Untersuchung zur Validität von Hörversuchen im Internet

Andreas Pysiewicz, Hauke Egermann

Fachgebiet Audiokommunikation, TU Berlin, E-Mail:andreas.pysiewicz@tu-berlin.de

Einleitung

Das Internet bietet Möglichkeiten, empirische Untersuchungen kosten- und zeiteffizient durchzuführen und dabei große, heterogene Stichproben zu generieren. Trotz einer weitgehenden Etablierung in einigen Wissenschaftsbereichen, existieren bis dato Vorbehalte gegenüber Web-basierten Hörversuchen und Onlinestudien im Bereich der auditorischen Wahrnehmungsforschung. Dies ist unter anderem der fehlenden Kontrolle über die Präsentationsmodalitäten (technisches Equipment, Störquellen etc.) geschuldet, die möglicherweise die Validität der Versuchsanlage gefährden. Zudem existieren bis dato nur wenige Studien, z.B. [1][2], die sich explizit mit dieser Problemstellung auseinandersetzten und im Web generierte und unter Laborbedingungen erhobene Daten gegenüberstellen.

Die durchgeführte Studie zielt auf diesen Missstand ab und versucht bisherige Befunde zu replizieren, die Internetexperimenten im Bereich der Musikwahrnehmung ein hohes Maß an interner Validität zusprechen. Ausgangspunkt war dabei die Hypothese, dass sich Einflüsse ausgewählter akustischer Eigenschaften (in Form von extrahierten Audio-Features) von Musikstücken auf ihre Eindrucksbewertung gleichermaßen im Web wie im Labor erfassen lassen. In einer Vergleichsstudie, die online (Testgruppe) und unter kontrollierten Laborbedingungen (Kontrollgruppe) durchgeführt wurde, bewerteten die Versuchsteilnehmer zufällig zugewiesene Musikstücke auf einer multidimensionalen Eindrucksskala. In einem Analysemodell sollte überprüft werden, ob sich diese Bewertungen voneinander unterscheiden, d.h. ob signifikante (Interaktions-)Effekte der Gruppenzugehörigkeit und der Audio-Features zu beobachten sind.

Methoden

Stichprobe: Die Rekrutierung von Teilnehmern für beide Experimente erfolgte über den studentischen E-Mail-Verteiler der TU Berlin. Am Laborversuch nahmen $N = 110$ Personen teil. Ihr durchschnittliches Alter lag bei 26,5 Jahren, ($SD = 5{,}6$). 37,3% der Testpersonen waren weiblich, 99,1% hatten einen höheren Bildungsstand (Abitur oder Hochschulabschluss), 39% gaben an, keine aktiven Musiker zu sein. Demgegenüber nahmen im Web $N = 1368$ Personen teil, die sich im durchschnittlichen Alter (26,3 Jahre, $SD = 7{,}6$), Geschlecht (37,5% weiblich) und Bildungsstand (96,5% mit Abitur oder Hochschulabschluss) unwesentlich von der Laborgruppe unterschieden. 51% der Online-Teilnehmer gaben an, nicht aktive Musiker zu sein. Die Stichproben wurden mittels *Propensity Scores* [3] auf Fallebene (Teilnehmermerkmale je gehörter Stimulus) statistisch gematcht, um maximale Homogenität zwischen Test- und Kontrollgruppe zu gewährleisten.

Stimuli: 20 Musikstücke aus dem Repertoire eines Internetradio-Dienstes (verschiedene Genres, hohe akustische Variabilität) wurden repräsentativ ausgewählt. Aus diesem Set wurde den Teilnehmern eine randomisierte Auswahl präsentiert.

Durchführung: Während die Abhörbedingungen (Wiedergabesystem, Ort der Teilnahme) im Web variierten, wurde das Wiedergabe-Setup für alle Teilnehmer im Labor konstant gehalten: Die Beschallung erfolgte hier über einen geschlossen Studio-Kopfhörer in einem den Standards für Abhörräume konformen Sound-Produktionsstudio. Der Versuchsablauf bestand für beide Gruppen aus drei Teilen: Der Einführung (Instruktionen, Motivationscheck, Soundcheck), den eigentlichen Musikbewertungen sowie einer abschließenden Befragung (zur Person und zur Teilnahme).

Messinstrument: Die Umsetzung erfolgte als Web-Anwendung, die den gesamten Versuchsablauf (von den Instruktionen bis zur Speicherung der Daten) bereitstellt. Die Hörerbewertungen wurden mithilfe von 27 Rating-Skalen erfasst, die aus bipolaren Adjektivpaaren (z.B. hell-dunkel, weich-hart) zur Eindrucksmessung mit einer vierstufigen Skalenausprägung bestanden.

