Verringerung der Höranstrengung von TV-Mischungen durch Vorverarbeitung einzelner Spuren während der Mischung

H. Baumgartner, A. Volgenandt, J. Rennies-Hochmuth
Fraunhofer Institut für digitale Medientechnologie (IDMT), Oldenburg,
E-Mail:Hannah.Baumgartner@idmt.fraunhofer.de

Einleitung

In Schleswig-Holstein initiierten der Landesseniorenrat und die Ärztekammer eine „Kampagne für deutliche Sprache“. Leserbriefe und Beschwerden wegen schlechter Sprachverständlichkeit im Rundfunk häufen sich bei den Programmverantwortlichen. Besonders aufwendig und budgetintensiv produzierte Spielfilme wie z.B. „Tatort“ oder „Polizeiruf“ stehen häufig in der „Sprachverständlichkeits-Kritik“. Die Problematik ist komplex, die Ursachen nicht eindeutig.

Wie eine Studie zeigte, bewerten schwerhörende Probanden die Höranstrengung für eine Original-Fernsehmischung im Mittel so, wie Normalhörende eine Mischung mit einem um 6dB verringerten SNR [1]. Bezüglich der präferierten Abhörlautstärke unterscheiden sich die beiden Gruppen n dieser Studie dagegen kaum. Sicherlich wären Dialoge leichter verständlich, wenn man Hintergrundgeräusche reduzieren würde - diese Vorgehensweise liegt aber nicht im Interesse der Filmschaffenden, da der Einsatz von Atmosphären, aufwendigen Soundeffekten und musikalischer Untermalung eine bessere Identifikation der Zuschauer mit der visuellen Szenerie beabsichtigt, Anhaltspunkte für die Orientierung in Zeit und Raum liefert, für Spannung sorgt und die emotionale Beteiligung fördert.

Um die Höranstrengung trotz aufwendiger Untermalung zu verringern wurden in dieser Studie Algorithmen getestet, welche vor dem Mixing sowohl auf Sprache als auch Hintergrundsignale automatisiert angewandt werden können. Mit Hilfe eines an das MUSHRA-Testdesign [2] angelehnten Verfahrens konnten normal- und schwerhörende Probanden die unterschiedlich prozessierten TV-Mischsignale bewerten und ihre Präferenz signalisieren. Durch den Einsatz der Algorithmen konnte die Höranstrengung für beide Zielgruppen deutlich verringert werden, ohne die Atmo in Sprachpausen zu reduzieren oder ganz auf animierende Hintergrund-Klangwelten zu verzichten. Außerdem konnte gezeigt werden, dass Probanden aus beiden Zielgruppen zu einem großen Teil die vorverarbeitete Mischung der Originalmischung vorziehen.

SI4B-kontrollierte Algorithmen

Die Algorithmen wurden zur Generierung erster Korrekturvorschläge für professionelle Anwender aus der Postproduktion entwickelt.

In vorangegangenen Studien wurde die Höranstrengung von TV-Ausschnitten mit diversen Hintergründen (Atmos) in unterschiedlichen Mischverhältnissen mit normalhörenden und schwerhörenden Probanden erfasst. Aus den Daten wurde ein Modell zur objektiven Schätzung der Höranstrengung erstellt.

