Automatische Segmentierung binauraler Raumimpulsantworten für die Modellierung von Sprachverständlichkeit

Omid Kokabi, Fabian Brinkmann, Stefan Weinzierl
kokabi@campus.tu-berlin.de, {fabian.brinkmann, stefan.weinzierl}@tu-berlin.de
Audio Communication Group, Technische Universität Berlin, Einsteinufer 17c, 10587, Berlin

Einleitung

Die beiden zentralen Wahrnehmungsaspekte binauraler Sprachwahrnehmung sind better-ear listening und binaural unmasking. Better-ear listening beschreibt den Umstand, dass bei der Wahrnehmung von Sprache maßgeblich das Ohrsignal ausgewertet wird, das den höheren Informationsgehalt aufweist. In Anwesenheit zusätzlicher, konkurrierender Störquellen – worunter auch einzelne starke Raumreflektionen des eigentlichen Sprachsignals fallen können – beschreibt binaural unmasking die sich ergebende Verbesserung der Verständlichkeit. Diese nimmt zu, je stärker sich die beim Zuhörer beobachteten interauralen Zeit- und Pegelunterschiede (ITD/ ILD) von Sprachquelle und Störquelle unterscheiden.
Beim Hören in geschlossenen Räumen beeinflusst das im Raum vorherrschende reflektierte Schallfeld zusätzlich die Verständlichkeit. Frühe Raumreflektionen, die kurz nach dem Direktschall eintreffen, führen hierbei zu einer Verbesserung der Verständlichkeit [1]. Späte Raumreflektionen bzw. diffuser Nachhall verursachen eine durch Verschmierung der zeitlichen Struktur des Sprachsignals bedingte Verschlechterung der Verständlichkeit.
Erfolgreiche Modelle zur Vorhersage binauraler Sprachverständlichkeit bilden better-ear listening und binaural unmasking mit hoher Genauigkeit nach [2, 3]. Better-ear listening ist hierbei mittels Auswertung des Signal-Rausch Verhältnisses (SNR) beider Ohrsignale implementiert. Binaural unmasking wird nach dem Vorbild der Equalization-Cancellation (EC) Theorie modelliert [4]. Als Eingangssignal für die Vorhersagemodelle dient die den Übertragungspfad von Sprachquelle zu Empfänger beschreibende binaurale Raumimpulsantwort (BRIR) bzw. das beim Empfänger eintreffende binaurale Sprachsignal.

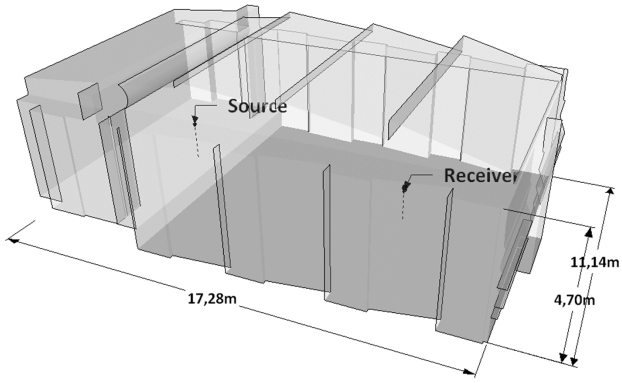

Abbildung 1: 3d Raummodell

Um den Einfluss von diffusem Nachhall in den genannten Prädiktionsmodellen zu berücksichtigen, wurde vorgeschlagen, die BRIR in einen frühen, nützlichen (engl. useful) und einen späten, nachteiligen (engl. detrimental) Anteil auzuteilen, bevor sie in das Prädiktionsmodell eingeht [5]. Mit diesem Useful/Detrimental (U/D) Ansatz konnte der Prädiktionsfehler bei der Vorhersage von Sprachverständlichkeit in geschlossenen Räumen verringert [5], bzw. der störende Einfluss einer einzelnen Raumreflektion auf das Sprachsignal überhaupt erst qualitativ nachgebildet werden [6]. In einer weiteren Arbeit wurde festgestellt, dass sich die U/D-Zeitgrenze auf den Prädiktionsfehler auswirkt, wobei die geringsten Fehler mit einer raumabhängigen U/D-Zeitgrenze ereicht wurden [7]. In der vorliegenden Arbeit wurde der Zusammenhang von Raum, U/D-Zeitgrenze und Prädiktionsfehler anhand eines Vergleichs gemessener und vorhergesagter Sprachrezeptionsschwellen (SRTs) für vier virtuelle, nachhallbehaftete Räume untersucht. Hierbei wurde der mittlere Prädiktionsfehlerkonnte anhand einer Vorhersage der raumabängigen U/D-Zeitgrenze durch raum- und empfängerabhängige Aspekte verringert. Die vorgeschlagene Methodik lässt sich auf beliebige Räume und Quelle-Empfänger-Kombinationen übertragen.

