Assessment of obstructive sleep apnea severity using audio-based snoring features

Biomedical Signal Processing and Control 86 (2023) 104942

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Assessment of obstructive sleep apnea severity using audio-based snoring features

Jiali Xie [a], Pedro Fonseca [a,b], Johannes van Dijk [a,c,d], Sebastiaan Overeem [a,c], Xi Long [a,b,*]

[a] Biomedical Diagnostics Lab, Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[b] Philips Research, High Tech Campus, 5656 AE Eindhoven, The Netherlands
[c] Sleep Medicine Center Kempenhaeghe, 5590 AB Heeze, The Netherlands
[d] Department of Orthodontics, Ulm University, Ulm, Germany

ARTICLE INFO

Keywords:
Sleep disordered breathing
Sleep apnea
OSA classification
Apnea-hypopnea index
AHI
Audio signal processing
Machine learning
Snoring
Sound features

ABSTRACT

Background and Objective: Snoring is a prima symptom of obstructive sleep apnea (OSA). Here, we add audio-based snoring features to improve the non-obtrusive assessment of sleep apnea, by estimating the apnea-hypopnea index (AHI) and classifying OSA severity.
Methods: We propose novel features to quantify temporal changes between snores (snore rate variability) and to describe trends in snore energy, based on the assessment of snore sounds from audio signals over the full night. We then combined those features with age, body mass index (BMI) and features described in literature. An extreme gradient boosting algorithm was trained with all these features on AHI estimation. The estimated AHI was then used to classify OSA severity.
Results: Audio-based estimated AHI showed a significant Spearman's correlation with the AHI based on gold-standard polysomnography ($R = 0.786$, $P < 0.0001$). Our results outperformed a model trained with solely previously described features in our dataset ($R = 0.676$, $P < 0.0001$) and a model trained with the combination of previously described features, age, and BMI ($R = 0.731$, $P < 0.0001$). The mean absolute error of AHI estimation was 7.26 events/h. Area under the receiver operating characteristic curve outcomes were 0.90, 0.87 and 0.93 for classifying patients with varying severity separated by the canonical thresholds of 5, 15 and 30 events/h respectively. The accuracy of classifying subjects to four classes (no, mild, moderate, and severe OSA) was 59.3 %.
Conclusion: Additional audio-based snore features can improve the performance of non-obtrusive AHI estimation and OSA severity classification methods.

1. Introduction

Obstructive sleep apnea (OSA) is the most common form of sleep disordered breathing and characterized by episodic partial or complete obstruction of the upper airway resulting in intermittent hypoxia and arousals from sleep [1]. It is a common sleep disorder with prevalence estimates ranging from 6 % to 17 % in the general adult population, which may still increase due to the rise of obesity rates [2,3]. OSA can have significant clinical consequences including daytime hypersomnolence, neurocognitive and metabolic dysfunction and increase risk of cardiovascular diseases [4]. The gold standard for diagnosing OSA and assessing severity is overnight polysomnography (PSG), with the Apnea-Hypopnea Index (AHI) as the primary outcome. However, PSG has several disadvantages including patient inconvenience coming from the attached sensors, an unfamiliar sleep environment, and high expenses from highly trained personnel and technical devices [5], which is only partly solved by simplified versions of the technique [6,7]. Therefore, a comfortable and economical method for OSA monitoring would be advantageous, especially if it can also be used long-term, e.g., for treatment monitoring.

Snoring is one of the most common and earliest symptoms of OSA occurring in 70 % to 95 % of all OSA patients [8,9]. It has been treated as a potential indicator for OSA monitoring by researchers for a long period [1,10]. Snoring sounds can be acquired by a low-cost non-contact microphone, which offers significant advantages compared with PSG and other sensors as it does not affect the patient's sleep quality and

* Corresponding author at: Department of Electrical Engineering, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.
E-mail address: x.long@tue.nl (X. Long).

https://doi.org/10.1016/j.bspc.2023.104942
Received 30 September 2022; Received in revised form 6 March 2023; Accepted 5 April 2023
Available online 21 April 2023

using it can be performed at home without additional cost from sleep specialists. Moreover, the snore signal is relatively easy to obtain, for example with the microphone embedded in modern smartphones. Compared to normal breathing, which can in theory also be recorded by microphones and used to characterize changes in airflow associated with disordered breathing, snoring is much louder and therefore easier to acquire in uncontrolled conditions.

A number of researchers have studied the relationship between snoring and OSA, and used audio signals of snore sounds during the whole night to directly classify subjects in OSA severity groups [11,12], or to directly estimate the AHI [13,14]. Fiz et al. [11] analyzed power spectral density parameters of snoring and classified 37 subjects into three classical severity groups (AHI < 5, 5 <= AHI < 15 and AHI >= 15) using logistic regression models. Their method showed a sensitivity (specificity) of 87 % (71 %) and 80 % (90 %) with AHI thresholds of 5 and 15. Mesquita et al. [12] analyzed the time interval between regular snores and applied a Bayesian classification algorithm to classify 34 subjects with AHI cut-points of 5 and 30, achieving classification accuracies of 88.2 % (with 90 % sensitivity, 75 % specificity) and 94.1 % (94.4 %, 93.8 %) respectively. Dafna et al. [13] extracted time and spectral related snoring features, and selected three from them to train a Gaussian mixture regression model to estimate the AHI. They achieved a Pearson's correlation R of 0.89, an AHI error of 7.35 events/h, and a diagnostic agreement of 77.3 % between OSA and non-OSA on 155 subjects. Ben-Israel et al. [14] developed five acoustic features to measure intra- and inter-snore properties and calculated AHI by a multivariate linear regression model, achieving a coefficient of determination (R^2) between estimated and PSG determined AHI of 0.81. Although these studies yielded promising results, they may not be robust enough yet for OSA severity classification and AHI estimation for clinical diagnosis [10]. One of the main reasons is the lack of additional validation of these methods in other cohorts, using different recording equipment and setups, in settings with different background noise. Besides, we expect between-subject snoring characteristics to vary, beyond the well-known heterogeneity in the OSA condition not captured well with a single AHI index. In fact, varying proportions of apnea and hypopnea events, the presence of central apnea events, as well as comorbidities may all influence snore characteristics. Therefore, it remains unknown if these methods generalize well to different datasets.

This study aims to improve the assessment of sleep apnea, by estimating AHI based on snoring features from the audio signal. Starting from the set of features described by Ben-Israel et al. [14,15], we first explored how the addition of age and BMI impacts AHI estimation performance. Furthermore, we added additional snoring features that may be less sensitive to the audio quality (frequency, amplitude) of recordings, and may thus improve robustness over different recording setups and conditions. We introduce a new concept called snore rate variability (SRV) as a proxy of respiratory rate variability (RRV), from which we derive time, frequency, and non-linear features. In addition, we exploit trends in the energy of snore signals (snore energy trends, SET) and investigate how the combination of these parameters can be used to estimate OSA severity.

2. Materials and methods

2.1. Data

We used a subset of the SOMNIA database [16], which was selected based on the availability of audio recordings and adequate synchronization between audio and PSG signals. Subjects younger than 18 years at the time of PSG, subjects without full night recordings, or patients treated with continuous positive airway pressure therapy were excluded. No further selection was applied in terms of sex, BMI, or AHI. A total of 172 subjects were included. Table 1 summarizes the demographic and OSA characteristics of included subjects.

All subjects underwent routine clinical PSG in the sleep lab of Kempenhaeghe Center for Sleep Medicine, Heeze, the Netherlands, between June and November 2017. Audio signals were recorded with a set of five microphones (Earthworks M23) placed around the patient in a room with low background noise. Two microphones were positioned above the subject's head, with a distance of approximately 70 cm and 130 cm. The third and fourth microphones were placed on the left and right side of the bed, about 60 cm from the center of the bed. Additionally, we placed a fifth microphone on the bedside table, approximately 100 cm from the center of the bed. From this set, only the microphone placed above each subject's head at a distance of 70 cm was used for analyses, because it yielded better snore detection results than the others in our previous study [17]. Time synchronization between audio and PSG signals was achieved by simultaneously recording a sine wave-based code with the audio recording device and with the PSG.

