FEHLERTOLERANTE MULTIMODALE FUSION VON N-BESTEN LISTEN VON SPRACHEINGABE UND ZEIGEGESTEN

Hartwig Holzapfel, Alex Waibel

Interactive Systems Labs
Universität Karlsruhe (TH)

{*hartwig,waibel*}*@ira.uka.de*

Abstract: Gestik wird häufig in Mensch-Mensch-Kommunikation eingesetzt, um Gesprochenes zu verdeutlichen (Redundanz) oder nichtredundante Informationen hinzu zu fügen. Derartige nichtredundante Information sind z.B. Zeigegesten, die zur Disambiguierung von Objekten dienen.

Fusion von Sprache und Gestik setzen wir bereits erfolgreich in unserem Projekt humanoide Roboter ein, um dem Benutzer eine natürlichere Schnittstelle zu bieten. In der vorliegenden Arbeit wird nun ein darauf aufsetzendes Verfahren beschrieben, das neben der Gestikauflösung auch Hypothesen aus der n-besten Liste der Spracherkenner-Ausgabe wählt, um das Fusionsergebnis zu verbessern. Hierzu wird eine Score-Funktion verwendet, die es erlaubt, Kombinationen der verschiedenen n-besten Listen zu bewerten und damit verschiedene Fusionsergebnisse zu ordnen. Die Ausgabe der Fusion stellt dabei ebenfalls eine n-besten Liste dar.

Die Fusion verwendet Informationen über die zeitliche Abfolge von Eingaben, um darauf mögliche Gruppierungen zu erstellen. Dabei bestimmen Fusionsregeln durch Constraints, ob verschiedene Eingaben aufgrund ihrer semantischen Repräsentation kombiniert werden können. Verschiedene Kombinationsmöglichkeiten entstehen erst durch die Verarbeitung von n-besten Listen. Die Score-Funktion, zur Bewertung der unterschiedlichen Ergebnisse, verwendet Konfidenzen von Sprach- und Gestikerkenner, sowie Wahrscheinlichkeiten für Fehldetektionen von Gesten.

Der Ansatz wurde mit zwei Basissystemen verglichen. Der erste Ansatz verwendet jeweils nur die beste Hypothese von Sprach- und Gestikerkenner, der zweite Ansatz versucht die am besten passende Geste aus einer n-besten Liste zu wählen. Eine Evaluation wird gezeigt, dass der neue Ansatz eine Verbesserung gegenüber den bestehenden Systemen darstellt.

1 Einleitung

Multimodale Benutzerschnittstellen [1] sind seit längerem ein interessantes Forschungsgebiet seit Bolts grundlegender Arbeit [2]. Seit einiger Zeit werden multimodale Benutzerschnittstellen eingesetzt, um dem Benutzer eine natürliche Interaktion mit humanoiden Robotern zu ermöglichen [3], [4].

Hierbei müssen verschiedene Eingabeströme verarbeitet werden, im Kontext der aktuellen Situation interpretiert werden, und eine passende Präsentationsform für Informationen gewählt werden.

Der Schwerpunkt dieser Arbeit liegt auf der Integration verschiedener Modalitäten. Es wurden multimodale Befehle untersucht, mit denen der Benutzer den Roboter instruieren kann, Lichter ein- und auszuschalten.

Die Spracheingabe wird durch einen Spracherkenner erkannt und mit Hilfe von kontextfreien semantischen Grammatiken in eine semantische Repräsentation konvertiert. Die Bewegungen des Benutzers werden durch eine Stereo-Kamera verfolgt und extrahiert. Durch den Einsatz einer Stereo-Kamera können Tiefeninformationen erfasst werden, wodurch Zeigegesten im 3-dimensionalen Raum möglich sind. Die Zeigerichtung wird in der Umweltdatenbank aufgelöst, welche eine Liste der Objekte zurück liefert, die am besten zu der Zeigerichtung passen.

