Experimente zur Wahrnehmbarkeit von Asynchronie in audio-visuellen Stimuli

Armin Kohlrausch, Steven van de Par

Philips Research Laboratories Eindhoven, Prof. Holstlaan 4, NL-5656 AA Eindhoven, Niederlande und
IPO-Center for User-System Interaction, TU Eindhoven, P.O. Box 513, NL-5600 MB Eindhoven, Niederlande

Zusammenfassung

In diesem Beitrag werden drei Experimente beschrieben, die der Frage nachgehen, wie Verzögerungen in einfachen audio-visuellen Stimuluspaaren wahrgenommen werden. Im ersten Experiment wurden mit einer adaptiven Methode direkt die Unterscheidbarkeitsschwellen zwischen synchronen und asynchronen Paaren gemessen, wobei die Schwellen für eine Verzögerung des Video und des Audio Signals getrennt bestimmt wurden. Die mittleren Schwellenwerte liegen bei - 30 ms (Video verzögert) und +90 ms (Audio verzögert). In den beiden folgenden Experimenten hatten die Versuchspersonen zu beurteilen, ob in den audio-visuellen Paaren der visuelle Stimulus oder der auditive Stimulus zuerst angeboten wurde oder ob beide Anteile synchron waren. Die Stimuli wurden mit relativen Verzögerungen zwischen -350 ms (Audio erst) und +350 ms (Video erst) angeboten. Über einen weiten Bereich von Verzögerungen (-100 ms bis + 200 ms) überwog das Urteil synchron, wobei allerdings dieselbe Asymmetrie (größere Toleranz im Falle von "Video erst" verglichen mit der Situation "Audio erst") wie im Detektionsexperiment auftritt.

Einleitung

In der alltäglichen Wahrnehmung unserer Umgebung erfahren wir durchgängig eine multi-sensorische Welt. Um die sensorische Information optimal zu benutzen können, ist es erforderlich die Information, die uns über die verschiedenen sensorischen Kanäle erreicht, zu integrieren. Eine solche Integration wird dadurch möglich, daß multisensorische Stimuli im allgemeinen eine spezifische räumliche, zeitliche und kontextuelle Beziehung haben, wenn sie vom selben Objekt hervorgerufen werden.

In diesem Beitrag wollen wir einen Faktor näher untersuchen, der bei dieser Integration eine bedeutende Rolle spielen kann, nämlich die zeitliche Synchronie zwischen auditorischen und visuellen Stimuli. Bei der Wahrnehmung natürlicher Vorgänge spielt die unterschiedliche Ausbreitungsgeschwindigkeit von Schall und Licht eine wichtige Rolle. Aufgrund dieser physikalischen Gesetzmäßigkeit kann der Schall eines Ereignisses niemals früher beim Beobachter eintreffen als das dazugehörige Lichtsignal. Einem Objektabstand von 10 m entspricht bereits eine physikalisch bedingte Asynchronie von etwa 30 ms.

Experimente zur Wahrnehmbarkeit von audio-visueller (AV) Asynchronie erfordern die Wiedergabe von AV Stimuli mit Apparaten, die es erlauben, die relative Verzögerungszeit zu variieren. Dabei wird davon ausgegangen, daß in der Originalaufnahme die visuellen und akustischen Stimuli synchron sind. Dixon und Spitz (1980) verwendeten die Aufzeichnung eines Hammers, der gegen einen Holzpflock geschlagen wurde, als Stimulus zur Messung der Wahrnehmbarkeit von Asynchronie. Dazu wurde, ausgehend von einer synchronen Darbietung, die Asynchronie langsam erhöht, bis die Versuchspersonen den Stimulus als asynchron beurteilten. Als zweiten Stimulus verwendeten diese Autoren ein Sprachsignal. Die mittleren Schwellenwerte lagen beim "Hammer"-Stimulus generell niedriger, hier wurden Schwellen von 188 ms (Video zuerst) und 75 ms (Audio zuerst) gefunden. Für den Sprachstimulus lagen die entsprechenden Werte bei 258 ms und 131 ms. Diese Ergebnisse wurden in einer neueren Studie von Miner und Caudell (1998) bestätigt, in der mit einer Reihe unterschiedlicher Impaktstimuli und mit Sprache gearbeitet wurde.

