

Communicative needs and respiratory constraints

Marcin Włodarczak*, Mattias Heldner*, Jens Edlund†

* Department of Linguistics, Stockholm University

† Speech, Music and Hearing, KTH Royal Institute of Technology
Stockholm, Sweden

{włodarczak, heldner}@ling.su.se, edlund@speech.kth.se

Abstract

This study investigates timing of communicative behaviour with respect to speaker’s respiratory cycle. The data is drawn from a corpus of multiparty conversations in Swedish. We find that while longer utterances (> 1 s) are tied, predictably, primarily to exhalation onset, shorter vocalisations are spread more uniformly across the respiratory cycle. In addition, nods, which are free from any respiratory constraints, are most frequently found around exhalation offsets, where respiratory requirements for even a short utterance are not satisfied. We interpret the results to reflect the economy principle in speech production, whereby respiratory effort, associated primarily with starting a new respiratory cycle, is minimised within the scope of speaker’s communicative goals.

Index Terms: breathing, multiparty conversation, speech production, multimodal feedback

1. Introduction

Language forms are inherently constrained by physical limitations of our bodies [1]. Perhaps nowhere is this dependency more readily visible than in the interaction of speech and breathing. Simply put, we need air to speak; in fact we need just enough air to say whatever we want to say, which requires tight coordination of speech and respiration, and poses a non-trivial problem for models of speech production.

Respiratory constraints on speech production have been addressed before. Previous research has primarily stressed adaptation of respiratory behaviour to speech production needs. Most importantly, respiratory patterns were observed to adapt to syntactic structure of utterances [2, 3, 4] with inhalations coinciding predominantly with syntactic constituent boundaries. In the same vein, inhalation depth was found to be positively correlated with duration of the utterance it precedes, at least in read speech [5, 6, 7], suggesting that breathing is involved in speech planning. Finally, in recent years some groundwork has been done on interactional aspects of breathing, especially in connection with adaptations of the breathing cycle to managing turn-taking in conversation [8, 9, 10, 11, 12]. The overall view of the relationship between speech and respiration in this body of work is that of language production requirements being the overriding factor shaping the observed respiratory patterns.

In this paper, we propose to look at this problem from a different angle. Namely, instead of studying adaptations of breathing to speech production needs as a one-way execution pathway, we investigate to what extent speech production itself takes advantage of the momentary respiratory state. We suggest that coordination of speech and breathing is an optimisation problem in which communicative needs are offset against respiratory ef-



Figure 1: Recording setup.

fort, associated with, among other things, starting a new respiratory cycle. The mechanism thus follows the economy principle observed in other areas of speech production [13]. Here, we are interested in temporal organisation of longer stretches of speech within a respiratory cycle compared to shorter verbal feedback expressions as well non-verbal ones. We predict diverging timing patterns in these three classes due to differences in their breathing requirements.

Specifically, short feedback expressions (henceforth SFEs, e.g. ‘mhm’, ‘aha’, ‘ja’) are fundamentally different from full dialogue turns: they are short, relatively quiet and have been described as unobtrusive [14]. Due to their brevity and low acoustic intensity, SFEs thus possibly require less air in the lungs. Consequently, unlike longer stretches of speech, which are expected to be most commonly preceded by an inhalation and therefore to be strongly tied to the exhalation onset, SFEs are more likely to be produced on residual breath and to be distributed more uniformly within the respiratory cycle. In addition, we include one important type of non-verbal feedback, head nods, which fulfils similar communicative functions (see [15] and references therein) but is completely free from physiological respiratory constraints. We predict non-verbal feedback to be produced more frequently towards the very end of a respiratory cycle when respiratory requirements for even a short verbal feedback expression cannot be met. In addition to timing of segment onsets, we also examine their position within speaker’s vital capacity (the maximum volume of air exhaled after a maximum inspiration) and expect nods and short feedback expression to be more frequent towards its bottom. Notably, al-

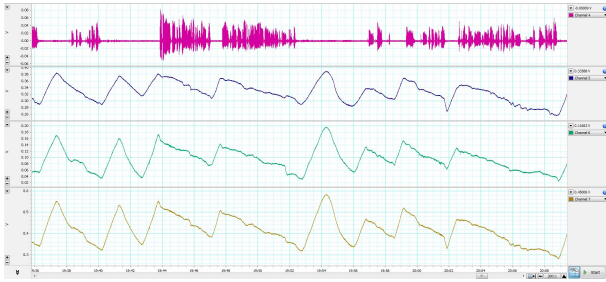


Figure 2: Speech recording (channel 1) and respiratory measurements from rib-cage and abdomen belts (channels 2-3) for one speaker. The bottom channel shows the weighted sum of the two belts.

though it is certainly possible to produce feedback expressions on ingressive air stream, they will not be analysed here.