Die Score-Funktion zur Bewertung und Vergleich verschiedener Fusionsergebnisse verwendet Konfidenzen von Spracherkenner und Gestik-Auflösung um die besten Kombinationen zu ermitteln. Abhängig von der Passgenauigkeit zwischen Zeigerichtung und Objektkoordinaten, wird eine Wahrscheinlichkeit für jedes Objekt berechnet. Da der Gestikerkenner auch Fehlerkennungen liefert, ist es nötig, dass eine Spracheingabe nicht zwingend mit einer Gestik kombiniert werden muss, sondern dass System auch eine Fehldetektion annehmen kann. Um dies korrekt durch die Score-Funktion erfassen zu können, werden Konfidenzen für die Detektion einer Geste verwendet.

1.1 Verwandte Arbeiten

Es existieren bereits verschiedene Systeme, die multimodale Fusion verwenden, wie z.B. [5], [6]. Hierfür gibt es verschiedene Ansätze, wie z.B. statistische Ansätze [7], salience-basiert [8], regelbasiert [9], und biologisch motivierte Ansätze [10].

Existierende Arbeiten mit Zeigestift-Eingabe [11] haben gezeigt, dass Sprach- und Gestikeingabe auf der Zeitebene korreliert sind, ebenso Untersuchungen mit mehreren Benutzern [12], Blickrichtung und Sprache [13], und Prosodie [14]. In unseren Experimenten haben wir festgestellt, dass das System viele Fehldetektionen von Gesten verarbeiten muss, was einen Unterschied zu Gestikeingabe auf 2-dimensionalen Oberflächen darstellt, wie z.B. bei Verwendung von Zeigestiften [11].

Wir erweitern existierende Arbeiten für multimodales Parsing [9] [15], um unter anderem Fehldetektionen verarbeiten zu können und Nutzen aus n-besten Listen ziehen zu können. Weiterhin kann die Fehlerrate dadurch gesenkt werden, dass Sprache als Hauptmodalität und Gestik zur Disambiguierung verwendet wird. Dadurch sind dann zwar keine Eingaben möglich, die nur durch Gestik ausgedrückt werden, allerdings war dies bisher auch nicht bei Benutzern zu beobachten.

1.2 Anwendungsszenario

Die hier beschriebene Fusionskomponente wurde auf einer humanoiden Roboter-Plattform getestet und evaluiert. Die Aufgaben des Roboters sind zur Zeit Hol- und Bringdienste, sowie die Steuerung von elektrischen Geräten, wie Ein- und Ausschalten von Lichtern, oder der Klimaanlage. Einsatzumgebung des Roboters ist eine Küche, in der er dem Menschen (weiterhin auch als Benutzer bezeichnet) unterstützen soll [16].

Die Interaktion des Benutzers mit dem System erfolgt durch Sprache und Gestik. Spracheingabe erfolgt wahlweise mit einem Nahbesprechungsmikrophon (Headset) oder über Raummikrophone, die auf dem Roboter montiert sind. Für beide Eingabemöglichkeiten wurden Experimente in englischer Sprache durchgeführt [17]. Inzwischen wurde auch eine deutsche Grammatik entwickelt, so dass weitere Experimente auf deutsch durchgeführt werden können. Zeigegesten des Benutzers werden über zwei Kameras, die auf dem Roboterkopf montiert sind erkannt. Der Benutzer kann sich dadurch frei bewegen, ohne an Verkabelungen gebunden zu sein. Abbildung 1 zeigt die Interaktion mit dem aktuellen System, sowie die neue Plattform, auf die zur Zeit umgestellt wird [18].

Der Benutzer bekommt Feedback durch Aktionen und Sprachausgabe des Systems. Hat das System eine Aufgabe erkannt (im vorliegenden Experiment Licht an-/ausschalten), fährt der Roboter zur Lampe und führt die entsprechende Aktion aus. Kann der Roboter die Eingabe nicht korrekt klassifizieren, oder die Aufgabe nicht ausreichend disambiguieren, wird vom Dialogmanager eine entsprechende Rückfrage gestellt.

