Über dieses Dokument

Inhalt

Das vorliegende Dokument ist die vierte von vier Application Notes über die Durchführung von Hörversuchen. Es bietet eine Einführung in die Auswertung der durch Hörversuche gewonnen Geräuschbeurteilungen. Das Dokument zeigt dazu einige grundlegende statische Methoden und gibt darüber hinaus Hinweise, was bei der Auswertung zu beachten ist.

Zielgruppe

Der nachfolgende Text wendet sich insbesondere an (potenzielle) Anwender und Anwenderinnen des ArtemiS SUITE Jury Testing Module SQala, die einen Einblick in die verschiedenen Auswertemethoden für Hörversuchsergebnisse benötigen.

Fragen?

Sie haben Fragen? Wir freuen uns über Ihre Rückmeldungen!
Fragen zum Inhalt dieses Dokument: Imke.Hauswirth@head-acoustics.com
Technische Fragen zu unseren Produkten: SVP-Support@head-acoustics.com

Hörversuche durchführen – Teil 4

Einleitung

Nach der Durchführung der Hörversuche wird mit der Auswertung der gewonnen Daten begonnen. Für diese Auswertung steht eine Vielzahl von statistischen Analysemethoden zur Verfügung. Zum einen dienen diese dazu, die Daten an sich zu untersuchen und auszuwerten, und zum anderen können mit diesen Berechnungen die Daten des Hörversuchs zusammengefasst und in eine übersichtliche Form gebracht werden.

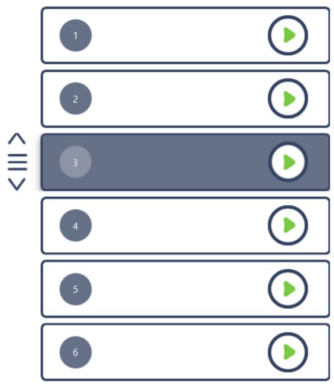
Bevor die Daten jedoch mithilfe der Statistik untersucht werden können, müssen sie zunächst in Zahlen „übersetzt“ werden. Wurde die Untersuchung mit einer entsprechenden Software z. B. mit dem Jury Testing Modul SQala durchgeführt, erhält der Versuchsleiter am Ende des Versuchs automatisch eine in Zahlen umgerechnete Angabe der Urteile. Unterschiedliche Testmethoden führen dabei zu unterschiedlichen Auswertungen bzw. Codierungen.

1. Hinweise zur Auswertung verschiedener Testarten

Ranking

Auswertung Ranking-Tests

Bei der Testmethode des Ranking werden nur Rangurteile abgegeben, d. h. es handelt sich um eine vergleichende Skalierung, bei der keine Informationen über den Abstand der einzelnen Ränge vorliegen. Bei der Auswertung darf der Versuchsleiter nicht vernachlässigen, dass jede Geräuschbeurteilung im hohen Maß von der Geräuschbeurteilung der anderen Geräusche abhängt. Durch die Mittelung der einzelnen Urteile der verschiedenen Teilnehmer entstehen zwar automatisch unterschiedliche Abstände, es muss aber für jeden Hörversuch individuell entschieden werden, ob es sinnvoll ist, diese Gewichtung für die weitere Auswertung zu übernehmen oder diese wieder in Rangurteile umzurechnen.