SRT Messung

Die hier betrachteten virtuellen Räume basieren auf einem realen, mittelgroßen Auditorium, skaliert auf ein Raumvolumen von $V = 1000$ m^3. Durch Skalierung der Raumabsorptionseigenschaften wurden vier geometrisch identische Räume mit systematisch variierter Nachhallzeit $T_{20,m}$ = 0,5 s, 1,0 s, 2,0 s und 4,0 s generiert. Pro Raum wurde eine BRIR in RAVEN simuliert [8].

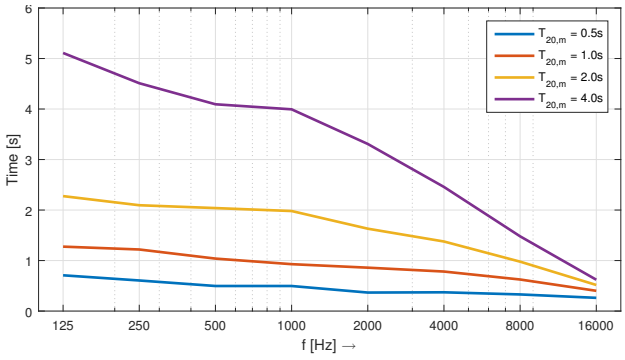

Abbildung 2: Am Empfängerort ausgewertete Nachhallzeitverläufe für die vier Räume

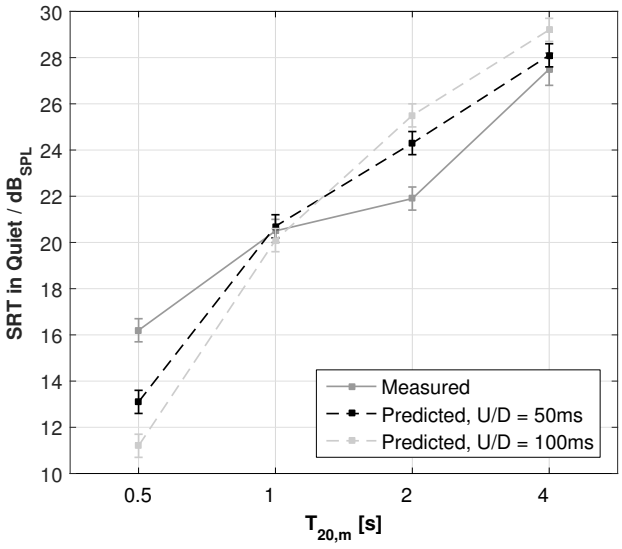

Abbildung 3: Gemessene und vorhergesagte SRTs mit raumunabhängiger U/D-Zeitgrenze

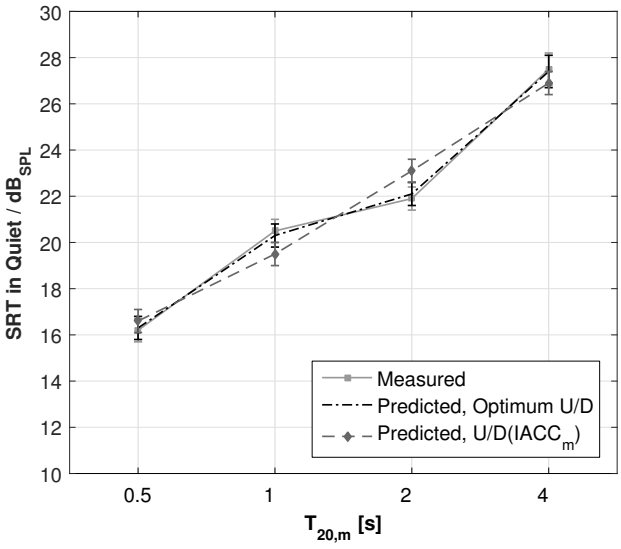

Abbildung 4: Gemessene und vorhergesagte SRTs mit optimaler und vorhergesagter raumabhängiger U/D-Zeitgrenze

