Hören mit zwei Ohren - Modellierung und Verbesserung von Sprachverständlichkeit

Jan Rennies

Fraunhofer IDMT, Hör-, Sprach- und Audiotechnologie, Oldenburg, E-Mail: jan.rennies@idmt.fraunhofer.de

Einleitung

Es ist ein erstaunliches Phänomen, wie gut das gesunde Gehör in der Lage ist, unterschiedliche Schallquellen auch in schwierigen akustischen Bedingungen zu trennen und so beispielsweise einem bestimmten Sprecher in Anwesenheit mehrerer Störschallquellen zu folgen. Diesen „Cocktail-Party-Effekt“ verdankt das Gehör einer hochwirksamen Verarbeitung von interauralen Pegel- und Phasenunterschieden der einzelnen Signale. Dieser Beitrag stellt effektive Modellansätze binauraler auditorischer Sprachverarbeitung vor, wobei besonders Hörsituationen beleuchtet werden, in denen die auditorische Quellentrennung durch Echos, Nachhall oder Hörminderung erschwert wird. Anschließend werden anhand mehrerer Anwendungsbeispiele die praktischen Einsatzmöglichkeiten binauraler Modelle veranschaulicht, etwa bei der Sprachkommunikation im Fahrzeuginnenraum oder mit Stereokommunikationssystemen. Im zweiten Teil des Beitrags wird vorgestellt, wie durch den Einsatz von Sprachverständlichkeitsmodellen Algorithmen zur Sprachverbesserung automatisch gesteuert werden können. Je nach Anwendungsfall müssen die Modelle hierfür nicht immer notwendigerweise über komplexe binaurale Verarbeitungsstufen verfügen, sondern bereits einfachere Modellansätze können ausreichen, um eine effektive Verbesserung der Sprachverständlichkeit in Echtzeit an eine gegebene Umgebungsakustik anzupassen. Dadurch ist es möglich das Sprachsignal nur dann zu modifizieren, wenn es nach Modellvorhersage zu schlecht verständlich ist, und damit auch bei variablen Hintergrundgeräuschen einen guten Kompromiss zwischen Natürlichkeit und Verständlichkeit herzustellen. Anwendungsbeispiele für dieses sogenannte „near-end listening enhancement“ sind Durchsagesysteme an Bahnhöfen oder Flughäfen, In-Car-Kommunikation oder Mobiltelefonie.

Modellierung von Sprachverständlichkeit in Räumen

Experimente und Modellansätze

Hörsituationen in realen Räumen zeichnen sich in der Regel durch komplexe Interaktionen der Raumakustik mit der relativen räumlichen Anordnung der Zielschallquelle und der Störschallquellen aus. In diversen Studien wurde daher untersucht, welchen Vorteil das auditorische System daraus ziehen kann, wenn Zielsprache und Störgeräusche nicht aus derselben Raumrichtung dargeboten werden. Abbildung 1 veranschaulicht diesen Effekt. Betrachtet werden Daten (schwarze Kreise) der Studie aus [1], in der Sprachverständlichkeitsschwellen (engl. speech reception thresholds, SRTs), d.h. Signal-Rausch-Abstände (SNR) bei einer Wortverständlichkeit von 50%, für einen frontalen Sprecher gemessen wurden, dessen Sprache durch ein stationäres, sprachgefärbtes Rauschen aus unterschiedlichen Azimuthrichtungen überlagert wurde. Die in Abbildung 1 gezeigten Daten zeigen, dass die Schwellen am höchsten sind, wenn Sprache und Rauschen aus derselben Richtung kommen. Sobald jedoch das Störgeräusch räumlich vom Zielsprecher getrennt wird, bricht die Schwelle drastisch ein, was in reflexionsarmen Umgebungen (rechts) einen Effekt von bis zu ca. 12 dB ausmachen kann. In realen Umgebungen wie einem Büro (links) ist der Effekt immer noch signifikant, jedoch deutlich reduziert.