Datenanalyse: Nach der Extraktion von Audio Features aus den verwendeten Stimulus-Signalen mittels MIRToolbox [3] konnte eine Verdichtung zu fünf unkorrelierten Hauptkomponenten als Prädiktoren für das statistische Analysemodell vorgenommen werden, welche die akustischen Merkmalen Intensität, Rhythmus, Timbre, Tongeschlecht und Tonhöhe repräsentieren. Zusätzlich wurde eine Dimensionsreduktion der 27 Rating-Items zu den fünf Hauptkomponenten Potenz (energisch, aufdringlich u.a.), Valenz (fröhlich, humorvoll u.a.), Variation (komplex, abwechslungsreich u.a.), Nüchternheit (rational, synthetisch u.a.), Unschärfe (verschwommen, durcheinander u.a.) als abhängige Bewertungsdimensionen vorgenommen.

Datenauswertung: Mithilfe linearer gemischter Regressionsmodelle (*linear mixed models*) wurde überprüft, ob sich der Einfluss der akustischen Merkmale auf die Musikbewertungen der Gruppen voneinander unterscheiden, d.h. ob signifikante (Interaktions-)Effekte der Gruppenzugehörigkeit und der akustischen Deskriptoren (Audio-Features) zu beobachten sind. Dabei wurden die fünf Audio-Feature-Komponenten sowie die Gruppenzugehörigkeit (Test- vs. Kontrollgruppe) als Prädiktoren für die Musikbewertungen verwendet. Der Einfluss der individuellen Stimuli und der Versuchspersonen wurden als Zufallseffekte im Modell berücksichtigt.

Ergebnisse

Die Modellauswertungen ergaben für drei der fünf Bewertungsdimensionen keine signifikanten Unterschiede zwischen Musikbewertungen im Labor und im Internet (Konfidenzniveau 95%). Für die Bewertungsdimensionen *Potenz* konnte ein signifikanter Effekt der Gruppenzugehö-

rigkeit aufgezeigt werden (b = -.12, t(332) = -3.42, p = .001), so auch für *Valenz* (b = .09, t(347) = 2.31, p = .02). In beiden Fällen unterschieden sich die Bewertungen der Laborteilnehmer signifikant von denen im Internet. Abbildung 1 zeigt die geschätzten B-Koeffizienten für diese beiden Dimensionen.

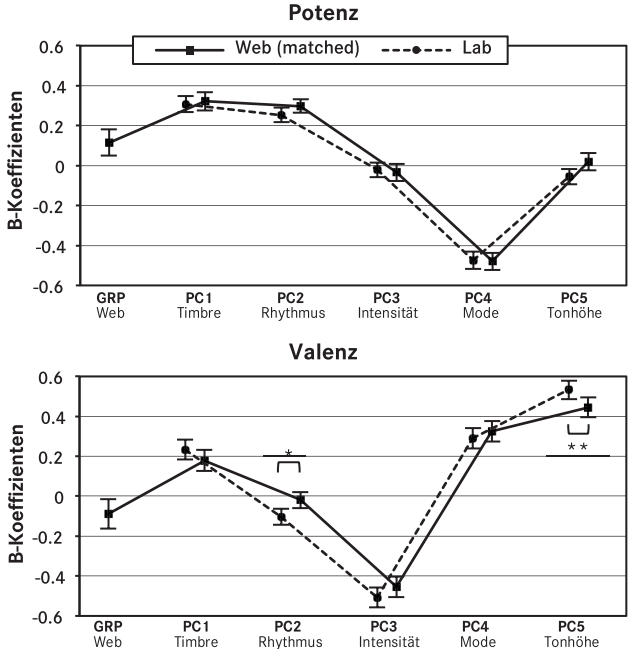

Abbildung 1: Geschätzte Regressionskoeffizienten für die Bewertungsdimensionen Potenz (oben) und Valenz (unten). Die Fehlerbalken zeigen das 95%-Konfidenzintervall, die gestrichelte Linie repräsentiert die Labor-Kontrollgruppe. Signifikante Interaktionen: * $p < .05$, ** $p < .01$.

