INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving prediction of speech activity using multi-participant respiratory state

Marcin Włodarczak[1], *Kornel Laskowski*[2,3], *Mattias Heldner*[1], *Kätlin Aare*[1,4]

[1]Stockholm University, Sweden,
[2]Carnegie Mellon University, Pittsburgh PA, USA
[3]Voci Technologies, Inc., Pittsburgh PA, USA
[4]University of Tartu, Estonia

wlodarczak@ling.su.se, kornel@cs.cmu.edu,
heldner@ling.su.se, katlin.aare@ut.ee

Abstract

One consequence of situated face-to-face conversation is the co-observability of participants' respiratory movements and sounds. We explore whether this information can be exploited in predicting incipient speech activity. Using a methodology called stochastic turn-taking modeling, we compare the performance of a model trained on speech activity alone to one additionally trained on static and dynamic lung volume features. The methodology permits automatic discovery of temporal dependencies across participants and feature types. Our experiments show that respiratory information substantially lowers cross-entropy rates, and that this generalizes to unseen data.

Key words: respiratory kinematics, interaction chronograms, stochastic turn-taking models

1. Introduction

Earlier work has firmly established the existence of respiratory turn-taking cues. In particular, several studies [1, 2, 3] have identified temporal compression, whereby the lag preceding and following the inhalation as well as the duration of the inhalation itself are reduced, a practice interpreted as a strategy to minimize the likelihood of pause interruption. In addition, exhalations were shown to be longer before speaker changes than before turn continuations [4], and preparing to take the turn was associated with an increased inhalation amplitude compared to the silent breathing pattern [2]. There were also some indications that turn-initial inhalation was deeper than inhalations made later in the turn [4, 5]; however, these findings were not confirmed by other studies [2], or were only observed in scripted dialogues [4]. By contrast, [2] found that inhalation tends to be deeper in turn-holding than in turn-changing pauses, although the size of this effect seemed to be rather small. By including an additional category of backchannel-like utterances in our own work we have been able to identify consistent variation in inhalation amplitude across turn-categories [3]. We also found that relatively increased lung volume at the inhalation onset cued speech inhalations, which we hypothesized to be a strategy for arriving at timely speaker transition.

While the results of all the above studies are potentially relevant for the prediction of speech in technological applications, the studies done so far have generally lacked formal evaluation of the results for online prediction of turn-taking. The only exception is Ishii et al. [2], who tested a three-stage prediction model. At each utterance offset, the model predicts: (1) *whether* the same speaker (versus another) is going to continue, (2) in case of speaker change, *which* of the three conversation partners is going to take the turn, and (3) *when* the next utterance is going to start (whether from the previous speaker or one of the listeners). They demonstrated that inclusion of respiratory features for prediction of turn-holding vs. speaker-change, as well as for prediction of next speaker in multiparty conversation, outperforms a random baseline. In addition, they found that for prediction of turn-holding vs. speaker-change, listeners' features outperformed speaker's features, and that fusion of the two feature sets resulted in the best performance. Regarding (3), inclusion of respiratory features was found to reduce prediction errors over a baseline of mean pause duration.

In the present paper, we use stochastic turn-taking (STT) modelling [6, 7] as a convenient means of answering the following questions:

Q1: Is there information in the breathing signal that is helpful for the prediction of speech activity in multiparty conversation?

Q2: How should the respiratory information be represented to maximize feature utility?

Q3: Is a participant's[1] breathing signal correlated with their interlocutors' future vocal activity?

Our experiments allow us to answer Q1 in the affirmative and Q3 in the negative, as well as to provide several guidelines with respect to Q2.

The resulting study is therefore an important extension of our earlier work on respiratory turn-taking cues [3]. In that work, we had modeled each conversation partner individually, without taking their interlocutors' behavior into consideration. Furthermore, we had predicted speech activity for each respiratory cycle, whereas in the current work predictions are made every 100 ms, from the beginning of the conversation to its end. In this second respect, the current article also extends the work in [2], where predictions were made only at turn landmark locations. In addition, unlike the authors of [2], who compared their system to a random baseline, the current article compares the contribution of breathing in the context of multi-participant speech activity to a baseline trained on multi-participant speech activity alone.

