Automatische Erkennung von Abschnitten mit kritischer Sprachverständlichkeit für Normal- und Schwerhörende in Film und Fernsehen

Moritz Wächtler[1,2], Jan Rennies[1,2], Birger Kollmeier[1,2,3]

[1] *Fraunhofer-Institut für Digitale Medientechnologie, Projektgruppe Hör-, Sprach- und Audiotechnologie, 26129 Oldenburg, Deutschland, E-Mail: jan.rennies@idmt.fraunhofer.de*

[2] *Cluster of Excellence, Hearing4all*

[3] *Carl-von-Ossietzky-Universität Oldenburg, 26129 Oldenburg, Deutschland*

Einleitung

Häufig betreffen Beschwerden über Filme und Fernsehprogramme die im Verhältnis zur Sprache zu hohe Lautstärke von Hintergrundgeräuschen bzw. Musik und die daraus resultierende mangelhafte Sprachverständlichkeit [1]. Aufgrund des relativ hohen Altersdurchschnitts der Zuschauer der öffentlich-rechtlichen Sender in Deutschland von ca. 60 Jahren [2] liegt es nahe, dass Effekte der Altersschwerhörigkeit (Presbyakusis) hierbei eine Rolle spielen. Eine Ursache für diese nicht zufriedenstellenden Mischverhältnisse von Sprache und sonstigen Signalteilen (Musik, Effekte, Atmosphäre etc.) lässt sich womöglich im Postproduktionsprozess finden. In diesem findet die Sichtung, Selektion, Nachbearbeitung (auch Nachvertonung) und Mischung des zuvor aufgenommenen Audiomaterials durch den Tonverantwortlichen statt. Dieser besitzt ein geübtes sowie in der Regel nicht-pathologisches Gehör, ist mit dem Inhalt der Sprachaufnahmen vertraut und verfügt zudem über sehr gute Abhörbedingungen (gute Raumakustik, hochwertige tontechnische Anlagen, ruhige Umgebung). Diese Faktoren erschweren eine Beurteilung der Sprachverständlichkeit des durchschnittlichen Konsumenten mitsamt seinen technischen Möglichkeiten und seiner Umgebung. Eine pauschal angewendete deutliche Absenkung des Pegels der sprachmaskierenden Signalkomponenten würde dem Konsumenten zwar das Verstehen der Sprache erleichtern, bedeutete aber zugleich einen zu starken Eingriff in die klangliche Ästhetik der Filmmischung.

Modelle für Sprachverständlichkeit könnten hier eine Abhilfe schaffen, da sie eine objektivierte und zudem zeiteffiziente Beurteilung des Materials erlauben. Zudem ist durch die Modellierung von Hörverlusten eine Anpassung an verschiedene Zielgruppen möglich. In dieser Studie erfolgte die Entwicklung und Evaluation eines modellbasierten Messtools, welches in der Lage ist, während der Mischung in der Postproduktion, also beim Vorliegen getrennter Spuren für Sprache und Störgeräusche, Abschnitte mit kritischer Sprachverständlichkeit bzw. Höranstrengung zu detektieren.

Da bis jetzt noch kein etabliertes subjektives Verfahren existiert, um Audiomaterial aus dem Rundfunkbereich bezüglich Sprachverständlichkeit bzw. Höranstrengung zu untersuchen, stellte sich des Weiteren die Frage, ob die hier durchgeführte subjektive Skalierung der Höranstrengung für diesen Einsatzbereich eine geeignete Methode darstellt.

