9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Exhalatory markers of turn completion

Marcin Włodarczak, Mattias Heldner

Department of Linguistics
Stockholm University
Stockholm, Sweden
{wlodarczak,heldner}@ling.su.se

Abstract

This paper is a study of kinematic features of the exhalation which signal that the speaker is done speaking and wants to yield the turn. We demonstrate that the single most prominent feature is the presence of inhalation directly following the exhalation. However, several features of the exhalation itself are also found to significantly distinguish between turn holds and yields, such as slower exhalation rate and higher lung level at exhalation onset. The results complement the existing body of evidence on respiratory turn-taking cues which has so far involved mainly inhalatory features. We also show that respiration allows discovering pause interruptions thus allowing access to *unrealised turn-taking intentions*.

Index Terms: breathing kinematics, multiparty conversation, turn-taking cues, hidden turn-taking events

1. Introduction

The present paper is an attempt to answer the question whether kinematic properties of exhalation following the speech offset can be used to infer if the speaker is going to yield the turn or continue speaking.

In previous work, we have shown that any participant's respiratory activity is indeed useful for prediction of future turn-taking behaviour of that participant (but not of the interlocutors) [1]. We have also investigated *inspiratory* turn-taking features, which distinguish between turn-holding and turn-yielding, both in terms of inhalation kinematics [2] and acoustics [3, 4]. Others have also concentrated almost exclusively on properties of the inhalation [5, 6].

The only existing piece of evidence suggesting that exhalation does indeed predict turn-taking behaviour is the *temporal compression* observed in turn-holding, whereby the interval between speech offset and the inhalation onset (as well as the inhalation itself) is reduced in duration [2, 5, 6].

At the same time, the exhalation itself has been suggested to function as a turn-yielding device. Most emphatically, Local and Kelly [7] make a distinction between *holding silences*, which fulfil a turn-holding function and *trail-off silences*, which signal turn-completeness and have a turn-yielding function. The two types are distinguished by presence or absence of exhalatory noise: while the former end "in glottal closure which is maintained through silence", the latter are "typically followed by audible out-breathing which does not terminate in glottal closure."

In order to test this hypothesis Ćwiek et al. [8] measured the acoustic similarity between pre- and post-pausal segments in breathing and non-breathing silences. They hypothesised that holding silences would involve an "articulatory hold" and would therefore results in greater similarity across the silent interval. Although this expectation was not borne out by the data, a strong tendency towards omitting the release of the prepausal plosive /p/ was observed. This form of anticipatory coarticulation is thus in line with the postulated articulatory holds.

In another study Ćwiek et al. [9] measured response times to resynthesised questions followed by either a breathing or a non-breathing pause. They found markedly longer response times and fewer interruptions during breathing pauses, suggesting again that respiratory noise functions as a turn-taking cue. These results are in line with the findings of Heldner and Włodarczak [10], who examined duration thresholds on perception of breathing and non-breathing pauses. They found that breathing pauses do indeed have significantly higher perception thresholds than non-breathing pauses and could, therefore, be used as turn-holding cues.

Notably, both Ćwiek et al.'s and Heldner and Włodarczak's results are at odds with Local and Kelly's account, which predicts that respiratory noise is a cue to turn termination. However, these studies also focused exclusively on inhalations.

By contrast, in this paper we explicitly focus on exhalatory kinematics. Specifically, we hypothesise that:

H1: Turn-holding is characterised by respiratory profiles with near-zero change in lung volume, corresponding to breath holds.

H2: Turn-yielding is characterised by quicker exhalation in order to achieve more audible respiratory noise.

H3: Turn-yielding exhalations are initiated at lower respiratory levels than turn holding exhalations.

H4: Turn-holding is characterised by shorter post-utterance exhalations.

H3 is motivated by the fact that speakers most likely want to complete whatever they planned to say before releasing the turn without starting a new respiratory cycle. Since there is mixed evidence for respiratory planning in spontaneous speech [11], it is likely that speakers might have to infringe on respiratory reserves to finish their thought. H4 has been already attested in previous literature but we include it here for the sake of completeness.

