EIN SPRACHÜBERGREIFENDER VERGLEICH DES PAUSENVERHALTENS NATÜRLICHER SPRECHER IN VERSCHIEDENEN SPRECHTEMPI MIT TTS-SYSTEMEN

Raphael Werner, Jürgen Trouvain, Bernd Möbius

Sprachwissenschaft und Sprachtechnologie, Universität des Saarlandes, Saarbrücken
rwerner@lst.uni-saarland.de

Kurzfassung: Die vorliegende Studie vergleicht die Pausensetzung in natürlicher und in synthetischer Sprache sprachübergreifend (Deutsch, Französisch, Englisch) mit Bezug auf Ort, Dauer und Anzahl der Pausen, sowie hörbare Atemgeräusche. Von den natürlichen Sprechern (drei Sprecher je Sprache) wurden Texte in fünf Sprechgeschwindigkeiten (von sehr langsam bis sehr schnell) vorgelesen. Im Vergleich zur Normalgeschwindigkeit bei natürlichen Sprechern haben die TTS-Systeme (vier Systeme je Sprache) tendenziell langsamere Artikulationsgeschwindigkeiten (in allen drei Sprachen), kürzere Pausendauern (Deutsch, Französisch) und mehr Pausen (Deutsch, Englisch). Die TTS-Systeme unterscheiden sich zusätzlich von den natürlichen Sprechern dadurch, dass sie gänzlich auf hörbare Atemgeräusche in Pausen verzichten. Darüber hinaus zeigten sich individuell verschiedene Strategien der Pausengestaltung der natürlichen Sprecher.

1 Einführung

In natürlicher gesprochener Sprache stellen Pausen elementare Bestandteile von Produktion und Perzeption dar. Text-to-Speech-Synthese (TTS) ist in der Produktion weniger eingeschränkt, so dass Pausen hier allein aus Gründen der Perzeption verwendet werden können. An welchen Stellen im Text und wie lange Pausen zu setzen sind, ist jedoch im Gegensatz zur segmentellen Lautstruktur nicht eindeutig vorgegeben. Diese Optionalität bei der Gestaltung von Pausen führt zu unterschiedlichen Strategien, wie syntaktisch bedingte prosodische Phrasengrenzen genutzt werden [1, 2]. Andere Ansätze wie die „performance structures“ von Gee & Grosjean [3] erweitern die prosodisch-syntaktische Betrachtungsweise noch um die Dimension der Länge der einzelnen Konstituenten zwischen den Pausen, die für Pausendauer und darin auftretendes Atemverhalten relevant sind [4]. Sprachsynthese-Systeme hingegen greifen oft auf Interpunktion als einzigen Hinweis zur Pausensetzung zurück. Allerdings gibt es bezüglich der Pausen außerhalb der Interpunktion große Unterschiede zwischen verschiedenen Systemen [1].

Änderungen des Sprechtempos sind in natürlicher Sprache häufig zu beobachten, z.B. als habituelle Tempi, die sich individuell stark unterscheiden können, als Ausdruck von Affektivität (Äußerungen mit Niedergeschlagenheit und Langeweile sind langsamer als solche mit aufgeregter Freude) oder auch Adaption an den Hörer (zu uns wohlbekannten Personen sprechen wir schneller als zu Unbekannten, Schwerhörigen und Nichtmuttersprachlern). Dabei werden bei Tempoerhöhung Pausen gekürzt oder ganz weggelassen, während andere Phänomene wie phrasenfinale Dehnung robuster sind [5].

Große Unterschiede zwischen natürlichen und synthetischen Sprechern werden bei Tempoveränderungen deutlich, wobei TTS-Systeme hauptsächlich Lautdauern linear ändern, natürliche Sprecher aber zusätzlich Veränderungen der Anzahl der prosodischen Phrasengrenzen und

```
Am nächsten Tag [1] fuhr ich nach Husum. [2] Es ist
eine Fahrt [3] ans [4] Ende der Welt. [5] Hinter Gie-
ßen [6] werden die Berge und Wälder [7] eintönig, [8]
hinter Kassel [9] die Städte [10] ärmlich [11] und bei
Salzgitter [12] wird das [13] Land flach [14] und öde.
[15] Wenn bei uns Dissidenten verbannt würden, [16]
würden sie ans Steinhuder Meer verbannt.
```

Abbildung 1 – Der verwendete deutsche Text mit Angabe der von mindestens einem der drei Sprecher genutzten Pausenorte. Zahlen in eckigen Klammern entsprechen der laufenden Nummerierung der Pausen, nicht der Anzahl der an dieser Stelle realisierten Pausen.

