Audiovisuelles Sprachverstehen in dynamischen Cocktailparty-Situationen bei jüngeren und älteren Erwachsenen

Alexandra Begau[1], Stephan Getzmann[1]

[1] *Leibniz-Institut für Arbeitsforschung an der TU Dortmund, 44139 Dortmund, E-Mail:begau@ifado.de*

Einleitung

„Cocktailparty"-Situationen zeichnen sich dadurch aus, dass sich eine Person in einer schwierigen Hörumgebung befindet, in der sie zeitgleich mehreren verschiedenen auditorischen Reizen ausgesetzt ist. Dies ist zum Beispiel der Fall, wenn mehrere Personen gleichzeitig sprechen. In solchen Situationen ist es notwendig, die Aufmerksamkeit auf den Zielreiz zu fokussieren und unwichtige Information auszublenden. Mit steigendem Alter lassen sich jedoch zunehmende Schwierigkeiten in solchen Situationen feststellen, insbesondere wenn es zu Sprecherwechseln kommt, die eine schnelle Umorientierung auf den relevanten Reiz erforderlich machen [1]. Grund dafür sind zum einen das Nachlassen der Hör- und Sehfähigkeiten auf sensorischer Ebene, aber auch mit dem Alter stärker werdende kognitive Einschränkungen. Diese betreffen unter anderem die gezielte Aufmerksamkeitsausrichtung sowie die effektive Inhibition irrelevanter Stimuli [2].

Redundante Informationen können hier hilfreich für das Sprachverstehen sein [3]. Im Falle eines Gesprächs zwischen zwei oder mehreren Personen ist diese Redundanz durch die parallel dargebotene visuelle und auditive Sprachinformation gegeben. Insbesondere Ältere profitieren in uneindeutigen Hörsituationen von der zusätzlichen (visuellen) Sprachinformation und integrieren diese effektiver als junge Personen. Dieser Befund zeigt sich sowohl in Verhaltensmaßen wie Reaktionszeiten und Antwortrichtigkeiten als auch in neurophysiologischen Maßen wie den ereigniskorrelierten Potenzialen (EKP), die aus dem Elektroenzephalogramm (EEG) abgeleitet werden und auf singuläre Ereignisse hin auftreten [4]. In verschiedenen Studien konnte die Relevanz der P1, N1 und P2 Komponenten für audiovisuelle Sprache herausgestellt werden. Die P1 reflektiert dabei die frühe Verarbeitung sensorischer Stimulation, während die N1 in Zusammenhang gebracht wird mit Prozessen der Integration zeitlicher und räumlicher Eigenschaften audiovisueller Sprache. Die P2 ist sensitiv für die Kohärenz des visuellen und auditorischen Sprachinhalts und wird eher mit der Integration der phonetischen Eigenschaften assoziiert [5]. Verschiedene Studien zur audiovisuellen Sprachverarbeitung zeigen eine Reduktion der Amplituden dieser Komponenten, wenn visuelle Sprachinformation dargeboten wird, die kongruent mit der auditiven Modalität ist [4, 5, 6]. Eine solche Modulation auditorischer EKP wird als Beleg für eine Erleichterung der Verarbeitung akustischer Sprachinformation durch komplementäre visuelle Information angenommen [7] und als Beispiel für multisensorische Integration im Sinne eines „analysis-by-synthesis" Mechanismus interpretiert [8]. Eine weitere, in diesem Zusammenhang relevante Komponente ist die N2, die ein Maß für die kognitive Kontrolle in Konfliktsituationen darstellt und in der Regel auf Stimuli folgt, die inhibitorische Aufmerksamkeitsprozesse erfordern [2]. Bei älteren Probanden wird häufig ein Fehlen der N2-Amplitude gemeinsam mit einer Reduktion der früheren P1- und N1-Amplituden beobachtet, welche mit einem inhibitorischen Defizit assoziiert werden [1, 2].

