The phonetics of speech breathing: pauses, physiology, acoustics, and perception

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von
Raphael Werner
geboren in Mellrichstadt

Saarbrücken, im Mai 2023

Dekan der Fakultät P: Univ.-Prof. Dr. Stefanie Haberzettl

Erstberichterstatter: Prof. Dr. Bernd Möbius

Zweitberichterstatter: Prof. Dr. Christine Mooshammer

Tag der letzten Prüfungsleistung: 04. Oktober 2023

Acknowledgements

The amount of work that led to this thesis would not have been possible without the help and support of many people and therefore the list of those to whom I am indebted is long. First, I am very grateful to my supervisors Bernd Möbius and Jürgen Trouvain for their guidance and support. You were the perfect match for this role for me – I profited immensely from Bernd's huge experience in the field and technical expertise and Jürgen's expert knowledge on pausology and when to take a pause. I would also like to thank Tine Mooshammer who kindly accepted to be my external reviewer.

I am deeply grateful to Susanne Fuchs for the immense impact she had on my work and me. This thesis would look very different without you, thank you for sharing your knowledge, your collaboration, and the data set you provided, when COVID annihilated any plans to make recordings. I am sure that collaborating on a topic with its absolute expert can be scary, but you made this extremely enjoyable.

I am also very grateful to all the people I met in Saarbrücken, where I not only found help but also very good friends. Iona Gessinger and Jens Neuerburg for making my start here as easy as possible. Jacek Kudera for sharing old Polish wisdom. Bistra Andreeva for being a great neighbor on the fifth floor. Ivan Yuen for sharing his knowledge in so many different fields. Cristina Deeg for running things smoothly in the background. Omnia Ibrahim for caring more about me than I did at times. In the project, I was very fortunate to not only have great supervision but also the greatest PhD colleagues I could have wished for. Beeke Muhlack and Mikey Elmers, who were nothing but supportive and a joy to work with. Beeke, thank you for always having an open ear and good advice. You deserve a medal for bearing with both Mikey and me. Mikey, I can't believe they're letting people like you and me actually get a PhD these days.

I am thankful to the people at the English linguistics department at Julius-Maximilians-Universität Würzburg and especially Barış Kabak, Marie-Christin Himmel, and Christina Domene Moreno – for sparking my interest in phonetics.

Thanks to all the wonderful people I got to meet at conferences, workshops, summer schools, or online or even collaborate with that I will not even attempt to list here.

I have a somewhat uncommon acknowledgement, but I feel very grateful to this project and thesis for being my escapism during unusual times. While the world was out of joint, I could keep coming back to this realm where things felt somewhat controllable.

Last but not least, I owe the deepest thanks to my family. Klara, Andi, and Luca, thank you for always having my back and taking care of things so I could focus on this work. Allie, Pascal, and Arlo, how you handled the last years is nothing short of inspirational. Anne, you remind me to enjoy the good times and pick me up during

the bad times. You make me a better person, thank you for putting up with me. Finally, I am forever grateful to my parents Doris and Michael for their support and believing in me. None of this would have happened without my Mom. For giving your children all the opportunities you were denied. Thank you – for everything. This is for you.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project “Pause-internal phonetic particles in speech communication” (project number: 418659027; project IDs: MO 597/10-1 and TR 468/3-1).

As an associate member of SFB1102 “Information Density and Linguistic Encoding” and its integrated graduate program, I profited greatly from many opportunities for scientific training and exchange (project number: 232722074).

Abstract

Speech is made up of a continuous stream of speech sounds that is interrupted by pauses and breathing. As phoneticians are primarily interested in describing the segments of the speech stream, pauses and breathing are often neglected in phonetic studies, even though they are vital for speech. The present work adds to a more detailed view of both pausing and speech breathing with a special focus on the latter and the resulting breath noises, investigating their acoustic, physiological, and perceptual aspects.

We present an overview of how a selection of corpora annotate pauses and pause-internal particles, as well as a recording setup that can be used for further studies on speech breathing. For pauses, this work emphasized their optionality and variability under different tempos, as well as the temporal composition of silence and breath noise in breath pauses.

For breath noises, we first focused on acoustic and physiological characteristics: We explored alignment between the onsets and offsets of audible breath noises with the start and end of expansion of both rib cage and abdomen. Further, we found similarities between speech breath noises and aspiration phases of /k/, as well as that breath noises may be produced with a more open and slightly more front place of articulation than realizations of /ə/. We found positive correlations between acoustic and physiological parameters, suggesting that when speakers inhale faster, the resulting breath noises were more intense and produced more anterior in the mouth. Inspecting the entire spectrum of speech breath noises, we showed relatively flat spectra and several weak peaks. These peaks largely overlapped with resonances reported for inhalations produced with a central vocal tract configuration. We used 3D-printed vocal tract models representing four vowels and four fricatives to simulate in- and exhalations by reversing airflow direction. We found the direction to not have a general effect for all models, but only for those with high-tongue configurations, as opposed to those that were more open. Then, we compared inhalations produced with the /ə/-model to human inhalations in an attempt to approach the vocal tract configuration in speech breathing. There were some similarities, however, several complexities of human speech breathing not captured in the models complicated comparisons.

In two perception studies, we investigated how much information listeners could auditorily extract from breath noises. First, we tested categorizing different breath noises into six different types, based on airflow direction and airway usage, e.g. oral inhalation. Around two thirds of all answers were correct. Second, we investigated how well breath noises could be used to discriminate between speakers and to extract coarse information on speaker characteristics, such as age (old/young) and sex (female/male). We found that listeners were able to distinguish between two breath noises coming from the same or different speakers in around two thirds of all cases. Hearing one breath noise, classification of sex was successful in around 64%, while for age it was 50%, suggesting that sex was more perceivable than age in breath noises.

Ausführliche Zusammenfassung

Gesprochene Sprache besteht aus einem kontinuierlichen Fluss von Sprachlauten, der durch Pausen unterbrochen wird. Phonetiker[1] sind in erster Linie an der Beschreibung der Segmente des Sprachstroms anhand artikulatorischer, akustischer und wahrnehmungsbezogener Kriterien interessiert. Da dieser Fluss für Pausen und Atmung unterbrochen wird, werden diese Aspekte in phonetischen Studien oft vernachlässigt, obwohl sie für das Sprechen unerlässlich sind: Auf Sprecherseite sind Pausen und Atmung notwendig für die Sprechplanung und das Auffüllen der Luft in der Lunge, das zur Sprachproduktion notwendig ist. Dies geschieht zusammen mit der primären Funktion der Atmung, den Körper mit dem für die Gehirn- und Muskeltätigkeit notwendigen Sauerstoff zu versorgen. Auf Zuhörerseite helfen sie bei der Sprachverarbeitung, indem sie strukturelle Grenzen aufzeigen und die Sprache zeitlich in kleinere Abschnitte aufteilen (Fuchs & Rochet-Capellan, 2021). Pausen können hörbare Atemgeräusche enthalten und lassen sich dementsprechend in Nicht-Atempausen und Atempausen aufteilen. Als solche sind sie Teil des größeren Bereichs der Pausen und pauseninternen phonetischen Partikeln (fortan PINTS). Diese Arbeit konzentriert sich auf zeitliche Aspekte von Atem- und Nicht-Atem-Pausen sowie auf akustische, physiologische und perzeptuelle Aspekte von Sprechatmungsgeräuschen.

Was die Pausen betrifft, so hat sich die Forschung bisher hauptsächlich mit den Fragen befasst, wo diese auftreten (z. B. Gee & Grosjean 1983; Goldman-Eisler 1958; Ferreira 1993), wie lang sie sind (z. B. Campione & Véronis 2002; Kirsner et al. 2003; Godde et al. 2021), und wie sie artikulatorisch ausgeführt werden (z. B. Gick et al. 2004; Ramanarayanan et al. 2013; Krivokapić et al. 2020). Atemgeräusche innerhalb von Pausen wurden weitgehend ignoriert und oft unter der irreführenden Bezeichnung *stille Pausen* subsumiert, obwohl sie phonetisch gesehen nicht stumm sind (Belz & Trouvain, 2019). Die wenigen Studien, die eine Unterscheidung zwischen der gesamten Pause und dem enthaltenen Atemgeräusch vorgenommen haben, fanden stille Abschnitte, sogenannte Edges, auf beiden Seiten des Atemgeräuschs (z. B. Grosjean & Collins 1979; Ruinskiy & Lavner 2007; Fukuda et al. 2018).

Was die Sprechatmung betrifft, so hat sich die Forschung hauptsächlich darauf konzentriert, wie Parameter des Lungenvolumens über das Sprechen variieren, d. h. in der Abfolge von Ausatmung zum Sprechen und dazwischen liegenden Einatmungen. Studien haben Zusammenhänge zwischen der Tiefe und/oder Dauer einer Einatmung und der Länge der folgenden (Gelfer et al., 1983; McFarland & Smith, 1992; Sperry & Klich, 1992; Hoole & Ziegler, 1997; Winkworth et al., 1995; Huber, 2008; Fuchs et al., 2013) und/oder der vorangehenden Äußerung (Kallay et al., 2019; MacIntyre, 2022) gefunden. Der Einfluss von verschiedenen Faktoren wie Alter (Boliek et al., 2009; Godde et al., 2021; Hoit & Hixon, 1987) oder körperlicher Aktivität (Fuchs et al., 2015b; Ng et al., 2022; Serré et al., 2021) auf die Sprechatmung wurde untersucht.

[1] Aus Gründen der Lesbarkeit wird bei Personenbezeichnungen hier die männliche Form gewählt. Die weibliche Form ist jedoch immer mitgemeint.

Atemmuster haben auch im Rahmen der Rhetorik (Barbosa et al., 2019) und im Hinblick auf Pathologien wie die Parkinsonkrankheit (Huber & Darling, 2011; Huber et al., 2012) oder COVID-19 (Nallanthighal et al., 2022) einige Aufmerksamkeit erhalten. Während diese Aspekte untersucht wurden, wurde die akustische Dimension, d. h. die hörbaren Atemgeräusche, die in den Atempausen auftreten, weitgehend vernachlässigt, obwohl diese Geräusche durch den Vokaltrakt geformt werden und somit Informationen über dessen Dimensionen und Konfigurationen liefern können (Whalen & Kinsella-Shaw, 1997). Einige wenige, meist kleinere Studien haben ihre spektralen Eigenschaften beim Sprechen (Kienast & Glitza, 2003), Schnarchen (Ng et al., 2008) oder Singen (Nakano et al., 2008) untersucht.

Was beide Bereiche, d. h. Pausen und Atemgeräusche, benötigen, sind mehr Studien, die sich auf feine phonetische Details konzentrieren. Für Pausen bedeutet dies, sie im Hinblick auf ihre Formbarkeit und die in ihnen auftretenden Partikeln, wie z. B. Atemgeräusche, zu analysieren. Bei den Atempausen verdient die zeitliche Zusammensetzung von stummen Abschnitten und Atemgeräuschen weitere Aufmerksamkeit. Bei der Sprechatmung ist viel zu wenig über die hörbaren Geräusche, die sie erzeugt, bekannt, weshalb diese Arbeit diesen Aspekt näher beleuchten wird. Wir werden uns auf Atemgeräusche konzentrieren, indem wir ihre spektralen Eigenschaften beschreiben, sie mit der Physiologie in Beziehung setzen und sie mit 3D-gedruckten Vokaltraktmodellen simulieren. Außerdem widmen wir uns der Perzeption und werden untersuchen, welche Informationen Atemgeräusche dem Hörer vermitteln können.

Daten und Annotation für Pausen und pauseninterne Partikeln In Kapitel 3 wurden zwei Datenquellen für die Erforschung von Pausen und PINTS betrachtet, nämlich bestehende Korpora und die Erstellung von eigenen, auf Atmungsfragen zugeschnittenen Aufnahmen. Dieses Kapitel basiert auf den Publikationen Werner et al. (2020) und Werner et al. (2022b).

Der erste Teil gab einen Überblick darüber, wie eine Vielzahl von pauseninternen Partikeln in verschiedenen Korpora annotiert wurde. Es wurde gezeigt, dass alle Korpora das Vorkommen von Pausen widergeben, allerdings mit Unterschieden in der Umsetzung und der verwendeten Symbole. Einige berücksichtigten sogar Füllpartikel (*äh/ähm*) oder Atemgeräusche. Die Korpora unterschieden sich jedoch darin, wie diese in der Praxis umgesetzt wurden, da Atemgeräusche manchmal überhaupt nicht erfasst wurden, manchmal wurde die gesamte Pause mit einem Intervall als Atempause versehen, oder sie wurden markiert, aber nicht als eigenes Intervall segmentiert. Ein gemeinsamer Rahmen zur Annotation von PINTS oder alternativ explizite Beschreibungen des Annotationsschemas für Pausen und PINTS könnten helfen, Annotationen zu vereinheitlichen und Transparenz und Reproduzierbarkeit zu fördern. Im zweiten Teil dieses Kapitels haben wir Pilotaufnahmen beschrieben, die dazu dienten, einen Aufbau für zukünftige Aufnahmen zu testen. Dabei zeichneten wir das Atemverhalten mittels Respiratorischer Induktionsplethysmographie (RIP), das glottale Verhalten mittels Elektroglottographie und das Audiosignal auf. Wir untersuchten verschiedene Sprechaufgaben, darunter eine mit möglichst wenigen Pausen, eine nach körperlicher Anstrengung und eine mit stummem Sprechen. Der vorgestellte Aufbau könnte zur Untersuchung der Sprechatmung und der Rolle des Knarrens in glottalen Füllpartikeln und des Atmungszyklus verwendet werden.

Atem- und Nicht-Atempausen in natürlicher und synthetischer gelesener Sprache In Kapitel 4 haben wir untersucht, wie sich unterschiedliche Sprechtempi auf das Pausenverhalten auswirken. Dieses Kapitel basiert auf den Publikationen Werner et al. (2020) und Werner et al. (2022b).

Im ersten Teil analysierten wir Pausen hinsichtlich Lage, Dauer und Anzahl, sowie Anwesenheit hörbarer Atemgeräusche. Dafür wurden natürliche und synthetische Sprache (Deutsch, Französisch, Englisch) berücksichtigt. Die menschlichen Sprecher lasen den Text in fünf Sprechgeschwindigkeiten von sehr langsam bis sehr schnell vor. Vier verschiedene Text-to-Speech-Synthese-Systeme produzierten den Text in den vier Sprachen in der Standardgeschwindigkeit. Neben der Beeinflussung der Sprechparameter, wie eine kürzere Artikulationsdauer, führte eine Erhöhung des Tempos zu kürzeren und weniger Pausen. So konnten wir diese Ergebnisse mit der synthetischen Sprache vergleichen: Dort stellten wir fest, dass die hier einbezogenen Systeme im Vergleich zu menschlichen Sprechern zu langsameren Artikulationsraten in Kombination mit kürzeren und häufigeren Pausen neigten. Diese Ergebnisse deuten darauf hin, dass diese Systeme noch verbesserungsfähig sind, wenn es darum geht, die computergenerierte Sprache an verschiedene Szenarien und Gesprächspartner anzupassen, die möglicherweise ein langsameres oder schnelleres Sprechen erfordern. Das richtige Gleichgewicht zwischen Artikulation und Pausen und das Hinzufügen von Atemgeräuschen zu den entsprechenden Stilen kann dazu beitragen, dass das Pausenverhalten der synthetischen Sprache natürlicher wird. Darüberhinaus beobachteten wir individuell verschiedene Pausenstratiegen der natürlichen Sprecher.

In der zweiten Studie nutzten wir die Tempoänderungen, um uns auf zwei Aspekte von Sprechpausen zu konzentrieren, nämlich ihre Optionalität, d. h. das Vorhandensein oder Nichtvorhandensein an einer bestimmten Stelle, und ihre Variabilität, d. h. die Variation der Dauer und das Vorkommen von Atemgeräuschen. Wir verwendeten hier Lesesprache produziert von 46 Sprechern aus sechs verschiedenen Sprachen (Arabisch, Tschechisch, Englisch, Französisch, Deutsch, Italienisch), die semantisch ähnliche Kurztexte vorlasen. Dabei fanden wir viele optionale Pausen, die nur in wenigen (langsameren) Bedingungen oder von wenigen Sprechern produziert wurden, während andere Pausen über Sprecher und Bedingungen hinweg weniger optional waren. An diesen weniger optionalen Stellen gab es eine große Anzahl von Pausen, die tendenziell länger waren als Pausen an anderen Orten und eher Atmungsgeräusche beinhalteten. Die Variabilität war hier hoch und hing mit dem Tempo zusammen. Darüber hinaus waren sie weitgehend, aber nicht ausschließlich, mit Interpunktion und Konjunktionen verbunden. Bei Tempowechseln passten die Sprecher die Anzahl der Atempausen, vor allem aber die Anzahl der Nicht-Atempausen an. In schnelleren

Tempi wurden Pausen, Einatmungen und stille Edges, um sie herum verkürzt.

Akustik und Physiologie von Atemgeräuschen Das übergeordnete Ziel von Kapitel 5 war die Beschreibung der Akustik von Sprachatemgeräuschen, die wir in drei Experimenten verfolgten. Ziel war es, die spärliche Literatur über die spektralen Eigenschaften und die akustisch-physiologischen Beziehungen von Atemgeräuschen zu ergänzen. Das Kapitel basiert auf den Publikationen Werner et al. (2021b), Werner et al. (2021a), Werner et al. (2022a), and Werner et al. (under review).

Im ersten Experiment verwendeten wir semispontane Sprachdaten, um die zeitliche Abstimmung des akustischen und kinematischen Signals der Atmung, letzteres via Oberkörperausdehnung über RIP, beim Hören und beim Sprechen zu untersuchen. Die Hörbedingung zeigte das erwartete Atemverhalten, d. h. weniger Atemzyklen und längere Einatmungen als bei Sprechatmung. Beim Sprechen stellten wir fest, dass die Ausdehnung des Brustkorbs tendenziell etwas früher erfolgte als die Ausdehnung des Bauches, allerdings gab es ein hohes Maß an Individualität. Das Einsetzen der Einatmungsgeräusche war tendenziell an die Ausdehnung von sowohl Bauch und Brustkorb gekoppelt.

Im zweiten Experiment untersuchten wir akustische und physiologische (über RIP) Merkmale hörbarer Einatmungsgeräusche und wie die beiden Aspekten miteinander korreliert sein könnten. Aus diesen Atemgeräuschen extrahierten wir die Parameter Center of Gravity, Intensität und die ersten drei Formanten. Dieselben Parameter wurden auch aus einigen Sprachlauten extrahiert, um Sprech- und Atemlaute vergleichen zu können mit dem Ziel, sich den Atemgeräuschen auf diesem Weg zu nähern. Dazu verwendeten wir die Affrikationsphasen der stimmlosen Plosive /t k/, den stimmlosen glottalen Frikativ /h/ und den neutralen Vokal /ə/, die von 31 Sprecherinnen des Deutschen produziert wurden. Wir fanden heraus, dass das Center of Gravity der Inhalationen den Aspirationsphasen von /k/ ähnelt, aber nicht denen von /t/ oder Realisierungen von /h/. Die Formantenwerte deuteten darauf hin, dass die Einatmungen im Vergleich zu /ə/ einen offeneren und etwas weiter vorne liegenden Artikulationsort hatten. Darüberhinaus stellten wir fest, dass Center of Gravity, Intensität und der erste Formant der Atemgeräusche positiv mit der Einatmungsgeschwindigkeit korreliert waren. Dies deutet darauf hin, dass, wenn die Sprecher schneller einatmeten, die resultierenden Atemgeräusche intensiver waren und weiter vorne im Mund erzeugt wurden. Die Öffnung des Kiefers und die Einatmungsgeschwindigkeit trugen also wesentlich zu den Einatmungsgeräuschen bei.

Im dritten Experiment untersuchten wir die spektralen Eigenschaften des gesamten Spektrums, um Atemgeräusche besser beschreiben zu können. Im ersten Teil dieses Versuchs fanden wir relativ flache Spektren mit mehreren schwachen Gipfeln in einer großen Stichprobe menschlicher Inhalationen, die von 100 männlichen und 34 weiblichen Sprechern der deutschen Sprache produziert wurden. Die Spektren wiesen eine abnehmende Steigung auf, d. h. eine höhere Intensität bei niedrigen Frequenzen und eine geringere Intensität bei hohen Frequenzen, sowie mehrere schwache Gipfel unterhalb von 3 kHz. Außerdem zeigten die Gipfel eine mäßige (bei weiblichen Sprechern) bis starke (bei männlichen Sprechern) Überlappung mit den Resonanzen, die von Hanna et al. (2018) für Inhalationen berichtet wurden, die von Sprechern des nicht-rhotischen australischen Englisch mit einer Vokaltraktkonfiguration von /ɜː/ produziert wurden. Dies deutet darauf hin, dass die bei der Sprechatmung angenommene Vokaltraktkonfiguration neutral und ähnlich wie /ə/ sein könnte. Im zweiten Teil haben wir 3D-gedruckte Vokaltraktmodelle verwendet, um synthetische Atemgeräusche zu erzeugen, für die die genaue Vokaltraktkonfiguration bekannt ist. Mit Modellen, die die vier Vokale /iː aː uː ə/ und die vier Frikative /x ç ʃ s/ abbildeten, simulierten wir Ein- und Ausatmungen durch Änderung der Luftstromrichtung. Dabei stellten wir fest, dass die Umkehrung des Luftstroms keinen generellen Effekt bei allen Vokaltraktkonfiguration hatte, sondern nur bei denen mit hoher Zungenlage, d. h. /iː ç s ʃ/. Bei allen vier war die Amplitude bei Ausatmung höher, bei /s ʃ/ war auch die Form des Spektrums betroffen. Die Unterschiede entstanden durch den gebündelten Luftstrom, der beim Ausatmen auf die Schneidezähne trifft, nicht aber beim Einatmen. Im dritten Teil haben wir die mit dem /ə/-Modell erzeugten Einatmungen mit menschlichen Einatmungen verglichen, um uns der Vokaltraktkonfiguration bei der Sprechatmung auf diese Weise zu nähern. Wir fanden einige Ähnlichkeiten, jedoch wurde der Vergleich der Spektren erschwert durch verschiedene Komplexitäten der menschlichen Sprechatmung, die in den Modellen nicht erfasst wurden.

Wahrnehmung von Atemgeräuschen In Kapitel 6 widmeten wir uns der Wahrnehmung von Atemgeräuschen und der Frage, wie viel Information Hörer aus den Atemgeräuschen auditiv extrahieren können. Dieses Kapitel basiert auf den Publikationen Werner et al. (2022d) and Werner et al. (2022c).

In der ersten Perzeptionsstudie wurden die Teilnehmer gebeten, verschiedene Atemgeräusche nach dem Hören dem jeweiligen Atemtyp zuzuweisen. Wir verwendeten natürlich vorkommende Stimuli aus einem Corpus. Die Stimuli gehörten sechs verschiedenen Atemtypen an, klassifiziert nach Luftstromrichtung und beteiligter Atemwege (Ausatmung: oral, nasal; Einatmung: oral, nasal, oral gefolgt von nasal, nasal gefolgt von oral). In einem Pilotversuch stellten wir fest, dass Phonetiker nicht besser abschnitten als Laien (je acht Teilnehmer pro Gruppe). Am Folgeexperiment nahmen 80 Laien teil. Insgesamt wurden ca. 66 % aller Atemgeräusche korrekt kategorisiert. Nasale Einatmungen wurden am häufigsten richtig zugewiesen, während dies bei nasalen Ausatmungen am seltensten der Fall war. Wir konnten keine helfende Funktion des Kontexts, d. h. des Hinzufügens von je 1 s Sprechkontext auf beiden Seiten des Atemgeräusches, bei der Kategorisierung feststellen. Es gab eine leichte, nichtsignifikante Tendenz, dass Stimuli mit Kontext häufiger korrekt kategorisiert wurden, allerdings interagierte dieser Faktor mit einigen Atemtypen, sodass Stimuli mit nasaler Einatmung und nasaler-gefolgt-von-oraler Einatmung ohne Kontext häufiger korrekt zugewiesen wurden als mit ihm. Der Grund dafür könnten Unterschiede in der Intensität zwischen Sprechen und Atmungsgeräuschen sein.

In der zweiten Perzeptionsstudie wollten wir herausfinden, wie gut Hörer Atemgeräusche nutzen können, um zwischen verschiedenen oder gleichen Sprechern zu unterscheiden und grobe Informationen über die Merkmale des Sprechers zu extrahieren, wie Alter (alt oder jung) und Geschlecht (weiblich oder männlich). Von sechs jüngeren (Altersspanne: 20–29; 3 weiblich, 3 männlich) und sechs älteren (Altersspanne: 59–65; 3 weiblich, 3 männlich) Sprechern haben wir jeweils fünf Atemgeräusche verwendet. Für eine hohe Vergleichbarkeit haben wir dafür nur Atemgeräusche mit oraler Einatmung verwendet, die im Sprechkontext häufig vorkommen. Den 33 Teilnehmern der Studie wurden zwei Aufgaben gestellt: Eine Unterscheidungsaufgabe, bei der sie zwei Atemgeräusche hörten und gefragt wurden, ob sie von demselben oder einem anderen Sprecher stammten. Eine Klassifizierungsaufgabe, bei der sie jeweils nur ein Atemgeräusch hörten und gefragt wurden, ob es von einem jungen oder alten bzw. männlichen oder weiblichen Sprecher stammte. Die Unterscheidungsaufgabe wurde zu ca. 64 % richtig beantwortet, wobei die niedrigsten Ergebnisse bei Kombinationen von unterschiedlichem Alter und gleichem Geschlecht erzielt wurden. Bei der Klassifizierungsaufgabe wurde die Altersgruppe des Sprechers zu ca. 50 % richtig beantwortet, während es beim Geschlecht ca. 66 % waren. Die Ergebnisse der beiden Aufgaben deuten darauf hin, dass Geschlechtsunterschiede deutlicher wahrnehmbar waren als Altersunterschiede, was auf unterschiedliche Vokaltraktlängen zurückzuführen sein könnte. Die stärkere Wahrnehmbarkeit von Geschlechtsunterschieden steht im Allgemeinen im Einklang mit normaler Sprache, auch wenn die hier verwendeten Atemstimuli sehr kurz, ingressiv und stimmlos waren.

Publications

Parts of this dissertation have appeared in the following publications:

Trouvain, J., & Werner, R. (2020). Comparing annotations of non-verbal vocalisations in speech corpora. In *Laughter and Other Non-Verbal Vocalisations Workshop* (pp. 69–72). Bielefeld.

Trouvain, J., & Werner, R. (2023). Muster der Sprechatmung in verschiedenen Sprechstilen – Eine Pilotstudie. In C. Draxler (Ed.), *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 95–102). Munich: TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=1178.

Trouvain, J., Werner, R., & Möbius, B. (2020). An acoustic analysis of inbreath noises in read and spontaneous speech. In *Speech Prosody* (pp. 789–793). Tokyo. URL: https://doi.org/10.21437/SpeechProsody.2020-161.

Werner, R., Fuchs, S., Trouvain, J., Kürbis, S., Möbius, B., & Birkholz, P. (2022a). Comparison of acoustic parameters of inhalations vs exhalations with 3D-printed vocal tract models. In *International Conference on Speech Motor Control* (pp. 253–254). Groningen. URL: https://doi.org/10.21827/32.8310/2022-115.

Werner, R., Fuchs, S., Trouvain, J., Kürbis, S., Möbius, B., & Birkholz, P. (under review). Acoustics of breath noises in human speech: Descriptive and 3D-modelling approaches. *Journal of Speech, Language, and Hearing Research*, *tbd*.

Werner, R., Fuchs, S., Trouvain, J., & Möbius, B. (2021a). Inhalations in speech: Acoustic and physiological characteristics. In *Interspeech* (pp. 3186–3190). Brno. URL: https://doi.org/10.21437/Interspeech.2021-1262.

Werner, R., Trouvain, J., Fuchs, S., & Möbius, B. (2021b). Exploring the presence and absence of inhalation noises when speaking and when listening. In *International Seminar on Speech Production (ISSP)* (pp. 214–217). New Haven.

Werner, R., Trouvain, J., & Möbius, B. (2020). Ein sprachübergreifender Vergleich des Pausenverhaltens natürlicher Sprecher in verschiedenen Sprechtempi mit TTS-Systemen. In A. Wendemuth, R. Böck, & I. Siegert (Eds.), *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 101–108). Magdeburg: TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=444.

Werner, R., Trouvain, J., & Möbius, B. (2022b). Optionality and variability of speech pauses in read speech across languages and rates. In *Speech Prosody* (pp. 312–316). Lisbon. URL: https://doi.org/10.21437/SpeechProsody.2022-64.

Werner, R., Trouvain, J., & Möbius, B. (2022c). Speaker discrimination and classification in breath noises by human listeners. In *Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)* (pp. 68–69). Prague.

Werner, R., Trouvain, J., Muhlack, B., & Möbius, B. (2022d). Perceptual Categorization of Breath Noises in Speech Pauses. In O. Niebuhr, M. S. Lundmark, & H. Weston (Eds.), *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 139–146). TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=1152.

Contents

1 Introduction **1**

2 Background **5**

2.1 Pauses and pause-internal particles 5

2.1.1 What is a pause? 5

2.1.2 Pause-internal particles 8

2.1.3 Pause locations and frequency 11

2.1.4 Pause duration 12

2.1.5 Relation to speech tempo 14

2.1.6 Pauses in synthetic speech 14

2.1.7 Pause articulation 15

2.2 Speech breathing 17

2.2.1 Breath cycle reorganization 17

2.2.2 Breathing physiology 20

2.2.3 Measuring speech breathing 23

2.2.4 Breath noise acoustics 25

2.2.5 Breath noise perception 28

2.2.6 Factors influencing speech breathing 29

3 Data and annotation for pauses and pause-internal particles **35**

3.1 Annotation of pauses, pause-internal particles, and non-verbal vocalizations: a look at existing corpora 35

3.1.1 Corpora 36

3.1.2 Annotation categories 37

3.1.3 Discussion 38

3.1.4 Conclusion 39

3.2 Pilot recordings of speech breathing 41

3.2.1 Setup 41

3.2.2 Observations 43

3.2.3 Discussion 46

3.2.4 Outlook for further recordings 49

4 Breath and non-breath pauses in natural and synthetic read speech 51

4.1 A cross-linguistic comparison of natural speakers' pausing behavior at different speech tempos with text-to-speech systems 51

4.1.1 Introduction 51

4.1.2 Methods 51

4.1.3 Results 53

4.1.4 Discussion 57

4.1.5 Conclusion 60

4.2 Optionality and variability of speech pauses in read speech across languages and rates 61

4.2.1 Introduction 61

4.2.2 Methods 62

4.2.3 Results 63

4.2.4 Discussion 66

4.2.5 Conclusion 69

5 Breath noise acoustics and physiology 71

5.1 Exploring the presence and absence of inhalation noises when speaking and when listening 71

5.1.1 Introduction 71

5.1.2 Methods 72

5.1.3 Preliminary analysis 72

5.1.4 Advanced analysis 75

5.1.5 Discussion 78

5.1.6 Conclusion 79

5.2 Inhalations in speech: acoustic and physiological characteristics . . . 80

5.2.1 Introduction 80

5.2.2 Methodology 81

5.2.3 Results 83

5.2.4 Discussion 86

5.2.5 Conclusion 88

5.3 Acoustics of breath noises in human speech: Descriptive and 3D-modeling approaches 89

5.3.1 Introduction 89

5.3.1.1 Speech breathing: physiology and acoustics 89

5.3.1.2 Change of airflow direction: in- vs. exhalations . . . 90

5.3.2 Inhalations in human speakers 91

5.3.2.1 Methods 91

5.3.2.2 Results and discussion 93

5.3.3 In- vs. exhalation in 3D-printed VT models 95

5.3.3.1 Methods 95

5.3.3.2 Results and Discussion 97
5.3.4 General Discussion 100
5.3.5 Conclusion 101

6 Breath noise perception 103
6.1 Perceptual categorization of breath noises in speech pauses 103
6.1.1 Introduction 103
6.1.2 Experiment 1 104
6.1.2.1 Method 104
6.1.2.2 Results 105
6.1.2.3 Discussion 105
6.1.3 Experiment 2 106
6.1.3.1 Method 106
6.1.3.2 Results 107
6.1.3.3 Discussion 108
6.1.4 Discussion of both experiments 109
6.1.5 Conclusion 110
6.2 Using breath noises for speaker discrimination and classification by human listeners 112
6.2.1 Introduction 112
6.2.2 Methods 112
6.2.3 Results 115
6.2.3.1 Discrimination task 115
6.2.3.2 Classification task 117
6.2.4 Discussion 118
6.2.4.1 Discrimination task 118
6.2.4.2 Classification task 119
6.2.5 Conclusion 120

7 General discussion 123

8 Conclusion 133

Appendices 137

A Appendix 138
A Nordwind und Sonne 138
B The North Wind and the Sun 138
C Picture description: Baseball 139

B Appendix 140
A By-speaker inhalation spectra 140

List of Figures 142

List of Tables 146

Bibliography 148

Chapter 1

Introduction

Speech is made up of a continuous stream of speech sounds that is interrupted by pauses. Phoneticians are primarily interested in describing the segments of the speech stream, using articulatory, acoustic, and perceptual criteria. As this stream is paused for pausing and breathing, these aspects are often neglected in phonetic studies, even though they are vital for speech: On the speaker side, pauses and breathing are necessary for planning and replenishing the air inside the lungs required for speech production. This is done along with breathing's primary function of supplying the body with the oxygen necessary for brain and muscle activity. On the listener side, they aid processing by indicating boundaries and temporally structuring speech into smaller chunks (Fuchs & Rochet-Capellan, 2021). Pauses may or may not include an audible breath noise, thus they can be distinguished into *non-breath pauses* and *breath pauses*. As such, they are part of the larger field of pauses and pause-internal particles (henceforth PINTS). This work focuses on temporal aspects of breath and non-breath pauses, as well as acoustic, physiological, and perceptual aspects of speech breath noises.

On the pausing side, research so far has mainly addressed the questions where they occur (e.g. Gee & Grosjean 1983; Goldman-Eisler 1958; Ferreira 1993), how long they are (e.g. Campione & Véronis 2002; Kirsner et al. 2003; Godde et al. 2021), and how they are performed articulatorily (e.g. Gick et al. 2004; Ramanarayanan et al. 2013; Krivokapić et al. 2020). What exactly a pause is can be rather fuzzy across studies, and consequently there are major differences in what particles, i.e. sounds, noises, or vocalizations, may or may not be included and if that should be distinguished at all, as reflected by the terms *silent*, *filled*, or *breath pause*. This is despite the fact that this can have a large impact on their causes, function, duration and location. Breath noises within pauses have largely been ignored, as they are often subsumed under the misnomer *silent pauses*, although not being phonetically silent (Belz & Trouvain, 2019). The few studies that have made a distinction between the entire pause and the contained breath noise found silent edges on either side of the breath noise (e.g.

Grosjean & Collins 1979; Ruinskiy & Lavner 2007; Fukuda et al. 2018).

Regarding speech breathing, research has mainly focused on how parameters related to lung volume vary over speech, i.e. exhalations, and the inhalations in between. Studies have found connections between depth and/or duration of an inhalation and the length of the following (Gelfer et al., 1983; McFarland & Smith, 1992; Sperry & Klich, 1992; Hoole & Ziegler, 1997; Winkworth et al., 1995; Huber, 2008; Fuchs et al., 2013) and/or preceding utterance (Kallay et al., 2019; MacIntyre, 2022). The influence of factors such as age (Boliek et al., 2009; Godde et al., 2021; Hoit & Hixon, 1987) or physical activity (Fuchs et al., 2015b; Ng et al., 2022; Serré et al., 2021) on speech breathing has been investigated. Respiratory patterns have also received some attention as part of rhetoric advice (Barbosa et al., 2019) and with regard to pathologies, such as Parkinson's disease (Huber & Darling, 2011; Huber et al., 2012) or COVID-19 (Nallanthighal et al., 2022). While these aspects have been studied, the acoustic dimension, i.e. the respiratory sounds, that occur in breath pauses, have largely been neglected, although these noises are shaped by the vocal tract and thus can provide information about its dimensions and configurations (Whalen & Sheffert, 1997). A few, typically smaller studies have looked at their spectral characteristics in speech (Kienast & Glitza, 2003), snoring (Ng et al., 2008), or singing (Nakano et al., 2008). An overview of previous research on pausing and speech breathing relevant to this work can be found in ch. 2.

What both areas, i.e. pauses and breath noises, need is more studies that focus on their fine phonetic detail. This means for pauses to analyze them with regard to their plasticity and particles that may occur within them, such as respiratory sounds. In breath pauses, the temporal composition of silent stretches and breath noises merits further attention. For speech breathing, far too little is known about the audible noises it creates, thus this work will zoom in on them. We will focus on breath noises by describing their spectral characteristics, correlating them to physiology, and simulating them with 3D-printed vocal tract models. Further, we will investigate what information breath noises may cue to the listener. Descriptions of pauses and breathing are relevant for numerous areas: A more detailed overview of breath pauses, regarding their temporal composition and acoustic characteristics, could improve models of speech production. Some styles of synthetic speech could profit from more natural pausing behavior and potentially enhance perceived naturalness by adding breath noises. Knowing more about the spectral characteristics of breath noises can help improve word aligners by outlining how they differ from speech sounds that may share some features. In addition, it could improve the non-invasive detection of diseases or disorders. Knowing how much information they may cue to listeners has implications for forensic phonetics and synthetic speech.

The major goal of this dissertation is to add to our understanding of pauses and speech breathing, which constitute the two main branches here. On the pausing side, non-breath and breath pauses are investigated. On the breathing side, this work further branches into the two fields of studying the production and perception of audible speech breath noises. Before the experimental part, a chapter will discuss the annotation of pauses and PINTS in different corpora (see ch. 3.1) and the recording of speech, pauses, and breathing (see ch. 3.2). This chapter is based on Trouvain & Werner (2020) and Trouvain & Werner (2023).

Breath and non-breath pauses The studies on pauses with and without breath noises are described in ch. 4. This chapter is based on Werner et al. (2020) and Werner et al. (2022b). In ch. 4.1, we use read speech in three different languages to study the plasticity of pauses as a function of different speech tempos. We also include text-to-speech synthesis systems to investigate how computer-generated speech at default speed compares to articulation and pausing rate of human speakers at different tempos. We find that higher tempos lead to shorter and fewer pauses, along with shorter articulation duration. Synthetic speech systems, in comparison, use slower articulations, while tending to use fewer and shorter pauses than the respective human speakers. In ch. 4.2, we utilize tempo changes to focus on two aspects of speech pauses, namely their optionality, i.e. presence or absence at a given location, and variability, i.e. variation in duration and involvement of breath noises. Across five languages, we find many optional pauses, taken only in few, slower conditions or by few speakers, while other pauses are less optional. At those less optional locations, speakers take many pauses, that tend to be longer, more variable, and more likely to involve breathing. With tempo changes, speakers adapt the number of breath pauses, but more so the number of non-breath pauses. At faster tempos, pauses, inhalations, and silent edges around them, are found to be shortened. When longer edges occur, they are typically non-symmetric, i.e. either the left or right edge becomes longer typically, but rarely both.

Breath noise acoustics and physiology The studies on breath noise acoustics can be found in ch. 5. The goal here is to further describe the acoustics of speech breath noises, both in general and in relation to the kinematics of the upper body. This part is based on Werner et al. (2021b), Werner et al. (2021a), Werner et al. (2022a), and Werner et al. (under review). In ch. 5.1, we use semi-spontaneous speech data to explore timing aspects of speech breathing, such as the temporal coordination of breath noises and torso expansion, as well as silent edges around breath noises. We find a close coupling between the onsets of inhalation noises and the expansion of both thorax and abdomen. In ch. 5.2, we approach speech inhalations by comparing them to reference speech sounds in a data set that includes audio and kinematic data. We make use of sounds that share potential similarities with breath noises, namely aspiration phases of voiceless plosives and the voiceless glottal fricative, and the neutral vowel /ə/. They are used as breath noises are made up of both noise and formant-like structure. We find inhalation noises to be similar to aspiration phases of /k/, but not /t/ or /h/, and to have formants that suggest a more open and slightly more front vocal tract (henceforth VT) configuration compared to /ə/. Further, we investigate correlations between breath noises and kinematic inhalation speed. We find that faster inhalations are associated with breath noises that have a higher center of gravity, intensity, and first formant. Jaw openness and inhalation speed are thus major contributors to inhalation noises. The comparison of breath noises and speech sounds is limited, as the latter are typically voiceless, thus complicating formant measures, and feature an ingressive air stream as opposed to speech sounds.

In ch. 5.3, we circumnavigate these problems: On the one hand, we do so by analyzing the spectral characteristics of the entire spectrum in a large set of human inhalations produced by 100 male and 34 female speakers of German. We find relatively flat spectra with a decreasing slope and several weak peaks below 3 kHz. They show overlap with experimental results for resonances with an open glottis and a central VT configuration, suggesting further similarity to /ə/. On the other hand, we create synthetic breath noises using 3D-printed VT models, for which we know the exact VT configuration. We simulate in- and exhalations by changing airflow direction in eight VT models of vowels and fricatives. Airflow direction has no general but a segment-specific effect, affecting noises from those VT models with a high tongue position. We link this to the concentrated air stream hitting the incisors in exhalation, but not inhalation. We also compare inhalations produced with the /ə/-model to human inhalations to further approach the VT configuration in inhalations. There are some similarities, however, several complexities of human speech breathing not captured in the models complicate this comparison.

Breath noise perception Ch. 6 describes the experiments on the perception of breath noises. The aim here is to investigate how much information listeners can auditorily extract from breath noises. This part is based on Werner et al. (2022d) and Werner et al. (2022c). In ch. 6.1, we use breath noises of different types, e.g. nasal-oral inhalation or nasal exhalations, to see how well listeners can distinguish between them. We find that listeners detect the correct answer in around two thirds of all cases. Further, one second of speech context on either side does not improve detection rates and phoneticians do not outperform lay listeners. In ch. 6.2, we investigate how well listeners can use breath noises to tell speakers apart and to extract coarse information on speaker characteristics. In the discrimination task, listeners are presented with two breath noises and asked if they are produced by the same or different speakers. In the classification task, one breath noise is played and participants are asked if it was produced by a young or old speaker and if the speaker is male or female. Discrimination is successful in a little less than two thirds of all cases. In the classification task, the speaker's sex is correctly detected in two thirds of all cases, while their age is guessed randomly. The results suggest that speaker sex is more perceivable than age.

Chapter 2

Background

This chapter provides an overview of research on pauses and speech breathing relevant to this work. As such, it will outline the state of knowledge in these two fields. Further, it will point out what limitations existing studies have and what general questions remain open that will be addressed in this thesis.

2.1 Pauses and pause-internal particles

2.1.1 What is a pause?

Attempt at a definition Defining what a pause is is both crucial and difficult.[1] This becomes readily accessible when looking at the different types of pauses, such as *silent*, *filled*, or *breath* pauses. Trouvain (2003) makes two major distinctions: The first one is about perceived, acoustic, and silent pauses. Listeners may miss acoustic pauses and not perceive them as such, and conversely perceive pauses at locations where there was no acoustic silence but other cues, although the majority of pauses should be both, acoustically present and perceivable. Acoustic pauses do not have to be silent and may include breath noises, as breath pauses are part of acoustic pauses. The second distinction he makes is between filled and unfilled pauses. This is typically used to distinguish pauses filled with filler particles, such as *uh* or *uhm*, from all other pauses. This view ignores that unfilled pauses may still be filled phonetically e.g. with breath noises or clicks. As the focus of this thesis is on speech breathing and pauses without filler particles, it will mainly deal with acoustic and perceived pauses. Within those there may be stretches of silence or other material, hence a distinction is drawn between *breath pauses*, in which the speaker typically inhales but may also exhale, and *non-breath pauses*.

[1] To make it a little simpler, we will mainly discuss pauses in the audio signal, as opposed to articulatory pauses or pause postures. Those are touched upon in ch. 2.1.7.

In Trouvain & Werner (2022), it is argued that defining a speech pause as "an interruption of speech" often falls short. This is exemplified using four types of speech interruptions:

(1) articulatory pauses,

(2) silent phases as listeners,

(3) gaps at turn changes in conversations,

(4) pauses in connected speech sections, e.g. while having the turn.

(1) Articulatory pauses typically occur in the closure phases of unvoiced stops. They do surface as silent stretches in the speech signal, but are part of the articulatory movement in the speech stream and thus not a pause (Hieke et al., 1983). (2) Silent phases as listeners are phases in conversations, where at least one interlocutor is not producing speech but listening to one or more other interlocutors. From the listener's perspective, these phases could be seen as some sort of pauses, while waiting to take the turn or producing feedback. However, this is very different from the within-speech pauses in terms of length and function. In conversations, there are often (3) gaps at turn changes, at which neither of the parties involved is speaking. The opposite of that are overlaps, i.e. two or more people speaking at the same time. In general, both gaps and overlaps tend to be avoided in conversation (Sacks et al., 1974; Stivers et al., 2009; Heldner, 2011). The best fitting concept of *pauses* for this thesis is that of (4) pauses in connected speech sections: They are produced by a speaker within their turn (in dialogues) or speech (in monologues) and typically contain silence, but may also include breath noises or clicks.[2] They are used to signal syntactic-prosodic breaks between sentences (in read speech) or utterances (in non-read speech), which is why the term *inter-pausal unit* is often used there, as it is easier to define than a sentence. When appearing in dialogues, these pauses would thus correspond to within-speaker silence in the conversation framework used by Wlodarczak & Heldner (2020), but not between-speaker silence.

Following Eklund (2004), Betz et al. (2023) argued for lengthening, i.e. elongated syllables due to hesitation, to be regarded as pauses within the execution of the articulatory plan and ongoing phonation. In this work, however, we will focus on (4) pauses in connected speech sections as discussed in Trouvain & Werner (2022).

Thresholds and other problems Using a minimum duration as a threshold for pauses is tempting. It works great as a heuristic for grouping data into pauses or non-pauses and makes the automatic treatment of them much easier: a given silent interval either exceeds the threshold or it does not. The threshold should then be set at a point that strikes a good balance between capturing as many true pauses as

[2] Other components could be possible, such as laughter, coughing, or filler particles. These will largely be ignored here.

possible and picking up as few non-pauses (e.g. silent phases of plosives) as possible. However, there is no such thing as a generally agreed-upon value for that and it may also differ widely between fields and studied subjects.

Matzinger et al. (2020, Box 1 on p. 6) included a great discussion of values in use as a minimum threshold: Values they listed ranged from 5 ms (Smiljanić & Bradlow, 2005) up to 400 ms (Derwing et al., 2004; Tavakoli, 2011), with 200 ms being a threshold that is often used for studying pauses in L2 speech (de Jong & Bosker, 2013; Kahng, 2014). Kirsner et al. (2003) included pauses greater than 20 ms and Kentner et al. (2023) those that exceeded 50 ms, while in Parlikar & Black (2012) all pauses shorter than 80 ms were excluded. Goldman-Eisler (1961) classified pauses of less than 250 ms duration as *very short pauses*, which according to her might also be called articulatory pauses, leaving those above the threshold as hesitation pauses. Hieke et al. (1983), Campione & Véronis (2002), and Kirsner et al. (2003) later argued against that: Hieke et al. (1983) reasoned that pauses between 130 and 250 ms cannot be attributed to articulation and are "psychologically functional", i.e. marking boundaries, emphasis, or repairs, and recommend "a minimum pause duration of somewhat over 0.10 sec". Fry (1979, p. 107) stated that a pause should be more than "about 150 ms" in duration, while acknowledging that shorter silences occur in relation to plosive articulation. In the examples he gave for that, the silent interval in relation to /k/ even is 115 ms. Lickley (2015, p. 456) mentioned English word-final stops to vary from 30 ms up to around 250 ms which could be reached in phrase-final stops and further complicates threshold-based approaches. In addition to minimum durations, some studies have also used maximum durations as an upper bound for pauses. Campione & Véronis (2002) lists Grosjean & Deschamps (1972) and Candea (2008) as examples, both of which have used 2000 ms for that.

Similarly, a classification into short and long pauses (or more groups) is sometimes done, e.g. by Goldman-Eisler (1961) and Kirsner et al. (2003), both of which agree on the cut-off point to be at or around 250 ms. Campione & Véronis (2002) reported brief (<200 ms), medium (200–1000 ms), and long pauses (>1000 ms). Both read and spontaneous speech featured brief and medium pauses, whereas long pauses were only reported for spontaneous speech. Bailly & Gouvernayre (2012) employed a distinction between short pauses (<200 ms), which they also call syntactic pauses, on the one hand and sentence-internal, sentence-final, and end-of-paragraph pauses on the other hand. Godde et al. (2021) distinguished between the two pause categories short and long. They defined different thresholds by speaker age (2nd to 7th grade and adults), which varied between 180 and 220 ms.

Using a threshold-based definition of a pause may entail problems, as outlined in Campione & Véronis (2002), who show the effects of pause thresholds by discussing the three possibilities: no threshold, low and high threshold (200 ms and 2000 ms) and low threshold only (200 ms). The average duration they observed for spontaneous and read speech changed substantially as a function of threshold. Using no threshold led to the pause duration on average being longer in read speech, in the low and high threshold condition they were almost equal, and with only a low threshold, pauses were longer in spontaneous speech. The type or absence of a threshold thus may lead to no or even opposite effects.

An alternative to using a minimum duration of silence in the acoustic correlate of a pause is to define a pause as a *perceived pause*, as used e.g. in Belz & Trouvain (2019) or Betz et al. (2023) and suggested in Trouvain & Werner (2022). It requires human annotators but it makes the pause more flexible, as pauses with a silent stretch shorter than 100 ms may still be perceived as pauses, as shown by de Pijper & Sanderman (1994) with untrained listeners.[3] Results may of course differ for trained annotators working with a tool like *Praat* (Boersma & Weenink, 2019). At the same time, it safeguards against false alarms by long stop closures that may exceed 100 ms or pauses that may not be perceivable as such. It can be very useful when dealing with faster speech rates as in ch. 4, as speakers tend to also shorten or omit pauses when asked to speak faster. While they shorten or vanish at higher rates, other cues of prosodic boundaries are more robust to increasing speech rates, such as final lengthening (Żygis et al., 2019). Along with silence and final lengthening, perceived pauses can also be triggered by intonational boundary tones, voice quality such as creaky voice, intensity drops, and syntactic information (Duez, 1993; Strangert, 1993; Swerts, 1998; Carlson et al., 2005; Turk & Shattuck-Hufnagel, 2007). Martin (1970) showed that pre-boundary lengthening can be a strong cue for prosodic breaks, as listeners reported pauses in locations where there was no stretch of silence in the signal. Perceived pauses can incorporate these pieces of information, however, it may make pause annotation more subjective and expensive with regard to time and labor costs.

2.1.2 Pause-internal particles

There are several components that may appear within a pause, which will be referred to as *pause-internal particles* (PINTS). Pauses typically have at least one stretch of silence. Whether that is a particle itself or rather the absence of particles is a debate that cannot be answered here. Instead, we will treat pauses then as pauses in connected speech, that typically have some acoustic-phonetically silent parts, but can also have other elements, such as breath noises, tongue clicks/percussives, or filler particles (e.g. *uh*/*uhm*). We will briefly mention two important PINTS, namely clicks and filler particles, but focus on breath noises.

Clicks may occur in pauses, also in a number of languages where they do not have phonemic status (see e.g. Wright 2007; Gold et al. 2013; Trouvain & Malisz 2016; Pinto & Vigil 2020; Kosmala 2020; Zellers 2022). *Percussives* are a similar phenomenon, which is produced by articulators separating and thus not always deliberate (Ogden, 2013) and may also show up in the course of breathing. Similarly, swallowing gestures

[3] Heldner (2011) found the perceptual threshold for acoustic silences to be at roughly 120 ms. However, their stimuli were all from speaker change situations, where the situation may be different.

followed by audible release may look very similar to silence followed by a click (Ogden, 2021).

Filler particles are another frequent PINT. Their shape can vary within (e.g. *uh* vs. *uhm*) (Laserna et al., 2014; Fruehwald, 2016; Wieling et al., 2016; Belz, 2021) but also between languages, either in vowel quality or in favoring *uh* vs. *uhm* (Candea et al., 2005; de Leeuw, 2007; Lo, 2020) or also by using lexical forms, such as e.g. *ano* or *eeto* in Japanese (Lickley, 2015). Parts of the research dealing with them comprise whether or not they can be regarded as lexical items (Clark, 2002; Kirjavainen et al., 2022), how non-native speakers use them (Belz et al., 2017; de Boer & Heeren, 2020; de Boer et al., 2022), and if and how much their occurrence can help recall and processing of speech (Fox Tree, 2001; Corley et al., 2007; Fraundorf & Watson, 2011; Loy et al., 2017; Bosker et al., 2019; Muhlack et al., 2021; Chen et al., 2022).

Breath noises in speech communication can usually be observed during articulatory activity but not when speakers let their articulation rest, for instance when they are listening (Trouvain et al., 2020). This is likely due to the reorganization of the breath cycle for speaking (see ch. 2.2.1). As speech is produced with an egressive airstream and inhalations are ingressive, the speech stream is stopped for inhalation, leading to a breath pause. The duration of breath (or inhalation) noises is quite variable (Trouvain et al., 2020). There seems to be a link between pause duration and the presence of (observable) breath noises: the longer the pause, the more likely it is to involve a breath noise. Or conversely, pauses that contain a breath noise tend to be longer than those without breathing. This was also observed by Godde et al. (2021), who categorized their pauses into short and long pauses (with thresholds between those varying by age between 180 and 220 ms) and found that breath pauses were almost always long pauses and made up around 70% of all long pauses that were employed in the reading task.

In audio book reading produced by one speaker, Bailly & Gouvernayre (2012) found average durations of breath noises to be 355 ± 145 ms (mean ± sd). In MacIntyre (2022), average breath durations across four speaking conditions were between 402 ms and 435 ms. Fuchs et al. (2013) found inhalation durations of 524 ms for short and 568 ms for long sentences. Fukuda et al. (2018) found more variability in breath duration for spontaneous speech (range: 60 to 1,350 ms) compared to read speech (130 to 670 ms). The duration of inhalations can be influenced by several linguistic and non-linguistic factors (see ch. 2.2.6).

Breath pauses contain breath noises, as well as short silent stretches around them, that are referred to as *edges*. This is visualized in Fig. 2.1, which shows a breath pause that is located within speech. The transition from speech to breath noise, however, is not immediate, as there are short stretches of silence before and after the breath noise, i.e. left (before) and right (after) edge. This is a typical breath pause, that consists of a breath noise that is sandwiched between silent edges on either side. Ruinskiy & Lavner (2007) speculated that edges emerge, as several muscle systems are involved in speech and breathing, hence the change in airflow direction from

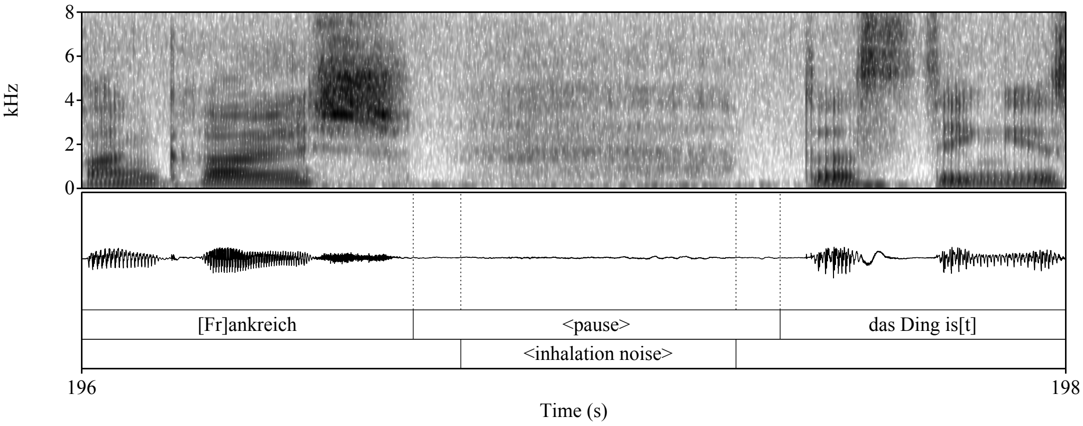

Figure 2.1: Spectrogram of a section from the Pool2010 corpus (Jessen et al., 2005) containing an inhalation noise located in a speech pause surrounded by silent edges.

egressive speech to ingressive inhalation cannot be instantaneous, resulting in these stretches of silence. They found them to be at least 20 ms long. Fukuda et al. (2018) stated that in spontaneous speech, they can become much shorter than 20 ms and almost disappear. In a small sample of four corpora, Trouvain et al. (2020) found average edge length to vary between 23 and 92 ms. There, the average left edge was longer than the right one in all four corpora, and spontaneous speech tended to have longer edges generally. However, the sample was small and average values alone may not be highly informative here. A concept similar to edges was already observed by Grosjean & Collins (1979) who divided a breath pause into *preinspiration*, *inspiration*, and *postinspiration*. It was not purely acoustic, though, as start and end of the inhalation were determined via inhalation slope of a pneumograph signal.

