SCHNELL GESPROCHENE SPRACHE ALS EINHEITEN-AUSWAHL-INVENTAR IN DER UNIT-SELECTION-SPRACHSYNTHESE

Donata Moers

Arbeitsbereich Sprache und Kommunikation, Universität Bonn
Fakultät für Linguistik und Literatur, Universität Bielefeld
dmo@sk.uni-bonn.de

Abstract: In diesem Beitrag wird ein neuer Ansatz zur Erzeugung schneller Sprache in der Unit-Selection-Sprachsynthese vorgestellt. Für die Erstellung eines eigenen Bausteininventars für schnell gesprochene Sprache wurde zunächst eine geeignete Sprecherin ausgewählt, die in der Lage war, bei maximaler Sprechgeschwindigkeit möglichst deutlich zu sprechen. Typischerweise bei schnellem Sprechen auftretende Phänomene wie zu starke Reduktion und Koartikulation konnten so weitestgehend vermieden werden. Anschließend wurden zwei unabhängige, bezüglich des linguistischen Inhalts jedoch identische Synthesekorpora aufgenommen: eines in normalem und eines in schnellem und möglichst deutlichem Sprechtempo. Auf Grundlage beider Korpora wurden Äußerungen in verschiedenen Sprechgeschwindigkeiten erzeugt und mittels Perzeptionstests mit verschiedenen Hörergruppen untersucht. Dabei wurde sowohl die Verständlichkeit mittels *Semantisch Unvorhersagbarer Sätze* (SUS, [1]), als auch der Gesamteindruck mit Hilfe eines *Mean Opinion Scores* (MOS) im Vergleich zu einem Formantsynthesesystem erhoben. Die Ergebnisse der perzeptiven Evaluation zeigen, dass 1) die Beurteilung in erster Linie von der Verständlichkeit abhängt, 2) die Gewöhnung an bestimmte Sprachausgabeanwendungen einen signifikanten Einfluss auf die Beurteilung der Gesamtqualität hat und 3) ein eigenes Inventar für schnell gesprochene Sprache in der Unit-Selection-Sprachsynthese bei hohen Sprechgeschwindigkeiten keinen Vorteil gegenüber einem normal gesprochenen Inventar aufweist.

1 Einleitung

Die phonetischen Charakteristika schneller Sprache unterscheiden sich von denen in normalem Tempo gesprochener Sprache: Je schneller jemand spricht, desto schlechter sind seine/ihre Äußerungen zu verstehen. Dies liegt vor allem daran, dass artikulatorische Zielstellungen nicht mehr vollständig realisiert werden. Damit einhergehende stärkere Koartikulation und Reduktion beeinträchtigen die Verständlichkeit schnell gesprochener Sprache [2]. Zur Durchführung des hier vorgestellten Vorhabens wurde daher zunächst eine Sprecherin ausgewählt, die in der Lage war, bei maximaler Sprechgeschwindigkeit möglichst deutlich zu sprechen [3], um solche Phänomene bei der Erstellung eines Korpus für schnell gesprochene Sprache weitestgehend zu vermeiden, da sie die Qualität der zu synthetisierenden Äußerungen negativ beeinflussen würden. Anschließend wurden zwei voneinander unabhängige, jedoch inhaltlich identische Synthesekorpora aufgenommen: eines in normaler (4 Silben/Sekunde) und eines in maximal schneller und deutlicher Sprechgeschwindigkeit (8 Silben/Sekunde).

Blinde und sehbehinderte Nutzer von Sprachausgabesystemen bevorzugen häufig eine hohe Sprechgeschwindigkeit, die über die in natürlichem schnellen Sprechen erreichbare Sprechgeschwindigkeit weit hinaus geht [4, 5]. Solch schnelle Sprache kann mit Formantsynthesesystemen ohne weiteres generiert werden; sie klingt jedoch sehr unnatürlich. Unit-Selection-Synthesesysteme dagegen können natürlicher klingenden Output liefern; eine adäquate Implementierung schnell gesprochener Sprache in derartigen Systemen erfolgte bisher jedoch nicht. Ein häufig gewählter Ansatz zur Generierung hoher Sprechgeschwindigkeiten in kor-

pusbasierter Sprachsynthese ist die Beschleunigung der Sprache mit Hilfe von Algorithmen zur Dauermanipulation. Dabei ist für die Erzeugung extrem hoher Sprechgeschwindigkeiten eine sehr starke Manipulation erforderlich, die dazu führt, dass die erzeugte Sprache sehr unnatürlich klingt. Die Nachahmung prosodischer Charakteristika natürlicher schneller Sprache, wie beispielsweise das Kürzen oder Auslassen von Pausen oder Phrasengrenzen, führt hingegen zu einer Verschlechterung der Verständlichkeit schneller Sprache [6] und ist daher nicht für deren Generierung geeignet. Durch die hier vorgestellte Verwendung eines schnell und deutlich gesprochenen Bausteininventars in der Unit-Selection-Sprachsynthese könnte somit vor allem die Natürlichkeit der generierten Sprache erhöht werden.