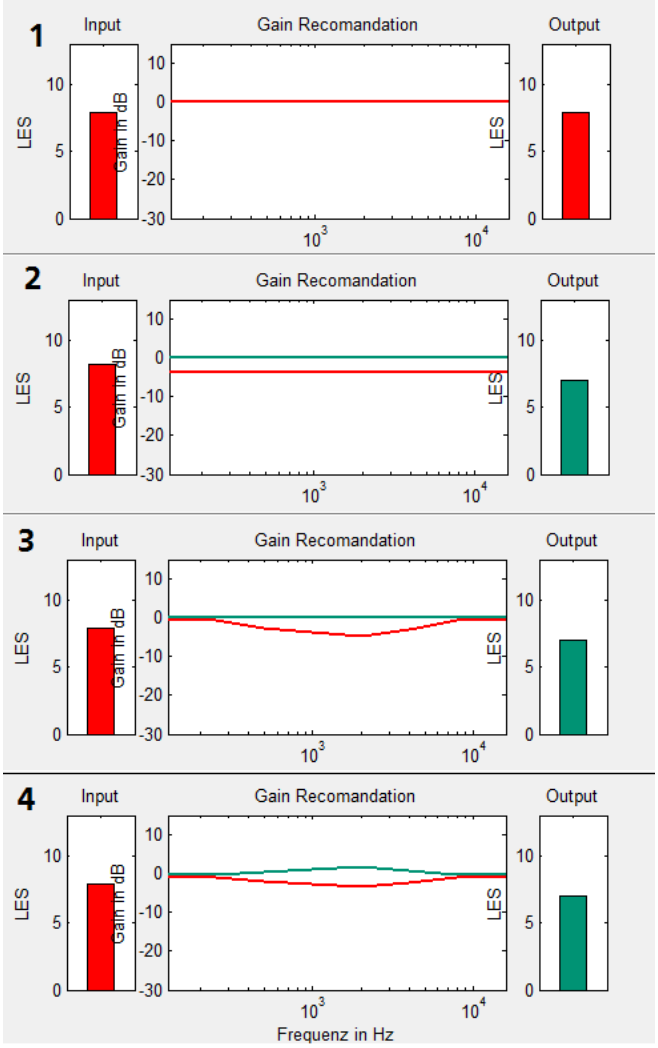

Abbildung 1: Kennlinien der SI4B-kontrollierte Algorithmen zur Verringerung der Höranstrengung in der Momentaufnahme - durch den Einsatz der Algorithmen können nicht akzeptable Höranstrengungswerte (Input: rot) in den „grünen Bereich“ (Output: grün) verschoben werden: unprocessed (1) mit schlechter Sprachverständlichkeit, breitbandige SNR-Verbesserung (2), SII-gainedNoise (3) und SII-gainedSpeech&Noise (4).

Beispielhaft ausgearbeitet wurden drei Verfahren: Die breitbandige SNR-Verbesserung (**Abbildung 1**, 2) zu Gunsten der aktuellen Sprachinformation. Die Absenkung des Hintergrundsignals insbesondere in Frequenzbereichen, die für Sprache besonders relevant sind (**Abbildung 1**, 3) und die Kombination aus Absenkung des Hintergrundsignals mit Anhebung des Sprachsignals (**Abbildung 1**, 4), jeweils in den für die aktuellen Sprachinformationen besonders relevanten Frequenzbereichen.

Als Basis der Algorithmen dienen die instantanen Zwischenergebnisse des SI4B-Sprachverständlichkeit-Modells (vgl. Abb.2).

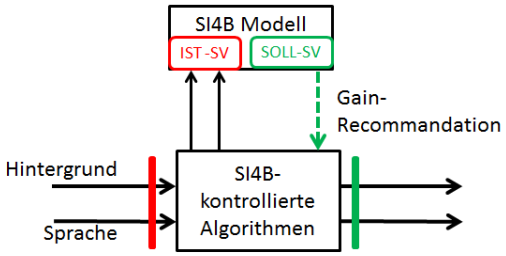

Abbildung 1: Die beabsichtigte Mischung wird vom Modell bzgl. ihrer Sprachverständlichkeit bewertet. Um eine gesetzte Ziel-Höranstrengung zu gewährleisten, berechnet das Modell die entsprechende frequenzabhängige Verstärkung für die SI4B-kontrollierten Algorithmen zur Verbesserung der Sprachverständlichkeit.

