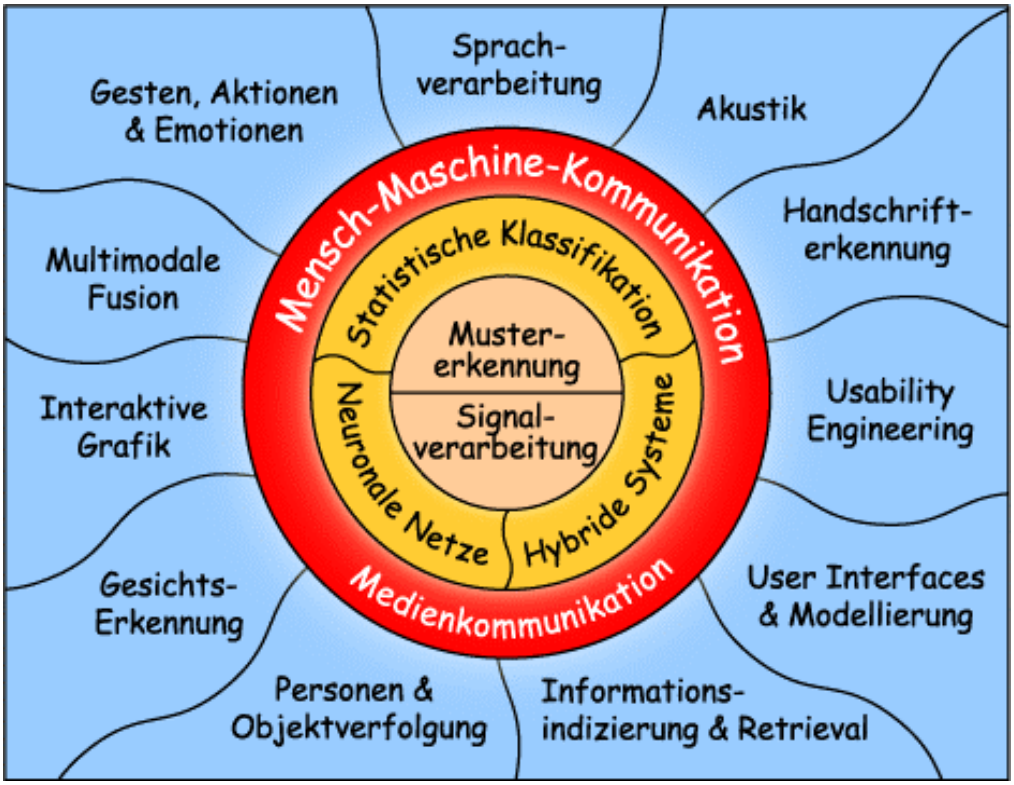
Multimodale Mensch-Maschine-Kommunikation in München: Stand der Forschung und Ausblick auf zukünftige Entwicklungen

Gerhard Rigoll

Lehrstuhl für Mensch-Maschine-Kommunikation, TU München, Arcisstr. 21, D-80333 München
Email: rigoll@tum.de



1. Einleitung

Ausgehend von den Anfängen der sechziger Jahre, hat der damalige Lehrstuhl für Elektroakustik, basierend auf seinen Kernkompetenzen, immer wieder neue Forschungsgebiete erschlossen, die sich in einem direkten Umfeld der Elektroakustik und Psychoakustik befanden, ohne dabei die Forschungskompetenzen in dem ursprünglichen Arbeitsgebiet aufzugeben. Da man sich in der Psychoakustik intensiv mit dem "System Mensch" beschäftigt und hierfür moderne Methoden der Signalverarbeitung zum Einsatz kommen, lag es nahe, diese Methoden auch auf andere Aspekte der menschlichen Kommunikation und Sinneswahrnehmung anzuwenden und so wurde im Verlauf der letzten 30 Jahre erfolgreich der Weg von der Psychoakustik zur multimodalen Mensch-Maschine-Kommunikation beschritten, bei der das Gebiet "Hören" immer noch eine der wichtigsten Modalitäten repräsentiert. Der heutige Lehrstuhl für Mensch-Maschine-Kommunikation versteht sich als ein Forschungsinstitut, das ein möglichst breites Methodenspektrum in dem sehr weiträumigen und vielfältigen Gebiet der multimodalen Mensch-Maschine-Kommunikation abdecken möchte, ohne dabei den Blick in die wissenschaftliche Tiefe zu verlieren, die notwendig ist, um entscheidende methodische Fortschritte auf diesem Gebiet zu erzielen.

