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

Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking

Marcin Włodarczak, Mattias Heldner

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

Abstract

This paper presents first results on using acoustic intensity of inhalations as a cue to speech initiation in spontaneous multiparty conversations. We demonstrate that inhalation intensity significantly differentiates between cycles coinciding with no speech activity, shorter ($<$ 1 s) and longer stretches of speech. While the model fit is relatively weak, it is comparable to the fit of a model using kinematic features collected with Respiratory Inductance Plethysmography. We also show that incorporating both kinematic and acoustic features further improves the model. Given the ease of capturing breath acoustics, we consider the results to be a promising first step towards studying communicative functions of respiratory sounds. We discuss possible extensions to the data collection procedure with a view to improving predictive power of the model.

Index Terms: breathing, multiparty conversation, turn-taking cues, acoustics, loudness

1. Introduction

Respiratory acoustics has mostly been investigated in medical studies, where noisy breathing might be an indication of chronic respiratory conditions such as bronchitis or asthma [1]. Speech technology was rather late to the table but is catching up fast with breathing noises proving increasingly useful for a wide range of applications from speech synthesis to automatic speech recognition to voice activity detection (VOD).

Most methods of automatic breath identification [2, 3, 4, 5, 6] rely predominantly on its spectral features. While spectral characteristics of respiratory sounds differ depending on where they are measured (chest wall, trachea, mouth) [7, 8], Nakano et al. [4] found spectral envelope of breathing noises measured at the mouth to be very stable with a major intensity peak at 1.6-1.7 kHz and smaller peaks between 850 and 1 kHz. Due to the low spectral variability, the existing breath identification techniques thus achieve very high accuracy.

Experiments with synthetic speech have demonstrated that inclusion of breath sounds improves its naturalness without impairing intelligibility [9], especially in longer texts, such as audiobooks [5]. At the same time, insufficient matching between breathing segments and speech (for instance when they do not come from the same speaker) might obliterate or even reverse the effect [10]. Respiratory sounds are also relevant for automatic speech recognition. For instance, [11] demonstrated that training ASR language models on spontaneous data, which includes phenomena such as audible breaths and laughters, improves accuracy of spontaneous speech recognition. Lastly, modelling breath sounds improves performance of VOD [6], which is otherwise often tricked into treating breaths as speech.

Figure 1: Recording setup.

Crucially, even though it is known that human judges are very good at detecting respiratory pauses in speech [12], few studies have addressed the ways in which breathing sounds enter into the processes of speech perception and communication. What little is known suggests that breath sounds improve listeners recall of synthetic speech [13], that they convey meanings linked to expressing preference [14], emotional content of speech [15], physical effort [16], and that they are employed for marking text structure in read texts [17]. Results in [18] also indicate that coarticulatory information in breath sounds is used for speaker normalisation (but see [19]).

Overall, there is some evidence that breathing is both perceptually salient and that it is used to mark affective and pragmatic meanings in speech. So far, however, there has been little work on whether breathing sounds are used as cues to turn-management. Admittedly, several authors, including the present ones, have suggested that breathing is a turn-taking cue [20, 21, 22, 23, 24], however the evidence was based on breathing kinematics measured with elastic belts wrapped around speakers' thorax. While this method delivers a gold standard in the detection of respiratory events [12], it is relatively invasive and difficult to incorporate into a spoken dialogue system, not least because it requires a relatively time-consuming calibration procedure.

For this reason, in this paper we make a first attempt at evaluating breathing acoustics as a potential turn-taking cue. Breathing sounds offer obvious advantages: they are easy to record (and are, in fact, captured in the process of a regular speech recording) and require no additional acquisition devices

http://dx.doi.org/10.21437/Interspeech.2016-344

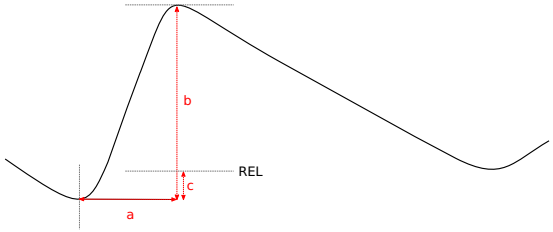

Figure 2: Kinematic features of the inhalation: duration (a), inhalation amplitude (b), inhalation minimum relative to REL (c), and slope (b / a).

or cumbersome calibration manoeuvres. Our working hypothesis is simple: we assume that inhalations preceding speech are much louder than in tidal breathing and before shorter, backchannel-like utterances. We, therefore, use acoustic intensity of inspiratory noises as a predictor for upcoming speech and compare it to kinematic measures collected with respiratory belts.

