MUSTER DER SPRECHATMUNG IN VERSCHIEDENEN SPRECHSTILEN – EINE PILOTSTUDIE

Jürgen Trouvain, Raphael Werner

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

Kurzfassung: Basierend auf einem Standardtext sind akustische, atmungskinematische und laryngographische Signale in sechs Sprechbedingungen untersucht worden: lautes Lesen, Vorlesen mit wenigen Pausen sowie nach physischer Anstrengung, stilles Lesen, lautloses Artikulieren und Nacherzählen. Die Ergebnisse zeigen teilweise deutliche Unterschiede zwischen den Bedingungen und zwischen individuellen Sprechern. Es zeigt sich, dass das komplexe Aufnahmesetting für die Erfassung der beim Sprechen stattfindenden Anpassung der Atmung in verschiedenen Sprechsituationen geeignet ist.

1 Einführung

Ziel dieser Pilotstudie ist es, Muster der Sprechatmung in verschiedenen Sprechstilen zu explorieren. Atmungsverhalten ist hochgradig anpassungsfähig, was beim Sprechen besonders deutlich hervortritt. Im Vergleich zur Ruheatmung zeigt Sprechatmung eine niedrigere Atemfrequenz, kürzere Einatmung sowie längere und viel variablere Ausatmungsphasen [1, 2]. Dabei passt sich – in aller Regel – die Einatmung an die Gestaltung der Sprechpausen an [3]. Umgekehrt können Pausen auch den Atmungsbedingungen unterworfen sein, z. B. bei Sprechen nach körperlicher Anstrengung [4] mit inter alia höherer Atemfrequenz und sehr deutlich hörbaren Ein- und Ausatmungsgeräuschen.

Der Fokus der vorliegenden Pilotstudie liegt auf der Erfassung der Sprechatmung bei Anpassung an verschiedene Sprechbedingungen bei fast identischem Sprechinhalt. Dabei werden akustische Signale und solche der Atmung (kinematisch) sowie der Phonation erfasst.

Die Kopplung phonatorischer und respiratorischer Ereignisse ist zwar zuweilen beim Singen untersucht worden [5, 6], aber selten beim Sprechen. Dabei wäre es wichtig, Details z.B. zu Knarrstimme zu erfassen [7], die im Deutschen oftmals am Äußerungsende vorkommt [8] und somit direkt mit nachfolgender Pause und Einatmung verbunden ist. Konträr zum Sprechen hinterlässt Atmung in Ruhehaltung keine oder kaum hörbare Spuren, so dass eine Überprüfung der tatsächlich stattgefundenen In- und Exhalation mittels kinematischer Daten verlässlicher und genauer ist als das akustische Signal alleine [2, 3].

2 Methoden

2.1 Versuchspersonen

Drei Vpn standen zur Verfügung: Vp1 (männlich, 55 Jahre, Muttersprache (L1): Deutsch); Vp2 (weiblich, 55 Jahre, L1: Deutsch); Vp3 (männlich, 33 Jahre, L1: Englisch).

2.2 Lese- und Sprechbedingungen

Der in der Phonetik oftmals verwendete Standardtext *Nordwind und Sonne* bzw. *The North Wind and the Sun* wurde in der jeweiligen Muttersprache in folgenden Bedingungen aufgenommen:

1. stilles Durchlesen (ST-LES),
2. normales Vorlesen (NORMAL),
3. Vorlesen mit möglichst wenig Pausen (WEN-PAU),
4. Vorlesen nach körperlicher Aktivität (AKTIV),
5. Nacherzählen in eigenen Worten (NACHERZ),
6. lautloses Artikulieren (LL-ART).

