Travaux neuchâtelois de linguistique, 2013, 59, 167-181

Audiovisuelle Sprechererkennung durch linguistisch naive Personen

Sibylle SUTTER und Volker DELLWO
Phonetisches Laboratorium, Universität Zürich

Human speech perception is not only based on acoustic speech signals but also on visual cues like lip or jaw movements. Based on this assumption we used a between-subject design to test listeners' speaker identification ability in a voice line-up after they were familiarized with a speaker under either of the following condition: (a) visual and degraded acoustic information, (b) degraded acoustic information only, and (c) visual information only. The results from this experiment indicate that listeners are able to perform the identification task to a considerable degree under all three experimental conditions. We conclude that listeners' identification ability of speakers based on degraded acoustic material is about as good as their identification ability based on visual speech cues. The combination of acoustic and visual cues does not enhance listeners' performance.

Gesprochene Sprache wird nicht nur auditiv sondern auch visuell wahrgenommen (audiovisuelle Sprachwahrnehmung; Rosenblum, 2005). Dies ist vor allem für die Sprachverständlichkeit von grosser Bedeutung. In Kommunikationssituationen, in denen Hörer die Sprecher nicht nur hören, sondern auch sehen können, ist die Verständlichkeit des Signals besonders unter schwierigen Hörbedingungen (Hintergrundlärm oder störende Sprachsignale von anderen Sprechern gesprochen) deutlich besser. Doch kann das visuelle Sprachsignal auch die Sprecheridentifizierungsleistung von Hörern verbessern? Eine solche Annahme ist aus zweifacher Sicht naheliegend. Als Hörer haben wir meist langjährige Erfahrung damit, welche Stimmen von welchen Sprechapparaten gebildet werden. Wenn wir das Bild einer Person sehen, die wir nicht kennen, haben wir daher häufig schon eine gewisse Vorstellung darüber, wie deren Stimme klingen könnte. Besser noch geht dies, wenn wir die Artikulationsbewegungen beobachten können. Dies ist schon durch eindeutige experimentelle Ergebnisse belegt. Mit der *Facial-Point-Light* Methode zeigte Rosenblum *et al*. (2006), dass Betrachter eine ihr vertraute Person aufgrund von ihrer Artikulationsbewegungen identifizieren können. Bei dieser Methode werden den zu erkennenden Personen leuchtende Punkte auf sichtbare Artikulatoren wie Lippen und Kiefer geklebt. Die Präsentation der Personen im Experiment erfolgt dann im Dunklen, weshalb das eigentliche Gesicht nicht sichtbar ist. Weitere Befunde für die Annahme das Individuen aufgrund von im Artikulationsprozess verwendeten Bewegungen erkannt werden können, zeigen Kamachi *et al*. (2003). In dieser Studie mussten Hörer einer Stimme einen von zwei präsentierten Sprechern (in Form eines

Videos) zuordnen, von dem sie glaubten, dass dies der Sprecher ist, der das Stimmsignal produziert hat. Die Studie zeigte, dass Hörer dies signifikant über einer Zufallsverteilung konnten, obwohl die eigentliche Effektgrösse eher klein war.

Zusammenfassend lässt sich sagen, dass man aufgrund vorausgehender Studien davon ausgehen sollte, dass Artikulationsbewegungen, ähnlich wie bei der Sprachverständlichkeit dazu beitragen sollten, dass auch Sprecher besser wiedererkannt werden. Mit anderen Worten könnte dies bedeuten: Sollten wir die Stimmen von Sprechern akustisch und visuell erlernen, können wir diese Stimmen möglicherweise besser memorisieren, als Stimmen von Sprechern, mit denn wir nur akustisch familiarisiert wurden. Es könnte weiterhin die Möglichkeit bestehen, dass wir Sprecher auch dann aufgrund ihrer Stimme wiedererkennen, wenn wir nur visuell mit ihnen familiarisiert wurden. Diese Hypothesen wurden in der vorliegenden Arbeit getestet.

1. Audiovisuelle Sprachwahrnehmung

1.1 Face Overshadowing Effect

Ergebnisse bisheriger Studien lassen annehmen, dass beim visuellen Stimulus eine Differenzierung vorgenommen werden muss. Cook & Wilding (1997) zeigten, dass sich das visuelle Signal auch negativ auf die auditive Wahrnehmung auswirken kann und zwar dann, wenn das Signal als statisches und nicht bewegtes Bild präsentiert wird. In ihrem Artikel ist vom *Face Overshadowing Effect* (FOE) die Rede. Damit ist eine Art *Überblendungseffekt* des visuellen Stimulus auf den auditiven gemeint. Die Studie zeigt, dass sich der visuelle Stimulus störend auf die auditive Wahrnehmung auswirkt. Bei gleichzeitiger Präsentation von auditivem und visuellem Stimulus verschlechtert sich in ihren Experimenten das Erinnerungsvermögen der Probanden in Bezug auf das Gesprochene, dass sie sich merken sollten. Diese Ergebnisse sind nicht unplausibel, denn die Präsentation von visuellen Signalen während der Familiarisierung mit der Stimme eines Sprechers kann durchaus dazu führen, dass die Aufmerksamkeit des Hörers vom akustischen Signal zugunsten des visuellen Signals gelenkt wird. Bei einer Wiedererkennung aufgrund von rein auditiver sprachlicher Information, fehlen dann dem Hörer wichtige Informationen (Paul Iverson, persönliche Kommunikation). Auch Legge, Grosmann & Pieper (1984) finden keinen positiven Einfluss des visuellen Stimulus auf die auditive Sprachperzeption. Den Versuchspersonen wird nebst der Präsentation eines auditiven Stimulus' ebenfalls ein statisches Bild des Sprechers gezeigt. Hier ist jedoch durchaus die Schlussfolgerung möglich, dass eine Steigerung der Sprecheridentifizierungsleistung durch

audiovisuelle Information nur bei einem bewegten visuellen Stimulus erfolgt.

