Automated Quality Control of Forced Oscillation Measurements: Respiratory Artifact Detection with Advanced Feature Extraction☆

Thuy T. Pham[a,b], Philip H.W. Leong[a], Paul D. Robinson[c,d], Thomas Gutzler[c], Adelle Jee[c], Gregory G. King[c], Cindy Thamrin[c]

[a]*Engineering & IT, University of Technology Sydney, Australia*

[b]*Electrical & Information Engineering, University of Sydney, Australia*

[c]*Woolcock Institute of Medical Research, University of Sydney, Australia*

[d]*The Children's Hospital at Westmead,NSW 2145, Australia.*

☆**Address for correspondence:**
Dr. Cindy Thamrin,
Airway Physiology and Imaging Group
Woolcock Institute of Medical Research
PO Box M77,
Missenden Rd, NSW 2050, Australia.
Phone: +61 2 9114 0440
Fax: +61 2 9114 0101
Email: cindy.thamrin@sydney.edu.au

Running head: Artifact detection in forced oscillation measurements.
Word count abstract: 235 words; main text: 4809 words (included appendix)

Abstract

The forced oscillation technique (FOT) can provide unique and clinically relevant lung function information with little cooperation with subjects. However, FOT has higher variability than spirometry, possibly because strategies for quality control and reducing artifacts in FOT measurements have yet to be standardized or validated. Many quality control procedures either rely on simple statistical filters or subjective evaluation by a human operator. In this study, we propose an automated artifact removal approach based on the resistance against flow profile, applied to complete breaths. We report results obtained from data recorded from children and adults with and without asthma. Our proposed method has 76% agreement with a human operator for the adult dataset and 79% for the pediatric dataset. Furthermore, we assessed the variability of respiratory resistance measured by FOT using within-session variation (wCV), between-session variation (bCV). In the asthmatic adults test dataset, our method was again similar to that of the manual operator for wCV (6.5 vs. 6.9%), and significantly improved bCV (8.2 vs. 8.9%). Our combined automated breath removal approach based on advanced feature extraction offers better or equivalent quality control of FOT measurements compared to an expert operator and computationally more intensive methods in terms of accuracy and reducing intra-subject variability.

New & Noteworthy

The forced oscillation technique (FOT) is gaining wider acceptance for clinical testing, however strategies for quality control of are still highly variable and require a high level of subjectivity. We propose an automated, complete breath approach for removal of respiratory artifacts from FOT measurements, using feature extraction and an interquartile range filter. Our approach offers better or equivalent performance compared to an expert operator, in terms of accuracy and reducing intra-subject variability.

Keywords

Forced oscillation technique, respiratory resistance, quality control, artifacts

Introduction

The forced oscillation technique (FOT) is a lung function test which provides detailed information about respiratory mechanics. FOT commonly involves superimposing small external pressure signals on the spontaneous, tidal breathing of the subject (12). Thus, FOT offers an advantage over traditional spirometry as it does not require forced maneuvers from the patient, which can be difficult for young children or in those with severe airway obstruction. Commercial FOT devices are becoming increasingly available, as are validation studies suggesting it has potential clinical utility (1; 20; 16; 3).

To become accepted as a clinical tool, there are still many barriers to overcome. FOT is known to have higher within- and between-test variability than spirometry (26), and it is difficult to disentangle variability due to physiological versus technical factors. Although recommendations for FOT measurements suggest practical strategies for reducing this variability, there are no specific guidelines as to how FOT quality control should be performed (20). It is also suggested that artifacts such as swallowing, glottal closure, leaks around the mouthpiece and noseclip should be excluded (20), however these are determined subjectively without quantifiable metrics or cut-offs.

Efforts to improve the quality of FOT measurements have included use of coherence (17) or statistical filters (25) and more recently, wavelet-based methods (2) to remove individual outlier points or windows of measurement. A complete breath method, where entire breaths rather than individual points associated with an artifact are removed, has been shown to be better than point-based statistical filters at reducing within- and between-session variability (24); however, artifacts still had to be identified manually by a subjective operator. In a previous technical exploratory study, we investigated the use of supervised machine learning methods to automatically and objectively detect artifacts, based on extraction of an exhaustive set of features from complete breaths (21).

In this study, we used the knowledge gained from our previous results to propose an automated artifact detection method which uses a specific set of features focused on the resistance versus flow (Rrs-flow) profile. We evaluated the performance of the method against different artifact removal techniques in pediatric and adult datasets, both in terms of agreement against a human operator as well as impact on within- and between-session variability.

Materials and Methods

Datasets

Adults: Details of the adult dataset have been previously published (26). Briefly, 10 healthy volunteers (mean(SD) age 32.2 (5.9) years, body mass index 23.2 (1.5)) and 10 patients with asthma (mean(SD) age 37.5 (11.6) years, body mass index 25.2 (4.6)) were recruited from Royal North Shore Hospital (St. Leonard, Australia) and the Woolcock Institute of Medical Research (Glebe, Australia). Subjects performed three technically acceptable FOT measurements during normal tidal breathing, wearing a nose clip, with cheeks supported and sitting in an upright position, each day over 7 visits within a 10-day period (consecutive days excluding weekends). Each subject was measured at the same of day every day to avoid diurnal variation. All subjects gave written, informed consent, and the study was approved by the Human Research Ethics Committee of Northern Sydney Central Coast Health.

Children: Data from 14 children were randomly selected from a larger epidemiological study (Ultrafine Particles from Traffic Emissions and Childrens Health, UPTECH); details have also been previously published (13,19). Briefly, eight- to eleven-year-old children (mean(SD) age 10.4 (1.1) years, weight 33.56 (6.73) kg, height 137.42 (6.47) cm) were recruited from public primary schools in the Brisbane Metropolitan Area (23% had doctor diagnosed asthma). FOT testing was performed as part of respiratory function assessment. Children were encouraged to breathe in a regular manner, avoid swallowing and maintain a tight mouthpiece seal. A series of technically acceptable FOT measurements were made with the child sitting upright, wearing a nose clip, with the cheeks and floor of the mouth supported by the child. The study was approved by the Queensland University of Technology Human Research Ethic Committee.

We randomly split each age group into two data sets: one for development and the other for test. Table 1 describes the development and test sets for children and for adults.

Measurements:

Respiratory system impedance (Zrs) was measured at 6, 11 and 19 Hz, using in-house built FOT devices conforming to current recommendations (20). Each FOT recording was one minute in total duration. Recordings were deemed acceptable by the technician if tidal volume and frequency appeared stable, with no obvious leaks and glottic closures from visual inspection of the volume trace. For the adult dataset, flow was measured using a screen-type pneumotachograph (R4830B series, flow range 0400 L/min, Hans Rudolph Inc., Shawnee, KS, USA) (26; 4). For the pediatric dataset, flow was also measured using a screen type pneumotachograph (R3700 series, flow range 0160 L/min, Hans Rudolph Inc, Shawnee, KS, USA) (13; 24).

The differential pressure was measured via a ± 2.5-cmH_2O silicon transducer (Sursense DCAL-4, Honeywell Sensing & Control, Golden Valley, MN, USA). Mouth pressure was measured by a similar transducer, with a range of ± 12.5-cmH_2O (Sursense DC005NDC4, Honeywell Sensing & Control, Golden Valley, MN, USA). Flow and pressure signals were digitally sampled at 396 Hz and digitally band-pass filtered with a bandwidth of ± 2 Hz centered around 6, 11 or 19 Hz. From Zrs, the respiratory system resistance (Rrs) and reactance (Xrs) were calculated for each frequency of interest separately, at $0.1s$ intervals as previous described (24) to allow a common reporting interval across the frequencies. Incomplete or partial breaths at the beginning or end of the recording were removed before any further processing, which helped ensure a balance between the inspiratory and expiratory contributions to each breath (24). For our filtering approach, we examined common variables obtainable from a FOT measurement, i.e. Rrs, Xrs, volume, pressure, and flow on a breath-by-breath basis.