Die letzte Bedingung wurde erst ab Vp3 hinzugefügt und wird auch als 'silent speech' bezeichnet. 'Silent speech interfaces' finden Anwendung z. B. in Szenarien, in denen aus Vertraulichkeit nicht laut gesprochen werden kann, die laute Umgebung das Gesprochene überdeckt, oder Menschen aus physiologischen Gründen nicht hörbar sprechen können [9]. Daher beschäftigen wir uns in dieser Studio mit zwei stillen Varianten (ST-LES und LL-ART), wobei bis auf ST-LES alle artikuliert sind und bis auf NACHERZ alle gelesen sind.

2.3 Signalerfassung

Wie in Abbildung 1 ersichtlich wurden drei Arten von Signalen aufgezeichnet: Akustik, kinematische Atmungsaktivität mittels respiratorischer Induktanzplethysmographie (RIP) und Elektroglottographie (EGG). Das akustische Signal wurde mit einem ECM-500L/SK Ansteckmikrofon (Monacor, Bremen) erfasst, das beim EGG-System enthalten war. Es wurde ca. 20 cm unterhalb des Kinns an der Kleidung befestigt. Die RIP-Signale wurden mit dem System *RespTrack* (Version 2.0; Columbi Computers, Stockholm) aufgezeichnet, wie es in [7] verwendet wird. Dazu kommen zwei Gurte zum Einsatz, jeweils einer am Brustkorb und am Bauch. Stimmlippenschwingungen wurde mit dem EGG-D800 (Laryngograph, Wallington, UK) über zwei Elektroden aufgezeichnet.

Die Signale wurden jeweils in der mitgelieferten Software aufgenommen, d.h. Audio- und EGG-Signal in *VoiceSuite10* und Atemsignal in *RTRecorder*. Die Kalibrierung der beiden Gürtel erfolgte durch das Isovolumenmanöver [10]. Wir verwenden hier nur das Summensignal. Bei Vp3 ist in der ersten Bedingung (ST-LES) kein verlässliches Atmungssignal aufgenommen worden, weswegen dies bei den Resultaten fehlt.

Die Ergebnisse beziehen sich innerhalb der Abschnitte des Lesens bzw. Sprechens auf die Messungen der Dauer der tatsächlich stattgefundenen Phasen der Einatmung und Ausatmung und in Bezug auf die Einatmung nicht nur auf deren hörbare Folgen, die im akustischen Signal sichtbar sind. Zu den so gewonnenen Parametern zählen die Gesamtdauer des Lesens/Sprechens in der jeweiligen Bedingung, die Häufigkeit und Dauer der Einatmungsphase und der Ausatmungsphase. Auf letzteren basiert die Dauer des Atemzyklus. Davon abgeleitet werden die Parameter Atemfrequenz (Anzahl der Einatmungsphasen pro Minute Lese-/Sprechzeit) sowie das relative Verhältnis von Ein- zu Ausatmung innerhalb eines Atemzyklus.

2.4 Ablauf der Aufnahme

Alle Vpn standen aufrecht, der Text lag auf einem Notenständer vor ihnen (außer bei NACHERZ). Für die Bedingung AKTIV wurden die Vpn entkabelt, um 5 Etagen im Gebäude hinunter

Abbildung 1 – Vp im schallbehandelten Raum vor der Aufnahme der ersten Bedingung und nach der Verkabelung des EGG-Bandes, des Mikrophons und der beiden RIP-Gurte.

und wieder hoch zu laufen (bzw. Seilspringen bei einer Vp). Die Pulsfrequenz der Vpn wurde mithilfe des Fingerpulsoximeters Pulox PO-200 (Novidion GmbH, Köln) an zwei Zeitpunkten festgestellt: in Ruhe (nach dem stillen Vorlesen) und nach der physischen Aktivität (vor Aufzeichnung von AKTIV). Bei allen Vpn war der Puls vor der physischen Aktivität niedriger als danach: 87 zu 129 bpm (Vp1), 71 zu 84 bpm (Vp2) und 74 zu 115 bpm (Vp3).