Pausen und in der Lautstruktur über Assimilationen, Reduktionen und Tilgungen von Lauten und Silben vornehmen [6].

Unterschiede gibt es außerdem in Bezug auf Atemgeräusche. Diese bieten in natürlicher Sprache Vorteile für Sprecher und Empfänger. In synthetischer Sprache kann dieser Effekt ebenfalls beobachtet werden [7], kann aber auch ausbleiben [8, 9, 10].

2 Methode

2.1 Material

Die in der vorliegenden Studie verwendeten Daten stammen aus dem BonnTempo-Corpus (BTC) [11], für das Texte im Umfang von ca. 80 Silben bzw. 50 Wörtern) in fünf Sprechgeschwindigkeiten vorgelesen wurden: sehr langsam (L1), langsam (L2), Normalgeschwindigkeit (NO), schnell (S1), sehr schnell (S2). Für die drei untersuchten Sprachen (Deutsch, Englisch, Französisch) wurden je drei Muttersprachler aus dem Korpus zufällig ausgewählt. Die Texte waren für natürliche und synthetische Sprecher identisch und bestanden entweder aus der deutschen Originalversion (Abbildung 1) oder einer sinngemäßen Übersetzung.

Für die synthetische Sprache wurden die im BTC verwendeten Texte von vier TTS-Systemen (IBM Watson TTS - IWTTS [12], Google Cloud TTS - GCTTS [13], Google Translate - GT [14], Oddcast TTS Demo - OCTTS [15]) in den drei untersuchten Sprachen per Web-Interface generiert.

2.2 Analyse

Die 57 Versionen ($3\ \mathit{Sprecher} \times 5\ \mathit{Tempi} \times 3\ \mathit{Sprachen}\ +\ 4\ \mathit{TTS–Systeme} \times 3\ \mathit{Sprachen}$) wurden in Praat [16] hinsichtlich der Parameter Artikulations- bzw. Sprechdauer, Pausendauer, Pausenorte, Pausenanzahl und hörbare Atemgeräusche untersucht.

Dabei wurden alle perzipierten Pausen unabhängig von Schwellenwerten bezüglich Mindestdauer berücksichtigt. Wenn eine Pause von einem Plosiv gefolgt wurde, wurden von der Pausendauer 30 ms abgezogen, um die Verschlussphase, mit der die nächste Artikulationsphase beginnt, nicht zur Pause zu zählen. Für alle synthetisch und natürlich gesprochenen Texte wurden außerdem die Artikulationsdauer (ohne Pausen) und die gesamte Sprechdauer (einschließlich Pausen) gemessen.

Tabelle 1 – Durchschnittliche Artikulationsdauern (in Sekunden) der natürlichen Sprecher in den fünf Sprechtempi (L1, L2, NO, S1, S2) und Artikulationsdauern der TTS-Systeme (IWTTS, GCTTS, GT, OCTTS), jeweils in Deutsch (D), Englisch (E) und Französisch (F).

	L1	L2	NO	S1	S2	IWTTS	GCTTS	GT	OCTTS
D	18,42	16,02	13,81	12,84	9,43	16,74	16,43	20,01	15,84
E	18,08	15,67	12,58	10,84	8,76	14,44	14,76	18,60	15,05
F	19,22	18,27	14,83	12,53	9,84	15,98	13,81	18,56	15,33

Tabelle 2 – Durchschnittliche Pausendauern (in Sekunden) der natürlichen und synthetischen Sprecher.

	L1	L2	NO	S1	S2	IWTTS	GCTTS	GT	OCTTS
D	3,74	2,83	2,23	1,34	0,14	1,86	1,71	1,79	1,69
E	5,49	4,22	1,79	1,23	0,41	1,91	1,76	1,90	2,82
F	6,50	6,03	3,82	1,40	0,24	2,16	1,67	2,26	1,74

3 Ergebnisse

3.1 Artikulationsdauer und Sprechdauer

Tabelle 1 gibt eine Übersicht über die Artikulationsdauern in den verschiedenen Sprechtempi der natürlichen Sprecher im Vergleich mit den TTS-Systemen. Die jeweiligen Sprechdauern lassen sich in Abbildung 2 als Summe aus Artikulations- und Pausendauern ablesen.

