EINZELFALLSTUDIE ZU GRENZEN DER VERSTÄNDLICHKEIT ULTRA-SCHNELLER SPRACHSYNTHESE

Anja Moos und Jürgen Trouvain

Institut für Phonetik, Universität des Saarlandes, Saarbrücken
{anmo|trouvain}@coli.uni-saarland.de

Abstract: Der Artikel beschreibt an Hand einer Einzelfallstudie die Fähigkeiten einer blinden Person, ultra-schnelle Sprachsynthese zu verstehen. Im Experiment wurde die Verständlichkeit von Texten in Geschwindigkeiten von 18-36 Silben pro Sekunde (s/s) untersucht, die von einer Formantsynthese produziert wurden. Eine gute bis sehr gute Verständlichkeit liegt bei Texten im Tempo bis 22 s/s vor, wohingegegen fast keine bis keine Verständlichkeit ab Tempostufe 32 s/s festzustellen ist. Die subjektive Einschätzung und die externe Überprüfung der Verständlichkeit stimmen dabei nicht immer überein, wobei die Versuchsperson ihre subjektive Verständlichkeit schlechter einschätzt als die Bewertung einer externen Überprüfung nahe legt. Abschließend werden Fragen zum "Training" ultra-schneller Sprachsynthese sowie Einflüsse von Prosodie auf deren Verständlichkeit diskutiert.

1 Einführung

Wie schnell kann synthetische Sprache im Extremfall sein, damit sie noch verstanden wird? Vorangegangene Studien [1, 2] mit blinden Personen, die seit mehreren Jahren täglich Sprachsynthese benutzen, und sehenden Nicht-Benutzern von Sprachsynthese haben gezeigt, dass "trainierte" Blinde sowohl stark beschleunigte natürliche Sprache als auch genauso schnelle Formantsynthese deutlich besser verstehen als Sehende.

Studie [2] zeigt, dass im Schnitt mit Formantsynthese generierte Texte mit einem Tempo bis zu 19 Silben pro Sekunde (s/s) als verständlich eingestuft wurden, von einzelnen Blinden sogar bis 22 s/s (vgl. Abbildung 1). Allerdings werden im Durchschnitt zeitkomprimierte Texte eines natürlichen Sprechers bei Geschwindigkeiten ab 19 s/s als weniger verständlich bewertet als mit Formantsynthese generierte Texte in vergleichbarem Tempo.

Im Gegensatz zu den Befunden bei den Blinden erreichte die sehende Kontrollgruppe die Grenze der guten Verständlichkeit bei 10 s/s für Formantsynthese, für zeitkomprimierte natürliche Sprache allerdings erst bei ca. 13 s/s. Dabei beträgt die normale Artikulationsgeschwindigkeit beim Vorlesen oder in Unterhaltungen zwischen 4 bis 8 s/s.

Abbildung 1 zeigt die durchschnittlichen Ergebnisse aus [2] für die subjektive Verständlichkeit bei den fünf (von 19) Blinden, die am besten abgeschnitten haben, verglichen mit denen der Sehenden (20 Vpn).

Die vorliegende Untersuchung widmet sich der Frage, bis zu welchem Tempo (in s/s) synthetische Sprache für trainierte Hörer noch verständlich ist. Dazu wird eine Einzelfallstudie vorgestellt.

