Bewertung von Messumgebung und Aufnahmequalität bei Online-Studien

Jule Pohlhausen, Inga Holube, Jörg Bitzer

Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg

Schlüsselwörter: Remote-Testverfahren, Referenzfreie Qualitätsbeurteilung, Umgebungsgeräusche, Nachhall, Übersteuerungen

Zusammenfassung

In der heutigen Zeit spielt die Einbindung von internetbasierten Testverfahren sowohl in die Hördiagnostik als auch in die Hörforschung eine wichtige Rolle. Die Möglichkeit, Patient_innen oder Proband_innen zuhause in ihrer alltäglichen Umgebung zu erreichen, zeichnet sich zunehmend als geeignete Erweiterung von klassischen Labormessungen ab. Jedoch sind Messumgebungen und Equipment individuell und weder vollständig dokumentier- noch kontrollierbar. Hieraus ergibt sich die Notwendigkeit, die Messumgebung möglichst unmittelbar zu bewerten, um einschätzen zu können, ob aus akustischer Sicht eine Messung möglich ist. Wesentlich ist, den störenden Einfluss durch Hintergrundgeräusche und Nachhall abzuschätzen. Im Hinblick auf Online-Studien, in denen die Antworten der Proband_innen aufgezeichnet werden, ist zudem eine Bewertung der Aufnahmequalität des verwendeten Mikrofons notwendig. Hierbei sind nicht nur verschiedene Mikrofontypen – intern, extern, Headset etc. – sondern auch die Mikrofoncharakteristik sowie mögliche Signalverarbeitungsalgorithmen z.B. zur Störgeräuschreduktion zu beachten. In einer Online-Studie wurden Hintergrundgeräusche, Sprache und Klatschen zum einen in einer möglichst ruhigen, zum anderen in einer geräuschvollen Umgebung aufgezeichnet. Diese Aufnahmen dienten als Testdaten für referenzfreie objektive Bewertungsmaße. Verglichen wurden objektive Bewertungsmaße aus der Literatur mit Messgrößen, die wenig Rechenleistung erfordern und eine echtzeitnahe Einschätzung ermöglichen. Ziel ist es, vor und möglichst auch während der Messung zu entscheiden, ob die Messumgebung geeignet sowie die Aufnahmequalität ausreichend gut ist und andernfalls auf zu laute Hintergrundgeräusche, Nachhall oder Verzerrungen hinzuweisen.

1. Einleitung

In der audiologischen Forschung sind Messungen mit Proband_innen essentiell. Jedoch hat die Corona-Pandemie für einen langen Zeitraum einen direkten persönlichen Kontakt zu Proband_innen deutlich erschwert. Zum Schutz aller Beteiligten sind deshalb Untersuchungs- und Testverfahren mit geringem bis keinem persönlichen Kontakt erstrebenswert. Ein wichtiger Lösungsansatz ist der Ausbau von Online-Studien als Erweiterung zu herkömmlichen Labormessungen. Online-Studien bezeichnen Testverfahren, z.B. zu psychologischen oder audiologischen Fragestellungen, die Proband_innen selbstständig zu Hause durchführen. Im Gegensatz zu der kontrollierbaren Messumgebung im Labor müssen bei Online-Studien die individuellen Rahmenbedingungen beachtet werden. Die Messumgebung sollte eine konzentrierte Durchführung zulassen, d.h. Störfaktoren wie Hintergrundgeräusche (Straßenlärm, Staubsauger, Lüftung etc.) oder weitere anwesende Personen sind zu vermeiden. Trotz der Möglichkeit, die Proband_innen um eine Messdurchführung allein und in einer möglichst ruhigen Umgebung zu bitten, ist zur Qualitätssicherung der Ergebnisse eine objektive Beurteilung der Messumgebung notwendig. Im Hinblick auf die Individualität im Messequipment bei Online-Studien, sowohl auf Seiten der Hardware als auch der Software, ist die Aufnahmequalität eine wichtige Größe.