In their study on (breath) pauses in school children from grade 2 (around 8 years old) to grade 7 (around 13 years old) and adults for control, Godde et al. (2021) used inhalation-to-phonation delay, which conceptually should include inhalation noise and right edge. They found this to be significantly affected by grade and highest in grade 2, while the adult group was similar to all other grades. It is hard to tease apart the edge contribution here, though, as it is typically much shorter than the breath noise. They also reported that adults made a longer pre-inspiratory pause, which should be equivalent to left edges. MacIntyre (2022, p. 106–142) was more interested in the timing of rhythm in four different speech styles, i.e. prose reading, poem reading, spontaneous speech, and automatic speech (counting, reciting the ten times table, the days of the week three times, and the months of the year two times). She used a metric that is related to edges, yet slightly different, namely acoustic landmarks which are based on estimated vowel onsets, before and after breathing, and breath onset and offset based on breathing signals (RIP). She found longer landmark-to-breath-onset periods than breath-offset-to-landmark. Again, this is slightly different from edges, as landmarks are based on estimated vowel onsets and hence the duration of the vowel is included in the landmark-to-breath-onset, but not in the breath-offset-to-landmark periods. After having focused on what particles may occur in pauses, the next section will delve into the aspect of how pauses relate to the surrounding speech by discussing how often and where they may show up, how long they may be, and how they relate to speech tempo.

2.1.3 Pause locations and frequency

Pauses are the most prominent cue for prosodic boundary marking (Petrone et al., 2017b; Matzinger et al., 2021).[4] Pauses do not occur at random points in the speech stream but are crucial for structuring utterances, both for the speaker and the listener. By chunking speech into smaller units, they serve speech comprehension (Zellner, 1994). On the speaker side, breathing and other physiological needs impose an upper limit for utterance length, but their appearance is also governed by other factors such as syntactic structure (Włodarczak & Heldner, 2017).

When reading a text, readers may often use punctuation as landmarks for pausing (Godde et al., 2021). While this plays an important role and covers a substantial share of the total pauses in a text, long and breath pauses especially, it does not cover all of them. In general, Krivokapić et al. (2020) stated that several factors make the occurrence of a pause more likely, such as higher complexity of the preceding or following syntactic/prosodic structure (e.g. Ferreira 1993; Grosjean et al. 1979), higher strength of the prosodic boundary (e.g. Ferreira 1993; Gee & Grosjean 1983; Petrone et al. 2017b; Zellner 1994), and longer length of the following or preceding utterance (e.g. Ferreira 1993; Fuchs et al. 2013; Kentner 2007; Zvonik & Cummins 2003). The duration of inter-pausal units was also incorporated in Gee & Grosjean's (1983) *performance structures*. In read speech, inhalations tend to occur at sentence or paragraph boundaries and are also deeper and longer when taken there (Conrad et al., 1983; Winkworth et al., 1994), which has a big impact on pauses and their duration there.

There is also more, finer detail to pause placement, as de Jong (2016) found that in (semi-)spontaneous speech, speakers were more likely to pause before lower-frequency nouns compared to higher-frequency nouns. In a similar vein, Goldman-Eisler (1958) had found that pauses tended to occur after words of high redundancy and before words of high information. Betz et al. (2023) found pause frequency to increase with cognitive load. It may also be influenced by language proficiency, as learners of a language typically produce more pauses (Trouvain et al., 2016b; Belz et al., 2017; Matzinger et al., 2020).

[4] This may be the reason why these two terms are frequently used synonymously. The other markers include final lengthening, intonational boundary tones, intensity drops, and voice quality and are briefly discussed in ch. 2.1.1.

The locations of pauses have a strong impact on the perceived fluency, especially when appearing within clauses, even more so than frequency and duration (Kahng, 2018). Adams & Munro (1973) showed that non-native speakers deviated from native speakers' pause placement schema, thus creating "nonsensical phrases". In a spontaneous speech task, Redford (2013) found that in comparison to adults, children made more pauses that were ungrammatical. Grammaticality was defined via syntactic completeness, so pauses following transitive verbs, determiners, conjunctions, final prepositions, and copulas were considered ungrammatical.

In addition to that, there is some degree of variation to pause placement. How a text is phrased prosodically is not pre-specified to the reader except for punctuation. It can vary from one speaker to another (Goldman-Eisler, 1961; Trouvain & Grice, 1999) but also within the same speaker (see e.g. Trouvain & Grice 1999). This optionality in pause usage leads to different strategies on how syntactically conditioned phrase boundaries are to be indicated (Trouvain & Möbius, 2018; Grice & Baumann, 2009).

2.1.4 Pause duration

The duration of the silent section of a pause is highly variable (Campione & Véronis, 2002; Matzinger et al., 2020). Comparing it across studies is complicated, as some choices can alter results drastically, such as whether or not researchers pick a minimum duration threshold, how high it is (see ch. 2.1.3), the type of speech (spontaneous vs. read) and the involved cognitive complexity, and whether or not non-breath and breath pauses are pooled together.

In addition, using mean values of pause duration may not always be very informative or appropriate: On the linear scale, the distribution of pause duration has been described as skewed or bimodal (or even tri-/multimodal), hence pause duration is often log-transformed to approach a normal distribution (e.g. Campione & Véronis 2002; Kirsner et al. 2003; Bailly & Gouvernayre 2012; Bosker et al. 2013; Godde et al. 2021; Šturm & Volín 2023). Kirsner et al. (2003) advocate for a log-transformation of pause duration data, as for this type of data the untransformed distributions are skewed, variances are large, and negative values not permissible. Using log-transformation of the pause durations, Campione & Véronis (2002) looked at pause duration in read and spontaneous speech in five languages, although spontaneous speech was French only. They reported a trimodal distribution of brief (<200 ms), medium (200–1000 ms), and long pauses (>1000 ms). For read speech, the first peak for brief pauses was centered around 100–150 ms and the second one around medium at 500–600 ms. For spontaneous speech, they were at 78 and 426 ms and showed a third peak for long pauses at 1585 ms, which was absent from spontaneous speech. Kirsner et al. (2003) reported that there were two log-normal functions in their data, whose medians would correspond to 3.95 and 6.3 on the log scale or 52 and 545 ms. In spontaneous speech or dialogues especially, pause duration can vary strongly, as there are more factors influencing them compared to read speech.

Syntax plays a role for pause duration in read speech, as the longest pause durations found by Bailly & Gouvernayre (2012) were between sentences and paragraphs, while sentence-internal pauses, e.g. at commas or ends of phrases, were shorter. Some studies have also found longer pauses to be correlated with more complex syntactic structure, although prosodic structure is very important, too (e.g. Ferreira 1993; Grosjean & Collins 1979; Gee & Grosjean 1983).

In experimental studies, pause duration was found to be related to the length of the upcoming inter-pausal unit, as well as its prosodic structure: Krivokapić (2007) found the prosodic structure to interact with length in determining pause duration. Length of both pre- and post-boundary phrases affected pause duration, as shorter phrases go with shorter pauses and longer phrases with longer pauses. The prosodic complexity, i.e. whether or not an intonation phrase branched into intermediate phrases, also had an effect, with post-boundary, branching phrases leading to shorter pauses. Fuchs et al. (2013) found longer sentences to lead to longer pauses, while syntactic complexity, here defined as the number of clauses, did not affect duration. Zvonik & Cummins (2003) found that the length of intonational phrases preceding and following a pause combine in making pauses longer.

Lickley (2015) stated that Ferreira (1993) and Ferreira (2007) incorporated both preceding and following information and made a distinction for pauses depending on their cause: They showed that pauses are influenced by prosodic, rather than syntactic structure. In these pauses, the preceding prosodic unit influenced pause duration. They then distinguished between these prosodic pauses on the one hand, and hesitation pauses that are taken for planning reasons on the other. Those pauses rather show planning difficulty concerning the next phrase and are thus influenced by complexity of the upcoming phrase. This could correspond to the two types of pauses, i.e. short and long pauses, that Kirsner et al. (2003) reported to be functionally independent, as speech task and presence of amnesia impacted the two types differently. Long, but not short pauses, were thus affected by story generation but not story recall and presence of amnesia, whereas L2 speakers produced both longer long and longer short pauses compared to L1 speakers. They further speculated that long pause duration may be influenced by "intention, attention, planning, topic change, and inspiration".

Trouvain et al. (2016b) observed that German and French speakers, when speaking the respective other language as L2, behaved differently with German speakers shortening and French speakers lengthening their pauses. Matzinger et al. (2020) found that L2 speakers produced more but not longer pauses. Findings may of course differ based on language proficiency.

2.1.5 Relation to speech tempo

Pauses and speech tempo are closely intertwined, or as Trouvain (2003) puts it "tempo modelling is, first and foremost, pause modelling." While it is possible to vary articulation rate without altering pauses, this is likely to result in perceptual mismatches.

In natural speech, tempo variation is frequent with both (strong) global, between-speaker differences in their habitual tempos and more local variations within the same speaker. Tempo may thus vary as an expression of affectivity (utterances produced by a despondent or bored speaker are typically slower than those produced with excitement) or as adaptations to the listener/s (people tend to speak faster to others they know and slower to strangers, hearing impaired, or non-native speakers). Changes of the global tempo in which a text is read lead to changes of the prosodic phrase structure. Local tempo variation, which constantly happens in the production process of reading a text, also implies changes of prosodic phrasing.

How are speech pauses affected by this? Generally, they are less robust to changes in speech tempo than other indicators of prosodic boundaries. With speech rate increasing, pauses may be shortened or omitted, while boundary indicators such as phrase-final lengthening are maintained at higher tempos (Żygis et al., 2019). Grosjean & Collins (1979) found that at faster speech tempos, pause number decreases and non-breath pauses especially vanish. Breath pauses also became shorter, with preinspiration, which is similar to left edges, and inspiration being shortened the most drastically. Postinspiration hardly shortened, as it had been very brief even in slow tempos.

Fletcher (2010) noted that different pauses characteristics, i.e. occurrence, type, and duration, do not vary symmetrically when speeding up or slowing down. According to Grosjean (1979) and Grosjean (1980), increasing or decreasing speech tempo affects the number of pauses more than their duration. Butcher (1981, p. 215–216) agreed with that view for slower speech rates, but finds pauses to be shorter at faster rates. Faster rates producing fewer and shorter pauses is also reported by Lane & Grosjean (1973) and Fletcher (1987). The pauses that do remain at faster rates are typically physiological or respiratory (Butcher, 1981; Grosjean & Collins, 1979; Fletcher, 2010). This leads to an increase in breath group size (Grosjean & Collins, 1979), although some speakers may use a different strategy of shortening breath groups and inhaling more often (Kuhlmann & Iwarsson, 2021). In rates from slow to fast speech, Matzinger et al. (2020) found decreases in pause duration, number of pauses, and the total ratio of time spent pausing compared to speaking.

2.1.6 Pauses in synthetic speech

This chapter should have conveyed so far that pauses are fundamental components of the production and perception of natural speech. In synthetic speech, on the other hand, production is far less restricted by factors such as planning, motor execution, or physiological limitations, making human perception the major limiting factor here. The idea of equipping a synthetic voice with breath noises may seem odd to some as there is no physiological necessity, but it may make the voice sound more natural and could thus find application in areas where naturalness of the voice is ranked high, e.g. in audio book readings (Braunschweiler & Chen, 2013). Similarly, Terzioglu et al. (2020) made a robot perform breathing motions and found it to improve most measures of how the robot was perceived, such as likeability, social presence, animacy, intention to use, attitude, anthropomorphism, and the perceived intelligence, sociability, adaptability and enjoyment.

Where and how long to pause is not fully specified to the reader (human or artificial) in a text, unlike the segmental sound structure of speech. Human readers may often use punctuation as landmarks for pausing, yet also use other sources of information such as prosody (see ch. 2.1.3). Text-to-speech synthesis systems often rely on punctuation as a pausing cue. How they deal with pausing in contexts of no punctuation differs widely between systems (Trouvain & Möbius, 2018). Changing the speech tempo (as described in ch. 2.1.5) may reveal strong differences between natural and synthetic speech: whereas synthetic speech tends to adapt sound durations linearly, natural speakers additionally modify the number of prosodic phrase boundaries and pauses and assimilate, reduce, or elide sounds and syllables (Trouvain, 2002).

The style of the synthetic speech also plays a role for modeling pauses there. Parlikar & Black (2012) analyzed pause durations within utterances in different corpora representing six speech styles, such as parliament proceedings, radio broadcasts, short sentences, political speeches, and audio books. They found that different styles feature different durations (average and distribution) which are better captured by style-specific models than e.g. pause models with a fixed duration.

Natural and synthetic speech also differ in the use of breath noises in pauses: in natural speech, they benefit the speaker and may also benefit the listener, whereas in synthetic speech, there is no physiological necessity, so a positive effect may (Whalen et al., 1995; Elmers et al., 2021b) or may not (Trouvain & Möbius, 2013; Braunschweiler & Chen, 2013; Bernardet et al., 2019) be observed solely on perceptual aspects.

2.1.7 Pause articulation

So far, pauses have been viewed only from an acoustic perspective, thus neglecting their articulatory dimension.[5] Krivokapić et al. (2020) stated that the idea of the vocal tract assuming a default position in pauses, that may be called *articulatory setting* had been postulated for a while (Honikman, 1964; Laver, 1978; Jenner, 2001)

[5] Since breath and non-breath pauses are not always distinguished, and a reliable distinction may also be difficult, e.g. in MRI, there is some overlap between the literature discussed here and in ch. 2.2.2.

and that they were first observed in kinematic data by Öhman (1967) and Laver (1978).

More recently, Gick et al. (2004) wanted to see if there were language-specific settings in x-ray films, i.e. if the inter-utterance rest position differed between English and French. They did find differences in five of seven articulatory settings, i.e. pharynx width, tongue body, tongue tip, as well as upper and lower lip, but not for velopharyngeal port width and jaw aperture. Their results suggest that there is a pause posture that articulators return to between utterances. This posture is language-specific and the accuracy with which it is achieved is comparable to that of speech targets. In both cases, the inter-utterance rest position may be related to that language's specific realization of /ə/. Differences between articulatory settings were also found for speakers of L1 and L2 American English with mixed L1 (Ramanarayanan et al., 2011). In addition, they found inter-speech pauses to be significantly different from absolute rest positions, with jaw height lower in the latter.

Ramanarayanan et al. (2009) found differences in speed of articulation decreases and increases depending on whether the pause was planned at a phrase juncture or not, with grammatical pauses showing strong decrease in comparison to speech before the pause. Articulator speed during and after the pause was found to be more varied in non-grammatical pauses, which suggests them to be less targeted. They interpret this as indicating that grammatical pauses may be choreographed centrally, while the ungrammatical ones are rather unplanned.

Ramanarayanan et al. (2010) and Ramanarayanan et al. (2013) looked at articulatory settings in spontaneous American English produced by five female speakers and the effects of speaking style (read vs. spontaneous) and position in utterance via rt-MRI. They differentiated three types of pauses, namely absolute rest, speech-ready, and inter-speech (which may include silent or filled pauses). They found significant differences for absolute rest compared to speech-ready and inter-speech pauses and a trend towards different degrees of variance in inter-speech pauses compared to the other two, which might indicate more control in execution. Articulatory settings also differed between read and spontaneous speaking styles, with higher jaw and lower tongue positions in spontaneous speech. Different types of pauses may thus involve different degrees of active control by the cognitive speech planning mechanism, with most planning involved in inter-speech pauses in read speech and least planning in absolute rest. They also noted that articulatory settings might be an agglomeration of several processes involved in speech production, such as respiration, cognitive load, physical effort, etc.

Katsika et al. (2014) used EMA and found that across eight speakers of Greek, acoustic pauses went together with a similar vocal tract configuration, visible in the trajectories of lip aperture and the tongue dorsum. They describe this *pause posture* as deviating from both the preceding and following constriction target, from which they deduce that it is not just a preparation for an upcoming event but related to the pause itself, as it involves an additional movement. Krivokapić et al. (2020) built on Katsika et al.'s (2014) findings and investigated pause postures, i.e. articulator movement occurring between the speech-related gestures before and after the pause that is not just an interpolation of the preceding and following gestures. They found pause postures in the EMA recordings of all eight speakers of American English participants here, with some differences in frequency between speakers and target words. Longer boundaries lead to more pause postures. They observed that spatial variability in pause postures was comparable to that of controlled targets, such as the pre-boundary target word and that shorter pauses showed more variability, which they interpreted as undershoot. From this, they deduce that pause postures have a controlled articulatory target. Finally, they noted that longer upcoming phrases lead to more pause postures, suggesting that they are related to planning. Krivokapić et al. (2022) analyzed pause postures in relation to speech planning. As in their previous work, they found longer upcoming utterances leading to higher rates of occurrence of pause postures but not to longer pause posture durations. They found that planning took place throughout the entire pause, rather than just in the later parts.

In an EMA study, Rasskazova et al. (2018) studied articulatory settings during inter-speech pauses and how they relate to the acoustic signal in eight speakers of German. They found acoustic pauses to be 251 ms longer on average than articulatory pauses. In their findings, they made out two articulatory pausing strategies: In the rest strategy, the tongue moved to the palate to rest there or articulators did not move after the pre-pausal segment. In the transition strategy, there was smooth transitory movement from the pre- to the post-pausal gesture without any phases of articulatory immobility. Rest is mostly found in longer acoustic pauses, whereas transition was more prevalent in shorter acoustic pauses. They did not find this to be affected by speech rate but individual speakers showed a clear preference for either strategy. Finally, they note that breathing patterns in the inter-speech intervals may be crucial for the articulatory preparation strategy (see ch. 2.2.2 for more on this).

2.2 Speech breathing

Breathing, i.e. repetitive cycles of inhalation and exhalation, is a vital activity of humans (and all aerobic organisms) necessary for gas exchange. In the context of speech production, breathing assumes its secondary function and becomes indispensable, as exhalations are used for articulating speech. The switch from rest to speech breathing has implications on different levels, including several parts of the body, as well as temporal and acoustic aspects.

2.2.1 Breath cycle reorganization

When breathing at rest, in- and exhalations are similar in length, with inhalations being only slightly shorter than exhalations (Conrad & Schönle, 1979), resulting in a sinusoidal breath wave. For one breathing cycle at rest, Fuchs & Rochet-Capellan

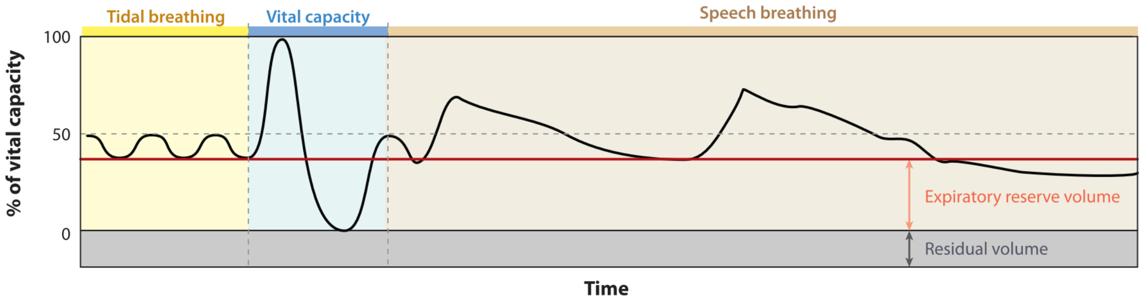

Figure 2.2: Respiratory kinematics for different breathing tasks. Tidal or rest breathing, vital capacity, i.e. maximal in- and exhalation, and speech breathing. The y-axis shows lung volume and the x-axis time. Figure taken from Fuchs & Rochet-Capellan (2021).

(2021) reported values of around 40–60% of its duration is spent inhaling and the remaining part exhaling. Switching from tidal breathing to speech has a reorganizing effect on the breathing cycle: When speaking, inhalations have a shorter duration and higher airflow velocity, while the exhalations, which are used for producing speech, become longer with a constant, slow decrease in lung volume (Conrad & Schönle, 1979). In the context of speech, inhalations only take up around 10% of the duration of one breath cycle, while the remaining 90% are used for speech production (Fuchs & Rochet-Capellan, 2021). The demand there is for inhalations to take in enough air for gas exchange and to power speech breathing in a relatively short time so as not to interrupt the speech stream for too long and for exhalations to produce speech for a sufficiently long time. In addition, the inhalation frequency is higher in speech breathing compared to tidal breathing, with 12.2 (range 7.2–19.3) vs. 20.0 (range 14.0–30.6) breaths per minute reported by Hoit & Lohmeier (2000). As a consequence, the inhalations become deeper in speech breathing to fulfill the gas exchange requirements in shorter time, which has consequences for the shape of the RIP signal, which turns into sawtooth-shaped (Fuchs & Rochet-Capellan, 2021). This reorganizing effect of speech on breathing is visualized in Fig. 2.2.

Respiratory patterns at rest and during speech are thus fairly different, but where the switch happens can be highly informative: Grosjean (1979) tested if this reorganization was only triggered by verbal speech or if it also occurred with non-verbal signing: He compared pausing and breathing in a spoken (American English) and a manual language (American Sign Language). For breathing, he found that signers retained the respiratory pattern from quiet breathing, while speakers reorganized their breath cycle as described above. As the production rate, i.e. words or signs per minute, increased, signers only inhaled a little more, probably due to the physical activity involved. For the speaker, however, he found a strong interrelation of speech and breathing, with much more frequent inhalation in higher speech rates. It only decreases again for the highest rate here, as the speaker changes strategy and underventilates for the course of the short text presented in this study. The percentage of respiratory cycle that was spent inhaling was constant between 30 and 40% for signers throughout different rates. For the speaker, only around 15% of the breath cycle was spent inhaling. At the highest rate, this number went down to 6%. In addition, signing and respiration were not coupled, with only 19% of inhalations happening in pauses, as opposed to 100% in speaking.

Further tackling the opposition of rest and tidal breathing, Conrad & Schönle (1979) looked at different tasks produced as *inner* speech, i.e. thinking about it without producing it, *subvocal* speech, participants moving their articulators as if they were speaking but without producing phonation, and actual speech tasks. They found that subvocal tasks showed respiratory patterns that were similar to speech, while in the inner tasks, respiratory behavior was more similar to tidal breathing, but already slightly different from that. In subvocal tasks, they found exhalation duration to be more adapted, i.e. lengthened, towards speech breathing than inhalation duration, i.e. not shortened as much, and also that behavior differed between different tasks that participants performed subvocally. Tasks with a higher difficulty seemed to be more transformed towards speech breathing. They thus suggested that there is a continuum from tidal to speech breathing rather than a categorical switch.

Włodarczak et al. (2015) and Włodarczak & Heldner (2017) looked at how communicative needs and respiratory constraints interacted to test how the breath cycle changed with the amount of speech produced: They made a distinction between *very short utterances* and other utterances. They defined the very short ones as utterances that were shorter than 1 second in duration to include feedback utterances and backchannels. They then looked at how their onsets aligned with the durations of the exhalations they were in: if it was right at the beginning, then participants may have inhaled just for that utterance, if it appeared later, then there was no new inhalation for that utterance. They found that long utterances were most often produced with a new respiratory cycle and became increasingly rare at later points of exhalation. For the short utterances, this tendency was much weaker: They tended to be started around the beginning of the exhalation, too, but often they were produced at a later point in the breath cycle up to around 75% of the exhalation's normalized duration. In Włodarczak & Heldner (2017), they additionally looked at other parameters of the respiratory cycle: they found that across the three cycle types, mean respiratory amplitude was substantially larger in inhalations for speech, whereas inhalations during quiet breathing and those for very short utterances were smaller and more similar to each other. Inhalation duration for very short utterances was substantially shorter than for quiet breathing but also much longer than for speech. Relating the inhalation duration and the relative timing of very-short-utterance onset in the exhalation suggests that the ones early in the exhalation may have been planned which is less likely for the later ones.

All in all, while the difference between tidal and speech breathing is continuous rather than categorical and there are several nuances to it, speech breathing does differ substantially from breathing at rest: Generally in speech breathing, the inhalation-to-exhalation duration ratio is smaller, inhalations are deeper, i.e. more air is inhaled there, and exhalations become longer to enable speech production. Additionally, expiratory pressures are more sustained and respiration becomes less frequent (for an overview of papers, see e.g. Binazzi et al. 2006, p. 233). Inhalations occur every three to four seconds in most speech situations (Winkworth et al., 1994; Rochet-Capellan & Fuchs, 2013b; Kuhlmann & Iwarsson, 2021), thus temporally structuring the flow of speech (Fuchs & Rochet-Capellan, 2021). The reorganization of the breathing cycle leads to these inhalations being rapid and deep so they typically become audible. This audible breath noise is generated along the involved parts of the body.

2.2.2 Breathing physiology

Lungs Human breathing is executed by changes in volume and pressure, making use of Boyle's law, which states that volume and pressure are inversely related. It follows from this law, that when the volume of an enclosed space is increased, the air pressure within it is decreased (Raphael et al., 2011, p. 57). Following this principle, respiration is performed via compression and expansion of the lungs, creating pressure gradients between the air inside the lungs and atmospheric pressure: For inhalation, chest and lungs are expanded, which causes air to flow in to equalize the negative pressure. For exhalation, thorax and lungs are contracted, which causes air to flow out to equalize the positive pressure that results from contraction (Raphael et al., 2011, p. 57).

The lungs themselves do not include muscles, hence compression and expansion cannot be actively controlled by them. Instead, other muscles move the torso, affecting the lungs, which is made possible by the pleura or pleural gap: a small, liquid-filled gap that connects the pleural visceralis, a membrane surrounding the lungs, with the pleural parietalis on the side of the rib cage. It keeps the membranes close to each other, while allowing them to slide along each other (Pouw & Fuchs, 2022). This connection is also referred to as *pleural linkage*. As a consequence, lungs and chest wall or torso form a functional unit, which has implications for both of them: opposing forces keep both of them from returning to their resting position, which for the pulmonary apparatus would be to collapse into a very small state and for the chest wall to expand. The linkage keeps the entire breathing apparatus in a resting, balanced state in which the lungs are a little expanded and the chest wall a little compressed and both forces neutralize each other (Hixon et al., 2020, p. 12–13).

A number of passive and active forces are important to breathing (Hixon et al., 2020, p. 13–14): Passive force is exerted by the natural recoil of muscles, cartilages, ligaments, and lung tissue, surface tension of the alveoli, and the pull of gravity. The connection between lungs and torso described earlier leads to direction and magnitude of the recoil: When there is more air in the lungs than at rest, the breathing apparatus recoils towards that state by compressing and thus exhaling. Conversely, when there is less air in the lungs than at rest, it recoils towards the resting state by expanding and thus inhaling. The further away from the resting state, the stronger the magnitude of the recoil.

Active force involved in breathing is exerted by a whole array of muscles of the chest wall (see Hixon et al. 2020, p. 14–22 for an overview). The largest muscle involved is the diaphragm, situated between and separating rib cage and abdomen. When it contracts, it moves downwards and thus expands the lungs, which leads to inhalation. External intercostalis muscles, situated between the ribs, simultaneously raise the rib cage, thus helping the lungs expand. Exhalation at rest is largely done by elastic recoil. In speech breathing, it is mainly performed by internal intercostalis, pulling the rib cage down, and abdominal muscles, compressing the abdominal cavity and pushing the diaphragm upward (Fuchs & Rochet-Capellan, 2021).

Lower airways Below the larynx, the lower airways include trachea, the two main bronchi, i.e. the first branching airways below the trachea, as well as the rest of the bronchial tree, branching into two smaller airways, which then further branch, down to the level of the alveoli. With the trachea considered the 0th generation of the tree, the main bronchi are 1st generation, and the bronchial tree branches further down to about 35 generations. The reduction in diameter with every generation leads to a combination of very small individual airway diameter of around 4 mm and a large total cross-sectional area after 6 or 7 generations. Therefore, acoustic fluctuations become very small there, such that generations below that can be neglected for the acoustics of the lower airway, with the exception of frequencies <50 Hz (Lulich, 2010).

Larynx The larynx, including the vocal folds, plays a big role in both in- and exhalations (Fink, 1974; Orlikoff et al., 1997): For inhalation, the larynx moves downward, also lowering the trachea, stretching out the laryngeal soft tissues near the hyoid bone. By this, the laryngeal soft tissues (vestibular and aryepiglottic folds and the preepiglottic body) are elongated vertically and thinned transversely. This results in the vocal folds being flattened against the side walls. For exhalation, these displacements are then reversed. This has an effect on glottal opening in breathing: in inhalations, the glottis is opened very wide, with adult males reaching a peak glottal area of 217 ± 54 mm^2 during slow and 228 ± 43 mm^2 during rapid inhalations. For females, the peak glottal area is 189 ± 32 mm^2 for slow breathing and 184 ± 25 mm^2 for fast breathing (Scheinherr et al., 2015).

Supraglottal airways Above the larynx, rest breathing is typically performed with an open velopharyngeal port and a closed mouth, i.e. nasally. Deviating from that norm may lead to several problems for the oral environment, craniofacial morphology (Harari et al., 2010; Inada et al., 2021), and lung function (Hallani et al., 2008).

The reorganization of the breathing cycle for speech also has an influence on breathing in the supraglottal airways: This may be the reason that around speech, speakers usually deviate from the pattern of purely nasal inhalations that is prevalent at rest.

Analyzing breath noises, i.e. the acoustic correlates of breath events around speech, Kienast & Glitza (2003) found oral inhalations to be the most frequent type to be used there. Nasal inhalations and sequential combinations, i.e. starting with nasal and switching to oral in the course of one inhalation or vice versa, were far less common and idiosyncratic to some speakers. This study, however, was based fully on the interpretation and perception of breath noises and hence could not test these findings via video or articulatory data.

Lester & Hoit (2014) examined the typical airway usage by healthy adults over several speech tasks (counting, paragraph reading, spontaneous speaking, conversation, sentence reading) and phonetic contexts. The contexts were combinations of the low vowel /ɑː/ (i.e. lips open, velopharyngeal port closed), bilabial stops (i.e. lips closed, velopharyngeal port closed), and nasals (i.e. lips closed, velopharyngeal port open), preceding and following the inhalations. Using nasal ram pressure, as well as audio and video recordings they categorized inhalations as nasal-only, oral-only, alternating between the two, or using both airways simultaneously. They found simultaneous oral and nasal inhalations to be the vastly predominant pattern across tasks and regardless of phonetic contexts. The reason they see for this is the benefits that come with this type of airway usage: minimizing upper airway resistance by making use of two airway channels instead of one while preserving some of the advantages that come with nasal breathing, such as filtering, humidifying, and warming the incoming air.

Potential support for velum lowering, along with mouth opening, in speech breathing may come from Gick et al. (2004), who wanted to see via x-ray films if the inter-utterance rest position differed between English and French. As the speech pauses they looked at were not differentiated by presence of inhalation, a fair amount of them may contain breathing.[6] They found differences for five of the seven articulatory parameters they investigated. However, they did not find any difference between the two languages for jaw aperture and velopharyngeal port width even though, unlike English, the French phoneme inventory features nasal vowels. They ascribed these similarities to physiological effects of inhalation that may be present in their pauses.

Similarly, Ramanarayanan et al. (2013) suggests that the articulatory settings they found in pauses may be an agglomeration of respiration, cognitive load, physical effort and several other factors involved in speech production.

Rasskazova et al. (2019) investigated the timing of acoustic, respiratory and articulatory events before speech initiation using electromagnetic articulography and RIP. They found speaker variability in the coordination of mouth opening and inhalation onset: two speakers started inhaling before opening their mouths, one speaker first opened their mouth and then started inhaling, and the other three started both at roughly the same time, hence coupling inhalation onset and mouth opening. Similarly, Scobbie et al. (2011) found pre-speech noises, i.e. largely clicks and inhalation,

[6] Neglecting breathing in speech pauses is not exclusive to articulatory studies, as many studies on pauses also have subsumed them under the term *silent* pauses, in opposition to *filled* pauses.

to occur around 250–500 ms before the onset of acoustic lexical content.

The problem with the majority of the articulatory studies mentioned here, thus excluding the study using airflow by Lester & Hoit (2014), is that often they do not differentiate between acoustic-phonetically silent non-breath pauses and breath pauses. Therefore, it can only be speculated, as is done by Gick et al. (2004), whether and to what degree the results have been affected by – and may thus be informative for – speech breathing. An exception to this is Breitbarth (2021), who examined tongue movement in pauses of 11 speakers of German via ultra sound. As around half of her pauses included oral inhalation, she analyzed those separately. However, she could not find systematic movement for those, as participants showed a lot of variance in which parts of the tongue they moved and where. In general, however, the tongue tended to be in a neutral, relaxed state for most participants.

2.2.3 Measuring speech breathing

With multiple parts of the body involved, breathing can be captured at several points and in several ways by different types of signals. A number of different instruments have been used for research on rest and speech breathing. For the latter, however, requirements are different and thus different devices may be more appropriate, as minimal interference with speech production is preferable (Raphael et al., 2011, p. 289).

Raphael et al. (2011, p. 289–291) mentions several devices: *Pressure-sensing tubes* can be used to capitalize on air pressure variations throughout the vocal tract while breathing and speaking. They can be applied to the area of interest, which may be intraoral, nasal, or subglottal. The latter is rather complicated and requires either tracheal puncture or a small balloon to be inserted into the esophagus for an indirect measurement. Tubes for intraoral and nasal pressure are comparatively easy to insert, either through the nose or the corner of the mouth. Measuring airflow is another useful approach for research on breathing which has the advantage that it also provides information on air volume, by multiplying flow by time. Common devices for this are the *pneumotachograph* or *Rothenberg mask*, which is a face mask that is divided by two chambers that allow measurements of oral and nasal airflow. While its interference with articulation is supposedly minimal, Lester & Hoit (2014) preferred a nasal cannula over a face mask, as those may compress the nasal pathways and thus increase nasal pathway resistance, thus influencing breathing. Fuchs & Rochet-Capellan (2021) agree that face masks come with drawbacks, such as affecting the acoustic signal, needing some adaptation, and limiting the degree of jaw motion.

Raphael et al. (2011, p. 291) go on to describe the *nasometer*, which is analogous to the pneumotachograph in distinguishing between nasal and oral information, but uses microphones to pick up the signal. The main principle is that there is one microphone for the nasal cavity and another one for the oral cavity, which are separated by a baffle. Acoustic intensity of the nasal signal can then be divided by the acoustic intensity of both nasal and oral signals to calculate nasalance, which is correlated with nasality. It has been used for studying participation of the nasal cavity in fricatives, e.g. /s/, (Rollins & Oren, 2020) and vowels (Carignan, 2018).

Raphael et al. (2011, p. 291–292) then mention several devices to track changes in lung volume, such as the *full body plethysmograph*, for which the participant stands in a completely sealed booth. Volume changes of the torso can thus be tracked as volume and pressure changes in the plethysmograph. Additionally, it allows for airflow to be measured via tubes with mouthpieces. The *pneumograph* also performs plethysmography, i.e. it tracks volume changes. It consists of wire coils that are put around a participant's rib cage and abdomen. An electric current travels through the coils and expansion or contraction of the torso leads to changes in inductance. A similar approach is the *magnetometer*, which also tracks rib cage and abdomen movements, by using two coils of wire. Here, the electric current passes through one of them to generate an electromagnetic field, thus inducing an electric current in the other coil.

The device that is most commonly used in recent work and has thus become the standard in speech breathing research is *(respiratory) inductance plethysmography* (RIP), which works like the pneumograph: It uses two flexible belts to track changes in the circumference of thorax and abdomen as a proxy for in- and exhalations. The varying inductance in the belts is converted into an analog signal with an amplitude proportional to changes in lung volume (Heldner et al., 2019). It is non-invasive, easy to use and easily combinable with other measures, while not interfering with breathing or speech production. The only drawback is that it is sensitive to gestures and other arm and shoulder movements (Fuchs & Rochet-Capellan, 2021). To keep the influence of this small, participants are often instructed to keep their arms on their sides, as e.g. in Rochet-Capellan & Fuchs (2013b), put them on a table (Wlodarczak & Heldner, 2020), or asked to hold something (MacIntyre, 2022, p. 88). Some amount and types of noise are bearable, though, as Serré et al. (2021) asked people to bike with their arms and legs and hence that signal showed up in the RIP signal, but was acceptable due to its low-amplitude and cyclical influence.

Binazzi et al. (2006) adopted a similar technique called *optoelectronic plethysmography* from Cala et al. (1996), which like RIP tracks the torso as a proxy for lung expansion and compression and can be used for measurements of lung volume, respiratory timing, and the kinematics of the respiratory system. The major difference is that here, markers are positioned on the chest wall and three-dimensional motion of them is monitored optically via infrared cameras.

From a phonetician's perspective, recording audio and investigating the breath's acoustic shape is arguably the most obvious. It has already been used in Henderson et al. (1965) as "perhaps the simplest and easiest method of locating breath intakes during speech". Włodarczak & Heldner (2016) also suggest it, as the advantages are that they are recorded along with regular speech without any extra work and there is no additional device needed for acquisition, hence no further calibration or synchronization is needed. In addition, there is no specific elicitation method needed, as is necessary for some rare features of speech, as virtually any speech task will also lead to speech breathing. With most microphones, speech breathing should be recorded without any adaptations to the setup, given that the distance to the microphone is not too big, as breath noises are less intense than speech. While this provides the acoustic results of breathing, there are many aspects that are missing or not clear in the acoustic signal. This approach has been used e.g. in Goldman-Eisler (1955) or Wang et al. (2010).

Another rather simple approach, which can be applied to several types of measurements is analyzing temporal aspects. This could be done by using the acoustic signal to look at durations of inhalations and the speech between them, also known as breath groups, as done in Wang et al. (2010). Or one could investigate timing aspects of in- and exhalations in the pneumograph signal, as done in Conrad & Schönle (1979).

It is also possible to combine several measurements into a multi-parametric analysis, as done by Yu & Demolin (2022). They simultaneously recorded the acoustic signal, as well was electroglottography, vital capacity via a spirometer, aerodynamic measures (intra-oral pressure and oral airflow), muscle activity via electromyography, kinematics of torso and jaw displacement, and brain activity via electroencephalography. The dangers with this approach are in losing temporal synchronization between measurements and creating artificial settings by hooking up participants to a whole array of equipment. Another multi-parametric setup was used by Petrone et al. (2017a), who combined acoustic, intraoral pressure, and RIP data.

2.2.4 Breath noise acoustics

Breath noises are what surfaces acoustically from the process of speech breathing: They are made audible by the breath cycle reorganization (see ch. 2.2.1) and modulated by physiological aspects of speech breathing (see ch. 2.2.2). Whalen & Sheffert (1997) stated that "[a]lthough the sound produced by inspiration is nonlinguistic, it is nevertheless shaped by the vocal tract. Thus such breath sounds have the potential to provide information about the dimensions and configuration of that vocal tract." That is because articulation and acoustics inform each other, although the acoustics are not indicative of one specific articulation (Atal et al., 1978).

Only few studies have analyzed the acoustics of breath noises in the context of speech. Nakano et al. (2008) investigated them in singing with the aim to improve breath detection algorithms. With regard to their acoustic characteristics, they found breath noises to have similar spectral envelopes within the same song and also within the same singer. They also note that the long-term average spectra of breath noises found in male singers have a peak at 1.6 kHz and those of female singers at 1.7 kHz. Along with these spectral peaks in the F2 region, they found secondary peaks that exist in the range of 850–1,000 Hz, which were, however, more prominent in female singers. Fukuda et al. (2018) also mentioned the formant-like structure of breath noises in their examples. Panne (2020) looked at formants in oral inhalations in 19 speakers of German. She found that F1 tended to not always be visible, unlike F2 which generally had a higher amplitude. On average, formant values for ten female speakers were 814 ± 179 Hz for F1, 1,958 ± 109 Hz for F2, 3,086 ± 311 Hz for F3, and 3,980 ± 137 Hz for F4. The values for nine male speakers were 629 ± 89 Hz for F1, 1,651 Hz ± 92 Hz for F2, 3,035 ± 547 Hz for F3, and 3,836 ± 548 Hz for F4. F1 and F2 were significantly higher in female speakers. Kienast & Glitza (2003) reported energy peaks in oral inhalations produced by ten speakers of German. Due to their experimental setup they only included frequencies from around 400 to 2,000 Hz. The first energy peak was typically somewhere around 500 Hz (one was at 1,500 Hz) and the second one between around 1,400 Hz and 2,000 Hz. In addition, half of the speakers only showed one energy peak, in which case it was the lower one in the region of 500 Hz that was missing. Link (2012) performed a similar study looking, among other things, at spectral characteristics of oral inhalations produced by eight female speakers of German. She found spectral peaks to be around 500 Hz, 1,000 Hz, 2,000 Hz, 3,200 Hz, and 4,100 Hz. Only three out of eight speakers showed all five of these maxima. The most robust ones, i.e. those found in all speakers, were those at around 500 and 2,000 Hz. Some other attempts to characterize the acoustics of breath noises include the analysis of their formants in snoring: Ng et al. (2008) found F1 to be much lower for benign vs. apneic snorers (360 vs. 724 Hz on average), while F2 (1,627 vs. 1,809 Hz) and F3 (2,840 vs. 2,955 Hz) did not differ as much. Breath noises have not exclusively been approached by examining their formant-like structure, but also their noisy characteristics. Ruinskiy & Lavner (2007) observed that the zero-crossing rate is lower in breath noises compared to the fricative /s/. Székely et al. (2019) referred to Ruinskiy & Lavner (2007) and Fukuda et al. (2018) in saying that zero-crossing rate could be used to differentiate breath sounds (mid zero-crossing rate) from unvoiced fricatives (high zero-crossing rate).

When analyzing breath noises, there may also be an influence of subglottal resonances that is stronger than in speech sounds. The degree of coupling between the supra- and subglottal tracts depends on the glottal state (Lulich, 2010): with an infinite glottal impedance, i.e. with a closed glottis, there is no coupling between the two tracts and the resulting poles are the natural frequencies of the supraglottal tract. With a partially open glottis, there is some coupling, which leads to the VT poles being shifted upward in frequency and new poles being introduced from the subglottal tract. This increase in frequency was also demonstrated experimentally for phonation in VT models with an increasing peak glottal area (Birkholz et al., 2019). In breathing, the glottis is opened much wider than in regular speech allowing for strong coupling between sub- and supraglottal airways.

Hanna et al. (2018) examined how the impedance spectrum measured through the lips is affected by the subglottal tract in varying states of glottal opening, i.e. fully closed, fully open, as well as intermediate states for phonation and respiration. In

male	530±60	880±55	1,335±145	1,735±70	2,210±75	2,565±95	3,660±270
female	660±30	1,020±45	1,490±45	1,820±65	2,460±55	2,790±160	3,680±455

Table 2.1: Impedance minima corresponding to resonances (in Hz; mean and standard deviation) reported for 7 male and 3 female participants inhaling with a VT configuration of /ɜː/ in Table I and II in Hanna et al. (2018).

their experimental part, they used impedance spectrometry to measure resonances and anti-resonances in ten Australian participants that were instructed to keep their tongue in the position of the /ɜː/ vowel. In non-rhotic Australian English, this vowel should be very close to the neutral VT configuration of /ə/, with the former being slightly more open. Participants were also instructed to keep their velum raised. As the velum may be hard to raise volitionally, some participants are reported to have pinched their noses to avoid nasal participation in Hanna et al. (2016). For inspiration in the seven male and three female participants, Hanna et al. (2018) found impedance minima, which correspond to resonances and can be seen in Table 2.1. They showed that combining the supra- and subglottal tract, as is done in respiration, effectively doubled the length of the tract, which leads to up to twice as many resonances and anti-resonances compared to the closed glottis condition.

In some cases, opening the mouth can lead to weak clicks, sometimes referred to as percussives (Ogden, 2013). The cause for those may be the sudden and strong vertical downwards movement of the larynx in inhalations in combination with an increased glottal opening, which can lead to sufficient negative pressure to generate tongue clicks (Fuchs et al., 2013; Trouvain, 2014). These can occur at different points in time along the breath noise and are probably related to mouth opening. Hence it can be assumed that when they appear at the beginning of a breath noise, mouth opening is present at the beginning of or throughout the entire inhalation, and when they occur later that the speaker inhales nasally first and then opens their mouth at a later point in time.

It should also be mentioned here, that noises related to inhalation cannot only occur in speech pauses for inhalation. A related phenomenon is ingressive pulmonic phonation, which uses the same airflow direction, unlike the majority of speech sounds which are produced with a pulmonic egressive air stream. Ingressive pulmonic phonation does occur in several contexts, but is still far less prevalent (Eklund, 2004), which may be related to the fact that vocal fold anatomy is better suited for phonation with a pulmonic egressive rather than ingressive air stream (Catford, 1977, p. 67–68). However, it plays an important role in non-verbal vocalizations, such as laughs, cries, and moans (Anikin & Reby, 2022). Virtually all inhalations encountered in speech pauses, however, are unvoiced.

2.2.5 Breath noise perception

Breathing can influence perception or cognition in two ways, i.e. by influencing the person breathing or the person perceiving the breath. The respiratory cycle contributes to structuring brain activity rhythmically, influencing perception, emotion, and cognition (Heck et al., 2017; Allen et al., 2022). For a person breathing, inhaling impacts cognition, as participants tend to inhale at the onset of cognitive tasks (Johannknecht & Kayser, 2022) and perform better when doing so as opposed to exhaling (Perl et al., 2019). This phase-locking effect between inhaling and task onset was also found by Zöllner et al. (2021) for reaction time measures in naming tasks.

More relevant to this work is the question how perceiving someone else's breath noise can affect the people perceiving them. In the context of speech breathing, aspects of breath noise perception can deal either with breath noises affecting the processing of speech or using information that is conveyed directly by the breath noise, such as information on the speaker or the airway used. These two are not separate, as the perceiver is likely to use both when listening to speech including breathing.

Since speech breathing can be shaped by several factors (see ch. 2.2.6), some of them might be perceivable to the human ear as cues, including speech-related information. In this regard, MacIntyre & Scott (2022) studied participants' ability to detect a silent gap next to breath noises over three experiments. Overall, they found an effect of breath position (preceding vs. within speech), such that listeners were more susceptible to the gap within speech, which might indicate sensitivity to speech rhythm. They also found that whether a gap was added before vs. after breath noises was important, too, as gaps were more likely to be detected after rather than before within-speech breaths. The length of the breath noise also played a role, with more listeners assuming a gap in the vicinity of long breath noises.

Breath noises may also contain information about other aspects beyond speech. They can cue formality, as in Korean, noisy inhalation with the quality of a retracted, sometimes bidental fricative and often a lateral place of articulation, is associated with more formal settings (Winter & Grawunder, 2012). Modifying breath noises may also convey emotions, as phonated breath noises that occur in laughs, cries, or moans enhance the perceived level of emotional intensity in comparison to typical, unphonated inhalations (Anikin & Reby, 2022). Although in the majority of them, phonation is absent, breath noises are also modulated by the filtering characteristics of the vocal tract and as such may contain information about the speaker. Whalen & Sheffert (1996) and Whalen & Sheffert (1997) found that the sex of the breather was perceivable in breath noises, while coarticulatory information about the upcoming vowel was more difficult to perceive.

We can assume several pieces of information to be present in breath noises, some of which are much more fine-grained than a distinction between male and female, because automatic approaches were able to utilize them: Neural networks have been shown to be able to identify speakers based on breath noises (Zhao et al., 2017; Lu et al., 2020; Ilerialkan et al., 2020). Breath noises may even be used for user authentication in smart phones or wearable devices (Chauhan et al., 2017). Beyond that, they can be used for the detection of pathologies or diseases such as COVID-19 (Mallol-Ragolta et al., 2022; Nallanthighal et al., 2022). Ilerialkan et al. (2020) even managed to classify the breather's current posture, i.e. high sitting, low sitting, standing, standing with hands behind head, lying, from breath noises.

2.2.6 Factors influencing speech breathing

Extralinguistic factors Speech breathing varies by age in children (Boliek et al., 2009; Godde et al., 2021) and adults (Hoit & Hixon, 1987).[7] As children get older, the amount of air expired per breath becomes larger and speech breathing becomes less variable and from an age of around ten years it is adult-like barring size-related differences. In adulthood, speech breathing remains relatively stable and starts to change again for speakers in their seventies or eighties: older speakers expend more air per syllable and thus tend to take deeper inhalations before speech onset, which may be due to uneconomical valving in the larynx (Hixon et al., 2020, p. 56–57). Both younger and older speakers may have to do more respiratory work to accomplish the same speech tasks as typical young and middle-aged adults due to smaller usable air volumes, lower recoil forces, higher airway resistance and lower muscle pressures (Huber & Stathopoulos, 2015). Aging typically affects the utterance length between inhalations with older adults producing shorter utterances than young adults (Huber, 2008). Gerstenberg et al. (2018) speculated that changes in the cardiovascular system lead to older speakers having shorter breath cycles and thus more frequent inhalations. Speech breathing can further be influenced by pathologies, such as Parkinson's disease (Huber & Darling, 2011; Huber et al., 2012) or voice disorders (Lowell et al., 2008). Sex or gender has no or only very small influence on speech breathing, after normalizing for differences in size (Hixon et al., 2020, p. 57).

Paralinguistic factors Fuchs et al. (2015b) examined how speaking and physical stress, here via biking, interact in influencing breathing rate. They found that speaking and biking imposed opposing forces on breathing rate: In conditions, where speaking was involved, breathing rate increased with increasing effort. In biking only, it also increased with increasing effort. However, participants inhaled much more often when there was no speech, as speech production otherwise required to keep breathing rate low. In speech conditions, they have to strike a balance between breathing for higher gas exchange requirements and following the changed respiratory pattern of speech.

[7] Different definitions of the terms *extralinguistic* and *paralinguistic* are in use, especially across disciplines. The division of this chapter is partly inspired by Trouvain (2003) and thus uses a similar interpretation of these terms.

Trouvain & Truong (2015) analyzed several prosodic parameters and breathing characteristics in read speech recorded before and after vigorous physical activity. They found that in the post-exercise condition, average pausing time was increased, as were duration and intensity of inhalations. Nearly all pauses in the post-exercise condition were breath pauses. Among the majority of speakers they also found audible exhalations in post-exercise, which were completely absent from the pre-exercise condition.

Ng et al. (2022) studied how 30 speakers of Cantonese, all majoring in sports science, changed their speech and breathing under different exercise conditions. They observed that in exercise, participants required more frequent and deeper breaths and talked faster, thus reducing their speech-to-pause time ratio. How cognitive load may influence respiratory behavior has not been extensively researched, but for tidal breathing it is associated with faster and more frequent breathing which may lead to overbreathing (Grassmann et al., 2016). Signs of trauma or depression may even be linked to different respiratory behavior in speech (Salah et al., 2021).

Language-related factors Speech and speech breathing are intertwined and thus there is a two-way influence of them interacting with and shaping each other (see Fuchs & Rochet-Capellan 2021). Since this work deals with breath noises, we will focus on the influence of speech on breathing here.

Human speech is necessarily shaped by breathing, as it imposes a maximum on length of chunks that can be produced before the next inhalation. Several studies have found breathing to appear every 3–4 s (Winkworth et al., 1994; Kuhlmann & Iwarsson, 2021; Rochet-Capellan & Fuchs, 2013a), which structures speech temporally (Fuchs & Rochet-Capellan, 2021). Sundaram & Narayanan (2003) found the mean number of words to be uttered between inhalations to be 7.46 ± 4.9. Linguistic factors have a strong influence on breath placement, with Conrad & Schönle (1979) stating that "speech is generally not stopped for the sake of gas exchange but for the sake of the speech-generating system itself." They also reported that inhalation occurred exclusively at syntactic boundaries. The placement of inhalations and pauses in speech show strong overlap. However, not all pauses involve inhalation, hence there is some distinction to be made. According to Henderson et al. (1965) breath pauses occur at syntactic junctures only in reading, whereas in spontaneous speech only two thirds of breath pauses were tied to that position. Grosjean & Collins (1979) agree in finding that the longer breath pauses were found at bigger syntactic junctures, while the shorter non-breath pauses were seen at minor syntactic breaks. Winkworth et al. (1994) found that the locations of inhalations in a read text showed high degree of inter-subject agreement and were governed by linguistic factors, as most were related to syntax and occurred at sentence and clause boundaries. Winkworth et al. (1995) reported for spontaneous speech that around 63% of inhalations were taken at structural boundaries. In monologues they followed clause structure more than in conversations.

Speech can also influence the duration and depth of inhalation. When looking at connected speech, i.e. several breath groups divided by several inhalations, one of the larger questions is whether the inhalations are rather predicted by the following utterance, i.e. *planning*, or the preceding utterance, i.e. *recovery*. The debate goes back to and has briefly been discussed in Winkworth et al. (1994) with regard to lung volume: They listed some studies findings lower termination lung volumes after longer utterances and are thus in favor of recovery (Hixon et al., 1973; Wilder, 1983; Hodge & Rochet, 1989). Other studies listed there, however, have found that lung volume levels may be influenced by the length of the utterances that is about to be produced (Gelfer et al., 1983; McFarland & Smith, 1992; Sperry & Klich, 1992). Winkworth et al. (1994) in the same paper speculate that speakers may pre-plan the amount of air in the lungs, and volumes in- and exhaled, in accordance with linguistic categories, by which they mean sentence and paragraph boundaries in comparison to within-sentence boundaries. Since then, Hoole & Ziegler (1997) found evidence for planning of sentence length and loudness to affect inhalations. Whalen & Kinsella-Shaw (1997) found a link between length of the upcoming sentence and inhalation duration, with longer sentences, i.e. containing more syllables, preceded by longer breath noises. Sentence complexity, i.e. number of clauses in the sentence, did not affect inhalation duration. In Fuchs et al. (2013), longer following sentences lead to deeper and longer inhalations, which is likely to be related to air volume requirements. Syntactic complexity did not affect length or depth of inspirations. Winkworth et al. (1995) and Huber (2008) have also found a connection between inhalation and length of the upcoming utterances and attribute this to a planning effect. Contrary to those, Kallay et al. (2019) did not find an effect of planning but rather of recovery, as in their study preceding utterance length had a stronger effect on duration and presence of an inhalation than following utterance length. MacIntyre (2022, p. 121) found length of the preceding and following utterance to positively influence breath duration.

The type of speech used can also have an influence on respiratory behavior beyond the location of breath pauses discussed above. Conrad & Schönle (1979) reported different speech tasks leading to specific speech respiration patterns. For spontaneous speech compared to reading, they noted a higher variability of expiration with a tendency to shorter duration, higher respiratory rate, and more variability in duration of inspiration. Rochet-Capellan & Fuchs (2013b) add that in spontaneous speech, breath locations are less bound to syntactic constituents. Amplitude and duration of inhalations are generally similar to read speech and also reflect the length of the upcoming breath group. They stated that the average breath group is longer in read speech and that generally most parameters are more variable in spontaneous speech. Winkworth et al. (1995) found no differences between read and spontaneous speech for average values of lung volume and breath group length. The ranges, however, were more variable in spontaneous speech. Rochet-Capellan & Fuchs (2013b) further stated that in spontaneous speech, planning and cognitive processes are more involved, which may also lead to the occurrence of filler particles, that may be related to breathing

(Schönle & Conrad, 1985). In addition, spontaneous speech may contain laughter, which has strong effects on respiration (Winkworth et al., 1995; Truong et al., 2019; Filippelli et al., 2001; Rathcke & Fuchs, 2022).