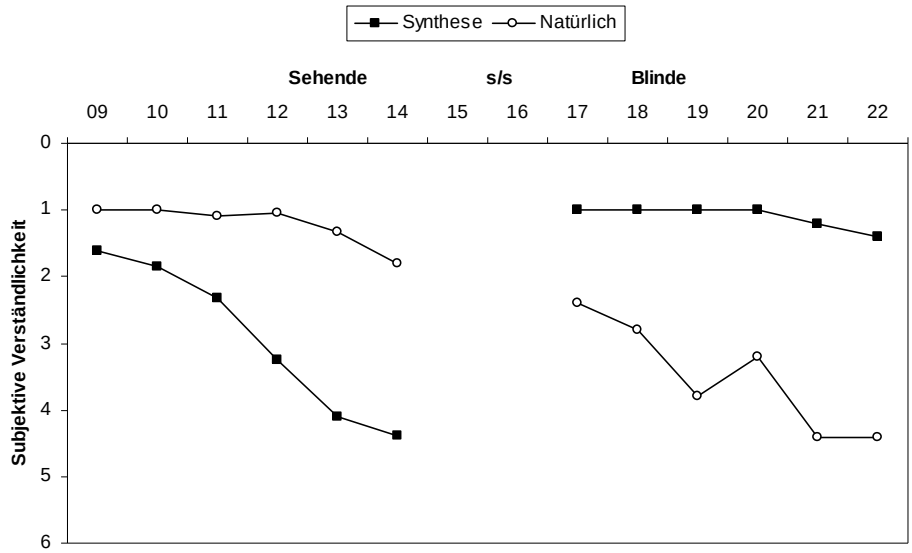

Abbildung 1 – Subjektive Verständlichkeit auf einer Skala von "alles verstanden" (1) bis "nichts verstanden" (6) für Stimuli von 9 Silben pro Sekunde (s/s) bis 14 s/s bei Sehenden (links) und bei einer Untergruppe von 5 (von 19) Blinden (von 17 s/s bis 22 s/s; rechts). Die Daten stammen aus Studie [2], bei denen Aufnahmen mit Texten im Normaltempo eines natürlichen Sprechers bzw. durch Formantsynthese erzeugt zeitlich komprimiert wurden.

2 Methode

2.1 Versuchsperson

Als Versuchsperson (Vp) fungierte eine Person, die im Alter von 13 Jahren erblindet ist. Seit 13 Jahren hat die Vp (tägliche) Erfahrung mit Formantsynthese.

2.2 Material

Als Stimuli für den Hörtest wurden 30 Nachrichtentexte im Umfang von je 80-100 Wörtern benutzt. Sie wurden mit der Screenreadersoftware JAWS [3] in Normalgeschwindigkeit erzeugt (ca. 5 s/s). Da bei diesem Programm keine stillen Pausen erzeugt werden, handelt es sich bei den angegebenen Geschwindigkeiten sowohl um die Sprechgeschwindigkeit (als Bruttogeschwindigkeit) als auch um die Artikulationsrate (als Nettogeschwindigkeit). Trotz der Option, Satzzeichen eigens vorlesen zu lassen, wurde hierbei von dieser Möglichkeit kein Gebrauch gemacht.

Zur Beschleunigung dieser Basisaufnahmen wurde die Standard-Software Praat [4] verwendet, die mittels der PSOLA-Methode [5] die zeitlichen Verhältnisse *linear* verändert, dabei die Grundfrequenz aber konstant hält. Hierzu ist festzuhalten, dass Tempoveränderung bei menschlicher Sprachproduktion in starkem Maße *nicht-linear* abläuft, was sich sowohl in Lautdauern als auch in Anzahl und Dauern von Pausen manifestiert. Zudem können bei schnellen Geschwindigkeiten lautliche Gesten so stark reduziert werden, dass Phoneme ganz verschwinden bzw. durch veränderte zeitliche Koordination neue Phoneme durch Assimilation entstehen. Auf prosodischer Ebene kann auch die Anzahl und die Ausformung von Satzakzenten durch Tempoveränderung betroffen sein (zu Effekten von Tempoänderung vgl. [6, 7]). Interessanterweise führen aber nicht die reduzierten hypo-artikulierten lautlichen Formen zu einer besseren Verständlichkeit bei sehr hohen Artikulationsgeschwindigkeiten, sondern hyper-artikulierte Formen (vgl. [6]). Es ist daher davon auszugehen, dass bei ultra-schnellen Geschwindigkeiten eine lineare Komprimierung des Sprachsignals – wie hier

angewendet – durch nicht stattfindende Reduktion sich günstig auf die Verständlichkeit auswirkt.

