15 Verkörperte Kommunikation mit kognitiven virtuellen Agenten

Nadine Pfeiffer-Leßmann, Stefan Kopp und Ipke Wachsmuth

15.1 Einleitung

Miteinander zu kommunizieren ist eine der komplexesten und wichtigsten Fähigkeiten des Menschen, deren Erforschung in der Künstlichen Intelligenz immer schon eine zentrale Rolle gespielt hat. Abgeleitet von dem lateinischen *communicari* („teilen, mitteilen") wird Kommunikation als Informationsaustausch zwischen Personen verstanden, oder meist etwas spezieller als wechselseitige, weitgehend beabsichtigte Informationsübertragung zwischen mindestens zwei Partnern. Die verbreitete, auf die Kommunikationstheorie von Shannon und Weaver zurückgehende Vorstellung, dass bei diesem Informationsaustausch ein „Sender" Informationen verschlüsselt, die ein „Empfänger" aufnimmt und entschlüsselt, hat dabei auch klassische Kommunikationsmodelle in der KI geprägt [77] und ist bis heute in der Mensch-Maschine-Interaktion stark verbreitet. Ein Vielzahl von empirischen Befunden in den kognitiven Disziplinen hat seitdem aber deutlich gemacht, dass dieses Modell für ein Verstehen und Modellieren der menschlichen Kommunikation – speziell der Kommunikation von Angesicht zu Angesicht – nicht ausreichend ist. Insbesondere die körperliche Einbettung kognitiver Agenten hat für ihre Kommunikation eine große Bedeutung, sowohl bezüglich der Signale, über die sie interagieren, als auch bezüglich der kognitiven Prozesse, die dabei zugrunde liegen. Beides wird durch den Begriff „verkörperte Kommunikation" (embodied communication [75]) in den Mittelpunkt gestellt und im vorliegenden Kapitel für die Mensch-Maschine-Kommunikation und die Erforschung virtueller verkörperter Agenten thematisiert.

Eine Vielzahl jüngerer Arbeiten in der Künstlichen Intelligenz und Robotik greift die Fragestellung der verkörperten Kommunikation in technischem Zusammenhang auf. Neben dem Aspekt der technischen Machbarkeit sind diese Arbeiten mit der Erwartung verbunden, durch die Entwicklung und den Test konkreter Modelle neue Erkenntnisse über das Funktionieren menschlicher Kommunikation zu gewinnen. Wie funktioniert beispielsweise das zeitliche und inhaltliche Zusammenspiel von Sprechen und Gestikulieren? Welche Rolle spielen Emotionen in der Kommunikation? Wie wird das Abwechseln im Dialog gesteuert? Wie koordinieren sich Interaktionspartner miteinander? Der Ansatz, technische Kommunikationspartner zu entwickeln, verlangt immer auch eine Behandlung der Frage, welche internen Mechanismen die entsprechenden Fähigkeit des Agenten erbringen können. Speziell widmen sich die Arbeitsrichtungen der „embodied conversational agents" [17] und der „virtual humans" [32] dem Ziel, Schlüsselaspekte der Verkörperung in der Kommunikation zu modellieren und auf der anderen Seite anthropomorphe Assistenzsysteme mit zahlreichen Anwendungsperspektiven zu erarbeiten. Oberflächlich betrachtet streben solche Systeme eine multimodale Interaktion an, welche ver-

bale und non-verbale Signale der menschlichen Konversation umfasst; d.h. Sprache, Mimik, Gesten oder Körperhaltung werden von beiden Interaktionspartnern eingesetzt, um zum Dialog beizutragen. Intern liegen komplexe Verhaltensmodelle und Agentenarchitekturen zugrunde, um z.B. mit der großen Varietät von kommunikativen Verhaltensweisen, ihrer feinen zeitlichen Synchronisierung und der Dynamik der Wechselwirkung zwischen Kommunikationspartnern umzugehen.

Im Folgenden werden Grundlagen der verkörperten Kommunikation samt wichtiger Fachbegriffe eingeführt. Danach werden technische Lösungen für die Modellierung einzelner Aspekte natürlicher verkörperter Kommunikation behandelt. Anschließend wird anhand des virtuellen Agenten Max ein Beispiel eines verkörperten Kommunikationspartners vorgestellt und so ein Einblick in ein komplexes Gesamtsystem vermittelt. Zum Schluss werden weiterführende Forschungsfragen auf dem Weg hin zu natürlich und verkörpert kommunizierenden Agenten thematisiert.

15.2 Grundlagen

Der Begriff „verkörperte Kommunikation“ (embodied communication) [75] betont die zentrale Rolle der Körperlichkeit kognitiver Agenten in ihrer Kommunikation. Eine genaue Analyse der Verwendung von körperlichen Signalen macht deutlich, wie vielschichtig und nebenläufig menschliche Kommunikation ist: Während des Sprechens produzieren wir einen Strom von körperlichen Signalen. Gleichzeitig nehmen wir dieselben Signale des Adressaten wahr und achten darauf, ob er aufmerksam ist und an den entsprechenden Stellen wie erwartet reagiert (z.B. einen Blick erwidert, mit einem Nicken Verstehen signalisiert oder durch ein Heben der Hand das Rederecht anfordert). Darüber hinaus passen sich Kommunikationspartner in vielfältiger Hinsicht aneinander an. Sie tendieren z.B. dazu dieselben Wörter zu verwenden, ihre Sprechgeschwindigkeit oder Tonhöhe anzugleichen oder ihre Körperhaltung zu imitieren [40]. Auch in ihren Gesten sind sie sensitiv für die Gesten anderer [10].

Diese Beobachtungen haben Konsequenzen für die Forschung in der Künstlichen Intelligenz, der humanoiden Robotik und der Mensch-Maschine-Kommunikation. Sie verlangen ein vollständigeres und korrekteres Verständnis von Interaktion, das – über symbolische Kommunikation hinaus – systematisch den Einbezug physischer Ressourcen verlangt. Als Herausforderung für die Konstruktion verkörperter kommunizierender Agenten impliziert dies den Entwurf operationaler Modelle, die spezifizieren, wie mentale Prozesse und Verkörperung in der Kommunikation und innerhalb eines kognitiven Agenten zusammenwirken.

15.2.1 Begriffe der Kommunikation

Dialog, Sprechakte und andere Handlungen

McTear definiert einen Dialog als eine gemeinsame, kooperative Handlung zwischen zwei oder mehreren Beteiligten [51]. In der Sprachwissenschaft besteht eine lange Tradition darin, Sprechen als eine Form des Handelns zu sehen [5, 68]. Austin betonte, dass verschiedene „Kräfte“ in einer Äußerung wirken: Der *lokutionäre* Akt der reinen Äußerung von Wörtern, der *perlokutionäre* Akt des gewünschten Effekts auf den Adressaten oder die Welt und den *illokutionären* Akt, eine bestimmte Aktion zur Herstellung des perlokutionären Ziels. Basierend auf dieser

Definition prägt Searle den Begriff des „Sprechakts“ (speech act), welcher sowohl einen propositionalen Inhalt als auch eine mentale Absicht der Verwendung umfasst. Andere Forscher haben für verwandte Konzepte Begriffe wie *communicative act* [1], *conversational move* [15] oder *dialogue move* [20] geprägt. [58] definieren einen kommunikativen Akt als die kleinste Einheit der Kommunikation, welche aus einem Signal und einer Bedeutung besteht. Zur Bedeutung gehört dabei ein propositionaler Inhalt sowie ein *Performativ*, welches die Aktion repräsentiert, die der Akt ausführt. Die Modelle unterscheiden hierbei zwischen verschieden Aktionen. Poggi und Kollegen etwa nehmen allgemein drei unterschiedliche Typen von Performativen an: Präsentation von Informationen (*inform*), Fragen (*query*) und Aufforderungen (*request*). Andere Taxonomien unterteilen in weit feinere Klassen, z.B. das DAMSL-Anotationssystem [21] oder die DIT++-Taxonomie von kommunikativen Funktionen [13]. Letztere umfasst mehr als 100 verschiedenen Funktionen, grob unterteilt in *General Purpose Communicative Functions* wie Informationstransfer, Fragen oder Direktiven und *Dialog Control Functions*, die sich auf Aktionen zur Koordinierung der Kommunikation mit Aushandung des Rederechts (*Turn-Taking*) oder der Etablierung von Informationen (*Grounding*) beziehen. Darüber hinaus erfüllen kommunikative Handlungen oft auch *sozio-emotionalen Funktionen*, die die Beeinflussung der Wahrnehmung und Einschätzung der eigenen Person durch andere betreffen [52].

Multimodalität

Laut Argyle laufen mehr als 65 Prozent des Austausches in einem Gespräch über nichtsprachliche Kanäle wie Gesten, Körperhaltung, Mimik oder Sprachmelodie ab [3]. Allgemein lassen sich *verbale* von *non-verbalen* Modalitäten unterscheiden, die letzteren können als vokalisch und nicht vokalisch differenziert werden. Sprache ist gut dafür geeignet, symbolische Informationen zu übertragen. Der Ausdruck in Gesicht und Stimme kann z.B. emotionale Zustände übermitteln. Gesten eignen sich besonders für die Vermittlung von visuellen räumlichen Informationen wie die Form eines Objekts oder den direkten Verweis auf ein externes Objekt oder eine Position (wie in Abb. 15.1 dargestellt). Durch ikonische Gestik kann ein Vorstellungsbild in verkörperter Form extern repräsentiert und übermittelt werden [50]. Ein solches gestisches Zeichen trägt Bedeutung durch bildliche Ähnlichkeit zu einzelnen wichtigen Aspekten des vorgestellten Bezugsobjekt in sich. Zudem können non-verbale Signale wie Körperhaltung, Mimik, Gestik und Blicke Intentionen vermitteln und bei der Steuerung des Dialogs (z.B. beim Aushandeln des Rederechts) mitwirken. *Paraverbale* Anteile einer Äußerung umfassen die Art und Weise des Sprechens (Stimmlage, Intonation, Lautstärke sowie Sprechtempo, Sprechpausen und Sprachmelodie) und können das Gesagte – oft im Zusammenspiel mit einem Gesichtsausdruck – modifizieren.

Turn-Taking

Turn-Taking stellt einen interaktiven Basismechanismus dar, um die Ablaufplanung des Rederechts während einer Konversation zu koordinieren [23], und sichert den kooperativen Ablauf einer Interaktion unabhängig davon, ob die inhaltlichen Ziele der Kommunikationspartner kooperativ sind [29]. Die Konversationsanalyse geht von einem kontextfreien, regelbasierten Mechanismus aus [66], mit dem lokal an übergaberelevanten Stelle der Sprecherwechsel erfolgt. Dagegen hat Duncan [23] im Rahmen eines signalbasierten Ansatzes dokumentiert, dass eine Aushandlung der Sprecherrolle durch interaktive Signale geleitet wird. Spätere Dialogtheorien [18, 30] greifen beides auf. Goodwin [31] hält einen signalbasierten Ansatz allein für nicht ausreichend, um die Phänomene des Turn-Takings zu erklären. Er sieht eine Aushandlung des Turns als zeitgebundenen kooperativen Prozess zwischen Sprecher und Hörer. Nach

Abbildung 15.1: *Multimodale Kommunikation mit Sprache und Gestik.*

Clark's Dialogtheorie [18] stellt Turn-Taking dazu eine Ebene des Dialogmanagements dar, auf der dezidierte Aktionen wirken, z.B. Ergreifen des Turns (take-turn), Anfordern des Turns (request-turn), Aufgabe des Turns (release-turn), Halten des Turns (hold-turn), Zuweisen des Turns (assign-turn).

