Erfassung der Höranstrengung fertiger TV-Mischungen

Rainer Huber, Hannah Baumgartner, Christian Rollwage, Stefan Goetze, Jan Rennies-Hochmuth

Fraunhofer IDMT, Hör-, Sprach- und Audiotechnologie, 26129 Oldenburg
E-Mail: Vorname.Nachname@idmt.fraunhofer.de

Einleitung

Die Überwachung der Sprachverständlichkeit und Höranstrengung von TV-Tonmaterial ist notwendig zur Qualitätssicherung von TV-Produktionen. Die Sicherstellung einer (nahezu) 100%igen Sprachverständlichkeit reicht dabei nicht aus, da auch bei solch hohen Sprachverständlichkeiten die zum Verstehen notwendige Höranstrengung auf Dauer unakzeptabel hoch sein kann. Die Höranstrengung stellt daher eine für die Qualitätssicherung sensitivere und relevantere Größe dar als die Sprachverständlichkeit. Ihre Bewertung durch einzelne Personen ist sehr subjektiv, d.h. sie kann individuell verschieden sein. Formale Hörtests mit vielen Testhörern auf der anderen Seite sind i.A. zu aufwändig. Technische Verfahren zur objektiven Bewertung bzw. Vorhersage der empfundenen Höranstrengung wären daher ein wertvolles Hilfsmittel bei der Qualitätssicherung von TV-Produktionen, aber auch für andere Anwendungen wie z.B. der Evaluation von Störgeräuschreduktions-Algorithmen. Grundsätzlich bieten sich hierfür referenzbasierte („double-ended“) und referenzfreie („single-ended“) Methoden an. Referenzbasierte Methoden vergleichen das zu bewertende Testsignal (z.B. Sprache mit Hintergrundgeräusch) mit dem ungestörten Sprachsignal als Referenz. Dieser Ansatz wird z.B. bei der instrumentellen Sprachqualitätsbewertung durch die ITU-T-Standardmethode POLQA [1] verfolgt. Ein ähnlicher Ansatz besteht in dem Vergleich von Nutz- und Störsignal(spektren), wie er z.B. vom Speech-Intelligibility-Index (SII) [2] angewandt wird. Ein Nachteil dieser genannten Ansätze ist, dass das reine, ungestörte Sprachsignal vorliegen muss bzw. Nutz- und Störsignal getrennt, was nicht immer der Fall ist, wie z.B. bei fertigen TV-Mischungen. Selbst wenn, wie beim Mischprozess in der TV-Produktion, Sprach- und Hintergrund-Tonspuren getrennt vorliegen, kann es sein, dass die Sprachspur bereits geringgradige Störungen wie leise Hintergrundgeräusche, Hall, schlechte Artikulation und/oder Verzerrungen enthält, die die Höranstrengung beeinflussen. Solche Störungen des Referenzsignals würden von referenzbasierten Verfahren nicht erfasst werden, da das Referenzsignal, so wie es ist, Prinzip-bedingt die optimale Qualität bzw. minimale Höranstrengung definiert.

Dieser Beitrag beschäftigt sich daher mit der Anwendung eines referenzfreien Verfahrens zur Vorhersage von Höranstrengung auf Audiosignale fertiger TV-Mischungen. Das Verfahren selbst wurde erstmalig in [3] vorgestellt. Auch seine Anwendung auf Audiosignale von TV-Mischungen wurde bereits in [4] beschrieben. Seitdem wurde das Verfahren jedoch um eine automatische Sprachaktivitätserkennung erweitert. Die damit erzielten verbesserten Ergebnisse werden in diesem Beitrag präsentiert.

Methode

Grundsätzlicher Ansatz des Verfahrens ist die Verwendung eines Teils eines automatischen Spracherkennungssystems, welches wiederum auf einem tiefen neuronalen Netz basiert. Störungen der Sprache wie z.B. Verzerrungen oder Hintergrundgeräusche führen zu einer erhöhten Erkennungsunsicherheit des Spracherkennungssystems, ähnlich wie bei der menschlichen Sprachwahrnehmung. Diese Unsicherheit lässt sich im Erkennungssystem wie folgt ablesen und quantifizieren: Das tiefe neuronale Netz erzeugt sogen. Phonem-posterior-Wahrscheinlichkeiten („Posteriorgramme“). Ein Posteriogramm stellt den zeitlichen Verlauf der Wahrscheinlichkeit für die Aktivität einzelner Phoneme dar (s. Abb. 1). Störungen der Sprache führen zu „Verschmierungen“ der Posteriorgramme (s. Abb. 1, unten). Der Verschmierungsgrad wird durch ein mathematisches Maß quantifiziert. Dieses Maß dient als Prädiktor für die Höranstrengung. Die Generierung der Posteriorgramme und das Maß zur Quantifizierung des Verschmierungsgrades werden im Folgenden beschrieben.