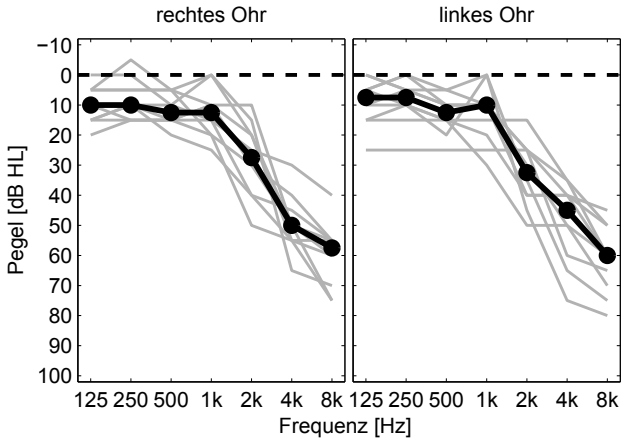

Abbildung 1: Audiogramme der älteren VPen. Die dicken schwarzen Linien geben interindividuelle Mediane an, während die dünneren grauen Linien individuelle Schwellen zeigen.

In diesem Beitrag wird zunächst ein Experiment vorgestellt, bei dem die Höranstrengungen von authentischem Rundfunkmaterial durch jüngere und ältere Versuchspersonen subjektiv beurteilt wurde. Es folgt die Beschreibung eines modellbasierten Frameworks zur Detektion von Signalabschnitten mit hoher Höranstrengung und eine Evaluation des Selbigen mit Hilfe der zuvor gewonnen subjektiven Daten.

Experiment

Methode

Elf jüngere (23 bis 29 Jahre) und zehn ältere (63 bis 77 Jahre) Versuchspersonen (VPen) nahmen an dem Versuch teil. Die jüngeren VPen hatten Reintonhörschwellen $\leq$ 15 dB HL bei Oktavfrequenzen von 125 bis 8000 Hz, mit Ausnahme einer VP, welche auf einem Ohr eine Schwelle von 20 dB HL bei 2000 Hz besaß. Die Hörschwellen der älteren VPen sind in Abbildung 1 dargestellt und lassen auf alterstypische Hochtonhörverluste schließen. Keine der VPen verwendete während des Versuchs Hörgeräte oder andere technische Hörhilfen.

29 verschiedene Tonausschnitte aus Fernsehproduktio-

nen, welche freundlicherweise vom Norddeutschen Rundfunk zur Verfügung gestellt wurden, kamen im Versuch zum Einsatz. Die Ausschnitte verfügten über Dauern von 5 bis 15 Sekunden und lagen aufgetrennt in Sprach- und Störgeräuschspur vor, wobei Letztere Musik, Umgebungsgeräusche oder auch andere Sprecher enthielt. Jeder der 29 Ausschnitte lag für den Versuch in zwei verschiedenen Signal-Rausch-Abständen (SNR) vor, wobei jede SNR-Variante zweimal präsentiert wurde (Test und Retest), wodurch sich eine Gesamtzahl von 116 Darbietungen pro VP ergab. Die Reihenfolge der Darbietungen wurde für jede VP randomisiert. Das dazugehörige Bildsignal wurde nicht gezeigt.

Die VPen wurden instruiert, nach jeder Signaldarbietung einzuschätzen, wie viel Anstrengung sie für das Verstehen der Sprache aufbringen mussten. Dafür stand ihnen eine 13-teilige kategoriale Höranstrengungsskala nach [3] zur Verfügung, deren Kategorien von „mühelos“ bis „extrem anstrengend“ betitelt waren. Jeder dieser Kategorien war dabei ein für die VPen nicht sichtbarer numerischer Wert in der Einheit Effort Scaling Categorical Unit (ESCU) zugeordnet, wobei 1 ESCU „mühelos“ und 13 ESCU „extrem anstrengend“ entsprachen.

Das Experiment fand in einer Wohnzimmerumgebung statt, wobei die Signale über Stereolautsprecher des Typs Yamaha NS-FS210 wiedergegeben wurden. Jede VP hatte im Vorfeld des eigentlichen Versuchs die Möglichkeit, sich mit Hilfe eines Testsignals einen angenehmen Wiedergabepegel einzustellen. Auf die nach EBU R 128 [4] bestimmte Lautheit des so eingestellten Testsignals wurden im Folgenden alle Signale normalisiert.

