# Vergleich von Messungen und Modellvorhersagen zur Sprachverständlichkeit in verschiedenen Störgeräuschsituationen

Wiebke Schubotz[1,2], Thomas Brand[1,2], Stephan D. Ewert,[1,2]
[1]*Carl von Ossietzky Universität Oldenburg*
[2]*„Exzellenzcluster Hearing4all“, E-Mail: wiebke.schubotz@uni-oldenburg.de*

## Einleitung

Sprache ist für Menschen ein zentrales Kommunikationsmittel, jedoch wird die Verständlichkeit eines Sprachsignals oft durch Störgeräusche (Maskierer) behindert. Je nach Situation weisen Maskierer unterschiedliche physikalische Eigenschaften (z.B. Frequenzgehalt, Amplitudenmodulationen, zeitliche Lücken) auf. In Bezug auf Sprachverständlichkeit (SV) können dabei drei Kategorien unterschieden werden: Energetische Maskierung (EM, [1]), Amplitudenmodulationsmaskierung (AM, [2]) sowie „informational masking“ (IM, [3]). Energetische Maskierung tritt auf, wenn Energie des Maskierers in denselben auditorischen Filter fällt, in dem auch Energie des Zielsignals liegt und diese verdeckt. Amplitudenmodulationsmaskierung tritt auf, wenn Modulationen des Maskierers Modulationen des Zielsignals innerhalb eines auditorischen Filters verdecken. Informational masking kommt zum Tragen, wenn die Information des Maskierers die Information des Zielsignals maskiert. Im Falle von Sprachverständlichkeitsmessungen tritt dies insbesondere auf, wenn Sprache als Maskierer genutzt wird (speech-on-speech masking). Um die Wirkung einzelner Maskiereigenschaften zu untersuchen und zu quantifizieren, wurden Messungen mit normalhörenden Probanden zur Sprachverständlichkeit und Sprachdetektion in verschiedenen Störgeräuschsituationen durchgeführt. Die Störgeräusche wurden dabei so generiert, dass sie gezielt einzelne Maskiereffekte beinhalteten. Darüber hinaus wurden vier Sprachverständlichkeitsmodelle betrachtet, um die gewonnenen Messergebnisse mit objektiven Vorhersagen der Modelle vergleichen zu können.

## Methode

Die präsentierten Stimuli setzten sich aus Sprachsignalen des Oldenburger Satztests OLSA [4] und sechs verschiedenen Störgeräuschen zusammen, die bei unterschiedlichen Signal-Rausch-Abständen (SNR) präsentiert wurden. EM wurde durch ein stationäres Rauschen (SSN) vermittelt, dessen Spektrum mit dem des International Speech Test Signal (ISTS, [5]) übereinstimmte. Durch Aufprägen verschiedener Einhüllenden auf das SSN wurden sukzessive Modulationen eingeführt. Die Einhüllenden waren ein 8-Hz Sinuston (SAM-SSN), sowie die Hilberteinhüllende eines breitbandigen Sprachsignals (BB-SSN). Die Modulationen wurden dabei auf den gesamten Bereich des SSN-Spektrums angewendet (kohärente Modulationen). Desweiteren wurde eine Maskiersituation generiert, in der kohärente Modulationen nur in bestimmten Frequenzbereichen auftraten (across-frequency shifted SSN, AFS- SSN). Dafür wurde das SSN in einem Bereich von 50Hz-12kHz in 32 Frequenzkanäle gefiltert und jeweils auf 4 Kanäle dieselbe Einhüllende aufgeprägt. Informational masking wurde durch zwei sprachähnliche Störgeräusche mit intakter Sprache eingeführt, durch das ISTS [5], welches sich aus sechs weiblichen Störsprecherinnen verschiedener Sprachen zusammensetzt und einer einzelnen weiblichen Störsprecherin (single talker, ST). In den SV-Messungen wurde in einem adaptiven Verfahren die SNR-Schwelle bestimmt, bei der 50% und 80% ($SRT_{50\%}$, $SRT_{80\%}$) der dargebotenen Sprache korrekt verstanden wurden. Bei der Detektion von Sprache wurde ein 1up-2down 2-alternative forced choice Verfahren verwendet, um die Schwelle auf der psychometrischen Funktion zu bestimmen, an der in 70.7% der Stimuli Sprache im Rauschen detektiert wurde. Als objektive Maße wurden vier Sprachverständlichkeitsmodelle betrachtet, (Speech Intelligibility Index (SII), extended SII (ESII), multi-resolution Speech Envelope Power Spectrum Model (mr-sEPSM), Short-Time Objective Intelligibility Measure (STOI)), die sich hinsichtlich der Betrachtung von zeitlichen Lücken innerhalb der Signale und der Analyse von Modulationen zum Teil stark unterschieden, siehe [6]-[9] für detaillierte Beschreibungen. Den Modellen ist gemein, dass es eine SNR-basierte Verarbeitung der Signale gibt, also speziell energetische Maskierung betrachtet wird, IM hingegen nicht berücksichtigt wird. Es ist zu erwähnen, dass der ESII das Spektrum des Maskierers in kurzen Zeitfenstern berechnet, also zeitliche Lücken beachtet werden, wohingegen der SII nur das Langzeitspektrum des Maskierers betrachtet. Das mr-sEPSM Modell betrachtet neben der Frequenz- auch die Modulationsebene eines Signals und nutzt die Gesamtleistung beider, um die Sprachverständlichkeit von Signalen im Störgeräusch vorherzusagen. Eine Analyse des Modulationsbereiches wird in keinem anderen der Modelle durchgeführt. Das STOI Modell analysiert Sprachsignale und maskierte Sprachsignale und berechnet einen Korrelationskoeffizienten zwischen beiden, anhand dessen Aussagen zur Sprachverständlichkeit getroffen werden können.

