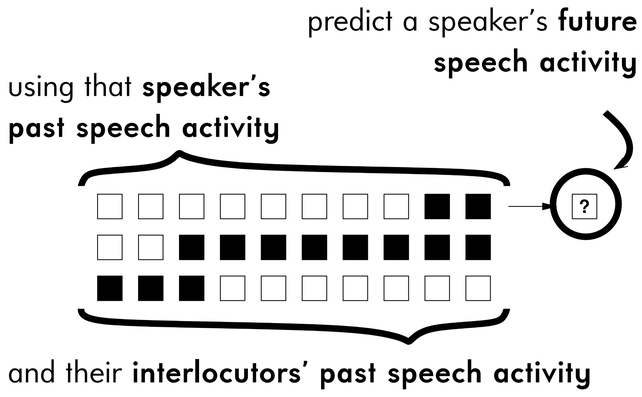
Improving prediction of speech activity using multi-participant respiratory state

Marcin Włodarczak[1], Kornel Laskowski[2,3], Mattias Heldner[1], Kätlin Aare[1,4]
{wlodarczak,heldner}@ling.su.se, kornel@cs.cmu.edu, katlin.aare@ut.ee

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

Stochastic turn-taking models



Questions

Is breathing helpful for predicting vocal activity?

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

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

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

Signal representation

How to quantize the breathing signal?

- **Average** of the breathing signal over a frame (B)
- **Slope** of the breathing signal over a frame (B')

In addition, we evaluate the **Z-normalised** versions of these features (B_z and B'_z).

In all experiments, we use ten previous 100-ms frames, i.e. **one second of history**.

Data

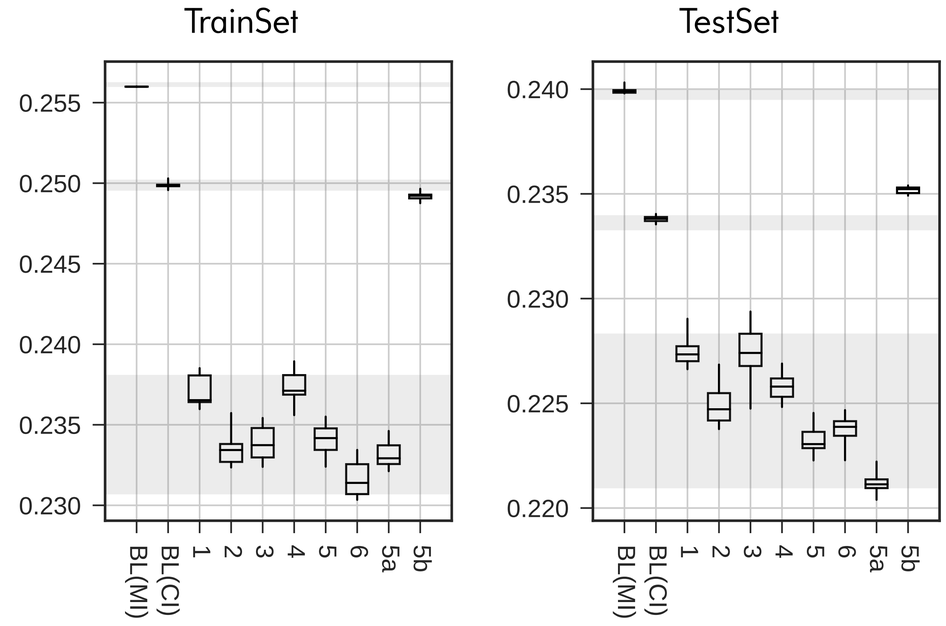
16 **three-party conversations** (8 in SWE, 8 in EST) without a predefined topic, each about 25 minutes long.

Breathing captured with two **respiratory belts** wrapped around the upper body (Respiratory Inductance Plethysmography).

Voice activity (VA) classification based on manually corrected intensity-based segmentations.

12 dialogues used for training and 4 for testing.

Results



Cross entropies (in bits per 100-ms frame, along the y-axis) for two baselines and eight respiration-sensitive systems (along the x-axis).

System comparison

System	Features	Performance
BL(MI)	participant's VA	—
BL(CI)	everyone's VA	Better than BL(MI)
1	B	Better than BL(CI)
2	B'	Better than 1
3	B, B'	No improvement
4	B_z	Better than 1
5	B'_z	Better than 2
6	B_z, B'_z	Better than 3
5a	particpant's B'_z	Best of all
5b	interlocutors' B'_z	Like BL(CI)

Answers

Yes, breathing does help to predict future vocal activity!

A1: Yes, inclusion of multi-participant respiratory history helps roughly as much as inclusion of interlocutors' vocal activity history

A2: Dynamic features (slope) outperform static features, Z-normalisation offers further improvement.

A3: No, breathing is only helfpful for the prediction of **this participant's** incipient vocal activity.