AN AUTOMATIC METHOD FOR SPEECH BREATHING ANNOTATION

Alexis Deighton MacIntyre[1]*, Raphael Werner*[2]

[1]*University of Cambridge,* [2]*Saarland University*
alexisdeighton.macintyre@mrc-cbu.cam.ac.uk, rwerner@lst.uni-saarland.de

Abstract: Breathing is central to speech planning and production; however, speech breathing is difficult to monitor and quantify without laborious and subjective manual annotation. Here, we describe a method for automatically detecting the beginning and end time points of speech-associated inhalations measured with inductive plethysmography, or breath belts. Unlike simpler approaches to breath detection, the technique introduced here employs slope analysis to improve temporal precision. First, inhalation events are identified by searching for roughly continuous, positive sloping segments. Inhalations are then rejected or modified based on slope height, duration, and grade, as well as contextual factors, such as the height or duration of neighbouring breaths. Finally, the respiratory time series can be optionally corroborated with acoustic recordings to further improve results. This approach is validated by two independent annotators using spontaneous and read English speech contributed by 10 individual speakers, including relatively noisy data. From a signal detection perspective, we estimate performance at 95% on average. The mean median error of detected breaths, when compared to human annotation, is 22.50 ms (IQR 37.71 ms). By comparison, a peak-finding method without acoustic calibration yields 91% accuracy with substantially larger errors (mean median 167.90 ms, IQR 381.45 ms). In conclusion, the proposed automatic method provides robust and temporally accurate annotation of the speech breathing time series.

1 Introduction

Fine respiratory control is foundational to speech production, and has been investigated across fields ranging from development [1], to disorder and disease such as Parkinson's [2]. Efforts to scientifically describe speech breathing date as early as the 1930s, with [3], for example, visually comparing the respiratory patterns of an adult male with a stammer to those of a neurotypical control. A common means to monitor speech breathing is to record torso displacement via respiratory inductance plethysmography [4]. This technique produces a linear signal with characteristic peaks corresponding to individual inhalations (Figure 1). At rest, the breathing signal resembles a quasi-sinusoidal profile, the phase of which can be reasonably estimated using basic signal processing techniques. In the context of speech, however, its quantification presents a special challenge: The signal is noisy, irregular, and prone to inter-individual idiosyncrasy. Hence, the automatic detection of speech-related inhalation requires a more specialised approach, especially for research areas where accurate timing is important, including prosody or conversational turn-taking. Here, we describe and evaluate a method to automatically annotate the speech breathing signal with the goal of maximising temporal precision and robustness to noise.

1.1 Automatic Approaches to Speech Breathing Detection

There are many approaches to the annotation of speech breathing in the literature, with most of them demanding considerable human judgement and labour. Some authors detect inhalation loci based on listening [5], and manual annotation of plethysmography is also possible [6]. Otherwise, there have been attempts to automatically identify the speech breathing time series: One proposal is to take the zero-crossings of the acceleration of the respiratory signal within an appropriate bandwidth (e.g., 0.05 Hz − 10 Hz) [7]. Another is to find the locations of values equivalent to 10% of the value of the velocity peak before and after the peak [8, 9]. Such procedures are appealingly simple, but the researcher is forced to either proceed with the understanding that many observations will likely be artefacts; or, invest time in manually inspecting and adjusting the breathing time series where needed. [10], for example, took a peak-finding algorithmic approach, but report that "about 100 [exhalations] have positive slopes [...] most likely caused by an error in the automatic segmentation of respiratory cycles which skipped over an inhalation". In sum, these automated techniques depend on levels of simplicity and stability typically missing from natural speech breathing data.

1.2 The Proposed Method

To save time, improve accuracy, and enhance reproducability, we introduce a new set of speech breathing-specific functions, the `SpeechBreathingToolbox`[1], developed in MATLAB [11]. The development of this toolbox was based on extensive visual inspection of breath belt data corroborated using the acoustic speech spectrogram. Essentially, the primary mechanism of our proposed method, and what differs from the previously described approaches, is its detailed slope analysis, making the `SpeechBreathingToolbox` robust to noise and adapted to the unusual profile of speech-related respiration. Moreover, given that inhalation duration, volume, and kinematic profiles vary between speakers [12], the algorithm flexibly estimates, rather than hard codes, many of its parameters.