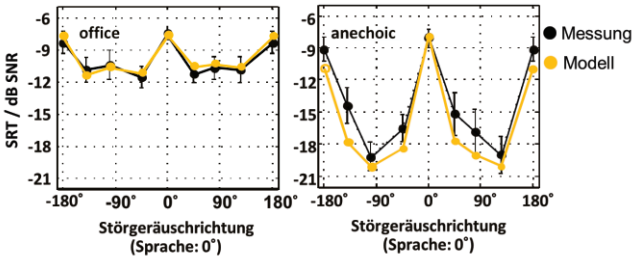

Abbildung 1: Gemessene (schwarz) und von BSIM vorhergesagt (orange) SRT in einer reflexionsarmen (rechts) und realen (links) Umgebung in Abhängigkeit des Störgeräuschazimuths bei einer frontalen Sprachquelle, entnommen aus [1].

Erklärt werden kann der zugrunde liegende effektive Verarbeitungsmechanismus unter der Annahme, dass das Gehör davon Gebrauch macht, dass sich die interauralen Pegel- und Laufzeitunterschiede eines frontalen Sprachsignals von denen eines räumlich getrennten Störgeräusches unterscheiden. Das binaurale System nutzt diese Unterschiede auf eine sehr effiziente Weise aus.

Zur effektiven Modellierung dieses Effektes wurden der sog. Equalization-Cancelation (EC) Mechanismus vorgeschlagen [2, 3]. In diesem Ansatz wird angenommen, dass das binaurale System in den einzelnen auditorischen Filtern unabhängig voneinander eine Optimierung des SNR vornimmt, indem es das Summensignal des linken und rechten Ohres relativ zueinander verzögert und verstärkt (equalization) und anschließend subtrahiert (cancelation). Bei entsprechender Wahl von Verzögerung und Verstärkung kann erreicht werden, dass die interauralen Pegel- und Laufzeitdifferenzen des Störgeräusches ausgeglichen werden, sodass durch die Subtraktion ein erhöhter SNR entsteht als an den jeweiligen Ohrsignalen.

Dieser Prozess wurde in verschiedenen Hörmodellen implementiert, wie z.B. im Binaural Speech Intelligibility Model (BSIM) in [1, 4], das schematisch in Abbildung 2 dargestellt ist (links).

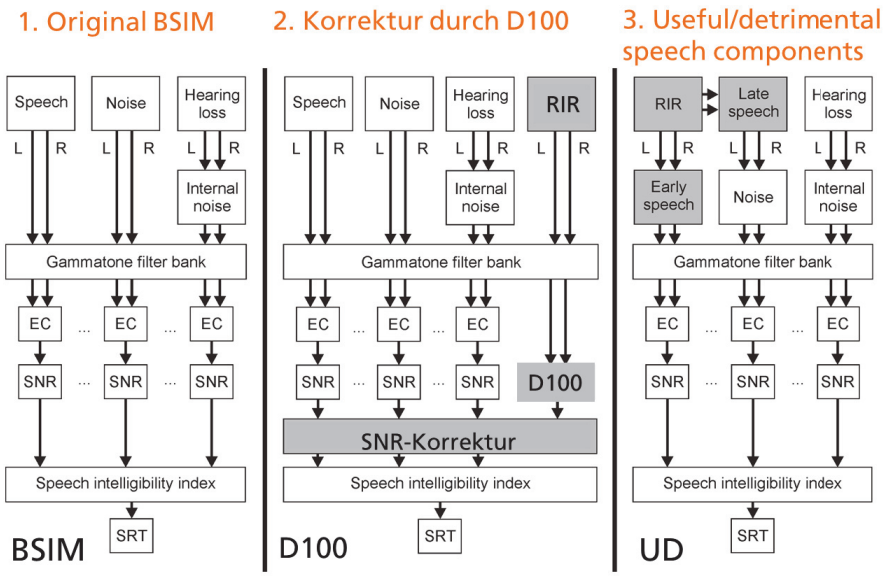

Abbildung 2: Schematische Darstellung des BSIM (links), der Erweiterung mit einer post-hoc Korrektur basierend auf dem Deutlichkeitsmaß (Mitte) und der Erweiterung basierend aus einer vor der binauralen Stufe stattfindenden Trennung von nützlichen und schädlichen Sprachanteilen (rechts), entnommen aus [5].