1.2 Bessere auditive Verständlichkeit durch visuelle Unterstützung

Zahlreiche Studien belegen, dass der bewegte visuelle Stimulus einen positiven Effekt auf die Sprachwahrnehmung hat (siehe auch Einleitung). Durch die gleichzeitige Präsentation von auditiven und visuellen Stimuli wird die Sprachverständlichkeit gesteigert (vgl. dazu bspw. Neti, Iyengar, Potamianos, Senior & Maison, 2000; Sheffert & Olson 2004). Dieses Phänomen ist jedoch nicht nur in wissenschaftlichen Studien sondern auch aus dem Alltag bekannt. Steht man an einem Bahnhof mit vorbeifahrenden Zügen oder in einer Disco mit lauter Musik, ist es einfacher, jemanden zu verstehen, wenn man dessen Gesicht respektive dessen Lippenbewegungen sehen kann. Die akustischen Signale können dabei fast gänzlich im Lärm untergehen und man ist trotzdem fähig, zu erahnen, was der andere einem mitteilen möchte. Die Studie von Neti *et al*. (2000) geht von der Annahme eines positiven Einflusses des visuellen Stimulus' auf den auditiven aus. Durch die visuelle Unterstützung wird die Sprachverständlichkeit gesteigert. Neti et al. erläutern in ihrem Artikel, wie dieses Phänomen für die Mensch-Computer-Interaktion genutzt werden kann. Auch in den Arbeiten von Sheffert & Olson (2004) zeigt sich der visuelle Stimulus unterstützend für die Sprachverständlichkeit.

1.3 Einfluss des auditiven Stimulus auf die visuelle Wahrnehmung

Alle diese Studien untersuchen den Effekt der visuellen Wahrnehmung auf die auditive. Bleibt die Frage, ob auch umgekehrt ein Effekt gefunden werden kann. Das heisst, ob auch der auditive Stimulus Einfluss auf die visuelle Wahrnehmung nimmt. In der Untersuchung von Joassin, Maurage, Bruyer, Crommelinck & Campanella (2004) wird diese Frage untersucht. Sie konzentrieren sich auf die Beeinflussung des auditiven Stimulus auf die visuelle Wahrnehmung und kommen zum Schluss, dass die Informationen von Gesicht und Stimme nicht gleich schnell verarbeitet werden, was zu einer gegenseitigen Beeinflussung der Perzeptionsarten führt. Für das Vorhandensein einer gegenseitigen Beeinflussung der auditiven und visuellen Wahrnehmung wird als Evidenz oft das Paradigma-Beispiel der audiovisuellen Sprachverarbeitung herangeführt: Der McGurk Effekt. In der Studie von McGurk & MacDonald (1976), die diesen Effekt beschreibt, wird gezeigt, dass sich visuelle und auditive Stimuli gegenseitig beeinflussen. Das akustische Sprachsignal wird durch die gleichzeitige Beobachtung der Lippenbewegung, auch wenn dies unbewusst geschieht, beeinflusst.

1.4 Ziel und Motivation

Die Wissenschaft scheint sich offensichtlich uneinig zu sein, ob die Integration von auditiver und visueller Wahrnehmung förderlich oder hemmend für unterschiedliche Sprachwahrnehmungsaufgaben ist. Mitunter scheint die Erkennungsleistung von Hörern auch abhängig davon zu sein, ob der visuelle Stimulus (d.h. zum Beispiel der Kopf eines Sprechers) nur als statisches oder bewegtes Bild präsentiert wird. Die Studien von Rosenblum et al. unterstreichen das Vorhandensein *cross-modaler* Informationen in Stimme und Gesicht. Sie zeigen, dass Informationen der Stimme einer Person in den Artikulationsbewegungen zu finden sind und das daher einem Hörer bekannte Personen nur aufgrund der Betrachtung ihrer Artikulationsbewegungen identifizierbar sind. Die auditive und visuelle Wahrnehmung stehen also in einem engen Zusammenhang. Es ist daher plausibel, dass idiosynkratrische Informationen eines Sprechers aus den Artikulationsbewegungen gelesen und zur Sprecheridentifizierung genutzt werden können. Cook & Wilding zeigen jedoch, dass sich visuelle Information negativ auf die Sprechererkennungsleistung auswirken kann.

In der vorliegenden Studie wurde die Sprecheridentifizierungsleistung von linguistisch-phonetisch naiven Hörern mit und ohne auditive Information mittels einer sogenannten *Voice Parade* getestet. Hörer wurden zunächst mit der Stimme eines Sprechers familiarisiert, welche sie darauf aus einer randomisierten hintereinander abfolgenden Präsentation unterschiedlicher akustischer Signale von Stimmen wiedererkennen mussten. Die Familiarisierung erfolgte in unserem Fall unter drei verschiedenen Bedingungen:

(a) Die Hörer wurden nur mit dem auditiven Signal der Stimme des Sprechers familiarisiert. (Audio Kondition [A])

(b) Die Hörer wurden mit dem auditiven und visuellen (Video des Sprecherkopfs) Signal des Sprechers familiarisiert. (Audio-Video Kondition [AV])

(c) Die Hörer wurden nur mit dem visuellen Signal des Sprechers familiarisiert. (Video Kondition [V])

2. Methoden

2.1 Versuchspersonen

65 Probanden nahmen Teil, davon 53 mit Schweizerdeutsch und 12 mit Hochdeutsch als Muttersprache. Je 22 Versuchspersonen wurden in der A- und in der AV-Kondition getestet sowie 21 in der V-Kondition. Die Rekrutierung erfolgte an der Universität Zürich. Daher handelt es sich bei

den Versuchspersonen zum grossen Teil um Studierende. Alle gaben an, keine einschränkenden Seh- oder Hörprobleme zu haben. Die Teilnahme wurde nicht vergütet.