Initiative

Während die Mechanismen des Turn-Takings auf der interaktionalen Ebene operieren, lässt sich auf der inhaltlichen Ebene ebenfalls ein Abwechseln der Einflüsse von Interaktionspartnern ausmachen. Man spricht von „Initiative" und unterscheidet mindestens zwei verschiedene Ebenen [19]: Auf linguistischer Ebene betrifft Initiative die Kontrolle im Dialogablauf, und Initiativwechsel treten oft zeitgleich mit der Übergabe oder Übernahme des Turns auf. Auf Problemlösungsebene bezieht sich Initiative auf das Vorantreiben des Dialogs zu einem Ziel. Je nach Rolle, Ziele und Kompetenz kann die Initiative nur bei einem Dialogpartner verbleiben oder dynamisch zwischen den Dialogpartnern wechseln. Ersteres (*system-initative*) ist häufig der Fall in technischen Sprachdialogsystemen. Letzteres (*mixed-initiative*) ist eine Voraussetzung für gleichberechtigtes Problemlösen und Kooperieren, mit verschachtelten Beiträgen der Interaktanten und asynchronen Initiativwechseln [34].

Feedback und Engagement

Während das klassische Sender-Empfänger-Modell von klar getrennten Rollen zu jedem Zeitpunkt ausgeht, funktioniert natürlicher Dialog anders. Dort produzieren die Kommunikationsteilnehmer Feedback- und Engagement-Signale, auch wenn sie gerade nicht im Besitz des Turns sind und sprechen [41]. Man spricht daher von „Backchannel-Signalen" [79], die eine wichtige Rolle im Prozess des Groundings (s.u.) spielen. Eine andere Rolle betrifft das Signalisieren von „Engagement", durch das ein oder mehrere Teilnehmer während einer gemeinsamen Interaktion eine Verbindung eingehen und aufrechterhalten [69]. Engagement-Signale verdeutlichen, dass ein Agent in eine Interaktion stark involviert ist, steuern eine Kooperation und fördern die Effizienz und Reibungslosigkeit einer Interaktion. Engagement kann sowohl durch verbale als auch durch non-verbale Signale gefördert werden. Im Bereich der non-verbalen Signale kann insbesondere der Aufmerksamkeitsfokus, repräsentiert durch die Blickfokussierung der Interaktionspartner, als zentrales Engagement-Signal eingesetzt werden [4, 23, 39].

Interpersonale Koordination

In der verkörperten Kommunikation kommt der gegenseitigen Anpassung von Interaktionspartnern großes Gewicht zu, da es oft mit automatischen Prozessen der Kopplung zwischen Wahrnehmung und Motorik in Verbindung gebracht wird [40]. In der Tat ist Verhaltensangleichung – vor allem non-verbale – in der Interaktion ein umfassend belegtes Phänomen, dem eine positive soziale Funktion zugeschrieben wird [44]. Einige Dialogtheorien – allen voran der Alignment-Ansatz nach [57] – nehmen an, dass gegenseitige Anpassung universell und automatisch ist und auch auf linguistischen Ebenen wirkt (von phonologischer über lexikalische bis hin zu syntaktischer und semantischer Angleichung). Ihr wird dementsprechend eine kommunikative Funktion zugeschrieben. Dies ist im Gegensatz zu klassischen Dialogtheorien zu sehen, die eine zentrale Rolle des kooperativen Etablierens („Grounding“) von Inhalten betonen. Demnach werden spezifische explizite Mechanismen, von Feedback bis hin zur relevanten nächsten Äußerung, eingesetzt und bei der Verarbeitung sowie die Produktion von Äußerungen berücksichtigt [12, 18].

15.3 Technische Ansätze

Die Ermöglichung verkörperter Kommunikation zwischen Menschen und Maschinen ist ein Ziel für die Künstliche Intelligenz ebenso wie für die Mensch-Maschine-Interaktion. Im Folgenden werden speziell die technischen Ansätze besprochen, die zur Realisierung kognitiver virtueller Agenten mit multimodalen Kommunikationsfähigkeiten entwickelt werden.

15.3.1 Dialogmanagement

Traditionelle Ansätze der Dialogmodellierung lassen sich in zustandsbasierte (*state-based*) und *Frame*-basierte Ansätze einteilen [51]. Die zustandsbasierten Ansätze bauen auf explizit repräsentierten Dialogzuständen auf und definieren durch eine Dialoggrammatik, wie durch einen Dialogakt (dialogue act) ein Zustand in einen anderen Zustand überführt werden kann. Sie werden in Szenarien eingesetzt, in welchen mögliche Dialogverläufe im Vorhinein bestimmt werden können. Frame-basierte Ansätze machen lediglich Annahmen über die auszutauschenden Inhalte (nicht deren Ordnung) und zeichnen sich daher durch eine größere Flexibilität aus. Modelle, nach denen der Dialogverlauf dynamisch durch zielgerichtete Entscheidungsprozesse der Konversationspartner entsteht, werden oft als agentenbasierte Ansätze konzipiert. Diese Ansätze betrachten Dialog als kooperativen Prozess zwischen zwei oder mehreren rationalen Agenten, denen allgemeine mentale Einstellungen wie Beliefs, Ziele, Pläne, sowie spezielle Commitments und Obligationen zugeschrieben werden [73]. Mit planbasierten Techniken werden Äußerungen in Form von Sprechakten wie intentionale Akte behandelt, die formal als Operatoren mit Rollen, Vorbedingungen, Constraints und Effekten dargestellt werden. Dafür muss der sich fortlaufend entwickelnde Dialogzustand in Form dieser Wissensstrukturen repräsentiert und verwaltet werden. Viele Dialogsysteme bauen dazu zum Beispiel auf dem *Information State Approach* auf [74]. Hier wird davon ausgegangen, dass jeder Agent Zustände verwaltet, die Informationen über den bisherigen wie evtl. zukünftigen Diskurs auf verschiedenen Ebenen repräsentieren und in den Entscheidungsprozess einbeziehen. Der Zustand des Dialogs aus Sicht des Agenten ist hierdurch klar definiert. Jede Aktion eines Dialogteilnehmers wird als Dialogakt verstanden, für den diese Wissensstrukturen anhand von festgelegten Regeln kontextabhängig aktualisiert werden.

Ein weiterer wichtiger Aspekt des Dialogmanagements ist die Regulierung des Konversationsablaufs, insbesondere durch das Turn-Taking. Klassische Modelle gehen von einer begrenzten Anzahl von konversationalen Zuständen aus (z.B. *NotPresent, Present, UserTurn, HoldTurn, AgentTurn*), zusammen mit Übergängen zwischen diesen, die durch kontextabhängige interaktionale Funktionen (z.B. *give-turn, want-turn, take-turn*) ausgelöst werden können [17]. Diese Ansätze wurden durch Übergangszustände (*Gap, Overlap*) sowie durch interaktionale Funktionen (*yield-turn*, *hold-turn*) erweitert [49]. Neuere Arbeiten setzen einen nicht-deterministischen endlichen Automaten mit einer Kostenmatrix und entscheidungstheoretischen Prinzipien ein, um Turn-Taking Verhalten zu modellieren [62]. Nach [72] reicht es nicht aus, kontextsensitive Regeln aufzustellen, nach denen das Rederecht genommen oder abgegeben wird. Vielmehr müssen auch antizipative Mechanismen in der Agentenarchitektur verankert werden. Aufbauend auf Arbeiten von Goodwin und Duncan stellt Thorisson das Ymir-Turn-Taking-Modell (YTTM) auf, ein Schichtenmodell, welches den kompletten Wahrnehmen-Handeln-Prozess modelliert. Weitere Arbeiten erfassen die Regeln des Turn-Takings durch maschinelles Lernen [38] oder modellieren Turn-Taking in einer Einheit mit Feedback und den zugrunde liegenden Verstehens- und Antwortprozessen [41].

15.3.2 Multimodale Verhaltensverarbeitung

Die Verarbeitung der kommunikativen Signale des Gegenübers folgt zumeist einer Erkennungspipeline von der Verarbeitung der Sensordaten der verschiedenen Modalitäten über Vorverarbeitungsschritte, das Extrahieren von Merkmalen, Prozesse der Segmentierung und Mustererkennung, bis hin zur multimodalen Integration und Interpretation (siehe Abbildung 15.2). Dieser Ansatz wurde zum Beispiel für Sprache und Gesten umgesetzt, aber auch Blickverhalten, Mimik sowie die Körperausrichtung können so verarbeitet werden. Sprache wird durch Mikrophone aufgenommen und weiter durch Spracherkenner verarbeitet. Für Gesten müssen die Position und Ausrichtung der Hände erkannt werden sowie die Winkel zwischen den einzelnen Fingern und damit die Postur der Hand. Die Erkennung wird z.B. mit Datenhandschuhen oder durch videobasierte Ansätze oder mit Infrarotkameras vorgenommen. Als Vorverarbeitungsschritte werden Filtermechanismen eingesetzt, zum einen um Rauschen zu reduzieren und zum anderen um ein Hand- bzw. Körpermodell auf die Daten zu projizieren, aus welchem dann Winkel und Positionsvektoren abgelesen werden können. Dazu gibt es kinematische Modelle, welche die Skelettstruktur der Gelenke und Gliedmaßen beschreiben, sowie dynamische Modelle, welche die Bewegung als Ergebnis von Kräften und Drehmomenten beschreiben. Der Vorteil der Körpermodelle besteht darin, dass fehlende Sensorinformationen ergänzt, Trajektorien vorhergesehen und Adaptationsprozesse eingesetzt werden können, um sich auf die Charakteristika des individuellen Benutzers einzustellen, sowie unmögliche Bewegungen und Posturen ausschließen zu können (Details siehe [70]).

Um aus den Sensordaten eine Geste identifizieren zu können, muss das Signal zunächst segmentiert werden, um die expressive Phase der Geste aus dem Datenstrom zu filtern. Zusammen mit den Daten des Spracherkenners können die Daten dann in einem n-dimensionalen Vektor aller zu einem bestimmten Zeitpunkt relevanten Sensordaten repräsentiert werden.

