Lautheitsbeurteilung von Musik: Methoden und Modellvergleiche

Florian Schmidt, Birger Kollmeier, Stefan Uppenkamp
Medizinische Physik, Universität Oldenburg, 26111 Oldenburg,
E-Mail: florian.schmidt@uni-oldenburg.de

Einleitung

Die Lautheit von Musik zu bestimmen, stellt einen vor gewisse Schwierigkeiten. Als Musik werden Hörszenarien mit Gesang, verschiedenen Klang- und Schlaginstrumenten, jedoch auch mit synthetischen Klängen oder Soundsamples und natürlich deren Kombination in Bands, Chören, Orchestern oder digital zusammengestellten Tonaufnahmen bezeichnet. Dieses breite Spektrum an Klangszenarien erfordert die Berücksichtigung diverser Einflussgrößen der Lautheit [1]. Darunter fallen spektrale Lautheitssummation, temporale Lautheitsintegration und Amplitudenmodulationseffekte. Hinzu kommen noch musikeigene Effekte, die auf höherer Ebene das Lautheitsurteil beeinflussen, z.B. Präferenzen und Kontexteffekte [2].

Zur objektiven Messung der Lautheit bieten sich eine große Anzahl an Lautheitsmaßen an [3], [4]. Im Wesentlichen unterscheidet man dabei frequenzgewichtete Pegel und psychoakustische Lautheitsmodelle, die sich am physiologischen Verarbeitungsprozess orientieren. Um die Qualität der Lautheitsmaße miteinander vergleichen zu können, benötigt man ein Vergleichsmaß. Dieses Maß liefern psychoakustische Messungen.

Methode

Ein Ansatz zur Qualitätsbeurteilung von verschiedenen Maßen ist es generell, unterschiedliche Objekte zu vermessen und im ersten Schritt die ermittelten Rangfolgen mit denen des Standardmaßes zu vergleichen (ordinale Ebene). Im zweiten Schritt bietet es sich an, die Abstände bzw. die Verhältnisse der Skalenwerte mit denen des Standardmaßes zu vergleichen (kardinale Ebene).

Aus den psychoakustischen Messmethoden wurde für diese Studie der Paarvergleich ausgewählt, da er einige Vorteile mit sich bringt. Im Gegensatz zu direkten Skalierungsmethoden ist die Aufgabenstellung bei einem Paarvergleich für die Versuchspersonen leichter zu bewältigen. In diesem Fall werden zwei Musikstücke nach ihrer insgesamt empfundenen Lautheit direkt miteinander verglichen. Im Vergleich dazu beinhaltet eine kategoriale Skalierung z. B. die Gefahr, dass Präferenz-Effekte die Messergebnisse der Lautheitswahrnehmung verzerren. Des Weiteren wird beim Paarvergleich kein Stimulus als Referenz bevorzugt, wie bspw. bei Adaptive Forced Choice Methoden, bei denen ein Referenzstimulus durch adaptive Vergleiche als Maßstab für die Teststimuli dient. Der Nachteil des Paarvergleichs ist der in der Regel hohe Zeitaufwand [5] und die Schwierigkeit, ein kardinales Skalenniveau zu konstruieren.

Die Auswahlhäufigkeit beim Paarvergleich liefert eine Schätzung der Rangfolge der Stimuli. Auch für die Lautheitsmaße können solche Rangfolgen gewonnen werden. Ein Rangkorrelationstest liefert dann mit dem Korrelationskoeffizienten ein Maß für die Übereinstimmung zwischen Lautheitsmaß und subjektiver Beurteilung. Dies gibt eine erste Schätzung für die Güte des Lautheitsmaßes ab.

Eine Korrelationsanalyse auf höherem Skalenniveau kann noch genauere Auskunft über die Qualität der Lautheitsmaße geben. Das Bradley-Terry-Luce Modell (BTL) ist dabei eines der gängigen Verfahren [6], [7], um aus den Ergebnissen eines Paarvergleichs ein quantitatives Skalenniveau zu gewinnen. Die Funktionsweise dieses Ansatzes soll im Folgenden kurz beschrieben werden.

Funktionsweise des BTL-Modells

Das BTL-Modell basiert einerseits auf der Annahme, dass eine Folge von Paarvergleichen einem Binomialprozess gleicht, so dass gilt:

$$B(n,k) = \binom{n}{k} {p_{AB}}^{k} \cdot (1 - p_{AB})^{n-k} \tag{1}$$

Die Anzahl der Vergleiche n und die dabei ermittelten Auswahlhäufigkeiten k ergeben sich unmittelbar aus dem Paarvergleich. Die Wahrscheinlichkeit p_{AB}, dass Stimulus A gegenüber Stimulus B ausgewählt wird, kann hingegen nur über Schätzverfahren bestimmt werden.