Vorteile audio-visueller Sprachinformation sollten sich nicht nur in vergleichsweise einfachen Hörumgebungen, sondern vor allem in Situationen mit mehreren wechselnden Sprechern zeigen. In unserer Studie untersuchten wir deshalb, wie hilfreich audiovisuelle Informationen im Rahmen von Cocktailparty-Situationen sind, in denen es zu dynamischen Wechseln des Zielsprechers kommt. Hierbei interessierte uns das Ausmaß und die Relevanz visueller Information, die zu einer Erleichterung der Sprachverarbeitung notwendig sind. Verwendet wurde eine Sprachverstehensaufgabe, bei der jüngeren und älteren Probanden kurze Zielwörter präsentiert wurden. Dies erfolgte entweder aus einer regulären (Standard-)Position oder einer abweichenden (Deviant-)Position. Wir erwarteten, dass insbesondere ältere Probanden im Vergleich zu Jüngeren von Informationen in der zusätzlichen Modalität profitieren.

Methode

Als Stimuli wurden Videos von den Gesichtern dreier junger Sprecherinnen verwendet, die kurze Wörter artikulierten. Zielreize waren die Wörter „Yes" und „No"; als Distraktoren wurden die Zahlen von „eins" bis „zehn" verwendet. Mittels einer horizontalen Anordnung von Lautsprechern und Monitoren wurden den Probanden gleichzeitig drei Videos aus drei Monitoren (-20°, 0°, 20° Azimuth) dargeboten, wovon immer nur eines das Zielwort beinhaltete und die beiden anderen Distraktorwörter präsentierten (s. Abbildung 1). Der Start der auditorischen Sprachreize erfolgte bei jedem Video zeitgleich 1500 ms nach Videobeginn. Ein Stimulusvideo (Gesamtlänge 2900 ms) begann und endete immer mit einer Einblendung der Sprecherin, deren Mund zu Beginn und Ende der Sprachäußerung geschlossenen war. Instruierte Blickrichtung war der zentrale Monitor, auf dem in 80% der Durchgänge das Zielwort erschien (Standard). In 20% der Durchgänge erfolgte die Zieldarbietung lateral (Deviant). Die Probanden sollten per Tastendruck indizieren, welches der beiden möglichen Zielwörter sie erkannt hatten. Die Stimuli wurden unter akustischen Freifeldbedingungen in drei audiovisuellen Bedingungen dargeboten. Dabei wurde alleine die visuelle Information variiert, während der auditorische Input über die Bedingungen hinweg gleichblieb. Die Variationen bestanden aus (a) audiovisuell kongruenten Lippenbewegungen, (b) Standbildern der Sprechergesichter, und (c) einer visuell unspezifischen Mundöffnung, bei der sich Mund und Lippen der Sprecherinnen bewegten, dabei jedoch kein sinntragendes Wort formuliert wurde. Während des Versuches wurde das EEG abgeleitet (1000 Hz Abtastrate; 64 Kanäle; 0.5 bis 30 Hz Bandpassfilter), die EKP auf den Start der auditorischen Sprachreize wurden über einem fronto-zentralen Elektroden-

cluster (FCz-Cz-Fz-FC1-FC2) gemittelt. Als Baseline-Intervall wurden 100 ms vor Videobeginn gewählt.

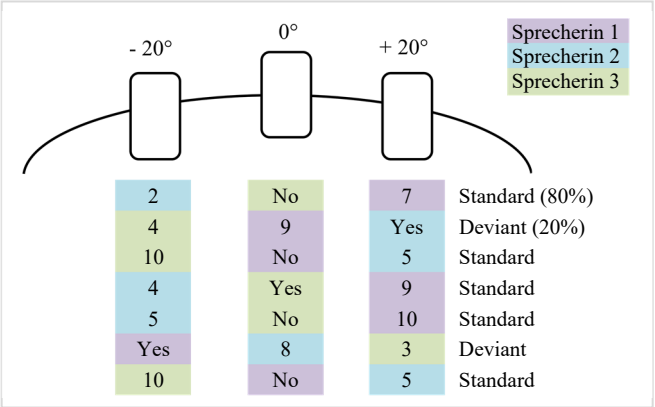

Abbildung 1: Schematische Darstellung der drei Monitor-/Lautsprecher-Kombinationen zur Darbietung der audiovisuellen Stimuli. Die Darbietung der Zielwörter („Yes“ bzw. „No“) und Distraktoren (Zahlwörter „1“ bis „10“) erfolgte zu 80% aus der zentralen Position (0°, Standard). In 20% der Fälle wurden Zielwörter aus der lateralen Position (+/-20°, Deviant) gezeigt. Blickrichtung war immer zentral.