Paarvergleich

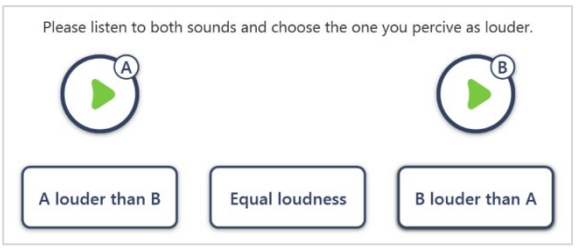
Auswertung Paarvergleich

Bei einem Paarvergleich wird zunächst auch nur eine vergleichende Skalierung durchgeführt (A ist besser als B). Auf einfache Weise können die gesammelten Einzeldaten zu eine Rangfolge zusammengesetzt werden (die Bewertung A>C, C>B ergibt die Reihenfolge A, C, B). Mit entsprechenden statistischen Hilfsmitteln kann darüber hinaus auch eine skalierte Reihenfolge berechnet werden, bei der auch die Unterschiede zwischen den Geräuschen ausgewertet werden können. Mithilfe dieser Skala können dann Korrelationsuntersuchungen durchgeführt werden. Zusätzlich können für den Paarvergleichtest verschiedene Auswertungen zur Urteilssicherheit und Reliabilität der Teilnehmer gemacht werden.

Inkonsistente Triaden

Hierzu zählt zum Beispiel die Untersuchung von Triaden. Ist Geräusch A besser bewertet worden als Geräusch B und Geräusch B besser als Geräusch C, sollte Geräusch A auch besser als Geräusch C bewertet werden. Ist dies nicht der Fall und kommt eine solche Inkonsistenz häufiger vor, muss geprüft werden, was die Ursache hierfür sein kann und ob das Testdesign geändert werden muss. Treten inkonsistente Triaden bei mehreren Teilnehmern auf, ist dies z. B. ein Hinweis darauf, dass die Teilnehmer überfordert sind bzw. die Testaufgabe nicht richtig vermittelt wurde. Weiterhin besteht die Möglichkeit, dass die Unterschiede zwischen den Geräuschen so gering sind, dass die Teilnehmer diese nicht reproduzierbar wahrnehmen können.

Beim Paarvergleich ist es sinnvoll, die einzelnen Geräuschpaare mehrmals (auch in umgekehrter Reihenfolgen, d. h. A - B und dann B - A) abzufragen. So kann die Wiederholbarkeit der Beurteilung für jeden einzelnen Teilnehmer überprüft werden. Diese gibt zusätzlich Aufschluss über die Lösbarkeit der Aufgabe und die Fähigkeiten der Teilnehmer.



Kategoriale Bewertung

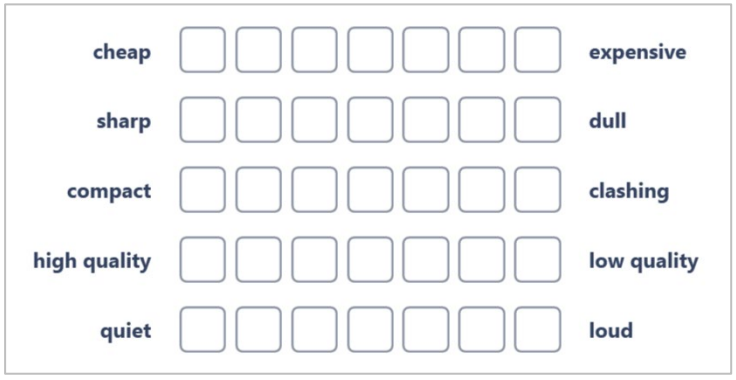
Auswertung kategoriale Bewertung

Die Beurteilung eines Geräuschs während eines Hörversuchs mit kategorialer Bewertung erfolgt mehr oder weniger unabhängig von den Beurteilungen der anderen Geräusche des Versuchs. Wurden die Kategorien so gewählt, dass sie als äquidistant[1] betrachtet werden können, kann man davon ausgehen, dass die gewonnen Daten intervallskaliert sind und somit auch die Größe der Unterschiede ausgewertet werden kann. Dies bietet den Vorteil, dass die Ergebnisse eines solchen Hörversuchs meist für eine Korrelationsanalyse mit den Ergebnissen aus messtechnischen Analysen herangezogen werden können. Auch bei Hörversuchen mit Kategorialskalierung sollten die Teilnehmer einige Geräusche mehrmals beurteilen, um Kontexteffekte zu minimieren und die intraindividuellen Unterschiede[2] zu überprüfen.