Turn taking in dialogue situations may influence speech breathing. McFarland (2001) even found conversational synchrony in respiratory behavior of conversation partners. Rochet-Capellan & Fuchs (2014) did not find that breathing cycles were systematically coordinated between two interlocutors. Instead, they observed that breathing was involved in turn-taking, as most successful turn takings happened after new inhalations, while unsuccessful attempts at turn-taking were started late in the exhalation phase. Rochet-Capellan et al. (2014) also saw a strong connection between the breathing kinematics and dialogue events. Wlodarczak & Heldner (2020) reported on several studies that have investigated respiratory behavior adjacent to turn-taking events. Among those, the most robust finding was inhalations being compressed in turn-holding, i.e. the rise in the breath signal was steeper and inhalations were shorter, while patterns for taking or yielding were less consistent.

Finally, Lester & Hoit (2014) tested if the airway used for breathing, i.e. nasal vs. oral vs. simultaneous oral-nasal vs. oral-nasal alternations, was affected by the phonetic context, i.e. combinations of preceding and following environment starting and ending in /ɑ b m/. They did, however, not find any significant effect, as the vast majority of inhalations were performed as simultaneous oral-nasal.

Idiosyncrasies It should be mentioned that breathing may differ by individual. Several studies have postulated a *personalité ventilatoire/respiratoire* or *respiratory personality*: Hixon et al. (1973) found high consistency for lung volumes in speech breathing of three out of six participants observed at three time points over two years. Benchetrit et al. (1989) examined how similar tidal breathing profiles were at two time points that were four to five years apart. They found stronger between- than within-participant variability on breathing cycle parameters. Benchetrit (2000) lists several parameters that show a strong degree of individuality, such as breathing frequency (breaths per minute), airflow, duration of in- and exhalation, and tidal volume. Shea & Guz (1992) stated that at rest, different people breathe differently and that this characteristic is quite stable within an individual even over time. In speech breathing during reading tasks, Winkworth et al. (1994) reported strong variability of lung volumes, which was equally strong between and within individuals. Kienast & Glitza (2003) found that oral inhalations were the most common type of speech breath, but there was also a fair share of idiosyncrasies involved, with by-speaker variability in terms of which other types (nasal, sequentially combined oral-nasal inhalation; oral, nasal, sequentially combined oral-nasal exhalation) they used most and how long those were. Most recently, Serré et al. (2021) investigated the ventilatory personality by looking at several breathing profile parameters (cycle duration, cycle symmetry, cycle shape) of German native speakers on two days under different limb movement conditions (arm movement, leg movement, none) while speaking or being quiet. They did find speaker-specific profiles that were consistent for all three parameters, over the different days and over conditions, i.e. also while performing light physical activity.

Chapter 3

Data and annotation for pauses and pause-internal particles

This chapter serves as a discussion of data for the analysis of pauses, PINTS, and breathing. Since my PhD project largely overlapped with the worldwide COVID-19 epidemic, a substantial amount of my work here was done on pre-existing data. The first part provides an insight into how several corpora annotated pauses and PINTS. In the second part of this chapter, a setup is presented that future studies could use for investigating audio, breathing, and glottal behavior.

3.1 Annotation of pauses, pause-internal particles, and non-verbal vocalizations: a look at existing corpora

Pauses and pause-internal particles, as well as non-verbal vocalizations are often regarded as clutter that occurs alongside speech and is peripheral at best to studies of speech. Some of them, such as laughter, clicks, and filler particles, may occur in read speech, but are more likely to be found in spontaneous speech or dialogues. However, natural speech as produced by humans necessarily includes pauses and inhalations.

This yields a great benefit for research on pausing and speech breathing, as they will occur in corpora of human speech regardless of whether or not the experimenters actively tried to elicit them. For reasons of time and costs, however, when annotating speech, decisions have to be made about what aspects to include, and conversely what to exclude, depending on specific or potential research angles and scopes. This means that often these phenomena are partly, or even completely, ignored. Many corpora, although not interested in them, still take into account that some of them may have importance in the speech signal. It is unclear, however, if and to what extent pauses, non-verbal vocalizations, and PINTS are annotated in speech corpora.

It is the aim of this exploratory study to compare corpora that include both spontaneous and scripted speech with regards to how they deal with a selection of PINTS and NVVs. The annotation scheme is borrowed from Trouvain & Belz (2019), who suggest six levels for annotating non-verbal vocalizations:

- speech activity (TURN),
- pauses (PAUSE),
- silence (SILENCE),
- respiratory activity (RESP),
- articulatory activity (ARTIC),
- residual phenomena (ELSE).

Here, we use all but TURN and SILENCE: TURN is not immediately relevant and is mainly determined by the type of corpus, as dialogue speech typically has some sort of turn annotation, while for monologues it is redundant. SILENCE exclusively serves to tease apart silent phases within PAUSE from non-silent phases. That means it is a very useful category for research on PINTS, and breath noises especially, but outside of that, it is not likely to be used. The PAUSE tier is proposed for pauses as audible interruptions of speech, containing both real silence and breath pauses but not closure phases of plosives. Filled pauses or filler particles are also marked on this tier. The RESP level is used for breath events. On ARTIC, non-lexical articulatory events such as clicks are marked. ELSE is then used to subsume laughing and vegetative events such as throat clearing, coughing, swallowing or affective expressions. Non-verbal noises such as knocking on the door or paper rustling can also be put here.

3.1.1 Corpora

Table 3.1 gives an overview of the eleven corpora under investigation, including different types of speech and transcription and/or aligned annotation. For the four categories relevant for PINTS, it marks presence or absence of at least an aspect of it, e.g. only silent pauses for PAU but no filler particles would still count as *y*. As the table simplifies the categories and there may still be strong differences within *y*, the following section will discuss them in more detail.

To test if and how a category was represented, we used the descriptions of the corpus in manuals, readme files, annotations guidelines, or published papers. In addition, we looked at samples from the corpora, except for BEA which was not available to us. Moreover, we text-searched through the annotations for potential label names of some phenomena, e.g. *b, br, inh* for breathing or *p, pau, pause* for pauses (see ch. 3.1.3 for a discussion of this approach).

Table 3.1: Overview of the corpora and the four inspected categories. *y* is for yes, this category or aspects of it are represented; for *n* that is not the case. *?* is for unclear cases.

Corpus	PAU	RESP	ARTIC	ELSE	source
BEA	y	y	?	y	Gósy (2014)
BonnTempo	y	n	n	n	Dellwo et al. (2004)
Buckeye	y	n	n	y	Pitt et al. (2007)
Diapix-FL	y	y	n	y	Wester et al. (2014)
Wildcat	y	y	y	y	Van Engen et al. (2010)
DIRNDL	y	y	n	n	Eckart et al. (2012)
GECO	y	n	n	y	Schweitzer & Lewandowski (2013)
GRASS	y	y	y	y	Schuppler et al. (2014)
IFCASL	y	y	y	n	Trouvain et al. (2016a)
LeaP German	y	y	n	y	Gut (2012)
Lindenstraße	y	y	y	y	IPDS (2006)

3.1.2 Annotation categories

PAU Many corpora mark speech pauses, often as *p*, *<P>*, *_p:_*, or sometimes with other symbols, such as *+*, *#* or a square mark, or just by leaving intervals empty. Buckeye also has *SIL* for non-segmental silence and BEA *[sil]* for silences that are "thought to be longer than usual" (Gósy, 2014).

For filler particles, some corpora attempt to take into account their variable shape with regard to different vowels and how much of a nasal portion there is. They thus annotate them as *ööö, uh, ah, eh, err, um, ähm, mh, mmh, mmm, mmm* (BEA, Buckeye, Wildcat, GECO, Lindenstraße) with many variations by language but also in the exact transcription. On the other hand, some corpora do not make these distinctions: Diapix-FL, that also uses *+* for silent pauses, uses *%* for filled pauses. IFCASL handles it similarly with *#* for pauses and *&* for filled pauses or hesitations. LeaP subsumes filler particles as part of *hp*, i.e. hesitation phenomena.

RESP Three corpora do not include the category RESP at all. Among the rest, there are some differences: many pool all breathing activity together using various labels for it, namely *§* (IFCASL), *h* (DIRNDL), *breath* (LeaP), *<A>* (Lindenstraße), or *<BR>* (Wildcat). Diapix-FL uses *<* for inbreath only. GRASS makes use of *<breathingIN>*, *<breathingOUT>* for in- and exhalations. For BEA, it is not discussed how they handle it in the paper, however, Fig. 6 there suggests that they have *[breath]* as a category. These labels further differ by whether or not they receive their own interval or if they are combined in the same intervals with e.g. speech, as is done in GRASS. In addition, Wildcat for example does annotate respiratory events as *<BR>* but not consistently, i.e. many are not annotated at all, so it may be related to a certain annotator or salience of the event. There, the same label is also used for annotating sighs.

ARTIC Most of the corpora do not have a category that is equivalent to ARTIC. Among those that do are Wildcat, where some clicks were found annotated as *<LS>*, GRASS that has *<smack>*, Lindenstraße *<Schmatzen>*, and IFCASL that subsumes all glottal activity as *q*. In BEA, this category is not discussed in the paper but some parts of it may potentially be annotated as *!*, which is used for non-verbal sounds.

ELSE The most frequent label used for ELSE is laughter: Buckeye uses *LAUGH* and *LAUGH_ WORD* for words produced while laughing. Analogously, Wildcat uses *<LG>* and *<LG_ word>*. GECO annotates laughter as *<LACHEN>* and also uses asterisks to indicate that words were produced while laughing. GRASS has *<laughter>*, LeaP has *laughter*, and Lindenstraße *<Lachen>*.

Other elements from this category are *VOCNOISE* and, analogous to how they handle laughing, *VOCNOISE_ WORD* from Buckeye. This includes vocal noises made by the speaker such as clearing their throat, sighing, etc. GECO uses *<THROAT-CLEARING>* and *<COUGH>* and GRASS *<singing>*, *<sigh>*, *<cough>*. Lindenstraße annotates for *<Ger"ausch>*, *<R"auspern>* (sic!), *<Husten>*, *<Schlucken>*. Two corpora also have a label that may overlap with ELSE, which is *!* for non-verbal sounds in BEA and # for non-speech in Diapix-FL. The same symbol (#) is used by LeaP to prefix non-articulatory noise, i.e. *<#Klicken>* there would indicate clicking that is coming from a different source, e.g. clicking a button.

3.1.3 Discussion

This look at how a selection of corpora deal with pauses, PINTS and NVVs has illustrated that there is quite some variability. All have some way of indicating audible interruptions of speech. Primarily, this is done by using some sort of symbol (often *p* or similar) or using empty intervals, especially for dialogues.

Indicating filler particles is also very common in the corpora investigated here. The only ones that do not have some sort of marking it are BonnTempo and DIRNDL, both of which exclusively contain read speech. The other corpora differ in how they label it – from using one symbol for all filler particles or hesitation phenomena to different ways of accounting for different fillers, e.g. by *uh* vs *um* or even more forms.

Respiration is still relatively commonly annotated, namely in eight out of eleven corpora. Again, corpora differ by how they use it. Most commonly, there is one label for all respiratory activity. GRASS is the most fine-grained here, having a label for in- and exhalation. However, breath events do not always receive their own intervals there, which is not ideal for automatic processing. It would be possible to annotate a higher level of detail here, e.g. by including direction of airflow (in- vs exhalation, as done by GRASS) and also whether it is oral, nasal, or alternating between them.

This however, comes at the cost of accuracy, as it is arguable how reliable that would be (see ch. 6.1).

ELSE is frequently used for laughter (six corpora), some of which even distinguish between laughter and laughing while speaking. Vegetative events like throat clearing, coughing, swallowing are still not too uncommon here. Two of the corpora subsume any non-verbal or non-speech sounds under one label here.

While the overall picture looks quite promising for pauses, PINTS and NVVs, there are of course more fine-grained differences. Even two corpora that use the same labels may have different approaches for the same category. Speech pauses are a great example, as big differences can be found there. Some researchers make use of a threshold, which then again can vary substantially. It may also make sense to let annotators rely on whether or not they perceive a pause (cf. ch. 2.1.1). *Filled pauses* are another example, where it is not clear whether that denominates the entire pause including a filler particle or just the filler particle within the pause (see Trouvain & Werner 2022, p. 62–64 for a discussion on the topic). Similarly, breath noises are also typically surrounded by stretches of silence with durations in the low double-digit ms range (Ruinskiy & Lavner, 2007; Fukuda et al., 2018). When breathing is annotated, it is usually not clear whether it is the entire breath pause or breath event that is included or just the breath noise. While these differences may seem small, they can create problems for automated analyses of large amounts of data.

It must be mentioned here that the analysis in this chapter is error-prone and thus more useful for a general overview than for definitive answers. Wherever available, we consulted a given corpus' manual, readme file, annotation guidelines, or paper. In addition to that, we looked at samples of the corpora, where available to us, i.e. all but BEA. There, we had to rely on the corpus description in the paper only. Moreover, we searched through the annotations for potential label names for some phenomena, e.g. *b, br, inh* for breathing or *p, pau, pause* for pauses. It should be clear that it is complicated to find a category when the exact label or even its existence in the corpus is unknown. As a consequence, it is impossible to claim perfect accuracy or completeness for this survey. Moreover, some of the categories are easier to find than others. Conversely, not finding a given event does not guarantee that it is never annotated: it may still be the case that it did not occur in the samples and was annotated with a label that I did not use when searching for it. In some cases, examples of a given phenomenon were found in the data, yet treated inconsistently. However, the difficulty of getting a definitive answer on some categories is quite telling in and of itself and makes a strong case for clearly describing one's annotation scheme for reasons of transparency and reproducibility.

3.1.4 Conclusion

The selected corpora are quite detailed with regard to the inspected categories. Although the inspected corpora differ in the labels used, they do take pauses, PINTS, and NVVs into account. However, problems can arise when they differ in the degree of granularity. Therefore, it is crucial to be explicit in the guidelines, manual, or readme files. Explicitly dealing with these events can also help standardizing their treatment within one corpus, as one annotator may be more sensitive to one event than another. Corpora are typically designed with some research idea in mind and some aspects will always have to be neglected in the annotation for reasons of costs and time. Adopting some standards for annotation could be beneficial for research. It is of course not realistic to expect corpora that are hardly interested in PINTS to annotate something as airflow direction and airway usage in breath noises or to tease apart filler particles like *um* into a vocalic and nasal section. Conversely, the annotations we have done within the PINTS project are likely to be seen by other researchers as omitting crucial aspects of more speech-related aspects. However, a potential solution could be having some common guidelines, on top of which other researchers could add their more fine-grained annotations.

One great way of handling this is given by GECO-FP (Belz, 2019), which uses the existing GECO corpus (Schweitzer & Lewandowski, 2013), on top of which it adds several tiers relevant for examining filler particles. The six added tiers include filler particles themselves, as well as details about their context, segmental setup, phonation type, and function. Making specialized corpora like this available to others, as done by the author for the case of GECO-FP, could help advance research in those areas of speech that are not contained in or satisfied with the standard annotation. Joining forces could be mutually beneficial for both researchers of speech aspects on the one side and pauses and PINTS on the other, as e.g. breathing and speech are intertwined and adapt to each other (see Fuchs & Rochet-Capellan 2021). An example where both linguistic and paralinguistic aspects could be important is how breath placement interacts with syntactic structure. Here, the syntactic boundaries play a role for where to inhale and different age groups may place their inhalations differently (Huber & Stathopoulos, 2015).

3.2 Pilot recordings of speech breathing

It was mentioned in ch. 3.1 that breathing is an inevitable part of human speech and as such occurs in existing recordings. However, when looking at speech breathing, it is often desirable to test different conditions, control certain factors, or measure different or additional relevant signals. Creating our own recordings gives us some more flexibility, control, and freedom. This chapter aims to give an overview of the setup used and discusses some observations.

3.2.1 Setup

Participants We had two pilot recording sessions. In the first session, two subjects (S1 and S2) were present. In the second session, only one subject (S3) participated. Subject one (S1) was 55 years old and male. Subject two (S2) was 55 years old and female. Both S1 and S2 were native speakers of German. Subject three (S3) was male, 33 years old and a native speaker of American English. Neither S1 nor S2 reported any respiratory problems. S3 reported having had asthma as a child and as a result having a lung reduced in size.

Hardware To track respiration, we used the RespTrack system (Version 2.0, Columbi Computers, Stockholm, Sweden; its predecessor is described in Heldner et al. 2019). This includes two belts to be worn around the rib cage and the abdomen respectively. For glottal behavior and audio, we used the EGG-D800 (Laryngograph, Wallington, UK). The audio signal was recorded with the ECM-500L/SK lavalier microphone (Monacor, Bremen, Germany) that came with the EGG-D800. The electrode cables and the 3.5 mm plug of the microphone were hooked up to the USB interface. Additionally, we used the finger pulse oximeter Pulox PO-200 (Novidion GmbH, Cologne, Germany) to determine participants' pulse rate at two time points, i.e. at rest and after physical exercise.

Software For each of the two devices we used the respective custom-built program on Windows. For EGG and audio that was VoiceSuite10. Before recording, microphone gain was set to +9 dB, as otherwise the audio recordings had been very low in intensity. The recording of audio and EGG was saved as a 2-channel wav file with a sampling frequency of 48 kHz. Since the microphone used is only made for a frequency range of 20-20,000 Hz, the highest frequencies may not be usable for analysis, which is not a problem for our interests, however.

Respiration was recorded with the program RTRecorder that came with the hardware RespTrack system. As audio was already picked up via VoiceSuite, we here monitored and recorded only 5 channels: channel one was for the abdomen belt (AB), channel three for the rib cage belt (RC), while channel two is the weighted sum of the two (SUM). Channel five (SYNC) was used exclusively for tracking the synchronization signal: when prompted, RespTrack can emit a beep that shows up in the SYNC and the audio channel, which can then be used to synchronize the separate recordings. We did not make use of any auxiliary signal in channel 4 (AUX), resulting in this signal to be flat. The resulting file was saved as a 5-channel wav file with a sampling frequency of 400 Hz.

Procedure Participants were recorded while standing and told to move their arms as little as possible. This is not uncommon in these types of recordings (e.g. Fuchs et al. 2015a; Kuhlmann & Iwarsson 2021) and is done to avoid arm movement to introduce artifacts in the RespTrack signal. Instead of overtly instructing participants, MacIntyre (2022) had them hold a large strip of foam to restrict arm movement. However, it must be noted that this is not how humans usually and naturally behave during speech (see Pouw & Fuchs 2022 or Serré 2022). This is important to note, as it is not yet entirely clear to what extent this restriction may have an influence on speech, as some individuals may be affected more than others (Cravotta et al., 2021).

After applying the EGG electrodes, the RespTrack, and the microphone on the participants and starting the recordings, we let the participants perform the isovolume maneuver for calibrating the SUM signal (Augousti, 1997). Subsequently we asked participants to inhale and exhale as much as they can to record their vital capacity.

We elicited two different types of speech here: reading and (semi-)spontaneous speech. The tasks started with the participants reading the standard text *Nordwind und Sonne* (see Appendix A) or *The Northwind and the Sun* (see Appendix B) quietly. This condition will be referred to as *QuietReading*. After this task, we measured the participants' pulse rate at rest, since this task was not very physically demanding. Not measuring right at the beginning also gives them some time to settle in so they would be less affected by previous activities or potential nervousness. This was followed by them reading the text aloud normally (*NormalReading*) and then with as few pauses as possible (*FewPausesReading*).

After this, the (semi-)spontaneous speech section started, where they were first asked to describe what they saw in a picture presented to them (see Appendix C). This spontaneous task is called *PictureDescription* and was followed by a semi-spontaneous speech task, where participants retold *Nordwind und Sonne* that they had read earlier (*Retelling*).

Once this was done, EGG electrodes and the lapel mic were taken off the participant. The RespTrack belts were left on them but unplugged from the cable. This was done so that participants could perform some physical activity, as they were asked to quickly walk down and back up the stairs from the 5th to the ground floor. As they came back, the equipment was put back on them, or checked if it still sat adequately and hooked up again respectively, and their pulse rate was again measured to see how much it was affected by the exercise. After this, they were asked to read *Nordwind und Sonne* one final time, now after having been physically active (*ReadingPhysicalActivity*). This was the last task and the recording was stopped afterwards. This part of the reading task was moved to the end for two reasons: first, to isolate the influence of physical activity on just this one task while avoiding it affecting other conditions. The second reason was that the equipment had to be largely removed and put back on, which is a source of error for measurements. Moving the task to the back thus reduced the danger of this influencing larger parts of the recordings and measurements. For S2, we had to move the *Retelling* task to the end of the recording, i.e. after *PhysicalActivityReading*. It may thus be slightly affected by the physical exercise.

Overall, this resulted in the recording of six conditions, that were either reading or (semi-)spontaneous: For reading there was *QuietReading* *NormalReading*, *FewPausesReading*, and *PhysicalActivityReading*. The (semi-)spontaneous conditions were *Retelling* and *PictureDescription*.

In the second session, we wanted to test silent speech, where the participant uses their articulators but does not produced audible speech. This is called *subvocal* in Conrad & Schönle (1979). This type of speech is relevant for a number of settings in which audible speech is not possible due to reasons of confidentiality, background noise, or physiological problems (Birkholz et al., 2018). For S3, only this condition will be discussed here. Further, we tested rope skipping as physical exercise.

3.2.2 Observations

Reading tasks Fig. 3.1 displays the four reading tasks for S1. The beginning of the initial breath after being asked to do the task was taken as the start and task completion as the end, i.e. either by reading out the final word or by signalling completion in the case of QuietReading. The amplitude of the SUM signal is scaled to each window for visibility. As a result, every signal for each condition by participant encompasses the whole vertical range of the window and thus the amplitude is not informative about absolute values.

For this participant, QuietReading is much shorter at around 24 s compared to the other conditions, which are roughly 41 to 54 s. The breath frequency, i.e. inhalations[1] per minute, is lowest for FewPausesReading (8.8) and QuietReading (12.3), higher for NormalReading (16.7), and highest for PhysicalActivityReading (20.1).

The shape of the breathing signal changes by condition: QuietReading exhibits the typical tidal breathing shape with long inhalations that are similar in length to exhalations. For all other tasks, the breath cycle is reorganized, showing very short inhalations and long exhalations that are used to produce speech. NormalReading and FewPausesReading are similar, the main difference being the few inhalations that are deeper to produce the long exhalations. One of the most striking changes between

[1] There are also some smaller peaks in the signal that look like inhalations, but this count only includes those peaks that did not happen within speech and thus cannot be actual inhalation but rather noise, gestures, or something else in the signal.

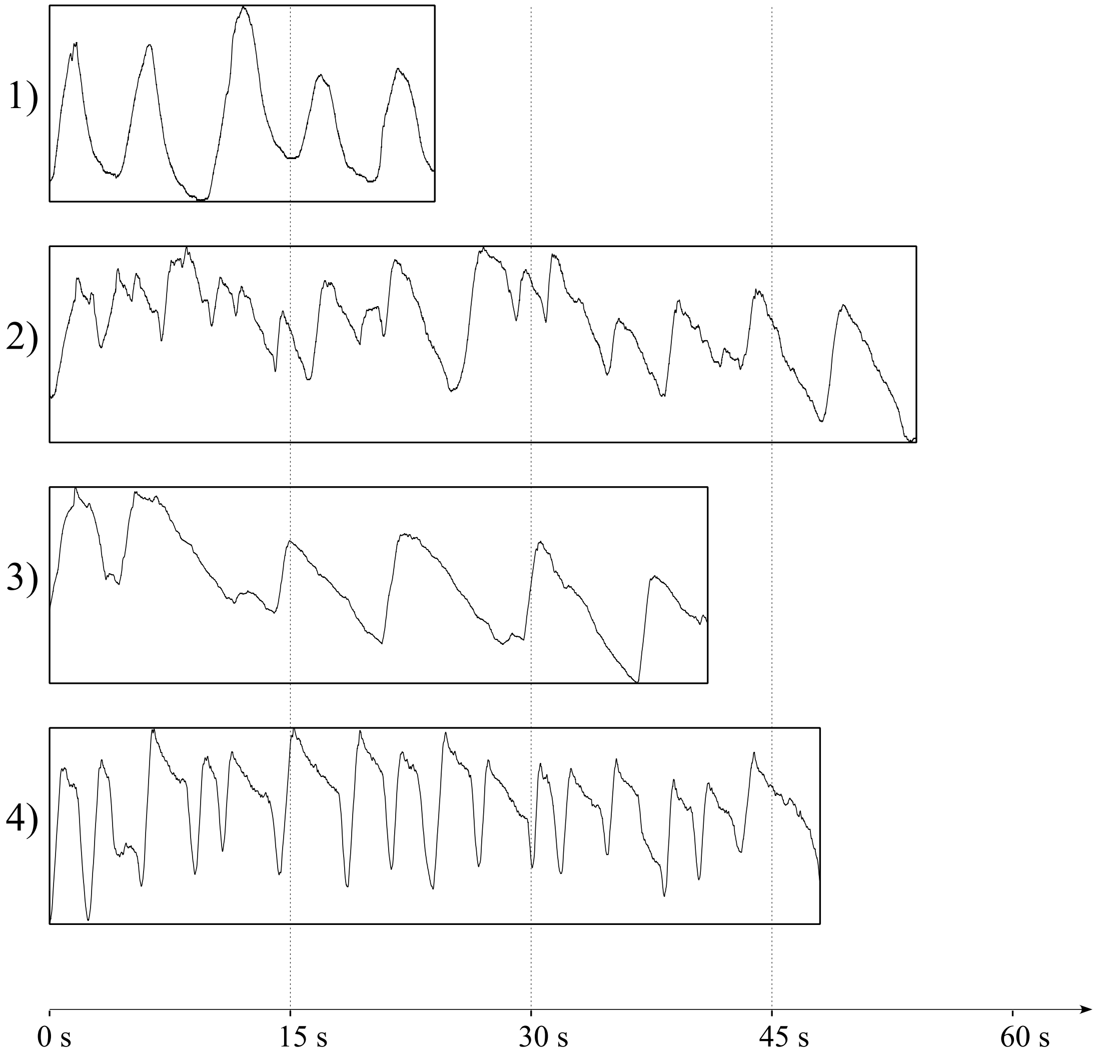

Figure 3.1: RespTrack SUM signals for the five conditions from S1. The x-axis shows time in seconds, the y-axis is in arbitrary units from -1 to 1 for torso expansion, scaled to the respective window. From top to bottom it includes 1) quiet reading, 2) normal reading, 3) reading with as few pauses as possible, and 4) reading after physical activity.

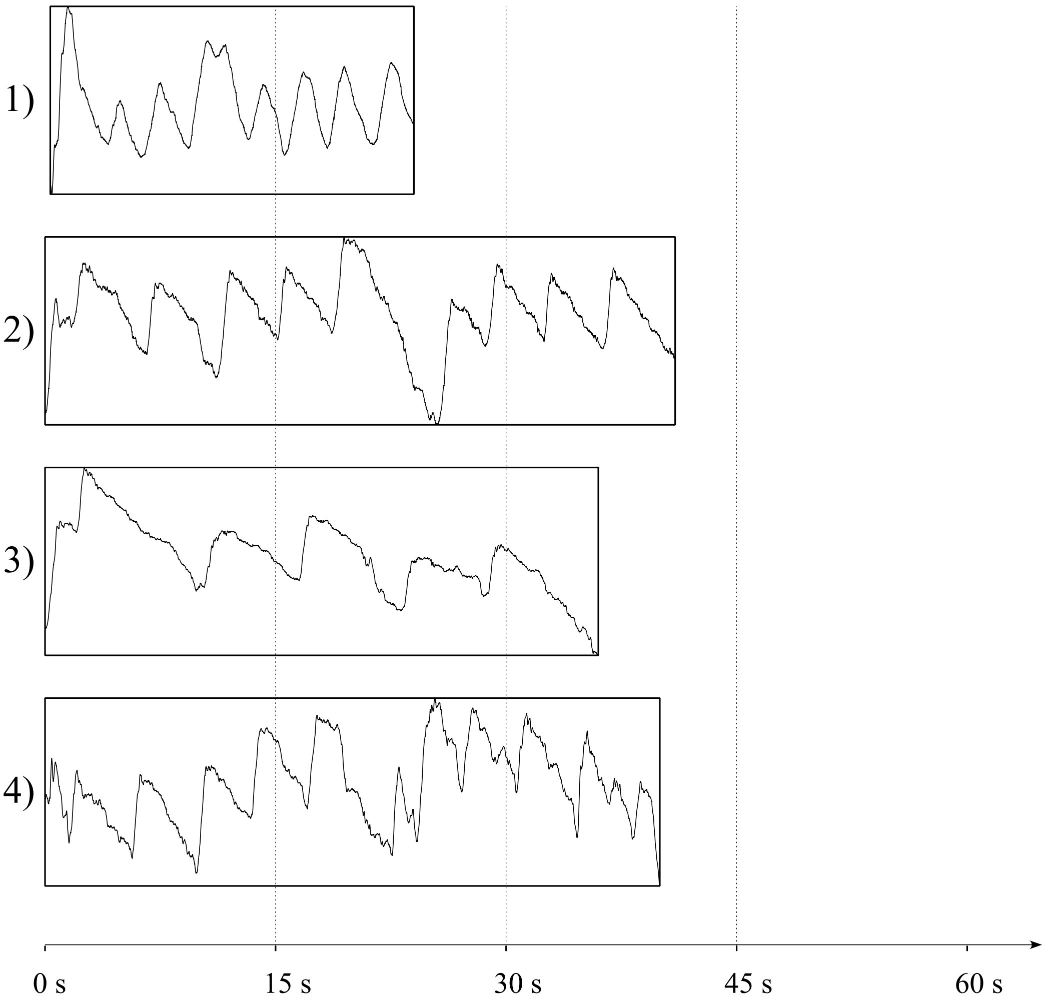

Figure 3.2: The RespTrack SUM signal for the four reading task conditions from Subject2 analogous to Fig. 3.1.

NormalReading and PhysicalActivityReading is the presence of unphonated exhalations. With the higher need for gas exchange in that condition, between speaking and inhaling the participant often had an exhalation phase. This is visible in Fig. 3.1 4), where the typical slow compression of the lungs for speech can be observed, followed by quick compression for exhalation, as shown by the steep decrease in the signal, which is in turn followed by lung expansion for inhalation, represented by a steep increase in the signal.

Fig. 3.2 shows the four reading conditions for S2. As for S1, QuietReading is the shortest at around 24 s compared to the other conditions, which range from 36 to 41 s. Here, there are 19.8 inhalations per minute in QuietReading, 14.6 inhalations for NormalReading, 10.1 for FewPausesReading, and 13.6 for PhysicalActivityReading.

The breathing signal in QuietReading here shows the typical sinusoidal shape. The switch to NormalReading shows the reorganization to the sawtooth shape, which is very similar throughout the task, with one long exhalation followed by an inhalation that are steeper than the rest. FewPausesReading features the lowest respiratory time quotient (RTQ, i.e. duration of inhalation divided by duration of exhalation; see Conrad & Schönle 1979), as the exhalations are very long here. PhysicalActivityReading here differs slightly from NormalReading with only few exercise-induced exhalations visible in the signal. However, the change is far less pronounced than it is in S1. NormalReading and PhysicalActivity are very similar for S2 with regard to total duration and the locations of inhalations.

As mentioned above, the silent speech condition was only included for S3. In the first 8.5 s of that condition, he did not show any respiratory activity. In the roughly 25 s, there were four inhalation phases, that were 1.5 s long on average. These inhalations were quite long and low in amplitude, making its RIP signal look more like rest than speech breathing. We assume that the absence of breathing in the first third is more likely caused by the unusualness of this speech task or the rare usage of it for the majority of people, rather than the silent speech task itself. Further recordings using more participants could show if the respiratory behavior is more similar to rest or speech tasks, as found by Conrad & Schönle (1979).

(Semi-)spontaneous speech tasks The (semi)-spontaneous tasks produced by S1 and S2 are depicted by participant in Fig. 3.3. The durations of the tasks are 52 and 81 s for S1 and 40 and 72 s for S2. The inhalation frequency for S1 is 16.1 for the Retelling task and 14.8 for PictureDescription. For S2, it is 12.1 (Retelling) and 9.1 (PictureDescription). What is new here compared to the reading tasks is that here there are short periods where the RespTrack signal is flat, which may be related to planning. In a more spontaneous speech task or dialogues, those might show up more frequently.

Fig. 3.4 shows the final seconds of the Retelling task produced by S1. It illustrates some aspects of how spontaneous speech typically differs from read speech: breath groups are very variable and due to planning and the repair (*die... der Stärkere*), there is even a monosyllabic breath group. A more cognitively demanding task is likely to introduce more phenomena related to planning, such as pauses and filler particles.

3.2.3 Discussion

The pilot recordings only include a very small number of subjects and are thus not apt for generalizations. However, they are useful for testing the equipment and making some observations. In the data here, there are some differences between the two speakers: For S2, the differences between the conditions were less pronounced, with the exception of FewPausesReading. The breath cycles in QuietReading were shorter in S2 as compared to S1 and thus more similar to NormalReading. FewPausesReading

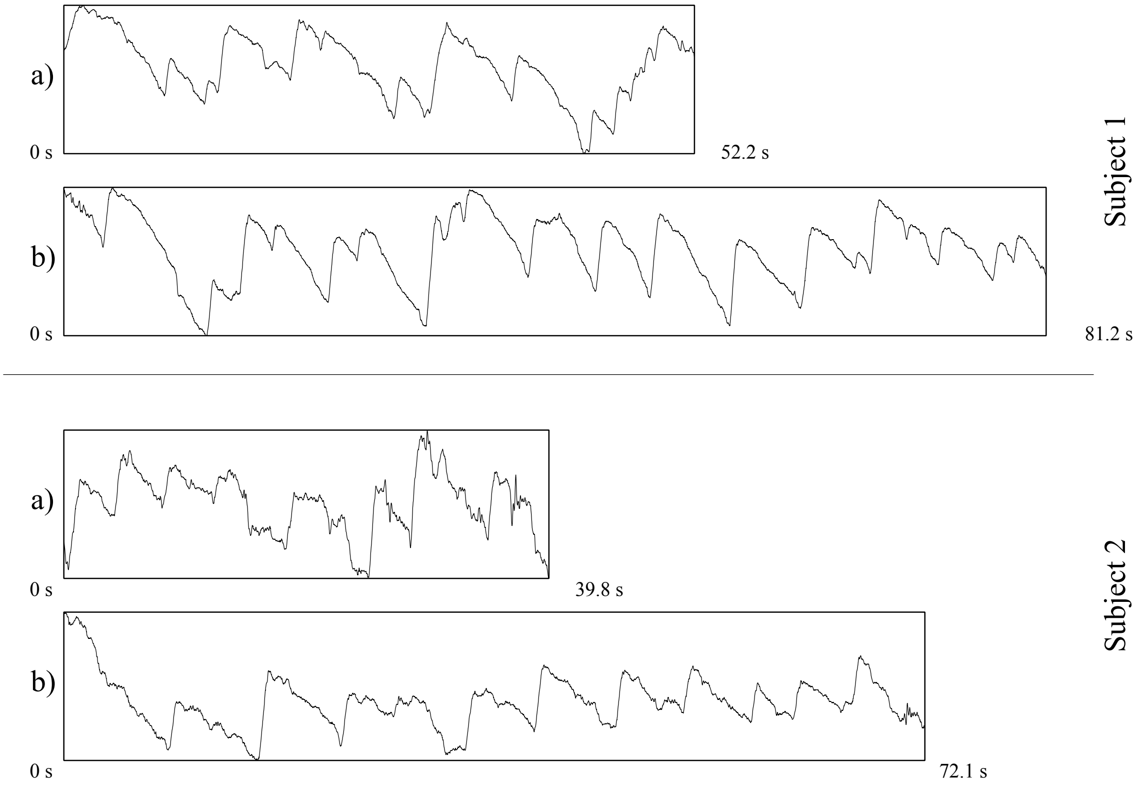

Figure 3.3: The RespTrack SUM signal for the two (semi-)spontaneous speech tasks by subject. The upper half shows S1, the lower half S2. Within a participant, the upper part (a) shows the Retelling and the lower one (b) the PictureDescription task.

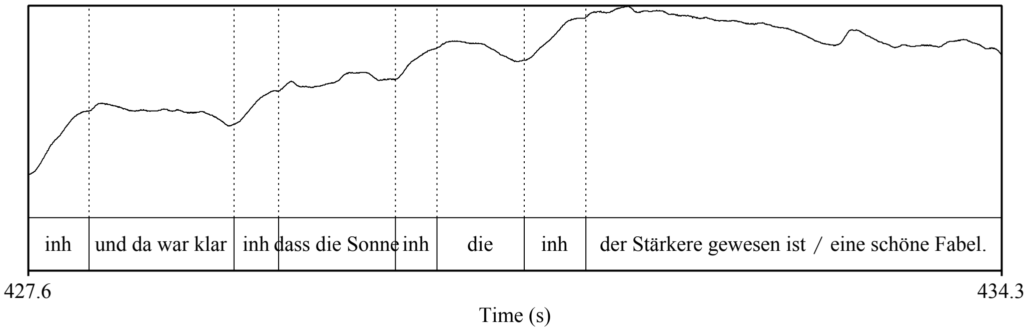

Figure 3.4: The RespTrack SUM signal for the final 6.7 s of the Retelling task as produced by S1.

was very similar between the participants. They differed, however, in PhysicalActivityReading, as for S2, it was quite similar to NormalReading. There were only few cases of an exhalation at the end of the breath group, which were prevalent in S1. The difference between participants for this condition could be related to how they handled the physical exercise which for S1 raised the pulse rate from 87 to 129 bpm, yet for S2 only changed it from 71 to 84 bpm. For S1, who was more affected by the physical exercise, NormalReading and PhysicalActivityReading only slightly differed in terms of the entire duration, as well as the number of breaths, and thus also the breath cycle length. However, the shape of the breathing signal changed with the higher need for gas exchange and the introduction of exhalations.

For features of spontaneous speech breathing, Conrad & Schönle (1979) list higher variability and a tendency to be shorter for exhalations, a higher rate of re-breathing, more variable durations of inhalations and deceleration in the final stage of the inhalation movement. These are of course hard to quantify here with such a small data set. However, comparing the NormalReading to the Retelling task, none of the participants breathed more in Retelling. However, Retelling is more semi-spontaneous than spontaneous, as participants had read the text at least twice before. S1 has a high density of inspiration towards the end of the Retelling, though, as visible in Fig. 3.4. There, the exhalation phases are also quite variable and comparably short. There certainly are some phases in Retelling where the participant is either pausing or producing a filler particle to plan their utterance which is absent from the reading tasks. For a more fine-grained analysis, more data is needed.

There are several aspects that can be improved and could be picked up in the follow-up recordings. A number of technical details should be improved: one of them is the gain in VoiceSuite, that was set to +9 dB, but still is not sufficient for all of the breath noises to be clearly visible or audible in the audio signal. Switching to other programs may be reasonable anyway, since the maximum time to be recorded in VoiceSuite is 30 min. In phases of no speech, the pulse of the participant was visible in the EGG signal. They may have been placed too far back and thus may have picked it up from the jugular veins. Improving the placement of the electrodes may make this problem go away, or alternatively some way of filtering out this repetitive noise could be used.

We here gave S1 and S2 the same task to elicit speech after physical activity, namely walking quickly from the 5th to the 1st floor and back up. Given how differently it affected the two participants' pulse rate and speech, it would be better to make the task more tailored to the respective person. This could be done by adapting the duration or intensity of the task until either the participants report a level of effort (e.g. moderate or strong) or to monitor their pulse throughout the tasks and have them reach a certain value. With rope skipping as done for S3, the exercise can be controlled more flexibly until a desired pulse value is reached.

3.2.4 Outlook for further recordings

One desirable aspect is having dialogues. It is generally a good idea to make the recordings as natural as possible and analyzing read speech covers only one type of speech. This would allow us to also look more into how breathing is adapted for speech there but also how it may be used for turn management (e.g. Włodarczak & Heldner 2016; Włodarczak & Heldner 2018; Wlodarczak & Heldner 2020). The EGG signal could be a useful tool for further research on the role of creak in the respiratory cycle (Aare et al., 2018) and glottal filled pauses (Belz, 2017). It could also help with looking into breath holds, i.e. phases where there is neither in- nor exhalation (Aare et al., 2020), as the glottis should be closed there, unless the pressure in lungs is equal to atmospheric pressure. In combination with physical activity, it could also help us learn about glottal adjustments to high subglottal pressure and speech with high physical load, as several parameters, such as f0 and voice quality measures, related to the glottis may be affected (Kirchhübel et al., 2011; Van Puyvelde et al., 2018; Weston et al., 2021).

The silent speech task, that we briefly touched upon here, could further be augmented by eye tracking or a different type of text presentation, so that the role of punctuation there could be studied. Otherwise, it is not possible to relate in- and exhalations to locations in the text.

As valuable as it can be to work with existing data sets, it is impossible to sample from all types of languages, dialects, populations, speech tasks, technical equipment, etc. There will thus always be limitations, as e.g. in the data used for ch. 5.2, that included only female participants, a low sampling rate, and one speech task. Thus, it is crucial to add to existing data sets for the sake of validity and generalizability.

Chapter 4

Breath and non-breath pauses in natural and synthetic read speech

4.1 A cross-linguistic comparison of natural speakers' pausing behavior at different speech tempos with text-to-speech systems

4.1.1 Introduction

In this study, we aimed to look at pausing behavior in read speech by human and synthetic speakers. On the human speech side, we used speech produced at five different tempos to see how it affects pausing behavior there. In addition, we could investigate how synthetic speech produced at the default tempo would fare in comparison.

Pausing locations in the text and the durations of those pauses are not prespecified to the human or artificial reader, contrary to the segmental sound structure. This optionality in pause usage leads to different strategies on how syntactically conditioned phrase boundaries are to be indicated (Trouvain & Möbius, 2018; Grice & Baumann, 2009), especially with changes in speech tempo, as increases typically lead to a shortening and/or omission of pauses.

4.1.2 Methods

Material The data we used for this study were taken from the BonnTempo-Corpus (BTC; Dellwo et al. 2004), for which short texts of around 80 syllables or 50 words (for an overview of text length in words and syllables see Table 4.5) were read aloud in 5 intended[1] speech tempos: *very slow*, *slow*, *normal*, *fast*, *very fast*. The text to

[1] They are called *intended* here, as participants were instructed to speak at a certain tempo, rather than explicitly specifying the tempo, e.g. by using a metronome. Participants' interpretation

> Am nächsten Tag [1] fuhr ich nach Husum. [2] Es ist eine Fahrt [3] ans [4] Ende der Welt. [5] Hinter Gießen [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.

Figure 4.1: The German BTC text read by participants. Square brackets indicate locations where at least one speaker paused. Numbers in brackets represent number of total pause locations used in that language, not number of pauses used in a given location.

be read is either the German original or close translations of it with high similarity in structure, content, and length. We here included three languages (German, English, French) for which we randomly selected three native speakers each from the corpus. For natural and synthetic speech, identical texts were used that were either the German original (see Fig. 4.1) or corresponding translations of it. On the synthetic speech side, the BTC texts were read aloud by four text-to-speech (TTS) synthesis systems in the three included languages via their web interface. Tempo was not specified or when possible set to normal. The four included TTS systems were: IBM Watson TTS (IWTTS)[2], Google Cloud TTS (GCTTS)[3], Google Translate (GT)[4], and Oddcast TTS Demo (OCTTS)[5].

Analysis In the resulting 57 read-out versions (3 speakers × 3 languages × 5 tempos + 4 TTS systems × 3 languages) the parameters speech duration (i.e. duration it took the participant to read the text from speech onset to offset, including pauses), articulation duration (i.e. speech duration excluding pauses), pause duration (i.e. sum of the durations of all pauses produced by the participant over the text), pause locations (i.e. where in the text the participant took a pause), number of pauses (i.e. how many pauses did the participant produce in reading the text), and number of breath pauses (i.e. number of pauses including audible breath noises) were measured in Praat (Boersma & Weenink, 2019). Breath noises occurring in speech pauses in a reading task are typically inhalations, hence these are used interchangeably here. The values for natural speakers given in the tables and figures are averaged over speakers of that language in this particular tempo.

of those tempos may thus vary.

[2] Accessed via https://text-to-speech-demo.ng.bluemix.net/ on 14/11/2019.

[3] Accessed via https://cloud.google.com/text-to-speech/ on 14/11/2019.

[4] Accessed via https://translate.google.com/ on 14/11/2019.

[5] Accessed via https://ttsdemo.com/ on 14/11/2019.

All perceived pauses were included regardless of their minimum duration. We chose to do so, as there is no generally agreed-upon minimum threshold, instead depending on purpose and approach of a study it can vary vastly from as low as 5 ms up to 400 ms (see ch. 2.1.1 for a discussion of this). Where a pause was followed by a plosive, we subtracted 30 ms from the pause duration to account for the closure phase initial to the post-pause speech.[6]

4.1.3 Results

Articulation duration and speech duration Table 4.1 provides an overview of the mean articulation durations in the varying speech tempos as produced by the natural speakers in comparison to the speech produced by the synthesis systems. The respective speech durations are also visualized in Fig. 4.2 as the sum of articulation and pause durations.

For the natural speakers, the expected tendency of a decrease in articulation duration with an increase in tempo can be observed for all languages. In the case of German, the TTS systems' articulation durations were roughly in the area of *slow*, for GT it was even slower than *very slow*. For English, TTS systems landed between natural speakers' mean articulation durations for *slow* and *normal*. GT, again, showed to be an exception that was slightly higher than *very slow*. For French, TTS systems were also between *slow* and *normal*, with the exceptions of GT, which was higher than *very slow*, and GCTTS, which was below *normal*.

Pause duration A comparison of the different pause durations, i.e. natural speakers in varying tempo conditions and TTS systems, can be seen in Table 4.2. In addition, pause durations are visualized in Fig. 4.2 along with articulation durations.

For natural speakers, pause durations decrease, as expected, with an increase in tempo. For German, some pause durations of natural speakers are strikingly shorter than those of English and French: This is the case for *very slow* and *slow*, whereas for the others at least one other language is comparable. The German-speaking TTS systems also have the shortest pause durations (with only one exception, namely GCTTS), but the differences to the other languages are relatively small in most cases. Comparing TTS systems' pause durations to those from real speakers, all German and French TTS systems produce pauses that are shorter than *normal* and longer than *fast*. For English, most are longer than *normal*, except for GCTTS which is very close to *normal*.

[6] This number is of course arbitrary and actual closure durations may vary by place of articulation and by voicing. We chose to subtract this number to approach pauses here not as stretch of silence (which would be disrupted by breath noises and clicks anyway) but rather as a combination of absence of audible speech and articulatory gestures directly related to speech sounds.

Table 4.1: Articulation duration (in seconds): human speakers are averaged over by the five speech tempos (very slow, slow, normal, fast, very fast); the synthetic speech side reports the actual number of the respective system (IWTTS, GCTTS, GT, OCTTS). All are split by language.

	very slow	slow	normal	fast	very fast	IWTTS	GCTTS	GT	OCTTS
Ger.	18.42	16.02	13.81	12.84	9.43	16.74	16.43	20.01	15.84
Eng.	18.08	15.67	12.58	10.84	8.76	14.44	14.76	18.60	15.05
Fre.	19.22	18.27	14.83	12.53	9.84	15.98	13.81	18.56	15.33

Table 4.2: Pause duration in seconds for natural and synthetic speakers.

	very slow	slow	normal	fast	very fast	IWTTS	GCTTS	GT	OCTTS
Ger.	3.74	2.83	2.23	1.34	0.14	1.86	1.71	1.79	1.69
Eng.	5.49	4.22	1.79	1.23	0.41	1.91	1.76	1.90	2.82
Fre.	6.50	6.03	3.82	1.40	0.24	2.16	1.67	2.26	1.74

Pause locations There were small differences between languages with regard to where pauses were placed in the text by at least one speaker: whereas German speakers used 16 different locations for pauses, English speakers made use of 19 and French speakers of 14 positions.

TTS systems in all languages exclusively put pauses in locations that were also used by natural speakers, i.e. they did not introduce any new positions. Often, but not always, they utilized punctuation as a cue for that. In the case of German, there were two potential pause locations without punctuation where two or all four respectively implemented a pause. In the latter case, the pause location was preceded by the conjunction *and*. The English-speaking TTS systems produced only one pause that was not triggered by punctuation. It was not adjacent to a conjunction either. However, it was in a position, where a comma would be possible and used quite frequently ("after Bristol [pause] the towns"). The synthetic French speech featured four pauses that were not adjacent to punctuation: however, three of those were followed by the conjunction *et*. The majority of those pauses was shorter than 100 ms and thus comparatively short.

Pause number Table 4.3 provides on overview of the number of pauses produced while reading the text. Comparing the TTS systems' numbers to *normal* as a reference, they are similar for German and English but a bit higher on the synthetic side. In French, however, natural speakers produced far more pauses than the respective TTS systems, with the exception of IWTTS that is quite close. In general, French natural speakers tended to pause more often than German or English speakers, especially in *normal*.

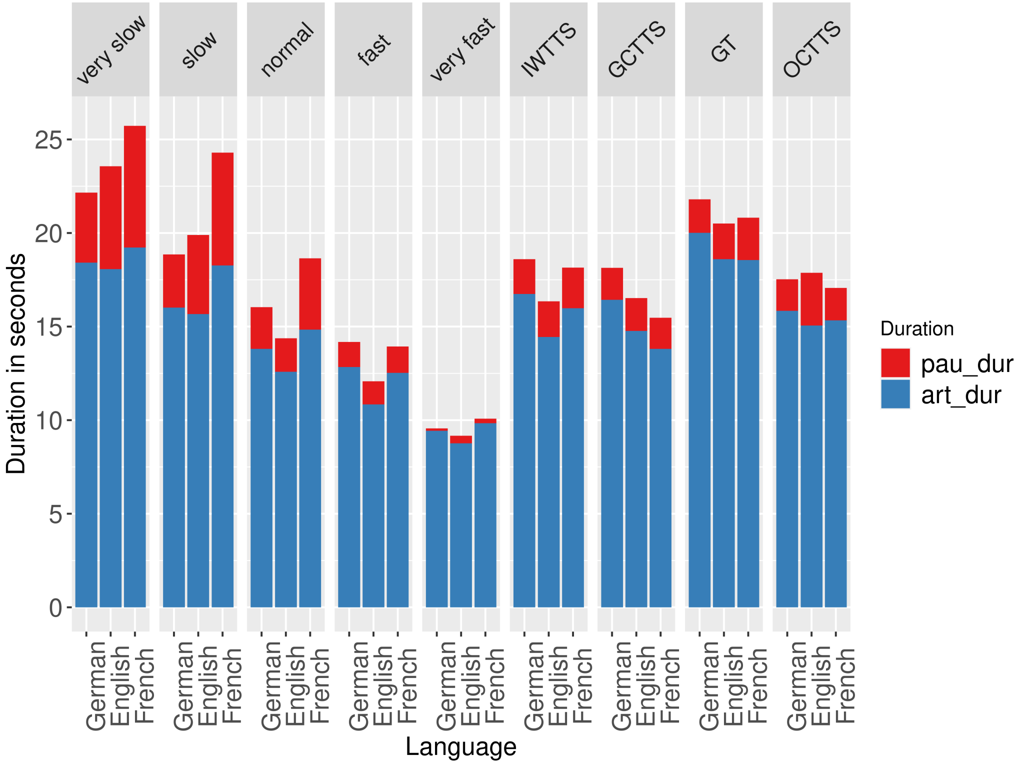

Figure 4.2: Comparison of articulation durations (art_dur) and pause durations (pau_dur) by languages and speech tempos or TTS systems respectively. Speech duration can be seen here as their sum.

Breath pauses Table 4.4 provides an overview of the number of pauses that include audible breath noises. Natural speakers and TTS systems are quite similar in this regard across languages. The biggest difference in breath pauses exists between natural and synthetic speech: none of the TTS systems included in this paper used breath noises. Fig. 4.3 visualizes how many pauses were taken: overall pause numbers for natural (for five speech tempos) and synthetic speakers as a sum of pauses with and without breath noises, split by language.

Table 4.3: Mean number of pauses produced by natural speakers in the five speech tempos and number of pauses produced by the TTS systems, split by the three languages.

	very slow	**slow**	**normal**	**fast**	**very fast**	**IWTTS**	**GCTTS**	**GT**	**OCTTS**
Ger.	7.7	7.3	5.7	4.3	0.7	7	7	6	6
Eng.	13.3	11.0	5.0	5.7	2.3	6	6	7	6
Fre.	11.0	10.3	9.3	5.7	0.7	8	5	6	5

Table 4.4: Number of pauses produced that included audible breath noises by languages and tempos. For natural speakers, the numbers are averaged over speakers.

	very slow	slow	normal	fast	very fast	IWTTS	GCTTS	GT	OCTTS
Ger.	5.7	4.3	4.0	3.0	0.3	0	0	0	0
Eng.	5.3	4.7	4.0	3.3	1.0	0	0	0	0
Fre.	5.3	5.3	5.3	2.3	0.7	0	0	0	0

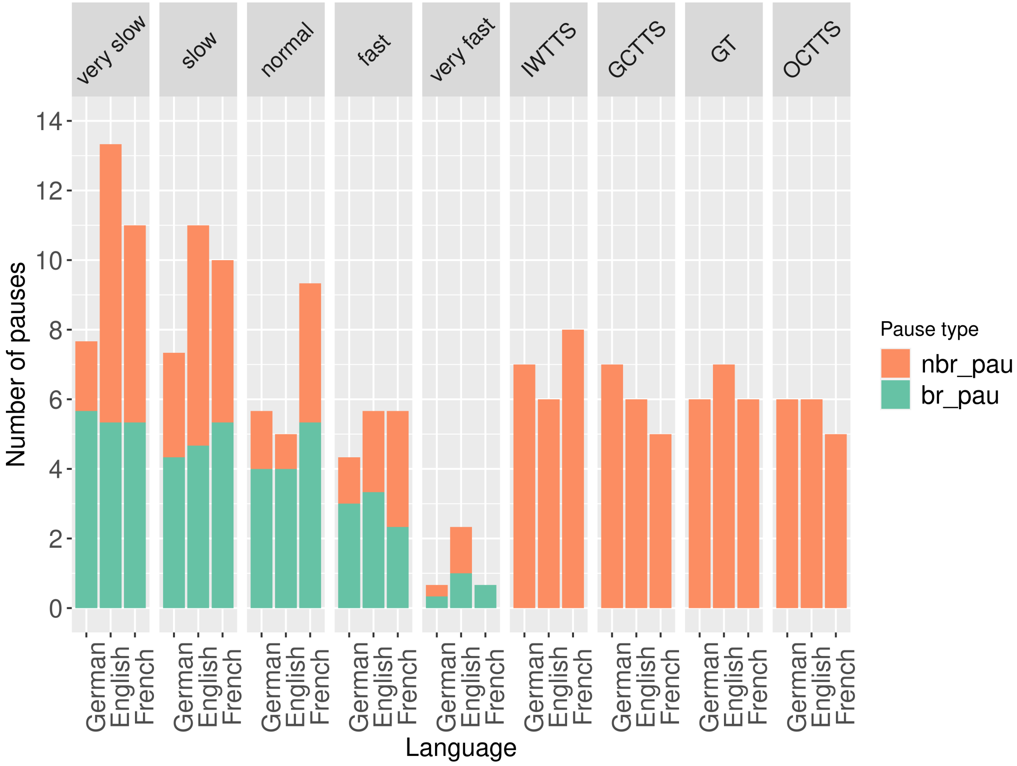

Figure 4.3: Number of pauses with (br_pau) and without (nbr_pau) audible breath noise by language and speech tempos, as well as in TTS systems.