Ziel des vorliegenden Beitrags ist es, Methoden zur objektiven Bewertung der Messumgebung sowie der Aufnahmequalität vorzustellen. Ein konkretes Anwendungsszenario ist die Qualitätsbewertung als Vortest und zudem parallel zu einem Online-Sprachtest (Hauptmessung). Erfolgt die Aufzeichnung der Antworten der Proband_innen über ein Mikrofon zur späteren Analyse mit einer automatischen Spracherkennung, ist die Detektion von Störfaktoren wie zu laute Hintergrundgeräusche oder Verzerrungen der Stimme wichtig. Mit dem Vortest erfolgt eine generelle Beurteilung, ob sich die Messumgebung und die Aufnahmequalität für die Durchführung des Experiments eignen. Die parallele Anwendung von Bewertungskriterien während der Hauptmessung dient zur Erkennung von kurzzeitig verschlechterten Messbedingungen.

2. Testmaterial

Für die Entwicklung und Definition von geeigneten objektiven Qualitätsmaßen bedarf es Testmaterial, das die individuelle Aufnahmesituation der Proband_innen zu Hause widerspiegelt. Deshalb wurde Testmaterial in einer Online-Studie über die Plattform Gorilla (www.gorilla.sc) mit der individuellen Soft- und Hardware der Proband_innen aufgenommen. Insgesamt haben 102 Proband_innen die Online-Studie gestartet. Dabei wurden 49 vollständige Datensätze erhoben. Das Alter der Proband_innen (66 % weiblich) lag zwischen 18-60 Jahren, mit einem Mittelwert von 30,9 ± 12,3 Jahren. Die Proband_innen nahmen Hintergrundgeräusche, Sprache und Klatschen zum einen in einer subjektiv wahrgenommenen möglichst ruhigen, zum anderen in einer geräuschvollen Umgebung auf. Zur Gestaltung einer geräuschvollen Umgebung wurden als Beispiele *Fenster öffnen, das Radio einschalten* oder *einen Staubsauger laufen lassen* genannt. Die Aufnahmedauer wurde jeweils auf 10 s begrenzt. Zur Bestimmung von Referenzgrößen wurden die Sprachaufnahmen händisch annotiert. Bei 26 Proband_innen war eine von der Testplattform zur Verfügung gestellte Echo-Unterdrückung (engl. echo cancellation, EC) aktiviert. Die Aktivierung der EC bewirkte eine zusätzliche, nicht vorhersagbare Aktivierung weiterer Signalverarbeitungsalgorithmen wie eine Störgeräuschunterdrückung.

3. Bewertung der Hintergrundgeräusche

Bei Online-Studien ist eine Vielzahl an akustischen additiven Störquellen denkbar. Hintergrundgeräusche stellen ein mögliches Ablenkungspotential dar und können zusätzlich maskierend auf das Zielsignal wirken. Als Bewertungsmaß eignet sich der Signal-Rausch-Abstand (engl. signal-to-noise ratio, SNR) mit der Stimme der Proband_innen als Zielsignal. Durch die getrennte Aufnahme von Hintergrund und Sprache kann unter der Annahme der Stationarität der Störung der SNR über die Differenz im äquivalenten Schalldruckpegel (engl. sound pressure level, SPL) dieser beiden Aufnahmen geschätzt werden; im Weiteren naive SNR-Schätzung genannt. Als Alternative wurde in Anlehnung an Chen [1] der Dynamikbereich (engl. dynamic range, DR) für die Sprachaufnahmen bestimmt. Statt der von Chen [1] vorgeschlagenen Differenz von Maximum und Minimum wurde zur Verbesserung der Robustheit das 90. und 10. Perzentil des blockweise berechneten SPL verwendet. Abbildung 1 vergleicht die naive SNR-Schätzung und den DR zu dem händisch annotierten Referenzwert SNR_{GT}.

