ZU MUSTERN DER PAUSENGESTALTUNG IN NATÜRLICHER UND SYNTHETISCHER LESESPRACHE

Jürgen Trouvain, Bernd Möbius

Sprachwissenschaft und Sprachtechnologie, Universität des Saarlandes, Saarbrücken
trouvain|moebius@coli.uni-saarland.de

Kurzfassung: Die vorliegende Studie vergleicht anhand eines vorgelesenen Textes die Pausensetzung in natürlicher Sprachproduktion (sechs Sprecher) und in der Sprachsynthese (vier Systeme). In die Auswertung gingen die Anzahl der Pausen, ihre Dauer, die Stelle der Pausen im Text und hörbare Atemgeräusche ein, und zwar jeweils an Satzgrenzen und an satzinternen Positionen. Die Pausenstrategie der Synthesesysteme weicht erheblich von der der Sprecher ab und erscheint selbst im Vergleich zu einer einfachen Heuristik, die auf Interpunktion beruht und zwei Arten von Pausen unterscheidet, als wenig adäquat. Eine besondere Herausforderung für die Modellierung stellt die Variation in Form einer optionalen Pausensetzung in der natürlichen Sprache dar.

1 Einführung

Die Qualitätsbeurteilung von Text-to-Speech-Synthese (TTS) beschränkt sich meist auf Einzelsätze. Damit entgeht man zwar der Problematik, wie längere Abschnitte mit mehreren Sätzen methodisch sauber zu bewerten sind, allerdings bleibt dabei auch unklar, wie man längere Abschnitte prosodisch verbessert. Zur prosodischen Bewertung gehört unter anderem auch die Bewertung der zeitlichen Gestaltung auf größeren Ebenen wie Gesamttext und Absatz und hier besonders die Strukturierung durch Pausen.

In der Phonologie werden Pausen im Zusammenhang mit prosodischer Phrasierung betrachtet. Dabei spiegeln prosodische Phrasengrenzen zumeist auch syntaktische Phrasengrenzen wider. Die häufig angenommene prosodische Hierarchie sieht am oberen Ende eine Differenzierung größerer und kleinerer prosodischer Phrasen vor. Man geht davon aus, dass die Beendigung größerer prosodische Phrasen, auch Intonationsphrasen genannt, in aller Regel durch Pausen markiert werden [1].

Gee & Grosjean [2] führen zusätzlich zu den syntaktisch motivierten auch rhythmisch bedingte Phrasengrenzen ein: wenn z.B. eine prosodische Phrase sehr lang ist und geteilt werden soll, ist es wahrscheinlich, dass die einzuführende Grenze etwa in der Hälfte der langen Phrase, aber an einer syntaktisch erlaubten Stelle gesetzt wird.

Eine flexible Modellierung von Phrasengrenzen und der zugehörigen Pausen würde es TTS-Systemen erlauben, die bei natürlichen Sprechern zu beobachtende Variation der prosodischen Gestaltung zumindest ansatzweise abzubilden (beim mehrmaligen Vorlesen desselben Textes neigen Sprecher zu einer leicht unterschiedlichen Lokalisierung ihrer Pausen, vgl. [3]). Darüberhinaus wäre es für TTS-Synthese von Vorteil, Änderungen des Sprechtempos vorzunehmen, was in erster Linie durch Hinzufügen bzw. Weglassen von Pausen sowie der zeitlichen Ausdehnung der Pausen erfolgt [4, 5].

Ein wichtiger Aspekt bei der Untersuchung und Modellierung von Pausen ist die Betrachtungsebene: aus der Perspektive der Sprachproduktion werden physiologische und akustische Korrelate von Pausen analysiert, aus der Sicht der Sprachperzeption müssen wahrgenommene Pausen nicht unbedingt akustischen Pausen entsprechen. Es ist durchaus möglich beim Zuhörer den Eindruck einer Pause hervorzurufen, ohne dass eine Stille oder Atmungsgeräusch da-

bei vorkommt. Zum Ende einer Phrase kommen häufig phrasen-finale Dehnung der letzten Silbe(n), Unterschiede in der Stimmqualität und die Grundfrequenz betreffende Ereignisse vor, die als Phrasengrenzmarker ausreichen.