Overall, the results indicate that the breathing signal provides additional information for predicting speaker state in conversation, which is fundamental to understanding the mechanisms of human interaction. While the applications are limited by the data acquisition method used (respiratory belts wrapped

[1]We henceforth use the term "participant" in place of the term "speaker" in order to stress that we are making predictions for all participants at all instants, regardless of whether they are speaking or not.

http://dx.doi.org/10.21437/Interspeech.2017-1176

around the interlocutors' upper body), our earlier work suggests that using respiratory acoustics is a promising alternative [8]. Other methods of measuring respiration remotely have also been recently proposed, see for instance [9, 10, 11].

2. Methods

2.1. Signal Collection

The material used in this study consisted of eight three-party recordings in Swedish and thirteen three-party recordings in Estonian. Each recording lasted for about 25 minutes and comprised conversation without a pre-defined topic. All participants were native speakers of the respective languages and, with the exception of one group, had known each other prior to the recording. No person participated in more than one conversation. The study has been approved by the Regional Ethical Committee in Stockholm (2015/63-31).

All recordings were made in the Phonetics Laboratory at Stockholm University. Each participant wore two elastic respiratory belts which measure expansion of the chest and the abdomen due to breathing. The respiratory signal was routed through custom-built processors (RespTrack) to an integrated data acquisition system (PowerLab hardware and LabChart software by ADInstruments). The relative contributions of the chest and abdomen belts to the total lung volue change had been established using the isovolume maneouver [12], and the weighted summed signal was used in subsequent analyses. The resulting per-participant continuous breathing signal was sampled at 20 kHz and stored alongside each participant's audio.

Audio was captured with close-talking directional microphones (Sennheiser HSP 4), and routed to PowerLab to allow for synchronization. The subjects were recorded while standing around a high bar table to minimize the impact of posture shifts on the respiratory signal. Video was recorded using GoPro Hero3+ cameras placed on the table, but was not used in the current work. For a more detailed description of the recording setup, see [13].

Of the twenty one dialogues thus recorded, six in Swedish and six in Estonian were placed in TRAINSET for training models. The remaining two in Swedish and two of the available seven in Estonian were placed in TESTSET; five Estonian dialogues were not used, in order to retain the same language balance in TRAINSET and TESTSET. Since no participants had taken part in more than one conversation, the two sets are disjoint in participants, rendering our analysis participant-independent.

2.2. Signal Representation

Following recording, intervals of vocalization were identified automatically using intensity-based segmentation in ELAN [14] and corrected manually.

To permit quantitative exploration of questions Q1–Q3, we first synchronize the vocal activity signal across the three participants in our data, and then discretize it in time. Discretization consists of splitting continuous time during each conversation into a contiguous sequence of 100-ms frames; a particular frame at discrete instant t for participant k is declared ■ if the kth participant was vocalizing for more than 50 ms of that 100-ms interval, and □ otherwise. This representation naturally leads to a two-dimensional matrix $\mathbf{Q}$ called a vocal interaction chronogram, whose number of rows for all of our data is $K \equiv 3$, whose number T of columns is the number of non-overlapping 100-ms frames, and whose entries are drawn from the set $\{\square, \blacksquare\}$. The tth column of $\mathbf{Q}$ will henceforth be denoted by the vector $\mathbf{q}_t$, and its value for the kth participant at the tth instant will be denoted by $\mathbf{q}_t[k]$. As additional shorthand, $\mathbf{Q}_a^b$ denotes the closed sequence of vectors from $\mathbf{q}_a$ to $\mathbf{q}_b$ inclusive.

A very similar synchronization and discretization-in-time is performed to obtain the matrices of static breathing information $\mathbf{B}$ and dynamic breathing information $\dot{\mathbf{B}}$. Each of these is also of dimensions $K \times T$, but their entries — $\mathbf{b}_t[k]$ and $\dot{\mathbf{b}}_t[k]$, respectively, for $1 \leq t \leq T$ and $1 \leq k \leq K$ — are drawn from $\Re$. Specifically, the value of $\mathbf{b}_t[k]$ is the average of the breathing signal for the kth participant over a 100-ms frame centered at the tth instant. In contrast, the value of $\dot{\mathbf{b}}_t[k]$ is the slope of the least-mean-square linear fit to the breathing signal for the kth participant over the same interval. While the discretization does reduce the temporal resolution of the signal, the time window used is considerably below the duration of typical respiratory events and is therefore assumed to not deleteriously impact our findings.

