Der Oldenburger Satztest mit synthetischer Stimme

Bianca Wiercinski, Theresa Nüsse, Inga Holube

Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg

Schlüsselwörter: Sprachverstehen, Sprachqualität, Text-To-Speech, Sprachsynthese, Oldenburger Satztest

Einleitung

Um das Sprachverstehen bei Personen mit und ohne Hörstörungen zu untersuchen, wurden bereits zahlreiche Verfahren entwickelt. Das verwendete Sprachmaterial wurde dabei entweder von ausgebildeten Sprechern oder von Laien in einem Tonstudio eingesprochen und anschließend meist aufwändig nachbearbeitet. Im vorliegenden Beitrag wurde der Frage nachgegangen, ob sich auch Sprachmaterial für die Durchführung von Sprachverständlichkeitstests eignet, das mittels Text-to-Speech (TTS) synthetisch erzeugt wurde. Dies würde den Aufwand bei der Generierung von Sprachmaterial erheblich reduzieren und eine beliebige Vergrößerung des Korpus der verwendeten Testsätze ermöglichen. Um zu untersuchen, inwieweit synthetische Sprache zu vergleichbaren Ergebnissen in der Sprachverständlichkeitsmessung führt wie die Aufzeichnung natürlicher Stimmen, wurde der Oldenburger Satztest (OLSA) mit weiblicher synthetischer Stimme evaluiert. Dafür wurde zunächst aus drei kommerziellen Produkten in einem Probandentest mit jungen Normalhörenden ein Synthesesystem mit einem möglichst natürlichen Klangbild ausgewählt. Mit dem gewählten Synthesesystem wurde der Korpus des weiblichen Oldenburger Satztests generiert. Eine nachträgliche (Pegel-)Anpassung der Wörter, wie sie bei der Entwicklung des Oldenburger Satztests angewendet wurde, erfolgte dabei nicht. Die Evaluationsmessungen wurden ebenfalls mit jungen Normalhörenden durchgeführt und orientierten sich an der Evaluation des OLSA mit natürlicher weiblicher Sprecherin (Ahrlich, 2013; Wagener et al., 2014). Den erhobenen Daten wurden anschließend Diskriminationsfunktionen angepasst sowie Unterschiede in der Listenäquivalenz und der Wortgruppenverständlichkeit analysiert.

Auswahl des TTS-Systems

Für die Auswahl eines geeigneten TTS-Systems zur Umsetzung des OLSA mit synthetischer Stimme wurden zwölf junge Normalhörende gebeten, in einem Hörversuch eine Vorauswahl aus drei TTS-Systemen zu beurteilen, indem sie testweise erzeugtes Sprachmaterial im Hinblick auf vier Attribute bewerteten: die Natürlichkeit, die Prosodie, den Sprachfluss und die Verständlichkeit sowie den Gesamteindruck. Hierfür wurde mit allen Systemen der gleiche Sprachkorpus erzeugt. Dieser bestand aus insgesamt 13 Sätzen aus verschiedenen Sprachverständlichkeitstests.

Die Bewertung der TTS-Systeme orientierte sich grob an der ITU-T Empfehlung für die Beurteilung von Sprachübertragung mittels des Mean Opinion Score (MOS; ITU-T P.800). Zur Orientierung wurden die angezeigten Skalenstufen in der Testdurchführung mit 1 = „gar nicht“ bis 5 = „vollkommen“ beschriftet. Zum Vergleich wurde zu jedem Stimulus auch die Variante mit natürlichem Sprecher dargeboten. Die Bewertung des Gesamteindrucks erfolgte separat. Hierfür wurde den Probanden eine Auswahl des aus dem vorherigen Versuchsteil bereits bekannten Sprachmaterials erneut dargeboten und durch eine Multiple-Choice-Frage der jeweilige Favorit ermittelt.

Abbildung 1 stellt die Ergebnisse des Hörversuchs für alle Probanden und alle Stimuli dar. Auch wenn die Bewertungen in den einzelnen Kategorien auf den ersten Blick nur geringe Unterschiede zwischen den drei Synthesesystemen zeigen, ist das Urteil in der Bewertung des Gesamteindrucks eindeutig: 8 von 12 Probanden entschieden sich für das Synthesesystem TTS-A, jeweils nur zwei Probanden wählten TTS-B bzw. TTS-C. Das hier mit TTS-A benannte TTS-System wurde daher erworben und damit der Sprachkorpus für die Evaluation des OLSA mit synthetischer Stimme generiert.

