PERCEPTUAL CATEGORIZATION OF BREATH NOISES IN SPEECH PAUSES

Raphael Werner, Jürgen Trouvain, Beeke Muhlack, Bernd Möbius

Language Science and Technology, Saarland University, Saarbrücken, Germany
rwerner@lst.uni-saarland.de

Abstract: This paper reports on two experiments that investigated how well listeners can discriminate between different types of breath noises. We used stimuli extracted from natural conversations and asked participants to assign them to one of six breath types (exhalations: oral, nasal; inhalations: oral, nasal, oral followed by nasal, nasal followed by oral). We further examined if phonetic knowledge, 2 seconds of speech context around the breath noises, and the breath type of the stimulus had an influence on the categorization. The results from Experiment 1 suggested an overall correct assessment rate of 74 %, a tendency for context to help with categorizing, and similar performance by phoneticians and lay people. Nasal inhalations were correctly categorized at very high rates, whereas oral exhalations seemed the most difficult. Experiment 2 further tested these findings and found an overall rate of 66 %. Nasal inhalations again stood out with very high rates, whereas nasal exhalations were lowest here. Although we matched the stimuli for context conditions, we found no significant effect for this factor. While there was a slight tendency for speech context to be beneficial, we found interactions of this factor with some breath types, such that for nasal inhalations and nasal followed by oral inhalations it was more helpful not to have the context.

1 Introduction

Breathing is an essential and frequent aspect of speech, with inhalations occurring on average every 3 to 4 seconds [1, 2]. The breathing events can occur in different shapes, as air flow direction (inhalation or exhalation) and airway usage (oral or nasal or sequential combinations thereof) can be altered. In this study, 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 [3], 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 [4]. At rest, nasal inhalation is the default and frequent mouth breathing in children is associated with dental and craniofacial problems [5, 6]. Exhalations, which in speaking are typically used in combination with phonation to produce speech, may also occur without it [4]. Breathing behavior in read and spontaneous speech may differ [7] and in conversation, breathing is also related to turn management [8].

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 [4, 7, 9]. In particular, correctly distinguishing between different breath noise categories could improve acoustic analyses of breath noises [10], forensic use of paralinguistic material [11], or automatic breath detectors [12]. Furthermore, it is useful for the detailed annotation of breath noises [13] and making synthetic speech sound more natural [14].

In this study, we examined two main questions: First, how well can listeners discriminate between different types of breath noises in an auditory perception task? Second, 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?

2 Experiment 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 [15]. 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 κ = .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. [3, 16]).

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 often embedded [12]. 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 [17]. 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 [18]. 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.

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. 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.

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.

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.

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 1) show some differences in their duration and intensity.[2]

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 [19]. Four participants indicated to have beginner's knowledge of Dutch and were kept in the study. Again, the stimuli were

[1]The stimuli have been made available online and can be found as supplementary material to this paper at http://pauseparticles.org/publication.html

[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. 2.1).

Table 1 – Overview of the stimuli used in Experiment 2. We used 4 breath noises for each type.

Breath noise type	Mean duration (& range) in ms	Mean intensity (& range) in dB
in:oral	559 (408–1050)	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 (327–1074)	46.2 (36.3–61.4)
ex:nasal	450 (417–548)	39.0 (30.3–47.9)

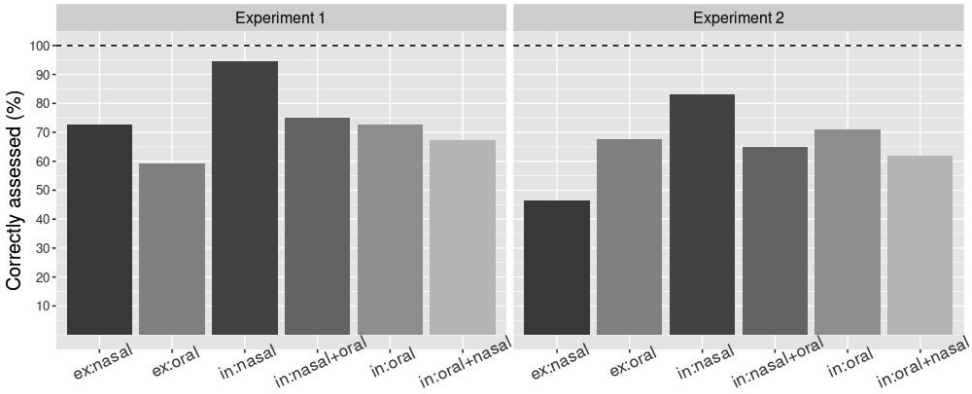

Figure 1 – Correct assessment of stimuli by the breath type. Results from Experiment 1 are plotted on the left, those from Experiment 2 is plotted on the right.

presented via Labvanced [17] and subjects were able to listen to a given stimulus a maximum of five times before they had to make a decision.