Given $\mathbf{Q}$, $\mathbf{B}$, and $\dot{\mathbf{B}}$ for any conversation, operationalizing question Q1 consists of comparing how well $\mathbf{q}_t[k]$ can be predicted, using $\mathbf{Q}_{t-1}^1$ alone, to how well it can be predicted when additionally exposed to $\mathbf{B}_{t-1}^1$ and/or $\dot{\mathbf{B}}_{t-1}^1$. In addition, we expect that a comparison of the contributions of $\mathbf{B}$ and/or $\dot{\mathbf{B}}$ may help to answer Q2. Note that in predicting $\mathbf{q}_t[k]$ we are predicting the incipient speech activity of the kth participant, which we refer to as the *target participant* for convenience. For tractability reasons, variable-duration conditioning histories such as $\mathbf{B}_{t-1}^1$ are truncated to $\mathbf{B}_{t-1}^{t-S}$, consisting of only the S most recent frames; in the current article, $S \equiv 10$, or one second of context. More specifically, we train a baseline model which ignores breathing information to provide $P\left(\mathbf{q}_t[k] = \blacksquare | \mathbf{Q}_{t-1}^{t-S}\right)$, and a breathing-sensitive model to provide $P\left(\mathbf{q}_t[k] = \blacksquare | \mathbf{Q}_{t-1}^{t-S}, \mathbf{B}_{t-1}^{t-S}\right)$. Both models yield the probability that the kth participant vocalizes[2] at the tth instant, but are conditioned on histories differing in feature type.

The predictions of both models for $P(\mathbf{q}_t[k])$ can be compared to $\mathbf{q}_t[k]$, that is what subsequently happened at t, to determine which model is *more accurate on average* for a dataset, by computing the cross entropy averaged over all K and all T in all chronograms in that dataset.

2.3. Probability Modeling

We propose to compute the probability that the kth participant vocalizes at the tth instant, conditioned on the recent vocalization and respiration history, using

$$P\left(\mathbf{q}_t[k] = \blacksquare | \mathbf{Q}_{t-1}^{t-S}, \mathbf{B}_{t-1}^{t-S}\right) \doteq f\left(\mathbf{Q}_{t-1}^{t-S}, \mathbf{B}_{t-1}^{t-S}\right), \quad (1)$$

estimating $f(\cdot)$ using a feed-forward neural network to accommodate both the discrete values of $\mathbf{Q}$ and the continuous values of $\mathbf{B}$ (and/or of $\dot{\mathbf{B}}$). The proposed neural network is quite simple, consisting of one hidden layer with J dot-product/tanh hidden units and one dot-product/sigmoid output unit, making the output of the network interpretable as a probability $\in (0, 1)$.

For the comparison with a breathing-insensitive baseline (BL) to be most meaningful, $f_{BL}(\cdot)$ in

$$P_{BL}\left(\mathbf{q}_t[k] = \blacksquare | \mathbf{Q}_{t-1}^{t-S}\right) \doteq f_{BL}\left(\mathbf{Q}_{t-1}^{t-S}\right) \quad (2)$$

should have the same form as $f(\cdot)$. Because we want to enable the learning of dependencies between breathing and vocaliza-

[2] By definition, $P(\mathbf{q}_t[k] = \square | \cdots) \equiv 1 - P(\mathbf{q}_t[k] = \blacksquare | \cdots)$, such that the sum of the probabilities of vocalizing and of not vocalizing is unity for any participant k at any instant t.

tion histories for any participant, the input feature space must explicitly represent each of the K participants. Consequently, all K participants should to be explicitly represented in the input vector to $f_{BL}(\cdot)$. Such a baseline has only been proposed for $K = 2$. We extend it to $K > 2$, and evaluate it for $K = 3$, in the following section.

