Multimodale Sprecherlokalisation für Mensch-Roboter-Interaktionen in einer Multi-Personen-Umgebung

Sascha Hohenner, Sebastian Lang, Marcus Kleinehagenbrock, Gernot A. Fink, Franz Kummert

Universität Bielefeld, Technische Fakultät, Postfach 100131, 33501 Bielefeld
{*sascha,slang,mkleineh,gernot,franz*}@*techfak.uni-bielefeld.de*

Zusammenfassung: Bei der natürlichen Mensch-Roboter-Interaktion kommt der Lokalisation eines Sprechers eine wichtige Rolle zu, da vor allem bei mobilen Robotern die Position eines Kommunikationspartners variabel ist. Befinden sich mehrere Personen im Sichtbereich des mobilen Roboters, muss zusätzlich entschieden werden, welche dieser Personen der aktuelle Kommunikationspartner ist und damit die Aufmerksamkeit des mobilen Roboters erhalten sollte. Auch hier kann die akustische Lokalisation als sehr wichtige Informationsquelle dienen. Aus den genannten Gründen wurde unser Gesamtsystem zur Personendetektion und -verfolgung, bei dem bisher nur eine Gesichtserkennung (videobasiert) und eine Beinerkennung (basierend auf Laserdaten) zum Einsatz kam, mit einem Verfahren zur akustische Sprecherlokalisation (*Cross-Powerspectrum Phase-Analysis*) erweitert. Die Signalaufnahme erfolgt dabei durch zwei auf unserem mobilen Roboter montierte Mikrofone. Eine Evaluation der akustischen Lokalisation im 2D-Fall (Sprecher und Mikrofone auf gleicher Höhe) hat gezeigt, dass dieses Verfahren eine sehr genaue und vor allem robuste Schätzung der Sprecherposition liefert. Im 3D-Fall ist eine eindeutige akustische Lokalisation mit nur zwei Mikrofonen jedoch nicht möglich. Allerdings kann durch die Integration in das Gesamtsystem die Sprecherposition unter Verwendung der Informationen aus den anderen Modalitäten eindeutig und robust bestimmt werden, wie eine Evaluation des erweiterten Gesamtsystems gezeigt hat. Außerdem ist es mit Hilfe der akustischen Lokalisation nun auch möglich, die Aufmerksamkeit des mobilen Roboters dynamisch auf den aktuellen Sprecher zu lenken, selbst wenn mehrere Person vor dem Roboter detektiert werden.

1 Einleitung

Die Entwicklung mobiler Roboter für verschiedene Anwendungen (z.B. als Haushaltshilfe, Laborassistent, etc.) ist ein aktuelles Forschungsthema. Um eine breite Akzeptanz solcher Systeme zu erreichen, müssen sie eine entsprechend natürliche Kommunikation ermöglichen. Ein Aspekt ist dabei die Detektion und Lokalisation menschlicher Kommunikationspartner, da es bei mobilen Robotern keine festgelegte Sprecherposition gibt. Erst dadurch wird eine natürliche Mensch-Roboter-Interaktion ermöglicht, da das System zunächst seine Aufmerksamkeit auf eine bestimmte Person lenken muss. Außerdem sollte ein mobiler Roboter in einer Umgebung mit mehreren potentiellen Kommunikationspartnern alle interessanten Personen detektieren und

[1] Die vorliegende Arbeit wurde im Rahmen des Sonderforschungsbereichs 360 "Situierte künstliche Kommunikatoren" sowie der Graduiertenkollegs 256 "Aufgabenorientierte Kommunikation" und 518 "Verhaltensstrategien und Verhaltensoptimierung" von der Deutschen Forschungsgemeinschaft (DFG) gefördert.

In K. Kroschel, editor, Elektronische Sprachsignalverarbeitung, volume 28 of Studientexte zur Sprachkommunikation, pages 162-169, Karlsruhe, 2003.

seine Aufmerksamkeit dynamisch auf den aktuellen Sprecher richten können. Dabei sollte er auch in der Lage sein, eine Entscheidung darüber zu treffen, ob der Sprecher mit ihm oder aber mit einer anderen Person kommunizieren will.
Im Folgenden wird ein entsprechendes Verfahren zur multimodalen Sprecherlokalisation vorgestellt, dass auf der Basis von Sprache sowie Bild- und Laserdaten Personen detektiert und (vor allem auf Basis akustischer Daten) entscheidet, welche dieser Personen der aktuelle Kommunikationspartner ist.