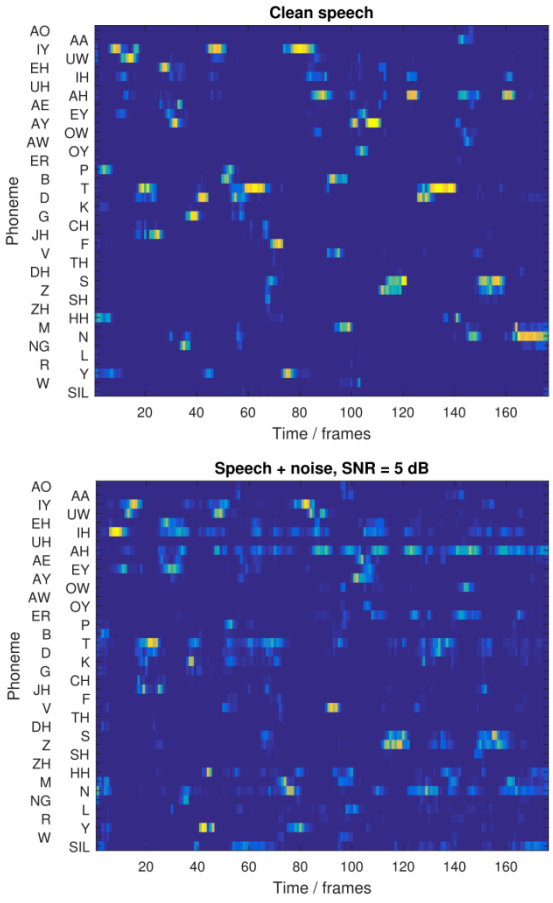

Abbildung 1: Posteriorgramme von reiner Sprache (oben) und derselben Sprachäußerung mit zusätzlichem Rauschen (SNR = 5 dB, unten).

Posteriorgramm-Generierung

Zur Generierung der Posteriorgramme wurde dasselbe automatische Spracherkennungssystem wie in [5] verwendet, das daher hier nur kurz beschrieben werden soll (für Details siehe [5]):

Als akustische Merkmale dienen die 10ms-Kurzzeit-Energien einer 40-kanaligen Mel-Filterbank. Es werden jeweils -15…+15 10ms-Zeitblöcke (= 310 ms) zusammengefasst und einem tiefen *Time-Delay Neural Network* (TDNN) als Merkmalsvektor übergeben. Dieses tiefe TDNN verfügt über sieben verdeckte Schichten mit jeweils 700 *rectified linear units.* Die Ausgabeschicht besteht aus 6448 Neuronen, d.h. eins je Triphon (ein Triphon ist eine Sequenz aus drei Phonemen). Trainiert wurde das Netz mit einer eigenen Sprachdatenbasis, bestehend aus 1000 Stunden ungestörter Sprache. Diese Datenbasis wurde durch Mischung mit verschiedenen Störgeräuschen bei unterschiedlichen SNR auf einen Umfang von ca. 8000 Stunden erweitert.

Posteriorgramm-Maß

Von der Ausgabe des tiefen TDNN, d.h. dem Posteriorgramm, wird die „mean temporal distance“ oder kurz „*M*-Measure“ nach Hermansky et al. [6] berechnet. Das *M*-Measure berechnet den durchschnittlichen mathematischen Abstand zwischen zwei Vektoren von Phonem-Posteriors $p_{t-\Delta t}$ und p_t (d.h. zwei Spalten des Posteriorgramms) mit einem zeitlichen Abstand Δt:

$$M(\Delta t) = \frac{1}{T-\Delta t}\sum_{t=\Delta t}^{T} D(p_{t-\Delta t}, p_t), \qquad (1)$$

T ist dabei die zeitliche Länge des analysierten Posteriorgramms (welche gleich der Länge der analysierten Audiodatei ist, d.h. ca. 10s in dieser Studie). D ist die symmetrische Kullback-Leibler-Divergenz zwischen zwei Vektoren x und y mit den Komponenten $x(i)$ und $y(i)$:

$$D(x,y) = \sum_{i=1}^{N} x(i)\log\left(\frac{x(i)}{y(i)}\right) + \sum_{i=1}^{N} y(i)\log\left(\frac{y(i)}{x(i)}\right) \qquad (2)$$

In dieser Studie ist N gleich der Dimensionalität der TDNN-Ausgabeschicht (6448) und M wurde berechnet für Δt =350 bis 800 ms (in 50ms-Schritten) und anschließend gemittelt. Dies ergibt den finalen Höranstrengungs-Prädiktor $\overline{M}$.