Ergebnisse

Abbildung 2 zeigt die Ergebnisse der Höranstrengungsmessung. Individuelle Ergebnisse wurden aus dem Mittel der beiden Werte für Test und Retest bestimmt. Über diese Mittelwerte wurden im Folgenden die abgebildeten interindividuellen Mediane gebildet. Die Fehlerbalken repräsentieren die Interquartilspannweite. Die Signale wurden für die Darstellung anhand der Mediane der jüngeren VP-Gruppe sortiert. Es ist zu erkennen, dass viele Bewertungen der Höranstrengung auf niedrige Kategorien entfielen. Die älteren VPen schätzten die Höranstrengung in vielen Fällen höher ein als die jüngeren (Median-Differenz: 1 ESCU), wobei allerdings eine deutliche Abhängigkeit vom Signal zu beobachten ist. Die Interquartilspannweiten erstreckten sich bei den Jüngeren über einen Bereich von 1 bis 5,5 ESCU (Median: 3 ESCU) und bei den Älteren von 1 bis 5 ESCU (Median: 3 ESCU). Statistische Analysen mit einem Signifikanzniveau von 0,05 zeigen, dass sich die Varianzen der interindividuellen Medianwerte von jüngeren und älteren VPen nicht signifikant unterschieden (Brown-Forsythe-Test, $p = 0{,}28$). Unter dieser Voraussetzung kann ein einseitiger Wilcoxon-Rangsummen-Test durchgeführt werden, welcher ergibt, dass die interindividuellen Mediane der älteren VPen signifikant höher waren als die der jüngeren ($p = 0{,}0013$).

In Tabelle 1 sind Einzahlwerte für den Zusammenhang zwischen Test und Retest für die jüngeren und älteren VPen dargestellt. Hierbei stellen ρ_S den Rangkorrelationskoeffizienten nach Spearman, Bias die mittlere vorzeichenbehaftete Differenz und ε den Root-Mean-Square-(RMS)-Fehler dar. Es sind Einzahlwerte basierend auf individuellen Ergebnissen und Medianergebnissen (in Klammern) angegeben. Für die Berechnung wurden alle Test-Retest-Paare ignoriert, die mindestens eine Bewertung von 1 oder 13 ESCU enthielten, um eine Verfälschung durch Boden- und Deckeneffekte zu vermeiden. Es zeigt sich ein starker Zusammenhang ohne systematische Fehler (Bias ≈ 0) zwischen den Mediandaten von Test und Retest für die jüngeren als auch die älteren VPen .