Calibration with the acoustic signal. We have seen that some speakers' plethysmographic signals do not conform to typical breathing patterns. One instance is with speakers whose apparent inhalation ends (i.e., signal peaks) seem to occur *after* speech has started (Figure 1; see discussion in [13]). Another issue arises with speakers whose chest movements during speech exhalation spuriously resemble the shape of an inhalation. Hence, we find algorithmic results are further improved by cross-referencing between the respiratory signal and corresponding acoustic speech data. Because speech breathing inconsistently registers as an acoustic trace, and is moreover often associated with ingressive noises, like pops and croaks, a straight-forward silence detection approach is not feasible. We therefore implement a method of multi-threshold silence detection; specifically, brief but loud sounds (e.g., a throat click), are allowed, as well as longer but low-intensity broadband sounds (e.g., the sound of the breath itself).

Evaluation. We evaluate `SpeechBreathingToolbox` by comparing its results with human manual annotation, as well as automatic annotations produced by the peak-finding algorithm used in `RespInPeace`, a speech breathing analysis tool developed for Python [14]. In the paper introducing `RespInPeace`, the authors report that 12.6% of breath events (from a corpus of 9 x 20-minute, three-party conversations) required manual adjustment [14]. Error rates in excess of 10% may be unacceptable when correcting a larger data set, however. Moreover, the temporal precision of this method is unclear. Thus, as it is intended to improve upon signal-crossing techniques for breath detection (e.g., instants of peak velocity), we use `RespInPeace` as a helpful

[1] All functions are packaged and distributed as the `SpeechBreathingToolbox` with code available for download from `https://github.com/alexisdmacintyre/SpeechBreathingToolbox`.

benchmark by which to compare the current method. In the following sections, we describe the corpus used to evaluate the proposed method, followed by the algorithmic steps in technical detail.

2 Methods and Materials

2.1 Evaluation Data Set

The `SpeechBreathingToolbox` is validated using a corpus produced by 10 individual English speakers (3 male and 7 female, ages 25-50) whose data were not used during the algorithmic development. The speakers were fitted with two breath belts (MLT1132, ADInstruments, Castle Hill, Australia) with one positioned at the abdomen and the other at the chest level. A cardioid dynamic microphone was placed on a stand in front of the speaker's mouth. The respiratory and acoustic speech signals were sampled together by the same acquisition device at 20 kHz. For the current analysis, we used 2 minutes of spontaneous speech and 1 minute 15 seconds of read speech contributed by each speaker. The spontaneous speech was elicited using text prompts with familiar, open-ended questions (e.g., "What is your favourite restaurant and why?"), and the read speech consisted of simplified popular articles that were edited for readability.

2.2 Data Preprocessing

Individual speakers exhibit wide variability in terms of chest and abdominal breathing patterns [8]. Where both the upper and lower breath belts were found to be in good alignment, the mean of the two belts was taken. Otherwise, only the single breath belt that most closely corresponded to the acoustic spectrogram was used. The respiratory signal was downsampled to 1 kHz. For smoothing and the removal of high frequency noise, a moving mean with a window of 20 milliseconds was found satisfactory. Other authors report band-pass or low-pass filtering the signal [9, 14]; however, when inspecting the filtered signal, it was determined that these more transformative methods distorted the signal, as is reported elsewhere [15]. Similarly, to avoid any unnecessary distortion of the respiratory signal, no baseline drift is subtracted, given this was not found to be problematic for the proposed method. Finally, the respiratory signal is re-scaled between $[0, 1]$.

2.3 Algorithms

2.3.1 Inhalation Onset and End Detection

1. The algorithm first determines all instances of continuous, positive-going slopes. This is performed by taking the moving average of the first derivative of the signal from non-overlapping windows. Transient non-positive segments associated with noise are tolerated at this early stage.

2. Preliminary inhalation ends are identified as the maximum value within each distinct positive slope. The corresponding inhalation onset is chosen as the latest value $\leq$ the 2.5^{th} percentile of the total peak height, rather than the absolute minimum, which was found to be less robust.

