INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Respiratory turn-taking cues

Marcin Włodarczak, Mattias Heldner

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

Abstract

This paper investigates to what extent breathing can be used as a cue to turn-taking behaviour. The paper improves on existing accounts by considering all possible transitions between speaker states (silent, speaking, backchanneling) and by not relying on global speaker models. Instead, all features (including breathing range and resting expiratory level) are estimated in an incremental fashion using the left-hand context. We identify several inhalatory features relevant to turn-management, and assess the fit of models with these features as predictors of turn-taking behaviour.

Index Terms: breathing, multiparty conversation, turn-taking cues, respiratory inductance plethysmography

1. Introduction

The proposition that breathing fulfils a communicative function in conversation can be most commonly found in the Conversation Analytical literature. Schegloff [1, pp. 105-106] formulates the idea succinctly: "It is tempting to dismiss breathing as merely a physiological prerequisite to talking, but this distracts from a variety of orderly practices which can inform the 'doing of breathing' in ways which achieve differing outcomes for the turn's construction and hearing. [...] The point here is that breathings - whether in or out - are practices; they can be done in various modalities [...]; they can be placed variously in the developing structure of the TCU [Turn Constructional Unit]." He then goes on to present some evidence of how respiration is used as turn-taking and turn-yielding cues (or, in his lingo, as pre-beginnings and post-completions).

Working in the same research paradigm, Local and Kelly [2] differentiated between *trail-off* and *holding silences*. The former are accompanied by an audible exhalation and mark turn yielding, the latter coincide with glottal closure and are a marker of turn-keeping.

Quantitative investigations of the interaction between breathing and turn-taking are rare. McFarland [3] analysed breathing in scripted spontaneous dialogues and found that listener's exhalations tended to increase in duration before speaker change. He also noted an increase in inhalation depth directly before turn-onset compared to the following inhalations in the turn, but that effect was only discernible in the scripted dialogue.

More recently, Rochet-Capellan et al. [4, 5] analysed the interaction between respiration and different turn configurations and found evidence of temporal compression during turn-keeping: inhalations themselves as well as the lag between speech offset and onset of the next inhalation, and between inhalation offset and speech onset were all shorter. This finding was interpreted as evidence of trying to minimise pause duration and consequently the risk of loosing the turn. They also found that unsuccessful interruptions (butting-ins) were shorter and less strongly tied to the onset of the exhalation. Finally, the temporal coordination of pre-speech inhalation onset to interlocutors breathing cycle depended on turn type: interruptions showed the most consistent (and earlier) peak towards the end of the other speaker's exhalation than smooth turn switches. Overall, half of the turns coincided with a single cycle and only 20% consisted of more than three cycles.

The topic has been addressed most comprehensively by Ishii et al. [6] using recordings of Japanese four-party conversations on a pre-assigned topic. The authors found significant differences between turn-taking and turn-keeping. Specifically, inhalations in turn-keeping were shorter, steeper and followed speech offset sooner than in turn-taking. Additionally, the next interpausal unit (IPU) followed the inhalation sooner when continuing after a turn-hold. No differences were found in amplitude, minimum and maximum lung volume levels. Finally, the amplitude and the peak volume of the inhalation were higher in next speakers than in the other listeners. While their paper was a noteworthy attempt at identifying respiratory turn-taking cues, the procedure employed involved performing a separate test for each feature considered thus possibly failing to detect the cumulative effects of minor changes in multiple predictors [7] and greatly reducing statistical power. This becomes even more troublesome given that the authors interpreted p-values between 0.05 and 0.1 as "significant trends". In addition, the variability in the data was greatly reduced by only considering speaker means. The paper also made certain assumptions about turn-taking which are not necessarily true. Most notably, prediction of turn initiation was only done in the vicinity of previous speaker's turn. This assumes, on the one hand, that the end-of-turn can be reliably predicted itself and, on the other hand, that contenders to the floor do not initiate speech at other locations within the turn. However, neither of these assumptions is uncontroversial. Last but not least, claims of predictive inference were complicated by the fact the some of the features were normalised by speaker's global parameters rendering the approach useless for online processing.