Dennoch sind die meisten perzipierten Pausen auch mit Stille und Atmungsgeräuschen verbunden. Hierbei ist zu beachten, dass Einatmung zwar häufig, aber nicht zwingend zu einem Einatmungsgeräusch führt [6]. Pausen mit Einatmungsgeräusch sind in aller Regel länger und kommen auch häufiger vor als solche ohne dieses Geräusch. Daher ist zu erwarten, dass bei einer hierarchisch höher stehenden Grenze, z.B. einer Grenze zwischen zwei Sätzen, eher eine Atempause erfolgt als bei einer niedriger stehenden Grenze, z.B. einer Grenze innerhalb eines Satzes.

Die einfachste Modellierung von TTS-Synthese für das Deutsche könnte theoretisch so aussehen, dass man sich ausschließlich nach der Interpunktion orientiert: Pausen werden bei Satzende-Punkten und Kommata eingefügt (die anderen Satzzeichen für den Moment ignorierend); eine "Komma-Pause" ist kürzer als eine "Satzgrenzpause"; es werden zwei Default-Dauerwerte definiert, die auf Beobachtungen von Produktionen natürlicher Lesesprache fußen. Atmungsgeräusche werden zunächst weggelassen, weil diese üblicherweise (noch) nicht in dem zu Grunde liegenden TTS-Material annotiert sind.

In der vorliegenden Studie haben wir vorgelesene Versionen desselben Textes bezüglich der darin vorgefundenen Pausen bezüglich der Parameter Pausenanzahl, Pausendauer, hörbare Atmung und Stelle der Pausen im Text untersucht. Das Material stammt zum einen von zufällig ausgewählten Sprechern eines Korpus mit Lesesprache und zum anderen von zwei verschiedenen TTS-Systemen.

2 Methode

2.1 Material

Als Text wurde die Kurzgeschichte "Die drei kleinen Schweinchen", die 13 Sätze enthält, ausgewählt für die Versionen von natürlichen Sprechern und von deutschsprachigen TTS-Systemen analysiert wurden. Die natürlich-sprachlichen Versionen stammen von 6 zufällig ausgewählten Sprechern des IFCASL-Korpus [7] (Muttersprachler des Deutschen).

Die synthetischen Versionen stammen von zwei verschiedenen TTS-Systemen: 1) zwei verschiedene Stimmen von MaryTTS [8]: bits3_hsmm_de_male und bits3_de_male_ unitselection_general; 2) zwei Versionen der bei Google benutzten Synthese, zum einen bei Home [9], zum anderen bei Translate [10]. Drei der vier Synthese-Versionen wurden durch ein Web-Interface generiert, eine Version wurde von der Firma Google zur Verfügung gestellt. Die vier synthetischen Versionen werden im Folgenden nicht namentlich genannt, sondern durchnummeriert, da lediglich ein allgemeiner Einblick in synthetische Proben von Interesse ist und keine Überprüfung spezifischer TTS-Systeme.

2.2 Analysen

Alle 10 Versionen (6 natürliche, 4 synthetische) wurden bezüglich der vorgefundenen Pausen mit Hilfe des Speech Editors Praat untersucht. Die Parameter sind im Einzelnen die Anzahl der Pausen, ihre Dauer, die Beteiligung hörbarer Atmung und die Stelle der Pausen im Text. In den (zahlreichen) Fällen, in denen einer Pause ein Plosiv folgt, z.B. wenn ein Satz mit "der, die, das, da" beginnt, wurden generell 50 ms von der Pausendauer als für die im Sprachsignal meist nicht erkennbare Bildungsphase des Plosivs (und damit der Artikulation) abgezogen.

3 Ergebnisse

3.1 Pausen an Satzgrenzen

Die Resultate zeigen, dass wie erwartet alle natürlichen Sprecher Pausen an den Satzgrenzen produzieren und diese nahezu immer mit Atemgeräuschen versehen. Diese Atempausen weisen mit durchschnittlich ca. 800 ms relativ lange Dauern auf (Abb. 1).

