Modellierung der Sprachverständlichkeit mit einem auditorischen Perzeptionsmodell

Tim Jürgens, Thomas Brand, Birger Kollmeier
Medizinische Physik, Universität Oldenburg
tim.juergens@uni-oldenburg.de

Einleitung

Störgeräusche führen zu einer Verminderung der Sprachverständlichkeit. Modelle, die versuchen Sprachverständlichkeit in rauschbehafteten Umgebungen vorherzusagen (wie z.B. der Speech Intelligibility Index), arbeiten meist auf der Grundlage von Langzeitspektren. Sie werten Rauschen und Sprachsignal getrennt voneinander bezüglich der Leistung in unterschiedlichen Frequenzkanälen aus und berechnen typischerweise die Verständlichkeit in Abhängigkeit vom Signal-Rausch-Verhältnis (signal-to-noise-ratio, SNR). Dies erlaubt keine detaillierte Betrachtung der Sprachverständlichkeit, die eine Vorhersage von Verwechslungen einzelner Phoneme einschließt.

Das hier vorgestellte Modell soll dies ermöglichen und beruht auf einem Ansatz von Holube und Kollmeier [1]. Es besteht aus einem auditorischen Perzeptionsmodell [2] und einem Dynamic-Time-Warp (DTW) Spracherkenner. Als Sprachmaterial diente der Oldenburger Logatom-Sprachkorpus (OLLO, [3]), der aus systematisch zusammengesetzten Phonemen besteht. Zusammen mit diesem Modell können so Trefferquoten und Verwechslungen für Vokale und Konsonanten bei unterschiedlichen SNR vorhergesagt werden. Die Vorhersagen wurden mit Sprachverständlichkeitstests an 10 Normalhörenden mit dem gleichen Sprachmaterial validiert.

Messungen

Es wurde die Verständlichkeit von 150 unterschiedlichen Logatomen an 10 klinisch normalhörenden Versuchspersonen bei 5 unterschiedlichen SNR gemessen. Dabei wurden Logatomaufnahmen des OLLO von demselben Sprecher mit normaler, ruhiger Sprachartikulation verwendet.
Die Darbietung erfolgte über Kopfhörer in einer Standard-Hörkabine. Der Sprachpegel betrug bei allen Messungen 60 dB SPL. Die Äußerungen wurden mit sprachähnlichem Rauschen (ICRA-1-Rauschen, [4]) bei unterschiedlichen SNR diotisch abgespielt. Der Test wurde geschlossen durchgeführt, d.h., das erkannte Logatom wurde von der Versuchsperson über eine Eingabemaske aus einer Auswahl aus 10 Konsonant-Vokal-Konsonant-Äußerungen (CVC) bzw. 14 VCV-Äußerungen ausgewählt. Es wurden nur Logatome als Antwortalternativen angeboten, die in den beiden Außenphonemen gleich zum akustisch dargebotenen Logatom waren und im Mittelphonem variierten.

Messergebnisse

Abb. 1 zeigt die Messergebnisse. Grundsätzlich werden CVCs besser erkannt als VCVs außer bei -20 dB SNR. Die Verständlichkeitsfunktion weist eine Sprachverständlichkeitsschwelle (Speech Reception Threshold, SRT) von (−12,2 ± 1,1) dB und eine maximale Steigung von (5,4 ± 0,6) %/dB auf.
Die detaillierte Auswertung der Phonemerkennung bei einer mittleren Verständlichkeit von etwa 50% ergab, dass Phoneme wie „s“, „ts“, und „ʃ“ mit nahezu 100% am besten erkannt wurden, wohingegen z.B. „v“ und „n“ fast nicht verstanden wurden. Für die Vokale ergab sich eine Dreiteilung: „ɔ“, „ʊ“, „o“ und „u“ wurden am schlechtesten erkannt und oft miteinander verwechselt. „a“ und „a:“ wurden nur untereinander signifikant verwechselt und wiesen mittlere Erkennungsraten auf und „ɛ“, „ɪ“, „e“ und „i“ wurden am besten erkannt.