Geübte blinde Hörer scheinen jedoch die weniger natürlich klingende Formantsynthese gegenüber einer natürlicher klingenden Diphonsynthese zu bevorzugen [4]. Dies liegt neben der reinen Gewöhnung an diese Art der Sprachsynthese möglicherweise daran, dass die für die Verständlichkeit der Sprache wichtigen Transitionen in der Formantsynthese ohne weiteres nachgebildet werden können. Um die für die Verständlichkeit wichtigen Transitionen und koartikulierten Kontexte zu berücksichtigen und so die Verständlichkeit der erzeugten schnell gesprochenen Sprache zu erhöhen, wurden so genannte *Phoxsy Units* [7] in das Unit-Selection-Synthesesystem BOSS [8] implementiert [9]. Auf Basis der zwei unabhängigen Synthesekorpora sowie mittels des auf Formantsynthese basierenden Screenreaders JAWS Eloquence [10] wurden dann *Semantisch Unvorhersagbare Sätze* (SUS, [1]) erzeugt und Perzeptionsexperimente zur Erhebung der Verständlichkeit sowie des Gesamteindrucks mit geübten und ungeübten Hörern durchgeführt.

2 Perzeptive Evaluation

Da davon auszugehen war, dass die Vertrautheit mit der Nutzung von auf Formantsynthese basierenden Sprachausgabeanwendungen bei der Beurteilung der Synthesequalität mittels eines Mean Opinion Scores (MOS) eine Rolle spielen würde, wurde die perzeptive Evaluation mit zwei verschiedenen Gruppen von Versuchspersonen durchgeführt. Eine dieser Hörergruppen bestand aus 21 blinden oder stark sehbehinderten Versuchspersonen (Durchschnittsalter 24 J., 38% weiblich), die bei der Arbeit am Computer auf die Anwendung von Sprachausgabeanwendungen angewiesen waren und über mindestens 2 Jahre Erfahrung in deren Nutzung verfügten (geübte Hörer). Die zweite Gruppe bestand aus 17 sehende Versuchspersonen (Durchschnittsalter 30 J., 35% weiblich), die über wenig oder gar keine Erfahrung in der Nutzung von Sprachausgabesystemen verfügten (ungeübte Hörer).

2.1 Material und Durchführung

Um die Verständlichkeit der synthetisierten Sprache zu erheben, wurden 20 SUS [1] sowohl mit dem normalen als auch mit dem schnell gesprochenen Syntheseinventar mittels des Unit-Selection-Synthesesystems BOSS [8] generiert. Sie umfassten jeweils 6 bis 7 einfache Wörter in grammatikalisch korrekten, inhaltlich jedoch sinnlosen Äußerungen. Die so erzeugten Sätze wurden anschließend mit Hilfe der Sprachverarbeitungssoftware Praat [11] linear auf 4 verschiedene Sprechgeschwindigkeiten beschleunigt, so dass alle Äußerungen in insgesamt 5 verschiedenen Geschwindigkeitsstufen als Stimuli zur Verfügung standen. Die niedrigste Tempostufe entsprach dabei dem normalen Sprechtempo von 4 Silben/Sekunde (sil/sek) und die Höchste einer Sprechgeschwindigkeit von 20 sil/sek; der Abstand zwischen den Tempostufen betrug jeweils 4 sil/sek. Für die Stimuli in normaler Sprechgeschwindigkeit kam nur das in normalem Tempo gesprochene Inventar zur Anwendung, da eine Verlangsamung der schnell gesprochenen Äußerungen für die geplante Erhebung als ungeeignet anzusehen war. Zusätzlich zu den auf Unit-Selection-Sprachsynthese basierenden Stimuli wurden 10 SUS in den entsprechenden 5 Geschwindigkeitsstufen mit dem auf Formantsynthese basierenden Screenreader JAWS [10] generiert. Um eine gleichmäßige Verteilung der Stimuli über alle

Sprechgeschwindigkeiten zu gewährleisten, wurde für das Experiment ein Blockdesign gewählt, bei dem jede Geschwindigkeitsstufe für die Beurteilung der Unit-Selection-Synthese jeweils 4 (2 x 2) der 20 SUS umfasste und für die Beurteilung der Formantsynthese jeweils 2 der mit ihr generierten 10 SUS.

