Towards Sleep Study Automation: Detection Evaluation of Respiratory-Related Events

Michal Borsky, *Member, IEEE*, Marta Serwatko, Erna S. Arnardottir, Jacky Mallett, *Member, IEEE*

***Abstract*— The diagnosis of sleep disordered breathing depends on the detection of several respiratory-related events: apneas, hypopneas, snores, or respiratory event-related arousals from sleep studies. While a number of automatic detection methods have been proposed, reproducibility of these methods has been an issue, in part due to the absence of a generally accepted protocol for evaluating their results. With sleep measurements this is usually treated as a classification problem and the accompanying issue of localization is not treated as similarly critical. To address these problems we present a detection evaluation protocol that is able to qualitatively assess the match between two annotations of respiratory-related events. This protocol relies on measuring the relative temporal overlap between two annotations in order to find an alignment that maximizes their F1-score at the sequence level. This protocol can be used in applications which require a precise estimate of the number of events, total event duration, and a joint estimate of event number and duration. We assess its application using a data set that contains over 10,000 manually annotated snore events from 9 subjects, and show that when using the American Academy of Sleep Medicine Manual standard, two sleep technologists can achieve an F1-score of 0.88 when identifying the presence of snore events. In addition, we drafted rules for marking snore boundaries and showed that one sleep technologist can achieve F1-score of 0.94 at the same tasks. Finally, we compared this protocol against the protocol that is used to evaluate sleep spindle detection and highlighted the differences.**

***Index Terms*— sleep disordered breathing, event detection, snoring, evaluation protocol, sequence alignment**

Abbreviations

AASM	American Academy of Sleep Medicine
AHI	Apnea-Hypopnea Index
AI	Artificial Intelligence
C	Confusion
DE	Duration Evaluation
SDC	Sórensen-Dice Coefficient
DL	Deep Learning
JI	Jaccard Index
FA	False Alarm
FN	False Negative
FP	False Positive
H	Hit
HE	Hypothesis Event
M	Miss
ML	Machine Learning
PDE	Presence and Duration Evaluation
PE	Presence Evaluation
RE	Reference Event
RERA	Respiratory Event-Related Arousal
SDB	Sleep Disordered Breathing
SI	Snore Index
SSD	Sleep Spindle Detection
TN	True Negative
TP	True Positive
κ	kappa

This paper was submitted for review on 1st of March 2021. "This study was supported by The Icelandic Research Fund No. 174067 and No. 175256, and NordForsk grant No. 90458."

Michal Borsky is with The Reykjavik University Sleep Institute, School of Technology, Reykjavik University, 102 Reykjavik, Iceland (e-mail: michalb@ru.is).

Marta Serwatko is with The Reykjavik University Sleep Institute, School of Technology, Reykjavik University, 102 Reykjavik, Iceland (e-mail: martas@ru.is).

Erna S. Arnardottir is with The Reykjavik University Sleep Institute, School of Technology, Reykjavik University, 102 Reykjavik, and with the Landspitali University Hospital, Reykjavik, Iceland. (e-mail: ernas@ru.is).

Jacky Mallett is with the Reykjavik University Sleep Institute, School of Technology, Reykjavik University, 102 Reykjavik, Iceland (e-mail: jacky@ru.is).

We would like to thank Dr. Thorarinn Gislason at Landspitali University Hospital, Reykjavik, Iceland, the principal investigator of the snoring dataset, for providing us with the study data.

I. Introduction

Sleep disordered breathing (SDB) is a condition that ranges from primary snoring to obstructive sleep apnea. It is clinically diagnosed by manually annotating a sleep study, for the presence of respiratory related SDB events such as snores, apneas, hypopneas, or respiratory event-related arousals (RERAs) [1]. The prevalence, risk factors and costs associated with diagnosing the various SDB forms [2] has fuelled work on automating its diagnosis and severity assessment, and modern artificial intelligence (AI) or deep learning (DL) solutions have attracted significant attention for this purpose, as reviewed in [3], and issues discussed in [4] [5]

Snoring is an important biomarker of SDB, but the severity of snoring is often evaluated subjectively [6], [7], [8]. Traditional approaches to automatic snoring detection have relied on handcrafted low-level temporal and spectral features and shallow learning algorithms [9], [10], [11]. Their advantage was that the features, and the model's decision, were easy to interpret, but this came at the expense of low discriminability (mid 80% frame accuracy). Alternative approaches relied on

low-level features and their descriptors [12], [13], [14] and more recently DL algorithms [15]. These approaches had a moderate-to-high discriminability ($\pm$90% frame accuracy), but this came at the expense of interpretability. The latest approaches have attempted to leverage the big data paradigm in order to extract latent features using purely data-driven approaches from a raw, or only very minimally processed waveform, then feeding this into a sequence-based DL model. Features can be extracted from the middle layers of an autoencoder [16], or occasionally as purely synthetic data generated by a data augmentation technique [17]. Their main advantage is that feature extraction is jointly optimized together with the model parameters completely automatically.

In broader machine learning (ML) applications, object detection is typically treated as a joint classification and localization problem. This approach has been popularized by several annual competitions held by the National Institute of Standards and Technology. The CLEAR 2006 [18], CLEAR 2007 [19], and the Rich Transcription 2007 [20] competitions spanned multiple ML fields and one of the main outcomes was the adoption of spatial coincidence measures for image recognition [21] [22], [23] [24], and temporal coincidence measures for sound event detection [24] [25], which was defined as a task to determine "the identity of sounds and their temporal position in the signal [26]". The acoustic event detection problems shares with the detection of respiratory-related events its reliance on time series signals, which strongly suggests that their evaluation should be based on temporal coincidence measures. A related problem in EEG measurements, sleep spindle detection (SSD) has used a temporal coincidence measure, the Jaccard index (JI), for several years [27] [28] [29], [30]. A still prevalent trend in snore detection, however, is to cut the time axis into fixed length segments with arbitrarily placed boundaries, also called frames, and attempt to determine simply if an event is present within the segment or not. This practice devolves a detection task into a classification problem as event localization then becomes dependant on the granularity of the segmentation, where segment lengths can range from tens of milliseconds to several seconds, and can contain multiple events or only a part of one. Relatively few approaches currently evaluate their algorithm's performance using temporally aware measures, [31], [15], [32], [33], [34].

This article proposes an evaluation protocol that is suited to quantify an agreement of annotation of respiratory related SDB events. It jointly assesses temporal localization and classification performance. Our objective is to improve SDB diagnosis, provide a method to asses reliability of labels for supervised AI training, and to assist in the reproduction and comparison of published AI solutions. We focus on snore events but we believe the protocol is also applicable to apneas, hypopneas, or RERAs. Our motivation is based on the fact that while snoring is very prevalent in the general population, it lacks a gold standard definition, and has received less attention than apneas and hypopneas, whilst still being important in the diagnosis of SDB disorders. To support our claims we apply the protocol on manual annotations of over 10,000 snore events marked by three experienced sleep technologists.