In a follow-up study [8], the authors used gaze- and respiration-related features to predict speaker change in multiparty conversations. Overall, classifiers using both information types performed better than gaze and respiration features separately. Nevertheless, gaze on its own was better at predicting whether there would be a speaker change or a turn hold and respiration was better at predicting who the next speaker would be in case of a speaker change.

Finally, Aare et al. [9] compared the inhalation amplitudes of the first two inhalations in single turns and found the turn initiating inhalation to be greater in amplitude.

The goal of the present paper is to provide a comprehensive account of respiration as a cue to turn-management in sponta-

http://dx.doi.org/10.21437/Interspeech.2016-346

neous multiparty conversation. Specifically, we want to identify the inhalatory features which most reliably predict whether an upcoming respiratory cycle is going to coincide with speech, with a backchannel or with no vocal activity from the speaker. Importantly, when estimating the features, we rely solely on dialogue history and do not require access to a global speaker model. To ensure maximal statistical power, we use a multivariable modelling approach.

2. Method

The study was based on the same material used in [10]. The description of the recording setup and data pre-processing is repeated below for completeness.

Eight recordings of three-party conversations in Swedish (with average length of 22:56 min, SD = 1:22 min) were used in the present study. In one half of the dialogues two of the speakers were males and in the other half two of the speakers were females. The topic and the course of interaction were not restricted in any way. All participants were native speakers of Swedish, with a median age of 25 (IQR = 4). With the exception of two conversations, all speakers knew each other prior to the recording.

Each participant's breathing was recorded using Respiratory Inductance Plethysmography, which measures changes in cross-sectional area of the rib cage and the abdomen by means of two elastic belts worn at the level of the armpits and the navel. Before the recording the individual contributions of each belt to the total lung volume change were assessed using the isovolume manoeuvre [11]. Participants were recorded standing at a bar table (105 cm in height), and were asked to avoid large torso movements, which would otherwise distort the respiratory trace.

The signal from the belts was sampled by RespTrack processors, designed and built at Stockholm University, and captured by PowerLab and LabChart (ADInstruments). The summed signal from the two belts corresponding to the total lung volume change was captured as well.

Cycles in the summed respiratory signal were identified automatically by replacing each sample value with a z-score calculated within a moving 10-second window, and locating signal maxima and minima which differed by at least 1 standard deviation in amplitude. The result was subsequently compared with manually corrected segmentations. Annotation errors (inhalations coinciding with speech), most likely due to large body movements were excluded from the analysis.

Laughter was detected automatically using a version of the algorithm described by [12] based on z-scored velocity and acceleration profiles. Manual inspection of the output of the laughter detector indicated that the method resulted in some false positives. However, as we were only using this technique for *data filtering*, this simply resulted in a smaller analysed sample.

Speech was collected using close-talking condenser microphones (Sennheiser HSP 4) and routed to PowerLab to allow synchronisation with the respiratory signal. Data collection took place in a sound-treated studio in Phonetics Laboratory, Stockholm University. The setup is described in greater detail in [13].

Voice activity detection was performed semi-automatically by manual correction of intensity-based segmentations done in ELAN [14]. Talkspurts shorter than 1 second were classified as *very short utterances* (VSUs). This class of utterances has previously shown to capture a large proportion of backchannel-like utterances [15].

Since this paper is concerned with *prediction* of dialogue participants' behaviour, we only used left-hand context for estimating features. Specifically, for each cycle we extracted the following features: (1) Inhalation duration, (2) Inhalation amplitude above the resting respiratory level (REL), (3) Inhalation slope, (4) Inhalation delay, and (5) Inhalation starting level with respect to REL.

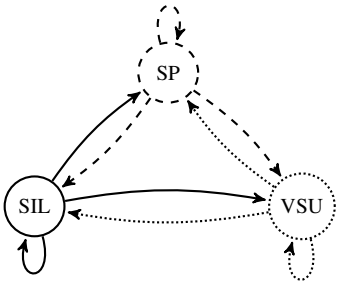

Figure 1: All possible transitions between respiratory cycle types: speech (SP), very short utterance (VSU) and silent (SIL). Separate models are fitted for each source state: SP (dashed lines), VSU (dotted lines) and SIL (solid lines).

