Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

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

Raphael Werner, Jürgen Trouvain, Bernd Möbius

Language Science and Technology, Saarland University, Saarbrücken (Germany)
{rwerner|trouvain|moebius}@lst.uni-saarland.de

Abstract

Most prosodic boundaries are optional regarding their location. Moreover, markers of prosodic boundaries such as speech pauses show great variability in the duration of pauses as a whole and that of breath noises occurring in these pauses. Optionality and variability of pauses are particularly observable when the speaking rate is varied. In this study we investigated pausing behaviour in six languages of the BonnTempo-Corpus with 46 speakers who read aloud a semantically similar short passage at five rates, from very fast to very slow. The general picture across languages shows that pause duration, as expected, correlates with rate, and breath noise duration correlates with total pause duration. The optionality of pauses is reflected by the number of pauses (within but also across rates) and by the importance of some locations. The variability is evident in pause and inhalation durations: pause durations and number of inhalations decrease at faster rates, whereas both increase when a pause is less optional at a given location in the text. We consider a closer look at details of pauses to be an important step for prosody modelling and essential for exploring and explaining stylistic variation in prosodic phrasing.

Index Terms: speech pauses, prosodic breaks, speech rate

1. Introduction

Prosodic phrasing of the same text is not fixed and underlies variation. It can vary from one speaker to another [1, 2] but also within the same speaker (e.g. [2]). In addition, a change of the global tempo leads to a change 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. It is typical for human speech to show variations in prosodic phrasing but atypical for synthetic speech. For the latter, a fixed orientation by punctuation is often applied – an approach that does not satisfy the many stylistic and individual degrees of variation of natural prosodic phrasing [3].

Using speech pauses as a proxy for prosodic boundaries (PBs) in this study, we aim to analyse two dimensions of prosodic phrasing behaviour: *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. It must be noted that multiple PB markers exist [4, 5, 6] of which pauses are the most prominent cue [7], which may be why the two terms are frequently used synonymously. PBs 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 PBs are used, as in (1); missing or moving a PB to a different position would change the meaning:

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

b) To govern people | use language.

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

(2) a) The president's advisers fear | an early announcement | would complicate his fundraising | and other activities.

b) The president's advisers fear an early announcement | would complicate his fundraising and other activities.

c) The president's advisers fear an early announcement would complicate his fundraising and other activities.

Punctuation-based modelling of prosodic phrasing would only predict option (2c) and ignore other options. More elaborate models of prosodic phrasing (e.g. [8]) consider additional factors, such as rhythmical balance, but still follow the "one size/prediction fits all" paradigm.

In addition to the optionality of pause locations (and PBs in general), huge variability in pause duration can be observed [9, 10]. The same holds for the duration of breath (or inhalation) noises [11]. 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. Both pause duration and inhalation are related to the length of the upcoming inter-pausal unit [12], while the prosodic structure interacts with length in determining pause duration [13].

Increasing speaking rate has an impact on pauses and breathing: At faster rates, non-breath pauses tend to disappear, breath pauses become less frequent, and breath group size increases [14], although some speakers may use a different strategy of shortening breath groups [15]. To some degree this can be seen in Fig. 1 and 2 where compared to normal, breath group durations get longer in faster conditions, due to fewer breath pauses, but also partly in the slower conditions due to slower articulation rate. In rates from fast to slow speech, pause duration, number of pauses and the total ratio of time spent pausing compared to speaking were found to increase [10]. Moreover, other PB markers, such as final lengthening, are more robust to increasing speech rate than pauses, which are less important at faster rates [16].

Optionality and variability of pauses can be exemplified with the speaker in Fig. 1 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. In contrast, the speaker in Fig. 2 shows more pause diversity, in terms of optionality and variability, with non-breath pauses 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 modelling, 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 this study, we focus on pauses in read speech across speakers, across rate categories, and across languages for semantically comparable texts.