The SOMNIA study was reviewed by the medical ethical committee of the Maxima Medical Center (Eindhoven, the Netherlands. File no: N16.074 and W17.128). Written informed consent was provided by all participants. The study met the ethical principles of the Declaration of Helsinki, the guidelines of Good Clinical Practice, and all current legal requirements. The Review Committee of Kempenhaeghe Center for Sleep Medicine approved the data-request for the current analysis.

Table 1
Subject demographic and diagnostic information.

Male/Female	109/63
Age, years	50.3 ± 14.9 (range: 18–86)
BMI, kg/m²	27.4 ± 4.7 (range: 20.0–45.2)
*AHI, events/hour of sleep**	18.4 ± 18.4 (range: 0–99.6)
*Formal OSA diagnosis (yes/no)**	95/77
AHI < 5	40
5 <= AHI < 15	60
15 <= AHI < 30	41
AHI > 30	31
Number of Snores	1245 ± 1208 (range: 9–5906)
Number of Apneas	10.7 ± 30.0 (range: 0–315, total: 1836)
Number of Hypopneas	95.3 ± 85.3 (range: 0–475, total: 16394)
Number of Mixed apneas	6.0 ± 25.6 (range: 0–254, total: 1025)
Number of Central apneas	8.6 ± 26.4 (range: 0–249, total: 1486)

*AHIs and final diagnosis obtained by diagnostic results provided by Kempenhaeghe Center for Sleep Medicine.
*Sample statistics indicate per-subject mean ± standard deviation, and between parenthesis the range and aggregated totals.

2.2. Methods

Fig. 1 presents a schematic overview of the methods applied. After receiving the audio signal, we first detected all sound events from the

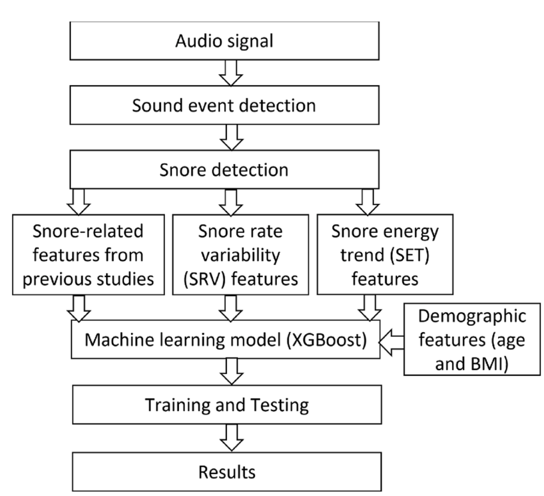

Fig. 1. Schematic overview of the methods used for AHI estimation and OSA severity classification.

audio recordings using the event detection method proposed by Dafna et al. [18]. We then employed the algorithm proposed in our previous paper [17] to classify all sound events into snore versus non-snore events. Detected snore events were further used to compute snoring features for AHI estimation. A combination of previously described features, a set of new features proposed in this work (detailed in the next section), and demographic characteristics, were used as input to a machine learning model for fivefold cross-validation. The complete method was implemented with Python-3.7 and the additional packages 'pyhrv' [19], 'sklearn' [20] and 'xgboost' [21].

2.2.1. Feature extraction

This section introduces the set of features used for AHI estimation and OSA severity classification and it includes 1) snore-related features described in previous studies, 2) newly developed snore rate variability features, 3) newly developed features extracted from snore energy trends, and 4) demographic features.

2.2.1.1. Baseline feature set. All five features described in the studies by Ben-Israel et al. [14,15] were implemented as the baseline feature set for modeling: inter-event silence, running variance, apneic phase ratio, pitch density, and 'Mel-cepstability'. A detailed description of those features can be found in the corresponding work [14,15]. A brief summary is provided here for convenience. Mel-cepstability measures the stability of the spectrum of the entire night. Running variance measures the inter-snore variability of the snore energy over the entire night. Apneic phase ratio presents the relative duration of the upper airway collapse, and it is the relative number of snore groups with energy variance larger than a specific threshold. Inter-event silence counts the number of long silences between snore events. Pitch density represents the stability of the tissue's vibration frequency.

2.2.1.2. Snore rate variability features. We hypothesize that the occurrence of sleep disordered breathing events disturbs the continuing occurrence of snore events. Fig. 2 shows such an example, where an obstructive apnea event and associated cessation of airflow interrupts a sequence of snore events, which resume after the obstruction is resolved. Exploiting the timing variability of snore sounds can thus help characterize the periodicity of the occurrence of these events and help estimate AHI without explicitly detecting apneas and hypopneas. We propose the concept of SRV to characterize the changes (e.g., regularity) of snore events. The idea of SRV was inspired by heart rate variability (HRV), a well-known technique in the analysis of autonomic changes in cardiac activity [22]. HRV characterizes variations in the time intervals between consecutive heartbeats, and in a comparably way SRV measures the variation in the time intervals between consecutive snore events. Similar to HRV, we exploit features that characterize variability in the time, frequency, and non-linear domains.

a. Snore-to-snore intervals

In contrast to HRV where, during sleep, interbeat intervals can mostly be measured continuously, snoring can occur in clusters separated by relatively long intervals, which should not be included in the time series of snore-to-snore intervals (SSIs) used to calculate SRV features. Thus, before SRV features are extracted, the first step is to calculate all SSIs, and then separate clusters of snores based on these distances. By doing this, all intervals longer than a certain threshold are excluded. After that, SRV features are calculated for each cluster. Detailed steps to compute an SSI series are given below.

i) Calculate the SSI between two consecutive snore events such that:
   i. $SSI = ST_{n+1} - ST_n (1)$

where ST_n stands for the middle time of the n^{th} snore event of a subject, and a snore event is a sound event classified as snore using the algorithm described in our previous work [17].

ii) Cluster snore events into different "snore groups" based on the value of consecutive SSIs. A snore group is defined as a group of snore events not separated by an interval larger than a given threshold. The reason to do this is that two consecutive snores can be apart from each other for a relatively long time (e.g., several minutes or even hours), meaning that the patient has simply stopped snoring. The clustering rules are defined as below, and are based on the study of Ben-Israel et al. [14,15].
   - Snore events are assigned to the same snore group if the SSIs are<60 s. The choice of 60 s reflects the fact that 99.5 % of all apnea and hypopnea events are shorter than this duration in our dataset.
   - A snore group must include at least five snore events to ensure sufficient samples per group to compute features.
   - Following the strategy of Ben-Israel et al. [14,15], a snore group cannot include more than twenty snore events, to have enough groups to calculate group-based features.

iii) Calculate SSI series for each snore group. Each snore group with n snore events results in an SSI series with $n-1$ SSIs. The full-night

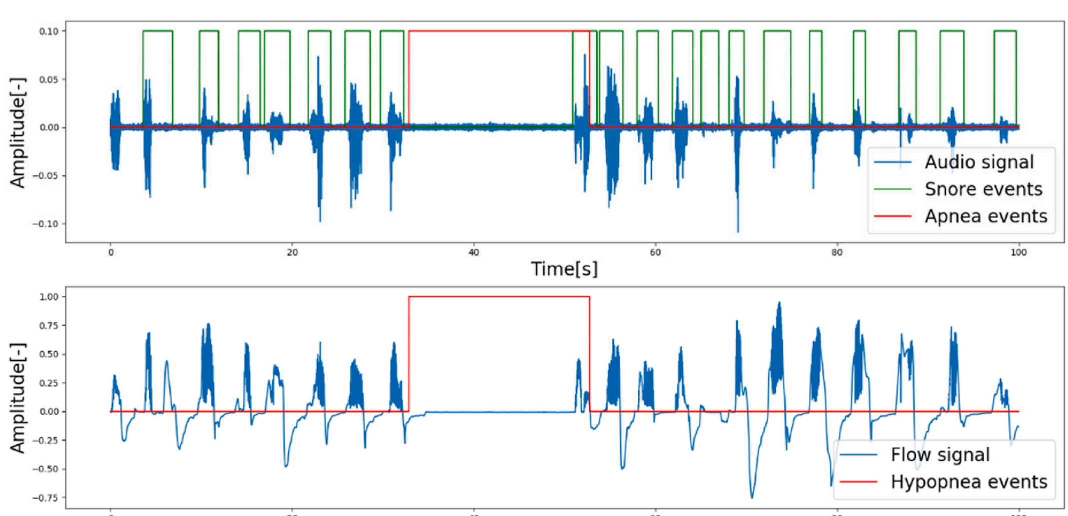

Fig. 2. An example where an obstructive apnea and associated complete cessation of airflow interrupts consecutive snore events.