2. Method

The material consisted of eight three-party conversation in Swedish (average duration 22:56 min, SD = 1:22). All participants were native speakers of Swedish (median age = 25, IQR = 4) and, apart from two conversation, had been colleagues or fellow students. The gender of the speakers was balanced across the conversations with two males in one half of the interaction

10.21437/SpeechProsody.2018-68

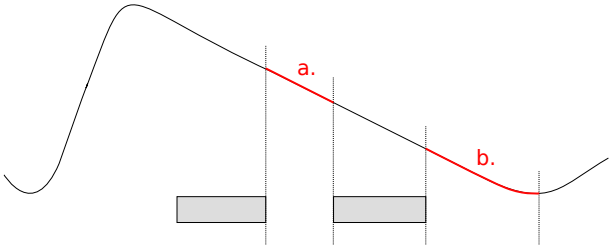

Figure 1: Exhalatory units terminated with speech (a) or with an inhalation (b). The shaded intervals represent stretches of speech by the same participant.

and two males in the other. The topic of the conversation was not restricted in any way.

The respiratory signal was collected using elastic respiratory belts worn across speakers' upper body (Respiratory Inductance Plethysmography, RIP) at the level of the armpits and the navel. In order to estimate the total lung volume change, the subjects were asked to perform the isovolume manoeuvre [12] prior to the recording. Throughout the recording session, the subjects were recording standing at a round bar table (105 cm in height). Since large torso movements cause severe distortions in the respiratory signal, the subjects were asked to stand as still as possible. The signal from the two respiratory belts was summed using RespTrack processors, custom-built hardware designed at Stockholm University, and collected with PowerLab (AD Instruments). Audio was recorded using close-talking condenser microphones with a cardioid polar pattern (Sennheiser HSP 4) and also routed to PowerLab for post-synchronisation. For a more detailed description of the setup, see [13].

Cycles in the respiratory signal were identified automatically by finding local minima and maxima in z-normalised signal separated by at least 1 standard deviation. Cycles including laughter, detected automatically using a modified version of the method outlined in [14], were excluded from the analysis. Stretches of speech were segmented using a simple intensity-based VAD [15] and corrected manually afterwards.

Given that we are concerned with an interaction between exhalation and turn-taking, a post-speech exhalation can be either bounded by an inhalation, initiating another respiratory cycle, or another stretch of speech by the same speaker. We adopt this, somewhat homogeneous, interval as the basic unit of our analysis and refer to it as the *exhalatory unit* (EU), see Fig. 2. In light of hypotheses H1-H4, we extracted the following parameters: (1) exhalatory slope (from least-squares regression), (2) lung volume level at the EU onset, and (3) EU duration. Measures related to lung volume (1 & 2) were normalised to speaker's range (estimated as the distance between the 5th and 95th percentiles of all peaks and valleys). In addition, (2) was expressed with respect to the *resting expiratory level* (REL), estimated at the median of all valleys. EU duration (3) was log-transformed to remove the skew typically found in durational data. Since we are interested in communicative function of breathing, only EUs longer than 120 ms, the minimal perceptual threshold for pause duration [16], were included in the analysis.

Since backchannels are generally assumed not to claim conversation floor [17], we have excluded from the analysis all EUs following segments of speech shorter than 1 second. This category of *very short units* (VSUs) has been previously demonstrated to correspond predominantly to backchannel-like categories [18]. It was subsequently observed that about 100 EUs have positive slope. These were most likely caused by an error in the automatic segmentation of respiratory cycles which skipped over an inhalation. Consequently, these instances were also excluded from the analysis and the final data set used in the analysis comprised 664 EUs. Finally, each EU was assigned to a category of *between-* or *within-speaker silences*, depending on whether the the silence was followed by a speaker change or more speech from the same participant.

Table 1: EU counts depending on the type of interval (between- or within-speaker silence) they coincide with and the type of event (inhalation or more speech) they are terminated with.

right bound	BSS	WSS	Σ
inhalation	288	73	361
speech	79	224	303
Σ	367	297	664

3. Results

Table 1 shows the distribution of EUs depending on the type of silence (between- or within-speaker silence) and the type of its right bound (silence or more speech). As expected, EUs coinciding with within-speaker silences are predominantly terminated by more speech from the previous speaker. In other words, in more than half of the cases, *when the speaker wants to keep the turn, she will continue without taking a breath*. By contrast, EUs corresponding to between-speaker silences tend to end with an inhalation. Indeed, the 79 instances of BSSs *ending with more speech by the same party* might seem counter-intuitive. These are instances in which a speech-bounded exhalatory pause coincides with speech by another participant. We propose that these are cases of *pause interruption* [19], in which the previous speaker has attempted to hold the turn but that intention has been nullified by another speaker starting during the pause. We will return to this point below.