Auch die TTS-Systeme zeigen ähnlich den natürlichen Sprechern Pausen an allen Satzgrenzen. Allerdings sind diese im Kontrast zu den natürlichen Sprechern ohne Atemgeräusch. Wie in Abbildung 1 ersichtlich sind die Satzgrenzpausen auch beträchtlich kürzer: ca. 420 ms für TTS 1 und 2 bzw. 640 ms für die beiden anderen Synthese-Versionen.

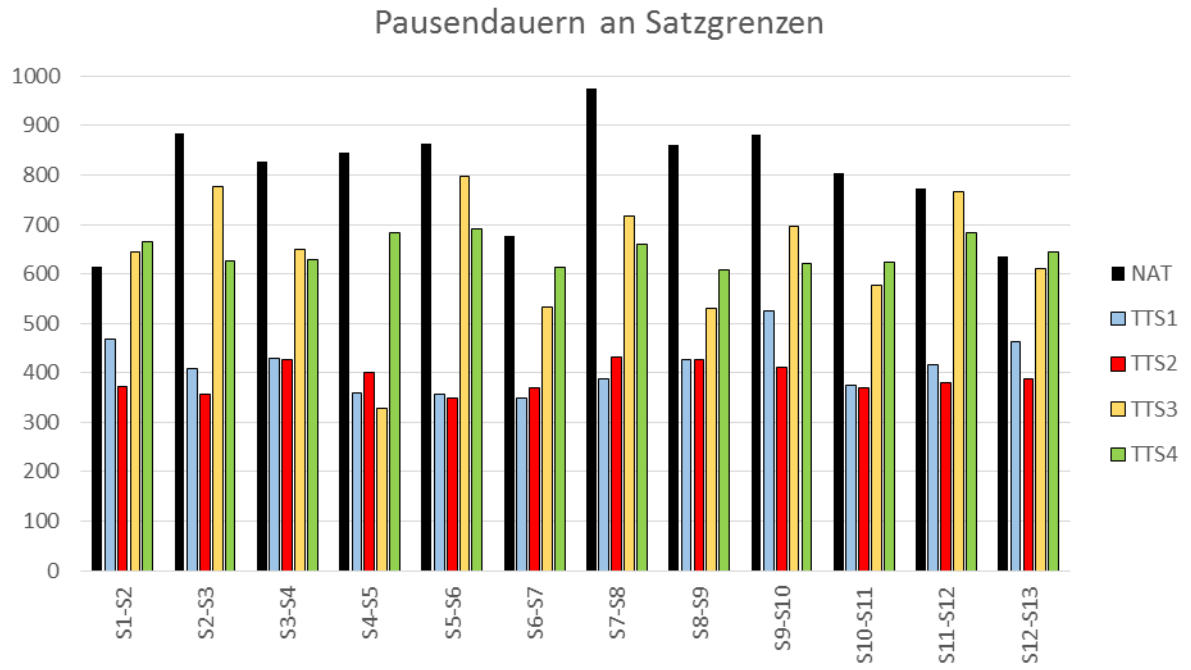

Abbildung 1 – Pausendauern (in ms) an Satzgrenzen (NAT = Mittelwerte der natürlichen Sprecher)

3.2 Pausen innerhalb von Sätzen

Alle natürlichen Sprecher produzieren auch Pausen innerhalb der Sätze. Je nach Sprecher schwankt die Anzahl zwischen 10 und 20. Bei diesen Binnenpausen gibt es leicht mehr Atempausen als Nicht-Atempausen (52:34, vgl. Tab. 1). Die Binnenpausen haben aber kürzere Dauern als die Satzgrenzpausen (ca. 300 ms im Schnitt).

Pausen innerhalb von Sätzen können an 23 verschiedenen Wortgrenzen festgestellt werden. Aber nur 11 dieser Stellen werden von der Mehrheit der Sprecher zur Pausierung genutzt (fett markierte Zahlen in Tab. 1). Eine weitere Differenzierung ist ersichtlich, wenn diese mehrheitlichen Binnenpausen bezüglich hörbarer Atembeteiligung analysiert werden. Manche Binnenpausen finden mehrheitlich *ohne* Atmung statt (z.B. S11.1), andere mehrheitlich *mit* Atmung (z.B. S2.2).