2 Hardware

Abbildung 1 - Unser mobiler Roboter

Die technische Plattform für unsere Software besteht aus einem Pioneer PeopleBot von ActivMedia (siehe Abb. 1) mit einem integrierten PC (Pentium III, 850 MHz), der zur Motoren- und Sensorensteuerung sowie für die Verarbeitung der akustischen Signale eingesetzt wird. Für die Bildverarbeitung wurde ein zusätzlicher PC (Pentium III, 500 MHz) eingebaut. Um akustische Signale erfassen zu können, wurden in einer Höhe von 106 cm zwei AKG-Fernfeldmikrofone auf dem Roboter befestigt (Abstand zwischen den beiden Mikrofonen: 28,1 cm). Zur Erfassung der oberen Körperhälfte und zur Gesichtserkennung möglicher Kommunikationspartner steht eine steuerbare Kamera (Sony EVI-D31) zur Verfügung, die auf dem Roboter in einer Höhe von 141 cm montiert ist. Zusätzlich hat der mobile Roboter einen Laser-Abstandsmesser auf einer Höhe von etwa 30 cm, der nicht nur zur Navigation, sondern auch für die Erkennung von Beinen eingesetzt wird.
Durch diesen Aufbau stehen insgesamt 3 Modalitäten (Sprache, Bild, Laserdaten) für die Lokalisation von Kommunikationspartnern zur Verfügung.

3 Verfahren zur akustischen Lokalisation

Die meisten der aktuellen Arbeiten zur akustischen Lokalisation basieren auf Mikrofonfeldern mit mindestens 3 Mikrofonen. Für die Lokalisation mit nur zwei Mikrofonen gibt es dagegen nur wenige Ansätze. Der Hauptgrund dafür liegt darin, dass eine eindeutige Lokalisation von Schallquellen in einem dreidimensionalen Raum mir nur zwei Mikrofonen i.d.R. nicht möglich ist, sofern nur akustische Daten verwendet werden.
Schnelle und robuste Verfahren stellen die *Generalized Cross-Correlation Method* (GCC) [1] oder *Cross-Powerspectrum Phase Analysis* (CSP, abgeleitet aus der GCC) [2] dar. Beide Verfahren lassen sich bereits für 2 Mikrofone einsetzen. Komplexere Methoden, wie z.B. *Spectral Separation and Measurement Fusion* [3] oder *Linear-Correction Least-Squares* [4], liefern zwar eine meist noch genauere und robustere Positionsschätzung und können darüber hinaus auch die Entfernung und Höhe einer Schallquelle bestimmen, benötigen dazu jedoch mindestens 3 Mikrofone und deutlich mehr Rechenzeit. Außerdem stehen in unserem Szenario auch andere Modalitäten zur Positionsschätzung von Personen zur Verfügung, so dass in unserem Fall für die akustische Lokalisation ein vor allem schneller Algorithmus mit einer hinreichend robusten und genauen Schätzung eingesetzt werden sollte. Deshalb verwenden wir die CSP-Analyse.
Für die Integration mehrerer Modalitäten zur Personenlokalisation gibt es verschiedene Ansätze, vor allem für die audio-visuelle Sprecherverfolgung (siehe z.B. [5, 6, 7]), aber auch

für die Kombination von Bild- und Laserdaten (siehe z.B. [8]). Der von uns gewählte und bereits für die Kombination von Bild- und Laserdaten in [9] präsentierte Ansatz basiert auf dem Verfahren des *Anchoring* [10].

4 Multimodale Sprecherlokalisation

Damit ein mobiler Roboter mit Menschen interagieren kann, muss dieser zunächst eine Person erkennen und seine Aufmerksamkeit auf sie richten, d.h. er muss die Person lokalisieren und eventuell auch verfolgen können. Befinden sich zudem mehrere Personen im Fokusbereich des Roboters, muss dieser außerdem entscheiden, wer der aktuelle Kommunikationspartner ist.

