Ein blindes Modell zur Vorhersage der binauralen Sprachverständlichkeit

Christopher Hauth, Thomas Brand

Medizinische Physik und „Cluster of Excellence Hearing4All“, Universität Oldenburg, 26129 Oldenburg,
E-Mail: christopher.hauth@uni-oldenburg.de

Einleitung

Modelle des Hörens werden genutzt, um die effektive Verarbeitung im menschlichen auditorischen System zu beschreiben. Im Gegensatz zum menschlichen auditorischen System benötigen diese Modelle jedoch in der Regel a priori Informationen. Diese ist häufig der separate Zugriff auf das Ziel- und Störsignal. Dadurch kann das resultierende Perzept, zum Beispiel eine Sprachverständlichkeit oder Detektion eines Tones im Rauschen, vorhergesagt werden. Jedoch wird nicht ersichtlich, ob die auditorische Verarbeitung tatsächlich so stattgefunden hat. Daher ist es notwendig, die Modelle so zu modifizieren, dass sie wie das menschliche auditorische System mit gemischten Signalen arbeiten.

Im Folgenden soll daher ein binaurales Sprachverständlichkeitsmodell vorgestellt werden, welches kein a priori Wissen über Sprache und Störgeräusch benötigt. Es kombiniert einen blinden Equalization-Cancellation (EC) Mechanismus [1] als front end mit dem Maß der „Speech to Reverberation Modulation Ratio“ (SRMR) [2] als back end und kann so den Gewinn durch räumliche Trennung von Ziel- und Störsignal auf die Sprachverständlichkeit in stationärem Rauschen beschreiben.

Hintergrund

Das binaurale Hören bietet Vorteile gegenüber dem monauralen Hören. Unter anderem können gegenüber einer diotischen Darbietung niedrigere Schwellen erzielt werden, wenn entweder das Zielsignal oder das Störsignal interaurale Unterschiede in der Laufzeit (ITD) oder Phase (IPD) aufweist. Dieser Effekt wird auch „binaural masking level difference“ (BMLD) genannt. Ein Modell, welches diesen Effekt beschreiben kann, ist das „Equalization-Cancellation“ (EC) Modell [1]. Das EC Modell kompensiert zunächst interaurale Pegelunterschiede (α) zwischen dem linken und rechten Ohr. Anschließend findet ein Ausgleich der ITD (τ) statt, so dass die Phasenlage des Störsignals interaural die gleiche ist. In einem dritten Schritt wird dann das Signal im linken Kanal von dem im rechten Kanal abgezogen, so dass das Störsignal durch destruktive Interferenzen abgeschwächt wird und so der Signal-zu-Rausch Abstand (SNR) vergrößert wird. Diese Operation lässt sich im Zeitbereich durch Gleichung (1) ausdrücken.

$$X_{EC}(t) = X_R(t) - \alpha X_L(t-\tau) \tag{1}$$

Das Modell wurde schon vielfach in Modellen des binauralen Hörens verwendet, vor allem in binauralen Sprachverständlichkeitsmodellen, in denen es mit dem Speech Intelligibility Index (SII) [3] kombiniert wurde [4][5].

Binaurales Sprachverständlichkeitsmodell (BSIM 2010)

In Abbildung 1 ist das binaurale Sprachverständlichkeitsmodell (BSIM 2010) nach [5] gezeigt. Auch dieses Modell benötigt a priori Wissen, da am Eingang Sprache und Rauschen jeweils separat für den rechten und linken Kanal vorliegen müssen.

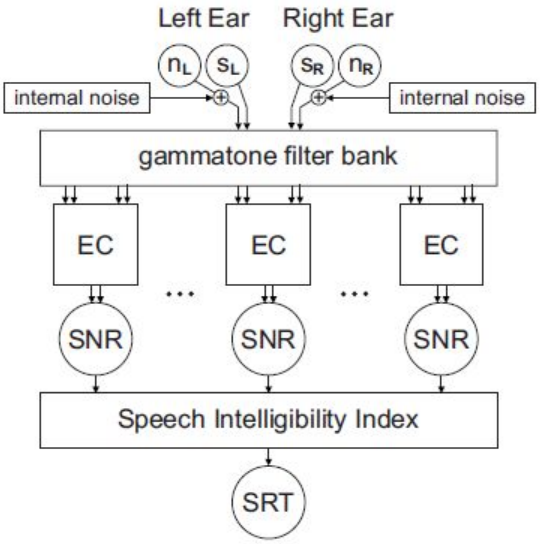

Abbildung 1: Schema des BSIM 2010 nach [5].