Sprachaktivitäts-Erkennung

Die Posteriorgramm-Berechnung erfolgt nur für Abschnitte des Audiosignals, in denen Sprachaktivität erkannt wird. Die verwendete automatische Sprachaktivitätserkennung basiert auf einem tiefen neuronalen Netz, das auch mit TV-Audiosignalen trainiert wurde [7]. Als Merkmalsvektoren am Eingang des neuronalen Netzes werden Mel-Frequenz-Cepstrum-Koeffizienten (*Mel-Frequency Cepstral Coefficients – MFCC*s) verwendet. (Für Details zur automatischen Sprachaktivitätserkennung siehe [7].)

Höranstrengungsdaten – Datensatz I

Stimuli

Es standen getrennte Tonspuren (Sprache/Hintergrund) von verschiedenen deutschen TV-Produktionen zur Verfügung. Aus diesem Material wurden 192 Audio-Clips mit einer mittleren Länge von ca. 10s herausgeschnitten und mit variierenden (z.T. auch unrealistischen) Mischungsverhältnissen gemischt, um einen möglichst großen Bereich der erwarteten Höranstrengung abzudecken.

Probanden und Bewertungsprozedur

20 normalhörende Probanden im Alter von 20-30 Jahren (Median = 25 Jahre), 10 davon männlich, 10 weiblich, nahmen an der Studie teil. Sie bewerteten die empfundene Höranstrengung der 192 Audio-Clips auf der in Abb. 2 gezeigten Skala auf einem Touch-Screen. Die Frage, die an die Probanden gestellt wurde, lautete: „Wie anstrengend ist es für Sie, die Sprache zu verstehen?“ Die rein akustische Darbietung erfolgte über Kopfhörer (Sennheiser HD 650) in einer Hörkabine. Die gewählten Höranstrengungskategorien wurden auf einen Zahlenwert aus dem Intervall [1, 13] abgebildet. (1 = „mühelos“ … 13 = „extrem anstrengend“) (vgl. Abb. 2)

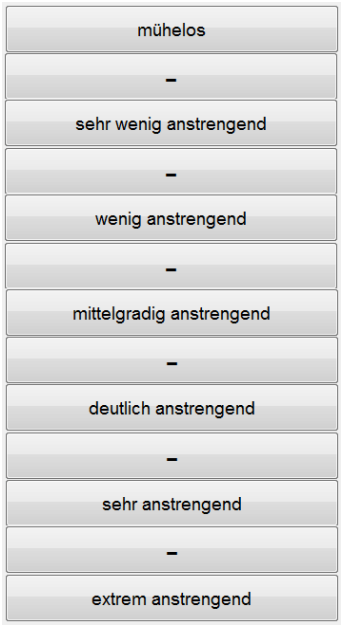

Abbildung 2: Bewertungsskala für Höranstrengung nach [8]

Höranstrengungsdaten – Datensatz II

Stimuli

Es standen wiederum getrennte Tonspuren (Sprache/ Hintergrund) von weiteren verschiedenen TV-Produktionen zur Verfügung. Aus diesem Material wurden 210 Audio-Clips mit einer mittleren Länge von ca. 10s herausgeschnitten und mit stark variierenden Mischungsverhältnissen gemischt.

Probanden und Bewertungsprozedur

20 normalhörende Probanden im Alter von 20-29 Jahren (Median = 23 Jahre), 9 davon männlich, 11 weiblich,

nahmen an der Studie teil. Die Bewertungsprozedur war dieselbe wie beim Datensatz I (s.o.).

Ergebnis

Subjektiv gemessene, über die Probanden gemittelte Höranstrengung-Bewertungen (LE-MOS - „Listening Effort-Mean Opionion Scores“) werden den entsprechenden Werten des Höranstrengungs-Prädiktors $\bar{M}$ in Abb. 3 in Form eines Scatter-Plots für beide Datensätze gegenüber gestellt. Für die vereinigte Datenmenge ergibt sich ein linearer Zusammenhang zwischen subjektiven Bewertungen und objektivem Maß $\bar{M}$ mit hoher Korrelation (r =0.92).