Sowohl Gesten- als auch Spracherkennung werden oft als Mustererkennungsprozesse aufgefasst. Dabei können Template-basierte wie auch merkmalsbasierte Ansätze eingesetzt werden [8]. Während bei Template-basierten Ansätzen die Klassifikationsklassen vorgegeben sind, wird bei den merkmalsbasierten Ansätzen von einer Menge maßgebender Merkmale ausge-

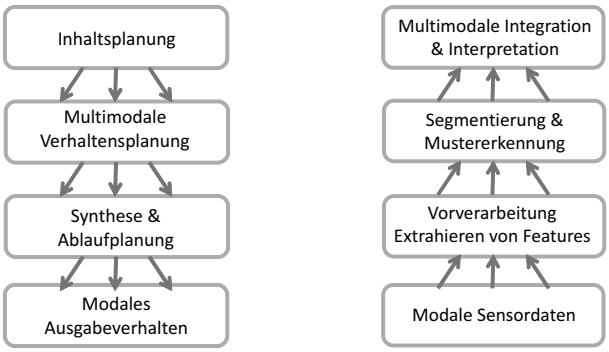

Abbildung 15.2: *Allgemeiner Aufbau einer Pipeline für die Verarbeitung und Generierung kommunikativen Verhaltens.*

gangen. [26] nutzt einen merkmalsbasierten Ansatz, der auf das Beschreibungssystem HamNoSys [60] aufsetzt. Dieses Framework wurde später für multimodale Eingaben erweitert und besonders für Echtzeitinteraktionen in der Virtuellen Realität angepasst [46].

Die Ansätze lassen sich anhand des Zeitpunkts, zu dem die Informationen der verschiedenen Modalitäten integriert werden, unterscheiden. Wenn die Daten sehr eng miteinander verbunden sind, können sie auf der Merkmalsebene früh zusammengeführt werden („early fusion"; beispielsweise Sprachsignale und Lippenbewegungen). Auf semantischer Ebene können Informationen später integriert werden („late fusion"; dabei werden die semantischen Informationen der einzelnen Modalitätskanäle zu einer übergeordneten Bedeutung zusammengeführt. Für die Bedeutungsspezifikation werden Attribut-Wert-Matrizen (AVMs) oder Frames genutzt [76]. Eine Alternative zum Frame-basierten Ansatz besteht darin, endliche Automaten zu verwenden [37]; dieser Ansatz wurde in [47] um zeitliche Aspekte erweitert, sodass multimodale Grammatiken mit erweiterten Transitionsnetzwerken (augmented transition networks) eingesetzt werden können.

15.3.3 Multimodale Verhaltensgenerierung

Analog zu der Pipeline der Verhaltensverarbeitung wird kommunikatives Verhalten in der Regel in mehreren Schritten generiert (siehe Abbildung 15.2). Die Struktur baut auf Arbeiten zur Sprach- bzw. Textgenerierung von [64] auf. Der erste Schritt besteht aus der Auswahl der zur übermittelnden Information (*Inhaltsplanung*) sowie der Anknüpfung zum vorherigen Diskurs (*Diskursplanung*). Dabei wird ein baumartiger Inhaltsplan erstellt, in welchem die Diskurseinheiten (zumeist Propositionen) die Blattknoten darstellen. Die Position der Blattknoten im Baum wird oft durch rhetorische Relationen (wie z.B. Reihung, Elaboration oder Kontrastierung) zwischen den einzelnen Diskurseinheiten festgelegt. Eine besondere Herausforderung ist die Planung von multimodalen Äußerungen mit aufeinander abgestimmten Teilen. Ähnlich wie bei der multimodalen Integration wird für die Verteilung des Inhalts und die Koordination der unterschiedlichen Modalitäten oft eine Grammatik eingesetzt. Dabei wird von einer Sprechakt-basierten Repräsentation ausgegangen, welche auch Diskursfunktionen und emotionale Funktionen beinhalten kann.

Der nächste Schritt betrifft die Wahl geeigneter Verhaltensformen (z.B. Wörter, Sätze oder Gesten) zur Realisierung multimodaler Äußerungen (*Mikroplanung* und *Realisierung* nach [64]). Dazu müssen in der Sprachgenerierung die Aufgaben der Lexikalisierung, Aggregation und der Generierung von referierenden Ausdrücken gelöst werden. Die Planung non-verbaler Aktionen geschieht meist Lexikon-basiert, wobei kontext-sensitiv aus vorgegebenen Schablonen ausgewählt wird [16]. Weiterführende Ansätze ermöglichen die Erzeugung neuer Aktionen durch Auswahl und Komposition verschiedener Anteile. So lassen sich z.B. ikonische Gesten aus Bestandteilen wie Handform, Trajektorie oder Händigkeit assemblieren, die wiederum aus Repräsentationen von räumlichen Informationen sowie zusätzlichen Kontextinformation (Sprechakttyp, Diskursstatus, vorherige Geste) abgeleitet werden können [9]. Für die Verhaltensumsetzung werden zumeist datenbasierte Modelle wie *Unit Selection* für Text-zu-Sprache-Systeme oder *Motion-Capturing* für non-verbales Verhalten eingesetzt [71]. Flexiblere Systeme setzen auch hier modellbasierte Techniken ein, um größere Teile des Verhaltens autonom und bedarfsgerecht zu generieren [43].

15.3.4 Emotionen

Emotionen übernehmen in der Kommunikation eine wichtige Signalfunktion. Sie geben Aufschluss über den inneren Stimmungszustand eines Agenten, können aber auch gezielt in der sozialen Interaktion eingesetzt werden um z.B. Empathie auszudrücken. Zudem beeinflussen die Emotionen auch die kognitiven Prozesse. So sind sie z.B. wichtig für Bewertung von Wahrnehmungen und Handlungserfolg. In Kognitionstheorien gelten sie sogar als eine Grundvoraussetzung für organisiertes Handeln und die Fähigkeit, zwischen verschiedenen Handlungsoptionen zu entscheiden [22]. Man charakterisiert Emotionen im allgemeinen durch eine positive oder negative Wertigkeit (Valenz) sowie ihre Stärke. Computationale Ansätze der Emotionsmodellierung lassen sich in zwei Grundrichtungen einteilen. In kommunikationsgetriebenen Ansätzen werden etwa die Gesichtsausdrücke eines animierten Agenten direkt nach der beabsichtigten Wirkung auf den Benutzer ausgewählt, z.B. [58]. Simulationsbasierte Ansätze bauen auf *Appraisal*-Theorien wie OCC [54] auf. Die OCC-Theorie sieht Emotionen als durch die bewertende Reaktion (*valenced reaction*) auf Ereignisse und Objekte im Licht von Zielen, Standards und Attituden entstehend, nimmt aber an, dass Emotionen vollständig kognitiv und kategorial sind. Andere Arbeiten setzen eine Simulation der Emotionsdynamik in kontinuierlichen Emotionsräumen ein, die durch Impulse von außen und innen beeinflusst wird [7]. Aktuelle Bestandsaufnahmen zu computationalen Ansätzen der Emotionsmodellierung finden sich bei [63] und – speziell das europäische Forschungsnetzwerk HUMAINE (Human-Machine Interaction Network on Emotion) betreffend – bei [67].

15.3.5 Kognitive Architektur

Um ein kohärentes Gesamtverhalten zu erzielen, ist eine Architektur erforderlich, die Wahrnehmungs- und Verhaltensprozesse mit kognitiven Prozessen vereinigt. Zu diesen zählen Aufmerksamkeit, Erinnerungen, Urteile, Vorstellungen, Antizipation, Planen, Entscheiden, Problemlösen und das Mitteilen von Ideen [80]. Für die verkörperte Kommunikation kommen Sprachverstehen und -generierung, Multimodalität, Dialogplanung sowie Prozesse der sozialen Kognition hinzu. Zu den bekanntesten kognitiven Architekturen zählen die auf Produktionssystemen basierenden wie z.B. SOAR [65]. Eine solche Architektur beruht auf Annahmen über die repräsentationalen Charakteristika und die Struktur von Gedächtnissen und den Prozessen, die

auf diesen Gedächtnissen operieren. In Analogie zu Wahrnehmungsvorgängen verläuft der Auswertungsprozess parallel, während Entscheidungs- und Problemlösevorgänge seriell erfolgen. Der Datenspeicher wird als Pendant des menschlichen Arbeitsgedächtnisses gesehen, während der Produktionenspeicher als Form eines prozeduralen Langzeitgedächtnisses dient. Produktionen selbst bestehen aus Regeln der Form, „wenn Bedingung, dann Aktion". Hybridsysteme wie Act-R [2] vereinen symbolisches Wissen mit subsymbolischen Prozessen, die auf dem Wissen operieren.

Des Weiteren werden oft Belief-Desire-Intention (BDI) Architekturen eingesetzt. Diese haben ihren Ursprung im Bereich der mentalen Modelle [11]. [28] und [24] betonen, dass ein System, welches in dynamischen, unsicheren und unvorhersehbaren Welten agieren soll, eine Repräsentation von *Beliefs*, *Desires* und *Intentions* enthalten muss. Die *Beliefs* repräsentieren deklarativ die Annahmen des Agenten. Die *Desires* bestehen aus den diversen übergeordneten Zielen des Agenten. Wird ein Ziel mit einer konkreten Handlungsabsicht verknüpft, d.h. mit einem Plan zur Erreichung des Ziels, so spricht man von einer *Intention*. Für den BDI-Ansatz existieren logische Semantiken [61,78] sowie eine Vielzahl von Implementierungen wie PRS [27], JAM [36], Jack [35] und Jadex [59] mit einem breiten Spektrum an erfolgreichen Anwendungen in verschiedenen Domänen, u.a. auch der konversationalen Agenten [73].

15.4 Verkörperte Kommunikation mit einem virtuellen Agent am Beispiel Max

Nachdem verschiedene Methoden zur Realisierung eines verkörperten kommunizierenden Agenten im Überblick vorgestellt wurden, soll nun eine konkrete Realisierung am Beispiel des virtuellen humanoiden Agenten „Max" verdeutlicht werden. Max wird seit 1999 an der Universität Bielefeld entwickelt und in diversen Interaktionsszenarien eingesetzt. Im Folgenden wird die Realisierung einer Kooperation zwischen Agent und menschlichem Partner im Rahmen einer gemeinsamen Montageaufgabe beschrieben, die auch vielfältige Prozesse der verkörperten Kommunikation einschließt.

15.4.1 Szenario

Der hier beschriebene Agent Max [42] kooperiert mit einem menschlichen Interaktionspartner, um den Zusammenbau eines Flugzeugmodells aus Teilen eines Konstruktionsbaukastens zu lösen. Dabei verhandeln die beiden Dialogpartner über einzelne Konstruktionsschritte und übernehmen Rollen als Instrukteur und Konstrukteur. Beide Rollen können sowohl vom Menschen als auch von Max eingenommen werden. Mensch und Max stehen sich in einer Virtual-Reality-Installation an einem virtuellen Tisch mit verschiedenen virtuellen Bausteinen gegenüber, wie in Abb. 15.3 gezeigt. Der Mensch sieht Max und die gesamte Szene mit einer Stereobrille dreidimensional und hört die synthetische Stimme von Max räumlich aus versteckten Lautsprechern. Max „sieht" den Menschen, dessen Blickrichtung, Hand- und Armbewegungen mit Hilfe eines Infrarot-Trackingsystems und kabelloser Datenhandschuhe, und er „hört" über Funkmikrophon dessen Sprache, die er mit einem Spracherkenner verarbeitet. Sowohl Mensch als auch Max können durch natürlichsprachliche Instruktionen und Gesten den Zusammenbau einzelner Modellbauteile veranlassen, der in physikgerechter Simulation ausgeführt wird.

