VERSTEHBARKEITS-UNTERSUCHUNG SYNTHETISCH GEGLÄTTETER INTONATIONSSTRUKTUREN IN ZEITSKALIERTER SPRACHE

Dirk Höpfner

BTU Cottbus, Institut für Elektronik und Informationstechnik
hoepfner@tu-cottbus.de

Abstract: Dieser Beitrag untersucht den Einfluss der Intonation auf die Verstehbarkeit zeitskalierter natürlicher Sprache. Mit Hilfe zweier Algorithmen wird in natürlichen Sprachbeispielen eines geschulten Sprechers die Grundfrequenzdynamik in drei Stufen von natürlich bis auf den Wert einer mittleren Grundfrequenz des Sprechers eingeebnet und anschließend mit vier Geschwindigkeitsstufen beschleunigt. In einem Hörtest wird evaluiert, inwieweit die Intonationsveränderungen Einfluss auf die Verstehbarkeit nehmen. Im Gegensatz zu häufig geäußerten Behauptungen und vor allem Wünschen von blinden und sehbehinderten Probanden erhöht sich die Verstehbarkeit von Sprachaufnahmen bei hohen Sprechraten und gleichzeitig monotonem Grundfrequenzverlauf gegenüber natürlicher Intonation nicht. Die Ergebnisse des Hörtest, der ausschließlich mit normal sehenden, in Bezug auf Sprachskalierung unerfahrenen Probanden durchgeführt wurden, wird nachfolgend diskutiert.

1 Einführung

Zeitskalierte Sprache ist eine wichtige Kommunikationskomponente für blinde und sehbehinderte Menschen. Sie nutzen diese Form der Sprachausgabe meist in synthetischer Form in Screenreadern ihres PC's, in Sprachausgaben von Menüs der Handys oder PDA's. Im Daisy-Hörbuch-Standard [1] ist eine beschleunigte Sprachsynthese für die interaktive Menüstruktur des Hörbuches sowie eine zeitskalierbare Wiedergabe des eigentlichen Hörbuches implementiert. Aber auch bei der täglichen Nutzung von wiederholten Informationsausgaben wie z.B. Menüerläuterungen in Navigationssystemen und Telefonservices sollten unterschiedliche Sprechraten nutzerabhängig wählbar sein.

Für die Zeitskalierung von Sprache wurde in [2] ein nichtlinearer Algorithmus entwickelt, der in Abhängigkeit stimmhafter oder stimmloser Laute bzw. Sprache und Sprachpause unterschiedliche Skalierungsfaktoren in einer bereits gespeicherten sprachlichen Äußerung anwendet. Ziel war es, möglichst alle Laute zu erhalten und weitgehend redundante Sprachanteile bei der Beschleunigung zu eliminieren. Somit wird gerade bei hohen Faktoren die Verstehbarkeit gespeicherter natürlicher Sprache gegenüber den linearen Skalierungsverfahren verbessert. Dieser Algorithmus wurde auch für die hier vorgestellten Untersuchungen genutzt.

Während der bisher zahlreich durchgeführten Verstehbarkeits- und Verständlichkeits-Untersuchungen zeitskalierter Sprache mit blinden Probanden [3], [4] tauchte immer wieder die Vermutung und die Hoffnung der Probanden auf, dass die beschleunigte Sprache mit abnehmenden Intonationschwankungen besser verstehbar sein würde. Dies resultiert in erster Linie aus den bisherigen Erfahrungen der blinden Probanden im Umgang mit der vielfach noch verwendeten und unnatürlich monoton wirkenden Formantsynthese. Betrachtet man Abb. 1a und 1b (Markierung) erkennt man die steileren Anstiege des GF-Verlaufs in den Nulldurchgangsstellen nach der Beschleunigung der Sprache auf die doppelte Sprechrate. Es könnte demnach realistisch erscheinen, dass ein flacherer Verlauf der Grundfrequenz F_0 (s. Abb. 1c) die Verstehbarkeit erhöhen könnte. Dies könnte durch eine geringere Amplitude des Grundfrequenzverlaufs um eine berechnete mittlere Frequenz bis hin zum vollständig monotonen Verlauf realisiert werden. Die Hypothese wäre, dass sich die Verstehbarkeit bei abnehmender Intonation erhöht.