Das Experiment wurde in ruhiger Umgebung durchgeführt. Zunächst wurden die Unit-Selection basierten Sätze der niedrigsten Geschwindigkeitsstufe vorgespielt. Diese Sätze dienten sowohl zur Eingewöhnung als auch zur Festlegung einer Baseline für die jeweilige Verständlichkeit und Beurteilung mittels MOS im Vergleich zu höheren Tempostufen. Ab Geschwindigkeitsstufe 2 hörten die Versuchspersonen jeweils 2 Sätze basierend auf dem normalen Bausteininventar und 2 Sätze basierend auf dem schnellen Bausteininventar in randomisierter Reihenfolge. Im Anschluss an die Präsentation der Unit-Selection basierten Stimuli wurden die mit JAWS [10] generierten Sätze in aufsteigender Geschwindigkeit präsentiert.

In beiden Teilen des Experiments konnte jeder Satz einmal wiederholt werden. Die Versuchspersonen sprachen den Wortlaut der Sätze nach und gaben anschließend ihren Gesamteindruck mit einem MOS entsprechend einer Schulnote von 1 = sehr gut bis 5 = mangelhaft wieder. Dabei wurden die Versuchspersonen darauf hingewiesen, dass auch die Natürlichkeit und die Stimmqualität bei der Beurteilung durch den MOS berücksichtigt werden sollten. Zusätzlich wurde erfasst, ob eine Wiederholung des Satzes notwendig war. Damit standen für die Auswertung der Ergebnisse folgende Variablen zur Verfügung:

- Hörer {1 = ungeübt, 2 = geübt}
- Inventar {1 = Unit-Selection normale Sprechgeschwindigkeit, 2 = Unit-Selection schnelle Sprechgeschwindigkeit, 3 = Formantsynthese)
- Sprechgeschwindigkeit {1 = 4 sil/sek, 2 = 8 sil/sek, 3 = 12 sil/sek, 4 = 16 sil/sek, 5 = 20 sil/sek)
- Wiederholung {1 = keine Wiederholung, 2 = einmalige Wiederholung}
- Anzahl der verstandenen Wörter {1, 2, 3, 4, 5, 6, 7}
- Beurteilung (MOS) {sehr gut = 1, gut = 2, befriedigend = 3, ausreichend = 4, mangelhaft = 5}

Als statistische Methode kamen bei der Auswertung binäre Entscheidungsbäume zur Anwendung, da sie eine Einbeziehung aller angeführten Variablen erlauben, unabhängig von deren Skalierung. Die oben aufgeführten Größen wurden für jeden Stimulus einzeln kodiert und in einer Datenbank zusammengefasst. Die Berechnung der Entscheidungsbäume erfolgte mit Hilfe der ctree-Funktion der Statistiksoftware R [12]. Zusätzlich wurde die Worterkennungsrate (WER) in Prozent der verstandenen Wörter ermittelt. Da erwartet wurde, dass zwischen den beiden Hörergruppen signifikante Unterschiede bezüglich der WER und des MOS für die jeweiligen Inventare und Synthesearten auftreten würden, wurden die Ergebnisse zunächst einzeln für die jeweilige Hörergruppe ermittelt, bevor anschließend eine Gesamtauswertung über beide Hörergruppen erfolgte. Die Ergebnisse werden im Folgenden entsprechend dieser Aufteilung vorgestellt.

2.2 Ergebnisse für geübte Hörer

Abbildung 1 zeigt den Entscheidungsbaum für die Vorhersage des MOS unter Einbeziehung der Variablen „Inventar“, „Sprechgeschwindigkeit“, „Wiederholung“ und „Anzahl der verstandenen Wörter“. Die Anzahl der verstandenen Wörter wird hier als oberstes Splitkriterium gesetzt. Dies zeigt, dass die Verständlichkeit den stärksten Einfluss auf die globale Beurteilung hat; Natürlichkeit und Stimmqualität scheinen hingegen eine weniger wichtige Rolle zu