Zunächst werden Sprache und Rauschen mit einer Gammaton Filterbank, bestehend aus 30 Filtern von 146 Hz bis 8346 Hz und einer Bandbreite von 1 ERB [6], gefiltert um die Frequenzselektivität auf der Basilarmembran nachzubilden. In jedem Frequenzband wird dann der EC Mechanismus angewendet, dessen Ausgangssignal dann mittels SII in eine Sprachverständlichkeitsschwelle (SRT), zum Beispiel für 50% korrekt verstandene Sprache, überführt wird. Dadurch, dass das Modell am Eingang Zugriff auf Sprache und Rauschen separat hat, kann der EC Mechanismus als SNR Maximierung realisiert werden, indem nach Ausgleich der ITD dasjenige α gesucht wird, welches den SNR maximiert (vgl. Gleichung (2)).

$$\frac{\delta}{\delta\alpha}\left(\frac{I(S_{EC})}{I(N_{EC})}\right) = 0\,, \tag{2}$$

Hierbei stellen $I(S_{EC})$ und $I(N_{EC})$ die Intensität von Sprache und Rauschen nach dem EC Mechanismus dar.

Blinder EC Mechanismus

Das BSIM 2010 hat a priori Wissen über Sprache und Rauschen und ist daher in der Lage den SNR zu maximieren.

Wenn der EC Mechanismus auf gemischte Signale angewendet wird, ist eine Maximierung des SNR schwer zu realisieren. Da die dominante Quelle in dem Mix von Sprache und Rauschen die Schätzung der Parameter dominiert, kann jedoch in der der EC Stufe eine Pegelminimierung erzielt werden. Allerdings muss die Annahme getroffen werden, dass der zugrunde liegende SNR negativ ist, da sonst die EC Parameter auf Grundlage der Sprache geschätzt würden. Das zöge keine Verbesserung der Sprachverständlichkeit nach sich. Diese Annahme ist gerechtfertigt, wenn die binaurale Demaskierung einen positiven Effekt auf die Sprachverständlichkeit bei negativem SNR hat und Schwellen für Normalhörende bei negativen SNRs zu finden sind. Bei positivem SNR hingegen kann ein Sprachverstehen von 100% auch mit monauralem Hören erzielt werden. Auch in [7] waren die Schwellen für 50% Sprachverstehen bei negativen SNR zu finden. Weiterhin kann eine SNR Abhängigkeit der BMLD gezeigt werden. In [8] wurde die Wortverständlichkeit für die Situation von interaural gleichphasigem Rauschen (S_0N_0) und interaural gegenphasigem Rauschen (S_0N_π) ermittelt. Hier zeigte sich, dass die BMLD verschwindet, wenn der SNR positiv wird. Eine Rekonstruktion der Darstellung der Ergebnisse von [8] für diese Konfiguration ist in Abbildung 2 gezeigt.

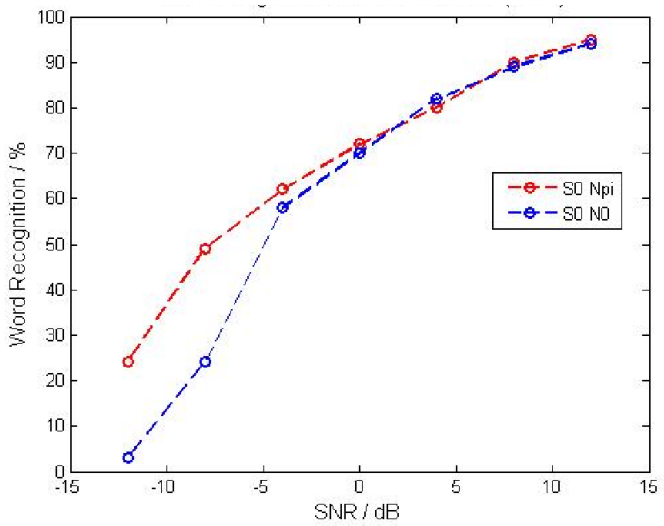

Abbildung 2: Wortverständlichkeit in Abhängigkeit des SNR für die Konditionen S_0N_0 (blau) und S_0N_π nach [8].

