TEMPOSTEUERUNG IN DER SPRACHSYNTHESE DURCH PROSODISCHE PHRASIERUNG

Jürgen Trouvain

Institut für Phonetik, Universität des Saarlandes
trouvain@coli.uni-sb.de

Abstract: Es ist vielfach wünschenswert, das Tempo synthetischer Sprache ändern zu können. In den meisten Synthesesystemen wird die Geschwindigkeit in linearer Weise angepasst, wohingegen in natürlicher Sprache Tempoänderungen nicht-linear vorkommen. Erste Modelle, die lediglich die Vorhersage und Realisierung prosodischer Phrasengrenzen berücksichtigen, werden für zwei langsame und zwei schnelle Tempi dargestellt. In Hörtests werden linear zeitskalierte mit modell-basierten Stimuli verglichen, die mit dem deutschsprachigen Synthesiser "Mary" [12] erzeugt wurden. Die Ergebnisse zeigen, dass die phrasen-modifizierten Versionen bei sehr langsamen Tempo den linearen bevorzugt werden. Bei beiden schnellen Geschwindigkeiten ist diese Tendenz geringfügig vorhanden, bei moderat langsamen Tempo aber nicht. Dies wirft die Frage auf, welche Benutzer welches Tempo bevorzugen.

1 Einführung

Die Präferenzen einzelner Hörer in Hinblick auf die Wiedergabe-Geschwindigkeit synthetischer Sprache können bisweilen stark divergieren. Faktoren, die dabei eine Rolle spielen bzw. spielen können, sind beispielweise Erfahrung im Umgang mit Sprachsynthese, Vertrautheit mit der gewählten Stimme, Alter des Hörers, Niveau der Sprachbeherrschung bei Nichtmuttersprachlern, Beeinträchtigungen des Gehörs, Informationsdichte, Texttypus, Dauer der synthetischen Äußerung sowie eine allgemeine Präferenz für schnelleres oder langsameres Sprechen. Es kann davon ausgegangen werden, dass Personen, die zum ersten Mal synthetische Sprache hören, diese mit langsamer Geschwindigkeit bevorzugen. Im Gegensatz dazu werden Personen, die täglich mit Sprachsynthese arbeiten, eher eine schnellere Geschwindigkeit präferieren.

Die Steuerung des Sprechtempos wird - falls überhaupt vorhanden - in den meisten Sprachsynthesesystemen durch eine *lineare* zeitliche Anpassung vollzogen. Sowohl die lautliche als auch die prosodische Struktur bleibt konstant, lediglich die Lautdauern werden dem gewünschten Zoomfaktor angepasst. Das Ergebnis dieser linearen zeitlichen Anpassung ist ähnlich (aber nicht genau gleich) einer Audiodatei, die mit höherer oder niederigerer Abtastrate abgespielt wird, wobei die Tonhöhe gleich gehalten wird.

In menschlicher Sprache sind die Unterschiede bezüglich des Sprechtempos hingegen als *nicht-linear* zu bezeichnen. Dieser nicht-lineare Charakter zeigt sich auf vielen Ebenen: bei der Anzahl der prosodischen Phrasengrenzen und der daraus resultierenden Pausen, der Anzahl der Satzakzente, Veränderungen der Lautstruktur in Form von Assimilationen, phonemischen Reduktionen sowie der Tilgung von Lauten und Silben. Auch die Dauer von Lauten und Silben ist in starkem Maße als nicht-linear anzusehen, da beispielsweise verschiedene Lautklassen verschieden "elastisch" sind. So werden z.B. Langvokale im Vergleich zu Plosiven stärker gedehnt und gestaucht. Ferner ist das Timing-Verhalten innerhalb von Lautsegmenten als nicht-linear anzusehen. Dazu seien die folgenden Faktoren genannt: unterschiedliche Elastizität von stationären gegenüber nicht-stationären Phasen bei Vokalen, spektrale Reduktion (target undershoot), Grad der koartikulatorischen Überlappung von Sprechgesten sowie der Grad der Beschleunigung der verschiedenen Artikulatoren.
All diese Faktoren bleiben bei einer linearen Veränderung der Zeitstruktur (bei konstant gehaltener Tonhöhe) von Sprachsignalen unbeachtet. Eine Folge davon ist, dass die Akzeptanz

bei Hörern erheblich absinkt. Bei langsamem Tempo entsteht ein "Leiern", das den Eindruck extremer Langeweile vermitteln (oder verstärken) kann. Bei schnellerem Tempo hingegen kann die Verständlichkeit erheblich leiden.