## Ergebnisse und Diskussion

Abb.1 zeigt die Messergebnisse der Sprachverständlichkeits- und Detektionsschwellenmessungen in den verschiedenen Störgeräuschsituationen. Der Verlauf der $SRT_{50\%}$ und $SRT_{80\%}$ Kurven ist ähnlich, jedoch liegen die $SRT_{50\%}$ Schwellen insgesamt niedriger. Die höchsten Maskierschwellen treten für SSN und AFS-SSN auf, sie sinken jedoch, wenn dem Maskierer kohärente Modulationen aufgeprägt werden. Für diese Konditionen liegen die Schwellen etwa 5 dB unterhalb der des stationären Störgeräusches. Die Schwellen der

sprachähnlichen Maskierer (ISTS, ST) liegen ebenfalls in diesem Bereich, wobei die Anzahl der Störsprecher nur einen geringfügigen Einfluss auf die Lage der Schwellen hat. Die Sprachdetektionsschwellen liegen bis zu 20 dB unterhalb der Sprachverständlichkeitschwellen, aber auch hier ist der Kurvenverlauf ähnlich: Die größte Maskierwirkung wird durch SSN und AFS-SSN hervorgerufen, wird aber geringer, wenn Modulationen eingeführt werden.

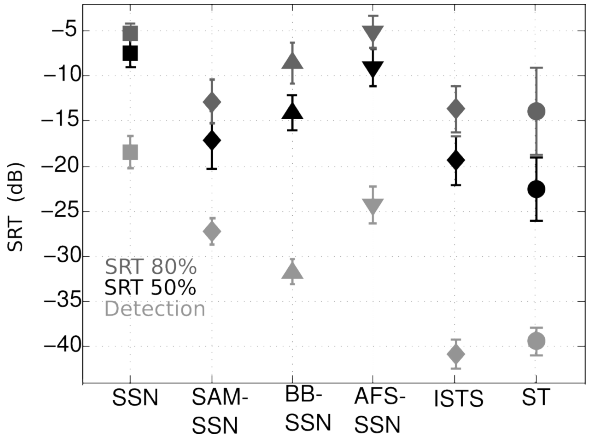


**Abbildung 1:** Gemessene Sprachverständlichkeits- und Detektionsschwellen mit dazugehörigen Standardabweichungen in den verschiedenen Maskiersituationen. Es zeigt sich für alle Messungen ein ähnlicher Kurvenverlauf, wobei die SSN Kondition jeweils die größte Maskierwirkung hat. Mit zunehmender Kohärenz der Modulationen (AFS-, BB-, und SAM-SSN, siehe Abschnitt Methode) sinken die Schwellen. Störsprecher (ISTS, ST) senken die SV-Schwellen ebenfalls.