Andererseits nimmt man beim BTL-Modell an, dass die Skalenwerte der Reize v_A von Stimulus A und v_B von Stimulus B in einem eindeutigen Zusammenhang zu p_{AB} stehen.

$$p_{AB} = \frac{v_A}{v_A + v_B} \tag{2}$$

Die Skalenwerte v sind Teil einer dadurch neu gewonnenen Verhältnisskala.

Um einschätzen zu können, wie gut die Verhältnisskala des BTL-Modells funktioniert, kann mit Zufallszahlen einPaarvergleich simuliert werden. Hierfür werden anfangs Test-Skalenwerte festgelegt, die gaußverteilt sind, um damit die Unsicherheit der Wahrnehmung durch die Versuchspersonen zu simulieren. Aus diesen Test-Skalenwerten werden gemäß ihrer Verteilung jeweils Stichproben gezogen, deren Umfang die Anzahl der Probanden simuliert. Nun können diese Werte wie bei einem Paarvergleichsexperiment miteinander verglichen und dem BTL-Modell zugeführt werden. Hieraus ergeben sich dann Schätzungen für die Test-Skalenwerte. Die Abweichung der Schätzung vom ursprünglichen Wert zeigt dann an, wie gut

das BTL-Modell unter den jeweiligen Bedingungen die Test-Skalenwerte aus dem Paarvergleich repräsentiert.

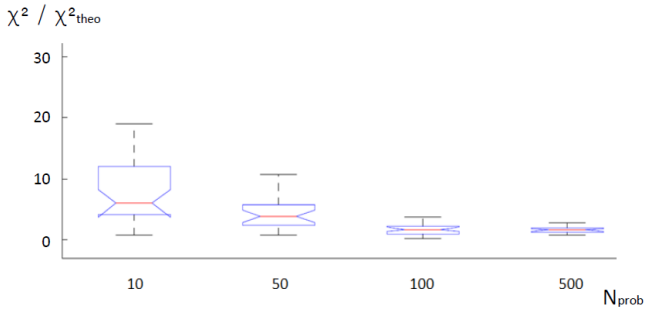

Abbildung 1: Güte der Schätzung von Test-Skalenwerte durch das BTL-Modell für unterschiedliche Probandenzahlen N_{prob}. Erst ab dem aus einem Anpassungstest gewonnenen Verhältnis $\chi^2/\chi^2_{theo} < 1$ kann die Schätzung als hinreichend genau angenommen werden.

Folgende Fragestellungen wurden untersucht:

1 Wie groß ist der Einfluss der Probandenzahl auf die Schätzqualität der BTL-Skalenwerte?

2 Spielt der Skalenwertabstand der Stimuli eine Rolle bei der Schätzqualität der BTL-Skalenwerte?

3 Verzerren zyklische Triaden die Schätzung des BTL-Modells?

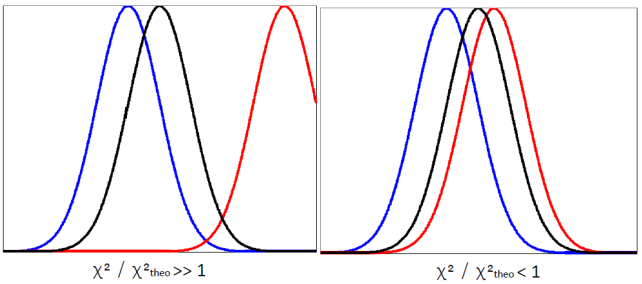

Abbildung 2: Schematische Darstellung dreier Skalenwerte, aus deren Verteilungen Stichproben für einen simulierten Paarvergleich gewonnenen werden können, zur Verdeutlichung von geeigneten BTL-Szenarien. Links: Der Abstand zwischen den blauen und schwarzen Skalenwerten zu dem roten Skalenwert ist so groß, dass der Überlapp der drei Verteilungen zu gering ist. Die Skalenwerte der BTL-Schätzung werden ungenau ($\chi^2/\chi^2_{theo} >> 1$). Rechts: Alle drei Skalenwerte sind nah genug beieinander, so dass die BTL-Schätzung brauchbare Ergebnisse liefert ($\chi^2/\chi^2_{theo} < 1$).