2.2 Material & Stimuli

Für die Stimulusproduktion wurden neun männliche Sprecher aufgenommen. Zwei Sprecher wurden nachträglich vom Experiment ausgeschlossen (undeutliche Aussprache und technische Probleme bei der Aufnahme). Die verbleibenden sieben Sprecher wurden für das Experiment verwendet. Alle Sprecher (Altersumfang: 20 bis 36) sind alle Schweizerdeutsch Muttersprachler (vier Aargauerdeutsch (Nordost Aargau), zwei Zürichdeutsch, einer eine Mischung aus Aargauer (Nordost Aargau)- und Zürichdeutsch und einer eine Mischung aus Aargauer (Nordost Aargau)- und Solothurnerdeutsch). Die Sprecher lasen einen Schweizerdeutschen Text (vgl. Appendix I) vor, der einen Kidnapping-Anruf simulierte. Die gesamte Aufnahme eines Sprechers dauerte rund 30 Sekunden.

Die Aufnahme der Sprecher erfolgte mit einer Sony Handycam 10.2 Mega Pixels. Sie wurde dem Sprecher auf einem Stativ im Abstand von ca. 1.5 Meter frontal gegenübergestellt. Es wurden nur die Gesichtspartien aufgenommen, so dass der Sprechapparat im Fokus der Aufnahme lag. Der Sprecher sass während der Aufnahme und es wurde darauf geachtet, dass möglichst keine Bewegungen mit dem Kopf gemacht wurden. Der Sprecher konnte den Text, welcher neben der Kamera aufgehängt war während der Aufnahme ablesen. Der Ton wurde mit einem Zoom H2 Handy Recorder aufgezeichnet. Den Probanden wurde folglich ein MPEG-4 Video in höchster Qualität, einer Basisbildrate von 24, Bitrate 6400 kBit/s in der Grösse von 1920x1080 vorgespielt. Die Mono-Tonspur wurde als 32-Bit-Integer-Datei (Little Endian) mit einer Abtastrate von 48.000 kHz aufgenommen.

Familiarisierungsmaterial: Die komplette Audio-, Audio-Video- und Videoaufnahmen (ca. 30 Sekunden) wurden dreimal hintereinandergeschnitten und als Familiarisierungsmaterial für die jeweiligen Konditionen verwendet. Pilotstudien zeigten, dass die Identifizierungsaufgabe für die A-Kondition mit qualitativ hochwertigen Aufnahmen zu einem Deckeneffekt führte (alle Versuchspersonen können die Aufgabe zu fast 100% lösen). Aus diesem Grund wurde das audio-Familiarisierungssignal degradiert, (a) durch einspielen eines Hintergrundgeräuschs mit +3dB SNR (*Multi Speaker Babble*[1]*;* 100 Sprecher in einer Kantine) und (b) durch telefonähnliche Bandpassfilterung (Pass

[1] Institute for Perception-TNO (1990): http://spib.rice.edu/spib/data/signals/noise/babble.html [Stand: 20.02.2012]

zwischen 300 bis 3500 Hz). Die durchschnittliche Intensität der Aufnahmen wurde auf 70 dB vereinheitlicht.

Voice-Parade Material: Für die Voice-Parade wurde der gelesene Text eines jeden Sprechers in 10 Sätze unterteilt. Die Parade bestand aus 140 Stimuli (10 Sätze x 7 Sprecher x 2 Durchgänge). Die Dauer der Stimuli betrug zwischen drei und vier Sekunden. Die beiden Durchgänge wurden für jeden Hörer individuell randomisiert und hintereinander präsentiert (*permute balanced*).

2.3 Ablauf

Nach der Familiarisierung mit einem Zielsprecher muss der Hörer aus einer ihm unbekannten Anzahl verschiedener Sprecher die Zielstimme wiedererkennen. Die Versuchspersonen werden randomisiert in drei Gruppen aufgeteilt und mit einem *Between Subject Design* getestet (Teilnahme jeweils nur an einer der drei Familiarisierungskonditionen: A, AV oder V). In der Familiarisierungsphase werden die Probanden mit einem Zielsprecher in einer der drei Kondition familiarisiert (Gruppe A: nur Audiosignal, Gruppe B: Audio-videosignal, Gruppe C: nur Videosignal). Während der Testphase müssen die Versuchsgruppen den Zielsprecher aus einer ihnen unbekannten Anzahl Sprecher wieder-erkennen. Allen drei Gruppen wird die identische Voice-Parade vorgespielt. Die Voice-Parade wird mittels Praat präsentiert. Den Hörern wird ein Stimulus vorgespielt, worauf sie auf einem Computerbildschirm mittels einer Maus eine Auswahl zwischen "ja, das ist der Sprecher" und "nein, das ist er nicht" treffen müssen. Die Antwort werden differenziert mit: "sicher", "weiss nicht recht", "nur geraten" (vgl. Appendix II).

