FORSCHUNG

DIALOG+ UND DIALOGSEPARATION ZUR VERBESSERUNG DER SPRACH-VERSTÄNDLICHKEIT IM RUNDFUNK

THERESA LIEBL, MICHAEL MEIER

Vielen Zuschauern bereitet die Verständlichkeit von Sprache im Fernsehen Probleme. Gerade ältere Menschen können den Dialogen häufig nur schwer folgen. Das IRT arbeitet schon seit einigen Jahren an Lösungen. Ein Ansatz, der dabei verfolgt wird, ist es, den Zuschauern sogenannte *Dialog+*-Mischungen mit einer verbesserten Sprachverständlichkeit anzubieten.

Bei Fernsehinhalten, zu denen Sprache und Hintergrund als Einzelspuren vorliegen, lassen sich *Dialog+*-Mischungen einfach generieren. Bei Archivmaterial, welches nur als fertige Stereo-Mischung vorhanden ist, ist dies schwieriger. Um verständlichere Varianten erzeugen zu können, müssen die Mischungen zuerst in ihre Einzelteile, also Sprache und Hintergrund, zerlegt werden. Ob eine solche Dialogseparation zur Erzeugung von *Dialog+*-Mischungen im Rundfunkumfeld geeignet ist, wurde am IRT in Zusammenarbeit mit den Rundfunkanstalten untersucht.

Dialogseparation

Es existieren bereits verschiedene Verfahren, mit denen eine fertige Stereo-Mischung wieder in ihre Bestandteile Sprache und Hintergrund zerlegt werden kann. Dabei wird die Sprache zwar nicht vollkommen sauber vom Rest der Mischung getrennt, was aber auch nicht nötig ist. Vor dem erneuten Zusammenmischen können die separierten Hintergrundgeräusche im Verhältnis zur Sprache abgesenkt und so die Sprachverständlichkeit verbessert werden. Diese Absenkung ist je nach Ausgangsmaterial und angewandter Separationsmethode jedoch nur in einem bestimmten Bereich möglich, da bei zu starker Absenkung Artefakte, welche bei der Separation entstehen, hörbar werden. Das Prinzip der Erzeugung von *Dialog+*-Mischungen mittels Dialogseparation ist in Abbildung 1 schematisch dargestellt.

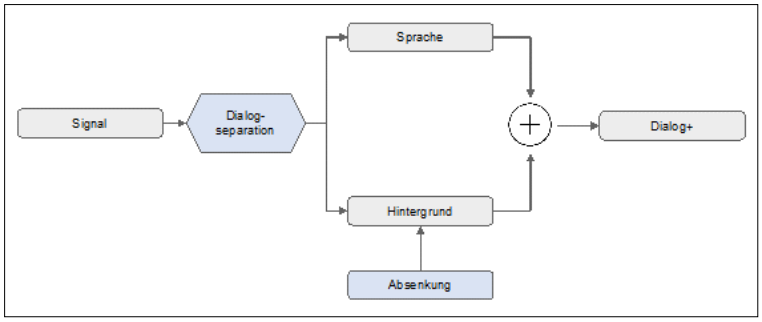

Abbildung 1: Schematische Darstellung der Erzeugung einer Dialog+-Mischung mittels Dialogseparation und Hintergrundabsenkung.

Grundsätzlich gibt es verschiedenste Verfahren, um aus komplexen Summensignalen wieder Einzelteile zu extrahieren; beispielsweise aus den Bereichen der Musikbearbeitung, der Erzeugung von Karaoke-Fassungen oder Speech-To-Text-Anwendungen. Diese basieren häufig auf einer Mischung eher klassischer Audiosignalverarbeitungsansätze, zudem finden neuerdings vermehrt Ansätze auf Basis von Machine Learning und Künstlicher Intelligenz Anwendung.

Hörversuch

Um das Potential und die Anwendbarkeit von *Dialog+* im Rundfunkumfeld näher zu untersuchen, wurde am IRT ein Hörversuch mit Fokus auf den potentiellen Mehrwert für Fernsehzuschauer durchgeführt. Ziel dabei war es, eine generelle Verbesserung der Sprachverständlichkeit durch automatisch erstellte *Dialog+*-Fassungen zu verifizieren. Darüber hinaus wurde untersucht, ob diese Verbesserung auch mittels vorangehender automatischer Dialogseparation erreicht werden kann.