3. The algorithm then iterates over each candidate inhalation to determine whether the onset and end of candidate inhalations should be modified. For example, the windowed cumulative percentage of total breath height is used to trim weakly graded segments near either extreme of the breath (e.g., if the current window reaches less than 6% of the total

peak height, the onset is moved up to the end of the current window). Another parameter is the presence of sudden change-points in the signal, like a notch, near the inhalation end. As ongoing adjustments change local and global statistics–for example, trimming an inhalation at a signal change point will affect its cumulative height–the script updates these measures with each iteration.

4. If any of the resultant inhalations occur within overly close proximity (e.g., < 300 ms by default), they may either be joined together or rejected according to contextual factors, such as their relative height or prominence.

5. Finally, the remaining inhalations are filtered according to global thresholds, such as minimum inhalation volume or inter-breath interval. These parameters can be estimated statistically (e.g., $\leq$ IQR below the first quartile of breath slope grades) or set as absolute values where desired.

2.3.2 Calibration with Acoustic Speech Recording

1. Values in the acoustic speech signal exceeding the upper and lower 10% are capped and a band pass filter with cutoff $2 - 1000$ Hz is applied to remove signal drift and reduce the acoustic presence of breathing, a broadband noise.

2. The slow amplitude modulation of the speech signal, the speech envelope, is extracted using a method described in [16, 17] and sampled at 1 kHz. The envelope is smoothed using a moving mean. Values exceeding the 97.5^{th} percentile are capped, and the envelope is re-scaled between $[0, 1]$.

3. Soft and hard thresholds are calculated for silent detection as T_{soft} and T_{hard}, respectively. Heuristically, we have found good results by dividing the envelope values into 50 quantiles and setting T_{soft} as the 24^{th} quantile (e.g., breath sounds), and T_{hard} to the 27^{th} quantile (e.g., ingressive noise and speech sounds). These thresholds are used to further simplify the envelope, where values $\leq T_{\text{soft}}$ are replaced with 0; values $\geq T_{\text{soft}}$ and $< T_{\text{hard}}$ are replaced with T_{soft}; and values $> T_{\text{hard}}$ are replaced with T_{hard}.

4. For each detected breath event, extract the corresponding thresholded speech envelope. A decision logic is implemented to accept or reject that breath:

   (a) If $> 50\%$ of the inhalation coincides with envelope values exceeding T_{hard}, and these suprathreshold values overlap with the greatest change in the respiratory signal, discard the breath as probable exhalation.

   (b) If non-zero speech envelope segments consist of a mix of T_{soft} and T_{hard} values, reject if the mid-section of this segment is mostly T_{hard}. Otherwise, very short (e.g., < 45 ms by default) T_{hard} segments associated with clicks or pops are permitted, as well as continuous T_{soft} segments where they are not interspersed with longer T_{hard} values.

   (c) If $> 75\%$ of the original total height of the inhalation slope or $> 66\%$ of total duration has been lost after the preceding adjustments, discard that breath event.

5. Otherwise, the remaining inhalation onsets and ends are adjusted when found to overlap with T_{hard} values in their corresponding speech envelope sections. If the new breath signal begins later than the original inhalation onset, the new onset will be moved up to the new local minimum. If the new breath signal ends sooner than the current inhalation end, the new end will be moved to the new local max.