SSI time series is obtained by concatenating the SSI series from all snore groups of that night.
b. SRV features

After obtaining the SSI per group and for the full night, features are extracted which can later be used in the estimation of AHI and classification of OSA severity. These include time-domain, frequency-domain, and non-linear features, and were inspired by comparable HRV features [23]. Table 2 lists time-domain features, frequency-domain features, and non-linear features extracted from the concatenated full-night SSI. For the frequency-domain features, the bands of VLF, LF and HF were empirically chosen by dividing the frequency bands of VLF, LF and HF typically used for HRV analysis by 5, since the heart rate (60–100 beats per minute) is in average 5 times faster than the respiratory rate (12–20 times per minute) during rest. All frequency-domain features and non-linear features were calculated using the Python package ‘pyhrv’ using the default parameters, with the exception of these frequency bands [19]. A total of 22 features were obtained from the full-night SSI. The time-domain SRV features described in Table 3 are computed separately based on the SSI time series of each separate snore group. To obtain a single value describing these features for the entire night, sample statistics of the resulting group-based features are then calculated, concretely: mean, variance, skewness, kurtosis, min, max, median, standard deviation, and interquartile range. A set of 45 features were obtained from the group-based SSI.

2.2.1.3. Snore energy trend features. Besides the SRV features, we also analyzed the amplitude characteristics of snores. Fig. 3 illustrates an example of snore events in three consecutive hypopneas, from which we can observe that the amplitude (and therefore the energy) of the audio signal decreases at the beginning and then progressively increases throughout the duration of the hypopnea event. In contrast with obstructive apneas, where there is a full cessation of airflow, – and therefore of snoring – during hypopneas snoring often occurs due to the airflow limitations that cause the actual disordered breathing event.

Table 2
Time-domain features, frequency-domain features, and non-linear features extracted from the concatenated full-night SSI.

Parameter	Unit	Description
Time-domain features		
SDSS	s	Standard deviation of SSIs of the whole night
cSS1		Count of successive SSIs that differ by more than 1 s
RMSSD	s	Root mean square of successive SSI differences of the whole night
SDSD	s	Standard deviation of successive SSI differences of the whole night
SRV triangular index		Integral of the density of the SSI histogram divided by its height
Triangular interpolation of SSI histogram (TISS)	s	The baseline width of the distribution measured as a base of a triangle, approximating the SSI distribution.
Frequency-domain features		
VLF relative power	%	Relative power of the very-low-frequency band (0.0006–0.008 Hz)
LF relative power	%	Relative power of the low-frequency band (0.008–0.03 Hz)
HF relative power	%	Relative power of the high-frequency band (0.03–0.08 Hz)
LF/HF	%	Ratio of LF-to-HF power
Non-linear features		
S	s	Area of the ellipse which represents total SRV
SD1	s	Poincaré plot standard deviation perpendicular the line of identity
SD2	s	Poincaré plot standard deviation along the line of identity
SD1/SD2	%	Ratio of SD1-to-SD2
DFA α1		Detrended fluctuation analysis, which describes short-term fluctuations of SRV
DFA α2		Detrended fluctuation analysis, which describes long-term fluctuations of SRV
SampEn		Sample entropy (entropy embedding dimension = 2, tolerance distance = 0.2 * standard deviation), which measures the regularity and complexity of a time series
MSE		Multiscale entropy (scale: 2, 4, 6, 8, 10), which measures the complexity of fluctuations over a range of time series

Table 3
Time-domain features based on SSI of each snore group.

Parameter	Unit	Description
SDSSG	s	Standard deviation of SSIs of each snore group
ASSG	s	Average SSI of each snore group
RMSSDG	s	Root mean square of successive SSI differences of each snore group
SDSDG	s	Standard deviation of successive SSI differences of each snore group
SS_Max_SS_Min	s	Difference between the maximum SSI and the minimum SSI of each snore group

We capture the trends in the energy of consecutive snores, in different features we dubbed snore energy trend (SET), as follows:

Based on the SET calculated for each snore, we computed features that describe how SET varies between consecutive snores. The calculation of the features can be done by following the steps, where root-mean square energy (RMSE) is used to represent the amplitudes of a snore event.

- First, calculate the root mean square energy (RMSE) of the signal in each detected snore event, and the SSI between consecutive snores for each snore group.
- Prepare two arrays comprising the RMSEs and the SSIs of each snore, and of the three following snore events, [$RMSE_0$, $RMSE_1$, $RMSE_2$, $RMSE_3$] and [SSI_0, SSI_1, SSI_2, SSI_3], setting SSI_0 to 0 since four snore events correspond to three intervals. The number of four snore events was chosen heuristically after visually inspecting examples of annotated hypopneas and detected snore events.
- Calculate a time stamp array for the four snore events as [SSI_0, $SSI_0 + SSI_1$, $SSI_0 + SSI_1 + SSI_2$, $SSI_0 + SSI_1 + SSI_2 + SSI_3$].
- Multiply RMSE array by 10^x with the smallest x which can make $min\{RMSEs\} \geq 1$.
- Use the RMSE array and time stamp array to fit a first-degree polynomial,

$$RMSEs = SET^{*}Timestamps + b \tag{2}$$

- Use the absolute value of the slope of the resulting fit as the SET of the current snore event.
- Calculate the SET for each group of 4 snores in the snore groups defined in the previous section. Each snore group with n snore sounds can thus have $n - 3$ values of SET.
- Concatenate all SETs calculated for all snore groups and compute the features in Table 4 to describe the final SET features for the entire night.

2.2.1.4. Demographic features. Besides the SRV and SET features, we also included age and BMI as demographic features. These were chosen due to their known relation with AHI [24,25].

2.2.2. AHI estimation and validation

Extreme gradient boosting (XGBoost) was used to train and validate models for AHI estimation. XGBoost, proposed by Chen and Guestrin [21], has been widely used in many fields, such as personal credit evaluation [26], neonatal sepsis prediction [27], and financial trading [28]. It is an advanced gradient tree boosting system with high performance and operational efficiency. Besides, XGBoost can compute feature

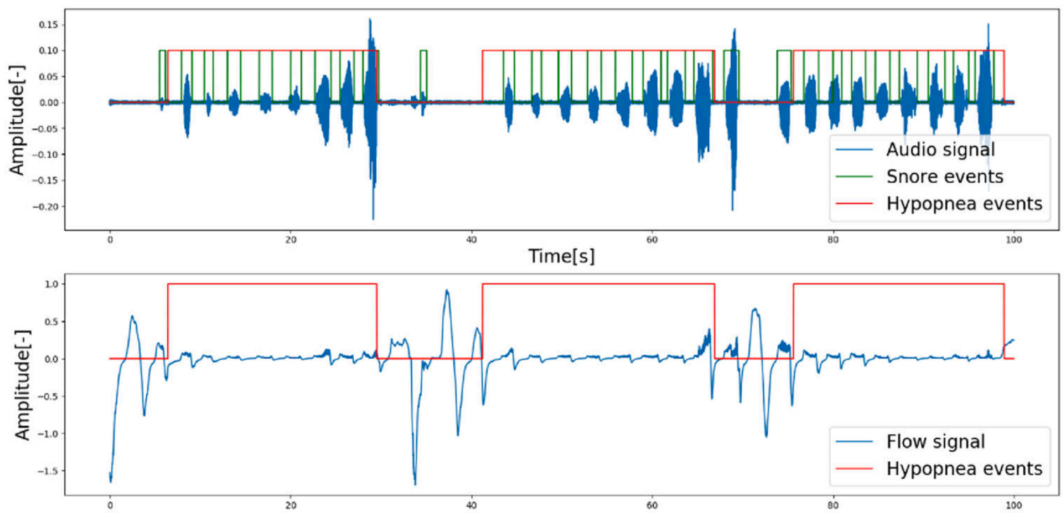

Fig. 3. An example of snore events during hypopneas: the amplitudes of the snore sound signal decreases at the beginning and then progressively increases towards the end of each hypopnea event.