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. 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 [20] in R [21]. 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 (AIC) [22]. 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: $glmer(correct \sim breathtype * context +$

$(1 + breathtype \mid participant) + (1 + context \mid breathnoise))$. 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. = 0.1612, SE = 0.4457, z = 0.362, p = 0.7176), revealed main effects for the breath types *ex:oral* (Est. = 1.3649, SE = 0.6397, z = 2.134, $p < 0.05$), *in:nasal* (Est. = 1.3315, SE = 0.6379, z = 2.087, $p < 0.05$), and *in:oral* (Est. = 1.4582, SE = 0.6467, z = 2.255, $p < 0.05$), all of which positively influenced correct assessment. The other breath types, as well as the *no-context* condition (Est. = -0.6364, SE = 0.3716, z= -1.713, p = 0.0867) 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. = 1.3675, SE = 0.5607, z = 2.439, $p < 0.05$) and *in:nasal+oral* & *no-context* (Est. = 2.9100, SE = 0.5883, z = 4.947, $p < 0.001$) both had a positive effect on correct identification.

Fig. 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 %.

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 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 clicked as answers the most or least respectively, as visible in Fig. 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.

4 General discussion

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 %

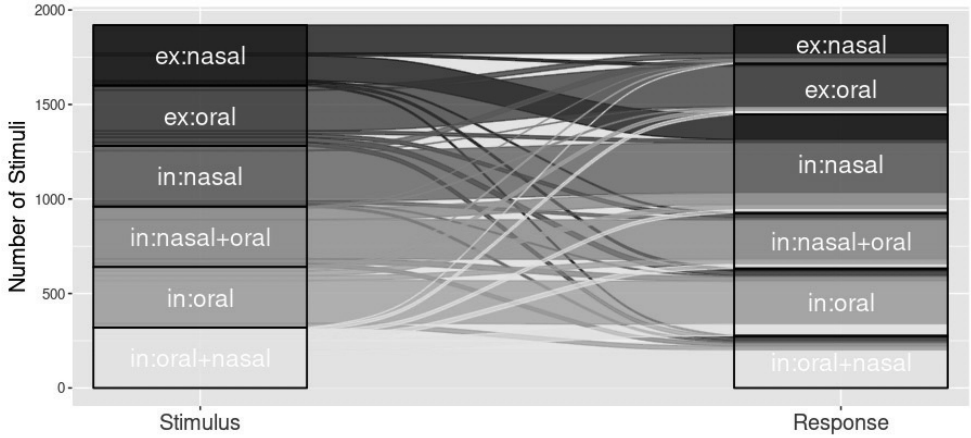

Figure 2 – Alluvial plot of the breath type of the stimulus (left) and type of the response (right). For correctly assessed stimuli, stimulus and response are the same, whereas otherwise the response will be something else.

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.

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. 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 findings [3], 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 [23]. 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*.

5 Conclusions

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 someone 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 [4]. We think that especially for forensic purposes, the findings of this study are relevant. In [4] 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. Additionally, when working with a tool like Praat [24], 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 Acknowledgements

This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID MO 597/10-1. We thank our student helpers Christin Weiß and Sven Kirchner for their assistance.

References

[1] ROCHET-CAPELLAN, A. and S. FUCHS: *The interplay of linguistic structure and breathing in German spontaneous speech*. In *Interspeech*, pp. 2014–2018. 2013.

[2] KUHLMANN, L. L. and J. IWARSSON: *Effects of Speaking Rate on Breathing and Voice Behavior*. *Journal of Voice*, 2021.