Im Text gibt es 13 Stellen mit Komma als Interpunktionszeichen. Es werden von diesen aber nur 10 Stellen und damit nicht all diese Stellen von der Mehrheit der Sprecher mit einer Pause markiert. Es gibt auch eine von allen gesetzte Binnenpause (S2.3), die ohne Interpunktionszeichen im Text vorkommt: vor der Konjunktion "und", die zwei Hauptsätze miteinander verbindet (8 bzw. 11 Silben lang).

Tabelle 1 - Platzierung der Pausen im Text mit Häufigkeit, Binnenpausen in eckigen Klammern. AP = Atempausen, NAP = Nicht-Atempausen, N = alle Pausen bei natürlichen Sprechern (max. 6), darin mehrheitlich gesetzte Pausen in Fettdruck, die synthetischen Versionen als 1/2 und 3/4.

Satz	**Text**	**AP**	**NAP**	**N**	**1/2**	**3/4**
S1	Die drei kleinen Schweinchen [1]	0	0	0	2	0
	gehen heim, [2]	2	2	**4**	2	1
	um ihre Häuser zu bauen.	**5**	1	**6**	2	2
S2	Das erste kleine Schweinchen baut [1]	0	0	0	2	0
	ein Haus aus Stroh, [2]	**5**	1	**6**	2	0
	das zweite baut ein Haus aus Holz [3]	**4**	2	**6**	2	0
	und das dritte baut ein Haus aus Backsteinen.	**5**	1	**6**	2	2
S3	Der Wolf entdeckt [1]	1	1	2	0	0
	die drei kleinen Schweinchen und beschließt, [2]	2	1	3	2	0
	das kleine Schweinchen mit dem Haus aus Stroh [3]	1	0	1	2	0
	als erstes zu fressen.	**5**	1	**6**	2	2
S4	Er klopft an der Türe, [1]	**4**	2	**6**	2	1
	aber das kleine Schweinchen lässt ihn nicht ins Haus.	**5**	1	6	2	2
S5	Der Wolf bläht daraufhin seine Backen auf [1]	3	0	3	0	0
	und bläst mit all seiner Kraft, [2]	3	0	3	0	0
	woraufhin das Haus weg fliegt.	**6**	0	**6**	2	2
S6	Das kleine Schweinchen [1]	0	0	0	2	0
	rennt zu seinem Bruder [2]	0	1	1	0	0
	mit dem Haus aus Holz.	**6**	0	**6**	2	2
S7	Der Wolf folgt dem Schweinchen und klopft wieder [1]	0	1	1	2	0
	an die Türe, [2]	3	3	**6**	2	0
	aber die kleinen Schweinchen lassen ihn nicht hinein.	**6**	0	**6**	2	2
S8	Der Wolf [1]	0	2	0	0	0
	bläst daraufhin seine Backen auf, [2]	2	3	**5**	2	0
	pustet mit all seiner Kraft [3]	2	0	2	0	0
	und das Haus fliegt davon.	**6**	0	**6**	0	2
S9	Die zwei kleinen Schweinchen [1]	0	0	0	2	0
	rennen zu ihrem Bruder [2]	0	1	1	0	0
	mit dem Haus aus [3]	1	0	1	0	0
	Backsteinen.	**6**	0	**6**	2	2
S10	Der Wolf klopft an die Türe, [1]	3	2	**5**	2	0
	aber die kleinen Schweinchen lassen ihn nicht hinein.	**6**	0	**6**	2	2
S11	Der Wolf bläst daraufhin seine Backen auf, [1]	1	5	**6**	2	0
	pustet mit all seiner Kraft, [2]	1	1	2	0	0
	aber das Haus fliegt nicht davon.	**6**	0	**6**	2	2
S12	Da beschließt der Wolf, [1]	2	2	4	2	0
	durch den Schornstein in das Haus zu steigen, [2]	**4**	2	**6**	2	0
	aber die kleinen Schweinchen haben einen Kessel mit kochendem Wasser vorbereitet.	**6**	0	**6**	2	2
S13	Der Wolf fällt hinein, [1]	2	3	**5**	2	0
	stößt einen Schrei aus [2]	0	1	1	2	0
	und rennt davon.	-	-	-	-	-
	Binnenpausen	52	34	86	36	2
	Satzgrenzpausen	68	4	72	24	24