Table 4
Snore energy trend parameters calculated over the concatenated set of SETs calculated from each snore group, 19 features are obtained in total.

Parameter	Description
SET_percentile	Percentile of SET at 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, respectively.
SET_ss	Sample statistics of all concatenated SET values: mean, variance, skewness, kurtosis, min, max, median, standard deviation, and interquartile range
SET_SampEn	Sample entropy of all concatenated SET values

importance scores that indicate the importance of each feature to the trained model.

Five-fold cross-validation was used to train and validate the models. Data from all subjects were randomly divided into five parts on the premise that the proportion of each OSA severity (in terms of AHI value) of subjects remains the same in each part. To be more specific, subjects in each severity group were divided into five parts, then one part (different each time) from every severity group was selected to form one fold. For each cross-validation iteration, a fold (different each time) was held out for testing while the remaining four folds were combined and used as training set. The training set of each iteration was further divided into a set exclusively used to fit the model (75 %) and a tuning set used to choose the hyperparameters (25 %). The parameters (epochs (10–90), learning rate (0.01–0.2), gamma (0.1–2), max depth of a tree (1–5), minimum sum of instance weight needed in a child (1–9), subsample ratio of the training instances (0.5–0.9), subsample ratio of columns when constructing each tree (0.5–0.9), L1 (0.1–100), and L2 (0.1–50) regularization term on weights) with the best performance of the tuning set were selected using the *RandomizedSearch* function from the Python package 'sklearn'. The parameter tuning method was applied for all five folds. After the training parameters were chosen, the fitting set and tuning sets were combined again to train the model. The corresponding hold-out test set was then used as input to the trained model to generate the predictions. After all cross-validation iterations were finished, we combined the predictions from all test sets to obtain measures of performance for all subjects in the dataset.

To compare our study with existing work, we used the feature set (Section 2.2.1.1, 5 features in total) proposed by Ben-Israel et al. [14,15] together with XGBoost to generate "baseline" results. We then combined those features with age and BMI (7 features in total) to train another model and evaluate the impact of age and BMI. Finally, we combined those baseline features, plus age and BMI with the new features proposed in the present study (SRV and SET features) to obtain the "extended" feature set (93 features in total) which was then used to train and test another XGBoost model. Then we compared the results obtained from the three models to analyze if our features contributed to the accuracy of AHI estimation. To further analyze the importance of individual features, we calculated the average feature importance score of all five folds models for each feature.

2.2.3. Evaluation of AHI and severity classification

To evaluate the accuracy of AHI estimation, we used Bland-Altman analysis, Spearman's correlation, and mean absolute error (MAE) between the estimated AHI and the reference AHI obtained by manual scoring of the PSG. In addition, the intraclass correlation coefficient (ICC) using two-way random-effects model which is often used to measure the agreement between different scorings was also calculated [29,30]. We excluded central apnea events when determining the reference AHI, because these are not related to airway obstructions and thus should bear no relation to the properties of snoring.

To evaluate the performance of OSA severity classification, we used the canonical thresholds to establish the severity classes, defined as non-OSA (AHI $<$ 5), mild (5 $<=$ AHI $<$ 15), moderate (15 $<=$ AHI $<$ 30), and severe (AHI $>=$ 30) [31]. Confusion matrices were derived from the classification results. In addition, we also evaluated the performance in binary tasks using different cut-off points, namely AHI $<$ 5 versus AHI $>=$ 5 (non-OSA versus OSA), AHI $<$ 15 versus AHI $>=$ 15 (non– and mild OSA versus moderate and severe OSA), and AHI $<$ 30 versus AHI $>=$30 (non–, mild, and moderate OSA versus severe OSA) in terms of Cohen's kappa, sensitivity, specificity, accuracy, and positive predictive value. Finally, we calculated receiver operating characteristic (ROC) curves for the three cut-off thresholds.

3. Results

3.1. AHI estimation

Using the baseline feature set, we obtained a Spearman's correlation of 0.676 ($p < 0.0001$) and an ICC of 0.594 (95 % confidence interval: 0.49–0.68) between the estimated AHI and the reference AHI, with an MAE of 8.74 events/h. After adding age and BMI to the baseline feature set, the Spearman's correlations, ICC, and the MAE increased to 0.731 ($p < 0.0001$), 0.703 (95 % confidence interval: 0.62–0.77), and 8.18 events/h. By adding our new snoring features to the extended feature set, we obtained a further improved Spearman's correlation of 0.786 (p

< 0.0001), an ICC of 0.733 (95 % confidence interval: 0.66–0.80), and a lower MAE of 7.26 events/h.

Fig. 4 and Fig. 5 illustrate the Bland-Altman and correlation plots comparing the AHI estimation results using the baseline feature set, baseline feature set with age and BMI, and extended feature set. The average bias ± 95 % limits of agreement were 0.54 ± 26.49 (baseline feature set), 1.0 ± 22.92 (baseline feature set + age & BMI), and 0.41 ± 21.42 (extended feature set).

3.2. OSA severity classification

Fig. 6 shows the confusion matrices for OSA severity classification for four classes. Using the baseline feature set, the model achieved an accuracy of 54.7 % (Cohen's kappa of 0.38), and among the incorrect classifications, 78.2 % of the subjects were misclassified to the adjacent severity classes. After including age and BMI, the accuracy actually decreased to 51.7 % (Cohen's kappa of 0.34), but the percentage of the subjects misclassified to the adjacent severity classes improved to 89.2 %. Using the extended feature set, 59.3 % of all subjects were correctly classified (Cohen's kappa of 0.44). Among the incorrectly classified subjects, 90.0 % of them were misclassified to the adjacent severity classes.

Table 5 indicates the performance in the different binary classification tasks for the three canonical severity thresholds (5, 15, and 30) while Fig. 7 illustrate receiver operating characteristic (ROC) curves for the three cut-off thresholds. Adding age and BMI does not change the results substantially, with the exception of an increase in specificity for the mild severity case. In contrast, the extended feature set achieved a higher performance in all tasks, but most notably for the AHI threshold of 5, with an increase in kappa from 0.42 to 0.56 and ROC-AUC which increased from 0.85 to 0.90, mainly driven by a simultaneous increase in the sensitivity and specificity to this task.

3.3. Result comparison

In Table 6, we compare the results of different feature sets for both AHI estimation and OSA severity classification. The results show that including age and BMI with the baseline feature set can enhance AHI estimation performance, but it has a minimal impact on the performance for severity classification. Using the extended feature set, we observed improved results for all metrics of performance.

3.4. Feature importance score

Figure 8 lists the top twenty features ranked by averaging the importance score obtained for each of the five cross-validation iterations. The new features proposed in this work for the extended feature set are prominent in the list, with five SRV features, nine SET features, age, and BMI. Four features from the baseline set are also in the list, only pitch density is missing.

4. Discussion

Non-obtrusive options for OSA detection and severity estimation would yield important clinical advantages. Our study aimed to exploit audio-based snoring features to estimate AHI and assess OSA severity. We added several new features to existing methods and validated this in a new real-life dataset against gold-standard PSG.