In Figure 2 we present distributions of slope, respiratory onset and duration of EUs coinciding with between- and within-speaker silences, further subdivided into subgroups depending on whether they are terminated with an inhalation or speech. It is evident that with respect to EU slope values cluster together depending on the right boundary type (inhalation / speech) rather than silence type (BSS / WSS). Specifically, inhalation-bounded EUs, whether or not coinciding with BSSs or WSSs are characterised by higher (less negative) values of slope. In particular, presence of near-zero values might suggests that these intervals are in fact *breath holds*. By contrast, EUs terminated by more speech from the speaker are likely to involve quicker expiratory movements.

EU duration shows a different pattern with the shortest durations in WSSs (with somewhat lower durations in inhalation-terminated EUs). BSSs, on the contrary, are characterised by substantially longer EUs, especially if bounded by more speech.

Finally, onset level shows the smallest differences between the types. There is, however, a tendency for inhalation-bounded between-speaker EUs to be started lower than than speech bounded within-speaker EUs.

In order to test the contribution of these features to predicting the silence class (BSS v. WSS), a logistic regression model was fitted with EU duration, onset respiratory level, slope as well as all one-way interactions with right bound type, which

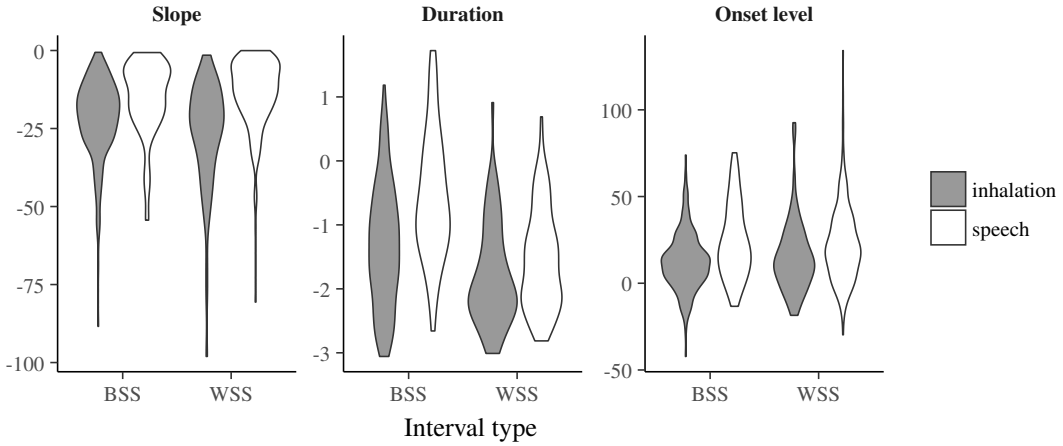

Figure 2: Distribution of slope, duration and respiratory onset level in EUs coinciding with between- and within-speaker silences.

was itself also included as a predictor. The interaction with respiratory onset level was not significant and was removed from the model, all other predictors were retained. Data points with leverage values greater than three times the mean were excluded and the model was re-fitted. The final model is summarised in Table 2.

Not surprisingly, presence of more speech by the same speaker is a very strong strong predictor for a within-speaker silence – it increases the odds for a WSS more than 12-fold. Given that the model included two-way interactions between right bound and slope and duration, the main effects for the two predictors give estimates for EUs bounded by an inhalation. Of these, only the effect of duration is significant – a increase of $\log_2$ s (i.e. doubling of duration) decreases the odds for a WSS by about half. In the absence of an interaction with right bound, the main effect of onset level indicates a weak tendency (exp(B) $= 1.018$) for higher EU onset levels to increase the likelihood of a WSS. In other words, the higher an EU is initiated in speaker's respiratory range, the less likely it is to involve a speaker change. The interaction terms indicate that in speech-bounded EUs both slope and duration significantly discriminate between WSSs and BSSs. Specifically, a one unit increase in slope (corresponding to 1% of speaker's lung capacity per second) increases the odds of a WSS by $1.002 \times 1.049 = 1.051$. By contrast, doubling of duration in speech-bounded EUs decreases the odds of a WSS by $0.496 \times 0.495 = 0.246$.