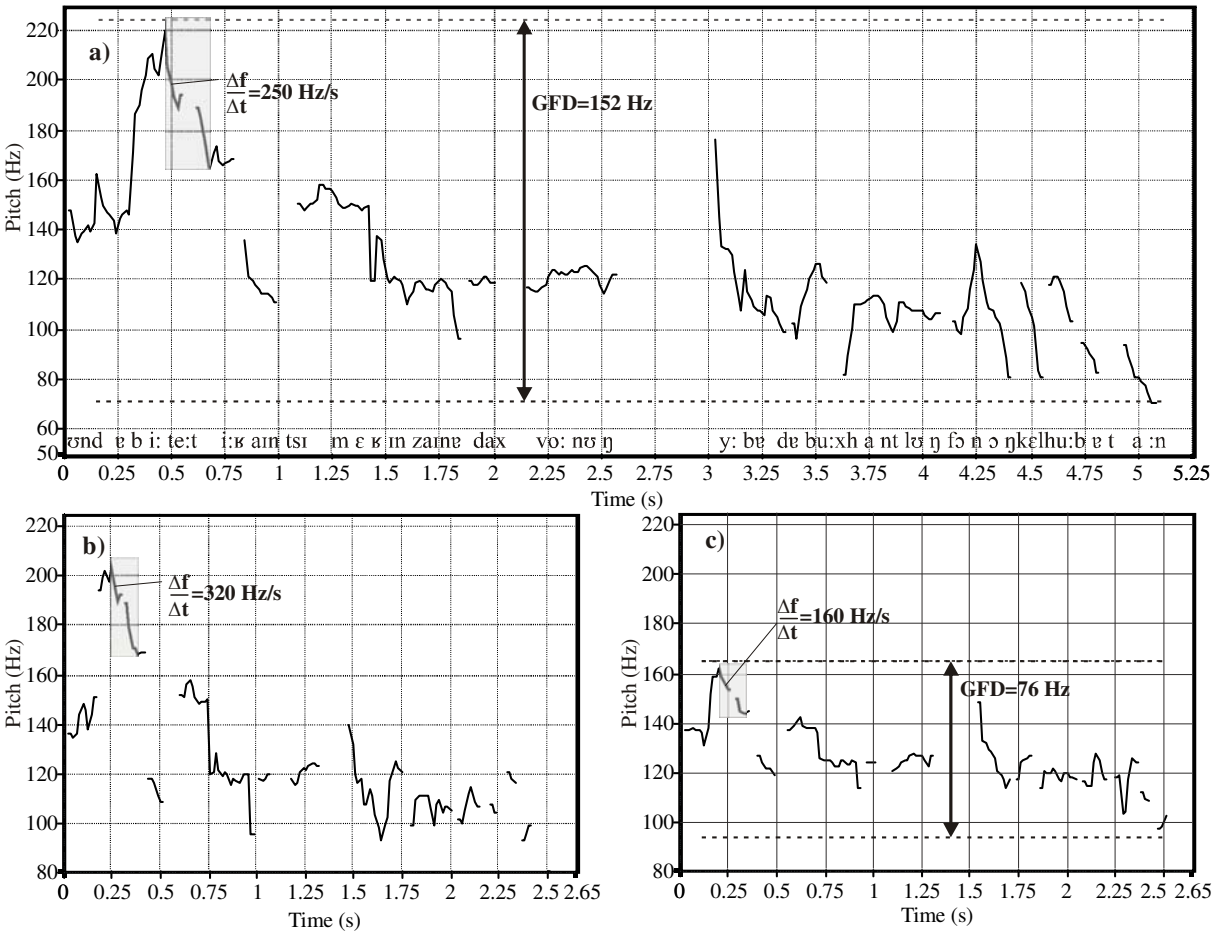

Abb. 1 - Gegenüberstellung der GF-Verläufe bei normaler Sprechrate von 4 Silb/s (a), nach der Beschleunigung auf 8 Silb/s mit überwiegender Kürzung stimmhafter Laute (b) und nach Verringerung der GF-Dynamik auf 50% (c)

2 Pitch-Shifting und zeitliche Skalierung der Sprachdaten

Für eine nutzerabhängige variable Änderung der GF-Dynamik während der Wiedergabe wurde ein eigener Algorithmus implementiert. Mit diesem Verfahren ist es möglich, die Grundfrequenz F_0 des Sprechers sowie den Grundfrequenzbereich kontinuierlich zu verändern. Nach der Einstellung der *Grundfrequenzdynamik GFD* in eine natürliche, mittlere und völlig monotone Stufe, wird das Sprachbeispiel mit einem nichtlinearen Skalierungsverfahren beschleunigt. Die Bearbeitungsschritte der im Hörtest verwendeten Sprachsignale sind in Abb. 2 erkennbar.

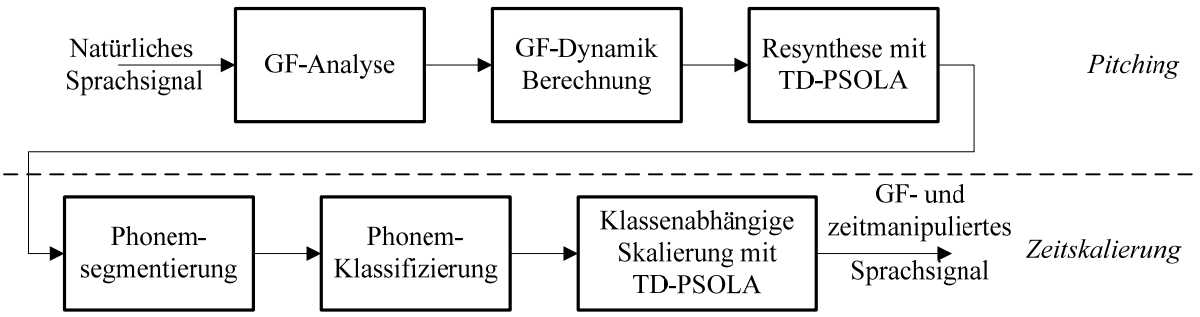

Abb. 2 - Ablauf der Manipulation des Sprachsignals: Verringerung der GF-Dynamik und zeitliche Skalierung