Die Schallquelle mit Richtcharakteristik eines Sängers (mittlerer Richtfaktor $Q_m \approx 1,5$) befand sich hierbei im Bereich der Bühne, der binaurale Empfänger im Zuhörerbereich im Abstand von ca. $d = 9$ m, was in etwa dem dreifachen Hallabstand entspricht [9]. Die Raumgeometrie, sowie die Nachhallzeitverläufe an der Empfängerposition zeigen Abbildungen 1 und 2. Für die vier Räume wurden zunächst mittels des Oldenburger Satztests (OLSA) SRTs in Ruhe gemessen [10, 11, 12]. Das ist der Schalldruckpegel in dB$_{SPL}$, der notwendig ist, damit eine Person 50% eines gesprochenen Satzes versteht, wobei eine Verschlechterung der Verständlichkeit durch einem Anstieg des SRTs abgebildet wird. Die hierbei zum Einsatz kommenden Testsätze bestehen aus jeweils fünf Wörtern mit fester, grammatikalisch korrekter Syntax aber unvorhersagbarer Semantik. Der OLSA sieht vor, dass der Versuchsperson pro Bedingung 30 Testsätze aus einem Korpus von 600 Testsätzen vorgespielt werden. Je nach Anzahl korrekt verstandener Worte wird der Wiedergabepegel für den nachfolgenden Satz nach einer vorgegebenen Adaptionsregel angepasst und konvergiert bei einem Verständlichkeitsniveau von 50 %. Der zugehörige Wiedergabeschalldruckpegel entspricht dem SRT in Ruhe. Die Testbedingungen wurden durch Faltung der OLSA-Testsätze mit den simulierten BRIRs der vier virtuellen Räume generiert. Zusätzlich wurden Audiogramme aller Versuchspersonen erhoben, um die gemessenen SRTs um interindividuelle Empfindlichkeitsunterschiede zu korrigieren. Am Versuch nahmen 18 normalhörende Personen teil (5 weiblich, Altersdurchschnitt 30,4 Jahre). Als Testumgebung diente der reflektionsarme Halbraum am Institut für Technische Akustik der TU Berlin.

SRT Prädiktion

Zur Vorhersage der gemessenen SRTs diente das zugängliche Prädiktionsmodell nach Jelfs [2, 13] und einer vom Autor implementierten Erweiterung nach dem U/D-Ansatz. Die Ausgabe des Prädiktionsmodells ist ein Einzahlwert in dB, welcher als Gewinn an Signal-Rausch Abstand zwischen binauralem Hören mit Kopf gegenüber monauralem, omnidirektionalem Hören ohne Kopf interpretiert werden kann. Durch Invertierung und Mittelwertskalierung über alle betrachteten Testbedingungen können die vorhergesagten Modellausgabewerte mit gemessenen SRTs verglichen werden. Dieses Vorgehen wurde u.a. auch in [2] angewandt. Bedingt durch die Mittelwertskalierung lassen sich allerdings nur relative SRT-Unterschiede zwischen verschiedenen Bedingungen korrekt vorhersagen.

Der U/D-Ansatz wurde durch Multiplikation der generierten BRIRs mit einem frühen und einem späten Zeitfenster implementiert. Das frühe Zeitfenster gewichtet alle Anteile der BRIR bis zur gewünschten (um die Schallaufzeit bereinigte) U/D-Zeitgrenze mit 1. Ab der U/D-Zeitgrenze wird die Gewichtung des frühen Anteils über einer Länge von 1 ms linear auf 0 reduziert. Das späte Zeitfenster beginnt bei der gewünschten U/D-Zeitgrenze mit einer Gewichtung von 0 und steigt dann linear über die Dauer von 1 ms auf 1 an und verbleibt dort bis zum Ende der BRIR. Beide Zeitfenster summieren sich damit zu 1. Früher und später Anteil der BRIR werden separat in das Modell eingegeben, wobei der frühe Anteil als Sprachsignal, der späte Anteil als Störquelle interpretiert wird.