Bei den natürlichen Sprechern zeigt sich erwartungsgemäß in allen Sprachen eine Abnahme der Artikulationsdauer mit steigendem Tempo. Die Artikulationsdauern der TTS-Systeme waren im Deutschen etwa auf dem Niveau von L2 der natürlichen Sprecher, bei GT war die Dauer länger als L1. Im Englischen waren die Artikulationsdauern der TTS-Systeme zwischen L2 und NO der natürlichen Sprecher. Ausnahme war auch hier GT, das sich etwas über der Dauer von L1 befand. Im Französischen bewegten sich die TTS-Systeme auf Höhe der Werte zwischen NO und L2 der natürlichen Sprecher mit Ausnahme von GT, das über L2 lag, und GCTTS, das unter NO war.

3.2 Pausendauer

Die verschiedenen Pausendauern im Vergleich, d.h. natürliche Sprecher in verschiedenen Tempi und TTS-Systeme, sind in Tabelle 2 zu sehen. Außerdem sind die Pausendauern zusammen mit den Artikulationsdauern in Abbildung 2 visualisiert.

Die Pausendauern nehmen, wie zu erwarten, bei den natürlichen Sprechern mit zunehmendem Tempo ab. Im Deutschen sind die Pausendauern der natürlichen Sprecher und auch der TTS-Systeme kürzer als die vergleichbaren Sprecher in den anderen Sprachen. In Bezug auf die Normalgeschwindigkeit gibt es bei den TTS-Systemen weniger Pausenzeit im Deutschen und Französischen.

3.3 Pausenorte

Bezüglich der Stellen im Text, an denen Pausen von mindestens einem Sprecher gesetzt wurden, gab es sprachübergreifend kleine Unterschiede. Während im Deutschen insgesamt 16 verschiedene Orte für eine Pause genutzt wurden, waren es im Englischen 19 und im Französischen 14.

Die TTS-Systeme haben ausschließlich Orte für Pausen verwendet, die auch von natürlichen Sprechern genutzt wurden. Dafür wurde häufig, aber nicht immer Interpunktion als Aus-