4. Discussion and conclusions

The results above indicate clearly that our initial hypotheses H1-H4 need to be further qualified by the type of event after the exhalation. In other words, the exhalatory unit behaves differently depending on whether it is followed by an inhalation or by more speech. Indeed, the results in Table 1 suggests a strong relationship between silence type and right EU boundary type but they are by no means universal. In general, EUs in WSSs were followed directly by more speech while EUs in BSS were, unsurprisingly, followed by an inhalation. At the same time exceptions to this rule were quite frequent. On the one hand, speakers obviously do sometimes inhale during while holding the turn – this was true in about 25% of all WSSs. More interestingly, however, about 20% of BSSs did not involve an inhalation. We hypothesised that this cases are in fact *pause interruptions*, where the previous speaker's intention to hold the turn was not realised due to turn-grabbing by a conversational partner. The distribution of kinematic features supports this tentative hypothesis.

Specifically, these instances are characterised by similar distributions of slope to inhalation-bound WSSs, with a marked tendency for near-zero exhalatory slopes. While we have not performed an independent check of the respiratory trances, a random inspection of a small subset of cases suggests that they do indeed correspond to breath holds. Thus, insofar as in all these cases the speaker's intention was to hold the floor, we have found supporting evidence for hypothesis H1.

By contrast, inhalation-bounded EUs show a tendency for more negative slope values, with little difference between the two categories. Thus, there is no evidence in favour of H2.

In addition, we have found supporting evidence in favour of H3 – exhalations in WSS are indeed initiated at higher lung volumes than in BSS.

The temporal compression in turn-holding has been previously attested in literature [5, 6, 2] and we have found it again here (thus confirming H4) – turn-holding is indeed associated with reduced duration of the interval between speech offset and inhalation onset. However, we have also found that speech-bounded EUs coinciding with BSSs are particularly long, longer in fact than BSSs involving a inhalation. This is also in line with the interpretation proposed above. Given that these instances involve pause interruption, the previous speaker is waiting (while holding her breath) to resume her turn when the interruption is over. In addition, EU durations in WSSs bounded by inhalation and WSSs bounded by speech, suggesting that pausing without an inhalation and releasing a turn and inhaling involve exhalations of comparable duration.

Perhaps the most unexpected finding of the present paper was identifying pause interruptions. While we do not have an independent annotation of speakers' turn-taking intentions, both the kinematic features and a manual inspection support this interpretation. Notably, pause interruptions are archetypal examples of *hidden events* [20], that is conversational events (in this case, turn-yielding intentions) which are obscured from view by a particular representation of the underlying process (here, a purely mechanistic turn-taking model determined fully

Table 2: Coefficients of the logistic regression model (95% BC_a bootstrap confidence intervals for odds ratio based on 3000 iterations). Model $\chi^2(6) = 294.21, p < .001$, $R^2 = .48$.

			95% CI		
	B	exp(B)	LL	UL	p
Intercept	-2.713	0.066	0.208	1.076	< 0.001
Right bound = Speech	2.526	12.500	14.857	160.718	< 0.001
Slope	0.002	1.002	0.976	1.021	0.9
Duration	-0.703	0.495	0.022	0.526	< 0.001
Onset level	0.018	1.018	1.004	1.032	0.01
Right bound = Speech × Slope	0.048	1.049	1.009	1.097	0.02
Right bound = Speech × Duration	-0.700	0.496	0.055	2.313	0.005

by temporal coordination of speech and silences across speakers, [21]). The present example shows the potential usefulness of the respiratory signal for uncovering such hidden events. We expect that respiratory activity will be also useful for other similar events, such as abandoned intentions to initiate a turn or failed intentions to release it. Indeed, it is possible that in some of the within-speaker silences in our material, the speaker did in fact attempt to release the floor but neither of the conversation partners was willing to take the turn or they provided a non-verbal response instead. At present, we are planning to obtain an independent annotation of turn-taking intentions, collected by means of parasocial consensus sampling [22] with a view to identifying other types of hidden events in spontaneous conversations.

In addition, similar to our work on inhalatory turn taking cues [2, 3], in future work, we are planning to compare the kinematic data to acoustic measurements of exhalatory sounds. The existing evidence suggests that turn-holds are characterised by articulatory holds, thus potentially leaving narrower constriction and producing louder air turbulence.

