Sensing of inspiration events from speech

SENSING OF INSPIRATION EVENTS FROM SPEECH: COMPARISON OF DEEP LEARNING AND LINGUISTIC METHODS

A PREPRINT

Aki Härmä[1][*], Ulf Grossekathöfer[1], Okke Ouweltjes[1] and Venkata Srikanth Nallanthighal[2]

May 22, 2023

ABSTRACT

Respiratory chest belt sensor can be used to measure the respiratory rate and other respiratory health parameters. Virtual Respiratory Belt, VRB, algorithms estimate the belt sensor waveform from speech audio. In this paper we compare the detection of inspiration events (IE) from respiratory belt sensor data using a novel neural VRB algorithm and the detections based on time-aligned linguistic content. The results show the superiority of the VRB method over word pause detection or grammatical content segmentation. The comparison of the methods show that both read and spontaneous speech content has a significant amount of ungrammatical breathing, that is, breathing events that are not aligned with grammatically appropriate places in language. This study gives new insights into the development of VRB methods and adds to the general understanding of speech breathing behavior. Moreover, a new VRB method, VRBOLA, for the reconstruction of the continuous breathing waveform is demonstrated.

Index Terms: Pathological speech sensing, speech breathing, respiratory health and fitness, automatic speech recognition

1 Introduction

In recent year we have seen increase in speech-based health sensing methods [1]. Often, the goal is to detect a condition, such as a respiratory infection [2], or a specific diagnostic condition such as OSA [3], Alzheimer's [4, 5] or Parkinson's disease [6]. One may also aim at estimating continuous values such as the age of a talker [7], or simultaneous physiological sensor signal from speech. For example, it would be very useful to be able to estimate the blood sugar level of a diabetic patient [8] or respiratory parameters of a talker [9] directly from the speech of a caller, for example, in a tele-health application. The current paper studies the problem of respiratory health sensing from speech using the VRB method. The Virtual Respiratory Belt, VRB, processing uses speech audio to model a signal captured using a chest-worn Respiratory Inductance Plethysmography, RIP, belt sensor. RIP measurements can be used to get an estimate of the respiratory parameters such as respiratory rate and tidal volume [10, 11]. McKenna *et al* demonstrated that it is also possible to predict, from RIP sensor data, the values of spirometry measurements during *speechlike* breathing [12].

Some of the early VRB work [9] was done using log-Mel spectrum data and RNN neural networks. A VRB task was also included in the 2020 Paralinguistic challenge [13]. Various CNN network architectures have been recently found useful in the parallel problem of covid-19 detection [14] and it was shown in a recent paper [anon] that the performance of the VRB modeling can be further improved by using pre-trained transformer networks such as Hubert [15] or Whisper [16]. In the current paper one such design is introduced, and we demonstrate its performance in the task of the detection of inhaling, or Inspiration Events, IEs, from read and spontaneous speech content. In addition, we introduce and demonstrate the performance of a novel modification, VRBOLA, of the conventional VRB method, which is using overlap-add processing in the waveform reconstruction.

[*] Currently with DACS, Maastricht University, The Netherlands. [1] Philips Research, Eindhoven, The Netherlands. [2] Philips Research, Bangalore, India