Die Verbesserungsalgorithmen reagieren mit einer Verzögerung von acht Blöcken (Blockgröße 2048 Samples) auf Veränderungen der modellierten instantanen Höranstrengung; das Ausmaß der Absenkung bzw. Anhebung durch die Verbesserungsalgorithmen ist dabei zu jeder Zeit abhängig von der vom Modell errechneten Instantan-Höranstrengung; in Relation zur modellierten „Ziel-Höranstrengung“ der Mischung – welche bei Schwerhörenden anders gesetzt ist als bei Normalhörenden. Den frequenzabhängigen Algorithmen liegt eine Filterbank mit Mittenfrequenzen bei [125, 250, 500, 1000, 2000, 4000, 8000, 16000] Hz zugrunde.

Experiment

Methode: Bei Audiosignalen unterschiedlicher Audioqualität liefert die MUSHRA–Methode (Multi Stimulus test with Hidden Reference and Anchor) [2] auch mit Probanden zuverlässige Ergebnisse, welche kein geschultes Ohr besitzen. Bei der von der ITU empfohlenen Variante findet eine Qualitätsskala von 0 - 100 (mangelhaft bis ausgezeichnet) Verwendung. Neben den parallel zu bewertenden Testsignalen und dem offensichtlichen Referenzsignal befindet sich in jedem Bewertungsdurchgang auch ein verstecktes Referenzsignal sowie ein Anker-Signal. Unter der Annahme, dass das versteckte Referenzsignal eine perfekte Qualität aufweist, sollte also auch mindestens eine Bewertung mit 100 Punkten abgegeben werden. Die Testperson hat die Möglichkeit, zwischen allen Signalen in beliebiger Reihenfolge hin und her zuschalten und die Hörproben zu vergleichen. Um eine Verfälschung der Ergebnisse durch Ermüdungserscheinungen der Probanden zu vermeiden, sollte die Anzahl der durchzuführenden Hörbeispiele überschaubar bleiben und die Dauer der verwendeten Testsignale 20s nicht überschreiten.

Für die hier vorgestellte Studie wurde der Original-MUSHRA Test leicht adaptiert (vgl. Abb. 3). Anstelle der Qualitätsskala von 0 bis 100 wurde eine Höranstrengungsskala von 1 bis 13 verwandt. Für jeden Vergleichsdurchgang wurde jeweils ein Hörbeispiel unterschiedlich aufbereitet: Als Testsignale wurden ein für hohen Höranstrengung (vorab geschätzter Höranstrengung-Skalenwert zwischen 8 und 12) gemischtes Signal angeboten (Original) und jeweils drei mit den vorgestellten Algorithmen zur Verbesserung der Sprachverständlichkeit prozessierte Varianten dieser „Originalmischung“. Als Referenz wurden Mischungsverhältnisse mit kaum hörbarem Hintergrund verwendet, als Anker eine extrem schlecht verständliche Mischung des Hörbeispiels.

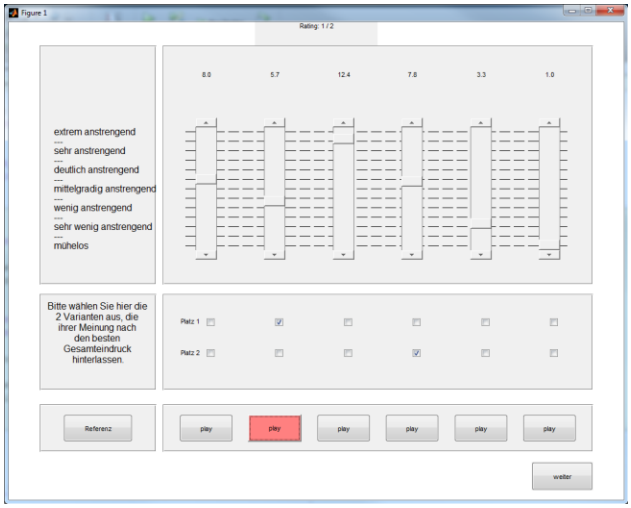

Abbildung 3: Nutzerschnittstelle des Höranstrengungstests mit Höranstrengungsskala von „mühelos“ bis „extrem anstrengend“.

