliche Aufmerksamkeit gewidmet. Von zunehmenden Anwendungsinteresse sind dabei besonders 'koverbale' Gesten, also Gesten, die gesprochene 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 Multimedia-Systeme, wie sie heute schon im virtuellen Entwurf eingesetzt werden, erheblichen Komfortgewinn bedeutete. Jedoch sind dafür noch Probleme zu bewältigen, für die es bislang kaum Lösungsvorschläge gibt. Die von der Natur her multimodalen Äußerungen eines Systemnutzers müssen als nebenläufige Sprach- und Gestenperzepte auf getrennten Kanälen technisch registriert und für

die Steuerung von Anwendungen integriert und interpretiert werden. Bei der Vorverarbeitung der in der Signalerfassung aufgenommenen Meßdaten kommt es zu Verzögerungen, die den Zeitpunkt, an dem die Meßergebnisse vorliegen, und den Meßzeitpunkt voneinander abweichen lassen. Zudem sind die Zeitkonstanten dieser Prozesse verschieden, das bedeutet, die zentrale Verfügbarkeit von Informationen aus der Signalvorverarbeitung ist zeitlich gestreut. Für die Interpretation der so erhaltenen Meßergebnisse ist es aber wichtig, den inhaltlichen Zusammenhang wieder herzustellen. Dafür ist zunächst einmal ihr zeitlicher Zusammenhang zu ermitteln. Technische Verfahren müssen das Zeitverhalten schon deshalb rekonstruieren, damit die Integration des Zeichenhaften (z.B. Zeigegeste) mit dem Signalgehalt (z.B. Zeigevektor im Moment des Zeigens) gelingen kann.

Anhaltspunkte ergeben sich durch Forschungsbefunde aus den Humandisziplinen, die zeigen, daß das menschliche Kommunikationsverhalten durch signifikant 'rhythmische' Muster geprägt ist (CONDON, 1986). Wenn eine Person spricht, bewegen sich oft viele Teile des Körpers (Arme, Finger, der Kopf etc.) zur gleichen Zeit und in enger zeitlicher Kopplung (Selbstsynchronität). Die Ausführung einer Geste läßt sich 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 dabei gesprochenen Wörtern zeitlich in enger Beziehung steht. Sprache und Körperbewegungen zeigen dabei charakteristische Periodizitäten; z.B. finden sich in allen germanischen Sprachen – bei flüssigem Sprechen – Korrelationen zwischen (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; der Stroke tritt kurz zuvor oder spätestens mit dem betonten Wort auf (MCNEILL, 1992). 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 kann zuweilen beobachtet werden (CONDON, 1986); (MCCLAVE, 1994), daß die Äußerungsrhythmik eines Sprechers vom Hörer in körperlichen Reaktionen übernommen wird (Interaktionssynchronität).

Ähnlich wie die rhythmische Koordination der Gliedmaßen bei der Lokomotion (SCHÖNER & KELSO, 1988) werden kommunikativen Rhythmen als koordinative Strategie des menschlichen Äußerungs- und Wahrnehmungsapparats gedeutet. Rhythmen scheinen eine Art Pulse oder "Taktschläge" bereitzustellen, die die Synchronisation von Körperbewegung und gesprochener Sprache bewerkstelligen. Durch erwartbare Periodizitäten bringen sie quasi vereinzelbare Prozeßeinheiten hervor, die dem Rezipienten das Segmentieren des übertragenen Signals erleichtert (MARTIN, 1979). Weitere Hinweise geben Untersuchungen zu temporalen Kontrollmechanismen für die Wahrnehmung und Bewußtseinsbildung im menschlichen Gehirn (PÖPPEL, 1997). Danach werden aufeinanderfolgende Wahrnehmungszustände zu größeren, bis drei Sekunden langen, Einheiten gebündelt. Dies gilt insbesondere für

Verbindungen zwischen den verschiedenen Sinnesmodalitäten. Ebenfalls gibt es Hinweise, daß die Abfolge von absichtlichen Bewegungen, zu denen auch die meisten Gesten zählen, zeitlich strukturiert und bis zu einem Zeitraum von 2 bis 3 Sekunden vorausgeplant ist.

2 Ansätze für die multimodale Integration mit Rhythmen

In der Arbeitsgruppe Wissensbasierte Systeme der Universität Bielefeld werden seit mehreren Jahren Möglichkeiten der Gestenerkennung für Mensch-Maschine-Schnittstellen und der multimodalen Integration von Gestik und Sprache erforscht; siehe z.B. (WACHSMUTH, 1999A). Die im Einleitungsabschnitt geschilderten Beobachtungen führten uns zu dem Gedanken, die Analyse kommunikativer Rhythmen zur Verbesserung der Leistungsfähigkeit technischer Mittlersysteme zwischen Mensch und Maschine auszunutzen. Gesprochene Sprache und Gestik sind zunächst einmal essentiell kontinuierliche Prozesse. Vor einer semantischen Analyse übermittelter Information sind also die folgenden logistischen Probleme zu lösen (SRIHARI, 1995):

- *Das Segmentierungsproblem:* Wie sind die Prozeßeinheiten zu determinieren, die das System in einem Zyklus verarbeiten soll?
- *Das Korrespondenzproblem:* Wie sind die Querbezüge zwischen den Modalitäten Gestik und Sprache zu determinieren?