II. Proposed Evaluation Protocol

The proposed evaluation protocol associates a temporal event overlap with a success, and non-overlapping events and parts of events with a failure. The protocol relies on a relative temporal overlap, instead of an absolute one, because it discriminates against overly long and too short predictions by taking the non-overlapping parts into consideration. It is also more straightforward to threshold. Within this framework, the event fractions can be thought of as standard confusion matrix labels that were generalized to a continuous domain. The true positive (TP) is equal to an overlap, and false negative (FN) and false positive (FP) correspond to the non-overlapping parts. The true negative (TN) is omitted because successfully localizing "nothing" is not the objective of event detection and its inclusion would give a skewed impression of the annotations' matches. An m-class detection is an (m+1)-class classification and localization problem in which the classes are $\{1,...,M\}$ or "nothing". This is a feature of event detection. Since multi-class detection is very common in SDB settings, e.g. scoring of multiple apnea and hypopnea types, the protocol was designed for both single- and multi-class detection. We first focus on the theoretical aspects of event detection, and then describe the implementation of the discussed ideas.

A. Respiratory-Related Event Detection

It is assumed that the reference signal is an event sequence $Ref = (RE^{(1)}, ..., RE^{(K)})$ of length K produced by an oracle, and the hypothesis is an event sequence $Hyp = (HE^{(1)}, ..., HE^{(L)})$ of length L that was a result of manual or automatic annotation. Figure 1 illustrates several examples of overlapping events, where the blue and red rectangles bound hypothesis events (HEs) and reference events (REs) respectively, with the hatched area showing their temporal overlap. The question is: "Which of the examples represent a successfully detected event?". The answer requires considering two separate issues. First, how many and which HEs can be matched with a single RE and vice versa. We refer to this as event alignment and there are two meaningful cases to explore: one-to-one, and one-to-many. Secondly, what degree of an overlap is required, which also branches into two cases: any non-zero value, and a specific minimum value is required.

The primary consideration in determining which overlap and alignment methodology is appropriate should be based on the intended application. To achieve this we defined three algorithms, each of which corresponds to one of the three applications in which the detection of SDB-related can assist in: 1) event number estimation, 2) total event duration estimation, and 3) joint event number and duration estimation.

A secondary consideration is the nature of event definitions: that is if there is a well defined start, end, and if there are restrictions for a minimum or maximum duration. Here we rely on the latest American Association of Sleep Medicine (AASM) definitions in their manual for the Scoring of Sleep and Associated Events v2.6. The essential issue with poorly defined boundaries is that event placement becomes dependent on the preferences of the scorer, which creates questions about what is an acceptable inter- and intra-scorer match. Apneas

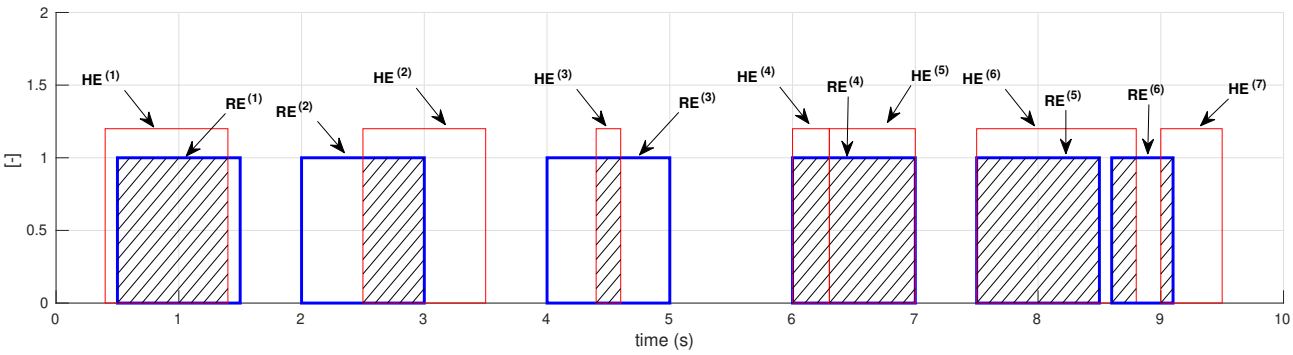

Fig. 1. Examples of various events with differing onset, duration, and alignment. The blue rectangles bound reference events $\boldsymbol{RE}^{(k)}$ while red rectangles bound hypothesis events $\boldsymbol{HE}^{(l)}$. The hatched area is the overlap. The question is which should be classified as a successful detection?

and hypopneas have well defined boundaries tied to signal values, RERAs are well defined but not tied to signal values, however at this time there is no definition for snoring. While there is no minimum defined duration for snoring; apneas, hypopneas and RERAs have a minimum defined duration of 10 seconds. None of these events have a definition for maximum duration. Clearly events that are too short need to be filtered out prior to the application of the protocol in order to conform to minimum duration rules, and the absence of a maximum duration condition generally plays no role, except perhaps in detecting signal corruption. However what can be defined as "too short" may be an open question. We believe these issues are best handled outside the annotation evaluation.

To highlight how these considerations impact this protocol, we examine each algorithm in terms of whether it has a clinical or a research use, what defines a success and an error, what is the correct approach to assess a match between two annotations, and what information the evaluation provides.

1) Algorithm 1 - Evaluation of Event Number Estimation: Clinical sleep medicine recognizes several indexes that measure the number of events per hour in order to diagnose SDB severity. For example, the apnea-hypopnea index (AHI) is used to diagnose sleep apnea severity, on a spectrum that labels patients with an AHI < 5 as normal, and patients with an AHI ≥ 30 as suffering from a severe disorder [35]. Counting the number of snores or RERAs has an analogous use in estimating the snore index (SI) [36] or respiratory disturbance index. A precise event number estimation can be achieved by correctly detecting the events' presence, but their precise boundary placement is of less importance. All events contribute toward the index regardless of their duration. It is only important that there is some overlap to confidently associate REs to HEs. A typical example is the $(HE^{(3)}, RE^{(3)})$ pair shown in Fig. 1. The events have a low overlap but there is little doubt the technologists aimed to score the same event. Such an annotation can be used for unsupervised or semi-supervised AI training for which it is expected the boundaries are automatically localized during training. In general, this evaluation is applicable whenever the objective is to detect the presence of events irrespective of their duration.

An evaluation for the purpose of an event number estimation is best done by using a one-to-one alignment and any non-zero overlap. This penalizes placing multiple events within a boundary of one, but does not penalize loose boundary placement. All aligned events of the same type represents a success. An error is an event that has no overlap with any other event, aligned events of differing types, or an event that fails to be aligned because its potential match was already aligned. This algorithm is referred to as presence evaluation (PE).