3. Baseline Development

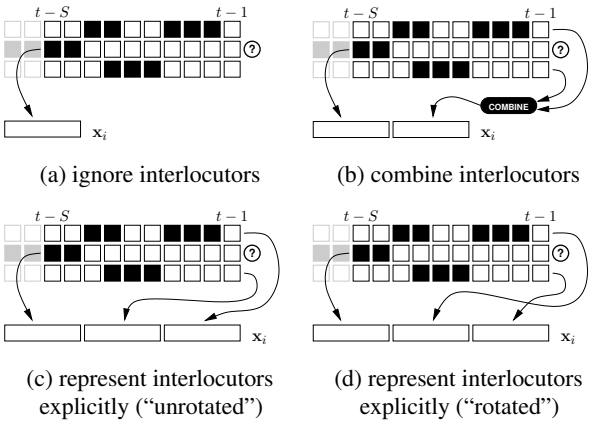
A number of STT models relying on only vocal activity history have been proposed in our earlier work. These models come in two flavors: the "mutually independent (MI) participants" model type assumes that each participant's vocal behavior is *not* informed by (i.e. "is independent of") their interlocutors' vocal behavior history, while the "conditionally independent (CI) participants" model type assumes that participants' vocal behavior is conditionally independent, given all K participants' joint vocal behavior history. With a conditioning history of S most recent frames, the input vector $\mathbf{x}_i$ to $f_{BL}(\cdot)$ for an MI model consists of S binary values as shown in Figure 1.a. When $f_{BL}(\cdot)$ is implemented as a Jelinek-Mercer-smoothed n-gram model as described in [15], the cross-entropies for TRAINSET and TESTSET are 0.256 bits/100ms and 0.239 bits/100ms, respectively; these are shown as baseline "1" in Figure 2. Baseline "2" represents an NN-based MI implementation [6] as described in Subsection 2.3; as for all other NNs in this article, we used $J = 32$ hidden units (decided using TRAINSET), with one hundred iterations of scaled conjugate gradient (SCG) pre-training[3] and one thousand iterations of SCG training. The performance of baselines "1" and "2" differs only negligibly[4].

For systems intended to operate on conversations with arbitrary K, a CI model using only vocal activity can be obtained by appending a fixed-length representation of the target participant's interlocutors' vocal activity history to the MI-model feature vector, as shown in Figure 1.b. When $K = 2$, this appendix can be the complete and explicit vocal activity history of the target participant's single interlocutor; when $K > 2$, a simple means of combining the $K - 1$ interlocutor histories is to form their inclusive-OR [16], thereby capturing whether *zero* or *at-least-one* interlocutor had been speaking at an instant $t - s$, $1 \leq s \leq S$.

An n-gram-based implementation of such a CI system [15] is depicted as baseline "3" in Figure 2. It can be seen that, for both TRAINSET and TESTSET, it reduces cross-entropies by approximately 0.006 bits/100ms. This is very similar to our observations made for other non-telephony conversational corpora [15, 17]. The performance of baseline "4" — the NN-based counterpart [6] to baseline "3" — indicates no relative advantage for either, as expected. Baseline "5" provides an alternative NN-based model, which also uses the feature vector construction method in Figure 1.b but the interlocutor portion of the vector contains the integer *number* of vocalizing interlocutors at instant $t - s$, a ternary variable for $K = 3$. As Figure 2.b shows, there

[3]We use this term to denote training for 100 iterations using every 1024th exemplar, then for 100 iterations using every 512th exemplar, etc, and finally for 100 iterations using every 2nd exemplar.

[4]For TRAINSET, baselines 2 and 4 exhibit slightly lower cross entropies than baselines 1 and 3, because the latter were smoothed with parameters selected to minimize cross entropy on TESTSET. For this same reason, baselines 2 and 4 exhibit slightly higher cross entropies than baselines 1 and 3 for TESTSET. This is the only occurrence in the current article where our TESTSET was used to make a system structure/parameter decision; in all other experiments, TESTSET can be treated as a truly held-out (non-development) data set, particularly since baselines "1" and "3" are not used in answering questions Q1–Q3.

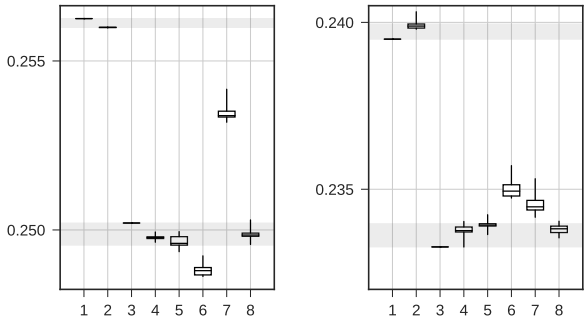


(a) ignore interlocutors (b) combine interlocutors

(c) represent interlocutors explicitly ("unrotated") (d) represent interlocutors explicitly ("rotated")

Figure 1: *Four alternative methods of marshalling a snippet* $\mathbf{Q}_{t-1}^{t-S}$ *of a chronogram into a fixed-length feature vector* $\mathbf{x}_i$*, shown for a history duration of* $S = 10$ *preceding 100-ms frames in chronograms with* $K = 3$ *participants. In the diagram, the target participant for whom the prediction is being made is the participant associated with the second row of the chronogram.*

appears to be no benefit to knowing how many interlocutors are vocalizing, just that at least one is. None of the three CI baselines "3", "4", and "5" model interlocutors explicitly, making these systems unsuitable for subsequently associating a participant's breathing activity with their vocalization activity.