These hypotheses are supported by results from a small pilot study based on two dyadic conversations in Estonian, which indicated that verbal backchannels (and other backchannel-like utterances) are indeed distributed more uniformly within the respiratory cycle than longer turns [16]. Here we extend on this study by using a larger and multimodal data set of multiparty conversations in Swedish. More recently, we also found that inspirations preceding backchannels resemble closely those during quiet breathing, suggesting that short feedback expressions do indeed have different respiratory demands than longer dialogue turns [17]. Finally, a study of breathing patterns in question-answer sequences found that long answers are more likely to be preceded by an inhalation than are short ones [18].

2. Method

Three recordings of three-party conversations in Swedish (27:18, 23:55 and 24 minutes long) were used in the present study. In two of the dialogues two of the speakers were males and in the third two speakers were females. The topic and the course of interaction were not restricted in any way. All participants were native speakers of Swedish.

Each participant’s breathing was recorded using Respiratory Inductance Plethysmography (RIP), 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 individual contributions of each belt to total lung volume change were assessed using the iso-volume manoeuvre [19]. Vital capacity and resting respiratory level (REL) were also estimated [20]. To account for drift in the data, the calibration procedure was repeated after the conversation. The resulting vital capacity was defined by extreme values of both measurements. Participants were recorded standing at a high table (95 cm), and were asked to avoid large torso movements, which would otherwise distort the respiratory trace. The setup is shown in Figure 1.

The signal from the belts was sampled by RespTrack processors, designed and built at Stockholm University, and captured by PowerLab (ADInstruments). The summed signal from the two belts corresponding to the total lung volume change was captured as well. A sample signal is presented in Figure 2.

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 differ by at least 1 standard devi-

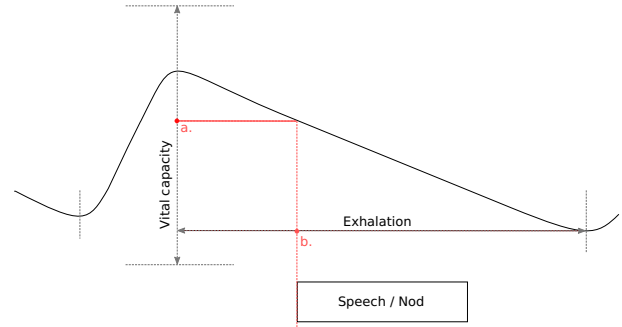


Figure 3: Position of a speech or nod segment onset with respect to vital capacity (a) and exhalation duration (b).

ation in amplitude. Annotation errors (inhalations coinciding with speech), most likely due to large body movements were excluded from the analysis.

Speech was collected with 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 [21].

Voice activity detection was performed semi-automatically by manual correction of intensity-based segmentations done in ELAN [22]. Talkspurts shorter than 1 second were classified as *very short utterances* (VSUs). This class of utterances was previously shown to capture a large proportion of short feedback expressions [23].

Video of each speaker’s head and torso was recorded using GoPro Hero3+ cameras placed on the table. Nods (head movement along the midsagittal plane) were marked manually in ELAN. The direction of movement (upwards or downwards) and the number of cycle repetitions were not labelled.

Subsequently, onsets of speech segments (both VSU and non-VSU) as well as of nods (whether or not accompanied by a VSU) were normalised with respect to: (1) duration of the exhalation, and (2) its position within speaker’s vital capacity (see Figure 3). The resulting values are thus expressed as respectively the proportion of exhalation and vital capacity at which a segment is initiated, with values ranging from 0 to 1.

Data of one participant were excluded from the analysis due to technical problems with the video recording, leaving eight speakers for the analysis. Altogether, the final data set included 570 (non-VSU) speech segments, 1034 VSUs, 141 nods and 167 nods coinciding with VSUs (nod+VSU).