Damit ein allfälliger Effekt nicht auf eine spezifische Stimme reduziert werden kann, werden die Versuchspersonen randomisiert mit einer von drei unterschiedlichen Sprechern (aus der Gruppe von sieben) familiarisiert.

Für die Familiarisierung werden die Versuchspersonen lediglich darauf hingewiesen, sich die Stimme gut einzuprägen. Sie bekommen keine weiteren Informationen zum Experiment. Nach dem dritten Anhören beziehungsweise Ansehen des Zielsprechers werden die Versuchspersonen darüber informiert, dass es sich um eine Sprecheridentifizierungsaufgabe handelt. Es wird eine Demo gezeigt, wie sie in der Folge ihre Antworten abgeben müssen. Ein Stimm-Sample aus dem Experiment wird vorgespielt, wobei die Hörer auf die Frage antworten müssen: "War das der Sprecher von vorhin?". Zur Auswahl stehen die Antworten "ja" und "nein", mit der Differenzierung "sicher", "weiss nicht recht", "nur geraten". Wenn die Hörer nach den zwei Demo-Samples keine Fragen haben, können sie einen Fragebogen zu den Personalien ausfüllen. Jeder Versuchsperson wird eine ID zugewiesen, mit der das Experiment anonymisiert wird. Erhoben werden

die sozialdemographischen Daten: Alter, Geschlecht, Muttersprache (ggf. zweite Muttersprache), Dialekt, Wohnkanton, Bildungsstand (Abschluss: Doktorat, Master, Bachelor, Matura, Volksschule). Ausserdem müssen die Versuchspersonen bestätigen, eingehend über das Experiment informiert worden zu sein, sowie freiwillig und aus eigenem Willen an der Untersuchung teilzunehmen. Nach dem Ausfüllen des Personalien-Fragebogens können die Probanden selbstständig das Experiment starten. In der Folge werden ihnen die 140 Stimuli der Voice-Parade vorgespielt. Nach jeder Stimulus Präsentation geben die Hörer ihre Antwort. Nach erteilen der Antwort wird umgehend der nächste Stimulus abgespielt. Bei der V-Kondition wurden alle Probanden gleichzeitig getestet. Die Familiarisierung mit dem Zielsprecher erfolgte mit der Übertragung auf eine Grossleinwand in einem Hörsaal der Universität Zürich. Die Antworten werden nicht wie beim Hauptexperiment von den Hörern direkt im Computer eingetragen, sondern manuell auf Fragebögen abgegeben.

2.4 Statistische Methoden

Die statistischen Analysen wurden mit Praat und SPSS (SPSS, Version 18.0, Chicago, Illinois) für Windows durchgeführt. Die erhobenen Daten wurden zuerst auf Normalverteilung durch optische Einschätzung beurteilt. Das Signifikantsniveau wird auf $p \leftarrow 0.05$ festgelegt. Als unabhängige Variable dient die Probandengruppen der A-, AV-Kondition sowie V-Kondition. Als abhängige Variable wird die *Identifizierungsperformanz IDP* der Hörer definiert. Diese wird durch die Richtigkeit der Antworten statistisch ermittelt.

Die Hörerperformanz werden mit %Correct und A' aus der Signaldetektionstheorie ermittelt. %Correct ermittelt sich aus dem Durchschnitt der korrekten Identifizierungen (Zielsprecherpräsentation mit ‚ja' beantwortet) und der korrekten Rückweisungen (Dummypräsentation mit ‚nein' beantwortet). Die daraus sich ergebende Zufallsverteilung liegt stets bei 50% korrekt. A' ist ein nicht-parametrisches Mass der Sensitivität einer Versuchsperson. Es berechnet sich aus der Fläche unterhalb der sogenannten ROC Kurve (receiver operating characteristics), die die Verteilung der Trefferrate über die Verteilung der Fehlerrate darstellt. Die Motivation für die Berechnung von A' ist, dass eine mögliche Hörerneigung zu *ja* oder *nein* Antworten in diesem Mass deutlich reduziert bis nicht mehr vorhanden ist.

3. Resultate

Die Frage ist, wie sich die audiovisuelle Wahrnehmung auf die Sprecheridentifizierungsperformanz von naiven Hörern beziehungsweise Betrachtern auswirkt. Es wird angenommen, dass die Versuchspersonen

der AV-Kondition am besten abschneiden im Vergleich zu den anderen beiden Konditionen A und V, da sie sowohl akustische Informationen aus der Stimme als auch visuelle Informationen aus den Artikulationsbewegungen nehmen können. Für die A- und V-Kondition wird ein ähnliches Resultat suggeriert, da bei der A-Kondition wesentliche Informationen im Frequenz-Bereich durch die Degradierung verloren gehen und der Hörer nur mit den Informationen aus dem Zeitbereich die Stimme beurteilen kann. Bei der V-Kondition kann der Betrachter des tonlosen Videos diese Informationen des Zeitbereichs aus den Artikulationsbewegungen ablesen.