2.1 GF-Analyse, Neuberechnung der GF-Dynamik

Ziel der GF-Manipulation ist die Einengung der *GF-Dynamik* um einen mittleren F_0-Wert der über einen bestimmten Zeitbereich des Sprachsignals berechnet wird.

Die Manipulation der Grundfrequenz F_0 bedarf zunächst der Berechnung des F_0-Verlaufs. Für den zu entwickelnden Algorithmus – mit der Forderung einer echtzeitfähigen manuellen Manipulation des GF-Verlaufs während der Wiedergabe – wurden zunächst zahlreiche Verfahren auf ihre Genauigkeit untersucht. Für die F_0- Verschiebung in einer möglichst korrekt intonationsmodifizierten synthetisierten Sprache ist eine hohe Genauigkeit der *GF-Analyse (GFA)* notwendig. Fehlberechnungen der GF-Marken können bei der Resynthese zu drastischen Sprüngen des GF-Verlaufs führen. In [5] wurden einige GFA-Verfahren hinsichtlich ihrer Genauigkeit evaluiert. Kein Verfahren lieferte fehlerfreie Ergebnisse. Bei ungestörten Signalen mit einer bandbegrenzenden Vorverarbeitung arbeitete das *Autokorrelationsverfahren (AKF)* mit einer ähnlichen Fehlerrate wie die drei weiteren Frequenzbereichsverfahren: *4. Wurzel des Leistungsdichtespektrum, Amplitudenspektrum* und *Cepstrum.* Bei den Verfahren handelt es sich um die sogenannten *Nichtlinearen Modifizierten Doppelten Spektraltransformations-Funktionen (NMDSF)*, denen eine nichtlineare Veränderung des Spektrums gemeinsam ist.

Für die GFA wurde ein Algorithmus ausgewählt, der in der frei verfügbaren Software *PRAAT* [6] implementiert ist und auf der Basis der AKF arbeitet. Die GFA wurde in [7] als sehr genau arbeitendes Verfahren für Frequenzen von 16 Hz (sog. „Strohbassregister“) bis zu 1800 Hz (schreiende Kleinkinder) vorgestellt, also auch sehr robust gegenüber tiefen Formantfrequenzen, die sich bis in den F_0-Bereich erstrecken können. Eigene Experimente mit den stärker spektral eingeebneten NMDS-Funktionen ergaben keine signifikanten Vorteile, bezogen auf die Genauigkeit oder die Berechnungsgeschwindigkeit bei ungestörter Sprache. Die Genauigkeit dieses Praat-Algorithmus überzeugte in der Resynthese des GF-manipulierten Signals gegenüber einem ähnlichen GFA-Algorithmus mit LPC- und Tiefpassfilterung in der Vorverarbeitung. Eine scriptbasierte Verwendungs- und die Implementierungsmöglichkeit der Pitchmarken an die nachfolgende GFD-Berechnung führten letztendlich zur Entscheidung für die Verwendung der Praat-GFA.

Die Pitchmarken werden nach der Analyse eines Sprachfiles einer einfachen Textdatei übergeben und konnten somit ohne großen Aufwand in der sich anschließenden Neuberechnung der GF für die Resynthese (s. Abb. 2) verwendet werden. Zunächst dient die F_0 zur Ermittlung der GFD und der Berechnung der neuen Grundfrequenzen in Abhängigkeit eines gewählten Skalierungsfaktors α.

Für die in einem bestimmten Zeitintervall ermittelte GF-Dynamik musste zunächst die mittlere Grundfrequenz F_{0m} errechnet werden, die als Bezugswert für die neuen minimalen und maximalen Grundfrequenzen diente:

$$F_{0m} = 2^{\left(\frac{\sum_{i=1}^{k-1} \log_2 \left(F_0(i)\right)}{k-1}\right)} \qquad 2.1$$