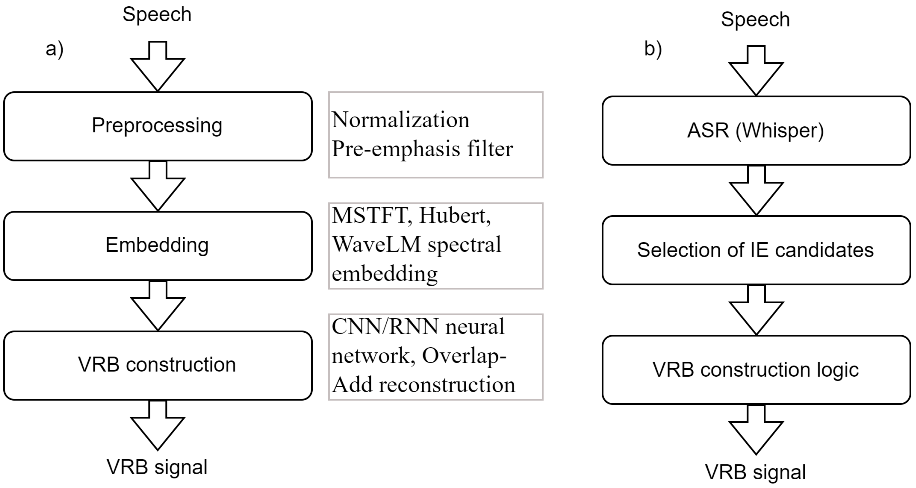


Figure 1: Block diagram for breathing waveform estimation from speech using a) VRB b) ASR and logic

Ingressive phonation, that is, speech vocalization during inspiration, is relatively rare [17], and the inspiration events in typical speech occur in pauses between words. Speech pauses can be detected using speech activity detection, SAD, algorithms. It is shown in this paper that by a selection of long pauses in speech using a good SAD can give a rough estimate of the occurrence of breathing events in speech.

There is a complex relationship between spoken language and our respiratory behavior influenced by breathing planning [18–21] which makes it challenging to estimate the VRB signal directly from speech, see discussion in [anon]. The breathing planning is expected to be guided by the rules of language so that IEs are placed where there is a natural pause in speech, for example, at end of a phrase or a sentence. In fact, some studies on read speech respiration have found that IEs in speech occur *almost exclusively at grammatically appropriate places* [18]. In conversational speech IEs occur also elsewhere, e.g., upto 13 % in [22]. One could propose to use the detection of grammatically appropriate places in speech as another method to detect IEs. In this paper, the idea is tested and compared against the VRB method.

This paper compares the performance of the proposed VRB algorithm to a method based on detection of speech pauses, and a method based on detection of *grammatically appropriate* places in speech, respectively. As expected, the VRB has the best performance in the detection of IEs in both read and spontaneous speech. The analysis of the results gives very interesting insights about speech breathing and design of algorithms for speech-based respiratory sensing. It also seems to show that, in the data used in this study, the assumption of *grammatical breathing* does not appear very strong but many IEs in both read and spontaneous speech are examples of *ungrammatical* breathing, i.e., inspiration events occurring in other parts of speech. Ungrammatical breathing occurs in both read speech and spontaneous speech but is more prevalent in spontaneous speech.

2 Breathing waveform estimation

The basic block diagram of a basic VRB algorithm is shown in Fig. 1a. The speech preprocessing block contains typical steps of signal level normalization by automatic gain control, pre-emphasis filtering to remove low frequency noises and to flatten the spectrum, and possibly an insertion of a background noise floor to make the models more robust for background noise conditions. In the experiments reported in [23] the signal was transformed to a Log-Mel-spectrogram representation. In [anon], it was demonstrated that the performance of both instantaneous breathing waveform modeling and forecasting breathing of events is significantly improved by using large pre-trained wave2vec transformer models, such as Hubert [15], in the preprocessing of the data before the neural network.

2.1 VRB wave modeling

The experiments of the current paper use a VRB model architecture that gave the highest Pearson correlation in another study [23] between the obtained VRB signal and the real belt signal. This model uses Hubert as the embedding model to generate vectors corresponding to 20ms processing windows at the sample rate of 16kHz. The vectors are stacked to a a matrix of 256 columns which are further processed using a 3-layer Gated Recurrent Unit, GRU, network with 64 hidden units, and finally one dense linear layer to produce K samples of the VRB signal corresponding to the speech signal.

Algorithm 1: Frame-based VRB modeling

```
1: Use pretrained Hubert model to embed the content in 30ms windows to $D = 1024$-tap vectors
2: Stack $K = 256$ vectors into a matrix $\mathbf{M}_{K \times D}$.
3: Train a cascade of a LSTM network and 1D CNN network to produce a $p$th frame of a VRB signal
   $b_p(k), k = 0, ...K - 1$.
```

The neural network was trained using the data from the dataset A introduced in Section 4. The model was optimized using the smooth L1 loss function, and the model optimization was performed by the Stochastic Gradient Descent, SGD, method, both implemented in the pytorch library [24], respectively.