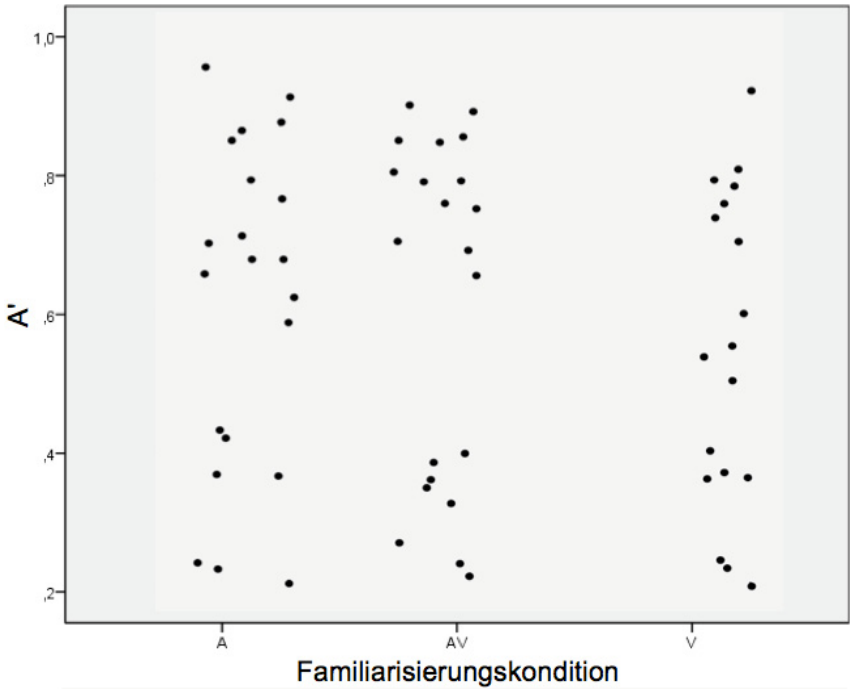

Abb. 1: Streudiagramm der A'-Werte aller Versuchspersonen

Wie das Streudiagramm in Abbildung 1 zeigt, verteilen sich die Ergebnisse der Probanden sehr stark. Jeder Punkt in der Graphik entspricht einer Versuchsperson respektive ihrer *IDP*. Klar ersichtlich wird in jeder Kondition die Zweiteilung der Probanden. Sie gruppieren sich sowohl ober- als auch unterhalb des Werts 0.5 für den A'. Um 0.5 zeigt sich eine grosse Lücke. Vor allem in den Konditionen A und AV scheinen die Versuchspersonen entweder erfolgreich den Test lösen zu können (jene die sich oberhalb von 0.5 gruppieren) oder Mühe zu haben, den Zielsprecher zu identifizieren (Punkte die sich unterhalb von 0.5 gruppieren).

Nach optischer Beurteilung des Streudiagramms kann ausgesagt werden, dass es stark hörerbedingt ist, ob die Aufgabe lösbar ist oder nicht. In allen drei Konditionen gibt es eine Gruppe von Versuchspersonen, die die Aufgabe recht gut können und eine Gruppe, die die Sprecheridentifizierungsaufgabe nicht oder nur schlecht lösen. In der A-Kondition erreichen rund 76 % einen A' > 0.5. Bei der AV-Kondition sind es

gar 80 %. Der A' liegt dabei bei beiden Konditionen im Schnitt bei 0.62 bei einer Standardabweichung von 0.24 (A) sowie 0.25 (AV-Kondition). Von den 21 Versuchspersonen der V-Kondition erreichen 13 einen A' > 0.5, was rund 60% der Teilnehmenden entspricht. Acht Probanden erreichten einen A' < 0.5 (38.09%) und können somit die Aufgabe nicht lösen.

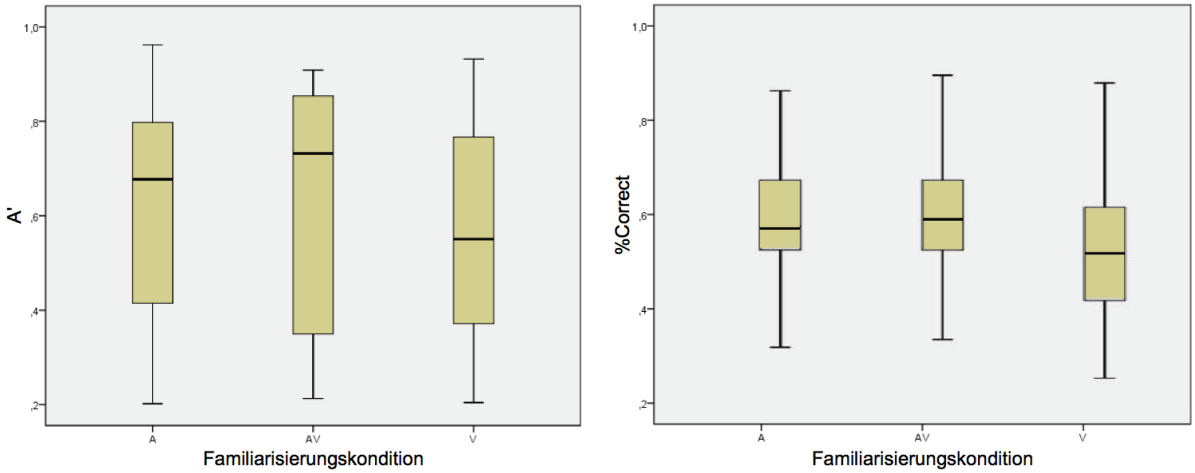

Abb 2: A' und %Correct der drei Versuchskonditionen

Wie die Boxplots links in Abbildung 2 zeigen, streuen sich die Ergebnisse enorm, während der Median in den Konditionen A und AV klar über 0.5 liegt und sich in der V-Kondition mehr oder weniger auf dem Wert der Zufallswahrscheinlichkeit einpendelt. Die Grafik zeigt deutlich die enorm starke Streuung in allen drei Konditionen. Weniger deutlich als im Streudiagramm zeigt sich hier die grosse Lücke (vor allem in den Konditionen A und AV) um den Zufallswahrscheinlichkeitswert von 0.5 herum. Es ist aber sehr schön ablesbar, dass sich die Streuung vom einen (1.0) zum anderen Extrem (0.0) zieht was die hohe Variabilität der Performanz der Versuchspersonen zeigt. Vergleicht man das Resultat der A'-Werte mit jenem von %Correct, zeigt sich ein interessantes Bild. Während beim A' die Sensitivität der Versuchspersonen ermittelt wird, zeigt %Correct die Zusammenfassung der korrekten Antworten. Interessant dabei ist zu erkennen, dass bei %Correct alle drei Versuchskonditionen über der Zufallswahrscheinlichkeit liegen. Und noch mehr: Die IDP scheint in allen drei Konditionen gleich zu sein.