10.21437/SpeechProsody.2022-64

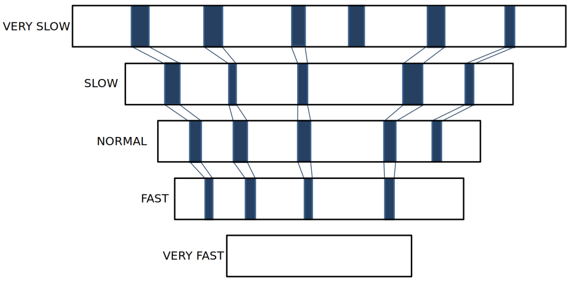

Figure 1: *Pauses of a "text book" speaker across five rates (read speech). White boxes: articulation time; blue boxes: breath pause time; shaded blue: non-breath pauses; pauses connected by dashed lines over rates: at the same location in the text.*

Table 1: *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

2. Methodology

We used data from the BonnTempo-Corpus (BTC) [17] for Czech, English, French, German, and Italian, as well as the Arabic Speech Rhythm Corpus [18] adding Egyptian Arabic with BTC's methods. The BTC versions use a short German text of four sentences, or close translations of it, and are thus very similar in many respects, whereas the Arabic one differs in terms of length, content, and punctuation (cf. Table 1). The texts were read aloud by native speakers at five intended speech rates: *very slow, slow, normal, fast, very fast*. We analysed 46 speakers producing 230 versions containing a total of 1710 pauses.

We did not use a fixed duration threshold for annotating pauses but included all perceived pauses (e.g. via final lengthening) that contained a silent period. We used the existing hand-labelled 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. Parameters analysed were pause duration and placement (with respect to the preceding word, which we here defined as a string of letters surrounded by whitespace or punctuation), 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 and after inhalations [19, 20]. 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 paper.

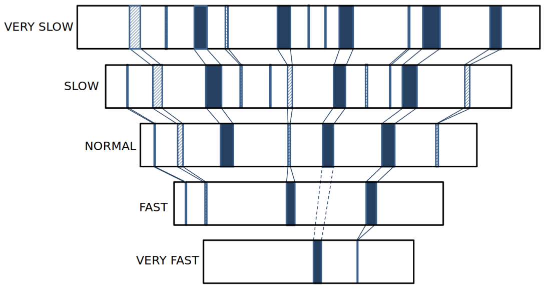

Figure 2: *Speaker from the BTC using different options for pause locations and more variability regarding breath pauses and non-breath pauses. Colour coding as in Fig. 1.*

3. Results

3.1. Pause location

Fig. 3 shows that some pauses are less optional than others, visualised by dots accumulating in vertical lines: while there is optionality for many pauses (e.g. the first eight 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 nine 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. 1), whereas French shows less uniform pausing behaviour 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 lead to 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 comparably many 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 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.

3.2. 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 (Table 2). There is much more flexibility regarding non-breath pauses across the rates than breath pauses. There are 15 times more