2 Ansätze zur nicht-linearen Steuerung des Sprechtempos

Es hat bislang mehrere Ansätze gegeben, nicht-lineare Aspekte bei der Steuerung des Sprechtempos bei der Sprachsynthese zu berücksichtigen. Im folgenden sollen sie in chronlogischer Ordnung kurz vorgestellt werden.

In den sogenannten Klatt-Regeln [8,1], einem additiv-multiplikativen Dauermodell, das eigentlich fürs Amerikanische Englisch entwickelt wurde, wird zur Verlangsamung eine kurze Pause zwischen einem Inhalts- und einem Funktionswort eingefügt und darüberhinaus einzelne Laute je nach Tempo ein wenig gedehnt oder gestaucht.

In einer deutschsprachigen TTS-Synthese [9] wird die globale Sprechgeschwindigkeit als einer von mehreren Faktoren in angepassten Klatt-Regeln behandelt. Die Folge davon ist, dass die Lautsegmente proportional zu ihrer inhärenten Dauer modifiziert werden.

Für einen französischsprachigen Synthesiser [3] wird ebenfalls vorgeschlagen, die Dauern von Pausen und Segmenten zu modellieren. Der Einfluss der Sprechgeschwindigkeit auf die Lautdauern geschieht aber unabhängig vom additiv-multiplikativen Dauermodell. Syntaktische Grenzen werden direkt auf Pausen abgebildet, die wiederum in obligatorische und optionale klassifiziert werden. Optionale Grenzen werden genutzt, um neue Pausen bei langsamen Tempo einzufügen oder sie werden bei schnellem Tempo übersprungen. Diese zumeist auf Interpunktion fußende Information bestimmt Häufigkeit und Dauer der Pausen.

Bei einem englischsprachigen Synthesesystem [10] wird der Schwerpunkt auf die Änderung der phonologischen Struktur gelegt, so dass prosodische Phrasen und Satzakzente modifiziert werden. Es wird argumentiert, dass Manipulationen auf der phonologischen Ebene besser das *wahrgenommene* Tempo beeinflussen, wohingegen das objektiv messbare Tempo mehr von den Lautdauern abhängt.

In [15] wird für eine französischsprachige Synthese neben der Anwendung von Regeln zur Re-Silbifizierung und der lautlichen Struktur auch zusätzliche Pausen und prosodische Phrasen eingefügt. Ein wichtiges Merkmal bei der Zuweisung der prosodischen Grenzen ist die Berücksichtigung sowohl rhythmischer als auch syntaktischer Gegebenheiten. Um die letztendlichen Lautdauern vorherzusagen wird Tempo als eigener Faktor im Dauermodell behandelt.

In einem Vergleich amerikanisch-englischer TTS-Synthesiser [7] wird zur Verlangsamung nach jedem Wort eine kurze Pause eingefügt, mit dem Ziel synthetische Sprache verständlicher zu machen, was in einem Hörtest auch gezeigt wird.

Obwohl kein direkter Zusammenhang zur synthetischen Sprache gegeben ist, sollte der nicht-lineare Ansatz der zeitlichen Skalierung in [5] erwähnt werden. Um bereits aufgenommene Sprachsignale zu stauchen, werden in erster Linie Pausen zeitlich reduziert (bis minimal 100 ms). Je nach Lautklasse werden die Laute wie folgt gestaucht: betonte Vokale am wenigsten, unbetonte Vokale durch einen mittleren Wert, Konsonanten je nach Betonung des benachbarten Vokals, und Konsonanten mehr als Vokale. Besonderer Wert wird darauf gelegt, dass sowohl die lautlichen Übergänge, in denen starke spektrale Änderungen vonstatten gehen, als auch bereits sehr kurz Laute nicht von der zeitlichen Skalierung betroffen sind. Perzeptionsexperimente zeigen, dass diese nicht-lineare Methode der Kompression einer linearen Methode überlegen ist.