Das zugrunde liegende Verfahren zur Detektion und Verfolgung von Personen auf der Basis von Bild- und Laserdaten wurde bereits in [9] vorgestellt. Daher wird im Folgenden der Schwerpunkt auf die akustische Lokalisation gelegt. Diese dient zum einen als weitere Modalität für die multimodale Lokalisation von Personen. Darüber hinaus kommt der akustischen Lokalisation jedoch eine zweite wichtige Rolle zu: die Detektion des aktuellen Kommunikationspartners und die damit verbundene Steuerung der Aufmerksamkeit.

4.1 Akustische Lokalisation

Für jede Schallquelle in einem Raum gilt allgemein, dass sich die Entfernungen d_1 und d_2 zwischen der Schallquelle s und den beiden Mikrofonen m_1 und m_2 um die Differenz $\Delta d := d_2 - d_1$ unterscheiden. Diese Entfernungsdifferenz resultiert in einer zeitlichen Verzögerung δ bei der Signalaufnahme zwischen dem rechten und dem linken Mikrofon. Ziel der akustischen Lokalisation ist es nun, auf der Basis dieser Laufzeitdifferenz den Winkel der Schallquelle zum mobilen Roboter zu berechnen. Das bedeutet, dass auf der Basis der Signale, die mit dem rechten und dem linken Mikrofon aufgenommen werden, zuerst die zeitliche Verschiebung zwischen diesen beiden Signalen berechnet werden muss.

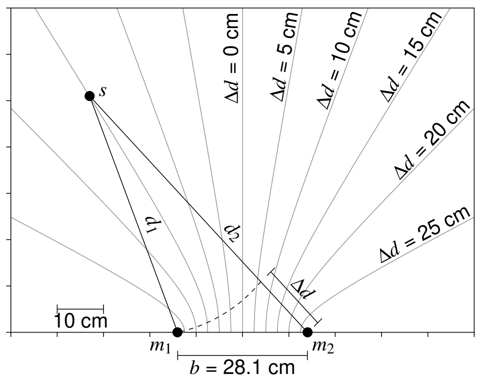

Abbildung 2 - Die Abstände $\boldsymbol{d_1}$ und $\boldsymbol{d_2}$ zwischen der Schallquelle $\boldsymbol{s}$ und den beiden Mikrofonen $\boldsymbol{m_1}$ und $\boldsymbol{m_2}$ differieren um den Betrag $\boldsymbol{\Delta d}$. Alle Schallquellen mit gleichem $\boldsymbol{\Delta d}$ befinden sich auf einem Hyperboloid.

Zur Berechnung dieser Laufzeitdifferenz wird bei dem Verfahren der *Cross-Powerspectrum Phase*-Analyse zunächst eine spektrale Korrelation zwischen dem rechten und dem linken Signal berechnet:

$$K(\tau) = FT^{-1}\left(\frac{\hat{S}_L(f)\,\hat{S}_R^*(f)}{|\hat{S}_L(f)|\,|\hat{S}_R(f)|}\right) \tag{1}$$

wobei $\hat{S}_L(f)$ und $\hat{S}_R(f)$ die Kurzzeitspektren des linken bzw. rechten Signals sind (Fensterlänge 43 ms bei einer Abtastrate von 48 kHz).

Ist nur eine Schallquelle vorhanden, lässt sich nun die Laufzeitdifferenz δ direkt durch das Argument τ mit der höchsten spektralen Korrelation $K(\tau)$ angeben:

$$\delta = \arg\max_{\tau}\; C(\tau) \tag{2}$$

Werden neben dem absoluten Maximum auch lokale Maxima aus Gleichung (2) berücksichtigt, so können sogar mehrere Schallquellen gleichzeitig lokalisiert werden.