Bei beiden Studien ist es schwierig, zwischen der Empfindlichkeit für Asynchronie an sich und der Neigung, eine bestimmte Asynchronie zu tolerieren, zu unterschieden. Ein solcher Unterschied wird z.B. in einer kürzlich veröffentlichten Norm der ITU (ITU, 1998) gemacht, in der sowohl Werte für die gerade wahrnehmbare AV Verzögerung ("detectability threshold") wie auch für den Akteptanzbereich angegeben sind. Diese Werte wurden allerdings indirekt aus Qualitätsbeurteilungen abgeleitet, wobei nicht eindeutig klar ist, inwieweit die gewählten Definitionen für die Wahrnehmbarkeitsschwellen den in der Psychophysik üblichen entsprechen. Ziel der hier vorgestellten Messungen ist es, mithilfe abstrakter AV Stimuli sowohl Detektionsschwellen wie auch Toleranzschwellen zu bestimmen.

Experiment 1: Detektionsschwellen

Die AV Stimuli wurden mit einem PC berechnet und präsentiert. Visuelle Stimuli wurden auf einem 17 Zoll Dell Monitor mit einer Bildfrequenz von 74.8 Hz und einer Auflösung von 1024x768 Pixel dargeboten. Die akustischen Signale wurden mit einer Samplefrequenz von 44.1 kHz berechnet und über Beyerdynamic DT 990 Kopfhörer wiedergegeben.

Der visuelle Stimulus bestand aus einer weißen Kreisscheibe, die sich mit linear zunehmender Geschwindigkeit auf eine weiße Platte herabbewegte, bis beide miteinander in Kontakt kamen. Danach bewegte sich die Scheibe mit linear abnehmender Geschwindigkeit wieder nach oben. Das akustische Signal bestand aus einem hart geschaltetem 500 Hz Sinuston, der entsprechend einer Exponentialfunktion (Zeitkonstante 30 ms) ausklang. In der synchronen Situation begann der Ton zum Zeitpunkt des Kontaktes zwischen Kreisscheibe und Platte.

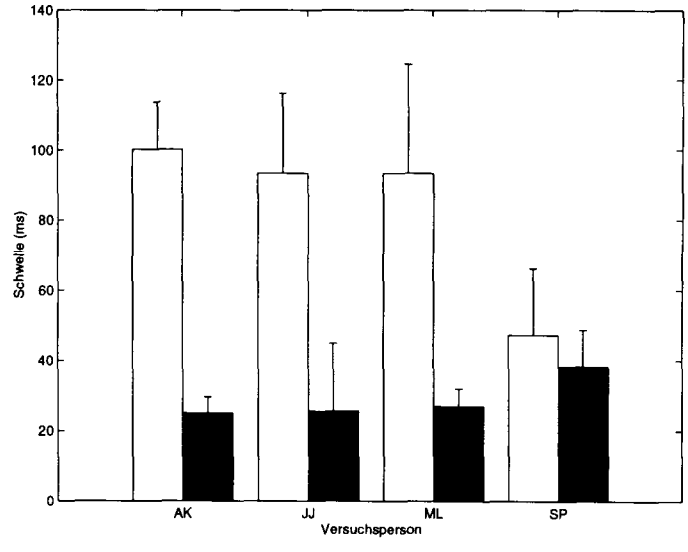

Abb. 1: Unterscheidbarkeitsschwellen zwischen synchronen und asynchronen AV Stimuli für 4 Versuchspersonen. Die offenen Balken zeigen Schwellen für den Fall "Video zuerst", die schwarzen Balken für den Fall "Audio zuerst".