Der Fokus der Untersuchung war ganz auf den potentiellen Mehrwert für den Zuschauer ausgerichtet. Unter der Maßgabe, dass die Separationsverfahren prinzipiell auch kurzfristig in der Praxis eingesetzt werden können, wurden drei Verfahren ausgewählt, welche die Bandbreite von verfügbaren kommerziellen Produkten und neueren Methoden möglichst gut abdecken. Zusätzlich sollte der Umfang auf einfach umsetzbare Abläufe beschränkt werden, auch wenn zu erwarten ist, dass in Zukunft noch verbesserte Methoden entwickelt und marktreif werden.

In einem öffentlichen Aufruf wurden Hörversuchs-Teilnehmer der Altersgruppe 50+ gesucht – der potentiell größten Zielgruppe für einen möglichen Dialog+-Dienst. Vorerfahrung mit Hörversuchen oder analytischem Hören waren bei den insgesamt 38 Teilnehmern explizit nicht gefordert. Der Hörversuch fand am IRT in einer „Wohnzimmerumgebung" statt (siehe Abbildung 2), um eine realistische Hörumgebung und Abhörsituation wie zu Hause nachzustellen. Dabei wurden ähnliche Konditionen beispielsweise in Bezug auf die Raumakustik oder die Wiedergabegeräte abgebildet. Die Audiowiedergabe erfolgte mittels einer qualitativ angemessenen Soundbar im mittleren Preissegment [1] – als Kompromiss zwischen dem meist schlechtesten und unberechenbarsten Wiedergabegerät, dem Flachbildfernseher, und qualitativ hochwertigen Lautsprechern.

Die Probanden sollten verschiedene *Dialog+*-Fassungen im Vergleich zum Original-Fernsehton bewerten. Die Bewertung erfolgte rein nach persönlicher Präferenz für einen potentiellen *Dialog+*-Dienst mittels der sogenannten *Hybrid Hedonic Scale*. Dabei handelt es sich um einen Akzeptanztest aus dem Bereich der affektiven Methoden, der sich bei Probanden ohne Erfahrungen mit Hörversuchen anbietet. Die Bewertungsskala ermöglicht kontinuierliche Angaben von „gefällt mir außergewöhnlich gut" *(like extremely)* über

„weder-noch" *(neither like nor dislike)* bis hin zu „gefällt mir überhaupt nicht" *(dislike extremely)* (siehe Abbildung 3). Bei der Bewertung der *Dialog+*-Fassungen sollten die Probanden auf eine verringerte Höranstrengung und eine bessere Sprachverständlichkeit im Vergleich zum Original achten. Kriterien wie beispielsweise die Gesamtklangqualität oder die Ausgewogenheit der Mischung konnten ebenfalls mit in die Bewertung einfließen.

Insgesamt standen den Probanden im Versuch fünf *Dialog+*-Varianten zum direkten Vergleich mit dem Original-Sendeton zur Verfügung (siehe Abbildung 3). Getestet wurden drei Dialogseparationsverfahren verschiedener Hersteller, sowie die Mischungen aus den Original-Einzelspuren (Stems) und ein vom IRT entwickelter Low-Anchor , also eine bewusst vergleichsweise schlechte Variante zur Einordnung. Für die neu erstellten Mischungen wurden zwei feste Mischungsverhältnisse für alle Varianten gewählt. In Vorversuchen wurde ermittelt, dass Absenkungen des Hintergrundes um –6dB und –9dB im Vergleich zur Originalmischung für die geplanten Untersuchungen am besten geeignet sind. Lediglich bei der Stem-Variante aus den Original-Einzelspuren wurde der Hintergrund lediglich um -6dB abgesenkt, da eine stärkere Absenkung durch die ideale Trennung von Sprache und Hintergrundsignal bei den anderen Varianten mittels Dialogseparation nicht erreichbar und somit nicht vergleichbar gewesen wären.