Unter der Annahme das der SNR negativ ist, kann ein blindes EC Modell mit dem SII kombiniert werden und somit ein halb blindes Modell realisiert werden, welches im Folgenden „instantBSIM“ genannt wird. Die EC Parameter werden auf Grundlage von gemischten Signalen geschätzt, so dass die dominante Quelle im Signal abgeschwächt wird und somit der Pegel minimiert wird

Evaluation

Für die Evaluierung der Modelle werden die Daten aus dem Sprachverständlichkeitsexperiment von [7] genutzt. Dabei werden die zwei räumlichen Konfigurationen „reflexionsarm“ (orig.: Anechoic) und „Seminarraum“ (orig.: Office) gewählt. In diesen beiden Konfigurationen wurden Sprachverständlichkeitsschwellen für Rauschen aus den Richtungen -140°, -100°, -45°, 0°, 45°, 80°, 125° und 180° gemessen, wobei das Sprachmaterial immer aus einer Richtung von 0° präsentiert wurde. Als Sprachmaterial wurde der Oldenburger Satztest (OlSa) [9] genutzt. Die Sätze weisen eine feste grammatikalische Struktur der Form: „Nomen, Verb, Zahlwort, Adjektiv und Objekt“ auf (z.B.: „Peter kauft vier nasse Dosen“). Das Rauschen wurde durch zufälliges Überlagern des Sprachmaterials generiert, wodurch sich ein stationäres Rauschen mit dem gleichen Langzeitspektrum wie das Sprachmaterial ergibt.

Ergebnisse I: instantBSIM vs. BSIM 2010

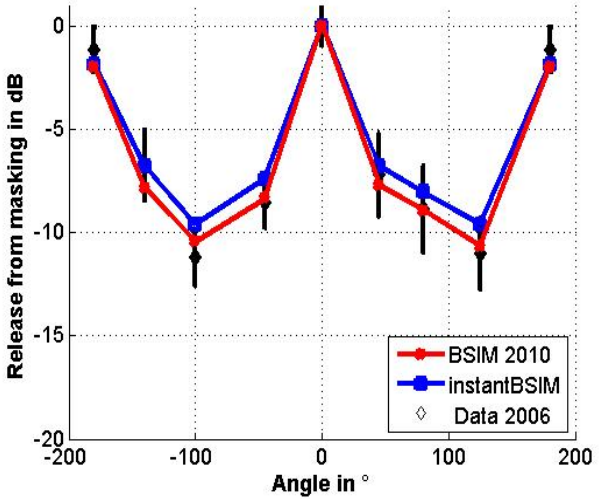

Abbildung 3: „Release from masking“ für Sprachverständlichkeit im reflexionsarmen Raum. Dargestellt sind Mittelwert und Standardabweichung der perzeptiven Daten (schwarz), sowie Modellvorhersagen des BSIM 2010 (rot) und instantBSIM (blau).

In Abbildung 3 sind die Ergebnisse für die reflexionsarme Umgebung, in Abbildung 4 die Ergebnisse für den Seminarraum gezeigt. Es zeigt sich, dass der blinde EC Mechanismus nahezu äquivalente Ergebnisse wie das BSIM 2010 liefert, wenn der SNR negativ ist. Im Vergleich zum BSIM 2010 wird der „release from masking“ in der reflexionsarmen Kondition um ca. 1 dB unterschätzt. Im Seminarraum ist die Abweichung kleiner als 1 dB. Daraus lässt sich schließen, dass eine Minimierung des Pegels bei negativem SNR einer Maximierung des SNR entspricht. Weiterhin zeigt sich, dass sich die binaurale Demaskierung

vollständig durch einen bottom-up Prozess modellieren lässt und ein top-down Prozess, der den SNR maximiert, nicht zwingend erforderlich ist.