Probanden: Insgesamt 20 normalhörende Versuchspersonen (Alter: 20-30 Jahre, 10w/10m; Median = 25Jahre) und 13 schwerhörende Probanden (Alter: 23 - 79 Jahre, 9w/4m; Median = 67 Jahre) nahmen an den Hörversuchen teil. Von allen Versuchspersonen wurde ein Audiogramm gemessen. Der durchschnittliche Pure Tone Average (PTA, gemittelter Hörverlust bei den Frequenzen 500, 1000, 2000 und 4000 Hz) der normalhörenden Kohorte lag bei etwa 2,5dB HL, der durchschnittliche PTA der schwerhörenden Gruppe bei 45dB HL.

Testmaterial: Als Testmaterial dienten 30 Clips von etwa 12s Dauer. Die Clips wurden aus Original-Fernsehproduktionen erstellt. Die ausgesuchten Clips bestanden immer aus Sprachsequenzen gemischt mit Atmo. Die Mischung der Clips - also das Verhältnis von Sprache zu Hintergrund - wurde für die Versuchsreihe manipuliert: Da die Testreihe die Verringerung der Höranstrengung durch den Gebrauch der Algorithmen belegen sollte, wurden die Clips zunächst in ein Mischungsverhältnis gebracht, welches zu erschwerten Abhörbedingungen führen sollte (vgl. Abb. 4). Bei den manipulierten, mithilfe des Modells bewerteten Mischungen wurden Höranstrengungs-Skalenwerte (LES) zwischen 8 und 12 angestrebt, also zwischen fast „deutlich anstrengend“ bis fast „extrem anstrengend“, um genügend Spielraum für etwaige Verringerungen der Höranstrengung bereit zu halten und die Algorithmen auch zweckdienlich einsetzen zu können. Die Lautheit der Clips entsprach den Empfehlungen der EBU zur Lautheitsnormalisierung **[3]**.
Als „Zielverständlichkeit“ wurde für Normalhörende als maximaler Wert der modellierte Höranstrengungs-Skalenwert von LES=7, für Schwerhörende von LES=5 festgelegt. Die Algorithmen sind so ausgelegt, dass sie - bei Anwesenheit von Sprache – alle die gleiche SNR-Verbesserung zur Folge haben. Die SNRs wurden aus der

Differenz der Lautheit der Sprache und der Lautheit der Hintergründe berechnet und entsprechend der anvisierten „Ziel-Verständlichkeit“ adaptiert (Sprache-zu-Hintergrund-Verhältnisse in dB LU) [3]. Die Verteilung der SNRs der einzelnen Clips zeigt das Histogramm in Abbildung 4.

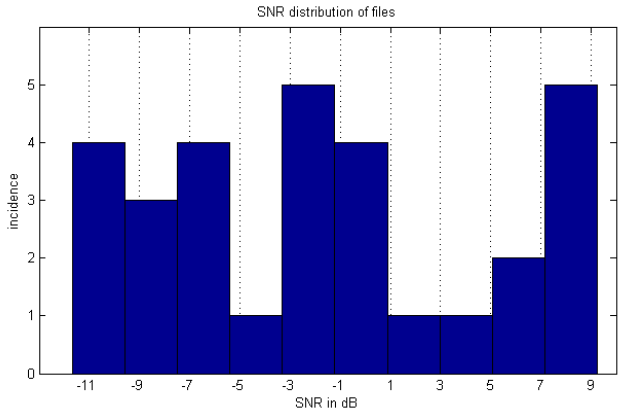

Abbildung 4: Histogramm: Sprache-zu-Hintergrund-Verhältnisse in dB LU für 30 unterschiedlich manipulierte TV-Clips.