2.2 Continuous wave reconstruction

In previous work the VRB method reconstructed the signal sample-by-sample, or in concatenated frames of N samples. It turns out that it is very useful to reconstruct the VRB signal using an overlap-add, OLA, method. In this paper, this method is called VRBOLA. The last step of Algorithm 1 produces a VRB signal fragment $b_p(k), k = 0, ...K - 1$ for a p^{th} block. In the VRBOLA method reconstruction is carried out by applying the overlap-add method using a sequence of window functions $w_p(t)$ so that the reconstruction of the final VRB signal is given by

$$b(t) = \sum_{p=-\infty}^{\infty} w_p(t) b_p(t - pS). \tag{1}$$

The window sequence is defined so that $w_p(t)$ is, e.g., a squared sine, in $t \in [pS,\ (p + 1)S - 1]$ and zero elsewhere. Overlap-add reconstruction reduces the effects of the errors that are usually larger close to the frame boundaries than in the center of the window.

In this paper, the focus is on the detection and characterization of inspiration events, IEs. In inspiration phase the belt sensor signal typically goes up, while during speech vocalization it goes down as lungs are depleted from air. The detection of one IE in the respiration signal is defined by a local maximum followed by local minimum in the waveform. In the following experiments, the same algorithm was used for the detection of IEs from the real belt waveform and the VRB signal estimated from speech.

Algorithm 2: Detection of IEs

```
1: Remove bias, sensor drift and high frequency noise by applying a third order Butterworth band-pass filter
   from 0.08 Hz to 1.0 Hz using forward-backward filtering.
2: Select local minima and maxima that have a minimal separation of 1s and that have normalized peak prominence
   that exceeds a threshold of 0.8.
3: The average breathing rate is taken from the mean of the distance between the maxima.
```

The average non-speech inspiration and expiration durations, for example, in Chronic Obstructive Pulmonary Disease, COPD, patients are 1s and 1.7s, respectively [25], which give a respiration rate of 0.2Hz. During speech, the average respiratory rate is approximately half of that [26] and the duration of an IE is may be less than 100ms. In this paper, the sample rate of the embedding data matrix $\mathbf{M}$ and the modeled belt signal $b(t)$ is 50Hz, corresponding to the 20ms input size of the Hubert model.

3 Respiratory sensing using ASR

In this paper, we use two databases of speech recordings with respiratory belt measurements. The dataset A has 500 subjects and it has been collected at a large medical center MAHE in India. The dataset B contains recordings of 40 talkers [27]collected in a research lab at Philips Research in Eindhoven, the Netherlands. The VRB model used in this paper has been trained using the dataset A, and all experiments reported in the paper have been performed on dataset B.

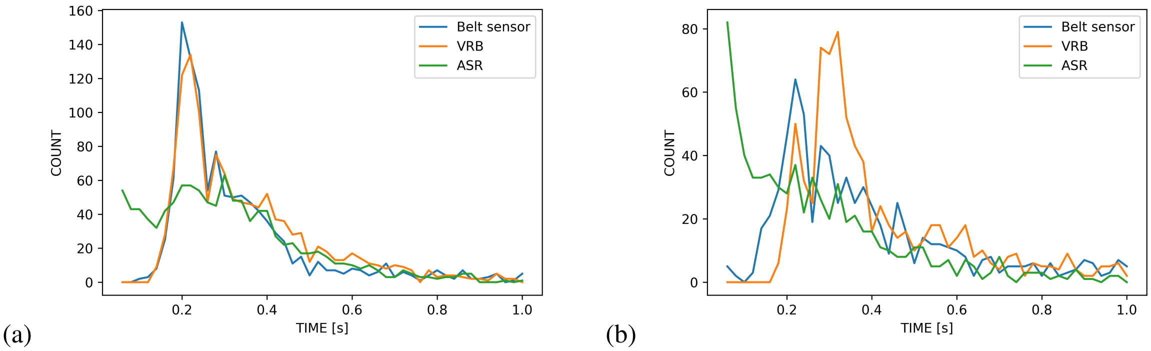


Figure 2: Histogram of inspiration event durations based on the belt sensor data, ASR-punct, and VRB, respectively in a) read b) spontaneous speech.