Semantisches Differenzial

Auswertung Semantisches Differenzial

Die Ergebnisse eines Hörversuchs mit semantischem Differenzial eignen sich ebenfalls meist für Korrelationsuntersuchungen und erlauben so eine umfangreiche Auswertung. Die Beurteilung eines Geräusches bzgl. mehrerer Beurteilungsitems erfordert mehr Zeit. Aus diesem Grund können in den meisten Hörversuchen dieser Art nicht alle Geräusche mehrmals abgefragt werden, da sonst der Test zu umfangreich wird und die Konzentration der Teilnehmer nachlässt. In den meisten Fällen ist es sinnvoll, zumindest einige Geräusche zweimal bewerten zu lassen, um so die Reliabilität und intraindividuellen Unterschiede der Teilnehmer zu überprüfen.



[1] Bei äquidistanten Kategorien werden die Abstände der vorgegebenen Kategorien ungefähr als gleich weit empfunden. Im gezeigten Beispiel mit „not“, „slightly“, „fairly“, „quite“ und „very“ sollte also der Abstand zwischen „not“ und „slightly“ als gleich groß wie der Abstand zwischen „quite“ und „very“ von den Teilnehmern wahrgenommen werden.

[2] Intraindividuelle Unterschiede sind die Unterschiede, die die Urteile einer Versuchsperson bei wiederholter Urteilsabgabe desselben Geräusches aufweisen.

2. Übersetzung der Hörversuchsergebnise in Zahlenwerte

Übersetzen der Beurteilungen in Zahlenwerte

Grundsätzlich müssen bei allen Testmethoden die Urteile der Teilnehmer in Zahlen übersetzt werden, sobald sie weiteren statistischen Auswertungen unterzogen werden sollen. Die Urteile einer fünfstufigen Kategorialskala erhalten z. B. die Zahlenwerte 1 bis 5. Bei einem semantischen Differenzial mit einer siebenstufigen, bipolaren Skala können z. B. die Zahlenwerte 1 bis 7 oder -3 bis 3 vergeben werden. Auch wenn die Skalen auf dem Bildschirm nicht immer in die gleiche Richtung weisen (die negativen Attribute stehen mal auf der linken mal auf der rechten Seite), kann es sinnvoll sein, die Zahlenwerte so zu vergeben, dass der höchste Wert immer dem positiven, der niedrigste immer dem negativen Skalenende entspricht (oder immer umgekehrt). Abbildung 1 zeigt hierzu ein Beispiel.

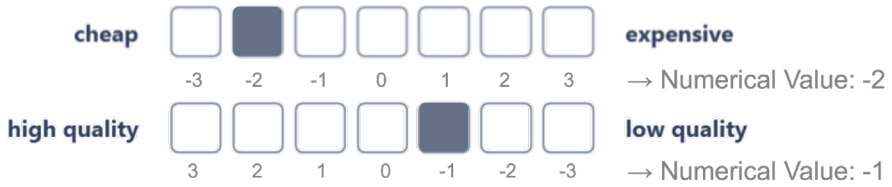


Abbildung 1: Übersetzung der Urteile in Zahlenwerte

Mit dem Jury Testing Modul SQala können Sie für jedes Attribut bzw. Attributpaar individuelle Werte für die einzelnen Skalenabschnitte festlegen.

Bei der folgenden Auswertung der in Zahlen umgerechneten Urteile muss berücksichtigt werden, dass diese Zahlen ursprünglich Urteile z. B. auf einer kategorialen Skala waren. Die eigentliche Urteilsabgabe darf durch die Umwandlung in Zahlenwerte, die nur der statistischen Auswertung dienen, nicht vergessen werden.

