Capturing respiratory sounds with throat microphones

Marcin Włodarczak and Mattias Heldner

1 Abstract

This paper presents results of a pilot study using throat microphones for recording respiratory sounds. We demonstrate that inhalation noises are louder before longer stretches of speech than in silent breathing and before shorter utterances (< 1 s). We thus replicate the results from our earlier study which used close-talking head-mounted microphones, while avoiding the associated data loss due to cross-talk. We also show that inhalations are louder within than before a speaking turn. Hence, the study provides another piece of evidence in favour of communicative functions of respiratory noises serving as potential turn-taking cues.

2 Introduction

While audible breathing is commonly associated with chronic pulmonary disorders, heavy exertion or being a former Jedi knight gone bad, inhalation and exhalation noises occurring in speech are louder than we are normally aware of. Any recording of spontaneous non-pathological speech provides ample evidence for this seemingly pedestrian observation.

However, in spite of the perceptual salience of respiratory noises, relatively little is known about their pragmatics. They have been shown to improve recall of synthetic speech (Whalen, Hoequist, & Sheffert, 1995), express dispreference (Kendrick & Torreira, 2015) and emotion (Yuan & Li, 2007) as well as to be used for marking thematic structure of read texts (Bailly & Gouvernayre, 2012). In addition, we recently presented some evidence based on detection thresholds for pauses accompanied and unaccompanied by inhalatory noise (Włodarczak & Heldner, 2016a) which suggests that the two types of pauses are perceptually different and could be employed to carry communicative meaning.

At the same time, respiration has been repeatedly claimed to be involved in turn management in spontaneous conversation (e.g. Local & Kelly, 1986; Schegloff, 1996). Indeed, there is some evidence supporting these claims using kinematic data (Ishii, Kumano, & Otsuka, 2015; McFarland, 2001; Rochet-Capellan & Fuchs, 2014; Włodarczak & Heldner, 2016c). However, given that human listeners are reported to be extremely good at detecting respiratory pauses (Wang et al., 2012) and that completely silent pauses unaccompanied by breathing are very rare (Grosjean & Collins, 1979; Trouvain, Fauth, & Möbius, 2016), turn-taking intentions are also likely to be reflected in respiratory acoustics.

In addition, respiratory sounds offers an attractive alternative to Respiratory Inductance Plethysmography (RIP), which is one of the most widely used methods of capturing respiration. RIP consists in two elastic belts worn around the chest and the abdomen, which measure changes in cross-sectional area of the thorax due to breathing. While this method undoubtedly delivers the gold standard in detection of respiratory events, it suffers from being somewhat invasive and from requiring time-consuming calibration procedures. As a result, it is also unsuitable for integration into a real-life spoken dialogue system. By contrast, breathing sounds have already proved useful in speech technology application ranging from voice activity detection (Fukuda, Ichikawa, & Nishimura, 2011) to speech synthesis (Braunschweiler & Chen, 2013; Sundaram & Narayanan, 2002), and automatic speech recognition (Butzberger, Murveit, Shriberg, & Price, 1992).

Motivated both by the relevance of respiratory sounds to understanding the underlying mechanisms of spontaneous conversation and by their potential applications, in an earlier study we measured sound pressure level (SPL) of inhalation in three different breathing cycle types (Włodarczak & Heldner, 2016b). The study found higher SPL in inhalations preceding longer stretches of speech than in shorter (< 1 s) vocalisations and in silent breathing (Figure 1). We also demonstrated that a logistic regression model using the SPL of the inhalation as the single predictor fitted the data approximately as well as a model incorporating several kinematic features collected using respiratory belts. Finally, as expected, SPL of inhalation was found to be mainly correlated with inhalation slope (change in lung volume per unit time) and negatively correlated with inhalation duration. In other words, short abrupt inhalations are louder, possibly due to increased ingressive airflow passing through a constriction in the vocal tract.

The results of the study were thus encouraging. The equivalence between kinematic and acoustic features of inhalations was particularly promising given than acoustic measures are much easier to extract and require little specialised and obtrusive equipment. However, recording respiratory noises with ordinary close-talking microphones turned out to be extremely sensitive to cross-talk effects. In our recording setup speakers stand quite close to one another, resulting in signal bleed between their microphones. Thus, measuring SPL of the relatively quiet respiratory noises is likely to pick up interlocutors' speech instead. For this reason, our analysis excluded all inhalations coinciding with speech of any interlocutor. Not surprisingly, this severely affected our sample sizes and left us with merely 339 data points.