3 Resultate

3.1 Dauer von Einatmung, Ausatmung und Atemzyklus

Bezüglich der Dauer der Einatmung (siehe Abbildung 2) fallen zwei Dinge auf: Erstens, im Vergleich zu den tatsächlich gesprochenen Varianten sind wie zu erwarten bei beiden nichthörbaren Varianten die Einatmungsdauern erheblich länger. Zweitens, innerhalb der gesprochenen Varianten ist die Streuung für AKTIV bei zwei der drei Vpn sehr gering.

Die Dauerwerte der Ausatmung (siehe Abbildung 3) sind wie zu erwarten viel länger und viel variabler als diejenigen der Einatmung. Dies betrifft vor allem die gesprochenen Varianten, aber auch LL-ART. Bei WEN-PAU werden hierbei wie beabsichtigt extreme Werte erzielt.

Die individuellen Unterschiede treten vor allem bei ST-LES zu Tage. Diese Unterschiedlichkeit tritt in den gesprochenen Varianten zurück. Bei AKTIV werden die Unterschiede zwischen den Vpn auf ein Minimum reduziert.

In Tabelle 1 sind die Dauerwerte der Atemzyklen für die 3 Vpn dargestellt. Nur zwei konsistente Änderungen über die Vpn hinweg sind erwähnenswert: zum einen die erwartbare extreme Dehnung der Atemzyklen bei WEN-PAU und zum anderen die längeren Atemzyklen bei NACHERZ gegenüber NORMAL, also zwischen Spontan- und Lesesprache.

3.2 Dauerverhältnis von Ein- zu Ausatmung

In Tabelle 2 sind die relativen Anteile der Einatmung für die 3 Vpn dargestellt. Der Wert 36,0 ist als 36,0 % Einatmungsdauer zu lesen, so dass 64,0 % für die Ausatmung der Gesamtdauer eines gemittelten Atemzyklus übrig bleiben.