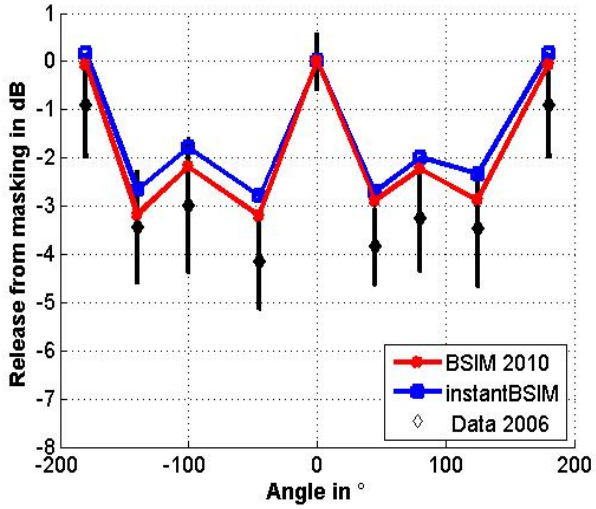

Abbildung 4: „Release from masking“ für Sprachverständlichkeit im Seminarraum. Dargestellt sind Mittelwert und Standardabweichung der perzeptiven Daten (schwarz), sowie Modellvorhersagen des BSIM 2010 (rot) und instantBSIM (blau).

Um ein vollständig blindes Modell zu realisieren, muss eine Alternative für den SII gefunden werden.

Speech to Reverberation Modulation Ratio (SRMR)

Die „Speech to Reverberation Modulation Ratio“ (SRMR) [2] ist ein Maß, welches ursprünglich für die Bewertung von Enthallungsalgorithmen entwickelt wurde. Weiterhin kann es zur Vorhersage von Sprachverständlichkeit genutzt werden. Für die Berechnung der SRMR wird das Signal zunächst in 23 Frequenzbänder zerlegt und die Hilbert Einhüllende berechnet. Dann wird in jedem Frequenzband eine 8 kanalige Modulationsfilterbank angewendet. Die jeweiligen Mittenfrequenzen sind 4, 6.5, 10.7, 17.6, 28.9, 47.5, 78.1, und 128 Hz. In jedem Frequenzband wird dann ein Verhältnis von Energie in tiefen zu Energie in hohen Modulationsfiltern berechnet, wobei jeweils die Summe über 4 Modulationsfilter berechnet wird (siehe Gleichung (3)).

$$SRMR(f) = \frac{\sum_{i=1}^{4} rms(env_i)}{\sum_{j=5}^{8} rms(env_j)} \quad (3)$$

Anschließend wird die SRMR über alle Frequenzen gemittelt, sodass das Maß in einem Einzahlwert resultiert.

Für reine Sprache resultiert das Maß in einem hohen Wert, da die Energie im Modulationsfilter mit der Mittenfrequenz von 4 Hz hoch ist. Wird nun Nachhall oder ein Störgeräusch auf das Sprachsignal aufgeprägt, nimmt das Verhältnis von Energie in tiefen Modulationsbändern zu Energie in hohen Modulationsbändern ab und das SRMR Maß nimmt einen niedrigeren Wert an.

Ein blindes Modell (blindBSIM)

Für ein blindes binaurales Sprachverständlichkeitsmodell wird der blinde EC Mechanismus mit dem SRMR Maß, wie in Abbildung 5 gezeigt, kombiniert. Hierbei wird zunächst das verrauschte Sprachsignal mittels Gammatone Filterbank (siehe BSIM 2010) gefiltert. Um die Duplex Theorie, nach der die binaurale Verarbeitung hauptsächlich in tiefen Frequenzen bis 1500 Hz stattfindet, zu berücksichtigen, wird bis 1500 Hz der blinde EC Mechanismus auf das Signal angewendet. Über 1500Hz bleiben der linke und rechte Kanal unverarbeitet, um ein Hören mit dem besseren Ohr zu gewährleisten. Der binaural verarbeitete Signalanteil wird dann mit den monaural unverarbeiteten Signalanteilen von linkem und rechtem Ohr kombiniert und das SRMR Maß berechnet. Das Maximum zwischen linkem und rechtem Ohr definiert dann die Sprachverständlichkeit.