Das Modell verarbeitet getrennt vorliegende Sprach- und Störgeräuschsignale zunächst durch eine Gammaton-Filterbank und führt anschließend in jedem auditorischen Filter den erwähnten EC-Prozess durch. Die ggf. durch diese binaurale Stufe verbesserten SNR werden anschließend durch den Sprachverständlichkeitsindex (engl. speech intelligibility index, SII [6]) in eine vorhergesagte SRT überführt.

Orangene Kreise in Abbildung 1 zeigen, dass die Vorhersagen von BSIM sehr gut mit den Messdaten übereinstimmen. Dies trifft auch auf die Daten der realen Hörumgebung und den reduzierten binauralen Gewinn zu (links). Dies ist leicht verständlich, wenn man sich vergegenwärtigt, dass in verhallten Umgebungen die „Equalization“ des Störgeräusches weniger gut funktioniert, da die beiden Ohrsignale durch den Raumhall dekorreliert werden.

Obwohl BSIM in vielen Hörsituationen wie geschildert gut mit im Experiment beobachteten Effekten übereinstimmt, ist es nicht für beliebige Hörsituationen anwendbar. In [5] wurde gezeigt, dass dies vor allem der Fall ist, wenn die Zielsprache ebenfalls in erheblichem Maße verhallt ist. Wie anhand des linken Schemas in Abbildung 2 angedeutet wird, wird in BSIM das vollständige Sprachsignal beim EC-Prozess und der Verarbeitung durch den SII als nützlich (*useful*) für die Sprachverständlichkeit angenommen wird. Der vielfach untersuchte schädliche (*detrimental*) Effekt des Nachhalls wird somit nicht erfasst. In [5] wurden daher Erweiterungen von BSIM vorgeschlagen, deren Schemata ebenfalls in Abbildung 2 dargestellt sind. Für eine Erweiterung (Mitte) wird eine post-hoc Korrektur durch das Deutlichkeitsmaß D_{te} vorgenommen, d.h., durch den Quotienten des frühen Energieanteils der Raumimpulsantwort (engl. room impulse response, RIR) bis zur Zeit *te* (hier 100 ms) und der Gesamtenergie der RIR.

In dieser Erweiterung wird der schädliche Einfluss des Nachhalls also auf Ebene des SII modelliert, der binaurale Verarbeitungsprozess (EC) jedoch weiterhin anhand des vollständigen Sprachsignals vorgenommen. Im Gegensatz dazu wird in einer weiteren (rechts) die Trennung von nützlichen und schädlichen Sprachanteilen bereits vor der binauralen Stufe vorgenommen. Hierbei wird der der späte, verhallte Sprachanteil dem externen Störgeräusch zugeschlagen.

Anhand der in [7] erhobenen Messdaten lassen sich die Modellansätze systematisch vergleichen. Es wurden SRT gemessen für eine frontale Sprachquelle (S_0), eine einzige Reflexion der Sprachquelle (hier ebenfalls frontal, R_0) und ein frontales (N_0), diffuses (N_D) oder laterales (N_{135}) Störgeräusch, das wie zuvor ein stationäres, sprachgefärbtes Rauschen war.

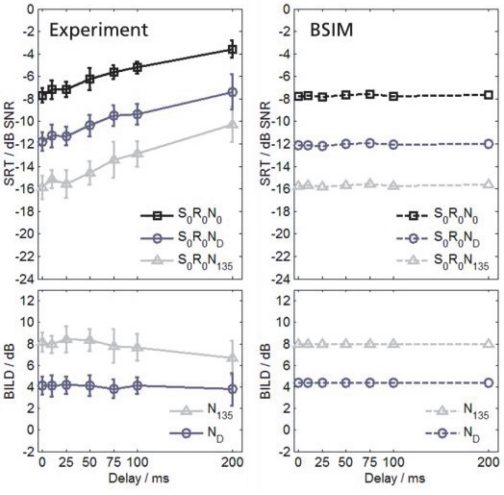

Abbildung 3: Experimentelle Daten (links) aus [7] und Vorhersagen des BSIM (rechts), entnommen aus [8].