The database B contains forty recordings of read speech and spontaneous conversational speech. In early experiments we found that the WhisperX algorithm [28] based on the Whisper ASR [16] and word-level alignment works well for the data. The processing model for VRB estimation using ASR is shown in Fig. 1b. After the ASR and word-level time-alignment the IE candidates are selected by the punctuation points in the output of WhisperX. The durations of IEs computed from the belt sensor data and the timestamps between punctuation points and the start of the next word in ASR data, and pause durations from the VRB signal, respectively, are shown in Fig. 2. The pauses in VRB follow similar statistics to the real belt measurements, which can be expected because the VRB model has been trained using similar belt data. The most common duration of an IE is 200-250ms.

Let us define two ASR-based methods for the estimation of IEs from speech data.

3.1 ASR-word method

The underlying idea is to assume that IEs occur in long word breaks in speech. The method uses the whisperX to detect word boundaries and count all word-to-word pauses that are longer than 150ms as IE candidates. This method is based on assumption that long pauses in speech are IEs. The limit of 150ms was found optimal for the best performance in IE detection accuracy in the dataset.

3.2 ASR-punct method

This method follows the observation in [18] that IEs usually occur at grammatically appropriate places, *grammatical stops*. First all ends of clauses and sentences are detected from the output of the whisperX algorithm. In most cases, the end points are indicated by punctuation. Next, the end point of each IE is selected to be the start point of the next word in the sequence after the grammatical stop.

4 Experiments

The dataset B has speech from 40 healthy volunteers in a research institute in the central Europe either reading ”The Rainbow” passage [29] or 1-2 minute free speech in English about their current project or the last holiday trip. The subjects were wearing a Respiratory Inductance Plethysmograph, RIP, belt while speaking, which measures a continuous signal corresponding to the circumference of the chest and abdomen of the subject. The VRB and VRBOLA methods were trained using the dataset A with 200 recordings collected with a medical center in India telling about a traditional fable in their local language.

Fig 3a shows an example of a true sensor belt signal and an VRB estimate from the VRB model. The IE detections based are shown in Fig. 3b. Fig. Fig. 3c shows the long word pauses and grammatical stops based on the time-aligned ASR output.

Based on the IE segments identified based on VRB, word pauses, or grammatical stops, we can compute performance metrics of a method in comparison to the *ground truth IEs* computed from the real sensor belt signal. True Positive detection is counted if an IE estimate temporally overlaps with the true IE and True Negative is counted when there are no IE estimates in regions between ground truth IEs. False Positives and Negatives are counts of IEs between ground truth IEs, and counts of missed ground truth IEs, respectively.

The results of read ”Rainbow” passage are shown in Table 1. The overall performance, by the F1 score, is the highest, 0.85, for the VRB detections followed by ASR-word and ASR-punct. The performance of ASR-punct is significantly reduced by a high number of false negatives (fn). This is an interesting observation because it is in opposition with the observation of [18] that IEs occur, almost exclusively, in grammatically appropriate points. By the relatively high specificity, 0.72, it indeed seems to be the case that grammatical stops are often points of IEs. But, the high number of false negatives is a clear indication of ungrammatical breathing, that is, a large number of IEs in this read speech dataset occur in other parts of read content than close to grammatical stops.

The ASR-word method based on long word pauses has a large number of both false positives and false negatives but overall performance with F1 score 0.74 is better than the ASR-punct method. Pauses in speech an IEs are indeed aligned, but not all long pauses are IEs.

	ASR-word	ASR-punct	VRB	VRBOLA
tp	960	752	885	929
tn	985	762	866	915
fp	343	319	101	93
fn	347	568	290	232
Sensitivity	0.73	0.57	0.75	0.80
Specificity	0.74	0.70	0.90	0.91
F1-score	0.74	0.63	0.82	0.85

Table 1: Performance of IE detection in read speech

The performance of the three methods, see Table 2, have a similar trend also in the spontaneous speech dataset with the VRB method giving the highest overall performance. However, the performance of the ASR-word method is here significantly closer to VRB. One possible reason may be in the recording setup where the subject was instructed to talk much and the interviewer was just triggering more talk with brief questions. Many subjects were speaking in a very relaxed way and had relatively longer IEs than in read speech. In the ASR-punct one can see again increase in false negatives, that is, ungrammatical breathing but also an increased number of false positives.