Leider lässt sich aus der berechneten Laufzeitdifferenz der Winkel vom mobilen Roboter aus zur Schallquelle nicht eindeutig bestimmen, da die Laufzeitdifferenz nicht nur vom Winkel zwischen Schallquelle und Mikrofonen, sondern auch von der Höhe und der Entfernung der Schallquelle abhängig ist. Es gibt also beliebig viele Positionen mit unterschiedlichen Winkeln, die in der gleichen Laufzeitdifferenz resultieren, wobei alle Positionen auf einem Hyperboloid liegen (siehe Abb. 2). Um eine eindeutige Positionsbestimmung im 3D-Fall nur mit akustischen Daten vornehmen zu können, sind also mehr als zwei Mikrofone notwendig.
Unter bestimmten Annahmen kann der Winkel jedoch auch mit nur zwei Mikrofonen eindeutig bestimmt werden:

1. Schallquelle und Mikrofone befinden sich in einer Ebene.
2. Der Abstand der Schallquelle zu den Mikrofonen ist deutlich größer als der Abstand zwischen den Mikrofonen selbst.

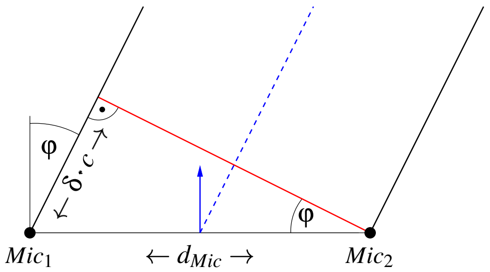

Abbildung 3 - Vereinfachte Geometrie zur Berechnung des Winkels φ zwischen der Schallquelle und den Mikrofonen Mic_1 und Mic_2.

Unter diesen Annahmen lässt sich der Winkel φ zwischen der Schallquelle und den Mikrofonen direkt aus der geschätzten Laufzeitdifferenz δ wie folgt berechnen (siehe dazu Abb. 3):

$$\varphi = \arcsin \frac{\delta \cdot c}{d_{Mic}} \tag{3}$$

wobei c die Schallgeschwindigkeit und d_{Mic} der Abstand zwischen den beiden Mikrofonen ist. Diese Gleichung wird für die Evaluation der akustischen Lokalisation in Abschnitt 5.1 verwendet.
Die genannten Voraussetzungen sind in unserem Szenario i.d.R. nicht gegeben. Allerdings stehen andere Modalitäten zur Lokalisation von Personen zur Verfügung, so dass die Schätzung der Laufzeitdifferenz in Kombination mit weiteren Personendaten eine robuste und genaue Sprecherlokalisation ergibt. Die genaue Verwendung der geschätzten Laufzeitdifferenz für die multimodalen Sprecherlokalisation wird in Abschnitt 4.4 beschrieben.

4.2 Gesichtserkennung

Für die Erkennung von frontalen Gesichtern verwenden wir das Verfahren von Viola und Jones [11]. Dabei werden im Grauwertbild der Kamera quadratische Bildausschnitte verschiedener Größen an allen möglichen Positionen als Gesichter oder Nicht-Gesichter klassifiziert. Trotz der großen sich ergebenden Anzahl von Bildausschnitten (in unserer Anwendung über 40.000 bei einer Bildgröße von 256×192) können mehrere Bilder pro Sekunde verarbeitet werden. Die hohe Effizienz des Verfahrens ergibt sich zum einen aus den einfach zu berechnenden Merkmalen und zum anderen aus der Verwendung einer Klassifikatorhierarchie, in der Einzelklassifikatoren aufsteigender Komplexität hintereinander geschaltet sind.

4.3 Detektion von Beinpaaren

Jede Messung des Laser-Abstandsmessers liefert ein Tiefenprofil innerhalb einer horizontalen Ebene auf der Höhe von Beinen. Dabei wird ein Bereich mit einem Öffnungswinkel von 180° vor dem Roboter erfasst. Die Winkelauflösung beträgt 0,5°, was in 361 Abstandswerten pro Messung resultiert. In solch einem Tiefenprofil ergeben Beinpaare ein charakteristisches Muster, welches wie folgt detektiert werden kann: Als erstes werden aufeinander folgende Messpunkte, die ähnliche Abstandswerte aufweisen, zu Segmenten zusammengefasst.