2) Algorithm 2 - Evaluation of Total Event Duration Estimation: The established practice of relying on the per hour indexes for SDB diagnosis has its limitations. Authors in [37] raise a concern current clinical practice does not distinguish between events which last 1 second and 10 seconds, figuratively speaking. For example, the literature agrees that the AHI is a poor predictor of risks associated with SDB such as cardiovascular problems [38] [39], quality of life [40], overall mortality [41], and many others [35] [42] [37]. The SI is also recognized as a weak predictor of disturbed sleep structure [43], perceived annoyance [44], or its relationship to the AHI [45], [46]. For this reason, scientists have started to advocate for using complementary measures, one of which is the ratio of total event duration with respect to sleep duration [47] [48]. The proposed measures are based on an assumption that adverse health outcomes may be positively correlated with the relative duration of breathing cessations, the resulting oxygen desaturation, or a combined apnea-hypopnea duration. A precise total event duration ratio can be estimated independently of the number of detected events. Multiple events can still be placed within a boundary of one, but their combined durations have to match and be perfectly overlapped. Typical examples of this are the $(RE^{(4)}, HE^{(4)})$ and $(RE^{(4)}, HE^{(5)})$ pairs from the Fig. 1.

Total event duration evaluation is best done by using a one-to-many alignment and any non-zero overlap. This penalizes imprecise boundary placement but does not penalize if multiple events are placed within a boundary of one. The duration of an overlap in seconds between any two events of the same type represents a success. An error is the duration of any non-overlapping event part, an overlap between events of differing types, or the duration of an unmatched event. This algorithm is referred to as duration evaluation (DE).

3) Algorithm 3 - Evaluation of Joint Event Number and Duration Estimation: The duration of individual SDB events is

another alternative to per-hour indexes. For example, authors in [41], [49] found that the durations of individual apneas and hypopneas can predict an all-cause mortality better than the AHI. Authors in [50] found that a composite sleep and pulmonary phenotype that contained apnea and hypopnea durations can predict hypertension. A precise individual event's duration can be estimated only by correctly detecting the presence of all events within a study and accurately localizing their boundaries. This annotation also has an important ML utility as it can be used for supervised AI training.

Evaluation for the purposes of joint event presence detection and duration estimation is best done by using a one-to-one alignment and a specific minimum overlap. This setup penalizes an imprecise boundary localization and placing multiple events within a boundary of one. Only aligned events of the same type that overlap by a defined degree represent a success. Errors are aligned events that do not pass the overlap threshold, any aligned events of differing types, an event that fails to align because its potential match was already aligned, or an event that has no overlap with any other event. This algorithm is referred to as presence and duration evaluation (PDE).

B. Implementation

The algorithms performs a series of passes over the scored annotations independent of each other as follows: 1) calculate the absolute overlap and identify event type match for all event pairs, 2) find an optimal alignment between the REs and HEs and associate aligned event pairs with decision labels, and 3) calculate chosen performance measures. Steps 1) and 3) are shared between all algorithms and thus discussed outside their descriptions. The protocol does not address cases when multiple events naturally overlap but are supposed to be scored separately, i.e. recordings of a subject and their partner snoring at the same time, as this is a deeper problem. It also assumes the events are of nonzero and finite duration and that the events are chronologically ordered. We advocate for using hit (H), miss (M), false alarm (FA), or confusion (C) decision labels instead of the more traditional TP, TN, FP, and FN labels, which are not suitable for multi-class event detection.

Algorithms performing the described protocol were implemented in Python 3.7, and can be downloaded from the Reykjavik University Sleep Institute website[1]. This repository also contains example Ref and Hyp annotations derived from both real-life and synthetic data and the algorithm outputs.

1) Step 1 - Calculate the Event Overlap and Identify Event Type Match: The first step is to calculate the absolute overlap in seconds, $o(RE^{(k)}, HE^{(l)})$, between all event pairs by using (1), and store the values in a matrix. Events that do not overlap with any other event have their overlap set to 0. The non-overlapping parts are referred to as a complement, $cmp(RE^{(k)}, HE^{(l)})$ and can be calculated using (2). The Sørensen-Dice coefficient (SDC) is then calculated by using (3) and the results are stored in a matrix. The on, off, and dur subscripts represent an onset, offset and duration of an event. The SDC is a measure of relative overlap. The output is an overlap matrix where $O_{k,l} = o(RE^{(k)}, HE^{(l)})$, an SDC matrix where $SDC_{k,l} = sdc(RE^{(k)}, HE^{(l)})$, and an identity matrix where $I_{k,l}$ = 1 if the event types match, and 0 if not.

$$o(RE^{(k)}, HE^{(l)}) = max(0, min(RE^{(k)}_{off} - HE^{(l)}_{off}) - max(RE^{(k)}_{on} - HE^{(l)}_{on})) \quad (1)$$

$$cmp(RE^{(k)}, HE^{(l)}) = RE^{(k)}_{dur} - o(RE^{(k)}, HE^{(l)}) + HE^{(l)}_{dur} - o(RE^{(k)}, HE^{(l)}) \quad (2)$$

$$sdc(RE^{(k)}, HE^{(l)}) = 2 * \frac{o(RE^{(k)}, HE^{(l)})}{RE^{(k)}_{dur} + HE^{(l)}_{dur}} \quad (3)$$

2) Step 2a - Duration Evaluation: The evaluation for the purpose of total event duration estimation uses the O and I matrices to directly assign event fractions with one of the decision labels. It is not necessary to find an optimal alignment because the application allows one-to-many alignment and any overlap represents a success. A hit is equivalent to the overlap between two events of the same type, an overlap between different event types is a confusion, a complement to an RE is a miss, and a complement to an HE is a false alarm. Flooring the overlap at 0 ensures that the total H, M, FA, and C values can be directly computed from the O and I matrices by using equations (4), (5), (6), and (7) respectively.

$$H = \sum_{k=1}^{K} \sum_{l=1}^{L} O_{k,l} * I_{k,l} \quad (4)$$

$$M = \sum_{k=1}^{K} (HE^{(k)}_{dur} - \sum_{l=1}^{L} O_{k,l}) \quad (5)$$

$$FA = \sum_{l=1}^{L} (RE^{(l)}_{dur} - \sum_{k=1}^{K} O_{k,l}) \quad (6)$$

$$C = \sum_{k=1}^{K} \sum_{l=1}^{L} O_{k,l} * (1 - I_{k,l}) \quad (7)$$

3) Step 2b - Presence Evaluation: The evaluation for the purpose of event number estimation takes the SDC and I matrices and finds an optimal event alignment that maximizes the total SDC at a sequence-level. Our implementation relies on finding all contiguous regions within the distance matrix. A contiguous region is defined as the largest possible non-empty set A of event indexes (k, l) such that $[\forall (k, l) \in A][\exists (m, n) \in A]$ such that $SDC_{k,l} > 0$, and $SDC_{m,n} > 0$, and (k, l) directly neighbours (m, n) along either of the axes, or $(k, l) = (m, n)$. The next step is to search within A for an index with a maximum SDC value, save it into a set of aligned indexes B, remove it from A, and then find and remove all competing indexes from A. A competing index is defined as an index that attempts to create an invalid alignment with any event that has already been aligned. The search-and-remove process continues until A is empty and the alignment moves on to the next region. This divide-and-conquer approach simplifies the problem because the exhaustive search-and-remove step is performed on a region that is in most practical cases only a small sample of all possible event pairs.

[1]Link will be made available upon publication.