Als Hilfestellung zur Bewertung sollten die Probanden sich in die Situation versetzen, wie sie einen potentiell neuartigen *Dialog+*-Dienst an ihrem Wiedergabegerät aktivieren. Dies zeigte eine am IRT entwickelte *Dialog+*-Demo (siehe [2] und [3]). Nach Erläuterungen und einer Einführungsphase durch die Versuchsleiter führte jeder Proband die Versuche allein im Raum und eigenständig durch. Pro Bewertungsdurchgang sollten alle fünf *Dialog+*-Varianten mit dem Original-Sendeton für ein Audiosignal und eine Absenkung bewertet werden. Die Reihenfolge war dabei für jeden Probanden zufällig gewählt, genauso wie die Zuordnung der einzelnen *Dialog+*-Varianten zu den anwählbaren Optionen A bis E. Bei der Bewertung konnten die Probanden in der Versuchsoberfläche frei zwischen allen Varianten wechseln.

Für die Untersuchung wurden von den Rundfunkanstalten zur Verfügung gestellte Sendebeiträge aus den Bereichen Dokumentation und Spielfilm ausgewählt, welche auch in Einzelspuren verfügbar waren. Der gewählte Content sollte unterschiedliche Hintergrundgeräusche und Sprachsituationen enthalten, um ein möglichst breites Spektrum realistischer Rundfunksignale abzudecken. Das zugehörige Fernsehbild wurde während des Versuchs über den mittleren Fernseher wiedergegeben (siehe Abbildung 2).

Ergebnisse des Hörversuchs

Die Ergebnisse zeigen den Mehrwert von *Dialog+*-Mischungen für den Zuschauer eindeutig – sowohl für die Stems-Variante, generiert aus den Original-Einzelspuren, als auch für die mittels Dialogseparation erzeugten Varianten. Einen Überblick über die Ergebnisse – für alle Signale, getrennt nach Absenkungen – zeigt die Abbildung 4. Hier ist deutlich zu erkennen, dass alle *Dialog+*-Varianten positiv bewertet wurden, im Gegensatz zum Low Anchor, der erwartungsgemäß tendenziell negative Bewertungen erhielt. Die Stems-Variante schnitt tendenziell am besten ab. Diese klaren Tendenzen bestätigten unsere Annahmen und zeigten auch, dass die Versuchspersonen die Aufgabenstellung verstanden haben und richtig umsetzen konnten.

Eine relativ hohe Streuung der Bewertungen ist bei Hörversuchen mit unerfahrenen Probanden üblich. Hinzu

(Quelle: IRT)

Abbildung 2: Hörversuchsaufbau in der IRT Wohnzimmerumgebung. Videowiedergabe erfolgte über den mittleren Bildschirm, Steuerung des Hörversuchs über das Laptop.

kommen unterschiedliche Erwartungshaltungen bei den Teilnehmern. Während einigen Probanden die *Dialog+*-Variante gut gefiel, war anderen der Unterschied zum Originalsignal noch nicht deutlich genug. Einige wenige Probanden empfanden die präsentierten Absenkungen bereits als zu stark. Diese Erkenntnisse deckten sich mit den Rückmeldungen der Teilnehmer im Anschluss-Gespräch. Außerdem zeigte sich, dass die Bewertung der einzelnen *Dialog+*-Varianten stark von der Qualität des Original-Testsignales abhängt. Testsignale, die bereits im Original und somit auch in den Einzelspuren eine schlechte Audioqualität aufweisen, können durch die *Dialog+*-Variante weiter verschlechtert werden.

Zwischen den beiden Absenkunken -6dB und -9dB zeigten sich keine systematischen Unterschiede oder direkte Abhängigkeiten zu bestimmten Signalcharakteristiken. Allerdings zeigt auch hier die erhöhte Streuung der -9dB Absenkung im Vergleich zur -6dB Variante, dass ein Teil der Versuchspersonen die stärkere Absenkung begrüßt, während ein anderer Teil die daraus resultierenden Artefakte bereits deutlicher wahrnimmt und negativ bewertet.