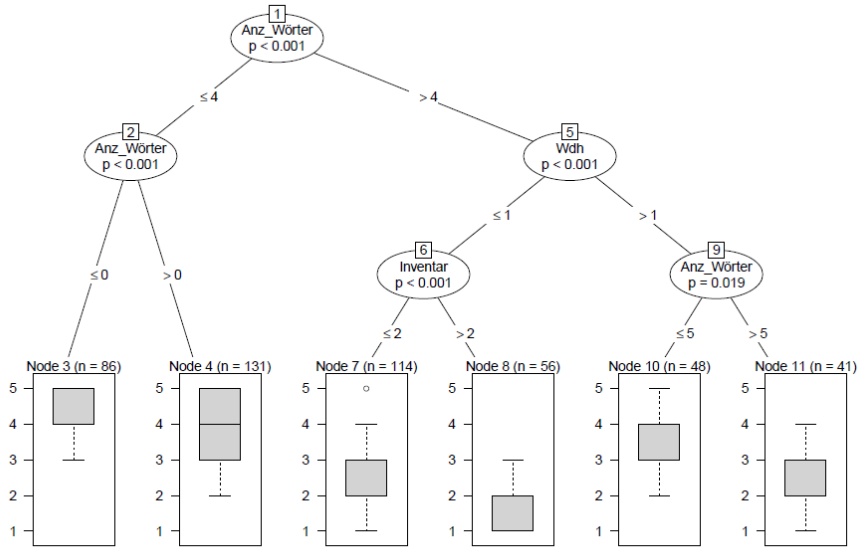

Abbildung 1 – Entscheidungsbaum für geübte Hörer: MOS in Abhängigkeit von Inventar, Sprechgeschwindigkeit, Anzahl der verstandenen Wörter und Wiederholung.

spielen. Wird die Anzahl der verstandenen Wörter mit der Beurteilung korreliert, ergibt sich dementsprechend ein hoch signifikanter Zusammenhang ($r = -0.729$). Die Korrelation der Sprechgeschwindigkeit mit der Beurteilung weist zwar ebenfalls einen signifikanten, aber niedrigeren Wert ($r = 0.565$) auf, der bei der Darstellung des Entscheidungsbaumes nicht berücksichtigt wird, da er keine zusätzlichen statistisch relevanten Informationen liefert.

Anhand des obersten Knotens wird zudem deutlich, dass ein signifikanter Unterschied zwischen Sätzen, von denen maximal 4 Wörter verstanden wurden, und Sätzen, von denen mehr als 4 Wörter verstanden wurden, besteht (Knoten 1, $p < 0.001$). Am linken Teil des Baumes ist zu erkennen, dass ein weiterer signifikanter Unterschied zwischen der Beurteilung von Sätzen besteht, von denen kein Wort verstanden wurde, und solchen Äußerungen, die nur teilweise ($4 \geq$ Anzahl verstandener Wörter > 0) verständlich waren (Knoten 2, $p < 0.001$). Im rechten Teil des Baumes folgt der Anzahl der verstandenen Wörter als nächstes Splitkriterium die Wiederholung. Zwischen dem MOS für Sätze, die bei der ersten Präsentation verstanden wurden, und Sätzen, die wiederholt werden mussten, besteht ein signifikanter Unterschied bezüglich der Beurteilung (Knoten 5, $p < 0.001$) zugunsten der nicht wiederholten Sätze.

Musste ein Satz wiederholt präsentiert werden, kommt als weiteres Kriterium wiederum die Anzahl der verstandenen Wörter zum Tragen: Der MOS für Sätze, von denen mehr als 5 Wörter verstanden wurden, ist signifikant besser (Knoten 9, $p < 0.05$) als für Sätze, von denen nur 5 Wörter verstanden wurden. Bei Äußerungen, die nicht wiederholt werden mussten (Wdh $\leq$ 1), folgt hingegen ein Split zwischen den beiden für die Unit-Selection-Synthese genutzten Inventaren und der Formantsynthese (Knoten 6, $p < 0.001$). Der MOS für die Formantsynthese fällt hier signifikant besser aus als die Beurteilung der Unit-Selection-Sprachsynthese. Da es sich um Beurteilungen durch (mit Formantsynthese) geübte Hörer handelt, ist dieses Ergebnis ein Hinweis darauf, dass hier ein Gewöhnungseffekt vorliegt, der zu höherer Verständlichkeit und damit einhergehend zu einem besseren MOS für diese Art der Sprachsynthese führt.

2.3 Ergebnisse für ungeübte Hörer

Abbildung 2 zeigt den Entscheidungsbaum für die Vorhersage des MOS der ungeübten Hörer unter Einbeziehung der Variablen „Inventar“, „Sprechgeschwindigkeit“, „Wiederholung“ und „Anzahl der verstandenen Wörter“. Die Anzahl der verstandenen Wörter wird hier ebenfalls