Die mittlere Grundfrequenz F_{0m} wird in der Gl. 2.1. zunächst logarithmisch und somit gehörrichtig berechnet. Angenommen die untere GF beträgt $F_{0u} = 50$ Hz und die obere GF beträgt $F_{0o} = 200$ Hz, so wäre das arithmetische Mittel $\overline{F_0} = 125$ Hz. Wahrnehmungskonform entspricht die obere Frequenz genau zweimal der Verdopplung der unteren Frequenz. Somit beträgt der Abstand beider Frequenzen zwei Oktaven. Die wahrgenommene mittlere Frequenz entspricht einer Oktave: $F_{0m} = 100$ Hz, wie mit Gl. 2.1 berechnet. Innerhalb eines wählbaren Zeitbereichs von $t = \frac{1}{16} \ldots 100$ s wird die Gesamtanzahl k aller stimmhaften i-Samples in diesem Bereich betrachtet. Eine zusätzliche Logarithmierung und anschließende Potenzierung verlangt allerdings erheblich mehr Rechenleistung und -zeit. Um hier Ressourcen zu sparen, wurden von mehreren männlichen und weiblichen Sprechern die arithmetischen und loga-

rithmischen Mittenfrequenzen berechnet und verglichen. Dabei entstand eine maximale GF-Differenz von $\Delta F_0 \approx 3\%$. Dieser Unterschied kann im unteren Hörbereich bis 500 Hz bei einer wahrnehmbaren Frequenzdifferenz ab 2 Hz [8] vernachlässigt werden. Somit lässt sich bei Anwendung des arithmetischen Mittels der Berechnungsaufwand minimieren. Aus der PRAAT-Prozedur werden wie in Tab.1 die Pitchmarken entnommen.

i	t_i (ms)	$T_0 = t_{i+1} - t_i$ (ms)	$F_0 = \frac{1}{T_0}$ (Hz)
1	0	0,0083	120
2	0,0083	0,0083	120
3	0,0166	0,0084	119
4	0,025	0,1851	5
5	0,2101	0,0056	179
6	0,2157	0,0055	182
7	0,2212	-	-

Tab. 1 - Ausgabe der ersten sechs Pitchmarken eines Sprachfiles mit dem PRAAT-Algorithmus

Die mittlere Grundfrequenz lässt sich, bezogen auf die Auftretensdauer des jeweiligen GF-Wertes, nun recht einfach berechnen:

$$\overline{F_0} = \frac{\sum_{i=1}^{k-1}\left[(t_{i+1} - t_i) \cdot \frac{1}{(t_{i+1} - t_i)} \cdot c(i)\right]}{\sum_{i=1}^{k-1}[(t_{i+1} - t_i) \cdot c(i)]} = \frac{\sum_{i=1}^{k-1} \cdot c(i)}{\sum_{i=1}^{k-1}[(t_{i+1} - t_i) \cdot c(i)]} \qquad 2.2$$

$$mit: c(i) = \begin{cases} 1, & wenn\ F_0 > F_{0min} \\ 0, & wenn\ F_0 \leq F_{0min} \end{cases}$$

In der fortlaufenden Aufzählung der Zeitmarken für die GF-Perioden sind auch die stimmlosen Sprachanteile enthalten (s. Tab. 1 bei $i = 4$). Mit dem Entscheidungsfaktor $c(i)$ wird anhand des unteren Schwellwertes F_{0min} festgelegt, ob es sich um einen stimmlosen ($c(i) = 0$) oder stimmhaften ($c(i) = 1$) Abschnitt zur nachfolgenden Pitchmarke handelt. Eine Manipulation der Grundfrequenz erfolgt in der Resynthese des Sprachsignals durch Stauchung oder Streckung des Signals zwischen den Pitchmarken. Bleibt man bei der weiteren Verarbeitung mit dem *Time Domain-Pitch Synchronous Overlapp And Add (TD-PSOLA)*-Verfahren im Zeitbereich, müssen zunächst die neuen Grundfrequenzperioden nach Formel 2.3. berechnet werden:

$$T_{0syn}(i) = \overline{T_0} \cdot 2^{\left(\beta \cdot \log_2 \frac{T_0(i)}{\overline{T_0}}\right)} \qquad 2.3$$

β stellt den Stauchungs- oder Streckungsfaktor der GFD dar. Ein Wert von $\beta = 0{,}5$ ist mit einer Stauchung der GFD um 50% verbunden. Mit diesen neu berechneten GF-Perioden wird die Resynthese vorgenommen.

2.2 Resynthese der GF-manipulierten Sprache mit TD-PSOLA

Die Anforderungen an die Resynthese des GF-manipulierten Sprachsignals waren ein möglichst natürlich klingendes Sprachsignal und eine echtzeitfähige Verarbeitung. In späteren Hörversuchen sollen sich die Probanden die günstigste GFD während der beschleunigten Wiedergabe von Sprachfiles in CD-Qualität einstellen können. Für die Manipulation kamen Zeitbereichs- und Frequenzbereichsverfahren in Frage. Die Short Time Fourier-Transformation (STFT) des Sprachsignals, also Manipulation der GF im Frequenzbereich mit anschließender inverser STFT - nach dem Vocoder Prinzip - wurde auf Grund des Berechnungs-Aufwandes und der gegenüber dem TD-PSOLA-Verfahren nicht besseren Qualitätser-

gebnissen [9] verworfen. Abb. 3 veranschaulicht die Vorgehensweise der GF-Manipulation, hier am Beispiel des Wortes „ein“.