Schwellen wurden mit einem adaptiven 2 IFC Verfahren gemessen. Innerhalb eines Trial wurde in einem Intervall der synchrone Stimulus als Referenz angeboten, während das Testintervall den asynchronen Stimulus enthielt. Die Größe der Verzögerung wurde mit einem 2-down 1-up Verfahren adaptiv gesteuert. Schwellen für "Video verzögert" und für "Audio verzögert" wurden getrennt bestimmt. Die Versuchsperson (Vp) erhielt nach jedem Intervall Rückmeldung darüber, ob ihre Anwort richtig oder falsch war. Vier Vpn nahmen an diesem Experiment teil und bestimmten die Schwellen jeweils vier mal.
Die Ergebnisse der vier Vpn sind in Abb. 1 dargestellt. Die offenen Balken geben die Schwellenwerte für die Situation "Audio verzögert" an, die schwarzen Balken zeigen die Werte für die Situation "Video verzögert". Für drei der vier Vpn zeigt sich eine deutliche Asymmetrie, mit höheren Werten (90 bis 100 ms) für "Audio verzögert" als für "Video verzögert" (25 bis 30 ms), während die Werte für SP für diese beiden Situationen mit 47 und 38 ms nahezu identisch sind.

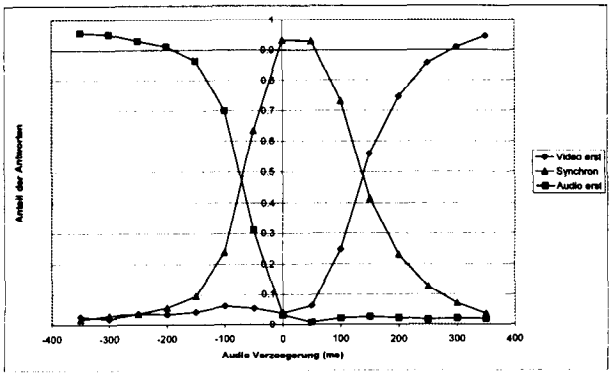

Abb. 2: Urteile zur wahrgenommenen (A)synchronie für dieselben Impaktstimuli wie in Abb. 1. Angegeben ist der relative Anteil der Antworten "audio zuerst", "synchron", "video erst" als Funktion der Audioverzögerung. Negative Verzögerungswerte entsprechen einer Verzögerung des Videosignals. Mittlere Daten von 9 Versuchspersonen.