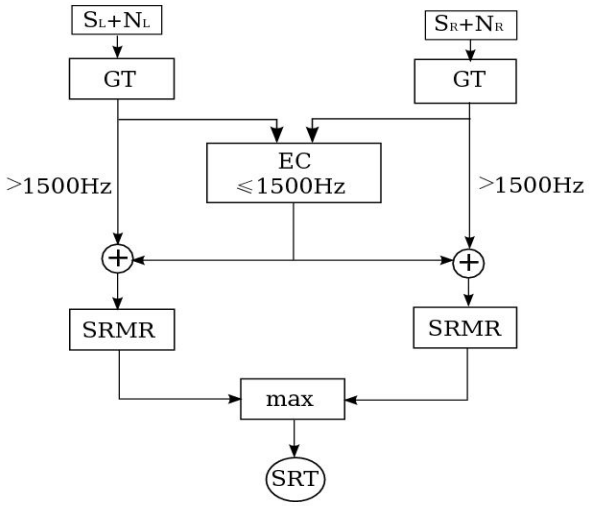

Abbildung 5: Schema des blinden binauralen Sprachverständlichkeitsmodell „blindBSIM“.

In [10] wurde das SRMR Maß verwendet, um den Effekt des besseren Ohres vorherzusagen. Daher wurde die SRMR jeweils für den linken und rechten Kanal berechnet um dann additiv mit der BMLD eine binaurale Sprachverständlichkeit vorherzusagen. In diesem Ansatz wird das SRMR Maß also auch auf die binaurale Komponente angewendet.

Ergebnisse II: blindBSIM

In Abbildung 6 sind die Ergebnisse des blindBSIM im Vergleich zu den perzeptiven Daten und den Modellrealisationen BSIM 2010 und instantBSIM gezeigt. Durch den Austausch des SII mit dem SRMR wird der „release from masking“ unterschätzt, wenn er klein ist (180°), aber überschätzt, wenn er groß ist (z.B. -100°). Prinzipiell wird der Trend der Daten aber richtig beschrieben. Eine Ausnahme bildet dabei die Kondition, in der das Rauschen aus 80° dargeboten wurde. Hier wird eine zu gute Sprachverständlichkeit vorhergesagt, der „release from masking“ also überschätzt.

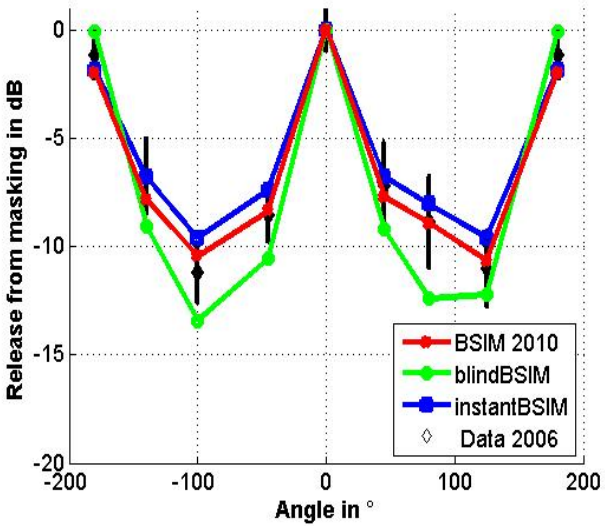

Abbildung 6: „Release from masking“ für Sprachverständlichkeit im reflexionsarmen Raum. Dargestellt sind Mittelwert und Standardabweichung der perzeptiven Daten (schwarz), sowie Modellvorhersagen des BSIM 2010 (rot), instantBSIM (blau) und blindBSIM (grün).