Given the literature on this topic, we started from a baseline set of features previously described for this task, i.e., those presented in the study of Ben-Israel et al. [14,15]. This study was chosen because it presented a detailed description that facilitated the implementation of features, while also reporting one of the highest performances in literature on comparable problems. However, the results obtained with this baseline feature set on our data set were substantially inferior to the results obtained in the previous paper (R = 0.676 versus R = 0.9 respectively). An important reason could be the differences between

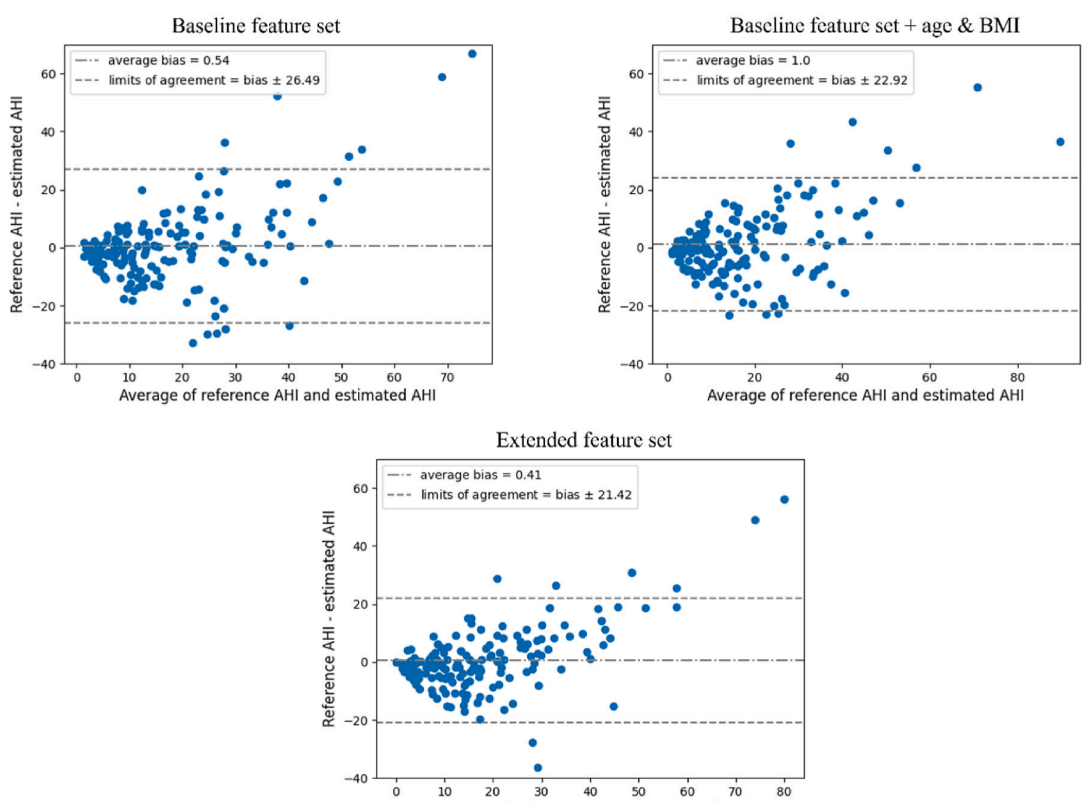

Fig. 4. Bland-Altman plots for AHI estimation.

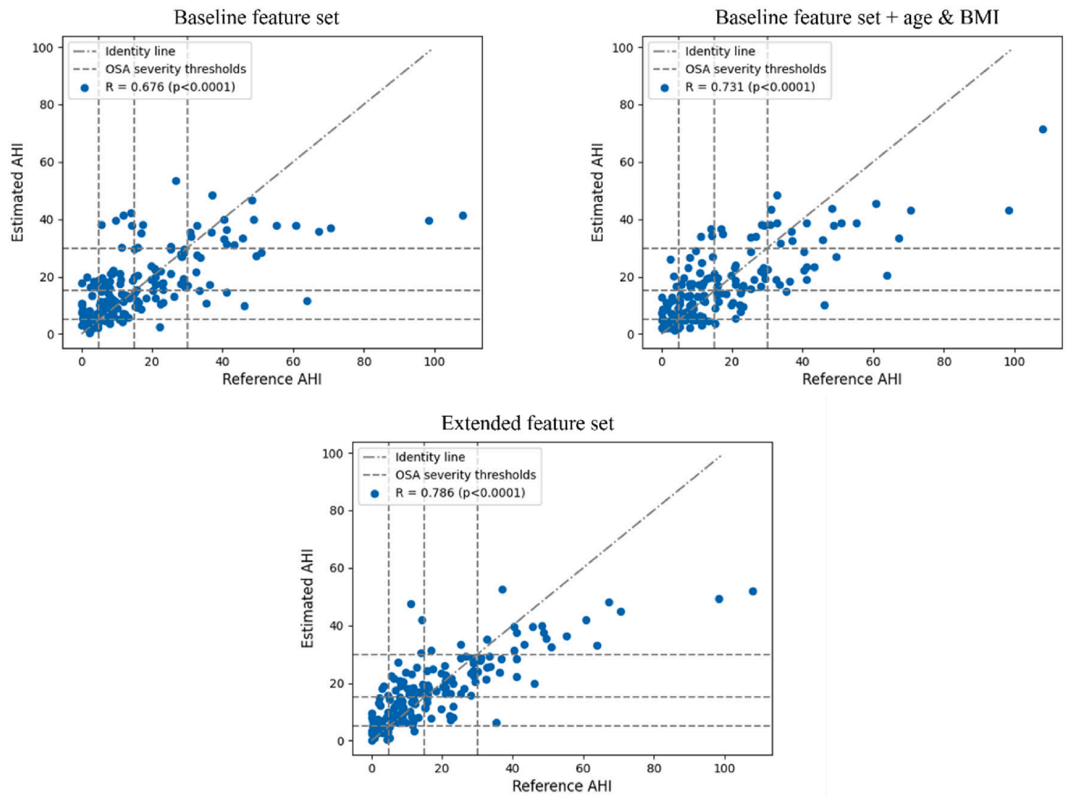

Fig. 5. Correlation plots for AHI estimation.

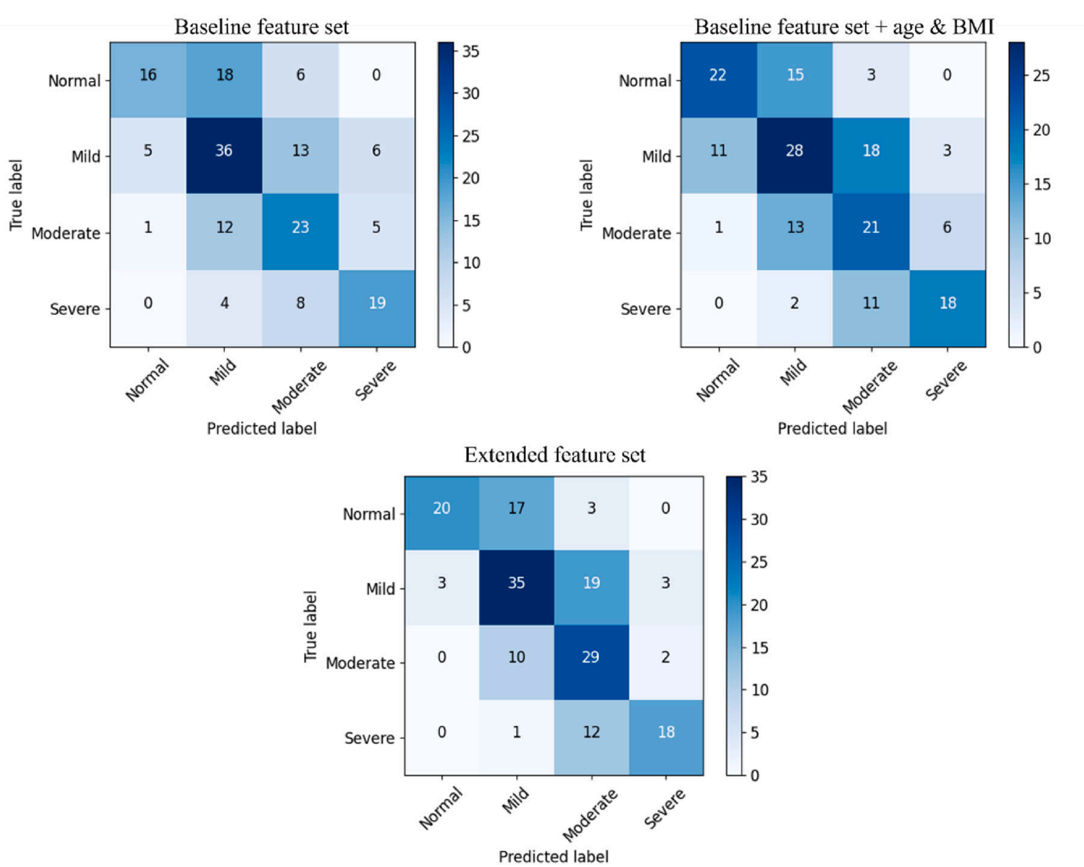

Fig. 6. Confusion matrices of OSA severity classification.