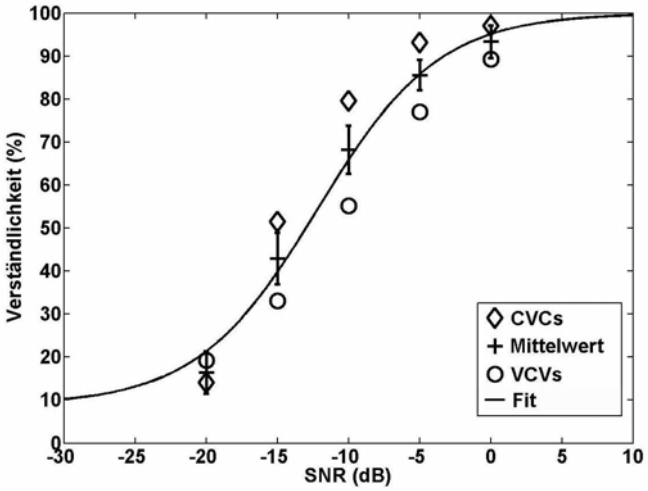

Abbildung 1: Verständlichkeitsfunktion für 10 Normalhörende, gemessen mit Logatomen in ICRA-1-Rauschen. Die Fehlerbalken zeigen die interindividuelle Standardabweichung.

Modellierung

Abb. 2 zeigt die Modellstruktur. Aus einem Zeitsignal wird durch das auditorische Perzeptionsmodell [2] eine „interne Repräsentation“ (IR) berechnet. Dabei wird ein spektral geformtes hörschwellensimulierendes Rauschen auf das Zeitsignal addiert und dieses in einer Gammatonfilterbank mit 27 Frequenzkanälen gefiltert. Darauf folgt ein Haarzellenmodell, das im Wesentlichen aus einer Halbwellengleichrichtung und Tiefpassfilterung besteht, fünf Adaptationsschleifen, die eine zeitliche Adaptation des Signals bewirken und eine Modulationsfilterbank mit vier Modulationsfrequenzkanälen. Aus dem so vorverarbeiteten Referenz-Sprachsignal wird mit einem DTW-Spracherkenner der „perzeptive Abstand“ zu einem in der gleichen Art und Weise vorverarbeiteten Sprachsignal des Vokabulars des Spracherkenners berechnet. Dasjenige Logatom mit dem geringsten Abstand zur Referenz wird dann vom Modell „erkannt“. Zur Erkennung zugelassen

werden nur diejenigen Logatome des geschlossenen Tests, die auch dem Menschen in den Messungen über die Eingabemaske angeboten wurden.

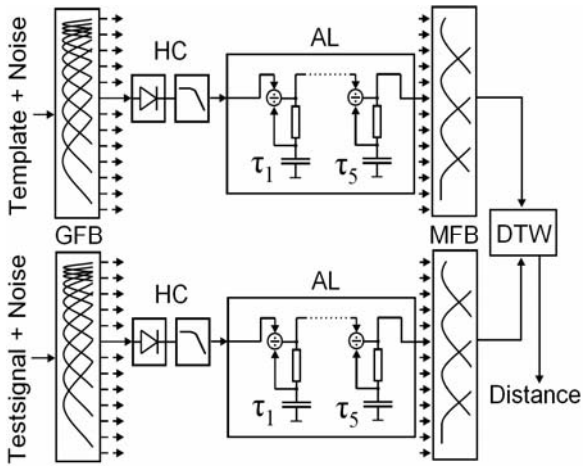

Abbildung 2: Schematische Darstellung des Modells. Das Modell berechnet die Distanz zwischen den in derselben Art und Weise vorverarbeiteten Zeitsignalen von Template (Teil des Vokabulars) und Testsignal. GFB: Gammatonfilterbank, HC: Haarzellenmodell, AL: Adaptationsschleifen, MFB: Modulationsfilterbank, DTW: Dynamic-Time-Warp-Spracherkenner

In der hier vorgestellten Modellvariation ist die Referenzaufnahme, die für die Spracherkennung verwendet wurde, identisch zur Testaufnahme. Dadurch, dass unterschiedliche Rauschsequenzen verwendet wurden, unterscheiden sich dennoch die jeweiligen IR. So wird zunächst die natürliche Variabilität von Sprache aus der Modellierung ausgeklammert und nur der Einfluss des Rauschens auf die Sprachverständlichkeit untersucht.

