On the generalizability of ECG-based obstructive sleep apnea monitoring: merits and limitations of the Apnea-ECG database

On the generalizability of ECG-based obstructive sleep apnea monitoring: merits and limitations of the Apnea-ECG database

Conference Paper · May 2018

CITATIONS
0

READS
103

7 authors, including:

Gabriele Papini
Technische Universiteit Eindhoven
21 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Pedro Fonseca
Philips
51 PUBLICATIONS **325** CITATIONS
SEE PROFILE

Merel M van Gilst
Technische Universiteit Eindhoven
7 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Sebastiaan Overeem
Technische Universiteit Eindhoven
213 PUBLICATIONS **10,402** CITATIONS
SEE PROFILE

On the generalizability of ECG-based obstructive sleep apnea monitoring: merits and limitations of the Apnea-ECG database.

Gabriele B. Papini[1,2,3], Pedro Fonseca[1,2], Jenny Margarito[2], Merel M. van Gilst[1,3], Sebastiaan Overeem[1,3], Jan W.M. Bergmans[1,2], *Senior Member, IEEE*, Rik Vullings[1]

***Abstract*— Obstructive sleep apnea syndrome (OSAS) is a sleep disorder that affects a large part of the population and the development of algorithms using cardiovascular features for OSAS monitoring has been an extensively researched topic in the last two decades. Several studies regarding automatic apneic event classification using ECG derived features are based on the public Apnea-ECG database available on PhysioNet. Although this database is an excellent starting point for apnea topic investigations, in our study we show that algorithms for apneic-epochs classification that are successfully trained on this database (sensitivity>85%, false detection rate<20%) perform poorly (sensitivity<55%, false detection rate>40%) in other databases which include patients with a broader spectrum of apneic events and sleep disorders. The reduced performance can be related to the complexity of breathing events, the increased number of non-breathing related sleep events, and the presence of non-OSAS sleep pathologies.**

I. INTRODUCTION

Obstructive sleep apnea syndrome (OSAS) is a public health issue that affects a significant fraction of the general population (5-15%). OSAS influences both the quality and the life expectancy of an untreated patient and often leads to daytime sleepiness and an increased risk for cardiovascular diseases. The gold standard used in the diagnosis of OSAS is night-time polysomnography (PSG), a sleep test consisting of an overnight measurement session during which several physiological signals are recorded. The diagnostic power of PSG is counterbalanced by several drawbacks such as its cost, its obtrusiveness and the impossibility to use it for long-term monitoring to better characterize OSAS. In the last two decades, there has been a trend to replace or complement PSG in OSAS monitoring with more portable and cost-effective tests, e.g. respiratory polygraphic tests replaced PSG in several European countries. The push towards new sleep monitoring devices boosted recently with the rise of wearable cardiac monitoring consumer devices, such as ECG patch and PPG based smartwatch [1], [2]. Given the affordability and broad availability of these devices and their relative convenience of use, research regarding features and algorithms to automatically characterize OSAS using solely cardiac signals has been promoted [3]–[7]. The Computer in Cardiology Challenge of 2000 was one of the starters of this research trend. One of the tasks of the competition was to identify one-minute epochs containing apneic events in 70 overnight recordings using only the ECG signal [3], [8]. The data used for the challenge were organized in a public database on PhysioNet ("Apnea-ECG database"), which, according to the citation index of the paper describing it [3], have been used over 150 times. Several papers regarding ECG-based apneic event detectors are based on this database, and report accuracies ranging from 80% to 90% in the classification of one-minute epochs as apneic or not. The common principle behind all of them is that the cardiac activity is affected by the presence of apneic events [4]–[7]. However, these events are not the only phenomena influencing cardiac activity during sleep [9]. For instance, many parasomnia events, such as limb movements and sleep terrors, are associated or followed by an arousal, just as most apneic events [10]. Arousals associated with non-apneic events can degrade the performance of apneic event detection algorithms and hence their presence can be crucial in databases used for development of apnea-related algorithms and features. The Apnea-ECG database does not have annotations for non-apneic events and the control group is composed of healthy subjects, indicating that patients suffering from other disorders are not included [3]. In addition to this, the Apnea-ECG database lacks more complex apneic cases, such as subjects suffering from central apnea, or other sleep-related disorders, or sleep comorbidities, such as insomnia. The absence of these cases can limit the applicability of algorithms exclusively developed on the Apnea-ECG database in real-world situations, e.g. for screening. As a consequence, even though the Apnea-ECG database has allowed a common testing ground to compare the performance of different algorithms, its usage as the only data source obscures the limitations and generalizability of apneic epoch classification solutions. For instance, Lado et al. showed that a threshold-based classification on a single HRV feature (inter-beat-interval low-to-high frequency band power ratio) produces unreliable results in case the training and testing are performed between different databases [11]. These differences could be related to the fact that a single feature may be insufficient to describe such a complex phenomenon as OSAS and the physiological differences between the different types of apneas and hypopneas.

In this paper we report the effect of testing an apneic-epoch classifier trained on the Apnea-ECG database on two other databases, including patients with mixed sleep disorders and a broad spectrum of events.