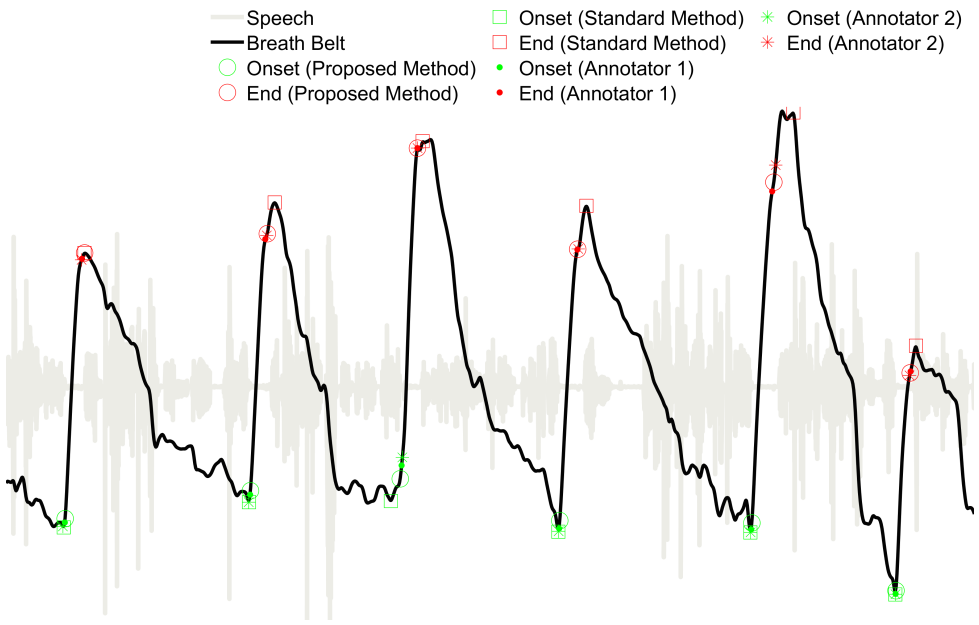

Figure 1 – The breath belt signal (black), acoustic speech signal (grey), and inhalation onsets and ends (green and red, respectively). The shape of the markers indicates their source.

3 Evaluation

To validate the `SpeechBreathingToolbox` functions described above ("proposed method"), the output was compared with manual annotations made independently by the authors in Praat [18] using the respiratory and acoustic speech signal, thus similar to the script. As a benchmark comparison with an automated technique from the literature, we report results from the peak-finding algorithm implemented by `RespInPeace` ("standard method") [14]. We generated these annotations in MATLAB following the procedure defined in [14] as closely as possible.

3.1 Measures

We evaluate performance using the signal detection metrics precision and recall, which characterise the proportion of false positives (i.e., returned inhalations that were absent in ground truth) and false negatives (i.e., inhalations present in ground truth that were not returned) relative to true positives, respectively. Precision and recall can be combined by taking their harmonic mean, resulting in the F1 score. We report mean precision, recall, and F1 across trials.

To empirically determine the temporal precision of each method, we paired each returned annotation across the three sources (human, proposed method, standard method) and calculated the Euclidean distance between paired time points (e.g., $Onset_{\text{human}}$ to $Onset_{\text{proposed method}}$) in milliseconds. Descriptive statistics are first calculated on a trial-by-trial basis, and then aggregated by taking the mean across trials.

3.2 Results

Signal Detection. The proposed method `SpeechBreathingToolbox` returned favourable results, with F1 of 0.95, meaning that the automatic technique detected the presence or absence of breath events similarly to human annotators. Moreover, precision and recall did not differ across speaking conditions (Table 1). The standard method fared, by comparison, poorly with spontaneous speech (F1 0.87) in comparison to reading (F1 0.92). For reference, the overall inter-human annotator F1 was 0.98, with little difference between reading or spontaneous speech.

Table 1 – Comparison of signal detection metrics between automatic methods.

	Proposed Method			Standard Method		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Spontaneous	0.96	0.95	0.96	0.86	0.9	0.87
Reading	0.94	0.97	0.95	0.96	0.89	0.92
Overall	0.95	0.96	0.95	0.93	0.89	0.91

Temporal precision. We turn now to error, which describes Euclidean distance, in milliseconds, between automatic results and human annotations (Table 2, Fig. 2). Overall, we find that the proposed method returns median errors of 39.33 ms (IQR 41.32 ms) for inhalation onsets and 36.08 ms (IQR 39.13 ms) for ends. By comparison, the median inter-human annotator error is 25.62 ms (IQR 29.03) for inhalation onsets and 22.73 ms (IQR 23.62 ms) for ends. This suggests a loss $\simeq$ 20 ms in temporal accuracy when automatic methods are employed; however, the median error for the standard method is 266.21 ms (IQR 675.80 ms) for inhalation onsets and 69.39 ms (IQR 87.09 ms) for ends. Hence, the proposed method yields a substantial improvement on the scale of tens of milliseconds, permitting a reasonable trade-off between human labour and temporal precision. Incidentally, we note that inhalation onsets are associated with larger errors than ends, and this applies to inter-annotator error as well as automatic method-human error. In particular, the standard method median error for inhalation onsets in spontaneous speech is 274.78 ms (IQR 758.85 ms).