5. Acknowledgements

This work was funded by Swedish Research Council project 2014-1072 *Andning i samtal* (*Breathing in conversation*) and Christian Benoît Award to the first author and Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034 *Hidden events in turn-taking* to the second author.

6. References

[1] M. Włodarczak, K. Laskowski, M. Heldner, and K. Aare, "Improving prediction of speech activity using multi-participant respiratory state," in *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 1666–1670.

[2] M. Włodarczak and M. Heldner, "Respiratory turn-taking cues," in *Proceedings of Interspeech 2016*, San Francisco, CA, 2016.

[3] ——, "Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking," in *Proceedings of Interspeech 2016*, San Francisco, CA, 2016, pp. 510–514.

[4] ——, "Capturing respiratory sounds with throat microphones," in *Nordic Prosody: Proceedings of the 12th Conference, Trondheim 2016*, J. E. Abrahamsen, J. Koreman, and W. A. van Dommelen, Eds. Frankfurt am Main: Peter Lang, 2017, pp. 191–190.

[5] A. Rochet-Capellan and S. Fuchs, "Take a breath and take the turn: How breathing meets turns in spontaneous dialogue," *Philosophical Transactions of the Royal Society B*, vol. 369, no. 1658, pp. 1–10, 2014.

[6] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Using respiration to predict who will speak next and when in multiparty meetings," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 2, pp. 20:1–20:20, 2016.

[7] J. Local and J. Kelly, "Projection and 'silences': Notes on phonetic and conversational structure," *Human studies*, vol. 9, no. 2, pp. 185–204, 1986.

[8] A. Ćwiek, M. Włodarczak, M. Heldner, and P. Wagner, "Acoustics and discourse function of two types of breathing signals," in *Nordic Prosody: Proceedings of the 12th Conference, Trondheim 2016*, J. E. Abrahamsen, J. Koreman, and W. A. van Dommelen, Eds. Frankfurt am Main: Peter Lang, 2017, pp. 83–92.

[9] A. Ćwiek, S. Neueder, and P. Wagner, "Investigating the communicative function of breathing and non-breathing "silent" pauses," in *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, C. Draxler and F. Kleber, Eds. München, Deutschland: Ludwig-Maximilians-Universität München, 2016, pp. 27 – 29.

[10] M. Heldner and M. Włodarczak, "Is breathing silence?" in *Proceedings of Fonetik 2016*, Stockholm, Sweden, 2016.

[11] M. Włodarczak and M. Heldner, "Respiratory constraints in verbal and non-verbal communication," *Frontiers in Psychology*, vol. 8, pp. 1–11, 2017.

[12] K. Konno and J. Mead, "Measurement of the separate volume changes of rib cage and abdomen during breathing," *Journal of Applied Physiology*, vol. 22, no. 3, pp. 407–422, 1967.

[13] J. Edlund, M. Heldner, and M. Włodarczak, "Catching wind of multiparty conversation," in *Proceedings of Multimodal Corpora 2014*, Reykjavík, Iceland, 2014.

[14] J. Urbain, R. Niewiadomski, M. Mancini, H. Griffin, H. Çakmak, L. Ach, and G. Volpe, "Multimodal analysis of laughter for an interactive system," in *Intelligent Technologies for Interactive Entertainment*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, A. Nijholt, D. Reidsma, and H. Hondorp, Eds. Berlin Heidelberg: Springer, 2013, vol. 9, pp. 183–192.

[15] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006, pp. 1556–1559.

[16] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *Journal of Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, 2011.

[17] V. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, Chicago, 1970, pp. 567–577.

[18] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Proceedings of Interspeech 2011*, 2011, pp. 2837–2840.

[19] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, 2011.

[20] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findngs and implications for automatic processing of multi-party conversation," in *Proceedings of EUROSPEECH*, 2001, pp. 1359–1362.

[21] S. Feldstein and J. Welkowitz, "A chronography of conversation: In defence of an objective approach," in *Noverbal behavior and communication*, 2nd ed., A. W. Siegman and S. Feldstein, Eds. Hillsdale, NJ: Erlbaum, 1987, pp. 435–499.

[22] L. Huang, L.-P. Morency, and J. Gratch, "Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, vol. 1, Toronto, Canada, 2010, pp. 1265–1272.