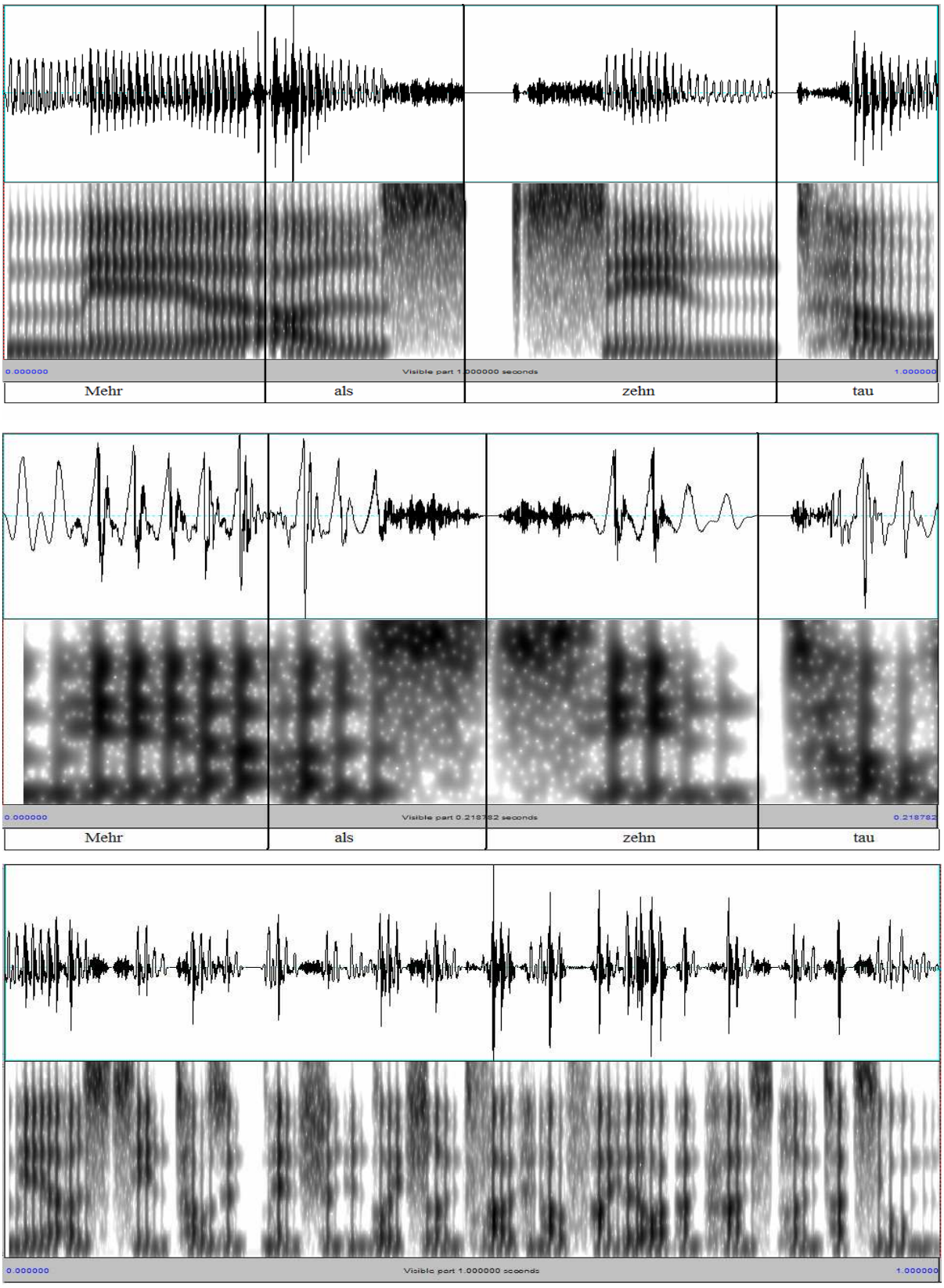

Abbildung 2 – Wellenform und Spektrogramm des Satzanfangs "Mehr als zehntau(send ...)". Oben: 1,000 sek im Normaltempo (ca. 5 s/s). Mitte: 0,219 sek im Tempo 22 s/s. Unten: 1,000 sek im Tempo 22 s/s.

Die Stimuli lagen in zehn verschiedenen Tempo-Stufen im Abstand von 2 s/s vor. Die "langsamste" Stufe lag bei 18 s/s, die schnellste bei 36 s/s. Pro Tempo-Stufe wurden drei Texte vorgespielt, wobei alle Stimuli in randomisierter Reihenfolge angeboten wurden.

Zur Illustration des Komprimierungseffekts dient Abbildung 2, bei der man das akustische Signal einzelner Silben zwischen der Normalgeschwindigkeit (oben) und der von Blinden noch als gut verständlichen ultra-schnellen Geschwindigkeit von 22 s/s (Mitte) vergleichen kann. Einen größeren Ausschnitt der schnellen Geschwindigkeit ist im unteren Teil erkennbar, in dem auch sichtbar wird, dass manche kurze Vokale nur wenige Perioden aufweisen – mitunter nur eine einzige Periode.

2.3 Perzeptionstest

Der Test wurde in einem ruhigen Büroraum via Laptop und Kopfhörer durchgeführt. Die ersten abgespielten Texte dienten als "Aufwärmphase".