Preprocessing

As a first step, we removed breaths containing data points which were physiologically implausible, i.e. those containing negative Rrs values (24). We also removed breaths corrupted by noise arising out of either nonlinearities in the pressure transducer or harmonics generated by nonlinearities in the respiratory system (18). These were defined as breaths having coherence values (see (5) and Appendix), C_{XY}, of pressure and flow less than 0.9 (17). C_{XY} and the impedance were calculated over $1/f$-second windows (where $f = 6$, 11 or 19 Hz), and ensemble-averaged every three windows with 50% overlap. For all three frequencies of interest, both the impedance and coherence were reported at intervals of 0.1 s. For the purposes of quality control, we primarily report our results for 6 Hz, although we also examined data at 11 and 19 Hz.

Feature Extraction of Rrs-flow Landmarks

In our previous work (21), we evaluated a list of potential features to separate respiratory artifacts from normal breaths. These include conventional statistical measures (e.g., minima, maxima, ranges, and variation) as well as more advanced features in time and frequency domains.

From a pool of 111 feature candidates including 11 commonly reported in the literature, we separately determined the top ten highest ranking candidates based on three different criteria (21). We found that "landmark" features used to characterize the shape of the Rrs-flow profile were consistent top performers across the methods. Thus, for the current study, feature selection focused on the Rrs-flow profile.

The within breath Rrs-flow curve provides a visual means of detecting glottal and laryngeal artifacts (24). Fig. 1 illustrates how to extract this landmark information from a complete breath. Point B and point

Z are two zero flow values for the higher and lower Rrs values. Point A and point D are at the maximum and minimum of Flow. Point CR, point CL and point E are at the maximum (right: positive Flow portion and left: negative Flow portion) and minimum of Rrs, respectively. Distance features from point Z to all other points are also calculated.

Interquartile range breath filter

As introduced in our exploratory work (21), a complete breath-based interquartile range filter (*IQR filter*) was proposed to replace the traditional standard deviation filter (e.g., *B-3SD* (24)). In this report, we used an IQR filter at two stages: a breath was marked as an artifact and discarded if its associated (*1*) Rrs, flow, Xrs, volume values and (*2*) landmark features extracted from the Rrs-flow profile had a value greater than a given upper threshold θ_H or less than a given lower threshold θ_L.

An IQR filter is described as follows. Let Q_1, Q_3, and IQR denote the 25^{th}, 75^{th} percentiles and the interquartile range of a variable, respectively. Let n_{IQR} be a number of interquartile intervals of any given variable away from its Q_1 and Q_3 values. The lower threshold $\theta_L = Q_1 - n_{IQR} \times IQR$ is the limit for values that are smaller than n_{IQR} away from Q_1; the upper threshold $\theta_H = Q_3 + n_{IQR} \times IQR$ is the limit for values that are greater than n_{IQR} away from Q_3. The effect of this filter can be adjusted using n_{IQR}, where an increased n_{IQR} reflects a less stringent rejection criterion. Previously, we used one parameter $n_{IQR} = 1$ across both age groups. In this study we investigated the effect of a wide range of n_{IQR} on filter performance.

The IQR filter implemented based on the use of landmark feature sets is termed *IQR-Landmark*. This is in contrast to our previous work using supervised learning to select from all features (21), which we refer to here as *IQR-SU*

Other filters

Previous work by Bhatawadekar et al. (2) proposed the use of wavelet decomposition for FOT artifact detection and removal. The method was based on the quantification of energy found in specific frequency bands and time locations to find differences between curves. We additionally examined the performance of this method against our *IQR-Landmark* filter, by also extracting wavelet coefficients from our FOT recordings using Eq. 1 (see Appendix). As per the Bhatawadekar et al. study, we used a three level decomposition with the Daubechies method (8) to obtain three coefficient vectors $cd1$, $cd2$, $cd3$ from the pressure signal, and then used their three recommended thresholds (i.e. $cd1^2 = 0.004$ $(cmH_2O)^2$; $cd2^2 = 0.023$ $(cmH_2O)^2$;

$cd3^2 = 0.07\ (cmH_2O)^2$) to detect artifacts. We also removed two neighbouring points from either side of the artifact point.

As this method was based on exclusion of points rather than complete breaths, we also compared two different implementations of the wavelet method: one as previously described (termed *Wavelet-point*), and one in which breaths containing artifacts detected by the wavelet method were excluded (termed *Wavelet-breath*).

Combined artifact detection

Finally, we examined the use of a composite detection algorithm, where we combined the use of the wavelet-based filter (2), i.e., *Wavelet-breath* with our *IQR-Landmark* filter. In preliminary investigations (results not shown), we determined that optimum performance was obtained using only the first level of derived wavelet coefficient $cd1$, previously found to be most sensitive and specific to high frequency artifacts such as light coughing which are often invisible on the recording. We applied a preset threshold of $cd1^2 = 0.004\ (cmH_2O)^2$ as per the work of Bhatawadekar et al (2). Thus, only the results for this combined algorithm are reported here for comparison, termed *IQR-Combined* . Specifically, the combined algorithm consists of three layers: *(1)* the pre-processing step, *(2)* the wavelet decomposition step, and *(3)* the IQR filter using landmark features (Fig. 2). Breaths that failed any threshold checking step were marked as artifacts and discarded (with complete-breath approach). The remaining breaths after three layers were considered to be *clean* data (i.e., without artifacts). -

Performance Measurement

Five automated filtering approaches are compared against the manual operator (*ground truth*) in our performance reports: our novel *IQR-Landmark*, *Wavelet-breath*, *Wavelet-point*, *IQR-SU*, and *IQR-Combined*. There was one manual operator for the adult (CT) and one for the pediatric (PDR) datasets, respectively; both researchers were experienced in analysing FOT waveforms. For comparison, we also report the results for raw unfiltered data and the manual operator (where not treated as ground truth). Performance of the filters was assessed using a number of measures:

Accuracy:

Breaths which were marked as artifacts by both our algorithm and the human operator (ground truth) were denoted as True Positives (TP), and breaths labelled as artifacts which did not agree with the ground truth we denoted as False Positives (FP). Breaths that the automated filtering approaches failed to label as artifacts but were annotated as such, were defined as False Negatives (FN). When the automated method and the annotation agreed a breath was not anomalous, it was counted as a True Negative (TN).

Sensitivity and specificity were defined as $\frac{TP}{TP+FN}$ and $\frac{TN}{TN+FP}$, respectively. The accuracy was calculated as $\frac{TP+TN}{TP+TN+FP+FN}$. F1-score (23), which is the harmonic mean of precision and sensitivity, has best value at 1 and worst at 0, is calculated as $\frac{2TP}{(2TP+FP+FN)}$. We investigated the effect of a wide range of nIQR on filter performance using receiver-operator characteristic (ROC) curves

Agreement:

Inter-rater reliability between a proposed method and human operators was assessed using unweighted Cohen's Kappa (6).

Within- and between-session variability:

Human-based artifact detection suffers heavily from subjective operators and using this as a gold standard may not reflect the true performance of a machine-based detection system. Hence, we additionally compared the variability of Rrs, via within-session coefficients of variation (wCV), between-session coefficients of variation (bCV) before and after discarding artifacts that are marked by clinicians versus our detection algorithms.