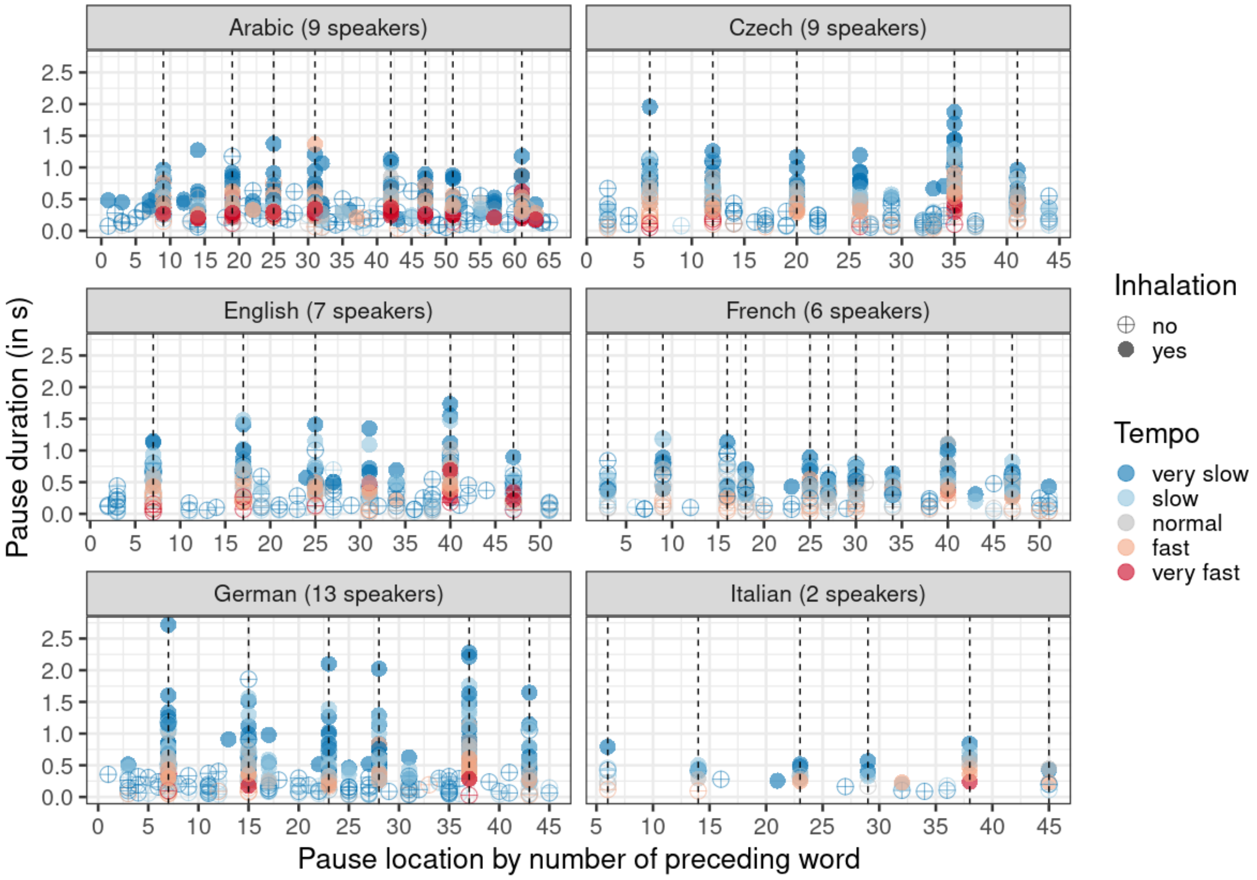

Figure 3: *Pauses and their durations by location in the text, sorted by language. Pause location is determined by the number of the preceding word. Whether a pause contains an inhalation is indicated by filled dots (inhalation) or empty, crossed dots (no inhalation). Dashed lines indicate locations with punctuation <, ; .>.*

Table 2: *Number of occurrence of breath and non-breath pauses by rate, pooled for all languages.*

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

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.

3.3. Duration of pauses, breath noises, and edges

Generally, as rate increases, pauses tend to be shorter. Fig. 4 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. 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 2. Conversely, not all faster rate inhalations are short, which might be related to participants reducing the number of pauses at these rates and then inhaling longer and/or more deeply when they do pause.

Fig. 5 shows how the durations of edges around inhalations vary by rate. When participants inhale in the faster rates, they shorten their edges. 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. Discussion

The findings concerning pause locations (Fig. 3) showed that pausing and punctuation are related in read speech as readers can use them as landmarks for pausing [21]. While there are many other locations for pauses, those are rather used at normal and slower rates. Breath pauses are more consistent in terms of location, as they prefer grammatically appropriate places in read speech, such as sentence, clause, or phrase boundaries [22]. 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