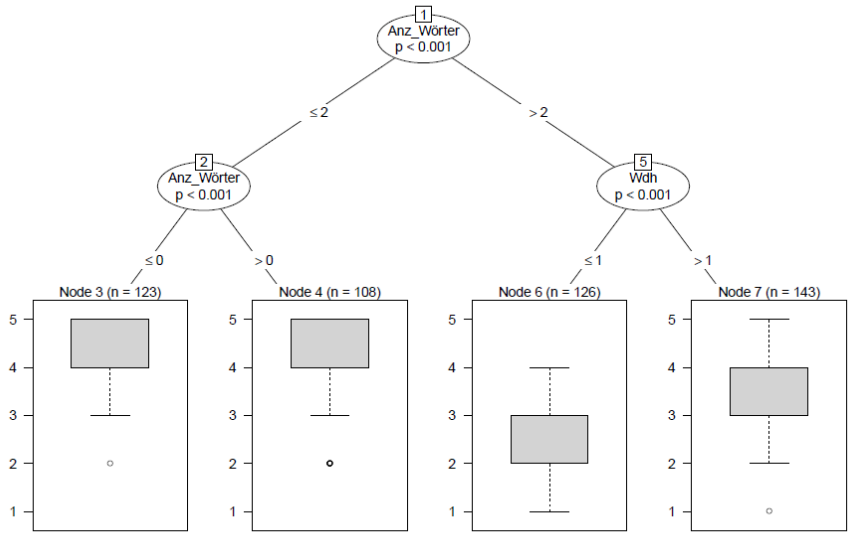

Abbildung 2 – Entscheidungsbaum für ungeübte Hörer: MOS in Abhängigkeit von Inventar, Sprechgeschwindigkeit, Anzahl der verstandenen Wörter und Wiederholung.

als oberstes Splitkriterium gesetzt. Dies zeigt, dass diese Größe auch bei ungeübten Hörern von allen zu berücksichtigenden Aspekten den stärksten Einfluss auf die Beurteilung hat. Wird die Anzahl der verstandenen Wörter mit der Beurteilung korreliert, ergibt sich dementsprechend auch hier ein signifikanter Zusammenhang ($r = -0.662$), ebenso wie für die Korrelation der Beurteilung mit der Sprechgeschwindigkeit ($r = 0.558$). Letztere wird bei der Berechnung des Entscheidungsbaumes aufgrund des kleineren Wertes wiederum nicht berücksichtigt, da sie keine zusätzlichen statistisch relevanten Informationen liefert.

Anhand des obersten Knotens wird deutlich, dass im Gegensatz zur Gruppe der geübten Hörer hier bereits ein signifikanter Unterschied zwischen Sätzen, von denen maximal 2 Wörter verstanden wurden, und Sätzen, von denen mehr als 2 Wörter verstanden wurden, auftritt (Knoten 1, $p < 0.001$). Am linken Teil des Baumes ist zu erkennen, dass ein weiterer signifikanter Unterschied zwischen der Beurteilung von Sätzen besteht, die gänzlich unverständlich waren, und solchen Äußerungen, die nur teilweise ($2 \geq$ Anzahl verstandener Wörter > 0) verstanden wurden (Knoten 2, $p < 0.001$). Im rechten Teil des Baumes folgt der Anzahl der verstandenen Wörter als zweitwichtigstes Kriterium auch hier die Wiederholung. Musste ein Satz wiederholt präsentiert werden, wurde er signifikant schlechter beurteilt als Sätze, die nicht erneut vorgespielt werden mussten (Knoten 5, $p < 0.001$). Im Gegensatz zur Gruppe der geübten Hörer besteht hier jedoch kein signifikanter Unterschied zwischen den Beurteilungen der einzelnen Inventare, was zunächst auf die Unerfahrenheit dieser Hörergruppe mit Sprachausgabesystemen im Allgemeinen zurückgeführt wurde. Bei der anschließenden Auswertung der Ergebnisse für beide Hörergruppen im Vergleich wurde dann aber deutlich, dass bei Einschränkung der berücksichtigten Variablen sehr wohl signifikante Unterschiede zwischen den Beurteilungen der verschiedenen Inventare durch die unterschiedlichen Hörergruppen auftreten.

2.4 Vergleich beider Hörergruppen

Abbildung 3 zeigt den Entscheidungsbaum für die Vorhersage des MOS unter Einbeziehung der Variablen „Inventar“ und „Hörer“. Die Zugehörigkeit zur Gruppe der ungeübten (1) oder geübten (2) Hörer wird hier als oberstes Splitkriterium gesetzt. Anhand des linken Teilbaumes wird deutlich, dass die Gruppe der ungeübten Hörer die auf dem in normalem Tempo gesprochenen Unit-Selection-Inventar basierenden Stimuli (1) als signifikant besser beurteilt als Stimuli, die entweder auf dem schnell gesprochenen Unit-Selection-Inventar (2) oder aber auf