Es ist bemerkenswert, dass die bislang verwendeten "nicht-linearen" Modelle entweder nur einen oder mehrere, aber nie *alle* der oben aufgeführten Aspekte berücksichtigen. Zudem scheinen die oben genannten Studien entweder nur *ein* schnelles und nur *ein* langsames

Tempo anzunehmen, oder die negative oder positive Beschleunigung sollte kontinuierlich betrachten werden, ohne dass Ober- und Untergrenzen bekannt wären. Weiterhin ist bemerkenswert, dass bis auf eine Studie [7] alle synthese-relevanten Studien auf eine formelle Evaluierung durch Hörer verzichten. Diese Studien basieren entweder auf Beobachtungen natürlicher Sprache bzw. phonetisch-linguistischer Intuition oder die Evaluierung wird an Hand von Sprachproduktionsdaten getestet. Es ist sicherlich von Vorteil, wenn die vorhergesagten Segmentdauern mit denen in den Produktionsdaten eines einzelnen Sprechers vergleichbar sind. Es ist aber nicht unbedingt gegeben, dass diese Daten zu den gleichen guten Ergebnissen mit wirklichen Hörern führen. Des weiteren kann angenommen werden, dass die Hörleistung unter ungünstigen Bedingungen, wie z.B. synthetische Sprache, abnimmt, und deswegen von Hörern Sprechgeschwindigkeiten bevorzugt werden, die langsamer sind als unter gewohnten und günstigen Bedingungen [14].

3 Temposteuerung durch Phrasierung

3.1 Vorüberlegungen

Mehrere Schlussfolgerungen der oben dargestellten Beobachtungen wären möglich: zum einen, dass man *alle* Ebenen in einem Temposteuerungsmodell berücksichtigt; zum zweiten, dass mehr als nur eine langsame bzw. schnelle Geschwindigkeit angegeben wird, denn Sprechen kann verschieden schnell und verschieden langsam vonstatten gehen; zum dritten, dass Perzeptionsexperimente durchgeführt werden. Allerdings gibt es Nachteile, wenn versucht wird, so viel wie irgend möglich versucht zu modellieren: man kann mit den Ergebnissen nicht erklären, welche Komponente für welches Ergebnis verantwortlich ist. Außerdem kann man davon ausgehen, dass nicht alle Komponenten gleich gut modelliert werden. Aus diesen Gründen scheint es angebracht, zunächst einmal auf den bzw. die wichtigsten Parameter bei der Temporegulierung einzugehen, um erste Hypothesen zu überprüfen und gegebenenfalls zu korrigieren, bevor weitere Komponenten untersucht werden. Seit [6] ist es allgemein anerkannt, dass Änderungen im Sprechtempo in erster Linie auf Änderungen in der Pausierung zurückzuführen sind. Daher sollte eine erste Modellierung je nach anvisiertem Tempo Dauer, Anzahl und Position der Pausen im Satz vorhersagen. Üblicherweise ist die Pausierung mit der Phrasierung eng verknüpft: in TTS-Systemen äußern sich prosodische Phrasengrenzen als Pausen, Dehnung von Lauten und Silben am Ende einer Phrase, sowie der Zuweisung eines Grenztones. Je nach Betrachtungsweise kann die Stärke einer prosodischen Grenze variieren: eine stärkere Grenze kann zu einer längeren Pausendauer, mehr phrasenfinaler Dehnung und einer ausgeprägteren Tonhöhenbewegung führen.

Grenze	vorhergesagte Position	Pausen-dauer	Faktor finale Dehnung	Grenzton
"2"	vor PP; vor Konj in koordin. NP o. AP	-	-	-
"3"	nach Vorfeld > 2 Tokens; vor "und"/"oder"	120	1,4 (Nukleus) 1,1 (Coda)	H-
"4"	Komma	200	0,6 (alle anderen)	H-%,
"6"	Satzende	410		H-^H%, L-%

Tabelle 1. Default-Vorhersage der Position prosodischer Grenzen je nach Grenzstärke und ihre jeweiligen Realisierungen in Pausendauer (ms), Faktor für finale Dehnung und Grenzton.