2. Method

The study was based on the same material used in [25]. 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 speakers were females. The topic and the course of interaction were not restricted in any way. All participants were native speakers of Swedish, with 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 (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 isovolume manoeuvre [26]. 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 recording 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.

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 deviation 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 [27].

Voice activity detection was performed semi-automatically by manual correction of intensity-based segmentations done in ELAN [28]. 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 backchannels and short feedback expressions [29].

Laughter was detected automatically using a version of the algorithm described by [30] 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.

Every cycle was subsequently assigned to one of three classes depending whether it coincided with no speech activity, VSUs or a longer (non-VSU) speech segment. Mean intensity of the inhalation was extracted using Praat and z-score normalised per speaker. To avoid adverse effects of crosstalk, inhalation coinciding with speech in other speakers' channels were excluded from the analysis. While this greatly reduced the sample size, it ensures validity of the obtained results.

To enable comparisons between acoustic and kinematic measures, the following features of the inhalation were extracted from the RIP signal: duration, amplitude, slope and the minimum lung volume with respect to the resting expiratory level (REL). Inhalation amplitude and inhalation starting level with respect to REL were expressed as percentages of speakers respiratory range, whose limits were estimated at 5^{th} and 95^{th} percentiles of all peaks and troughs in the respiratory cycles observed so far. REL itself was estimated at the median level of troughs in the previous 20 cycles. The features are represented schematically in Figure 2.

All in all, the analysed sample consisted of 82 silent cycles, 180 speech cycles 77 VSU cycles.

3. Results

Distributions of inhalation intensity (z-scored per speaker) are plotted in Figure 3. Although the three distributions overlapped to a large extent, a clear tendency could be seen for inhalations in silent cycles to be quieter than before speech, with VSU cycles falling more or less half-way between the two.

A multinomial logistic regression model with cycle type as

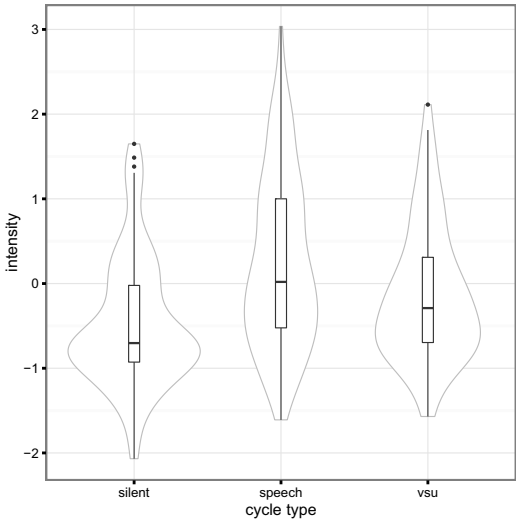

Figure 3: Distribution of inhalation intensity (z-scored per speaker) of the inhalation across cycle types

Table 1: Coefficients of the multinomial logistic regression model predicting whether an inhalation was followed by speech, a VSU or a silent exhalation based on inhalation intensity (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	0.941	2.563	1.883	3.510	0.000
	Inhalation intensity	0.978	2.660	1.818	3.899	0.000
VSU	Constant	0.109	1.116	0.771	1.599	0.535
	Inhalation intensity	0.537	1.711	1.136	2.624	0.009

Note. $R^2 = .05$, (McFadden), .12 (Cox & Snell), .10 (Nagelkerke).
Model $\chi^2(2) = 37.334, p < .001$

dependent variable and intensity as independent variable was fitted to the data. The model is summarised in Table 1, which indicates that inhalation intensity significantly distinguished between silent cycles and the other two cycle types. The effect was particularly pronounced for speech cycles: for each unit increase of intensity (i.e. a difference of one standard deviation) the probability of the cycle coinciding with speech increased 2.7 times. For VSU the odds ratio was smaller but still substantial and equalled 1.7.

As noted in the introduction, acoustic measures of breath offer an attractive alternative to capturing respiratory kinematics by means of RIP or magnetometers. For this reason, below we compare the results of the acoustic analysis with kinematic measures of respiratory activity. As could be expected, acoustic intensity was mainly correlated with slope of the inhalation (0.43), and consequently also with inhalation duration (-0.38). By contrast, the correlation with inhalation amplitude (0.04) and the initial lung volume of the inhalation relative to REL (0.02) was weak.