The list B contains event pairs that are labelled as a hit if their type matches and a confusion otherwise. Misses are unaligned events from the reference sequence and false alarms are unaligned events from the hypothesis. Their total number can be calculated as $M = (||Ref|| - H - C)$ and $FA = (||Hyp|| - H - C)$, where the $||.||$ denotes the sequence length.

4) Step 2c - Presence and Duration Evaluation: The evaluation for the purpose of joint event presence and duration estimation takes the SDC and I matrices and finds an optimal event alignment in exactly the same manner as we proposed for the PE algorithm. The event pairs in the alignment list are labelled as a hit if their type matches and their $sdc(.)$ value exceeds the defined threshold, and as a confusion if their type does not match but their the value still exceeds the defined threshold. All unaligned events from the reference set, or aligned ones that did not pass the threshold, are labelled as a miss, and false alarms are labelled in an analogous fashion. The recommended minimum SDC threshold is $sdc_{thr} = 2/3$, as it corresponds to a balanced overlap and complement. The interpretation of this condition is that the annotations are required to agree to a higher degree than to disagree. Higher values means the overlap dominates, whereas lower value means the complement dominates.

5) Step 3 - Calculate Performance Measures: The final step is to calculate performance measures to quantify the performance. The H,M,FA,C labels are well suited for calculating the F1-score which is preferred by spindle detection and many ML fields. Its main advantage is the independence of which annotation is a reference and hypothesis and the omission of the TN from the formula. Its main disadvantage is its per-class definition, which makes its application for a multi-class detection more cumbersome. For these reasons, we argue for using the F1-score for a single-class detection or when both annotations are conceptually a hypothesis.

An alternative is to define an error rate as a ratio of the total error $(M+FA+C)$ with respect to the reference length $(H+M+C)$. This measure is preferred by various natural language processing fields [25]. Its main advantage is that one measure quantifies an agreement even for a multi-class detection, but the operator must decide which sequences are the reference. It is worth mentioning that if there are more errors than the reference length, then the error rate > 1.

Cohen's kappa κ_c is a popular classification measure used to quantify inter- and intra-raters agreement. Extending it to a detection task requires solving the missing data issue. The authors in [51] analyzed three variants of κ_c that can deal with the issue: 1) remove all missing data, 2) use Gwet's kappa κ_g, and 3) use regular category kappa κ_r. The problem with removing missing data is that the removed data corresponds to misses and false alarms which leaves only hits and confusions to quantify an agreement. If the task is a single-class detection, then the κ_c is either 1 or 0. The κ_g suffers from the same problem, but its value for a a single-class detection is undefined $(\frac{0}{0})$. The κ_r appropriately treats data with one missing value as disagreements, but it includes data with both missing values into the agreement calculation, which skews the results by identifying "nothing". Due to these reasons we argue for using F1-score or error rates instead.

6) Final Notes: The crux of the protocol lies in using the SDC as the temporal coincidence measure, the search-and-remove optimization, and having three evaluation algorithms for different applications. If precision and recall are generalized to a continuous domain using the complement definition (2) as an explicit sum of FP and FN fractions, then a straightforward manipulation shows that $sdc(.)$ corresponds to an F1-score between the events at a continuous level. The produced alignment represents a sequence-level maximum of the F1-score. However, the same alignment would be obtained if the protocol used JI, as SDC and JI can be expressed using one generalized formula. If the F1-score is used as the final performance measure, then the value we obtain by using the protocol is the maximum that can be obtained while maintaining the one-to-one alignment condition.

The difference between this protocol and the SSD is that SSD finds an optimal alignment on an event-by-event basis, which leaves a possibility that some events will remain unaligned when compared to the sequence-level alignment we propose. As a consequence, all performance measures obtained by using the SSD will be equal or lower than the values obtained by using our approach. SSD has no equivalent to the DE algorithm. Also, we argue that using the 0 threshold is the correct setup to perform event presence detection evaluation, as opposed to using a non-zero threshold.

III. Snoring Annotation Agreement Analysis

This protocol can be applied to a dataset that contains at least two concurrent annotations of SDB events. The National Sleep Research Resource (NSRR) is a public repository of sleep studies [52] but none of the studies that are made available meet these criteria. The Wisconsin Sleep Cohort [53] contains singular annotation of SDB events. The Munich-Passau Snore Sound Corpus [54] contains re-scored segments of individual snores that was designed for the purpose of identifying obstructions location and not snore events detection. Consequently, the protocol is at first applied on a toy dataset to demonstrate its utility and then on internal datasets of snore events to assess the reliability of snores annotations. Regardless, we think that the protocol is also applicable to apneas, hypopneas, and RERAs. The agreement is quantified using the F1-scores and error rates. The PDE threshold was set to $sdc_{thr} = 2/3$, which corresponds to the JI threshold of $ji_{thr} > 1/2$, which was used for the SSD. In addition, we assess the agreement using the SSD approach under equivalent conditions to compare the two protocols. The observed snore reliability assessment is compared against reported reliability of annotating sleep spindles using the Montreal Archive of Sleep Studies (MASS) [55] and the DREAMS corpus [56].

A. Results and Discussion

1) Example Toy Data: Fig. 2 illustrates the protocol's behavior on the examples from Fig. 1. The plot contains the SDC matrix and the decision labels for the PE and PDE algorithms and their equivalents by using the SSD protocol. The rows correspond to REs, the columns to HEs, and the overlapping events have their cells highlighted in yellow. The simplified

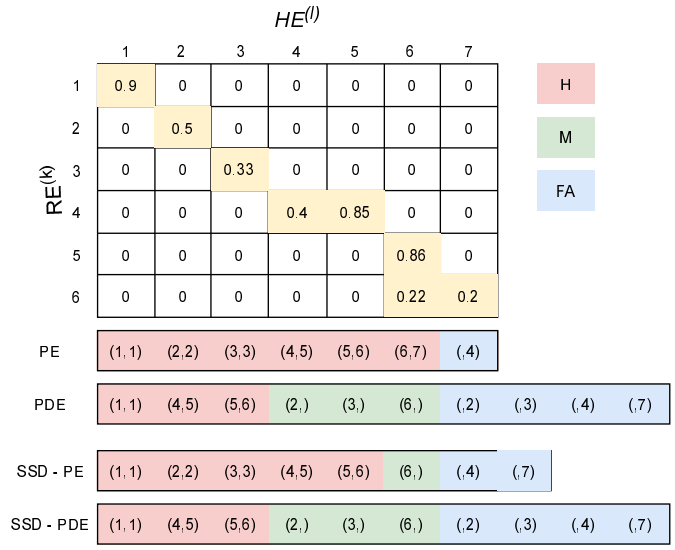

Fig. 2. Demonstration of the protocol on the examples from Fig. 1. The figure plots the distance matrix with non-zero values highlighted. Decision labels for the presence (PE) and presence and duration evaluation (PDE) are below, together with their sleep spindle detection (SSD) equivalents. The hits are in red and use the (k, l) notation. The misses and false alarms are in green and blue, and use the sparse index notation, $(k,)$ or $(, l)$, which reflects a lack of an event to align with.

TABLE I

Agreement between the examples from the Fig. 1. Calculated using the detection (DE), presence (PE), and presence and detection evaluation (PDE) as per this protocol and sleep spindle detection (SSD).