3.2 Default-Fall

Bevor eine Modellierung verschiedener Geschwindigkeiten stattfindet, lohnt es sich, den Default-Fall zu betrachten, den es zu ändern gilt. Das Modell beruht auf dem deutschsprachigen Diphon-basierten Sprachsynthesesystem "Mary" [12]. Im Normfall werden in "Mary" die prosodische Phrasengrenzen wie in Tabelle 1 vorhergesagt. Hier gilt es zu beach-

ten, dass momentan in "Mary" die folgenden Etiketten für die prosodischen Phrasengrenzen verwendet werden, die auf den ToBI-Konventionen beruhen [11]: "2", "3", "4", "6". Die phrasen-finale Dehnung wird im Dauermodell behandelt, das auf multiplikativ-additiven Regeln nach Klatt [8] beruht und fürs Deutsche angepasst wurden [4]. Die intrinsische Dauer der Silbenkerne (meist die Vokale) in phrasen-finaler Silbe wird mit Faktor 1,4 bei allen Grenzstärken multipliziert. Wenn keine Grenze vorliegt, wird Faktor 0,6 angewandt. Konsonanten in Coda-Position werden in phrasen-finaler Stellung mit 1,1 multipliziert.

3.3 Modell 1

Insgesamt sollen vier Geschwindigkeiten modelliert werden: jeweils eine extreme und eine moderate Form für langsam und schnell. Ein erstes Modell sieht die folgenden Zusätze und Änderungen vor:

- für langsame Geschwindigkeiten sollen zusätzlich zu den "normalen" Phrasengrenzen kleinere prosodische Phrasengrenzen eingefügt werden: nach jeder Nominalphrase (NP) und jeder Adjektivphrase (AP) (beabsichtigter Effekt: mehr Pausen und mehr phrasen-finale Dehnung)

- für schnellere Geschwindigkeiten werden Grenzen der Stärke "3" übersprungen

- die Dauern von Pausen sollen zum einen abhängig sein von der Stärke der Phrasengrenze und zum anderen vom anvisierten Tempo (siehe Tabelle 2)

	sehr schnell	*relativ schnell*	*default*	*relativ langsam*	*sehr langsam*
Grenze	60%	80%	100%	120%	140%
"2"	-	-	-	100 (120)	120 (200)
"3"	40 (20)	80	120	180 (200)	300 (410)
"4"	50	100	200	300 (410)	700
"6"	100	200	410	620 (700)	1000

Tabelle 2. Pausendauern in msec in Abhängigkeit der prosodischen Grenzstärke und des gewünschten Sprechtempos; in Klammern jeweils die Dauer von *modell 1* [13], falls von *modell 2* abweichend.

All diese Änderungen gegenüber der Default-Verarbeitung werden im folgenden als *modell 1* bezeichnet. In einem ersten Test [13] hat *modell 1* die Hypothesen bestätigt, den linear bearbeiteten Versionen überlegen zu sein. Eine Ausnahme bildet die *relativ langsame* Version, die wahrscheinlich auf Grund einer "übergenerierenden" Phrasierung deutlich abgelehnt wurde.

3.4 Modell 2

Modell 2 zielt darauf ab, die vermuteten Defzite wie sie im ersten Test für die *relativ langsame* Version (120%) aufgetreten sind, zu beseitigen, und auch *modell 1* weiter zu verfeinern. Die relativ schnellen Artikulationsphasen werden verlangsamt und die zu vielen bzw. zu langen Pausen werden gekürzt. Weiterhin wird versucht, die *sehr langsame* Version durch eine verstärkte phrasen-finale Dehnung zu verbessern. Deshalb wurden folgende Änderungen am Modell vorgenommen, die zu *modell 2* führen:

- Einfügen von "2"er-Grenzen nach NP und AP nur, wenn die neuen "Mikro-Phrasen" auch einen Akzent enthalten

- zusätzlicher Faktor 1,5 für jeden Silbenreim (Nukleus plus Coda) in akzentuierten Wörtern bei phrasen-finaler Dehnung (alle Geschwindigkeiten)

- Pausendauer für *relativ langsam* ebenfalls mit einem Faktor 1,5 in Bezug auf Normalfall multipliziert (Änderungen siehe Tabelle 2)

- für schnellere Geschwindigkeiten werden Grenzen der Stärke "3" *nicht* übersprungen