In Abb. 1 ist das Simulationsergebnis für die erste Frage dargestellt. Bei vier verschiedenen Probandenzahlen: 10, 50, 100, 500 wurden 30mal zufällig (gaußverteilt) 10 Test-Skalenwerte gewählt und mithilfe eines χ^2-Anpassungstests miteinander verglichen. Wenn $\chi^2/\chi^2_{theo} < 1$, dann können die geschätzten Skalenwerte als hinreichend genau angenommen werden. Das Ergebnis zeigt, dass erst bei etwa 100 Versuchspersonen dieses Kriterium erfüllt wird. Dies unterstreicht noch einmal den hohen Aufwand des Paarvergleichs. .

In der Literatur wird angemerkt, dass Stimuli-Vergleiche mit einseitig dominierenden Ausgängen für das BTL-Modell unbrauchbar sind [5]. Dies kann simuliert werden, indem der Bereich, aus dem die Test-Skalenwerte gewählt werden, schrittweise erhöht wird. Abbildung 2 zeigt schematisch das Ergebnis dieser Simulation. Zu sehen sind drei Test-Skalenwerte (rot, blau, schwarz) mit gleicher Streuung. Aus diesen Verteilungen werden nun Stichproben im Umfang von n=1000 gezogen und in einem Paarvergleich ausgewertet. Auch hier ist χ^2/χ^2_{theo} das Qualitätsmaß der BTL-Schätzung. Es wird deutlich, dass sich nur bei nahe beieinander liegenden Skalenwerten mit überlappenden Verteilungen Werte mit $\chi^2/\chi^2_{theo} < 1$ ergeben.

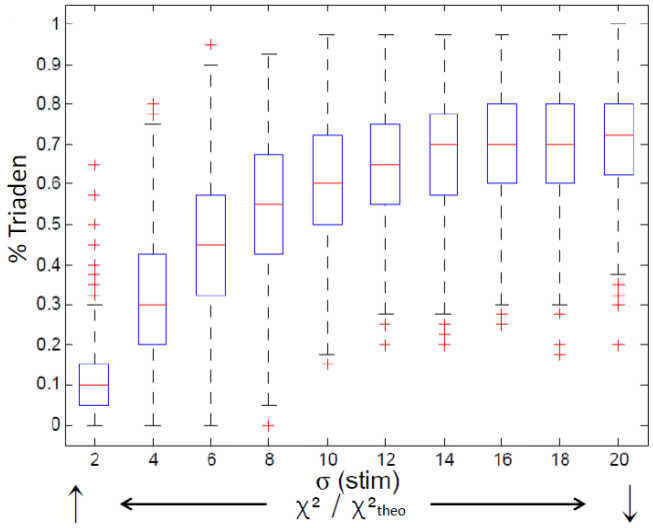

Abbildung 3: Die Simulation der relativen Anzahl der zyklischen Triaden für eine Gruppe von n=1000 Versuchspersonen für unterschiedlich stark streuende Skalenwerte. Die Streuung σ(stim) soll die Messgenauigkeit einer Versuchsperson simulieren. Je ungenauer eine Versuchsperson einen Skalenwert angibt, umso höher steigt die Zahl der auftretenden zyklischen Triaden. Gleichzeitig nimmt aber auch die Qualität der durch das BTL-Modell geschätzten Skalenwerte zu.

Zyklische Triaden werden oft als Ausschlusskriterium für Versuchspersonen herangezogen [8], denn sie stellen ein Maß für die Beurteilungskonsistenz dar. Eine zyklische Triade wird dann gezählt, wenn eine Versuchsperson Stimuli A, B und C in folgender Weise widersprüchlich bewertet: A>B, B>C, und C>A. Sie können aber auch dann auftreten, wenn die Beurteilung dreier Stimuli jeweils einer größeren Unsicherheit unterworfen ist, so dass deren Verteilungen sich stark überlappen. Aus der vorherigen Simulation ergab sich aber gerade, dass ein solcher Überlapp im Prinzip ein gutes Szenario für eine BTL-Bewertung liefert. Durch die Variation der Streuung der Test-Skalenwerte kann dieser Überlapp simuliert und schrittweise erhöht werden. Abbildung 3 zeigt, dass bei größer werdendem Überlapp die Anzahl der zyklischen Triaden stark ansteigt. In gleicher Weise nimmt das Verhältnis χ^2/χ^2_{theo} ab. Zyklische Triaden treten also gerade gehäuft in Szenarien auf, die sich für die Verwendung des BTL-Modells gut eignen.