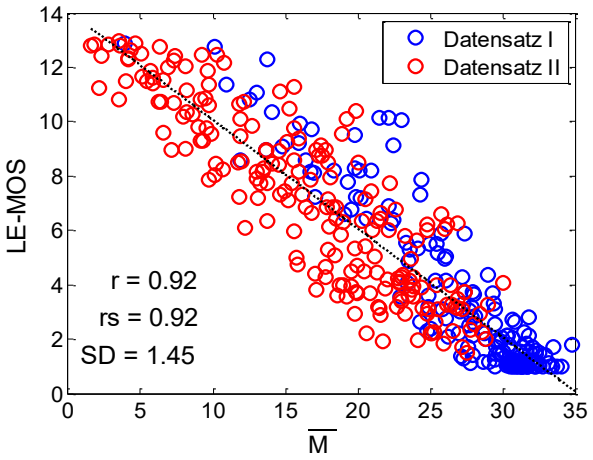

Abbildung 3: Ergebnis der Höranstrengungsvorhersage für beide Datensätze. Über die Probanden gemittelte subjektive Bewertungen („Listening Effort Mean Opion Scores“ – LE-MOS) auf der Ordinate sind über den entsprechenden Werten des instrumentellen Höranstrengungs-Prädiktors $\bar{M}$ auf der Abszisse aufgetragen. Blaue Kreise: Datensatz I. Rote Kreise: Datensatz II. Schwarze, gepunktete Gerade: Ausgleichsgerade nach linearer Regression. r: Pearson-Korrelationskoeffizient. rs: Spearman-Rangkorrelations-koeffizient. SD: Standardabweichung von einer perfekten Vorhersage in LE-MOS-Skaleneinheiten

Bewertung von Audio-Clips mit Sprache im Hintergrund

Datensatz I umfasste ursprünglich 42 weitere Audio-Clips, die Sprache im Hintergrund (zumeist „voice-over-voice“) enthalten. Die Höranstrengung solcher Signale wird vom Prädiktor typischerweise unterschätzt, da das Verfahren nicht zwischen Nutz- und Störsprache unterscheiden kann. Im LE-MOS – $\bar{M}$ – Scatterplot (Abb. 4) zeigt sich dies durch Datenpunkte rechts-oberhalb der übrigen Datenpunkte.

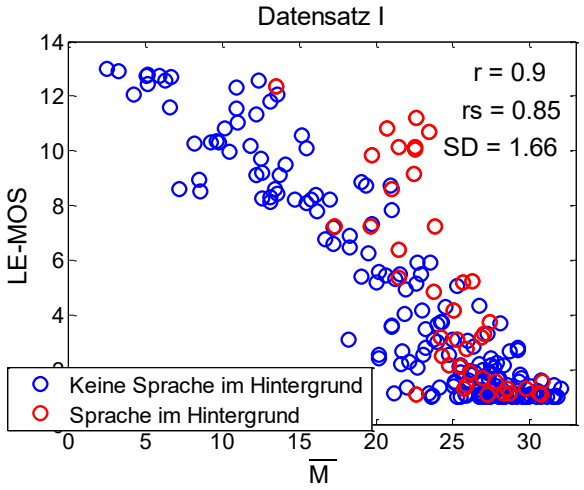

Abbildung 4: Ergebnis der Höranstrengungsvorhersage für den vollständigen Datensatz I inkl. Audio-Clips mit Sprache im Hintergrund (rote Kreise). Sonst wie Abb. 3.

Diskussion

Die Anwendung der referenzfreien instrumentellen Methode zur Vorhersage von Höranstrengung nach [3] bzw. [5] scheint auch für den Einsatz bei Rundfunk-Anwendungen sehr vielversprechend zu sein, insbesondere, da die bisherigen Ergebnisse der Höranstrengungsvorhersage ohne jegliches Training oder Optimierung der Methode auf die vorliegenden Daten erzielt wurden. Weitere Verbesserungen der Ergebnisse sind denkbar, wenn der Spracherkenner auch mit typischen Hintergrund-„Stör“-Signalen, wie sie im Rundfunkmaterial auftreten, trainiert würde.

Eine Verbesserung der Ergebnisse gegenüber [4] konnte durch den Einsatz einer automatischen Sprachaktivitäts-erkennung erzielt werden. Die Korrelation zwischen subjektiven Bewertungen und Vorhersagen stieg dadurch von r = 0.89 auf r = 0.92.