Versuchskonditionen: Für jeden der anvisierten Höranstrengungswerte zwischen LES = 8 und LES = 12 wurden jeweils sechs Original Fernseh-Clips – in ihrem SNR entsprechend adaptiert – angeboten. Jeder dieser Clips wurde innerhalb eines MUSHRA-Durchgangs in unterschiedlichen Varianten angeboten:

- Versteckte Referenz: Clip mit sehr gutem SNR, leicht verständlich
- Versteckter Anker: Clip mit extrem schlechtem SNR, schwer verständlich
- Original-Clip: adaptierter Clip, gemischt entsprechend der Ziel-Höranstrengung zwischen LES = 8 und LES = 12
- Processing 1 (BBG): Hintergrund wird breitbandig abgesenkt; der Grad der instantanen Absenkung des Hintergrundes wird vom SI4B-Modell bezüglich der gewünschten Ziel-Höranstrengung kontrolliert.
- Processing 2 (SIIgainedNoise): Hintergrund wird in den für die aktuellen Sprachinformationen besonders relevanten Frequenzbändern abgesenkt. Der Grad der instantanen Absenkung wird vom SI4B-Modell bezüglich der gewünschten Ziel-Höranstrengung kontrolliert.
- Processing 3 (SIIgainedSpeech&Noise): Hintergrund wird in den für die aktuellen Sprachinformationen besonders relevanten Frequenzbändern abgesenkt, Sprache in den für die aktuellen Sprachinformationen besonders relevanten Frequenzbändern verstärkt. Der Grad der Adaption des Vorder- und des Hintergrundes wird vom SI4B-Modell bezüglich der gewünschten Ziel-Höranstrengung kontrolliert.

Bei 30 Clips und sechs Varianten ergeben sich 180 Hörproben, die bzgl. der Höranstrengung verglichen und bewertet wurden. Die Reihenfolge der Darbietungen innerhalb einer Sitzung war für jede/n Probandin/en randomisiert.

Durchführung: Der Hörversuch bestand aus einem Termin von etwa anderthalb bis zwei Stunden Dauer. Zu Beginn des Termins wurde das Audiogramm gemessen. Das eigentliche Experiment fand in einer Hörkabine statt; die Testsignale wurden über Kopfhörer dargeboten. Die Probanden konnten zunächst mithilfe zweier Beispiel-Hörproben eine für sie angenehme Abhörlautstärke einstellen. Für die Bewertung der Höranstrengung wurde die Wiedergabelautstäke individuell auf diesen „angenehmen“ Pegel justiert.
Zwei MUSHRA-Bewertungsdurchgänge mit Signalen, die nicht in die Auswertung einflossen, dienten als Trainingsphase. Die Probanden wurden instruiert, sich zunächst die Referenz anzuhören und anschließend die Auswahl der unterschiedlich prozessierten Hörbeispiel-Varianten bezüglich der empfundenen Höranstrengung zu vergleichen und zu bewerten. Zusätzlich wurden die Probanden aufgefordert, ihre Präferenz bezüglich der angebotenen Signale zu benennen. ***„Bitte wählen Sie die zwei Varianten, die Ihrer Meinung nach den besten Gesamteindruck hinterlassen.“***

Ergebnisse

Individuell bevorzugte Abhörpegel: Die Abhörpegel (Interquartilbereiche) liegen für

- normalhörende Probanden zwischen 56.5 und 64.5 dB SPL (median = **62.5 dB SPL**)
- schwerhörende Probanden zwischen 69 und 75.5 dB SPL (median = **70.5 dB SPL**)

Bewertung der Höranstrengung: Die Probanden mussten 30 verschiedene Clips in jeweils sechs verschiedenen Mischungs- und Verarbeitungs-Varianten bezüglich der Höranstrengung beurteilen: Das sehr gut verständliche Referenz-Signal(0), das für eine bestimmte Ziel-Höranstrengung gemischte Signal(1), die mit den SI4B-kontrollierten Sprachverständlichkeitsverbesserungs-Algorithmen verarbeiteten Signale(2,3,4) und das extrem schlecht verständliche Anker-Signal(5).