Ausgehend von den Binnenpausen, die mehrheitlich produziert werden, können optimale Zwischenpausenabschnitte angenommen werden. Wie in Abbildung 2 zu sehen variiert die Länge dieser Abschnitte zwischen 5 und 29 Silben mit einem Median-Wert von 11 Silben und einem Durchschnittswert von 12 Silben. Auffällig sind die Extremwerte: es gibt Zwischenpausenabschnitte mit 20 bzw. 22 Silben (in Fettdruck), die von *allen* Sprechern ohne Pausen produziert wurden. Der Abschnitt mit 29 Silben wurde von zwei Sprechern ohne Pausen gesprochen, die anderen Sprecher haben an verschiedenen Stellen eine bis drei Pausen produziert.

Die drei kleinen Schweinchen gehen heim, [13] um ihre Häuser zu bauen. [8] Das erste kleine Schweinchen baut ein Haus aus Stroh, [12] das zweite baut ein Haus aus Holz [8] und das dritte baut ein Haus aus Backsteinen. [11] Der Wolf entdeckt die drei kleinen Schweinchen und beschließt, das kleine Schweinchen mit dem Haus aus Stroh als erstes zu fressen. [**29**] Er klopft an der Türe, [6] aber das kleine Schweinchen lässt ihn nicht ins Haus. Der Wolf bläht daraufhin seine Backen auf und bläst mit all seiner Kraft, [18] woraufhin das Haus weg fliegt. [7] Das kleine Schweinchen rennt zu seinem Bruder mit dem Haus aus Holz. [15] Der Wolf folgt dem Schweinchen und klopft wieder an die Türe, [13] aber die kleinen Schweinchen lassen ihn nicht hinein. [12] Der Wolf bläst daraufhin seine Backen auf, [11] pustet mit all seiner Kraft und das Haus fliegt davon. [13] Die zwei kleinen Schweinchen rennen zu ihrem Bruder mit dem Haus aus Backsteinen. [**20**] Der Wolf klopft an die Türe, [7] aber die kleinen Schweinchen lassen ihn nicht hinein. [13] Der Wolf bläst daraufhin seine Backen auf, [11] pustet mit all seiner Kraft, aber das Haus fliegt nicht davon. [15] Da beschließt der Wolf, [5] durch den Schornstein in das Haus zu steigen, [10] aber die kleinen Schweinchen haben einen Kessel mit kochendem Wasser vorbereitet. [**22**] Der Wolf fällt hinein, [5] stößt einen Schrei aus und rennt davon.[9]

Abbildung 2 – Der verwendete Text mit Anzahl der Silben des vorangehenden Abschnitts zwischen zwei Pausen, die von der Mehrheit der Sprecher produziert wurden (siehe Tab. 1)

Die Versionen der TTS-Systeme unterscheiden sich bezüglich der Binnenpausen markant von den natürlichen Sprechern. Die Versionen 1 und 2 weisen relativ viele Pausen auf (an 18 Stellen), die Versionen 3 und 4 extrem wenige Pausen (an nur zwei Stellen). Dies ist im deutlichen Gegensatz zu den 10 Stellen der mehrheitlich gesetzten Binnenpausen der natürlichen Sprecher.

Die Binnenpausen bei TTS 1 und 2 sind im Schnitt auch noch etwas länger als die Pausen an Satzgrenzen (Abb. 3a,b). Erschwerend kommt hinzu, dass manche Binnenpausen an Stellen auftreten, welche die natürlichen Sprecher gar nicht benutzen (S1.1, S6.1, S9.1) oder bei nur einem natürlichen Sprecher (S3.3, S7.1, S13.2) auftauchen.

Bei TTS 3 und 4 gibt es sehr wenige Pausen innerhalb der Sätze. Diese sind extrem kurz (weniger als 100 ms, Abb. 3a,b), was zu sehr langen pausenfreien Intervallen führt.