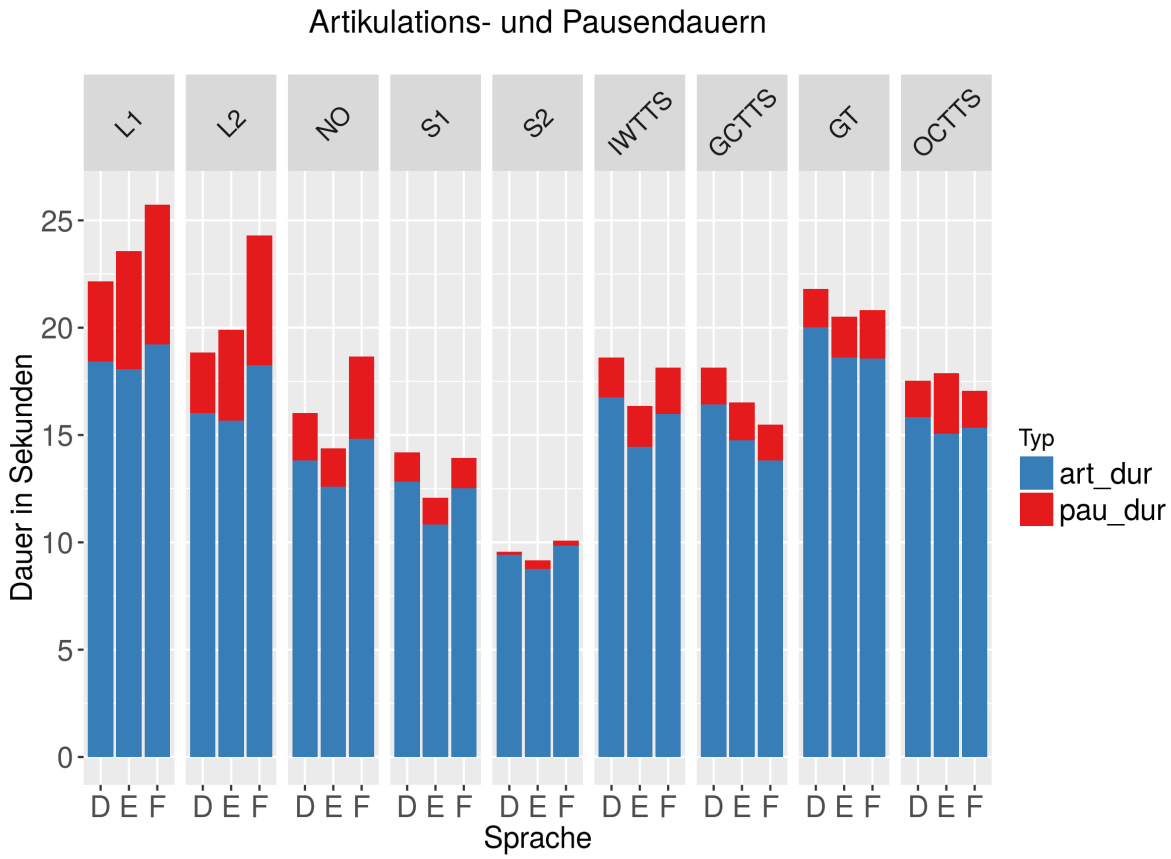

Abbildung 2 – Vergleich der Artikulations- (art_dur) und Pausendauern (pau_dur) nach Sprache und Sprechtempo bzw. TTS-System.

löser für eine Pause verwendet. So gibt es im Deutschen zwei mögliche Pausenorte ohne Interpunktion, an denen zwei bzw. alle vier TTS-Systeme eine Pause setzten, wobei bei einer Stelle die Konjunktion *und* vorausging. Bei den englischsprachigen Systemen gab es nur eine Pause, die nicht durch Interpunktion ausgelöst wurde und die auch nicht in der Nähe einer Konjunktion ist. In synthetischer französischer Sprache gab es wiederum vier Pausen, die nicht an Interpunktionen verortet waren. Diesen folgte dafür in drei der vier Fälle die Konjunktion *et*. Der Großteil dieser Pausen war außerdem kürzer als 100 ms und damit vergleichsweise kurz.

3.4 Pausenanzahl

Tabelle 3 gibt eine Übersicht über die Anzahl der Pausen. Die Pausenanzahl der TTS-Systeme sind im Vergleich zur Normalgeschwindigkeit der natürlichen Sprecher im Deutschen und Englischen ähnlich, aber etwas höher. Im Französischen treten bei den natürlichen Sprechern deutlich mehr Pausen auf als bei den TTS-Systemen, wobei generell die französischen Sprecher auch mehr Pausen gesetzt haben als die deutschen und englischen.

Tabelle 3 – Durchschnittliche Pausenanzahl der natürlichen Sprecher in den fünf Sprechtempi sowie Pausenanzahl in den TTS-Systemen, jeweils in den drei Sprachen.

	L1	L2	NO	S1	S2	IWTTS	GCTTS	GT	OCTTS
D	7,7	7,3	5,7	4,3	0,7	7	7	6	6
E	13,3	11,0	5,0	5,7	2,3	6	6	7	6
F	11,0	10,3	9,3	5,7	0,7	8	5	6	5

Tabelle 4 – Anzahl der Pausen mit hörbaren Atemgeräuschen in verschiedenen Sprachen, Sprechtempi (Durchschnitt der natürlichen Sprecher) und TTS-Systemen (IWTTS, GCTTS, GT, OCTTS).

	L1	L2	NO	S1	S2	IWTTS	GCTTS	GT	OCTTS
D	5,7	4,3	4,0	3,0	0,3	0	0	0	0
E	5,3	4,7	4,0	3,3	1,0	0	0	0	0
F	5,3	5,3	5,3	2,3	0,7	0	0	0	0