Untersucht wurden $N = 22$ junge Probanden im Alter von 20 bis 34 Jahren ($M = 25.45$ Jahre) sowie $N = 20$ ältere Probanden im Alter von 55 bis 70 Jahren ($M = 64.50$ Jahre). Bei allen Probanden wurden vor Versuchsbeginn die Seh- und Hörfähigkeiten überprüft und kognitive Leistungstests durchgeführt.

Es wurden Varianzanalysen mit Messwiederholung und dem Zwischensubjektfaktor Alter (Jung vs. Alt) sowie den Innersubjektfaktoren visuelle Information (audiovisuell kongruent vs. Standbild vs. visuell unspezifisch) und Zielposition (Standard vs. Deviant) berechnet, abhängige Variablen waren die Amplituden der P1, N1, P2, N2. Zur Reduktion der Einflüsse (nicht-auditiver) Verarbeitungsprozesse, die auf den Beginn der visuellen Sprachinformation hin auftraten (vgl. Abbildung 4), wurden Differenzmaße zwischen den mittleren („peak-to-peak“) Amplituden der einzelnen EKP gebildet. Für die Analyse der Reaktionszeiten und Antwortrichtigkeit wurden Differenzwerte für Deviant- minus Standardposition berechnet, um den Einfluss eines Sprecherwechsels auf das Sprachverstehen abzubilden. Paarweise Post-Hoc Vergleiche wurde mithilfe von Tukey‘s HSD vorgenommen. Als Effektstärkemaße berichten wir ω_{part}^2 und Hedge’s g.

Ergebnisse

Reaktionszeiten. Es zeigte sich ein Haupteffekt der visuellen Information ($F_{80,2} = 107.33$, $p < .001$, $\omega_{part}^2 = .84$), der durch das Alter moduliert wurde (Interaktion Alter*visuelle Information $F_{80,2} = 4.73$, $p = .011$, $\omega_{part}^2 = .15$). Die Reaktionszeitdifferenz zwischen Deviant- und Standardposition zeigt, dass bei Stimuli aus der Deviantposition längere Reaktionszeiten auftraten. Diese Unterschiede waren bei Jüngeren bei audiovisuell kongruenter Information (443.74 ms) größer als bei visuell unspezifischer Information (214.88 ms, $p < .001$, $g = 3.04$) und beim Standbild (243.17 ms, $p < .001$, $g = 3.48$; Abbildung 2). Auch bei Älteren zeigten sich größere Unterschiede bei audiovisuell kongruenter Information (494.43 ms) im Vergleich zu unspezifisch visueller Information (209.29 ms, $p < .001$, $g = 2.44$) und dem Standbild (167.84 ms, $p < .001$, $g = 2.14$). Der Unterschied zwischen audiovisuell kongruenter Information und dem Standbild fiel bei Älteren größer aus (326.59 ms vs. 200.57 ms).

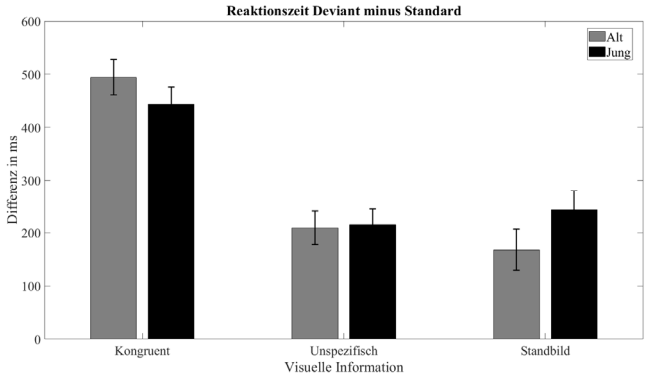

Abbildung 2: Differenzen der Reaktionszeiten bei Deviant- minus Standardposition relativ zum Beginn der Lippenbewegung, aufgeschlüsselt nach Altersgruppe und visueller Information.