Table 5
OSA severity estimation results with three cut-points (5, 15, 30).

Severity threshold	Cohen's kappa	Sensitivity (%)	Specificity (%)	Accuracy (%)	Positive predictive value (%)
Baseline feature set					
Mild (AHI >= 5 vs AHI < 5)	0.42	95.5	40.0	82.6	84.0
Moderate (AHI >= 15 vs AHI < 15)	0.51	76.4	75.0	75.6	68.8
Severe (AHI >= 30 vs AHI < 30)	0.54	61.3	92.2	86.6	63.3
Baseline feature set + age & BMI					
Mild (AHI >= 5 vs AHI < 5)	0.48	90.9	55.0	82.3	86.7
Moderate (AHI >= 15 vs AHI < 15)	0.53	77.8	76.0	76.7	70.0
Severe (AHI >= 30 vs AHI < 30)	0.54	58.1	93.6	87.2	66.7
Extended feature set					
Mild (AHI >= 5 vs AHI < 5)	0.56	97.7	50.0	86.6	86.6
Moderate (AHI >= 15 vs AHI < 15)	0.58	84.7	75.0	79.1	70.9
Severe (AHI >= 30 vs AHI < 30)	0.61	58.1	96.5	89.5	78.3

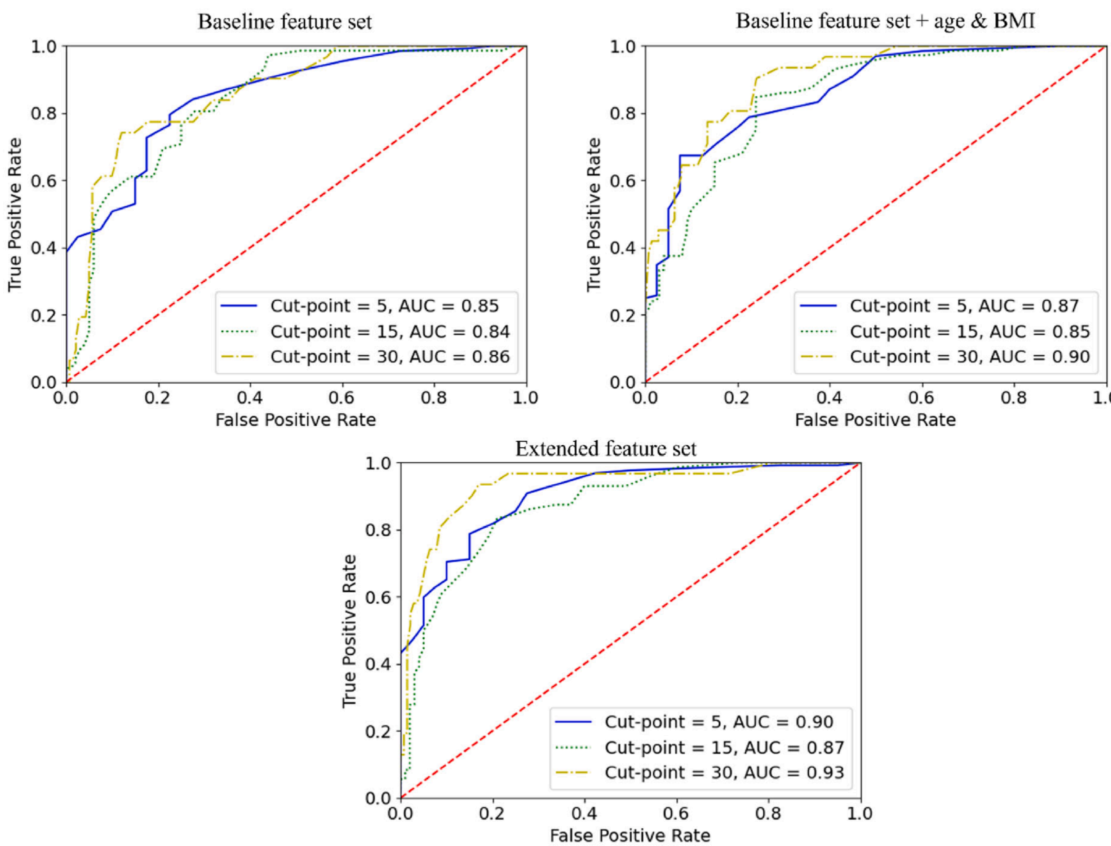

Fig. 7. Receiver Operating Characteristic curves of different cut-points.

Table 6
Comparison of the results using different feature sets.

Feature set	Baseline	Baseline + age & BMI	Extended
AHI estimation			
Spearman's correlation	0.676	0.731	0.786
ICC (95 % confidence interval)	0.594 (0.49–0.68)	0.703 (0.62–0.77)	0.733 (0.66–0.80)
MAE (events/h)	8.74	8.18	7.26
Limits of agreement	0.54 ± 26.5	1.0 ± 22.9	0.41 ± 21.4
OSA severity classification (4 classes)			
Accuracy (%)	54.7	51.7	59.3
Cohen's kappa	0.38	0.34	0.44

datasets, including recording setups and subject characteristics with respect to severity and distribution of sleep disordered breathing events (e.g., predominance of hypopneas over obstructive apneas). For example, we found that the minimum number of detected snore events on a single participant in the dataset of Ben-Israel et al. [14] was 127. In contrast, we included 29 subjects with<127 detected snore events. This may hint at important differences in the characteristics of the participants, or possibly even at the characteristics of the snore events themselves, which may be less prominent in the overall audio recording and thus more difficult to detect, let alone characterize. We can also not fully exclude the possibility that our implementation did not exactly mimic the previously reported analysis, in the absence of a complete description of all feature parameters. To mitigate this as much as possible, we tuned several features (e.g., thresholds for apneic phase ratio, removing outlier during the calculation of running variance) to make sure that

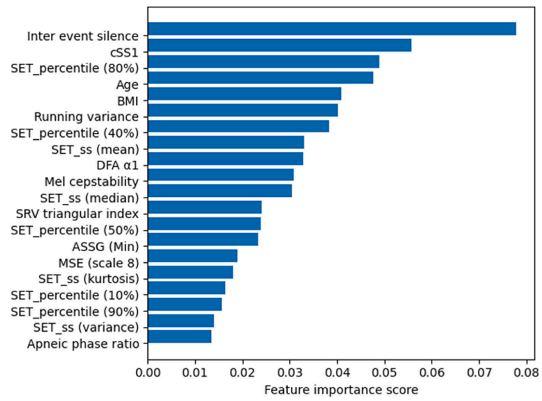

Fig. 8. Features with the top twenty average importance scores.

these features kept their discriminative power in our dataset.

To the best of our knowledge, while age and BMI have been shown to be correlated with AHI, they have not been used in combination with snoring features to train machine learning models for AHI estimation. To investigate the potential impact of incorporating age and BMI on AHI estimation and OSA severity classification, we added them to the baseline feature set in our study. Our findings suggest that these parameters can enhance AHI estimation performance, but with limited impact in the overall performance for OSA severity classification.