In the adult dataset, wCV quantified the variability from three recordings performed on the same day while bCV was obtained from 7-10 days per subject. In the pediatric dataset, wCV was computed from any number of recordings performed on the same day in each subject; it was not possible to compute bCV. For each filter, wCV and bCV were compared to the values for manual operator using paired t-tests.

Acceptability:

The discard rate is the percentage of filtered data in the total data input. As the first filter layer is standard practice (17) for any further data processing, the number of artifacts discarded by this layer is reported separately to facilitate comparison. For point-based approaches, the discard rate was reported in number of points; for complete-breath approaches, number of breaths is used.

Results

Comparisons of agreement and accuracy of filters against manual operator

In terms of comparison against the manual operator as ground truth, we examined the receiver-operator characteristic of the proposed filter across a range of n_{IQR} values (from $0.5 \rightarrow 3$ with 0.5 steps) for both adult and pediatric data. We found that $n_{IQR} = 1.5$ gave the best performance in adult data, whereas $n_{IQR} = 2.5$ gave the best performance for pediatric data. For adults, the positive rate fell below 0.45 when $n_{IQR} > 1.5$ or the false positive rate increased over 0.3 when $n_{IQR} < 1.5$. For children, when $n_{IQR} > 2$ the positive rate fell below 0.75 while $n_{IQR} < 2$ the false positive rate increased over 0.4. This agreed with the compared Rrs-flow profile between children and adults using the human removal (Fig. 3). Thus, we determined to use age-group oriented n_{IQR} to achieve the best performance (i.e., 2 for children and 1.5 for adults).

With the chosen n_{IQR} values, the combined method achieved 76% (adult) and 79% (pediatric) agreement with the manual operator. The performance metrics for the filters studied are shown in Table 2. Note that since the manual operator labelled acceptability in terms of breaths and not points, metrics were not available for the wavelet-point method.

Comparisons of variability and acceptability between filters

As mentioned, the agreement might not reflect the true performance of a test method. For example, Fig. 4 illustrates examples of artifacts in a recording that were missed by the operator but detected by our proposed method, i.e. contribution of the second and/or the third layer. The inter-rater comparison was observed to be poor, with Cohen's kappa = 0.473 (95% CI 0.411 to 0.534).

Table 3 and 5 show the variability of filtered *Rrs* profiles across test methods in comparison with the unfiltered data and filtering by a manual operator, for the development and test datasets, respectively. Of note, in the asthmatic adults test dataset, our proposed automated method yielded similar variability to that of the manual operator for wCV (6.5 vs 6.9%), and significantly improved bCV (8.2% vs 8.9%). In the pediatric test dataset, the wCV of our method was again similar to the manual operator (8.2% compared to 8.6%). The percentage of breaths that were removed by the first preprocessing layer from raw data sets were only about 1% (pediatric) and 2% (adult) (Table 3). The remaining breaths that were kept by our method were 69% (pediatric) and 73% (adult) of the total raw input (the manual method kept about 77% in both cases). While the *Wavelet-point* method kept 99% (pediatric) and 97% (adult) of total raw data points, *Wavelet-breath* only kept 78% (pediatric) and 98% (adult) of raw breaths. Without the wavelet layer, *IQR-Landmark* produced 74% (pediatric) and 81% (adult).

In the adult test dataset, i.e., those with asthma, the above performance was maintained (Table 4 and Table 5). Our method kept 66% of breaths compared with 69% of the human method. The accuracy of children test set was 89.1%, higher than 82.7% of the development performance.

Effect of combined filter at 11 and 19 Hz

We also examined the performance of the proposed combined filter when applied to FOT data at 11 and 19 Hz, in the two adult (development and test) datasets. In particular, if a breath was flagged as an artifact at 6 Hz, we looked at the proportion of breaths that were also flagged at 11 and 19 Hz, respectively. We found that for both datasets, out of the breaths identified as artifacts at 6 Hz, 85% were also classified as artifact at 11 Hz, and 83% were also classified as artifact at 19 Hz (true positives). In contrast, out of those breaths not considered artifacts at 6 Hz, only 15% was classified as artifact at 11 Hz, and 17% at 19 Hz (false negatives). Concordance between 6 and 11 Hz was moderate with kappa = 0.501 for the healthy dataset and 0.472 for the asthma dataset, and between 6 and 19 Hz was kappa = 0.464 for the healthy dataset and 0.454 for the asthma dataset.

Discussion

Summary of results

In this work, we propose a new technique for respiratory artifact removal, based on a novel scheme which involves extracting landmark features from the resistance versus flow profile and use of an interquartile range filter. We found that partly combining the method with the previously published wavelet detection method resulted in slightly higher accuracies and lower variability particularly in children.

We tested the different filtering methods using real data collected from a variety of subjects: children, healthy and asthmatic adults. A high degree of agreement between our method and the manual work was observed and several breaths containing artifacts missed by the manual operator were detected by our method. Possible reasons for human error include subjectivity in determining outlying Rrs vs flow loops, and the superimposition of multiple breaths in the software display potentially obscuring problematic breaths. Finally, within- and between-session variability was used to assess the performance of each filtering method in the absence of ground truth, i.e. without assuming the manual operator as gold standard. The combined method resulted in similar or lower variabilities compared with the operator, with a slightly higher exclusion rate. Though using the *IQR-Landmark* scheme produced a similar variation, a much lower exclusion rate than the operator implies that it may have missed several artifacts that were recognized by the human.

Comparison with other methods

In the past, quality control of forced oscillation data has often been done on the basis of measures such as coherence, i.e. the degree of correlation between the oscillatory flow and pressure waves, where coherence values less than 0.95 were typically excluded (17). However, this has known limitations: coherence is highly dependent on windowing and other signal processing settings (17), is potentially biased in the presence of nonlinearities (18), and has limited meaning when assessing the time course of impedance using single sliding windows (1). Furthermore both the literature (28) and anecdotal evidence suggests it is often much reduced in disease, particularly at low frequencies.

We have previously proposed using a complete breath approach to identify and exclude respiratory artifacts (24), in contrast to the more typical individual data point rejection using statistical filters based on number of standard deviations from the mean Rrs or Xrs value (25). Comparing to either 3SD or 5SD filtering, we found thata complete breath-based approach resulted in lower within- and between-session variability in children. We also proposed removal of transient artifacts based on the distinct deviations observed in the oscillatory flow and admittance signals, and in the Rrs-flow profile (24). Specifically, mouthpiece leak artifacts manifest as a marked increase in oscillatory flow and a pronounced spike in the magnitude of admittance. Other artifacts often contain depressions or gaps in the oscillatory flow signal but are best identified by examining the Rrs-Flow profile (e.g. spikes in Rrs at or near zero flow) (22,11,24). However, these observations were made subjectively, with no quantitative criteria or threshold to determine exclusion. The results of the present study represent a first step towards more objective and automated criteria for quality control of FOT measurements, based on a complete breath strategy. It employed an intuitive approach to detecting anomalies from the Rrs-flow profile, for the first time using landmark features to identify outliers.

The recent use of wavelet decomposition applied to the pressure profile of the breath (2) was effective at excluding light coughing, swallowing and vocalization artifacts. Although the wavelet method had high performance in sensitivity and specificity (over 90%), its evaluation was limited to simulated artifacts by trained subjects, and its performance on real world data was unknown. In this study, using retrospective clinical FOT data, we found that partially incorporating the wavelet approach into our proposed algorithm, particularly that component which detects artifacts invisible to the operator from the FOT recording, resulted in superior accuracies and similar variabilities compared to either method alone.