Table 2 – Comparison of error (milliseconds) between proposed and standard methods with human-produced manual annotations. Measures are aggregated by taking the mean across trials.

	Proposed Method Error (ms)					
	Spontaneous					
	Median	IQR	Mean	SD	Min.	Max.
Breath Onset	25.43	37.06	37.10	36.90	4.80	113.13
Breath End	20.53	45.88	41.24	54.63	4.40	162.43
	Reading					
	Median	IQR	Mean	SD	Min.	Max.
Breath Onset	27.68	42.74	47.64	57.43	5.95	183.4
Breath End	15.83	21.36	35.97	56.87	4.70	184.25
	Standard Method Error (ms)					
	Spontaneous					
	Median	IQR	Mean	SD	Min.	Max.
Breath Onset	274.78	758.85	535.30	690.43	24.73	1933.37
Breath End	80.98	88.45	113.12	120.61	21.73	393.2
	Reading					
	Median	IQR	Mean	SD	Min.	Max.
Breath Onset	253.35	551.21	531.22	706.63	56.00	2127.4
Breath End	52.00	85.05	87.26	95.38	20.15	294.75

4 Discussion and Conclusion

To annotate the speech breathing time series, an effective and trustworthy automatic method should balance signal detection with good temporal precision. Based on our evaluation, we find that the proposed method, `SpeechBreathingToolbox`, offers good improvement over the

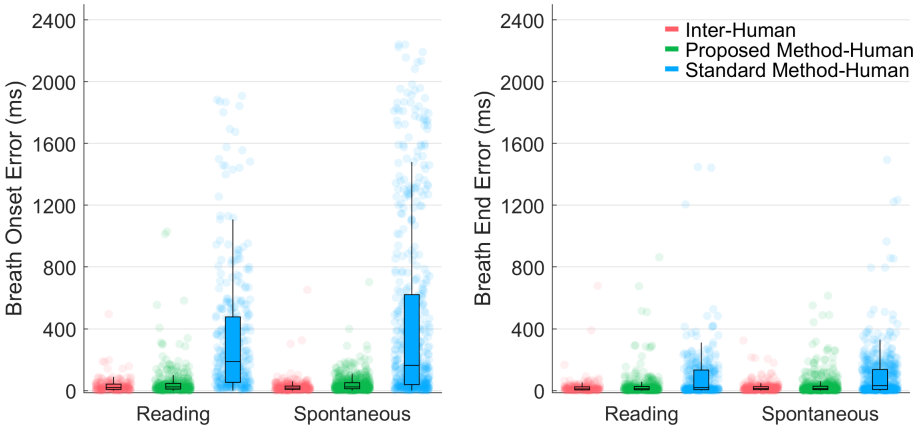

Figure 2 – Comparison of the distributions and Tukey boxplots of individual annotation errors for inhalation onsets (left panel) and ends (right panel).

standard method on both fronts. Importantly, we found no evidence for a qualitative difference on the basis of speaking style, meaning that even less controlled, spontaneous speech can be used without concern for additional noise. Although we performed no corrective procedures on the data, the performance could be further improved using minimal post-processing steps (e.g., removing obvious outliers).

Unlike `RespInPeace`, `SpeechBreathingToolbox` currently does not address breath holds or pauses [14], wherein an individual halts speech but does not inhale–this functionality should be implemented in future versions of the toolbox. Similarly, we used breath belt data that were not calibrated using measures of lung volume or breathing range, and therefore do not address spatial aspects of breathing movements in detail. Researchers interested in respiratory capacity may wish to calibrate the breath belt signal with equipment such as a mask-type spirometer.

In conclusion, our objective was to develop a fast, effective, and easy to use package of scripts, enabling researchers to sift through great amounts of speech breathing data. In comparing the output of `SpeechBreathingToolbox` to human annotation, we find high agreement concerning how many and when, precisely, inhalation events happen during speech production. Overall, the proposed method seems to produce reliable, objective annotations that are comparable to those made by human annotators, but with a fraction of the time and labour.