Abbildung 15.3: *Face-to-Face-Dialogsituation im Kooperationsszenario mit Max*

Abbildung 15.4 zeigt den Ablauf einer Beispielsinteraktion, in welcher Max zusammen mit seinem Partner einen Propeller aus Modellbauteilen baut. Die Bilder illustrieren das gezeigte Verhalten sowie die Bauteile in verschiedenen Momentaufnahmen der Interaktion (nach [48, S. 298]).

15.4.2 Kognitive Architektur: Beispiel

Die Grundlage für die Realisierung des künstlichen Kommunikationspartners Max ist die in Bielefeld entwickelte CASEC-Architektur (**C**ognitive **A**rchitecture for **S**ituated **E**mbodied **C**ommunicators). Diese verbindet symbolverarbeitende und verhaltensbasierte Ansätze in einer hybriden Systemarchitektur. Für die Modellierung des mentalen Zustands des Agenten baut CASEC auf einer BDI-Implementierung [36] auf, welche die Grundlage für die Modellierung des Zusammenspiels von Zielen, Wissen und Ereignissen bietet. Um reaktive Verarbeitungsprozesse zu berücksichtigen, wurden weitere Ebenen einer kognitiven Architektur konzipiert und implementiert: An Stelle der starren Menge der Beliefs tritt ein dynamisches Arbeitsgedächtnis mit Aktivierungsfunktionen, welches in die BDI-Struktur eingebettet ist und auf Modelle der Psychologie zurückgreift [6,53]. Subsymbolische Aktivierungsfunktionen werden auf die Ziel- und Intentionsmodellierung übertragen und ihre Einflüsse auf den kognitiven Entscheidungsprozess berücksichtigt. CASEC baut auf den folgenden Prinzipien auf:

- **Nebenläufige** Realisierung von Wahrnehmungs-, Schlussfolgerungs- und Handlungskomponente (*Perceive, Reason, Act*) durch **modulare** Struktur
- Ausführung reaktiver und deliberativer Verhaltensweisen mittels **Behaviors** mit individuellen Prioritätswerten, **kontextsensitive Zielsteuerung** durch den Einsatz sowohl von **zielbasierten** als auch von **reaktiven Plänen**
- Ständiger Fluss von Informationen und Aktivierungen innerhalb einer zentralen Verarbeitungsschleife durch ein **dynamisches Arbeitsgedächtnis**
- Modellierung **expliziter mentaler Zustände** durch ein **erweitertes BDI**-Modul (Obligationen, gemeinsame Pläne, *Constraint*-basierte Repräsentation)

1 Interlokutor: *Lass uns einen Propeller bauen.*
2 Max: *Okay.*
3 Max: *Lass uns eine Schraube...*
4 Interlokutor: *Welche Schraube?*
5 Max: *... in das mittlere Loch einer Leiste stecken.*
6 Max: *Eine Schlitzschraube.*
7 Interlokutor veranlasst Aktion [Schraube in erstes Loch einer Dreierleiste]
8 Max: *Nein. Das war das falsche Loch.*
9 Max [*macht die Handlung rückgängig*]
10 Max: *Nochmal. Eine Schraube muss mit dem mittleren Loch einer Leiste verbunden werden.*
11 Interlokutor veranlasst die Aktion [gelbe Schraube in die Mitte einer Dreierleiste]
12 Max: *Gut. Jetzt stecke diese <Zeigegeste> Schraube in eine andere Dreierleiste.*
13 Interlokutor: *Mach du das.*
14 Max: *Okay, dann mach ich das.*
15 Max [*verbindet die Teile*]
16 Max: *Und jetzt muss diese Leiste quer <Geste> zu der Leiste gedreht werden.*
17 [Eine lange Pause tritt auf. Der Interlokutor reagiert nicht.]
18 Max: *Okay, dann übernehme ich das.*
19 Max [dreht die Bauteile]
20 Max: *Super. Wir haben soeben einen Propeller richtig zusammen gebaut.*

Abbildung 15.4: *Beispiel-Interaktion aus dem kooperativen Montage-Szenario*

Um die Prinzipien umzusetzen, besitzt die Architektur folgenden strukturellen Aufbau (Abbildung 15.5). Die Sektoren *Perceive* und *Act* repräsentieren die Physis des Agenten, durch die er mit der Umwelt wechselwirkt. Sie bietet multimodale Ausdrucksmöglichkeiten (Sprache, Gestik, Mimik) und verankert seine sensorische Wahrnehmung und ausführende Aktorik in der Umwelt. Der direkte Informationsfluss zwischen den beiden Sektoren ermöglicht schnelles, reaktives Verhalten. Ein *Reason*-Sektor ermöglicht deliberatives, planbasiertes Verhalten durch ein enges Zusammenspiel des klassischen *Perceive-Reason-Act*-Zyklus. Die Wahrnehmungs-, Schlussfolgerungs- und Handlungskomponenten sind nebenläufig realisiert, so dass die deliberative und reaktive Verarbeitung parallel erfolgen. Ein *Mediator* im *Act*-Sektor übernimmt die Aufgabe, reaktive und deliberative Verhaltensstränge zusammenzuführen und zieht in Betracht, welche Modalitäten gerade frei bzw. im Einsatz sind. Im Konfliktfall besteht die Entscheidungsgrundlage aus Prioritätswerten, die die Dringlichkeit und Angemessenheit eines Verhaltens in einer vorliegenden Situation repräsentieren. Reaktive Verhaltensweisen realisieren bspw. unmittelbare Systemreaktionen mit hoher Priorität, aber auch untergeordnete *Secondary Behaviors* wie Lidschlag und Atmen.

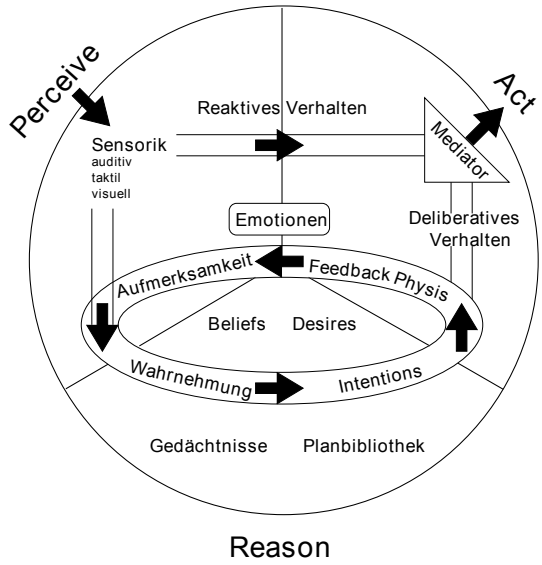

Abbildung 15.5: *Struktur der CASEC-Architektur.*

Die interne kognitive Verarbeitung des Agenten geschieht in einer „deliberativen Schleife". Wahrnehmung besteht damit nicht aus der reinen Erfassung sensorischer Daten, sondern aus einer aktiven, situationsabhängigen Filterung und Verarbeitung perzipierter Sensoreindrücke. Kognitive Prozesse werden so nicht als abgelöster interner Vorgang betrachtet, sondern an die Physis gekoppelt, mit einer stärkeren Betonung der prozeduralen Komponente.

Der Kern des deliberativen Moduls folgt dem BDI-Ansatz. Als verhaltensauslösender Antrieb dienen explizit repräsentierte Ziele (*Desires*), die sowohl intern als auch von außen aufgeworfen werden können. Durch eine *Constraint*-basierte Repräsentation können abstrakte Ziele, Teilentscheidungen sowie eingegangene *Commitments* dargestellt werden. Die Intentionsbildung in der kognitiven Schleife wird aufgrund der vorliegenden Beliefs, der aktuellen Ziele sowie alternativer Handlungsmöglichkeiten bestimmt. Für die erforderliche Modellierung des Commitment-Verhaltens im Rahmen einer Kooperation ist die BDI-Struktur so erweitert, dass Obligationen und gemeinsame Pläne unterstützt werden.

Handlungsoptionen liegen in Form von Plänen vor, die durch Vorbedingungen, Kontextbedingungen, erreichbare Konsequenzen und eine Prioritätsfunktion beschrieben werden. Die Planbibliothek besteht aus simplen Plankonstrukten, die einfache Aktionen direkt in entsprechende *Behaviors* umsetzen können. Bei Bedarf werden darüber hinaus eigenständige Planer angestoßen, die einen komplexeren Plan ausarbeiten. Planselektion findet dabei einerseits auf der Ebene der kognitiven Intentionsbildung statt, andererseits prioritätsbasiert durch den Mediator auf der Ebene der direkten Aktionsausführung. Sowohl die aktiv ausgeführten Intentionen als auch die aktuell anliegenden und möglicherweise konkurrierenden Behaviors werden bei den zurückfließenden Feedbackinformationen berücksichtigt. Die so entstehende Schleife verdeutlicht eine der Kernideen der Architektur, nämlich dass ein ständiger Strom von Informationen zwischen den Sektoren umläuft, der sowohl aktuelle Sensor- und Aktorinformationen als auch interne Zustände einbezieht.

15.4.3 Interaktionssteuerung

Während die reaktiven Verhaltensweisen für die Generierung von Engagementsignalen und Feedbackmechanismen Einsatz finden, spielt die deliberative Komponente eine zentrale Rolle für die Interaktionssteuerung. Der gemeinsame Montageprozess zwischen Max und Mensch vereint Kommunizieren und Handeln zu einer zielgerichteten Interaktion, die auf Basis der Sprechakttheorie bzw. der Theorie kommunikativer Akte als Abfolge intentionaler Akte verstanden werden kann. Das Interaktionsmanagement bei Max geschieht daher planbasiert; kommunikative wie manipulative Akte werden als Aktion-Plan-Operatoren dargestellt und in der Handlungsplanung der BDI-Komponente verarbeitet.

Genauer verfügt der Agent über prozedurales Wissen in Form aufgabenorientierter Pläne, welche die als nächstes auszuführenden Konstruktionsschritte repräsentieren und hierarchisch geschachtelt werden können. So besitzt er beispielsweise einen Skelettplan, der den generellen Ablauf der kooperativen Montage eines bestimmten Aggregats vorgibt und die Koordination der daraus abgeleiteten Handlungsoptionen übernimmt. Zunächst initiiert dieser Plan einen Montageplaner, der die auszuführenden Konstruktionsschritte aus dem Langzeit-Montagewissen des Agenten ableitet, welche dann als Unterziele sukzessive aufgeworfen und situationsabhängig in der Interaktion abgearbeitet werden.