Antwortrichtigkeit. Insgesamt war die Rate der korrekten Antworten auf Zielreize aus der Standardposition höher als aus Deviantpositionen (95.24% vs. 86.35%). Zudem lag ein Haupteffekt der visuellen Information auf die Differenz der Antwortrichtigkeit vor ($F_{80,2} = 20.29$, $p < .001$, $\omega_{part}^2 = .48$; Abbildung 3). Bei audiovisuell kongruenter Information (-14.11%) war diese Differenz stärker ausgeprägt als bei visuell unspezifischer Information (-7.02%, $p < .001$, $g = -1.08$) und beim Standbild (-5.52%, $p < .001$, $g = -1.30$).

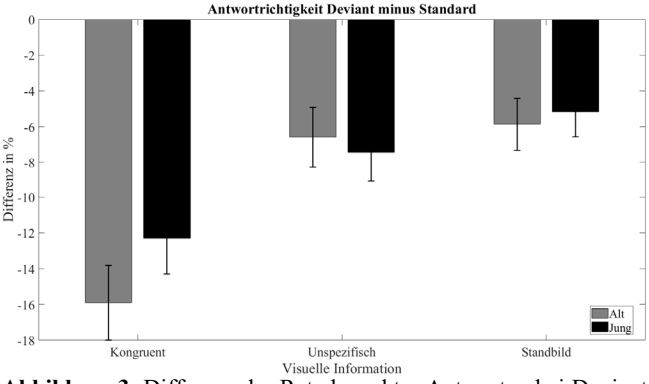

Abbildung 3: Differenz der Rate korrekter Antworten bei Deviant- minus Standardposition in Prozent, aufgeschlüsselt nach Altersgruppe und visueller Information.

ERPs. Für die P1-N1 „peak-to-peak“ Amplitude ergab sich ein Haupteffekt der visuellen Information ($F_{80,2} = 7.89$, $p < .001$, $\omega_{part}^2 = 0.25$). Im Post-Hoc Vergleich zeigten sich geringere Amplituden bei audiovisuell kongruenten Reizen (1.52 µV) als beim Standbild (1.97 µV, $p = .037$, $g = -.40$) und visuell unspezifischen Reizen (2.18 µV, $p < .001$, $g = -.60$).

Für die N1-P2 „peak-to-peak“ Amplitude ergab sich ebenfalls ein Haupteffekt der visuellen Information ($F_{80,2} = 5.08$, $p = .008$, $\omega_{part}^2 = .16$). Die Amplituden waren bei visuell unspezifischen Reizen (-3.35 µV) größer als bei audiovisuell kongruenten Reizen (-2.86 µV, $p = .017$, $g = -.36$) und dem Standbild (-2.71 µV, $p = .016$, $g = -.47$).

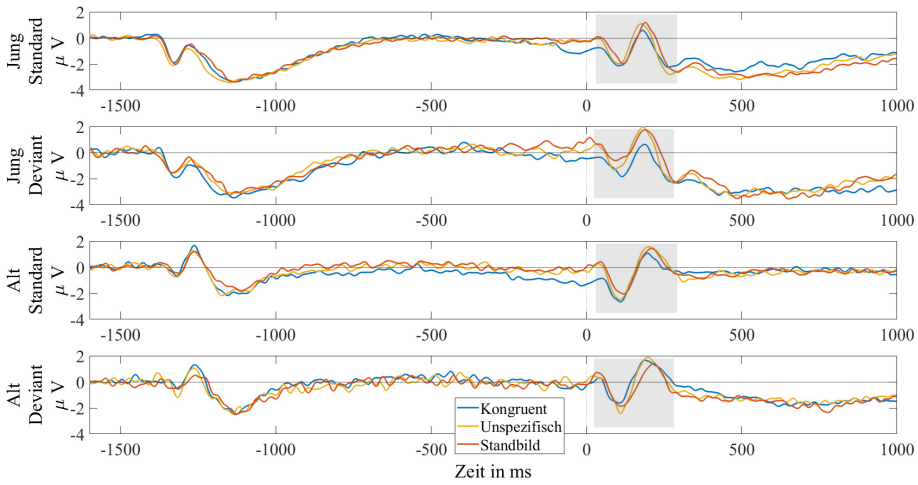

Abbildung 4: Ereigniskorrelierte Potenziale gemittelt über ein fronto-zentrales Elektrodencluster, getrennt für junge und alte Probanden und für Zielreize aus der Standardposition und devianten Positionen. Bei Zeitpunkt -1500 ms begannen die Videosequenzen, Zeitpunkt 0 ms stellt den Beginn des (akustischen) Sprachreizes dar. Die analysierten EKP-Komponenten liegen im grau schattierten Bereich.