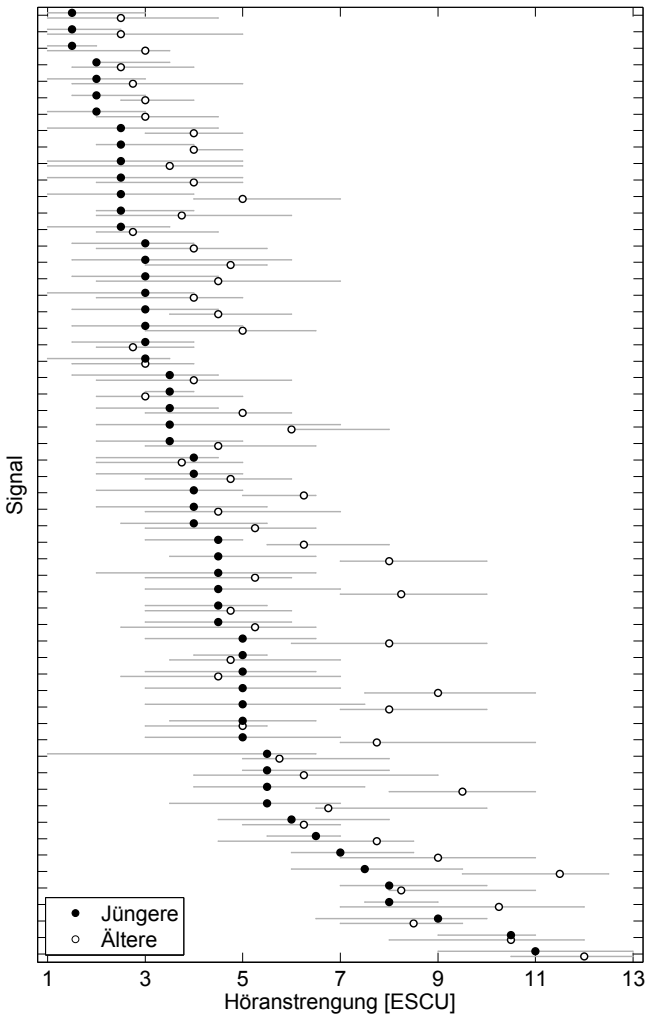

Abbildung 2: Mediane der kategorialen Höranstrengung für jüngere und ältere VPen für alle 58 im Versuch getesteten Signale (29 Ausschnitte bei je zwei SNR-Varianten). Die Signale wurden anhand der Mediane der jüngeren VPen sortiert. Die Fehlerbalken geben Interquartilspannweiten an.

Tabelle 1: Test- und Retest-Statistiken für die jüngeren und älteren VPen.

	Jüngere	Ältere
ρ_S	0,72 (0,82)	0,68 (0,80)
Bias [ESCU]	0,0 (-0,1)	0,1 (-0,1)
ε [ESCU]	1,6 (0,9)	1,9 (1,0)

Modellierung

Zum Zwecke der Modellierung wurden der Speech Intelligibility Index (SII, [5]), das Short-Time Objective Intelligibility Measure (STOI, [6]) und das Lautheitsmodell nach EBU R 128 [4] herangezogen und in ein Framework eingebettet. Dieses agiert hier als ein binärer Klassifikator, teilt Signalabschnitte also in die Kategorien „anstrengend“ und „nicht anstrengend“ ein. Das Framework verarbeitet linken und rechten Kanal zunächst separat. Die erste Stufe des Frameworks besteht aus einem spektralen Filter, welches die Übertragungsfunktion von den im Experiment verwendeten Lautsprechern und dem Versuchsraum nachbildet. Es folgt eine sogenannte makroskopische Blockeinteilung (Länge 2 Sekunden), um in längeren Rundfunkprogrammen einzelne Zeitabschnitte genauer betrachten zu können (diese wird in dieser Evaluation nicht durchgeführt, stattdessen wird ein Signal als ein Block betrachtet). Alle folgenden Analyseschritte werden blockweise durchgeführt. Anschließend wird mit einem RMS-basierten Sprachaktivitätsdetektor geprüft, ob der aktuelle Block über nennenswerte Sprachanteile verfügt. Ist dies nicht der Fall, so erfolgt keine Analyse durch ein Sprachverständlichkeitsmodell. Andernfalls werden mikroskopische Signalabschnitte (Länge ca. 26 ms), die keine Sprache enthalten (erneut RMS-basierte Entscheidung), aus Sprache und Störgeräusch entfernt, um daraufhin das Signal mit dem Sprachverständlichkeitsmodell zu analysieren. Das Ergebnis sind zwei Modellindizes, von denen im Folgenden nur noch der größere Verwendung findet. Unterschreitet dieser einen Modellschwellwert θ, so wird der aktuelle makroskopische Block als „anstrengend“ deklariert, andernfalls als „nicht anstrengend“. Für das EBU-Lautheitsmodell wurde die Lautheitsdifferenz zwischen Sprache und Störgeräusch als Modellindex verwendet.