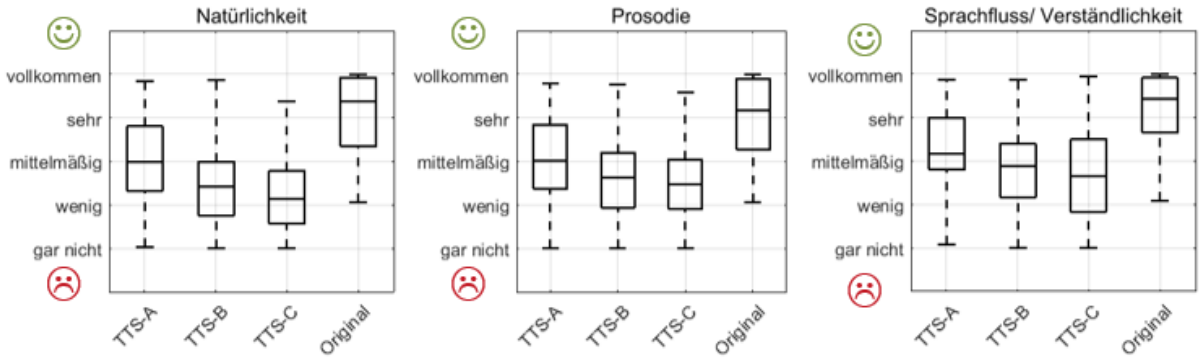

Abbildung 1: Ergebnisse des Hörversuchs zur Auswahl eines TTS-Systems mit möglichst natürlichem Klangbild. TTS-X: Text-to-Speech Systeme, Original: natürliche Stimme. Links: Natürlichkeit, Mitte: Prosodie, rechts: Sprachfluss und Verständlichkeit. Median, Quartile und Extremwerte für alle 13 Stimuli und alle zwölf Probanden

Generierung des Sprachmaterials

Der verwendete Sprachkorpus bestand aus insgesamt 150 Sätzen, die sich aus der Basismatrix des OLSA von Wagener et al. (1999) zusammensetzten. Die verwendete Synthesesoftware bot die Möglichkeit, neben Sprechgeschwindigkeit und Tonlage der Synthesestimme auch die Betonung einzelner Wörter und Satzteile anzupassen. Hierfür standen zu jedem Wort zehn Varianten mit unterschiedlicher Betonung zur Verfügung. Die Auswahl der Kombinationen mit der natürlichsten Betonung erfolgte nach subjektivem Urteil. Dabei wurden drei Kriterien berücksichtigt:

1. Vermeidung von Verzerrungen (durch Frequenzsprünge, hörbare „Schnitte“)
2. Korrekte Betonung der einzelnen Wörter (Bsp. rote Sessel)
3. Natürliche Satzmelodie

Vorab wurde die Sprechgeschwindigkeit der Synthesestimme bestmöglich an die der natürlichen Sprecherin angeglichen, indem sie soweit herabgesetzt wurde, wie es das System ohne hörbare Verzerrungen zuließ. Auf gleiche Weise wurde auch das Frequenzspektrum der Synthesestimme möglichst an die Stimme der natürlichen Sprecherin angepasst. Im Endergebnis wiesen beide Stimmen eine ähnliche Verteilung der Grundfrequenzen im Bereich von 85 bis 450 Hz auf, wobei die mediane Grundfrequenz der Synthesestimme mit 186 Hz um weniger als einen Ganzton tiefer lag als die der natürlichen Sprecherin (204 Hz). Eine ähnliche Abweichung voneinander wiesen die gemittelten Langzeitspektren der beiden Stimmen auf. Die Sprechgeschwindigkeit der natürlichen weiblichen Sprecherin betrug im Mittel etwa 167 Silben pro Minute. Im Vergleich dazu sprach die Synthesestimme mit 175 Silben pro Minute etwas schneller. Beide weiblichen Stimmvarianten wiesen eine deutlich geringere Sprechgeschwindigkeit auf als der männliche Sprecher des OLSA (208 Silben pro Minute).

