Audio-visuelle Interaktion

Armin Kohlrausch

Philips Research Laboratories Eindhoven, Prof. Holstlaan 4, NL-5656 AA Eindhoven, Niederlande und
Department of Technology Management, TU Eindhoven, P.O. Box 513, NL-5600 MB Eindhoven, Niederlande

Zusammenfassung
In diesem Beitrag wird eine Übersicht über aktuelle grundlagen- sowie anwendungsorientierte Forschungsfragen auf dem Gebiet der audio-visuellen Interaktion gegeben. Als ein konkretes Beispiel für diese Interaktion wird die Frage behandelt, in welcher Weise die zeitliche Relation zwischen auditorischen und visuellen Stimuli (Asynchronie) die Wahrnehmung beeinflusst. Hierzu werden, neben einer kurzen Übersicht über die Literatur, eigene Daten zur Wahrnehmbarkeit von festen Verzögerungen zwischen Bild und Schall vorgestellt sowie methodische Aspekte besprochen.

Einleitung
In der alltäglichen Wahrnehmung unserer Umgebung erfahren wir durchgängig eine multi-sensorische Welt. Die Integration der Information in den verschiedenen sensorischen Modalitäten wird dadurch ermöglicht, dass multisensorische Stimuli im allgemeinen eine spezifische räumliche, zeitliche und kontextuelle Beziehung haben, wenn sie vom selben Objekt hervorgerufen werden. Bei technischen Systemen ist diese enge Kopplung nicht immer gegeben, da z.B. durch aufwendige Kodierungsverfahren oder durch Internetübertragung die zeitliche Relation zwischen Bild und Schall verändert werden kann. Neben der *Integration* von akustischen und visuellen Stimuli spielt auch die *Interaktion* zwischen ihnen eine wichtige Rolle. Solche Interaktionen sind nicht nur für die Kognitionsforschung von grossem Interesse, sondern müssen auch in technischen Anwendungen berücksichtigt werden, um unerwünschte Effekte zu vermeiden. Um technische Systeme so gut wie möglich auf die perzeptiven und kognitiven Fähigkeiten des menschlichen Benutzers abstimmen zu können, ist daher eine enge Zusammenarbeit zwischen experimenteller Wahrnehmungsforschung und den verschiedenen Anwendungsbereichen notwendig. Mit diesem Vortrag will ich dazu beitragen, eine solche Zusammenarbeit zu stimulieren.

Interaktion zwischen akustischen und visuellen Stimuli
Ein bekanntes Beispiel für die gegenseitige Beeinflussung von Stimuli in verschiedenen Modalitäten ist der sogenannte McGurk Effekt (McGurk und MacDonald, 1976). Zwei kurze Konsonant-Vokal Äußerungen werden gleichzeitig angeboten, eine akustisch und eine visuell. Die zwei Äußerungen enthalten denselben Vokal, aber unterschiedliche Konsonanten. Aufgabe der Versuchspersonen ist es, anzugeben, welche Äußerung sie wahrgenommen haben. Besteht der visuelle Stimulus aus der Silbe [ga] und der akustische aus der Silbe [ba], gaben nur 2 % der erwachsenen Versuchsteilnehmer an, [ba] wahrgenommen zu haben. Die Antwort von 98 % der Teilnehmer war [da], obwohl diese Silbe weder akustisch noch visuell angeboten wurde. Eine Demonstration des McGurk Effektes ist unter folgender Internetadresse zu finden:
`http://www.media.uio.no/personer/arntm/McGurk_english.html`

Eine andere starke Interaktion zeigt sich bei der Anbietung kurzer optischer und akustischer Stimuli (Shams et al., 2000). Wird ein kurzer optischer Stimulus von zwei akustischen Clicks begleitet, wird von der überwiegenden Zahl der Versuchspersonen berichtet, dass sie zwei optische Ereignisse wahrnehmen. Die Autoren schließen aus ihren Experimenten, dass solche Interaktionen über die Grenzen der sensorischen Modalitäten hinweg eher die Regel als die Ausnahme seien, und dass ein Verstehen dieser Interaktionen daher ein integraler Aspekt des Verstehens von Wahrnehmungsvorgängen sei. Auch für diesen Effekt gibt es eine Demonstration unter folgender Internetadresse:
`http:\\mag.bidmc.harvard.edu/soundInducedIllusoryFlash2/index.html`