Die erhobenen Variablen (Alter, Geschlecht, Ausbildung, Muttersprache beziehungsweise Dialektnähe) zeigen keine Unterschiede in Bezug auf die IDP der Probanden. Es muss davon ausgegangen werden, dass diese Variablen keinen Einfluss haben auf die Identifizierungsfähigkeit der Probanden in diesem Experiment.

4. Diskussion

Die drei Zielsprecher wurden mehr oder weniger in demselben Masse wiedererkannt. Die Unterschiede der korrekten Antworten innerhalb der drei Zielsprecher sind minim und nicht signifikant. Die feine Tendenz eines Sprechers, der etwas schlechter identifiziert wurde, im Vergleich zu den anderen beiden Zielsprechern, zeigt, dass es nicht nur hörerbedingt ist, wie gut man eine Stimme wiedererkennen kann, sondern dass das Ergebnis auch von der Stimme des Zielsprechers abhängig sein kann. Möglicherweise kann man sich je nach Stimmcharakteristika eine Stimme besser oder schlechter merken. Im vorliegenden Experiment handelt es sich zudem um degradierte Tonaufnahmen. Beim Anhören der Stimmsamples sind klare Unterschiede zu erkennen. Im Vergleich zu den gut identifizierten Zielsprechern ist es akustisch viel schwieriger, den dritten Sprecher aus dem *Babble Speech Noise* heraus zu hören. Möglicherweise wird deshalb dieser Sprecher am schlechtesten identifiziert. Dass bei diesem Sprecher eine etwas höhere IDP in der AV-Kondition ablesbar ist, im Vergleich zur A-Kondition, könnte daran liegen, dass der visuelle Stimulus sich positiv auf den auditiven auswirkt. Durch den visuellen Stimulus wird das Gesagte verständlicher, beziehungsweise kann die Stimme des Sprechers besser herausgehört werden, weil man durch das Lippenlesen dem Gesagten besser folgen kann. Bei den anderen beiden Zielsprechern tritt dieser Effekt nicht ein. Bei ihnen ist der Wert der korrekten Antworten in der A-Kondition leicht höher als jener in der AV-Kondition. Dieses Ergebnis spricht gegen den von Cook & Wilding definierten *FOE* (vgl. Cook & Wilding, 2001), unterstreicht jedoch die Resultate von Sheffert & Olson (2004), die den Einfluss des visuellen Stimulus auf die auditive Wahrnehmung positiv werten. Vergleicht man alle drei Konditionen miteinander kann erstaunliches festgestellt werden: Die Versuchspersonen aller drei Konditionen schneiden ungefähr gleich ab. Das heisst, dass ein degradiertes Tonsignal für den Hörer mehr oder weniger denselben Informationsgehalt liefert, den er zur Identifizierung der Stimme braucht, wie aus den Artikulationsbewegungen gelesen werden kann. Fügt man beide Konditionen zusammen, werden die Informationen jedoch nicht akkumuliert. Auch die Versuchspersonen der AV-Kondition schneiden in demselben Masse ab. Sie können den Vorteil nicht nutzen, sowohl das akustische als auch das visuelle Signal zu hören und zu sehen. Es ist jedoch auch kein Überblendungseffekt des einen Stimulus auf den anderen auszumachen, was zu einer schlechteren Performanz geführt hätte.

Die Verteilung der Ergebnisse zeigt, dass die Performanz der Hörer in diesem Experiment nicht von deren muttersprachlichen Dialekten abhängig ist. Die Resultate zeigen ebenfalls, dass die *IDP* bei der Sprecheridentifizierung unabhängig davon ist, mit welcher Kondition das Experiment gelöst wird. In den einzelnen Dialektgruppen schneiden alle

Probanden in allen Kategorien ungefähr gleich ab. Folglich scheint in diesem Experiment der Dialekt keinen Einfluss auf die Performanz der Hörer zu haben. Interessant wäre an diesem Punkt weiter zu testen, ob dasselbe Resultat erzielt würde, wenn das Experiment mit Hörern durgeführt würde, die kein Schweizerdeutsch verstehen. Es sind dabei verschiedene Möglichkeiten in Betracht zu ziehen. Eine bessere Performanz der Versuchspersonen, die die Sprache der Sprecher nicht verstehen, könnte so interpretiert werden, dass man sich besser auf die Stimmmerkmale konzentrieren kann, wenn man nicht durch den Inhalt des Gesagten abgelenkt ist. Fiele das Resultat genau umgekehrt aus, würde das wohl daran liegen, dass man sich eine Stimme besser merken kann, wenn man mit der gesprochenen Sprache vertraut ist. Beide Ergebnisse werden als plausibel betrachtet, müssten jedoch erst in einem weiteren Schritt getestet werden.