4 Experimente

4.1 Methode

Mit dem Sprachsynthesesystem "Mary" wurden für einen Text Stimuli in vier verschiedenen Tempi erzeugt, wobei weniger akzeptable Formen unkorrigiert gelassen wurden, um möglichst realistische Benutzungsbedingungen zu haben. Die vier Tempi *sehr schnell, relativ schnell, relativ langsam, sehr langsam* entsprechen 140%, 120%, 80%, 60% der Gesamtdauer der erzeugten Äußerung im Default-Fall. Für jedes Tempo wurden mehrere Modllierungen miteinander verglichen: zum einen rein *linear* zeit-skalierte Versionen mit beibehaltener F0-Charakteristik, zum anderen *hybride* Versionen, bei denen die Grenzen und Pausen neu modelliert wurden. Für die hybriden Versionen wurden zunächst die Grenzen gemäß des Modells neu vorhergesagt, danach die Pausendauern angepasst, und im letzten Schritt die so gewonnene Äußerung mit dem Speech Editor CoolEdit® linear zeit-skaliert. Die Sprechgeschwindigkeiten inklusive Pausen in Silben pro Sekunde (s/s) sind die folgenden: 3.66 s/s (140%); 4.27 s/s (120%); 6.41 s/s (80%); 8.23 s/s (60%).

Bei dem ausgewählten Text handelt es sich um einen Absatz eines Nachrichtentextes (2 Sätze, 36 Wörter, 74 Silben), der mit dem Text aus Test 1 vergleichbar, aber nicht identisch ist. Der Grund für die Auswahl eines Textes statt mehrerer Einzelsätze liegt darin, dass a) Phrasierung eher über längeren Äußerungen als über kürzere zum Tragen kommen, und dass b) Benutzer in vielen Anwendungen mehr als nur Einzelsätze zu hören bekommen.

	Bezugs dauer		und Höhe		des Arbeits-losengeldes		sollen nicht pauschal ge-kürzt werden	.	Ein-schnitte		soll es aber	
A		"2"		"-"		"3"		"6"		"-"		"2"
	1010	0	597	0	1817	176	2612	646	692	0	912	0
B		"2"		"2"		"3"		"6"		"2"		"2"
	902	140	485	140	1540	346	2209	1153	690	120	737	140

Tabelle 3. Der Textanfang in der *sehr langsamen* Geschwindigkeit (140%) in der linearen Bearbeitung (A) und nach *modell 2* (B). Für jede Textportion sind die prosodischen Grenzen mit ihrer Stärke sowie die Dauern der Pausen und der Artikulationsphasen in msec eingetragen. Neu eingefügte Grenzen der Stärke "2" in "B" entsprechen keiner Grenze "-" in "A".

Der Hörtest fand zweigeteilt statt, einer für die Vergleiche mit den langsamen Versionen (test 2 a-d), ein anderer für die schnellen Versionen (test 2 e-k). In den Fällen, in denen beim ersten Test die hybride Version der linearen überlegen war (140%, 80%, 60%) wurde erneut ein Vergleich zwischen dem damals verwendeten *modell 1* und *linear* vorgenommen. Ebenso wurde ein Vergleich zwischen dem veränderten Modell (*modell 2*) und der linearen Version vorgenommen. Um herauszufinden, ob die Veränderungen am Modell sich in den Präferenzen der Hörer niederschlägt, wurden auch beide hybriden Versionen gegeneinander getestet. Für jedes der 10 Vergleichspaare wurden zwei Stimuli erzeugt, wobei die Position jeder Version jeweils ausgetauscht wurde.

An jedem der beiden Teil-Experimente nahmen 10 Muttersprachler teil. Zum "Aufwärmen" wurde der Text zunächst einmal im Default-Tempo vorgespielt. Danach wurden die Stimuli in randomisierter Reihenfolge über Lautsprecher in einer ruhigen Büroumgebung dargeboten. Die Versuchspersonen sollten nach jedem Stimulus notieren, ob sie die erst- oder die zweitgespielte Version bevorzugen, wobei in jedem Fall eine Antwort gegeben werden musste (forced choice).

4.2 Ergebnisse

Die Ergebnisse für die langsamen Geschwindigkeiten (Tabelle 4) sind teilweise entgegen den Erwartungen. Im Gegensatz zu dem Ergebnis aus dem ersten Test ist *modell 1* im Vergleich

zu *linear* in den *sehr langsamen* Versionen (140%) nicht mehr der Favorit (test 2a), wohingegen *modell 2* deutlich gegenüber *linear* bestehen kann (test 2b). Das gute Abschneiden von *modell 2* in Beziehung zu *modell 1* kommt auch im direkten Vergleich der beiden hybriden Versionen zum Ausdruck, in dem *modell 1* klar präferiert wird (test 2c).