## Modellvorhersagen

Abb.2 zeigt die SV-Vorhersagen der Modelle im Vergleich zu den $SRT_{50\%}$ Schwellen der Probandenmessungen (human data). Es zeigt sich, dass die Modelle weder den Verlauf der Kurven, noch die Schwellen exakt vorhersagen können. SII, mr-sEPSM und STOI sagen generell zu hohe Schwellen vorher, wobei Amplitudenmodulationen keinen Einfluss haben (SII) oder als störend betrachtet werden und deshalb die vorhergesagten Schwellen steigen (mr-sEPSM, STOI). Die besten Voraussagen liefert das ESII- Modell, es sagt die SAM-SSN Schwelle exakt vorher und beschreibt das Sinken der SRTs mit zunehmender Modulation des Maskierers. Die SV-Schwellen der sprachähnlichen Maskierer werden beim ESII jedoch unterschätzt, was durch die fehlende Betrachtung von IM erklärt werden kann.

## Zusammenfassung

Wenn Sprache im Störgeräusch präsentiert wird, können verschiedene Maskiereigenschaften zur Verdeckung von Sprache führen. Die Messungen haben gezeigt, dass die größte Maskierwirkung bei Sprachverständlichkeit und Sprachdetektion durch energetische Maskierung hervorgerufen wird. Aufgeprägte Amplitudenmodulationen senken die Schwellen, wobei diese tiefer liegen, je kohärenter die Modulationen im Spektrum des Maskierers sind. Für sprachähnliche Maskierer sinken die Schwellen ebenfalls, es spielt jedoch keine Rolle, ob es einen oder mehrere Störsprecher gibt. Existierende SV-Modelle sagen die gemessen $SRT_{50\%}$ Schwellen nur qualitativ vorher. Der Einfluss von Amplitudenmodulationen und Ähnlichkeit von Maskierer und Zielsignal wird zum Teil sehr stark überschätzt.

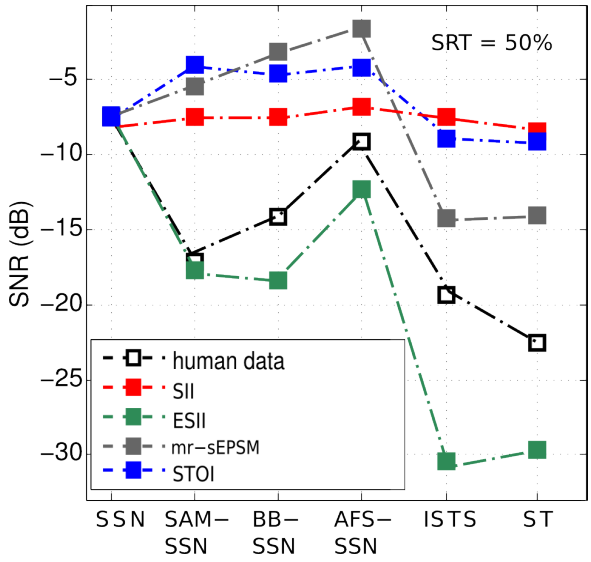


**Abbildung 2:** Modellvorhersagen der $SRT_{50\%}$ Schwelle für die verschiedenen Maskiersituationen. Generell weichen die Vorhersagen stark von den gemessenen Schwellen ab, die SRTs werden oft zu hoch vorhergesagt. Die besten Vor-hersagen macht ESII.

## Literatur

[1] Durlach, N.I., Mason, C.R., Kidd, G. Jr., Arbogast, T.L., Colburn, H.S., Shinn-Cunningham, B.G.(2003).“Note on informational masking“, J. Acoust. Soc. Am, 113, 2984-2987

[2] Houtgast, T. (1989).“Frequency selectivity in amplitude-modulation detection”, Acoust. Soc. Am. 85(4), 1676-1680

[3] Stone, M. A., Füllgrabe, C., and Moore, B.C.J.(2012). „Notionally steady background noise act primarily as a modulation masker of speech“, J. Acoust. Soc. Am. 132(1), 317-326

[4] Wagener, K. und Brand, T. und Kollmeier, B.(1999). “Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests.“, Zeitschrift für Audiologie, 38:86-95

[5] Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (2010).“Development and analysis of an International speech Test Signal (ISTS).”, Int. J. Audiol. 49, 891-903

[6] ANSI (1997). ANSI S3.5-1997, “American National Standard Methods for Calculation of the Speech Intelligibility Index” (American National Standards Institute, New York).

[7] Rhebergen, K.S., Versfeld, N.J., and Dreschler, W.A. (2006), “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise”, J. Acoust. Soc. Am. 120, 3988-3997

[8] Jørgensen, S., Ewert, S.D., and Dau, T.(2013).“A multi-resolution envelope-power based model for speech intelligibility.“, J. Acoust. Soc. Am. 134, 436-446

[9] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011).“An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech.“, IEEE, Vol.19, No.7, 2125-2136