Webbasierte Hörversuche unter unkontrollierten Abhörbedingungen am Beispiel einer Audioqualitäts-Evaluierung von Lautheitstools

Thomas Rauscher[1]

[1] *Institut für Rundfunktechnik GmbH, 80939 München, E-Mail: thomas.rauscher@irt.de*

Einleitung

In dieser Untersuchung werden mit Hilfe einer browserbasierten Hörversuchs-Software die Ergebnisse zweier Versuchs-Szenarien gegenübergestellt. Als Ausgangsbasis dient ein Laborexperiment, das unter der Prämisse einer Erweiterung des Probanden-Panels durch einen Feldversuch ergänzt wird. Anhand eines im IRT durchgeführten Hörversuchs zur Bestimmung der Audioqualität von Lautheitsprozessoren für die Online-Distribution soll diese Möglichkeit einer Panel-Erweiterung überprüft und im Folgenden unterstützt durch statistische Ergebnisse validiert werden. Ziel ist es, die grundlegende Fragestellung zu beantworten, ob die Bewertungen und Ergebnisse des Laborexperiments unabhängig von äußeren Einflüssen im Feldexperiment reproduzierbar sind.

Durchführung

Basierend auf einem ABX-Vergleich nach ITU-R BS.1116 [1] sollten die teilnehmenden Probanden je zwei Stimuli zu einem unbearbeiteten Referenzsignal vergleichen und beurteilen. Zusätzlich wurde den Probanden die Möglichkeit geboten, die qualitativen Unterschiede zwischen dem zu bewertenden und dem Referenzsignal mittels spezieller Attribute näher zu bestimmen. Die Wahl der Attribute wurde an typische Artefakte angelehnt, die im Grenzfall einer unsachgemäßen Parametrisierung eines Lautheitsprozessors oder Dynamikalgorithmus´ entstehen. Als Messwiederholungsplan ausgelegt bietet dieses Versuchsdesign den Vorteil einer geringen Probandenzahl zur Erzielung valider Ergebnisse. Die für diese Evaluierung gewählte Methode zielt auf die Beurteilung kleiner Unterschiede ab. Als Nachteil erweist sich jedoch die fehlende Möglichkeit eines direkten Vergleiches der Ergebnisse oder einer absoluten Bewertung.

Der Hörversuch wurde mit Hilfe einer im IRT entwickelten Hörversuchssoftware umgesetzt. Ausschlaggebend für diese Entwicklung stellt die Einbindung verschiedener Methoden und Standards sowie die Umsetzung komplexerer Szenarien und eine Anpassung an jeweilige Bedürfnisse dar. Als Client-Server-Architektur ausgelegt besteht der größte Vorteil der Software in einer browserbasierten Anwendungsmöglichkeit und der individuellen Gestaltung der Rahmenbedingungen. Die Oberfläche der Software ist in Abbildung 1 dargestellt. Neben den Attributen stand auch ein zusätzliches Kommentar-Feld für Anmerkungen bzw. zur Angabe individueller Attribute zur Verfügung. Als Audiosignale wurden 7 verschiedene Beispiele (siehe Tabelle 1) mit je einer maximalen Dauer von ca. 20 sec. gewählt. Die Beispiele wurden nebst einem speziellen Chor-Stück aus unterschiedlichen Genres wie E- und U-Musik, Radio- und TV-Inhalten ausgewählt, um eine möglichst große Bandbreite an senderelevanten Inhalten zur Verfügung zu stellen. Im Wesentlichen handelt es sich um Signale, bei deren Bearbeitung durch einen Lautheitsprozessor sowohl eine Anpassung der Programmlautheit (IL_k), der Loudness Range (LRA) als auch des Spitzenpegels erfolgen sollte mit den Zielwerten -16 LUFS für IL_k bzw. 10 LU für LRA.