Bei den *relativ langsamen* Versionen (120%) schneidet *modell 2* in test 2d besser ab als *modell 1* in test1. Dennoch wird nicht der Level der *linearen* Version erreicht. Es ist weiterhin bemerkenswert, dass *modell 2* in der 120%-Versionen gerade die Hälfte der Punkte erreicht wie *modell 2* in den 140%-Versionen (test 2b vs. test 2d).

Für die *relativ schnellen* Versionen (80%) *modell 1* zwar gegenüber *linear* leicht überlegen ist, aber hier ist die sehr hohe Anzahl an inkonsistenten Antworten beachtenswert (test 2e). Somit wird die Tendenz aus dem ersten Test nicht voll bestätigt. Ein ähnliches Bild bietet der Vergleich von *modell 2* vs. *linear* (test 2f). Zwischen den beiden hybriden Modellen besteht im direkten Vergleich demnach auch kein Unterschied (test 2g).

Im Vergleich *linear-modell 1* (test 2h) bei den *sehr schnellen* Versionen (60%) gibt es eine Tendenz zu Gunsten von *modell 1*, die die Erwartungen aus test 1 bestätigt. Im Vergleich *linear-modell 2* (test 2i) gibt es die Tendenz zu Gunsten des hybriden Modells nicht. Konsequenterweise ist *modell 1* auch *modell 2* im direkten Vergleich überlegen (test 2k).

140%	test 1	test 2a	test 2b	test 2c	**120%**	test1	test 2d		
linear	17	50	20	-	*linear*	83	60		
modell 1	83	10	-	10	*modell 1*	17	-		
modell 2	-	-	80	90	*modell 2*	-	40		
inkonsistent	33	40	40	40	inkonsistent	33	40		
80%	test 1	test 2e	test 2f	test 2g	**60%**	test1	test 2h	test2i	test2k
linear	23	40	45	-	*linear*	40	30	55	-
modell 1	77	60	-	55	*modell 1*	60	70	-	70
modell 2	-	-	55	45	*modell 2*	-	-	45	30
inkonsistent	46	80	50	50	inkonsistent	53	60	50	40

Tabelle 5. Antworten der Versuchspersonen in Prozent für die einzelnen Vergleiche: "sehr langsamen" Versionen (140% der Defaultdauer), "relativ langsam" (120%), "relativ schnell" (80%) und "sehr schnelle" (60%). Für jeden Vergleich ist auch der prozentuale Anteil an inkonistenten Bewertungen angegeben.

Bemerkenswert für alle durchgeführten Tests ist die Anzahl der inkonsistenten Antworten, die bei den langsamen Versionen zwischen 33% und 40% liegen, d.h. die Mehrheit (67% - 60%) nimmt Unterschiede zwischen den verschiedenen Versionen wahr. Diese mehrheitliche Diskriminationsfähigkeit ist bei den schnellen Äußerungen selten der Fall, wo die inkonsistenten Beurteilungen zwischen 46% und 80% betragen.

4.3 Diskussion

Für die *sehr langsamen* Äußerungen hat es sich in Test 2 erneut gezeigt, dass einer der beiden nicht-linear behandelten Versionen der Vorzug vor der linearen Version gegeben wird. Allerdings überrascht es auf den ersten Blick, dass *modell 1* im erneuten Test schlechter abschneidet als *linear*. Das mag dieselben Ursachen haben wie beim Vergleich *modell 1-linear* im ersten Test [13], in dem die 120%-Versionen (aber nicht die 140%-Versionen) deutlich abgelehnt wurden: zu viele und/oder zu lange Pausen an möglicherweise ungünstigen Positionen im Satz. Dieses Ergebnis (test 1 vs. test 2a) zeigt aber auch, dass es unbedingt erforderlich ist, mehr als nur *einen* Text zu testen, auch wenn es sich um einen längeren Text handelt.