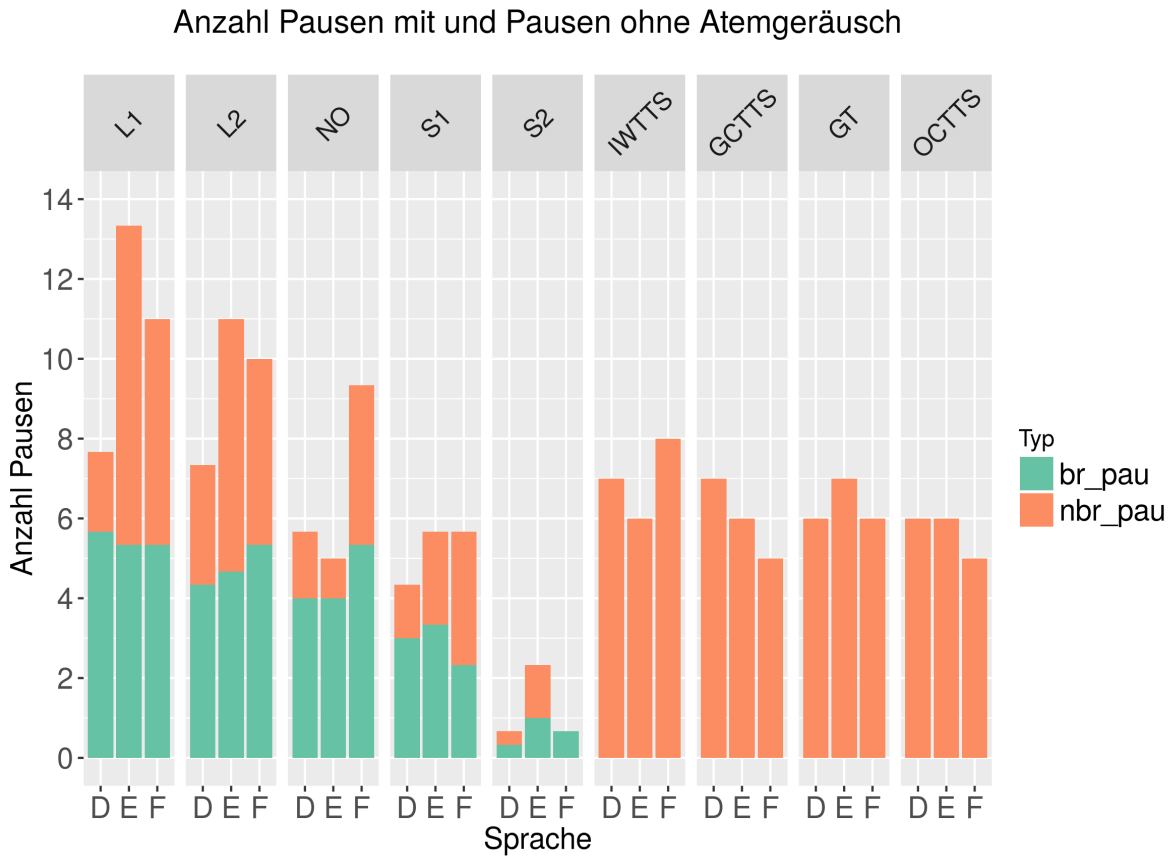

Abbildung 3 – Anzahl der Pausen ohne (nbr_pau) und mit Atemgeräusch (br_pau) nach Sprachen und Sprechtempi sowie in den TTS-Systemen.

3.5 Hörbare Atemgeräusche

Tabelle 4 gibt eine Übersicht über das Vorkommen von Pausen mit hörbarem Atemgeräusch. Die natürlichen Sprecher und TTS-Systeme sind sich untereinander in Bezug auf hörbare Atemgeräusche in Pausen über die Sprachgrenzen hinweg sehr ähnlich, während bei diesem Parameter der größte Unterschied zwischen natürlicher und synthetischer Sprache deutlich wird: Die untersuchten TTS-Systeme verwendeten keine Atemgeräusche.

Zusammenfassend bietet Abbildung 3 einen Vergleich der Anzahl der Pausen zwischen natürlichen (für die fünf Sprechtempi) und synthetischen Sprechern als Summe der Pausen ohne und mit Atemgeräusch in den drei Sprachen.