Nach jedem Text sollte auf einer 6-Punkte-Skala subjektiv bewertet werden, wie gut der Inhalt verstanden wurde (*subjektive Verständlichkeit*). Die Skala reicht von "alles verstanden" (1), über "fast alles verstanden" (2), "mehr als die Hälfte verstanden" (3), "weniger als die Hälfte verstanden" (4), "fast nichts verstanden" (5) bis "nichts verstanden" (6).

Außerdem erstellte die Vp eine knappe inhaltliche Zusammenfassung, welche die Vp in einen separaten PC eintippte. Diese Aufgabe diente einer nicht subjektiven Überprüfung des Textverständnisses (*externe Bewertung*). Dabei bewerteten die Autoren unabhängig voneinander auf einer 6-Punkte-Skala (ähnlich der Schulnotenskala) die schriftlichen Zusammenfassungen.

Zusätzlich musste die Vp in einem zweiten Durchgang Teilsätze (Länge: 6-9 Wörter) der jeweiligen Stimuli nachsprechen (mit Mikrophonaufnahmen). Die Verständlichkeit wurde durch die Anzahl der korrekt wiederholten Wörter gemessen (*Nachsprechbewertung*), bei der zum einen alle Wörter und zum anderen nur die Inhaltswörter berücksichtigt wurden.

3 Ergebnisse

3.1 Subjektive Verständlichkeit

Die Antworten zur subjektiven Verständlichkeit wurden über die drei Texte pro Tempostufe gemittelt. Abbildung 3 zeigt, dass es bei der subjektiven Einschätzung der Vp einen kontinuierlichen Abfall in der Verständlichkeit von 18 s/s bis 34 s/s gibt. Bis 22 s/s wurden die Stimuli als "fast alles verstanden" eingestuft, ab 32 s/s hat die Vp "fast nichts verstanden".