Unter der Voraussetzung, daß ein grundlegender Takt im sprachlich-gestischen Äußerungsverhalten des Menschen besteht, könnte durch Verwertung von Segmentierungshinweisen, wie Gesten-Stroke und Sprechtakt, der kommunikative Rhythmus systemseitig reproduziert und u.U. antizipiert werden. Dies würde dabei helfen, die Korrespondenzen der zeitlich gestreuten Sprach- und Gestenperzepte wieder herzustellen und dadurch die semantische Analyse multimodaler Information erleichtern.

In einem ersten technischen Ansatz im Projekt VIENA (Virtuelle Entwurfsumgebung und Agenten) wurde das Prinzip der kommunikativen Rhythmen zur Bestimmung von zusammengehörigen Worten und Zeigegesten ausgenutzt (LENZMANN, 1998); siehe auch (WACHSMUTH, 1999B). Zum Beispiel kann das System bei Instruktionen wie "*make - <Geste> this - chair - green*" die über Spracherkenner und Datenhandschuh registrierten Eingaben zusammenführen, um entsprechende Änderungen in einer computergrafisch visualisierten Szene zu berechnen. Die Korrespondenz von Zeigegesten und bei der Sprachanalyse determinierten 'Gestenplätzen' (das sind Informationsplatzhalter, die Erwartungen bezüglich ergänzender Objekt- und Richtungsspezifikationen formalisieren) ist u.a. durch den zeitlichen Abstand geleitet: Je kleiner der Abstand, desto besser passen beide Teile zusammen. Der Integrationsinstanz des Systems (dem sogenannten Koordinator) ist ein 2-Sekunden-Rhythmus aufgeprägt

(illustriert in Abb. 2-1); er wird durch das erste Ansprechen des Mikrophons angestoßen und sorgt dafür, daß die in einem solchen "Takt" registrierten Ereignisse a priori als zusammengehörig betrachtet werden. Der angestoßene Rhythmus ("swing") klingt aus ("subside"), wenn keine Eingabeereignisse mehr registriert werden, und geht bis zu einer Folgeinstruktion in einen Wartezustand ("wait") über. Der mit Agententechniken realisierte Ansatz unterstützt außerdem offene Eingaben, indem nach Ablauf eines Taktes automatisch eine Integration der Ereignisse vorgenommen wird und so keine explizite Markierung des Eingabeendes erforderlich ist; die Segmentierung erfolgt allein durch den im Koordinator evozierten Rhythmus.

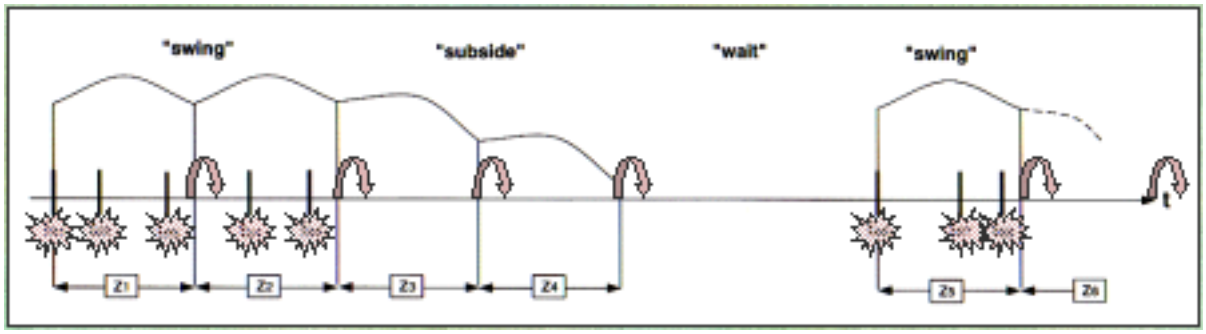

Abbildung 2-1 : *Rhythmusprinzip für die multimodale Integration (Projekt VIENA)*