	DE		PE		PDE	
	F1	error	F1	error	F1	error
This protocol	0.74	0.49	**0.92**	0.16	0.46	1.16
SSD	-	-	**0.76**	0.5	0.46	1.16

(k, l) notation is used for brevity. The hits are colored in red and marked as (k, l). The misses and false alarms are in green and blue, and use a sparse marking $(k,)$ and $(, l)$, that reflects the lack of an event to match with. It was assumed all events are of the same type. Table I summarizes the F1-scores.

The main difference occurs in the region $\{(5, 6), (6, 6), (6, 7)\}$ for the PE. Both approaches will select $(5, 6)$ at first and then block $(6, 6)$ to adhere to the one-to-one rule. Our approach will then select $(6, 7)$ despite $sdc(6, 6) > sdc(6, 7)$, since it maximizes the sequence-level F1-score. The indexes $(5, 6)$ and $(6, 7)$ are labelled as hits, $(6,)$ is a miss, and the PE F1-score = 0.92. On the other hand, the SSD will eliminate $(6, 7)$ because $sdc(6, 6) > sdc(6, 7)$. The index $(5, 6)$ is labelled as a hit, $(6,)$ is a miss, $(, 7)$ is a false alarm, and the PE F1-score = 0.76.

2) Snore Annotation based on AASM: This analysis data comes from a dataset which was previously used to compare sensors in terms of their suitability for annotating snoring [57]. The study was approved by the National Bioethics Committee and Data Protection Agency of Iceland, protocol no. 10-048, on June 26, 2018. The studies were annotated by two sleep technologists, who are referred to as "SA" and "SB", and who marked the onset and offset of snore events in the same 2-h long segments for 10 patients. The article reported an inter-scorer's correlation of 0.966. The sleep technologist SB then annotated different, 1-h long segments from 5 patients twice, and the reported intra-scorer correlation was 0.99.

TABLE II

Summary of the number of annotated snore events, their mean and variance ($\mu \pm \sigma$) duration (s), and the total period duration for the SA-18h, SB-18h, SB-4h.1, and SB-4h.2 annotations.

	SA-18h	SB-18h	SB-4h.1	SB-4h.2
snore events	6831	8106	2720	2754
snore dur. (s)	1.2±0.8	0.9±0.5	1.0±0.5	1.3±0.5
period dur.	18h		3h 58m	

TABLE III

An inter-scoring agreement between the SA-18h and SB-18h annotations. Calculated using the detection (DE), presence (PE), and presence and detection evaluation (PDE) as per this protocol and sleep spindle detection (SSD).

	This protocol			SSD	
subject	DE	PE	PDE	0	1/2
1	0.77	**0.98**	0.84	0.98	0.84
2	0.51	**0.51**	0.35	0.51	0.35
3	0.54	0.58	0.49	0.58	0.49
4	0.85	0.92	0.81	0.92	0.81
5	0.73	0.58	0.50	0.58	0.50
6	0.52	**0.90**	0.39	**0.89**	0.39
7	0.95	0.9	0.88	0.91	0.88
8	0.96	0.96	0.93	0.96	0.93
9	0.97	0.97	0.96	0.97	0.96
avg(.)	**0.74**	**0.88**	0.74	0.88	0.74
avg(1,4,7,9)	0.87	**0.96**	0.88	0.96	0.88

One subject recorded only 15 snore events and was removed from the analysis, leaving 9 and 4 subjects to calculate an inter- and intra-scorer's agreement. The total duration of the 2-h long segments from 9 subjects was 18h, and this analysis period is referred to as "18h". The total duration of the 1-h long segments from 4 subjects was 3h 58m, but to distinguish between the first and second pass, these periods are referred to as "4h.1" and "4h.2", respectively. To fully identify the annotator and the analysis period, the article refers to them as "SA-18h", "SB-18h", "SB-4h.1", and "SB-4h.2". Their detailed information is summarized in Table II.

The results for the inter-scorer's agreement analysis between the SA-18h vs. SB-18h annotations is summarized in the Table III. The average F1-score for subjects {1,4,7,9} is included to allow comparison against the SB-4h.1 vs SB-4h.2 intra-scoring agreement analysis which was limited to only these subjects. The PE F1-score ranged from 0.51 to 0.98, indicating there were significant differences in the acoustic representation of snore events among subjects and their perception by the technologists. Overall the mean PE F1-score = 0.88, which indicates that current snore annotation practices and definitions are sufficient for counting snore events. The average DE and PDE F1-score was 0.74 which means there was a lower agreement on boundary placement and it opens a

question of how useful these annotations are for event duration estimation. There was also a small difference between the PE F1-scores between this protocol and SSD for subject 6, which demonstrates that the discussed difference in the alignment methodology influences real annotations.

TABLE IV
An intra-scoring agreement between the SB-4h.1 and SB-4h.2 annotations. Calculated using the detection (DE), presence (PE), and presence and detection evaluation (PDE) as per this protocol and sleep spindle detection (SSD).

	This protocol			SSD	
Subject	DE	PE	PDE	PE	PDE
1	0.68	0.99	0.61	0.99	0.61
4	0.92	0.94	0.90	0.94	0.90
5	0.87	0.92	0.85	0.92	0.85
9	0.84	0.96	0.91	0.96	0.91
avg(.)	0.81	**0.96**	0.80	0.96	0.80

The intra-scorer's analysis between SB-4h.1 and SB-4h.2 is summarized in Table IV, using the same methodology as before. The primary observation was that employing one sleep technologist did not improve the snore annotation consistency. The mean intra-rater's F1-score for the PE was 0.96, which was exactly the same as the inter-rater's for these subjects. The mean F1-scores for DE and PDE decreased.

3) Snore Annotation based on Internal Rules: The previous analysis indicated that a lack of a definition of snore events in the AASM manual negatively affected the boundary placement consistency and decreased the DE and PDE scores. Our secondary concern was that human scoring does not operate at a granularity of hundreds of microseconds, when processing audio signals, as machines are able to. To take these concerns into consideration and to explore the ceiling for human performance, a two-phase annotation protocol with consistent annotation rules was carried out. Both phases included re-annotating the previous dataset with more precise annotation rules, with a 6 months gap in between. A snore event was defined as in [57]; a sound which is synchronous with breathing, protuberant from the background and has an audible oscillatory component. This definition was congruent with definitions that focus on snoring acoustics [58], [59] and snoring mechanics [60], [61]. All snore events were annotated cycle by cycle and given the same label, regardless of whether they contained only the inspiration or expiration part, or spanned the whole breathing cycle. All non-snore events were excluded, e.g. catathrenia events (groaning), loud breathing or other environmental sounds. Snore events were annotated by looking at a 10 (s) long window, but when necessary a 1 (s) long window was used, from the onset of the oscillation sound-wave towards the end. In case of mixed events, i.e. a mixture of loud breathing and a snore event, only a distinct snore wave was annotated. Therefore, much more attention was paid to the event onset and offset than in the previous scoring of the data.