Table 1: Cycle type counts

	Cycle type		
Previous cycle	Silent	Speech	VSU
Silent	1789	205	571
Speech	153	269	186
VSU	589	188	364

Inhalation duration was expressed in $\log_2$ ms. Inhalation delay was measured from offset of (own) preceding speech to onset of inhalation. Inhalation amplitude and inhalation starting level (henceforth inhalation minimum) above REL were expressed as percentages of speakers' respiratory range, whose limits were estimated at the 5^{th} and 95^{th} percentiles of all peaks and troughs in the respiratory cycles observed so far. REL itself was estimated as the median level of troughs in the previous 20 cycles. To the best of our knowledge, this is the first attempt at estimating speaker's respiratory range and REL in a fully automatic and incremental fashion.

Every respiratory cycle was then assigned to one of three classes depending on whether it coincided with no speech activity, a shorter (VSU) or a longer (non-VSU) speech segment. Subsequently, we fitted three separate multinomial logistic regression models predicting the class of each cycle depending on whether the previous cycle itself coincided with no speech activity, VSUs or a longer (non-VSU) speech segment. We refer to the three models as: the *silence model*, the *speech model* and the *VSU model*, depending on the class of the previous cycle. Effectively, the models cover all possible transitions between speaker states (see Figure 1). The overall model was split into three submodels to facilitate incorporating the immediately preceding context and to ease interpretation. The logistic regression models were fitted following the hierarchical procedure outlined in [16]. The distribution of transitions between respiratory cycle types is shown in Table 1.

If we assume that speakers generally do not vocalise during inhalations, and furthermore that inhalations introduce noticeable gaps into speech, these transitions can be used as estimations of different turn-taking events. The speech model

predicts whether a speaker who has been speaking during the previous respiratory cycle is going to continue speaking (i.e. turn-keeping), to be silent (i.e. turn-yielding), or to produce backchannel-like activity. Similarly, the silence model predicts whether a speaker who has been silent during the previous respiratory cycle is going to speak (i.e. turn-taking) or to produce a backchannel. The VSU model, finally, predicts whether a speaker who has produced a backchannel in the previous cycle is going to produce a longer stretch of speech (turn-taking) or be silent.

3. Results

First, we inspected the distributions of the inhalatory features for the different transitions (figures omitted due to space constraints). We found, among other things, that inhalations preceding turn-keeping transitions were characterised by shorter inhalation duration, a shorter inhalation delay, a slightly higher inhalation starting level, and higher inhalation slope compared both to those in turn-yielding and to those in turn-taking. Inhalation amplitude did not seem to differentiate much between turn-keeping and turn-yielding. We also found that inhalations preceding turn-taking were primarily characterised by a higher inhalation starting level, while inhalation durations and slopes were somewhere in-between those in turn-keeping and turn-yielding. In addition, we observed that the transition from VSU to speech cycles stood out in several respects. Amplitude was highest overall here, duration was only slightly longer than and slope was almost as high as that in turn-keeping.

Next, we looked into the results from the three multinomial logistic regression analyses (shown in Tables 2-4). These tables show the 'final models' including only the features that significantly improved the models in terms of reduction of -2 $\times$ log-likelihood. We arrived at these models by hierarchical entry of predictors [16]. We first entered inhalation duration as this was one of the two most robust features in [6], and this improved all three models significantly compared to a model where only the constant was included. We entered inhalation amplitude next, and this also improved all models significantly. In the third step, we entered inhalation delay (which is irrelevant by definition in the silence model) and this improved the speech and VSU models. In the last step, we entered inhalation minimum, and this improved the VSU and the silence models, but not the speech model. Slope was not included in the models due to its high correlation with duration and amplitude, which is likely to bias individual parameter estimates.