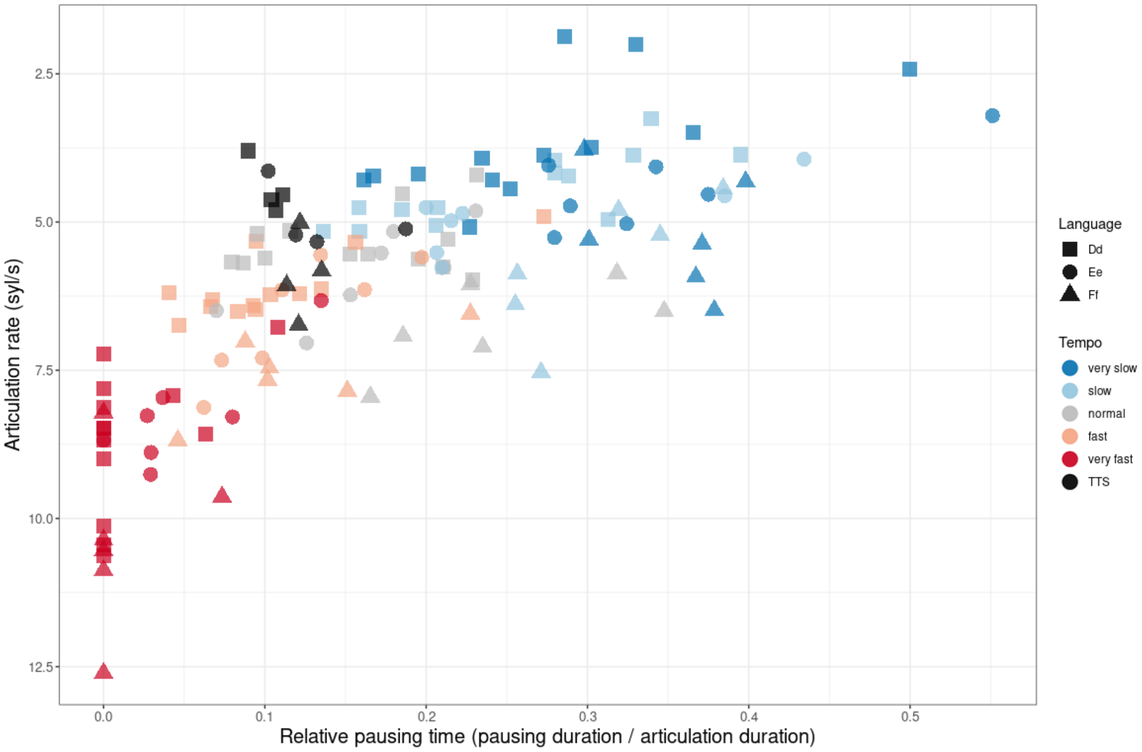

Figure 4.4: Articulation rate (in syllables per second, on flipped y-axis) and relative pausing time (pause duration divided by articulation duration) in real speakers (colored) and TTS systems (black). Language is indicated by shape, for natural speakers colors indicate tempo.

Combination with data from ch. 4.2 For TTS, it is important to strike a balance between articulation rate and pausing frequency and duration, as they may be related in real speakers' articulation. We here combine the synthetic speech data from this study with the German, English, and French speakers from ch. 4.2[7] to see how the TTS systems would behave in relation to more speakers.

For this we used articulation rate, i.e. syllables per second, and the relative pausing time, i.e. pausing duration divided by articulation duration (see Fig. 4.4).

4.1.4 Discussion

Along with the mean values presented above, differences can also be observed in the pausing strategies individual speakers employ. Fig. 4.5 visualizes the pausing behavior of a single German speaker who behaves like the ideal textbook example: with virtually every increase in tempo, the number of pauses decreases and so does the duration of many individual pauses with tempo increase. In the faster tempos, pauses are omitted, in *very fast*, the speaker does not paue at all. Moreover, all pauses

[7] The natural speakers are from the exact same corpus, they are just added here to have more data points.

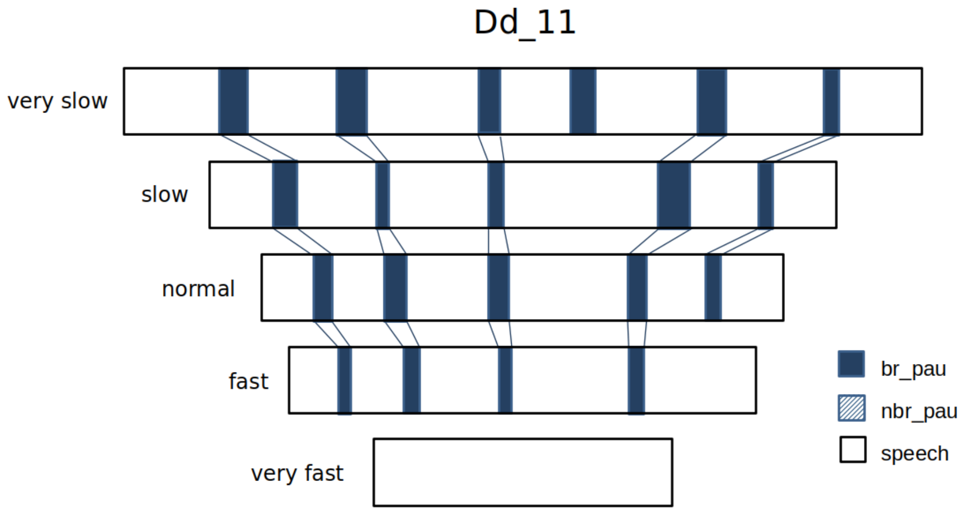

Figure 4.5: Visualized pausing behavior of one German speaker (Dd_11): the long, unfilled squares represent the entire speech duration for the respective tempo, the dark blue squares within them represent the duration of a given breath pause. The length of all boxes roughly corresponds to their durations. Connections between pauses over tempos signify that the same location was used for a pause (see Fig. 4.1 for pause locations).

are also used for inhalations, as there is not a single non-breath pause here.

Fig. 4.6 paints a different picture with more variability: whereas some of the general tendencies can largely be observed here as well, such as retention of some *important* pauses, reduction in pause number and shortening of pause duration with an increase in tempo, there are some differences. In *slow*, some pauses are introduced that were not present in *very slow*. Moreover, this speaker makes use of non-breath pauses, partly in the same locations as breath pauses. Compared to the *ideal speaker* in Fig. 4.5, there is more inconsistency and variability with regards to pause realization, as is also likely to be the case when looking at a larger number of participants and at natural speech.

In this study, speech tempo variations lead to changes in the articulation duration, speech duration, and pause duration, as well as the number of pauses with and without audible breath noises. Articulation and pause duration decreased in a relatively constant way, as speech tempo increased. The two slower and normal tempos were relatively similar with regard to breath pauses (partly even in *fast*), with a stark decrease setting in with the faster tempos and with *very fast* especially. This may be related to the fact that speech tempo was only intended, not explicitly set.

Articulation rate in TTS systems tends to be slower than the *normal* rate of natural speakers. The French version produced by GCTTS is the only TTS version that is faster than the respective *normal* articulation rate.

For the overall duration and number of pauses, however, the German and English

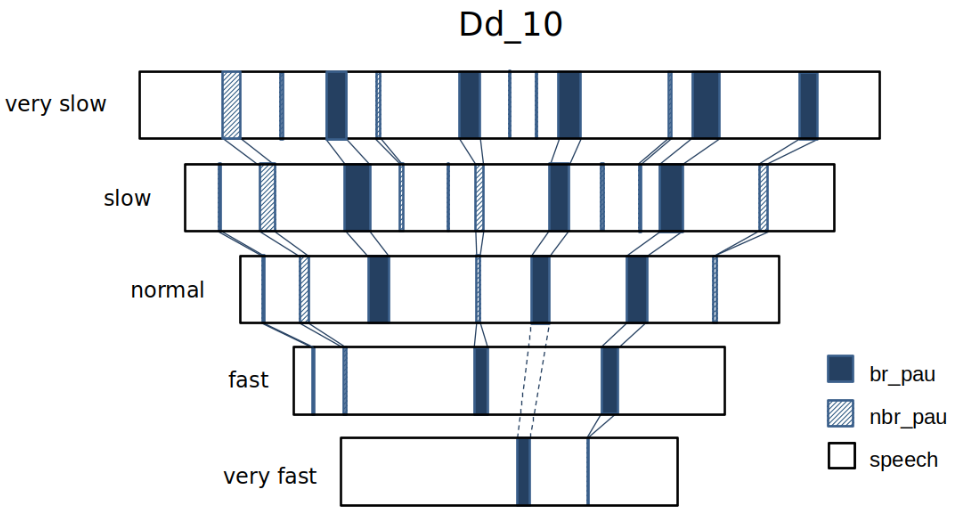

Figure 4.6: Pausing behavior of another German speaker (Dd_10): the squares hatched in light-blue represent non-breath pauses (i.e. without audible breath noise). When pauses are connected by a dashed line, it means that they were not used in the next tempo level but picked up again in a higher tempo. The rest is equivalent to Fig. 4.5

TTS systems fall between natural speakers' *normal* and *fast*. Pause durations within TTS systems were relatively constant across languages and for English mostly above *normal* and for German and French (partly far) below. The number of pauses produced by TTS systems was higher than *normal* for English and German. in French, natural speakers in *normal* took far more pauses than the TTS systems, which is due to the high numbers of natural speakers rather than the TTS systems. The strongest difference can be found for breath pauses, as none of the included TTS systems made use of them at all. Natural speakers showed strong similarity across languages in the respective tempos.

TTS systems being used by different people in a variety of areas makes it hard to establish a single, universally preferred articulation speed. As this study found that TTS systems tended to have a slower articulation duration than the respective *normal* speech tempo, adapting the pauses (which were shorter than *normal* for German and French) could be beneficial. Whether or not breath noises should be used in synthetic speech depends on a number of factors such as potential benefits to information processing of the human listener, as well as the respective setting or task in which human and computer interact. Personal preferences and experiences of course contribute to this, too.

The combination of data from this study with more data points from ch. 4.2 relates articulation rate and how much speakers (natural and synthetic) paused compared to their articulation duration. It suggests that some TTS systems do not strike the balance between them and thus stand out. The most striking ones are the German TTS systems, however, the German natural speakers also tend to have a slower articulation rate as well. French TTS, which seems to fall neatly within the area covered by natural speech, still does not perfectly integrate into the colored cloud, where French natural speakers are mostly at the lower end, i.e. with higher articulation rates. For English, TTS seems to be doing well with only one or two systems being somewhat distant from the natural English speakers.

Overall, this shows that these four TTS systems, not all of which are cutting-edge anymore, do relatively well, although they could blend in more with natural speech by increasing articulation rate slightly or pausing a little more, or ideally do a combination of both. Whether synthetic speech should only copy natural speech as best as possible or whether other factors come into play there should be considered, but is beyond the scope of this work.

4.1.5 Conclusion

This study looked at read speech and how parameters related to articulation and pausing change as a function of tempo in natural speakers, with the goal of locating synthetic speech in those tempo steps. We did so for three different languages, namely German, English, and French. For human speakers, we found the expected effects of increasing tempo, such as shorter articulation, speech, and pause durations, and less pauses with and without inhalations. In TTS systems, articulation duration is generally longer than in natural speech, while the duration of pauses tended to be shorter than at the reference tempo *normal*. The number of pauses was still higher in TTS, indicating that the pauses taken there were comparably short. Finally, none of the TTS systems in this study produced any breath noises.

In this study, we only covered a tendency of average values as observed in very few natural speakers. Future studies should include more speakers and also delve more into the variance or range of results to find *acceptable* ranges rather than a single number. Most importantly, adapting synthetic speech in the proposed directions, i.e. articulating a bit faster with slightly longer pauses that could potentially include breath noises, needs to be confirmed by perception studies first. There, the synthetic speech should be tested for specific domains, as a high naturalness may be preferred in some settings, such as audio book reading, while in other cases, such as a driving assistant in a car, clarity of instruction is arguably more important. By contrast, a ticket machine at a train station should be very understandable but also reasonably fast. Hence, TTS should be tailored at the specific needs and environmental conditions of its application (see e.g. Parlikar & Black 2012).

4.2 Optionality and variability of speech pauses in read speech across languages and rates

4.2.1 Introduction

Prosodic phrasing of the same text is not fixed and underlies variation, both between (Goldman-Eisler, 1961; Trouvain & Grice, 1999) and within speakers (Trouvain & Grice, 1999). In addition, prosodic boundaries can be divided into those that are *obligatory*, e.g. for disambiguation, and those that are *optional*. In (second language) teaching, several example sentences with obligatory prosodic boundaries are used, as in (1); missing or moving a prosodic boundary to a different position would alter their meaning:

1. (a) To govern | people use language.
   (b) To govern people | use language.

However, most prosodic boundaries are optional, as exemplified in (2):

1. (a) The president's advisors fear | an early announcement | would complicate his fundraising | and other activities.
   (b) The president's advisors fear an early announcement | would complicate his fundraising and other activities.
   (c) The president's advisors fear an early announcement would complicate his fundraising and other activities.

Purely punctuation-based modeling of prosodic phrasing would only predict option (2c) and ignore other alternatives. More elaborate models of prosodic phrasing (e.g. Gee & Grosjean 1983) consider additional factors, such as rhythmical balance, but still follow the 'one size/prediction fits all' paradigm.

In this study, we used speech pauses as a proxy for prosodic boundaries (PBs) to analyze two dimensions of prosodic phrasing behavior: *optionality*, i.e. absence/presence of a pause in a given location, and *variability*, i.e. variation in pause duration and the involvement of breath noises. We analyzed recordings of read speech from different languages using semantically comparable texts, that were performed at five different intended tempos, as changes in tempo have strong effects on the prosodic phrase structure (see ch. 2.1.5).

Optionality and variability of pauses surface particularly when varying speech tempo. They can be exemplified with the speaker in Fig. 4.5 who exhibits a highly consistent pausing scheme using only breath pauses and reducing number and duration of pauses as speech rate increases, although not linearly for individual pauses. This speaker also does not introduce any new pauses at faster rates. In contrast, the speaker in Fig. 4.6 shows more pause diversity, in terms of optionality and variability, with non-breath pauses being used here, new pauses emerging in the slow compared to very slow rate, and some breath pauses turning into non-breath pauses and vice versa.

Predicting PBs is at the core of prosody modeling, however, so far the *optionality* of PBs has not been in its focus. The missing optionality is also reflected in the prediction of PBs in synthetic speech which leads to the same PBs for all speech styles and rates. The paradigm 'one style fits all' still seems to be prevalent. In contrast to synthetic speech where PBs are always used at the same location, human speakers do not always use the same PBs when producing the same text. Human speakers make use of the optionality of phrasing on three levels:

- whether to use a PB at all at a given word boundary,
- to apply a certain strength/level of PB, e.g. a major or a minor boundary,
- how to realize a given PB of a given strength, e.g. with a longer or a shorter pause or no pause at all.

4.2.2 Methods

We used data from the BonnTempo-Corpus (BTC; Dellwo et al. 2004) for Czech, English, French, German, and Italian, as well as the Arabic Speech Rhythm Corpus (Ibrahim et al., 2020). The latter adds Egyptian Arabic to the set of languages, sticking close to BTC's methods. As described in ch. 4.1.2, the original BTC versions are very similar in many respects, however, the Arabic one differs in terms of length, content, and punctuation, as can be seen in Table 4.5. The texts were read aloud by native speakers at five intended speech rates: *very slow*, *slow*, *normal*, *fast*, *very fast*. The number of speakers per language is given in Table 4.5. We decided to exclude two speakers (Dd_20, Dd_21) from the analysis because their breathing was extremely soft so it was not possible to reliably annotate their breathing. We analyzed 46 speakers producing 230 versions containing a total of 1,710 pauses.

As in ch. 4.1, we did not use a fixed duration threshold for annotating pauses but included all perceived pauses (e.g. via final lengthening) that contained at least a silent period or breath noise. We used the existing hand-labeled corpus annotations and added our additional aspects of interest manually, such as inhalations, additional pauses, and preceding word number. The segmentation was clear for the vast majority with the exception of a few cases where the breath noise was very soft. In the original annotation when voiceless plosives followed after pauses, around 50 to 100 ms of silence were annotated as belonging to the plosive to account for the acoustically silent closure phases. We analyzed the following parameters: pause duration and placement, presence and duration of breath noise within the pauses, as well as duration of left and right edges. The latter are short periods of silence typically found right before

Table 4.5: Overview of material used from the corpora: number of speakers, words, and syllables by language.

	speakers	words	syllables
Arabic	9	66	178
Czech	9	47	93
English	7	55	77
French	6	53	93
German	13	49	76
Italian	2	49	106

and after inhalations (Ruinskiy & Lavner, 2007; Fukuda et al., 2018). We annotated edges for every breath pause, which can lead to some edges being very short, as in some cases they can practically disappear at faster rates. For the analysis, we did descriptive statistics only in this experiment. For pause placement, we assigned every word, which we here defined as a string of letters surrounded by white space or punctuation, its number of occurrence in the text so that a given pause's location number refers to the word it follows.

4.2.3 Results

Pause location Fig. 4.7 shows that some pauses are less optional than others, visualized by dots accumulating in vertical lines: while there is optionality for many pauses (e.g. the first 8 locations in Arabic), every language has a number of pause locations that stand out by being preferred for pausing, tending to have longer pauses with many exceeding 500 ms, and being more likely to involve breathing (as e.g. location 9 in Arabic). From the original BTC versions, Czech, English and German pauses accumulate 6 to 7 of those lines (cp. the German speaker in Fig. 4.5), whereas French shows less uniform pausing behavior with about 10 lines, although some of them are quite short and do not contain many breath pauses. Italian might follow the majority of the BTC, but with two speakers the tendency is not clear. The longer Arabic data set features about 10 lines and also differs regarding pause locations and inhalations in *very fast*, as here each line also includes an inhalation from that condition.

The clearer lines are closely related to punctuation <, ; .> and conjunctions (e.g. *and* in English) that reflect boundaries between larger syntactic structures. Czech and English and to some degree Arabic and German show a larger number of vertical pause lines than there are punctuation marks in the text. The French text contains a comparably high number of punctuation marks (n=10), and there are few pauses that do not coincide with them. Overall, breath pauses seem closely related to punctuation and conjunctions and rarely appear outside of these locations.

To illustrate the relative importance of some pauses, we looked into the three bigger

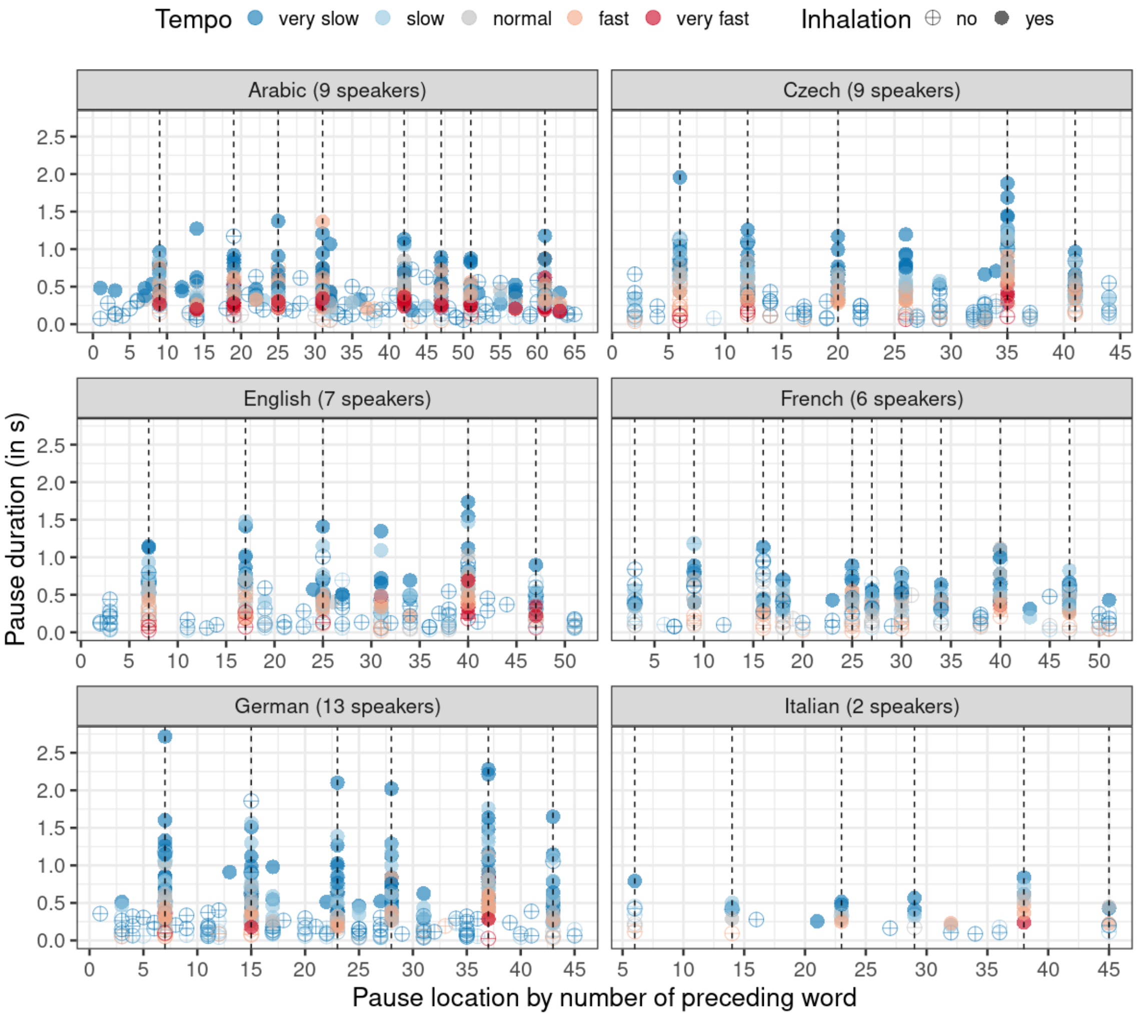

Figure 4.7: Pauses and their durations by location in the text, sorted by language. Pause location is determined by the number of the preceding word. Whether a pause contains an inhalation is indicated by filled dots (inhalation) or empty, crossed dots (no inhalation). Dashed lines indicate locations with punctuation <, ; .>.

Table 4.6: Number of occurrence of breath and non-breath pauses by rate, pooled for all languages.

	very slow	slow	normal	fast	very fast
breath pause	285	253	225	158	62
non-breath pause	314	187	110	95	21
total	599	440	335	253	83

pause accumulations towards the end in Czech, i.e. locations 26 (no punctuation), 35 (full stop), and 41 (comma). We looked at mean pause duration, number of pauses taken compared to potential pauses here (9 speakers × 5 rates = 45), and number of breath pauses compared to number of pauses taken in this location. Location 26 (no punctuation) has a mean pause duration of 497 ms, 35 (of 45) pauses taken here of which 28, i.e. 80%, are breath pauses. At 35, mean duration is 882 ms, 41 pauses were taken with 37 (90%) involving inhalation. At location 41, pauses have a mean duration of 482 ms with 34 pauses taken here, of which 21 (62%) are breath pauses.

Number of breath and non-breath pauses As rate increases there is a clear tendency towards fewer breath pauses and particularly fewer non-breath pauses (see Table 4.6). There is much more flexibility regarding non-breath pauses across the rates than breath pauses. There are 15 times as many non-breath pauses at *very slow* compared to *very fast*, but for the same relationship in breath pauses this factor is only 4.5. For most rates there are more breath pauses than non-breath pauses. The ratio of breath to non-breath pauses at the *normal* rate is 2:1, and 3:1 for *very fast*. However, for *very slow* there are more non-breath pauses than breath pauses.

Duration of pauses, breath noises, and edges Generally, as rate increases, pauses tend to be shorter. Fig. 4.8 shows breath and non-breath pauses and that varying rate also has an effect on the duration of inhalation, which tends to become shorter as rate increases. The pauses from faster rates are all rather close to the reference line, i.e. the inhalation fills a large portion of the pause, thus also resulting in shorter edges. In the other conditions, pauses appear more distant from the line, too. At slower rates, pauses can also be found at relatively short durations since participants generally pause more when reading at a slow or very slow rate, as can be seen in Table 4.6. Conversely, not all faster rate inhalations are short, which might be related to a possible strategy of reducing the number of pauses at these rates and then inhaling longer and/or more deeply when pausing for it. In general, though, with increasing tempo the point cloud moves closer to the line and extends less in the upper, i.e. longer inhalation duration, regions.

After Fig. 4.8 has already indirectly shown that edges are shortened in faster tempos, Fig. 4.9 puts the focus on edges only: It supports the interpretation that the durations of edges around inhalations vary by rate. When participants inhale in the

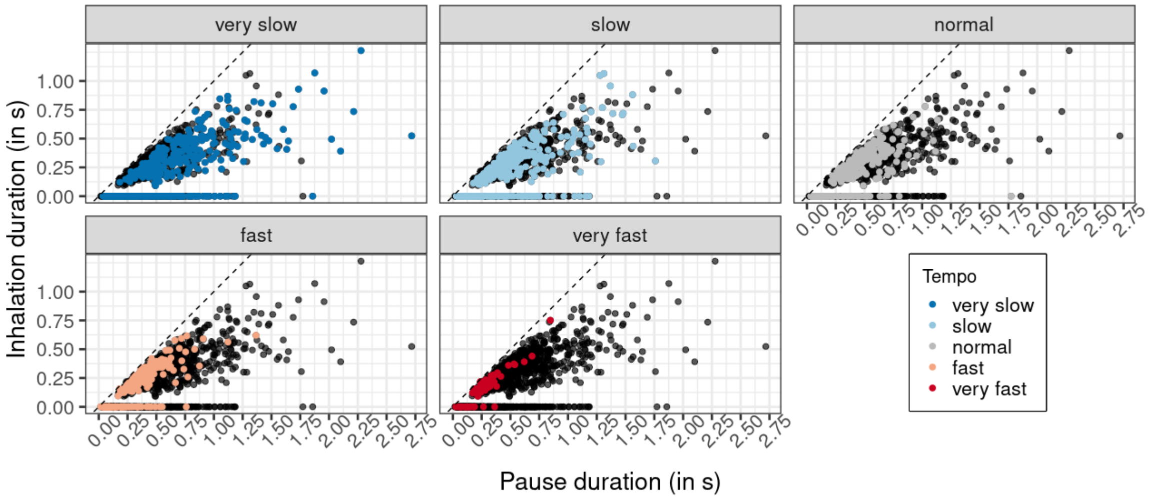

Figure 4.8: Inhalation duration relative to pause duration split by rate for all languages. The dashed reference line indicates where both durations would be equally long. Color indicates tempo, black dots represent all inhalations from the other tempos for reference. The dots on the x-axis (where y=0 s) are non-breath pauses.

faster rates, they tend to shorten their edges. They do so up to very short, and in extreme cases unidentifiably short, durations in the faster tempos. Moreover, there seems to be a difference between right and left edge duration: while right edges rarely exceed durations of 300 ms, left edges remain relatively frequent up to around 600 ms.

4.2.4 Discussion

The findings concerning pause locations (Fig. 4.7) showed that pausing and punctuation are related in read speech as readers can use them as landmarks for pausing (Godde et al., 2021). While there are many other locations for pauses, those are rather used at normal and slower rates. As in the studies on read speech by Conrad et al. (1983) and Winkworth et al. (1994), the inhalations here tended to occur at sentence, clause, or paragraph boundaries and were also deeper and longer when taken there. The less uniform pausing pattern in French is likely to be related to the different usage of commas: the text contained commas after place names (e.g. "après Lisieux, les montagnes") resulting in a high number of commas, which was not the case in the other BTC languages, for example English ("after Lincoln the hills").

At the seemingly less optional locations, pauses show high variability that seems related to speech rate. Even though pauses do often coincide with punctuation and conjunctions like *and*, simply using those as a trigger is not sufficient but needs to take into account the larger syntactic structure, as exemplified by the three cases of *and* in the second sentence of the English version (location in brackets for comparison

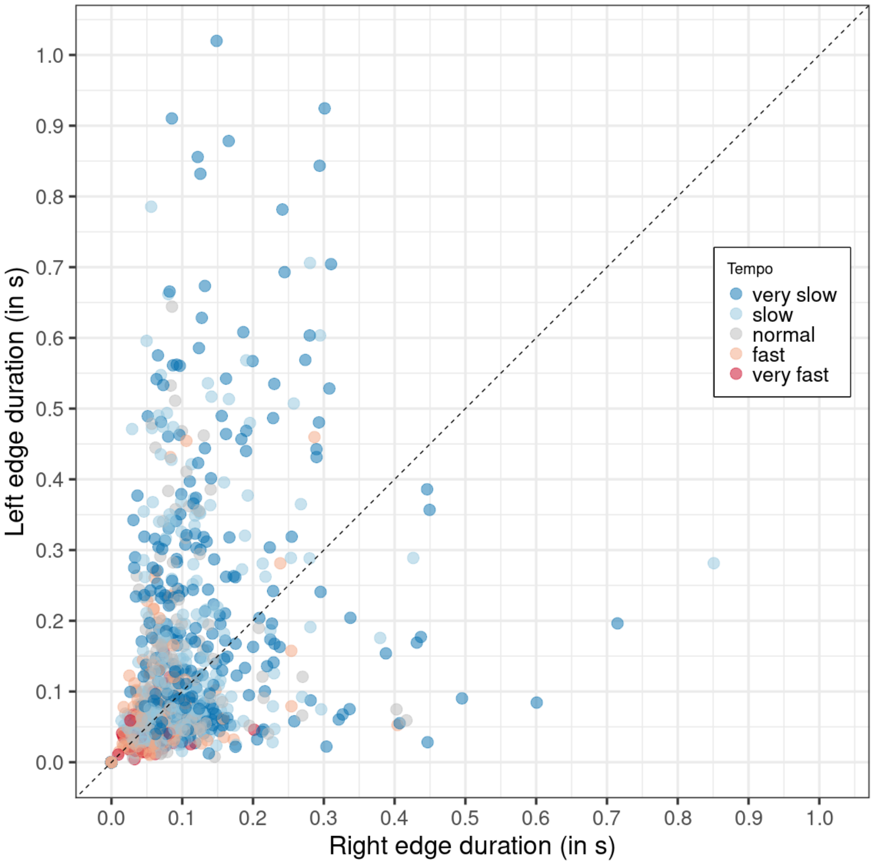

Figure 4.9: Duration of left vs. right edge in breath pauses in all six languages pooled together. The dashed reference line indicates equal durations. Five breath pauses (all from the same speaker) with left edges between 1.3 s and 2.1 s. were excluded from the plot.

with Fig. 4.7): "after Lincoln the hills (21) and woods become monotonous, after Bristol the towns get boring (31) and near Saintsbury the countryside becomes flat (38) and desolate." This also becomes clear in the similarity of the fourth bigger pause line in Czech and English, which have no punctuation there, to the same location in German, which does have punctuation. The closer look at the last three bigger pause accumulations in Czech further illustrated that punctuation alone is not sufficient for pause modeling. Incorporating syntactical (cf. Truckenbrodt 2007) or prosodic structures (cf. Ferreira 1993) could be beneficial, especially for cross-linguistic comparisons.

When comparing inhalation duration in relation to pause duration (Fig. 4.8) to previous results in Trouvain et al. (2020), the non-professional German speakers there seemed to behave quite similarly to the ones analyzed here, although the text used there was longer. The few long pauses (breath and non-breath) that appear along with some shorter pauses at the fast and very fast rate might hint at different strategies for increasing rate. This should become clearer in longer texts, as the one used here was short and many speakers attempted to produce *very fast* without any pausing. Conversely, it should be borne in mind that participants' interpretations of intended speech rates may vary: while the faster rates may be more naturally limited by how fast a given speaker can produce speech, the slower ones may be less uniform and may thus lead to very long pauses in extreme cases.

Comparing the inhalation edges (Fig. 4.9) to previous findings in ch. 5.1, the left edge tendency is not there, but those results were preliminary and the speech analyzed there was semi-spontaneous without intended rate variations as conditions, which limits comparability. Left edges tending to be longer in the present study could be related to the breathing apparatus' elastic recoil, exerting passive force on the lungs to diminish in size after inhalation (e.g. Hixon et al. 2020, p.13–14). With the inhalations, they increase their subglottal pressure, which makes it harder to employ pauses and further delay speech and exhalation onset. Thus, when speakers use relatively long pauses (in normal and slower conditions), they tend to increase inhalation and left edge duration, while the right edge remains rather short. Along with these physiological reasons, this may also have functional reasons, as audible inhalations may also be used as a cue for upcoming speech (Włodarczak & Heldner, 2016). Both aspects may of course go hand in hand in speech.

A similar picture was reported by (MacIntyre, 2022, p. 137), who found a metric similar to left edge (acoustic landmark to breath onset, i.e. preceding vowel onset to breath onset) to be longer than the metric similar to right edge (breath offset to landmark, i.e. breath offset to onset of following vowel) across four different speech styles including counting, poems, prose, and spontaneous. This metric, however, is slightly different and the equivalent to the left edge there always includes the vowel and is thus likely to be longer anyways. Similarly, Bailly & Gouvernayre (2012) noted that breath noises typically ended close to phonation onset of the upcoming utterance, i.e. right edges were short, but the distribution of left edges was multimodal. The delay between end of the previous phrase and onset of the breath noise was found to be very short for sentence-internal pauses, but large for sentence-final pauses and end-of-paragraph pauses. They noted that the size of the delays are an important cue for signaling (dis)continuity of the theme at hand and may have more influence than the type of pause involved. The results are also in line with Grosjean & Collins (1979), as they noted preinspiration, which is conceptually similar to left edges, to be much longer than postinspiration, which is similar to right edges. At faster tempos, preinspiration was found to be shortened much more, making them similar in length there.

A similarity to the previous findings is that edges seem to be of the same length only when short: When short, i.e. up to around 150 ms, they look relatively symmetrical but as one edge gets longer, the other one tends to remain short in comparison. Pause modeling needs to take these differences into account, as listeners are susceptible to short gaps around breath noises and even more so for gaps following vs. preceding inhalations (MacIntyre & Scott, 2022).

In general, it should be borne in mind that the text used here is short and that participants' interpretations of intended speech rates may vary. Furthermore, adjusting speech rate towards the fast extreme may be more straightforward and thus more uniform than when approximating the slow extreme.

Another point to consider when comparing different languages and their pausing behavior is how the pauses interact with the speech produced. Coupé et al. (2019) investigated how different languages encode information. They found that the 17 languages they considered showed quite some variation in the speech rate, i.e. syllables per second, while their information rate, i.e. bits per second by multiplying speech rate and information density, stayed fairly constant across languages. In their study, however, they discarded any pauses longer than 150 ms. Studying how all pauses interact with speech and information rates across languages may be an interesting topic for the future.

4.2.5 Conclusion

This study has shed light on the optionality and variability of prosodic boundaries and pauses and how these may interact. Reading aloud the same text at different rates, speakers show preferences for certain pauses that often coincide with punctuation. These preferred pauses generally involve a highly variable pause duration that varies with speech rate. Purely punctuation-based pause modeling misses this variability and while it may be able to reproduce pause location selection behavior for the majority of pauses it would not account for other pauses that are unrelated to punctuation or do have punctuation but not necessarily a pause. Differences in punctuation complicate comparability over languages, as in some languages readers are not very constrained by punctuation while in others orthographic breaks are indicated at virtually every prosodic break. For authentic pause modeling across languages and beyond read speech, punctuation-independent linguistic perspectives are needed. Furthermore, the findings on the duration of pauses, inhalations, and edges are important for modeling pauses and the placement of audible breaths within them.

These findings have implications for modeling pauses in natural and synthetic speech, especially for situations or target audiences where slower speech is more appropriate. So far, pausing in synthetic speech differs from natural speech by using relatively short pauses (and only non-breath pauses) with a slow articulation rate, mostly triggered by punctuation (cf. ch. 4.1). A more human-like pausing pattern would need a better integration of the optionality of pause locations, more variability of pause duration, and a consideration of breath noises, which in turn could benefit listeners by improving recollection of content (Elmers et al., 2021a,b).

Chapter 5

Breath noise acoustics and physiology

5.1 Exploring the presence and absence of inhalation noises when speaking and when listening

5.1.1 Introduction

Breath noises in speech communication can usually be observed in the proximity of articulatory activity but not when speakers let their articulation rest, for instance when they are listening (Trouvain et al., 2020). The question is why inhalation is made audible only in active (or planned) articulation in comparison to listening, even though respiration is permanently at work. This question can only be addressed by an analysis of acoustic and physiological (respiratory) data. Thus, the aim of this exploratory study is to look more closely at breathing and the interplay between synchronously recorded acoustic and kinematic respiratory signals. For this, we aim to investigate the temporal properties of breathing, especially inhalation noise, using acoustic and respiratory data. In particular, we compare general aspects of breathing when listening and when speaking. For speech breathing, we further examine the interplay between inhalation noises and articulation and how they are temporally aligned with expansions of the abdomen and the rib cage.

For this, we also looked at *edges*, i.e. short stretches of silence that surround the breath noise and together with it form a breath pause. We aimed to find a link between the timing of the edges in the acoustic signal and the respiratory activities of RC and/or AB in the inhalation phase in breath pauses in speech.

In a first exploratory step, we had a closer look at 4 speakers and their respiratory behavior, observed a pattern that can be seen in Fig. 5.1 and schematized the findings in Fig. 5.2. In the final part, we tried to test this general observation with speech data featuring synchronously recorded acoustic and kinematic (respiratory) signals from 14 speakers (including the other 4).

5.1.2 Methods

This exploratory study builds on German material originally elicited for a different study (Rochet-Capellan & Fuchs, 2013a). All participants analyzed here were female and engaged in two tasks, listening to a fable (LN) and re-telling this fable (SN). Along with the audio, kinematic data were also collected via Respiratory Inductance Plethysmography (RIP): one elastic band was placed at the level of the axilla to measure movement of the rib cage (RC) and the other one at the level of the umbilicus to measure movement of the abdomen (AB). By that, compression and expansion of AB and RC during in- and exhalation can be monitored. Participants were told to stand still during the experiments to avoid the RIP signal picking up arm movement.

In post-processing, the RIP signal was transformed into on- and offsets of in- and exhalations resulting in time-aligned annotation for these events split into phases of inhalation and exhalation. Inhalation onsets were detected automatically at 10% of the velocity peak, while offsets were detected at 90%. The remaining stretch from the end of inhalation to the start of the next inhalation is regarded as exhalation and thus potentially includes phases of breath holding.

We further annotated audible inhalation noises from the speech signal to relate them to the on- and offsets of articulation, AB, and RC. It is important to note that by articulation here we refer to the acoustic result of speech production, which is therefore based solely on the audio signal. Since it is not possible to reliably annotate breath noises in LN (due to less loud breathing and masking from the fable being played), analysis of audible breath noises is only done in SN, whereas the analysis of LN is fully based on kinematic data.

The dataset used here includes 108 corresponding files (54 LN, 54 SN) from 14 participants. Analyses are based on the duration and coordination of these intervals as annotated in Praat TextGrids (Boersma & Weenink, 2019).

5.1.3 Preliminary analysis

Inhalation in listening vs. in speaking A first comparison between the kinematic respiration patterns of both tasks (for an illustration of a sample see Fig. 5.3) reveals that subjects have shorter and more variable breath cycles (inhalation phase plus following exhalation phase) in LN than in SN. Consequently, there are more breath cycles per minute in LN than in SN (see Table 5.1).

In addition, the duration of the inhalation phase is substantially longer in LN (as opposed to SN) while the exhalation phase in SN is much longer and more variable in its duration. This can be seen in Table 5.1 where the respiratory time quotient (RTQ: duration of inhalation divided by duration of exhalation), as proposed by Conrad

Table 5.1: *Overview of the respiratory time quotient (RTQ) for abdomen (AB) and rib cage (RC), mean durations of breath cycles, and number of mean breath cycles per minute while speaking and listening.*

	speaking	**listening**
RTQ AB: mean (sd)	0.18 (0.41)	0.59 (0.90)
duration of AB breath cycle	5.45 (2.79)	3.68 (1.37)
AB breath cycles per minute	11.0	16.3
RTQ RC: mean (sd)	0.16 (0.13)	0.61 (0.72)
duration of RC breath cycle	5.26 (2.47)	3.67 (1.37)
RC breath cycles per minute	11.4	16.3

& Schönle (1979), is much lower in SN than in LN, reflecting the less symmetric respiratory behavior there. The high standard deviation of AB in speaking is partly caused by two RTQs that are around 1.2, created by a very short speaking phase between 2 exhalations. Thus, in both conditions exhalations are generally longer than inhalations but in speech breathing, the imbalance is much stronger. The RTQs are largely in agreement with those usually reported for breathing during speech (Conrad & Schönle, 1979). Fuchs & Rochet-Capellan (2021) reported I-fractions of 0.1 for speech and those 0.4–0.6 at rest. I-fractions are slightly different from RTQs, as RTQs are calculated as duration of inhalation divided by duration of exhalation. I-fractions are computed as duration of inhalation divided by duration of inhalation+exhalation, i.e. the entire breath cycle. As a consequence, numbers for RTQs are systematically higher than those for I-fractions. Additionally, the duration of a breath cycle is shorter in listening.

As expected, only few breath noises in LN were observable (in the phases before and after the playback of the fable to be listened to; during the playback an observation was not possible). The few instances of breath noises in LN were very soft compared to those in SN. In addition, in SN all inhalation phases were acoustically reflected by a salient breath noise.

Timing of AB and RC activities Interestingly, several speakers shifted their AB phase in relation to RC to an earlier time point when speaking (as compared to listening) as illustrated schematically in Fig. 5.2.

When inspecting the temporal structure of the acoustic signal (articulation phases and inhalation noises) together with the kinematic signal (RC and AB) the following pattern was observed for two of the four speakers (as illustrated in Fig. 5.1): The end of articulation in the acoustic signal [1] seems to be aligned with the start of AB [7], whereas the start of an articulation phase [2] seems to be aligned with the end of RC [6]. In contrast, the start of the inhalation noise [3] seems to be synchronous with the start of RC [5] whereas the end of the inhalation noise [4] and the end of AB [8] seem to be synchronized.

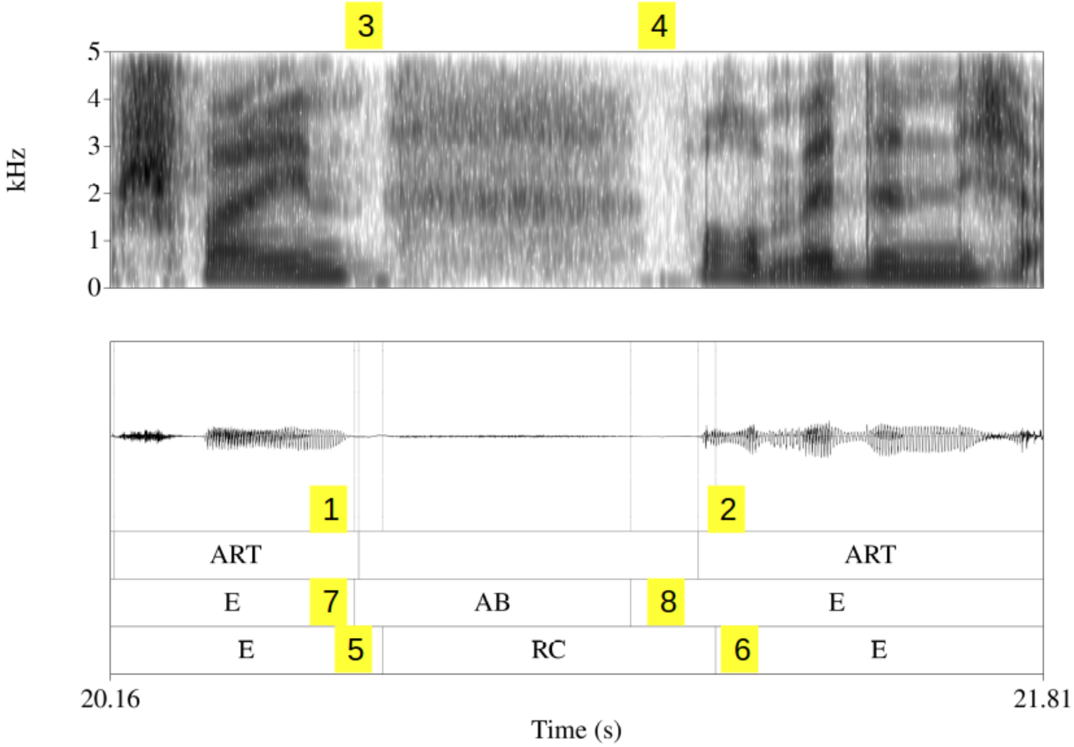

Figure 5.1: The temporal alignment of articulation phases, inhalation noises and AB and RC in an inhalation phase between two articulation phases in a speech signal (left, for numbers see text, speaker S04).

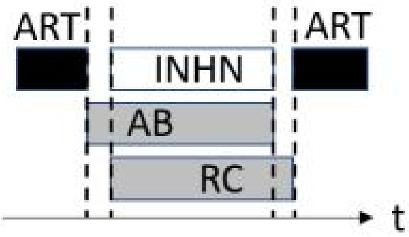

Figure 5.2: A schematic view of the temporal alignment of articulation phases (ART) and the abdominal (AB) and rib cage expansion (RC) in relation to an inhalation noise (INHN) in speech.

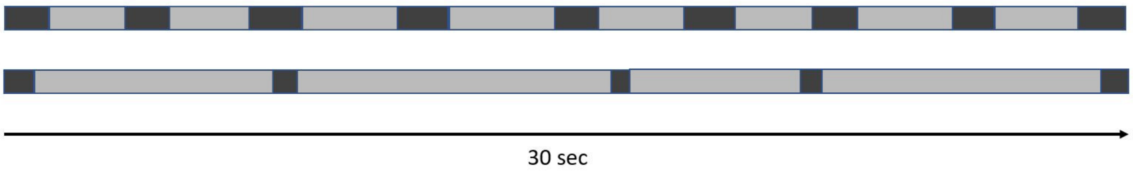

Figure 5.3: Exhalation (light grey) and inhalation phases (black) in two 30-sec excerpts from the inspected kinematic data of one speaker in both conditions (top: LN, bottom: SN).

5.1.4 Advanced analysis

To test the observed patterns described in the previous section, the following questions will be addressed:

(1) Are AB & RC of similar length?

(2) Does AB inhalation start earlier than RC inhalation?

(3) Does AB inhalation generally begin when the preceding articulation ends? Does RC inhalation generally end when the following articulation starts?

(4) Is inhalation only audible when both AB & RC are synchronously inhaling, i.e. from when the later contributor (presumably RC) starts to when the earlier contributor (presumably AB) ends?

(5) How do the durations of the edges left and right of inhalation noises relate to each other?

To test observations (1) to (4), for every inhalation noise in speaking we calculated the difference between the respective events. The values for LN are based on kinematics while listening. The results can be seen in Fig. 5.4 and 5.5 showing the difference of the respective subtraction.

Durations of AB and RC For observation (1) we used the duration values of AB and RC. In Fig. 5.4, Δdur shows that RC generally tends to be slightly longer than AB in both SN and LN. For LN, it is less clear with more variance (probably coming from longer total durations) and values closer to 0.

Start and end of AB inhalation relative to RC inhalation Δstart and Δend from Fig. 5.4 show results for (2): to see if AB was typically earlier than RC we compared the start and end times of both. While the start times show a difference that is close to 0 with a tendency towards RC being later for SN and the opposite for LN, the difference is clearer when looking at the end times: here RC is later than AB for SN but for LN the results are close to 0 with little variance. For SN, the difference between Δstart and Δend can be explained by Δdur that showed RC being longer. Further inspection of the SN data revealed that the pattern illustrated in Fig. 5.2 with start of AB first, RC second is valid for 6 out of the 14 speakers, whereas 3 show the exact opposite pattern. The remaining 5 speakers had no clear tendency. Fig. 5.2 also includes the pattern that AB ends before RC ends. 10 speakers follow this pattern. Among the 4 with the opposite behavior here were also the 3 'outliers' with the opposite starting pattern.

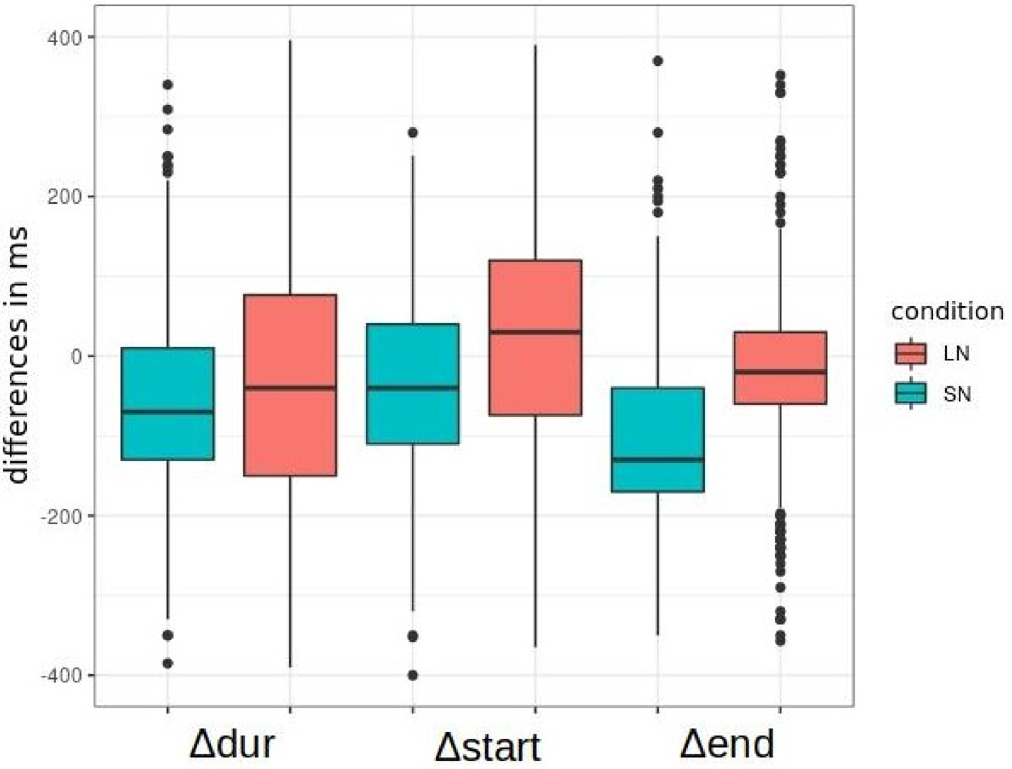

Figure 5.4: Differences between durations and start/end times of inhalation in the two conditions. Boxes 1 and 2 show differences between duration of AB minus duration of RC (Δdur), boxes 3 and 4 show differences between starting times of AB minus RC (Δstart), and boxes 5 and 6 show differences between end times of AB minus RC (Δend). Positive values thus show a longer/later AB, while negative values indicate a longer/later RC.

AB and RC inhalations relative to the articulation Observation (3) is addressed in the boxes 1 and 2 of Fig. 5.5: Concerning the start of AB and the end of the preceding articulation, the inhalation in AB starts a little later than the end of the articulation, leading to a short gap here. For the end of RC and the beginning of the preceding articulation, the difference is less clear with the box being at 0 but with a positive median, suggesting a slightly smaller gap than for end of articulation and start of AB.

AB and RC inhalations relative to the inhalation noise Observation (4) is about inhalations only being audible when both AB & RC are expanding. To test this we looked at two time points: First, the difference between the onset of audible inhalation and the begin of inhalation in AB or RC (whichever started later to ensure both were expanding; Fig. 5.5, box 3); second, the difference between the offset of audible inhalation and the end of AB or RC (whichever ended earlier; Fig. 5.5, box

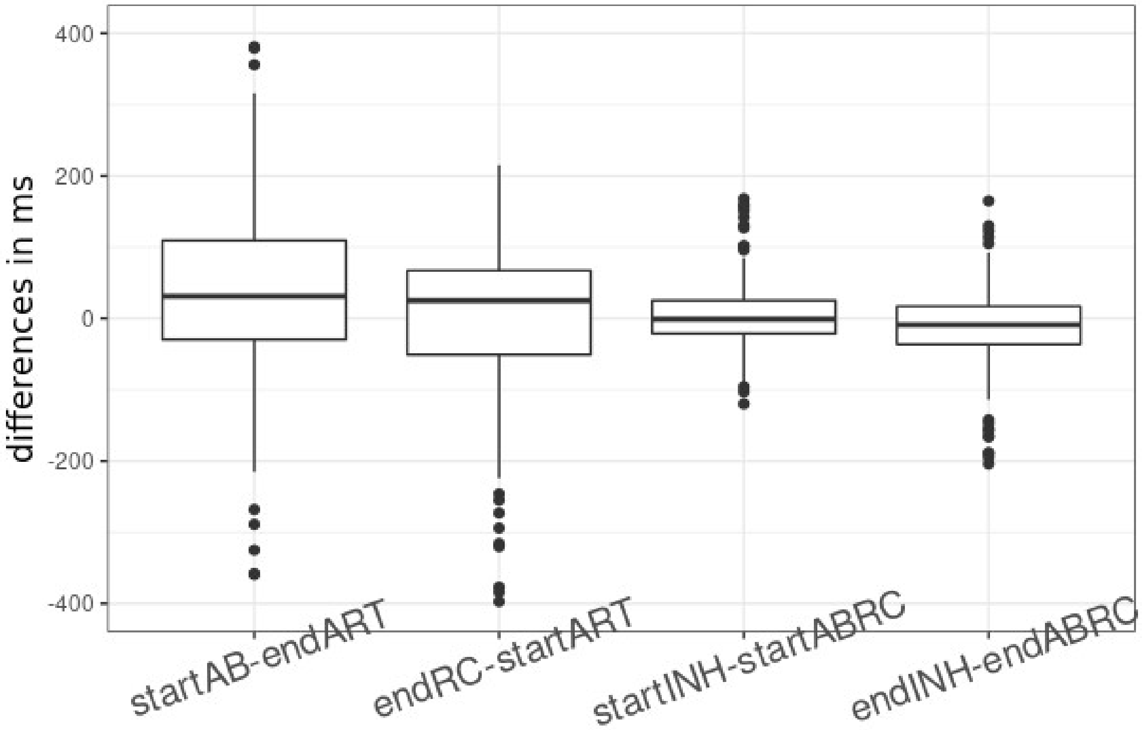

Figure 5.5: Differences between starting times (start) and end times (end) of inhalation noises, acoustic articulation, AB breathing, and RC breathing. In 'startABRC' the value is taken from the one starting later and in 'endABRC' from the one ending earlier to have both synchronously expanding.

4). The results of both subtractions are very close to 0 with little variance, suggesting that there is only a very small gap between those events happening. This suggests that there is a link between both AB and RC being synchronously active and an audible breath noise being produced even though AB and RC are displaced slightly, with RC being later than AB.

Timing of edges As concerns the edges surrounding a breath noise (5), it can be seen in Fig. 5.6 that they have a similar duration on both sides, with a slight tendency for longer edges following an inhalation. The mean duration for left edges is 116 ± 107 ms and 160 ± 164 ms for right edges.

Most inhalation noises (79%) are accompanied by edges that are shorter than 250 ms both left and right. Only 7% have one or two edges that are longer than 500 ms. There are hardly any combinations of both edges being long, meaning that an inhalation noise is typically not surrounded by two longer silent phases. For all except four cases, at least one edge always remains shorter than 250 ms. As a consequence, the inhalation noises here are only central when both edges are short – otherwise, one edge is longer.

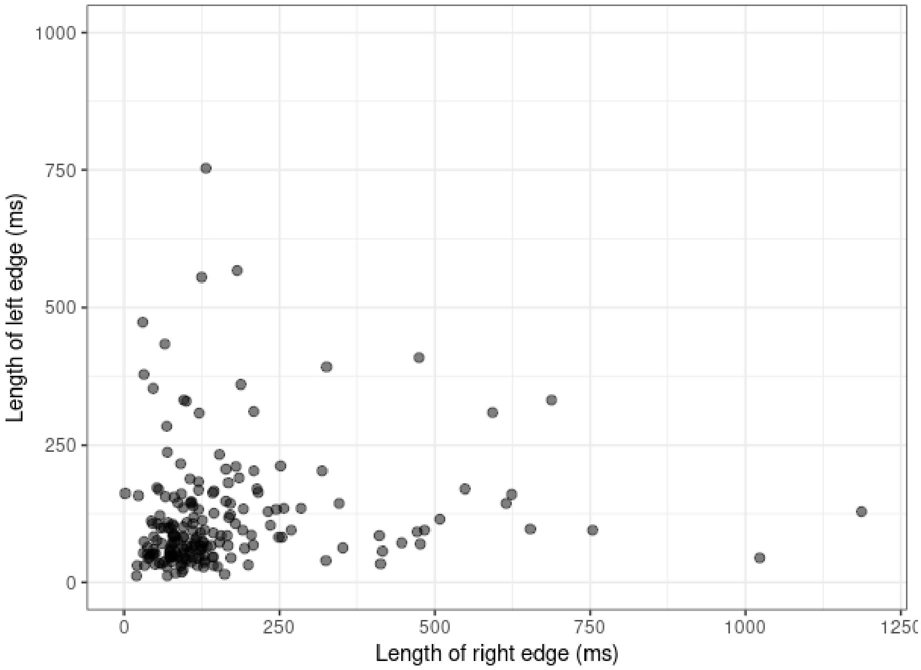

Figure 5.6: Corresponding sections of silence (edges) left and right of an audible inhalation noise.