Messmethode

Die Evaluationsmessungen des OLSA mit synthetischer Stimme orientierten sich an der Evaluation des OLSA mit natürlicher weiblicher Sprecherin (Ahrlich, 2013; Wagener et al., 2014). Jedoch wurde den Probanden für einen direkten Vergleich der Verständlichkeit beider Stimmvarianten während der Messungen die natürliche und die synthetische Stimme im Wechsel präsentiert. Die Messung erstreckte sich über zwei Termine, die jeweils mit einer Trainingsphase begannen. In dieser wurden zunächst die Schwellen für ein Sprachverstehen von 50 % von acht (bzw. vier im zweiten Termin) Testlisten mittels adaptiver Anpassung des Signal-Rausch-Abstands (SNR) ermittelt.

Darauf folgten die eigentlichen Evaluationsmessungen. Hierbei wurde das Sprachverstehen für zehn (bzw. 14 beim 2. Termin) Testlisten bei festem SNR bestimmt. Die Messungen wurden bei den SNR -6, -8.5 und -11 dB durchgeführt. Die Reihenfolge der SNR, sowie die der Testlisten und der präsentierten Sprecherstimmen wurde dabei vollständig randomisiert.

Als Störgeräusch wurde ein durch 30-fache Überlagerung des zeitlich verschobenen Sprachmaterials für beide Sprecherstimmen jeweils individuell erzeugtes Rauschen genutzt, dessen Langzeitspektrum dem der zugrundeliegenden Stimme entsprach. Das Rauschen wurde während aller Messungen bei einem konstanten Pegel von 65 dB SPL dargeboten. Es startete jeweils 500 ms vor dem Sprachstimulus und endete 500 ms nach dessen Ende.

An dem Hörversuch nahmen 48 junge normalhörende Probanden teil. Als Kriterium für die Normalhörigkeit diente die Definition der Norm DIN EN ISO 8253-3. Die Probanden verfügten über keinerlei Erfahrung mit dem OLSA und waren zum Zeitpunkt der Messung eingeschriebene Studierende an einer Universität oder Fachhochschule in Oldenburg. Die Messdurchführung erfolgte monaural auf dem nach gemessenem PTA-4 (gemittelter Hörverlust bei 500, 1000, 2000 und 4000 Hz) besseren Ohr.

Ergebnisse und Diskussion

Den empirischen Daten wurden zunächst für jeden Probanden und beide Sprechervarianten individuelle Diskriminationsfunktionen mit den Parametern L_{50} (SNR für ein Sprachverstehen von 50 %) und s_{50} (Steigung in diesem Punkt) angepasst. Anschließend ergab sich aus den Medianen der erhaltenen L_{50}- und s_{50}-Werte der Funktionen für beide Stimmen eine Gesamtdiskriminationsfunktion.

Für das TTS-System betrug der L_{50} im Median -8,7 dB. Die natürliche Stimme war mit einem medianen L_{50} von -9,0 dB nur geringfügig besser verständlich. Die Steigung der Gesamtdiskriminationsfunktionen im L_{50} war für beide Varianten gleich und betrug 13 %/dB. Somit konnten die Ergebnisse von Ahrlich (2013; L_{50} = -9,3 dB, s_{50} = 13 %/dB) trotz des veränderten Studiendesigns mit Sprecherwechsel und anderen in der Messung verwendeten SNRs sehr gut reproduziert werden. Auch die Messvariante mit synthetischer Stimme weicht nur unwesentlich von den Literaturwerten ab. Für die berechneten L_{50} ergab der Wilcoxon-Test dennoch einen auf dem 5%-Niveau signifikanten Unterschied für die beiden Stimmvarianten (p = 0,001). Die berechneten s_{50}-Werte unterschieden sich hingegen nicht signifikant voneinander (p = 0,47).

Die Analyse der Listenäquivalenz zeigte für beide Stimmvarianten eine ähnliche Streuung der angepassten Diskriminationsfunktionen für die einzelnen Testlisten. Die größte Abweichung zwischen zwei Testlisten betrug dabei in beiden Fällen für die gemittelten L_{50}-Werte 0,5 dB. Die s_{50}-Werte wichen zwischen den Testlisten um 1-2 %/dB voneinander ab. Darüber hinaus wurde auch die Verständlichkeit der verschiedenen Wortgruppen des OLSA (Name, Verb, Zahlwort, Adjektiv und Objekt) untersucht. Hierfür wurden die Messdaten nach Wortgruppen sortiert und erneut Diskriminationsfunktionen angepasst (siehe Abbildung 2). Hierbei wies das synthetische Sprachmaterial mit einer Spanne von 1,1 dB zwischen den ermittelten L_{50} eine geringere Streuung auf als die Variante mit natürlicher Stimme (1,9 dB).