Qualitätswahrnehmung von audio-visuellen Stimuli
Bei der Qualitätsbeurteilung von Fernseh- und Multimediasystemen interessiert die Frage, inwieweit die wahrgenommene Qualität in einer Modalität von der angebotenen Qualität in der jeweils anderen Modalität abhängt. Zu diesem anwendungsrelevanten Komplex gibt es überraschenderweise nur eine geringe Zahl von publizierten Untersuchungen, von denen zwei kurz besprochen werden sollen.
Beerends und de Caluwe (1999) verwendeten für ihre Untersuchungen Werbevideos, bei denen das Bild von HiFi Musik begleitet wurde. Sie fanden einen relativ deutlichen Einfluss der angebotenen Bildqualität auf die wahrgenommene Audioqualität. Bei hoher (niedriger) Bildqualität wurde die Audiobewertung besser (schlechter) beurteilt, verglichen mit der Situation, in der das Audiosignal ohne Bild angeboten wurde. Der Einfluss von Audioqualität auf wahrgenommene Bildqualität war dagegen sehr viel schwächer. Die Dominanz des Bildkanals in diesem Experiment wurde auch deutlich sichtbar beim Einfluss auf die Gesamtqualität. Die Urteile zur Gesamtqualität hatten eine Korrelation von 0,9 mit den Urteilen zur jeweiligen Bildqualität, während die Korrelation zwischen Gesamt- und Audioqualität nur bei 0,35 lag.
Eine ähnliche Untersuchung wurde von Rimell et al. (1998) veröffentlicht. Der Hauptunterschied zur vorherigen Studie lag in der Natur der Stimuli. Hier wurde die Aufnahme eines Sprechers verwendet, so dass Bild und Schall auch in ihrer zeitlichen Dynamik eng aneinander gekoppelt waren. Diese Autoren fanden deutliche Interaktionseffekte zwischen den beiden Modalitäten, die aber, im Gegensatz zu Beerends und de Caluwe, nahezu symmetrisch waren: Die angebotene Bildqualität hatte einen starken Einfluss auf die wahrgenommene Audioqualität, aber andererseits hatte die angebotene Audioqualität auch einen starken Einfluss auf die wahrgenommene Bildqualität. Die naheliegendste Erklärung für diese Diskrepanz zwischen den beiden Studien ist die Natur der Stimuli. Im Falle des Sprachstimulis kann von einer sehr starken perzeptiven Fusion zwischen Bild und Schall ausgegangen werden (die sich z.B. auch im Bauchsprecheffekt zeigt), wodurch die stärkere Interaktion verständlich ist.

(A)synchroniewahrnehmung in audio-visuellen Stimuli

Die zeitliche Relation zwischen der akustischen und der optischen Komponente in audio-visuellen Stimuli hat einen starken Einfluss auf die Wahrnehmung. Da durch digitale Kodierung und Übertragung von Multimediainhalten spezifische Verzögerungen für die Information in der jeweiligen Modalität auftreten können, sind die perzeptiven Folgen von asynchroner Wiedergabe auch für zahlreiche Anwendungsgebiete relevant.

In psychologischen Experimenten zur Asynchroniewahrnehmung werden häufig relativ einfache audio-visuelle (AV) Stimuli wie Kombinationen von Lichtblitzen und Clicks mit einer bestimmten physikalischen Verzögerung zwischen A und V verwendet. Bei der Frage nach der zeitlichen Reihenfolge wird von der Versuchsperson nach jeder Stimulusdarbietung ein Urteil abgegeben, welcher der beiden Anteile eher kam, A oder V. Alternativ kann gefragt werdem, ob die beiden Anteile synchron oder asynchron wahrgenommen wurden. Eine Variante dieser letzten Methode erlaubt drei Antwortalternativen, A erst, synchron, V erst. Fasst man die Ergebnisse solcher Studien zusammen, zeigt sich a) ein relativ grosser Bereich relativer Verzögerungen, in dem ein AV Stimulus als synchron beurteilt wird (von ca. 50 ms Audiovorlauf bis ca. 100 ms Audioverzögerung), und b) eine Asymmetrie in dem Sinne, dass Verzögerungen der akustischen Komponente eher zu einem Synchron-Urteil führen als entsprechende Verzögerungen der optischen Komponente. Diese Asymmetrie lässt sich mit dem Begriff des Punktes der subjektiven Gleichzeitigkeit fassen, der im Mittel bei einer physikalischen Verzögerung des Audio Signals gegenüber dem Video Signal um ca. 40 ms liegt.

Diese Asymmetrie zeigt sich auch bei mehr anwendungsorientierten Fragestellungen. Von Rihs (1995) wurde untersucht, wie die wahrgenommene Qualität einer Videosequenz von der Verzögerung zwischen A und V abhängt. Es zeigt sich ein breites Maximum der Qualitätsbeurteilung rund um eine Audioverzögerung von ca. 40 ms. Die Qualitätsurteile fielen bei größeren Audioverzögerungen relativ langsam ab, während bei kleineren Audioverzögerungen (entsprechend zunehmender Videoverzögerung) die wahrgenommene Qualität stark abnahm. Diese Ergebnisse haben zu einer Empfehlung der ITU (International Telecommunication Union) für die zulässige Verzögerung zwischen Bild und Schall bei Anwendungen im Fernsehbereich geführt (ITU, 1998).