4 Diskussion und Zusammenfassung

Neben den gezeigten Durchschnittswerten gibt es auch Unterschiede bei den individuellen Pausenstrategien. Abbildung 4 zeigt das Pausenverhalten eines einzelnen deutschen Sprechers, der sich quasi idealtypisch verhält. Auffällig ist dabei, dass sich bei fast jeder Erhöhung des Tempos

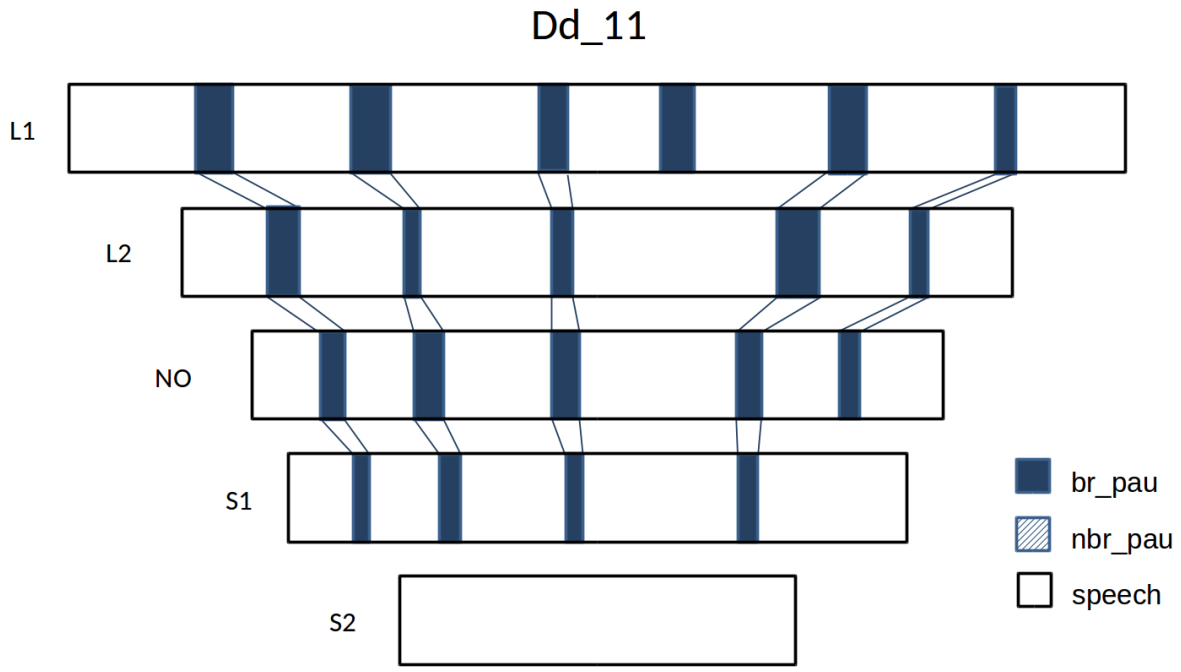

Abbildung 4 – Das Pausenverhalten eines deutschen Sprechers (Dd_11) visualisiert: Die langen schwarzen Rechtecke zeigen die Gesamtdauern des Lesens in den verschiedenen Tempi, die darin befindlichen dunkelblauen Rechtecke die Pausendauer. Die Länge der Kästen entspricht den jeweiligen Dauern. Verbindungen zwischen den Tempi deuten an, dass der gleiche Pausenort verwendet wurde (vgl. Abbildung 1).