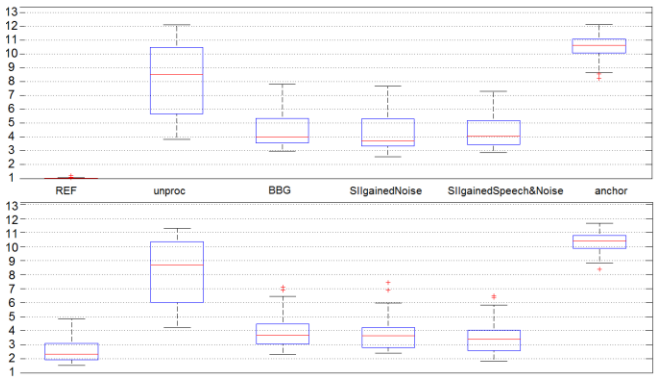

Abbildung 5: Boxplots der über alle Hörproben gemittelten Höranstrengungsbewertungen (y-Achse) für normalhörende (oben) und schwerhörende (unten) Probanden bzgl. der verschiedenen Verarbeitungs-Konditionen (x-Achse). Die roten Linien markieren die Mediane der Bewertungen als LES. Die mittlere vom Modell prognostizierte Höranstrengung liegt bei LES=10, die tatsächliche von Menschen bewertete mittlere Höranstrengung um LES=8,5 (unverarbeitet). Die Bewertungen für Referenz und Anker platzieren sich gut an den Extrema.

Abbildung 5 zeigt die über alle Hörproben gemittelten Probandenbewertungen für normalhörende (oben) und schwerhörende (unten) Probanden bzgl. der verschiedenen Verarbeitungskonditionen. Die 30 verschiedenen unverarbeiteten Clips setzen sich aus jeweils sechs Hörproben, gemischt für die Ziel-Höranstrengungen LES=[8,9,10,11,12], zusammen. Daraus ergibt sich eine mittlere, vom Modell prognostizierte Höranstrengung von LES=10. Die Probanden bewerten die Höranstrengung der Hörproben im Mittel geringer; die durchschnittliche Bewertung liegt bei etwa LES=8,5 (Median) für die für die Ziel-Höranstrengung gemischten unverarbeiteten Signale(1). Die Bewertungen für Referenz(0) und Anker(5) platzieren sich gut an den äußeren Rändern der Skala: Normalhörende bewerten die Höranstrengung der sehr gut verständlichen Referenz im Mittel mit LES=1, Schwerhörende im Mittel mit LES=2.4, bei einem Interquartilbereich von etwa 2 bis 3. Die Bewertung des schlecht verständlichen Anker-Signals(5) erfolgt in beiden Gruppen mit im Mittel LES=10,5, (Interquartilbereich = [10, 11]) Die durch die Verarbeitung mit den SI4B-Modell-kontrollierten-Algorithmen (2,3,4) erreichte verringerte Höranstrengung liegt im Mittel bei LES=3,5 für normalhörende Probanden und LES=4 für schwerhörende Probanden – die Interquartilbereiche verteilen sich auf etwa 1,5 Skalenwerte, was zeigt, dass die Mischungen bzgl. der Ziel-Höranstrengungen von den Algorithmen gut adaptiert wurden.

Interindividuelle Unterschiede: Die Bewertung der Höranstrengung zeigt sich in beiden Gruppen sehr individuell. Die interindividuellen Unterschiede bei der Bewertungen sind deutlich: Bei den normalhörenden Probanden liegen die Bewertungen des unprozessierten Signals in einem Bereich zwischen etwa 1,5 und 4 LE-Distanzpunkten. Die Gruppe der schwerhörenden Probanden antwortet einheitlicher, ihre Bewertungen liegen in einem Bereich zwischen 1,5 und 3 LE-Distanzpunkten.

Präferenz der Verarbeitung: Die Probanden wurden aufgefordert, ihre Präferenz bezüglich der angebotenen Signale zu benennen. Die Histogramme in **Abb. 6** zeigen die unterschiedlichen Präferenzen der Probanden. Der Referenzmix mit sehr guter Sprachverständlichkeit kommt sowohl bei normalhörenden (oben) als auch bei den schwerhörenden (unten) Probanden sehr gut an: Die schwerhörenden Probanden setzten eine eindeutige Präferenz auf den leicht verständlichen Referenzmix, die normalhörenden priorisieren die mit den SI4B-kontrollierten Algorithmen prozessierten Hörproben und den Referenzmix(0) annähernd gleich. Anker(5) und das unverarbeitete Signal(1) - gemischt für hohe Höranstrengung - wurden erwartungsgemäß nicht bevorzugt.