In unseren Experimenten verwendeten wir als visuellen Stimulus eine weisse Scheibe, die auf eine (sichtbare) Platte herabfiel und nach der Berührung wieder reflektiert wurde. Das akustische Signal bestand aus einem kurzen 500-Hz Ton, der hart eingeschaltet wurde und mit einer Zeitkonstante von 30 ms ausklang. Mit diesem Stimulus wurden zunächst adaptive Schwellenmessungen mit einem 2AFC Verfahren zur Wahrnehmbarkeit von Asynchronie durchgeführt (Kohlrausch und van de Par, 2000). Hier zeigte sich, dass die Schwelle bei Videoverzögerung (Audiovorlauf) im Mittel bei ca. 30 ms lag, während bei Audioverzögerung die Schwellen bei ca. 85 ms (Check Werte) lagen.

In einem weiteren Experiment wurden verschiedene Methoden zur Beurteilung der wahrgenommenen Reihenfolge bzw. der Asynchronie dieses AV Stimulus angewendet (van de Par und Kohlrausch, 2002). Hierbei zeigte sich, dass die beiden Methoden mit den Antwortkategorien "Synchron-Asynchron" , bzw. "Video erst-Synchron-Audio erst" sehr ähnliche Ergebnisse liefern. Die Kurven mit den Antworten "synchron" als Funktion der relativen Verzögerung waren jeweils nahezu identisch und die Ergebnisse stimmten gut mit Literaturdaten überein, insbesondere lag der Punkt der subjektiven Gleichzeitigkeit bei positiven Audioverzögerungen

Abweichend davon verhielten sich die Resultate für die Methode, bei der die zeitliche Reihenfolge beurteilt wurde (Antwortkategorien "Audio erst-Video erst"). Zunächst einmal lag hier der Punkt der subjektiven Gleichzeitigkeit deutlich anders, und zwar häufig bei negativen Audioverzögerungen, also bei Audiovorlauf. Weiterhin zeigte sich, dass die gemessenen Werte stark durch das von der Versuchsperson gewählte Kriterium zu beeinflussen waren. Aufgrund dieser Ergebnisse liegt die Vermutung nahe, dass die Methode des zeitlichen Reihenfolgeurteils weniger stabile Werte ergibt als die anderen verwendeten Methoden. Dieser Eindruck wurde auch durch die Ergebnisse einer ausführlichen Literaturübersicht unterstützt (Kohlrausch, 2000).

Zusammenfassung

In diesem kurzen Überblick habe ich einige Aspekte zur Erforschung audio-visueller Interaktionen vorgestellt, wobei aufgrund des begrenzten Raumes allerdings vieles, z.B. neurophysiologische Beobachtungen, unerwähnt bleiben musste. Interessierte Leser seien deshalb auf den Beitrag in dem Buch hingewiesen, in dem die Ergebnisse des Vorkolloquiums und auch dieses Plenarvortrages ausführlich dargestellt werden, und auf das an anderer Stelle dieses Konferenzbandes genauer eingegangen wird.

Literatur

Beerends, J. and de Caluwe, F. (1999). The influence of video quality on perceived audio quality and vice versa, J. Aucio Eng. Soc. **47**, 355–362.

ITU (1998). Relative timing of sound and vision for broadcasting. Recommendation ITU-R BT 1351.1

Kohlrausch, A. (2000). Perceptual consequences of audio-visual asynchrony, in *IPO Annual Progress Report* **35**, 140–149.

Kohlrausch, A. und van de Par, S. (2000). Experimente zur Wahrnehmbarkeit von Asynchronie in audio-visuellen Stimuli, in *Fortschritte der Akustik, DAGA 2000*, 317–318.

McGurk, H. und MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, pp. 746–748.

van de Par, S. und Kohlrausch, A. (2002). Some methodological aspects for measuring asynchrony detection in audio-visual stimuli, in *Proceedings of the 3rd EAA Convention*, Sevilla, Sept. 2002, in press.

Rihs, S. (1995). The influence of audio on perceived picture quality and subjective audio-video delay tolerance, in *Proceedings of the MOSAIC workshop Advanced Methods for the Evaluation of Television Picture Quality*, R. Hamberg and H. de Ridder, eds., ch. 13, pp. 133–137, Institute for Perception Research, Eindhoven, The Netherlands.

Rimell, A.N., Hollier, M.P. und Voelcker, R.M. (1998). The influence of cross-modal interaction on audio-visual speech quality perception, in *105rd Convention of the Audio Engineering Society, San Francisco, September 1998, Preprint No. 4791 (H-6)*.

Shams, L., Kamitani, Y. und Shimojo, S. (2000). What you see is what you hear, *Nature* **408**, 788.