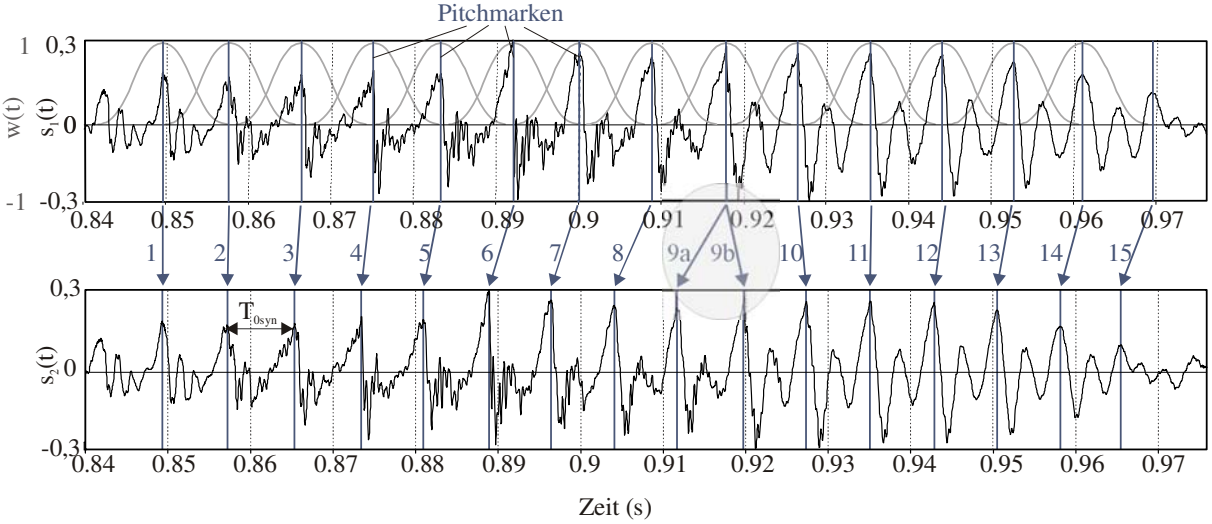

Abb. 3 - Verschiebung der GF im Sprachsignal (oben Original $\boldsymbol{s_1(t)}$, unten gepitchtes Signal $\boldsymbol{s_2(t)}$)

Die GF-Periodendauer wird zwischen den einzelnen Pitchmarken in $s_2(t)$ überwiegend verringert. Das Sprachsignal wird an der Pitchmarke mit der Hanningfunktion $w(t)$ und der Zielperiodenbreite T_{0syn} gefenstert und benachbarte Pitchmarken werden im synthetisierten Signal $s_2(t)$ mit monotoner Intonation überblendend zusammenaddiert. Als problematisch könnte sich bei der Synthese der neuen GF-Periode T_{0syn} die Streckung mit TD-PSOLA auswirken. Bei einer Periodenverlängerung werden die benachbarten Amplitudenspitzen nicht ausreichend durch die Fensterfunktion ausgeblendet, was bei einer maximalen Verbreiterung der GF-Periode von 50% gegenüber der Originalperiode jedoch noch nicht störend wirkt. In dem vorgestellten Beispiel in Abb.1a und 1c wird z.B. bei Halbierung der GFD die GF von $F_{0\max} = 220$ Hz auf 165 Hz eingeengt, was einer Verbreiterung der Periodendauer von 4,5 ms auf 6,1 ms entspricht, also einer Vergrößerung um 36%. Da in dem gezeigten Beispiel in Abb. 3 die Periodendauer überwiegend verringert wird, kommt es gleichzeitig zur zeitlichen Dezimation des Sprachsignals. Dem wird durch einfaches Kopieren von GF-Segmenten (Pitchmarke 9a+9b) entgegengewirkt.

2.3 Zeitskalierung

Das Kopieren und Entfernen von Pitchsegmenten ist auch das Prinzip der nachfolgenden TD-PSOLA Anwendung im Zeitskalierungsalgorithmus von [2] (s. Abb. 2).