	ASR-word	ASR-punct	VRB	VRBOLA
tp	513	370	512	532
tn	517	389	488	507
fp	269	312	136	135
fn	207	266	168	175
Sensitivity	0.71	0.58	0.75	0.75
Specificity.	0.66	0.55	0.78	0.79
F1-score	0.68	0.56	0.77	0.78

Table 2: Performance of IE detection in spontaneous speech

The comparison of the conventional concatenated VRB processing, and the new VRBOLA method proposed in this paper shows a clear advantage for the proposed overlap-add processing. The VRBOLA method increases the computational requirements but it seems to increase the overall performance probably due to elimination of transients in frame borders.

5 Discussion

Ingressive speech is rare in most languages and speech is largely phonetically controlled expiration. Consequently, inspiration takes place, almost exclusively, in pauses between words, or in nonverbal vocalization during gasping or yawning. The results of the experiment reported in this paper shows that detection of long pauses in speech can give a rough detection of inspiration events, IEs, in speech. However, not all pauses are used for inhaling and talkers are able to ventilate in very brief, less than 100ms, pauses between words.

It is often suggested that IEs take mostly place in grammatically appropriate places in speech. In a study by Winksworth *et al.* [18] breathing in read speech is almost completely aligned with syntactic stops. The results of the current paper do not support this observation but nearly half of the IEs seem to take place in other parts of speech. In this paper we use the term ungrammatical breathing for this phenomenon and it was shown that it is common in both read and spontaneous speech. Obviously, not all paragraph, sentence, clause, and phrase boundaries are IEs. In addition, ambiguities in ASR and parsing, and ungrammatical speech may also lead to well-known errors, and spontaneous speech may have ungrammatical passages of words.

Compared to the word pauses and grammatical stops, the best performance in the detection of IEs was with the Virtual Respiratory Belt, VRB, algorithm trained to model respiratory behavior of a talker from speech. The F1 score in the correct detection of IEs was here over 0.85 while in the methods based on word pauses or grammatical stops F1 score was below 0.75.

The spontaneous speech in the current paper consists of speech fragments from a conversation but the actual turn-taking was not included in the analysis. Breathing in conversation is also influenced by the turn-taking behavior, see, e.g., [30], and in future work it would be interesting to explore the use of the talk of the other party so support IE detection.

6 Acknowledgements

The authors wish to thank anonymous reviewers for their useful comments and [anon] for the financial support of this work.

References

[1] N. Cummins, A. Baird, and B. W. Schuller, “Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning,” *Methods*, vol. 151, pp. 41–54, Dec. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1046202317303717

[2] M. Albes, Z. Ren, B. W. Schuller, and N. Cummins, “Squeeze for Sneeze: Compact Neural Networks for Cold and Flu Recognition,” in *Interspeech 2020*. ISCA, Oct. 2020, pp. 4546–4550. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2020/abstracts/2531.html

[3] E. Goldshtein, A. Tarasiuk, and Y. Zigel, “Automatic Detection of Obstructive Sleep Apnea Using Speech Signals,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1373–1382, May 2011.

[4] F. Haider, S. De La Fuente, and S. Luz, “An assessment of paralinguistic acoustic features for detection of Alzheimer’s dementia in spontaneous speech,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[5] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s Dementia Recognition through Spontaneous Speech: The ADReSS Challenge,” in *Proc. Interspeech*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[6] L. Moro-Velazquez, J. Villalba, and N. Dehak, “Using x-vectors to automatically detect Parkinson’s disease from speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 1155–1159.

[7] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, “End-to-end deep neural network age estimation.” in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 277–281.