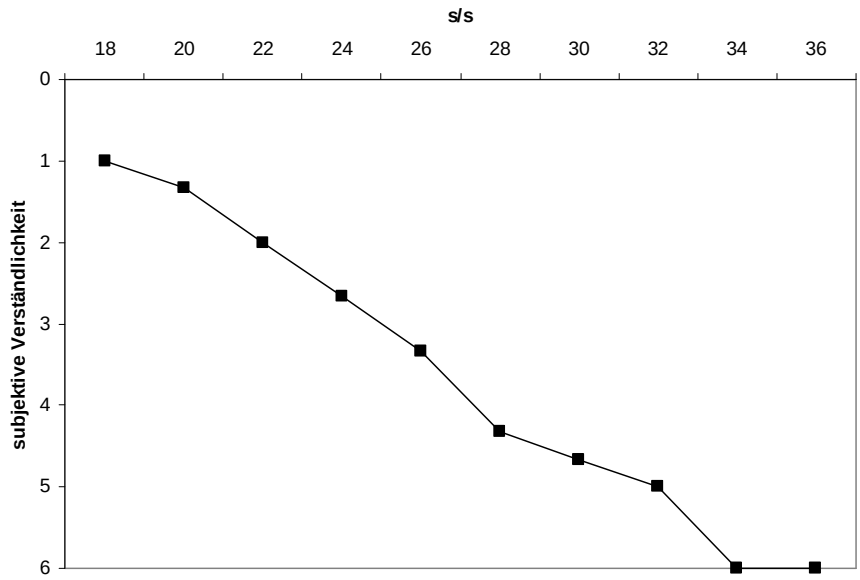

Abbildung 3 – Subjektive Verständlichkeit der Vp auf einer Skala von "alles verstanden" (1) bis "nichts verstanden" (6) für Stimuli von 18 Silben pro Sekunde (s/s) bis 36 s/s

3.2 Objektive Maße

Der kontinuierliche Abfall der Verständlichkeit spiegelt sich auch in beiden objektiven Maßen wieder, allerdings mit leicht besseren Einstufungen: Bis 28 s/s wurden bei der Nachsprechaufgabe mehr als 50% der Wörter korrekt wiedergegeben (vgl. Abbildung 4). Ebenso gut fielen die Bewertungen der inhaltlichen Zusammenfassungen aus.

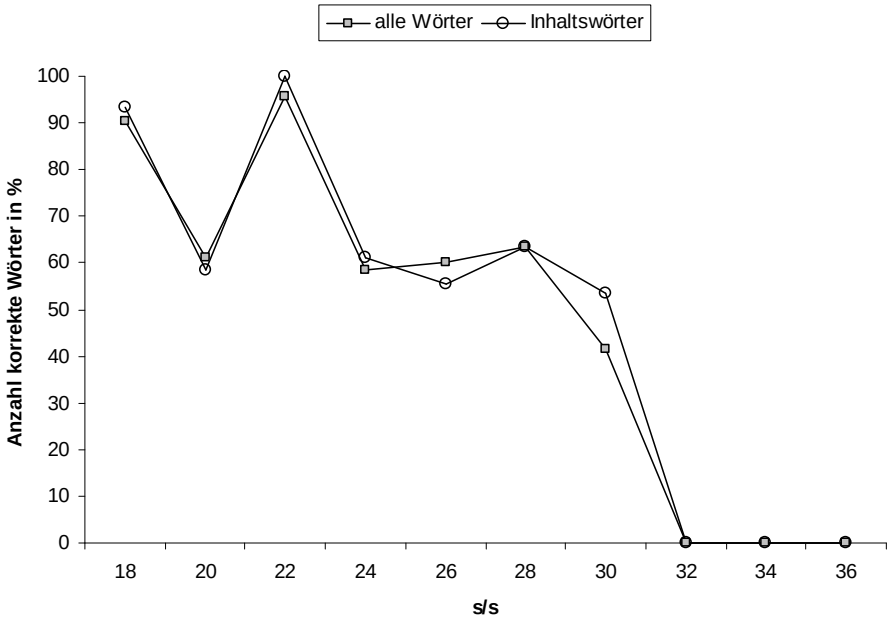

Abbildung 4 – Objektive Maße: Nachsprechbewertung in Prozent korrekt wiedergegebener Wörter bzw. Inhaltswörter.