5.1.5 Discussion

We worked with a coupled approach of observing patterns in the data and then trying to test them by looking at the respective times in the data. While this shows general tendencies, there are different strategies at work here, especially for the cooardination of AB and RC. This high degree of individuality was also observed for prephonatory chest wall posturing by Hixon et al. (1988).

When comparing the relation of breathing and articulation, we compared kinematic data for breathing and the speech signal for articulation. The gaps we found there might thus be due to a delay between articulatory and acoustic onset (cf. Rasskazova et al. 2019). As concerns acoustic and kinematic inhalation, it appears to be the case that the acoustic inhalation is closely coupled to breathing, happening synchronously at both abdomen and rib cage. This study cannot answer why that is and it should further be studied if that also applies to speakers with a clear preference for either AB or RC. The edges we found here are partly longer than the example numbers given in previous studies which is 20 or 40 ms (Ruinskiy & Lavner, 2007; Fukuda et al., 2018), with some of them here even exceeding a duration of 1 s. The reason for this is that we defined edges to be the time between preceding articulation and inhalation noise (left edge) and inhalation noise and following articulation (right edge). This thus includes potential hesitations that are not as clearly attributable to motor control reasons as edges of 20 ms length. However, since it is not clear where the boundary between an edge in a narrow sense and a hesitation following inhalation lies, we decided to include them.

In an attempt to analyze temporal aspects of these breath noises and the respective kinematic coordination, we used intervals for rib cage or abdomen expansion or compression. For this, we used the discretized, binary segmentation into either inhalation or exhalation. The segmentation was done automatically with 10% and 90% of the velocity peaks used as on- and offsets of inhalations, as described in Rochet-Capellan & Fuchs (2013a). In other words, this turned continuous signals, i.e. the breath kinematics, into discrete units. The annotation of breath noises, which did a similar thing to the audio signal, is not always completely clear either with regards to onset and offset of breath noises, especially in some speakers who tend to breathe more quietly. In addition, most differences in Fig. 5.4 were small with a lot of variance.

5.1.6 Conclusion

In summary, it has been shown that when retelling a fable as compared to listening to it, participants have fewer breath cycles which in turn are longer but also more variable in their duration. When listening, the ratio of duration of inhalation to duration of exhalation is about 6:10, whereas in speaking it is less than 2:10.

As expected, both articulatory phases and inhalation noises seem to be closely coupled to the activity of RC and AB, which often leads to short near-silent gaps around the inhalation noise. It appears to be the case that an audible inhalation noise is only generated when both AB & RC are expanding at the same time.

Finally, it has been shown that the edges left and right of the breath noise are generally short and have a similar duration (both edges are <250 ms in about 80% of the cases). When one of them is longer, the other typically remains relatively short, meaning that the inhalation noise in the speaking condition is only central when neither of the edges is long. This aspect should be investigated in a different experimental setting where the cognitive load is higher and/or the elicited speech is more spontaneous as opposed to pseudo-spontaneous data in our study. The question as to why edges of silence can be found on both sides of an audible inhalation noise remains open. It may be related to motor control when switching from exhaling (i.e. speech) to inhaling and vice versa. This should also be investigated in speakers who show a clear preference for either abdominal or thoracical breathing. The findings reported here are all based on younger, female participants who were standing during the experiment. It would be worthwhile to verify them by using a more diverse group of participants and a different experimental setup, as breathing movements can vary by age, sex, and posture (Kaneko & Horie, 2012). Furthermore, breath noises were only regarded as either present or absent in our study, but a closer look at their spectral properties may yield important insights.

5.2 Inhalations in speech: acoustic and physiological characteristics

5.2.1 Introduction

The aim of this study was to examine the acoustic characteristics of breath noises in speech by comparing them to similar speech segments. Although breath noises are so crucially related to speech, their spectral properties remain understudied. However, knowing these properties and particularly, how they may be similar to or different from speech, would improve word aligners, which may confuse breath noises with speech events and as a consequence reorganize the whole speech stream accordingly. Another goal of this study was to use these acoustic characteristics of breathing and try to correlate them to the corresponding kinematic properties. If inhalations in the audio and the physiological signals are found to be linearly related, it may be possible to draw certain conclusions about the underlying respiratory signals even without directly measuring them, as attempted in Mostaani et al. (2021).

As outlined in ch. 2.2.1, the breath cycle is reorganized in speech: Accordingly, the inhalation phases become short and rapid, resulting in the typically audible speech inhalations. As the reorganization affects the RIP signal, we assume differences on this level to also affect the resulting noises. Breath noises can be generated by any constriction in the vocal tract that may result from different coordinations of the respiratory system along with the supraglottal airways, the glottis, the velum, the lips, and the tongue (see ch. 2.2.4). We think that comparing breath noises to speech sounds may shed more light on where this may be.

We here assume that, in contrast to rest breathing which is usually done via the nasal cavity with a closed mouth, the larger part of the breath noises would be done via oral inhalation. Inhalation through the mouth is accompanied by the lowering of the jaw and the opening of the glottis (a closed glottis would not allow air to pass to the lungs and a closed mouth would only allow nasal breathing). While the jaw needs to be lowered, it is unclear whether the tongue simply rests on the jaw, as in the configurations for /a/ or /ə/ or whether it is affected by coarticulation with the preceding and following segments. Another crucial factor for audible noise is the coordination between inhalation and glottal aperture. In principle, noise can be generated due to a glottal constriction, similar to the production of /h/, but with ingressive instead of egressive airflow. At rest, the vocal tract is closed and the glottis is open, but in running speech the glottis can also be closed before inhalation, because it is surrounded by segments involving phonation. With a preceding voiced segment, the glottis needs to move from a closed to an open configuration for inhalation.

We will focus on three main questions here:

1. Do breath noises have similar spectral properties as /h/ and aspirations of stops?

If noise is generated at the glottal level, we may expect similarities with the voiceless glottal fricative. If noise is generated in the upper vocal tract, we may find similarities with aspiration noise in stops. We do not consider sibilants, because for the production of these sounds, the jaw needs to be in a high position so that the lower incisors can function as an obstacle source.

2. Does breath noise reveal similar formant structures as /ə/-realizations?

Initial inspections of the breath noise revealed a formant structure, even in the absence of phonation. This would support a vowel-like vocal tract configurations during the lowering of the jaw. We chose /ə/ because we do not assume any specific articulatory target for either /ə/ or breath noises.

3. Is there a relation between the speed of inhalation and certain acoustic parameters? Specifically, does speed of inhalation reveal a correlation with the first formant frequency (corresponding to the lowering of the jaw), center of gravity, and acoustic intensity?

We aim to investigate the relation between spectral properties and the physiological breathing signal since speech breathing is often related to the opening of the mouth.

5.2.2 Methodology

Material We used a subset of the material described in Rochet-Capellan & Fuchs (2013a) where five fables were retold in German by each subject. All 31 participants analyzed in the present study were female native speakers of German with a mean age of 25 years (age range: 21–32 years, normal body mass index).

For this study, we only included the material in which the participants were speaking (as opposed to listening) with a normal (as opposed to loud) volume. The files generally consisted of three phases: Speech phases with a mean duration of 41.2 s ± 12.3 s were preceded by pre-speech inactivity (7.6 s ± 2.6 s) and followed by post-speech inactivity (8.0 s ± 2.8 s).

The data include audio as well as kinematic data collected via Respiratory Inductance Plethysmography (RIP). Movement of the rib cage (RC) and the abdomen (AB) were measured by placing one elastic band at the level of the axilla and another one at the umbilicus. Compression and expansion of the AB and RC are monitored to infer in- and exhalations, respectively. To avoid the RIP signal picking up arm movement, participants were asked to stand still while recording. We used the sum of AB and 2×RC movements to get a more realistic representation of the lung volume (see Rochet-Capellan & Fuchs (2013a) for a detailed justification).

Due to recording conditions, the audio files were sampled with 11,025 Hz resulting in a frequency range from 0 to around 5,500 Hz. This, however, does not pose a major problem for the acoustic analysis as this range is sufficient for the segments inspected here, i.e. breath noises, glottal fricatives, /ə/, and aspiration phases of plosives Zellers & Schuppler (2020).

Annotation In this study, we only focus on audible inhalation noises; therefore, we hand-annotated them to separate them from the surrounding edges of silence (Ruinskiy & Lavner, 2007; Fukuda et al., 2018). The breath noises were categorized according to their position: the label *inh* was used for inhalation within speech, *inh-ini* when immediately preceding speech initiation, and *n-inh* when outside of speech phases, i.e. in articulatory inactivity. *inh* would thus appear in speech pauses, *inh-ini* when transitioning from rest to speech, and *n-inh* when participants were at rest, but still breathed audibly. We decided to separate *inh-ini* from *inh* as they might differ based on the involvement of inhalation noises in turn-taking (McFarland, 2001; Rochet-Capellan & Fuchs, 2014), which are typically louder than those in tidal breathing (Włodarczak & Heldner, 2016). Thus, there might be an intensity difference based on 'turn' taking (even though there is no real dialogue situation in the data here) or speech initiation (Scobbie et al., 2011), as speech planning is connected to both inhalation duration and depth (Fuchs et al., 2013).

As for speech segments, we annotated aspiration phases of fortis plosives (as *p-asp, t-asp, k-asp*, depending on place of articulation), voiceless glottal fricatives (/h/; voiced variants not included), and mid central vowels (/ə/; the more open /ɐ/ was not included). These were chosen because of their potential similarity to breath noises either due to the glottal opening in production (aspirations and glottal fricatives) or the neutral configuration of the vocal tract (/ə/). *p-asp* was removed due to a small number of data points. /ə/ and *n-inh* are only used in the section on formants. Overall, we found 690 instances of *inh*, 138 *inh-ini*, 101 *n-inh*, 259 /h/, 185 *k-asp*, 537 *t-asp*, and 675 /ə/.

The assessment of the airway used (nasal or oral or combined) was not part of the experiment when the material was collected. This categorization based on audio alone does not seem to be reliable (see ch. 6.1); for this reason a distinction between nasal and oral was not included here. For the time being, we adopt the findings of Lester & Hoit (2014) for a prevalence of around 90% or more depending on the task for simultaneous usage of nasal and oral airways (over nasal only, oral only, and alternating nasal and oral) as a working hypothesis for breath noises.

Even though intensity is highly sensitive to several factors (such as distance to microphone, acoustic conditions, ambient noise), it is included here. Preserving the same distance to the microphone was controlled for in the experimental setup (cf. Rochet-Capellan & Fuchs 2013a). All participants in this study were female, so biological sex as a potential factor is eliminated. This could, because of the typical height differences also lead to differences in lung size, which would then result in higher

Table 5.2: Mean duration and standard deviation of breathing events and speech segments (*h, k-asp, t-asp*) without /ə/ in ms.

Segment type	mean	sd
inh	408	150
inh-ini	535	241
speech segments	68	36

respiratory flow that could surface as louder breathing, as is the case in auscultation Oliveira & Marques (2014). To account for local speaker-dependent differences we normalized the intensity of the examined segments by subtracting the mean intensity of the segment from the mean intensity of the entire speech activity in the respective file. As a result of this normalization, the 'normalized intensity' will be lower for more intense and higher for less intense segments. We initially tried out another more local intensity normalization technique, by using 1 s of speech to the left and 1 s of speech to the right of the respective event, getting their mean and using that as a reference. Since the results from this measurement were similar to those of the technique we ended up using and this is far more time-consuming, we decided to use the trade-off variant.

Procedure The acoustic data were extracted from the audio signal using a Praat (Boersma & Weenink, 2019) script and the kinematic breathing signals using MATLAB (2017b). Acoustic parameters were taken as averages over the duration of the segment, except for the formants, where the parameters were extracted and averaged for the central third of the segment to control for potential coarticulation effects. Formant objects were created with maximum frequency set to 5,500 Hz and 5 formants. From the temporal segmentation, we obtained the corresponding lung volume (sum signal) at the onset and offset of the segment. From these temporal (x) and displacement (y) events, we calculated the respiratory slope for each segment, using the formula $slope = (y_2 - y_1)/(x_2 - x_1)$. Inhalation slope thus corresponds to the speed of inhalation.

All statistical results reported here come from linear mixed effects models calculated with the lme4 (Bates et al., 2015) package in R (RStudio Team, 2022). For the formant data, the Pillai score was calculated to measure vowel overlap (Hay et al., 2006; Nycz & Hall-Lew, 2013), with lower values indicating higher degrees of overlap.

5.2.3 Results

Duration and intensity Both types of inhalation are longer than the speech segments (Table 5.2). Inhalations right before speech (*inh-ini*) tend to be longer than those sandwiched between speech (*inh*).

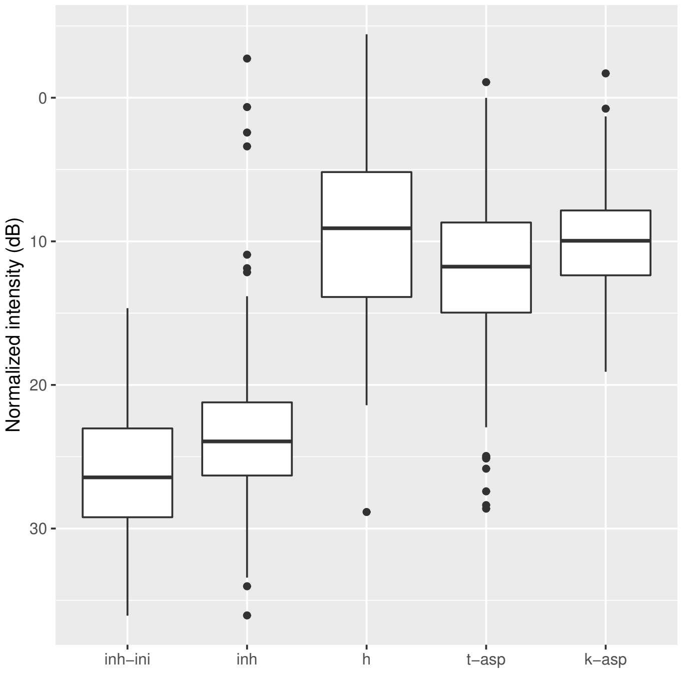

Figure 5.7: Normalized intensity (dB) of the respective segments. The y-axis is reversed here, since lower values represent higher intensity due to normalization.

For intensity (Fig. 5.7), there is a separation between breath segments and speech segments, with the latter being more intense. Within the breath noises, *inh* tends to be slightly louder than *inh-ini*. This is reflected in the model $lmer(norm\_int \sim segment + (1 + segment \mid speaker))$ using *inh* as the intercept (23.62, t=47.3, p<0.001), suggesting that *inh-ini* is less loud (2.46, t=5.39, p<0.001), whereas the speech segments are all louder: *h* (-14.18, t=-21.71, p<0.001), *t-asp* (-11.21, t=-18.63, p<0.001), *k-asp* (-13.28, t=-22.96, p<0.001). The difference between the two types of inhalation may be related to shorter durations, assuming that a similar amount of air is being inhaled.

Center of Gravity The differences in center of gravity (CoG) between the inhalation and speech segments can be seen in Fig. 5.8. We used the log-transformed CoG to account for linearity of the residuals in the model $lmer(logCoG \sim segment + (1 \mid speaker))$ with *inh* as intercept (7.23, t=451.08, p<0.001), showing an effect for *h* (-0.51, t=-20.94, p<0.001) and *t-asp* (0.28, t=14.27, p<0.001) but no significant effect for *inh-ini* (-0.04, t=-1.1, p=0.253) and *k-asp* (-0.05, t=-1.82, p=0.069).

Formants Fig. 5.9 shows a vowel chart plotting F1 and F2 values for inhalations surrounded by speech (*inh*) and in phases of articulatory inactivity (*n-inh*) compared to /ə/-realizations. While F2 tends to be less variable and generally higher in *inh*, i.e. slightly more front than /ə/, F1 separates them with *inh* involving a more open vocal

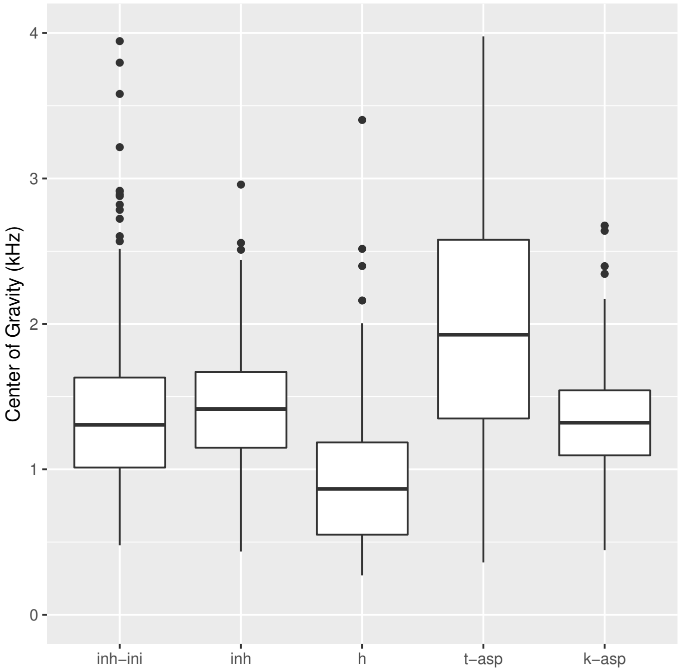

Figure 5.8: Center of Gravity of the respective segments in kHz.

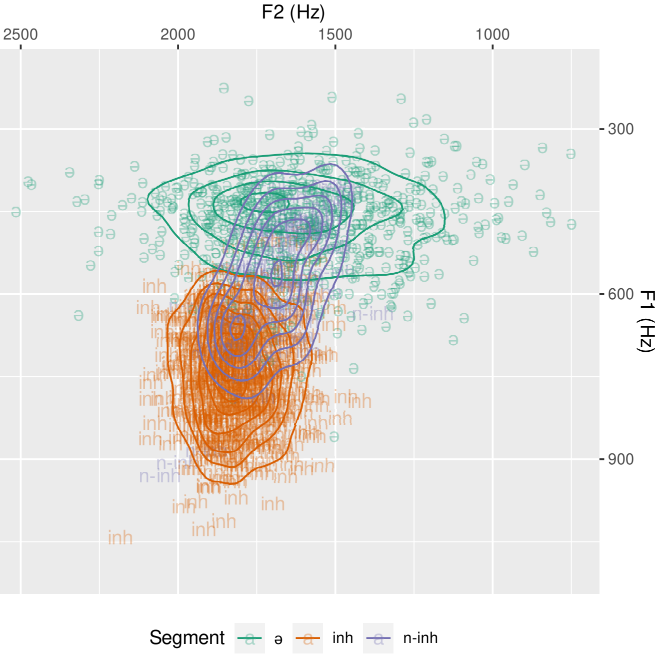

Figure 5.9: F1-F2 vowel chart for inhalations within speech (*inh*, orange) and outside of speech (*n-inh*, blue) compared to /ə/ (green). Ellipses refer to density.

tract than the neutral vowels. While there is virtually no density overlap for these two types of inhalation, inhalations outside of speech occupy a position that overlaps with both the *inh* and the /ə/ regions with a separate peak in each. To measure the overlap between these three segments, we calculated Pillai scores using F1, F2, and F3: There are similar degrees of overlap between both *inh* and *n-inh* (V=.28, F(1, 789)=102.44***) and *n-inh* and /ə/ (V=.21, F(1, 774)=70.03***). The degree of overlap is lowest between *inh* and /ə/ (V=.74, F(1, 1363)=1319.1***).

Relations between inhalation speed and acoustic properties We assume that deeper inhalations are related to faster inhalations since in articulatory kinematics, movement velocity and displacement are positively correlated (Ostry & Munhall, 1985). Taking a deep breath involves lowering the jaw to allow air to pass through the vocal tract without large obstructions. In the current data set, jaw motion was not obtained, but the first formant might approximate the degree of jaw opening (larger jaw opening for higher F1 values). To test the effect of inhalation speed on intensity, CoG, and F1, we subset the data to include only *inh* with the centralized inhalation slope as a continuous predictor. For each of those three parameters, a separate model was run with speaker-specific random slopes for inhalation slope as random effects on 30 speakers (n=667) without further normalization.

For intensity, the model $lmer(norm\_int \sim slope\_cen + (1 + slope\_cen \mid speaker))$ returns an intercept of 27.03 (t=26.53, p<0.001) and shows an effect on intensity for slope (-2.27, t=-4.58, p<0.001). The effect for faster inhalation or higher volume intake, which leads to more intense inhalation noise, is visualized in Fig. 5.10, left.

For CoG, the model $lmer(logCoG \sim slope\_cen + (1 + slope\_cen \mid speaker))$ outputs an intercept of 6.80 (t=198.9, p<0.001) and shows an effect for slope (0.30, t=10.6, p<0.001). Shorter inhalations or situations where larger amounts of air are inhaled in the same time thus lead to a higher center of gravity (Fig. 5.10, middle).

The model for inhalation slope and F1 $lmer(F1 \sim slope\_cen + (1 + slope\_cen \mid speaker))$ finds an intercept of 742.34 (t=60.64, p<0.001) and reveals a significant effect (34.73, t=6.16, p<0.001) for slope. Their relationship is visualized in Fig. 5.10, right. z-transformation was only carried out for visualizing the data, i.e. pooling all speakers together while making sure the correlation is not a by-product of individual differences in breathing slope or F1. The output is encouraging and suggests that a large part of the breathing noise is related to the motion of the jaw.

5.2.4 Discussion

The present study found intensity to be higher for inhalations within speech (*inh*) than for speech initiation (*inh-ini*). We assume this is caused by a steeper inhalation slope that was shown to increase intensity in *inh*. The CoG of *inh* is not significantly different from those of *inh-ini* or *k-asp*, which may have implications for modeling the 'place of articulation' for breath noises. Higher inhalation speed also increases

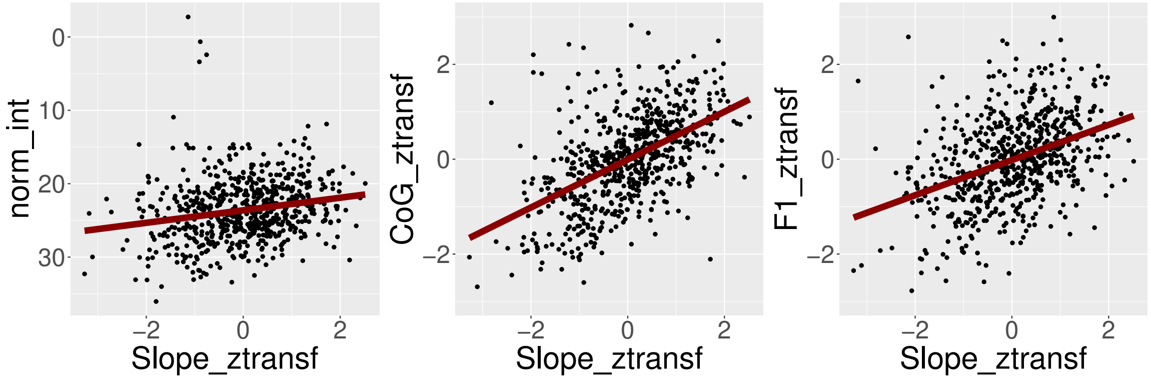

Figure 5.10: Correlations between inhalation slope and normalized intensity (norm_int, left), inhalation slope and CoG (middle), and inhalation slope and F1 (right) for all *inh* data. All speakers are pooled together. To account for different anatomical properties, data were z-transformed by speaker. norm_int was not transformed as it had already undergone a normalization procedure; its y-axis is reversed to visualize the underlying positive correlation (cf. Fig. 5.7 and ch. 5.2.2).

CoG. When looking at F1, F2, and F3, we find F1 to separate *inh* from a neutral configuration of the vocal tract as seen in /ə/. This difference may be linked to the wider mouth opening in inhalations. This idea is also supported by the positive correlation between inhalation slope (as a measure of inhalation speed) and F1.

We found a relationship between inhalation slope and the acoustic parameters of the inhalation noise, suggesting that degree of jaw lowering and inhalation speed are major contributors to the creation of these noises. For confirmation, these findings should be tested by including articulatory data (e.g. electromagnetic articulography or real-time magnetic resonance imaging) or modeled with the help of vocal tract models. Our findings held in speaker-wise examination and should thus also be applicable to male speakers whose breathing is expected to be similar (Hixon et al., 2020, p. 57).

A limitation of this study is the absence of the source component, i.e. vocal fold vibration, from the source-filter model since breath noises are typically voiceless. However, even being unable to derive a clear position of the tongue in the vocal tract through breathing noise, there is still a positive relationship between inhalation slope and F1 within a given speaker. This indicates that each speaker has a specific lung volume and vocal tract properties that are anatomically determined. Extracting formants seems to work in breath noises even though F2 is generally more prominent than F1. Finally, the results for CoG suggest that averages of F1 and F2 could be merged here. Future studies should try to use parameters other than CoG, e.g. coefficients of the Discrete Cosine Transform, as that may be a better fit to describe the spectra of breath noises.

Some inhalations contain spikes in the acoustic signal, typically in the beginning or somewhere in the middle part. These may be related to mouth opening, so it may be the case that they appear early in those breath noises starting with an open mouth and centrally for those that begin with nasal breathing and then switch over to mouth breathing. Articulatory data could help finding out whether these unintentional 'clicks' or percussives (Ogden, 2013) are related to mouth opening. One way of looking into those breath noises is by using the electro-optical palatograph as described in Birkholz et al. (2012) or Stone & Birkholz (2020). Like conventional electropalatography, it tracks tongue contact to a pseudopalate. However, it additionally has optical sensors in the same pseudo-palate that allow for recording distance between palate and tongue. This could give great insight into mouth opening in breath noises in general and also help finding the source of these spike within breath noises.

5.2.5 Conclusion

The present study found that inhalations within speech share spectral properties (CoG) with the aspiration phase of /k/-realizations and generally involve a more open vocal tract (higher F1) than /ə/-realizations. Intensity, CoG, and F1 were found to be positively correlated with inhalation speed.

These findings have implications for models of speech production in general, the automatic classification of breath noises, and the non-invasive detection of abnormal behavior for the diagnosis of diseases or disorders. Future work may benefit from extending the participant pool to also include males and older speakers. Breath noises should also be evaluated in tasks other than the pseudo-spontaneous setting used here. Another area of interest is coarticulation between breath noises and surrounding speech. Finally, categorizing breath noises based on airway usage may lead to different outcomes concerning the acoustic parameters.

5.3 Acoustics of breath noises in human speech: Descriptive and 3D-modeling approaches

5.3.1 Introduction

In this study, we focus on breath noises produced in speech pauses with regard to three aspects: First, we will provide general (average) spectral descriptions of inhalation noises that are produced by a large number of human speakers. These spectral properties cannot be easily interpreted, because the airflow direction in inhalation is ingressive, while most spectra reported in phonetics are realized with egressive airflow. Therefore, the second part of this paper is dedicated to a better understanding of how spectra of in- and exhalations differ when keeping everything else but airflow direction constant. This was done using 3D-printed vocal tract (VT) models. Finally, we compare human to model inhalations, aiming to approach the VT configuration of speakers when inhaling. Our ultimate goal is to estimate the potential VT contributions to the spectral characteristics of inhalation noise. Providing such a description of inhalation noise is not only important for the recognition of breath noise in speech corpora, it is also an important baseline for clinical applications.

Previous research has investigated how breathing interacts with speech and linguistic factors. These studies typically used parameters related to the changes in lung volume, e.g. via respiratory inductance plethysmography (RIP), placement of inhalation peaks in the spoken utterance, inhalation depth or breath cycle duration. While some research has been done on the role and behavior of breathing in speech, the supraglottal physiology and mechanisms, as well as the resulting breath noises have been largely neglected so far and thus remain under-researched.

5.3.1.1 Speech breathing: physiology and acoustics

As outlined in ch. 2.2.1, switching from tidal breathing to speech typically reorganizes the breathing cycle: at rest, in- and exhalations are similar in duration, with inhalations being only a little shorter than exhalations. When speaking, inhalations have a shorter duration and higher airflow velocity, while the exhalations, which are used for producing speech, become longer with a constant slow rate decrease in lung volume (Conrad & Schönle, 1979). At rest, inhalations take up 40–60% of the duration of one breathing cycle (one in- and exhalation), which in speech breathing is reduced to around 10% (Fuchs & Rochet-Capellan, 2021).

This reorganization also affects how air is inhaled: we assume that the pattern of nasal inhalation at rest changes to oral or simultaneous oral-nasal around speech (Lester & Hoit, 2014), so speakers are able to take in enough air for gas exchange and to power speech breathing in a relatively short time so as not to interrupt the speech stream for too long. Yet, not much is known about changes in the vocal tract other than the mouth *opening*. The general VT configuration in inhalations or more fine-grained aspects like the degree of mouth opening or the behavior of the tongue remain largely unknown, as articulatory studies on speech inhalation are limited.

Fig. 2.1 shows a typical example of an inhalation noise embedded in a speech pause, surrounded by short edges of silence around them (Ruinskiy & Lavner, 2007). It has a weak formant-like structure, as well as noise in the frequency range of up to 4–5 kHz. Some studies that have looked at the acoustics of breath noises can be found in ch. 2.2.4.

In breath noises, there may be a strong influence of subglottal resonances, as the degree of coupling between supra- and subglottal tracts depends on the glottal state (Lulich, 2010). A high degree of coupling can be expected as the glottis is opened very wide in inhalations (Scheinherr et al., 2015).

A highly relevant study was conducted by Hanna et al. (2018), who examined how the impedance spectrum measured through the lips is affected by the subglottal tract in varying states of glottal opening, i.e. fully closed, fully open, as well as intermediate states for phonation and respiration. They used impedance spectrometry to measure impedance minima, which correspond to resonances, and impedance maxima, which correspond to anti-resonances, in ten Australian participants that were instructed to keep their tongue in the position of the /ɜː/ vowel. In Australian English, this vowel can be expected to be close to the VT configuration of /ə/, with the former being slightly more open. Participants were also instructed to keep their velum raised. As the velum may be hard to raise volitionally, some participants are described to have pinched their noses to avoid nasal participation in Hanna et al. (2016). For inspiration in the seven male and three female participants, Hanna et al. (2018) reported impedance minima, i.e. resonances, which can be seen in Table 2.1. Their results showed that combining the supra- and subglottal tract, as is done in respiration, effectively doubled the length of the tract, which leads to up to twice as many resonances and anti-resonances compared to the closed glottis condition.

This study aims to have a close look at the spectral characteristics of breath noises in speech, as a comprehensive study of them is lacking and many experiments are based on a small number of participants.

5.3.1.2 Change of airflow direction: in- vs. exhalations

One specific property that has an effect on the spectral characteristics of breath noise and its interpretation, is airflow direction. The majority of speech sounds are produced with a pulmonic egressive air stream, while ingressive pulmonic phonation is far less prevalent (Eklund, 2008). This may be related to the fact that vocal fold anatomy is better suited for phonation with a pulmonic egressive rather than ingressive air stream (Catford, 1977, 67–68). Ingressive phonation does play an important role in non-verbal vocalizations, such as laughs, cries, and moans (Anikin & Reby, 2022). When vocalizing those, phonated inhalations enhanced the perceived level of emotional intensity compared to regular, unphonated inhalations. Virtually all inhalations encountered in speech pauses, however, are unvoiced.

Catford (2001, p. 18–21) described for the two voiceless fricatives [f] and [s] that they can also be produced with pulmonic suction, i.e. while inhaling (as expressed by the IPA symbol [↓]) rather than exhaling. The inhalation would then affect the sounds [f↓] and [s↓] differently: [f] and [f↓] would not be very different, as the channel through which air flows is not changed much as a function of direction, hence the turbulent airflow and the resulting sound are similar. For [s] and [s↓], however, the egressive version flows through the narrow gap between the tongue and alveolar ridge and leaves that channel in a high-velocity turbulent jet. This jet then hits the teeth, making it even more turbulent, which adds an additional high-frequency component to the sound of [s]. The ingressive version [s↓] involves comparably slow, non-turbulent airflow past the teeth before it reaches the narrow channel between the tongue and the alveolar ridge. There, it is accelerated and becomes turbulent. In this case, however, there is no obstacle present, like the teeth in the egressive version, so no added turbulence and thus the resulting sounds differ.

5.3.2 Inhalations in human speakers

5.3.2.1 Methods

Material We here used data from two sources: One is the Pool 2010 corpus (Jessen et al., 2005), from which we used semi-spontaneous speech produced by 100 male, native speakers of German (average age: 39 years; range: 21–63). Each speaker was recorded in two conditions: a Lombard condition, in which they heard white noise via headphones while speaking, and a 'normal', non-Lombard condition. The data used here were taken from the non-Lombard speech condition. For this task, speech was elicited in a setup similar to the game *Taboo*, i.e. speakers were asked to describe pictures to a conversation partner without using several terms. The recordings were made with a sampling rate of 16 kHz.

The other source comprises 34 female, native speakers of German (average age: 25 years, range: 20–33) from the data set described in Rochet-Capellan & Fuchs (2013a). The participants produced semi-spontaneous speech, each retelling five short stories (fables). The sampling rate in these recordings was 11,025 Hz. For the data set of female speakers, we here only used frequencies up to 4.5 kHz and thus downsampled it to a sampling rate of 9 kHz for two reasons: The audio files had shown a strong intensity decline in the higher frequency regions before and downsampling to a sampling rate divisible by 50 allowed us to have the same spectral resolution of 50 Hz across data sets for better comparability (see below for details).

We used on- and offset of noise in the audio signal to annotate inhalations in both data sets. This resulted in 1,892 inhalations produced by male speakers in the first source and 749 inhalations from female speakers in the second source. Both sources included stretches of participant inactivity in the audio files, for instance when the speaker was quiet between tasks. We only used inhalations occurring in speech pauses, i.e. those preceded and followed by speech, since we focus on speech inhalations and breathing at rest may differ. This also excludes breath noises that may be related to turn-taking. All 2,641 inhalations produced by human speakers were extracted via a Praat script (Boersma & Weenink, 2019) using rectangular windows. On average, the 134 speakers contributed a mean number of 19.7 ± 13.2 inhalations (range: 1–61). The mean duration of these breath noises was 467 ± 225 ms for the male speakers and 410 ± 147 ms for the female speakers.

Data analysis Although other studies have used formants to describe breath noises (Nakano et al., 2008; Werner et al., 2021a), here we used the averaged power spectral density (PSD) of the sounds, because formants are inherently difficult to determine in voiceless speech and breath noise and not uniquely defined due to the spectral zeroes introduced by the coupling with the subglottal system. In addition, formant values of unvoiced vowels, for instance in whispered speech, seem to constantly deviate from voiced vowels (Heeren, 2015), thus complicating interpretation. Using PSD also gives us the advantage of analyzing the entire spectrum, rather than extracting just single values. We obtained PSDs via `pwelch` in MATLAB, using a Hamming window. For the male speakers, recorded with a sampling rate of 16 kHz, we used a window length of 320 points with a 50% overlap between windows. This results in a spectral intensity value every 50 Hz from 0–8 kHz. For the female data, we followed the same procedure but adjusted the window length to 180. With the data downsampled to 9 kHz, we also have a spectral resolution of 50 Hz here to make the data sets comparable. We removed the measurement points at 0 and 50 Hz, as these low-frequency components tend to be problematic and differences there could arise from different recording setups.

After gathering all the spectra for all human speakers, we merged the two data sets and normalized their amplitudes together: In short, we obtained the sum of spectral components for every spectrum and divided it by the number of components, i.e. the number of spectral intensity values per spectrum. We chose the median of these sums as the arbitrary target sum, which we then subtracted from the sum of every individual spectrum to compute the difference between the target sum and the sum of an individual spectrum. As a final step, we subtracted this difference from every component. This allowed us to maintain the shape of each spectrum while shifting them in amplitude towards the same sum of components. Combining data from two sources and a larger number of speakers makes this even more necessary, as recording setups are likely to vary. For the normalization, it was necessary to have the same frequency range for male and female data, namely 100–4,500 Hz. The unnormalized inhalation spectra averaged by sex are shown in Fig. 5.11. It can be seen that for male speakers, the spectrum hardly has any peaks above 4,500 Hz. Therefore, we considered it reasonable to trade the information above 4.5 kHz for the sake of better comparability.

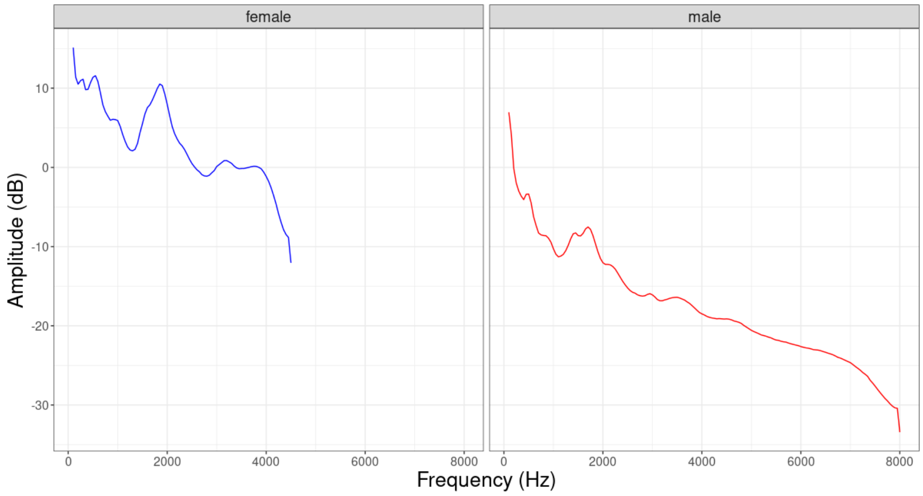

Figure 5.11: Unnormalized inhalation spectra averaged by sex for female (100–4,500 Hz) and male participants (100–8,000 Hz).

5.3.2.2 Results and discussion

The inhalation spectra averaged by speaker can be seen in Appendix B for female (Fig. B.1) and male participants (Fig. B.2). While there is some by-speaker variation, inhalations show some common patterns between speakers.

Fig. 5.12 shows all the human inhalation spectra as well as the average by sex. Both of them are relatively flat in comparison to speech sounds with a decreasing slope from higher intensity for low frequencies to lower intensity for higher frequencies. They do have several weak peaks: seven in the male and eight in the female data. The strongest of these can be seen below 2 kHz, it is highest at around 1.85 kHz for women and 1.7 kHz for men. In both cases, there is also a slightly weaker peak a little below, at around 1.6 kHz for females and 1.45 kHz for males. Both speaker groups also show a peak in the region of 500 Hz, highest for female speakers around 550 Hz and for male speakers around 500 Hz. In addition, both spectra have very weak peaks at around 1 kHz, 2.2 kHz, 3 kHz, and 3.5–4 kHz. Only the female speakers have an additional peak at 300 Hz which is absent from the male speakers. Moreover, there is not a lot of variation between individual inhalations, as within one sex they are all very similar. We suspect the differences in the spectra by sex to be caused by differences in average body height between the two speaker groups, as this is generally related to airway size. When normalizing for differences in these volumes and tract lengths, speech breathing is generally the same for men and women (Hixon et al., 2020, p. 57).

The spectra shown here are similar to those reported by Nakano et al. (2008), who found spectral peaks at about 1.6 kHz for males and 1.7 kHz for females, although both are slightly higher in our data. The secondary peak, which they found to be stronger for female than for male speakers at around 850–1,000 Hz, occurs at lower frequencies, i.e. around 500 Hz, in our data. Similar to their findings, it is slightly

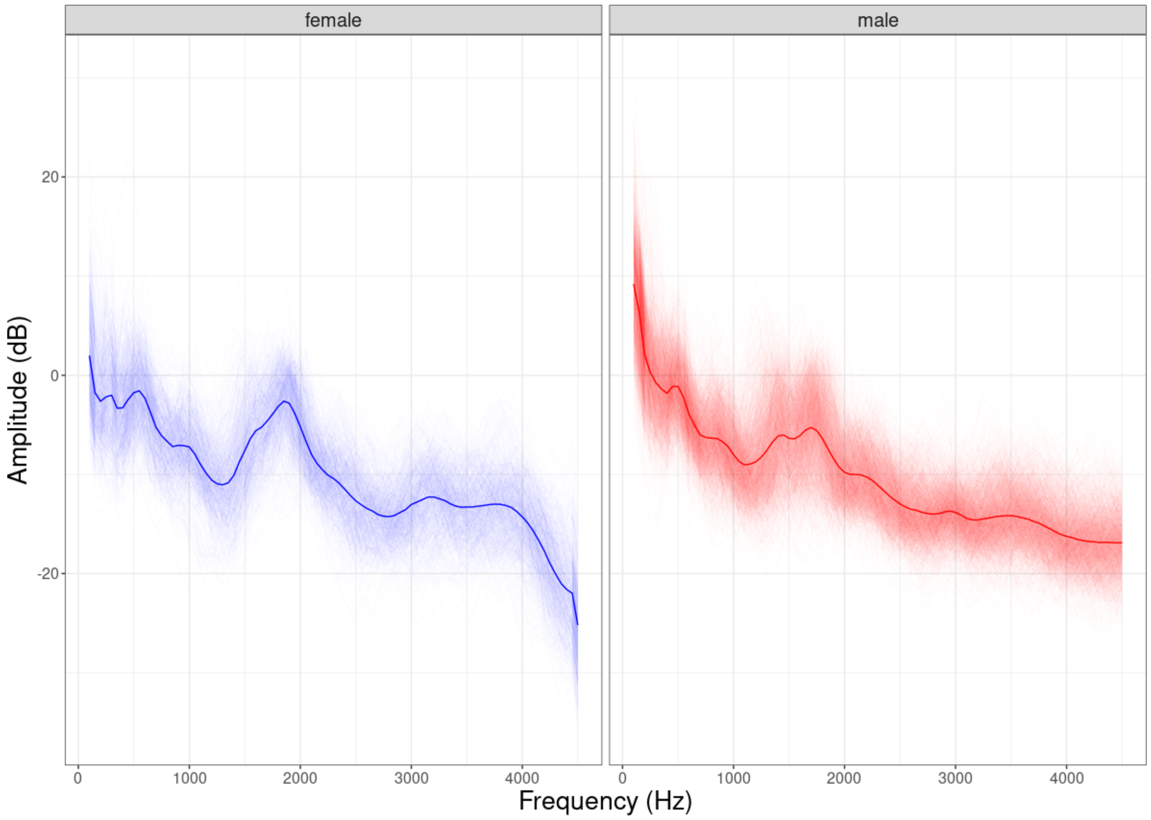

Figure 5.12: All human inhalation spectra for female (blue) and male (red) data. The average spectrum per sex is overlaid in bold.

stronger for female speakers.

Fig. 5.13 shows the average spectra by sex in relation to the resonances reported for male and female speakers inhaling with a VT configuration of /ɜː/ in Hanna et al. (2018). It is striking that for male speakers, all of the resonances they reported, except for the sixth, also align with peaks in our data. For female speakers, the data do not overlap as clearly, but are still similar. The sixth resonance, as for male speakers, should also be higher according to our data. However, most of the shaded resonances are close to peaks in our spectra. The peak at 300 Hz, which only showed up in our female speakers, is absent from their findings. A possible reason for why the resonances align better for male speakers may be that both our study and Hanna et al. (2018) have more male speakers (100 and 7) than female speakers (34 and 3). In addition, our data were elicited under more natural conditions than theirs which were recorded in a very controlled setting, for which participants were asked to keep their velum, VT configuration, and glottal opening constant for several seconds. Our data are likely to include more articulatory movements, as well as some inhalations happening partly or entirely through the nasal tract.

In sum, our human inhalation spectra suggest that the male speakers, and to some degree the female speakers, may inhale similarly to the inhalation condition in Hanna et al. (2018), i.e. with a central VT configuration and a widely opened glottis, coupling

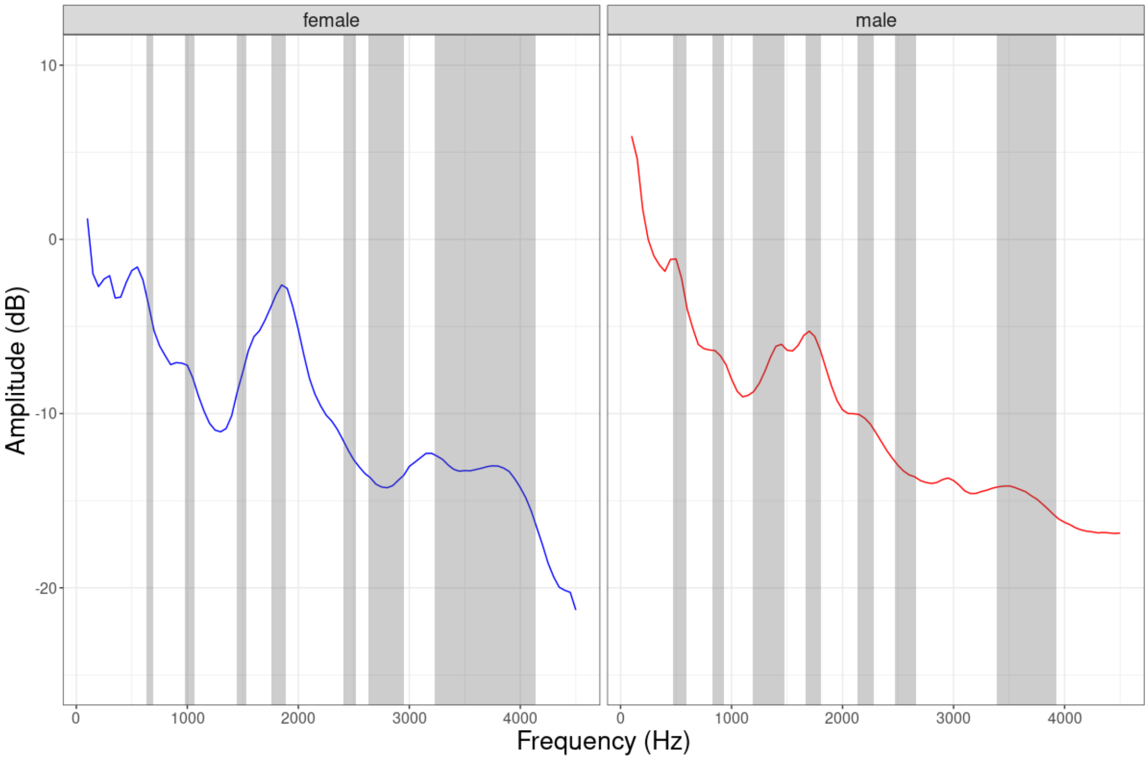

Figure 5.13: Averaged human inhalation spectra for female (blue) and male (red) data. The regions shaded in grey indicate resonance bands for inhalations with the VT configuration of the vowel /ɜː/, as reported for male speakers in Table I and female speakers in Table II by Hanna et al. (2018).

the supra- and subglottal tracts.

5.3.3 In- vs. exhalation in 3D-printed VT models

In this part, we focused on the effect that a reversion of the air stream direction, i.e. ingressive vs. egressive, has on the acoustic characteristics of the resulting noise. For this, we used 3D-printed VT models to study the effect of direction in the exact same VT configuration.

5.3.3.1 Methods

Material We used 3D-printed VT models (see Fig. 5.14), as described in more detail in Birkholz et al. (2020). For these, a male and a female native German speaker were asked to produce several sustained speech sounds, while capturing a volumetric MRI scan of their VTs. Maxilla and mandible shapes were included via plaster models. The nasal cavity is excluded from all models. The VT models were 3D-printed using polylactic acid, a commonly used filament material. This results in the VT walls being hard in comparison to a human VT's soft walls.

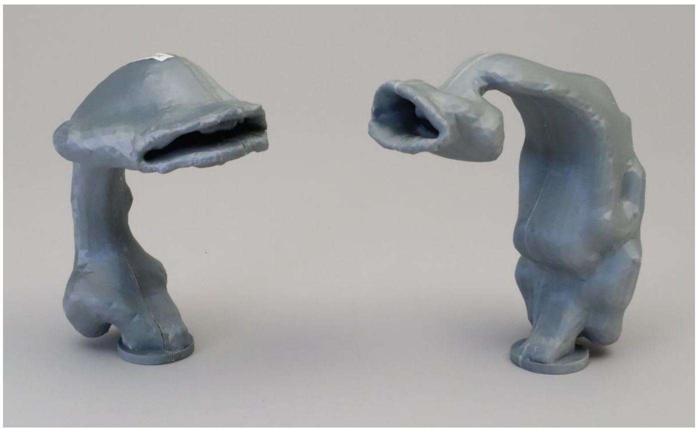

Figure 5.14: Two of the 3D-printed vocal tracts corresponding to a male speaker producing the sounds /aː/ (left) and /ʃ/ (right).

We here use a subset of these VT configurations, representing several sounds, namely four vowels /iː aː uː ə/ and four fricatives /x ç ʃ s/. We chose /ə/ because we assume a similar configuration for inhalations. The peripheral vowels /iː aː uː/ are used as reference, and the fricatives were chosen because breath noises typically have fricative-like acoustics (Székely et al., 2019). To imitate in- and exhalations, the VT models were supplied with static airflow through a constantly open glottis (diameter: 10 mm; glottal area: 78.54 mm^2) at three fluid power levels in two airflow directions. The power levels were 500 mW, 1000 mW, and 2000 mW and were chosen to roughly simulate quiet breathing, loud breathing, as well as an intermediate level. The open glottis was connected to the artificial lung via a polylactic acid trachea of 20 cm length (diameter: 17 mm) and a bronchial horn 7 cm in length. Below the glottis, the trachea is tapered from 17 mm to 10 mm over a length of 30 mm. The generated noises were recorded with a microphone with a sampling rate of 48 kHz for 10 s each. Overall, this results in 96 audio recordings of modeled breath noises (8 vocal tract configurations × 2 directions × 2 model speakers × 3 power levels).

Data analysis To obtain the spectra, we used similar methods as described in ch. 5.3.2.1. Given the sampling rate of 48 kHz, we set the window length in `pwelch` to 960 samples with a 50% overlap between windows, which resulted in a spectral resolution of 50 Hz. For the analysis, we chose to filter out all measurement points greater than 10 kHz. We also removed the measurement points at 0 and 50 Hz as in ch. 5.3.2.1.

We calculated the Discrete Cosine Transform (DCT) coefficients 0–3 to further characterize and compare the sound spectra via the RStudio (RStudio Team, 2022) package and function `emuR::DCT` (Winkelmann et al., 2021). This decomposes the signal into a set of half-cycle cosine waves and the resulting coefficients express how similar a given spectrum is to the respective cosine wave (Jannedy & Weirich, 2017). DCT0 corresponds to its mean amplitude and DCT1 to its slope, while the higher-

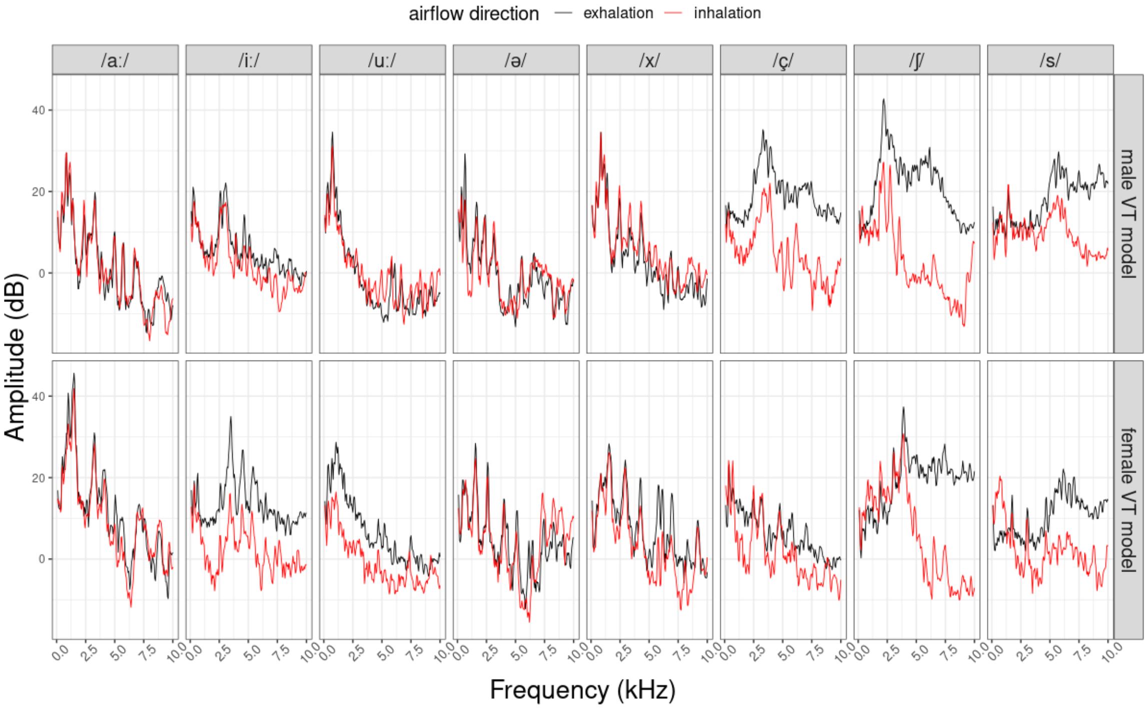

Figure 5.15: Spectra (100—10,000 Hz) for exhalation (black) and inhalation (red) by VT configuration and VT model (male and female). Every spectrum averages over three power levels.

order coefficients represent increasingly finer spectral detail. We then fitted a separate linear mixed effects model for each DCT coefficient with *direction* (2 levels: in- vs. exhalation) and *vocal tract configuration* (8 levels), as well as their interactions, as predictors. The models also included random intercepts for speaker and power level. We used `lme4` (Bates et al., 2015) for model fitting and `emmeans` (Lenth, 2021) for pairwise post-hoc comparisons between in- and exhalations for each configuration. All models had at least one significant interaction between direction and configuration except for the one for DCT3, so we used an additive model there. To avoid singular fit warnings, for DCT2 we used the linear model $lm(DCT2 \sim direction * VTconfig)$ without random effects. For DCT0 and DCT1 we used the following model formulae: $lmer(DCTi \sim direction * VTconfig + (1|speaker) + (1|condition))$, with i being 0 or 1. In the case of DCT3, $*$ was replaced by $+$ and i was 3.

5.3.3.2 Results and Discussion

Airflow direction: in- vs. exhalations in VT models The resulting spectra can be seen in Fig. 5.15. For every combination of speaker, VT configuration, and direction, the three power levels are averaged for better readability, as they mainly differ in amplitude. The plot shows that for some VT configurations, a change in airflow direction entails stronger spectral differences than for others.

Table 5.3: Significant contrasts between airflow direction by VT configuration in the VT models (exhalation - inhalation).

	VT configuration	Est.	SE	df	t-ratio	p-value
DCT0	/iː/	10.58	2.95	77	3.58	0.0444
	/ç/	13.50	2.95	77	4.58	0.0018
	/ʃ/	22.42	2.95	77	7.60	<0.0001
	/s/	11.95	2.95	77	4.05	0.0108
DCT1	/ʃ/	-12.44	1.12	77	-11.12	<0.0001
	/s/	-9.08	1.12	77	-8.11	<0.0001
DCT2	/ʃ/	-5.74	1.12	80	-5.12	0.0002

We found no general effect of reversing airflow direction on the spectrum in any of the four statistical models for DCT0–3. Instead, we found differences between in- and exhalation that were specific to some VT configurations. The statistical output for the pairwise post-hoc comparisons between directions that were significant (after Tukey adjustment) using an α level of 0.05 can be found in Table 5.3. DCT0 was significantly higher for /iː ç ʃ s/ in exhalation. DCT1 values were significantly higher in inhalation for /ʃ s/. For DCT2, the results from the linear model suggested that /ʃ/ was the only VT configuration that showed an effect of direction, as it was higher in inhalation. For DCT3, there were no significant interactions between direction and VT configuration, thus we do not have any significant contrasts between directions.