Für die Gerätesteuerung wird ein kommerzielles Paket eingesetzt, mit dem über X10 (Kommunikation über Stromnetz), Funk und Infrarot Geräte ferngesteuert werden können.

a) b)

Abbildung 1 - a) Interaktion mit dem Entwicklungssystem b): Integration auf der neuen Zielplattform

Zwei Beispiele sollen die Interaktion mit dem System (mit Sprache und multimodal) verdeutlichen. Beispiel 1 - nur Sprache
User: please switch on the light
System: which light do you want me to switch on
User: the big lamp

Beispiel 2 - Sprache + Gestik
User: please switch on the light
System: which light do you want me to switch on
User: this lamp + Zeigegeste

2 Eingabeverarbeitung und Systemkomponenten

Für Spracherkennung verwenden wir das Janus Recognition Toolkit (JRTk) [19], mit dem Ibis single pass-decoder [20]. Mit Ibis können kontextfreie Grammatiken als Sprachmodell verwendet werden, welche auch gleichzeitig zur Generierung der Semantik verwendet werden. Die Ausgabe des Spracherkenners ist ein Parsebaum, der durch die Konvertierungsregeln der kontextfreien Grammatik in eine typisierte Merkmalsstruktur (TFS) [21] umwandelt.

Zeigegesten des Benutzers werden durch einen Gestikerkenner mit Stereo-Kameras erkannt [22]. Die Ausgabe des Gestikerkenners ist ein Vektor mit sechs Elementen, die die Zeigerichtung im dreidimensionalen Raum repräsentieren. Dieser Vektor wird verwendet, um das Ziel der Geste in der Umweltdatenbank aufzulösen, wobei eine semantische Repräsentation des referenzierten Objektes als Ergebnis dieses Prozesses entsteht.

Die multimodale Fusion verwendet die Ausgaben von Gestikauflösung und Spracherkennung, die jeweils n-besten Listen produzieren und leitet die gemeinsame Bedeutung der Eingaben an die Dialogverarbeitung weiter. Die Ausgabe ist ebenfalls wieder eine n-besten Liste von fusionierten Repräsentationen. Um diese n-besten Listen zu erstellen, muss für jedes Fusionsergeb-

nis eine Score-Funktion berechnet werden, wodurch die Ordnung der Ausgabe festgelegt wird. Bei der multimodalen Fusion werden keine syntaktischen deiktischen Referenzen wie 'this' oder 'that' verwendet, da diese Wörter leicht durch die Spracherkennung verwechselt werden können, sondern mit einem informationsbasiertem Ansatz realisiert. Dabei wird auf Kompatibilität von Objektbeschreibungen auf semantischer Ebene im gegebenen Sprachkontext geprüft.

3 Score-Funktion

Durch die Score-Funktion sollen die besten Kombinationen und Fusionsergebnisse ermittelt werden. Einfluss auf den Gesamt-Score haben die Score-Funktionen der Erkennerkomponenten und eine Bewertung der Fusionsgüte. Die Fusionsgüte ist abhängig von der zeitlichen Korrelation (welche insbesondere bei Zeigegesten und Sprache ein wichtige Rolle spielt), der semantischen Bedeutung der Eingaben und deren gemeinsame Interpretationsfähigkeit. Die Score-Funktion des Spracherkenners basiert auf Standardmaßen der Spracherkennung, die durch Kombination des akustischen Modells und des Sprachmodells durch eine logarithmische Likelihood-Funktion berechnet werden. In die Score-Funktion der Gestikerkennung fließt die Wahrscheinlichkeit mit ein, dass eine Geste erkannt wurde (Konfidenz für Gestikdetektion), und die Genauigkeit mit der die Gestik auf ein Objekt im Umweltmodell referenziert.