Die P2-N2-Amplitude war bei Älteren geringer als bei Jüngeren (1.91 µV vs. 3.33 µV, Haupteffekt Alter: $F_{40,1} = 12.38$, $p = .001$, $\omega_{part}^2 = .21$). Bei Darbietung aus der Standardposition traten geringere Amplituden auf als aus Deviantpositionen (2.42 µV vs. 2.81 µV, Haupteffekt Zielposition: $F_{40,1} = 6.37$, $p = .016$, $\omega_{part}^2 = .11$). Zudem lag ein Haupteffekt der visuellen Information vor ($F_{80,2} = 17.70$, $p < .001$, $\omega_{part}^2 = .44$), der durch das Alter moduliert wurde (Interaktion visuelle Information*Alter: $F_{80,2} = 6.72$, $p = .002$, $\omega_{part}^2 = -.47$; Abbildung 5).

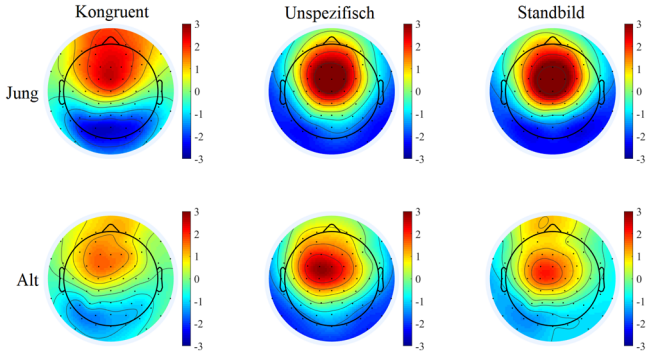

Abbildung 5: Topographien der P2-N2-Amplitudendifferenzen bei Darbietung aus der Standardposition für junge (oben) und alte (unten) Probanden und der Art der visuellen Information. Die mittlere P2-Amplitude befindet sich im Zeitfenster 180-200 ms, die der N2 bei 292-312 ms.

Die Post-Hoc Analyse des Interaktionseffekts ergab, dass die Amplituden bei den Jüngeren bei audiovisuell kongruenten Reizen (2.58 µV) geringer waren als beim Standbild (3.76 µV, $p < .001$ $g = -1.15$) und bei visuell unspezifischen Reizen (3.64 µV, $p < .001$, $g = -1.03$). Währenddessen waren die Amplituden bei den Älteren bei audiovisuell kongruenten Reizen (1.62 µV, $p = .005$, $g = -.77$) und beim Standbild (1.69 µV, $p = .006$, $g = -.71$) kleiner als bei visuell unspezifischen Reizen (2.42 µV).

Diskussion

Die Verhaltensdaten zeigen, dass kongruente audiovisuelle Sprachinformation im Vergleich zum Standbild nur dann einen Vorteil bot, wenn der Zielreiz aus der erwarteten Position präsentiert wurde. Bei einem unerwarteten Wechsel der Zielposition war sie jedoch nachteilig. In Hinblick auf die Reaktionszeiten war dieser Effekt bei Älteren besonders stark ausgeprägt. Grund dafür könnte die Inkongruenz der audiovisuellen Sprachinformation sein, die dadurch entstand, dass der Zielreiz plötzlich aus einer anderen Richtung erschien als aus derjenigen, in die gerade geschaut wurde. Dieser Nachteil der audiovisuellen Sprachinformation bestand auch im Vergleich zur visuell unspezifischen Information, die im Vergleich zum Standbild jedoch keinerlei Mehrwert für die zuhörende Person bot. Diese Befunde sind in Übereinstimmung mit verschiedenen Studien, die Vorteile audiovisuell kongruenter Sprachinformation gegenüber rein auditiver Sprache [4] sowie Nachteile inkongruenter Sprachinformation zeigen konnten [9].