Die augenblickliche Situation sowie den Verlauf der Interaktion mit dem menschlichen Partner erfasst Max deklarativ in Form von Beliefs. Aus Sicht des Agenten stellt sich dieser Verlauf als Folge von Situationen dar, in denen jeweils eine bestimmte (kommunikative oder manipulative) Aktion zu neuen Objektkonstellationen in der Szene sowie zu veränderten mentalen Zuständen der Interaktionspartner führt. Der Agent muss diese Entwicklungen erfassen, um beispielsweise Fragen wie „Welche Schraube hast du gemeint?“ verstehen und beantworten zu können. Ein Gedächtnis erfasst dazu den Dialogverlauf in den folgenden Aspekten:

- die *Ziele*, die in der Interaktion verfolgt werden; oftmals gibt es mehrere Ziele, die sich überlagern und in bestimmten Beziehungen zueinander stehen
- den propositionalen *Inhalt*, der im Diskurs ausgetauscht wird
- die *Obligationen* (Verpflichtungen), die für die Kommunikationspartner im Verlauf des Dialogs entstehen, z.B. auf eine Frage zu antworten
- die Übernahme der *Initiative*, d.h. die Kontrollausübung auf den Dialog durch das Aufwerfen neuer Ziele
- den Besitz der Sprecherrolle (*Turn*)

Diese Aspekte und ihr Einfluss auf das Interaktionsverhalten des Agenten werden in der kognitiven Architektur wie folgt modelliert: Spezielle Pläne übernehmen die Aushandlung des Rederechts, Ziele werden mit dem jeweiligen Initiator versehen und fließen in die Intentionsbildung ein, und der Inhalt eines Dialogaktes oder einer Montageaktion führt zu veränderten Beliefs. Obligationen stellen eine besondere Unterart von Zielen dar, die als *Social Attitudes* zusätzlich in den BDI-Ansatz eingeführt wurden [73]. Sie betreffen Verpflichtungen, die ein kooperativer Dialogpartner entsprechend gewisser Normen eingeht, die aber mit seinen eigentlichen Zielen in Widerspruch stehen können.

Um dem Verlauf des Montageprozesses folgen zu können und die dabei neu entstandenen Aggregate auch sprachlich referenzierbar zu machen, verfügt der Agent über detailliertes Langzeitwissen über die Modellbauteile sowie die Zieldomäne (z.B. Flugzeug-Bauen). Zum einen

benötigt Max dieses Wissen für seinen Montageplaner, der die Konstruktionsschritte bestimmt, die für den Zusammenbau des vom Menschen nachgefragten Montageziels auszuführen sind. Zum anderen nimmt Max fortwährend das Montagegeschehen in der Szene wahr und aktualisiert schritthaltend eine Beschreibung, die Informationen über die Objekte und deren eingegangene Verbindungen enthält. Das Konzeptwissen über Aggregate und die Rollen einzelner Komponenten ist über assoziative Verknüpfungen an die Deliberation angebunden. Wenn der Agent prüft, welches Objekt sich für den Bau eines Aggregats eignet, wird relevantes Langzeitwissen aktiv und so für weitere Inferenzen verfügbar.

Um die Handlungsplanung dynamisch an die Konstruktion des entstehenden Aggregats anpassen zu können, werden Pläne möglichst allgemein formuliert. Wenn Max einen Plan für den Bau eines Propellers entwirft, so gibt dieser den allgemeinen Typ der dafür zu verwendenden Bauteile an, legt sich aber nicht auf spezielle Typen oder gar konkrete Objektinstanzen fest. Dadurch ist der Agent in der Lage, sich auf unterspezifizierte Absichten seines Konstruktionspartners und deren Konkretisierung in der laufenden Interaktion einzustellen. Max verwendet eine Constraint-basierte Repräsentation, die es erlaubt, möglichst viele Optionen offen zu lassen, und die gleichzeitig die geltenden und ausgehandelten Einschränkungen explizit macht. Die Constraints gelten dabei innerhalb des gesamten Handlungskontextes, werden an Unterpläne vererbt und können auch versprachlicht werden. In Abbildung 15.8, die den Ablauf einer multimodalen Äußerungsplanung zeigt, sind Constraints bei der Satzplanung dargestellt.

15.4.4 Sprach- und Gestenverarbeitung

Im Rahmen der multimodalen Verhaltensverarbeitung perzipiert Max seinen menschlichen Partner in der realen Welt (Bewegung, Blickrichtung, Gestik) über Sensor-Marker, Datenhandschuhe, einen Eye-Tracker und durch ein Mikrophon, das Daten an seine Sprachverarbeitungskomponente liefert. Die virtuelle Umwelt perzipiert er über zwei visuelle Sensoren, die jeweils einen Blickwinkel auf die Szene definieren und sich mit den Augen des Agenten bewegen.

Spracheingaben des menschlichen Interaktionspartners analysiert Max in mehreren Verarbeitungsschritten, beginnend mit dem Registrieren, dass etwas gesagt wurde, und der Verarbeitung des Signals durch einen Spracherkenner [25]. Die Daten des Spracherkenners werden dann weiter verarbeitet, um die syntaktischen Kategorien Subjekt, Prädikat und Objekt zuzuordnen (*Part-of-Speech-Tagging*). Es folgen die semantische Analyse und die Auflösung von Referenzen. Parallel dazu wird die Äußerung nach Hinweisen auf die Intention des Sprechers und das Performativ (*inform, query, request, propose*) des Sprechers ausgewertet. Die gewonnenen Daten füllen die Felder einer Frame-artigen Wissensstruktur, die in Form eines Dialogakts sämtliche für die deliberative Verarbeitung relevanten Aspekte beinhaltet: illokutionärer Akt, perlokutionärer Akt, propositionaler Inhalt, Turn-Taking-Funktion, Bezug zu vorherigen Äußerungen sowie den Adressaten.

Für die Referenzauflösung wird ein *Constraint Satisfaction*-Verfahren eingesetzt [55]. Die Referenzauflösung wie auch die Intentionserkennung erfordern den Einbezug von Kontextwissen über den mentalen Zustand des Agenten, den Partner, die aktuelle Szene, sowie über die Diskurshistorie. Aus diesem Grund ordnet Max mit Hilfe spezieller Pläne jede Äußerung des Menschen in einen relevanten Diskurskontext ein. Dabei wird überprüft, ob sich die Äußerung direkt auf aktuelle Vorhaben oder Ziele des Agenten bezieht oder auf vorherige Dialogbeiträge, die in der Diskurshistorie gespeichert sind. Dies geschieht durch die Berechnung einer Korrelation

zwischen dem Inhalt des Dialogaktes und der aktuellen Intention des Agenten. Wenn der Agent einen passenden Kontext findet, so wirft er als Unterziel die Obligation auf, die Äußerung in diesem Kontext zu bearbeiten. Durch diese Assoziation eines Dialogaktes mit einem Ziel-Kontext ist Max in der Lage, mit mehreren Handlungssträngen gleichzeitig umzugehen.

Um multimodale Eingaben zu verarbeiten, durchsuchen spezielle Detektoren fortwährend die Sensordaten aus Bewegungstracking und Datenhandschuhen nach Mustern, die auf eine Geste hindeuten (aufbauend auf [45]). Entsprechend der Merkmalsstruktur einzelner Gesten sind die Detektoren als hierarchische Netze organisiert. Erkannte Gesten werden dann im Kontext der zeitnahen verbalen Äußerung sowie der aktuellen virtuellen Szene ausgewertet und multimodal integriert. Diese Integration geschieht nach temporalen und semantischen Aspekten und wird von erweiterten Zustandsübergangsautomaten durchgeführt.

15.4.5 Turn-Taking

Als zentrale Form des Dialogmanagements besitzt Max Turn-Taking-Kompetenzen. Max erkennt und interpretiert Turn-Taking-Signale des menschlichen Interlokutors. Dazu wird das Sprachsignal an einen *is-speaking*-Detektor geleitet. Dieser reagiert erst ab Äußerungen einer gewissen Länge, sodass Feedbacksignale wie „*hmmm*“ und „*okay*“ nicht als Sprechakt im Sinne eines Turns gewertet werden. Zusätzlich zur Sprachverarbeitung werden Detektoren für Gestenerkennung und multimodale Integration eingesetzt. Die Detektoren werden in Abhängigkeit vom konversationalen Zustands aktiviert und deaktiviert, beziehen den aktuellen konversationalen Zustand in ihre Berechnungen mit ein und operieren kontextsensitiv. Die erkannten Gesten, Blickrichtungen und sprachlichen Signale werden auf konversationale Funktionen abgebildet. Innerhalb der kognitiven Komponente spielen die konkreten Ausprägungen der Signale keine Rolle, die Signale können aber die Dringlichkeit ihrer Funktion aufzeigen.

Es sind drei Zeitpunkte im Verlaufe der Interaktion vorgesehen, zu denen der Agent entscheiden kann, auf eine Unterbrechung einzugehen. Abbildung 15.6 zeigt die relevanten Zustände beim Aushandeln des Turns in zeitlicher Abfolge sowie die aktiven Pläne des Agenten. Desweiteren werden der Wechsel des konversationalen Zustands und der konverationalen Rollen sowie die Sprechhandlungen der Agenten angezeigt. Zunächst wird ein wahrgenommenes Turn-Taking-Signal von einem Plan verarbeitet, der Verhaltensweisen (z.B. Blicke und Gesten) anstößt, die dem Gegenüber vermitteln, dass der Agent den Interlokutor wahrgenommen hat. Um dem regelbasierten Charakter der Turn-Taking-Mechanismen gerecht zu werden, werden für Max *Conclude*-Pläne eingesetzt, welche auf die Signale *Taking-Turn*, *Wanting-Turn*, *Giving-Turn* oder *Yielding-Turn* „lauern“. Sie sind jeweils durch Vorbedingungen so eingeschränkt, dass ein kontext-sensitives Verhalten in Abhängigkeit vom konversationalen Zustand und der konversationalen Funktion des wahrgenommenen Signals in Form einer Obligation generiert wird. Der erste Entscheidungszeitpunkt ist unabhängig vom Inhalt der Äußerung des Gegenübers. Die Entscheidung, sich zu diesem frühen Zeitpunkt unterbrechen zu lassen, ist damit allein abhängig von der Dominanzbeziehung zwischen den Konversationspartnern und den Dringlichkeiten, mit welchen die Interlokutoren jeweils ihr Rederecht verfolgen. So sieht man in der Beispielsinteraktion (Abbildung 15.6), wie der Interlokutor Max mit einer Zwischenfrage unterbricht. Dies führt zur Detektion der konversationalen Funktion *TakingTurn* und der Verarbeitung durch den entsprechenden Conclude-Plan. Max entscheidet sich seinen Turn zu unterbrechen und macht dies durch eine *GivingTurn*-Handlung deutlich („*Ja*?“). Nachdem der Sprechakt des Interlokutors interpretiert wurde, besteht erneut die Möglichkeit, dass Max sich unterbrechen lässt.