In phase 1, a snore event saturated hour was chosen from each sleep study and the events were annotated according to the aforementioned rules. Six months later, in phase 2, one study was randomly chosen as a training study (subject 5) and the remaining 8 sleep studies were annotated again. The sleep technologist was blinded to the previous annotation while re-annotating. The total amount of annotated data was 8h 53m. The technologist for this task is denoted as "SC", and the analysis periods are referred to as "8h.1" and "8h.2" to distinguish between the phases. The same naming convention was used to distinguish between the annotations: "SC-8h.1", "SC-8h.2". Table V summarizes the annotation statistics.

TABLE V
Summary of the number of annotated snore events, their mean and variance ($\mu \pm \sigma$) duration (s), and the total period duration for the SC-8h.1 and SB-8h.2 annotations.

	SC-8h.1	SC-8h.2
snore events number	5597	5912
snore dur. $\mu \pm \sigma$ (s)	1.2±0.6	1.0±0.5
analysis period dur.	8h 53m	

The intra-scoring agreements are summarized in the Table VI. The sleep technologist was much more consistent in annotating snore events boundaries. The mean DE F1-scored improved to 0.9. The PE F1-score improved to 0.94 and the PDE F1-score to 0.88. In general, values around 0.9 are considered a very good match, indicating these annotations were suitable for all discussed clinical and research related applications. An in-depth look on a per-subject basis revealed a notable trend for Subject 1. The sleep technologist achieved nearly perfect PE (F1-score = 0.99) but only mediocre DE (F1-score = 0.79), which meant that there was no doubt which events were snore events and which were not, but their boundaries were much harder to localize consistently.

B. Snore Index Estimation and its Variability

The analyses demonstrated that sleep technologists achieved various levels of consistency in identifying presence of snore events. To explore the F1-score relation to an annotation quality for snore event counting, the SI was calculated for each individual annotation. In addition, the index was calculated from the consensual annotation (Cons.), when only

TABLE VI
An intra-scoring agreement between the SC-8h.1 and SC-8h.2 annotations. Calculated using the detection (DE), presence (PE), and presence and detection evaluation (PDE) as per this protocol and sleep spindle detection (SSD).

	This protocol			SSD	
Subject	DE	PE	PDE	PE	PDE
1	**0.79**	**0.99**	0.79	0.99	0.79
2	0.78	0.84	0.78	0.84	0.78
3	0.97	0.97	0.97	0.97	0.97
4	0.97	0.97	0.97	0.97	0.97
6	0.80	0.84	0.80	0.84	0.80
7	0.86	0.88	0.86	0.88	0.86
8	0.93	0.95	0.93	0.95	0.93
9	0.91	0.97	0.91	0.97	0.91
avg(.)	0.90	0.94	0.88	0.94	0.88

TABLE VII

The snore index (-/h) estimation based on each individual annotation and their consensus (Cons.), when only the repeatedly identified events counted towards the consensus index calculation. Each annotation has its onset summarized.

	Subject								
	1	2	3	4	5	6	7	8	9
onset	01:00	01:00	01:00	01:00	01:00	01:00	01:00	01:00	01:00
SA-18h	761	346	436	382	134	**575**	293	436	689
SB-18h	760	131	188	337	55	**587**	250	426	681
Cons.	749	122	184	334	55	**526**	247	414	669
onset	00:44	-	-	23:44	-	-	01:30	-	05:14
SB-4h.1	896	-	-	635	-	-	511	-	805
SB-4h.2	895	-	-	703	-	-	487	-	770
Cons.	895	-	-	630	-	-	463	-	764
onset	00:44	01:39 / 03:51	01:50 / 05:39	23:44	-	02:22 / 07:05	01:30 / 03:29	02:16 / 04:50	05:14
SC-8h.1	895	329	577	703	-	592	454	748	770
SC-8h.2	894	395	565	709	-	721	560	743	785
Cons.	894	304	559	692	-	557	449	714	756

the repeatedly identified events were counted. The question was how significant were the differences. The values for each subject and annotations are summarized in Table VII.

Annotation quality is often compared using the SI statistics which have confusing interpretation value. The proposed protocol accumulates misses and false alarms together. Operating directly on a per-hour index assumes the errors have a compensatory quality where misses and false alarms cancel each other out. Achieving a perfect intra- or inter-score becomes a matter of making an equal number of errors, rather than not making any. Such an analysis reports on a systemic bias of one sleep technologist over another, and not on an agreement in identifying event presence. This behavior was best demonstrated by looking at SA-18h and SB-18h annotations for Subject 6. The sleep technologists achieved a near perfect SI match (575 vs. 587 snores/hr), but the number of repeatedly marked events was only 526. This means a SI calculated from the consensus was 8-10% lower then from either individual annotation, and 9% lower than their mean. Likewise, their disagreement was higher then the snore index variance would imply. The example practically demonstrated that the proposed protocol is superior to per-hour index variability analysis.

Table VII contains the analysis onset for each subject and annotation. All ”18h” periods started at 1AM, but the periods for ”4h” were randomly chosen, and the periods for ”8h” analysis were chosen based on event saturation, sometimes even splitting the period into two. The table reports on both timestamps in these cases. It was likely the observed differences across annotations for the same subject were affected by a natural event saturation during the night, and not only due to sleep technologist inconsistencies.

Spindle detection provides a good reference to compare the observed reliability results against. The F1-scores reported in the literature range from 0.61 to 0.67 for inter-scoring [28], [27], [29], and 0.72 for intra-scoring [28]. The cited works use a very low Jaccard threshold ji_{thr} = 0.2, which corresponds to the SDC threshold sdc_{thr} = 0.28. Our analysis used sdc_{thr} = 2/3. We observed F1-scores of 0.74 in the inter-scoring setup, and after we created additional snore annotation rules, the values increased to 0.88 for intra-scoring. There were multiple likely factors at play that explain the superior intra-scoring agreement. First, there is a common recognition of what snoring sounds like, which ameliorates its lack of an exact definition. Second, the sleep technologist could reinforce the visual cues by a simultaneous listening, whereas spindle annotation relies on visual cues only. Third, an acoustic signal recorded in a lab environment is rather clean, whereas EEG is by nature a more noisy signal.

IV. Conclusion

This paper proposes a protocol to evaluate an agreement in the task of detecting SDB events in a sleep study. While the field of SDB event detection is well researched, it lacks a generally accepted protocol to compare competing AI solutions or to calculate agreement between human scorers. The article also studied the reliability of manual annotation of snore events in private datasets that were annotated according to the AASM manual and internally developed rules. We demonstrated that a lack of gold standard rules for annotating snore events has a relatively small effect on the ability of human scorers to detect the presence of snore events. The absence of rules however, had a more pronounced and negative effect on localizing event boundaries. We also demonstrated that trained sleep technologists can achieve much better intra- and inter-scorer’s agreement in annotating snore events than is reported for annotating sleep spindles.