Das obenstehende Bild zeigt dabei die momentan bearbeiteten Forschungsgebiete am Lehrstuhl.

2. Stand der Technik bei der Mensch-Maschine-Kommunikation

Bei einem Rückblick auf die wesentlichen Fortschritte der Mensch-Maschine-Kommunikation in den letzten 20-30 Jahren muss sicherlich die in diesem Zeitraum stattgefundene enorme Entwicklung der automatischen Spracherkennung erwähnt werden, die wiederum auf dem erfolgreichen Einsatz der stochastischen Modellierung auf der Basis der Hidden-Markov-Modelle (HMM) zurückzuführen ist. Gerade in diesem Zeitabschnitt konnte die Überlegenheit dieser statistischen Technik gegenüber den anderen, traditionelleren Techniken, die z.B. auf regel- und wissensbasierten Verfahren beruhten, eindrucksvoll demonstriert werden. Die entscheidenden Vorteile der HMM-Technik liegen dabei in den effizienten Modellierungsmöglichkeiten von Sprach-Untereinheiten, den selbstorganisierenden automatischen Lernverfahren, den eleganten Möglichkeiten zur Einbindung statistischer Sprachmodelle auf höheren semantischen Ebenen und insbesondere der Fähigkeit der HMMs zur gleichzeitigen Segmentierung und Erkennung von zusammenhängenden Teilmustern.

Interessanterweise sind genau diese Eigenschaften auch die entscheidenden Ursachen dafür, dass die Technik der stochastischen Modellierung von Mustern sich auch in den letzten Jahren bei der Realisierung von vielen anderen Modalitäten der Mensch-Maschine-Interaktion durchsetzen konnte /1/. Dies wurde zunächst erfolgreich demonstriert am Beispiel der automatischen Handschrifterkennung, bei der die HMM-Technik zu Beginn der 90er Jahre erstmalig eingesetzt wurde und sich dann zügig in diesem Forschungsgebiet etabliert hat /2/. Dies liegt nicht zuletzt daran, dass die Gegebenheiten und Randbedingungen im Bereich der Handschrifterkennung denen der automatischen Spracherkennung sehr ähnlich sind und sich daher wieder genau die beschriebenen Vorteile der HMMs durchsetzen konnten. Auch hier erfolgt wiederum eine Modellierung der Wörter durch Buchstaben aus entsprechenden Untereinheiten und die gleichzeitige Erkennung und Segmentierung von zusammenhängenden Teilmustern ist von großer Bedeutung. Auf der höheren semantischen Ebene lassen sich auch hier praktisch dieselben Sprachmodelle wie bei der Spracherkennung einsetzen. Es konnte in den darauffolgenden Jahren gezeigt werden, dass die HMM-Technik nicht nur in der Handschrifterkennung, sondern auch generell im Bereich des „Pen-Computing“, bei dem die intuitive Eingabe über einen Schreibstift erfolgt, sehr erfolgreich eingesetzt werden kann, etwa bei der Erkennung von Piktogrammen oder handgeschriebenen Skizzen, die beispielsweise höherwertige semantische Symbolfolgen oder vordefinierte „Shortcuts“ innerhalb von Dialogen darstellen können.

Etwas erstaunlicher ist es da schon, dass auch im Bereich der rein visuellen Kommunikation zwischen Mensch und Maschine die stochastische Modellierungstechnik ähnliche Erfolge vorweisen konnte. Diese Entwicklung hat noch einige Jahre später als bei der Handschrifterkennung eingesetzt und sich insbesondere bei der Gestikerkennung als ausgesprochen erfolgreich bewährt. Entscheidend war dabei die Tatsache, dass – weniger bei der klassischen Bildverarbeitung – aber eher bei der visuellen Mensch-Maschine-Kommunikation der Faktor „Zeit“ eine wesentliche Rolle spielt und sich beispielsweise bei der Erkennung dynamischer Gesten in der Notwendigkeit zur Verarbeitung von Videosignalen widerspiegelt. Dadurch, dass es gelungen ist, jede Einzelaufnahme einer Gestik in einen stark reduzierten Merkmalsvektor zu überführen, konnte der Einsatz der stochastischen Sequenzmodellierung relativ reibungslos auf die Gestenerkennung ausgeweitet werden. Heutzutage setzen die meisten Systeme zur Gestenerkennung diese Technologie ein, die auch noch auf die Erkennung von Aktionen und Aktivitäten von Benutzern in intelligenten Umgebungen ausgeweitet werden konnte, beispielsweise bei der Bewegungs- und Verhaltenserkennung in Besprechungen.