3.1 Score-Berechnung für Zeigegesten

In die Berechnung der Score-Funktion fließt zunächst die Konfidenz für die Detektion einer Geste ein. Wir haben diese in unserem Experiment durch eine einfache relative Häufigkeit über die Anzahl der korrekt detektierten Gesten geteilt durch die Anzahl aller gemeldeten Gesten approximiert.

$$conf(g) = P(corr_{detection}) = \frac{f_{correct}}{f_{detections}} \tag{1}$$

Entsprechend kann die Wahrscheinlichkeit für eine Fehldetektion berechnet werden.

$$P(err_{detection}) = 1 - P(corr_{detection}) = 1 - \frac{f_{correct}}{f_{detections}} \tag{2}$$

Um nun die Wahrscheinlichkeit zu bestimmen, dass der Benutzer ein bestimmtes Objekt obj_i mit der Zeigegeste referenziert, wird der Fehlerwinkel zwischen der erkannten Zeigerichtung und der direkten Linie zum Objekt verwendet.

Der Fehlerwinkel ist immer kleiner/gleich 180 Grad. Die Wahrscheinlichkeit, dass der Benutzer ein bestimmtes Objekt obj_i referenziert bezeichnen wir mit dem Term $P(point(g) = obj_i))$. Er wird bestimmt durch

$$P(point(g) = obj_i)) = 1 - \frac{errangle(g, obj_i)}{180} \tag{3}$$

Aus den zwei Gleichungen 1 und 3 können wir nun den Score $PScore$ einer Zeigegeste bestimmen als

$$PScore(g, obj_i) = \| \log(conf(g) * P(point(g) = obj_i)) \| \tag{4}$$

3.2 Score-Funktion für das Fusionsergebnis

Der Score des Fusionsergebnisses wird über die Summe der (logarithmischen) Score-Funktionen der Erkennerkomponenten berechnet. Zusätzlich entscheiden zeitliche Korrelation und semantische Kompatibilität über die Fusion. In unseren Experimenten wurde für die zeitliche Zusammengehörigkeit eine Rechteckfunktion gewählt, mit Wahrscheinlichkeit 0 für Ereignisse

die zu weit auseinander liegen und Wahrscheinlichkeit 1 für Ereignisse die bis 800 Millisekunden auseinander liegen. Eine Benutzerstudie hat gezeigt, dass Gestik- und Spracheingabe zeitlich sehr nahe bei einander liegen bzw. sich überlappen [23]. Mit dem heuristischen Wert von 800ms werden fast alle Eingaben korrekt abgedeckt, mit wenigen Ausnahmen, in denen die Gestikerkennung grobe zeitliche Fehler liefert. Dabei sind auch Zeitungenauigkeiten der Erkennerkomponenten mit berücksichtigt. Um die Berechenbarkeit der Score-Funktion aufrecht zu erhalten, werden Ereignisse, mit Fusionswahrscheinlichkeit 0, ausgelassen. Um zu ermitteln, ob die Semantik der Ereignisse kombinierbar ist, werden die Constraint-basierten Regeln der Fusionskomponente verwendet [23].

Die Score-Funktion bildet eine Heuristik, mit der verschiedene Kombination aus n-besten Listen der Gestikauflösung und Spracherkennung bewertet werden, wobei wieder eine n-besten Liste als Ausgabe erstellt wird. Der interessante Fall ist daher, wenn eine Spracheingabe mit einer oder mehreren Gestikeingaben fusioniert werden soll. Es müssen dabei unter anderem Ergebnisse verglichen werden, die im einen Fall Sprache und Gestik fusionieren, im anderen Fall Sprache ohne Gestik betrachten. Im ersten Fall müssen die Scores für Sprachhypothese und Gestik kombiniert werden, im zweiten Fall muss der Score für die Sprachhypothese mit der Wahrscheinlichkeit für die Fehldetektion einer Geste kombiniert werden.