Das linke Teilbild von Abbildung 3 zeigt die mit normalhörenden Probanden gemessenen SRT in Abhängigkeit der Verzögerung der Reflexion relativ zum Direktschall, welche zwischen 0 ms (keine Reflexion) und 200 ms variiert wurde. Die Daten für die $S_0R_0N_0$-Kondition (schwarz) zeigen einen Anstieg mit ansteigender Verzögerung, was mit dem aus monauralen Messungen bekannten, schädlichen Effekt eines späten Echos übereinstimmt. Interessanterweise ergibt sich für diffuses (dunkelgrau) und laterales (hellgrau) Störgeräusch eine parallele Verschiebung der Schwellen nach unten, d.h., der Gewinn durch die binaurale Verarbeitung ist für eine frontale Reflexion unabhängig von der zeitlichen Integration der Sprachinformation der Reflexion (siehe binaural intelligibility difference, BILD, im unteren linken Teilbild).
Die Vorhersagen der unmodifizierten Version des BSIM sind in den rechten Teilbildern von Abbildung 3 dargestellt. Wie erwartet wird der binaurale Gewinn korrekt vorhergesagt (vergleiche Verzögerung von 0 ms). Jedoch wird der Anstieg der Schwellen mit steigender Reflexionsverzögerung nicht vorhergesagt.

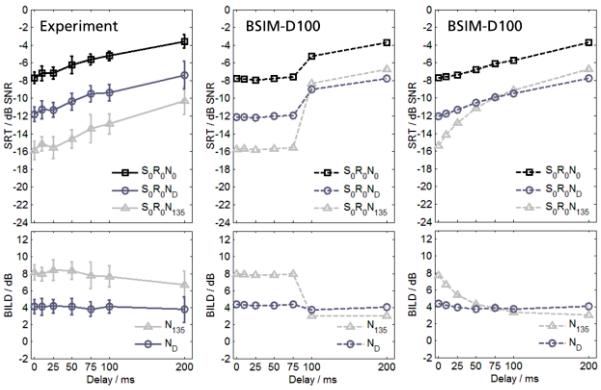

Abbildung 4: Experimentelle Daten (links) aus [7] und Vorhersagen vom mit einer post-hoc Korrektur erweiterten BSIM (Mitte und rechts), entnommen aus [8]. In der Mitte wurde eine scharfe Trennung zwischen nützlichen und schädlichen Sprachanteilen vorgenommen, im rechten Teilbild wurde dieser Übergang linear gewichtet [8].

Die Vorhersagen des durch post-hoc Korrektur erweiterten BSIM (BSIM-D100) sind in Abbildung 4 dargestellt (Mitte). Es zeigt sich eine nicht mit dem Experiment übereinstimmende, sprunghafte Abhängigkeit von der Reflexionsverzögerung. Dies ist verständlich, da ja das Deutlichkeitsmaß eine Verzögerung entweder als nützlich (bei Verzögerungen kleiner als $te = 100$ ms) oder als schädlich (bei Verzögerungen kleiner als $te = 100$ ms) modelliert wird. Führt man einen linear gewichteten Übergang zwischen nützlich und schädlich ein (rechtes Teilbild in Abbildung 4, Details siehe [7]), verschwindet das sprunghafte Verhalten und die Vorhersagen für N_0 und N_D stimmen gut mit den experimentellen Daten überein. Für das laterale Störgeräusch (N_{135}) wird jedoch der Einfluss der Reflexionsverzögerung stark überschätzt. Im Gegensatz hierzu werden die beiden wesentlichen Effekte, d.h., der Anstieg der Schwellen mit steigender Reflexionsverzögerung und die parallele Verschiebung der Schwellen für N_D und N_{135} korrekt vorhergesagt (Abbildung 5).

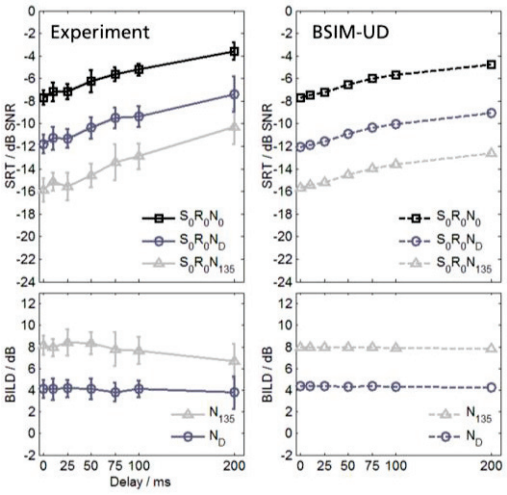

Abbildung 5: Experimentelle Daten (links) aus [7] und Vorhersagen des erweiterten BSIM-UD (rechts), entnommen aus [8].