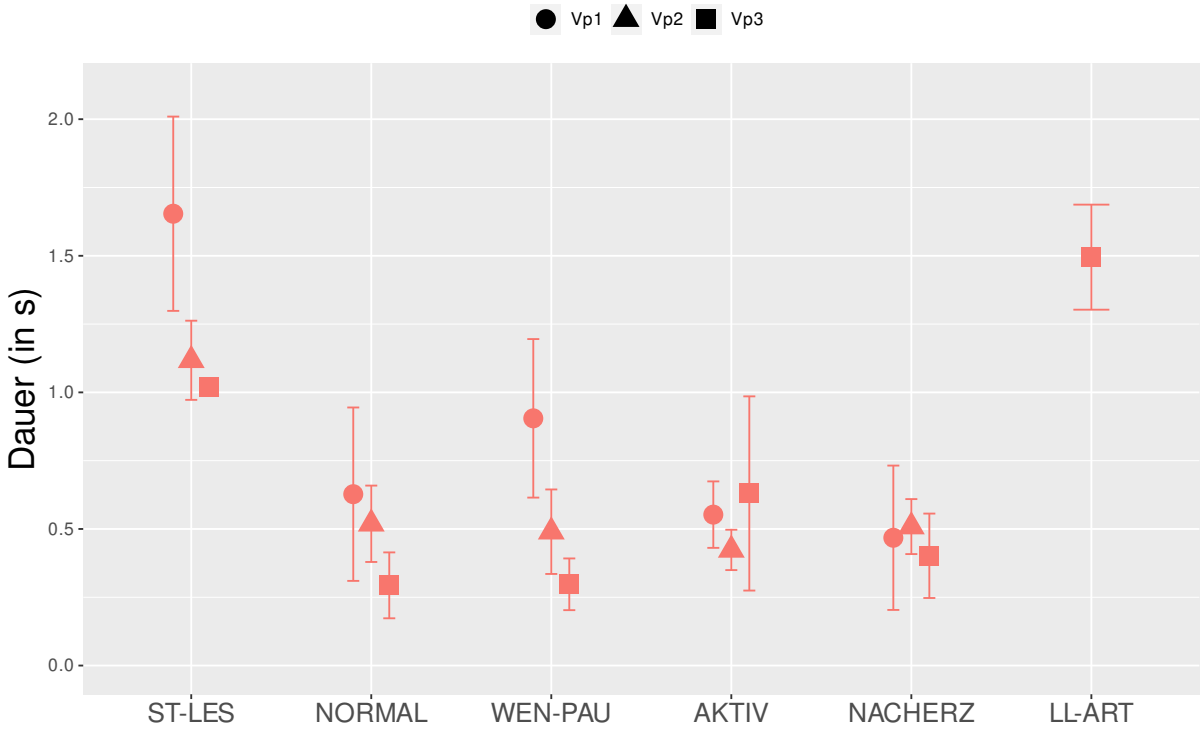

Abbildung 2 – Einatmungsdauern (Mittelwerte und Standardabweichungen) der drei Versuchspersonen in Sekunden, aufgeteilt in die sechs Bedingungen.

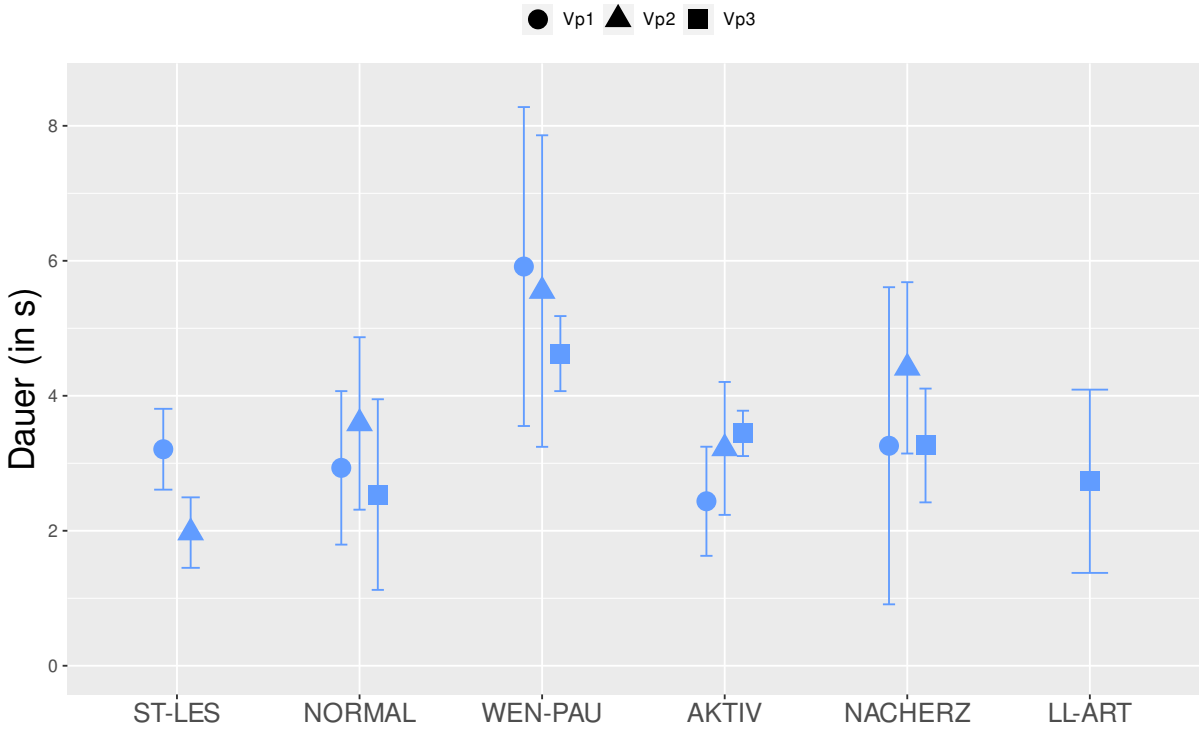

Abbildung 3 – Ausatmungsdauern (Mittelwerte und Standardabweichungen) der drei Versuchspersonen in Sekunden, aufgeteilt in die sechs Bedingungen.

