Sprechen hören und sehen

Sascha Fagel, FG Kommuniaktionswissenschaft

Die Fähigkeit zu sprechen ist eine, wenn nicht sogar die Fähigkeit, die den Menschen von den Tieren unterscheidet. Mit der Sprache haben sich auch geistige Fähigkeiten entwickelt und entwickeln sich immer neu, wenn ein Mensch sprechen lernt. Heutzutage wird die Welt vom Fluss der Informationen geprägt. Während für Computer immer schnellere Netzwerke entwickelt werden und immer effizientere Datenspeicherung, ist und bleibt für uns Menschen die sprachliche Kommunikation die wichtigste Art des Informationstransfers. Trotz der seit Jahrzehnten rasant verlaufenden Entwicklung im Bereich der Mikroelektronik haben wir dem Computer unsere Art der Kommunikation noch nicht beibringen können. Sätze wie: „Computer! Zeig mir die Kometenannäherungen der letzten zehn Jahre!“ gehören immer noch in die Welt der Science Fiction. Wir Menschen müssen für den Umgang mit der Maschine neue Kommunikationsformen erlernen. Aber oft genug versteht der Computer trotz intensiver Bemühungen doch nicht, was wir von ihm wollen. Und ebenso ist das, was der Computer ausgibt, für uns nicht selten unverständlich.

Wie viel Information selbst in kürzesten Sprachäußerungen enthalten sein kann, wird am Beispiel der lapidaren Antwort an der Gegensprechanlage deutlich: "Ich bin's." Wir erfahren nicht nur - falls sie uns bekannt ist - die Identität der Person, obwohl sie uns ihren Namen gar nicht nennt. Wir erfahren möglicherweise auch, in welchem emotionalen Zustand sie sich befindet, ob sie außer Atem ist oder vielleicht erkältet. An der Stimme eines Menschen können wir noch viel mehr erkennen. Beispielsweise ob er männlich oder (sie) weiblich ist, ob er jung oder alt ist oder woher er kommt. Die Kommunikationswissenschaft bemüht sich herauszufinden, welche Eigenschaften der Stimme und der Sprechweise verantwortlich dafür sind, was wir an Informationen herauslesen können. Kennen wir diese Eigenschaften, können wir sie einerseits nutzen, um einem Computer beizubringen, uns besser zu verstehen. Andererseits können wir die Sprachausgabe eines Computers ähnlich reichhaltig gestalten wie die natürliche Sprache.

Aber Sprechen ist noch mehr. Beim Sprechen bewegen wir z.T. sichtbar unsere Sprechwerkzeuge, die Artikulatoren. Diese visuelle Information nutzen wir täglich, um unser Gegenüber besser zu verstehen. Ist der akustische Kanal bei der mündlichen Kommunikation gestört, hilft uns der optische Kanal, die fehlende Information zu ergänzen (audio: http://fourier.kgw.tu-berlin.de/125Jahrfeier/Verstaendlichkeit.mp3, audiovisuell: http://fourier.kgw.tu-berlin.de/125Jahrfeier/Verstaendlichkeit.wmv). Dabei ist äußerst hilfreich, dass gerade diejenigen Laute, die sich akustisch sehr ähnlich und deshalb leicht zu verwechseln sind (z.B. [m], [n] oder [f], [s]), sich optisch besonders deutlich voneinander unterscheiden.

Am Institut für Sprache und Kommunikation haben wir ein audiovisuelles Sprachsynthesesystem zur künstlichen Erzeugung hör- und sichtbarer Sprachäußerungen entwickelt. Es kann derzeit sechs artikulatorische Parameter visualisieren (Demo: http://fourier.kgw.tu-berlin.de/Displacer-Slider-Demo/woman.wrz). Die Steuerung des Artikulation des Sprachsynthesesystems erfolgt über ein Modell der Artikulation, das von den Sprechbewegungen einer Sprecherin abgeleitet ist (Sprecherin: http://fourier.kgw.tu-berlin.de/125Jahrfeier/Caroline.jpg, Messaufbau: http://fourier.kgw.tu-berlin.de/125Jahrfeier/Experiment.jpg,). Die Sprecherin hat bestimmte Äußerungen gesprochen und dabei wurden die Positionen der Empfängerspulen aufgezeichnet. Wichtig ist dabei, dass man so die Stellung der Artikulatoren auch für Äußerungen vorhersagen kann, die nicht aufgezeichnet wurden. So kann das Sprachsynthesesystem audiovisuelle Sprachausgabe aus einem beliebigen Text erzeugen.