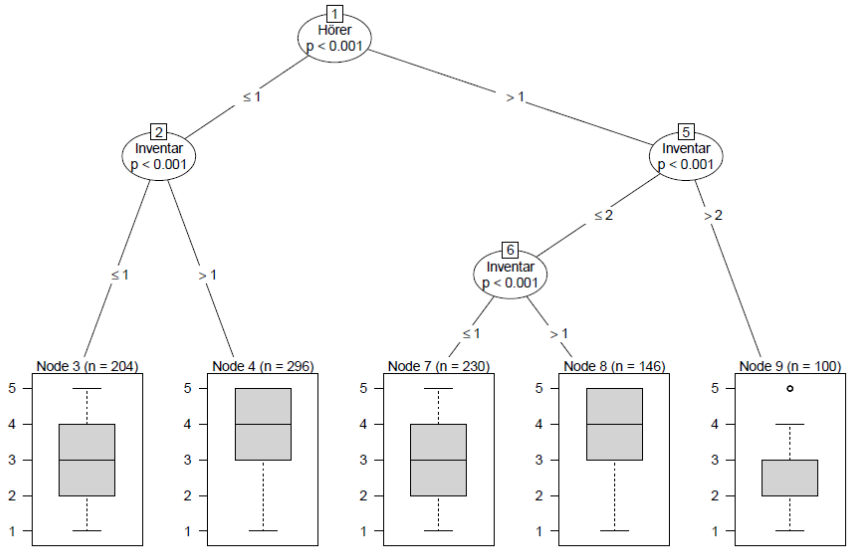

Abbildung 3 – Entscheidungsbaum für beide Hörergruppen: MOS in Abhängigkeit von Gruppenzugehörigkeit und Inventar.

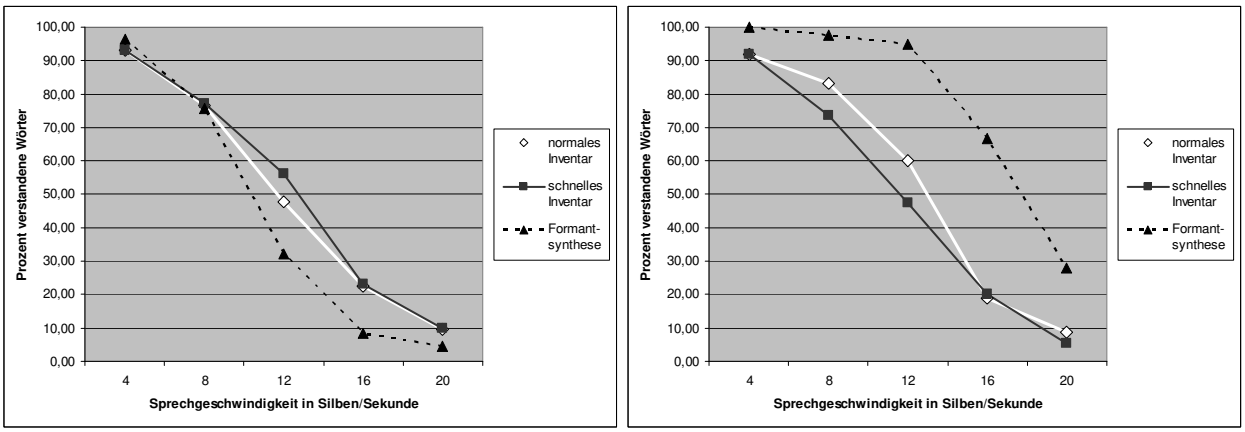

Abbildung 4 – Worterkennungsrate für normal gesprochenes Unit-Selection-Inventar (weiße Linie), schnell gesprochenes Unit-Selection-Inventar (graue Linie) und Formantsynthese (schwarze Linie) in Abhängigkeit von der Sprechgeschwindigkeit für ungeübte (links) und geübte Hörer (rechts).

Formantsynthese (3) basieren (Knoten 2, $p < 0.001$). Dagegen ist im rechten Teil des Baumes, der die Gruppe der geübten Hörer repräsentiert, zu beobachten, dass zunächst ein signifikanter Unterschied zwischen den Beurteilungen der auf Formantsynthese basierenden Stimuli und den Beurteilungen der auf Unit-Selection-Synthese basierenden Stimuli zugunsten der Formantsynthese existiert (Knoten 5, $p < 0.001$). Es folgt für die auf Unit-Selection basierenden Stimuli dann ein weiterer Split, der einen signifikanten Unterschied zwischen dem MOS für die auf dem in normalem Tempo gesprochenen Inventar basierenden Stimuli und dem MOS für die auf dem schnellen Inventar basierenden Stimuli aufzeigt (Knoten 6, $p < 0.001$). Genau wie ungeübte Hörer beurteilen auch geübte Hörer die auf dem in normalem Sprechtempo erstellten Inventar basierenden Stimuli besser als die auf schneller Sprache basierenden Stimuli.