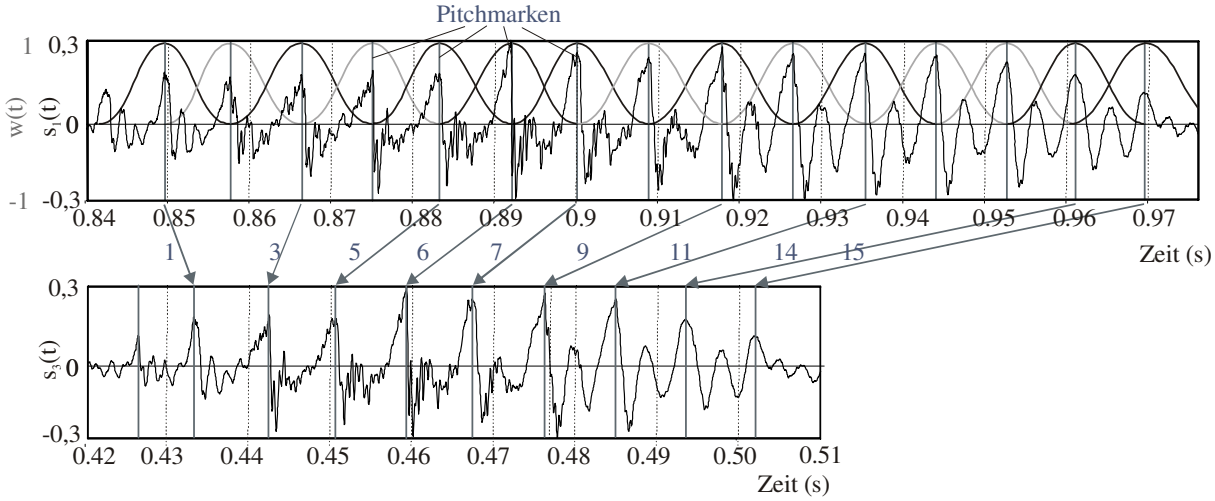

Abb. 4 - Zeitskalierung mittels TD-PSOLA (oben Original $\boldsymbol{s_1(t)}$, unten beschleunigtes Signal $\boldsymbol{s_3(t)}$)

Die Skalierung findet phonemabhängig, demzufolge nichtlinear statt. Das bedeutet, bei einem eingestellten Skalierungsfaktor von z.B. $\alpha = 2$ (doppelte SR) werden die Vokale mit $\alpha_V = 2{,}5$ und die Konsonanten mit $\alpha_K = 1{,}7$ beschleunigt, vorausgesetzt beide Lautformen besitzen das gleiche Längenverhältnis im Sprachsignal. Diese optimalen Verhältnisse wurden bereits in [2] ermittelt.

Weiterhin werden nach Sprachpausen und Sprache unterschieden und hier die Sprachpause stärker komprimiert. In Abb. 4 wird die Zeitskalierung mittels TD-PSOLA verdeutlicht und die Ähnlichkeit mit dem Pitching in Abb. 3 sichtbar. In dem Abschnitt wird die Sprechgeschwindigkeit ungefähr verdoppelt. Auf Grund der lautabhängigen Skalierungen wird jedoch nicht jedes zweite Pitchsegment verworfen (s. Pitchmarken 5, 6, 7).

3 Hörtest mit GF-manipulierten und beschleunigten Sprachdaten

3.1 Sprachmaterial und Probanden

Ziel des Tests sollte die Messung der Verstehbarkeit von skalierter Sprache mit unterschiedlichen Geschwindigkeiten und GF-Dynamiken sein. Dazu sollten vier Geschwindigkeiten von 8, 10, 12 und 14 Silben/s mit jeweils den jeweiligen GF-Dynamiken 100%, 50% und 0% präsentiert werden.

Zur Erstellung des sprachlichen Testmaterials wurden 12 Sätze von einem Schauspieler in zwei Varianten gesprochen: mit natürlicher Intonation und mit möglichst monotoner Intonation. Nach Messung der GF-Verläufe der monotonen Version konnte festgestellt werden, dass die GF gegenüber der natürlich normalen Intonation nur um ca. 50% eingeengt war. Die Stufen 50% und 0% GF-Dynamik wurden von der natürlich monotonen Variante erstellt, so dass sich für die Sprachdarbietung eine natürlichere Qualität gegenüber einer vollständigen Verringerung der GFD mit TD-PSOLA ergab.

Die Inhalte der zu verstehenden Texte besaßen unterschiedliche Schwierigkeitsgrade aus den Bereichen Spielbeschreibung, Filmerzählung, populärwissenschaftliche Universitätsstudien sowie Nachrichten aus der Informatik.

Insgesamt konnten für den Hörtest 54 normal sehende Probanden von 20-45 Jahre gewonnen werden. Da überwiegend Studenten daran teilnahmen, war das Alter von 22-27 Jahre sehr häufig vertreten. Blinde Personen, die eventuell bereits Erfahrung mit zeitskalierter Sprache besitzen, wirkten an dem Hörtest zur Verringerung der Parameterstreuung nicht mit, worauf später noch eingegangen wird.