[8] J. Sidorova, P. Carbonell, and M. Čukić, “Blood Glucose Estimation From Voice: First Review of Successes and Challenges,” *Journal of Voice*, vol. 36, no. 5, pp. 737.e1–737.e10, Sep. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0892199720303349

[9] V. S. Nallanthighal, A. Härmä, and H. Strik, “Deep Sensing of Breathing Signal During Conversational Speech,” in *Interspeech 2019*. ISCA, Sep. 2019, pp. 4110–4114. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/1796.html

[10] C. Heyde, H. Leutheuser, B. Eskofier, K. Roecker, and A. Gollhofer, “Respiratory Inductance Plethysmography—A Rationale for Validity during Exercise,” *Medicine & Science in Sports & Exercise*, vol. 46, no. 3, pp. 488–495, Mar. 2014. [Online]. Available: https://journals.lww.com/00005768-201403000-00008

[11] B. Laufer, S. Krueger-Ziolek, P. D. Docherty, F. Hoeflinger, L. Reindl, and K. Moeller, “Tidal volume via circumferences of the upper body: a pilot study,” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2019, pp. 3559–3562, Jul. 2019.

[12] V. S. McKenna and J. E. Huber, "The Accuracy of Respiratory Calibration Methods for Estimating Lung Volume During Speech Breathing: A Comparison of Four Methods Across Three Adult Cohorts," *Journal of speech, language, and hearing research: JSLHR*, vol. 62, no. 8, pp. 2632–2644, Aug. 2019.

[13] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2042–2046. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/schuller20_interspeech.html

[14] M. Effati and G. Nejat, "A Performance Study of CNN Architectures for the Autonomous Detection of COVID-19 Symptoms Using Cough and Breathing," *Computers*, vol. 12, no. 2, p. 44, Feb. 2023. [Online]. Available: https://www.mdpi.com/2073-431X/12/2/44

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," Jun. 2021, arXiv:2106.07447 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2106.07447

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[17] R. Eklund, "Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech," *Journal of the International Phonetic Association*, vol. 38, no. 3, pp. 235–324, Dec. 2008, publisher: Cambridge University Press.

[18] 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.

[19] M. Włodarczak and M. Heldner, "Respiratory Constraints in Verbal and Non-verbal Communication," *Frontiers in Psychology*, vol. 8, May 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5434352/

[20] O. Egorow, T. Mrech, N. Weißkirchen, and A. Wendemuth, "Employing Bottleneck and Convolutional Features for Speech-Based Physical Load Detection on Limited Data Amounts," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 1666–1670. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2502.html

[21] S. Fuchs and A. Rochet-Capellan, "The respiratory foundations of spoken language," *Annual Review of Linguistics*, vol. 7, no. 1, pp. 13–30, 2021.

[22] Y.-T. Wang, J. R. Green, I. S. B. Nip, R. D. Kent, and J. F. Kent, "Breath group analysis for reading and spontaneous speech in healthy adults," *Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP)*, vol. 62, no. 6, pp. 297–302, 2010.

[23] V. S. Nallanthighal, Z. Mostaani, A. Härmä, H. Strik, and M. Magimai-Doss, "Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings," *Neural Networks*, vol. 141, pp. 211–224, Sep. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0893608021001179

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[25] W. Chatila, T. Nugent, G. Vance, J. Gaughan, and G. J. Criner, "The Effects of High-Flow vs Low-Flow Oxygen on Exercise in Advanced Obstructive Airways Disease," *Chest*, vol. 126, no. 4, pp. 1108–1115, Oct. 2004.