Gleichung 5 liefert den Score für ein Sprache-Gestik-Paar, bei dem nur das Sprachereignis übernommen wird, und die Gestik als Fehldetektion interpretiert wird. Die Score-Funktion ist immer positiv, da jeweils die Betragsfunktion auf die Logarithmen der Wahrscheinlichkeiten angewendet werden. Die Faktoren α und β in Gleichung 5 und 6 dienen zum Ausgleich und unterschiedlicher Gewichtung. In den Experimenten lagen die Scores für die Spracherkennung im Bereich zwischen 3000 und 13000. α und β wurden jeweils auf 10000 gesetzt.

$$Score_{speechonly}(s, g) = Score(s) + \alpha * \| \log(P(err_{detection})) \| \tag{5}$$

Können Sprache und Gestik fusioniert werden, wird die Score-Funktion wie in Gleichung 6 dargestellt, berechnet. Hierbei liefert obj_i (das Ziel der Gestik) die Information, die mit dem Sprachereignis verschmolzen wird.

$$Score(s, g, obj_i) = Score(s) + \beta * PScore(g, obj_i) \tag{6}$$

4 Experimente und Evaluation

Obige multimodale Fusion (im folgenden System C genannt) wurde auf einem Datenset getestet, das mit der oben beschriebenen Roboterplattform gesammelt wurde. Das Datenset umfasst ca. 500 Benutzereingaben, die Befehle an den Roboter enthalten, Aktionen auszuführen, wovon 102 Benutzereingaben für die multimodale Fusion relevant sind, mit Gestik- und Sprachereignissen, als Ausgabe der Erkennerkomponenten.

Der Ansatz sollte mit einer Fusion der besten Spracherkenner- und Gestikerkennerhypothese als untere Schranke verglichen werden (System A) und einer regelbasierten Fusion (System B), die die Spracherkennerausgabe mit der am besten passenden Gestikhypothese aus der n-besten Liste fusioniert. In vorangegangenen Experimenten wurden auf dem Datenset Wortfehlerrate (WER) und Satzfehlerrate (SER) evaluiert. Auf dem Datenset wurde eine WER von 23.55% und eine SER von 33,13% erziehlt [17].

Die einfache Fusion (System A) schlägt insbesondere dann fehl, wenn Objekte im Raum nahe beisammen sind. Dann kann die Zeigegeste nicht genau genug aufgelöst werden, um das Objekt allein durch visuelle Information zu bestimmen. Wenn sich Koordinaten von Objekten überlappen ist dies gar unmöglich. Es muss daher zusätzliche Information aus der Spracheingabe verwendet werden, um die Geste im Kontext zu interpretieren. In einem vorangegangenen

Experiment wurde bei der Auswahl von drei nahe beieinander stehenden Objekten (eine Lampe, zwei Tassen), zu 56% das korrekte Ergebnis nicht als bestes Ergebnis, aber in der n-besten Liste erkannt. Der erste faire Vergleich kann daher erst mit System B gezogen werden, das referenzierte Objekten aus der n-besten Liste auswählt. In diesem Fall erfolgt die Auswahl aus der n-besten Liste darüber, ob das referenzierte Objekt im gegebenen Sprachkontext Sinn macht. Hiermit konnten, im Gegensatz zur einfachen Fusion, alle Eingaben korrekt fusioniert werden, bei denen der Spracherkenner das korrekte Ergebnis lieferte und das korrekte Ziel der Geste in der n-besten Liste vorhanden war. Eine weitere Verbesserung konnte durch die oben beschriebene Fusion in System C erzielt werden, die neben der n-besten Liste der Gestik auch die n-besten Liste der Spracherkennung betrachtet.