Mit den Ergebnissen aus diesem Versuchsaufbau lassen sich keine signifikanten Unterschiede zwischen den Dialogseparationsverfahren aufzeigen. Dies ist allerdings auch auf die hier gewählte Methode und Fragestellung zurückzuführen und bedeutet im Umkehrschluss nicht zwingend, dass keine Unterschiede

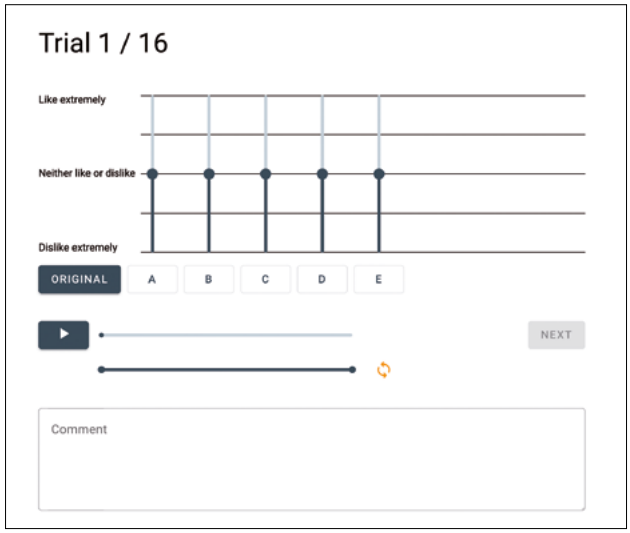

Abbildung 3: Benutzeroberfläche zur Steuerung des Hörversuchs

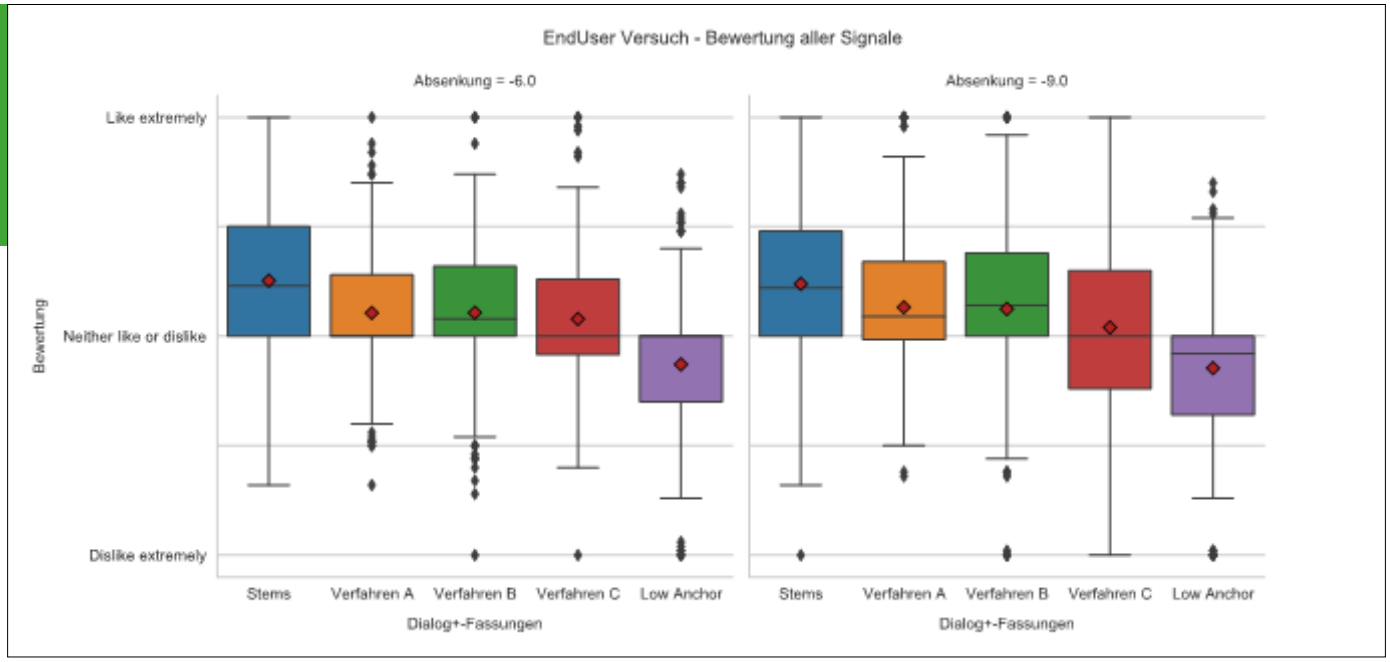

Abbildung 4: Boxplot-Diagramm über die Bewertungen aller Signale, aufgeteilt nach Verfahren und Absenkung. Dargestellt werden Mittelwert als roter Punkt, Median als Linie und Quantile als Box. Whisker zeigen die Extremwerte auf Basis des 1,5-fachen Interquartilsabstands, Bewertungen außerhalb werden als Punkt dargestellt.

vorhanden sind. Für genauere Erkenntnisse wäre eine dedizierte Untersuchung hinsichtlich der Qualität einzelner Verfahren nötig. Die Ergebnisse zeigen aber auch, dass andere Faktoren wie die Art und Stärke der Absenkungen einen deutlich höheren Einfluss auf die wahrgenommene Qualität haben.