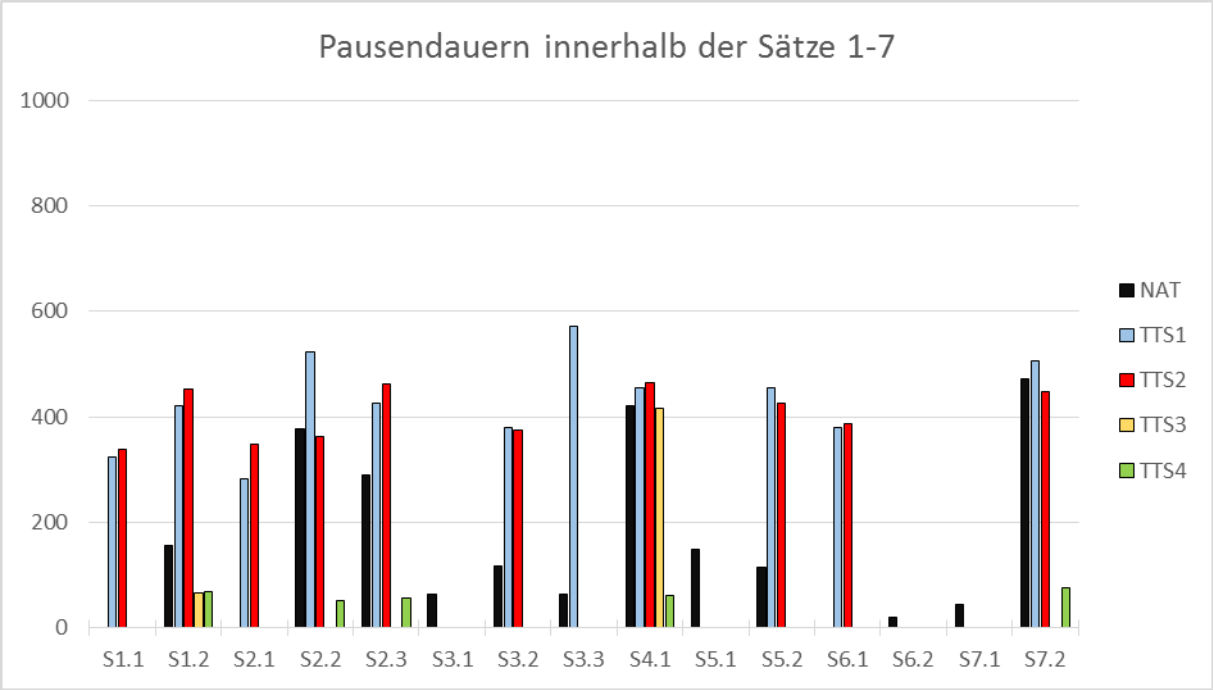

Abbildung 3a – Pausendauern (in ms) innerhalb der ersten sieben Sätze (NAT = Mittelwerte der natürlichen Sprecher)