(a) TRAINSET (b) TESTSET

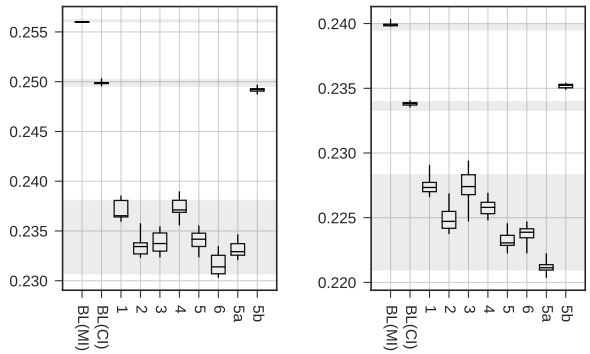
Figure 2: *Cross entropies (in bits per 100-ms frame, along the* y*-axis) for 8 baseline systems shown along the* y *axis. Progression from baseline "1" through to "8" as described in the text. The shaded areas denote extrema of IQR's of MI (1, 2) and CI (3, 4, 5, 8) baselines, respectively.*

Our first attempt at representing all $K = 3$ participants' vocal activity histories in a feature vector is shown in Figure 1.c: the vector is enlarged relative to Figure 1.b and each of the $K - 1 = 2$ interlocutors' history is explicitly stored in it, in addition to the history of the target participant. Unfortunately, a baseline like this (denoted "6" in Figure 2) is sensitive to the ordering of the participants in the chronogram. To see this, it suffices to apply the marshalling method in Figure 1.d to TRAINSET during training, but retain Figure 1.c during testing as for baseline "6"; cross-entropies for this case are shown as baseline "7". It can be seen that this renders TRAINSET cross entropies almost as bad as if interlocutor information had not

been included at all; for TESTSET, neither baseline "6" nor "7" is as good as "4". The solution we propose is to duplicate TRAINSET feature vectors for training, marshalling one copy as in panel (c) of Figure 1 and the other as in panel (d). This yields our final baseline whose performance on unseen data (cf. Figure 2.b) is no worse than baseline "4", with the benefit that it represents all participants' vocal activity history explicitly.

4. Results

Given our final CI baseline, breathing history is easily marshalled into feature vectors in exactly the same way as vocal activity history. The results are shown in Figure 3, for both TRAINSET and TESTSET; the MI and CI baselines are denoted "BL(MI)" and "BL(CI)", and correspond to baselines "2" and "8" from the preceding section.



(a) TRAINSET (b) TESTSET

Figure 3: *Cross entropies (in bits per 100-ms frame, along the y-axis) for 8 respiration-sensitive systems (along the x-axis); baseline "BL(MI)" and "BL(CI)" correspond to baselines "2" and "8" in Figure 2. The shaded areas denote extrema of IQR's of MI and CI baselines (retained from Figure 2) as well as systems "1" through "5a".*

System "1" extends BL(CI) by exploiting $\mathbf{B}_{t-1}^{t-S}$ in addition to $\mathbf{Q}_{t-1}^{t-S}$. For TRAINSET, the resulting cross entropy reduction is approximately 0.012 bits per 100 ms, or twice as large as the gap between the MI and CI baselines. This effect generalizes to TESTSET, where its magnitude is 0.006 bits per 100 ms, similar to the difference between BL(CI) and BL(MI).

System "2" uses $\dot{\mathbf{B}}$ instead of $\mathbf{B}$. Figure 3 shows that for both TRAINSET and TESTSET, this offers a considerable improvement. Evidently, the information found in the instantaneous slope of lung volume is more relevant to vocal activity prediction than is information found in the instantaneous value of lung volume—despite the fact that the neural network can learn to compute rate of change from $S \equiv 10$ consecutive values. System "3" uses both $\mathbf{B}$ and $\dot{\mathbf{B}}$, but the experiments in Figure 3 suggest that $\mathbf{B}$ provides no additional information over and above $\dot{\mathbf{B}}$: it merely provides opportunity for the neural network to overfit to TRAINSET.

Systems "4" through "6" are structurally identical to systems "1" through "3", except that the respiration signal chronograms are Z-normalized using chronogram-row-specific statistics. This makes these systems acausal, since at each instant t the mean and variance used to normalize each $\mathbf{b}_t[k]$ had been computed using instants $t' < t$ but also $t' \geq t$. Nevertheless, these experiments indicate the performance that could be obtained if these statistics were available through other means. It appears that Z-normalization is helpful, and generalizes to TESTDATA.