When going into details in these tables, we found that they generally supported our observations from the distribution of features. In the speech model, inhalation duration and delay made significant contributions to the prediction of the outcome speech (vs silent) according to the Wald statistic (the p-values shown in Tables 2-4). Similarly, inhalation delay made a significant contribution to the prediction of VSU (vs silent). The exp(B) values showed that an increase in duration with one unit (i.e. a doubling in duration) will decrease the odds for speech by 0.404. Similarly, an increase in delay with one unit will decrease the odds for VSU by 0.685. In other words, the shorter the duration and delay, the more likely will speech be as output. In the VSU model, all predictors made significant contributions to the prediction of speech, while only duration and delay contributed to the prediction of VSU. What we learn from the exp(B) numbers, however, is that a change in one unit of amplitude and inhalation starting level has a much smaller impact on the outcome than a change in duration and delay. Finally, in the silence model amplitude and inhalation starting level made significant contributions to the prediction of speech as well as to the prediction of VSU, while duration only contributed to prediction of VSU. Again, we learn that a change in one unit of duration has a larger impact on the outcome than a change in amplitude and starting level.

4. Discussion and conclusions

This paper has shown that several inhalatory features can be used as cues to turn-taking behaviour, even if the different features sometimes cue different distinctions. This in itself indicates context-sensitivity of respiration and the necessity of incorporating the previous cycle into predictive models.

The most robust features appear to be those related to temporal compression in turn-keeping, i.e. inhalation duration and inhalation delay. Thus, previous findings of temporal compression in turn-keeping [4, 5, 17] could be replicated. Contrary to previous findings [17], we have shown that inhalation amplitude and inhalation starting level also make significant contributions to predictions of turn-taking behaviour. We speculate that the lack of significant differences in [17] was due to lack of statistical power caused by problematic use of statistical testing as well as exclusion of short feedback expression from the analysis.

Indeed, inhalation amplitude was a particularly significant predictor of speech when following a VSU cycle. This is intuitively plausible: VSU function as feedback and grounding devices and as indicators of speaker's readiness to take the turn. One or more VSU coupled with a pronounced (and perceptually salient) inhalation might be used as an effective means of competing for the floor. By contrast, the decreased inhalation amplitude in VSU cycles following silent breathing is likely to be associated with modest respiratory requirements of short utterances. In [18], we suggested that short backchannel-like utterances need little respiratory planning and can be initiated at any point during the exhalation. The present result might be another aspect of the same phenomenon: if a speaker needs to produce a feedback expression, he or she can simply interrupt the inhalation (resulting in its lower amplitude) and start vocalising.

In a similar vein, the somewhat surprising finding that speech cycles tend be initiated at levels above REL becomes less puzzling when interpreted as a means of facilitating fast speaker transitions. From the point of view of a respiratory system, speakers have two possibilities to start speaking in a timely manner. They can either modify their breathing early in anticipation of an upcoming speaker change or interrupt their respiratory pattern in a more abrupt fashion by cutting the exhalation short and initiating a pre-speech inhalation. The latter possibility is likely to produce exactly the pattern we see in our data, namely inhalations started before reaching the resting lung configuration. We consider this strategy to be a much more plausible proposition in the light of little evidence of synchronisation of respiration between speakers found in literature [19, 3].

In conclusion, we have identified several novel inhalatory features predictive of turn-taking behaviour. All features were estimated in an incremental fashion not relying on global speaker models, in principle making them available for dialogue managers. Future work will include formal evaluation of the models for online prediction of turn-taking.

5. Acknowledgements

This work was funded by the Swedish Research Council project 2014-1072 *Andning i samtal* (*Breathing in conversation*).

Table 2: Coefficients of the speech model (95% BC_a bootstrap confidence intervals for odds ratio based on 3000 iterations). The reference category is *silent*.

				95% CI		
		B	exp(B)	LL	UL	p
Speech	Constant	-0.568	0.567	0.318	1.027	0.048
	Inhalation duration	-0.907	0.404	0.288	0.572	0.000
	Inhalation amplitude	0.005	1.005	0.993	1.018	0.405
	Inhalation delay	-0.378	0.685	0.576	0.814	0.000
VSU	Constant	-0.035	0.966	0.541	1.711	0.902
	Inhalation duration	-0.163	0.849	0.601	1.149	0.348
	Inhalation amplitude	-0.007	0.993	0.981	1.006	0.268
	Inhalation delay	-0.278	0.757	0.641	0.900	0.000

Note. $R^2 = .07$ (McFaddden), .16 (Nagelkerke), .14 (Cox & Snell).
Model $\chi^2(6) = 98.70, p < .001$

Table 3: Coefficients of the VSU model (95% BC_a bootstrap confidence intervals for odds ratio based on 3000 iterations). The reference category is *silent*.