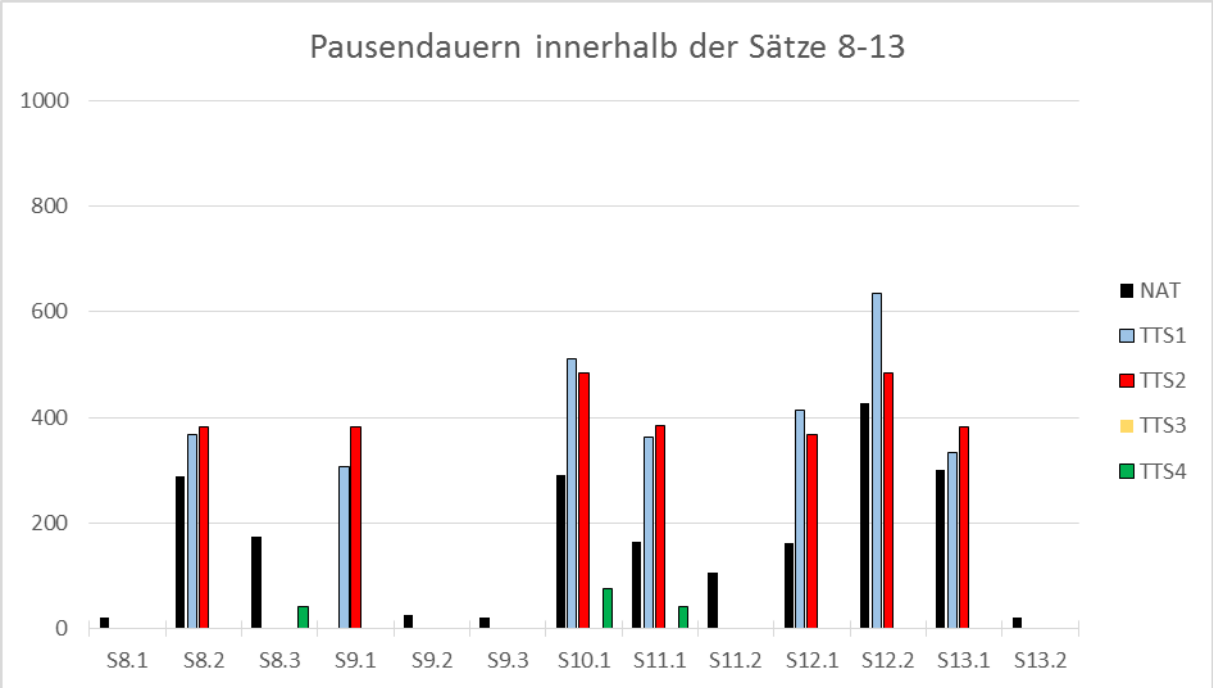

Abbildung 3b – Pausendauern (in ms) innerhalb der letzten sechs Sätze (NAT = Mittelwerte der natürlichen Sprecher)

4 Diskussion

Die natürlichen Sprecher zeigen eine große Variation bezüglich der Dauern ihrer Pausen. Dabei scheint es sprecherübergreifende Tendenzen zu geben, Pausen an bestimmten Stellen länger zu gestalten als an anderen Stellen. Ein genaues Verständnis davon, unter welchen Umständen Pausen an manchen Stellen gedehnt werden, muss aber noch entwickelt werden.

Die Pausendauergestaltung bei den Synthesen sieht gänzlich anders aus. TTS 3 und 4 generieren Pausen an Satzgrenzen mit ca. 600 ms und kommen dem Mittel der natürlichen Sprecher (800 ms) näher als TTS 1 und 2 mit nur 400 ms. Allerdings kommen TTS 3 und 4 innerhalb von Sätzen quasi ohne akustische Pausen aus (andere Parameter zur Pausenperzeption sind durchaus vorhanden), was das Verständnis längerer Sätze erschwert. Die natürlichen Sprecher unterscheiden in ihrer Lesesprache klar zwischen Satzende- und Binnenpausen. Bei TTS 1 und 2 oszillieren die Pausendauern stets um 400 ms, ganz gleich, ob es sich um Binnengrenzen oder solche am Satzende handelt. Eine Differenzierung wäre wahrscheinlich sinnvoll.

Die Anzahl der Pausen ist starker Variation unterworfen, wobei dies nur die Binnengrenzen betrifft; die Satzendegrenzen werden regelmäßig mit einer Pause in Lesesprache versehen. Bei natürlichen Sprechern, die wenige Binnenpausen setzen, tun dies an relativ wichtigen Stellen, an denen die "Viel-Pausierer" dies auch tun. Es gibt aber auch etliche Textstellen, bei denen optional (zumeist kurze) Pausen produziert werden können. TTS 1 und 2 hingegen zeigen dabei interessanterweise solche Optionen auf, die nicht von den natürlichen Sprechern gewählt werden (aber mit der Standarddauer).

5 Zusammenfassung und Ausblick

Die vorliegende Studie hat an Hand eines vorgelesenen Textes Pausen natürlicher Sprecher mit solchen synthetischer Sprecher verglichen. Im Gesamtblick erstaunt es ein wenig, wie Pausen in den inspizierten TTS-Systemen behandelt werden und wie wenig dabei natürliche Sprecher Modell zu stehen scheinen.

Am Anfang des Aufsatzes ist eine simple Modellierung vorgeschlagen worden: lange Pausen am Satzende (z.B. 800 ms) und kurze Pausen bei Kommata (z.B. 300 ms). Es ist anzunehmen, dass eine solche Gestaltung zu einem adäquateren Ergebnis für die Zuhörer führt.

Eine detailliertere Modellierung könnte Differenzierungen enthalten bezüglich der Dauern sowohl der langen als auch der kurzen Pausen, aber auch der Berücksichtigung von Atemgeräuschen. So haben beispielsweise Whalen et al. [11] für Formantsynthese gezeigt, dass Einatmungsgeräusche einen positiven Effekt auf die gedächtnisbezogene Behaltensleistung der nachfolgenden Sätze haben können.