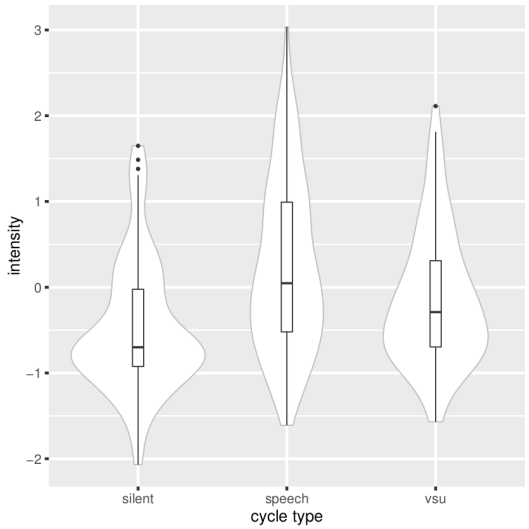

Figure 1. Distribution of inhalation sound pressure level (z-scored per speaker) in SILENT, SPEECH and VSU cycles (reproduced from Włodarczak & Heldner, 2016b).

In an attempt to deal with this issue, in the present paper we revisit the question of respiratory acoustics with *throat microphones*. Throat microphones are accelerometers (piezoelectric transducers) attached to the neck which pick up surface vibrations due to vocalisation. The have been mostly used in ambulatory voice monitoring and voice source studies (e.g. Askenfelt, Gauffin, Kitizing, & Sundberg, 1977; Mehta et al., 2015; Wokurek & Pützer, 2013). They are also robust to cross-talk and while capturing clear and highly audible respiratory noises.

3 Method

For this pilot study, three dialogues were recorded, one in English and two in Swedish. All speakers were native or highly proficient users of the respective languages. The speakers were recorded in a sound-treated room in the Phonetics Laboratory at Stockholm University. All participants knew each other prior to the recording. They were instructed to talk on a topic of their choice for about 20 minutes and were rewarded for their participation with a cinema ticket.

Respiratory movements were recorded using Respiratory Inductance Plethysmography (RIP): each participant wore a set of two elastic belts, one around the chest at the level of the armpits and the other around the abdomen, at the level of the navel. The belts were connected to RespTrack belt processors (designed and manufactured in the Phonetics Laboratory, Stockholm University), which measure inductance changes in the coils sewn into the belts as the subjects breathe in and out. The signal from the RespTrack processors was captured by a physiological data acquisition system (AD Instruments PowerLab) and recorded using LabChart. To minimise distortions of the signal due to movement, speakers were recorded standing around a round bar table.

Speech was recorded using directional close-talking microphones (Sennheiser HSP 4). In addition, respiratory sounds (as well as mostly unintelligible speech) were recorded using throat (contact) microphones attached to the speaker's necks (see Figure 2). The throat microphones were also made in the Phonetics Laboratory at Stockholm University, and were attached just below the Thyroid cartilage using cosmetic glue. To reduce noise in the signal due to cable movement, the cable was secured with a piece of surgical tape. The throat microphones were connected to a Shure ULX-D digital wireless system. All audio signals were routed to a Motu 8M audio interface and recorded using the REAPER software. Additionally, the speech signals were routed to PowerLab to allow for post synchronisation.

Figure 2. Throat microphone attached to a speaker's neck.

Stretches of speech and silence were detected automatically using the speech activity detection described in Laskowski (2011). In addition, interactional labels were calculated, following the computational model proposed in Jaffe and Feldstein (1970). In particular, joint pauses (stretches where both speakers remained silent) were classified as either between-speaker silences (BSS) or within-speaker silences (WSS), depending on whether a speaker change occurred during the pause.

Inhalations and exhalations were also identified automatically based on detection of local minimal and maxima in the RIP signal. Periods of laughter were detected automatically based on velocity and acceleration profiles of the respiratory signal and excluded from the analysis. Also excluded were respiratory cycles in which inhalations coincided with speech. While some of those were instances of ingressive speech, they mostly corresponded to erroneous speech detections and/or breathing cycle segmentations. The remaining respiratory cycles were classified into three classes: SILENT (if they coincided with no speech activity), SPEECH (if

they coincided with longer stretches of speech) or VSU (if they coincided with stretches of speech shorter than 1 second). The last category has previously been shown to largely correspond to short feedback expressions (Edlund, Heldner, Al Moubayed, Gravano, & Hirschberg, 2010; Heldner, Edlund, Hjalmarsson, & Laskowski, 2011).