Tabelle 1 – Mittlere Dauerwerte eines Atemzyklus in Sekunden pro Vp und Lese-/Sprechbedingung.

	ST-LES	NORMAL	WEN-PAU	AKTIV	NACHERZ	LL-ART
Vp1	4,86	3,56	6,82	2,99	3,73	
Vp2	3,09	4,11	6,04	3,64	4,92	
Vp3		2,83	4,92	4,07	3,67	4,23

Tabelle 2 – Relative Dauer der Einatmung innerhalb eines Atemzyklus pro Vp und Bedingung in Prozent.

	ST-LES	NORMAL	WEN-PAU	AKTIV	NACHERZ	LL-ART
Vp1	34,0	17,6	13,3	18,5	12,6	
Vp2	36,2	12,6	8,1	11,6	10,3	
Vp3		10,4	6,0	15,5	11,0	35,4

Der Anteil der Einatmung von den stillen Varianten, wo die Einatmung ca. ein Drittel eines Atemzyklus ausmacht, um die Hälfte bis zwei Drittel in den laut gesprochenen Varianten reduziert. Bei WEN-PAU reduziert sich die Dauer bei allen Vpn im Vergleich zu allen anderen Bedingungen.

3.3 Atemfrequenz

Die in Tabelle 3 ausgewiesenen Dauerwerte der Atemfrequenz für die 3 Vpn zeigen kaum eindeutige Richtungen an. Eine Ausnahme ist WEN-PAU, wo jede Vp weniger häufig geatmet hat als bei den anderen Bedingungen, wenn man den einen Datenpunkt bei LL-ART außer Acht lässt. Dieser zeigt eine ähnlich extreme Richtung wie WEN-PAU an.

Tabelle 3 – Atemfrequenz als Anzahl der Inhalationsphasen pro Minute pro Vp und Bedingung.

	ST-LES	NORMAL	WEN-PAU	AKTIV	NACHERZ	LL-ART
Vp1	12,3	16,7	8,8	20,1	16,1	
Vp2	19,8	14,6	10,1	13,6	12,1	
Vp3		23,9	15,2	14,6	17,8	9,6

3.4 Gesamtdauer

Betrachtet man die Gesamtdauer der Aufnahmen in der jeweiligen Lese-/Sprechsituation (zusammengefasst in Tabelle 4), dann fallen wiederum die stillen Varianten als die kürzeren auf im Vergleich zu den laut gesprochenen.

Generell ist NACHERZ bezüglich der Gesamtdauer wegen der offen gestalteten Aufgabe nicht mit den anderen Varianten vergleichbar. WEN-PAU zeigt wie zu erwarten die kürzesten Gesamtdauern bei den laut gesprochenen Varianten.

3.5 Globale Atmungsgestaltung

Die globale Pausengestaltung lässt sich an den Atmungsdaten in Abbildung 4 darstellen, in denen die Daueunterschiede in In- und Exhalationsphasen von Vp1 vergleichend in den verschiedenen Bedingungen dargestellt sind. ST-LES (1) unterscheidet sich von den vier laut gesprochenen Bedingungen (2-5) durch Gesamtdauer, Atemfrequenz sowie Konturformen. WEN-PAU (3) zeigt eine niedrigere Atemfrequenz sowie schnelleres Einatmen als NORMAL (2). AKTIV (4) zeigt uniformere Atemzyklen als die anderen Lese-/Sprechbedingungen sowie sehr schnelle

Tabelle 4 – Gesamtdauer der Lese-/Sprechaufgabe in Sekunden pro Vp und Bedingung.