Die Evaluierung längerer Texte mit mehreren Sätzen hingegen ist ein nicht gelöstes Problem, das für natürliche als auch für synthetische Sprache gilt. Es ist daher nicht verwunderlich, dass es nur wenige Untersuchungen zu Pausen in der Sprachsynthese gibt; zu den Ausnahmen zählen Parlikar & Black [12] sowie Braunschweiler & Chen [13].

Im Allgemeinen wäre eine bessere Modellierung von Pausen in der Sprachsynthese wünschenswert. Da die Auswahl der geeigneten Pausenstellen dabei sowohl von der syntaktischen Struktur als auch von der Phrasenlänge abhängen, könnte sich eine Modellierung nach dem Vorbild von Gee & Grosjean [5] eignen. Basierend auf den aktuellen Befunden kann die Hypothese aufgestellt werden, dass Hörer im Vergleich zu den manipulierten Versionen die existierenden TTS-Versionen als weniger adäquat empfinden. Ob beim Zuhören einer pausenoptimierten TTS-Synthese tatsächlich mehr Informationen im Gedächtnis behalten werden, muss allerdings noch mit Tests nachgewiesen werden. Die Berücksichtigung der genannten Punkte ist für eine Folgestudie geplant, bei der die hier generierten TTS-Versionen mit solchen TTS-Versionen verglichen werden, die nach natürlichen Vorbildern bezüglich der Pausen manipuliert wurden.

Literatur

[1] GRICE, M. & BAUMANN, S.: An Introduction to Intonation – Functions and Models. In: TROUVAIN, J. & GUT, U. (Hgg.). *Non-Native Prosody. Phonetic Description and Teaching Practice*. Berlin, New York: De Gruyter (= Trends in Linguistics. Studies and Monographs [TiLSM] 186), S. 25-51, 2007.

[2] GEE, J. & GROSJEAN, F.: Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology* 15, S. 411-458, 1983.

[3] TROUVAIN, J. & GRICE, M.: The effect of tempo on prosodic structure. *Proceedings 14th International Congress of Phonetic Sciences* (ICPhS), San Francisco, S. 1067-1070, 1999.

[4] TROUVAIN, J.: Temposteuerung in der Sprachsynthese durch prosodische Phrasierung. *Tagungsband 13. Konferenz Elektronische Sprachsignalverarbeitung* (ESSV), Dresden, S. 294-301, 2002.

[5] MOOS, A. & TROUVAIN, J.: Einzelfallstudie zu Grenzen der Verständlichkeit ultraschneller Sprachsynthese. *Tagungsband 19. Konferenz Elektronische Sprachsignalverarbeitung* (ESSV), Frankfurt/M., S. 207-214, 2008.

[6] TROUVAIN, J. & MÖBIUS, B.: Einatmungsgeräusche vor synthetisch erzeugten Sätzen — Eine Pilotstudie. *Tagungsband 24. Konferenz Elektronische Sprachsignalverarbeitung* (ESSV), Bielefeld, S. 50-55, 2013.

[7] TROUVAIN, J., BONNEAU, A., COLOTTE, V., FAUTH, C., FOHR, D., JOUVET, D., JÜGLER, J., LAPRIE, Y., MELLA, O., MÖBIUS, B. & ZIMMERER, F.: The IFCASL corpus of French and German non-native and native read speech. *Proceedings 9th Language Resources and Evaluation Conference* (LREC), Portorož, S. 1333-1338, 2016.

[8] http://mary.dfki.de/

[9] persönliche Anfrage bei Google London

[10] https://translate.google.com/?hl=de

[11] WHALEN, D.H., HOEQUIST, CH.E. & SHEFFERT, S.: The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America* 97, S. 3147-3153, 1995.

[12] PARLIKAR, A. & BLACK, A.: Modeling pause-duration for style-specific speech synthesis. *Proceedings Interspeech*, Portland (OR), 4 S., 2012.

[13] BRAUNSCHWEILER, N. & CHEN, L.: Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. *Proceedings of Speech Synthesis Workshop*, Barcelona, 6 S., 2013.