The direction differences were thus mostly found in sibilants, and /ʃ/ especially. For the mean amplitude, as expressed by DCT0, we found differences for four VT configurations, all of which featured a high tongue position. Here, we assumed that the tongue height led to a concentrated airstream hitting the incisors. While this obstacle source amplified the signal in exhalation, with a reversed airstream, there was no concentrated airstream hitting the incisors, which is why the signal was much weaker in inhalations. /s/, but also /ʃ iː ç/, thus followed Catford's (2001) prediction, as they showed high-frequency components in exhalation, which they did not have in inhalation.

It should be mentioned that the 3D-printed VTs are based on MRIs of a single male and female speaker, respectively. Therefore, we could not determine if differences between the two with regard to a change in airflow direction were based on sex or idiosyncratic differences. This can be seen to varying degrees for /iː uː ç ʃ/.

VT configuration: human vs. model inhalations We here compare human inhalations with model inhalations produced with an underlying /ə/-VT configuration. This builds on our previous findings on airflow direction in unconstricted VTs and on the similarity of human inhalation to experimental findings on inhalations with the VT of a central vowel. For this, we used the human inhalations from ch. 5.3.2

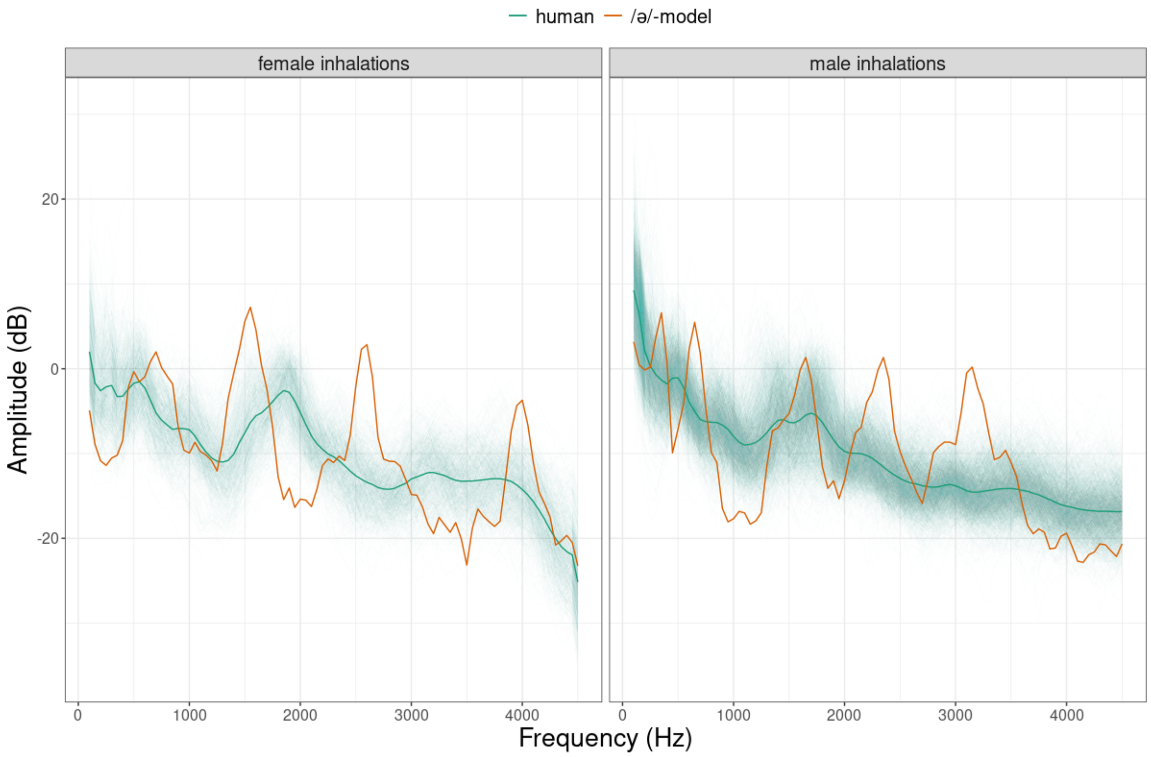

Figure 5.16: Human inhalation spectra (green; all spectra and average) vs. model inhalation with a /ə/ VT configuration (orange; averaged over three power levels each). The data are split by sex.

and the /ə/ inhalation part from the models, thus excluding exhalations. We only included data up to 4.5 kHz, as this was the maximum for the female human data. Afterwards, we removed the first two measurement points at 0 and 50 Hz and normalized the amplitude within each of the two data sets as described in ch. 5.3.2.1. Rather than using quantitative methods, we chose to employ a qualitative approach to compare spectra instead, as several differences between human and model VTs complicate the comparison and need to be taken into account.

The comparison between human inhalation and model inhalation with a /ə/ VT can be seen in Fig. 5.16. In the /ə/ models, there are stronger but also fewer peaks compared to human inhalations, with four major peaks for the female model and five for the male model in the frequency range up to 4.5 kHz. There are some similarities regarding their locations. The two main peaks in human inhalations, i.e. around 500 Hz and below 2 kHz, have corresponding peaks in the respective model data. In the male data, this is visible for the higher of the two especially, whereas the model has two peaks below and above 500 Hz. In the female data, the two peaks are closer to each other, i.e. the lower one is shifted upwards in frequency while the higher one is close to 1.5 Hz. The human inhalations only have weaker peaks beyond 2 kHz, while for the model /ə/, the peaks remain almost as or equally strong in higher frequencies.

A possible reason for why model /ə/ and human inhalations align better in the male data may be, beside the difference in sample size of human inhalations, that the synthetic subglottal tract and glottis that were used in our setup were the same throughout the recordings and only the supraglottal VT was changed for male and female models. In human speakers, the length of the subglottal tract differs by sex, with 19.5 cm for male and 16.0 cm for female participants on average (Hanna et al., 2018). In human participants, the size of the glottal opening in breathing also differs by sex (Scheinherr et al., 2015).

In addition to that, there are several ways in which inhalations in the VT models differ from human inhalations, which might complicate the comparison based on acoustics and led to our decision of comparing them qualitatively. First of all, the VT models do not have a nasal tract. Following Lester & Hoit (2014), the majority of human speech inhalations we have in our data set may be simultaneously nasal and oral, which could affect the spectral properties by enlarging the VT's surface area and volume. In addition to that, there may of course also be purely nasal inhalations or alternations of oral and nasal airway usage. Another level of complexity not captured in the models is vertical larynx movement (Fink, 1974; Orlikoff et al., 1997). In human inhalation, the larynx moves downward, lowering the trachea, stretching out the laryngeal soft tissues and flattening the vocal folds against the side wall. The displacements are then reversed in exhalation. With the models being static, they do not account for this vertical larynx movement modifying the lengths of the supra- and subglottal cavities, which may affect the spectral properties of inhalations. The lack of movement in the models may also differ from human inhalation, where speakers are unlikely to hold the exact same configuration in their vocal tract throughout the inhalation, as opening and closing gestures of the mouth and the velopharyngeal port may fall into that phase, as well as potentially coarticulatory adaptation to surrounding gestures. Moreover, the VT models were printed using hard plastic, which is different from the soft, fleshy human VT. Although these differences are not substantial, using soft VT models may have led to higher resonance frequencies and bandwidths below 2 kHz (Birkholz et al., 2022), thus making the peaks less sharp. Finally, the VT models are based on the anatomy of a single speaker per sex. Even if these two speakers represented the respective average VT, we would still expect a substantial degree of variation between speakers.

5.3.4 General Discussion

In the first part of this paper, we characterized male and female human inhalations, which were quite similar within a sex group. Their spectra were relatively flat but showed several weaker peaks that aligned with resonances for inhalation with a VT setting of a central vowel reported by Hanna et al. (2018). In the second part of this paper we found that reversing airflow direction does not yield a general effect on the resulting audio signal. We did, however, find an effect on the mean amplitude of the spectrum for VT configurations with a high tongue position, as well as on the slope for both sibilants and DCT2 for /ʃ/. Following findings from the first two parts, we compared the spectra of human inhalations and model inhalations produced with a /ə/ VT configuration. There, we found that the human inhalation spectra show similarity to modeled /ə/ inhalations. However, many aspects of the physiology of speech breathing are under-researched, which makes it difficult to neatly tease all the factors apart that potentially contribute to the spectral properties.

Therefore, more articulatory studies with a focus on speech breathing are necessary to learn more about the VT that speakers assume in speech inhalations. There are numerous influences on the acoustics beyond the supraglottal configuration, including the subglottal tract, glottal opening, nasal participation, and larynx lowering. Since the velum and the larynx are difficult, or even impossible, to track via electromagnetic articulography, real-time magnetic resonance imaging is arguably the best method to do this. Modeling inhalations by means of 3D-printed VT models helps us learn more about how speech breathing is performed. However, there are several complexities and aspects of temporal coordination the models cannot capture.

In this study, we chose to use the power spectral densities of breath noises to investigate their spectral properties and thus decided against using parameters such as formants or center of gravity. The reasoning behind this was that, given the scarcity of studies on the spectral properties of breath noises in comparison to speech, we wanted to examine the spectrum as a whole, rather than extracting single values from it. In addition, we were not convinced that they would be a good fit for our research, as formants are difficult to determine in the absence of phonation. Center of gravity is most informative for those fricatives that have strong concentrations of energy in certain frequency regions, which was not the case for the inhalation spectra investigated here. With the description of inhalation spectra presented here, future studies could explore which other parameters work best to describe inhalation noises.

5.3.5 Conclusion

In this study, we examined the spectral properties of speech inhalations with the goals to describe human inhalations, to investigate the effect of reversing airflow direction in in- and exhalations in VT models, and to approach the VT configuration in human inhalations by comparing them to modeled inhalations produced using VT models whose underlying configurations we know. We found that human inhalations have relatively flat spectra with decreasing slope and several weak peaks. The peaks showed moderate (female data) to strong (male data) overlap with resonances found for participants inhaling with the VT configuration of a central vowel by Hanna et al. (2018) that arise due to coupling of subglottal and supraglottal tracts. For the VT models, results suggested that airflow direction has a segment-specific rather than a general effect on the acoustic properties, which affected especially the realizations of /ç/, /iː/, and sibilants. We further found similarities between human inhalations and modeled inhalations with the VT configuration of /ə/. However, several differences between the model and human inhalation, as well as aspects of the physiology of speech breathing that are yet to be researched, complicated the comparison.

Along with their general importance for speech science, a better understanding of the acoustics of breath noises has implications for a variety of areas. In speech technology, it may help make automatic speech segmentation and alignment get better at differentiating breathing from speech sounds and make breath noises in synthetic speech more natural. In clinical applications, it could contribute to improving the automatic detection of pathologies. If the sex difference we saw is mainly caused by height, the spectral properties of breath noises may aid in narrowing down the list of suspects in forensic applications. To test our acoustics-based findings, future studies should examine the supraglottal physiology of speech breathing. Ultimately, relating acoustic and kinematic aspects could help making inferences about a speaker's speech breathing behavior from audio signals.

Chapter 6

Breath noise perception

6.1 Perceptual categorization of breath noises in speech pauses

6.1.1 Introduction

Breath noises can occur in different shapes, as airflow direction (inhalation or exhalation) and airway usage (oral or nasal or sequential combinations thereof) can be altered. In this study, we wanted to test if those can be auditorily distinguished by listeners. To this end, we included six different breath types: oral (*ex:oral*) and nasal exhalation (*ex:nasal*), as well as inhalations that are oral – and possibly nasal at the same time – (*in:oral*), only nasal (*in:nasal*), oral followed by nasal (*in:oral+nasal*), and nasal followed by oral (*in:nasal+oral*). Ideally, we could have also identified simultaneous oral-nasal inhalations (Lester & Hoit, 2014), in which speakers open the velopharyngeal port to inhale through both the oral and nasal tract at the same time; however, with our methods this was not possible and so oral inhalations may include some degree of participation of the nasal tract.

In speech, oral inhalations are probably the most frequent (Kienast & Glitza, 2003). At rest, nasal inhalation is the default and frequent mouth breathing in children is associated with dental and craniofacial problems (Harari et al., 2010; Inada et al., 2021). Exhalations, which in speaking are typically used in combination with phonation to produce speech, may also occur without it (Kienast & Glitza, 2003). Breathing behavior in read and spontaneous speech may differ (Trouvain & Belz, 2019) and in conversation, breathing is also related to turn management (Wlodarczak & Heldner, 2020).

When these different types of breaths are analyzed in phonetic studies, it is usually done on the basis of audio data, as those are generally easily available and the most used in phonetics. Since often there are no complementary data that would help with the classification into breath noise types, such as video or articulatory data, annotators are restricted to using perceptual or automatic annotation methods. Correct perceptual categorization of breath noises by their type is thus important for studying speech respiration in general and in combination with speech preparation (Kienast & Glitza, 2003; Trouvain & Belz, 2019; Scobbie et al., 2011). In particular, correctly distinguishing between different breath noise categories could improve acoustic analyses of breath noises (see ch. 5), forensic use of paralinguistic material (Braun, 2020), or automatic breath detectors (Fukuda et al., 2018). Furthermore, it is useful for the detailed annotation of breath noises (Trouvain & Werner, 2022) and making synthetic speech sound more natural (Székely et al., 2020).

In this study, we addressed two main questions:

1. How well can listeners discriminate between different types of breath noises in an auditory perception task?

2. Do the factors phonetic knowledge (phoneticians vs. lay people), speech context (presence vs. absence of one second of speech before and after the breath noise), or the type of the breath noise (e.g. *in:oral, ex:nasal*, etc.) have an effect on correct categorization?

6.1.2 Experiment 1

This study was designed as a pilot to get an idea of the overall correct identification of breath noises and to see the influence of the factors phonetic knowledge, audio context and breath noise type on it.

6.1.2.1 Method

We annotated breath noises in 20 speakers (10m, 10f; aged 20 to 65) from a freely available Dutch audio-visual dialog corpus (van Son et al., 2008). Complementary to the audio signal, mouth opening in the video signal was used as a visible cue for oral contribution. Two experienced raters annotated a total of 812 breath noises reaching an inter-rater agreement of 92% (Cohen's $\kappa = .88$) on a 20% subset of the data; none of the ambiguous cases were used. We are aware that using the videos and inter-rater agreement do not lead to a perfect ground truth, but other methods such as face masks, EMA, or MRI could have an influence on breathing behavior or the audio quality of the material to be used for perception tasks (cf. Lester & Hoit 2014; Fuchs & Rochet-Capellan 2021).

These stimuli for the six most frequent breath types in our data were extracted from natural conversations and prepared in two conditions: with and without context, i.e. with/without including one second of speech before and after the breath noise. The no-context stimuli thus included only the respective breath noise without any silent stretches from the speech pause in which they are typically embedded (Ruinskiy

& Lavner, 2007; Fukuda et al., 2018). Conversely, the context stimuli included the breath noise in the middle and may, within the 2-second context span, also include speech by either or both interlocutors as well as silent stretches. From each type and condition, we selected four noises to present to participants in a web-based experiment via Labvanced (Finger et al., 2017). It should be mentioned that some breath types (*in:oral* & *in:nasal*) are much more frequent than others in natural data. Audible exhalations are quite rare in regular, fluent speech as opposed to speech under physical load where they become frequent (Trouvain & Truong, 2015). Our additional requirements of the breath noise being clear of any other noises (such as the interlocutor speaking) and no other breath noises occurring within the 2-second context span restricted our choice of suitable stimuli, especially in the exhalation categories.

In this experiment, every breath noise was used only once and thus the resulting 48 independent stimuli were presented to two groups of people: eight phoneticians and eight lay persons (none of whom spoke Dutch). Participants could listen to a given stimulus up to five times and then had to assign it to one of the six breath noise types.

6.1.2.2 Results

Due to the small number of participants, we focused on descriptive statistics only in this experiment. We found the assessment of breath noises to be correct in 73.6% of the cases. Individual participants' scores ranged from 56.3% to 83.3%. While context seemed to help (76.8% correct with context compared to 70.3% without it), there was no difference between phoneticians (74.0%) and lay persons (73.2%) and no tendency for an interaction between the two conditions context and phonetician. There were some stronger differences between the different types of breath noises (see Fig. 6.1, left): correct identification was highest in *in:nasal* (94.5%), while *in:nasal+oral* (75.0%), *in:oral* (72.7%), and *ex:nasal* (72.7%) were close to the overall mean and *in:oral+nasal* (67.2%) and *ex:oral* (59.4%) reached the lowest values.

6.1.2.3 Discussion

Surprisingly, potential differences between phoneticians and lay persons (in their audio equipment, phonetic knowledge, or being used to perception experiments) did not translate into differences in correct assessment of the stimuli. Context, which does seem to make a difference, may help on a smaller, e.g. nasal inhalations may be more frequently found adjacent to nasal sounds, or larger scale, e.g. audible exhalations may appear more often outside of fluent speech sections. However, the difference between the two levels of this factor is relatively small and the participant number is low so it should be tested with more subjects. Additionally, the stimuli for with/without context were different, independent breath noises so differences could also have been driven by individual items. As for the correctness by breath type, there is not a clear pattern emerging but *in:nasal* and *ex:oral*, which were the most and least correctly assessed breath types, were also the ones that were given as an answer the most and least in general, regardless of the stimulus heard.

6.1.3 Experiment 2

The main goal of Experiment 2 was to increase the number of participants to validate previous findings and to further examine the influence that context and breath type of the stimulus may have on correct categorization. To do that, we did not include the phonetic knowledge variable in Experiment 2. Additionally, we wanted to see if breath noise intensity and/or duration affected whether or not it was assessed correctly.

6.1.3.1 Method

The material and methodology were similar to Experiment 1 but with some modifications: To study the influence of context, stimuli in Experiment 2 were matched, i.e. each breath noise we used had one version with and one without context. Therefore, we created two different stimulus lists so that each subject was exposed to only 24 breath noises to avoid them encountering the same noise twice, both with and without context.[1] As we used stimuli from real conversations again, the stimuli (see Table 6.1) show some differences in their duration and intensity.[2]

Table 6.1: Overview of the stimuli used in Experiment 2, regarding their duration (dur.) and intensity (int.). We used 4 breath noises for each type.

Breath type	**Mean dur. (& range) in ms**	**Mean int. (& range) in dB**
in:oral	559 (408–1,050)	43.6 (38.7–47.7)
in:nasal	446 (305–526)	42.5 (31.0–50.7)
in:oral+nasal	661 (443–812)	44.1 (35.0–47.5)
in:nasal+oral	587 (459–719)	46.5 (41.7–53.0)
ex:oral	712 (327–1,074)	46.2 (36.3–61.4)
ex:nasal	450 (417–548)	39.0 (30.3–47.9)

We recruited and paid 80 native speakers of German (41f, 38m, 1 non-binary) with a mean age of 34.0 years (range: 18–72) as participants via Prolific.[3] Four participants

[1] The stimuli can be accessed via https://cloud.hiz-saarland.de/s/2ktDEJPk95GHLjS.

[2] While we did try to use diverse stimuli from every group, differences in intensity and duration could be a result of sampling and/or natural differences between breath noise types. We tried to use unaltered stimuli wherever possible but in two cases we had to make minor modifications to two of the contexts: in one *ex:oral* stimulus, we cut off the first 500 ms in the beginning, as there was a noise that could have been interpreted as a breath noise, and in another *ex:oral* stimulus there was mainly silence around the breath noise followed by very loud laughter which we reduced by 20 dB. The second of these was also used in its modified form in Experiment 1. We still used those stimuli as our choice, especially for exhalations, was limited (cf. 6.1.2.1).

[3] Accessed via https://www.prolific.co on 18/01/2022.

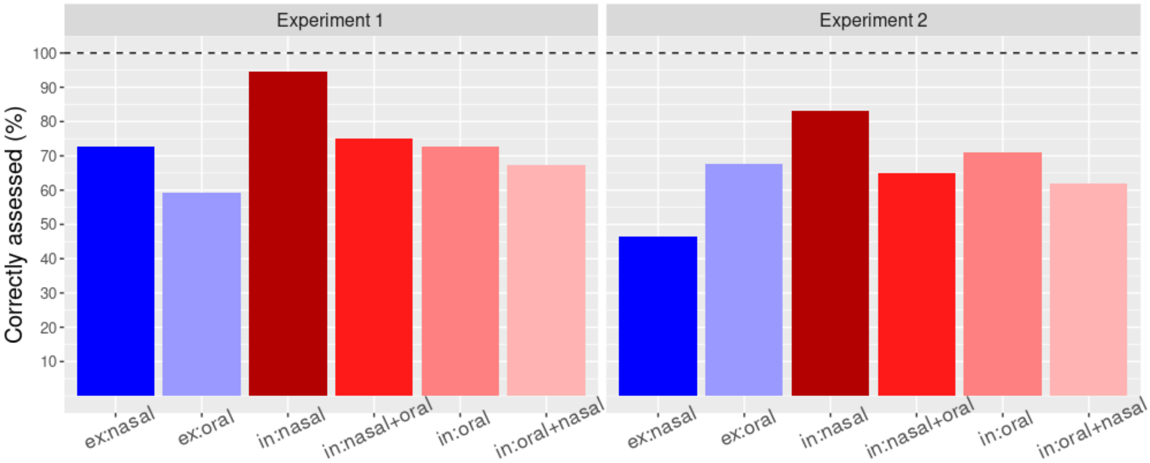

Figure 6.1: Correct assessment of stimuli by the breath type. Results from Experiment 1 are plotted on the left, those from Experiment 2 are plotted on the right.

indicated to have beginner's knowledge of Dutch and were kept in the study. Again, the stimuli were presented via Labvanced (Finger et al., 2017) and subjects were able to listen to a given stimulus a maximum of five times before they had to make a decision.

6.1.3.2 Results

Overall, participants classified breath noises correctly 65.8% of the time. Individual participants ranged from 25.0% to 91.7% correctly assessed stimuli. When taking into account the factor context, 66.7% of stimuli with context were classified correctly, whereas 65% of the no-context stimuli were assigned to the right category. Looking at correct classification by breath type (see Fig. 6.1, right) we find some differences: *ex:oral* (67.5%), *in:nasal+oral* (65.0%), *in:oral* (70.9%), and *in:oral+nasal* (61.9%) are all relatively close to the overall mean of 65.8%. *in:nasal* (83.1%) is higher and *ex:nasal* (46.6%) lower than the rest.

We then analyzed the data using generalized linear mixed-effects models (GLMMs) from the lme4 package (Bates et al., 2015) in RStudio (RStudio Team, 2022). The GLMMs used a binomial family and logit link. Decisions on models were made bottom-up starting with a random effect structure only and gradually adding fixed effects. As random effects we had intercepts for subjects and items, as well as by-subject random slopes for *breathtype* and by-item random slopes for *context* (as the variable *breathnoise* on its own contains the 24 individual breath noises but not whether or not there is context). Models were compared using the Akaike information criterion (Akaike, 1973). The dependent variable was whether or not an answer to a stimulus was *correct* (binary: yes, no) and potential predictors were *context* (binary: context, no-context), *breathtype* (6 levels), *breathduration* (continuous), and *intensity* (continuous).

The resulting model had the following structure: $glmer(correct \sim breathtype * context + (1 \mid participant) + (1 + context \mid breathnoise))$.[4] Adding intensity and/or duration of the breath noises did not improve the model. The model, using the alphabetical default of *ex:nasal* with *context* as intercept (Est. = 0.1561, SE = 0.3995, z = 0.391, p = 0.6960), revealed main effects for the breath types *ex:oral* (Est. = 1.3063, SE = 0.5732, z = 2.279, $p < 0.05$), *in:nasal* (Est. = 1.2677, SE = 0.5675, z = 2.279, $p < 0.05$), and *in:oral* (Est. = 1.4262, SE = 0.5719, z = 2.494, $p < 0.05$), all of which positively influenced correct assessment. The other breath types, as well as the *no-context* condition (Est. = -0.6244, SE = 0.3705, z= -1.685, p = 0.0919) did not reach significance as main effects. There were, however, two interactions between these two predictors that turned out significant: the interactions between *in:nasal* & *no-context* (Est. = 1.3312, SE = 0.5598, z = 2.378, $p < 0.05$) and *in:nasal+oral* & *no-context* (Est. = 2.2326, SE = 0.5396, z = 4.137, $p < 0.001$) both had a positive effect on correct identification.

Fig. 6.2 shows the breath types of the stimuli in boxes on the left and the answer type given by participants on the right. It gives an overview of how many items of a given stimulus move to the same type in the answer (representing a correct answer) but also how many migrate over to a different type in the answer (wrong answers). Some 'migrations' are more frequent than others as can be seen by the thickness of the line. Further, it shows how often a certain type was chosen as an answer regardless of the given stimulus (via the height of the black box on the right). The plot suggests that the most frequent misassessments were *ex:nasal* as *in:nasal* (with 41.6% of answers given for the stimulus type *ex:nasal*). In addition, *in:nasal+oral* was miscategorized as *in:nasal* (19.4%) relatively often and *in:oral+nasal* as *in:oral* (19.1%), while all other migrations remained below 10%.

6.1.3.3 Discussion

The context effect we had suspected from the previous experiment did not turn out to be as strong and not significant in the GLMM. We did, however, find differing results by breath type again: The two purely nasal breath events stand out with either higher (for inhalation) or lower (for exhalation) than average assessment rate. An explanation for *ex:nasal* scoring so low could be in its characteristics, as this type has short durations and low intensity (cf. Table 6.1). Yet, this would not account for *in:nasal* being the most correctly assessed type, as it is equally short and only a little more intense. It may be related to the fact that these two breath types were generally

[4] Please note that in the original publication, the model structure was $glmer(correct \sim breathtype * context + (1 + breathtype \mid participant) + (1 + context \mid breathnoise))$. As this produced a singular fit warning, the by-participant random slope was removed here. Importantly, this has only marginal influence on model output.

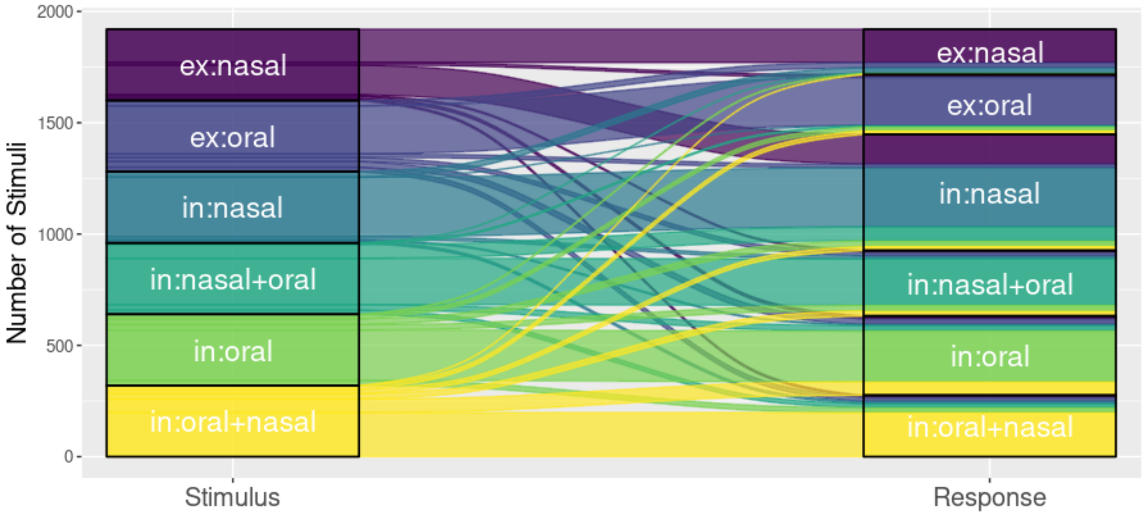

Figure 6.2: Alluvial plot of the breath type of the stimulus (left) and type of the response (right). For correctly assessed stimuli, stimulus and response are the same, whereas otherwise the response will be something else.

clicked as answers the most or least respectively, as visible in Fig. 6.2. Since in this experiment, there were only 4 items per breath type (presented with and without context), individual items may have had an influence on the results.

6.1.4 Discussion of both experiments

Overall, correct classification rate was higher in Experiment 1 than in Experiment 2 (73.6% vs. 65.8%). It is not clear where the difference comes from but it may have come from the different recruitment methods (voluntary participants known to at least one of the authors vs. paid participants via an online recruitment platform). While there was a slight tendency for a context effect in Experiment 1 with a 6.5% difference in correct assessment, the difference was only 1.7% in Experiment 2. Experiment 2 is more apt to test that effect with matched stimuli and a higher participant number. Yet, the logistic regression model did not find a main effect for context suggesting that the effect is either not there or very small. The difference in Experiment 1 could have been driven by individual stimuli, as they were not matched there.

Correct identification by breath type was lower for most types in Experiment 2 compared to Experiment 1, which is in accordance with the lower overall rates. The biggest differences from Experiment 1 to 2 could be observed in *in:nasal* (94.5% vs. 83.1%), *in:nasal+oral* (75.0% vs. 65.0%) and especially *ex:nasal* (72.4% vs. 46.6%). Since only 4 (or 8 for Experiment 1) individual breath noises were used as stimuli, effects of particular examples being more typical or salient than others may not be ruled out. *in:nasal* stood out in both experiments with its high correct categorization rates. Fig. 6.2 suggests that in Experiment 2 at least, this may have been caused by many responses being *in:nasal*, regardless of the stimulus. This might be related to participants assuming *in:nasal* as the default for breathing, which it is at rest hence they should have been exposed the most to this type of breathing.

The interaction effects seen in Experiment 2 suggest that while in general having speech context leads to slightly and non-significantly higher correct assessment rates, for some breath types having no context is actually more helpful. One reason for this may be found in the experimental setup: we tried to keep the stimuli we used as close to the originals as possible and thus only made two minor modifications (cf. 6.1.3.1). While we did account for the intensity of the respective breath noise, we did not include the intensity of the speech in the surrounding context. As this intensity may differ depending on speaker, recording, or the speech produced, a relatively intense context may make it harder for the listeners to identify the breath noise or even lead to them lowering the volume of their headphones.

Breath noises are typically not very intense, which makes it hard to incorporate them in perception studies. Also, the two experiments were conducted online so there is little control over participants and their audio settings and equipment. Another point to mention is that by Lester & Hoit's (2014) findings, a large part of our oral inhalations may have been simultaneous oral-nasal inhalations. This may have had an influence as listeners are susceptible to differences in degree of nasality in speech (Santos et al., 2021). It may have contributed to why nasal inhalations were clicked as answers the most, even though nasal exhalation stimuli were the biggest group to be misinterpreted as *in:nasal*.

6.1.5 Conclusion

The aim of the experiment was to test how well listeners can discriminate between different types of breath noises and which factors influence that rate. Overall, we found assessment to be correct for around two thirds of the stimuli. Whether or not a participant had phonetic knowledge did not make a difference in Experiment 1. We found differences in correct categorization based on the breath type of the stimulus. Context by itself did not significantly affect correct assessment but it interacted with two breath types where having no context was beneficial.

There is only a small number of studies that have examined different types of breath noises, such as Kienast & Glitza (2003). We think that especially for forensic purposes, the findings of this study are relevant. In Kienast & Glitza (2003) for instance, two trained annotators assessed breath noises (using six slightly different categories than we did), whereas we tested breath noise categorization on a large number of untrained people. It is still important to note that the overall correct categorization rate is not very high and that there seem to be some systematicities that can create problems for categorizations, such as nasal exhalations being frequently interpreted as nasal inhalations. Whether or not the overall rate reported here is usable or reliable enough for a given annotation depends on its purpose and granularity. However, when translating these results into an annotation scenario, the correct identification rate is expected to be higher. Annotators have more control over how often and which part exactly they listen to. In addition, when working with a tool like Praat (Boersma & Weenink, 2019), there is visual information, too. Further experiments on the perception of breath noises could simulate that with trained annotators. They could also use a more controlled experimental design by not using stimuli extracted from natural conversations, instead eliciting them with the help of nasometry or similar devices, provided that audio quality remains usable and the breathing behavior natural.

6.2 Using breath noises for speaker discrimination and classification by human listeners

6.2.1 Introduction

Audible breath noises may bear great potential to forensic purposes for several reasons. Around speech, they occur very frequently due to the reorganization of the breathing cycle (see ch. 2.2.1). When performing effortful actions, breathing may also surface as audible noises (Trouvain & Truong, 2015). When disguising their voices, breathing as a vital function may be less controlled consciously and thus less affected than speech (Zhao et al., 2017). Additionally, neural networks have shown promising results on speaker identification by breath noises (Lu et al., 2020; Zhao et al., 2017) and may also be able to extract a speaker's posture (Ilerialkan et al., 2020) and help identifying people sick with COVID-19 (Mallol-Ragolta et al., 2022; Nallanthighal et al., 2022). Research into individuality of breathing also suggests that, while there is variability, the respiratory patterns show quite some within-speaker stability over time in rest breathing (Benchetrit et al., 1989), but also in speech breathing, even under light physical activity (Serré et al., 2021). Yet, breath noises' potential has remained largely untapped for forensic purposes, with few exceptions (e.g. Kienast & Glitza 2003; Braun 2020). This experiment aims at investigating how usable breath noises are for human listeners in discriminating between speakers and in classifying the speaker by asking two questions:

1. When hearing two audible inhalations, how successful are human listeners at deciding whether or not the two breath noises came from the same speaker?

2. When hearing only one audible inhalation, how often will human listeners classify a) their sex (female vs. male) and b) their age (young vs. old) correctly?

6.2.2 Methods

Materials For this purpose, we used breath noises annotated in dyadic dialogues taken from an audio-visual corpus (van Son et al., 2008). The annotation procedure has been described previously in ch. 6.1.2.1. Following Lester & Hoit's (2014) finding that the majority of speech inhalations may involve both the oral and nasal tract simultaneously, we chose to use only oral (and probably simultaneously nasal) inhalations. In addition to frequency, this decision was also made to increase comparability by keeping the airway usage in the breath noises constant.

Since discrimination of age is tricky even in phonated speech (Jessen, 2007) and the perceptual cues potentially most exploitable there, such as pitch, voice quality, articulation and speech rate (cf. Moyse 2014; Harnsberger et al. 2008; Hartman 1979), we opted for a binary distinction into the two extremes *young* and *old*, rather than age

age	sex	duration (in s)		intensity (in dB)	
		mean	sd	mean	sd
old	female	0.437	0.108	45.34	5.79
	male	0.574	0.208	39.87	4.15
young	female	0.416	0.103	44.44	4.23
	male	0.470	0.104	45.19	7.47

Table 6.2: The breath noises used as stimuli in the experiment.

on a continuum and asking to assign a precise age. This also bypasses the tendency of subjects to place their age guesses more centrally on the scale than they are, i.e. to overestimate the age of younger speakers' stimuli and to underestimate it in older speakers' stimuli (Moyse, 2014). For this, we chose the six youngest (ages: 20, 20, 21, 23, 29, 29) and six oldest participants (ages: 59, 62, 62, 62, 64, 65) from our annotated corpus. Both age groups consisted of 3 *female* and 3 *male* speakers each thus balancing the set also by sex. From every speaker we extracted 5 noises that were relatively salient, i.e. not too short or too soft. An overview of those 60 breath noises can be seen in Table 6.2. As we used naturally occurring stimuli[5], they differed slightly in their duration and intensity, as most strikingly visible in the old, male speakers who on average produced less intense stimuli. In turn, the duration was higher there which may make up for the lower intensity.

For the discrimination task, every stimulus was created by concatenating a breath noise (i.e. breath1), 500 ms of silence, and another breath noise (i.e. breath2), either from the same or a different speaker. We chose 500 ms because shorter stretches of silence made the pair of two inhalations sound like in- and exhalations. There are twelve speakers who can be pooled by sex and age, resulting in four groups of three speakers each. We chose to create seven stimuli per group that would have someone from this group as breath1. Per group the pairs are as follows: Every noise from a given speaker will be presented in combination with a different breath noise from the same speaker, which leads to three uses as breath1 per group. Per group, there are also three comparisons with the other groups, i.e. one with every other group, which adds another three. There is also one comparison per group with a different speaker from the same group (e.g. male+old vs. male+old, but not the same speaker), adding thus one stimulus. These seven stimuli per group multiplied by the number of groups, namely four, lead to 28 stimuli in total. With all instances of breath1 balanced, we also made sure that no speaker would be over- or underrepresented in the stimuli, so that every speaker appears either four or five times in total in the stimuli for this task, including both as breath1 or breath2. Every individual breath noise occurs only once in the stimuli. In the classification task, every stimulus consisted of just one breath noise, taken from the pool of 60 noises (5 taken each from 12 speakers).

[5] Audio files have been made available at https://cloud.hiz-saarland.de/s/PgZ3PQ3PM9DJnoF.

Participants In total, we recruited and paid 42 participants via Prolific[6] who completed all tasks. From those, we excluded some participants, so that we only included those who indicated no hearing difficulties (2 participants excluded), sitting in a quiet environment throughout the experiment (0 participants excluded), and wearing headphones throughout the entire experiment (7 participants excluded) in the questionnaire. This resulted in including 33 speakers (22 female, 10 male, 1 other; age range: 20–71, median: 31) in the experiment.

Procedure The stimuli and the answering interface were presented via Labvanced (Finger et al., 2017) in two different ways for the two tasks: In the speaker discrimination task related to Question 1, participants heard two breath noises that were separated by 500 ms of silence. After hearing them one to five times, participants were asked if they thought the two noises were produced by the same speaker, to be answered in multiple choice design: yes vs. no, and how confident they were of their answer on a 5-point Likert scale from *not at all* to *very confident*. In total, every participant heard 14 pairs, i.e. half of the data, for this task.

In the speaker classification task, participants were presented with one breath noise at a time and asked what they assumed the sex and age of the person who produced it was. For sex, they were asked to answer in a multiple choice design whether they thought the speaker was *female* or *male*. The question for age between *young* and *old* was presented analogously. For each of these two questions, participants were also asked to specify on a 5-point Likert scale how confident they were of their answer, similar to the one used in the discrimination task. In total, every participant heard 20 stimuli for this task.

Statistics The dependent variables were analyzed using generalized linear mixed-effects models (GLMMs) via the `lme4` package (Bates et al., 2015) and evaluated using the `lmerTest` package (3.1-3; Kuznetsova et al. 2017) in RStudio (RStudio Team, 2022) with R (4.2.1; RCore Team 2022). Models were run using a logit link, the bobyqa optimizer, and increased iterations to $2 \cdot 10^5$ to avoid convergence issues. For the discrimination task, the outcome was thus binary (correct: yes, no). The predictors were the relation between the two breath noises in the stimuli regarding sex (different, same_male, same_female) and age (different, same_young, same_old). In the classification task, we ran two models that were either about sex or age being guessed correctly. Their outcome again was binary, i.e. either sex correct (yes, no) or age correct (yes, no). Since participants heard only one breath noise here, the predictors were speaker's sex (male, female) and age (young, old). All models here include random intercepts for participant and stimulus, as including random slopes led to non-singular fits.

We here tested the theoretically-motivated models by comparing the respective

[6] Accessed via https://www.prolific.co on 17/05/2022.

		sex			
		same_male	same_female	different	total
age	same_old	79.5% (73)	76.8% (69)	60.0% (35)	74.6% (177)
	same_young	73.1% (67)	53.1% (64)	65.0% (40)	63.7% (171)
	different	31.8% (22)	35.3% (34)	63.8% (58)	49.1% (114)
	total	70.4% (162)	59.3% (167)	63.2% (133)	64.3% (462)

Table 6.3: Correctness rate of the discrimination task by speaker sex and age in percent. The numbers in brackets indicate number of stimuli per cell.

null model (without any predictors) to the full model with and without predictors using the Akaike Information Criterion (AIC) (Akaike, 1973). A model is assumed to have a better fit, if AIC decreases by at least two points. We used a significance of $\alpha = 0.05$ for the models. After fitting the model, we used the function `ggemmeans` from the package `ggeffect` to predict probabilities from the model in more easily interpretable percentages, rather than log-likelihood, as well as their 95% confidence intervals.

6.2.3 Results

6.2.3.1 Discrimination task

Overall, participants were successful at deciding whether two breath noises were produced by the same speaker in 64.29 ± 11.85% of all cases. Table 6.3 provides an overview of the correctness rate by how age and sex of the speakers in the respective stimulus relate to each other.[7] There are strong differences for different combinations in the table: Most combinations of same age and same sex seem to work well together with rates between 73 and 80%, with the exception being stimuli from young, female speakers. The numbers for same age but different sex are lower with 60 or 65%, respectively. In the other direction, i.e. same sex but different age, the numbers drop dramatically to around 30–35%. When both age and sex are different, the rate is at around 64%, thus very close to the overall mean.

The best-fit model for the discrimination task included the predictors speaker age, speaker sex, as well as their interaction. Using the combination of different sex and different age as an intercept (Est. (log-odds) = 0.58, SE = 0.31, z = 1.85, p = 0.06), neither level of speaker age, i.e. neither same_young nor same_old, turned out significant. Sex, however did have a significant, negative[8] effect with both same_female

[7] For the table, it is important to bear in mind that participants were asked one question (do the breath noises come from the same or different speakers), to which they replied yes or no. In either case, that answer can be correct or incorrect. The table shows how often the answer given was correct, regardless of what the answer was.

[8] As this model includes an interaction, these effects cannot be interpreted by themselves but only including the interaction.

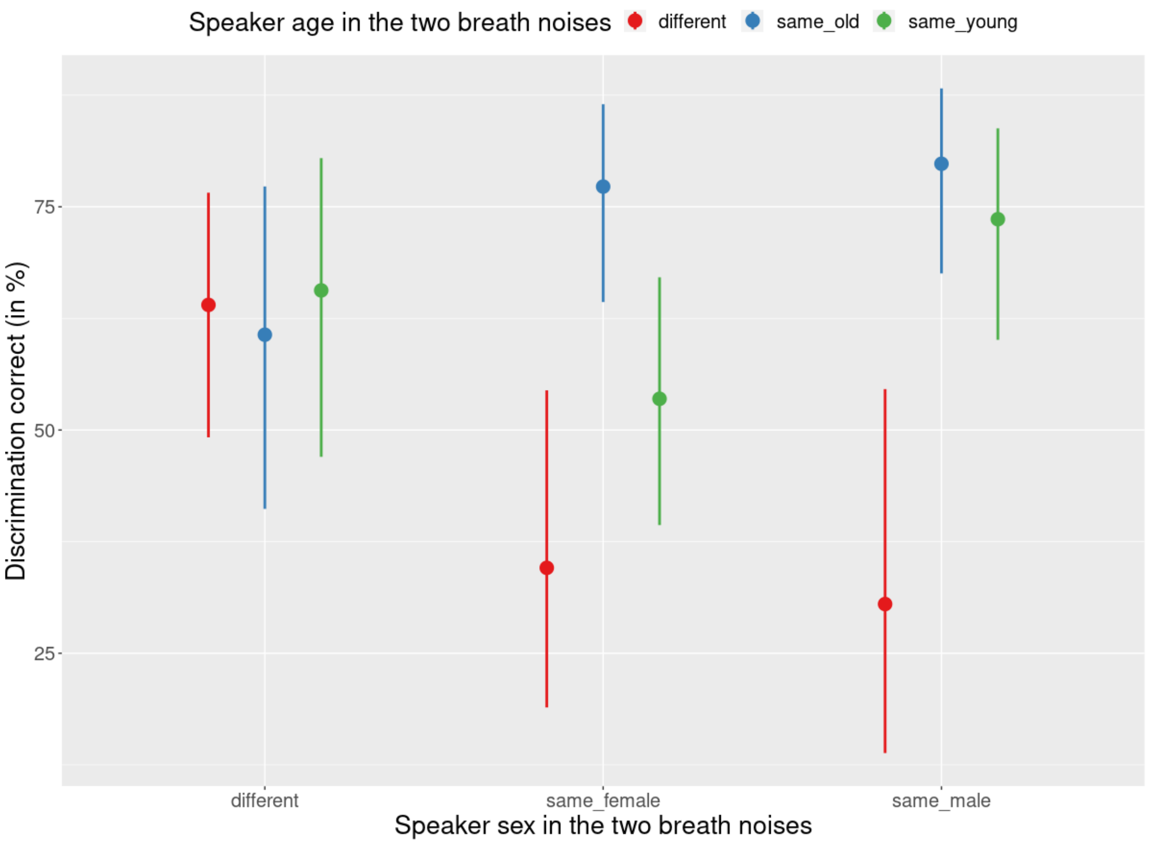

Figure 6.3: Plot showing the model predictions for the discrimination task (outcome: discrimination correct in % on the y-axis) by the relation of the speaker age in the two breath noises (in color) and the speaker sex (on the x-axis). The points are averages, the lines around them visualize the 95% confidence intervals.

(Est. (log-odds) = -1.21, SE = 0.52, z = -2.34, $p < 0.05$) and same_male (Est. (log-odds) = -1.40, SE = 0.6003, z = -2.33, $p < 0.05$) turning out significant. The interactions same_old:same_female (Est. (log-odds) = 2.00, SE = 0.73, z = 2.74, $p < 0.01$), same_old:same_male (Est. (log-odds) = 2.34, SE = 0.79, z = 2.95, $p < 0.01$), and same_young:same_male (Est. (log-odds) = 1.78, SE = 0.78, z = 2.28, $p < 0.05$) turned out as significant. This becomes clearer when plotting the model output, as done in Fig. 6.3: Looking at the part on the left, i.e. with different speaker sex, all points are very close to each other and thus not different from the intercept, which is different sex and different age, i.e. the red line on the left. Comparing the other two red lines to the intercept, i.e. looking at different age but same sex, it can be seen that the points are much lower. The plot also shows how much higher the combinations of same sex and same age are from the intercept, with the exception of same_female and same_young. Participants' overall confidence in their answer had an average of 3.50. Confidence differed only slightly depending on whether the answer was correct or not, with a mean of 3.57 for correct answers and 3.39 for incorrect answers.

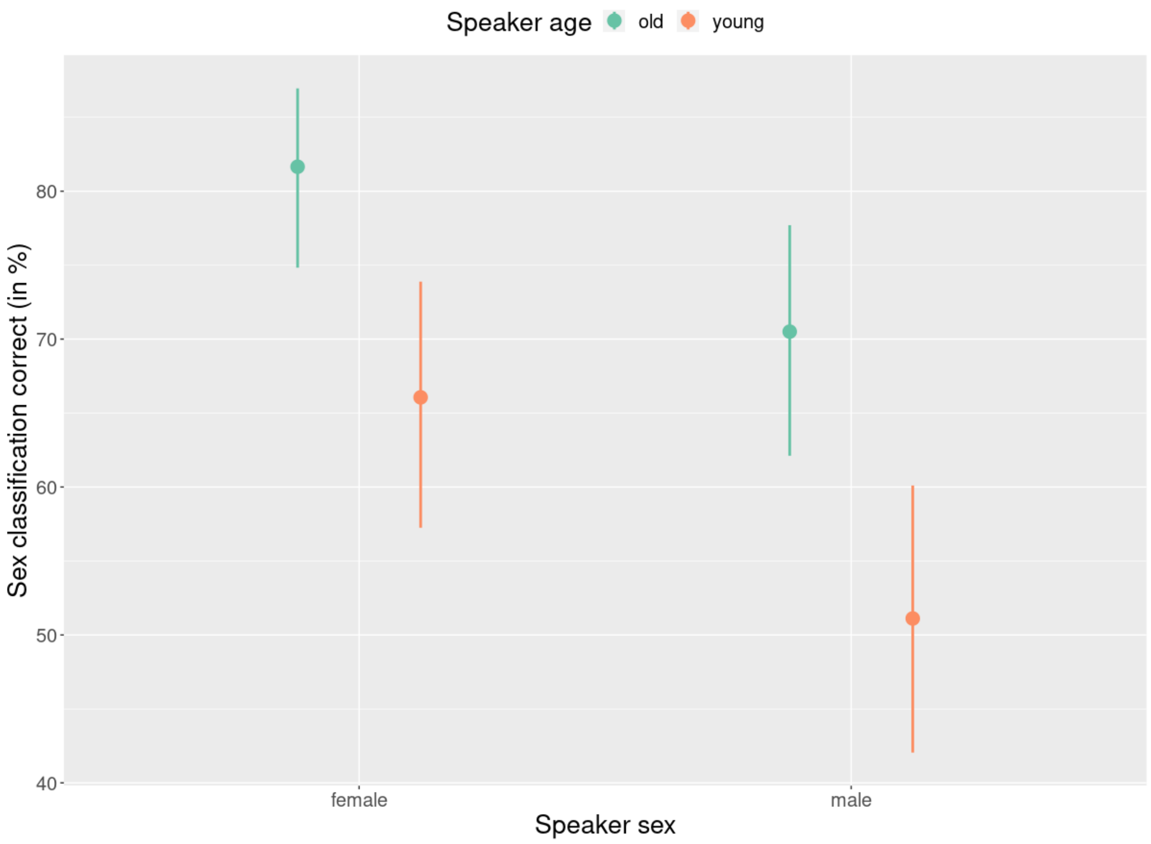

Figure 6.4: Plot showing the model predictions for the classification task regarding speaker sex (outcome: classification of speaker sex correct in % on the y-axis) by speaker age (in color) and speaker sex (on the x-axis).

6.2.3.2 Classification task

Participants were more successful at guessing a speaker's sex than their age. While the overall mean for age being guessed correctly is 50.15 ± 9.06% (331 *correct* vs. 329 *incorrect* answers), it is 66.67 ± 13.50% for sex (440 *correct* vs. 220 *incorrect*).

For classifying sex correctly, the best-fit model was additive, i.e. including the speaker's age and sex but no interaction between them. Using female+old speakers as intercept (Est. (log-odds) = 1.49, SE = 0.21, z = 7.25, $p < 0.001$), there was an effect of both age and sex: changing sex to male (Est. (log-odds) = -0.62, SE = 0.20, z = -3.05, $p < 0.01$) and age to young (Est. (log-odds) = -0.83, SE = 0.20, z = -4.04, $p < 0.001$) have significant, negative effects. Model predictions are plotted in Fig. 6.4. The plot also visualizes why the interaction did not increase model fit: both speaker age and speaker sex have effects that only add up but do nothing beyond that. We were not able to fit a model for whether age was guessed correctly, as all possible models returned singular fit warnings.

On average, participants were a little more confident in their answers regarding sex (mean = 3.21), than they were in their answers on age (3.01). Confidence in their age estimation remained very similar, regardless of whether their answer was correct or not, with a mean of 2.98 for correct answers and 3.04 for incorrect answers to the speaker age question. For sex, there was a bigger, yet still small, difference: mean confidence was at 3.31 for correct answers and at 2.99 for incorrect answers regard speaker sex.

Since the two variables speaker age and sex also allow for combination, it is worthwhile looking into how often not just one of them but both were actually guessed correctly, which happened in 34.70 ± 11.5% of all cases. How stimuli (as combinations of speaker age and sex) were classified can be seen in the alluvial plot in Fig. 6.5. Stimuli can be seen in the boxes on the left, participants' answers to them on the right. When the connection goes from a box on the left to the same box on the right, the response is correct, otherwise it is incorrect. In this case, sex or age may still be correct. This visualization reveals that a large portion (57.5%) of old+female stimuli were classified as young+female, while only 23.0% were correctly classified as old+female. Moreover, it shows that in the case of young+male stimuli, the streams leading to the four possible combinations are more or less even in size (ranging between 23.0 and 28.7%), suggesting that in this case, a high degree of pure guessing was involved. For old+male and young+female stimuli, the biggest portion is classified correctly, namely 45.2 and 47.8% respectively.

6.2.4 Discussion

6.2.4.1 Discrimination task

The fact that most combinations of same sex and same age work well together, as visible in Table 6.3 and Fig. 6.3, is likely due to the fact that there, both cues, i.e. speaker sex and age, point in the same direction for participants' answer, namely that they are perceiving the same speaker (or different speakers from the same sex and age combinations). The combination of different sex and different age being lower than that may be explained by a possible tendency of participants to assume the same speaker, potentially related to single breath noises only carrying a limited set of information on the speaker and thus sounding similar. Overall, both the Table and the Figure suggest that sex differences were more perceivable than age differences: When speakers have the same sex but a different age, the correctness rate of the answers drops even far below chance level. Vice versa, the combination of different sex and same age seems to be more perceivable as coming from different speakers. Having no theoretically motivated reason to assume that participants may behave very differently for young, female speakers, as opposed to all other combinations of same sex and same age, this difference may be due to the stimuli, rather than an actual effect. As we used naturally occurring stimuli, there may have happened to be more within-speaker or less between-speaker variability in the case of this group which ended up complicating the task. It may also be related to the participants (that were

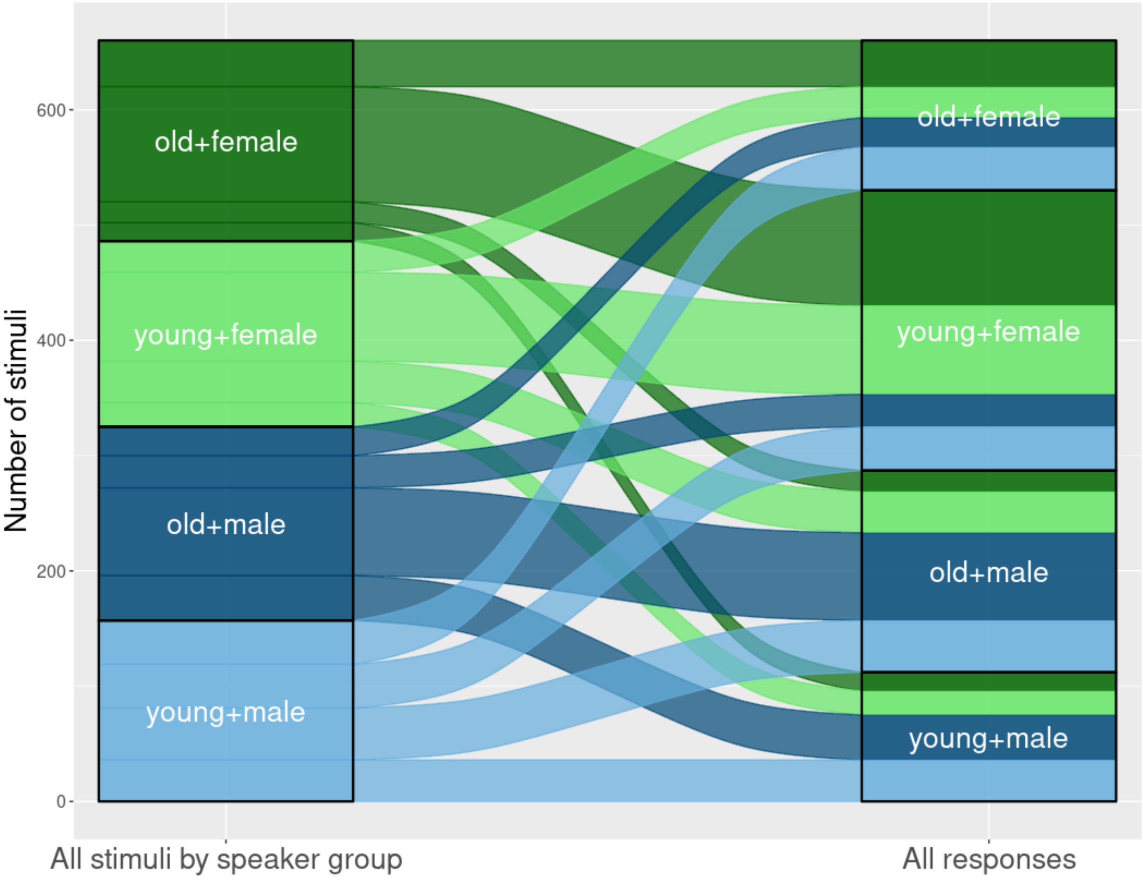

Figure 6.5: Alluvial plot for the four groups defined by age and gender. The boxes on the left side show the stimuli, the boxes on the right the responses.

largely young and female) reacting differently to different stimuli, as some studies suggest this to introduce bias (Moyse, 2014). However, this own-age bias is disputed and if present, should have contributed to the young+female group being the most correct, which it was not. Whether what has been seen with young+female stimuli is a real effect or an artifact of the stimuli or participants should thus be tested with other stimuli and more participants. If this effect is not shown, the groups same_female and same_male could also be merged into same sex as opposed to different sex. The same applies to same_old and same_young. We here decided to keep them separate due to this study's exploratory nature. The numbers in brackets in the Table and the wide confidence intervals suggest that these findings are to be taken cautiously due to the small number of participants. Many of the confidence intervals also overlap with the 50% line. Whether adding more participants narrows down these ranges and thus gives a clearer picture remains to be tested.