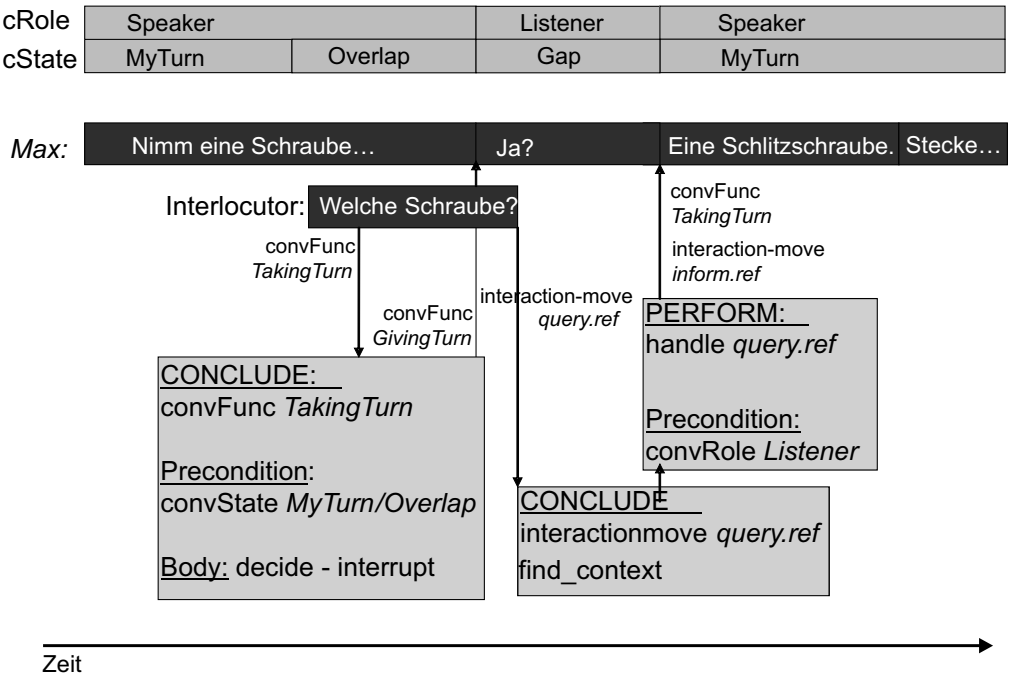

Abbildung 15.6: *Turn-Verarbeitung: Zeitlicher Ablauf.*

Zu diesem zweiten Entscheidungspunkt beruht die Entscheidung allein auf dem bis dahin erkannten Performativ des Dialogakts; handelt es sich um eine Frage (*query*) oder eine direkte Anweisung (*request*), so sollte sich der Agent direkt seinem Konversationspartner zuwenden. Im Falle der Performative *inform* und *propose* fährt Max zunächst mit seiner Äußerung fort. In Abhängigkeit der Relevanz des registrierten Sprechakts lässt er sich aber noch zu einem dritten Zeitpunkt unterbrechen, nachdem die Äußerung in den intentionalen Kontext des Agenten eingeordnet wurde. Besteht ein offensichtlicher Konflikt mit bestehenden Plänen und Intentionen, so kann der Agent direkt darauf einzugehen. Die letzte Möglichkeit des Agenten, auf eine Unterbrechung einzugehen, besteht nach Beendigung seines Turns. In der Beispielinteraktion (Abbildung 15.6) wurde der Inhalt der Zwischenfrage erkannt und einem passenden Kontext zugeordnet. Max geht auf die Frage ein, indem er seinen Interlokutor darüber aufklärt, dass eine Schlitzschraube gemeint war. Danach führt Max seinen ursprünglichen Turn fort.

Möchte der Agent selbst eine Äußerung machen, so muss er sich an konversationale Regeln halten und die Rolle des Sprechers mit seinem Interaktionspartner aushandeln. Während der Interaktion modelliert Max intern explizit seine Sichtweise des konversationalen Zustands (*MyTurn*, *OthersTurn*, *Gap* oder *Overlap*) sowie der Rollenverteilung der Interlokutoren (*uninvolved*, *speaker* oder *listener*). Die Rolle des Konversationspartners wird aufgrund des overten Verhaltens beurteilt. Für sich selbst setzt Max nur dann die Rolle „Sprecher“, wenn er eine längere Äußerung vornehmen möchte; so kann er auch kurzes konversationales Feedback geben, ohne sich in der Rolle des Sprechers zu sehen. Um die Rolle des Sprechers anzunehmen, muss Max zuvor den konversationalen Zustand *MyTurn* herbeiführen. Besteht der konversationale Zustand aus *OthersTurn*, *Gap* oder *Overlap*, so existieren verschiedene Handlungspläne, wel-

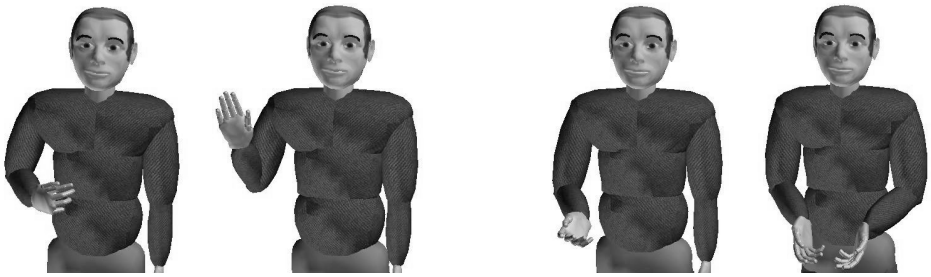

Abbildung 15.7: *Gesten für Wanting-Turn, Taking-Turn, Giving-Turn (Stufe 1), Giving-Turn (Stufe 2).*

che durch Vor- bzw. Kontextbedingungen kontext-sensitiv modelliert sind und beispielsweise die Produktion von Gesten anstoßen, deren Ausdrucksstärke iterativ gesteigert wird (siehe Abbildung 15.7), bis der Agent den Turn bekommen hat oder die Pläne als gescheitert gelten. Möchte Max gezielt sein Gegenüber unterbrechen, so findet ein gesonderter Plan Einsatz, welcher unabhängig von der Rollenverteilung und dem konversationalen Zustand ausgeführt wird.

15.4.6 Multimodale Verhaltensgenerierung

Wirft die Deliberation des Agenten eine Handlungs- bzw. Kommunikationsintention auf, übernehmen spezielle Generator-Pläne die Ableitung von geeigneten Aktionssequenzen sowie deren Ausdifferenzierung in entweder multimodale Äußerungen oder manipulative Aktionen. Der Ablauf der Äußerungsplanung ist in Abbildung 15.8 dargestellt und startet von einem kommunikativen Ziel, das sich direkt aus der aktuellen Intention des Agenten ergibt (Inhaltsplanung). Die folgenden Schritte werden von einem speziellen Planer vorgenommen, der sich an systemischen Grammatiken [33] orientiert. Der Planer wählt den passenden Dialogakttyp aus und expandiert einzelne Konzepte schrittweise gemäß ihrer semantischen Rolle. Er operiert auf einer Repräsentation, die sich aus einem Template-Knoten für die Aktion (das Performativ) sowie weiteren Knoten (*MessageNodes*) für die involvierten Konzepte zusammensetzt, aus denen die Konstituenten des Satzes (z.B. Subjekt, Prädikat, Objekt, Attribut) generiert werden können.

In der Satzplanung werden die MessageNodes mit den abstrakten Konzepten gefüllt, die sie repräsentieren sollen und die durch Merkmalswissen in Form von Attribut-Wert-Listen angereichert sind. Diese Knoten enthalten Realisierungsanweisungen, die einzelne Generierungsentscheidungen repräsentieren und die syntaktische Struktur der Äußerung Schritt für Schritt einschränken. Solche Anweisungen können sich beispielsweise auf das Einfügen von Konstituenten (*+Verb, +Subjekt*), deren Reihenfolge (*Subjekt > Verb*) oder die Realisierung von grammatischen Beugungen beziehen. Dadurch können Subjekt-Verb-Kongruenz oder Attribut-Nomen-Kongruenz realisiert werden, auch kann der Typ eines Objektes die Auswahl des lexikalischen Eintrags des mit dem Objekt assoziierten Verbs beeinflussen. Während der Satzplanung baut der Agent multimodale, referierende Ausdrücke aus sprachlichen und gestischen Anteilen auf – insbesondere deiktischen Gesten, mit denen direkt auf Objekte in der Umwelt verwiesen werden kann, sowie ikonischen Gesten, die räumlich-visuelle Informationen darstellen können. Max kann Gesten explizit durch die Formeigenschaften der bedeutungstragenden Phase (wie z.B. Handform oder Bewegung) oder indirekt durch die Angabe einer kommunikativen