Diskussion und Anwendungsbeispiele

Die der Sprachverarbeitung zugrunde liegende binaurale Verarbeitung ist ein hocheffizienter, periphärer Prozess, der auf eine optimale Nutzung der frühen / nützlichen Sprachanteile abzielt. Dabei interagieren räumliche und zeitliche Verarbeitung miteinander. Durch systematische Variation der Ebene, auf der eine Trennung von nützlichen und schädlichen Sprachanteilen vorgenommen wird, ist es möglich zu untersuchen, an welcher Stelle das auditorische System die entsprechenden Informationen verarbeitet. Hierbei zeigt sich, dass ein Mechanismus, der bereits in der binauralen Verarbeitung die frühen von den späten Sprachanteilen trennt, einer effektiven post-hoc Korrektur überlegen ist. Weiterführende Studien zeigen, dass die räumlich-zeitliche Verarbeitung bei normalhörenden Nicht-Muttersprachlern („normale periphäre Verarbeitung, kognitive Beeinträchtigung“) identisch der bei normalhörenden Muttersprachlern ist [9] (es stellt sich lediglich eine generelle Erhöhung aller Schwellen ein, unabhängig von der zeitlich-räumlichen Konfiguration). Im Gegensatz dazu sind bei schwerhörenden Muttersprachlern („pathologische periphäre Verarbeitung, normale kognitive Verarbeitung“) sowohl die zeitliche als auch die periphäre Verarbeitung beeinträchtigt [9]. All dies deutet darauf hin, dass die binaurale Verarbeitung auf einer sehr tiefen Ebene der Hörbahn stattfindet.

Die Unterschiede zwischen den beiden Varianten des BSIM-D100 (Abbildung 4) zeigen, dass eine scharfe Trennung zwischen „nützlich“ und „schädlich“, wie sie standardmäßig in der Raumakustik angewendet wird, in Situationen mit starken Echos nicht haltbar ist.

Die hohe quantitative Vorhersagegenauigkeit des erweiterten BSIM-UD100 macht das Modell attraktiv für praktische Anwendungen, in denen eine instrumentelle Bewertung von Sprachverständlichkeit hilfreich ist, wie z.B. spatial conferencing systems, Entwicklung von mehrkanaligen Signalverbesserungsalgorithmen oder Maskiersysteme zur Verbesserung der Privatsphäre.

Adaptive Verbesserung von Sprachverständlichkeit durch Vorverarbeitung

Anwendungsszenario und Algorithmenschemata

In vielen Anwendungen ist es wünschenswert Sprachverständlichkeit nicht nur quantitativ zu bestimmen, sondern auch zu verbessern. Ein Beispiel hierfür sind Durchsagesysteme. Hier liegt das wiederzugebende Sprachsignal häufig in ungestörter Form vor (z.B. als Audiokonserve einer Bahnhofsdurchsage), während das Störgeräusch am Ort des Empfängers (z.B. am Bahngleis) kaum vorhersehbar, stark instationär und nicht beeinflussbar ist. In solchen Anwendungsszenarien lässt sich Verständlichkeit dadurch verbessern, dass das Sprachsignal vor der Wiedergabe vorverarbeitet wird. Dies sind die sogenannten „near-end listening enhancement (NELE)“ Algorithmen, die in diesem Abschnitt behandelt werden sollen.

Einige der in der Literatur vorgeschlagenen NELE Algorithmen, wie der in [10] vorgeschlagene AdaptDRC Algorithmus, passen sich dabei adaptiv an die aktuelle Hörsituation an. Das Schema des Algorithmus ist in Abbildung 6 dargestellt.

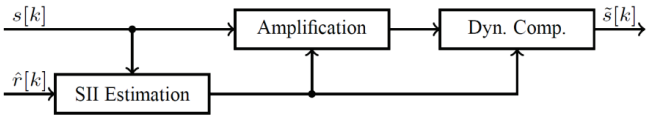

Abbildung 6: Schema des AdaptDRC Algorithmus.