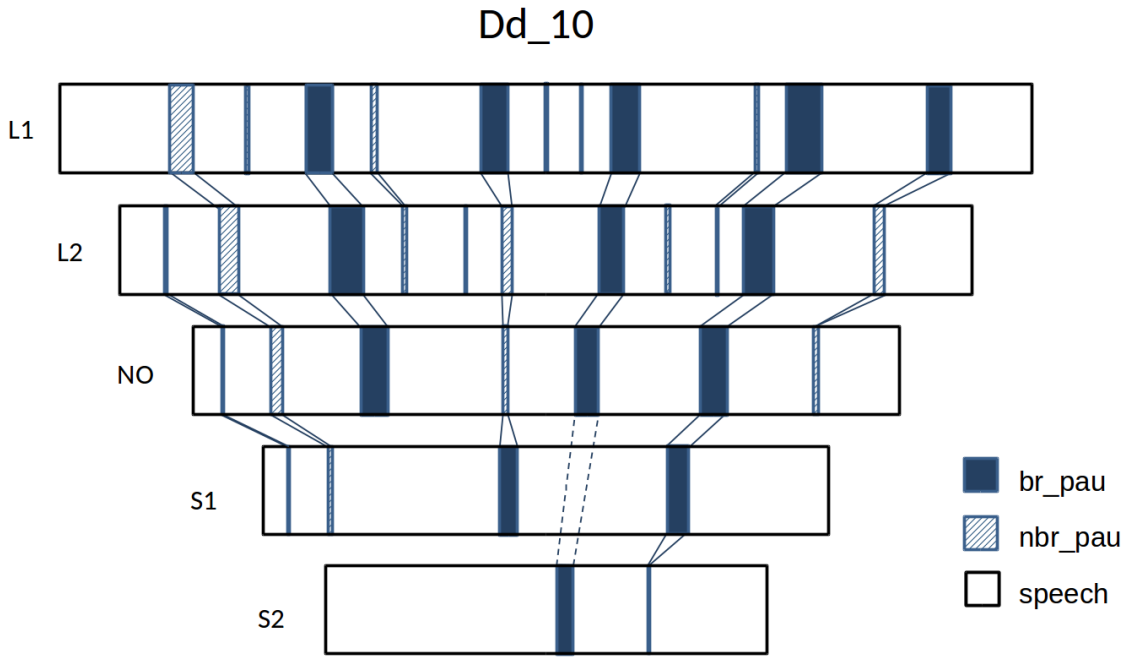

Abbildung 5 – Das Pausenverhalten eines weiteren deutschen Sprechers (Dd_11). Die blau schraffierten Kästen stehen für Pausen ohne hörbares Atemgeräusch. Die gestrichelte Verbindungslinie zwischen Pausen deutet an, dass dieser Pausenort im benachbarten Tempo nicht genutzt wurde, danach aber wieder auftaucht.

die Anzahl der Pausen verringert und auch die Dauer der Pausen abnimmt. Pausen fallen in den schnelleren Tempi weg, beim schnellsten Tempo ist sogar überhaupt keine Pause vorhanden. Zudem verwendet der Sprecher ausschließlich Pausen mit hörbarem Atemgeräusch.

Ein anderes Bild und mehr Variabilität zeigen sich in Abbildung 5. Während auch hier universelle Tendenzen wie Erhaltung einiger *wichtigerer* Pausen, Verringerung der Pausenanzahl und Verkürzung der Pausendauer bei Tempoanstieg größtenteils zu erkennen sind, tauchen hier in L2 Pausen auf, die in L1 nicht realisiert wurden. Außerdem sind hier Pausen ohne Atemgeräusch zu finden, zum Teil an ähnlichen Orten wie Atempausen. Im Vergleich zum idealtypischen Sprecher gibt also mehr Inkonsistenz und Variabilität bei der Pausengestaltung, wie sie womöglich mehrheitlich in natürlicher Sprache auftritt.

Variationen des Sprechtempos brachten in dieser Studie Veränderungen der Artikulations-, Sprech- und Pausendauer sowie der Anzahl von Pausen mit und ohne hörbare Atemgeräusche mit sich. Artikulations- und Pausendauer nahmen bei Steigerung des Tempos relativ konstant ab. Bezüglich der Anzahl der Pausen unterschieden sich die Pausen mit Atemgeräuschen in den beiden langsameren Tempi und im Normaltempo (teilweise auch in S1) vergleichsweise wenig voneinander und nahmen erst bei den schnelleren Tempi, besonders bei S2, stark ab.

Tendenziell ist bei den TTS-Systemes die Artikulationsgeschwindigkeit, ausgedrückt durch die Artikulationsdauer, langsamer als die Normalgeschwindigkeit der natürlichen Sprecher. GCTTS im Französischen ist das einzige TTS-System mit schnellerer Artikulation als die jeweilige Normalgeschwindigkeit.

Bezüglich Gesamtzeit der Pausen und ihrer Anzahl verhalten sich hingegen die TTS-Systeme für Englisch und Deutsch zwischen den Werten für NO und L1 der natürlichen Sprecher. Die Pausendauern waren innerhalb der TTS-Systeme sprachübergreifend relativ konstant und im Englischen großteils über den NO-Werten, im Deutschen und Französischen (teilweise weit) darunter. Die Pausenanzahl der TTS-Systeme war im Deutschen und Englischen etwas höher als die NO-Werte. Im Französischen setzten die natürlichen Sprecher weitaus mehr Pausen als die TTS-Systeme, was hier aber eher an sehr hohen Werten der natürlichen Sprecher lag. Der deutlichste Unterschied ist in den Atempausen zu finden, da keines der untersuchten TTS-Systeme Atemgeräusche verwendete. Hier waren sich die natürlichen Sprecher in den jeweiligen Tempi sprachübergreifend sehr ähnlich.