Wie die Regressionskoeffizienten in Abbildung 1 aufzeigen, bewerteten die Teilnehmer im Internet Attribute der Dimension Potenz stärker als die Teilnehmer der Kontrollgruppe, sie neigten aber zu niedrigeren Valenz-Bewertungen als die Laborteilnehmer. Neben diesen Stimulus unabhängigen Unterschieden konnten für Valenz auch signifikante Interaktionseffekte zwischen der Gruppenzugehörigkeit und den Prädiktoren *Rhythmus* (b = .07, t(478) = 2.28, p = .02) und *Tonhöhe* (b = -.10, t(963) = -2.73, p = .006) beobachtet werden. Demnach zeigten die Laborteilnehmer negativere Reaktionen auf ausgeprägt rhythmische Stimuli und bewerteten höhere Tonhöhen positiver als die Teilnehmer am Online-Experiment.

Um zu überprüfen, ob sich diese gefundenen Gruppenunterschiede auf bestimmte (unkontrollierte) Bedingungen zurückführen lassen, die für die Teilnahme am Online-Experiment spezifisch sind, wurden weitere erhobene Variablen in das Analysemodell aufgenommen. Dies umfasste neben dem verwendeten Wiedergabesystem, die Bearbeitungszeit und die selbsteigestufte Ernsthaftigkeit und Konzentriertheit der Teilnahme. Während für das Wiedergabesystem kein signifikanter Effekt nachgewiesen werden konnte, erwies sich die Ernsthaftigkeit der Teilnehmer als geeigneter Prädiktor. So konnten die Gruppenunterschiede für die Valenz-Bewertungen auf den im Labor etwas höheren Grad der Ernsthaftigkeit (gegenüber dem Internet) zurückgeführt werden. Sowohl der Gruppeneffekt (b = .06, t(2172) = 1.06, p = .288), als auch die Interaktion der Gruppe mit dem Rhythmus-Deskriptor (b = -.06, t(2172)=-1.84, p = .070) war unter Berücksichtigung der Ernsthaftigkeit der Teilnahme nicht mehr signifikant. Jedoch konnte keine der erhobenen Variablen die Interaktion zwischen der Gruppe und der Tonhöhe für Valenz oder den signifikanten Gruppenunterschied für Potenz-Bewertungen erklären.

Zusammenfassung

Die Analyse ergab nur eine geringe Anzahl signifikanter Unterschiede zwischen den Musikbewertungen der Web- und der Laborgruppe. Für vier aus 30 Effekten in unserer Modellanalyse wurden statistisch signifikante Unterschiede zwischen Teilnehmern im Web und im Labor nachgewiesen. In diesem Zusammenhang konnte die grundlegende Bedeutung der Erhebung von einer größtmöglichen Anzahl (unkontrollierter) Störvariablen verdeutlicht werden, welche die Varianz zwischen den Experimentalgruppen erklären können. So konnten die Bewertungsunterschiede in zwei von vier Fällen auf die Wirkung einer unkontrollierten (aber gemessenen) Drittvariablen (Die Ernsthaftigkeit der Teilnahme) zurückgeführt werden. Damit konnten zudem die Befunde in [1] repliziert werden.

Zusammengenommen bestätigen unsere Ergebnisse weitgehend, dass Experimente im Internet valide Ergebnisse liefern können und sich grundsätzlich zur Untersuchung von Fragestellungen der Musikwahrnehmung eignen. Neben der erhöhten externen und ökologischen Validität und der gegenüber dem Laborexperiment effizienteren Sammlung von Stichproben größeren Umfangs konnten wir mit unserer Studie den empirische Nachweis liefern, dass die interne Validität von Web-basierten Experimenten mit der von traditionellen Laborexperimenten vergleichbar ist, im Besonderen, wenn potentielle Störeinflüsse in der Analyse berücksichtigt werden. Unsere Studie motiviert somit den weiteren Einsatz von Online-Methoden im Bereich der auditiven Wahrnehmungsforschung.

Danksagung

Diese Studie entstand mit Unterstützung der Audiovisor GmbH und wurde gefördert vom FG Audiokommunikation der TU Berlin unter Leitung von Prof. Dr. Stefan Weinzierl.

Literatur

[1] H. Egermann, F. Nagel, E. Altenmüller, R. Kopiez: Continuous measurement of musically-induced emotion: A web experiment. International Journal of Internet Science, 4(1), 2009, 4–20

[2] M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler, J. Herre: An Experiment about Estimating the Number of Instruments in Polyphonic Music: A Comparison between Internet and Laboratory Results. Presented at ISMIR 2013, 2013

[3] P.R. Rosenbaum, D.B. Rubin: The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70, 1983

[4] O. Lartillot, P. Toiviainen: A Matlab Toolbox for Musical Feature Extraction from Audio. Proceedings of the 10th DAFx-07, Bordeux, 2007