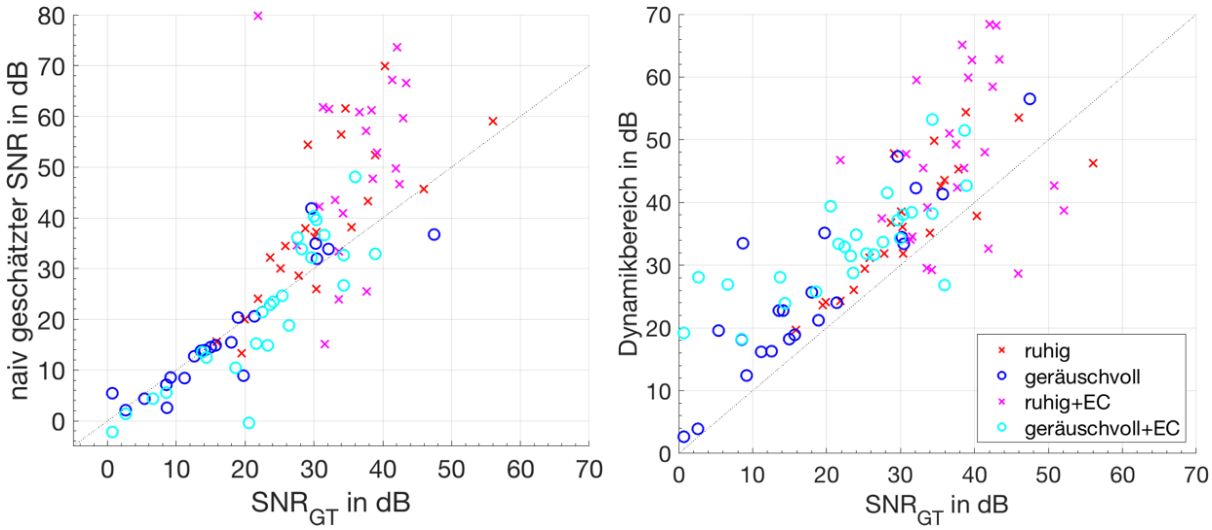

Abbildung 1: Naive SNR-Schätzung (links) und Dynamikbereich DR (rechts) in Abhängigkeit vom Referenzwert SNR_{GT} in dB für die Sprachaufnahmen in ruhiger Umgebung (Kreuze) sowie in einer möglichst geräuschvollen Umgebung (Kreise) jeweils ohne und mit EC.

Insgesamt erschwerte die EC eine Bewertung der Messumgebung und der Aufnahmequalität. Die naive SNR-Schätzung ist für Aufnahmen ohne EC ausreichend genau und eignet sich für den Vortest zur Bewertung der Hintergrundgeräusche. Die teilweise sehr hohen Überschätzungen der naiven SNR-Schätzung betreffen vor allem Aufnahmen in ruhiger Umgebung mit einem $SNR_{GT} > 30$ dB, d.h. einer ausreichend guten Messumgebung, und sind vernachlässigbar.

4. Bewertung des Nachhalls

Die Messumgebung ist grundlegend gekennzeichnet durch die vorherrschende Raumakustik. Ein Raum mit zu viel Nachhall verschlechtert die Sprachverständlichkeit, da leisere Sprachanteile wie Konsonanten von lauteren Sprachanteilen wie Vokalen maskiert werden [2]. Die wichtigste akustische Kenngröße eines Raumes ist die Nachhallzeit τ. Die Nachhallzeit ist definiert als die Zeit, in der nach Abschalten einer stationären, breitbandigen Quelle die Schallenergiedichte um 60 dB abgesunken ist. In der Praxis ist die Schätzung von τ durch den SNR limitiert, sodass meistens die Abklingzeiten von −5 dB auf −25 dB (τ_{20}) bzw. auf −35 dB (τ_{30}) relativ zum Anfangspegel bestimmt und auf −60 dB extrapoliert werden. Maßgebend für τ sind das Raumvolumen und die schallabsorbierenden Eigenschaften des Raums. Die Energieabklingkurve (engl. energy decay curve, EDC) kann über die Rückwärtsintegration der quadrierten Raumimpulsantwort *h(t)* bestimmt werden [3]