Building on the baseline feature set, we manually crafted features that describe the timing and energy content of snores, designing them to be more robust to the quality of the audio acquisition. We introduced SRV as a proxy for RRV. Given the limited literature on changes in RRV during sleep, studying SRV could serve as a starting point for RRV analysis. Based on observations of the properties of snoring in some subjects in our dataset, we hypothesized that changes in airway mechanics during snoring could alter SRV, although to our knowledge, this relationship has not been explored. In a pragmatic manner, to analyze SRV, we utilized features inspired by the field of HRV analysis. Similarly, we derived time, frequency, and non-linear features. In addition, we exploited how the energy of snore signals vary during hypopneas. Combining these with the baseline set, age, and BMI to form an extended feature set, we obtained an improved Spearman's correlation of 0.786 for AHI estimation, and an accuracy of 59.3 % for OSA severity classification. We observed that SET features dominated the importance score list. One possible reason might be that there were more hypopnea events (16394) than apnea events (1836) in our dataset, and 51.0 % of the hypopnea events did overlap with snore events. As illustrated in Fig. 3, flow has a decrease period when entering a hypopnea and an increase period at the end of a hypopnea. It has been shown that sound intensity is dependent on respiratory flow [32,33]. The SET features could describe this characteristic of hypopnea events co-occurring with snore sounds. Therefore, those features may show higher importance score with our hypopnea dominated dataset.

Despite improved performance, it is relevant to highlight that four out of the five original features from the baseline set described in literature ranked consistently high in the feature importance estimation from our classifier, indicating that they are not only suitable, but actually very valuable for this task. As the binary classification tasks and the ROC curves indicate, the biggest improvement seemed to originate from distinguishing participants with an AHI below versus above 5, suggesting that the added robustness of our SRV and SET features improved the estimation especially for the non-OSA and mild-severity groups. After an inspection of our dataset, we found that there is only an aggregated total of 9 apneas, and 611 hypopneas among all subjects in the non-OSA group. We speculate that the reason why the extended feature set performs better at distinguishing participants in this group is related to the addition of our SET features, particularly useful at describing variations in snoring during hypopneas, predominant in participants without OSA and with mild OSA.

Even though the proposed features contributed to an overall improvement in performance, there are relevant limitations to our study. First, we observed the presence of several outliers, clearly visible in Fig. 5. After investigating those, we found that two participants (on the top right of the correlation plots in Fig. 5) had an extremely high AHI (98.4 events/h and 108.1 events/h) while the outliers in the top middle also had a high AHI (around or above 50 events/h). As evident from the scatter plot, most of the participants in our dataset have a relatively low AHI, with only eight subjects having an AHI bigger than 50 events/h. This might limit the capability of our model to learn characteristics associated with higher severities and provide an accurate estimation of AHI in these cases. This can probably be alleviated by extending the dataset with more examples of such subjects. After further inspection, we found that some of the outliers did have a large number of apnea or hypopnea events without clear neighboring snore events, which will obviously violate the basic assumption behind our approach. In this case it is arguable whether our method can ever deliver an accurate estimation of AHI in such conditions, and this may be the most important limitation: even though the method does not rely on the detection of individual sleep disordered breathing events, it leverages the characteristics of snoring as a function of the severity of the airflow limitations associated with these events; when snoring is absent, this relation cannot be established.

A second limitation is that the SRV and SET features require sufficient detected snores to be computed accurately and as such, they cannot be extracted from OSA patients with very few snore events. For example, if some subjects with a moderate to high AHI only have very few snore groups, and by coincidence, all those snore groups do not contain apnea or hypopnea events, then both SRV and SET features extracted from those snore groups will not be correlated to their OSA severity.

Finally, our features depend on the accurate detection of snore events. For example, if several snore events among a group of regular snore events are missed, then SRV and SET features will exhibit an irregularity that is not related to the presence of sleep disordered breathing events, and it is possible that the AHI will be overestimated. In addition, SET features could be affected if a snoring event happens together with some noise with high energy. This suggests that a quiet recording environment, or an algorithm that is able to separate noise and snore, would be required when deploying such a model in real-world situations where environmental noise is difficult to suppress.

In comparison with other OSA screening methods, using audio analysis of snoring may have several advantages. For example, compared with often-used questionnaires [34,35], our accuracy in binary classification tasks for AHI thresholds 5 (86.6 %) and 15 (70.9 %) were higher than those (Berlin: 71 %, 46 %; STOP:77 %, 48 %; STOP-BANG: 79 %, 56 %) reported by Cowan et al. [34]. Furthermore, the AUCs of binary classification tasks for all three thresholds 5, 15, and 30 (0.90, 0.87, and 0.93, respectively) were higher than the multivariate apnea prediction index survey (0.699, 0.671, and 0.761, respectively) reported by Wilson et al. [35]. Besides, our method will not be affected by subjective assessments by the patient or relatives, or criteria that may lead to false negative estimations, such as low age, regular weight, and lack of awareness of some symptoms (e.g., sleepiness and snoring).

Compared with other surrogate methods, for example based on ECG [36,37], and wearables such as PPG [38,39], an audio signal is relatively easy to obtain, for example with smartphones having sensitive microphones. The relatively low complexity of the algorithm could potentially enable and embedded implementation in such devices [40]. In addition, audio sensors such as microphones would not need to be attached to the human body, making it less obtrusive and more comfortable. Moreover, as snoring is usually louder than environmental sounds, it is relatively robust to noise. In addition, unlike features like Linear Prediction derived Cepstral Coefficients and MFCCs which can be affected by sampling frequency [41,42], we expect SRV features to be more robust to the characteristics of the microphones, since they rely only on an accurate detection of the occurrence of snoring but not the audio properties of the snoring itself. Besides, we expect SET features to also be robust to these factors, as they only require an accurate identification of the timing and the signal energy of the snoring event. This should be further confirmed in future studies with recordings from microphones with different quality, resolution, sampling frequency, and in positions more suitable for night-to-night use, for example on a night table next to the bed. Furthermore, those features are robust to pre-processing techniques which do not change the energy of the audio signals. Finally, we have shown a sensitivity to the composition of the included datasets for these methods, not only from the recording standpoint, but also clinical characteristics such as the balance between apneas and hypopneas. This underscores the importance of validating published algorithms in different datasets, which further supports efforts to enable access to different datasets and cooperation between research groups.

Future work can take multiple directions. One direction could be to further validate the generalizability and robustness of the new features using larger datasets that contain different microphone settings and a wider spectrum of apneic events and sleep disorders. Another direction could be to investigate the physiological basis of SRV and its relation to RRV during sleep, and in the presence of sleep disordered breathing. A better understanding of the physiological basis behind this phenomenon would allow us to better choose certain parameters of our features, such as the frequency bands of VLF, LF, and HF, and possibly drive a further improvement in AHI estimation accuracy.

CRediT authorship contribution statement

Jiali Xie: Writing – original draft, Methodology, Software, Validation, Formal analysis, Investigation, Visualization. **Pedro Fonseca:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Johannes van Dijk:** Methodology, Writing – original draft. **Sebastiaan Overeem:** Conceptualization, Writing – original draft, Supervision. **Xi Long:** Methodology, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Currently, the included dataset is not publicly available. Data can be made available in collaboration with researchers, depending on reasonable request and respecting privacy regulations

Acknowledgements

The authors wish to express their gratitude to Xavier Aubert for the help of synchronizing the audio signals and PSG signals, and all somnologists, sleep technicians, and nurses at the Center for Sleep Medicine Kempenhaeghe for help in collecting study data.

Funding

Parts of this study were supported by the Open Technology Program from STW/NOW (OSA+ project no. 14619), OPZuid (Bedsense, 2015) and an Impulse grant in the Eindhoven MedTech Innovation Center (e/MTIC) cooperation.

Ethics approval

Maxima Medical Center (Eindhoven, the Netherlands, File no: N16.074).

Data availability

Currently, the included dataset is not publicly available. Data and the scripts can be made available in collaboration with researchers, depending on reasonable request and respecting privacy regulations.