Im Gegensatz zu ungeübten Hörern schneidet jedoch die Formantsynthese bei den Beurteilungen durch geübte Hörer insgesamt deutlich besser ab als die Unit-Selection-Sprachsynthese. Dieser Umstand spiegelt sich in der WER, die mit den Beurteilungen stark korreliert, deutlich wider. Abbildung 4 zeigt die WER in Prozent für das in normalem Tempo gesprochene Unit-Selection-Inventar (weiße Linie), das schnell gesprochene Unit-Selection-Inventar (graue Linie) und die Formantsynthese (schwarze Linie) in Abhängigkeit von der

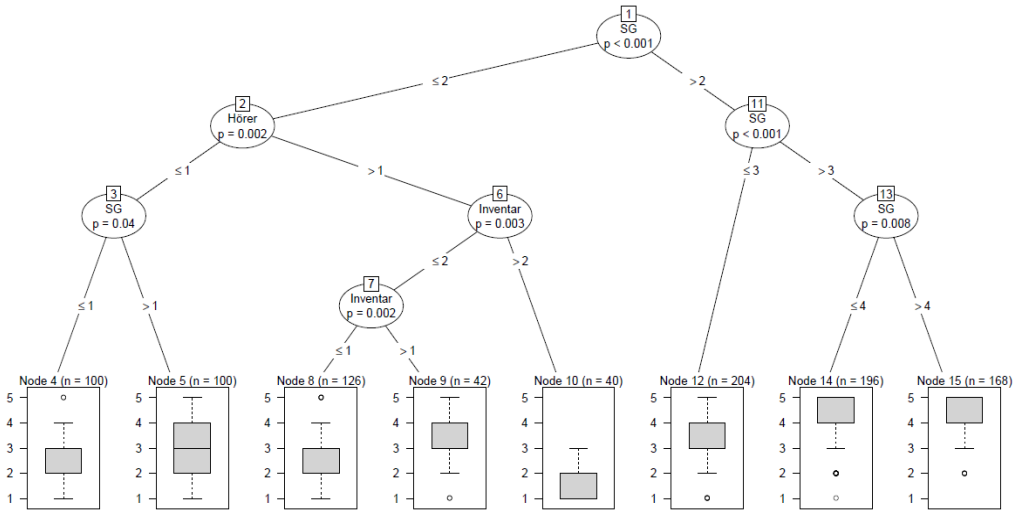

Abbildung 5 – Entscheidungsbaum für beide Hörergruppen: MOS in Abhängigkeit von Sprechgeschwindigkeit, Gruppenzugehörigkeit und Inventar.

Sprechgeschwindigkeit für ungeübte (links) und geübte Hörer (rechts). Es ist deutlich zu erkennen, dass die WER für die Formantsynthese bei geübten Hörern oberhalb der WER für beide Unit-Selection-Inventare liegt, wohingegen sie bei ungeübten Hörern darunter verläuft.

Weiterhin zeigt der Entscheidungsbaum, bei dem neben „Hörer“ und „Inventar“ auch die Sprechgeschwindigkeit (SG) als Einflussgröße berücksichtigt wurde (Abbildung 5), dass der signifikanteste Unterschied beim MOS in Abhängigkeit von der Sprechgeschwindigkeit für beide Hörergruppen zwischen Geschwindigkeitsstufe 2, die mit 8 sil/sek der in natürlichem Sprechen schnellst möglichen Sprechgeschwindigkeit entspricht, und Geschwindigkeitsstufe 3 (12 sil/sek) liegt (Knoten 1, $p < 0.001$). Dies zeigt, dass Sprache, die in einem Sprechtempo jenseits der natürlich artikulierbaren Sprechgeschwindigkeit vorliegt, als signifkant schlechter beurteilt wird als Sprache, die auf einem Inventar in einer natürlichen Sprechgeschwindigkeit basiert.

3 Diskussion

Für die Evaluation synthetisierter schnell gesprochener Sprache wurden so genannte *Semantisch Unvorhersagbare Sätze* [1] sowohl mit einem Unit-Selection-Synthesesystem unter Verwendung eines in normalem Tempo gesprochenen Inventars sowie eines schnell gesprochenen Inventars, als auch mit einem Formantsynthesesystem generiert.

Die perzeptive Evaluation der synthetisierten Sprache zeigte zunächst, dass für beide befragten Hörergruppen die Beurteilung in Form eines *Mean Opinion Scores* signifikant von der Anzahl der verstandenen Wörter sowie der notwendigen Wiederholung der Äußerungen beeinflusst wurde. Dies kann möglicherweise darauf zurückgeführt werden, dass im Allgemeinen bei der Verwendung von Sprachsynthesesystemen in erster Linie die Verständlichkeit der generierten Sprache von Bedeutung ist und Natürlichkeit sowie Stimmqualität eine untergeordnete Rolle spielen; zahlreiche Versuchspersonen aus beiden Teilnehmergruppen merkten nach Durchführung des Experiments an, dass sie sich bei schlechter Verständlichkeit für eine schlechtere Gesamtbeurteilung entschieden hätten, obwohl die Stimmqualität und Natürlichkeit bei der Unit-Selection-Synthese deutlich besser seien als bei der Formantsynthese.