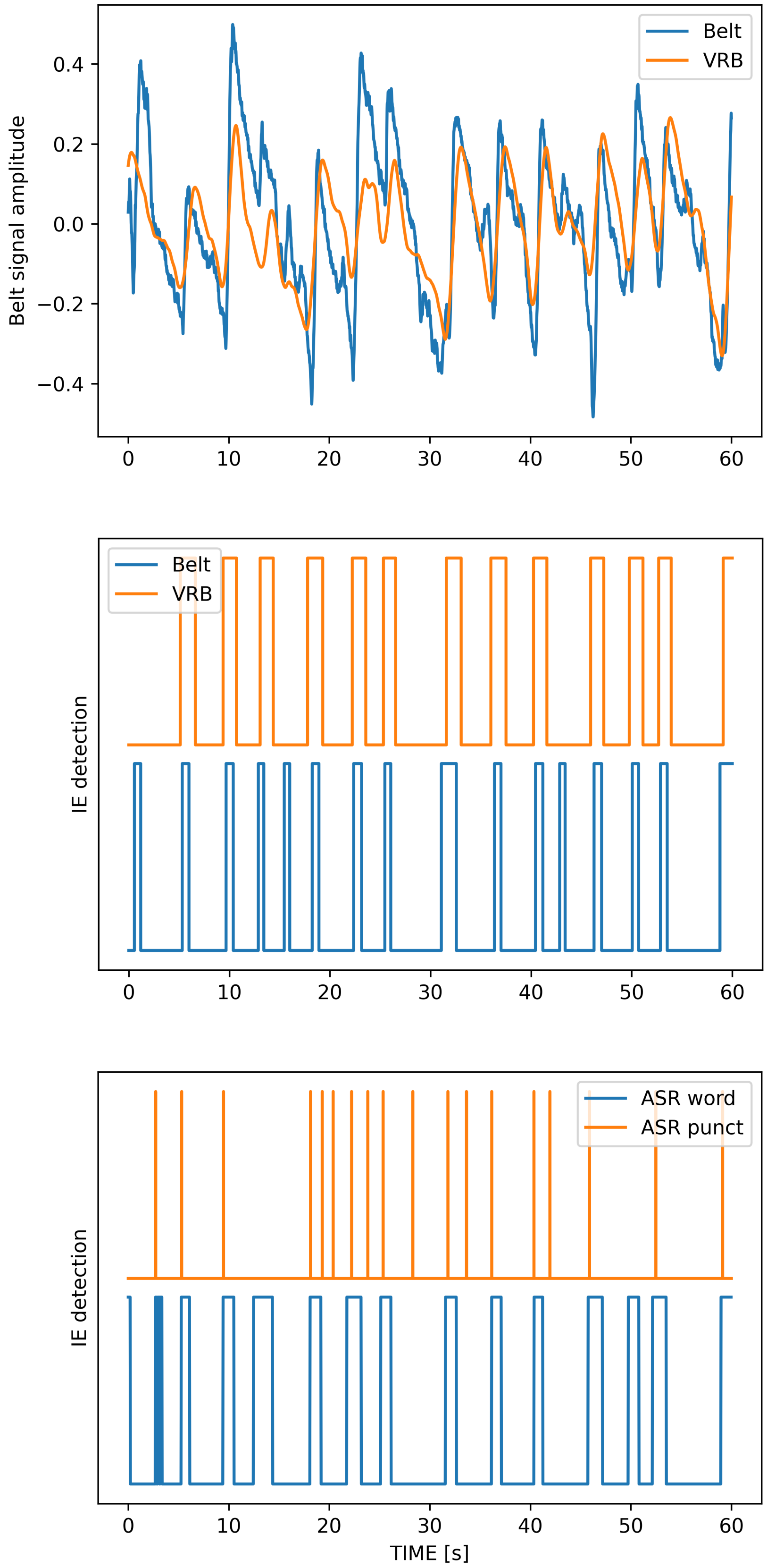
[26] M. Włodarczak, M. Heldner, and J. Edlund, "Breathing in conversation: An unwritten history," in *2nd European and the 5th Nordic Symposium on Multimodal Communication, Tartu, Estonia, August 6-8, 2014*, 2015, pp. 107–112.

[27] V. S. Nallanthighal, A. Härmä, and H. Strik, "Deep Sensing of Breathing Signal During Conversational Speech," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 4110–4114. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/1796.html

[28] M. Bain and T. Han, "WhisperX," 2022, publication Title: GitHub repository. [Online]. Available: https://github.com/m-bain/whisperX

[29] G. Fairbanks, "The rainbow passage," *Voice and articulation drillbook*, vol. 2, 1960.

[30] M. Wlodarczak and M. Heldner, "Breathing in Conversation," *Frontiers in Psychology*, vol. 11, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.575566



(a)

(b)

(c)

Figure 3: Examples of a) sensor belt signal and estimated VRB, b) detected IEs in the belt and VRB signals, and c) word pauses and punctuation points at the output of ASR.