Eine Einschränkung der Methode betrifft die Bewertung von Sprache, die von konkurrierender Sprache im Hintergrund („voice-over-voice“) überlagert wird. Hier unterschätzt die Methode zumeist die empfundene Höranstrengung. Dies ist erklärbar, da die Methode nicht zwischen Nutz- und Stör-sprache unterscheiden kann. Die entsprechenden Posterior-gramme von zwei Sprechern weisen nur geringradige Verschmierungen auf und die Phoneme (bzw. Triphone) werden vom automatischen Spracherkennungssystem weiterhin überwiegend erkannt. Erst bei einer Überlagerung von vielen Sprechern („babble noise“) ist mit einer erhöhten Erkennungsunsicherheit des Systems zu rechnen.

Der bei einzelnen (verständlichen) Störsprechern zusätzlich auftretende Effekt des „informational masking“, der neben der energetischen Maskierung zu einer weiteren Erhöhung der Höranstrengung beiträgt, kann durch den hier verwendeten instrumentellen Ansatz nicht abgebildet werden.

Eine mögliche Abmilderung dieser Einschränkung des Verfahrens könnte durch eine unabhängige Detektion von Sprache im Hintergrund vermittelt werden. In einem solchen

Fall könnte ein Korrekturwert (Malus) zur Höranstrengungsschätzung hinzuaddiert werden. Folgende Klassifizierung der Signale ist hierfür nötig: (1) nur Nutzsprache, (2) Sprache mit nicht-Sprache im Hintergrund, (3) Nutzsprache mit Störsprache im Hintergrund.

Eine weitere Einschränkung der Methode trifft die Tatsache, dass das referenzfreie Höranstrengungsmodell im Moment noch als „Mono"-Modell agiert und so den Einfluss einer räumlichen Verteilung der Quellen außer Acht lässt. Insbesondere mit Blick auf Surround und 3D-Produktionen ist das ein Mangel, den es in Form einer Binauralisierungsstufe noch zu beheben gilt.

Schlussfolgerungen

Die vorgestellte referenzfreie, instrumentelle Methode zur Schätzung der empfundenen Höranstrengung hat sich insgesamt als geeignet zur Anwendung auf Rundfunk-Audiomaterial herausgestellt. Bei der Bewertung von Audioclips mit Sprache im Hintergrund muss berücksichtigt werden, dass die Methode die tatsächliche Höranstrengung tendenziell unterschätzt. Die Methode bietet Verbesserungspotenzial, z.B. durch Trainieren des Spracherkenners mit Sprache mit rundfunktypischen Hintergrundgeräuschen und der Erweiterung durch eine Binauralisierungsstufe, welche der Räumlichkeit verschiedener Produktionsformate gerecht wird.

Danksagung

Diese Studie wurde durchgeführt im Rahmen des vom BMBF geförderten Projekts SITA; FKZ: 01/S17017

Literatur

[1] ITU-T Rec. P.863. Perceptual Objective Listening Quality Assessment. Geneva, Switzerland

[2] ANSI S3.5–1997. American National Standard Methods for the Calculation of the Speech Intelligibility Index. New York: ANSI, 1997

[3] Huber, R., Spille, C., Meyer, B.T.: Single-Ended Prediction of Listening Effort Based on Automatic Speech Recognition. In Proceedings Interspeech 2017, 1168-1172

[4] Huber, R., Baumgartner, H., Moritz, N., Goetze, S.: Automatische Überwachung der Sprachverständlichkeit im Rundfunkmaterial. In Proceedings 30. Tonmeistertagung 2018

[5] Huber, R., Pusch, A., Moritz, N., Rennies, J., Schepker H., Meyer, B.T.: Objective Assessment of a Speech Enhancement Scheme with an Automatic Speech Recognition-Based System. In Proceedings ITG Conference on Speech Communication (2018), 86-90

[6] Hermansky, H., Variani, E., Peddinti, V.: Mean temporal distance: Predicting ASR error from temporal properties of speech signal. In Proceedings ICASSP 2013, 38th IEEE Int. Conf. Acoust. Speech Signal Process. doi: 10.1109/ICASSP.2013.6639105

[7] Moritz, N., Drefs, J., Baumgartner, H., Rennies, J.: Sprachaktivitätserkennung basierend auf Deep Neural Networks für Anwendung in Film und Fernsehen. In Fortschritte der Akustik. DAGA 2016, S.960-963; DEGA, Berlin

[8] Schulte, M., Meis, M., Wagener, K.: Listening Effort and Speech Intelligibility. In Proceedings 8th EFAS Congress / 10th Congress of the German Society of Audiology, 2007.