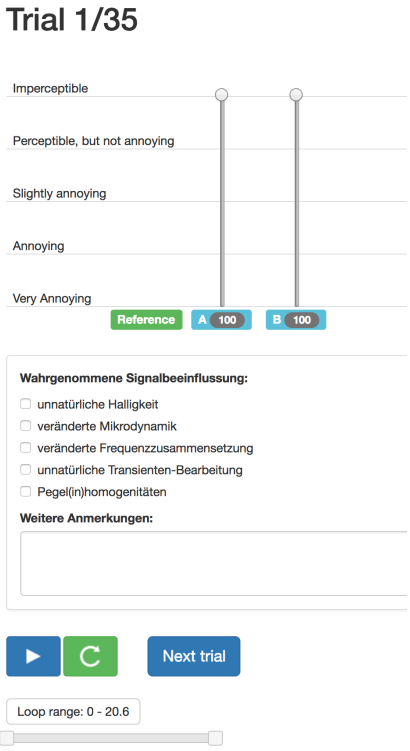

Abbildung 1: Oberfläche der IRT-eigenen Hörversuchssoftware in Anlehnung an ITU-R BS.1116 und deren Erweiterung durch Attributangaben zur Spezifizierung der wahrgenommenen Signalbeeinflussung nebst einem zusätzlichen Feld für Anmerkungen.

Tabelle 1: Übersicht der im Hörversuch verwendeten Audiobeispiele (Soundfile) unter Angabe der Lautheitswerte der Originaldateien und dem jeweiligen Genre. (IL_k = Programmlautheit, LRA = Loudness Range, TP_{max} = maximaler TruePeak-Wert).

Soundfile	IL_k [LUFS]	LRA [LU]	TP_{max} [dBTP]	Genre
„1“	-19,0	25,9	-0,1	Chor
„2“	-19,5	23,2	0,0	E-Musik
„3“	-21,6	18,0	-0,2	Radio
„4“	-22,5	9,9	0,0	TV/Doku
„5“	-23,4	11,4	-0,8	TV/Serie
„6“	-26,3	17,3	-0,1	U-Musik
„7“	-14,4	18,5	0,0	E-Musik

Es wurden deshalb keine Signale ausgewählt, deren Anpassung lediglich einer Pegelabsenkung bedürfe, da hier keine Artefakte oder Qualitätsbeeinflussungen zu erwarten sind.

Das Laborexperiment wurde in einem akustisch optimierten Raum mit definierten Abhörbedingungen (Fremdgeräuschpegel ≤ 22 dB(A), Nachhallzeit T_{60} = 0,25 s @ 1 kHz) eingerichtet. Nach der Durchführung dieses Laborexperiments mit 20 Probanden, bestehend aus Experten und hörunerfahreneren Personen (Consumer), wurde der Versuch als Feldexperiment anhand einer Online-Version umgesetzt. Dabei diente eine erweiterte Fassung der Versuchsanleitung den Feld-Probanden der selbständigen Durchführung des Versuches. Zur Erhöhung der Validität wurden zusätzlich zur Randomisierung des Versuchsablaufs weitere Vorkehrungen getroffen, um äußere Einflüsse zu minimieren. Auf der einen Seite wurde der gesamte Versuch auf eine reine Kopfhörerwiedergabe beschränkt. Hierzu war es ebenfalls nötig, den Probanden vergleichbare Wiedergabepegel bereitzustellen. Dies wurde anhand eines speziellen Einpegelsignals (Referenz-Sprachsignal) ermöglicht, welches die Probanden auch nach Unterbrechung des Versuches verwenden konnten, um eine adäquate Wiedergabelautstärke einzustellen. Den Probanden wurden auf der anderen Seite Hörbeispiele für jedes Attribut sowie einem Trainings-Durchlauf zur Eingewöhnung an die Bewertungsskala dargeboten. Zudem wurde ein Fragebogen (anonymisierte Abfrage persönlicher Daten wie Alter und Hörerfahrung und technischer Daten wie Interface und Kopfhörer) und eine detaillierte Anleitung inklusive Hinweis auf Einhaltung von Pausen präsentiert. Da für die Feld-Evaluierung hauptsächlich Tonschaffende der Rundfunkanstalten adressiert waren, bestand dieses Panel vorwiegend aus Experten.