6.2.4.2 Classification task

The descriptive results already indicated that participants were more successful at classifying speaker sex than age when hearing a breath noise. While sex was perceived correctly in two thirds of all cases, the rate for age was at chance level. This is supported by the GLMMs that for classification of sex returned both speaker sex and age as significant predictors. For age classification, model fitting was not successful, which may be related to the participants struggling with this subtask.

Age classification being more difficult than sex is in line with findings from Jessen (2007). That is the case, even though we used a very coarse, binary distinction instead of a more continuous approach, which would have been much harder. However, it may have been complicated by us using chronological age, rather than biological age. Chronological age represents someone's age as predicted by their date of birth, while biological age takes into account the individual aging of organs and physiological mechanisms relevant for speech production. It can thus be influenced by speakers' lifestyles and consumption habits, potential vocal fold overuse or abuse, or stress (cf. Jessen 2007). Fig. 6.5 helps explore some patterns that may or may not generalize to bigger studies, such as many stimuli from old+female being perceived as young+female, or the pure guessing for young+male stimuli.

6.2.5 Conclusion

This study has shown that listeners are sensitive to speaker information in breath noises: They are more or less able to tell whether two breath noises came from the same speaker or not, and are successful at classifying a speaker's sex when hearing just one breath noise. Age classification, however, did not work beyond guessing. While it seems that breath noises may be useful in discriminating speaker sex, they seem hardly exploitable for the distinction of chronological age. Since we cannot rule out that biological age differences in the speaker set may have confounded the results, future studies should take this into account. These findings, however, are largely exploratory and need further testing. There, with the direction laid out by this work, it is crucial to increase the number of stimuli and participants. This should help make the findings more clear, as the confidence intervals in Fig. 6.3 and Fig. 6.4 are still very wide in general and for some combinations especially.

In addition to that, there are other potential confounding factors such as a speaker's height or weight. However, this opens up potential for further research to see if the sex effect is really driven by differences in sex or mainly by differences in size. Speech breathing (in terms of lung use and breathing pattern) should generally be the same in males and females when normalizing for height (Hixon et al., 2020, p. 57). It remains to be tested if that would also translate into similar breath noises or if the filtering introduced by male vs. female vocal tracts (see ch. 5.3.2) would result in perceptually different breath noises.

Perceivable differences by sex but not age may be related to differences in vocal tract length, which differs by sex (Stevens, 2000, p. 25–26). All young speakers in this study were at least 20 years old and thus postpubescent. Age differences may be more salient in discriminating children vs. adults, via stronger differences in vocal tract morphology. The absence of a prepubescent sexual dimorphism in the vocal tract (Barbier et al., 2015), however, may make the perception of sex differences via breath noises very hard there.

Along with the relevance to forensic phonetics, these findings have implications for naturalistic synthetic speech that aims to exploit the potential advantages of including breath noises into speech synthesis (Elmers et al., 2021b; Whalen et al., 1995). Since human listeners are more sensitive to speaker sex as opposed to age, this should be paid attention to especially when creating breath noises to go with synthetic speech. Further research could look into listeners' reaction of splicing in *matching* (i.e. from the same speaker/sex) and non-matching breath noises into synthetic or natural speech. The results do suggest that splicing in the same off-the-shelf breath noise into different artificial voices is to be avoided. The findings also add to the body of literature dealing with applying PINTS for forensic purposes (e.g. Braun 2020; Muhlack et al. 2022).

Generally, all stimuli that were used in this experiment were produced under the same recording setup and may thus have higher comparability than may be the case in real-world forensic applications, where many other factors in the recording conditions may complicate comparisons. Depending on the task, however, participants here were only exposed to two or even a single breath noise, and had to base their decision on that. When given more material and potentially exploiting long-term breath features, as in Serré et al. (2021), or looking for idiosyncratic use of breath types (Kienast & Glitza, 2003; Werner et al., 2022d) there may be great potential in using breath noises for forensic purposes.

Chapter 7

General discussion

The overarching goal of this thesis was to investigate speech pauses and audible breath noises, focusing on fine phonetic detail. This goal was approached in four main steps. First, we showed how several corpora annotated pauses and pause-internal particles and presented a recording setup for speech breathing. Second, we analyzed pauses with regard to their optionality and variability under different tempo conditions and compared pausing in human and computer-generated speech. In breath pauses, we considered the temporal composition of silence and respiratory sounds. Third, we zoomed in on the breath noises themselves by describing their spectral characteristics, correlating them to physiology, and simulating them with 3D-printed vocal tract models. Finally, we tested how much information breath noises may cue to listeners about the breath type or the speaker's sex and age.

Data and annotation for pauses and pause-internal particles Ch. 3 dealt with different sources of data for research on pauses and pause-internal particles (PINTS), namely using existing corpora and setting up custom-made recordings. Ch. 3.1 gave an overview of how different PINTS were annotated in a selection of corpora. It was shown that most of them had some way of indicating pauses or audible interruptions of speech and several even accounted for filler particles or breath noises. However, there was quite some variation of how those were implemented in practice, with some assigning an interval to the entire breath pause and one identifying the breath noises, but marking it with a symbol in the same interval with an utterance. While it is obvious that different researchers have different angles and scopes, we advocate for a common, standardized framework for the annotation of PINTS. With this, annotators would not have to include all categories, but instead may use a common foundation, on which particular elements may be stacked depending on the research interests and level of detail, as done with GECO-FP (Belz, 2019) on GECO (Schweitzer & Lewandowski, 2013). This way, researchers with a higher degree of granularity regarding e.g. breath noises could add them to the pre-existing pause annotation, knowing under which specifications that was done. It would be desirable if some general categories and ideally even their symbols could be agreed upon and unified. A potential scheme could be based on Trouvain & Belz (2019). If this is too idealistic, corpora should have a manual, readme file, annotation guidelines, or paper in which they explicitly state their scope and scheme. There, they could outline if there was a pause threshold, how high it was, if all breath noises were annotated or just some that were salient, if filler particles were included and if they were annotated by themselves or as part of pauses, etc. This could not only help transparency and reproducibility, as well as research across corpora, but also the different annotators working on the same corpus.

Ch. 3.2 was dedicated to the description of pilot recordings, where we tested a setup for future work. We used a combination of respiratory inductance plethysmography (RIP), electroglottography (EGG), and audio recording. To control how much physical exercise affected participants, we used a finger pulse oximeter. We tried several speech tasks that had effects on speech and breathing behavior. Building on this setup may generate interesting insights into speech breathing research and how creak relates to the respiratory cycle (Aare et al., 2018) and is used in glottal filled pauses (Belz, 2017). The part on silent speech, that was inspired by Conrad & Schönle (1979), could further be augmented by eye tracking or a different form of text presentation, so that the role of punctuation there could be studied. This could help research on silent speech interfaces that are used in a multitude of settings, where speakers cannot speak audibly for reasons of confidentiality, background noise, or physiological reasons (Birkholz et al., 2018).

Breath and non-breath pauses in natural and synthetic read speech The main question of ch. 4 was how pausing behavior differed across several speech tempos. In ch. 4.1, we explored pauses regarding location, duration, and number, as well as audible breath noises. We included natural (three speakers each; from BonnTempo-Corpus, Dellwo et al. 2004) and synthetic speech (four text-to-speech systems each) in the three languages German, French, and English. The human readers read the text in five tempos, i.e. from very slow to very fast. Along with affecting speech parameters, increasing tempo led to shorter and fewer pauses. This allowed us to compare these results to the synthetic speech: There, we found that the systems included here tended to employ slower articulation rates (all three languages) combined with shorter (German, French) and more frequent pauses (German, English), when compared to human speakers. These results suggest that synthetic speech has room for improvement when adjusting computer-generated speech for different scenarios and interlocutors that may require slower or faster speech. Striking the right balance between articulation and pausing, along with including style-specific requirements in the modeling of pause duration (Parlikar & Black, 2012), may help synthetic speech become more natural. In some styles that give strong weighting to naturalness, breath noises could be added for enhancement. In the natural speakers, we found individual strategies regarding the use and duration of non-breath and breath pauses that different readers may follow, when changing speech tempo.

In ch. 4.2, we made use of the tempo changes to analyze the optionality, i.e. presence or absence, and variability, i.e. variation in duration and involvement of breath noises, of speech pauses. For this, we analyzed read speech across five speech tempos, six languages (Arabic, Czech, English, French, German, Italian), and 46 speakers from the BonnTempo-Corpus (Dellwo et al., 2004) and its Arabic expansion (Ibrahim et al., 2020). As tempo increased, speakers produced fewer breath and particularly fewer non-breath pauses. We found many pauses to be optional, i.e. taken only in few, typically slower conditions or by few speakers, while some were less optional. At those locations, a high number of pauses occurred that tended to be longer than pauses elsewhere and more likely to involve breathing. Variability was high there and related to tempo. These pauses were closely linked to punctuation and conjunctions, although that did not account for all of them. As speakers altered their speech tempo, they modified the number of breath pauses, but particularly dropped non-breath pauses when speaking faster. Increasing tempo led to shorter pauses, inhalations, and also edges, as speakers tried to minimize the time spent pausing there. The findings about right edges being shorter add to what has been reported by Grosjean & Collins (1979), Bailly & Gouvernayre (2012), and MacIntyre (2022), although using differing metrics. A more fine-grained description of (breath) pauses could add to modeling them more accurately, which could help advance naturalistic synthetic speech. This could also have perceptual benefits, given that MacIntyre & Scott (2022) have shown that listeners are sensitive to short gaps around breath noises.

Taken together, these two studies on pausing have shown that they follow certain patterns, e.g. when altering speech tempo. They have also emphasized the optionality and variability of pauses. They further add to research on breath and non-breath pauses in read speech (e.g. Winkworth et al. 1994; Godde et al. 2021) in corroborating the finding that speakers use punctuation as landmarks for pausing and breathing especially. What our studies add is the notion of how malleable pauses are in terms of presence at a given location, duration, and inclusion of a breath noise, as exemplified by different speech rates. While punctuation is a major predictor of pauses, it alone does not cover all pauses. Disregarding those human pause patterns when gearing text-to-speech systems for different needs and situations may create mismatches for synthetic speech, for which human perception should be the ultimate criterion. How much deviation from these patterns is acceptable could be tested in perception studies. In both experiments, we also observed individual strategies for pausing and breathing in text reading. However, the texts used in both studies were quite short and using longer texts or different speech styles may create further insight into how speakers adapt to tempo changes more generally. In addition, the tempos were varied by asking participants to use a certain tempo and hence the interpretation of them differed across participants. On the faster side, they may have been limited by a natural bound to how fast they were able to speak but for slower speech, participants are likely to have varied more. Future recordings could try to elicit different tempos more naturally or exert more control on how fast participants are speaking.

Given the bimodal distributions that have been reported for pause duration, the fact that breath and non-breath pauses are often not distinguished, and the average duration of a breath noise and edges, this thesis advocates for further pause research to distinguish between non-breath pauses and breath pauses. Although that may not completely correspond to the two modes, it could help disentangling different aspects and causes of pauses. Kirsner et al. (2003) stated that the short and long pauses are functionally independent and speculated that long pause duration may be influenced by "intention, attention, planning, topic change, and inspiration".

Moreover, future studies on pausing in read speech should incorporate more approaches like Šturm & Volín (2023) to look beyond punctuation in pause prediction. Along with overt text structure, which is arranged by the written unit, they included covert text structure, reflecting information on word class, syntax, and speech rhythm. They found overt text organization to influence both predicted pause duration and rate, i.e. number of pauses at a given location. Covert text structure, ranging from locations where pauses are blocked to stronger breaks, may have only influenced the pause rate and not duration, but helped explain the whole picture. Their approach, also adopted in Kentner et al. (2023), was inspired by a recent coding manual presented in Franz et al. (2022) that was made for the annotation of several degrees of prosodic boundaries based on syntactic and lexical structure, as well as punctuation. In addition, it is important to bear in mind that read speech is just a fraction of all speech types. Studies like Betz et al. (2023) are crucial for ecological validity and to stress the cross-modal nature of hesitations, that can be found in speech and gesture.

Breath noise acoustics and physiology The experiments of ch. 5 were all centered around the aim of further describing the acoustics of speech breath noise. It was aimed at adding to the sparse literature on the spectral characteristics and acoustics-physiology interactions in speech breath noises. To this end, we conducted three studies focusing on the temporal alignment of acoustic and kinematic signal of the upper body, acoustic parameters and how they correlate with kinematics, and describing the spectrum of breath noises, as well as simulating them in 3D-printed vocal tract models. In ch. 5.1, we explored the temporal coordination of acoustic and respiratory events when listening and when speaking. We used audio and kinematic (RIP) data of 14 female speakers. The listening condition showed the expected respiratory behavior, i.e. fewer breath cycles and longer inhalations. For speaking, we found that rib cage expansion (in relation to abdomen) tended to happen slightly earlier, but there was a high degree of individuality. We found that the onset of inhalation noises tended to be coupled to the expansion of both abdomen and thorax.

In ch. 5.2, we examined acoustic and physiological characteristics of audible inhalations and how they may correlate. To this end, we used audio and RIP data of 31 female speakers. We extracted several parameters from these breath noises, namely center of gravity, intensity, and the first three formants (F1, F2, F3), to compare inhalations to a selection of speech sounds as reference. Inhalations preceding speech were less intense but longer than those in speech pauses. We found the center of gravity of inhalations to be similar to the aspiration phase of /k/, but not to that of /t/ or /h/. Inhalations tended to have higher F2 values but particularly higher F1 values than realizations of /ə/, suggesting a more open and slightly more front place of articulation for inhalations. Center of gravity, intensity, and the first formant were also positively correlated with inhalation slope, suggesting that when speakers inhaled faster, the resulting breath noises were more intense and produced more towards the front of the mouth. A limitation of this study is that inhalations are typically produced without phonation, thus complicating the use of formant measures, and with an air stream direction that is opposed to nearly all speech sounds, exacerbating comparisons. Further, their spectra are relatively flat, making center of gravity less informative than in fricatives, such as /s/ or /ʃ/.

To circumnavigate these problems, we designed the experiments of ch. 5.3, where we departed from analyzing formants in our approach. The reason for this, along with the limitations mentioned above, was that they were not very strong and F1 especially was not always visible. Reasons for this may be the high amount of airflow through the glottis leading to a larger bandwidth (Ishikawa & Webster, 2023) and simultaneous usage of both oral and nasal airway that may be assumed for a substantial part of speech inhalations (Lester & Hoit, 2014), which might contribute to F1's larger bandwidth (Styler, 2017), further weakening its amplitude (Park, 2002). Therefore, we here considered the entire spectrum. In a large sample of human inhalations produced by 100 male and 34 female speakers, we found relatively flat spectra with decreasing slopes and several weak peaks. The spectra were similar to those reported by Nakano et al. (2008) and the peaks mostly aligned with the resonances reported by Hanna et al. (2018) for inhalations produced with a vocal tract configuration of /ɜː/ by speakers of non-rhotic Australian English. This suggests that speech breathing may be performed with a neutral vocal tract configuration, close to /ə/. Further, we simulated in- and exhalations by changing the airflow direction in the 3D-printed VT models that we used to model breathing. There, we found that reversing airflow did not have a general effect for all VT configurations, but only those with high-tongue configurations, as opposed to those that were more open. /s/ and /ʃ/ especially changed the most, which we reasoned to be related to a concentrated air stream hitting the incisors in exhalation, but not in inhalation. Finally, we compared the /ə/-model spectra to human inhalations to approach the vocal tract configuration that is assumed during speech breathing. We found some similarities between the two, although several complexities of human breathing not captured in the models complicated the comparisons.

Taken together, in this part we have made use of different approaches to describe the acoustics of speech breath noises, which had not been researched extensively (see e.g. Kienast & Glitza 2003; Nakano et al. 2008). A better understanding of them may advance models of speech production, improve speech segmentation, which may otherwise confuse breath noises with e.g. fricatives, and improve automatic speaker classifications of breath noises, as well as the non-invasive detection of pathologies. In addition, it may make speech synthesis more natural, in combination with a more detailed modeling of the entire breath pause. The findings support our view that in speech breathing, the vocal tract assumes a configuration close to or more open than /ə/. This suggested VT configuration in inhalations should be taken into account by studies on articulatory settings and pause postures (e.g. Gick et al. 2004; Krivokapić et al. 2020) to tease apart non-breath and breath pauses. To validate this and to add more detail, articulatory studies are needed (see *Future work* below). By including physiological aspects, these studies also add to the body of literature on speech preparation that highlight their acoustics (Scobbie et al., 2011) and coordination of acoustics and physiology (Rasskazova et al., 2019). As an auditory categorization into different types of breath noises, such as nasal, oral, or nasal-oral in- or exhalation, did not seem convincingly reliable (see ch. 6.1), we here subsumed all of them together. Following Lester & Hoit (2014), this may mean that the majority of them were simultaneous oral-nasal inhalations. However, it could also mean that we lumped together different types of breath noises that may have different spectral properties. Future research could try to disentangle what implications the usage of different airways has on the acoustics of the breath noises.

Breath noise perception The main question of the experiments on breath noise perception in ch. 6 was how much information listeners could auditorily extract from speech breath noises. This was motivated by the goal to find out how reliable a perceptual categorization of breath noises was, as used in studies like Kienast & Glitza (2003). In addition, as neural networks have become quite successful at using breath noises for speaker identification (Lu et al., 2020; Zhao et al., 2017) and extracting pathologies (Mallol-Ragolta et al., 2022; Nallanthighal et al., 2022), we wanted to identify how much detail the human ear could perceive. For this, we ran two experiments that were aimed at testing how well listeners could distinguish different types of breath noises and to what degree they could use them to discriminate between speakers or detect a speaker's sex and age.

In ch. 6.1, we conducted online perception experiments to investigate the auditory categorization of different breath noise types, i.e. in- and exhalations produced with the oral or nasal airway or combinations thereof, and found in the pilot that phoneticians did not perform better than lay people. In the follow-up experiment with 80 lay participants, the breath types were correctly categorized in around two thirds of all cases. By and large, the presence of one second of audio context before and after the breath noise did not help significantly. Instead, two types (nasal inhalation and nasal+oral inhalation) were even less often correct when presented with context, which may be related to differences in amplitude between speech and breath noise.

When intensity is adjusted between them to eliminate intensity differences as a source of disturbance, the extra information contained in the context may help potentially. This may operate on a smaller scale, e.g. through coarticulatory influence, or a larger scale, e.g. audible exhalations may occur more frequently outside rather than within fluent speech sections. Some types, such as nasal and oral inhalation, were more often correctly detected than others, especially nasal exhalation. Overall, this study has implications for other studies (e.g. Kienast & Glitza 2003), forensic speaker profiling interested in individual characteristics, and annotations (e.g. Schuppler & Ludusan 2020) that rely on auditory categorizations of breath noises. Even though phoneticians did not outperform lay people in the experiment, annotators working with a tool like Praat may still do so.

Ch. 6.2 dealt with how well listeners could use breath noises to distinguish between different speakers and to extract coarse speaker characteristics, such as speaker age (old or young) and sex (female or male). We included 33 participants in this study. Discrimination was correct in 64% of all cases, with results indicating that sex was more perceivable than age. In the classification task, sex was correctly detected in two thirds of all cases, while age was pure guesswork, with correct detection at chance level of 50%. Perceivable differences by sex but not age may be related to differences in vocal tract length, which differs by sex (Stevens, 2000, p. 25–26). Sex classification was more often successful for female speakers' stimuli and much better for older speakers' stimuli regardless of their sex. Future studies should test if this was just an artifact or an actual effect that could be related to sex differences becoming more prominent with age. Overall, both tasks suggest that sex was more perceivable than age, which is in line with findings for speech (Jessen, 2007). Although perceivable, sex identification in these experiments was not as successful as in Whalen & Sheffert (1996) and Whalen & Kinsella-Shaw (1997), where it was close to perfect, but based on small numbers of speakers and participants.

Since the two studies were rather exploratory, their effects should first be replicated with more participants and a new set of stimuli. In the first study, it would be desirable to have a more reliable ground truth of how participants inhaled, rather than the approximation to it that we used, namely using video and audio. However, most ways of categorizing breath types reliably would deteriorate the audio quality. In the second experiment, several aspects should be further tested: Along with chronological age, future studies could incorporate the speakers' biological age, i.e. including how their way of life may have influenced their aging and thus breathing (Hixon et al., 2020, p. 57), as it does for speech (Jessen, 2007). In addition, it should be tested if the sex differences found here were caused by differences in vocal tract length as a consequence of different heights, rather than actual sex difference. Future studies on the perception of breath noises could also further test how beneficial they may be for processing speech. Given how much speech influences speech breathing, as outlined in ch. 2.2.6, it could be investigated if listeners also utilize those cues. One could test if listeners expect the speaker to produce a longer utterance when they hear them inhale longer and deeper, or if they use breath noises for managing their expectations regarding taking, holding, or yielding of a turn.

Taken together, these two studies have implications for forensic work and naturalistic synthetic speech. The sensitivity for information contained in breath noises that we found in listeners may be helpful in discriminating between speakers in real-life forensic work. The audio quality there may be worse than in laboratory settings, however, there may be more material to work with than just single breath noises as used here, making it possible to capitalize on the breathing individuality (Kienast & Glitza, 2003; Serré, 2022). For synthetic speech aiming to use breath noises, our results imply that computer-generated breath noises should be tailored to the respective voice, rather than using the same off-the-shelf noise for different voices and situations.

Future work All parts of this work taken together have added to a more detailed view of pauses and speech breathing and advocate for distinguishing between breath noises and the pauses in which they are embedded. This could enrich speech production models and help make pause models and synthetic speech more natural, possibly by using synthesized breath noises in appropriate contexts, like audio books (Braunschweiler & Chen, 2013). For computer-generated respiratory sounds, as well as forensic phonetics, it is also important how much information is perceivable from audible breathing. The findings on the acoustics of breath noises have several implications for the non-invasive detection of pathologies (e.g. Mallol-Ragolta et al. 2022; Nallanthighal et al. 2022) and automatic breath detection (e.g. Ruinskiy & Lavner 2007; Dumpala & Alluri 2017; Fukuda et al. 2018; Székely et al. 2019). Beyond acoustics, they are relevant for articulatory studies, which could profit from teasing apart influences of silent pauses and inhalation gestures. The physiological coordination above the larynx is an area of speech breathing research that is severely under-researched.

Future studies could complement the existing types of measurements outlined in the part on supraglottal airways in ch. 2.2.2. We here used acoustics to make inferences about vocal tract configuration in speech breathing indirectly, but it could also be measured directly using articulatory approaches, such as real-time magnetic resonance imaging (rt-MRI), electro-optical palatography (Birkholz et al., 2012), or electromagnetic articulography (EMA). Of particular interest here are velopharyngeal port width, jaw opening, and tongue behavior in several areas. Articulatory studies could look into the vocal tract configuration of speech inhalations. To this end, EMA could be useful for several aspects, such as mouth opening and tongue movement. Electro-optical palatography could be used to study the causes of spikes in the acoustic signal that occur in some breath noises, typically in the beginning or around the middle of their total duration. This way one could track the distance between palate and tongue and see if the spikes are really related to mouth opening. These two methods are particularly interesting if the acoustic dimension is to be analyzed, too. However, tracking the velum is notoriously difficult or impossible with this method

(Rebernik et al., 2021). rt-MRI could be a better fit, as one could clearly track velum movement and thus directly monitor the velopharyngeal opening (Oh & Lee, 2018; Oh et al., 2021). Additionally, one could study how large the cross-sectional area is in different regions of the vocal tract from the lips to the larynx, as shown in Kim et al. (2014, p. 224). The drawback here is that it would all be under the assumption that the supine position does not change inhalation behavior. While in speech, it is assumed that non-critical articulators may be affected by the supine position, while critical articulators are not (Tiede et al., 2000), it is less easy to categorize articulators into critical and non-critical for breathing. In the context of speech, however, differences should be small and rt-MRI has been used for studying articulatory settings (Ramanarayanan et al., 2013) and pause postures (Krivokapić et al., 2022).

Ideally, acoustic measures could also be correlated to articulatory measures for two purposes: First, one could verify the findings on their acoustics. However, the challenge here is that some of these measurements, e.g. rt-MRI, severely affect the acoustic quality of the recording, making EMA a more appropriate approach here. Second, further support for correlating acoustic and articulatory, or generally physiological measures, could advance respiration research, as future studies could cut out the middleman of articulatory data and make these inferences directly from the acoustics. This could be done similarly to Nallanthighal et al. (2021), but also extended to other kinematic measures, thus advancing telediagnosis. Here, advances in the automatic processing of kinematic data for speech breathing make it possible to work with larger amounts of speech breathing data (MacIntyre & Werner, 2023).

Another aspect that has largely been neglected here is that of coarticulation. Speech inhalations may be affected by the speech surrounding them, similar to how speech sounds influence each other (Farnetani & Recasens, 2010). Although Lester & Hoit (2014) found the phonetic context to not affect the type of inhalation, e.g. simultaneous oral-nasal inhalation, it may still affect more fine-grained aspects, such as tongue movement. As this aspect has not been researched yet, it remains an open question whether this could also work vice versa, with inhalation gestures exerting coarticulatory influence on the surrounding speech segments. Whalen & Sheffert (1996) and Whalen & Kinsella-Shaw (1997) did not find perceivable, coarticulatory information of surrounding vowels in breath noises. It has been observed in the acoustic domain in (Panne, 2020), but without clear articulatory patterns (Breitbarth, 2021).

A related aspect is adopting a dynamic view on speech inhalations: In this work, we treated breath noises as stable entities and did not look at how they may change throughout the inhalation. This was done to simplify the parameters, however, the articulators may move throughout the inhalation or move to an inhalation target (similar to pause postures, see ch. 2.1.7), where they remain stable, and then move back towards the end. This would also entail a dynamic nature of the breath noise. Fink (1974) stated that "[i]n general the larynx and the lungs are constantly varying in shape and size with respiration". This could be complemented by a number of supraglottal processes, such as jaw lowering, tongue movement, and velar opening.

As phonetic parameters of speech, such as formants or spectral characteristics, have been more and more treated as dynamic rather than stable recently (see e.g. Kirkham et al. 2019; Carignan et al. 2020; Sóskuthy 2021; Wikse Barrow et al. 2022), such an approach may further enrich our understanding of speech breathing. Now that the general spectral characteristics of breath noises have been described for a large number of speakers here, future work could test what parameters affect it, treating the whole spectrum, and not just extracted values, as a dependent variable, as done in Puggaard-Rode (2022).

In the course of studying the articulation of speech breathing, future studies could highlight the vocal tract configuration found in non-breath and breath pauses and try to tease apart the different settings. Potentially, there are targeted *inhalation postures* in accordance with pause postures or inhalation may play a larger role in pause postures than previously assumed. This could give more insight into the causes and articulatory correlates of the silent edges that show up in the acoustic signal around breath noises. This way, one could also examine very short edges as opposed to the longer ones we found in these studies. Speculatively, the underlying causes for them might differ with the short ones only emerging for reasons of motor control and the longer ones due to planning reasons. These potential differences could surface as different articulatory executions.

Since the spectrum of speech inhalations can be influenced by a number of factors, such as subglottal resonances or degree of nasal contribution, future research dealing with speech breathing could look more into this. If one wants to keep using the audio signal, one could restrict the usage of the nasal or oral airway, e.g. with a nose clip or a closed mouth, and then compare those with each other and with simultaneous oral-nasal inhalations. Data recorded via this method could be used to analyze the different types of inhalations and to measure the degree of nasal contribution in simultaneous oral-nasal inhalations, thus building on and validating Lester & Hoit (2014). This would, however, only indirectly show how much nasality affects the spectrum in inhalations. Alternatively, one could make use of the separation of microphone channels for nasal and oral airflow in the nasometer. This would result in a more realistic picture of airway contribution to speech breathing. A general shortcoming of this work is that large parts of it are exploratory and thus require further validation by future studies. Ideally, this work will lead to more research in the directions taken here, but also of pauses, PINTS, and speech breathing in general.

Chapter 8

Conclusion

The present work contributes to a better understanding of speech pauses and of the breath noises found therein in particular. We first considered existing corpora and how they annotated pauses and pause-internal particles (PINTS). Then, a recording setup to study speech breathing across different conditions was presented. For studies on pauses, this work has added more detail to their optionality and variability, as well as their temporal composition when involving breathing. While previous studies on speech breathing have largely focused on parameters related to lung behavior, we have put the acoustic dimension of breathing, i.e. *breath noises*, in the focus of investigation. We simulated in- and exhalations in 3D-printed vocal tract models, described the spectral characteristics of naturally occurring breath noises, and tried to approach their related vocal tract configuration. Further, we explored how much information listeners can extract from breath noises with regard to how and by whom they were produced.

In the third chapter, we discussed two types of data: existing corpora and their annotation regarding pauses and PINTS and pilot recordings. We found that most of them did indicate pauses and even filler particles and breath noises were frequently considered. How exactly that was implemented in terms of segmentation and symbols used differed, however, which was why we advocate for a standardized framework. The pilot recordings exemplified a recording setup that may be useful for investigating speech breathing in different conditions, as well as the role of creaky voice in the respiratory cycle and filled pauses.

The fourth chapter shed further light on the plasticity of pauses. It has been demonstrated how pause number and duration change as a function of speech tempo. Comparing the tempo steps to synthetic speech produced at default speed allowed us to unravel how fitting pauses to speech can be improved there. Further, the speech tempo was used to emphasize the optionality and variability of pauses. On the other hand, several pauses related to punctuation and conjunctions were found to be less optional, longer, and more likely to involve breathing. Variability of duration was high there and related to speech tempo.

In the fifth chapter, the focus was put on breath noises and their acoustic characteristics. First, we explored the coupling of the acoustic dimension of breathing with its kinematic dimension, for which we used volume changes in the torso as a proxy for in- and exhalation. We found close coupling of the inhalation noise onset with both rib cage and abdomen expansion. Second, we set out to describe the spectral characteristics of the breath noises and relate them to the kinematics. To this end, we compared them to speech sounds using several parameters to approach the vocal tract configuration assumed in speech breathing. We found inhalations to be similar to /ə/ but more open and slightly more front. In addition, we found several parameters of the breath noise to be positively correlated with how fast speakers inhaled. Third, we described the entire spectrum of breath noises in a large sample of human inhalations. The spectra were relatively flat with decreasing slopes and weak peaks, that may also suggest a vocal tract configuration similar to /ə/. We also used 3D-printed vocal tract models to simulate noises of in- and exhalation by reversing the direction of airflow. We found that this affected sounds with a high tongue position more than those with a more open configuration, which we also assume for speech breathing. Finally, the inhalations produced with the vocal tract model representing /ə/ were compared to human inhalation. The spectra did show some similarities but differences between human and model inhalation complicated this comparison.

The sixth chapter dealt with the perception of breath noises. First, we tested how well listeners could auditorily categorize breath noises into different types, such as oral inhalation or nasal-oral exhalation. From a set of six different types, they were able to give the correct answer in around two thirds of all cases. Adding speech context around the breath noise did not help and phoneticians did not outperform lay people on this task. Some types, such as nasal and oral inhalation, were correctly answered more often than others. Second, we aimed to test how well listeners could use breath noises to discriminate between different speakers and to extract information to very roughly classify the speaker, such as old or young age and male or female sex. Discrimination was correct in slightly less than two thirds of all cases, with sex being more detectable than age. In the classification task, sex was correctly detected in two thirds of all cases, while age detection was random. In both, sex was thus more perceivable than age.

This work adds to research on pauses by stressing their optionality and variability, as well as the temporal composition when containing breath noises. It further contributes to research on speech breathing, for which we mainly used audio and the kinematic signal of torso expansion. In the acoustic domain, we have made a first extensive description of inhalation spectra. In the future, further factors influencing them should be analyzed, such as different speaker groups, pathologies, and emotional or physical stress. Beyond that, the articulatory execution of breathing in the speech stream still holds great potential. Combining the two domains could help validate our findings and inferences that were formed on the acoustic signal and pave the way for a better understanding of the role of supraglottal articulators in speech breathing. The findings from that could then be used to test how supraglottal articulation influences the resulting breath noises. While in this work, we mainly used single values of acoustic parameters to represent a given breath noise, thus assuming some stability in them, in the future a more dynamic view could be adopted. This could also shed further light on potential coarticulatory influence. With several parts of the body involved and a number of factors influencing it, speech breathing offers many angles and multiple ways of measuring.

Moreover, the perception of breath noises merits further research. This includes how much information listeners can extract from them, how they may aid listeners in processing speech, and to what degree they may be helpful or usable in synthetic speech. For the latter, it may also aid in designing the most appropriate speech breathing for the respective style. For forensic phonetics, it can help paint a realistic picture of what listeners can and cannot extract from them.

This work thus adds to the existing body of research by employing an in-depth analysis of pauses, as well as the breath noises within them. This includes the temporal composition of pauses and breath noises, the plasticity of pauses, as well as the spectral characteristics of breath noises, how they are related to physiology, and what information they can cue for listeners. We hope that this work will motivate further work on pauses and speech breathing!

Appendices

Appendix A

Appendix

A Nordwind und Sonne

Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam. Sie wurden einig, dass derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind den Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus. Da musste der Nordwind zugeben, dass die Sonne von ihnen beiden der Stärkere war.

B The North Wind and the Sun

The North Wind and the Sun were disputing which was the stronger when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shone out warmly and immediately the traveller took off his cloak. And so the North wind was obliged to confess that the Sun was the stronger of the two.

C Picture description: Baseball

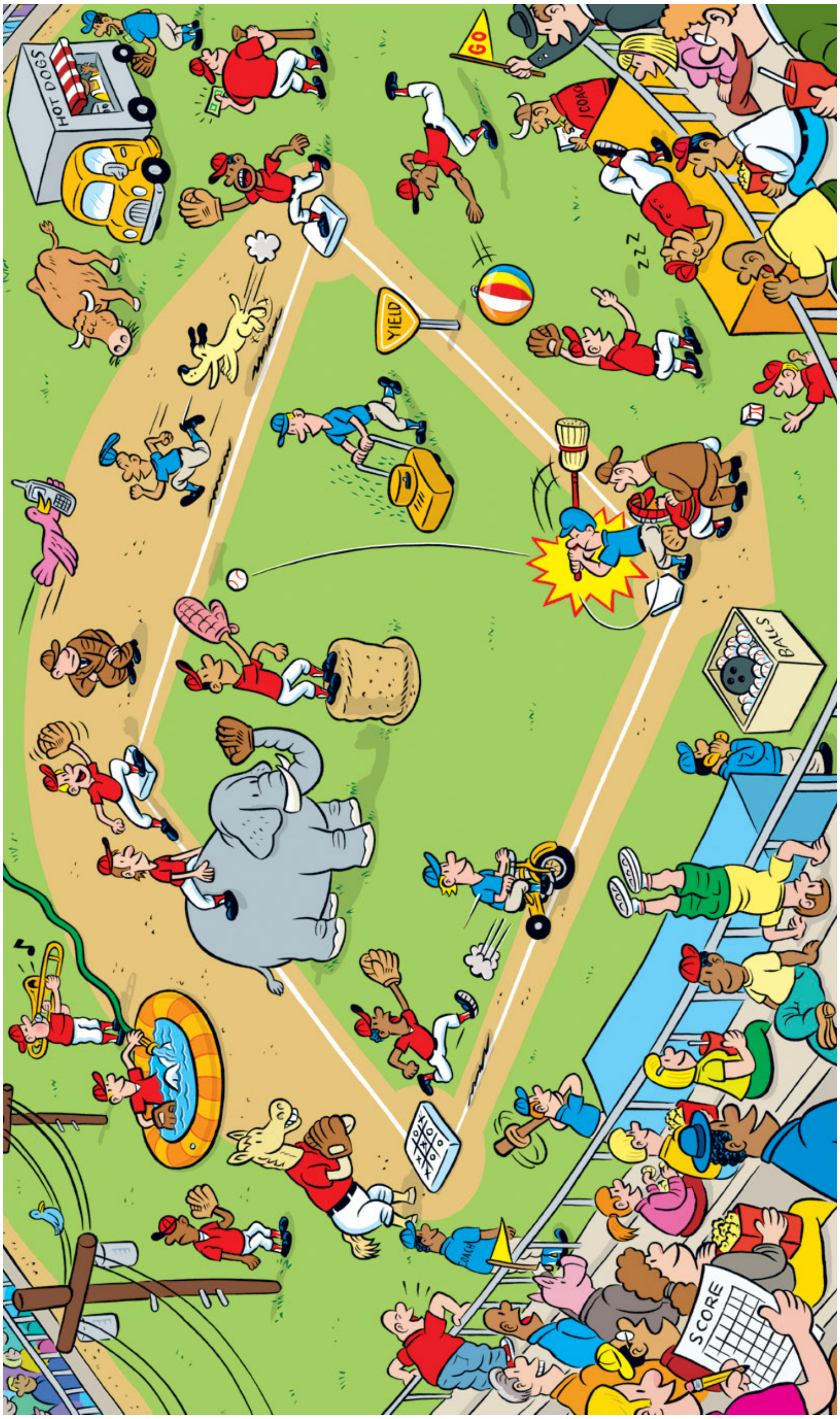

Figure A.1: Downloaded from https://www.behance.net/gallery/1341531/Whats-Wrong-with-these-pictures.

Appendix B

Appendix

A By-speaker inhalation spectra

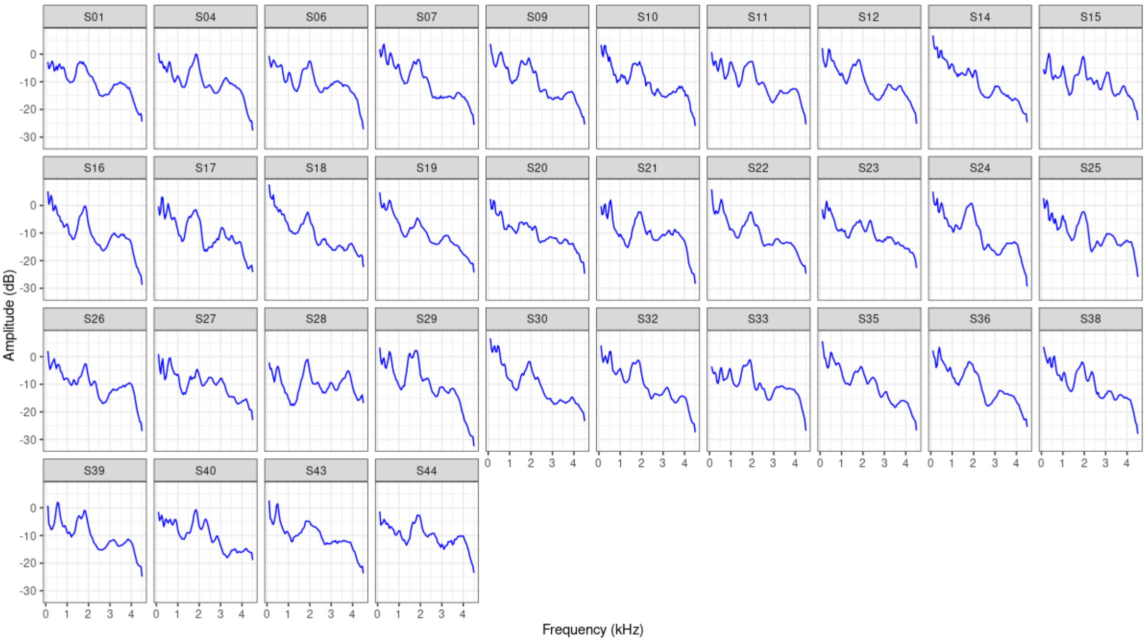

Figure B.1: Inhalation spectra averaged by speaker for the female data (100-4,500 Hz)

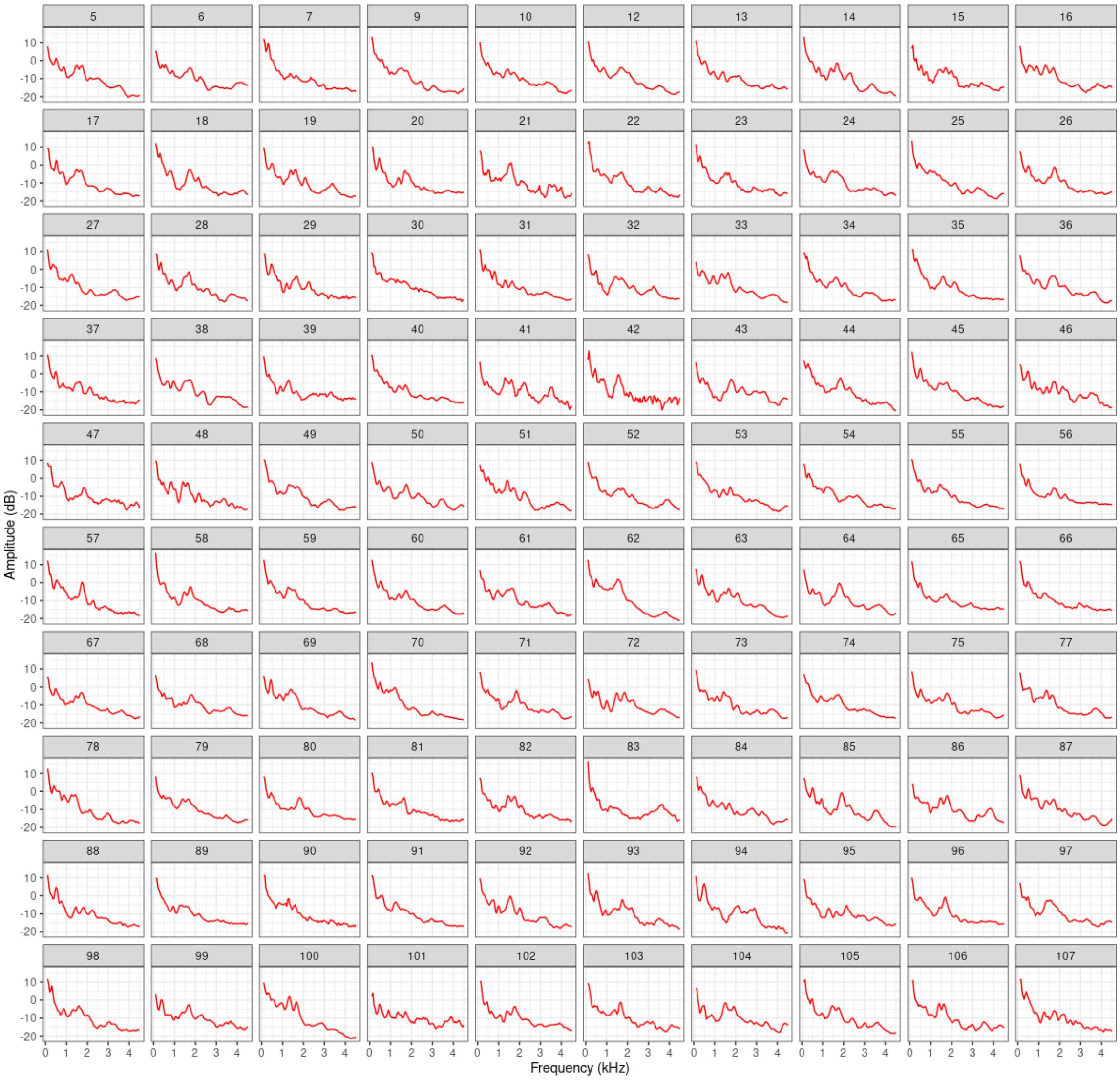

Figure B.2: Inhalation spectra averaged by speaker for the male data (100-4,500 Hz)

List of Figures

2.1 Spectrogram of a section from the Pool2010 corpus (Jessen et al., 2005) containing an inhalation noise located in a speech pause surrounded by silent edges. 10

2.2 Respiratory kinematics for different breathing tasks. Tidal or rest breathing, vital capacity, i.e. maximal in- and exhalation, and speech breathing. The y-axis shows lung volume and the x-axis time. Figure taken from Fuchs & Rochet-Capellan (2021). 18

3.1 RespTrack SUM signals for the five conditions from S1. The x-axis shows time in seconds, the y-axis is in arbitrary units from -1 to 1 for torso expansion, scaled to the respective window. From top to bottom it includes 1) quiet reading, 2) normal reading, 3) reading with as few pauses as possible, and 4) reading after physical activity. 44

3.2 The RespTrack SUM signal for the four reading task conditions from Subject2 analogous to Fig. 3.1. 45

3.3 The RespTrack SUM signal for the two (semi-)spontaneous speech tasks by subject. The upper half shows S1, the lower half S2. Within a participant, the upper part (a) shows the Retelling and the lower one (b) the PictureDescription task. 47

3.4 The RespTrack SUM signal for the final 6.7 s of the Retelling task as produced by S1. 47

4.1 The German BTC text read by participants. Square brackets indicate locations where at least one speaker paused. Numbers in brackets represent number of total pause locations used in that language, not number of pauses used in a given location. 52

4.2 Comparison of articulation durations (art_dur) and pause durations (pau_dur) by languages and speech tempos or TTS systems respectively. Speech duration can be seen here as their sum. 55

4.3 Number of pauses with (br_pau) and without (nbr_pau) audible breath noise by language and speech tempos, as well as in TTS systems. . . 56
4.4 Articulation rate (in syllables per second, on flipped y-axis) and relative pausing time (pause duration divided by articulation duration) in real speakers (colored) and TTS systems (black). Language is indicated by shape, for natural speakers colors indicate tempo. 57
4.5 Visualized pausing behavior of one German speaker (Dd_11): the long, unfilled squares represent the entire speech duration for the respective tempo, the dark blue squares within them represent the duration of a given breath pause. The length of all boxes roughly corresponds to their durations. Connections between pauses over tempos signify that the same location was used for a pause (see Fig. 4.1 for pause locations). 58
4.6 Pausing behavior of another German speaker (Dd_10): the squares hatched in light-blue represent non-breath pauses (i.e. without audible breath noise). When pauses are connected by a dashed line, it means that they were not used in the next tempo level but picked up again in a higher tempo. The rest is equivalent to Fig. 4.5 59
4.7 Pauses and their durations by location in the text, sorted by language. Pause location is determined by the number of the preceding word. Whether a pause contains an inhalation is indicated by filled dots (inhalation) or empty, crossed dots (no inhalation). Dashed lines indicate locations with punctuation <, ; .>. 64
4.8 Inhalation duration relative to pause duration split by rate for all languages. The dashed reference line indicates where both durations would be equally long. Color indicates tempo, black dots represent all inhalations from the other tempos for reference. The dots on the x-axis (where y=0 s) are non-breath pauses. 66
4.9 Duration of left vs. right edge in breath pauses in all six languages pooled together. The dashed reference line indicates equal durations. Five breath pauses (all from the same speaker) with left edges between 1.3 s and 2.1 s. were excluded from the plot. 67

5.1 The temporal alignment of articulation phases, inhalation noises and AB and RC in an inhalation phase between two articulation phases in a speech signal (left, for numbers see text, speaker S04). 74
5.2 A schematic view of the temporal alignment of articulation phases (ART) and the abdominal (AB) and rib cage expansion (RC) in relation to an inhalation noise (INHN) in speech. 74
5.3 Exhalation (light grey) and inhalation phases (black) in two 30-sec excerpts from the inspected kinematic data of one speaker in both conditions (top: LN, bottom: SN). 74

5.4 Differences between durations and start/end times of inhalation in the two conditions. Boxes 1 and 2 show differences between duration of AB minus duration of RC (Δdur), boxes 3 and 4 show differences between starting times of AB minus RC (Δstart), and boxes 5 and 6 show differences between end times of AB minus RC (Δend). Positive values thus show a longer/later AB, while negative values indicate a longer/later RC. 76
5.5 Differences between starting times (start) and end times (end) of inhalation noises, acoustic articulation, AB breathing, and RC breathing. In 'startABRC' the value is taken from the one starting later and in 'endABRC' from the one ending earlier to have both synchronously expanding. 77
5.6 Corresponding sections of silence (edges) left and right of an audible inhalation noise. 78
5.7 Normalized intensity (dB) of the respective segments. The y-axis is reversed here, since lower values represent higher intensity due to normalization. 84
5.8 Center of Gravity of the respective segments in kHz. 85
5.9 F1-F2 vowel chart for inhalations within speech (*inh*, orange) and outside of speech (*n-inh*, blue) compared to /ə/ (green). Ellipses refer to density. 85
5.10 Correlations between inhalation slope and normalized intensity (norm_int, left), inhalation slope and CoG (middle), and inhalation slope and F1 (right) for all *inh* data. All speakers are pooled together. To account for different anatomical properties, data were z-transformed by speaker. norm_int was not transformed as it had already undergone a normalization procedure; its y-axis is reversed to visualize the underlying positive correlation (cf. Fig. 5.7 and ch. 5.2.2). 87
5.11 Unnormalized inhalation spectra averaged by sex for female (100–4,500 Hz) and male participants (100–8,000 Hz). 93
5.12 All human inhalation spectra for female (blue) and male (red) data. The average spectrum per sex is overlaid in bold. 94
5.13 Averaged human inhalation spectra for female (blue) and male (red) data. The regions shaded in grey indicate resonance bands for inhalations with the VT configuration of the vowel /ɜː/, as reported for male speakers in Table I and female speakers in Table II by Hanna et al. (2018). 95
5.14 Two of the 3D-printed vocal tracts corresponding to a male speaker producing the sounds /aː/ (left) and /ʃ/ (right). 96
5.15 Spectra (100—10,000 Hz) for exhalation (black) and inhalation (red) by VT configuration and VT model (male and female). Every spectrum averages over three power levels. 97

5.16 Human inhalation spectra (green; all spectra and average) vs. model inhalation with a /ə/ VT configuration (orange; averaged over three power levels each). The data are split by sex. 99

6.1 Correct assessment of stimuli by the breath type. Results from Experiment 1 are plotted on the left, those from Experiment 2 are plotted on the right. 107

6.2 Alluvial plot of the breath type of the stimulus (left) and type of the response (right). For correctly assessed stimuli, stimulus and response are the same, whereas otherwise the response will be something else. . 109

6.3 Plot showing the model predictions for the discrimination task (outcome: discrimination correct in % on the y-axis) by the relation of the speaker age in the two breath noises (in color) and the speaker sex (on the x-axis). The points are averages, the lines around them visualize the 95% confidence intervals. 116

6.4 Plot showing the model predictions for the classification task regarding speaker sex (outcome: classification of speaker sex correct in % on the y-axis) by speaker age (in color) and speaker sex (on the x-axis). . . . 117

6.5 Alluvial plot for the four groups defined by age and gender. The boxes on the left side show the stimuli, the boxes on the right the responses. 119

A.1 Downloaded from https://www.behance.net/gallery/1341531/Whats-Wrong-with-these-pictures. 139

B.1 Inhalation spectra averaged by speaker for the female data (100-4,500 Hz) 140

B.2 Inhalation spectra averaged by speaker for the male data (100-4,500 Hz) 141

List of Tables

2.1 Impedance minima corresponding to resonances (in Hz; mean and standard deviation) reported for 7 male and 3 female participants inhaling with a VT configuration of /ɜː/ in Table I and II in Hanna et al. (2018). 27

3.1 Overview of the corpora and the four inspected categories. *y* is for yes, this category or aspects of it are represented; for *n* that is not the case. *?* is for unclear cases. 37

4.1 Articulation duration (in seconds): human speakers are averaged over by the five speech tempos (very slow, slow, normal, fast, very fast); the synthetic speech side reports the actual number of the respective system (IWTTS, GCTTS, GT, OCTTS). All are split by language. . 54

4.2 Pause duration in seconds for natural and synthetic speakers. 54

4.3 Mean number of pauses produced by natural speakers in the five speech tempos and number of pauses produced by the TTS systems, split by the three languages. 55

4.4 Number of pauses produced that included audible breath noises by languages and tempos. For natural speakers, the numbers are averaged over speakers. 56

4.5 Overview of material used from the corpora: number of speakers, words, and syllables by language. 63

4.6 Number of occurrence of breath and non-breath pauses by rate, pooled for all languages. 65

5.1 *Overview of the respiratory time quotient (RTQ) for abdomen (AB) and rib cage (RC), mean durations of breath cycles, and number of mean breath cycles per minute while speaking and listening.* 73

5.2 Mean duration and standard deviation of breathing events and speech segments (*h, k-asp, t-asp*) without /ə/ in ms. 83

5.3 Significant contrasts between airflow direction by VT configuration in the VT models (exhalation - inhalation). 98

6.1 Overview of the stimuli used in Experiment 2, regarding their duration (dur.) and intensity (int.). We used 4 breath noises for each type. . . 106
6.2 The breath noises used as stimuli in the experiment. 113
6.3 Correctness rate of the discrimination task by speaker sex and age in percent. The numbers in brackets indicate number of stimuli per cell. 115

Bibliography

Aare, K., Gilmartin, E., Włodarczak, M., Lippus, P., & Heldner, M. (2020). Breath holds in chat and chunk phases of multiparty casual conversation. In *Speech Prosody* (pp. 779–783). URL: https://doi.org/10.21437/SpeechProsody.2020-159.

Aare, K., Lippus, P., Włodarczak, M., & Heldner, M. (2018). Creak in the respiratory cycle. In *Interspeech* (pp. 1408–1412). URL: https://doi.org/10.21437/Interspeech.2018-2165.

Adams, C., & Munro, R. R. (1973). The relationship between internal intercostal muscle activity and pause placement in the connected utterance of native and non native speakers of English. *Phonetica*, *28*, 227–250. URL: https://doi.org/10.1159/000259457.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory* (pp. 267–281). URL: https://doi.org/10.1007/978-1-4612-1694-0_15.

Allen, M., Varga, S., & Heck, D. H. (2022). Respiratory rhythms of the predictive mind. *Psychological Review*, (pp. 1–52). URL: https://doi.org/10.1037/rev0000391.

Anikin, A., & Reby, D. (2022). Ingressive phonation conveys arousal in human nonverbal vocalizations. *Bioacoustics*, *31*, 680–695. URL: https://doi.org/10.1080/09524622.2022.2039295.

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, *63*, 1535–1555. URL: https://doi.org/10.1121/1.381848.

Augousti, A. T. (1997). 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*, 283–291. URL: https://doi.org/10.1088/0031-9155/42/2/002.

Bailly, G., & Gouvernayre, C. (2012). Pauses and respiratory markers of the structure of book reading. In *Interspeech* (pp. 2218–2221). URL: https://doi.org/10.21437/Interspeech.2012-591.

Barbier, G., Boë, L.-J., Captier, G., & Laboissière, R. (2015). Human vocal tract growth: a longitudinal study of the development of various anatomical structures. In *Interspeech* (pp. 364–368). URL: https://doi.org/10.21437/Interspeech.2015-156.

Barbosa, P. A., Niebuhr, O., & Neitsch, J. (2019). Revisiting rhetorical claims of breathing for persuasive speech. In *1st International Seminar on the Foundations of Speech (SEFOS)* (pp. 103–105).

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*. URL: https://doi.org/10.18637/jss.v067.i01.

Belz, M. (2017). Glottal filled pauses in German. In *Proceedings of Disfluency in Spontaneous Speech* (pp. 5–8). Stockholm.

Belz, M. (2019). Dokumentation der Forschungsdaten GECO-FP v.1. URL: https://doi.org/10.18452/20794.

Belz, M. (2021). *Die Phonetik von äh und ähm. Akustische Variation von Füllpartikeln im Deutschen.*. Berlin: J.B. Metzler. URL: http://link.springer.com/10.1007/978-3-662-62812-6.

Belz, M., Sauer, S., Lüdeling, A., & Mooshammer, C. (2017). Fluently disfluent? Pauses and repairs of advanced learners and native speakers of German. *International Journal of Learner Corpus Research*, *3*, 118–148. URL: https://doi.org/10.1075/ijlcr.3.2.02bel.

Belz, M., & Trouvain, J. (2019). Are 'silent' pauses always silent? *International Congress of Phonetic Sciences (ICPhS)*, (pp. 2744–2748).

Benchetrit, G. (2000). Breathing pattern in humans: Diversity and individuality. *Respiration Physiology*, *122*, 123–129. URL: https://doi.org/10.1016/S0034-5687(00)00154-7.

Benchetrit, G., Shea, S. A., Dinh, T. P., Bodocco, S., Baconnier, P., & Guz, A. (1989). Individuality of breathing patterns in adults assessed over time. *Respiration Physiology*, *75*, 199–209. URL: https://doi.org/10.1016/0034-5687(89)90064-9.