Experiment 2: Toleranzschwellen für AV Asynchronie

In diesem Experiment wurden dieselben AV Stimuli wie in Experiment 1 verwendet, aber eine andere Fragestellung gewählt: Die Vpn sahen einen isolierten AV Stimulus und hatten danach anzugeben, ob ihrem Urteil nach 1) Video zuerst kam, 2) Audio und Video synchron waren, oder 3) Audio zuerst kam. Insgesamt wurden 15 verschiedene AV Verzögerungen angeboten und, verteilt auf 6 Blöcke, von jeder Vpn 60 mal beurteilt. An diesem Experiment nahmen 9 Vpn teil.
In Abb. 2 sind die gemittelten Daten aller Vpn dargestellt. Die drei Kurven geben den relativen Anteil der drei möglichen Antworten als Funktion der Audio Verzögerung an, wobei negative Verzögerungen ein Verauslaufen des Audio Signals angeben. Die Antworten der Kategorie "synchron" (Dreiecke) zeigen einen deutlichen Bias in Richtung positiver Audio Verzögerungen. Als quantitatives Maß für diesen Bias haben wir das gewichtete Mittel dieser Kurve berechnet, das sich als Punkt der subjektiven Gleichzeitigkeit interpretieren läßt und für die gemittelten Daten bei +39 ms liegt. Für die einzelnen Vpn variiert diese Größe zwischen 8 und 66 ms.
Die Schnittpunkte der mittleren mit den beiden anderen Kurven ergeben ein Maß für den Toleranzbereich der wahrgenommenen Synchronie. Dieser Bereich erstreckt sich in Abb. 2 von -76 bis +148 ms, also insgesamt über 224 ms. Dieser letzte Wert liegt für die einzelnen Vpn zwischen 113 und 413 ms.
In einem weiteren Experiment wurden wiederum die Toleranzschwellen bestimmt, allerdings besaß der AV Stimulus keine so deutlich ausgeprägte Zeitstruktur. Der visuelle Stimulus bestand wiederum aus einer weissen Kreisscheibe, die sich senkrecht auf und ab bewegte. Die (vertikale) Position veränderte sich als Funktion der Zeit entsprechend einer Gaussglocke mit einer Standardabweichung von 25 ms. Dieselbe Gaussfunktion wurde als Einhüllende eines 500 Hz Tones verwendet, so dass bei synchroner Darbietung die höchste Position der Kreisscheibe mit dem Einhüllendenmaximum des Tones zusammenfiel. Alle weiteren experimentellen Details entsprachen dem des vorherigen Experimentes.
Die gemittelten Daten in Abb. 3 zeigen dieselben Trends, die bereits in Abb. 2 zu sehen waren. Unterschiede bestehen allerdings in der geringeren Höhe und der (hier größeren) Breite der mittleren Kurve und in den geringeren Steigungen der Kurven. Der Punkt subjektiver Gleichzeitigkeit liegt hier bei 53 ms und variiert für die einzelnen Vpn zwischen 18 und 97 ms. Die Extremwerte werden in beiden Messungen von denselben Vpn erreicht. Der Toleranzbereich in den gemittelten Daten erstreckt sich über 262 ms, bei den einzelnen Vpn liegt er zwischen 122 und 493 ms.
Zusammenfassend läßt sich sagen, daß die in den genannten früheren Studien gefundenen relativ hohen Schwellenwerten eher die Toleranz als die absolute Empfindlichkeit der Vpn angeben. Weiterhin zeigt sich, daß die mit unserer adaptiven Schwellenmethode gefundenen Wahrnehmbarkeitsschwellen von -30 und +90 ms um einiges unter den im ITU Standard genannten Werten von -45 und +125 ms liegen. Dieser Unterschied kann sowohl auf der Verwendung eines relativ unkritischen Stimulus (Sprache) wie auch auf dem im ITU Standard verwendeten Schwellenkriterium (Qualitätsabnahme) beruhen.

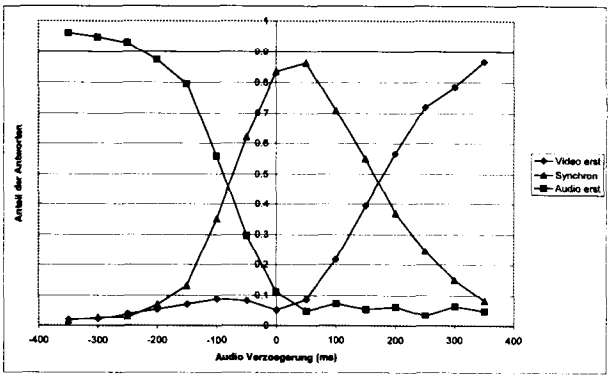

Abb. 3: Urteile in Hinblick auf wahrgenommene (A)synchronie für AV Stimuli mit weniger ausgeprägter zeitlicher Dynamik als die in Abb. 2 verwendeten Stimuliwie in Abb. 1. Angegeben ist der relative Anteil der Antworten "audio zuerst", "synchron", "video erst" als Funktion der Audioverzögerung. Negative Verzögerungswerte entsprechen einer Verzögerung des Videosignals. Mittlere Daten von 9 Versuchspersonen.

Literatur

Dixon, N.F. und Spitz, L. (1980). The detection of audiovisual desynchrony. Perception **9**, 719–721.

ITU-R (1998). Relative timing of sound and vision for broadcasting. Rec. ITU-R BT. 1359.

Miner, N. und Caudell, T. (1998). Computational requirements and synchronization issues for virtual acoustic displays. Presence **7**, 396–409.