$$EDC(t) = \int_t^{\infty} h^2(\kappa)\mathrm{d}\kappa. \tag{1}$$

Die Impulsantwort kann durch Anregung des Raumes mit Klatschen angenähert werden [4]. Klatschen ist ein einfaches und leicht verfügbares Anregungssignal. Jedoch ist es im Vergleich zu den üblichen Anregungssignalen bei der Messung von raumakustischen Größen weniger genau reproduzierbar und weist typischerweise keinen flachen Frequenzgang auf, sondern einen Abfall zu tiefen Frequenzen unter 500 Hz. Papadakis und Stavroulakis [4] empfehlen, mindestens viermal zu klatschen und anschließend die entsprechenden raumakustischen Parameter zu mitteln. Die Klatschaufnahmen enthalten eine unbekannte Anzahl an Klatschereignissen. Folglich ist eine Detektion der Klatschstartzeitpunkte mit anschließender Segmentierung notwendig (s. Abbildung 2). Das Klatschkriterium definiert 80 % der absoluten Maximalamplitude der jeweiligen Aufnahme, da nicht alle Klatschereignisse innerhalb einer Aufnahme die gleiche Spitzenamplitude erreichten. Zudem wurde ein zeitlicher Mindestabstand von 500 ms zwischen zwei Klatschereignissen eingeführt. Um den Einfluss von ggf. tieffrequenten Störgeräuschen auf die Schätzung von τ_x zu reduzieren, wurde ein Hochpass auf die Klatschaufnahme angewendet. Dazu wurde ein Butterworth-Hochpassfilter 2. Ordnung mit einer Grenzfrequenz von 500 Hz verwendet.

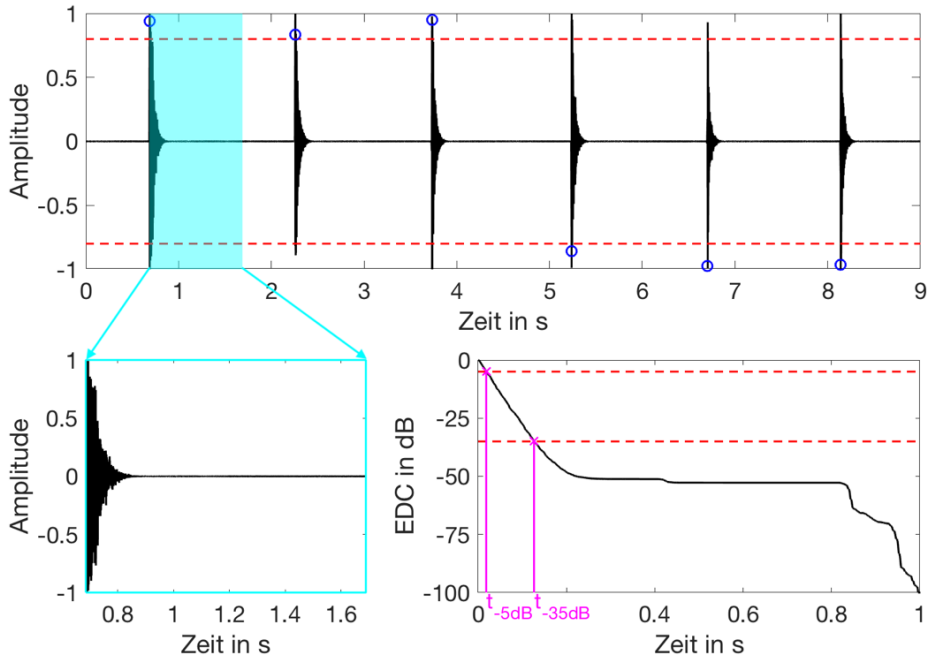

Abbildung 2: Analyse einer Klatschaufnahme mit sechs Klatschereignissen (oben): Die roten Linien kennzeichnen das Klatschkriterium für die Spitzenamplituden. Die blauen Kreise markieren die detektierten Klatschzeitpunkte. Der hellblaue Bereich enthält das erste Segment und wird unten links vergrößert dargestellt. Die entsprechende EDC ist unten rechts dargestellt. Markiert wurden die Abklingwerte -5 und -35 dB und die zugehörigen Zeitpunkte t_{-5dB} und t_{-35dB} zur Schätzung von τ_{30}.

Die Auswertung der Klatschaufnahmen ergab, dass für die meisten Aufnahmen die Unterschiede zwischen τ_{20} und τ_{30} vernachlässigbar gering sind. Dies ist ein Indikator für einen ausreichenden SNR des Klatschens zu Hintergrundgeräuschen. Die mittlere τ_{30} lag meist zwischen 0,2-0,5 s und entspricht damit den zu erwartenden Werten für gewöhnliche Wohnzimmer [2]. Aufnahmen mit sanftem Klatschen führten zu unrealistisch hohen τ_x. Folglich ist eine genaue Instruktion ggf. mit Demonstration der gewünschten Handhaltung notwendig.