In previous exploratory work (21), we utilized several supervised learning selection algorithms to evaluate different features suitable for use with IQR filters. Using completely automated selection algorithms, similar variability was observed compared to a manual operator. The method was completely automated in that it required no a priori input for the n_{IQR} parameter, allowing it to operate independently of the target population characteristics, especially age. However, this came at a high computational cost due to the learning algorithms. The method also required a preset number of top ranking candidate features (e.g., 10). Our current proposed method is far less computationally intensive, and uses only features which are intuitive and potentially physiologically meaningful as it is based on the Rrs-flow profile. We propose that the n_{IQR} be customized to the target population of interest, and report the optimum n_{IQR} for an adult and pediatric test group.

Significance of findings

The improvements in within- and between-variability offered by the quality control methods examined in this study may be small compared to the natural physiological variability measured by FOT. However, they become important when the variability is a signal of interest that can provide insight into pathological states, via the use of simple and advanced analyses of variability (26; 22; 14) or the emerging interest in the flow-independent variability between end-inspiratory and end-expiratory resistance (7). In such cases, the sensitivities of such analyses can be refined by good quality control methods, to enhance discriminative power. For example, in our dataset, we see an improvement in wCV of approximately 0.5%. This may be small compared to the natural physiological variability of FOT (as deduced from the manually filtered wCV). However, it becomes significant when compared to the difference in wCV between health and asthma of approximately 2%, and would dramatically improve the ability to discriminate between the groups.

More importantly, we have shown that it is possible to implement an objective, automated method of quality control which performs just as well or slightly better than an expert manual operator. This is an advancement on our previous approach (24) showing significantly better performance than simple filtering methods but was still a subjective method relying on an expert manual operator.

Limitations

Most commercial FOT systems employ multi-frequency signals. We have focused our quality control approach on 6 Hz, as it or 5 Hz is the most common frequency of primary interest reported in the literature. We did not evaluate how the proposed filter compared against manual quality control at other frequencies, as we would recommend always taking the quality of the primary frequency of interest into account. When we compared the performance of the filter at 11 and 19 Hz to 6 Hz, we found that breaths were more likely to be excluded at 6 Hz than at 11 and 19 Hz. There were proportionally fewer breaths excluded at 11 and 19 Hz that were not already excluded at 6 Hz. Thus, there was moderate concordance between 6 Hz and the higher frequencies. In practice, impedance at lower frequencies are more susceptible to the effects of breathing, however the effects of glottal interference may tend to manifest at higher frequencies. The higher sensitivity to detect artifacts at 6 Hz could be due to the observation that resistance spikes at 11 and 19

Hz were generally smaller (and perhaps more difficult to detect) than at 6 Hz, or simply the fact that the algorithm was optimised using data from 6 Hz.

In terms of applicability, the test datasets we examined exhibited a mild to medium range of airway obstruction, ranging in Rrs from 1.7 to 8 cmH_2OsL^{-1}. Thus our method will need to be tested for applicability across a wide range of obstruction, e.g., severely obstructed patients or during an exacerbation. However, we note that our performance metrics remained mostly high (approval rate $\geq 75\%$) regardless of median Rrs in both the children and adult datasets. There was also a low correlation between approval rate and Rrs as reported previously (21).

In our previous work (21), we found that features associated with Xrs did not rank highly compared to Rrs in predicting manual operator decisions in the same datasets. Minimum and range of Xrs were within the top 10 ranking features across all those examined, and outlying values were taken into consideration in the combined filter (Figure 2, Layer 2), but we did not examine detailed landmark features in e.g. the Xrs versus flow or volume profiles. However, it is worth noting that these results may only be relevant to the healthy and asthma populations we examined.

Further work will also be needed to determine how our method will perform in other diseases, e.g. acute respiratory distress syndrome (10; 15), or chronic obstructive pulmonary disease, where abnormalities in Xrs may be more important than Rrs, but may also be confounded by expiratory flow limitation (11).

Finally, we only relied on one manual operator for each dataset and did not examine inter-rater variability. This may have underestimated within- and between-session variability for manual exclusion, as well as the differences with and between the filters being tested.

Conclusions

Lack of standardization in FOT has contributed to diversity in FOT setups, signal processing and quality control approaches across manufacturers and laboratories. This has been a barrier to its adoption into widespread clinical usage, despite decades of studies showing promising physiological and clinical relevance. Our work shows that the resistance versus flow profile is a useful target for automated exclusion of artifacts on a breath-by-breath basis. The ability to remove common artifacts using objective and automatable criteria is an important step towards overcoming this barrier, as these approaches can be eventually incorporated into commercial software to guide the user and minimize inter-operator variability. These approaches are also especially desirable in emerging applications of FOT such as in epidemiological field testing (13) and home monitoring (9; 27).

Appendix and Equations

Wavelet decomposition coefficients (8) and spectral coherence (5) was calculated as below.

Wavelet decomposition: Let $s(t)$ be a curve which can be presented by coefficients $C\left(a,b\right)$(1).

$$C\left(a,b\right)=\frac{1}{\sqrt{a}}\int\limits_{-\infty}^{+\infty}s(t)\psi_{a,b}(t)dt \tag{1}$$

where $\psi_{a,b}\left(t\right)=\psi\left(\frac{t-b}{a}\right)$ is an expanded or contracted and shifted version of a unique wavelet function $\psi(t)$ a and b are the scale and the time localization, respectively.

In this work, we implemented a three level decomposition with the Daubechies method (8) to obtain $cd1$, $cd2$, $cd3$ using Matlab packages (The MathWorks Inc., Natick, MA, 2000). The Daubechies wavelets are orthogonal wavelets defining a discrete wavelet transform (DWT).

Spectral coherence: Let C_{XY} be the spectral coherence between signals X and Y. C_{XY} is defined by the Welch method (5) as in Eq. 2.

$$C_{XY}(\omega)=\frac{P_{XY}(\omega)}{\sqrt{P_{XX}(\omega).P_{YY}(\omega)}} \tag{2}$$

where ω is frequency, $P_{XX}(\omega)$ is the power spectrum of signal x, $P_{YY}(\omega)$ is the power spectrum of signal y, and $P_{XY}(\omega)$ is the cross-power spectrum for signals x and y. When $P_{XX}(\omega)=0$ or $P_{YY}(\omega)=0$, then also $P_{XY}(\omega)=0$ and we assume that $C_{XY}(\omega)$ is zero.To estimate power and cross spectra, let $\mathfrak{F}_x(\omega)$ and $\overline{\mathfrak{F}_x(\omega)}$, denote the Fourier transform and its conjugate of signal x, respectively, i.e. $\mathfrak{F}_x(\omega)=\int\limits_{-\infty}^{+\infty}x(t).e^{-j\omega t}dt$. The power spectrum is then: $P_{XX}(\omega)=\mathfrak{F}_x(\omega).\overline{\mathfrak{F}_x(\omega)}$; $P_{YY}(\omega)=\mathfrak{F}_y(\omega).\overline{\mathfrak{F}_y(\omega)}$; and $P_{XY}(\omega)=\mathfrak{F}_x(\omega).\overline{\mathfrak{F}_y(\omega)}$.

Acknowledgements

T. Pham was supported by Endeavour Scholarships and Fellowships (Prime Minister's Australia Postgraduate Scholarship) and the Faculty of Engineering & Information Technologies, The University of Sydney, under the Faculty Research Cluster Program. C. Thamrin was supported by the Co-operative Research Centre for Asthma, and funded by a National Health and Medical Research Council of Australia Career Development Fellowship. We would like to thank the UPTECH study team for the provision of the pediatric dataset. The UPTECH study is a collaborative research project between Queensland University of Technology (QUT), the Woolcock Institute of Medical Research (WIMR), and Queensland Institute of Medical Research (QIMR). The study was funded through a Linkage Grant LP0990134 by the Australian Research

Council, QLD Department of Transport and Main Roads (DTMR) and QLD Department of Education, Training and Employment (DETE). We would like to thank Dr. Sophie Timmins for the provision of the adult dataset. We also like to thank Dr. Diep N. Nguyen for technical discussions and suggestions to our work. There is no conflict of interest.