Even if many events will likely always remain in the ”gray zone”, the protocol offers opportunities to study human-to-human scoring consistency, to determine what is an acceptable AI performance, or what ML architectures are suitable to automate the task of sleep study annotation. We recognize that a low number of subjects is a limitation of our analysis. The number of annotated events is over 10,000. Sleep scoring is an extremely time intensive activity when done manually, we hope that this algorithm will assist other sleep centres with their development of automated analyses, and this will also allow the algorithm to be tested on more subjects gathering more data on its performance.

REFERENCES

[1] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. D. Clifford, "A review of signals used in sleep analysis," *Physiol. Meas.*, vol. 35, no. 1, pp. R1–57, Jan. 2014.

[2] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. M. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *The Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687 – 698, 2019.

[3] F. Mendonça, S. S. Mostafa, A. G. Ravelo-García, F. Morgado-Dias, and T. Penzel, "A review of obstructive sleep apnea detection approaches," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 825–837, 2019.

[4] M. Bianchi, K. Russo, H. Gabbidon, T. Smith, B. Goparaju, and M. B. Westover, "Big data in sleep medicine: prospects and pitfalls in phenotyping," *Nature and Science of Sleep*, vol. Volume 9, pp. 11–29, Feb. 2017.

[5] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G. Yang, "Big data for health," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1193–1208, 2015.

[6] J. R. Stradling and J. H. Crosby, "Predictors and prevalence of obstructive sleep apnoea and snoring in 1001 middle aged men." *Thorax*, vol. 46, no. 2, pp. 85–90, 1991. [Online]. Available: https://thorax.bmj.com/content/46/2/85

[7] Y. Endeshaw, T. B. Rice, A. V. Schwartz, K. L. Stone, T. M. Manini, S. Satterfield, S. Cummings, T. Harris, M. Pahor, and for the Health ABC Study, "Snoring, Daytime Sleepiness, and Incident Cardiovascular Disease in The Health, Aging, and Body Composition Study," *Sleep*, vol. 36, no. 11, pp. 1737–1745, 11 2013. [Online]. Available: https://doi.org/10.5665/sleep.3140

[8] D. Li, D. Liu, X. Wang, and D. He, "Self-reported habitual snoring and risk of cardiovascular disease and all-cause mortality," *Atherosclerosis*, vol. 235, no. 1, pp. 189–195, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021915014002317

[9] M. Cavusoglu, M. Kamasak, O. Erogul, T. Ciloglu, Y. Serinagaoglu, and T. Akcam, "An efficient method for snore/nonsnore classification of sleep sounds," *Physiological Measurement*, vol. 28, no. 8, pp. 841–853, jul 2007.

[10] X. Sun, J. Young Kim, Y. Won, J.-J. Kim, and K.-A. Kim, "Efficient snoring and breathing detection based on sub-band spectral statistics," *Bio-medical materials and engineering*, vol. 26, pp. S787–S793, 09 2015.

[11] T. L. Nguyen and Y. Won, "Sleep snoring detection using multi-layer neural networks," *Bio-medical materials and engineering*, vol. 26, pp. S1749–S1755, 09 2015.

[12] E. Dafna, A. Tarasiuk, and Y. Zigel, "Automatic Detection of Whole Night Snoring Events Using Non-Contact Microphone," *PLoS ONE*, vol. 8, no. 12, Dec. 2013.

[13] ——, "Automatic detection of whole night snoring events using non-contact microphone," *PLOS ONE*, vol. 8, no. 12, pp. 1–14, 12 2014.

[14] W. D. Duckitt, S. K. Tuomi, and T. R. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," *Physiological Measurement*, vol. 27, no. 10, pp. 1047–1056, sep 2006.

[15] V. R. Swarnkar, U. R. Abeyratne, and R. V. Sharan, "Automatic picking of snore events from overnight breath sound recordings," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Seogwipo: IEEE, Jul. 2017, pp. 2822–2825.

[16] H. E. Romero, N. Ma, G. Brown, A. V. Beeston, and M. Hasan, "Deep learning features for robust detection of acoustic events in sleep-disordered breathing," in *Proceedings of ICASSP 2019*, 05 2019, pp. 810–814.

[17] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snore-GANs: Improving Automatic Snore Sound Classification With Synthesized Data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 300–310, Jan. 2020.

[18] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 Evaluation," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen and J. Garofolo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4122, pp. 1–44, series Title: Lecture Notes in Computer Science.

[19] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 Evaluation," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer, 2008, pp. 3–34.

[20] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The Rich Transcription 2007 Meeting Recognition Evaluation," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, vol. 4625, pp. 373–389, iSSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *arXiv:1409.0575 [cs]*, Jan. 2015, arXiv: 1409.0575.

[22] V. E. Vivek and N. Sudha, "Robust Hausdorff distance measure for face recognition," *Pattern Recognition*, vol. 40, no. 2, pp. 431–442, Feb. 2007.

[23] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 720–735.

[24] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint Measurement of Localization and Detection of Sound Events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2019, pp. 333–337.

[25] H. Bredin, "pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3587–3591.

[26] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009.

[27] S. C. Warby, S. L. Wendt, P. Welinder, E. G. S. Munk, O. Carrillo, H. B. D. Sorensen, P. Jennum, P. E. Peppard, P. Perona, and E. Mignot, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature Methods*, vol. 11, no. 4, pp. 385–392, Apr. 2014.

[28] S. L. Wendt, P. Welinder, H. B. Sorensen, P. E. Peppard, P. Jennum, P. Perona, E. Mignot, and S. C. Warby, "Inter-expert and intra-expert reliability in sleep spindle scoring," *Clinical Neurophysiology*, vol. 126, no. 8, pp. 1548–1556, Aug. 2015.

[29] K. Lacourse, J. Delfrate, J. Beaudry, P. Peppard, and S. C. Warby, "A sleep spindle detection algorithm that emulates human expert spindle scoring," *Journal of Neuroscience Methods*, vol. 316, pp. 3–11, Mar. 2019.

[30] C. O'Reilly and T. Nielsen, "Automatic sleep spindle detection: benchmarking with fine temporal resolution using open science tools," *Frontiers in Human Neuroscience*, vol. 9, Jun. 2015.

[31] R. Nonaka, T. Emoto, U. R. Abeyratne, O. Jinnouchi, I. Kawata, H. Ohnishi, M. Akutagawa, S. Konaka, and Y. Kinouchi, "Automatic snore sound extraction from sleep sound recordings via auditory image modeling," *Biomedical Signal Processing and Control*, vol. 27, pp. 7 – 14, 2016.

[32] H. J. Davies, T. Nakamura, and D. P. Mandic, "A Transition Probability Based Classification Model for Enhanced N1 Sleep stage Identification During Automatic Sleep Stage Scoring," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, Jul. 2019, pp. 3641–3644.

[33] C. F. Goh, L. B. Samuclsson, M. H. Hall, G. G. Lee Seet, and K. Shimada, "Semi-automatic snore detection in polysomnography based on hierarchical clustering," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, 2018, pp. 1116–1122.

[34] A. K. Ng, T. San Koh, K. Puvanendran, and U. Ranjith Abeyratne, "Snore signal enhancement and activity detection via translation-invariant wavelet transform," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 10, pp. 2332–2342, 2008.