Die Erstellung eines umfassenden Systemprototyps, der von der Erkennung komplexerer Gesten über die Sprach-Gestik-Integration bis zur Anbindung an eine Zielapplikation des virtuellen Konstruierens reicht, ist das Kernziel des SGIM-Projekts (Sprach- und Gesten-Interfaces für Multimedia). In den hier entwickelten verfeinerten Ansätzen ist die Methode `rhythmInfo` für die Verarbeitung rhythmusartiger Information verantwortlich. Sie erwartet als Parameter den Zeitpunkt eines Taktschlags, der die Grenze eines semantischen Segments andeutet. In der Methode werden alle Signalperzepte, deren Assertationszeit älter als die Taktzeit ist, aus dem Arbeitsgedächtnis des Systems entfernt; es wird damit zyklisch von nicht mehr relevanter Information befreit. Rhythmusartige Information wird mit der Message-Klasse `RhythmMessage` im System kommuniziert. Als einzige Komponente enthält die Klasse einen Zeitstempel, mit dem der genaue Zeitpunkt des Taktschlags mitgeteilt wird. Der Zeitpunkt, der damit festgelegt ist, liegt je nach Signaltyp innerhalb oder direkt am Anfang einer neuen semantischen Einheit. Als Reaktion auf den Erhalt einer `RhythmMessage` wird in der jetzigen Implementierung die Methode `rhythmInfo` des Integrators aufgerufen. `RhythmMessage` ist die Oberklasse aller Nachrichtenklassen, mit denen Rhythmusinformation versendet wird. Die Teilklasse `GestureSegmentationCue` ist dabei für alle Gestensegmentierungshinweise zuständig. Als Komponente enthält sie eine Markierung, die angibt, ob der Zeitstempel die Grenze zweier semantischer Einheiten oder die expressive Phase einer Geste bezeichnet. Als Beispiel eines Gestensegmentierungshinweises arbeiten wir

zum Beispiel mit einer Klasse `HandTensionMessage`, die sich zunutze macht, daß zwischen je zwei ausgeprägten Gesten sich die Hand kurzfristig entspannt, was über die Meßsignale eines Datenhandschuhs feststellbar ist. Eine Interaktionssequenz mit dem SGIM-System ist in Abb. 2-2 veranschaulicht. Ein regelbasiertes Rahmensystem, in dem die zeitliche Integration symbolischer Information aus unterschiedlichen Modalitäten realisiert wird, ist in (SOWA, FRÖHLICH & LATOSCHIK, 1999) beschrieben.

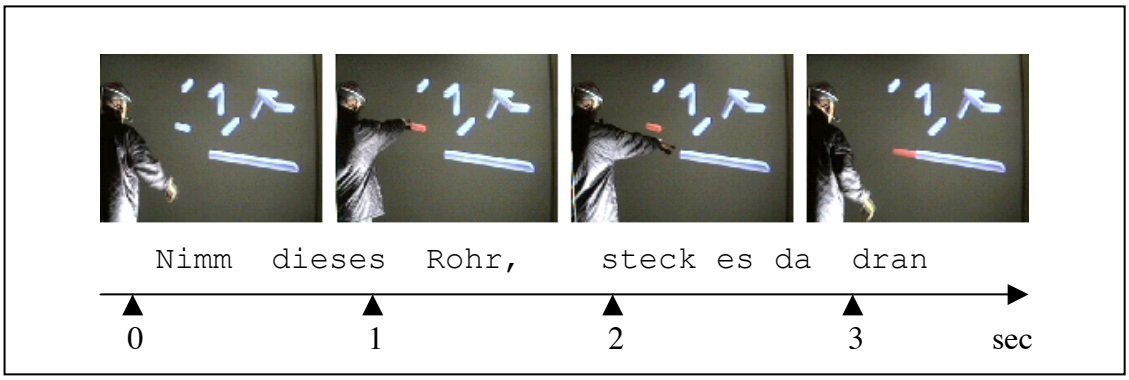

Abbildung 2-2 : *Sprachlich-gestische Eingabe an der Interaktionswand (Projekt SGIM)*

Dank und Hinweise

Dank gebührt den Mitgliedern meiner Arbeitsgruppe, auf deren Forschungsbeiträgen die hier vorgestellten Arbeiten wesentlich beruhen. 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 wird seit 1996 im Forschungsverbund "Multimedia NRW: Die Virtuelle Wissensfabrik" vom Wissenschaftsministerium Nordrhein-Westfalen gefördert.

Literatur

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

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

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

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

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

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.

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, presented at GW'99: 3rd Internat. Gesture Workshop, 17-19 March, 1999, Gif-sur-Yvette.

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

WACHSMUTH, I. (1999A). Mensch-Maschine-Kommunikation mit Gestik und Sprache, ersch. in: W.-D. Miethling & J. Perl (Hrsg.), *Sport und Informatik VI (S. 167-178)*. Köln: Sport und Buch Strauss.

WACHSMUTH, I. (1999B). Communicative rhythm in gesture and speech, presented at GW'99: 3rd Internat. Gesture Workshop, 17-19 March, 1999, Gif-sur-Yvette, to appear.