Subsequently, another multinomial logistic regression model was fitted with inhalation duration and slope as predictors of cycle type. The other kinematic features were not included in the model since they did not significantly improve the model fit. The final model provides comparable fit to the model using inhalation intensity as the sole predictor (R^2 = .07 (McFadden), .14 (Cox & Snell), .13 (Nagelkerke), model $\chi^2(4) = 45.383, p < .001$).

Finally, we fit a full model with both kinematic and acoustic features included. The resulting model improved significantly on both models (R^2 = 0.09 (McFadden), 0.18 (Nagelkerke), 0.16 (Cox & Snell), model $\chi^2(6) = 58.597$, $p < .001$). Due to relatively high collinearity between the predictors, we do not report estimates of individual parameters as they are likely to be biased. However, assuming that the degree of collinearity in the data is representative of its value in the population, it should not adversely affect estimate of the overall model fit.

4. Discussion

Deciding whether a dialogue participant is going to start speaking within the immediate future has been one of the key problems in dialogue research. Respiration provides a particularly viable solution to this problem. However, standard ways of capturing breathing kinematics by means of respiratory belts (whether RIP or magnetometers) is not without drawbacks. In our experience, setting up a three-party recording with respiratory belts takes from 20 to 30 minutes. In addition, some participants are noticeably uncomfortable with the calibration procedure. Using respiratory acoustics instead of kinematic measures alleviates all these concerns. Capturing breathing sounds is fast, calibration-free and unobtrusive. In fact, respiratory noises are part of any normal speech recording and special care must be taken *to remove them* for specific applications, such as VOD. Indeed, the very fact that VOD algorithms can be misled to treat inhalations as speech segments testifies to their perceptual salience.

Consequently, in the present paper we presented results on acoustic inhalation intensity as a cue to speech initiation. As expected, we found inhalations preceding speech to be louder than those in tidal breathing and before backchannel-like segments. Importantly, the model using two kinematic features of the inhalation, namely its duration and slope, fitted the data approximately as well as the acoustic model. Finally, fusion of the two models results in substantially better fit.

Notably, although the initial size of the analysed material amounted to over nine hours worth of respiratory data, the resulting sample was quite small due to stringent filtering criteria. For this reason, the results are only a preliminary account of respiratory acoustics for turn-taking. In addition, the sample came from a large number of speakers and is consequently sensitive to substantial inter-speaker variability. That we nevertheless get statistically significant results testifies to relative strength of the effect under investigation.

We consider the results to be a preliminary towards studying communicative functions of breathing acoustics, particularly in connection with turn-management in spontaneous conversation. On that view, our results could be thought of as exemplifying "intentionality" of respiratory planning, posited elsewhere by Huber [31]. Huber showed that participants modify their respiratory kinematics depending on the type of cue to elicit loud speech. In other words, breathing is a purposeful activity governed by speaker's communicative and physiological goals. Similarly, Włodarczak et al. [32] showed that communicative constraints interact with physiological constraints to produce respiratory patterns accompanying backchannels and longer stretches of speech. Here we claim that a similar degree of goal orientation can be involved in signalling upcoming speech.

Admittedly, at present we have no evidence, nor do we want to claim, that producing audible inhalations is a *conscious* action on the part of the speaker. In other words, in semiotic terms the information contained in the breath noise may be indicated

rather than displayed or signalled [33]. The greater loudness might as well be produced by increased airflow passing through a relatively narrow constriction in the vocal tract. On that view, it might be brought about by the need to inhale quickly in preparation for speech while minimising articulatory effort linked to reaching a more open vocal tract configuration. For this reason, similar to our earlier work [32], we leave the question of traditionally conceived intentionality open and prefer to discuss it instead in terms of context sensitivity and mutual dependence of communicative requirements and physiological constrains. We consider it to be a more fruitful and ecologically valid perspective than abstract concepts of intentionality decoupled from its physical context.

5. Conclusions and future work

The present paper is, to the best of our knowledge, the first study of breath acoustics as a cue to speech initiation in spontaneous multiparty dialogues. As expected, the inhalatory noise was louder before speech than in tidal breathing. We have demonstrated that models using acoustic and kinematic features provide similar fit to the data. In addition, using both feature sets provides a substantial improvement over individual models. The results are thus a first step towards a more comprehensive account of turn-management functions of breath acoustics and its evaluation for speech technology applications.

The preliminary character of the results is mainly due to a relatively small sample size resulting from a strict filtering criteria necessitated by crosstalk in the acoustic signal. For this reason, we are planning another data collection effort ensuring better channel separation and, consequently, less data loss in the resulting sample. Among other possibilities, we are currently exploring possibilities of recording respiratory sounds using contact microphones attached directly to speaker's neck. We hope that this technique will allow better classification of the respiratory cycle type.