Subsequently, the mean sound pressure level of each inhalation was extracted from the throat microphone signal and, for comparison, from the close-talking microphone. To allow comparison across different speakers, per-speaker *z*-scores were calculated.

The final analysed data sample consisted of 313 silent cycles, 205 speech cycles and 320 VSU cycles. Notably, because throat microphones are robust to cross-talk effects, no filtering similar to that used in Włodarczak and Heldner (2016b) was necessary. As a result, the three two-party dialogues yielded a larger sample than eight three-party conversations analysed previously.

4 Results

The left panel of Figure 3 shows violin plots of *z*-scored sound pressure level in the three types of respiratory cycles from throat microphones. In these plots kernel density estimates (mirrored around the ordinate) are overlayed over box and whiskers plots, allowing inspection of both distribution shapes and basic descriptive statistics. Significance of the differences was ascertained by means of an ANOVA ($F(2, 835) = 45.5$; $p < 0.001$). A post-hoc test (Tukey's HSD) revealed significant pairwise differences between all cycle types ($p < 0.001$).

The results obtained using throat microphones are thus largely compatible with our earlier findings. At the same time, they are free from the adverse effects of cross-talk requiring strict filtering and leading to large data loss.

Notably, when the same procedure is applied to the airborne, close-talking microphones without the data filtering step, the effect disappears completely (see right panel of Figure 3). In other words, when airborne microphones are used, the costly filtering procedure is necessary, lest the signal should indeed largely include sound bleed between microphones.

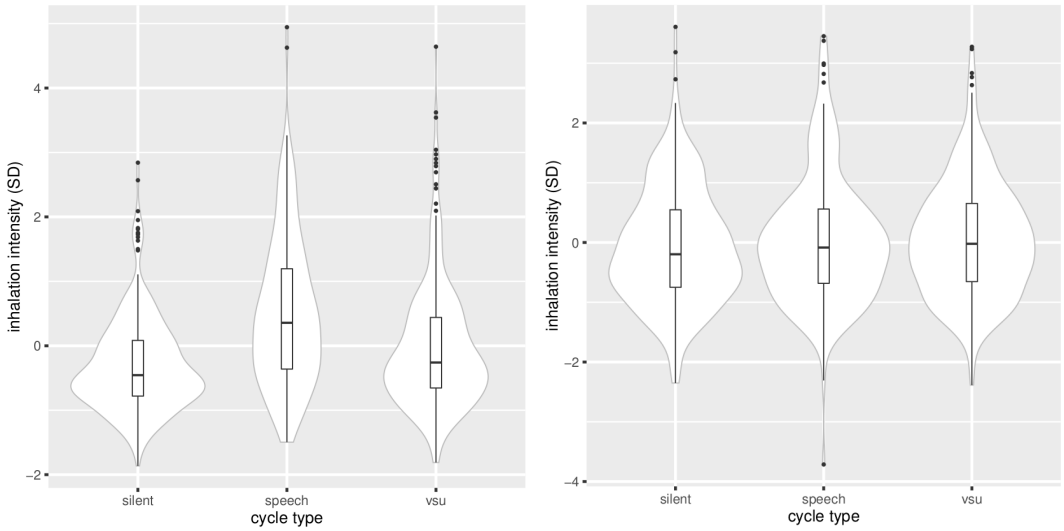

Figure 3. Distribution of inhalation sound pressure level (z-scored per speaker) in SILENT, SPEECH and VSU cycles from throat microphones (left) and airborne microphones (right panel).

As the main motivation of the present paper was to study of respiratory acoustics for turn-taking in spontaneous conversation, in Figure 4 we compare inhalation SPL in within- and between-speaker silences. Effectively, the plot compares the SPL of inhalation noises before (BSS) and inside (WSS) a turn. Additionally, we split the data depending on whether the following cycle coincided with speech or VSUs. As can be appreciated from the figure, the inhalations are louder within the turn than before one.