Diskussion

Im ersten Teil wurden die Auswirkungen eines blind betriebenen EC Mechanismus untersucht. Im Gegensatz zum nicht blinden EC Mechanismus kann der SNR nicht maximiert werden. Stattdessen kann aber eine Abschwächung der dominanten Quelle erzielt werden, sodass der Pegel minimiert wird. Es hat sich gezeigt, dass nahezu die gleichen Ergebnisse für beide Modellansätze erzielt werden können, sofern der SNR negativ ist. Das bedeutet, dass für negative SNR eine Minimierung des Pegels und eine Maximierung des SNR nahezu äquivalente Ergebnisse liefern. In [10] wurde ein binaurales Sprachverständlichkeitsmodell vorgestellt, welches die EC Parameter mittels eines Lokalisationsmodelles schätzt. Dabei mussten allerdings die Annahmen getroffen werden, dass das Zielsignal aus einer Richtung von 0° kommt und nur eine Störquelle vorhanden ist. Das impliziert einen top-down Mechanismus, der Sprachsignal und Störsignal identifiziert und die Störquelle abschwächt. Diese Annahmen werden in dem Ansatz des instantBSIM nicht benötigt. Hier muss lediglich die Annahme eines negativen SNR getroffen werden, damit der EC Mechanismus die Parameter auf Grundlage des Störsignals schätzen kann. Diese Annahme ist auch durch Erkenntnisse von [8] gerechtfertigt. Daraus resultiert ein binuaraler bottom-up Mechanismus, der die Störquelle abschwächt und so die BMLD erklären kann.

Ein komplett blindes Modell der binauralen Sprachverständlichkeit kann durch Kombination des blinden EC Mechanismus mit dem SRMR Maß realisiert werden. Allerdings ist das verwendete SRMR back-end nur limitiert einsetzbar, da es Modulationen analysiert und in diesem konkreten Fall für stationäres Rauschen funktioniert. Eine ähnliche Leistung kann aber für sprachähnliche Störgeräusche oder langsam modulierte Störgeräusche nicht erwartet werden, da dann Zielsignal und Störsignal ähnliche Modulationen aufweisen.

Literatur

[1] Durlach, N.I. (1963): Equalization and Cancellation Theory of Binaural Masking Level Differences,” The Journal of the Acoustical Society of America 35(8), 1206–1218

[2] Falk, T. H., Zheng, C., and Chan, W.-Y. (2010), “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” Audio, Speech, and Language Processing, IEEE Transactions on 18(7), 1766–1774

[3] ANSI (1997), “Methods for the calculation of the speech intelligibility index,” American National Standard S3.5-1997 (Standards Secreteriat, Acoustical Society of America)

[4] Lavandier, M. and Culling, J. F. (2010), “Prediction of binaural speech intelligibility against noise in rooms,” The Journal of the Acoustical Society of America 127(1), 387–399

[5] Beutelmann, R., Brand, T., and Kollmeier, B. (2010), “Revision, extension, and evaluation of a binaural speech intelligibility model,” The Journal of the Acoustical Society of America 127(4), 2479–2497

[6] Glasberg, B. R. and Moore, B. C. (1990), “Derivation of auditory filter shapes from notched-noise data,” Hearing Research 47, 103 – 138

[7] Beutelmann, R. and Brand, T. (2006), “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,” The Journal of the Acoustical Society of America 120(1), 331–342

[8] Licklider, J. C. R. (1948), “The influence of interaural phase relations upon the masking of speech by white noise,” The Journal of the Acoustical Society of America 20(2), 150–159.

[9] Wagener, K., Brand, T., Kühnel, V., and Kollmeier, B. (1999a,b,c), “Entwicklung und Evaluation eines Satztests für die Deutsche Sprache ,” Z. Für Audiologie, Audiological Acoust. 38, 4–15.

[10] Cosentino, S., Marquardt, T., McAlpine, D., Culling, J.F., and Falk, T. H. (2014), “A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals,” The Journal of the Acoustical Society of America 135(2), 796–807.