References

1. **Bates JHT, Irvin CG, Farre R, and Hantos Z.** Oscillation mechanics of the respiratory system. *Comprehensive Physiology*, 1:1233–1272, 2011.
2. **Bhatawadekar SA, Leary D, Chen Y, Ohishi J, Hernandez P, Brown T, McParland C, and Maksym GN.** A study of artifacts and their removal during forced oscillation of the respiratory system. *Ann Biomed Eng*, 41(5): 990–1002, 2013.
3. **Brown NJ, Thorpe CW, Thompson B, Berend N, Downie S, Verbanck S, Salome CM, and King GG**. A comparison of two methods for measuring airway distensibility: nitrogen washout and the forced oscillation technique. *Physiol Mea*, 25(4): 1067–1075, 2004.
4. **Brown NJ, Xuan W, Salome CM, Berend N, Hunter ML, Musk AB, James AL, King GG**. Reference equations for respiratory system resistance and reactance in adults. *Respiratory physiology & neurobiology*,172(3): 162-168, 2010.
5. **Challis R and Kitney R**. Biomedical signal processing (part 3 of 4): The power spectrum and coherence function. *Med Biol Eng Comput*, 28(6): 509–524, 1990.
6. **Cohen J** A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20(1): 37-46, 1960.
7. **Czövek D, Shackleton C, Hantos Z, Taylor K, Kumar A, Chacko A., Ware RS, Makan G, Radics B, GinglZ, et al..** Tidal changes in respiratory resistance are sensitive indicators of airway obstruction in children. *Thorax*,71(10): 907-915, 2016.
8. **Daubechies I and Bates BJ**. Ten lectures on wavelets. *J Acoust Soc Am*, 93(3):1671–1671, 1993.
9. **Dellaca RL, Gobbi A, Pedotti A, and Celli B**. Home monitoring of within-breath respiratory mechanics by a simple and automatic forced oscillation technique device. *Physiological measurementJ*, N11, 2010.
10. **Dellaca RL, Olerud MA, Zannin E, Kostic P, Pompilio PP, Hedenstierna G, Pedotti A, Frykholm P**. Lung recruitment assessed by total respiratory system input reactance. *Intensive care medicine*, 2164, 2009.
11. **Dellaca RL, Pompilio PP, Walker PP, Duffy N, Pedotti A, and Calverley PMA**. Effect of bronchodilation on expiratory flow limitation and resting lung mechanics in COPD. *Eur Respir J*, 33(6): 1329–1337, 2009.
12. **DuBois AB, Brody AW, Lewis DH, Burgess BF, et al.**. Oscillation mechanics of lungs and chest in man. *J Appl Physiol*, 8(6): 587–594, 1956.

13. **Ezz WN, Mazaheri M, Robinson P, Johnson GR, Clifford S, He C, Morawska L, and Marks GB.** Ultrafine particles from traffic emissions and childrens health (UPTECH) in Brisbane, Queensland (Australia): Study design and implementation. *International journal of environmental research and public health*, 12(2):1687–1702, 2015.

14. **Gobbi A, Dellaca RL, King GG, Thamrin C**. Toward predicting individual risk in asthma using daily home monitoring of resistance. *American journal of respiratory and critical care medicine*, 195(2): 265-267, 2017.

15. **Kaczka DW, Lutchen KR, Hantos Z,**. Emergent behavior of regional heterogeneity in the lung and its effects on respiratory impedance. *Journal of Applied Physiology*, 110(5): 1473-1481, 2011.

16. **King GG** Cutting edge technologies in respiratory research: Lung function testing. *Respirology*, 16(6):883–890, 2011.

17. **Lorino H, Mariette C. , Karouia M. , and Lorino AM.** Influence of signal processing on estimation of respiratory impedance. *Journal of Applied Physiology*, 74(1):215–223, 1993.

18. **Maki BE**. Interpretation of the coherence function when using pseudorandom inputs to identify nonlinear systems. *IEEE transactions on biomedical engineering*, 33(8): 775–779, 1986.

19. **Mazaheri M, Clifford S, Jayaratne R. , Mokhtar MAM, Fuoco F. , Buonanno G. , and Morawska L.** School childrens personal exposure to ultrafine particles in the urban environment. *Environmental science & technology*, 48(1):113–120, 2013.

20. **Oostveen E. , MacLeod D. , Lorino H, Farre R, Hantos Z, Desager K. , Marchal F. , et al.** The forced oscillation technique in clinical practice: methodology, recommendations and future developments. *European Respiratory Journal*, 22(6):1026–1041, 2003.

21. **Pham TT, Thamrin C, Robinson PD, and Leong PHW.** Respiratory artefact removal in forced oscillation measurements: A machine learning approach. *Biomedical Engineering, IEEE Transactions on*, in press, 2016.

22. **Que C. , Kenyon CM, Olivenstein R, Macklem PT, and Maksym GN**. Homeokinesis and short-term variability of human airway caliber. *J Appl Physiol*, 91(3): 1131–1141, 2001.

23. **Van Rijsbergen CJ** . Information Retrieval. *Butterworth-Heinemann*. 2nd ed., Newton, MA, USA, 1979.

24. **Robinson PD, Turner M, Brown NJ, Salome C, Berend N. , Marks GB, and King GG**. Procedures to improve the repeatability of forced oscillation measurements in school-aged children. *Respir Physiol Neurobiol*, 177(2): 199–206, 2011.

25. **Schweitzer C. , Chone C. , and Marchal F**. Influence of data filtering on reliability of respiratory impedance and derived parameters in children. *Pediatr Pulmonol*, 36(6):502–508, 2003.

26. **Timmins SC, Coatsworth N, Palnitkar G, Thamrin C, Farrow CE, Schoeffel RE, Berend N, Diba C, Salome CM, and King GG.** Day-to-day variability of oscillatory impedance and spirometry in asthma and COPD. *Respiratory physiology & neurobiology*, 185(2):416–424, 2013.

27. **Timmins SC, Diba C, Thamrin C, Berend N, Salome CM, and King GG.** The feasibility of home monitoring of impedance with the forced oscillation technique in chronic obstructive pulmonary disease subjects. *Physiological measurement*, 34(1):67–81, 2012.

28. **Wouters EFM, Verschoof AC, Polko AH, Visser BF**. Impedance measurements of the respiratory system before and after salbutamol in COPD patients. *Respiratory medicine*, 83(4):309–313, 1989.

Figure Legends

Figure 1. *7 points* proposed to determine *thresholdary landmarks* (dotted) for a *Rrs* against Flow curve from one breath of a child. Features extracted by landmarks for this breath are Euclidean distances between points (dotted).

Figure 2. Combined Respiratory artifact detection scheme. Rrs is resistance values of input breaths. Cxy is the spectral coherence between pressure and flow values of breaths. $Cd^2{}_1$ is the squared first level wavelet decomposition of pressure values. R, F, X, V are resistance, flow, reactance, volume values. R, F, X, V are checked if in their normal range. Fea is the advanced feature set extracted (from the relationship between Rrs and $Flow$ values) is checked with their threshold ranges.

Figure 3. Example of Rrs-flow profile of a measurement. (a): adult data. (b): children data. Solid lines are accepted breaths and dotted lines are discarded data by manual operator.

Figure 4. Example artifacts in a recording that were missed by the operator but detected by Layer 2 (square markers) and/or Layer 3 (diamond markers). The breath in bold indicates an artifact that was detected and excluded by the operator.