Für einen Vergleich mit System B wurden absolute Verbesserungen gemessen, in Teilen in denen sich die Systeme unterscheiden. Da die Systeme sich insbesondere darin unterscheiden, dass System C aus der n-besten Liste der Spracherkennung auswählen kann, wurde gemessen, wie oft eine Verbesserung dadurch erzielt wurde, dass die korrekte Spracherkennerhypothese aus der Liste ausgewählt wurde.

Gegenüber System B wurde in System C in 2% der Fällen die korrekte semantische Repräsenation aus der n-besten Liste ausgewählt, wo System B Fehler erzeugte. In keinem der Fälle wurde das Ergebnis im Vergleich zum zweiten System verschlechtert.

Eingaben insgesamt	≈ 500
Multimodale Eingaben	102
Korrekte Fusion System B	74%
Korrekte Fusion System C	76%

Abbildung 2 - Prozentsätze für die korrekte Fusion in System B und System C

5 Diskussion der Ergebnisse und Ausblick

Die Korrektrate für die Fusion bei System B liegt bei 74%, wie in auch in Tabelle 2 gezeigt. Davon entfallen allerdings nur $\approx$ 17% relativ auf reine Spracherkennungsfehler bei korrekter Gestik. Durch die 2% erreichte absolute Verbesserung konnten die Fehler, die nur auf die Spracherkennung zurückzuführen sind, halbiert werden. Die relativ geringe Verbesserung ist auch darauf zurückzuführen, dass die generierte Semantik relativ einfach ist und auch Artikel statt Pronomen in der Deixisauflösung erlaubt sind. Würde man eine differenziertere Semantik erzeugen, würde vermutlich auch die Verbesserung steigen, da sich die Elemente in der n-besten Liste häufig nur durch die Pronomen bzw. Artikel unterscheiden, wie z.B. 1.: 'switch on the lamp' 2.: 'switch on this lamp'. Die relativ gute Spracherkennungsrate, der für die Fusion relevanten Daten, liegt daran, dass viele Spracheingaben im Dialogkontext eine hohe Auftretenswahrscheinlichkeit haben. Durch andere Daten, mit weniger vorhersehbaren Eingaben, wäre ebenfalls eine größere Verbesserung durch die Fusion möglich.

Es konnte gezeigt werden, dass das System C eine Verbesserung gegenüber System B darstellt und eine Fusion auf n-besten Listen Sinn macht. Allerdings stellen Datensammlungen grundsätzlich einen großen Aufwand dar. Durch die nicht sehr große Datenmenge, die für die vorliegende Evaluation verwendet werden konnte, können die vorgestellten Zahlen nur Tendenzen liefern. In der Zukunft wollen wir die Experimente ausweiten und auf neuen Daten evaluieren. Verbleibende Fehler der Fusionskomponente sind auf zu weit auseinander liegende Zeitstempel von Gestik und Sprache (aufgrund von Erkennungsfehlern) zurückzuführen. Die restlichen Fehler aus Tabelle 2 sind reine Erkennungsfehler, welche durch die Fusionskomponente nicht korrigiert werden können (z.B. Fehlerkennung einer Geste). Die zeitliche Korrela-

tion wurde, auf Grund zu geringer Datenmenge, in dieser Arbeit nicht optimiert, soll aber in Zukunft auf einer größeren Datenmenge verbessert werden.

6 Acknowledgements

Wir bedanken uns bei Petra Gieselmann für Unterstützung bei der Datensammlung, Christian Fügen und Kai Nickel für Hilfe bei Sprach- und Gestikerkenner und Carolin Anne Linnenbank für die Transkription der Daten.
Diese Arbeit wurde finanziert durch die DFG im Rahmen des Sonderforschungsbereiches 588 - Humanoide Roboter, und durch die Europäischen Kommission im Rahmen des Projektes CHIL (contract #506909).

Literatur

[1] R. Sharma, V. Pavlovic, and T. Huang, "Toward multimodal human-computer interface," in *Proceedings of the IEEE*, vol. 86, May 1998, pp. 853 – 869.