				95% CI		
		B	exp(B)	LL	UL	p
Speech	Constant	-2.512	0.081	0.050	0.134	0.000
	Inhalation duration	-0.816	0.442	0.311	0.628	0.000
	Inhalation amplitude	0.026	1.026	1.015	1.037	0.000
	Inhalation delay	-0.273	0.761	0.653	0.894	0.000
	Inhalation min	0.044	1.045	1.027	1.064	0.000
VSU	Constant	-0.791	0.453	0.316	0.639	0.000
	Inhalation duration	-0.541	0.582	0.453	0.755	0.000
	Inhalation amplitude	0.005	1.005	0.997	1.014	0.198
	Inhalation delay	-0.165	0.848	0.745	0.958	0.006
	Inhalation min	0.009	1.009	0.996	1.023	0.152

Note. $R^2 = .06$ (McFadden), .12 (Nagelkerke), .11 (Cox & Snell).
Model $\chi^2(8) = 128.12, p < .001$

Table 4: Coefficients of the silence model (95% BC_a bootstrap confidence intervals for odds ratio based on 3000 iterations). The reference category is *silent*.

				95% CI		
		B	exp(B)	LL	UL	p
Speech	Constant	-2.702	0.067	0.045	0.101	0.000
	Inhalation duration	-0.183	0.833	0.593	1.153	0.188
	Inhalation amplitude	0.011	1.011	1.000	1.021	0.025
	Inhalation min	0.038	1.039	1.027	1.050	0.000
VSU	Constant	-0.952	0.386	0.298	0.508	0.000
	Inhalation duration	-0.564	0.569	0.455	0.708	0.000
	Inhalation amplitude	-0.007	0.993	0.986	1.000	0.046
	Inhalation min	0.011	1.012	1.002	1.021	0.012

Note. $R^2 = .03$ (McFadden), .05 (Nagelkerke), .04 (Cox & Snell).
Model $\chi^2(6) = 105.73, p < .001$

6. References

[1] E. A. Schegloff, "Turn organization: One intersection of grammar and interaction," *Studies in Interactional Sociolinguistics*, vol. 13, pp. 52–133, 1996.

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

[3] D. H. McFarland, "Respiratory markers of conversational interaction," *Journal of Speech, Language and Hearing Research*, vol. 44, no. 1, pp. 128–143, 2001.

[4] A. Rochet-Capellan, G. Bailly, and S. Fuchs, "Is breathing sensitive to the communication partner?" in *Proceedings of Speech Prosody 2014*, Dublin, Ireland, 2014.

[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, "Analysis of respiration for prediction of "who will be next speaker and when?" in multi-party meetings," in *Proceedings of the 16^{th} ACM International Conference on Multimodal Interaction (ICMI 2014)*, Istambul, Turkey, 2014, pp. 18–25.

[7] F. E. Harrell, Jr., *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*, ser. Springer Series in Statistics. New York: Springer, 2001.

[8] R. Ishii, S. Kumano, and K. Otsuka, "Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*, 2015, pp. 99–106.

[9] K. Aare, M. Włodarczak, and M. Heldner, "Inhalation amplitude and turn-taking in spontaneous estonian conversation," in *Proceedings from Fonetik 2015*, M. Svensson Lundmark, G. Ambrazaitis, and J. van de Weijer, Eds., Lund, Sweden, 2015, pp. 1–5.

[10] M. Włodarczak and M. Heldner, "Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking," in *Proceedings of Interspeech 2016*, San Francisco, CA, 2016.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] A. Field, J. Miles, and Z. Field, *Discovering statistics using R*. Los Angeles: Sage, 2012.

[17] R. Ishii, K. Otsuka, K. Shiro, and J. Yamato, "Predicting who will be the next speaker and when in multi-party meetings," NTT, Tech. Rep., 2015.

[18] M. Włodarczak, M. Heldner, and J. Edlund, "Communicative needs and respiratory constraints," in *Proceedings of Interspeech 2015*, Dresden, Germany, 2015.

[19] B. Garssen, "Synchronization of respiration," *Biological Psychology*, vol. 8, no. 4, pp. 311–315, 1979.