3. Statistische Untersuchung der Hörversuchsergebnisse

Statistische Mittelung

Wenn die Urteile der einzelnen Teilnehmer in Zahlenwerten vorliegen, können diese zunächst grafisch aufgetragen und verglichen werden. Dies vermittelt einen ersten Eindruck der Beurteilung und hilft bei der Entscheidung, welche statistischen Tests eingesetzt werden können und ob eine Mittelung der Urteile verschiedener Teilnehmer durchgeführt werden kann.

Histogramm und Normalverteilung

Mit einem Histogramm der abgegebenen Beurteilungen ist eine derartige Einschätzung meist sehr einfach möglich. Ein solches Histogramm zeigt für jedes bewertete Geräusch die Anzahl der Beurteilungen für die jeweilige Antwortkategorien an.

Sind die Beurteilungen der Teilnehmer normalverteilt (wie in Abbildung 2 schematisch dargestellt), kann eine Mittelung ohne signifikanten Informationsverlust erfolgen.

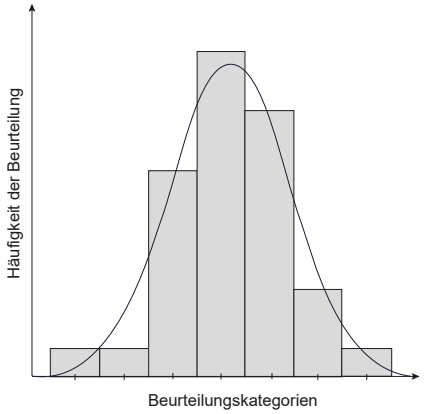


Abbildung 2: Histogramm mit normalverteilten Beurteilungen

Clusteranalyse

Besitzt die Verteilung stattdessen zwei oder mehrere Maxima, kann es sinnvoll sein, die Teilnehmer in zwei oder mehr Gruppen einzuteilen, in denen dann eine Mittelung durchgeführt werden kann (sogenanntes Clustern). Dies muss anhand der Daten für jeden Hörversuch individuell entschieden werden. Verschiedene Statistik-Programme stellen dem Benutzer entsprechende Analysemethoden zur Verfügung, die bei der Auswertung helfen (Clusteranalyse).

Mittelwertberechnung

Grafische Analysen können noch für weitere Überprüfungen herangezogen werden, um zu kontrollieren, ob eine Mittelung der Daten sinnvoll ist. Beispielsweise lässt sich durch eine geeignete Auftragung der Beurteilungen der verschiedenen Teilnehmer überprüfen, ob die Skalenausnutzung der Teilnehmer vergleichbar war.

Medianwert

Neben der Berechnung des arithmetischen Mittelwerts werden auch häufig der Medianwert und die Interquartilbereiche bestimmt. Der Medianwert ist der Wert, der von 50% der Beurteilungen überschritten und von den restlichen 50% unterschritten wird. Der Medianwert wird im Gegensatz zum arithmetischen Mittelwert von Extremwerten[3] kaum beeinflusst, daher eignet er sich meist gut für die Untersuchung von Hörversuchen, an denen nur wenige Personen teilgenommen haben. In diesem Fall liegen nur wenige Datenpunkte vor, was dazu führen kann, dass ein Ausreißer die Berechung des arithmetischen Mittelwertes stark verfälscht. Im Allgemeinen kann das arithmetische Mittel verwendet werden, wenn die Beurteilungen im Histogramm normalverteilt sind. Ist dies nicht der Fall, sollte der Medianwert berechnet werden.

[3] Extremwerte sind Urteile, die sehr weit von den anderen entfernt liegen.