Es ergeben sich keine relevanten Unterschiede zwischen der Zählung aller Wörter und der Zählung der Inhaltswörter (außer bei 30 s/s). Allerdings treten starke Schwankungen zwischen den Tempostufen 18 s/s und 22 s/s einerseits und 20 s/s und 24 s/s und schneller andererseits auf. Der Grund für diese großen Unterschiede liegt darin, dass bei 20 s/s und bei 24 s/s jeweils einer von drei Sätzen, die nachzusprechen waren, gar nicht verstanden wurden im Gegensatz zu den beiden anderen in der jeweiligen Tempostufe. Dadurch fällt der Prozentwert korrekt erkannter Wörter um jeweils etwa 30 %.

Die Kurzzusammenfassungen der Texte in den Tempostufen von 18 s/s bis 24 s/s wurden als "gut" bis "sehr gut" bewertet (vgl. Abbildung 5). Von 26 s/s bis 28 s/s sind die Zusammenfassungen noch "gut" bis "befriedigend", bevor sie ab 30 s/s deutlich schlechter werden und ab 32 s/s nicht mehr ausreichend sind.

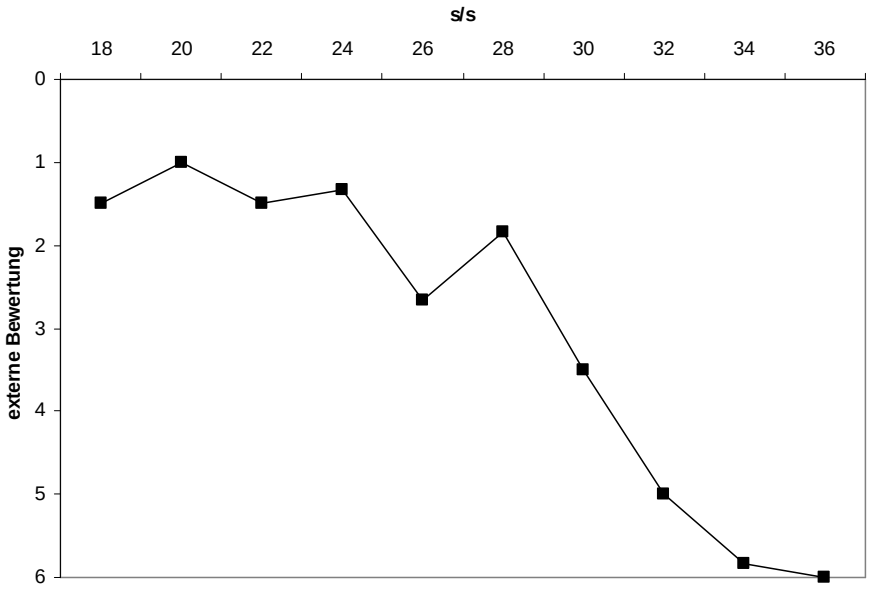

Abbildung 5 – Objektives Maß: Externe Bewertung der Textzusammenfassung auf einer Schulnotenskala.

4 Diskussion

Die vorliegende Fallstudie zeigt, bis zu welchen extremen Geschwindigkeiten Hörer in der Lage sind, ultra-schnelle Sprache zu verstehen. Die Ergebnisse der subjektiven und der objektiven Bewertungen lassen den Schluss zu, dass für die untersuchte Vp Texte mit einer Artikulationsgeschwindigkeit bis 22 s/s gut verständlich sind. Somit handelt es sich um eine Bestätigung des Befundes aus [2], bei der die selbe Vp neben anderen Vpn ebenfalls Texte bis 22 s/s noch als gut verständlich eingestuft hat (vgl. Abbildung 1).