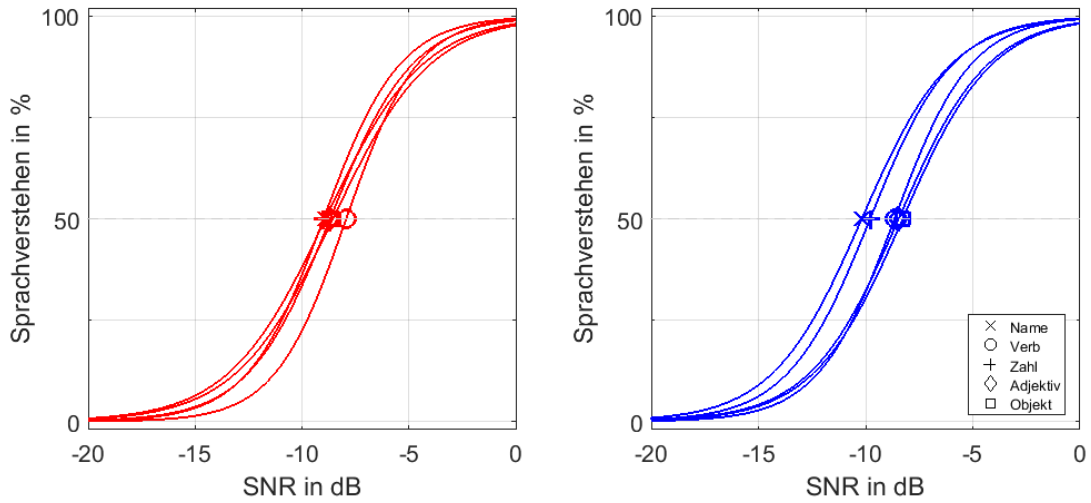

Abbildung 2: Diskriminationsfunktionen für die einzelnen Wortgruppen. Rot, links: TTS-System, Blau, rechts: natürliche Stimme. Mediane über alle gemessenen Testlisten aller Probanden

Zusammenfassung und Ausblick

In der vorliegenden Studie wurde der OLSA mit synthetischer weiblicher Stimme evaluiert, um der Frage nachzugehen, ob sich mittels TTS erzeugtes Sprachmaterial für die Durchführung von Sprachverständlichkeitstests eignet. In einem Hörversuch wurde dafür das Sprachmaterial der Synthesestimme und der natürlichen weiblichen Stimme listenweise im Wechsel präsentiert. An die erhobenen Daten wurden Diskriminationsfunktionen angepasst. Die medianen L_{50} der beiden Gesamtdiskriminationsfunktionen unterschieden sich dabei um 0,3 dB; die Steigung beider Kurven war mit 13 %/dB identisch. Auch die Listenäquivalenz lieferte vergleichbare Werte für beide Stimmvarianten. Für die einzelnen Wortgruppen ergab sich für die Sprachsynthese eine geringere Streuung als für die natürliche Stimme.

Synthetische Sprache eignet sich demnach für die Verwendung in der Sprachaudiometrie. Durch die Generierung des Sprachmaterials mittels TTS-System sowie den Verzicht auf die Optimierungsmessung und Pegelanpassungen reduzierte sich der Aufwand der Testentwicklung erheblich. Für Folgestudien wäre somit die Generierung neuer, vergrößerter Sprachkorpora denkbar. Dies würde es ermöglichen, häufige Sprachverständlichkeitsmessungen durchzuführen ohne einzelne Testlisten wiederholen zu müssen.

Literatur

Ahrlich, M. (2013) Optimierung und Evaluation des Oldenburger Satztests mit weiblicher Sprecherin und Untersuchung des Effekts des Sprechers auf die Sprachverständlichkeit. Bachelorarbeit, Universität Oldenburg

ITU-T P.800. ITU-T (08/1996). P.800 Methods for objective and subjective assessment of quality: Methods for subjective determination of transmission quality (https://www.itu.int/rec/T-REC-P.800-199608-I/en, aufgerufen am 26.05.18)

Wagener, K.; Kühnel, V.; Kollmeier, B. (1999) Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests. Z Audiol 38 (1) 1-32

Wagener, K.; Hochmuth, S.; Ahrlich, M.; Zokoll, M.; Kollmeier, B. (2014). Der weibliche Oldenburger Satztest. In: 17. Jahrestagung der DGA.