[35] A. Malhotra, I. Ayappa, N. Ayas, N. Collop, D. Kirsch, N. Mcardle, R. Mehra, A. I. Pack, N. Punjabi, D. P. White, and D. J. Gottlieb, "Metrics of sleep apnea severity: beyond the apnea-hypopnea index," *Sleep*, vol. 44, no. 7, 07 2021.

[36] L. B. L. Benoist, S. Morong, J. P. van Maanen, A. A. J. Hilgevoord, and N. de Vries, "Evaluation of position dependency in non-apneic snorers," *European Archives of Oto-Rhino-Laryngology*, vol. 271, no. 1, pp. 189–194, Jan 2014. [Online]. Available: https://doi.org/10.1007/s00405-013-2570-5

[37] D. A. Pevernagie, B. Gnidovec-Strazisar, L. Grote, R. Heinzer, W. T. McNicholas, T. Penzel, W. Randerath, S. Schiza, J. Verbraecken, and E. S. Arnardottir, "On the rise and fall of the apnea-hypopnea index:

A historical review and critical appraisal," *Journal of Sleep Research*, vol. 29, no. 4, p. e13066, 2020.

[38] R. Li, M. Rueschman, D. J. Gottlieb, S. Redline, and T. Sofer, "A composite sleep and pulmonary phenotype predicting hypertension," *EBioMedicine*, vol. 68, p. 103433, Jun 2021.

[39] N. S. Marshall, K. K. Wong, S. R. Cullen, M. W. Knuiman, and R. R. Grunstein, "Sleep apnea and 20-year follow-up for all-cause mortality, stroke, and cancer incidence and mortality in the Busselton Health Study cohort," *J Clin Sleep Med*, vol. 10, no. 4, pp. 355–362, Apr 2014.

[40] S. R. Patel, D. P. White, A. Malhotra, M. L. Stanchina, and N. T. Ayas, "Continuous positive airway pressure therapy for treating sleepiness in a diverse population with obstructive sleep apnea: results of a meta-analysis," *Arch. Intern. Med.*, vol. 163, no. 5, pp. 565–571, Mar. 2003.

[41] M. P. Butler, J. T. Emch, M. Rueschman, S. A. Sands, S. A. Shea, A. Wellman, and S. Redline, "Apnea–hypopnea event duration predicts mortality in men and women in the sleep heart health study," *American Journal of Respiratory and Critical Care Medicine*, vol. 199, no. 7, pp. 903–912, 2019, pMID: 30336691.

[42] S. Tam, B. T. Woodson, and B. Rotenberg, "Outcome measurements in obstructive sleep apnea: beyond the apnea-hypopnea index," *Laryngoscope*, vol. 124, no. 1, pp. 337–343, Jan. 2014.

[43] K. E. Macarthur, T. D. Bradley, C. M. Ryan, and H. Alshaer, "Dissociation between objectively quantified snoring and sleep quality," *Am. J. Otolaryngol.*, vol. 41, no. 1, p. 102283, Jan. 2020.

[44] R. Fischer, T. S. Kuehnel, V. Vielsmeier, F. Haubner, S. Mueller, and C. Rohrmeier, "Snoring: is a reliable assessment possible?" *Eur. Arch. Otorhinolaryngol.*, vol. 277, no. 4, pp. 1227–1233, Apr. 2020.

[45] H. Alshaer, R. Hummel, M. Mendelson, T. Marshal, and T. D. Bradley, "Objective relationship between sleep apnea and frequency of snoring assessed by machine learning," *J. Clin. Sleep Med.*, vol. 15, no. 3, pp. 463–470, Mar. 2019.

[46] ——, "Objective relationship between sleep apnea and frequency of snoring assessed by machine learning," *Journal of Clinical Sleep Medicine*, vol. 15, no. 03, pp. 463–470, 2019. [Online]. Available: https://jcsm.aasm.org/doi/abs/10.5664/jcsm.7676

[47] C. Ma, Y. Zhang, J. Liu, and G. Sun, "A novel parameter is better than the AHI to assess nocturnal hypoxaemia and excessive daytime sleepiness in obstructive sleep apnoea," *Sci. Rep.*, vol. 11, no. 1, p. 4702, Feb. 2021.

[48] A. Kulkas, P. Tiihonen, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients having similar apnea–hypopnea index," *Medical & Biological Engineering & Computing*, vol. 51, no. 6, pp. 697–708, Jun 2013.

[49] A. Muraja-Murro, A. Kulkas, M. Hiltunen, S. Kupari, T. Hukkanen, P. Tiihonen, E. Mervaala, and J. Töyräs, "The severity of individual obstruction events is related to increased mortality rate in severe obstructive sleep apnea," *J. Sleep Res.*, vol. 22, no. 6, pp. 663–669, Dec. 2013.

[50] R. Li, M. Rueschman, D. J. Gottlieb, S. Redline, and T. Sofer, "A composite sleep and pulmonary phenotype predicting hypertension," *EBioMedicine*, vol. 68, no. 103433, p. 103433, Jun. 2021.

[51] A. D. Raadt, M. J. Warrens, R. J. Bosker, and H. A. L. Kiers, "Kappa coefficients for missing data," *Educational and Psychological Measurement*, vol. 79, no. 3, pp. 558–576, 2019, pMID: 31105323.

[52] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018.

[53] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. Hla, "Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study," *WMJ*, vol. 108, no. 5, pp. 246–249, Aug. 2009.

[54] C. Janott, M. Schmitt, Y. Zhang, K. Qian, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Snoring classified: The Munich-Passau Snore Sound Corpus," *Computers in Biology and Medicine*, vol. 94, pp. 106–118, Mar. 2018.

[55] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 2014.

[56] S. Devuyst, T. Dutoit, P. Stenuit, and M. Kerkhofs, "Automatic sleep spindles detection — overview and development of a standard proposal assessment method," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 1713–1716.

[57] E. S. Arnardottir, B. Isleifsson, J. S. Agustsson, G. A. Sigurdsson, M. O. Sigurgunnarsdottir, G. T. Sigurdarson, G. Saevarsson, A. T. Sveinbjarnarson, S. Hoskuldsson, and T. Gislason, "How to measure snoring? a comparison of the microphone, cannula and piezoelectric sensor," *Journal of Sleep Research*, vol. 25, no. 2, pp. 158–168, 2016.

[58] D. Pevernagie, R. M. Aarts, and M. De Meyer, "The acoustics of snoring," *Sleep Medicine Reviews*, vol. 14, no. 2, pp. 131–144, Apr. 2010.

[59] C. Janott, C. Rohrmeier, M. Schmitt, W. Hemmert, and B. Schuller, "Snoring - An Acoustic Definition," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, Jul. 2019, pp. 3653–3657.

[60] N. Gavriely and O. Jensen, "Theory and measurements of snores," *Journal of Applied Physiology*, vol. 74, no. 6, pp. 2828–2837, Jun. 1993.

[61] K. P. Strohl, J. P. Butler, and A. Malhotra, "Mechanical properties of the upper airway," *Comprehensive Physiology*, vol. 2, no. 3, pp. 1853–1872, Jul. 2012.