	ST-LES	NORMAL	WEN-PAU	AKTIV	NACHERZ	LL-ART
Vp1	24,4	53,8	41,0	47,8	52,2	
Vp2	24,2	41,1	35,8	39,7	39,8	
Vp3	20,8	30,1	19,8	37,1	40,4	25,1

Einatmung und tiefe Ausatmung bei jedem Ausatmungszyklus. Beides ist auch sehr gut in jeder Sprechpause hörbar. NORMAL (2) und NACHERZ (5) unterscheiden sich in der Art, wie häufig Pausen mit Einatmung verknüpft und wie die Ausatmungsphasen gestaltet werden.

Auf lokaler Ebene zeigt die zeitliche Alignierung der respiratorischen mit den akustischen Signalen, dass in den Bedingungen WEN-PAU und AKTIV die Einatmungsphasen länger sind als in NORMAL und NACHERZ. Sprechpausen nach Anstrengung bei AKTIV (4) zeigen zuerst eine längere unphonierte Ausatmung, gefolgt von einer Einatmung mit stark erhöhten Amplitudenwerten (akustisch und kinematisch).

3.6 Kontur des RIP-Signals

Die Kontur des Atemsignals ist bei ST-LES (1) Vp-übergreifend sehr ähnlich der Ruheatmung, d.h. wellenförmig und mit ähnlicher Länge bei Ein- und Ausatmung. Hier ist die Reorganisation des Atemzyklus für Sprechatmung nicht in Kraft gesetzt ist, da keine Sprache produziert wird.

Bei den anderen Bedingungen ändert sich die Kontur: So ist bei NORMAL (2) die typische Sägezahnform zu erkennen, die gekennzeichnet ist durch kurze, tiefe Einatmung und in die Länge gezogene Exhalationsphasen, die genutzt werden, um zu sprechen. Der Hauptunterschied zwischen dieser Bedingung und WEN-PAU (3) liegt darin, dass die Ausatemphasen sehr gestreckt werden, um Atempausen zu meiden, wodurch die Ausatmung etwas weniger steil ist, um länger als gewöhnlich sprechen zu können. Im Vergleich zu NORMAL (2) ändert sich die Kontur des Atemsignals besonders beim Vorlesen nach körperlicher Anstrengung, da durch den erhöhten Sauerstoffbedarf hier unphonierte Ausatmung zu sehen ist, was in den anderen Bedingungen gar nicht vorkommt. In der Kontur äußert sich das durch einen raschen, starken Abfall am Ende der Ausatmungsphase. Dieses Phänomen ist ungleich auf die Sprecher verteilt: Während es bei Vp1 sehr ausgeprägt bei quasi jedem Atemzyklus zu beobachten ist, kommt es bei Vp2 und Vp3 nur einmal bzw. dreimal Anfang und Ende vor. Vp1 wies durch die physische Aktivität auch die stärkste Veränderung in der Pulsrate auf.

NACHERZ (5) ist dem normalen Vorlesen relativ ähnlich. Da diese Bedingung die einzige semi-spontansprachliche ist, d.h. die einzige ohne direkte Vorgabe, was zu sprechen ist, gibt es hier einige Häsitationserscheinungen wie z. B. Füllpartikeln. Die Kontur wird dadurch allerdings kaum beeinflusst.

LL-ART (6) wurde wie oben erwähnt nur von einer Person (Vp3) produziert: In den ersten 8,5 s lässt sich gar keine Atemaktivität feststellen. Im Rest der Zeit finden sich dann vier Einatemphasen, die mit im Durchschnitt 1,5 s Dauer ziemlich lang und flach sind, wodurch die Kontur eher der Ruheatmung als der Sprechatmung entspricht. Die Abwesenheit von Atmung im ersten Drittel hat vermutlich eher mit der Ungewöhnlichkeit bzw. seltenen Verwendung von lautlosem Artikulieren als mit der Artikulationsart an sich zu tun.