Ab der Tempostufe 24 s/s stimmen subjektive und objektive Bewertung nicht mehr gut überein, was erst ab 32 s/s wieder der Fall ist, bei dem Tempo, bei dem "fast nichts" mehr verstanden wird (Abbildung 6). Der "Mismatch" zwischen subjektiven und objektiven Maßen der Verständlichkeit an dieser Stelle zeigt, dass es unterschiedliche Auffassungen zu dem allgemeinen Konzept von Verständlichkeit gibt, die man methodisch unterschiedlich angehen kann, auch wenn hier immer von Textverständlichkeit und nicht von Wort- oder Lautverständlichkeit ausgegangen wurde. Die Lösung dieses methodischen Problems wäre aber Voraussetzung für die genaue Beantwortung der Frage, wie schnell tempo-manipulierte Sprache noch sein darf, dass sie von bestimmten Personen verstanden wird.

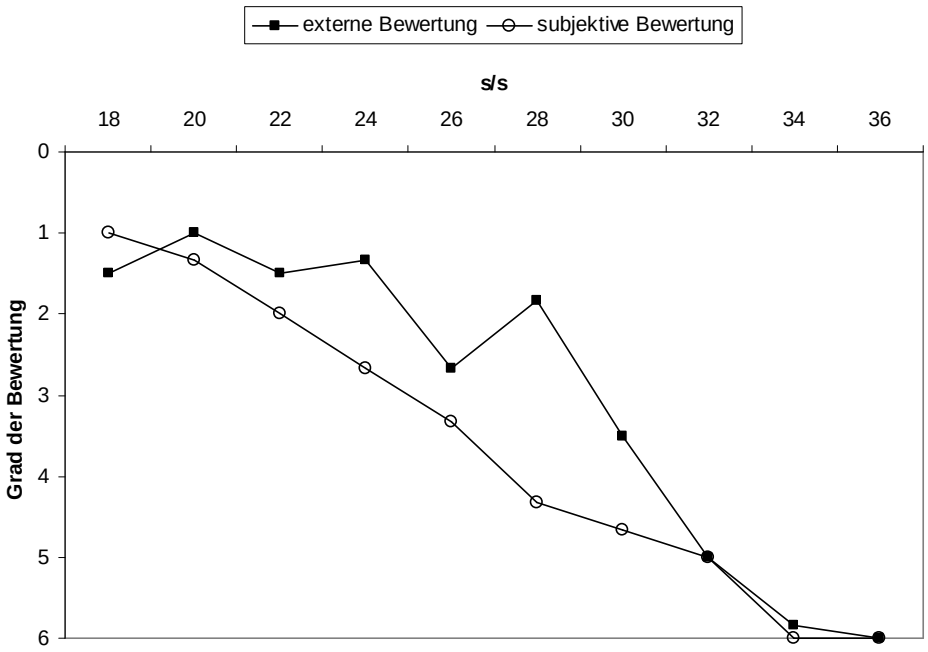

Abbildung 6 – Subjektive Bewertung (aus Abbildung 3) und externe Bewertung (aus Abbildung 5) zusanmmengefasst.

Auch wenn diese Frage nicht in der gewünschten Genauigkeit beantwortet wird, so kann festgehalten werden, dass es sich um eine außergewöhnliche Hörfertigkeit handelt, die unsere Vp mit etlichen blinden Menschen teilt. Ebenso steht fest, dass diese Hörfertigkeit nicht auf Erfahrung mit menschlicher Sprachproduktion zurückzuführen ist, da menschliche Artikulationsgeschwindigkeit üblicherweise 8-10 s/s nicht überschreitet.