Raumunabhängige U/D-Zeitgrenze

In Abbildung 3 sind die gemessenen SRTs sowie die mit dem verwendeten Modell vorhergesagten SRTs für alle vier betrachteten Räume dargestellt, wobei für die Vorhersage zwei unterschiedliche, raumunabhängige U/D-Zeitgrenzen von 50 und 100 ms verwendet wurden. Es zeigt sich, dass der gemessene und prädizierte SRT in Ruhe mit zunehmender Nachhallzeit steigt. Allerdings überschätzt die Prädiktion den tatsächlichen SRT-

Tabelle 1: Optimale U/D-Zeitgrenzen, raumakustische Parameter und vorhergesagte, raumabhängige U/D-Zeitgrenzen

Condition	Optimum U/D-limits [ms]	Room acoustic parameters D/R [dB]	$C80_m$ [dB]	$IACC_m$	U/D-limit [ms] predicted by D/R [dB]	$C80_m$ [dB]	$IACC_m$
$T_{20,m} = 0.5$ s	59	-1.6	6.4	0.43	62	61	57
$T_{20,m} = 1.0$ s	90	-7.4	-0.7	0.22	97	97	99
$T_{20,m} = 2.0$ s	142	-10.9	-4.9	0.08	118	118	127
$T_{20,m} = 4.0$ s	122	-14.0	-8.7	0.06	136	137	131

Anstieg zwischen Räumen mit Nachhallzeiten unterhalb von $T_{20,m} = 2$ s und unterschätzt den SRT-Anstieg zwischen den Räumen mit $T_{20,m} = 2$ s und $T_{20,m} = 4$ s Nachhallzeit. Der beobachtete Prädiktionsfehler ist damit raumabhängig, wobei die größten Abweichungen zwischen Messung und Prädiktion bei geringen und mittleren Nachhallzeiten im Bereich zwischen 1-2 s zu beobachten sind. Nachhall ist in diesen Bedingungen offensichtlich weniger nachteilig als vom Model mit raumunabhängiger U/D-Zeitgrenze interpretiert.

Zwei im Zusammenhang mit Nachhall relevante, raumabhängige Wahrnehmungsaspekte die hierfür ursächlich sein könnten sind binaurale Nachhallunterdrückung und Raumadaption. Beide Effekte beschreiben Beobachtungen zu teilweiser Unterdrückung von Nachhall in der auditorischen Verarbeitung durch binaurales Hören (= binaurale Nachhallunterdrückung) bzw. durch zeitliche Adaption an den Nachhallkontext (= Raumadaption). In beiden Fällen wurde deutliche Raumabhängigkeiten beobachtet, wobei sich die stärksten Unterdrückungseffekte bei 1-2 s Nachhall gezeigt haben [14, 15, 16, 17]. In Räumen mit mehr bzw. weniger Nachhall verringerten sich diese Unterdrückungseffekte. Beide Effekte können vom verwendeten Prädiktionsmodell nicht abgebildet werden. Bei Nachhallzeiten im Bereich 1-2 s wurden im Experiment die größten Abweichungen hinsichtlich SRT-Anstieg zwischen Messung und Prädiktion beobachtet. Es wird geschlossen, dass diese beiden raumabhängigen Wahrnehmungsaspekte (teilweise) für die beobachteten Abweichungen zwischen Messung und Prädiktion ursächlich sind.

Raumabhängige U/D-Zeitgrenzen

Um die genannten Wahrnehmungsaspekte zumindest quantitativ im Modell zu berücksichtigen wurde nun untersucht, inwieweit sich raumabhängige U/D-Zeitgrenzen aus am Zuhörerort beobachteten raumakustischen Parametern vorhersagen lassen. Lassen sich hierbei Zusammenhänge finden, lässt sich der U/D-Ansatz mit raumabhängigen U/D-Zeitgrenzen auf beliebige Räume bzw. Quelle-Empfänger Kombinationen anwenden.

Hierzu wurden zunächst optimale U/D-Zeitgrenzen für jeden der vier betrachteten Räume ermittelt, die zu einer optimalen Nachbildung (mittlerer Betragsfehler MAE < 1 dB) der gemessenen SRTs mit dem verwendeten Prädiktionsmodell führen. Diese sind für alle vier betrachteten Räume in Tabelle 1, Spalte 2 abgebildet.