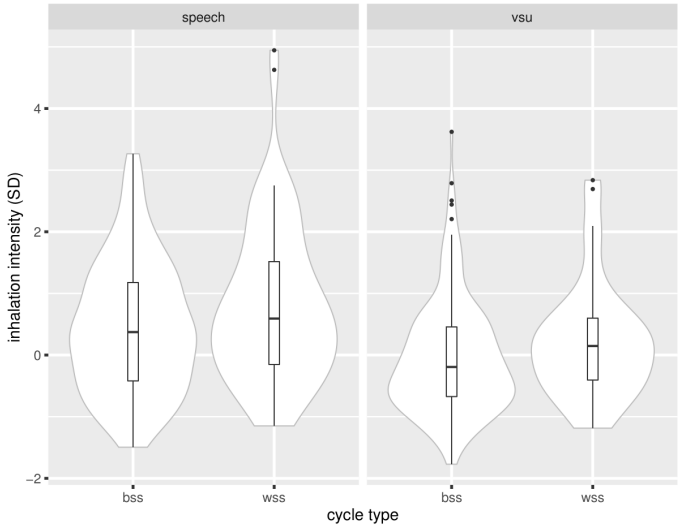

Figure 4. Distribution of inhalation sound pressure level (z-scored per speaker) before between- (BSS) and within-speaker silences (WSS) in SPEECH and VSU cycles.

The effects were again checked for significances with an ANOVA ($F(3, 381) = 12.49$; $p < 0.001$). The main effects of both silence type (BSS/WSS) and cycle type (VSU/speech) were significant ($p = 0.03$ and $p < 0.001$, respectively), but the interaction between them was not ($p = 0.78$).

5 Discussion and conclusions

The present paper is another in a series of studies in which we explore the communicative functions of breathing. Building on our earlier results based on kinematic data (Włodarczak & Heldner, 2016c), in Włodarczak and Heldner (2016b) we used respiratory sounds collected with regular close-talking microphones. As expected, the results indicated that inhalations before speech were indeed louder than those before VSUs, which in turn were louder than those before silent cycles. Thus, increased inhalation intensity could potentially cue upcoming speech. However, due to heavy crosstalk we were forced to exclude large portions of the data from the analysis. Namely, we filtered out all inhalations coinciding with other speakers' speech. The amount of data loss is not surprising given that inhalations are commonly started while the previous speaker is still speaking (Torreira, Bögels, & Levinson, 2015) and within-speaker pauses are likely to coincide with interlocutor's feedback (e.g. Goodwin, 1986).

This paper extends and improves on our previous results. Specifically, by using contact microphones mounted directly on the speakers' necks we were able to reproduce our earlier findings. However, because the throat microphones are largely free from crosstalk, no data exclusion was necessary. As a result, although this pilot study was based on a relatively small sample of only three dyadic interactions, the analysed sample was larger than that from eight three-party conversations. In addition, we demonstrated that airborne microphones are indeed strongly affected by crosstalk and that when no care is taken to ward it off, any existing differences in inhalation loudness are completely obscured.

Last but not least, inhalations inside a turn were found to be louder than before turn onsets. Increased intensity could thus potentially also function as turn-holding device, especially given our earlier results indicating that pauses accompanied by inhalatory noise have higher detection thresholds than silent pauses (Włodarczak & Heldner, 2016a). Notably, the difference between inhalations accompanying within- and between-speaker silences was of interest to us already in the previous paper, but the comparison was not possible since the sample size, small to start with, was far too small to allow for further partitioning of the data. While we had some indirect evidence in favour of this finding–in-turn inhalations are steeper and shorter (Włodarczak & Heldner, 2016c), which in turn is associated with increased loudness (Włodarczak & Heldner, 2016b), to the best of our knowledge this is the first time a direct evidence has been presented.

Clearly, the signal from the throat microphone is not what listeners (and speakers) have access to in the course of an ordinary conversation. However, throat microphones provide a simple work-

around to the technical problems of capturing clear respiratory noises. They also represent a more viable solution for speech technology applications. More fundamentally, however, given a systematic (whether linear or not) relationship between loudness levels measured on the neck and at the mouth, listeners may indeed be able to normalise the signal they hear and extract the relevant cues in spite of the supraglottal filtering of breath.

Overall, this pilot study has demonstrated the usefulness of throat microphones to studies of respiratory acoustics. We are planning a follow-up study using a higher number of dialogues. Additionally, although the bulk of existing studies have concentrated on features of the inhalation, exhalation features are also likely to be relevant for signalling turn-taking intentions. We intend to fill this gap in the future.

6 Acknowledgements

The research presented here was funded in part by the Swedish Research Council project 2014-1072 *Andning i samtal* (*Breathing in conversation*).

7 References

Askenfelt, A., Gauffin, J., Kitizing, P., & Sundberg, J. (1977). Electroglottograph and contact microphone for measuring vocal pitch. *QPSR, 18*(4), 13–21.

Bailly, G., & Gouvernayre, C. (2012). Pauses and respiratory markers of the structure of book reading. In *Proceedings of Interspeech 2012*. Portland, OR, USA: ISCA.