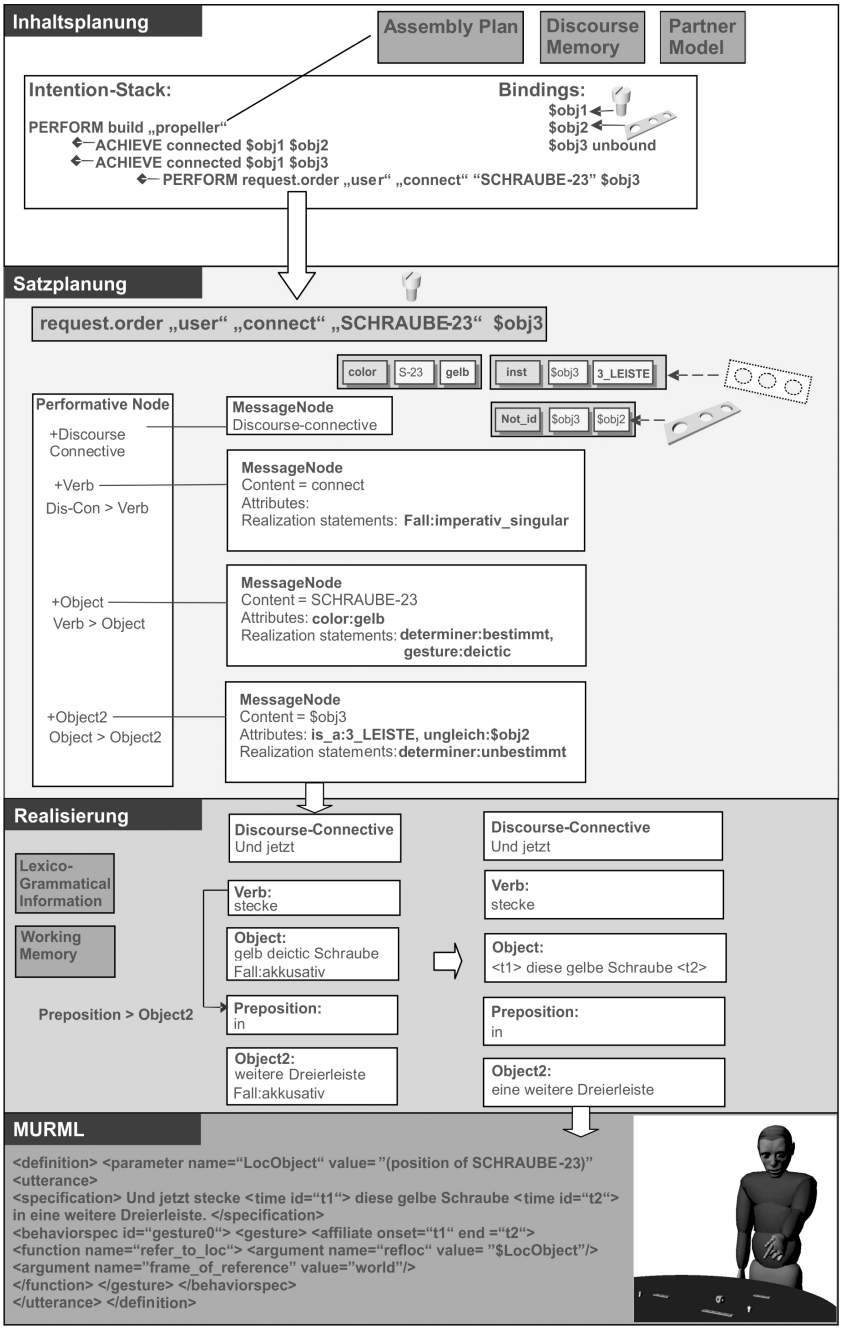

Abbildung 15.8: *Beispielablauf einer multimodalen Äußerungsplanung.*

Funktionen spezifizieren [43]. In Abbildung 15.8 wird im MessageNode der definit bestimmten Schraube (SCHRAUBE-23) eine deiktische Geste vorgesehen. Als letzten Unterschritt der Satzplanung leitet Max die Funktion der Äußerung im Diskurs sowie das Diskurssegment, zu dem sie beiträgt, aus seiner aktuellen Intentionen ab. Ist die Äußerung Teil einer bereits vorgeplanten längeren Sequenz, so fügt er passende Konnektoren ein, wie zum Beispiel „*als erstes*" zu Beginn einer Sequenz, „*und dann*", „*als nächstes*" für Zwischenschritte und „*als letztes*" für den Abschluss der Sequenz.

In der Realisierungsphase wird für jeden Knoten ein sprachlicher Ausdruck aus dem Lexikon ausgewählt und entsprechend den jeweiligen grammatikalischen Informationen angepasst. Für multimodale Äußerungen wird dabei sichergestellt, dass die Gesten synchron zu den zugehörigen verbalen Elementen geplant werden. So geplante Äußerungen werden in MURML, einer Auszeichnungssprache auf XML-Basis, formuliert und dann in synthetische Sprache und simultane Gesten- bzw. Gesichtsanimationen umgesetzt [43]. Um sprachbegleitend zu gestikulieren, erzeugt Max seine Hand- und Armbewegungen jedes Mal neu und speziell an den Kontext angepasst. Die Gestenplanung bearbeitet dazu zunächst Fragen des Timings, der Auswahl der Gliedmaße und der Einpassung in den räumlichen Kontext (z.B. die Zielposition bei Zeigegesten). Die Motorplanung erstellt dann in Realzeit direkt ausführbare Animationen.

Komplexe Äußerungen entstehen schrittweise aus mehreren sogenannten *Chunks*, einzelne Pakete aus jeweils einer kurzen Intonationsphrase und begleitenden non-verbalen Aktionen. Solche Pakete werden nacheinander kontinuierlich verbunden und mit Phasen wechselseitiger Ablösung produziert. Diese Phasen sind notwendig, um die Synchronität z.B. zwischen Sprache und Geste sicherzustellen, für die sich beide Modalitäten an die jeweils andere anpassen müssen. Innerhalb eines Pakets adaptiert sich die Geste in ihrem Bewegungsablauf an die Anfangszeit und Dauer der zugehörigen Wörter, kann also erst geplant werden, wenn Zeitinformationen aus der Sprachsynthese vorhanden sind. Ist der Körper in der Simulation nicht in der Lage, rechtzeitig zu einer Gestenausführung zu gelangen, wird die Aussprache der zugehörigen Intonationsphrase hinausgezögert. Die kognitive Komponente wird dabei durch ständiges Feedback über den Planungs- und Ausführungsstand einzelner Pakete und ihrer sprachlichen wie gestischen Anteile informiert. Die Flexibilität seines Sprachsynthesesystems erlaubt es Max, Betonung an jeder gewünschten Stelle sinnvoll zu erzeugen und mit dem Gestenplanungsprozess abzustimmen. Auf diese Weise kann Max im richtigen und natürlich wirkenden Miteinander sprechen und gestikulieren.

15.4.7 Physis, Emotionen, Bewegungsgenerierung

Um natürliche Ausdrucksmöglichkeiten samt multimodalen Ausdrücken, sprachlichen Äußerungen, emotionalen Reaktionen und situationsangepasstem Blickverhalten zu beherrschen, verfügt Max über eine anthropomorphe Erscheinung. Die äußerliche Physis von Max besteht aus einer mehrteiligen Hülle, die sowohl seinen Körper als auch sein Gesicht umfasst. Der Körper wird durch ein bewegliches kinematisches Skelett gesteuert, das insgesamt 68 Segmente und 103 Freiheitsgrade in 57 Gelenken enthält, davon allein 25 Freiheitsgrade (16 Gelenke) in den Fingern jeder Hand.

Das Gesicht von Max besteht aus einer verformbaren „Haut" mit simulierten Muskeleffekten. Zur Darstellung der Sprechmimik wird der zu sprechende Satz als Abfolge von Lauten durch eine Liste aus Phonemen modelliert; für deutsche Sprache werden 39 Phoneme einbezogen, die

auf 11 sog. „Viseme“ (visuelle Phoneme) abgebildet werden. Viseme beschreiben die Gesichtsstellung (Mund, Lippen etc.) bei der Artikulation eines oder mehrerer Phoneme und werden ineinander überführt. Mit dem Beginn eines jeden Lautes beginnt zeitgleich die Animation des Visems, so dass der Mund synchron zum Sprechen bewegt wird.

Für das Engagementverhalten zeigt Max zusätzlich ein leicht verständliches Feedback seines inneren Zustandes in Form expressiver Gesichtsausdrücke; z.B. kann Max verständnislos oder nachdenklich schauen. Zudem können mimische Elemente auch das von Max Gesagte unterstreichen (freundliches Gesicht bei dem Angebot von Hilfe). Beide Arten der Gesichtsbewegung lassen sich kombinieren. Für die Berechnung des emotionalen Zustands des Agenten wird im Gegensatz zu den kommunikationsgetriebenen Ansätzen eine kontinuierliche Simulation der Emotionsdynamik [7] eingesetzt. Sowohl Impulse von außen (durch seine Perzeption) als auch von innen (ein Ziel wird erreicht oder verfehlt) wirken auf den momentanen emotionalen Zustand von Max. Da sich aufeinander folgende Impulse aufsummieren, können auch starke Emotionen entstehen; bei Ausbleiben von Impulsen strebt das System zurück in den neutralen Zustand.

Eine weitere Form des Engagement wird durch die Modellierung interaktiven Blickverhaltens umgesetzt, welches den aktuellen Aufmerksamkeitsfokus des Agenten verdeutlicht und bis hin zur Herstellung gemeinsamer Aufmerksamkeit reicht [56].

15.5 Zusammenfassung und Ausblick

Neben einem Überblick über grundlegende Fragestellungen und aktuelle technische Ansätze wurde am Beispiel des künstlichen Agenten „Max“ ein konkretes System beleuchtet, das „verkörperte Kommunikation“ mit Computersimulationen in virtueller Realität realisiert.

An diesem Beispiel wurde gezeigt, wie virtuelle Agenten mit synthetischer Stimme und einem animierten Körper sprechen, gestikulieren und Gesichtsausdrücke zeigen können. Analog können sie Sprache, Gestik und Blickrichtung des Menschen als Eingaben verarbeiten. Doch die Entwicklung verkörperter Agenten, die tatsächlich mit menschlichen Partnern interagieren sollen, verdeutlicht, dass verkörperte Kommunikation mehr als multimodale Kommunikation ist. Sie verlangt die enge Verbindung von Wahrnehmen und Handeln in der Interaktion auf zwei Ebenen. Auf der Kommunikationsebene müssen verbale und non-verbale Signale im Sinne einer kohärenten multimodalen Konversation zusammenwirken. Entsprechende Ansätze zur Verhaltensverarbeitung, Verhaltensgenerierung und Dialogmodellierung wurden vorgestellt. Gleichzeitig müssen diese auf der Architekturebene durch eine kognitiv plausible Verbindung von reaktivem und deliberativem Verhalten umgesetzt werden. Am Beispiel Max wurde konkret gezeigt, wie dazu kognitive Architektur, Emotionsmodellierung, Sensorik und Physis zusammenspielen.

Die Erforschung natürlicher verkörperter Kommunikation birgt noch viele Herausforderungen für die Entwicklung intelligenter interaktiver Systeme. Die wohl größte besteht in der Umsetzung der vielschichtigen Dynamik menschlicher verkörperter Kommunikation. Ein wichtiger Baustein dafür ist die Fähigkeit der inkrementellen und parallelen Verarbeitung, sowohl in der Wahrnehmung und Interpretation der Signale des menschlichen Partners als auch der Generierung und Anpassung des Agentenverhaltens. Jüngere Arbeiten verfolgen dazu Ansätze der kontinuierlichen Verhaltensadaptation (z.B. in der Sprach- und Gestengenerierung [14]), mit der

künstliche Kommunikationspartner bereits während einer eigenen Äußerung auf das Feedback des Adressaten reagieren können. Mit der verkörperten Kommunikation geraten dabei kognitive Mechanismen und Architekturprinzipien in den Blick, die auch für die sprachbasierten Systeme der Zukunft prägend sein können.

Literaturverzeichnis

[1] Allwood, J. (1976). Linguistic communication in action and co-operation: A study in pragmatics. In *Gothenberg Monographs in Linguistics*, volume 2, pages 637–663. Department of Linguistics, University of Gothenberg.

[2] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., und Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060.

[3] Argyle, M. (1988). *Bodily Communication*. Methuen & Co., New York, 2nd edition.

[4] Argyle, M. und Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press, Cambridge, UK.

[5] Austin, J. L. (1962). *How to do things with words*. Oxford University Press, Oxford, UK.

[6] Baddeley, A. (2007). *Working Memory, Thought, and Action*. Oxford University Press, New York.

[7] Becker, C., Kopp, S., und Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. In *Affective Dialogue Systems*, LNAI 3068, pages 154–165, Berlin. Springer.

[8] Benoit, C., Martin, J.-C., Pelachaud, C., Schomaker, L., und Suhm, B. (2000). Audio-visual and multimodal speech-based systems. In Gibbon, D., Mertins, I., und Moore, R., editors, *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation*, pages 102– 203. Dordrecht, The Netherlands: Kluwer.

[9] Bergmann, K. und Kopp, S. (2009). Increasing expressiveness for virtual agents – autonomous generation of speech and gesture for spatial description tasks. In Decker, K., Sichman, J., Sierra, G., und Castelfranchi, editors, *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 361–368. Ann Arbor, MI: IFAAMAS.