3.2 Hörtestablauf

Grundsätzlich ist bei der inhaltlichen Abfrage wahrgenommener sprachlicher Äußerungen zwischen Verständlichkeits- und Verstehbarkeitstests zu unterscheiden. Ein Verständlichkeitstest beschäftigt sich mit der Erkennbarkeit segmentaler Lautsprache wie einzelne Phoneme, Silben, Wörter oder Wortgruppen. Beim Verstehbarkeitstest geht es um das inhaltliche Verstehen eines Satzes oder mehrerer Sätze. Dies entspricht in der Regel auch der Anwendung zeitskalierter Sprache, wie z. B. die Ausgabe von komplexen Anweisungen, das Verlesen von Nachrichten oder das Hören literarischer Werke. Ein Problem bei der Verstehbarkeit ist dessen Messbarkeit. Eine mündliche Wiederholung der wahrgenommenen Sprache und gleichzeitige Audiospeicherung wäre eine Möglichkeit. Die anschließende Auswertung durch einen Versuchsleiter beherbergt eine Reihe von Fehlermöglichkeiten zwischen der Äußerung des Probanden und der Verständlichkeit der Audioaufzeichnung beim Auswerter. Eine schriftliche Protokollierung direkt durch den Probanden wäre zu aufwendig und würde die Dauer des Hörtests von maximalen und sinnvollen 30 min übersteigen. Aus diesem Grund können nur Stichproben des Gehörten abgefragt werden und es wurde ein Multiple-Choice Test ausgewählt. Die Probanden mussten nach einem gehörten maximal 30 s langen Abschnitt vier

Fragen beantworten. Für die Beantwortung wurden vier Möglichkeiten vorgelegt, von der eine Antwort richtig war. Die richtige Antwort wurde in einigen Fällen auch nicht wörtlich vom Sprecher geäußert, so dass es neben der Verständlichkeit auf die logischen Fähigkeiten der Probanden ankam. Jeder Proband hatte demnach 48 Fragen auf einem Fragebogen zu beantworten.

Die Sprachbeispiele wurden in drei Blöcke aufgeteilt, so dass sich die Sätze bei keinem der Probanden wiederholten. Die zeitskalierte und GF-manipulierte Sprache wurde dem Probanden über Studio-Kopfhörer in CD-Qualität vorgespielt.

Dem eigentlichen Verstehbarkeitstest war ein Verständlichkeitstest als Trainingsphase vorgelagert, da viele Probanden bisher noch keine Erfahrung mit zeitskalierter Sprache hatten. In vergangenen Versuchen wirkte sich eine zu kurze Trainingsphase erheblich auf die Ergebnisse aus. Der Hörtest dauerte im Durchschnitt 32 Minuten.

4 Hörtestergebnisse

Unter Ausblendung der Ergebnisse eines fachlich sehr anspruchsvollen IT-Textes, bei dem in allen Sprechraten schlechte Verstehbarkeitswerte ermittelt wurden, konnten folgende Häufigkeiten richtig gegebener Antworten (Tab. 2) festgestellt werden:

SR (Silb/s)	8	10	12	14	SR1 8, 10	SR2 12, 14
GFD=100 %	76,3	75,6	85,4	72,7	75,9	77,8
GFD=50%	83,3	81,5	74,7	61,1	82,4	69,9
GFD=0%	88,1	86,9	59,1	63,4	87,3	61,4

Tab. 2 - Relative Häufigkeiten (in %) richtiger Antworten unterschieden in GFD und SR (Sprechrate) im Verstehbarkeitstest

Betrachtet werden zunächst die Ergebnisse der einzeln getesteten Sprechraten von 8-14 Silb/s. Die nach X^2-Test sehr signifikanten Ergebnisse ($p<0{,}01$) bei einer GFD von 50% und 0% und der Abfall der Verstehbarkeit zu höheren Sprechraten hin können die anfangs gestellte Hypothese nicht aufrecht erhalten. Bei komplett erhaltener Intonation (GFD=100%) bleibt die Verstehbarkeit auch bei den höheren Sprechraten von 12 und 14 Silben/s erhalten. Über alle Sprechraten ergibt sich keine signifikante Abweichung. Klassifiziert man die vier Sprechraten in eine moderat beschleunigte Sprechrate SR1 (8 und 10 Silb/s) und eine sehr schnelle Sprechrate SR2 (12 und 14 Silb/s) werden die Unterschiede zwischen den unterschiedlichen GF-Dynamiken noch besser herausgearbeitet (s. Tab. 2 rechts).

Auffallend ist weiterhin eine bessere Verstehbarkeit bei SR1 bei abnehmender GF-Dynamik. Die Ursache dafür ist jedoch nicht nur in der Intonation zu suchen, sondern insbesondere dem Trainingseffekt während des Versuches zuzuordnen. Der sehr langwierige Trainings- und Lerneffekt bei zeitskalierter Sprache besitzt großen Einfluss auf die Hörtestergebnisse. Die Gewöhnungsphase kann nicht innerhalb eines Hörtests abgeschlossen werden. Dies ist ein Prozess, der sich über Wochen bzw. Monate ausdehnen kann [10]. Aus diesem Grund kann es trotz ungewohnter Beschleunigung der Sprache bis zu sehr schnellen 12 Silb/s zu ansteigenden Verstehensleistungen kommen. Umso stärker ist Verstehensabhname bei Berücksichtigung des Trainingseffektes aber abnehmender Intonation zu gewichten.