Segmentgrenzen zeichnen sich somit durch große Sprünge im Tiefenprofil aus und lassen auf Objektgrenzen schließen. Beine werden in der Regel durch genau ein Segment erfasst. Nach der Segmentierung werden die Segmente abhängig von einem Satz von Merkmalen als Beine oder Nicht-Beine klassifiziert (Details in [9]). Im letzten Schritt werden erkannte Beine, abhängig von ihrem Abstand, zu Beinpaaren gruppiert. Das Verfahren ist schnell und robust und liefert wesentliche Information über den Standort von Personen.

4.4 Integration der Modalitäten

Die akustische Lokalisation, die Gesichtserkennung und die Detektion von Beinen ist in drei voneinander unabhängigen Prozessen realisiert. Sie liefern asynchron Personendaten, die integriert werden müssen. Zu diesem Zweck verwenden wir das von Fritsch et al. entwickelte *Multimodal Anchoring* [9], welches das *Standard-Anchoring* [10] auf mehrere Modalitäten erweitert. *Anchoring* formalisiert den Aufbau von Verknüpfungen zwischen Symbolen und Sensordaten, die sich auf ein und dasselbe physikalische Objekt beziehen. Diese Verknüpfungen (genannt *Anchors*) müssen dynamisch sein, da ein Symbol fortlaufend den aktuellen sensorischen Beobachtungen (genannt *Perzepte*) des entsprechenden Objektes zugeordnet werden muss.

Für die Sprecherlokalisation repräsentieren wir jede Person durch die drei Symbole *Sprache*, *Gesicht* und *Beinpaar*. Ein Kompositionsmodell beschreibt dabei die räumlichen Relationen der einzelnen Komponenten und garantiert, dass Perzepte nur geeigneten Personen zugeordnet werden. Eine Person kann nur über Gesichts- oder Beinpaarperzepte aufgebaut werden, nicht aber über Sprachperzepte, da diese mit zwei Mikrofonen ohne zusätzliche Information nicht eindeutig in 3D lokalisiert werden können: Alle Sprachquellen, die in dem selben Laufzeitunterschied resultieren, liegen auf einem Hyperboloid (siehe Abschnitt 4.1).

Um zu entscheiden, ob ein Sprachperzept einer Person zugeordnet werden kann, für die bereits eine Position im Raum (Abstand, Höhe und Winkel im Roboter-Koordinatensystem) über Gesicht und Beine ermittelt wurde, wird angenommen, dass das Sprachperzept den selben Abstand und die selbe Höhe wie der Mund der betrachteten Person hat. Mit diesen Informationen (Abstand und Höhe) sowie der von der akustischen Lokalisation gelieferten Laufzeitdifferenz lässt sich ein hypothetischer Winkel des Sprachperzepts errechnen. Ist die Differenz zwischen Personenwinkel und hypothetischem Sprachperzeptwinkel klein (also mit dem Kompositionsmodell vereinbar), wird das Sprachperzept dem entsprechenden Anchor der Person zugeordnet.

5 Evaluation

Im Folgenden wird zunächst die Evaluation der akustischen Sprecherlokalisation beschrieben. Anschließend erfolgt die Evaluation der um die akustische Komponente erweiterten multimodalen Sprecherlokalisation.

5.1 Akustische Lokalisation

Für die Evaluation der akustischen Sprecherlokalisation wurde folgender Aufbau gewählt: Fünf verschiedene Testpersonen wurde nacheinander jeweils auf 12 verschiedenen Positionen vor dem Roboter positioniert (bei ca. $0°$, $10°$, $20°$, $40°$, $60°$ sowie $80°$ mit einer Entfernung von 1 bzw. 2 Metern). Dabei wurde der Roboter so aufgestellt, dass sich der Mund der Testperson und die beiden Mikrofone in etwa auf derselben Höhe befanden (um Gleichung (3) zur Winkelberechnung anwenden zu können). Auf jeder Position wurde von den Testpersonen ein Satz (Sprechdauer etwa 8 Sekunden) vorgelesen. Während der Sprechdauer wurde ca. alle 50 ms die Laufzeitdifferenz mit der CSP-Analyse berechnet und darauf basierend eine Winkelberechnung mit Gleichung (3) vorgenommen.