[10] Bergmann, K. und Kopp, S. (2012). Gestural alignment in natural dialog. In *Proc. of the Int. Conference of the Cognitive Science Society (CogSci 2012)*.

[11] Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Havard University Press.

[12] Brennan, S. E. und Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

[13] Bunt, H. (2011). Multifunctionality in dialogue. *Computer, Speech and Language*, 25:222–246.

[14] Buschmeier, H., Bergmann, K., und Kopp, S. (2010). Adaptive Expressiveness – Virtual Conversational Agents That Can Align to Their Interaction Partner. In van der Hoek, W., Kaminka, G. A., Luck, M., und Sen, S., editors, *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 91–98. International Foundation for Autonomous Agents and Multiagent Systems.

[15] Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., und Anderson., A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*,

23(1):13–31.

[16] Cassell, J., Stone, M., und Yan, H. (2000a). Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the International Natural Language Generation Conference*, pages 171–178, Mitzpe Ramon, Israel.

[17] Cassell, J., Sullivan, J., Prevost, S., und Churchill, E., editors (2000b). *Embodied Conversational Agents*. The MIT Press, Cambridge (MA).

[18] Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge, UK.

[19] Cohen, R., Allaby, C., Cumbaa, C., Fitzgerald, M., Ho, K., Hui, B., Latulipe, C., Lu, F., Moussa, N., Pooley, D., Qian, A., und Siddiqi, S. (1998). What is initiative? *User Modeling and User-Adapted Interaction*, 8:171–214.

[20] Cooper, R. und Larsson, S. (1999). Dialogue moves and information states. In *Proceedings of the Third International Workshop on Computational Semantics*, pages 398–400.

[21] Core, M. G. und Allen, J. F. (1997). Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines.

[22] Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, New York.

[23] Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

[24] Elio, R. (2002). Belief-desire-intention agency in a general cognitive architecture. *Cognitive Science Quarterly*, 2:321–340.

[25] Fink, G., Schillo, C., Kummert, F., und Sagerer, G. (1998). Incremental speech recognition for multimodal interfaces. In *Proceedings 24th Annual Conference of the IEEE Industrial Electronics Society*, pages 2012–2017.

[26] Fröhlich, M. und Wachsmuth, I. (1998). Gesture recognition of the upper limbs – from signal to symbol. In Sowa, T. und Wachsmuth, I., editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 173–184. Springer.

[27] Georgeff, M. und Lansky, A. (1987). Reactive reasoning and planning. In *Proceedings of the Sixth National Conference of Artificial Intelligence*, pages 677–682, Menlo Park, California, USA. AAAI, AAAI Press/MIT Press.

[28] Georgeff, M., Pell, B., Pollack, M., Tambe, M., und Wooldridge, M. (1999). The Belief-Desire-Intention Model of Agency. In Müller, J., Singh, M. P., und Rao, A. S., editors, *Proceedings of the 5th International Workshop on Intelligent Agents V: Agent Theories, Architectures, and Languages (ATAL-98)*, volume 1555, pages 1–10, Heidelberg, Germany. Springer-Verlag.

[29] Goffman, E. (1983). The interaction order. *American Sociological Review*, 48:1–17.

[30] Goodwin, C. (1981a). Achieving Mutual Orientation at Turn Beginning. *Conversational Organization: Interaction between speakers and hearers*, pages 55–89.

[31] Goodwin, C. (1981b). *Conversational organization: Interaction between speakers and hearers*. Academic Press New York.

[32] Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., und Badler, N. (2002). Creating interactive virtual humans: Some Assembly Required. *IEEE Intelligent Systems*, pages 54–63.

[33] Halliday, M. (1985). *An Introduction to Functional Grammar*. Arnold.

[34] Horvitz, E. (2007). Reflections on Challenges and Promises of Mixed-initiative Interaction. *AI Magazine*, pages 19–22.

[35] Howden, N., Rönnquist, R., Hodgson, A., und Lucas, A. (2001). JACK Intelligent Agents-Summary of an Agent Infrastructure. In *Proceedings of the 5th ACM International Conference on Autonomous Agents*.

[36] Huber, M. J. (1999). JAM: A BDI-theoretic Mobile Agent Architecture. In *Proceedings of the International Conference on Autonomous Agents*, pages 236–243, Seattle, WA.

[37] Johnston, M. und Bangalore, S. (2001). Finite-state methods for multimodal parsing and integration. In *Proc. of the ESSLLI Summer School on Logic, Language, and Information, Helsinki, Finland*, pages 1–6.

[38] Jonsdottir, G. R., Thorisson, K. R., und Nivel, E. (2008). Learning smooth, human-like turn-taking in realtime dialogue. In *Proceedings of the 8th international conference on Intelligent Virtual Agents*, IVA '08, pages 162–175, Berlin, Heidelberg. Springer-Verlag.

[39] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:1–47.

[40] Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52(6):587–597.

[41] Kopp, S., Allwood, J., Ahlsen, E., Grammer, K., und Stocksmeier, T. (2008). Modeling embodied feedback in a virtual human. In Wachsmuth, I. und Knoblich, G., editors, *Modeling Communication With Robots And Virtual Humans*, pages 18–37. Springer-Verlag, Berlin.

[42] Kopp, S., Jung, B., Lessmann, N., und Wachsmuth, I. (2003). Max – A Multimodal Assistant in Virtual Reality Construction. *KI-Künstliche Intelligenz*, 4/03:11–17.

[43] Kopp, S. und Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52.

[44] Lakin, J. L., Jefferis, V. E., Cheng, C. M., und Chartrand, T. L. (2003). The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, 27(3):145–162.

[45] Latoschik, M. (2001a). *Multimodale Interaktion in Virtueller Realität am Beispiel der virtuellen Konstruktion*, volume 251 of *DISKI*. Akademische Verlagsgesellschaft Aka GmbH, Berlin.

[46] Latoschik, M. E. (2001b). A gesture processing framework for multimodal interaction in virtual reality. In *Proc. of 1st International Conference on Computer Graphics, Virtual Reality and Visualization in Africa (Afrigraph 2001)*, pages 95–100.

[47] Latoschik, M. E. (2002). Designing transition networks for multimodal VR interactions using a markup language. In *Proc. of the 4th IEEE Int. Conf. on Multimodal Interfaces (ICMI 2002)*, pages 411–416. IEEE Computer Society.

[48] Lessmann, N., Kopp, S., und Wachsmuth, I. (2006). Situated interaction with a virtual human – perception, action, and cognition. In Rickheit, G. und Wachsmuth, I., editors, *Situated Communication*, pages 287–323. Mouton de Gruyter, Berlin.

[49] Lessmann, N., Kranstedt, A., und Wachsmuth, I. (2004). Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max. In *AAMAS 2004 Workshop on Embodied Conversational Agents: Balanced Perception and Action*, pages 57–64.

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

[51] McTear, M. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys*, 34(1):90–169.

[52] Mehrabian, A. und Ferris, S. R. (1967). Inference of attitutes from non-verbal communication in two channels. *Journal of Consulting Psychology*, 31:248–252.

[53] Oberauer, K. (2003). The Multiple Faces of Working Memory: Storage, Processing, Supervision, and Coordination. *Intelligence*, 31:167–193.

[54] Ortony, A., Clore, G. L., und Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

[55] Pfeiffer, T. und Latoschik, M. E. (2004). Resolving object references in multimodal dialogues for immersive virtual environments. In *Proceedings of the IEEE Virtual Reality 2004*, Illinois: Chicago.

[56] Pfeiffer-Lessmann, N. und Wachsmuth, I. (2008). Toward alignment with a virtual human – Achieving joint attention. In Dengel, A., Berns, K., und Breuel, T., editors, *KI 2008: Advances in Artificial Intelligence*, Berlin. Springer-Verlag.

[57] Pickering, M. J. und Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Science*, 27:169–226.

[58] Poggi, I. und Pelachaud, C. (2000). Performative facial expression in animated faces. In [17].

[59] Pokahr, A., Braubach, L., und Lamersdorf, W. (2005). Jadex: A BDI Reasoning Engine. In Bordini, R., Dastani, M., Dix, J., und Seghrouchni, A. E. F., editors, *Multi-Agent Programming*, pages 149–174. Springer Science+Business Media Inc., USA.

[60] Prillwitz, S., Leven, R., Zienert, H., Hanke, T., und Henning, J. (1989). *HamNoSys version 2.0: Hamburg Notation System for Sign Languages: An introductory guide (Vol. 5).* Hamburg: Signum Press.

[61] Rao, A. und Georgeff, M. (1998). Decision procedures of BDI logics,. *Journal of Logic and Computation*, 8(3):293–344.

[62] Raux, A. und Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 629–637.

[63] Reichardt, D. M., editor (2011). *KI – Künstliche Intelligenz.* Volume 25, Issue 3 / August 2011 (Special Issue on Emotion and Computing).

[64] Reiter, E. und Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

[65] Rosenbloom, P., Laird, J., und Newell, A. (1993). *The Soar Papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA, USA.

[66] Sacks, H., Schegloff, E., und Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

[67] Scherer, K., Bänzinger, T., und Roesch, E., editors (2010). *Blueprint for Affective Computing*. Oxford University Press.

[68] Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press, New York.

[69] Sidner, C., Lee, C., und Lesh, N. (2003). Engagement Rules for Human-Robot Collaborative Interactions. In *IEEE International Conference on Systems, Man & Cybernetics (CSMC)*, volume 4, pages 3957–3962. IEEE Press.

[70] Sowa, T. (2006). *Understanding Coverbal Iconic Gestures in Shape Descriptions*. PhD thesis, Doctoral dissertation, DISKI series, Vol. 294, Berlin: AKA/Infix.

[71] Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., und Bregler, C. (2004). Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of SIGGRAPH '04*, pages 506–513.

[72] Thorisson, K. R. (2002). Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In *Multimodality in Language and Speech Systems*, pages 173–207. Kluwer Academic Publishers, The Netherlands.

[73] Traum, D. (1996). Conversational agency: the TRAINS-93 dialogue manager. In *Proc. 11th Workshop on Language Technology: Dialogue Management in Natural Language Systems*, Universiteit Twente, Enschede, The Netherlands.

[74] Traum, D. und Larsson, S. (2003). The information state approach to dialogue management. In Smith, R. und Kuppevelt, J., editors, *Current and New Directions in Dialogue*. Kluwer.

[75] Wachsmuth, I., Lenzen, M., und Knoblich, G., editors (2008). *Embodied Communication in Humans and Machines*. Oxford University Press.

[76] Waibel, A., Vo, M. T., Duchnowski, P., und Manke, S. (1996). Multimodal interfaces. *Artificial Intelligence Review*, 10:299–319.

[77] Winograd, T. (1972). *Understanding Natural Language*. Academic Press, New York.

[78] Wooldridge, M. (2000). *Reasoning about Rational Agents*. MIT Press, Cambridge, MA, USA.

[79] Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the 6th Regional Meeting of the Chicago Linguistics Society*, pages 567–578. University of Chicago.

[80] Zimbardo, P. G. (1995). *Psychologie*. Springer-Verlag.