3.7 Phonatorische Besonderheiten

Phonatorisch auffällig sind nur zwei der 3 Vpn mit unterschiedlicher Verwendung von Knarrstimme. Vp1 knarrt an Satzenden sehr gut hörbar und im EGG sichtbar, allerdings nicht bei NACHERZ und bei AKTIV. Vp3 zeigt Knarrstimme einerseits vor Silben, wenn diese mit einem Vokal beginnen, andererseits bei Häsitationen in den Füllpartikeln.

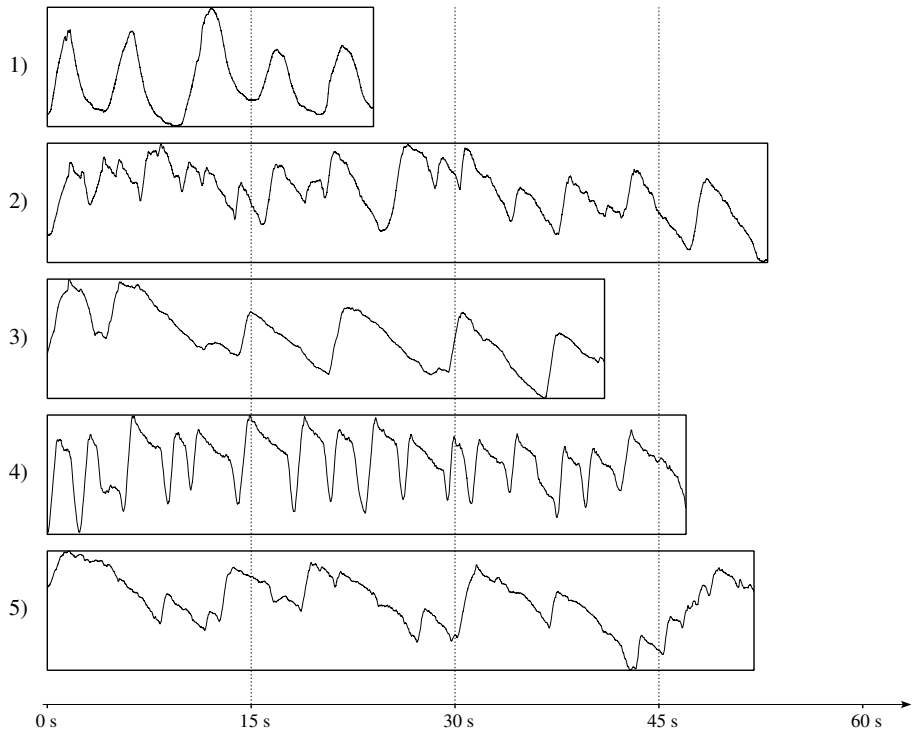

Abbildung 4 – Relative RIP-Summensignale von Vp1 in den ersten 5 Bedingungen; x-Achse: Zeitverlauf in Sek; y-Achse: relative Auslenkung beim Einatmen (nach oben) und Ausatmen (nach unten).

4 Diskussion

4.1 Methodik

Die Durchführung und Kontrolle der relativ komplexen Aufnahmekonstellation mit drei Signalerfassungen und sechs verschiedenen Lese-/Sprechbedingungen sind mit zwei Aufnahmeleitern ohne Probleme durchführbar. Für spätere Aufnahmen soll zusätzlich zu dem mit der EGG-Hardware mitgelieferten Mikrofon ein weiteres verwendet werden, das eine bessere Schallpegelsteuerung erlaubt. Ebenso wäre zur Erfassung vertikaler Larynxbewegungen ein EGG-Setting mit vier statt nur zwei Elektroden besser geeignet. Ein zu behebendes Problem stellt noch die Vermeidung unbeabsichtigt erfasster Herzschlagsignale beim EGG dar.