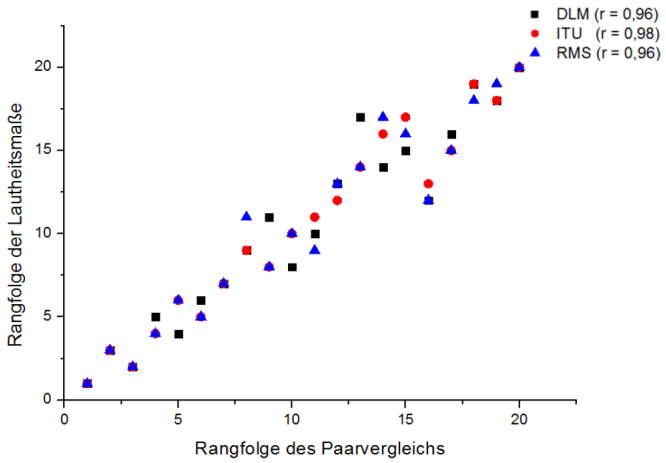

Abbildung 4: Rang-Korrelationsdiagramm zwischen den Lautheitsmaßen und der gemessenen Lautheitswahrnehmung für 20 verschiedene Musik-Stimuli. Schwarz: Dynamisches Lautheitsmodell von Fastl und Chalupper, Rot: ITU-R BS.1770-2, Blau: RMS des Schalldruckpegels. Der Rangkorrelationskoeffizient r stellt ein Maß für die Qualität der Lautheitsmaße dar.

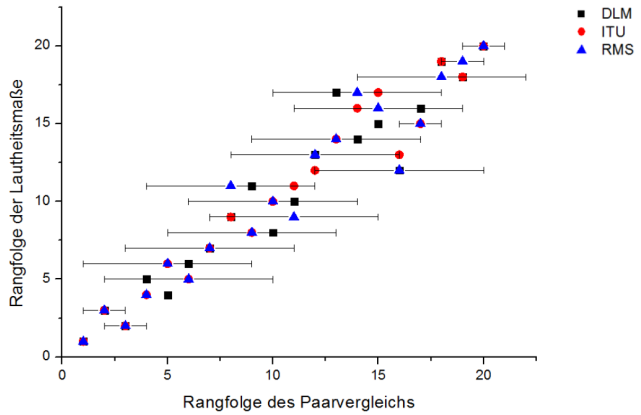

Abbildung 5: Rang-Korrelationsdiagramm zwischen den errechneten Lautheitsmaßen und der gemessenen Lautheitswahrnehmung mit zusätzlich ermittelter Genauigkeit der Rangposition mittels BTL-Analyse.

Ergebnisse

Für den Paarvergleich wurden nun 20 Stimuli aus fünf verschiedenen Musik-Genres ausgewählt: Rock, Klassik, Pop, Jazz und Hip Hop. Aus jedem Genre wurden jeweils zwei Musikstücke ausgewählt, aus denen jeweils zwei 10s-Ausschnitte entnommen wurden. Dabei wurde darauf geachtet, dass bei allen Ausschnitten der Dynamikbereich etwa 2 dB betrug. Jeder Stimulus wurde nun unterschiedlich gepegelt. Es wurde darauf geachtet, dass der Präsentationspegel der musiktypischen Hörgewohnheit entsprach (zwischen 55 und 95 dB SPL). Der Paarvergleich wurde mit 10 Versuchspersonen durchgeführt. Ein Vergleich beinhaltete die Präsentation des ersten Stimulus, gefolgt von einer kleinen Pause, bis dann der zweite Stimulus präsentiert wurde. Alle Stimuli wurden in beide Richtungen miteinander verglichen, so dass 20x19 Vergleiche vorlagen. Die Versuchspersonen wurden gezwungen sich zu entscheiden, welchen Stimulus sie als lauter einschätzen würden.

Neben dem Schalldruckpegel (RMS) wurden der frequenzgewichtete Pegel ITU-R BS.1770-2 (ITU) [9] und das Dynamische Lautheitsmodell (DLM) von Fastl und Chalupper [10] als miteinander konkurrierende Lautheitsmaße verglichen.

Aus den Lautheitsmaßen und den Auswahlhäufigkeiten lassen sich jeweils Rangfolgen ermitteln, die in einem Korrelationsdiagramm aufgetragen werden können (Abb.4). Es ist zu erkennen, dass die Rangfolgen aller drei Lautheitsmaße eine gute Übereinstimmung mit der Rangfolge des Paarvergleichs liefern. Die Bewertung der ITU-Gewichtung scheint geringfügig näher an der subjektiven Beurteilung zu liegen als die beiden anderen Maße.