Anwendung der Ergebnisse

Die Untersuchung hat gezeigt, dass der Dienst *Dialog+* einen deutlichen Mehrwert bieten kann. Folglich sind nun Überlegungen wichtig, wie *Dialog+* den Zuschauern zur Verfügung gestellt werden kann. Die Anwendungsmöglichkeiten und Ausspielwege für eine einmal erzeugte *Dialog+*-Fassung sind vielfältig. Diese kann beispielsweise „klassisch", als weitere Tonspur in einem DVB Signal übertragen werden, falls ein freier Kanal zur Verfügung steht. Oder mit Hilfe des HbbTV2-Standards kann im Live-Fernsehsignal die Verfügbarkeit von *Dialog+* signalisiert und bei Bedarf als extra Tonspur über eine Internetverbindung geladen und synchron zum Fernsehbild wiedergegeben werden (siehe [2] [3]). Dabei ist es für eine bessere Personalisierung auch möglich, verschiedene Mischungen anzubieten. Ebenfalls ist eine Anwendung im Browser mittels Webplayer [4] verfügbar oder eine Integration in die entsprechenden Apps möglich. Alle Varianten bieten Vor- und Nachteile hinsichtlich Kosten, Zielpublikum und Aufwand für die Umsetzung. In der Praxis ist daher wohl eine Mischung der Möglichkeiten sinnvoll, welche flexibel und individuell auf den konkreten Bedarf und die Zielsetzung abgestimmt sind.

Fazit

Die Untersuchung zeigt, dass durch *Dialog+* grundsätzlich eine Verbesserung der Sprachverständlichkeit im Fernsehen erzielt werden kann. Varianten aus Original-Einzelspuren bieten oft den größeren Mehrwert, aber auch durch Dialogseparation erzeugte Mischungen können zu einem vergleichbaren Ergebnis führen. Um den individuellen Anforderungen und Wünschen der Zuschauer gerecht zu werden, welche sich in den Versuchsergebnissen zeigen, sollten mehrere *Dialog+*-Mischungen angeboten werden. Dafür gibt es mehrere, kurzfristig umsetzbare Lösungen. Ein verbessertes Fernseherlebnis durch *Dialog+*-Mischungen ist folglich bereits jetzt möglich.

Danksagung

Die Autoren bedanken sich ganz herzlich bei Maike Richter für ihre tatkräftige Unterstützung bei der Planung, Durchführung und Auswertung des Hörversuchs.

Bild: IRT

MICHAEL MEIER

ist Ingenieur im Bereich Metadata & Accessibility am Institut für Rundfunktechnik in München und dort schwerpunktmäßig zum Thema Sprachverständlichkeit tätig. www.irt.de

Bild: IRT

THERESA LIEBL

ist Ingenieurin im Bereich Metadata & Accessibility am Institut für Rundfunktechnik in München und dort schwerpunktmäßig zum Thema Sprachverständlichkeit tätig. www.irt.de

Literatur

[1] T. Liebl, S. Wakan, O. Curdt: „Audio quality evaluation of soundbars using the multiple stimulus ideal profile method". Fortschritte der Akustik - DAGA, 2018.

[2] A. Tai, T. Liebl, R. Mies: „Forschungsfeld Barrierefreiheit". FKT, Juni 2020.

[3] IRT Mediathek: „Demo Video: *Dialog+* via HbbTV 2". URL: https://www.youtube.com/watch?v=dbSqn7xtxcQ [Zugriff am 2. Oktober 2020]

[4] IRT Lab: „Alternative audio tracks in ARD Player". URL: https://lab.irt.de/alternative-audio-tracks-in-ard-player/ [Zugriff am 2. Oktober 2020].