Bei der geringen Anzahl Männern, die getestet wurde, ist es nicht möglich, eine signifikante Aussage zur geschlechtsspezifischen Performanz der Versuchspersonen zu machen. Nach den eingehend diskutierten Studien gibt es aber keinen Grund zur Annahme, dass sich die Identifizierungsfähigkeit von Frauen und Männern unterscheiden sollte. Es wird zwar in verschiedenen Experimenten getestet, ob es Differenzen in der Identifizierungsperformanz gibt, wenn der Stimulus von einer weiblichen oder einer männlichen Stimme stammt (vgl. Sheffert & Olson, 2004; Joassin *et al.*, 2004; Belin *et al.*, 2000 und weitere), nicht aber ob sich die Performanz von männlichen und weiblichen Versuchspersonen unterscheidet. Eine neuropsychologische Studie von Lattner, Meyer & Friederici (2005), welche sich mit der Sprachperzeption beschäftigt, eingehend aber nicht berücksichtigt wurde, da der Fokus nicht auf der audiovisuellen Wahrnehmung liegt, sondern auf der Beurteilung der Natürlichkeit einer Stimme, hat diesen Aspekt jedoch berücksichtig und untersucht, ob es Unterschiede in der Performanz von Frauen und Männern gibt. Lattner et al. eruierten in ihrer Studie, inwiefern sich die Sprachperzeption von männlichen und weiblichen Hörern unterscheidet, die eine Stimme auf ihre Natürlichkeit beurteilen müssen (vgl. Lattner *et al.*, 2005). Den Versuchspersonen werden natürliche und manipulierte Stimmen von Frauen und Männern präsentiert. Die Studie zeigt, dass es 90% der Hörer gelingt, die natürlichen von den manipulierten Stimmen zu unterscheiden. Die gute Performanz der Hörer ist unabhängig davon, ob diese männlich oder weiblich sind. Hinsichtlich diesem Ergebnis und der Tatsache, dass keine der eingehend zitierten Untersuchungen den Gender-Aspekt in der Performanz von Stimmbeurteilungen berücksichtigt, darf angenommen werden, dass es keine signifikanten geschlechtsspezifischen Unterschiede in der Sprachperzeption gibt.

5. Conclusion

Die Hypothese, dass der visuelle Stimulus den auditiven bei der Sprechererkennung positiv beeinflusst, kann nicht signifikant gezeigt werden, eine feine Tendenz zu dieser Annahme ist jedoch aus den Resultaten ablesbar. Die vorliegende Untersuchung zeigt, dass es extrem hörerbedingt ist, wie gut man bei einer Identifizierung von Sprechern sowohl bei der auditiven, audiovisuellen als auch bei der visuellen Kondition abschneidet. Es scheint für einige Hörer der A-Kondition kein Problem zu sein, den Zielsprecher aus verschiedenen Tonaufnahmen herauszuhören. Es gibt aber eine ähnlich grosse Anzahl Hörer, die Probleme damit haben. Dasselbe Bild zeigt sich in der AV- und V-Kondition. Einige Versuchspersonen schneiden nach der Familiarisierung mit einem Video des Zielsprechers sehr gut ab und können ihn von den anderen Sprechern unterscheiden. Jedoch zeigt sich auch eine grosse Gruppe an Hörern, denen dies nicht gelingt. Die Versuchspersonen der AV-Kondition können die zusätzliche visuelle Information nicht zur Steigerung ihrer IDP nutzen. Ihre *IDP* ist nicht höher als jene der A-Gruppe. Die V-Kondition zeigt jedoch, dass auch mit einem fehlenden Tonsignal eine Stimmidentifizierung möglich ist. Nimmt man %Correct, zeigt sich, dass die visuellen Informationen, die durch die Artikulationsbewegungen gesendet werden, ungefähr gleich viel Stimminformation beinhaltet wie ein degradiertes Tonsignal. Die Kombination aus beidem, ein degradiertes Tonsignal und ein Videosignal der Artikulationsbewegungen, liefert jedoch nicht die *doppelte* Information. Das zeigt das Ergebnis von %Correct, welches in allen drei Konditionen gleich zu sein scheint.

Die grosse Variabilität des Ergebnisses zeigt jedoch, dass dieses Experiment sehr hörerspezifisch ist. Es müsste ein Versuchsdesign konzipiert werden, bei dem jeder Proband in jeder Kondition getestet wird. Dabei muss jedoch beachtet werden, dass sich beim Hörer kein *Lerneffekt* einstellt sowohl in Bezug auf die Stimuli als auch in Bezug auf den Versuchsablauf. Ein weiterer spannender Punkt, der im Bereich der audiovisuellen Sprachwahrnehmung untersucht werden könnte, ist die Frage, ob man sich besser auf die Stimme eines Sprechers konzentrieren kann, wenn man dessen Sprache nicht versteht oder ob es genau umgekehrt der Fall ist. Interessant wäre dabei zu sehen, ob sich Unterschiede in der *IDP*_zeigen zwischen den drei Konditionen. Eine weitere Variable, die mit dem vorliegenden Experiment nicht getestet wurde, ist die Frage nach dem Langzeitgedächtnis. Die Probanden lösten die Sprecheridentifizierungsaufgabe unmittelbar nach der Familiarisierung mit dem Zielsprecher. Wie lange kann man sich eine einmalig gehörte Stimme merken? Und ist es möglich eine Stimme nach einer gewissen Zeitperiode wieder zu erkennen, wenn die Stimme beim ersten Kontakt degradiert oder verstellt war?