3. Results

Distributions of speech and nod segment onsets within speaker’s vital capacity are plotted as kernel density estimates in the left panel of Figure 4. Hence, in that figure the abscissa corresponds to lung volume change obtained for maximum inhalation followed by maximum exhalation. As can be seen, most data clusters around 25% of that interval (mean = 0.24, SD = 0.11). Admittedly, these values are lower than commonly reported lung levels found in speech [20], suggesting drift in the data or calibration issues. For this reason, we suggest that the results in Figure 4 be treated in relative rather than absolute terms. When analysed in that manner, an interesting pattern in the data emerges. Namely, nods and nod+VSU composites are initiated at lower lung volumes (on average at 0.2 and 0.22, respectively),

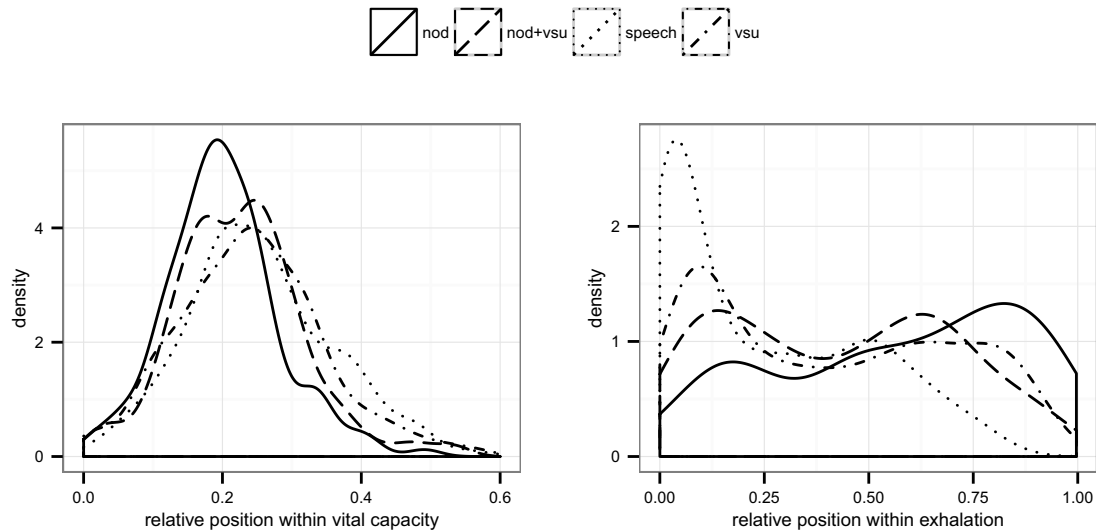


Figure 4: Kernel density estimates of nod, nod+VSU, speech and VSU segments onset timing relative to speaker’s vital capacity (left) and exhalation duration (right).

followed by VSUs (0.24) and speech segments (0.26). The results are thus in line with the reduced respiratory requirements of non-verbal feedback, whether or not accompanied by speech, than for speech-only segments. In addition, the nod+vsu distribution shows signs of bimodality with peaks corresponding broadly to those in the nod and the VSU distributions.

The distributions were compared by means of ANOVA and were found to be significantly different ($F(3, 1908) = 12.329, p < 0.001$). Pairwise comparisons between segment types using Tukey’s HSD test revealed statistically significant differences between speech and nod; VSU and nod ($p < 0.001$); speech and VSU+nod ($p < 0.01$), as well as VSU and speech ($p < 0.05$). Neither the nod+VSU and nod nor the nod+VSU and VSU classes were distinguishable statistically.

A compatible picture can be seen in the right panel of Figure 4, where position of segment onsets is normalised to expiration duration. Predictably, longer speech segments are started predominantly right at the beginning of the exhalation. After that their likelihood drops sharply, and they are extremely rare in the latter half of the expiration. While VSUs also show a tendency to be started towards the beginning of the exhalation, the peak is smaller and is followed by a plateau extending up to about 75% of the exhalation. Perhaps most interestingly, nods diverge from a uniform, flat distribution (one-sample Kolmogorov-Smirnov test, $p < 0.01$) and show a steady increase in frequency towards the end of the exhalation. Finally, the nod+VSU class shows a clear bimodal distribution with one peak aligned with exhalation onset and another occurring around 0.7 of the exhalation. Here, again, the nod+VSU distribution approximates the summed distributions of unimodal nods and VSUs. The four distributions were compared by means of two-sample Kolmogorov-Smirnov test with all comparisons except for *nod+VSU* vs. *VSU* being statistically significant at $p < 0.001$ (with Bonferroni correction).