Welcher Art die Visualisierung ist, hängt dabei nicht von der Steuerung der Artikulation ab. In einem Fall handelt es sich um einen dreidimensionalen künstlichen Kopf. Man kann z.B. auch in einem Foto bestimmte Merkmalspunkte definieren und diese dann entsprechend der Artikulationssteuerung verschieben, also das Bild deformieren (Merkmalspunkte: http://fourier.kgw.tu-berlin.de/125Jahrfeier/Merkmalspunkte.jpg, Online-Sprachsynthese: http://fourier.kgw.tu-berlin.de).

Die Vorteile sichtbarer natürlicher Sprache zur Verbesserung der Sprachverständlichkeit lassen sich auch bei synthetischer Sprache nutzen. Wir haben experimentell nachgewiesen, dass die visuelle Verständlichkeit des Synthesesystems nicht wesentlich von der natürlicher Sprache abweicht. Wir können mit dem audiovisuellen Sprachsynthesesystem aber auch einiges, was mit natürlicher Sprache nicht möglich ist. Wir können z.B. Artikulation sichtbar machen, die normalerweise nicht sichtbar ist (Demo: http://fourier.kgw.tu-berlin.de/Displacer-Slider-Demo/woman-transparent.wrz. Als zukünftige Anwendung bietet sich beispielsweise ein Sprechtrainer an.

Die audiovisuelle Sprachwahrnehmung ist ein sehr komplexer Prozess. Gesichert ist die Erkenntnis, dass optische Reize die Hörwahrnehmung beeinflussen, wir also tatsächlich etwas anderes hören, wenn visuelle Information hinzukommt (Demo: http://fourier.kgw.tu-berlin.de/125Jahrfeier/McGurk125.wmv). Die umgekehrte Beeinflussung der visuellen Wahrnehmung durch akustische Reize untersuchen wir z.Z. in einem Experiment.

Doch welche weiteren Informationen außer der rein sprachlichen enthält hör- und sichtbare gesprochene Sprache? Unser emotionaler Zustand z.B. äußert sich sowohl in unserer Stimme als auch in unserem Gesicht. Ebenso wie beim reinen Sprachverstehen ergänzen sich hier der optische und der akustische Kanal. Wir erhalten also linguistische und paralinguistische auditive und visuelle Informationen, die wir gemeinsam verarbeiten.

	linguistisch	paralinguistisch
auditiv	„Ich denke an A...elie.“	Er hört sich traurig an.
visuell	„I... de....e an Amelie.“	Er sieht nachdenklich aus.

Dabei gibt es sich überschneidende aber auch exklusive Informationen. Man kann beispielsweise ein Lächeln sehen und auch in der Stimme hören. Stirnrunzeln hingegen ist nur sichtbar, eine heisere Stimme jedoch kann man nicht sehen.

Beim Sprechen erzeugen wir also nicht nur Sprachschall, sondern auch sichtbare Sprechbewegungen. Dieser Vorgang läuft nicht nur parallel akustisch und optisch ab, vielmehr handelt es sich bei der Sprachproduktion um einen physiologischen Prozess, der sich akustisch und optisch manifestiert. Analog ist die Sprachwahrnehmung nicht nur Hören, sondern auch Sehen. Auditive und visuelle Sprachwahrnehmung findet dabei nicht getrennt statt, sondern ist direkt nach den Sinnesorganen Auge und Ohr ein stark vernetzter Prozess.

Unter dieser Voraussetzung versucht die Kommunikationswissenschaft, das komplexe System der mündlichen Kommunikation besser zu verstehen. Und eine wichtige Vision für die Zukunft besteht darin, die gewonnenen Erkenntnisse für eine bessere Kommunikation zwischen Mensch und Maschine zu nutzen. Weitere mögliche Anwendungen außer dem erwähnten Sprechtrainer sind virtuelle Agenten, Bildtelefonie, Computerspiele, Animationsfilme und der Einsatz als Werkzeug für Experimente zur audiovisuellen Sprachwahrnehmung.