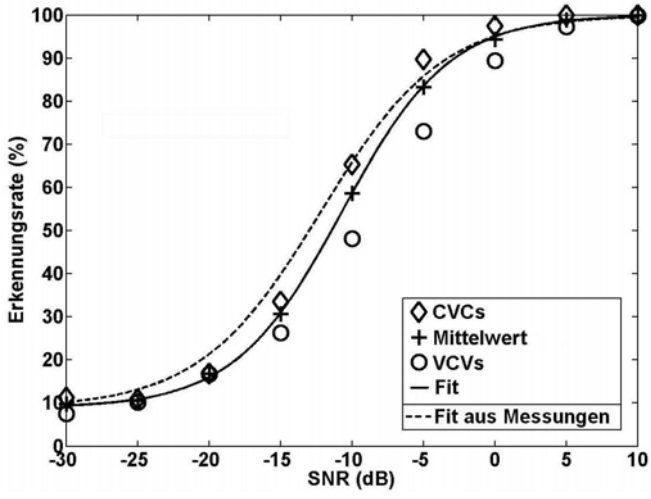

Abbildung 3: Modellierte Verständlichkeitsfunktion berechnet mit Logatom-Aufnahmen in ICRA-1-Rauschen. Zum Vergleich ist der Fit an die psychometrische Funktion von Normalhörenden aus Abb.1 eingefügt.

Abb.3 zeigt die Ergebnisse der Modellierung im Vergleich zu den über die Versuchspersonen gemittelten Messergebnissen. Es ist festzustellen, dass die Reihenfolge der Verständlichkeitsfunktionen von CVCs und VCVs, genauso wie die Steigung der psychometrischen Funktion mit 6,0 %/dB korrekt vorhergesagt werden kann. Die SRT des Modells liegt mit -10,7 dB SNR im Rahmen der beim Menschen gemessenen interindividuellen Standardabweichung. Eine detaillierte Auswertung der Verwechslungen für Konsonanten ergab, dass ebenfalls die Phoneme „s“, „ts“, und „ʃ“ vom Modell am besten erkannt wurden, allerdings mit 70 bis 76 % mit weit niedrigeren Trefferquoten als die, die der Mensch gezeigt hat. Andererseits wurden „v“ und „n“ vom Modell weitaus besser erkannt als vom Menschen. Die beim Menschen beobachtete dreiteilige Gruppierung in der Erkennung der Vokale konnte mit dem Modell nicht nachgebildet werden. Insgesamt weisen die Trefferquoten des Modells für die Konsonanten zwar große Ähnlichkeiten auf, allerdings zeigen sie keine so große Differenz, wie sie zwischen den am besten und am schlechtesten erkannten Phonemen beim Menschen bestehen.
Dies könnte ein Hinweis darauf sein, dass hochfrequente Anteile bei der menschlichen Spracherkennung besser genutzt werden können als bisher im Modell berücksichtigt.
In einer Modellvariation mit Berücksichtigung der natürlichen Sprachvariabilität wurde eine Verschiebung der psychometrischen Funktion um 12 dB hin zu höheren SNR beobachtet.

Zusammenfassung

Das hier vorgestellte Modell ist in der Lage die Sprachverständlichkeit von Nonsensäußerungen für Normalhörende im sprachähnlichen Rauschen vorherzusagen. Dies gelingt allerdings nur, wenn zum Training und zur Erkennung identische Sprachaufnahmen verwendet werden. Ein Vergleich der Antwortraten für einzelne Phoneme ergab große Ähnlichkeiten bei der Erkennung der Konsonanten. Das Modell zeigt aber nicht so große Differenzen der Trefferquoten zwischen den am besten und am schlechtesten erkannten Phonemen wie in den Sprachverständlichkeitstests beobachtet wurde.

Danksagung

Wir danken dem HearCom Projekt und SFB/TR 31 ‘Das aktive Gehör’ (URL: http://www.uni-oldenburg.de/sfbtr31) für die Finanzierung dieser Studie.

Literatur

[1] Holube, I. and B. Kollmeier, *Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model.* J Acoust Soc Am, 1996. **100**(3): p. 1703-16.
[2] Dau, T., *Modeling auditory processing of amplitude modulation.* Journal of the Acoustical Society of America, 1997. **101**: p. 3061 (A).
[3] Wesker, T., et al., *Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines*, *Interspeech 2005.*
[4] Dreschler, W.A., et al., *ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment.* Audiology, 2001. **40**: p. 148–157.