Given that nods in the analysis above have been seen to occur more freely within the exhalatory cycle, we now investigate to what extent nods are likely to co-occur with inhalations

as well. To this end, we include 82 instances of nods and 78 instances of the nod+VSU class whose onset coincides with inhalations and we normalise its relative position between -1 and 0. The resulting distribution is presented in Figure 5. In that figure 0 on the abscissa corresponds to exhalation onset, while -1 and 1 correspond to inhalation onset and exhalation offset, respectively. Not surprisingly, the likelihood of nods overlapping with VSUs tends to increase towards the end of an exhalation. More interestingly, however, unimodal nods are most frequent around respiratory cycle boundaries, and least frequent right before the inhalation offset.

4. Discussion

In Section 1 we hypothesised that coordination of speech and breathing is not a one-way execution pathway in which breathing behaviour is modified freely to realise an arbitrary speech motor program. If this were the case, most utterances should be associated with a new respiratory cycle. Indeed, the results outlined in the previous section do not support this view. Specifically, while longer utterances, whose duration exceeds one second, do tend to be associated with a new respiratory cycle, this tendency is much weaker for shorter utterances (< 1 s), which correspond to a large extent to short feedback expressions and which are initiated fairly frequently up to 75% of the exhalation duration. Furthermore, nods, which are not restrained by respiratory requirements and could thus be expected to be distributed uniformly across the breathing cycle, are in fact most frequent around exhalation offset and least frequent around inhalation offset.

The findings are thus in line with the economy principle [13]. On this view, if the respiratory demands of the upcoming utterance are satisfied by the current respiratory state, the utterance is produced on residual breath without necessarily starting a new respiratory cycle (cf. [18]). Indeed, the increasing frequency of nods around breathing cycle boundary indicates that visual feedback is produced in place of verbal feedback when

lung levels are too low for sustaining even a short vocalisation. The distributions of segment onsets within the vital capacity provided consistent, if somewhat less clear, evidence.

Importantly, in addition to preserving respiratory effort, the mechanisms described above allow for more immediate and appropriate timing of utterances (cf. [18]). This is particularly true of short feedback expressions, which can be executed immediately, avoiding the delay introduced by initiating a new cycle. The same goal is achieved by producing a nod in place of a vocal feedback towards the very end of the expiration or during the inspiration. Notably, the lowest likelihood of producing a nod is shortly before the inhalation onset, suggesting that nods are dispreferred when a vocal segment is imminently possible.

Finally, given that distributions of the nod+VSU class approximates the summed distributions of nods and VSU produced separately, our data provides no evidence in favour of a special role of such multimodal composites. In other words, nods overlapping with VSUs occur in positions where either a nod or a short vocalisation are permissible on their own. This, in turn, suggests that the pragmatic function of a combination of a nod and a VSU is in fact little different from that of each of its components.

5. Conclusions and future work

In this paper we aimed to demonstrate that studying breathing in interaction is highly instructive for understanding mechanisms governing speech production. Distributions of speech and nod onsets within the the respiratory cycle suggest existence of temporal patterns consistent with an economy principle. In short, within the limits of their communicative goals (e.g. producing feedback) speakers seem to adapt their behaviour in such a way that respiratory effort (i.e. the need for a new respiratory cycle) is minimised. Consequently, communicative needs, respiratory constraints and momentary lung volume jointly shape the coordinative respiratory patterns.

In addition, the study has presented first results on timing of short feedback expression (operationalised as VSUs) in relation to the breathing cycle. Indeed, respiratory basis of one of the most striking and pervasive features of spontaneous conversation, that of short feedback expressions used as a basic grounding mechanism [24] has been so far almost completely overlooked. The oversight is perhaps to be partly explained by the special status of backchannels as conversational moves which do not involve taking the floor [24]. For this reason, in most existing work on respiration in interaction (for instance, [8, 9, 11]), backchannels have been classified together with “quiet breathing” cycles. At the same time, our own work suggests that respiratory cycles coinciding with short feedback expressions do in fact differ both from silent breathing and speech breathing and that, furthermore, conflating them with silent breathing might bias the obtained results [17]. We plan to follow on this potentially fruitful line of research in the future.

Due to relatively small size of the data, the results presented above need to be necessarily regarded as preliminary. The material was collected as part of a new project focusing on interactional functions breathing in Swedish multiparty dialogues, started in January 2015. Collection and annotation of more data are currently under way. In the future we plan to address other respiratory mechanisms underlying turn management in multiparty spontaneous dialogue with a view to both understanding fundamental mechanisms of speech production and improving existing speech technology solutions.