Ergebnisse

Die Box-Plots in Abbildung 2 zeigen die Bewertungen der nach dem Post-Screening verbleibenden 13 Labor- und 8 Feld-Probanden, aufgeschlüsselt nach den jeweiligen bearbeiteten Soundfiles (Programmen). Die Skala stellt gemäß der statistischen Beurteilungsmethode in ITU-R BS.1116 die Differenz zwischen dem zu bewertenden Stimulus und der versteckten Referenz dar. Positive Bewertungen bedeuten demnach eine irrtümliche Bewertung dieser versteckten Referenz durch den Probanden. Es wird zunächst deutlich, dass vor allem das Chor-Signal („1“) außerhalb der Bewertungskriterien dieser Methode liegt, da kaum ein Algorithmus in der Lage ist, dieses Audiobeispiel mit guter Audioqualität zu prozessieren. Demgegenüber werden die beiden TV-Signale („4“ und „5“) größtenteils mit guter Audioqualität, d.h. ohne erkennbare Artefakte bearbeitet. Die übrigen Signale weisen je nach Algorithmus und Programmart meist erkennbare (perceptible) bis störende (annoying) Artefakte auf. In der direkten Gegenüberstellung der beiden Versuchsszenarien fällt eine tendenziell strengere Bewertung durch die Feld-Probanden auf, was auf den prozentual höheren Anteil an Experten zurückführbar sein könnte. Basierend auf den Angaben der Probanden bezüglich ihrer Hörerfahrung konnte dies statistisch jedoch nicht nachgewiesen werden. Als Kernaussage kann jedoch festgehalten werden, dass die Tendenz der programmabhängigen Bewertungs-Unterschiede in beiden Szenarien in vergleichbarem Maße ausgeprägt ist. Die Differenzen in der Prozessierungsqualität der einzelnen Systeme werden ebenfalls in beiden Szenarien gleicherweise deutlich.

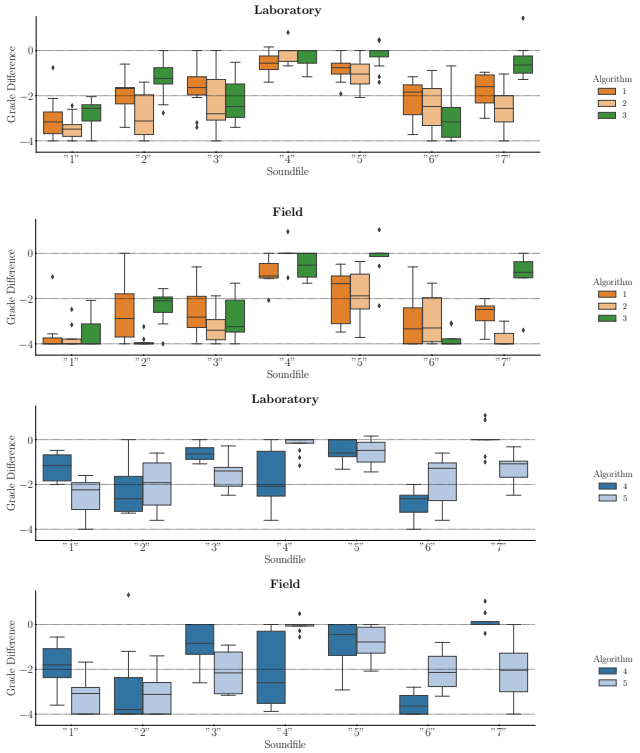

Abbildung 2: Explorative Statistik: grafische Darstellung der Bewertungen der Algorithmen 1, 2 und 3 bzw. (farblich separiert) 4 und 5 mittels Box-Plot aufgeschlüsselt nach dem jeweiligen Programm aus beiden Hörversuchs-Szenarien. Jeweils oben: Labor-Versuch, unten: Feld-Versuch. Die Skala entspricht der Differenz zwischen zu beurteilendem Stimulus und der versteckten Referenz.