In a final set of experiments, we compare the relative contribution of target participant's vs. their interlocutors' respiration history. Starting with the best-generalizing system — that labeled "5" which exploits the Z-normalized dynamic respiration signal $\dot{\mathbf{B}}$ — we supress either the interlocutors' respiration history (in system "5a") or the target participant's respiration history (in system "5b"). It can be seen in the cross-entropies for TESTSET that the target participant's respiration history ("5a") offers an even bigger improvement over the CI baseline than does inclusion of all participants' respiration history ("5"). Meanwhile, excluding the target participant's respiration history ("5b") renders the system not significantly different from the CI baseline which ignores breathing altogether.

5. Discussion and conclusions

The experiments described in this paper permit us to answer Q1 in the affirmative: the multi-participant breathing signal contains information which can be leveraged to improve prediction of incipient participant-attributed vocal activity. Specifically, multi-participant respiration history offers roughly as much cross entropy reduction as the inclusion of interlocutors' vocal activity history. Furthermore, we note that since no TESTSET participants were present in TRAINSET, these findings generalize to unseen data and are participant-independent.

The experiments also support a tentative answer to Q2: dynamic lung volume features (i.e. slope) offer an advantage over the static ones, a trend which also generalizes to unseen participant-independent data. Furthermore, Z-score normalization was shown to be beneficial, and the Z-normalized slope was the best-performing feature of the ones we have tried. In fact, the improvement due to the inclusion of Z-scored slope is about twice as large as that due to the inclusion of interlocutor vocal activity history.

We have also contrasted the contribution of target participants's breathing history to that of their interlocutors'. Our experiments show that each participant's breathing is helpful, but only for the prediction of *that* participant's incipient speech activity; it appears unhelpful for the prediction of others' incipient speech activity. Note that this is irrespective of which participants happen to be currently speaking; unlike [2], our concept of "participant" includes both current speakers and current listeners. This answers Q3 in the negative.

In answering these questions, we have extended the stochastic turn-taking framework to make it possible to explicitly attribute features of any type to participants in conversations with more than two participants; previously, this could only be done for dyadic conversations. In future work, we will inspect the trained models for the type of patterns that have been learned and evaluate the impact of the size of the context on model performance. We will also evaluate the combination of breathing with other continuous features such as intensity and fundamental frequency.

6. Acknowledgements

This work was funded in part by the Swedish Research Council grant 2014-1072 *Andning i samtal* (*Breathing in conversation*) to the first author. Computing resources at Carnegie Mellon University were made accessible courtesy of Florian Metze.

7. References

[1] 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.

[2] 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.

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

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

[5] 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.

[6] K. Laskowski, "Exploiting loudness dynamics in stochastic models of turn-taking," in *Proceedings of the 4th IEEE Workshop on Spoken Language Technology (SLT2012)*, Miami, FL, 2012, pp. 79–84.

[7] K. Laskowski and A. Hjalmarsson, "An information-theoretic framework for automated discovery of prosodic cues to conversational structure," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2015)*, Brisbane, Australia, 2015, pp. 5376–5380.

[8] 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, pp. 510–514.

[9] J. Fei and I. Pavlidis, "Thermistor at a distance: unobtrusive measurement of breathing," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 988–998, 2010.

[10] J. Xia and R. A. Siochi, "A real-time respiratory motion monitoring system using kinect: Proof of concept," *Medical Physics*, vol. 39, no. 5, pp. 2682–2685, 2012.

[11] I. van Dijk and A. Heinrich, "Vital signs camera," in *Proceedings of the international workshop on computer vision applications (CVA)*, P. H. N. de With and P. Shrestha, Eds., Eindhoven, the Netherlands, 2011, pp. 107–109.

[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] 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] K. Laskowski, J. Edlund, and M. Heldner, "Incremental learning and forgetting in stochastic turn-taking models," in *Proceedings Interspeech 2011*, Florence, Italy, 2011, pp. 2069–2072.

[16] ——, "A single-port non-parametric model of turn-taking in multiparty conversation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, 2011, pp. 5600–5603.

[17] K. Laskowski, "Transfer cross entropy for fast sociometric inference in longitudinal collections of multi-party conversation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, 2012, pp. 2189–2192.