Aufgrund der bislang noch geringen Anzahl der Versuchspersonen (N = 12) ist es noch nicht möglich mit dem BTL-Modell hinreichend genaue Skalenwerte zu schätzen, die die Übereinstimmung der Abstände und Verhältnisse der Skalenwerte der Lautheitsmaße gegenüber der subjektiven Beurteilung zu überprüfen ermöglichen würden. Allerdings lässt sich mit dem BTL-Modell untersuchen, welche Rangfolgevertauschungen der Lautheitsmaße wirklich signifikant sind oder möglicherweise nur als statistische Artefakte des Paarvergleichs zu interpretieren sind. Hierfür müssen die Stimuli gruppiert werden, um eindeutige Vergleiche zu vermeiden (gemäß der Simulationsergebnisse zur zweiten Frage). Zusätzlich zu den durch das BTL-Modell geschätzten Skalenwerten lassen sich auch für jeden Skalenwert Konfidenzintervalle bestimmen, mit deren Hilfe die Sicherheit der Schätzung jedes einzelnen Rangs bestimmt werden kann. In Abb. 5 wird deutlich, dass ein Großteil der Rangvertauschungen auf der Ungenauigkeit des Paarvergleichs beruht. Einige Rangvertauschungen haben ihre Ursache allerdings in der Fehlerhaftigkeit der errechneten Lautheitsmaße.

Ausblick

Die bisherigen Ergebnisse der Studie zeigen, dass es ist mit der beschriebenen Vorgehensweise möglich ist, Lautheitsmaße erfolgreich auf ihre Qualität zu untersuchen. Signifikante Rangvertauschungen bieten einen Angriffspunkt für eine Analyse, aus welchem Grund ein Musikstück von diesem Lautheitsmaß falsch eingeschätzt wurde. Dies ermöglicht die Modifizierung des Maßes.

Eine höhere Auflösung für eine Analyse würde allerdings die Erhöhung des ordinalen Skalenniveaus der Lautheitswahrnehmung liefern. Dieses Ziel sollte im weiteren Verlauf der Studie angestrebt werden. Hierfür sollte die Versuchspersonenzahl erhöht werden. Dabei sollte darauf geachtet werde, dass der Paarvergleich auf die für das BTL-Modell relevanten Vergleiche der gruppierten Stimuli reduziert wird, um die Datenerhebung möglichst effizient zu machen.

Des Weiteren sollten weitere Analyse-Tools ausprobiert werden, die es evtl. auf ökonomischere Weise ermöglichen das Skalenniveau anzuheben (bspw. zusätzlich die Durchführung einer direkten Skalierung).

Literatur

[1] Rennies, J.: Comparison of loudness models for time-varying sounds. Acta Acustica united with Acustica 96 (2010), 383-396

[2] Fucci, D.: Children Scaling Rock Music. Acoustical Society of America 138th Meeting Lay Language Papers (1999)

[3] Vickers, E.: Metrics for Quantifying Loudness and Dynamics. Audio Engineering Society 129th Convention (2010)

[4] Scovenborg, E.: Evaluation of Different Loudness Models with Music and Speech Material. Audio Engineering Society 117th Convention (2004)

[5] Ellermeier, W., Hellbrück, J., Kohlrausch, A., Zeitler, A.: Kompendium zur Durchführung von Hörversuchen in Wissenschaft und industrieller Praxis. DEGA, Berlin, 2008

[6] Ellermeier, W.: Scaling the Unpleasantness of Sounds According to the BTL Model: Ratio-Scale Representation and Psychoacoustical Analysis. Acta Acustica united with Acustica 90 (2004), 101-107

[7] Tsukida, K.: How to Analyze Paired Comparison Data. UWEE Technical Report Number UWEETR-2011-0004 (2011)

[8] Kendall, M.G.: On the Method of Paired Comparisons. Biometrika 31 (1940), 324-345

[9] International Telecommunication Union: Recommendation ITU-R BS.1770-2 - Algorithms to measure audio programme loudness and true-peak audio level. Electronic Publication, Geneva, 2011

[10] Chalupper J., Fastl H.: Dynamic Loudness Model (DLM) for Normal and Hearing-Impaired Listeners. Acta Acustica united with Acustica 88 (2002), 378-386