Tables

Table 1: Data Sets used in this work.

	Adults		Children	
Characteristics	**Development**	**Test**	**Development**	**Test**
Dataset source	Timmins et al (26)	Timmins et al (26)	Ezz et al (13)	Ezz et al (13)
Collection Site	Sydney, NSW	Sydney, NSW	Brisbane, QLD	Brisbane, QLD
Diagnosis	healthy	asthmatics	healthy/ asthmatics	healthy/ asthmatics
No. of subjects	9	10	9	5
No. of measurements	261	285	69	31
No. of breaths	3067	3947	1110	580

Table 2: Comparisons of filters against the manual operator during development. *IQR-Landmark* and *IQR-SU* are our works related to our current proposed, *IQR-Combined*. Others are the existing. Positives are artifacts. True positive breaths are breaths rejected by both machine-based and manual removal. F1-score is the harmonic mean of precision and sensitivity.

	Healthy Adults				Children			
Method	Accuracy[a]	F1[a]	Sensitivity[a]	Specificity[a]	Accuracy[a]	F1[a]	Sensitivity[a]	Specificity[a]
IQR-Landmark[b]	0.753	0.545	0.640	0.787	0.693	0.525	0.842	0.655
Wavelet-breath[c] (2)	0.584	0.335	0.453	0.623	0.431	0.341	0.730	0.356
IQR-SU[d] (21)	0.763	0.571	0.683	0.787	0.731	0.553	0.824	0.708
IQR-Combined[e]	0.781	0.569	0.626	0.828	0.827	0.632	0.734	0.851

[a] Removals by a specialist is considered ground truth.
[b] A single filter approach with landmark features and $nIQR = 1.5$ for adults and 2.5 for children (where relevant).
[c] A complete breath rejection approach using the wavelet coefficient thresholding detecion.
[d] A single filter approach with features selected by a supervised learning technique (21) and $n_{IQR} = 1$ for both age groups.
[e] A multi-filter approach (comprising a wavelet and *IQR-Landmark*) with $n_{IQR} = 1.5$ (adults) or 2.5 (children).

Table 3: Comparison of filtered *Rrs* profiles between filters during development. *IQR-Landmark* and *IQR-SU* are our works related to our current proposed, *IQR-Combined*. Others are the existing. *wCV* and *bCV* are in %. *P* values are from paired t-tests (two-tailed). $\%_{out}$ is the percentage of remaining *breaths* (against the total raw input, unit in %) after being filtered by methods except for *Wavelet-point* which is in percentage of the raw data points. $\%_{\text{discarded-by-preprocessing}}$ is the percentage of artifacts that were removed in the preprocessing step (a common step for all test filters).

	Healthy Adults					**Children**		
Method	**wCV**	P-value wCV[a]	**bCV**	P-value bCV[a]	$\%_{out}$	**wCV**	P-value wCV[a]	$\%_{out}$
Unfiltered (raw data)	*5.25*	-	*6.69*	-	*100.0*	*13.62*	-	*100.0*
Manual (reference)	**5.14**	-	**6.31**	-	**76.9**	**11.66**	-	**77.2**
IQR-Landmark[b]	4.56	0.08	5.76	0.18	80.6	12.69	0.57	74.5
Wavelet-point (2)	5.43	0.34	6.84	0.46	97.1	13.96	0.30	98.9
Wavelet-breath[c]	5.93	0.20	7.82	0.34	98	11.9	0.85	77.8
IQR-SU[d]	4.69	0.20	5.91	0.05	67.8	12.25	0.80	60.0
IQR-Combined[e] (proposed)	4.57	0.11	5.75	0.17	72.8	13.27	0.32	69.6
$\%_{\text{discarded-by-preprocessing}}$					1.9			2.6

[a] compared to *Manual operator*, significant if $P < 0.05$.
[b] A single filter approach with landmark features and $nIQR = 1.5$ for adults and 2.5 for children (where relevant).
[c] A complete breath rejection approach using the wavelet coefficient thresholding detecion by the research group (2).
[d] A single filter approach with features selected by a supervised learning technique (21) and $n_{IQR} = 1$ for both age groups.
[e] A multi-filter approach (comprising a wavelet and *IQR-Landmark*) with $n_{IQR} = 1.5$ (adults) or 2.5 (children).

Table 4: Comparisons of filters against the manual operator with out-of-sample data. *IQR-Landmark* and *IQR-SU* are our works related to our current proposed, *IQR-Combined*. Others are the existing. Positives are artifacts. True positive breaths are breaths rejected by both machine-based and manual removal. F1-score is the harmonic mean of precision and sensitivity.

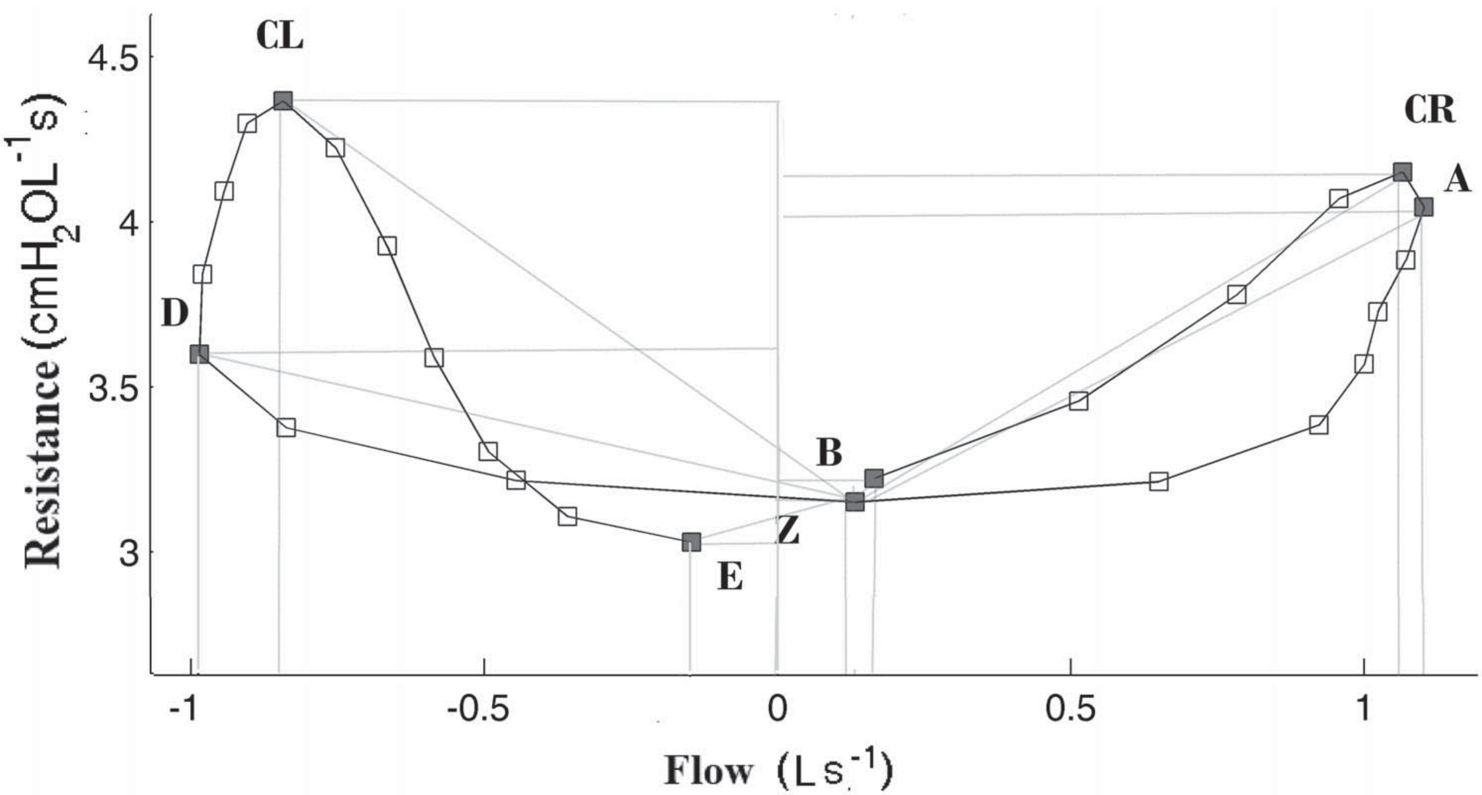
	Asthma Adults				**Children**			
Method	Accuracy[a]	F1[a]	Sensitivity[a]	Specificity[a]	Accuracy[a]	F1[a]	Sensitivity[a]	Specificity[a]
IQR-Landmark[b]	0.719	0.610	0.715	0.720	0.738	0.398	0.848	0.725
Wavelet-breath[c] (2)	0.596	0.435	0.506	0.636	0.369	0.179	0.674	0.334
IQR-SU[d] (21)	0.736	0.606	0.661	0.769	0.747	0.412	0.870	0.733
IQR-Combined[e]	0.731	0.609	0.683	0.752	0.891	0.588	0.761	0.906