Die Intonation als Teil der Prosodie sowie als melodische Erscheinung der Sprache hat eine semantische Bedeutung in ihren unterschiedlichen segmentalen Ebenen. In der suprasegmentalen Ebene unterstützt eine Klangverfärbung der Stimme, also die Änderung der GF und deren Oberwellen, emotionale Ausdrucksformen. Diese sind hier jedoch nicht Betrachtungsgegenstand des Tests. Prosodiemerkmale, darunter die Intonation, zeichnen sich auch durch eine bedeutungsunterscheidende Struktur aus. Innerhalb der Wortebene zeichnet die Prosodie verantwortlich für die Distinktion der Wortkomponenten und sogar für lexikalische Betonungen.

Versteht man diese Akzente und Markierungen auch als Information, die dem Rezipienten übermittelt werden soll, so führen Veränderungen der Intonation zwangsläufig zu Informationsverlusten oder Fehlinterpretationen. Eine Geschwindigkeitserhöhung des Sprachmaterials ist zusätzlich auf diese Kontext unterstützenden Akzentuierungen angewiesen, da gerade die Verständlichkeit auf der Silben- und Phonemebene erschwert wird. Demnach könnte eine Monotonisierung der Intonation zu Verstehbarkeits-Verlusten im Satzgefüge führen. Gerade der kontextuelle Zusammenhang der Laut- bzw. Wortgruppen wird durch die Intonation bekräftigt und ist bei der zeitlichen Skalierung von Sprache nicht zu vernachlässigen.

5 Zusammenfassung, Ausblick

Die Hypothese, dass beschleunigt wiedergegebene Sprache bei monotonem Grundfrequenzverlauf gegenüber der natürlichen Intonation verständlicher wahrgenommen wird, konnte in diesem Versuch nicht bestätigt werden. Bleibt die Frage nach dem Wunsch und der Ursache der scheinbar höheren Erkennungsraten blinder und mit beschleunigter Sprache erfahrener Personen. Diese beruht ausschließlich auf der Erfahrung derjenigen Personen, die z. B. täglich die in Userinterfaces eingebaute skalierte Sprache nutzen. Ihnen wird jedoch in derzeit erhältlichen Systemen überwiegend die skalierbare Formantsynthese angeboten, die - technologisch bedingt - nur eine geringe natürliche Intonation aufweist.

Für weitere Untersuchungen diesbezüglich, sollten auch blinde Personen in den Probandenkreis mit einbezogen werden. Dieses setzt allerdings einen großflächigen Feldversuch voraus, um eine möglichst repräsentative Anzahl an blinden Probanden zu erreichen, die Erfahrungen mit skalierter Sprache haben, bzw. darauf trainiert sind. Hier bietet sich ein Online-Versuch im Internet an.

Weiterhin kann aus den Ergebnissen geschlussfolgert werden, dass eine natürlichere Intonation in Sprachsynthese-Systemen mit der hohen Wiedergabegeschwindigkeiten, wie von blinden Nutzern vorliebend genutzt, die Verstehbarkeit steigern würde.

6 Literatur

[1] Daisy Consortium: ANSI/NISO Z39.86 Specifications for the Digital Talking Book. URL: http://www.daisy.org/z3986/ v. 15.06.2009

[2] Höpfner, D.: *Nichtlinearer Zeitskalierungsalgorithmus für gespeicherte natürliche Sprache*. Lacroix, A. (Hrsg.): Elektronische Sprachsignalverarbeitung. Frankfurt/Main, 2008, S.199-207

[3] Moos, A.; Trouvain, J.: *Einzelfallstudie zu Grenzen der Verständlichkeit ultraschneller Sprachsynthese.* Lacroix, A. (Hrsg.): Elektronische Sprachsignalverarbeitung. Frankfurt /Main, 2008, S. 207-214

[4] Höpfner, D.: *Untersuchungen zeitskalierter Sprachwiedergabe mit normal sehenden, sehbehinderten und blinden Probanden.* Fellbaum, K. (Hrsg.): Elektronische Sprachsignalverarbeitung. Cottbus, 2007, S. 235-242

[5] Vary, P.; Heute, U.; Hess. W.: Digitale Sprachsignalverarbeitung. Teubner Verlag, Stuttgart, 1998

[6] Boersma, P.; Weenink, D.: *Praat. Doing Phonetics by Computer.* URL: http://www.fon.hum.uva.nl/praat/, Netherland, v. 01.07.2009

[7] Boersma, P.: *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.* Proceedings of the Institute of Phonetic Sciences 17, Amsterdam, 1993, S. 97-110

[8] Zwicker, E.; Fastl, H.: *Psychoacoustics: Facts and Models.* Springer Verlag, Berlin, Heidelberg, New York, 1999

[9] Moulines, E.; Laroche, J.: *Non-parametric techniques for pitch-scale and time-scale modification of speech.* Speech Communication, Vol.16, Netherland, 1994, S.175-205

[10] Friedman, H. L.: *Further Research on Speeded Speech as an Educational Medium.* Report No. 5-0801 U.S. Office of Education, AIR, Silver Spring, 1967