Lastly, this paper has been only concerned with cues to speech initiation. However, breath could also be potentially relevant as a cue to turn-yielding, where an audible *exhalation* could be an indication that the speaker is finished speaking. We are planning to explore this hypothesis in the future.

6. Acknowledgements

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

7. References

[1] P. Forgacs, A. R. Nathoo, and H. D. Richardson, "Breath sounds," *Thorax*, vol. 26, no. 3, pp. 288–295, 1971.

[2] P. Price, M. Ostendorf, and C. Wightman, "Prosody and parsing," in *Proceedings of the DARPA Workshop on Speech and Natural Language*, 1989, pp. 5–11.

[3] C. W. Wightman and M. Ostendorf, "Automatic recognition of prosodic phrases," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91)*. IEEE, 1991, pp. 321–324.

[4] T. Nakano, J. Ogata, M. Goto, and H. Yuzuru, "Analysis and automatic detection of breath sounds in unaccompanied singing voice," in *Proceedings of the 10 th International Conference on Music Perception and Cognition (ICMPC 10)*, Sapporo, Japan, 2008, pp. 387–390.

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

[6] T. Fukuda, O. Ichikawa, and M. Nishimura, "Breath-detection-based telephony speech phrasing," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011, pp. 2625–2628.

[7] N. Gavriely, Y. Palti, and G. Alroy, "Spectral characteristics of normal breath sounds," *Journal of Applied Physiology*, vol. 50, no. 2, pp. 307–314, 1981.

[8] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clinical medicine: Circulatory, respiratory and pulmonary medicine*, vol. 2, pp. 45–58, 2008.

[9] S. Sundaram and S. Narayanan, "Spoken language synthesis: Experiments in synthesis of spontaneous monologues," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002, pp. 203–206.

[10] J. Trouvain and B. Möbius, "Einatmungsgeräusche vor synthetisch erzeugten Sätzen – Eine Pilotstudie," in *Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013)*, ser. Studientexte zur Sprachkommunikation, P. Wagner, Ed., vol. 65. Dresden: TUDpress, 2013, pp. 50–55.

[11] J. Butzberger, H. Murveit, E. Shriberg, and P. Price, "Spontaneous speech effects in large vocabulary speech recognition applications," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 339–343.

[12] Y.-T. Wang, I. S. B. Nip, J. R. Green, R. D. Kent, J. F. Kent, and C. Ullman, "Accuracy of perceptual and acoustic methods for the detection of inspiratory loci in spontaneous speech," *Behavior research methods*, vol. 44, no. 4, pp. 1121–1128, 2012.

[13] D. H. Whalen, C. E. Hoequist, and S. M. Sheffert, "The effects of breath sounds on the perception of synthetic speech," *The Journal of the Acoustical Society of America*, vol. 97, pp. 3147–3153, 1995.

[14] K. H. Kendrick and F. Torreira, "The timing and construction of preference: A quantitative study," *Discourse Processes*, vol. 52, pp. 255–289, 2015.

[15] C. Yuan and A. Li, "The breath segment in expressive speech," *Computational Linguistics and Chinese Language Processing*, vol. 12, no. 1, pp. 17–31, 2007.

[16] R. Pellegrini and M. R. Ciceri, "Listening to and mimicking respiration: Understanding and synchronizing joint actions," *Review of Psychology*, vol. 19, no. 1, pp. 17–27, 2012.

[17] G. Bailly and C. Gouvernayre, "Pauses and respiratory markers of the structure of book reading," in *Proceedings of Interspeech 2012*, Portland, OR, 2012.

[18] D. H. Whalen and S. M. Sheffert, "Normalization of vowels by breath sounds," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 133–143.

[19] ——, "Perceptual use of vowel and speaker information in breath sounds," in *Proceedings of ICSLP 96*, H. T. Bunnell and W. Idsardi, Eds., 1996, pp. 2494–2497.

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

[21] 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.

[22] 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.

[23] 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.

[24] M. Włodarczak and M. Heldner, "Respiratory properties of backchannels in spontaneous multiparty conversation," in *Proceedings of the 18 International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, UK, 2015.

[25] ——, "Respiratory turn-taking cues," in *Proceedings of Interspeech 2016*, San Francisco, CA, 2016.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] J. E. Huber, "Effect of cues to increase sound pressure level on respiratory kinematic patterns during connected speech," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 3, pp. 621–634, 2007.

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

[33] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.