Letztendlich hat dann die stochastische Modellierungstechnik auch eine der letzten Domänen der klassischen Bildverarbeitung erobert, indem es gelang, durch zweidimensionale Varianten der zunächst nur eindimensionalen Hidden-Markov-Modelle auch statische Einzelbilder zu verarbeiten, bei denen es insbesondere auf Möglichkeiten zur räumlichen Bildverzerrung angekommen ist, wie beispielsweise bei der Gesichtserkennung. Die Herleitung von effizienten Trainings- und Dekodiertechniken, auch für diesen zweidimensionalen Fall, war sicherlich eine der entscheiden Voraussetzungen hierfür. So hat sich der Bereich der Gesichtserkennung als ein weiteres interessantes Betätigungsfeld der Mensch-Maschine-Kommunikation etabliert, mit dem Hintergrund der Zugangsberechtigung und Personalisierung von Benutzerschnittstellen über die Erkennung des aktuellen Benutzers. Die HMM-Technik ist hierbei in den letzten Jahren zu einem ernsthaften Konkurrenten der ursprünglichen dort angewandten Technik geworden (s. /3/), beispielsweise der elastischen Graphen und der Eigenfaces.

3. Ausblick auf zukünftige Entwicklungen

Eines der bisher noch am wenigsten gelösten Probleme der Mensch-Maschine-Kommunikation ist die multimodale Verarbeitung verschiedener synchroner oder asynchroner Eingabekanäle, beispielsweise bei der Kombination von Sprache und Gestik oder Sprache und Zeigegesten bzw. Touchscreen. Dort werden bisher meistens relativ einfache regelbasierte Ansätze verwendet, mit den bekannten Nachteilen dieser Methoden, insbesondere bei der Behandlung von nicht berücksichtigten Kombinationsmöglichkeiten. Hier wäre eine algorithmische Lösung, die auf Methoden des maschinellen Lernens basiert, deutlich wünschenswerter. Auch hier bieten sich wiederum statistische Mustererkennungsverfahren an, im einfachsten Fall beispielsweise sog. Multistream-HMMs, welche die verschiedenen Eingabekanäle in unterschiedlichen Vektorsequenzen verarbeiten, die sich dann einfach mit separaten Gaußmodellen und individuell einstellbaren Gewichtungen dieser Gauß-Parameter zu Emissionswahrscheinlichkeiten der HMMs berechnen lassen können. Komplexere Strukturen, auch geeignet für asynchrone Datenströme und Mischungen aus numerischen Merkmalen und semantischen Merkmalen sind hier momentan Gegenstand der aktuellen Forschungen und werden in den nächsten Jahren zu den erwarteten Durchbrüchen bei der echten multimodalen Mensch-Maschine-Kommunikation entscheidend beitragen können.

/1/ Rigoll, G.; Müller, S.: Statistical Pattern Recognition Techniques for Multimodal Human Computer Interaction and Multimedia Information Processing. Survey Paper, Int. Workshop "Speech and Computer", Moscow, Russia, October 1999, pp. 60-69.

/2/ Rigoll, G.; Kosmala, A.; Willett, D.: A Systematic Comparison of Advanced Modeling Techniques for Very Large Vocabulary On-Line Cursive Handwriting Recognition. In Seong-Whan Lee (Editor), Advances in Handwriting Recognition, Chapter 2, pp. 69-78. World Scientific, 1999.

/3/ Eickeler, S.; Müller, S.; Rigoll, G.: Recognition of JPEG Compressed Face Images Based on Statistical Methods. Image and Vision Computing Journal, Special Issue on Facial Image Analysis, 18(4):279-287, March 2000.