4.2 Atemsignale

Zusätzlich zu den erwartbaren Ergebnissen gilt es festzustellen, dass die Unterschiede zwischen den individuellen Sprechern stärker als erwartet ausgefallen sind. Bei den Lese-/Sprechbedingungen ist zu Tage getreten, bei welchen Parametern sich welche Bedingungen in welchem Umfang in der Organisation der Atmung unterscheiden. Dabei sind auch Gruppen von Bedingungen wie z. B. stille vs. laut gesprochene zu beachten. Es empfiehlt sich daher in Folgestudien die interindividuellen als auch situationsbedingten Unterschiede stärker in den Fokus zu stellen.

4.3 Phonation

Die Erfassung phonatorischer Ereignisse mittels EGG in Ergänzung zum Mikrofonsignal lässt eine genauere Untersuchung verschiedener Phänomene wie Knarrstimme in präpausaler Position, vor vokalischen Silbenanfängen, vor allem bei Häsitationspartikeln zu. Diese sollen in Folgestudien näher betrachtet werden.

5 Schlussfolgerungen

Wie auf der Basis von [1] und [2] zu erwarten, gibt es etliche Unterschiede bezüglich Atmung und Pausen zwischen den Sprechbedingungen, was Parameter sowohl auf der globalen als auch lokalen Ebene anbelangt. Von Vorteil ist, dass EGG und Respitrack aufgenommen werden, um detailreichere Daten zu gewinnen. Somit rechtfertigt diese Pilotstudie Aufnahmen mit mehr Teilnehmern und den hier getesteten Bedingungen, die über die bloße Gegenüberstellung von Lese- und Spontansprache hinausgehen. Vor allem stilles Mitartikulieren (wie bei 'silent speech interfaces') lässt neue Erkenntnisse bezüglich Atmung und Pausengestaltung erwarten. Bei Leseaufgaben wäre auch die Erfassung von Eyetrackingdaten eine sinnvolle Ergänzung.

Literatur

[1] CONRAD, B. und P. SCHÖNLE: *Speech and respiration. Arch. Psychiatr. Nervenkr.*, 226, S. 251–268, 1979.

[2] WERNER, R., J. TROUVAIN, S. FUCHS, und B. MÖBIUS: *Exploring the presence and absence of inhalation noises when speaking and when listening*. In *Proc. 12th International Speech Production Seminar*, S. 214 – 217. New Haven, CT, USA, 2021.

[3] FUCHS, S. und A. ROCHET-CAPELLAN: *The respiratory foundations of spoken language*. *Annual Review of Linguistics*, 7, S. 1408–1412, 2020.

[4] TROUVAIN, J. und K. TRUONG: *Prosodic characteristics of read speech before and after treadmill running*. In *Proc. Interspeech*, S. 3700–3704. Dresden, 2015.

[5] IWARSSON, J.: *Effects of inhalatory abdominal wall movement on vertical laryngeal position during phonation*. *Journal of Voice*, 15, S. 384–394, 2001.

[6] TERNSTRÖM, S., S. D'AMARIO, und A. SELAMTZIS: *Effects of the lung volume on the electroglottographic waveform in trained female singers*. *Journal of Voice*, 34, S. pp. 485.e1–e.21, 2020.

[7] AARE, K., P. LIPPUS, M. WŁODARCZAK, und M. HELDNER: *Creak in the respiratory cycle*. In *Proc. Interspeech*, S. 1418–1422. Hyderabad, 2018.

[8] TROUVAIN, J. und K. TRUONG: *Three influences on glottalization in read and spontaneous german speech*. In *AIPUK 34*, S. 177–284. Kiel, 1999.

[9] BIRKHOLZ, P., S. STONE, K. WOLF, und D. PLETTEMEIER: *Non-invasive silent phoneme recognition using microwave signals*. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(12), S. 2404–2411, 2018. doi:10.1109/TASLP.2018.2865609.

[10] AUGOUSTI, A. T.: *A theoretical study of the robustness of the isovolume calibration method for a two-compartment model of breathing, based on an analysis of the connected cylinders model*. *Physics in Medicine and Biology*, 42(2), S. 283–291, 1997.