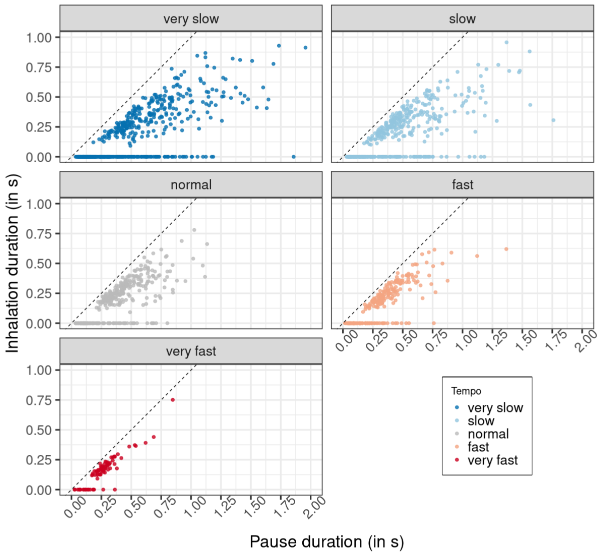

Figure 4: *Inhalation duration relative to pause duration split by rate for all languages. Dashed reference line: both durations would be equally long. Dots at y=0 s: non-breath pauses.*

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 with Fig. 3): *"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 modelling. Incorporating syntactical structures (cf. [23]) could be beneficial, especially for cross-linguistic comparisons.

When comparing inhalation duration in relation to pause duration (Fig. 4) to previous results in [11], the non-professional German speakers there seemed to behave quite similarly to the ones analysed 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. 5) to previous findings in [24], the left edge tendency is not there, but those results were preliminary and the speech analysed 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. [25], pp. 13f.), which makes it harder to sustain 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.

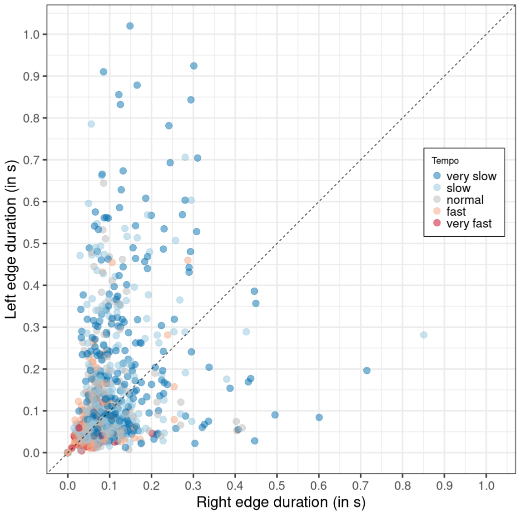

Figure 5: *Duration of left vs right edge in breath pauses in all six languages. The dashed reference line indicates equal durations. Five breath pauses (all from the same speaker) with left edges between 1.3 s and 2.1 s. were excluded from the plot.*

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.

5. Conclusions

This study has shed light on the optionality and variability of PBs 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 modelling misses this variability and while it may be able to reproduce pause location selection behaviour 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 modelling 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 modelling pauses and the placement of audible breaths within them.

These findings have implications for modelling 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 [26]. 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 [27, 28].

6. Acknowledgements

This research was funded in part by the Deutsche Forschungsgemeinschaft - Project-ID MO 597/10-1 and TR 468/3-1.

7. References

[1] F. Goldman-Eisler, “The Distribution of Pause Durations in Speech,” *Language and Speech*, vol. 4, no. 4, pp. 232–237, 1961.

[2] J. Trouvain and M. Grice, “The Effect of Tempo on Prosodic Structure,” in *International Conference of Phonetic Sciences*, 1999, pp. 1067–1070.

[3] A. Parlikar and A. W. Black, “Modeling Pause-Duration for Style-Specific Speech Synthesis,” in *Interspeech*, 2012, pp. 446–449.

[4] D. Duez, “Acoustic correlates of subjective pauses,” *Journal of Psycholinguistic Research*, vol. 22, no. 1, pp. 21–39, 1993.

[5] E. Strangert, “Speaking style and pausing,” *Phonum*, vol. 2, pp. 121–137, 1993.

[6] R. Carlson, J. Hirschberg, and M. Swerts, “Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates,” *Speech Communication*, vol. 46, no. 3-4, pp. 326–333, 2005.

[7] T. Matzinger, N. Ritt, and W. T. Fitch, “The Influence of Different Prosodic Cues on Word Segmentation,” *Frontiers in Psychology*, vol. 12, 2021.

[8] J. P. Gee and F. Grosjean, “Performance structures: A psycholinguistic and linguistic appraisal,” *Cognitive Psychology*, vol. 15, no. 4, pp. 411–458, 1983.