References

[1] J.A. Dempsey, S.C. Veasey, B.J. Morgan, C.P. O'Donnell, Pathophysiology of sleep apnea, Physiol. Rev. 90 (1) (2010) 47–112.
[2] C.V. Senaratna, et al., Prevalence of obstructive sleep apnea in the general population: a systematic review, Sleep Med. Rev. 34 (2017) 70–81.
[3] J. Collen, C. Lettieri, E. Wickwire, A. Holley, Obstructive sleep apnea and cardiovascular disease, a story of confounders!, Sleep Breath. 24 (2020) 1299–1313.
[4] S.P. Patil, H. Schneider, A.R. Schwartz, P.L. Smith, Adult obstructive sleep apnea: pathophysiology and diagnosis, Chest 132 (1) (2007) 325–337.
[5] D.J. Gottlieb, N.M. Punjabi, Diagnosis and management of obstructive sleep apnea: a review, JAMA - J. Am. Med. Assoc. 323 (14) (2020) 1380–1400.
[6] M. Bruyneel, V. Ninane, Unattended home-based polysomnography for sleep disordered breathing: current concepts and perspectives, Sleep Med. Rev. 18 (4) (2014) 341–347.
[7] M. Bruyneel, W. Libert, L. Ameye, V. Ninane, Comparison between home and hospital set-up for unattended home-based polysomnography: a prospective randomized study, Sleep Med. 16 (11) (2015) 1434–1438.
[8] K.F. Whyte, M.B. Allen, A.A. Jeffrey, G.A. Gould, N.J. Douglas, Clinical features of the sleep apnoea/hypopnoea syndrome, QJM 72 (1) (1989) 659–666.
[9] V. Hoffstein, S. Mateika, D. Anderson, Snoring: Is it in the ear of the beholder? Sleep 17 (6) (1994) 522–526.
[10] H. Jin, et al., Acoustic analysis of snoring in the diagnosis of obstructive sleep apnea syndrome: a call for more rigorous studies, J. Clin. Sleep Med. 11 (7) (2015) 765–771.
[11] J.A. Fiz, R. Jané, J. Solà-Soler, J. Abad, M.Á. García, J. Morera, Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index, Laryngoscope 120 (4) (2010) 854–862.
[12] J. Mesquita, J. Solà-Soler, J.A. Fiz, J. Morera, R. Jané, All night analysis of time in terval between snores in subjectswith sleep apnea hypopnea syndrome, Med. Biol. Eng. Comput. 50 (2012) 373–381.
[13] E. Dafna, A. Tarasiuk, Y. Zigel, OSA severity assessment based on sleep breathing analysis using ambient microphone, Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS (2013) 2044–2047.
[14] N. Ben-Israel, A. Tarasiuk, Y. Zigel, Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults, Sleep 35 (9) (2012) 1299–1305.
[15] N. Ben-Israel, A. Tarasiuk, Y. Zigel, Nocturnal sound analysis for the diagnosis of obstructive sleep apnea, In: 2010 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC'10, 2010, pp. 6146–6149.
[16] M.M. Van Gilst et al., Protocol of the SOMNIA project: An observational study to create a neurophysiological database for advanced clinical sleep monitoring, BMJ Open 9 (11) (2019).
[17] J. Xie, et al., Audio-based snore detection using deep neural networks, Comput. Methods Programs Biomed. 200 (2021) 105917.
[18] E. Dafna, A. Tarasiuk, Y. Zigel, Automatic detection of whole night snoring events using non-contact microphone, PLoS One 8 (12) (2013).
[19] P. Gomes, P. Margaritoff, H. Placido da Silva, pyHRV: development and evaluation of an open-source python toolbox for heart rate variability (HRV), in: On electrical, electronic and computing engineering (icetran), 2019, pp. 822–828.
[20] F. Pedregosa, et al., Scikit-learn: machine learning in python Fabian, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[21] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2016), pp. 785–794.
[22] B. Xhyheri, O. Manfrini, M. Mazzolini, C. Pizzi, R. Bugiardini, Heart rate variability today, Prog. Cardiovasc. Dis. 55 (3) (2012) 321–331.
[23] F. Shaffer, J.P. Ginsberg, An overview of heart rate variability metrics and norms, Front. Public Heal. (2017) 258.
[24] T. Leppänen, J. Töyräs, E. Mervaala, T. Penzel, A. Kulkas, Severity of individual obstruction events increases with age in patients with obstructive sleep apnea, Sleep Med. 37 (2017) 32–37.
[25] A. Romero-Corral, S.M. Caples, F. Lopez-Jimenez, V.K. Somers, Interactions between obesity and obstructive sleep apnea: implications for treatment, Chest 137 (3) (2010) 711–719.
[26] H. Li, Y. Cao, S. Li, J. Zhao, Y. Sun, XGBoost model and its application to personal credit evaluation, IEEE Intell. Syst. 35 (3) (2020) 52–61.
[27] Z. Peng, et al., A Continuous Late-Onset Sepsis Prediction Algorithm for Preterm Infants using Multimodal Physiological Signals from a A Continuous Late-Onset Sepsis Prediction Algorithm for Preterm Infants using Multimodal Physiological Signals from a Patient Monitor, IEEE J. Biomed. Health Inform. 27 (1) (2023) 550–561.

[28] J. Nobre, R.F. Neves, Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets, Expert Syst. Appl. 125 (2019) 181–194.

[29] N.M. Punjabi, N. Shifa, G. Dorffner, S. Patil, G. Pien, R.N. Aurora, Computer-assisted automated scoring of polysomnograms using the somnolyzer system, Sleep 38 (10) (2015) 1555–1566.

[30] U.J. Magalang, E.S. Arnardottir, N. Chen, P.A. Cistulli, T. Gíslason, D. Lim, T. Penzel, R. Schwab, S. Tufik, A.I. Pack,, Agreement in the scoring of respiratory events among international sleep centers for home sleep testing, J. Clin. Sleep Med. 12 (1) (2016) 71–77.

[31] American Academy of Sleep Medicine Task Force, Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force, Sleep 22 (5) (1999) 667.

[32] H. Pasterkamp, S.S. Kraman, G.R. Wodicka, Respiratory sounds: advances beyond the stethoscope, Am. J. Respir. Crit. Care Med. 156 (3) (1997) 974–987.

[33] A.Y.M. Jones, R.D. Jones, K. Kwong, Y. Burns, The effect on sound generation of varying both gas flow rate and the viscosity of sputum-like gel in a simple tubular model, Lung 178 (2000) 31–40.

[34] D.C. Cowan, et al., Predicting sleep disordered breathing in outpatients with suspected OSA, BMJ Open 4 (4) (2014) e004519.

[35] G. Wilson, et al., Screening for sleep apnoea in mild cognitive impairment: the utility of the multivariable apnoea prediction index, Sleep Disord. 2014 (2014) 945287.

[36] V.P. Rachim, G. Li, W.Y. Chung, Sleep apnea classification using ECG-signal wavelet-PCA features, Biomed. Mater. Eng. 24 (6) (2014) 2875–2882.

[37] L. Chen, X. Zhang, C. Song, An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram, IEEE Trans. Autom. Sci. Eng. 12 (1) (2015) 106–115.

[38] M. Deviaene, et al., Sleep apnea detection using pulse photoplethysmography, Comput. Cardiol. 45 (2018) 1–4.

[39] G.B. Papini, P. Fonseca, M.M. van Gilst, J.W.M. Bergmans, R. Vullings, S. Overeem, Wearable monitoring of sleep-disordered breathing: estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography, Sci. Rep. 10 (1) (2020) 13512.

[40] J.K. Chiang, Y.C. Lin, C.W. Lin, C.S. Ting, Y.Y. Chiang, Y.H. Kao, Validation of snoring detection using a smartphone app, Sleep Breath. 26 (2022) 81–87.

[41] C. Sanderson, K.K. Paliwal, Effect of different sampling rates and feature vector sizes on speech recognition performance, in: IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 1, 1997, pp. 161–164.

[42] O. Lowna Baroi, M.S. Abrar Kabir, A. Niaz, M. Jahidul Islam, M. Jakaria Rahimi, Effects of filter numbers and sampling frequencies on the performance of MFCC and PLP based Bangla isolated word recognition system, Int. J. Image, Graph. Signal Process. 11 (11) (2019) 36–42.