Zur Evaluation des Frameworks wurden die subjektiven Daten dichotomisiert. Im Zuge dessen wurden Signale mit Median-Höranstrengungswerten ≥ 5 ESCU als „anstrengend“, alle anderen als „nicht anstrengend”gekennzeichnet. Das Vorliegen von subjektiven Daten und Modellergebnissen in binärer Form ermöglicht eine Evaluation mit sogenannten Receiver-Operating-Characteristic-(ROC)-Kurven. Hierbei wird der Modellschwellwert θ vom kleinsten bis zum größten in den Daten vorkommenden Modellindex variiert und Richtig-Positiv- und Falsch-Positiv-Rate (TPR und FPR) werden für jedes θ gegeneinander aufgetragen. Die Richtig-Positiv-Rate beschreibt den Anteil der durch das Framework korrekterweise als „anstrengend“ identifizierten Signale. Die Falsch-Positiv-Rate gibt hingegen den Anteil der fälschlicherweise durch das Framework als „anstrengend“ deklarierten Signale an. Als Referenz bzw. Goldstandard dienten die subjektiven Medianergebnisse. Abbildung 3 zeigt die ROC-Kurven für die Daten von jüngeren und älteren VPen und die Vorhersagen von SII, STOI und EBU-Modell. Die Winkelhalbierende der Koordinatenachsen (hellgrau, gepunktet) gibt die Leistung eines fiktiven Modells an, dessen Vorhersagen rein zufällig sind, welches also über keinerlei Diskriminationsfähigkeit verfügt. Die jeweils linke obere Ecke der Abbildungen (TPR = 1, FPR = 0) kann als Optimalpunkt angesehen werden, da dort eine perfekte Klassifikation erreicht wird. Tabelle 2 zeigt die minimalen Euklidischen Abstände der ROC-Kurven zu diesem Optimalpunkt für die drei Modelle. Aus Abbildung 3 und Tabelle 2 lässt sich erkennen, dass für die Daten der jüngeren VPen das EBU-Modell am besten abschneidet, während für die älteren VPen das STOI die beste Diskriminationsleistung zeigt.

Tabelle 2: Minimale Euklidische Abstände der ROC-Kurven zum Optimalpunkt (TPR = 1, FPR = 0). Kleinere Werte sind besser.

	Jüngere	Ältere
SII	0,30	0,20
STOI	0,24	0,15
EBU R 128	0,21	0,25

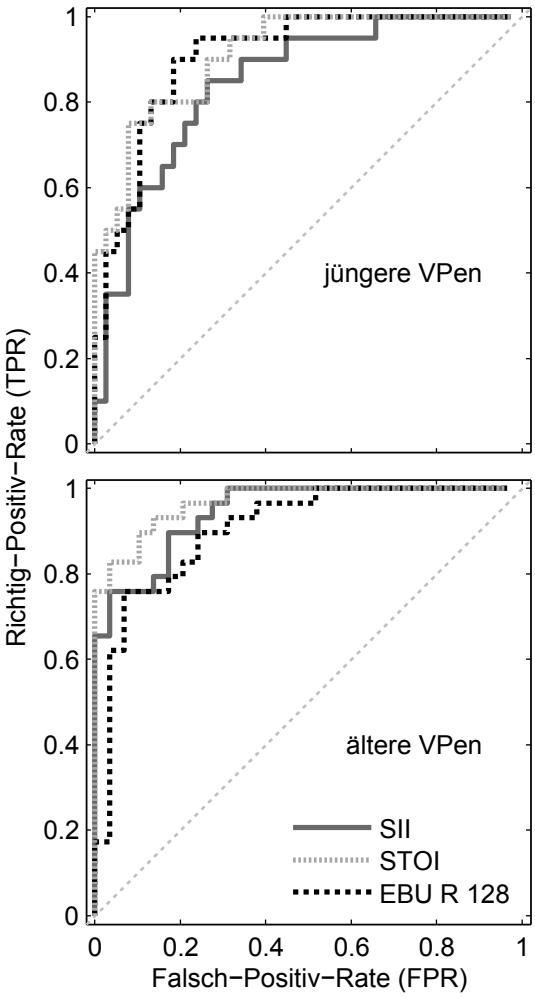

Abbildung 3: Receiver-Operating-Characteristic-Kurven für die Daten von den jüngeren (oben) und den älteren (unten) VPen.