5. Detektion von Übersteuerungen

Übersteuerungen zählen zu den nichtlinearen Verzerrungen. Ein übersteuertes Signal erfährt eine Begrenzung oder aus dem Englischen „Clipping“ auf einen bestimmten Amplitudenbereich. Das Ziel war nicht eine exakte zeitliche Detektion, sondern eine Rückmeldung über das generelle Vorhandensein von Übersteuerungen. Eine Möglichkeit zur Detektion von Übersteuerungen bietet die Analyse des Amplitudenhistogramms. Durch Übersteuerungen treten entweder einseitig oder beidseitig – für positive und negative Amplituden – Überhöhungen bei den Maximalwerten der Signalamplitude auf. In Anlehnung an den Clipping-Koeffizienten nach Aleinik und Matveev [5] erfolgte eine vereinfachte Definition des Clipping-Koeffizienten

$$C = \frac{\max\left(N_{min}, N_{max}\right)}{N_{median}}, \tag{2}$$

mit der absoluten Anzahl an Samples *N,* welche sich in dem Wertebereich um das Minimum, das Maximum sowie den Median $\tilde{x}$ der Signalamplituden *x(n)* befinden (s. Abbildung 3).

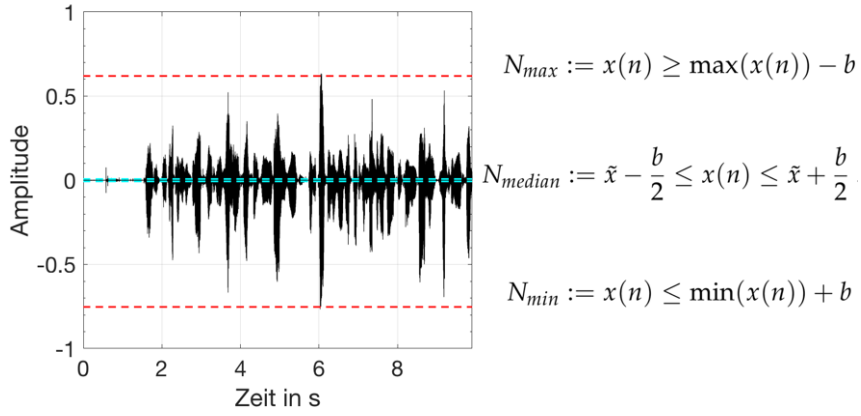

Abbildung 3: Der Definitionsoperator := beschreibt die Bedingungen für die Wertebereiche aus Gleichung 2. N ergibt sich jeweils aus der Summe aller Samples, die die Bedingung erfüllen. Die Breite des Wertebereichs b wurde empirisch auf 1 % vom Abstand der maximalen zur minimalen Amplitude gesetzt.

Für Signale ohne Übersteuerung ist $N_{median} \gg N_{min}$ und $N_{median} \gg N_{max}$ zu erwarten und somit $C \rightarrow 0$. Mit künstlichem Hard-Clipping (Clipping-Schwelle beim 99. Perzentil der absoluten Amplitudenwerte) steigt *C* deutlich. Die Evaluation des eingeführten Clipping-Koeffizienten *C* anhand der Sprachaufnahmen in ruhiger Umgebung ergab eine zuverlässige Detektion von übersteuerten Aufnahmen.

6. Literatur

[1] Chen, F. (2016). „Modeling Noise Influence to Speech Intelligibility Non-Intrusively by Reduced Speech Dynamic Range“. Interspeech 2016. ISCA, S. 1359–1362. DOI: 10.21437/Interspeech.2016-9.

[2] Kuttruff, H. (2009). Room acoustics. 5. Auflage. London & New York: Spon Press/Taylor & Francis.

[3] Schroeder, M. (1965). „New Method of Measuring Reverberation Time“. J. Acoust. Soc. Am 37, S. 409–412.

[4] Papadakis, N. M. und G. E. Stavroulakis (2020). „Handclap for Acoustic Measurements: Optimal Application and Limitations“. Acoustics 2(2), S. 224–245. DOI: 10.3390/acoustics2020015.

[5] Aleinik, S. und Y. Matveev (2014). „Detection of Clipped Fragments in Speech Signals“. World Academy of Science, Engineering and Technology 86, S. 703–709.