Im zweiten Schritt wurde eine Regressionsanalyse durchgeführt, wobei die ermittelten, optimalen U/D-Zeitgrenzen als abhängige Variable eingingen. Als unabhängige Variable dienten an der Empfängerposition zusätzlich ausgewertete raumakustische Parameter (Direkt/Diffusschallverhältnis D/R, mittleres Klarheitsmaß $C80_m$ und mittlerer interauraler Kreuzkorrelationskoeffizient $IACC_m$). Letztere sind in Tabelle 1, Spalte 3-5 für alle vier betrachteten Räume gelistet.

Für alle raumakustischen Parameter wurden signifikante Regressionsgleichungen gefunden ($F(1, 70) > 120$, $p < .001$), wobei die höchste Varianzaufklärung durch $IACC_m$ erzielt wurde ($IACC_m$: $r^2_{adj.} = 0.72$; D/R: $r^2_{adj.} = 0.62$; $C80_m$: $r^2_{adj.} = 0.62$). Die zugehörigen Regressionsgleichungen lauten: U/D = 143 – 201 ($IACC_m$) ms, U/D = 52.1 – 6.0 (D/R) ms, und U/D = 93.4 – 5.0 ($C80_m$) ms. Die Standardfehler liegen bei 19 ms ($IACC_m$), bzw. 21 ms (D/R und $C80_m$). Die mit diesen Regressionsgleichungen aus den raumakustischen Parametern vorhergesagten raumabhängigen U/D-Zeitgrenzen sind in Tabelle 1, Spalte 6-8 gelistet. In Abbildung 4 sind die gemessenen, sowie die mit dem verwendeten Modell vorhergesagten SRTs für alle vier betrachteten Räume dargestellt. Für die Vorhersage wurden die mittels $IACC_m$ vorhergesagten raumabhängigen U/D-Zeitgrenzen verwendet. Zudem sind die prädizierten SRTs mit optimalen U/D-Zeitgrenzen, die in Regressionsanalyse als abhängige Variable eingingen, dargestellt. Wie aus Abbildung 4 hervorgeht, verringert sich der Prädiktionsfehler bei Verwendung der raumabhängigen U/D-Zeitgrenzen im Vergleich zur Vorhersage mit raumunabhängigen U/D-Zeitgrenzen. Der mittlere Betragsfehler (MAE) in dB über alle vier betrachteten Räume mit raumunabhängiger und mittels raumakustischer Parameter vorhergesagter raumabhängiger U/D-Zeitgrenzen ist in Tabelle 2 dargestellt. Zusätzlich wurde die gleiche Methodik auf zwei Testbedingungen (S0/S90) á 4 Quelle-Empfänger Kombinationen eines externen Datensatzes [5], nachfolgend RS11 genannt, angewandt. Für die Vorhersage der raumabhängigen U/D-Zeitgrenzen wurden in diesem Fall ebenfalls die im Experiment ermittelten Regressionsgleichungen verwendet. Wie aus Tabelle 2 hervorgeht, lässt sich der mittlere Prädiktionsfehler mit den vorhergesagten, raumabhängigen U/D-Zeitgrenzen sowohl für die vier betrachteten Räume im Experiment, als auch für die acht verschiedene Quelle-Empfänger Kombinationen in einem virtuellen Raum mit ca. 2 s Nachhall (RS11) reduzieren.

Tabelle 2: Mittlerer Betragsfehler (MAE) in dB mit raumunabhängigen und raumabhängigen U/D-Zeitgrenzen

	MAE [dB] with fixed U/D-limits [ms]		MAE [dB] with room/ receiver dependent U/D-limits predicted by		
	50	100	D/R ($r^2_{adj.} = 0.62$)	$C80_m$ ($r^2_{adj.} = 0.62$)	$IACC_m$ ($r^2_{adj.} = 0.72$)
Experiment	1.9	2.9	1.3	1.3	1.2
RS11 data (S0/S90)	2.6/2.5	2.0/ 2.2	1.2/ 1.1	1.5/ 2.0	0.4/ 1.5
Ø	2.3	2.4	1.2	1.6	1.0