Basierend auf den so geschätzten Winkeln wurde für jeden Sprecher auf jeder Position ein mittlerer Winkel und eine Varianz berechnet. Da es sehr schwierig war, die Testperson exakt auf dem gewünschten Winkel zu positionieren (schon ein leichtes Neigen der Testperson führt zu einer Abweichung vom angestrebten Winkel), wurden die geschätzten Winkel anstelle der angestrebten Winkel zur Berechnungen der Varianzen verwendet. Nach den separaten Berechnungen für jeden Sprecher wurden der mittlere Winkel und die Varianz noch einmal für jede Position über alle Sprecher gemittelt.

	Winkel					
Entfernung	0°	10°	20°	40°	60°	80°
1 Meter	-0.9° (0.56)	9.1° (0.34)	18.9° (0.21)	38.2° (0.50)	57.7° (0.40)	74.0° (2.62)
2 Meter	-0.3° (0.81)	9.2° (0.37)	19.3° (0.27)	38.8° (0.22)	57.5° (0.64)	73.3° (2.18)

Tabelle 1 - Ergebnisse der akustischen Sprecherlokalisation bei einer Entfernung von 1 bzw. 2 Metern: Geschätzte Sprecherpositionen und (in Klammern) Varianz der Schätzungen.

In Tabelle 1 sind die Ergebnisse der Evaluation aufgeführt. Zunächst fällt auf, dass der mittlere geschätzte Winkel vor allem bei kleineren Winkeln ($0°$ bis $40°$) fast konstant um $1°$ vom angestrebten Winkel abweicht. Der Grund dafür liegt höchst wahrscheinlich darin, dass der Roboter bei der Durchführung der Evaluation nicht ganz korrekt aufgestellt wurde. Unter dieser Annahme liefert die akustische Lokalisation für Winkel bis etwa 60° eine sehr genaue und konstante Schätzung der Sprecherposition (d.h. berechnete und tatsächliche Laufzeitdifferenz stimmen nahezu überein): Der mittlere geschätzte Winkel ist nahezu identisch mit dem angestrebten Winkel bei einer niedrigen Varianz. Außerdem arbeitet die Lokalisation für eine Entfernung sowohl von 1 als auch 2 Metern etwa gleich gut. Erst ab einem Winkel von $60°$ besteht eine deutlichere Differenz zwischen den berechneten und den angenommenen Winkeln, wobei die Konstanz der Winkelschätzung erst bei einem Winkel von $80°$ deutlich nachlässt (die Varianz steigt deutlich an). Der Hauptgrund für die nachlassende Genauigkeit dürfte dabei in der leichten Richtcharakteristik der beiden Mikrofone liegen.

Insgesamt hat diese Evaluation gezeigt, dass die Schätzung der Laufzeitdifferenz (mit der darauf basierenden Winkelberechnung im 2D-Fall) sehr genaue und konstante Ergebnisse liefert. Somit ist dieses Verfahren dazu geeignet, wichtige Informationen im Kontext einer multimodalen Sprecherlokalisation zu liefern. Dies wird auch durch die folgende Evaluation des Gesamtsystems gezeigt.

5.2 Aufmerksamkeitssteuerung

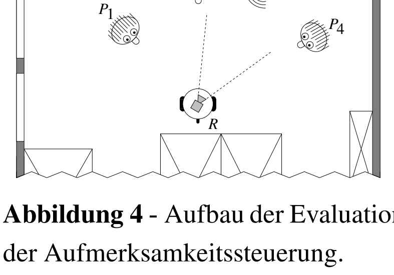

Abbildung 4 - Aufbau der Evaluation der Aufmerksamkeitssteuerung.

Für die Evaluation der Aufmerksamkeitssteuerung wurden jeweils vier Personen (P_1 bis P_4) auf vordefinierte Positionen vor dem Roboter (R) positioniert, so dass verschiedene Winkel und Abstände berücksichtigt wurden (Abb. 4). Die Personen hatten die Aufgabe, in einer festgelegten Reihenfolge für jeweils 10 Sekunden zu sprechen, jede Person drei Mal. Dabei sprach jede Person beim ersten Mal zum Roboter, beim zweiten Mal zu einer der anderen Personen und beim dritten Mal wieder zum Roboter.