Diskussion und Fazit

Auch wenn das SI4B-Höranstrengungsmodell die Höranstrengung von Mischsequenzen mit einer mittleren Abweichung von der menschlichen Bewertung von etwa 1-2 LES-Einheiten schon gut voraussagt, ist eine weitere Verbesserung der Vorhersagegüte durch weiteres Training absehbar. Die SI4B-Modell-kontrollierten Algorithmen, die das Mischungsverhältnis der Film- oder Fernsehmischung nur in Anwesenheit von Sprache manipulieren, lassen die Atmosphären eines Films und die künstlerischen Freiheiten

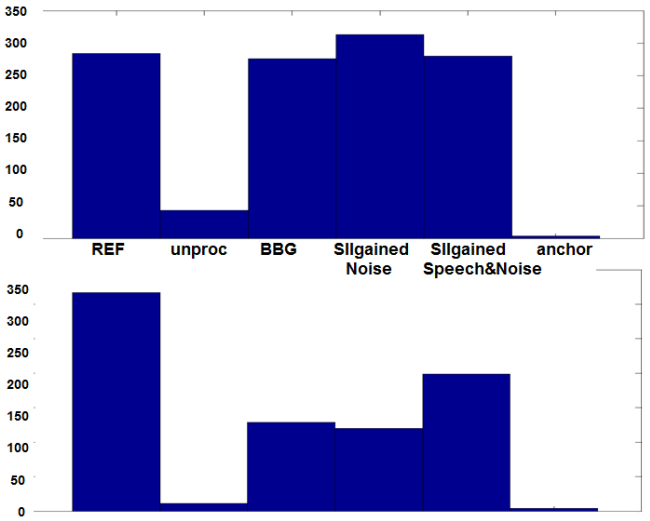

Abbildung 6: Histogramme der Probandenreferenzen für normalhörende (oben) und schwerhörende (unten) Probanden: Normalhörende Probanden bevorzugen neben dem leicht verständlichen Referenzmix die mit den SI4B-kontrollierten Verbesserungsalgorithmen prozessierten Hörbeispiele. Die Höhe der Balken verdeutlicht die Anzahl der Nennungen.

des Sounddesigns nahezu unbehelligt. Die Höranstrengung lässt sich mit Hilfe der vorgestellten Algorithmen sowohl für Normal- als auch für Schwerhörende um 4 bis 5 LES-Einheiten verringern, ohne die Atmosphären des Films zu beeinträchtigen. Sowohl normalhörende als auch schwerhörende Probanden nehmen die Klangveränderung des Signals zu Gunsten einer verringerten Höranstrengung gut an. Als Führungsgröße eines Adaptions-Algorithmus zeigt sich die modellbasierte Schätzung der Höranstrengung als hervorragend geeignet.

DANKSAGUNG

Die Studien wurden im Rahmen des Kooperationsprojekt „Objektive Analyse, Visualisierung und Korrektur von Sprachverständlichkeit in Broadcastanwendungen für Normal- und Schwerhörende/ Speech Intelligibility for Broadcast“ - kurz: SI4B durchgeführt, und gefördert durch das Bundesministerium für Wirtschaft und Energie [Förderkennzeichen ZF4072002SS5].

Literatur

[1] Rennies, J. et al., Höranstrengung von TV-Mischungen in Abhängigkeit von charakteristischen Hintergrundsignalen, DAGA 2017.

[2] Rec. ITU-R BS.1534-1 RECOMMENDATION ITU-R BS.1534-1 Method for the subjective assessment of intermediate quality level of coding systems.

[3] EBU Tech Doc 3341 ‘Loudness Metering: ‘EBU Mode’ metering to supplement loudness normalisation in accordance with EBU R 128.