[3] LESTER, R. A. and J. D. HOIT: *Nasal and Oral Inspiration During Natural Speech Breathing*. *Journal of Speech, Language, and Hearing Research*, 57(3), pp. 734–742, 2014.

[4] KIENAST, M. and F. GLITZA: *Respiratory Sounds as an Idiosyncratic Feature in Speaker Recognition*. In *15th International Congress of Phonetic Sciences*, pp. 1607–1610. 2003.

[5] HARARI, D., M. REDLICH, S. MIRI, T. HAMUD, and M. GROSS: *The Effect of Mouth Breathing Versus Nasal Breathing on Dentofacial and Craniofacial Development in Orthodontic Patients*. *Laryngoscope*, 120(10), pp. 2089–2093, 2010.

[6] INADA, E., I. SAITOH, Y. KAIHARA, and Y. YAMASAKI: *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), pp. 1–10, 2021.

[7] TROUVAIN, J. and M. BELZ: *Zur Annotation nicht-verbaler Vokalisierungen in Korpora gesprochener Sprache*. *Elektronische Sprachsignalverarbeitung*, pp. 280–287, 2019.

[8] WLODARCZAK, M. and M. HELDNER: *Breathing in Conversation*. *Frontiers in Psychology*, 11, pp. 1–17, 2020.

[9] SCOBBIE, J. M., S. SCHAEFFLER, and I. MENNEN: *Audible aspects of speech preparation. 17th International Congress of Phonetic Sciences*, pp. 1782–1785, 2011.

[10] WERNER, R., S. FUCHS, J. TROUVAIN, and B. MÖBIUS: *Inhalations in Speech: Acoustic and Physiological Characteristics*. In *Interspeech*, pp. 3186–3190. 2021.

[11] BRAUN, A.: *Nonverbal Vocalisations – A Forensic Phonetic Perspective*. In *Laughter and Other Non-Verbal Vocalisations Workshop*, pp. 19–23. 2020.

[12] FUKUDA, T., O. ICHIKAWA, and M. NISHIMURA: *Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. Speech Communication*, 98, pp. 95–103, 2018.

[13] TROUVAIN, J. and R. WERNER: *A phonetic view on annotating speech pauses and pause-internal phonetic particles*. In C. SCHWARZE and S. GRAWUNDER (eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion*. Gunter Narr Verlag, 2022.

[14] SZEKELY, E., G. E. HENTER, J. BESKOW, and J. GUSTAFSON: *Breathing and Speech Planning in Spontaneous Speech Synthesis*. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 7649–7653. 2020.

[15] VAN SON, R. J., W. WESSELING, E. SANDERS, and H. VAN DEN HEUVEL: *The IFADV corpus: A free dialog video corpus. Language Resources and Evaluation Conference*, 2(1), pp. 501–508, 2008.

[16] FUCHS, S. and A. ROCHET-CAPELLAN: *The Respiratory Foundations of Spoken Language. Annual Review of Linguistics*, 7(1), pp. 13–30, 2021.

[17] FINGER, H., C. GOEKE, D. DIEKAMP, K. STANDVOSS, and P. KÖNIG: *LabVanced: A Unified JavaScript Framework for Online Studies*. In *International Conference on Computational Social Science*. 2017.

[18] TROUVAIN, J. and K. P. TRUONG: *Prosodic characteristics of read speech before and after treadmill running*. In *Interspeech*, pp. 3700–3704. 2015.

[19] *Prolific*. 2014. URL `https://www.prolific.co`. Accessed: 18/01/2022.

[20] BATES, D., M. MÄCHLER, B. M. BOLKER, and S. C. WALKER: *Fitting linear mixed-effects models using lme4. Journal of Statistical Software*, 67(1), 2015.

[21] R CORE TEAM: *R: A language and environment for statistical computing (version 4.1.2)*. 2021.

[22] AKAIKE, H.: *Information theory and an extension of the maximum likelihood principle*. In E. PARZEN, K. TANABE, and G. KITAGAWA (eds.), *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer New York, 1998.

[23] DOS SANTOS, T. D., J. S. PARDO, and T. BRESSMANN: *Interlocutor accommodation of gradually altered nasal signal levels in a model speaker. Phonetica*, 78(1), pp. 95–112, 2021.

[24] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer (version 6.1.09)*. 2009. URL `http://www.praat.org`.