Mit der vorliegenden Untersuchung kann gezeigt werden, dass *cross-modale* Information von Stimme und Gesicht bestehen müssen. Die eingangs in den untersuchten Studien aufgezeigte positive Beeinflussung des visuellen Stimulus auf die auditive Wahrnehmung kann nicht signifikant bestätigt werden. Ein *FOE* des visuellen Stimulus auf den auditiven wurde jedoch nicht gefunden. Die Resultate zeigen, dass die Information, welche aus den Artikulationsbewegungen gelesen werden können, mehr oder weniger jenen entsprechen, die aus einem degradierten Stimmsignal noch zu hören sind und für die Identifizierung dieser Stimme nötig sind. Verbindet man diese beiden Stimuli, indem sowohl ein akustisches sowie auch ein visuelles Signal gesendet wird, ist jedoch keine bessere Identifizierungsperformanz beim Rezipienten auszumachen. Die Stimuli scheinen sich nicht zu akkumulieren.

Mit der aufgezeigten Konzeptionierung des Experiments kann festgestellt werden, dass die Performanz beim Wiedererkennen von Stimmen extrem hörerspezifisch ist. Mit den erhobenen sozialdemographischen Daten der Versuchspersonen kann keine signifikante Aussage darüber gemacht werden, was diese hörerspezifische Performanz erklären lässt. Weder Alter, Geschlecht, Muttersprache noch Dialekt scheinen einen bedeutsamen Einfluss auf die Sprecheridentifizierungsfähigkeit zu haben. Weitere Studien sind nötig, um die Komponenten zu finden, welche es Hörern ermöglicht, eine familiarisierte Stimme aus verschiedenen Stimmaufnahmen wieder zu erkennen. Ausserdem muss das Experiment so konzeptioniert werden, dass jeder Proband in allen Konditionen getestet wird, um zu eruieren, ob innerhalb der Sprecher eine bessere Performanz in der AV-Kondition gegenüber der A- und V-Kondition auszumachen ist.

Danksagung

Die Autoren möchten sich bei Adrian Leemann und einem annonymen Gutachter für wichtige Änderungsvorschläge bedanken.

Bibliographische Angaben

Armstrong, H. A. & McKelvie, S. J. (1996): The effect of face context on recognition memory for voices. In: Journal of Experimental Psychology: General, 123(3), 259-270.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P. & Pike, B. (2004): Thinking the voice: neural correlates of voice perception. In: TRENDS in Cognitive Sciences, 8(3), 129-135.

Campanella, S. & Belin, P. (2007): Integrating face and voice in person perception. In: TRENDS in Cognitive Sciences, 11(12), 535-543.

Cook, S. & Wilding, J. (2001): Earwithness testimony: Effects of exposure and attention on the face overshadowing effect. In: British Journal of Psychology, 92(4), 617-629.

— (1997): Earwithness testimony 2: Voices, Faces and Context. In: Applied cognitive Psychology, 11(6), 527-541.

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M. & Campanella, S. (2004): When audition alters vision: an event-related potential study of the cross-modal interactions between faces and voices. In: Neuroscience Letters, 369, (2), 132-137.

Kamachi, M., Hill, H., Lander, K. & Vatikiostis-Bateson, E. (2003): ‚Putting the Face to the Voice': Matching Identity across Modality. In: Current Biology, 13, (19), 1709-1714.

Lattner, S., Meyer, M. E., Friederici, A. D. (2005): Voice Perception: Sex, Pitch, and the Right Hemisphere. In: Human Brain Mapping, 24, (1), 11-20.

Legge, G. E., Grosmann, C. & Pieper, C. M. (1984): Learning unfamiliar voices. In: Journal of Experimental Psychology: Learning, Memory, & Cognition, 10, (2), 298-303.

McAllister, H. A., Dale, R. H., Bregman, N. J., McCabe, A. & Cotton, R. (1993): When eyewitnesses are also earwitnesses: effects on visual and voice identifications. In: Basic and Applied Social Psychology, 14, 161-170.

McGurk, H. & MacDonald, J. D. (1976): Hearing lips and seeing voices. In: Nature, 264, (5588), 746-748.

Neti, C., Iyengar, G., Potamianos, G., Senior, A., & Maison, B. (2000): Perceptual interfaces for information interaction: joint processing of audio and viusal information for human-computer interaction. In: Processing of the International Conference on Spoken Language, 3, 11-14.

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale & S., Lee, J. (2006): Hearing a face: Cross-modal speaker matching using isolated visible speech. In: Perception & Psychophysics, 38, (1), 84-93.

Rosenblum, L. D. (2005): Primacy of Multimodal Speech Perception. In: David B. Pisoni & Robert E. Remez (Hg.), The Handbook of Speech Perception, Malden, Oxford, Victoria (Blackwell Publishing), 51-78.

Sheffert, S. M. & Olson, E. (2004): Audiovisual spreech facilitates voice learing. In: Perception & Psychophysics, 66, (2), 352-361.

Appendix I: Stimulus Text

Mir händ nach de Schuel uf ihri Tochter Lisa gwartet.

Wänn Sii sii läbend weder wänd haa, dänn losed Sii jetzt mal guet zue:

Morn, am sächsi, leged Sii en Koffer mit vierzgtuusig Franke bim Becker Huusmaa hinder d'Abfalltonne.

Sii werdet beobachtet. Chömmed Sii elläi.

Fahred Sii dänn mit Ihrem Auto in Richtig Wald.

Fahred Sii bis zu de Schrebergärte.

Stopped Sii det, stelled Sii de Motor ab und bliibed Sii im Wage sitze.

Sii ghöred dänn vo eus.

Mer bestimmed d'Spiilregle.

Und no öpis: kei Polizei.

Appendix II : Test Interface