Interquartilbereich

Der Interquartilbereich umschließt den Medianwert und zeigt den Bereich an, in dem die mittleren 50 % der Urteile liegen. Von den anderen 50 % der Urteile liegen 25 % unter dem Interquartilbereich und 25 % darüber. Der Interquartilbereich gibt somit direkt Aufschluss darüber, wie stark die Urteile der einzelnen Teilnehmer streuen. Mit sogenannten Boxplots können der Medianwert, die Interquartilbereiche und somit die Streuung der Daten sehr übersichtlich dargestellt werden (Abbildung 3). Dies ist sehr nützlich, um die Datenverteilung (normalverteilt um den Medianwert oder schief) abzulesen. Um die Lage der außerhalb des Interquartilbereiches liegenden Werte zu visualisieren, kann der Boxplot durch die Darstellung von Antennen (engl. Whiskers) ergänzt werden. Allerdings ist der zugrundeliegende Wert dieser Antennen nicht einheitlich definiert. Häufig wird das 1,5-fache des Interquartilbereichs dargestellt, Urteile außerhalb dieses Bereichs werden dann meist als Ausreißer bezeichnet. In einigen Fällen stehen die Antennen-Enden auch für die Maxima und Minima der Urteile.

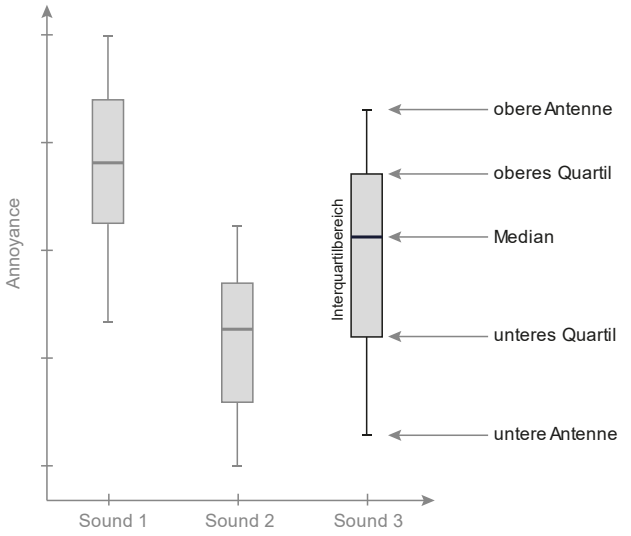


Abbildung 3: Darstellung von Geräuschbewertungen im Boxplots

Standardabweichung und Konfidenzintervall

Weitere häufig verwendete statistische Größen sind die Standardabweichung und das Konfidenzintervall. Bei der Berechnung der Standardabweichung wird die mittlere Abweichung vom arithmetischen Mittelwert bestimmt. Wie der Interquartilbereich ist die Standardabweichung ein Maß für die Streuung der Urteile. Je kleiner dieser Wert desto ähnlicher wurde das Geräusch von allen Teilnehmern bewertet. Das Konfidenzintervall gibt einen Bereich an, in dem das Ergebnis eines erneut durchgeführten Tests voraussichtlich liegen wird. So zeigt beispielsweise das 95%-Konfidenzintervall den Bereich, in dem das Ergebnis eines zusätzlichen Tests mit einer Wahrscheinlichkeit von 95 % liegen wird.

Statistik richtig anwenden

Die bereits angesprochene grafische Auswertung kann zusätzliche Hinweise geben, ob die Urteile eines Teilnehmers sich sehr deutlich von denen der anderen Teilnehmern unterscheiden (d. h. nicht nur in der Skalenausnutzung, sondern in der Kurvenform). Die Urteile dieser Teilnehmer müssen dann eventuell gesondert betrachtet werden und dürfen nicht in die Berechnung des Mittelwerts miteinbezogen werden. Das Ausschließen von Teilnehmern darf nicht leichtfertig angewendet werden. Der Versuchsleiter darf die Daten eines Versuchs nicht mithilfe der Statistik so verändern, dass das gewünschte Ergebnis „herbeigerechnet“ wird.

4. Weiterführende Auswertungen

Korrelationsanalyse

Nachdem die Urteile der Teilnehmer zu einem Mittelwert bzw. Medianwert zusammengefasst wurden, kann eine Korrelations- und Regressionsanalyse stattfinden. Dazu werden neben den Urteilen aus dem Hörversuch zusätzliche Daten für jedes Geräusch benötigt, z. B. die Ergebnisse aus messtechnischen Analysen. Liegen diese für jedes beurteilte Geräusch als Einzahlwerte vor, kann mithilfe der Korrelationsanalyse die Ähnlichkeit des Kurvenverlaufs aus den Ergebnissen des Hörversuchs und der messtechnischen Analyse bestimmt werden.