Das Sprachsignal wird durch lineare, frequenzabhängige Verstärkung und Dynamikkompression verarbeitet. Der Grad der Signalmodifikation wird dabei gesteuert durch Kurzzeitschätzungen des SII, wofür eine Schätzung des beim Empfänger vorherrschenden Störgeräusches vorliegen muss. Hierdurch wird erreicht, dass der Algorithmus nur dann in das Signal eingreift, wenn die Verständlichkeit schlecht ist. Für kleine Werte des SII werden eine progressive Verstärkung von höheren Frequenzen sowie eine steigende Kompression der Signaldynamik vorgenommen. Der AdaptDRC Algorithmus arbeitet dabei unter der Randbedingung, dass der rms-Pegel des Sprachsignals nicht verändert werden darf (Details siehe [10]). Eine in [11] vorgeschlagene Erweiterung des Algorithmus (AdaptDRCplus) erlaubt hingegen eine adaptive Erhöhung des rms-Pegels des Sprachsignals um bis zu 6,5 dB. Bei dieser Algorithmenvariante darf jedoch der sampleweise Spitzenpegel nicht erhöht werden, was durch hartes Peak Clipping erreicht wird. Dies bedeutet, dass die Erhöhung des des rms-Pegels auf Kosten von nichtlinearen Signalverzerrungen erfolgt (Details siehe [11]).

Experimentelle Validierung

Um die Wirksamkeit der Algorithmen zu prüfen, wurden Sprachverständlichkeitsuntersuchungen mit normalhörenden Probanden durchgeführt. Bei einem festen Sprachpegel von 60 dB SPL wurden bei verschiedenen SNR Wort-erkennungsraten gemessen. Abbildung 7 zeigt Messdaten aus [10] (grau) und [11] schwarz für ein Störgeräusch aus einem Fahrzeuginnenraum (links) und einer Cafeteria (rechts).

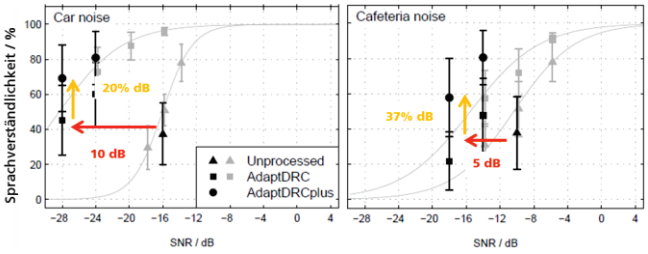

Abbildung 7: Sprachverständlichkeit als Funktion des SNR für ein Fahrgeräusch und ein Cafeteria-Szenario. Dreiecke stellen die unverarbeitete Situation dar. Quadrate und Kreise repräsentieren Ergebnisse für den AdaptDRC und AdaptDRCplus Algorithmus.

In beiden Studien zeigte sich eine erhebliche Verbesserung der Sprachverständlichkeit durch die Vorverarbeitung mit dem AdaptDRC Algorithmus (Dreiecke vs. Quadrate in Abbildung 7). Die äquivalente SNR-Verbesserung betrug abhängig vom Störgeräusch zwischen 5 dB (Cafeteria) und 10 dB (Fahrgeräusch). Die adaptive Pegelerhöhung bewirkte eine weitere Verbesserung der Worterkennungsraten (Quadrate vs. Kreise) von 20-37%.