Bernardet, U., Kanq, S.-H., Feng, A., DiPaola, S., & Shapiro, A. (2019). Speech breathing in virtual humans: An interactive model and empirical study. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)* (pp. 1–9). IEEE. URL: https://doi.org/10.1109/VHCIE.2019.8714737.

Betz, S., Bryhadyr, N., Türk, O., & Wagner, P. (2023). Cognitive load increases spoken and gestural hesitation frequency. *Languages*, *8*, 71. URL: https://doi.org/10.3390/languages8010071.

Binazzi, B., Lanini, B., Bianchi, R., Romagnoli, I., Nerini, M., Gigliotti, F., Duranti, R., Milic-Emili, J., & Scano, G. (2006). Breathing pattern and kinematics in normal subjects during speech, singing and loud whispering. *Acta Physiologica*, *186*, 233–246. URL: https://doi.org/10.1111/j.1748-1716.2006.01529.x.

Birkholz, P., Dächert, P., & Neuschaefer-Rube, C. (2012). Advances in combined electro-optical palatography. In *Interspeech* (pp. 703–706). URL: https://doi.org/10.21437/Interspeech.2012-220.

Birkholz, P., Gabriel, F., Kürbis, S., & Echternach, M. (2019). How the peak glottal area affects linear predictive coding-based formant estimates of vowels. *The Journal of the Acoustical Society of America*, *146*, 223–232. URL: https://doi.org/10.1121/1.5116137.

Birkholz, P., Hasner, P., & Kurbis, S. (2022). Acoustic comparison of physical vocal tract models with hard and soft walls. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8242–8246). URL: https://doi.org/10.1109/ICASSP43922.2022.9746611.

Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., & Fleischer, M. (2020). Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties. *Scientific Data*, *7*, 1–16. URL: https://doi.org/10.1038/s41597-020-00597-w.

Birkholz, P., Stone, S., Wolf, K., & Plettemeier, D. (2018). Non-invasive silent phoneme recognition using microwave signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*, 2404–2411. URL: https://doi.org/10.1109/TASLP.2018.2865609.

de Boer, M. M., & Heeren, W. F. L. (2020). Cross-linguistic filled pause realization: The acoustics of *uh* and *um* in native Dutch and non-native English. *The Journal of the Acoustical Society of America*, *148*, 3612–3622. URL: https://doi.org/10.1121/10.0002871.

de Boer, M. M., Quené, H., & Heeren, W. F. L. (2022). Long-term within-speaker consistency of filled pauses in native and non-native speech. *JASA Express Letters*, *2*, 035201. URL: https://doi.org/10.1121/10.0009598.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer. URL: http://www.praat.org/.

Boliek, C. A., Hixon, T. J., Watson, P. J., & Jones, P. B. (2009). Refinement of speech breathing in healthy 4- to 6-year-old children. *Journal of Speech, Language, and Hearing Research*, *52*, 990–1007. URL: https://doi.org/10.1044/1092-4388(2009/07-0214).

Bosker, H. R., van Os, M., Does, R., & van Bergen, G. (2019). Counting 'uhm's: How tracking the distribution of native and non-native disfluencies influences online language comprehension. *Journal of Memory and Language*, *106*, 189–202. URL: https://doi.org/10.1016/j.jml.2019.02.006.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 159–175. URL: https://doi.org/10.1177/0265532212455394.

Braun, A. (2020). Nonverbal vocalisations – a forensic phonetic perspective. In *Laughter and Other Non-Verbal Vocalisations Workshop* (pp. 19–23).

Braunschweiler, N., & Chen, L. (2013). Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *ISCA Speech Synthesis Workshop* (pp. 1–6). Barcelona. URL: http://ssw8.talp.cat/papers/ssw8_OS1-1_Braunschweiler.pdf.

Breitbarth, Z. (2021). Eine Ultraschall-Analyse der Zungenbewegungen in Sprechpausen (Unpublished MA-Thesis).

Butcher, A. (1981). *Aspects of the speech pause: Phonetic correlates and communicative functions*. Ph.D. thesis Universität Kiel.

Cala, S. J., Kenyon, C. M., Ferrigno, G., Carnevali, P., Aliverti, A., Pedotti, A., Macklem, P. T., & Rochester, D. F. (1996). Chest wall and lung volume estimation by optical reflectance motion analysis. *Journal of Applied Physiology*, *81*, 2680–2689. URL: https://doi.org/10.1152/jappl.1996.81.6.2680.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Proceedings Speech Prosody 2002* (pp. 199–202). URL: https://doi.org/10.1.1.12.561.

Candea, M. (2008). *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. Etude sur un corpus de récits en classe de français*. Ph.D. thesis Université de la Sorbonne nouvelle - Paris III. URL: https://theses.hal.science/tel-00290143.

Candea, M., Vasilescu, I., & Adda-Decker, M. (2005). Inter- and intra-language acoustic analysis of autonomous fillers. In *Disfluency in Spontaneous Speech Workshop* (pp. 47–51).

Carignan, C. (2018). Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels. *The Journal of the Acoustical Society of America*, *143*, 2588–2601. URL: https://doi.org/10.1121/1.5034760.

Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., Frahm, J., & Harrington, J. (2020). Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Laboratory Phonology*, *11*. URL: https://doi.org/10.5334/LABPHON.214.

Carlson, R., Hirschberg, J., & Swerts, M. (2005). Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, *46*, 326–333. URL: https://doi.org/10.1016/j.specom.2005.02.013.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Edinburgh University Press. URL: https://doi.org/10.2307/412751.

Catford, J. C. (2001). *A practical introduction to phonetics*. (2nd ed.). Oxford University Press.

Chauhan, J., Hu, Y., Seneviratne, S., Misra, A., Seneviratne, A., & Lee, Y. (2017). BreathPrint. In *Annual International Conference on Mobile Systems, Applications, and Services* (pp. 278–291). URL: https://doi.org/10.1145/3081333.3081355.

Chen, X., Liesenfeld, A., Li, S., & Yao, Y. (2022). Effects of filled pauses on memory recall in human-robot interaction in Mandarin Chinese. In D. Harris, & W.-C. Li (Eds.), *HCII 2022, LNAI 13307* (pp. 3–17). URL: https://link.springer.com/10.1007/978-3-031-06086-1_1.

Clark, H. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111. URL: https://doi.org/10.1016/S0010-0277(02)00017-3.

Conrad, B., & Schönle, P. (1979). Speech and respiration. *Archiv für Psychiatrie und Nervenkrankheiten*, *226*, 251–268. URL: https://doi.org/10.1007/BF00342238.

Conrad, B., Thalacker, S., & Schönle, P. (1983). Speech respiration as an indicator of integrative contextual processing. *Folia Phoniatrica et Logopaedica*, *35*, 220–225. URL: https://doi.org/10.1159/000265766.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668. URL: https://doi.org/10.1016/j.cognition.2006.10.010.

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, *5*. URL: https://doi.org/10.1126/sciadv.aaw2594.

Cravotta, A., Prieto, P., & Busà, M. G. (2021). Exploring the effects of restraining the use of gestures on narrative speech. *Speech Communication*, *135*, 25–36. URL: https://doi.org/10.1016/j.specom.2021.09.005.

Dellwo, V., Aschenberner, B., Wagner, P., Dancovicova, J., & Steiner, I. (2004). BonnTempo-corpus and Bonntempo-Tools: A database for the study of speech rhythm and rate. In *Interspeech* (pp. 777–780). URL: https://doi.org/10.21437/Interspeech.2004-294.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 655–679. URL: https://doi.org/10.1111/j.1467-9922.2004.00282.x.

Duez, D. (1993). Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research*, *22*, 21–39. URL: https://doi.org/10.1007/BF01068155.

Dumpala, S. H., & Alluri, K. N. R. K. R. (2017). An algorithm for detection of breath sounds in spontaneous speech with application to speaker recognition. In A. Karpov, R. Potapova, & I. Mporas (Eds.), *Speech and Computer. SPECOM.* (pp. 98–108). volume 10458 LNAI. URL: https://doi.org/10.1007/978-3-319-66429-3{_}9.

Eckart, K., Riester, A., & Schweitzer, K. (2012). A discourse information radio news database for linguistic analysis. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked Data in Linguistics* (pp. 65–76). Springer. URL: https://doi.org/10.1007/978-3-642-28249-2_7.

Eklund, R. (2004). *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Ph.D. thesis Linköping, Sweden. URL: http://www.anst.uu.se/a/pehel169/Computer_tools_course/Eklund_Thesis.pdf.

Eklund, R. (2008). Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *Journal of the International Phonetic Association*, *38*, 235–324. URL: https://doi.org/10.1017/S0025100308003563.

Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. (2021a). Evaluating the effect of pauses on number recollection in synthesized speech. In *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 289–295). Berlin: TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=1131.

Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. (2021b). Take a breath: Respiratory sounds improve recollection in synthetic speech. In *Interspeech* (pp. 3196–3200). Brno. URL: https://doi.org/10.21437/Interspeech.2021-1496.

Farnetani, E., & Recasens, D. (2010). Coarticulation and connected speech processes. In W. Hardcastle, F. Gibbon, & J. Laver (Eds.), *The Handbook of Phonetic Sciences: Second Edition* (pp. 316–352). URL: https://doi.org/10.1002/9781444317251.ch9.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, *100*, 233–253. URL: https://doi.org/10.1037/0033-295X.100.2.233.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*, 1151–1177. URL: https://doi.org/10.1080/01690960701461293.

Filippelli, M., Pellegrino, R., Iandelli, I., Misuri, G., Rodarte, J. R., Duranti, R., Brusasco, V., & Scano, G. (2001). Respiratory dynamics during laughter. *Journal of Applied Physiology*, *90*, 1441–1446. URL: https://doi.org/10.1152/jappl.2001.90.4.1441.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. In *2017 International Conference on Computational Social Science IC2S2*.

Fink, B. R. (1974). Folding mechanism of the human larynx. *Acta Oto-Laryngologica*, *78*, 124–128. URL: https://doi.org/10.3109/00016487409126336.

Fletcher, J. (1987). Some micro and macro effects of tempo change on timing in French. *Linguistics*, *25*, 951–968. URL: https://doi.org/10.1515/ling.1987.25.5.951.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences: Second Edition* (pp. 523–602).

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, *29*, 320–326.

Franz, I., Knoop, C. A., Kentner, G., Rothbart, S., Kegel, V., Vasilieva, J., Methner, S., Scharinger, M., & Menninghaus, W. (2022). Prosodic phrasing and syllable prominence in spoken prose. A validated coding manual. URL: https://doi.org/10.31219/osf.io/h4sd5.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*, 161–175. URL: https://doi.org/10.1016/j.jml.2011.03.004.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*, 41–49.

Fry, D. B. (1979). *The physics of speech*. Cambridge: Cambridge University Press. URL: https://doi.org/10.1017/CBO9781139165747.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*, 29–47. URL: https://doi.org/10.1016/j.wocn.2012.08.007.

Fuchs, S., Petrone, C., Rochet-Capellan, A., Reichel, U. D., & Koenig, L. L. (2015a). Assessing respiratory contributions to f0 declination in German across varying speech tasks and respiratory demands. *Journal of Phonetics*, *52*, 35–45. URL: https://doi.org/10.1016/j.wocn.2015.04.002.

Fuchs, S., Reichel, U., & Rochet-Capellan, A. (2015b). Changes in speech and breathing rate while speaking and biking. In *International Congress of Phonetic Sciences (ICPhS)*. Glasgow. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1005.pdf.

Fuchs, S., & Rochet-Capellan, A. (2021). The respiratory foundations of spoken language. *Annual Review of Linguistics*, *7*, 13–30. URL: https://doi.org/10.1146/annurev-linguistics-031720-103907.

Fukuda, T., Ichikawa, O., & Nishimura, M. (2018). Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. *Speech Communication*, *98*, 95–103. URL: https://doi.org/10.1016/j.specom.2018.01.008.

Gee, J. P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*, 411–458. URL: https://doi.org/10.1016/0010-0285(83)90014-2.

Gelfer, C. E., Harris, K. S., Collier, R., & Baer, T. (1983). Is declination actively controlled? In *Vocal Fold Physiology*.

Gerstenberg, A., Fuchs, S., Kairet, J. M., Frankenberg, C., & Schröder, J. (2018). A cross-linguistic, longitudinal case study of pauses and interpausal units in spontaneous speech corpora of older speakers of German and French. In *Proceedings of the International Conference on Speech Prosody* (pp. 211–215). Poznán. URL: https://doi.org/10.21437/SpeechProsody.2018-43.

Gick, B., Wilson, I., Koch, K., & Cook, C. (2004). Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, *61*, 220–233. URL: https://doi.org/10.1159/000084159.

Godde, E., Bailly, G., & Bosse, M.-L. (2021). Pausing and breathing while reading aloud: Development from 2nd to 7th grade in French speaking children. *Reading and Writing*, *35*, 1–27. URL: https://doi.org/10.1007/s11145-021-10168-z.

Gold, E., French, P., & Harrison, P. (2013). Clicking behavior as a possible speaker discriminant in English. *Journal of the International Phonetic Association*, *43*, 339–349. URL: https://doi.org/10.1017/S0025100313000248.

Goldman-Eisler, F. (1955). Speech-breathing activity — a measure of tension and affect during interviews. *British Journal of Psychology*, *46*, 53–63. URL: https://doi.org/10.1111/j.2044-8295.1955.tb00524.x.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*, 96–106. URL: https://doi.org/10.1080/17470215808416261.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, *4*, 232–237. URL: https://doi.org/10.1177/002383096100400405.

Gósy, M. (2014). BEA - A multifunctional Hungarian spoken language database. *Phonetician*, *105*, 50–61.

Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J. M., & Van den Bergh, O. (2016). Respiratory changes in response to cognitive load: A systematic review. *Neural Plasticity*, *2016*, 1–16. URL: https://doi.org/10.1155/2016/8146809.

Grice, M., & Baumann, S. (2009). An introduction to intonation – functions and models. In J. Trouvain, & U. Gut (Eds.), *Non-Native Prosody. Phonetic Description and Teaching Practice* (pp. 25–51). Berlin, New York: De Gruyter. URL: https://doi.org/10.1515/9783110198751.1.25.

Grosjean, F. (1979). A study of timing in a manual and a spoken language: American sign language and English. *Journal of Psycholinguistic Research*, *8*, 379–405. URL: https://doi.org/10.1007/BF01067141.

Grosjean, F. (1980). Comparative studies of temporal variables in spoken and sign languages: a short review. In H. Dechert, & M. Raupach (Eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler* (pp. 307–312). The Hague: Mouton.

Grosjean, F., & Collins, M. (1979). Breathing, Pausing and Reading. *Phonetica*, *36*, 98–114.

Grosjean, F., & Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica*, *26*, 129–156. URL: https://doi.org/10.1159/000259407.

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, *11*, 58–81. URL: https://doi.org/10.1016/0010-0285(79)90004-5.

Gut, U. (2012). The LeaP corpus. URL: https://doi.org/10.1075/hsm.14.03gut.

Hallani, M., Wheatley, J. R., & Amis, T. C. (2008). Enforced mouth breathing decreases lung function in mild asthmatics. *Respirology*, *13*, 553–558. URL: https://doi.org/10.1111/j.1440-1843.2008.01300.x.

Hanna, N., Smith, J., & Wolfe, J. (2016). Frequencies, bandwidths and magnitudes of vocal tract and surrounding tissue resonances, measured through the lips during phonation. *The Journal of the Acoustical Society of America*, *139*, 2924–2936. URL: https://doi.org/10.1121/1.4948754.

Hanna, N., Smith, J., & Wolfe, J. (2018). How the acoustic resonances of the subglottal tract affect the impedance spectrum measured through the lips. *The Journal of the Acoustical Society of America*, *143*, 2639–2650. URL: https://doi.org/10.1121/1.5033330.

Harari, D., Redlich, M., Miri, S., Hamud, T., & Gross, M. (2010). The effect of mouth breathing versus nasal breathing on dentofacial and craniofacial development in orthodontic patients. *Laryngoscope*, *120*, 2089–2093. URL: https://doi.org/10.1002/lary.20991.

Harnsberger, J. D., Shrivastav, R., Brown, W. S., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, *22*, 58–69. URL: https://doi.org/10.1016/j.jvoice.2006.07.004.

Hartman, D. E. (1979). The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders*, *12*, 53–61. URL: https://doi.org/10.1016/0021-9924(79)90021-2.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*, 458–484. URL: https://doi.org/10.1016/j.wocn.2005.10.001.

Heck, D. H., McAfee, S. S., Liu, Y., Babajani-Feremi, A., Rezaie, R., Freeman, W. J., Wheless, J. W., Papanicolaou, A. C., Ruszinkó, M., Sokolov, Y., & Kozma, R. (2017). Breathing as a fundamental rhythm of brain function. *Frontiers in Neural Circuits*, *10*, 1–8. URL: https://doi.org/10.3389/fncir.2016.00115.

Heeren, W. F. L. (2015). Vocalic correlates of pitch in whispered versus normal speech. *The Journal of the Acoustical Society of America*, *138*, 3800–3810. URL: http://dx.doi.org/10.1121/1.4937762.

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, *130*, 508–513. URL: https://doi.org/10.1121/1.3598457.

Heldner, M., Włodarczak, M., Branderud, P., & Stark, J. (2019). The RespTrack system. In *International Seminar on the Foundations of Speech (SEFOS 2019)* (pp. 16–18). Sønderborg.

Henderson, A., Goldman-Eisler, F., & Skarbek, A. (1965). Temporal patterns of cognitive activity and breath control in speech. *Language and Speech*, *8*, 236–242. URL: https://doi.org/10.1177/002383096500800405.

Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with "articulatory" pauses. *Language and Speech*, *26*, 203–214. URL: https://doi.org/10.1177/002383098302600302.

Hixon, T. J., Goldman, M. D., & Mead, J. (1973). Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung. *Journal of Speech and Hearing Research*, *16*, 78–115.

Hixon, T. J., Watson, P. J., Harris, F. P., & Pearl, N. B. (1988). Relative volume changes of the rib cage and abdomen during prephonatory chest wall posturing. *Journal of Voice*, *2*, 13–19. URL: https://doi.org/10.1016/S0892-1997(88)80052-3.

Hixon, T. J., Weismer, G., & Hoit, J. D. (2020). *Preclinical speech science: Anatomy, physiology, acoustics, and perception*. (3rd ed.).

Hodge, M. M., & Rochet, A. P. (1989). Characteristics of speech breathing in young women. *Journal of Speech, Language, and Hearing Research*, *32*, 466–480. URL: https://doi.org/10.1044/jshr.3203.466.

Hoit, J. D., & Hixon, T. J. (1987). Age and speech breathing. *Journal of Speech and Hearing Research*, *30*, 351–366. URL: https://doi.org/10.1044/jshr.3003.351.

Hoit, J. D., & Lohmeier, H. L. (2000). Influence of continuous speaking on ventilation. *Journal of Speech, Language, and Hearing Research*, *43*, 1240–1251. URL: https://doi.org/10.1044/jslhr.4305.1240.

Honikman, B. (1964). Articulatory settings. In D. Abercrombie, D. Fry, P. MacCarthy, N. C. Scott, & J. Trim (Eds.), *In Honour of Daniel jones* (pp. 73–84). London: Longman. URL: http://www.jbe-platform.com/content/journals/10.1075/hl.28.1.09jen.

Hoole, P., & Ziegler, W. (1997). A comparison of normals' and aphasics' ability to plan respiratory activity in overt and covert speech. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, *35*, 77–80. URL: http://www.phonetik.uni-muenchen.de/~hoole/pdf/resp_fipkm.pdf.

Huber, J. E. (2008). Effects of utterance length and vocal loudness on speech breathing in older adults. *Respiratory Physiology and Neurobiology*, *164*, 323–330. URL: https://doi.org/10.1016/j.resp.2008.08.007.

Huber, J. E., & Darling, M. (2011). Effect of Parkinson's disease on the production of structured and unstructured speaking tasks: Respiratory physiologic and linguistic Considerations. *Journal of Speech, Language, and Hearing Research*, *54*, 33–46. URL: http://pubs.asha.org/doi/10.1044/1092-4388%282010/09-0184%29.

Huber, J. E., Darling, M., Francis, E. J., & Zhang, D. (2012). Impact of typical aging and Parkinson's disease on the relationship among breath pausing, syntax, and punctuation. *American Journal of Speech-Language Pathology*, *21*, 368–379. URL: https://doi.org/10.1044/1058-0360(2012/11-0059).

Huber, J. E., & Stathopoulos, E. T. (2015). Speech Breathing Across the Life Span and in Disease. *The Handbook of Speech Production*, (pp. 11–33). URL: https://doi.org/10.1002/9781118584156.ch2.

Ibrahim, O., Asadi, H., Kassem, E., & Dellwo, V. (2020). Arabic speech rhythm corpus: Read and spontaneous speaking styles. In *International Conference on Language Resources and Evaluation (LREC)* (pp. 5337–5342). Marseille.

Ilerialkan, A., Temizel, A., & Hacıhabiboğlu, H. (2020). Speaker and posture classification using instantaneous Intraspeech breathing features. *CoRR*, *abs/2005.1*. URL: https://doi.org/https://doi.org/10.48550/arXiv.2005.12230.

Inada, E., Saitoh, I., Kaihara, Y., & Yamasaki, Y. (2021). Factors related to mouth-breathing syndrome and the influence of an incompetent lip seal on facial soft tissue form in children. *Pediatric Dental Journal*, *31*, 1–10. URL: https://doi.org/10.1016/j.pdj.2020.10.002.

IPDS (2006). *Video task scenario: Lindenstraße – the Kiel corpus of spontaneous speech*. Technical Report 4 Institut für Phonetik und Digitale Sprachsignalverarbeitung Universität Kiel.

Ishikawa, K., & Webster, J. (2023). The formant bandwidth as a measure of vowel intelligibility in dysphonic speech. *Journal of Voice*, *37*, 173–177. URL: https://doi.org/10.1016/j.jvoice.2020.10.012.

Jannedy, S., & Weirich, M. (2017). Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *The Journal of the Acoustical Society of America*, *142*, 395–405. URL: https://doi.org/10.1121/1.4991347.

Jenner, B. (2001). 'Articulatory setting': Genealogies of an idea. *Historiographia Linguistica*, *28*, 121–141.

Jessen, M. (2007). Speaker classification in forensic phonetics and acoustics. In *Speaker Classification I* (pp. 180–204). Berlin, Heidelberg: Springer Berlin Heidelberg. URL: http://link.springer.com/10.1007/978-3-540-74200-5_10.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12*, 174–213. URL: https://doi.org/10.1558/sll.2005.12.2.174.

Johannknecht, M., & Kayser, C. (2022). The influence of the respiratory cycle on reaction times in sensory-cognitive paradigms. *Scientific Reports*, *12*, 1–17. URL: https://doi.org/10.1038/s41598-022-06364-8.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*, 113–132. URL: https://doi.org/10.1515/iral-2016-9993.

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *Workshop on Disfluency in Spontaneous Speech* (pp. 17–20).

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*, 809–854. URL: https://doi.org/10.1111/lang.12084.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*, 569–591. URL: https://doi.org/10.1017/S0142716417000534.

Kallay, J. E., Mayr, U., & Redford, M. A. (2019). Characterizing the coordination of speech production and breathing. In *Proceedings of the International Congress of Phonetic Sciences 2019* (pp. 1412–1416). Melbourne: NIH Public Access.

Kaneko, H., & Horie, J. (2012). Breathing movements of the chest and abdominal wall in healthy subjects. *Respiratory Care*, *57*, 1442–1451. URL: https://doi.org/10.4187/respcare.01655.

Katsika, A., Krivokapić, J., Mooshammer, C., Tiede, M., & Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics*, *44*, 62–82. URL: https://doi.org/10.1016/j.wocn.2014.03.003.

Kentner, G. (2007). Length, ordering preference and intonational phrasing: evidence from pauses. In *Interspeech* (pp. 2637–2640). Antwerp volume 2. URL: https://doi.org/10.21437/Interspeech.2007-693.

Kentner, G., Franz, I., Knoop, C. A., & Menninghaus, W. (2023). The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase. *Journal of Phonetics*, *97*, 101225. URL: https://doi.org/10.1016/j.wocn.2023.101225.

Kienast, M., & Glitza, F. (2003). Respiratory sounds as an idiosyncratic feature in speaker recognition. In *International Congress of Phonetic Sciences (ICPhS)* (pp. 1607–1610). Barcelona.

Kim, J., Kumar, N., Lee, S., & Narayanan, S. (2014). Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *International Seminar on Speech Production (ISSP)* (pp. 222–225). Cologne.

Kirchhübel, C., Howard, D. M., & Stedmon, A. W. (2011). Acoustic correlates of speech when under stress: Research, methods and future directions. *International Journal of Speech, Language and the Law*, *18*, 75–98. URL: https://doi.org/10.1558/ijsll.v18i1.75.

Kirjavainen, M., Crible, L., & Beeching, K. (2022). Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of um. *Language and Speech*, *65*, 263–289. URL: https://doi.org/10.1177/00238309211011201.

Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., & Groarke, E. (2019). Dialect variation in formant dynamics: The acoustics of lateral and vowel sequences in Manchester and Liverpool English. *The Journal of the Acoustical Society of America*, *145*, 784–794. URL: https://doi.org/10.1121/1.5089886.

Kirsner, K., Dunn, J., & Hird, K. (2003). Fluency: Time for a paradigm shift. In R. Eklund (Ed.), *Disfluency in Spontaneous Speech Workshop* (pp. 13–16). Göteborg, Sweden.

Kosmala, L. (2020). On the distribution of clicks and inbreaths in class presentations and spontaneous conversations: blending vocal and kinetic activities. In *Laughter and Other Non-Verbal Vocalisations Workshop* (pp. 76–79). Bielefeld.

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*, 162–179. URL: https://doi.org/10.1016/j.wocn.2006.04.001.

Krivokapić, J., Styler, W., & Byrd, D. (2022). The role of speech planning in the articulation of pauses. *The Journal of the Acoustical Society of America*, *151*, 402–413. URL: https://doi.org/10.1121/10.0009279.

Krivokapić, J., Styler, W., & Parrell, B. (2020). Pause postures: The relationship between articulation and cognitive processes during pauses. *Journal of Phonetics*, *79*, 100953. URL: https://doi.org/10.1016/j.wocn.2019.100953.

Kuhlmann, L. L., & Iwarsson, J. (2021). Effects of speaking rate on breathing and voice behavior. *Journal of Voice*, *tbd*. URL: https://doi.org/10.1016/j.jvoice.2021.09.005.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. URL: https://doi.org/10.18637/JSS.V082.I13.

Lane, H., & Grosjean, F. (1973). Perception of reading rate by speakers and listeners. *Journal of Experimental Psychology*, *97*, 141–147. URL: https://doi.org/10.1037/h0033869.

Laserna, C. M., Seih, Y.-T., & Pennebaker, J. W. (2014). Um . . . Who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*, 328–338. URL: https://doi.org/10.1177/0261927X14526993.

Laver, J. (1978). The concept of articulatory settings: an historical survey. *Historiographia Linguistica*, *5*, 1–14.

de Leeuw, E. (2007). Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics*, *19*, 85–114. URL: https://doi.org/10.1017/S1470542707000049.

Lenth, R. V. (2021). emmeans: Estimated marginal means, aka least-squares means.

Lester, R. A., & Hoit, J. D. (2014). Nasal and oral inspiration during natural speech breathing. *Journal of Speech, Language, and Hearing Research*, *57*, 734–742. URL: https://doi.org/10.1044/1092-4388(2013/13-0096).

Lickley, R. J. (2015). Fluency and disfluency. In *The Handbook of Speech Production* (pp. 445–474). John Wiley & Sons, Ltd. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118584156.ch20.

Link, L. (2012). Individualtypische Aspekte des Atemgeräusches (Unpublished MA-Thesis).

Lo, J. J. H. (2020). Between äh(m) and euh(m) : The distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals. *Language and Speech*, *63*, 746–768. URL: https://doi.org/10.1177/0023830919890068.

Lowell, S. Y., Barkmeier-Kraemer, J. M., Hoit, J. D., & Story, B. H. (2008). Respiratory and laryngeal function during spontaneous speaking in teachers with voice disorders. *Journal of Speech, Language, and Hearing Research*, *51*, 333–349. URL: https://doi.org/10.1044/1092-4388(2008/025).

Loy, J. E., Rohde, H., & Corley, M. (2017). Effects of disfluency in online interpretation of deception. *Cognitive Science*, *41*, 1434–1456. URL: https://doi.org/10.1111/cogs.12378.

Lu, L., Liu, L., Hussain, M. J., & Liu, Y. (2020). I sense you by breath: Speaker recognition via breath biometrics. *IEEE Transactions on Dependable and Secure Computing*, *17*, 306–319. URL: https://doi.org/10.1109/TDSC.2017.2767587.

Lulich, S. M. (2010). Subglottal resonances and distinctive features. *Journal of Phonetics*, *38*, 20–32. URL: http://dx.doi.org/10.1016/j.wocn.2008.10.006.

MacIntyre, A. D. (2022). *The analysis of breathing and rhythm in speech*. Ph.D. thesis University College London.

MacIntyre, A. D., & Scott, S. K. (2022). Listeners are sensitive to the speech breathing time series: Evidence from a gap detection task. *Cognition*, *225*, 105171. URL: https://doi.org/10.1016/j.cognition.2022.105171.

MacIntyre, A. D., & Werner, R. (2023). An automatic method for speech breathing annotation. In *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 103–110). TUDpress, Dresden.

Mallol-Ragolta, A., Cuesta, H., Gomez, E., & Schuller, B. (2022). Multi-type outer product-based fusion of respiratory sounds for detecting COVID-19. In *Interspeech* (pp. 2163–2167). Incheon. URL: https://doi.org/10.21437/Interspeech.2022-10291.

Martin, J. G. (1970). On judging pauses in spontaneous speech. *Journal of Verbal Learning and Verbal Behavior*, *9*, 75–78. URL: https://doi.org/10.1016/S0022-5371(70)80010-X.

Matzinger, T., Ritt, N., & Fitch, W. T. (2021). The influence of different prosodic cues on word segmentation. *Frontiers in Psychology*, *12*. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.622042/full.

Matzinger, T., Ritt, N., & Tecumseh Fitch, W. (2020). Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLoS ONE*, *15*, 1–20. URL: https://doi.org/10.1371/journal.pone.0230710.

McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language, and Hearing Research*, *44*, 128–143. URL: https://doi.org/10.1044/1092-4388(2001/012).

McFarland, D. H., & Smith, A. (1992). Effects of vocal task and respiratory phase on prephonatory chest wall movements. *Journal of Speech and Hearing Research*, *35*, 971–982. URL: https://doi.org/10.1044/jshr.3505.971.

Mostaani, Z., Srikanth Nallanthighal, V., Harma, A., Strik, H., & Magimai-Doss, M. (2021). On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1345–1349). IEEE. URL: https://doi.org/10.1109/ICASSP39728.2021.9414756.

Moyse, E. (2014). Age estimation from faces and voices: A review. *Psychologica Belgica*, *54*, 255–265. URL: https://doi.org/10.5334/pb.aq.

Muhlack, B., Elmers, M., Drenhaus, H., Trouvain, J., van Os, M., Werner, R., Ryzhova, M., & Möbius, B. (2021). Revisiting recall effects of filler particles in German and English. In *Interspeech* (pp. 1678–1682). URL: https://doi.org/10.21437/Interspeech.2021-1056.

Muhlack, B., Trouvain, J., & Jessen, M. (2022). Acoustic characteristics of filler particles in German. In *Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)* (pp. 87–88). Prague.

Nakano, T., Ogata, J., Goto, M., & Hiraga, Y. (2008). Analysis and automatic detection of breath Sounds in unaccompanied singing voice. In *International Conference on Music Perception and Cognition (ICMPC)* (pp. 387–390). Sapporo.

Nallanthighal, V. S., Harma, A., & Strik, H. (2022). COVID-19 detection based on respiratory sensing from speech. In *Interspeech* (pp. 2498–2502). Incheon. URL: https://doi.org/10.21437/Interspeech.2022-11209.

Nallanthighal, V. S., Mostaani, Z., Härmä, A., Strik, H., & Magimai-Doss, M. (2021). Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings. *Neural Networks*, *141*, 211–224. URL: https://doi.org/10.1016/j.neunet.2021.03.029.

Ng, A. K., Koh, T. S., Baey, E., Lee, T. H., Abeyratne, U. R., & Puvanendran, K. (2008). Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea? *Sleep Medicine*, *9*, 894–898. URL: https://doi.org/10.1016/j.sleep.2007.07.010.

Ng, S.-I., Ma, R.-S., Lee, T., & Sum, R. K.-W. (2022). Acoustical analysis of speech under physical stress in relation to physical activities and physical literacy. In *Speech Prosody* (pp. 200–204). Lisbon. URL: https://doi.org/10.21437/SpeechProsody.2022-41.

Nycz, J., & Hall-Lew, L. (2013). Best practices in measuring vowel merger. In *Meetings on Acoustics* (p. 060008). San Francisco volume 20. URL: https://doi.org/10.1121/1.4894063.

Ogden, R. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association*, *43*, 299–320. URL: https://doi.org/10.1017/S0025100313000224.

Ogden, R. (2021). Swallowing in conversation. *Frontiers in Communication*, *6*, 1–19. URL: https://doi.org/10.3389/fcomm.2021.657190.

Oh, M., Byrd, D., & Narayanan, S. S. (2021). Leveraging real-time MRI for illuminating linguistic velum action. In *Interspeech* (pp. 3964–3968). Brno. URL: https://doi.org/10.21437/Interspeech.2021-1823.

Oh, M., & Lee, Y. (2018). ACT: An Automatic Centroid Tracking tool for analyzing vocal tract actions in real-time magnetic resonance imaging speech production data. *The Journal of the Acoustical Society of America*, *144*, EL290–EL296. URL: https://doi.org/10.1121/1.5057367.

Öhman, S. (1967). Peripheral motor commands in labial articulation. *Speech Transmission Laboratory Quarterly Progress Status Report (STL-QPSR)*, *8*, 30–63.

Oliveira, A., & Marques, A. (2014). Respiratory sounds in healthy people: A systematic review. *Respiratory Medicine*, *108*, 550–570. URL: https://doi.org/10.1016/j.rmed.2014.01.004.

Orlikoff, R. F., Baken, R. J., & Kraus, D. H. (1997). Acoustic and physiologic characteristics of inspiratory phonation. *The Journal of the Acoustical Society of America*, *102*, 1838–1845. URL: https://doi.org/10.1121/1.420090.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *The Journal of the Acoustical Society of America*, *77*, 640–648. URL: https://doi.org/10.1121/1.391882.

Panne, L. (2020). Individualtypische Respirationsmerkmale im forensisch-phonetischen Kontext - Eine Untersuchung an Atemgeräuschen in der Lesesprache (Unpublished MA-Thesis).

Park, H. (2002). Time course of the first formant bandwidth. *Annual Meeting of the Berkeley Linguistics Society*, *28*, 213. URL: https://doi.org/10.3765/bls.v28i1.3836.

Parlikar, A., & Black, A. W. (2012). Modeling pause-duration for style-specific speech synthesis. In *Interspeech* (pp. 446–449). URL: https://doi.org/10.21437/Interspeech.2012-154.

Perl, O., Ravia, A., Rubinson, M., Eisen, A., Soroka, T., Mor, N., Secundo, L., & Sobel, N. (2019). Human non-olfactory cognition phase-locked with inhalation. *Nature Human Behaviour*, *3*, 501–512. URL: https://doi.org/10.1038/s41562-019-0556-z.

Petrone, C., Fuchs, S., & Koenig, L. L. (2017a). Relations among subglottal pressure, breathing, and acoustic parameters of sentence-level prominence in German. *The Journal of the Acoustical Society of America*, *141*, 1715–1725. URL: https://doi.org/10.1121/1.4976073.

Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I., & Höhle, B. (2017b). Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. *Journal of Phonetics*, *61*, 71–92. URL: https://doi.org/10.1016/j.wocn.2017.01.002.

de Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America*, *96*, 2037–2047. URL: https://doi.org/10.1121/1.410145.

Pinto, D., & Vigil, D. (2020). Spanish clicks in discourse marker combinations. *Journal of Pragmatics*, *159*, 1–11. URL: https://doi.org/10.1016/j.pragma.2020.01.009.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymon, W., Hume, E., & Foster-Lussier, E. (2007). Buckeye corpus of conversational speech (2nd release). URL: www.buckeyecorpus.osu.edu.

Pouw, W., & Fuchs, S. (2022). Origins of vocal-entangled gesture. *Neuroscience & Biobehavioral Reviews*, *141*. URL: https://doi.org/10.1016/j.neubiorev.2022.104836.

Puggaard-Rode, R. (2022). Analyzing time-varying spectral characteristics of speech with function-on-scalar regression. *Journal of Phonetics*, *95*, 101191. URL: https://doi.org/10.1016/j.wocn.2022.101191.

Ramanarayanan, V., Bresch, E., Byrd, D., Goldstein, L., & Narayanan, S. S. (2009). Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation. *The Journal of the Acoustical Society of America*, *126*, EL160–EL165. URL: https://doi.org/10.1121/1.3213452.

Ramanarayanan, V., Byrd, D., Goldstein, L., & Narayanan, S. (2010). Investigating articulatory setting — pauses, ready position, and rest — using real-time MRI. In *Interspeech* (pp. 1994–1997).

Ramanarayanan, V., Byrd, D., Goldstein, L., & Narayanan, S. S. (2011). An MRI study of articulatory settings of L1 and L2 speakers of American English. In *International Seminar on Speech Production (ISSP)*. Montreal.

Ramanarayanan, V., Goldstein, L., Byrd, D., & Narayanan, S. S. (2013). An investigation of articulatory setting using real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *134*, 510–519. URL: https://doi.org/10.1121/1.4807639.

Raphael, L. J., Borden, G. J., & Harris, K. S. (2011). *Speech science primer: Physiology, acoustics, and perception of speech*. (6th ed.). Lippincott Williams & Wilkins.

Rasskazova, O., Mooshammer, C., & Fuchs, S. (2018). Articulatory settings during inter-speech pauses. In *Phonetik und Phonologie (P&P)* (pp. 161–164).

Rasskazova, O., Mooshammer, C., & Fuchs, S. (2019). Temporal Coordination of Articulatory and Respiratory Events Prior to Speech Initiation. In *Interspeech* (pp. 884–888). URL: https://doi.org/10.21437/Interspeech.2019-2876.

Rathcke, T., & Fuchs, S. (2022). Laugh is in the air: An exploratory analysis of laughter during speed dating. *Frontiers in Communication*, *7*. URL: https://doi.org/10.3389/fcomm.2022.909913.

RCore Team (2022). R: A language and environment for statistical computing. URL: https://www.r-project.org/.

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, *12*, 1–42. URL: https://doi.org/10.5334/LABPHON.237.

Redford, M. A. (2013). A comparative analysis of pausing in child and adult storytelling. *Applied Psycholinguistics*, *34*, 569–589. URL: https://doi.org/10.1017/S0142716411000877.

Rochet-Capellan, A., Bailly, G., & Fuchs, S. (2014). Is breathing sensitive to the communication partner? In *Speech Prosody* (pp. 613–617). URL: https://doi.org/10.21437/SpeechProsody.2014-111.

Rochet-Capellan, A., & Fuchs, S. (2013a). Changes in breathing while listening to read speech: The effect of reader and speech mode. *Frontiers in Psychology*, *4*, 906. URL: https://doi.org/10.3389/fpsyg.2013.00906.

Rochet-Capellan, A., & Fuchs, S. (2013b). The interplay of linguistic structure and breathing in German spontaneous speech. In *Interspeech* (pp. 2014–2018). Lyon. URL: https://doi.org/10.21437/Interspeech.2013-478.

Rochet-Capellan, A., & Fuchs, S. (2014). Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*. URL: https://doi.org/10.1098/rstb.2013.0399.

Rollins, M., & Oren, L. (2020). Effects of nasal emission and microphone placement on nasalance score during /s/. In *Proceedings of Meetings on Acoustics* (p. 060001). volume 42. URL: https://doi.org/10.1121/2.0001353.

RStudio Team (2022). RStudio: Integrated development environment for R. URL: http://www.rstudio.com/.

Ruinskiy, D., & Lavner, Y. (2007). An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals. *IEEE Transactions on Audio, Speech and Language Processing*, *15*, 838–850. URL: https://doi.org/10.1109/TASL.2006.889750.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of tTurn taking for conversation. *Language*, *50*, 696–735. URL: https://doi.org/10.1016/B978-0-12-623550-0.50008-2.

Salah, A. A., Salah, A. A., Kaya, H., Doyran, M., & Kavcar, E. (2021). The sound of silence: Breathing analysis for finding traces of trauma and depression in oral history archives. *Digital Scholarship in the Humanities*, *36*, ii2–ii8. URL: https://doi.org/10.1093/llc/fqaa056.

Santos, T. D. d., Pardo, J. S., & Bressmann, T. (2021). Interlocutor accommodation of gradually altered nasal signal levels in a model speaker. *Phonetica*, *78*, 95–112. URL: https://doi.org/10.1515/phon-2019-0105.

Scheinherr, A., Bailly, L., Boiron, O., Lagier, A., Legou, T., Pichelin, M., Caillibotte, G., & Giovanni, A. (2015). Realistic glottal motion and airflow rate during human breathing. *Medical Engineering and Physics*, *37*, 829–839. URL: https://doi.org/10.1016/j.medengphy.2015.05.014.

Schönle, P. W., & Conrad, B. (1985). Hesitation vowels: A motor speech respiration hypothesis. *Neuroscience Letters*, *55*, 293–296. URL: https://doi.org/10.1016/0304-3940(85)90451-3.

Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., & Pessentheiner, H. (2014). GRASS: the Graz corpus of read and spontaneous speech. In *International Conference on Language Resources and Evaluation (LREC)* (pp. 1465–1470). Reykjavik.: European Languages Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/394_Paper.pdf.

Schuppler, B., & Ludusan, B. (2020). An analysis of prosodic boundary detection in German and Austrian German read speech. In *Speech Prosody* (pp. 990–994). Tokyo. URL: https://doi.org/10.21437/SpeechProsody.2020-202.

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Interspeech* (pp. 525–529). Lyon. URL: https://doi.org/10.21437/Interspeech.2013-148.

Scobbie, J. M., Schaeffler, S., & Mennen, I. (2011). Audible aspects of speech preparation. In *International Congress of Phonetic Sciences (ICPhS)* (pp. 1782–1785). Hong Kong.

Serré, H. (2022). *Speech-breathing-limb movement interaction and coordination in a narrative task*. Ph.D. thesis Université Grenoble Alpes.

Serré, H., Dohen, M., Fuchs, S., Gerber, S., & Rochet-Capellan, A. (2021). Speech breathing: variable but individual over time and according to limb movements. *Annals of the New York Academy of Sciences*, *1505*, 142–155. URL: https://doi.org/10.1111/nyas.14672.

Shea, S., & Guz, A. (1992). Personnalité ventilatoire — An overview. *Respiration Physiology*, *87*, 275–291. URL: https://doi.org/10.1016/0034-5687(92)90012-L.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America*, *118*, 1677–1688. URL: https://doi.org/10.1121/1.2000788.

van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. In *International Conference on Language Resources and Evaluation (LREC)*. Marrakech. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/132_paper.pdf.

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*. URL: https://doi.org/10.1016/j.wocn.2020.101017.

Sperry, E. E., & Klich, R. J. (1992). Speech breathing in senescent and younger women during oral reading. *Journal of Speech and Hearing Research*, *35*, 1246–1255. URL: https://doi.org/10.1044/jshr.3506.1246.

Stevens, K. N. (2000). *Acoustic Phonetics*. (30th ed.). MIT Press.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*, 10587–10592. URL: https://doi.org/10.1073/pnas.0903616106.

Stone, S., & Birkholz, P. (2020). Articulation-to-speech using electro-optical stomatography and articulatory synthesis. In *International Seminar on Speech Production (ISSP)*. Providence.

Strangert, E. (1993). Speaking style and pausing. *Phonum*, *2*, 121–137.

Šturm, P., & Volín, J. (2023). Occurrence and Duration of Pauses in Relation to Speech Tempo and Structural Organization in Two Speech Genres. *Languages*, *8*, 23. URL: https://doi.org/10.3390/languages8010023.

Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, *142*, 2469–2482. URL: https://doi.org/10.1121/1.5008854.

Sundaram, S., & Narayanan, S. (2003). An empirical text tranformation method for spontaneous speech synthesizers. In *Eurospeech* (pp. 1221–1224). Geneva.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*, 485–496. URL: https://doi.org/10.1016/s0378-2166(98)00014-9.

Székely, , Henter, G. E., Beskow, J., & Gustafson, J. (2020). Breathing and speech planning in spontaneous speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7649–7653). Barcelona. URL: https://doi.org/10.1109/ICASSP40776.2020.9054107.

Székely, , Henter, G. E., & Gustafson, J. (2019). Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6925–6929). Brighton: IEEE. URL: https://doi.org/10.1109/ICASSP.2019.8683846.

Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal*, *65*, 71–79. URL: https://doi.org/10.1093/elt/ccq020.

Terzioglu, Y., Mutlu, B., & Sahin, E. (2020). Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration. *ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 343–357). URL: https://doi.org/10.1145/3319502.3374829.

Tiede, M., Masaki, S., & Vatikiotis-Bateson, E. (2000). Contrasts in speech articulation observed in sitting and supine conditions. In *Seminar on Speech Production* (pp. 25–28). Kloster Seeon.

Trouvain, J. (2002). Temposteuerung in der Sprachsynthese durch prosodische Phrasierung. In *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 294–301). TUDpress, Dresden.

Trouvain, J. (2003). *Tempo variation in speech production implications for speech synthesis*. Ph.D. thesis. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.639.7921&rep=rep1&type=pdf.

Trouvain, J. (2014). Laughing, breathing, clicking - the prosody of nonverbal vocalisations. In *Speech Prosody* (pp. 598–602). Dublin.

Trouvain, J., & Belz, M. (2019). Zur Annotation nicht-verbaler Vokalisierungen in Korpora gesprochener Sprache. In *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 280–287). Dresden: TUDpress, Dresden.

Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B., & Zimmerer, F. (2016a). The IFCASL corpus of French and German non-native and native read speech. In *International Conference on Language Resources and Evaluation (LREC)* (pp. 1333–1338). Portorož.

Trouvain, J., Fauth, C., & Möbius, B. (2016b). Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech. In *Speech Prosody* (pp. 31–35). Boston. URL: https://doi.org/10.21437/speechprosody.2016-7.

Trouvain, J., & Grice, M. (1999). The effect of tempo on prosodic structure. In *International Congress* (pp. 1067–1070). San Francisco.

Trouvain, J., & Malisz, Z. (2016). Inter-speech clicks in an Interspeech keynote. In *Interspeech* (pp. 1397–1401). San Francisco. URL: https://doi.org/10.21437/Interspeech.2016-1064.

Trouvain, J., & Möbius, B. (2013). Einatmungsgeräusche vor synthetisch erzeugten Sätzen — eine Pilotstudie. In *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 50–55). Bielefeld: TUDpress, Dresden.

Trouvain, J., & Möbius, B. (2018). Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache. In *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 334–341). Ulm. URL: https://www.essv.de/paper.php?id=426.

Trouvain, J., & Truong, K. P. (2015). Prosodic characteristics of read speech before and after treadmill running. In *Interspeech* (pp. 3700–3704). Dresden. URL: https://doi.org/10.21437/Interspeech.2015-734.

Trouvain, J., & Werner, R. (2020). Comparing annotations of non-verbal vocalisations in speech corpora. In *Laughter and Other Non-Verbal Vocalisations Workshop* (pp. 69–72). Bielefeld.

Trouvain, J., & Werner, R. (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. In C. Schwarze, & S. Grawunder (Eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion* (pp. 55–73). Tübingen: Narr.

Trouvain, J., & Werner, R. (2023). Muster der Sprechatmung in verschiedenen Sprechstilen – Eine Pilotstudie. In C. Draxler (Ed.), *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 95–102). Munich: TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=1178.

Trouvain, J., Werner, R., & Möbius, B. (2020). An acoustic analysis of inbreath noises in read and spontaneous speech. In *Speech Prosody* (pp. 789–793). Tokyo. URL: https://doi.org/10.21437/SpeechProsody.2020-161.

Truckenbrodt, H. (2007). The syntax–phonology interface. In P. D. Lacy (Ed.), *The Cambridge Handbook of Phonology* (pp. 435–456). Cambridge University Press. URL: https://doi.org/10.1017/CBO9780511486371.019.

Truong, K. P., Trouvain, J., & Jansen, M.-P. (2019). Towards an annotation scheme for complex laughter in speech corpora. In *Interspeech* (pp. 529–533). URL: https://doi.org/10.21437/Interspeech.2019-1557.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*, 445–472. URL: https://doi.org/10.1016/j.wocn.2006.12.001.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, *53*, 510–540. URL: https://doi.org/10.1177/0023830910372495.

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, *9*, 1–25. URL: https://doi.org/10.3389/fpsyg.2018.01994.

Wang, Y.-T., Green, J. R., Nip, I. S., Kent, R. D., & Kent, J. F. (2010). Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatrica et Logopaedica*, *62*, 297–302. URL: https://doi.org/10.1159/000316976.

Werner, R., Fuchs, S., Trouvain, J., Kürbis, S., Möbius, B., & Birkholz, P. (2022a). Comparison of acoustic parameters of inhalations vs exhalations with 3D-printed vocal tract models. In *International Conference on Speech Motor Control* (pp. 253–254). Groningen. URL: https://doi.org/10.21827/32.8310/2022-115.

Werner, R., Fuchs, S., Trouvain, J., Kürbis, S., Möbius, B., & Birkholz, P. (under review). Acoustics of breath noises in human speech: Descriptive and 3D-modelling approaches. *Journal of Speech, Language, and Hearing Research*, *tbd*.

Werner, R., Fuchs, S., Trouvain, J., & Möbius, B. (2021a). Inhalations in speech: Acoustic and physiological characteristics. In *Interspeech* (pp. 3186–3190). Brno. URL: https://doi.org/10.21437/Interspeech.2021-1262.

Werner, R., Trouvain, J., Fuchs, S., & Möbius, B. (2021b). Exploring the presence and absence of inhalation noises when speaking and when listening. In *International Seminar on Speech Production (ISSP)* (pp. 214–217). New Haven.

Werner, R., Trouvain, J., & Möbius, B. (2020). Ein sprachübergreifender Vergleich des Pausenverhaltens natürlicher Sprecher in verschiedenen Sprechtempi mit TTS-Systemen. In A. Wendemuth, R. Böck, & I. Siegert (Eds.), *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 101–108). Magdeburg: TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=444.

Werner, R., Trouvain, J., & Möbius, B. (2022b). Optionality and variability of speech pauses in read speech across languages and rates. In *Speech Prosody* (pp. 312–316). Lisbon. URL: https://doi.org/10.21437/SpeechProsody.2022-64.

Werner, R., Trouvain, J., & Möbius, B. (2022c). Speaker discrimination and classification in breath noises by human listeners. In *Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)* (pp. 68–69). Prague.

Werner, R., Trouvain, J., Muhlack, B., & Möbius, B. (2022d). Perceptual Categorization of Breath Noises in Speech Pauses. In O. Niebuhr, M. S. Lundmark, & H. Weston (Eds.), *Elektronische Sprachsignalverarbeitung (ESSV)* (pp. 139–146). TUDpress, Dresden. URL: https://www.essv.de/paper.php?id=1152.

Wester, M., Lecumberri, M. L. G., & Cooke, M. (2014). DIAPIX-FL: a symmetric corpus of problem-solving dialogues in first and second languages. In *Interspeech* (pp. 509–513). URL: https://doi.org/10.21437/Interspeech.2014-126.

Weston, H., Koenig, L. L., & Fuchs, S. (2021). Changes in Glottal Source Parameter Values with Light to Moderate Physical Load. In *Interspeech* (pp. 3350–3354). URL: https://doi.org/10.21437/Interspeech.2021-1881.

Whalen, D., & Kinsella-Shaw, J. (1997). Exploring the relationship of inspiration duration to utterance duration. *Phonetica*, *54*, 138–152. URL: https://doi.org/10.1159/000262218.

Whalen, D. H., Hoequist, C. E., & Sheffert, S. M. (1995). The effects of breath sounds on the perception of synthetic speech. *The Journal of the Acoustical Society of America*, *97*, 3147–3153. URL: https://doi.org/10.1121/1.411875.

Whalen, D. H., & Sheffert, S. (1996). Perceptual use of vowel and speaker information in breath sounds. In *International Conference on Spoken Language Processing (ICSLP)* (pp. 2494–2497). URL: https://doi.org/10.1109/icslp.1996.607319.

Whalen, D. H., & Sheffert, S. (1997). Normalization of vowels by breath sounds.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *6*, 199–234. URL: https://doi.org/10.1163/22105832-00602001.

Wikse Barrow, C., Włodarczak, M., Thörn, L., & Heldner, M. (2022). Static and dynamic spectral characteristics of Swedish voiceless fricatives. *The Journal of the Acoustical Society of America*, *152*, 2588–2600. URL: https://doi.org/10.1121/10.0014947.

Wilder, C. N. (1983). Chest wall preparation for phonation in female speakers.

Winkelmann, R., Jaensch, K., Cassidy, S., & Harrington, J. (2021). emuR: Main package of the EMU speech database management system.

Winkworth, A. L., Davis, P. J., Adams, R. D., & Ellis, E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research*, *38*, 124–144. URL: https://doi.org/10.1044/jshr.3801.124.

Winkworth, A. L., Davis, P. J., Ellis, E., & Adams, R. D. (1994). Variability and consistency in speech breathing during reading. *Journal of Speech, Language, and Hearing Research*, *37*, 535–556. URL: https://doi.org/10.1044/jshr.3703.535.

Winter, B., & Grawunder, S. (2012). The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics*, *40*, 808–815. URL: https://doi.org/10.1016/j.wocn.2012.08.006.

Włodarczak, M., & Heldner, M. (2016). Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking. (pp. 510–514). San Francisco. URL: https://doi.org/10.21437/Interspeech.2016-344.

Włodarczak, M., & Heldner, M. (2017). Respiratory constraints in verbal and non-verbal communication. *Frontiers in Psychology*, *8*. URL: https://doi.org/10.3389/fpsyg.2017.00708.

Włodarczak, M., & Heldner, M. (2018). Exhalatory turn-taking cues. In *Speech Prosody* (pp. 334–338). Poznán. URL: https://doi.org/10.21437/SpeechProsody.2018-68.

Wlodarczak, M., & Heldner, M. (2020). Breathing in conversation. *Frontiers in Psychology*, *11*, 1–17. URL: https://doi.org/10.3389/fpsyg.2020.575566.

Włodarczak, M., Heldner, M., & Edlund, J. (2015). Communicative needs and respiratory constraints. In *Interspeech* (pp. 3051–3055). Dresden. URL: https://doi.org/10.21437/Interspeech.2015-620.

Wright, M. (2007). Clicks as markers of new sequences in English conversation. In *International Congress of Phonetic Sciences (ICPhS)* (pp. 1069–1072).

Yu, S., & Demolin, D. (2022). Multi-parametric analysis of the respiratory activity in speech production. In *International Conference on Speech Motor Control* (p. 82). Groningen.

Zellers, M. (2022). An overview of discourse clicks in Central Swedish. In *Interspeech* (pp. 3423–3427). Incheon. URL: https://doi.org/10.21437/Interspeech.2022-583.

Zellers, M., & Schuppler, B. (2020). Microprosodic variability in plosives in German and Austrian German. In *Interspeech* (pp. 656–660). Shanghai. URL: https://doi.org/10.21437/Interspeech.2020-2353.

Zellner, B. (1994). Pauses and the Temporal Structure of Speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41–62).

Zhao, W., Gao, Y., & Singh, R. (2017). Speaker identification from the sound of the human breath. URL: https://doi.org/https://doi.org/10.48550/arXiv.1712.00171.

Zöllner, A., Mooshammer, C., Rasskazova, O., & Fuchs, S. (2021). Breathing affects reaction time in simple and delayed naming tasks. In *International Seminar on Speech Production (ISSP)* (pp. 218–221). Providence.

Zvonik, E., & Cummins, F. (2003). The effect of surrounding phrase lengths on pause duration. In *Eurospeech* (pp. 777–780).

Żygis, M., Tomlinson, J., Petrone, C., & Pfütze, D. (2019). Acoustic cues of prosodic boundaries in German at different speech rate. In *International Congress of Phonetic Sciences (ICPhS)* (pp. 999–1003).