Regressionsanalyse

Bei der Regressionsanalyse werden die Daten aus dem Hörversuch und die Daten aus der messtechnischen Analyse in einem XY-Plot aufgetragen und der mathematische Zusammenhang zwischen den Achsen berechnet. Der Grad der Übereinstimmung dieser mathematischen Formel mit den eigentlichen Daten ist das Bestimmtheitsmaß R^2. Ein hohes Bestimmtheitsmaß sagt aus, dass die Ergebnisse des Hörversuchs sehr gut mithilfe der gefunden mathematischen Formel und den Ergebnissen aus der messtechnischen Analyse wiedergegeben werden können. Abbildung 4 zeigt ein einfaches Beispiel einer linearen Regressionsanalyse. Auf der X-Achse des Diagramms sind die für die Geräusche berechneten Einzahlwerte eine messtechnischen Analyse aufgetragen. Die Y-Achse stellt die Mittelwerte der Teilnehmerurteile dar. Die Beurteilung durch die Teilnehmer wird im gezeigten Beispiel durch die berechneten Analysewerte gut wiedergegeben.

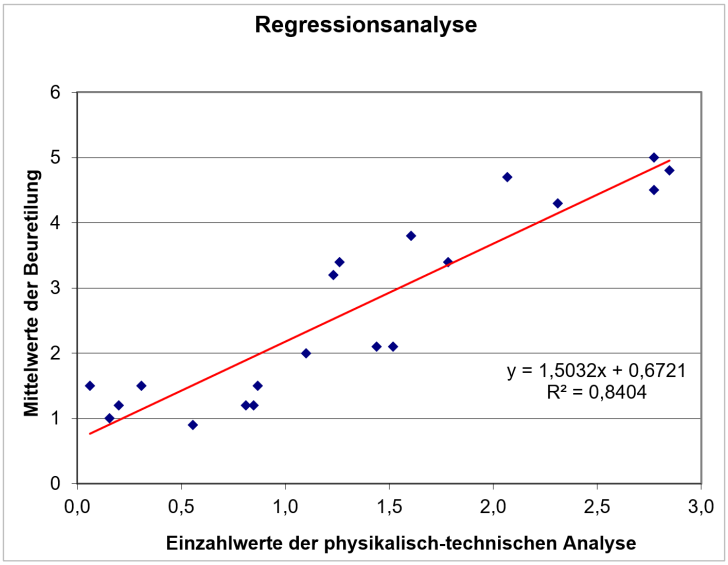


Abbildung 4: Beispiel für das Ergebnis einer Regressionsanalyse

Verwendung von Metriken

Für ein ausreichend hohes Bestimmtheitsmaß müssen unter Umständen die Ergebnisse mehrerer messtechnischer Analysen kombiniert werden. In ArtemiS SUITE kann dies mittels des Metrik-Projekts automatisiert erfolgen. Hinweise zur Metrikerstellung sind in der Application Note „Metrikerstellung“ zusammengestellt.

Entwicklung robuster Metriken

Grundsätzlich sollte bei der Metrikentwicklung eine hohe Korrelation zwischen den Ergebnissen des Hörversuchs und den Ergebnissen der messtechnischen Analysen erreicht werden. Es ist allerdings nicht zielführend, die Hörversuchsergebnisse mit einer großen Anzahl von Einzahlwerten aus vielen verschiedenen messtechnischen Analysen abzubilden. Bei der Metrikentwicklung muss der Einfluss jedes Einzahlwerts systematisch überprüft und mit Hinblick auf den kausalen Zusammenhang bzgl. der Geräuschqualität für die zu bewertenden Geräusche ausgewählt werden. Meist ist es