Ebenso waren die Beurteilungen für Sprechgeschwindigkeiten, die auf natürliche Weise mit schnellem Sprechen zu erreichen sind, signifikant besser als die Beurteilungen für Sprechgeschwindigkeiten, die nur mit synthetischer Sprache und unter Einsatz starker Dauermanipulation zu erreichen sind. Dies unterstützt die Beobachtung, dass Sprechgeschwindigkeiten jen-

seits der natürlich gegebenen Grenze der artikulatorisch möglichen Sprechgeschwindigkeit in erster Linie in Abhängigkeit von ihrer Verständlichkeit beurteilt werden und Attribute wie Stimmqualität und Natürlichkeit bei unnatürlich hoher Sprechgeschwindigkeit nur noch eine untergeordnete Rolle spielen.

Die mit dem in normalem Tempo gesprochenen Unit-Selection-Inventar generierten Sätze wurden von ungeübten Hörern als signifikant besser beurteilt als Äußerungen, die entweder mit dem schnell gesprochenen Bausteininventar oder aber mit Formantsynthese generiert wurden. Im Gegensatz dazu beurteilten geübte Hörer jedoch die mit Formantsynthese erzeugten Äußerungen als signifikant besser als die mit Unit-Selection-Synthese generierten Stimuli. Dies könnte auf die durch häufige Nutzung entstandene Gewöhnung an das verwendete Formantsynthesesystem zurückzuführen sein. Möglicherweise ließe sich bei einem über einen längeren Zeitraum andauernden Training mit Unit-Selection-Sprachsynthese eine ähnlich hohe Worterkennungsrate bei besserer Stimmqualität und Natürlichkeit erreichen.

Zusätzlich wurden, wie bei der Gruppe der ungeübten Hörer, Stimuli, die mit dem schnell gesprochenen Bausteininventar synthetisiert wurden, jedoch als signifikant schlechter beurteilt als Stimuli, die mit dem in normalem Tempo gesprochenen Unit-Selection-Inventar generiert wurden. Die Ergebnisse zeigen, dass die Verwendung eines unabhängigen schnell gesprochenen Bausteininventars für die Synthese sehr hoher Sprechgeschwindigkeiten mit einem Unit-Selection-Synthesesystem keinen Vorteil gegenüber der Verwendung eines in normalem Tempo gesprochenen Inventars aufweist.

Literatur

[1] Benoît, C., Grice, M., und Hazan, V., „The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences“, Speech Communication 18: 381-392, 1996.

[2] Krause, J.C., und Braida, L.D., “Investigating Alternative Forms of clear speech: the effects of speaking rate and speaking mode on intelligibility”, Journal of the Acoustical Society of America 112: 2165–2172, 2002.

[3] Moers, D., und Wagner, P., “Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis”, Proc. Interspeech 2009, Brighton, UK, 2009.

[4] Moos, A. und Trouvain, J., “Comprehension of Ultra-Fast Speech – Blind vs. 'Normally Hearing' Persons”, in Proc. ICPhS XVI: 677–684, Saarbrücken, 2007.

[5] Moers, D., Wagner, P., und Breuer, S., “Assessing the Adequate Treatment of Fast Speech in Unit Selection Speech Synthesis Systems for the Visually Impaired”, in Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn, 2007.

[6] Janse, E., “Production and Perception of Fast Speech”, Dissertation, Universiteit Utrecht, 2003.

[7] Breuer, S. und Abresch, J., „Phoxsy: Multiphone Segments for Unit Selection Speech Synthesis“, in Proc. Interspeech 2004 - ICSLP, Jeju Island, Korea, 2004.

[8] Klabbers, E. et al., “Speech synthesis development made easy: The Bonn Open Synthesis System”, Proc. Eurospeech, Aalborg, 2001.

[9] Moers, D., Wagner, P., und Möbius, B., „Erzeugung schnell gesprochener Sprache in der Unit Selection Sprachsynthese“, Proc. ESSV 2010, Berlin, 2010.

[10] Freedom Scientific, „JAWS Screen reading software“, http://www.freedomscientific.com, 01.07.2011.

[11] Boersma, P. und Weenink, D., „Praat: doing phonetics by computer“ (Version 5.2.21), http://www.praat.org/, 01.07.2011

[12] R Development Core Team, „R: A Language and Environment for Statistical Computing“, R Foundation for Statistical Computing, Vienna, Austria, 2011, http://www.R-project.org, 01.07.2011