Diskussion und Anwendungsbeispiele

Die im AdaptDRC Algorithmus realisierte, SII-gesteuerte Verstärkung und Dynamikkompression kann in gegebenen Störgeräuschumgebungen zu erheblichen Verbesserungen der Verständlichkeit bei normalhörenden Probanden führen, ohne dass dabei der rms-Pegel des Sprachsignals erhöht wird. Die adaptive Verarbeitung reagiert dabei auf zeitliche Änderungen von Störgeräuschen und ermöglicht so einen guten Kompromiss zwischen Verständlichkeit und Natürlichkeit, da nur dann in das Signal eingegriffen wird, wenn es nach Schätzung des SII notwendig ist. Dabei ist der erreichbare Nutzen stark von der Art des Störgeräusches abhängig. In der Praxis besteht nicht notwendigerweise der Anlass für eine (eher akademische) Beschränkung des Ausgangspegels auf den Eingangspegel. Lässt man zusätzlich zur AdaptDRC Verarbeitung noch eine ebenfalls SII-gesteuerte, moderate Pegelerhöhung zu, ohne dabei den Spitzenpegel zu erhöhen, so steigt die Sprachverständlichkeit weiter signifikant an. Erste Studien zeigen, dass auch schwerhörende Probanden von den vorgestellten NELE-Algorithmen profitieren können, jedoch treten hier starke interindividuelle Schwankungen und ein im Mittel kleinerer Gewinn auf [12].

NELE-Verfahren haben vielseitige Einsatzmöglichkeiten. Neben den erwähnten Durchsagesystemen für Bahnhöfe, Flughäfen oder öffentliche Verkehrsmittel sind sie z.B. auch für Kommunikationssysteme im Automotive-Bereich (Freisprechen, Navigation, In-Car-Communication), in Telefonsystemen oder in Hörgeräten einsetzbar.

Danksagung

Die in diesem Beitrag vorgestellten Arbeiten wurden unterstützt durch DFG SFB TRR 31 und das Niedersächsische Ministerium für Wirtschaft und Kultur. Großer Dank gilt Birger Kollmeier, Jesko Verhey, Jens Appell, Thomas Brand, Anna Warzybok, Henning Schepker, Simon Doclo, David Hülsmeier, Jakob Drefs und vielen weiteren Weggenossen für die Unterstützung und großartige Zusammenarbeit.

Literatur

[1] Beutelmann, R. und Brand, T. (2006). „Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 120, 331-342.

[2] Durlach, N. I. (1963). „Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. 35, 1206-1218.

[3] Durlach, N. I. (1972). „Binaural signal detection: Equalization and cancellation theory," in Foundations of Modern Auditory Theory, edited by J.Tobias (Academic, New York), Vol. II, pp. 371-462.

[4] Beutelmann, R., Brand, T. und Kollmeier, B. (2010). „Revision, extension, and evaluation of a binaural speech intelligibility model," J. Acoust. Soc. Am. 127, 2479-2497.

[5] Rennies, J., Brand, T. und Kollmeier, B. (2011). „Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," J. Acoust. Soc. Am. 130, 2999–3012.

[6] ANSI (1997). „Methods for the Calculation of the Speech Intelligibility Index," Standards Secretariat, Acoustical Society of America, American National Standard, S3.5-1997.

[7] Warzybok, A., Rennies, J., Brand, T., Doclo, S. und Kollmeier, B. (2013). „Effects of spatial and temporal integration of a single early reflection on speech intelligibility," J. Acoust. Soc. Am. 133, 269-282.

[8] Rennies, J., Warzybok, A., Brand, T. und Kollmeier, B. (2014). „Modeling the effects of a single reflection on binaural speech intelligibility," J. Acoust. Soc. Am. 135, 1556-1567.

[9] Warzybok, A. (2012). „The combined effects of binaural hearing and reverberation on speech intelligibility in noise," Dissertation, Universität Oldenburg.

[10] Schepker, H., Rennies, J. und Doclo, S. (2015). „Speech-in-noise enhancement using a two-stage amplification and dynamic range compression algorithm controlled by the speech intelligibility index," J. Acoust. Soc. Am., 138, 2692-2706.

[11] Drefs, J., Rennies, J., Schepker, H. und Doclo, S. (2015). „Weiterentwicklung und Evaluation eines Algorithmus zur SII-basierten Sprachverständlichkeits-verbesserung in störgeräuschbehafteter Umgebung," Fortschritte der Akustik - DAGA 2015, Nürnberg, 1031-1034.

[12] Hülsmeier, Rennies, Drefs, Schepker, Doclo (2015). „Evaluation eines Algorithmus zur SII-basierten Sprachverständlichkeitsverbesserung in störgeräusch-behafteter Umgebung mit schwerhörenden Probanden," Fortschritte der Akustik - DAGA 2015, Nürnberg, 154-157.