Da TTS-Systeme in verschiedenen Bereichen und von verschiedenen Personen angewendet werden, ist es denkbar schwierig, eine allgemein gültige bevorzugte Artikulationsgeschwindigkeit festzulegen. Bei der hier festgestellten langsameren Artikulationsdauer könnte sich in der synthetischen Sprache eine Anpassung der Pausendauern bei Normalgeschwindigkeit als nützlich erweisen. Ob Atemgeräusche in synthetischer Sprache verwendet werden sollten, hängt schließlich nicht nur von möglichen Vorteilen für die Informationsverarbeitung des menschlichen Rezipienten, sondern auch von persönlichen Präferenzen und Erfahrungen ab.

Literatur

[1] TROUVAIN, J. und B. MÖBIUS: *Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache*. In *29. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, S. 334–341. Ulm, 2018.

[2] GRICE, M. und S. BAUMANN: *An introduction to intonation – functions and models*. In J. TROUVAIN und U. GUT (Hrsg.), *Non-Native Prosody. Phonetic Description and Teaching Practice*, S. 25–51. De Gruyter, Berlin, New York, 2009. doi:10.1515/9783110198751.1.25.

[3] GEE, J. P. und F. GROSJEAN: *Performance structures: A psycholinguistic and linguistic appraisal*. *Cognitive Psychology*, 15(4), S. 411–458, 1983. doi:10.1016/0010-0285(83)90014-2.

[4] FUCHS, S., C. PETRONE, J. KRIVOKAPIĆ, und P. HOOLE: *Acoustic and respiratory evidence for utterance planning in German*. *Journal of Phonetics*, 41(1), S. 29–47, 2013. doi:10.1016/j.wocn.2012.08.007.

[5] ŻYGIS, M., J. TOMLINSON, C. PETRONE, und D. PFÜTZE: *Acoustic cues of prosodic boundaries in German at different speech rate*. In *Proceedings of 19th International Congress of Phonetic Sciences*, S. 999–1003. Melbourne, 2019.

[6] TROUVAIN, J.: *Temposteuerung in der Sprachsynthese durch prosodische Phrasierung*. In *13. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, S. 294–301. 2002.

[7] WHALEN, D. H., C. E. HOEQUIST, und S. M. SHEFFERT: *The effects of breath sounds on the perception of synthetic speech*. *The Journal of the Acoustical Society of America*, 97(5), S. 3147–3153, 1995. doi:10.1121/1.411875.

[8] TROUVAIN, J. und B. MÖBIUS: *Einatmungsgeräusche vor synthetisch erzeugten Sätzen - eine Pilotstudie*. In *24. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, S. 50–55. Bielefeld, 2013.

[9] BRAUNSCHWEILER, N. und L. CHEN: *Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS*. In *8th ISCA Workshop on Speech Synthesis*, Nr. July, S. 1–6. 2013. URL `http://ssw8.talp.cat/papers/ssw8_OS1-1_Braunschweiler.pdf`.

[10] BERNARDET, U., S. H. KANQ, A. FENG, S. DIPAOLA, und A. SHAPIRO: *Speech Breathing in Virtual Humans: An Interactive Model and Empirical Study*. *2019 IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*, S. 1–9, 2019. doi:10.1109/VHCIE.2019.8714737.

[11] DELLWO, V., I. STEINER, B. ASCHENBERNER, J. DANKOVI, und P. WAGNER: *BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate*. In *Proceedings of Interspeech 2004*, S. 777–780. Jeju Island, Korea, 2004.

[12] *IBM Watson TTS*. Abgerufen am 14.11.2019. URL `https://text-to-speech-demo.ng.bluemix.net/`.

[13] *Google Cloud TTS*. Abgerufen am 14.11.2019. URL `https://cloud.google.com/text-to-speech/`.

[14] *Google Translate*. Abgerufen am 14.11.2019. URL `https://translate.google.com/`.

[15] *Oddcast TTS Demo*. Abgerufen am 14.11.2019. URL `http://ttsdemo.com/`.

[16] BOERSMA, P. und D. WEENINK: *Praat: doing phonetics by computer (Version 6.0.04)*. 2019. URL `http://www.praat.org/`.