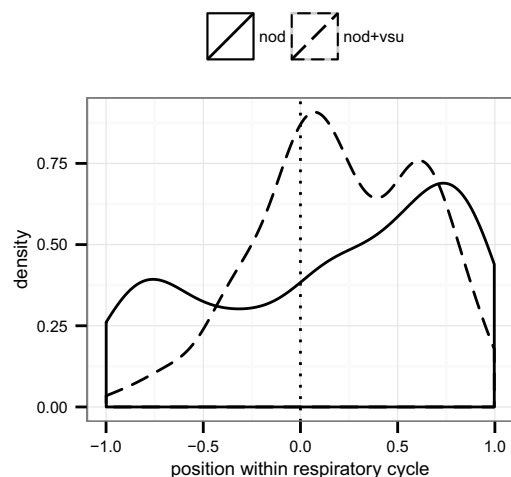


Figure 5: Distribution of nod and nod+VSU onset timing normalised to breathing cycle duration. 0 on the abscissa corresponds to exhalation onset, -1 and 1 correspond to inhalation onset and exhalation offset, respectively.

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

- [1] P. Lieberman, *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press, 1988.
- [2] —, *Intonation, perception, and language*. Cambridge, MA: MIT Press, 1967.
- [3] A. Henderson, F. Goldman-Eisler, and A. Skarbek, “Temporal patterns of cognitive activity and breath control in speech,” *Language and Speech*, vol. 8, no. 4, pp. 236–242, 1965.
- [4] F. Grosjean and M. Collins, “Breathing, pausing and reading,” *Phonetica*, vol. 36, no. 2, pp. 98–114, 1979.
- [5] S. Fuchs, C. Petrone, J. Krivokapić, and P. Hoole, “Acoustic and respiratory evidence for utterance planning in German,” *Journal of Phonetics*, vol. 41, no. 1, pp. 29–47, 2013.
- [6] D. H. Whalen and J. M. Kinsella-Shaw, “Exploring the relationship of inspiration duration to utterance duration,” *Phonetica*, vol. 54, no. 3–4, pp. 138–152, 1997.
- [7] A. L. Winkworth, P. J. Davis, E. Ellis, and R. D. Adams, “Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors,” *Journal of Speech, Language and Hearing Research*, vol. 37, no. 3, pp. 535–556, 1994.
- [8] D. H. McFarland, “Respiratory markers of conversational interaction,” *Journal of Speech, Language and Hearing Research*, vol. 44, no. 1, pp. 128–143, 2001.
- [9] M. M. Rahman, A. A. Ali, K. Plarre, M. al’Absi, E. Ertin, and S. Kumar, “mConverse: Inferring conversation episodes from respiratory measurements collected in the field,” in *Proceedings of the 2nd Conference on Wireless Health*, San Diego, CA, 2011, pp. 1–10.

- [10] G. Bailly, A. Rochet-Capellan, and C. Vilain, "Adaptation of respiratory patterns in collaborative reading," in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 1653–1657.
- [11] 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.
- [12] 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 16th ACM International Conference on Multimodal Interaction (ICMI 2014)*, Istanbul, Turkey, 2014, pp. 18–25.
- [13] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*, W. J. Hardcastle and A. Marchal, Eds. Springer, 1990, pp. 403–439.
- [14] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010, pp. 3054–3057.
- [15] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: an overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.
- [16] K. Aare, M. Włodarczak, and M. Heldner, "Backchannels and breathing," in *Proceedings of FONETIK 2014*, Stockholm, Sweden, 2014, pp. 47–52.
- [17] M. Włodarczak and M. Heldner, "Respiratory properties of backchannels in spontaneous multiparty conversation," in *18th International Congress of Phonetic Sciences 2015*, Submitted.
- [18] F. Torreira, S. Bögels, and S. C. Levinson, "Breathing for answering: the time course of response planning in conversation," *Frontiers in Psychology*, vol. 6, pp. 1–11, 2015.
- [19] 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.
- [20] T. J. Hixon, G. Wismer, and J. D. Hoit, *Preclinical Speech Science. Anatomy, Physiology, Acoustic, Perception*. San Diego: Plural Publishing, 2014.
- [21] J. Edlund, M. Heldner, and M. Włodarczak, "Catching wind of multiparty conversation," in *Proceedings of Multimodal Corpora 2014*, Reykjavík, Iceland, 2014.
- [22] 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.
- [23] 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.
- [24] 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.