[a] Removals by a specialist is considered ground truth.
[b] A single filter approach with landmark features and $nIQR = 1.5$ for adults and 2.5 for children (where relevant).
[c] A complete breath rejection approach using the wavelet coefficient thresholding detecion.
[d] A single filter approach with features selected by a supervised learning technique (21) and $n_{IQR} = 1$ for both age groups.
[e] A multi-filter approach (comprising a wavelet and *IQR-Landmark*) with $n_{IQR} = 1.5$ (adults) or 2.5 (children).

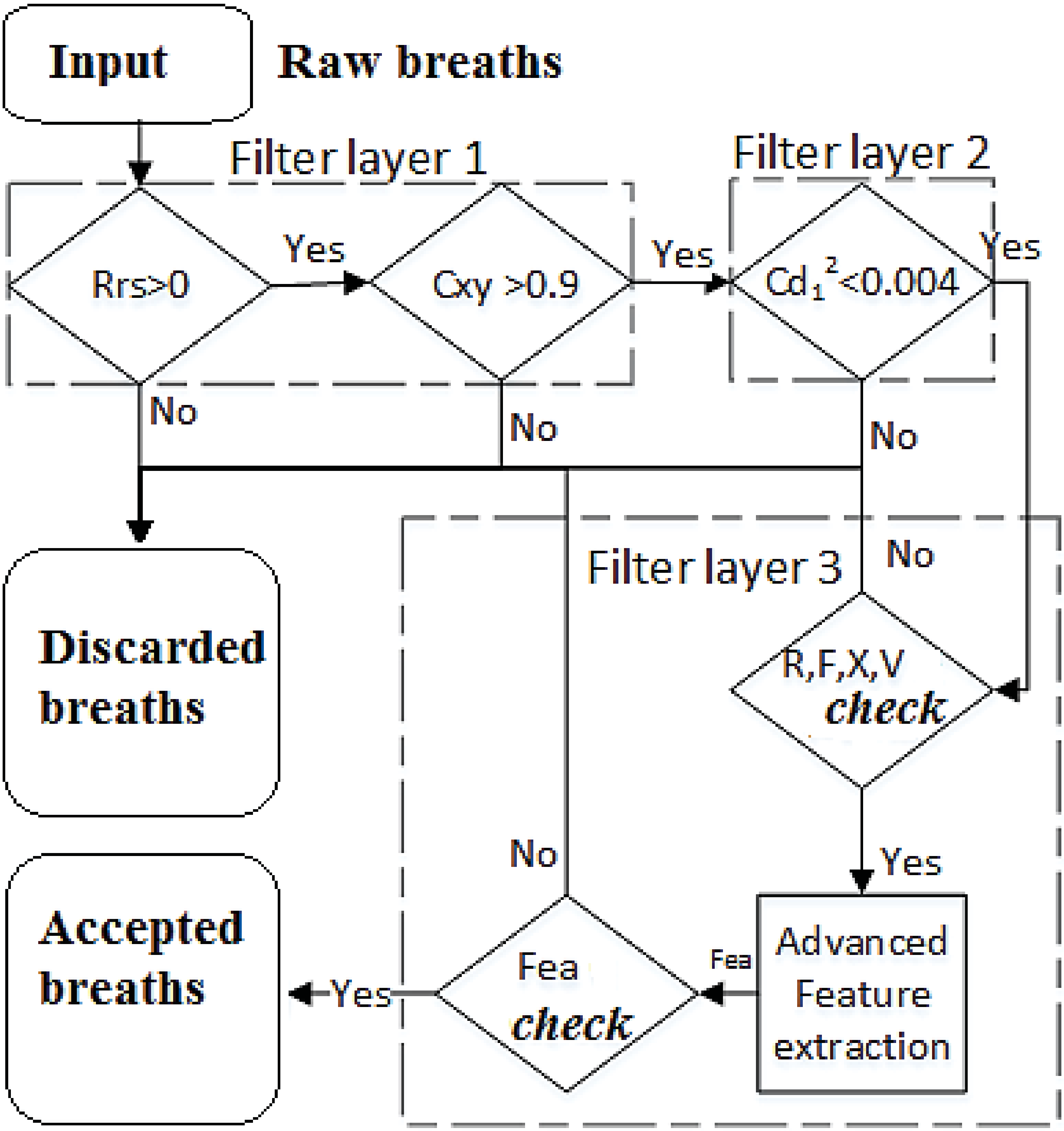
Table 5: Comparisons between filters during out-of-sample tests using the Rrs profile. *IQR-Landmark* and *IQR-SU* are our works related to our current proposed, *IQR-Combined*. Others are the existing. *wCV* and *bCV* are in %. *P* values are from paired t-tests (two-tailed). $\%_{out}$ is the percentage of remaining *breaths* (against the total raw input, unit in %) after being filtered by methods except for *Wavelet-point* which is in percentage of the raw data points. $\%_{\text{discarded-by-preprocessing}}$ is the percentage of artifacts that were removed in the preprocessing step (a common step for all test filters).