[2] R. A. Bolt, ""put-that-there": Voice and gesture at the graphics interface," in *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. ACM Press, 1980, pp. 262–270.

[3] *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, 2004.

[4] *Third IEEE International Conference on Humanoid Robots - Humanoids*, Karlsruhe, Germany, 2003.

[5] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "Match: An architecture for multimodal dialogue systems," in *Proceedings of ACL*, 2002.

[6] N. Reithinger, J. Alexandersson, T. Becker, A. Blocher, R. Engel, M. Löckelt, J. Müller, N. Pfleger, P. Poller, M. Streit, and V. Tschernomas, "Smartkom - adaptive and flexible multimodal access to multiple applications," in *Proceedings of ICMI 2003*, Vancouver, B.C. (USA), 2003.

[7] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration - a statistical view," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334–341, 1999. [Online]. Available: citeseer.ist.psu.edu/wu99multimodal.html

[8] J. Eisenstein and C. M. Christoudias, "A salience-based approach to gesture-speech alignment," in *HLT-NAACL*, 2004.

[9] M. Johnston, "Unification-based multimodal parsing," in *COLING-ACL*, 1998, pp. 624–630. [Online]. Available: citeseer.ist.psu.edu/article/johnston98unificationbased.html

[10] I. Wachsmuth, "Communicative rhythm in gesture and speech," in *Gesture-Based Communication in Human-Computer Interaction: International Gesture Workshop, GW'99*, Gif-sur-Yvette, France, March 1999.

[11] S. L. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in *CHI*, 1997, pp. 415–422. [Online]. Available: citeseer.ist.psu.edu/oviatt97integration.html

[12] P. R. Cohen, R. Coulston, and K. Krout, “Multimodal interaction during multiparty dialogues: Initial results,” in *International Conference On Multimodal Interfaces (ICMI)*, 2002.

[13] M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi, “Where is ”it”? event synchronization in gaze-speech input systems,” in *Proceedings of the 5th International Conference on Multimodal Interfaces*. ACM Press, 2003, pp. 151–158.

[14] S. Kettebekov, M. Yeasin, and R. Sharma, “Prosody based co-analysis for continuous recognition of coverbal gestures,” in *International Conference on Multimodal Interfaces (ICMI’02)*, Pittsburgh, USA, 2002.

[15] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith, “Unification-based multimodal integration,” in *Proceedings of the ACL*. Somerset, New Jersey: Association for Computational Linguistics, 1997, pp. 281–288.

[16] P. Gieselmann, C. Fuegen, H. Holzapfel, T. Schaaf, and A. Waibel, “Towards multimodal communication with a household robot,” in *Proceedings of the International Conference on Humanoid Robots*, 2003.

[17] C. Fuegen, H. Holzapfel, and A. Waibel, “Tight coupling of speech recognition and dialog management dialog-context dependent grammar weighting for speech recognition,” in *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[18] T. Asfour, A. Ude, K.Berns, and R. Dillmann, “Control of armar for the realization of anthropomorphic motion patterns,” in *The second IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS 2001)*, 2001, pp. 22–24.

[19] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, “The karlsruhe-verbmobil speech recognition engine,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP-97*, Munich, Germany, 1997.

[20] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, “A one pass- decoder based on polymorphic linguistic context assignment,” in *Proceedings of the ASRU Workshop*, Madonna di Campiglio, Trento, Italy, December 2001.

[21] B. Carpenter, *The Logic of Typed Feature Structures*. Cambridge University Press, 1992.

[22] K. Nickel and R. Stiefelhagen, “Pointing gesture recognition based on 3d-tracking of face, hands and head orientation,” in *Proceedings of the Fifth International Conference on Multimodal Interfaces (ICMI)*, Vancouver, Canada, Nov. 5-7 2003.

[23] H. Holzapfel, K. Nickel, and R. Stiefelhagen, “Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures,” in *Proceedings of the International Conference on Multimodal Interfaces*, 2004.