Statistische Auswertung

Die unabhängigen statistischen Auswertungen der beiden Szenarien zeigten jeweils in Bezug auf eine Gegenüberstellung der Mittelwerte und Varianzen mittels Varianzanalyse (ANOVA [2]) ein vergleichbares Resultat. Die Nullhypothese $\mu_1 = \ldots = \mu_n$ gleicher Mittelwerte der System-Bewertungen kann für beide Szenarien verworfen werden. Ebenso kann die Nullhypothese der gleichen Mittelwerte der Programm-Bewertungen ($\mu_1 = \ldots = \mu_n$) innerhalb eines Systems verworfen werden und zwar in gleichem Maße für alle Systeme und für beide Szenarien. Ausschlaggebend für eine Erweiterung des Probanden-Panels durch das Feldexperiment ist jedoch die Nullhypothese, dass für beide Szenarien die Mittelwerte der jeweiligen Programm-Bewertungen innerhalb eines Systems gleichwertig sind ($\mu_1 = \mu_2$). In 80% der Fälle ist diese Forderung gegeben wie in Tabelle 2 dargestellt. Aufgrund der nicht vollständig gegebenen Normalverteilung wurde diese Hypothese – bis auf eine Ausnahme – durch einen Kruskal-Wallis-Test als nicht-parametrische Alternative zu einer Auswertung mittels ANOVA bestätigt. Eine Erweiterung des Probanden-Panels scheint für dieses Experiment geglückt zu sein. Ein Einfluss der äußeren Bedingungen auf die Versuchsergebnisse wie verwendetes Equipment, Abhörumgebung oder Probanden-bezogene Effekte ist nicht nachweisbar.

Tabelle 2: Vergleich der Mittelwerte aus beiden Versuchs-Szenarien durch Varianzanalyse (ANOVA): F-Statistik mit oberem F-Wert von 4,38; Signifikanzniveau 0,05. H_0: $\mu_1=\mu_2$. Die rot markierten F-Werte kennzeichnen ein Verwerfen der Nullhypothese.

Soundfile	System/Algorithmus				
	1	2	3	4	5
„1“	1,54	1,43	4,35	3,31	2,98
„2“	2,30	8,52	10,25	0,57	4,59
„3“	3,85	3,36	2,29	1,53	3,02
„4“	1,54	0,60	1,49	0,43	0,97
„5“	7,52	4,57	0,00	1,44	0,88
„6“	4,12	1,01	3,48	7,55	0,57
„7“	8,82	9,91	1,63	0,33	3,04

Für eine tiefergehende und weiterführende statistische Analyse wurde als Post-hoc-Test ein multipler Paarvergleich nach Tukey (vgl. [3]) vorgenommen wie in Abbildung 3 dargestellt (beispielhaft für Algorithmus 1 bis 3). Die Gegenüberstellung zeigt eine scheinbar homogenere Bearbeitung durch Algorithmus 1 über alle Signale hinweg, während Algorithmus 3 teils sehr gute Ergebnisse erzielt, allerdings mit größeren Diskrepanzen in Bezug auf die unterschiedlichen Programme. In der Tendenz ist diese Erkenntnis in beiden Szenarien identifizierbar mit sichtbar größeren Gruppen-Konfidenzintervallen und tendenziell strengeren Bewertungen im Feldversuch.

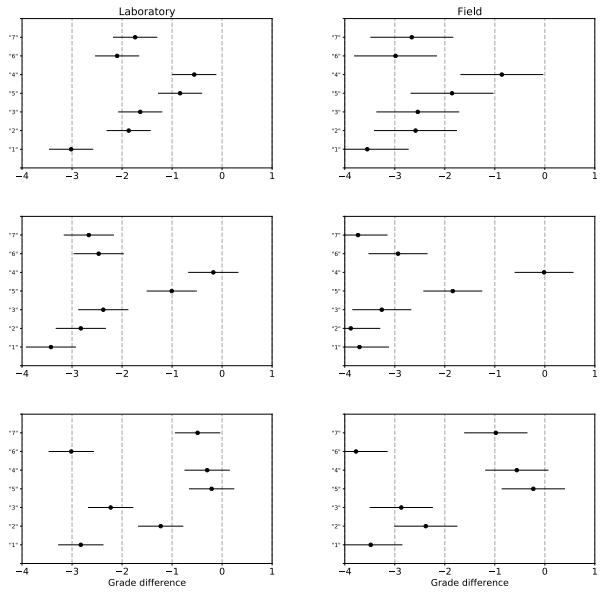

Abbildung 3: Grafische Darstellung der Post-hoc-Analyse mittels multiplem Paarvergleich nach Tukey durch Gegenüberstellung der Bewertungs-Mittelwerte unter Angabe der Gruppen-konfidenzintervalle der jeweiligen Programme innerhalb der Algorithmen 1, 2 und 3. Links: Labor-Versuch, rechts: Feld-Versuch.