Diskussion

Es wurde die subjektive Höranstrengung von TV-Ausschnitten durch jüngere und ältere VPen bewertet.

Bezüglich der Mediandaten zeigen sich bei beiden Gruppen große Übereinstimmungen zwischen den Ergebnissen von Test und Retest, wodurch geschlussfolgert werden kann, dass die Skalierung der Höranstrengung eine geeignete Methode zur Bewertung von TV-Material darstellt.

ROC-Analysen zeigen, dass die drei hier verwendeten Modelle in der Lage sind, Signale mit hoher Höranstrengung zu detektieren. Bei einer Gesamtbetrachtung der Daten von älteren und jüngeren VPen stellt sich kein Modell als eindeutig überlegen heraus, wobei insbesondere für die Zielgruppe älterer Hörer SII und STOI vorteilhaft gegenüber dem EBU-Modell zu sein scheinen. Selbst das Lautheitsmodell aus EBU R 128 zeigt trotz seiner vergleichsweise simplen Struktur für die hier betrachteten Signalabschnitte gute Ergebnisse. Eventuell lässt sich dies darauf zurückführen, dass das EBU-Modell bereits wichtige Komponenten für eine Bewertung von Sprache im Störgeräusch, wie ein Gating (Ausschluss niedrig ausgesteuerter Signalabschnitte), eine rudimentäre Frequenzgewichtung und eine RMS-Bestimmung, enthält.

Das hier vorgestellte Framework verwendet eine einfache RMS-basierte Entscheidungsstufe, um Signalabschnitte mit Sprache zu detektieren. Da auch in der Sprachspur mitunter Störgeräusche vorhanden sind (etwa Nebengeräusche vom Drehort), könnte mit einem fortgeschritteneren Algorithmus zur Sprachaktivitätserkennung das Framework robuster gestaltet werden.

Der hier verfolgte erste Ansatz scheint vielversprechend, um eine objektive Bewertung von Sprachverständlichkeit für Film und Fernsehen zu ermöglichen. Insbesondere Modelle aus der Audiologie sind dabei aussichtsreich, da sie sich in zukünftigen Studien direkt auf die Zielgruppe schwerhörender Probanden anpassen lassen, wodurch im besten Fall eine weitere Verbesserung der Vorhersagegenauigkeit erreicht werden kann.

Literatur

[1] Hildebrandt, E. (2014). Sprachverständlichkeit im Fernsehen: Vorstellung von ausgewählten Teilaspekten zu diesem Thema im Kontext der Entwicklung einer Production Guideline. Diplomarbeit, Universität für Musik und darstellende Kunst Wien.

[2] Giersch, V. (2008). Ein nur noch seltenes Paar: Öffentlich-rechtlicher Rundfunk und Jugend – Strategien gegen den Generationenabriss. ARD-Jahrbuch 08.

[3] Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D. und Spriet, A. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. The Journal of the Acoustical Society of America, 127(3):1491-1505.

[4] EBU (2014). Loudness Normalisation and Permitted Maximum Level of Audio Signals. Recommendation R 128 (European Broadcasting Union, Genf, Schweiz).

[5] ANSI (1997). Methods for calculation of the speech intelligibility index. American National Standard S3.5-1997 (Standards Secretariat, Acoustical Society of America, New York, USA).

[6] Taal, C. H., Hendriks, R. C., Heusdens, R. und Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech. IEEE Transactions on Audio, Speech, and Language Processing, 19(7):2125-2136.