[9] E. Campione and J. Véronis, “A large-scale multilingual study of silent pause duration,” *Speech Prosody*, pp. 199–202, 2002.

[10] T. Matzinger, N. Ritt, and W. Tecumseh Fitch, “Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates,” *PLoS One*, vol. 15, no. 4, pp. 1–20, 2020.

[11] J. Trouvain, R. Werner, and B. Möbius, “An Acoustic Analysis of Inbreath Noises in Read and Spontaneous Speech,” in *Speech Prosody*, 2020, pp. 789–793.

[12] S. Fuchs, C. Petrone, J. Krivokapić, and P. Hoole, “Acoustic and respiratory evidence for utterance planning in German,” *Journal of Phonetics*, vol. 41, no. 1, pp. 29–47, 2013.

[13] J. Krivokapić, “Prosodic planning: Effects of phrasal length and complexity on pause duration,” *Journal of Phonetics*, vol. 35, no. 2, pp. 162–179, 2007.

[14] F. Grosjean and M. Collins, “Breathing, Pausing and Reading,” *Phonetica*, vol. 36, pp. 98–114, 1979.

[15] L. L. Kuhlmann and J. Iwarsson, “Effects of Speaking Rate on Breathing and Voice Behavior,” *Journal of Voice*, 2021.

[16] M. Żygis, J. Tomlinson, C. Petrone, and D. Pfütze, “Acoustic cues of prosodic boundaries in German at different speech rate,” 2019, pp. 999–1003.

[17] V. Dellwo, I. Steiner, B. Aschenberner, J. Dankovi, and P. Wagner, “BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate,” in *Interspeech*, 2004, pp. 777–780.

[18] O. Ibrahim, H. Asadi, E. Kassem, and V. Dellwo, “Arabic speech rhythm corpus: Read and spontaneous speaking styles,” 2020, pp. 5337–5342.

[19] D. Ruinskiy and Y. Lavner, “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*, vol. 15, no. 3, pp. 838–850, 2007.

[20] T. Fukuda, O. Ichikawa, and M. Nishimura, “Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition,” *Speech Communication*, vol. 98, pp. 95–103, 2018.

[21] E. Godde, G. Bailly, and M.-L. Bosse, “Pausing and breathing while reading aloud: development from 2nd to 7th grade in French speaking children,” *Reading and Writing*, 2021.

[22] A. L. Winkworth, P. J. Davis, E. Ellis, and R. D. Adams, “Variability and Consistency in Speech Breathing During Reading: Lung Volumes, Speech Intensity, and Linguistic Factors,” *Journal of Speech and Hearing Research*, vol. 37, no. 3, pp. 535–556, 1994.

[23] H. Truckenbrodt, “The syntax–phonology interface,” in *The Cambridge Handbook of Phonology*, P. Lacy, Ed. Cambridge University Press, 2007, p. 435–456.

[24] R. Werner, J. Trouvain, S. Fuchs, and B. Möbius, “Exploring the presence and absence of inhalation noises when speaking and when listening,” in *International Seminar on Speech Production*, 2021, pp. 214–217.

[25] T. J. Hixon, G. Weismer, and J. D. Hoit, “Preclinical Speech Science: Anatomy, Physiology, Acoustics, and Perception,” *Plural Publishing*, 2020.

[26] R. Werner, J. Trouvain, and B. Möbius, “Ein sprachübergreifender Vergleich des Pausenverhaltens natürlicher Sprecher in verschiedenen Sprechtempi mit TTS-Systemen,” in *Elektronische Sprachsignalverarbeitung*, 2020, pp. 101–108.

[27] M. Elmers, R. Werner, B. Muhlack, B. Möbius, and J. Trouvain, “Evaluating the effect of pauses on number recollection in synthesized speech,” in *Elektronische Sprachsignalverarbeitung*, Berlin, 2021, pp. 289–295.

[28] ——, “Take a Breath: Respiratory Sounds Improve Recollection in Synthetic Speech,” in *Interspeech*, 2021, pp. 3196–3200.