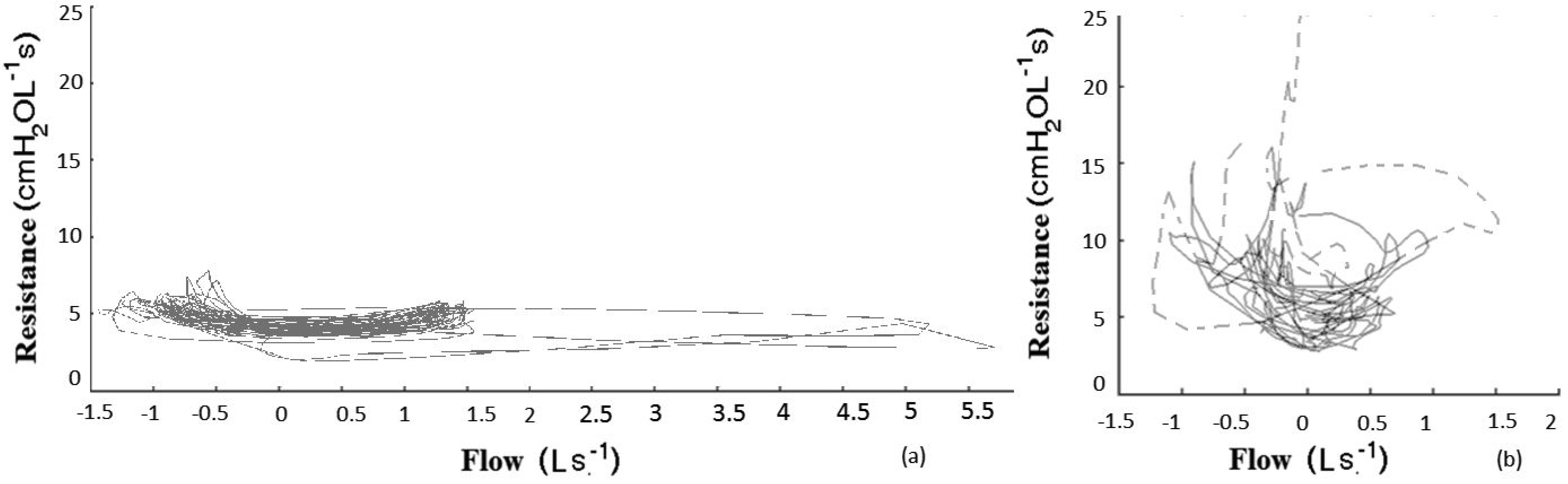
	Asthma Adults					**Children**		
Method	**wCV**	P-value wCV[a]	**bCV**	P-value bCV[a]	$\%_{out}$	**wCV**	P-value wCV[a]	$\%_{out}$
Unfiltered (raw data)	*6.25*	-	*7.95*	-	*100.0*	*8.41*	-	*100.0*
Manual (reference)	**6.86**	-	**8.86**	-	**68.9**	**8.55**	-	**89.8**
IQR-Landmark[b]	6.52	0.13	8.22	0.05	80.6	8.30	0.66	74.5
Wavelet-point (2)	6.64	0.57	8.15	0.09	98.7	9.06	0.21	79.1
Wavelet-breath[c]	7.51	0.37	8.35	0.24	97.9	10.12	0.23	33.3
IQR-SU[d]	7.93	0.26	6.68	0.05	63.7	8.62	0.86	67.1
IQR-Combined[e](proposed)	6.46	0.12	8.18	**0.03**	65.6	8.22	0.62	66.9
				$\%_{\text{discarded-by-preprocessing}}$	2.5			0.9

[a] compared to *Manual operator*, significant if $P < 0.05$.
[b] A single filter approach with landmark features and $nIQR = 1.5$ for adults and 2.5 for children (where relevant).
[c] A complete breath rejection approach using the wavelet coefficient thresholding detecion by the research group (2).
[d] A single filter approach with features selected by a supervised learning technique (21) and $n_{IQR} = 1$ for both age groups.
[e] A multi-filter approach (comprising a wavelet and *IQR-Landmark*) with $n_{IQR} = 1.5$ (adults) or 2.5 (children).

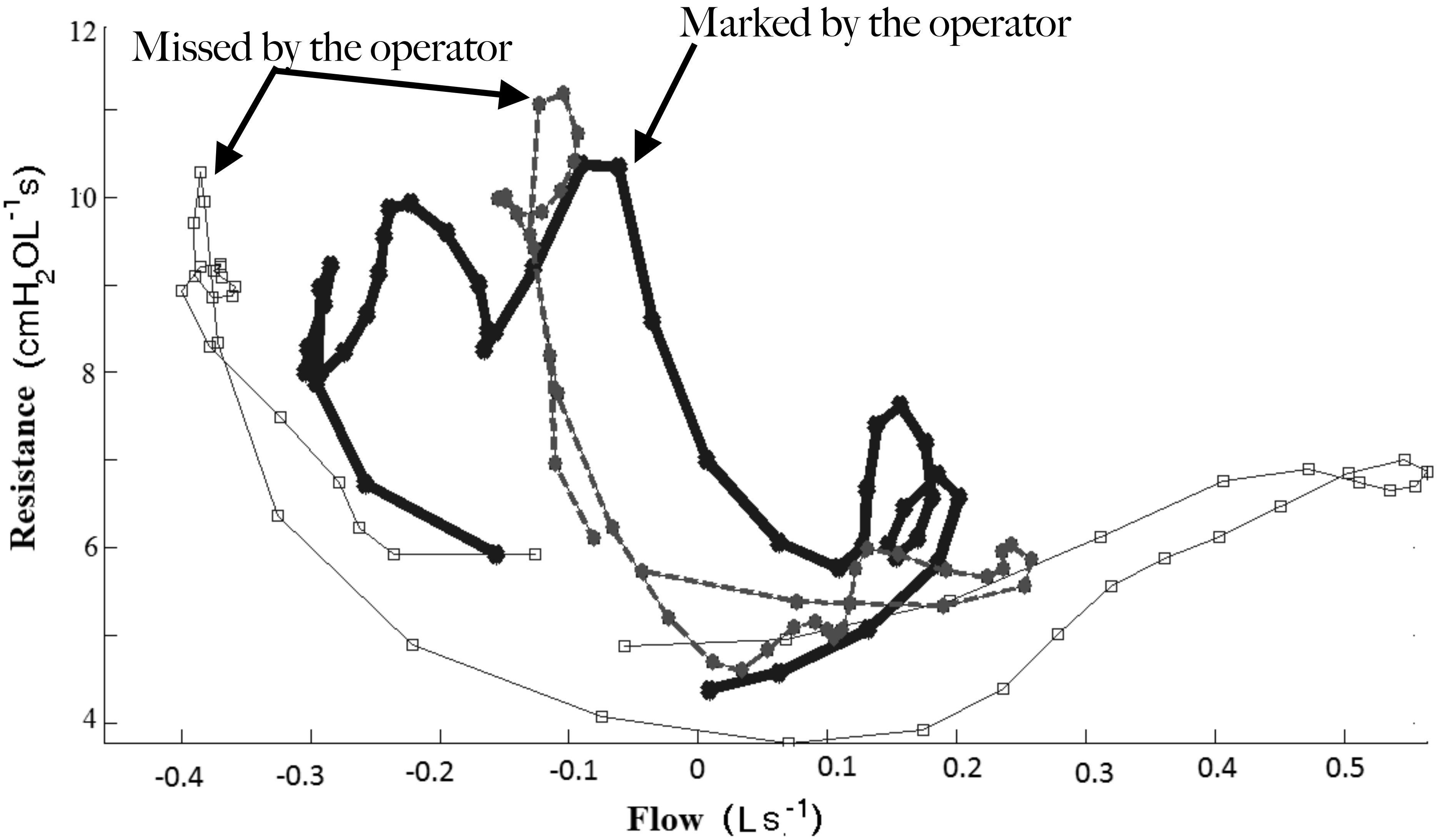


CL
CR
A
D
B
Z
E
4.5
4
3.5
3
-1
-0.5
0
0.5
1
Resistance (cmH2OL-1s)
Flow (L s-1)


Input
Raw breaths
Filter layer 1
Rrs>0
Yes
Cxy >0.9
Yes
Filter layer 2
Cd1^2<0.004
Yes
No
No
No
Discarded breaths
Filter layer 3
No
R,F,X,V check
Yes
No
Accepted breaths
Yes
Fea check
Fea
Advanced Feature extraction



Resistance (cmH$_2$OL^{-1}s)
0
5
10
15
20
25
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
Flow (Ls^{-1})
(a)
Resistance (cmH$_2$OL^{-1}s)
0
5
10
15
20
25
-1.5
-1
-0.5
0
0.5
1
1.5
2
Flow (Ls^{-1})
(b)


Missed by the operator
Marked by the operator
12
10
8
6
4
Resistance (cmH2OL-1s)
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
Flow (L s-1)