Bei den *relativ langsamen* Versionen hat weder *modell 1* noch *modell 2* die erwartete Überlegenheit in Bezug zur *linearen* Veränderung gebracht. Möglicherweise besteht ein Zusammenhang zwischen dem relativ langsamen Tempo für den Hörer der Diphon-Synthese und der

Artikulationsgeschwindigkeit, mit der das zugrunde liegende Diphon-Material produziert wurde. Diphon-Aufnahmen sind üblicherweise durch eine langsamere und akzentuiertere Sprechweise gekennzeichnet. Eine künstliche Verlangsamung durch "langsame Phrasierung" scheint daher bei den Hörern kein Gefallen zu finden.

Interessanterweise gibt es bei den *schnellen* Versionen leichte Präferenzen für die phrasenveränderten Modelle. Diese Tendenzen sind wie zu erwarten nicht so stark ausgeprägt wie bei *sehr langsam*. Es scheint sich als Vorteil herauszustellen, dass durch eine Verkürzung der Pausen die Stauchung der Artikulationsphasen nicht so stark ist wie bei den *linearen* Versionen. Um weitere Verbesserungen in diesen Tempo-Kategorien zu erreichen, ist es notwendig, die Dauermodellierung genau auf diese Tempi hin auszurichten, was mehr Kenntnisse über die Dauereigenschaften von Lauten unter verschiedenen Tempobedingungen verlangt. Weitere Verbesserungen könnte man erwarten, wenn besser geeignetes Diphon-Material (bzw. Material anderer Bausteingröße) bereitgestellt wird. Wie oben angedeutet entsteht durch die ursprünglich leicht hyperartikulierte Sprechweise des Diphon-Materials vor allem bei schneller Sprechweise der Eindruck von unpassender Hyper-Artikulation. Auch mit der Auswahl von Bausteinen, die größer als Diphone sind, sind Verbesserungen vorstellbar, wenn man bedenkt, dass sich die Koartikulation bei schneller Sprechweise möglicherweise stark verändert: zum einen durch einen stärkeren Grad der Koartikulation, zum anderen durch eine weitere zeitliche Ausdehnung auf vorhergehende bzw. nachfolgende Laute und Silben.

Im Gegensatz zum schnelleren Sprechen kann man langsame Sprachsynthese erfolgreich mit einer längeren *relativen* Pausendauer modellieren. Dazu sollte die Anzahl der eingefügten Pausen unter Berücksichtigung der Position im Satz erhöht werden, sodass die Artikulationsgeschwindigkeit lediglich moderat erhöht wird. So ist bei den 140%-Versionen die Artikulationsgeschwindigkeit von *linear* bei 3,9 Silben pro Sekunde (s/s) und der Anteil der Pausen an der Gesamtsprechzeit beträgt lediglich 6 %; die entsprechenden Zahlen von *modell 2* sind 4,7 s/s und 22 %. Eine zu langsame Artikulation kann den manchmal berichteten Effekt von Langeweile synthetischer Sprache verstärken.

5 Abschließende Diskussion und Zusammenfassung

Bei den *schnellen* Geschwindigkeiten werden die nicht-linear veränderten Versionen den linearen Versionen meistens leicht bevorzugt. Aber auf Grund der Geringfügigkeit der Ergebnisse und des hohen Anteils an inkonsistenten Hörer-Urteilen, kann man nicht behaupten, dass eine wirkliche Verbesserung durch die vorgenommen Änderungen stattgefunden hat. Die Modellierung der Vorhersage und der Realisierung von Phrasengrenzen *alleine* scheint nicht auszureichen, kann aber möglicherweise in Verbindung mit Änderungen auf anderen Ebenen zu wirklichen Verbesserungen führen.

Die Tatsache, dass der lineare Ansatz bei *relativ langsamen* Tempo den hier dargestellten nicht-linearen Ansätzen überlegen sind, lässt die Frage aufkommen, ob das von den Entwicklern vorgegebene Default-Tempo und das von Benutzern präferierte Tempo dasselbe ist. Es gibt bei der Frage nach der bevorzugten persönlich präferierten Geschwindigkeit viele Faktoren, die hierbei eine Rolle spielen können: neben der Natürlichkeit auch die Verständlichkeit, aber auch der Eindruck der Langeweile, und wie oben angedeutet die ursprüngliche Geschwindigkeit, mit der die Aufnahmen produziert wurden.