Zusammenfassung

Zwei, im Zusammenhang mit Nachhall relevante raum- bzw. empfängerabhängige Wahrnehmungsaspekte (binaurale Nachhallunterdrückung und Raumadaption) hatten mutmaßlich Einfluss auf die gemessenen SRTs. Beide Wahrnehmungsaspekte werden im aktuellen Prädiktionsmodell nicht berücksichtigt. Die Raumabhängigkeit beider Aspekte führt zu einem beobachteten raumabhängigen Prädiktionsfehler. Mittels raum-/ empfängerabhängiger U/D-Zeitgrenze, vorhergesagt aus am Empfänger beobachteten raumakustischen Parametern lässt sich dieser Prädiktionsfehler teilweise kompensieren. Der mittlere Prädiktionsfehler verringert sich dadurch um 1,3 – 1,4 dB.

Literatur

[1] Bradley, J. S.; Hiroshi Sato und M. Picard (2003): „On the importance of early reflections for speech in rooms." In: *The Journal of the Acoustical Society of America*, 113(6) S. 3233–3244.

[2] Jelfs, Sam; John F. Culling und Mathieu Lavandier (2011): „Revision and validation of a binaural model for speech intelligibility in noise." In: *Hearing research*, 275(1) S. 96–104.

[3] Beutelmann, Rainer; Thomas Brand und Birger Kollmeier (2010): „Revision, extension, and evaluation of a binaural speech intelligibility model." In: *The Journal of the Acoustical Society of America*, 127(4) S. 2479–2497.

[4] Durlach, Nathaniel I. (1963): „Equalization and Cancellation Theory of Binaural Masking-Level Differences." In: *The Journal of the Acoustical Society of America*, 35(8) S. 1206–1218.

[5] Rennies, Jan; Thomas Brand und Birger Kollmeier (2011): „Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet." In: *The Journal of the Acoustical Society of America*, 130(5) S. 2999–3012.

[6] Rennies, Jan (2014): „Modeling the effects of a single reflection on binaural speech intelligibility." In: *The Journal of the Acoustical Society of America*, 135(3) S. 1556–1567. doi:10.1121/1.4863197.

[7] Leclère, Thibaud; Mathieu Lavandier und John F. Culling (2015): „Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation." In: *The Journal of the Acoustical Society of America*, 137(6) S. 3335–3345.

[8] Schröder, Dirk und Michael Vorländer (2011): „RAVEN: A real-time framework for the auralization of interactive virtual environments." In: *Forum Acusticum*.

[9] Brinkmann, Fabian et al. (2017): „A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations." In: *Journal of the Audio Engineering Society*, 65(10) S. 841–848.

[10] Kuehnel, Volker; Birger Kollmeier und Kirsten Wagener (1999): „Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests." In: *Zeitschrift für Audiologie*, 38 S. 4–15.

[11] Wagener, Kirsten; Thomas Brand und Birger Kollmeier (1999): „Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests." In: *Zeitschrift für Audiologie/Audiological Acoustics*, 38 S. 44–56.

[12] Wagener, K.; T. Brand und B. Kollmeier (1999): „Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests." In: *Zeitschrift fr Audiologie/Audiological Acoustics*, 38 S. 8695.

[13] Søndergaard, P. und P. Majdak (2013): „The Auditory Modeling Toolbox." In: *The Technology of Binaural Listening*. Berlin, Heidelberg: Springer, S. 33–56.

[14] Gelfand, S. A. und I. Hochberg (1976): „Binaural and monaural speech discrimination under reverberation." In: *Audiology*, 15(1) S. 72–84.

[15] Moncur, John P. und Donald Dirks (1967): „Binaural and monaural speech intelligibility in reverberation." In: *Journal of speech and hearing research*, 10(2) S. 186–195.

[16] Nábělek, Anna K. und Pauline K. Robinson (1982): „Monaural and binaural speech perception in reverberation for listeners of various ages." In: *The Journal of the Acoustical Society of America*, 71(5) S. 1242–1248.

[17] Zahorik, Pavel und Eugene J. Brandewie (2016): „Speech intelligibility in rooms: Effect of prior listening exposure interacts with room acoustics." In: *The Journal of the Acoustical Society of America*, 140(1) S. 74–86.