Es wurden keine Einschränkung gemacht, wie die Testpersonen zu stehen haben. Damit ergab sich implizit, dass, im Gegensatz zu dem in Abschnitt 5.1 beschriebenen Experiment, der Mund nicht auf der Höhe der Mikrofone positioniert war. Das beschriebene Experiment wurde drei Mal durchgeführt, wobei insgesamt neun verschiedene Testpersonen teilgenommen haben.

In dem Experiment wurde die Richtung der Aufmerksamkeit des Roboters über die Zeit aufgezeichnet. Abbildung 5 zeigt den Verlauf der Aufmerksamkeit für die drei Durchführungen des Experimentes. Es ist zu erkennen, dass der Roboter immer in der Lage war, seine Aufmerksamkeit auf den aktuellen Sprecher zu richten. Allerdings wurde in einigen Fällen die Aufmerksamkeit zu lange aufrecht erhalten (Zeitschritt 8, 1. und 3. Experiment), oder die Aufmerksamkeit wechselte zwischendurch fälschlicherweise zu einer nicht sprechenden Person (Zeitschritt 4, 3. Experiment; Zeitschritt 5, alle Experimente). Es ist hierbei zu beachten, dass in jedem dieser fehlerbehafteten Fälle Person P_2 involviert war. Der Grund dafür ist, dass in Richtung von $0°$ Sprache lokalisiert wurde, obwohl die dort positionierte Person P_2 nicht gesprochen hatte. Ursache hierfür sind die vom Roboter selbst erzeugten Geräusche, die als Geräuschquelle in entsprechender Richtung gedeutet werden. Fünf der sechs beschriebenen Fehler sind während der Zeitschritte 5 bis 8 aufgetreten. In dieser Phase adressierten die Testpersonen nicht den Roboter sondern eine der anderen Personen, wodurch der Lautstärkepegel des Sprachsignals geringer war. Das beobachtete Problem lässt sich beheben, indem zusätzlich eine Sprechaktivitäts-Detektion durchgeführt wird. Dies ist Teil unserer weiteren Arbeit.

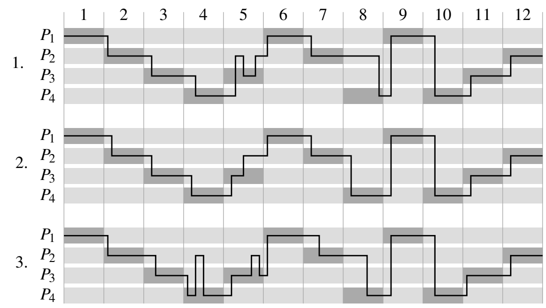

Abbildung 5 - Diagramm für die drei Experimente zur Aufmerksamkeitssteuerung. Jeder Person ist eine Spur zugeordnet (hellgrau), die dunkelgrau gefärbt ist, während die Person spricht. Die durchgezogene Linie gibt an, welche Person sich im Aufmerksamkeitsfokus des Roboters befindet.

Wie in Abbildung 5 zu sehen ist, ist jeder Aufmerksamkeitswechsel um etwa 2 Sekunden verzögert. Dies ist im *Multimodal Anchoring* begründet: Der *Anchor* für Sprache bleibt nach dem letzten Sprachperzept noch 2 Sekunden erhalten, bevor er entfernt wird, um mögliche Sprechpausen berücksichtigen zu können. Erst danach kann eine andere Person die Aufmerksamkeit erlangen. Für das Anchoring wurden in dem Experiment folgende Werte ermittelt: Die Sprecherlokalisation erfolgte mit einer Rate von 5,5 Hz und die Gesichtsdetektion mit 9,6 Hz. Der Laser-Abstandsmessers stellte Daten mit einer Frequenz von 4,7 Hz zur Verfügung, während die zur Beinpaar-Erkennung benötigte Zeit vernachlässigbar ist. Den Personen, die aktuell im Fokus der Aufmerksamkeit waren, konnten im Mittel 69,5% der generierten Sprachperzepte, 71,4% der Gesichtsperzepte und 99,9% der Beinperzepte zugeordnet werden. Über die Gesichtsperzepte konnten die Körpergrößen der Testpersonen mit einer Genauigkeit von ± 5 cm gemessen werden, was ausreichend, war um Sprache in 3D genau genug zu lokalisieren. Das Experiment hat gezeigt, dass die Aufmerksamkeitssteuerung unter Verwendung von nur zwei Mikrofonen im Rahmen des multimodalen Anchorings sehr gut funktioniert und eine geeignete Basis für weitergehende Forschung im Bereich natürlicher Mensch-Roboter-Interaktion auf mobilen Robotern darstellt.