Auf neuronaler Ebene deutete die verringerte P1-N1-Amplitude auf eine erleichterte Sprachverarbeitung bei der Darbietung audiovisuell kongruenter Stimuli hin. Dies ist konsistent mit verschiedenen Studien, die zeigen konnten, dass geringere P1-/N1-Amplituden mit erleichterter Verarbeitung assoziiert sind (z.B. [6]). Da sich bei visuell unspezifischer Information gegenüber dem Standbild keine Amplitudenunterschiede zeigten, ist davon auszugehen, dass die zusätzliche visuelle Information sprachhaltig sein muss, um tatsächlich eine Erleichterung zu erzielen. Die stärkeren N1-P2-Amplituden bei visuell unspezifischen Reizen gegenüber audiovisuell kongruenten Reizen und Standbildern deuten auf Unterschiede in der Integration der auditiven und visuellen Sprachinformation je nach Art der dargebotenen visuellen Zusatzinformation hin. Diese stellt in unserer Studie nur in der kongruenten Darbietung einen Mehrwert dar. Dagegen bietet die unspezifische Mundöffnung vor Einsetzen der auditiven Sprache einen Hinweisreiz darauf, dass der relevante Sprachreiz beginnen wird, ein unterstützendes Lippenlesen ist währenddessen aber nicht

möglich. Diese Annahme ist konsistent mit früheren Befunden, die zum einen zeigten, dass die N1-Amplitude durch visuelle Sprachreize moduliert wird, die der auditiven Information vorausgehen [8]. Zum anderen konnte bereits gezeigt werden, dass die P2 sensitiv auf die phonetische Verbindung beider Modalitäten reagiert [5].

Mit Blick auf die P2-N2 zeigten sich in beiden Altersgruppen verringerte Amplituden bei audiovisuell kongruenter im Vergleich zu visuell unspezifischer Sprachinformation. Darüber hinaus zeigte sich bei Älteren eine Reduktion der P2-N2 beim Standbild im Vergleich zu visuell unspezifischer Information. Eine mögliche Erklärung hierfür ist die Assoziation der P2 mit der sprachspezifischen Verknüpfung auditiver und visueller Information. Eine geringere P2-Amplitude ließ sich dementsprechend bei der effektiven Verarbeitung audiovisuell kongruenter Sprachreize beobachten [5, 6]. Bei den Älteren zeigte sich außerdem eine deutliche Reduktion der N2, welche wiederum mit einem altersbedingten inhibitorischen Defizit in Verbindung gebracht wird [1, 2]. Die Ergebnisse stützen zudem die Annahme von [4], dass ergänzende visuelle Information älteren Probanden möglicherweise dabei helfen kann, dieses Defizit zu kompensieren.

Zusammenfassend konnten wir zeigen, dass die im Alter häufig eingeschränkte Fähigkeit, Sprache in schwierigen Hörsituationen zu verstehen, durch das Bereitstellen visueller Information unterstützt werden kann. Die in natürlichen Sprachumgebungen auftretenden, häufig unerwarteten Wechsel zwischen Sprechern scheinen die Vorteile audiovisueller Sprache aber ins Gegenteil zu kehren. Besonders älteren Personen scheint es Schwierigkeiten zu bereiten, die dann auftretende Inkohärenz zwischen visueller und auditiver Sprachinformation aufzulösen.

Literatur

[1] Getzmann, S., Falkenstein, M. & Wascher, E.: ERP correlates of auditory goal-directed behavior of younger and older adults in a dynamic speech perception task. Behavioural Brain Research 278 (2015), 435-445

[2] Stothart, G. & Kazanina, N.: Auditory perception in the aging brain: the role of inhibition and facilitation in early processing. Neurobiology of Aging 47 (2016), 23-34

[3] Bronkhorst, A. W.: The cocktail-party problem revisited: early processing and selection of multi-talker speech. Attention, Perception, and Psychophysics 77 (2015), 1465-1487

[4] Winneke, A. H. & Phillips, N. A.: Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. Psychology and Aging 26 (2011), 427-438

[5] Baart, M., Stekelenburg, J. J. & Vroomen, J.: Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia 53 (2014), 115-121

[6] Baart, M.: Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. Psychophysiology 53 (2016), 1295-1306

[7] Campbell, R.: The processing of audio-visual speech: empirical and neural bases. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (2008), 1001-1010

[8] Van Wassenhove, V., Grant, K. W. & Poeppel, D.: Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences 102 (2005), 1181-1186

[9] Klucharev, V., Möttönen, R. & Sams, M.: Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Cognitive Brain Research 18 (2003), 65-75