Ein langfristiges Training mit Sprachsynthese scheint hier eine entscheidende Rolle zu spielen. Dieser Befund deutet auf eine viel größere neuronale Plastizität bei der Sprachperzeption hin als bislang angenommen. Blinden kommt hierbei auch die Nutzung des visuellen Kortex zu Gute, wie auch für unsere Vp in einer anderen Studie nachgewiesen [8]. Ob ein intensives Training für alle Personen zu solchen außergewöhnlichen Fähigkeiten führt, müssen zukünftige Studien klären. Dabei sind Variablen wie die Dauer des Trainings (unsere Vp "trainiert" bereits seit 13 Jahren), Sehfähigkeit und auch Zeitpunkt der Erblindung zu berücksichtigen.

Eine weitere Fragestellung betrifft den Einfluss der Prosodie auf die Verständlichkeit ultraschneller Sprachsynthese. Dabei ist nicht nur an die Beseitigung falsch gesetzter Satzakzente und eine Erweiterung des Tonhöhenumfangs zu denken, sondern auch an eine nicht-lineare Veränderung der Lautdauern sowie die Einfügung sehr kurzer Pausen zur besseren Dekodierung von Phrasengrenzen (vgl. [7] und [9]).

Eine Verbesserung der Prosodie betrifft nicht nur sehr schnelle Sprache, die mit Formantsynthese erzeugt wird, sondern auch tempo-skalierte natürliche Sprache, wie sie auf Hörbüchern und anderen Audio-Dateien vorliegen. Für die Erzeugung ultra-schneller Sprache mittels Unit-Selection-Synthese kann eine zusätzliche Datenbank mit schnell artikulierter Sprache eines natürlichen Sprechers eine Verbesserung bringen [9, 10].

Da unter Blinden (in Deutschland) jedoch Formantsynthese sehr weit verbreitet ist und innerhalb dieser Benutzergruppe ein sehr starker Bedarf an Temporegulierung besteht (vgl. [10]), sollte auf Forschung mit dieser "altertümlichen" Synthesemethode nicht verzichtet

werden, zumal auch hier Erkenntnisgewinne zu erwarten sind, wie Menschen tempo-manipulierte Sprache im Allgemeinen verarbeiten.

Literatur

[1] Trouvain, J.: On the comprehension of extremely fast synthetic speech. *Saarland Working Papers in Linguistics* 1, 2007, pp. 5-13. (http://scidok.sulb.uni-saarland.de/sulb/portal/swpl/)

[2] Moos, A. & Trouvain, J.: Comprehension of ultra-fast speech – blind vs. "normally hearing" persons. Proceedings *16th International Congress of Phonetic Sciences* (ICPhS), Saarbrücken, 2007, pp. 677-680. (http://www.icphs2007.de/conference/Papers/1186/1186.pdf)

[3] JAWS (Job Access With Speech) Screenreader software http://www.freedomsci.de besucht am 21.01.06.

[4] Praat version 4.5 http://www.fon.hum.uva.nl/praat/ besucht am 10.01.07

[5] Charpentier, F., Moulines, E.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Proc. *Eurospeech* (2), 1989, pp. 13-19.

[6] Janse, E.: *Production and Perception of Fast Speech*. Lot, Utrecht. 2003.

[7] Trouvain, J.: *Tempo Variation in Speech Production. Implications for Speech Synthesis.* (Phonus 8, Reports in Phonetics). Universität des Saarlandes, Saarbrücken, 2004.

[8] Moos, A., Hertrich, I., Dietrich, S., Trouvain, J. & Ackermann, H.: Perception of ultra-fast speech by a blind listener – does he use his visual system? 8th *International Speech Production Seminar* (ISSP), Strasbourg, 2008.

[9] Moers, D., Wagner, P. & Breuer, St.: Assessing the adequate treatment of fast speech in unit selection speech synthesis systems for the visually impaired. Präsentation beim *6th ISCA Workshop on Speech Synthesis*, Bonn, 2007.

[10] Nishimoto, T., Sako, S., Sagayama, S., Ohshima, K., Oda, K. & Watanabe, T.: Effect of learning on listening to ultra-fast synthesized speech. Proceedings *28th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society*, 2006, pp. 5691-5694.