References

[1] HOIT, J. D., T. J. HIXON, P. J. WATSON, and W. J. MORGAN: *Speech breathing in children and adolescents*. *Journal of Speech and Hearing Research*, 33(1), pp. 51–69, 1990. doi:10.1044/jshr.3301.51.

[2] SOLOMON, N. P. and T. J. HIXON: *Speech breathing in Parkinson's disease*. *Journal of Speech and Hearing Research*, 36(2), pp. 294–310, 1993. doi:10.1044/jshr.3602.294.

[3] SETH, G.: *An experimental study of the control of the mechanism of speech, and in particular of that of respiration, in stuttering subjects*. *British Journal of Psychology*, 24(4), pp. 375–388, 1934.

[4] FUCHS, S. and A. ROCHET-CAPELLAN: *The Respiratory Foundations of Spoken Language*. *Annual Review of Linguistics*, 7(1), pp. 13–30, 2021. doi:10.1146/annurev-linguistics-031720-103907.

[5] BAILLY, G. and C. GOUVERNAYRE: *Pauses and respiratory markers of the structure of book reading*. *13th Annual Conference of the International Speech*

Communication Association 2012, INTERSPEECH 2012, 3, pp. 2215–2218, 2012. doi:10.21437/interspeech.2012-591.

[6] BARBOSA, P. A., S. MADUREIRA, M. A. S. FONTES, and P. MENEGON: *Speech Breathing and Expressivity: An Experimental Study in Reading and Singing Styles*. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12037 LNAI, pp. 393–398. 2020. doi:10.1007/978-3-030-41505-1_37.

[7] BAILLY, G., A. ROCHET-CAPELLAN, and C. VILAIN: *Adaptation of respiratory patterns in collaborative reading*. In *Interspeech 2013*, pp. 1653–1657. ISCA, ISCA, 2013. doi:10.21437/Interspeech.2013-428.

[8] ROCHET-CAPELLAN, A. and S. FUCHS: *Changes in breathing while listening to read speech: The effect of reader and speech mode*. *Frontiers in Psychology*, 4, p. 906, 2013. doi:10.3389/fpsyg.2013.00906.

[9] ROCHET-CAPELLAN, A. and S. FUCHS: *Take a breath and take the turn: How breathing meets turns in spontaneous dialogue*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 2014. doi:10.1098/rstb.2013.0399.

[10] WŁODARCZAK, M. and M. HELDNER: *Exhalatory turn-taking cues*. In *Speech Prosody 2018*, no. June, pp. 334–338. ISCA, ISCA, 2018. doi:10.21437/SpeechProsody.2018-68.

[11] MATLAB: *version 9.0 (R2016a)*. The MathWorks Inc., Natick, Massachusetts, 2016.

[12] SERRÉ, H., M. DOHEN, S. FUCHS, S. GERBER, and A. ROCHET-CAPELLAN: *Speech breathing: variable but individual over time and according to limb movements*. *Annals of the New York Academy of Sciences*, 1505(1), pp. 142–155, 2021.

[13] WHALEN, D. H. and J. M. KINSELLA-SHAW: *Exploring the Relationship of Inspiration Duration to Utterance Duration*. *Phonetica*, 54(3-4), pp. 138–152, 1997. doi:10.1159/000262218.

[14] WŁODARCZAK, M.: *RespInPeace : Toolkit for Processing Respiratory Belt Data*. In *Proceedings from Fonetik 2019*, pp. 115–118. Stockholm, 2019. doi:10.5281/zenodo.3246019.

[15] NOTO, T., G. ZHOU, S. SCHUELE, J. TEMPLER, and C. ZELANO: *Automated analysis of breathing waveforms using BreathMetrics: A respiratory signal processing toolbox*. *Chemical Senses*, 43(8), pp. 583–597, 2018. doi:10.1093/chemse/bjy045.

[16] OGANIAN, Y. and E. F. CHANG: *A speech envelope landmark for syllable encoding in human superior temporal gyrus*. *Science advances*, 5(11), p. eaay6279, 2019.

[17] MACINTYRE, A. D., C. Q. CAI, and S. K. SCOTT: *Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques*. *The Journal of the Acoustical Society of America*, 151(3), pp. 2002–2026, 2022.

[18] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer*. 2019. URL `http://www.praat.org/`.