In einer abschließenden Auswertung unter Anwendung eines exakten Tests nach Fisher (vgl. [4]) wurden die Häufigkeiten der Angaben jedes einzelnen Attributs innerhalb eines Systems in Abhängigkeit der Programme verglichen. Die Nullhypothese geht hierbei von einer Unabhängigkeit der Attribut-Angaben vom jeweiligen Versuchs-Szenario aus. Wie in Abbildung 4 erkennbar, ist die Wahrscheinlichkeit einer zufälligen Unabhängigkeit in den meisten Fällen größer dem Signifikanzniveau von 5%, und die Nullhypothese muss größtenteils verworfen werden. Es ist somit davon auszugehen, dass die Attribut-Angaben der Probanden aus den beiden Versuchs-Szenarien korrelieren.

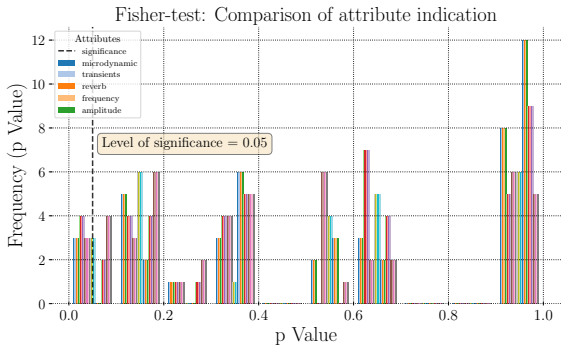

Abbildung 4: Vergleich der Attribut-Häufigkeiten der beiden Hörversuchs-Szenarien: p-Wert und dessen Häufigkeit ausgewertet für jedes Attribut innerhalb eines Systems in Abhängigkeit des Programms gem. einem exakten Test nach Fisher unter der Annahme unabhängiger Häufigkeiten (Nullhypothese).

Fazit und Ausblick

Die statistischen Resultate lassen die Schlussfolgerung zu, dass sich aufgrund der hohen Übereinstimmung (80%) der Bewertungen die Ergebnisse aus beiden Versuchs-Szenarien gegenseitig ergänzen, eine Erweiterung des Probanden-Panels somit erfolgreich vorgenommen werden kann. Aufgrund der Vorkehrungen in Bezug auf die Wiedergabesituation, der Vorgaben in der Anleitung und der dediziert ausgewählten (Experten-)Zielgruppe sind kaum Einflüsse der äußeren Rahmenbedingungen erkennbar. Letzteres könnte auch ein Indiz für die strengeren, tendenziell schlechteren Bewertungen der Feld-Probanden sein. Die hier vorgestellten Erkenntnisse sind die Ergebnisse einer ersten Versuchsreihe und bedürfen einer Verifikation durch weitere Hörversuche mittels geeigneter erweiterter oder geänderter Fragestellungen. Als übergeordnetes Ziel dieser Evaluierung könnte mit Hilfe der vorgestellten browserbasierten Software-Lösung eine Consumer-Evaluierung zu dieser Thematik unternommen werden. Im Vorfeld ist hierzu ein geeignetes Versuchsdesign beispielsweise als Zufallsgruppenplan zu eruieren. Des Weiteren ist es unabdingbar, ausreichend valide Vorversuche zur Abschätzung verschiedener Einflüsse wie Abhörumgebung, Art des Endgerätes oder Nutzerpräferenzen durchzuführen.

Literatur

[1] ITU: Recommendation ITU-R BS.1116-3. Methods for the subjective assessment of small impairments in audio systems. Genf, 2015

[2] Bech, S., Zacharov, N.:Perceptual Audio Evaluation. Wiley & Sons, Chichester, 2006

[3] Borz, J., Schuster C.: Statistik für Human- und Sozialwissenschaftler. Springer-Verlag, Berlin Heidelberg, 2010

[4] Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition. Chapman & Hall, Boca Raton, 2011