6 Zusammenfassung

In diesem Beitrag wurde die Erweiterung unseres Systems zur multimodalen Personendetektion und -verfolgung für unseren mobilen Roboter um ein Verfahren zur akustischen Lokalisation

(*Cross-Powerspectrum Phase-Analysis*) vorgestellt. Eine Evaluation dieses Verfahrens hat gezeigt, dass damit eine sehr genaue und robuste Schätzung der Sprecherposition mit nur zwei Mikrofonen im 2D-Fall ermöglicht wird.

Das Problem der nicht eindeutigen akustischen Lokalisation im 3D-Fall mit nur zwei Mikrofonen wurde durch die Integration in das Gesamtsystem gelöst. Die akustische Lokalisation liefert dabei unter Verwendung anderer Informationen (Detektion von Beinpaaren und Gesichtserkennung) wichtige Informationen für eine eindeutige und robuste Bestimmung der Sprecherposition. Außerdem hat eine Evaluation gezeigt, dass unser System mit Hilfe der akustischen Lokalisation nun auch in der Lage ist, die Aufmerksamkeit eines mobilen Roboters dynamisch auf den aktuellen Sprecher zu lenken, selbst wenn mehrere Personen vom Roboter detektiert werden.

Literatur

[1] C. Knapp, G. Carter: *The Generalized Correlation Method for Estimation of Time Delay*, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Bd. ASSP-24, Nr. 4, 1976, S. 320–327.

[2] D. Giuliani, M. Omologo, P. Svaizer: *Talker Localization and Speech Recognition Using a Microphone Array and a Cross-Powerspectrum Phase Analysis*, in *Proc. Int. Conf. on Spoken Language Processing*, Bd. 3, 1994, S. 1243–1246.

[3] B. Berdugo, J. Rosenhouse, H. Azhari: *Speakers' direction finding using estimated time delays in the frequency domain*, *Signal Processing*, Bd. 82, 2002, S. 19–30.

[4] Y. Huang, J. Benesty, G. Elko, R. Mersereau: *Real-Time Passiv Source Localization: A Practical Linear-Correction Least-Square Approach*, *IEEE Trans. on Speech and Audio Processing*, Bd. 9, Nr. 8, 2001, S. 943–956.

[5] H. Okuno, K. Nakadai, H. Kitano: *Social Interaction of Humanoid Robot Based on Audio-Visual Tracking*, in *Proc. Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 2002.

[6] J. Vermaak, A. Blake, M. Gangnet, P. Perez: *Sequential Monte Carlo fusion of sound and vision for speaker tracking*, in *Proc. Int. Conf. on Computer Vision*, Bd. 1, 2001, S. 741–746.

[7] Y. Matsusaka, S. Fujie, T. Kobayashi: *Modeling of Conversational Strategy for the Robot Participating in the Group Conversation*, in *Proc. European Conf. on Speech Communication and Technology*, 2001, S. 2173–2176.

[8] S. Feyrer, A. Zell: *Robust Real-Time Pursuit of Persons with a Mobile Robot Using Multisensor Fusion*, in *Proc. Int. Conf. on Intelligent Autonomous Systems*, 2000, S. 710–715.

[9] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, G. Sagerer: *Multi-Modal Anchoring for Human-Robot-Interaction*, *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, Bd. 43, Nr. 2–3, 2003, S. 133–147.

[10] S. Coradeschi, A. Saffiotti: *Perceptual anchoring of symbols for action*, in *Proc. Int. Conf. on Artificial Intelligence*, 2001, S. 407–412.

[11] P. Viola, M. Jones: *Robust Real-time Object Detection*, in *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, 2001.