Ein weiterer nächster Schritt wären Hörtests mit schnell artikulierter Synthese mit Personen, die auch ein schnelles Tempo bei Sprachsynthese bevorzugen. Dies beträfe auch das Testen der Verständlichkeit, über die bei Präferenztests nichts herausgefunden wird. Es mag durchaus sein, dass die als natürlicher wahrgenommene Äußerung der verständlicheren vorgezogen wird, wenn nicht ausdrücklich nach der Verständlichkeit gefragt oder darauf hingewiesen wird.

Es hat sich gezeigt, dass Beobachtungen der natürlichen Sprache nicht im Verhältnis 1:1 auf synthetische Sprache übertragbar sind. So hat es sich z.B. als sinnvoll erwiesen, mehr als nur 2 Tempi auszuwählen, denn (extrem) "langsam" und (moderat) "langsam" scheinen sich nicht gleich zu verhalten.

Weitere Schritte bestehen darin, die hier vorgestellten Modelle weiter zu verfeinern, auf andere Komponenten auszuweiten und durch weitere Hörtests zu überprüfen. Eine solche Weiterentwicklung verlangt mehr Kenntnisse über die Vorhersage von Phrasengrenzen bei verschiedenen Tempi (siehe auch [2]), den Zusammenhang zwischen finaler Dehnung und Pausendauer sowie günstige Verhältnisse zwischen Artikulationsgeschwindigkeit und Pausendauer.

Nach jetzigem Stand scheint *modell 2* für *sehr langsames* Tempo angemessen. Obwohl viele bei Temposteuerung für Sprachsynthese zunächst an schnelle Geschwindigkeiten denken, so scheint es – wie in der Einführung angedeutet - bei mehreren Anwendungssituationen und Benutzerprofilen sinnvoll, synthetisierte gesprochene Information *langsam* darzubieten.

Literatur

[1] Allen, J., Hunnicutt, S., & Klatt, D. (1987): From Text to Speech: The MITalk System. Cambridge University Press, Cambridge.

[2] Atterer, M. (2002): Assigning prosodic structure for speech synthesis: a rule-based approach. Proc. Prosody 2002, Aix-en-Provence, 147-150.

[3] Bartkova, K. (1991): Speaking rate modelization in French application to speech synthesis. Proc. ICPhS Aix-en-Provence (3), 482-485.

[4] Brinckmann, C. & Trouvain, J. (erscheint): The Role of Duration Models and Symbolic Representation for Timing in Synthetic Speech. International Journal of Speech Technology Research.

[5] Covell, M., Withgott, M., Slaney, M.: MACH1 (1998): Nonuniform time-scale modification of speech. Proc. ICASSP Seattle.

[6] Goldman-Eisler, F. (1968): Psycholinguistics. Experiments in Spontaneous Speech. London & New York: Academic Press.

[7] Higginbotham, D.J., Drazek, A.L., Kowarsky, K., Scally, C. & Segal, E. (1994): Discourse comprehension of synthetic speech delivered at normal and slow presentation rates. Augmentative and Alternative Communication 10, 191-202.

[8] Klatt, D.H. (1979): Synthesis by rule of segmental durations in English sentences. In: Lindblom, B. & Öhmann, S. (eds): Frontiers of Speech Communication Research, London etc: Academic Press, 287-299.

[9] Kohler, K. (1990): Zeitstrukturierung in der Sprachsynthese. ITG-Fachber. 105, 165-170.

[10] Monaghan, A.I.C. (1991): Accentuation and speech rate in the CSTR TTS system. Proc. ISCA Workshop on "Phonetics and Phonology of Speaking Styles" Barcelona, 41/1–5.

[11] Price, P.J, Ostendorf, S., Shattuck-Hufnagel, S. & Fong, C. (1991): The use of prosody in syntactic disambiguation. Journal of the Acoustical Society of America 90 (6), 2956-2970.

[12] Schröder, M. & Trouvain, J. (erscheint): The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technolgy Research.

[13] Trouvain, J. (2002): Tempo control in speech synthesis by prosodic phrasing. Proc. Konvens 2002 Saarbrücken.

[14] Uchanski, R.M., Choi, S.S., Braida, L.D., Reed, C.M. & Durlach, N.I. (1996): Speaking clearly for the hard of hearing IV: further studies of the role of speaking rate. Journal of Speech and Hearing Research 39, 494-509.

[15] Zellner-Keller, B. (im Druck): Prediction of temporal structures for various speech rates. In: Campbell et al. (eds): Progress in Speech Synthesis II, Berlin, Heidelberg: Springer.