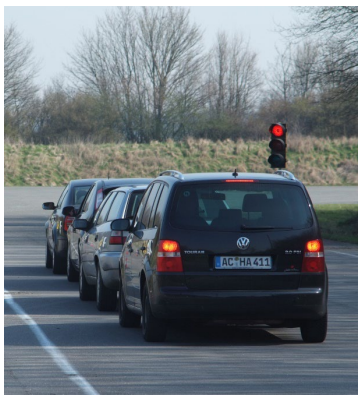
für die Erstellung einer robusten Metrik[4] sinnvoller, nur eine kleine Anzahl von Analysen zu verwenden, um eine Überanpassung des Modells zu verhindern. Sofern eine robuste Metrik erstellt werden konnte, lassen sich im Folgenden die Beurteilungen von Geräuschen mit ähnlicher Geräuschcharakteristik rechnerisch mithilfe der mathematischen Formel und den Ergebnissen der messtechnischen Analyse vorhersagen. Um die Qualität der Vorhersage zu bestätigen, können zusätzliche Validierungshörversuche durchgeführt werden.

Haupkomponenten-Analyse

Die Ergebnisse eines Hörversuchs mit semantischem Differenzial sind sehr umfangreich, weil die Teilnehmer ihre Beurteilung auf mehreren Attributskalen abgeben. Um die Datenmenge zu reduzieren, werden die Ergebnisse dieser Testmethode häufig einer Hauptkomponenten-Analyse unterzogen. Mithilfe einer solchen Analyse kann bestimmt werden, welche Attributpaare zusammengefasst werden können und wie groß ihr Einfluss auf die Beurteilung ist. Sobald einige Attributpaare zusammengefasst werden können, muss die Regressionsanalyse nur noch für die übergeordneten Hauptkomponenten und nicht mehr für jedes Attributpaar einzeln durchgeführt werden. Außerdem kann man die für die Gesamtbeurteilung entscheidenden Komponenten finden. Falls weitere Hörversuche mit ähnlichen Geräuschen durchgeführt werden sollen, besteht die Möglichkeit, auf einige der Attributpaare, die zusammengefasst werden konnten, zu verzichten. Dies erlaubt es dann, neue Attributpaare abzufragen, die zusätzliche Informationen liefern.

Untersuchung von nicht-stationären Signalen

Eine weitere Besonderheit stellt die Auswertung von Hörversuchen dar, in denen nicht-stationäre Geräusche beurteilt wurden. Nicht-stationäre Geräusche verändern sich in Abhängigkeit von der Zeit (z. B. das Fahrgeräusch bei einem Ampelstart oder einer Vorbeifahrt). Wenn ein Teilnehmer ein solches Signal beurteilen soll, muss er seinen Geräuscheindruck, der sich wie das Signal mit der Zeit ändern kann, in ein Urteil zusammenfassen. Diese „interne“ Mittelung durch die Teilnehmer wird im Allgemeinen nicht dem arithmetischen Mittelwert der Einzelurteile entsprechen. Und auch der Mittelwert einer messtechnischen Analyse wird in vielen Fällen nicht den Eindruck der Teilnehmer widerspiegeln.



Perzentilwerte

Im Fall von nicht-stationären Signalen hat sich die Berechnung von Perzentilwerten bewährt. Die Berechnung der Perzentilwerte ist eine statistische Auswertung, bei der die Werteverteilung der messtechnischen Analyse untersucht wird. Wenn Sie in ArtemiS SUITE den Wert *10* in die Perzentilwert-Tabelle auf der Eigenschaftenseite einer Analyse eintragen, wird der Einzahlwert ermittelt, der von 10% der Analyseergebnisse überschritten wird usw. Für zeitabhängige Analysen in Kombination mit dem Eintrag *5* bestimmt ArtemiS SUITE den Wert, der während des ausgewerteten Zeitraums in 5% der Zeit überschritten wird. In Abbildung 5 ist ein Beispiel einer zeitabhängigen Lautheitsanalyse dargestellt. Neben der zeitabhängigen Lautheitskurve sind die 5%-, 10%- und 50%-Perzentilwerte gekennzeichnet. Mit ArtemiS SUITE