Braunschweiler, N., & Chen, L. (2013). Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *Proceedings of the 8th ISCA Speech Synthesis Workshop* (pp. 1–6). Barcelona, Spain.

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceeding HLT '91 Proceedings of the workshop on Speech and Natural Language* (pp. 339–343). Stroudsburg, PA, USA: Association for Computational Linguistics.

Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., & Hirschberg, J. (2010). Very short utterances in conversation. In *Proceedings from Fonetik 2010* (pp. 11–16). Lund, Sweden.

Fukuda, T., Ichikawa, O., & Nishimura, M. (2011). Breath-detection-based telephony speech phrasing. In *Proceedings of Interspeech 2011* (pp. 2625–2628). Florence, Italy: ISCA.

Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human studies, 9*, 205–217. doi: 10.1007/BF00148127

Grosjean, F., & Collins, M. (1979). Breathing, pausing and reading. *Phonetica, 36*(2), 98–114. doi: 10.1159/000259950

Heldner, M., Edlund, J., Hjalmarsson, A., & Laskowski, K. (2011). Very short utterances and timing in turn-taking. In *Proceedings of Interspeech 2011* (pp. 2837–2840). Florence, Italy: ISCA.

Ishii, R., Kumano, S., & Otsuka, K. (2015). Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings. In *ICMI '15- Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 99–106). New York, NY, USA: ACM.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue*. New York, NY, USA: Academic Press.

Kendrick, K. H., & Torreira, F. (2015). The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes, 52*(4), 255–289. doi: 10.1080/0163853x.2014.955997

Laskowski, K. (2011). *Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation (Doctoral dissertation)*. Carnegie Mellon University, Pittsburgh, PA, USA. Retrieved from http://www.cs.cmu.edu/~kornel/pubs/laskowskiTHESIS.pdf

Local, J., & Kelly, J. (1986). Projection and 'silences': Notes on phonetic and conversational structure. *Human studies, 9*, 185–204. doi: 10.1007/BF00148126

McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language and Hearing Research, 44*(1), 128–143. doi: 10.1044/1092-4388(2001/012)

Mehta, D. D., Van Stan, J. H., Zanartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., . . . Hillman, R. E. (2015). Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update. *Frontiers in Bioengineering and Biotechnology, 3*. doi: 10.3389/fbioe.2015.00155

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*(1658), 1–10. doi: 10.1098/rstb.2013.0399

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (pp. 52–133). Cambridge: Cambridge University Press.

Sundaram, S., & Narayanan, S. (2002). Spoken language synthesis: Experiments in synthesis of spontaneous monologues. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis* (pp. 203–206). Santa Monica, CA, USA.

Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology, 6*, 284. doi: 10.3389/fpsyg.2015.00284

Trouvain, J., Fauth, C., & Möbius, B. (2016). Breath and non-breath pauses in uent and dis uent phases of German and French L1 and L2 read speech. In *Proceedings of Speech Prosody 2016* (pp. 31–35). Boston, MA, USA.

Wang, Y. T., Nip, I. S., Green, J. R., Kent, R. D., Kent, J. F., & Ullman, C. (2012). Accuracy of perceptual and acoustic methods for the detection of inspiratory loci in spontaneous speech. *Behavior research methods, 44*(4), 1121–1128. doi: 10.3758/s13428-012-0194-0

Whalen, D. H., Hoequist, C. E., & Sheffert, S. M. (1995). The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America, 97*(5 Pt 1), 3147–3153. doi: 10.1121/1.411875

Włodarczak, M., & Heldner, M. (2016a). Is breathing silence? In *Proceedings Fonetik 2016*. Stockholm, Sweden: KTH Speech, Music and Hearing.

Włodarczak, M., & Heldner, M. (2016b). Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking. In Proceedings Interspeech 2016 (pp. 510–514). San Francisco, USA: ISCA. doi: 10.21437/Interspeech.2016-344

Włodarczak, M., & Heldner, M. (2016c). Respiratory turn-taking cues. In Proceedings Interspeech 2016 (pp. 1275–1279). San Francisco, USA: ISCA. doi: 10.21437/Interspeech.2016-346

Wokurek, W., & Pützer, M. (2013). Correlation analysis between acoustic source, electroglottogram and neck vibrations signals. In *Proceedings of the Eighth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2013)* (pp. 89–92). Florence, Italy.

Yuan, C., & Li, A. (2007). The breath segment in expressive speech. *Computational Linguistics and Chinese Language Processing, 12*(1), 17–31.