[4] Hinweise zum Erstellen von robusten Metriken gibt die folgende Veröffentlichung: Fiebig, Kamp; “Development of metrics for characterizing product sound quality”, Proceedings Aachen Acoustics Colloquium 2015, 123-133.

können die Einzahlwerte entweder in der Diagrammlegende oder in einer Einzahlwerttabelle ausgegeben werden.

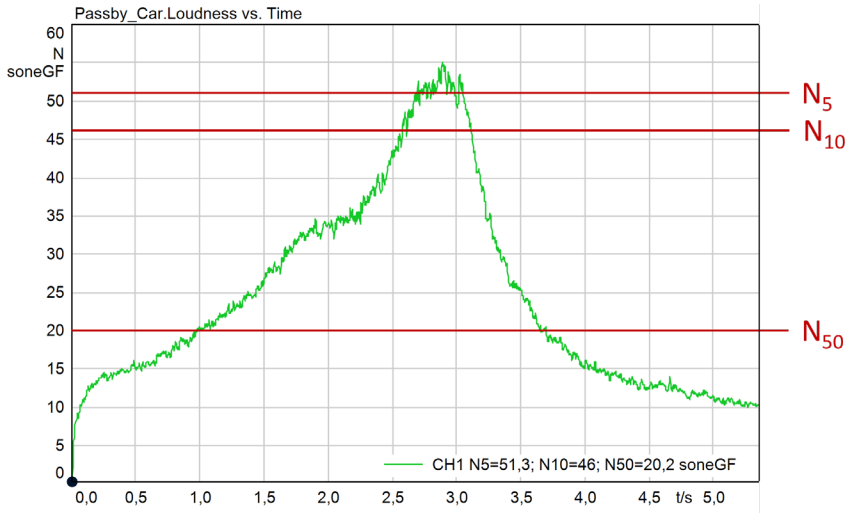


Abbildung 5: Beispiel für Perzentil-Werte

N5-Lautheit

Perzentilwerte korrelieren in vielen Fällen deutlich besser mit den Ergebnissen aus Hörversuchen als der arithmetische Mittelwert. Untersuchungen der Lästigkeit von Verkehrslärm zeigten, dass der 5%-Perzentilwert der Lautheit (N5) sehr gut mit der Beurteilung des Lärms in einem Hörversuch korreliert. Dieser Wert liegt höher als der durchschnittliche Lautheitswert, aber die lauten Anteile des Verkehrslärms fallen auch bei der Beurteilung durch die Teilnehmer sehr viel stärker ins Gewicht. Die DIN 45631/A1 schreibt daher die Berechnung der N5-Lautheit als Einzahlwert für zeitvariante Geräusche vor.

Für die Korrelationsuntersuchung sollten immer verschiedene Perzentilwerte bestimmt werden, um mehr über die von den Teilnehmern durchgeführte Gewichtung zu erfahren und den geeigneten Perzentilwert zu finden.

Zusammenfassend sollte bei der Auswertung folgendes beachtet werden: Jede mathematische Operation (Mittelwertbildung, Ausschluss eines Teilnehmers usw.) muss mit Bedacht ausgewählt und durchgeführt werden. Außerdem muss jede durchgeführte Maßnahme sorgfältig dokumentiert werden, um festzuhalten auf welcher Basis die Ergebnisse entstanden sind. Nur so kann eine aussagekräftige Interpretation der Ergebnisse erfolgen. Eine umfangreichere Einführung in die statistitische Auswertung von Testurteilen findet sich z. B. in Büchern über Testmethoden und deren Evaluation für Human- und Sozialwissenschaftler.