*This work is financially supported by STW/IWT in the context of the OSA+ project (No. 14619)

[1]G.B. Papini, P. Fonseca, M.M. van Gilst, S. Overeem, J.W.M. Bergmans, R. Vullings are with the Dept. of Electrical Engineering, TU/e, De Zaale, 5612 AZ Eindhoven, NL. g.p.papini@tue.nl

[2]G.B. Papini, P. Fonseca, J. Margarito and J.W.M Bergmans are with Philips Research, HTC, 5656 AE Eindhoven, NL.

[3]G.B. Papini, M.M. van Gilst and S. Overeem are with the Sleep Medicine Centre Kempenhaeghe, Sterkselseweg 65 5591 VE Heeze, NL

II. METHODOLOGY

A. Databases

The Apnea-ECG database consists of 70 recordings divided into two equinumerous subsets, one for training and one for testing apneic-epoch classifiers [3]. The data separation in the training and test set used during the challenge is maintained in our study. Besides the ECG signal, the database provides annotations for each one-minute epoch of each recording indicating whether there is an obstructive apnea, hypopnea or mixed apnea event.

In addition, we use the St. Vincent's University Hospital/University College of Dublin Sleep Apnea (UCD) database [12] and the Sleep and OSA Measuring with Non-Invasive Applications (SOMNIA) database. The UCD database contains 25 overnight PSG recordings from adults with suspected sleep breathing disorders. For this research, the ECG modified lead V2 (128 Hz) and the respiratory event annotations (obstructive apnea and hypopnea, central apnea and hypopnea, and mixed apnea) are used. The average $\pm$ standard deviation apnea-hypopnea index (AHI) is 24$\pm$20 (minimum 6). Central and obstructive hypopneas are grouped in a single hypopnea class since their split classification is still under investigation [13]. The UCD database is more complex than the Apnea-ECG database regarding apneic events because it also includes central apnea and periodic breathing episodes.

The SOMNIA database stems from a currently ongoing data collection in the Sleep Medicine Center Kempenhaeghe, the Netherlands, and is planned to comprise 1000 overnight PSG recordings of patients undergoing PSG as part of standard sleep diagnostic procedure. For our study, a subset of 57 subjects (30 males) is selected from the SOMNIA database. The subjects have an average age of 50$\pm$16 years, weight 82$\pm$21 kg, height 173$\pm$13 cm and BMI 28$\pm$16 kg/m^2. Each overnight recording includes ECG modified lead II signal (512 Hz) and events annotated by a single clinical annotator, including limb movements (single and periodic limb movements); arousals; central, mixed, obstructive apneas; obstructive hypopnea (no split between obstructive and central) and snore. In addition, apneic events are labeled as *unsure respiratory event* in case of uncertainty. The SOMNIA subset used does not contain periodic breathing events, such as Cheynes-Stokes breathing. The patients in the SOMNIA subset predominately have an OSAS diagnosis without any comorbidities (21), but there are also subjects with OSAS comorbid with other (single or multiple) pathologies (9). In addition to OSAS, the database includes diagnosis—alone or as comorbid disorders—of parasomnias (8), sleep movement disorders (12), insomnias (14), snoring (1), undifferentiated somatoform disorder (1) and behavioral-related sleep problems (3). Only one subject is not diagnosed with any sleep disorder. The average AHI is 14$\pm$14 for the complete subset and 22$\pm$14 for OSAS subjects. This database is the most complex of the three because of the combination of apneic events and other sleep disorders.

One-minute epoch annotations are derived for the UCD and the SOMNIA databases by labeling them as apneic-epochs if they include a breathing related event, with the exception of snoring, similar to other research using the Apnea-ECG database [6], [7].

B. Classification algorithm training and testing

Features based on heart rate variability (HRV) and surrogate respiratory effort, based on ECG derived respiration (EDR) are obtained from the ECG signals using previously published algorithms for heart beat localization and EDR extraction [14], [15]. All features are obtained for non-overlapping windows of one minute in order to match the Apnea-ECG database epoch annotations, except where indicated (with $*$), where the same window size as described in literature is used. A total of 51 HRV and 8 EDR features are extracted, including: HRV time and frequency domain ($>$0.15 Hz) [16], [17], Hilbert transform [4], Detrended fluctuation analysis (DFA) ($*$) [18], [19], and EDR time and frequency domain ($>$0.15 Hz) [14]. Frequency related features are calculated for only the spectrum above 0.15 Hz due to the windows chosen size. The features are not normalized per subject, but scaled by removing the median and dividing for their 5$^{\text{th}}$ to 95$^{\text{th}}$ percentile (%ile) range. They are then filtered using a centered moving average (size 7 epochs), as proposed by de Chazal et al., in order to attenuate calculation errors and to consider the effect of apneic events on adjacent epochs [5].

The Apnea-ECG database training set is used for feature and classifier selection via 10-fold cross-validation. The features that best characterize apneic-epochs are selected by recursive feature elimination (RFE) using a logistic regression model (L1 penalty) as external estimator [20]. A selected feature set is composed of the features recurring, after RFE, in all the cross-validation splits. Two starting numbers of features, 10 and 40, are chosen to determine two feature sets, each of them used to train a separate apneic-epoch classifier. The lowest starting number of features is used to boost the generalizability of the algorithm developed [7]. The largest is chosen to show the relation between the results and the dimension of feature space. After feature selection, for each feature space, a classifier is automatically chosen between a linear, a quadratic, and a logistic regression classifier as the one with the best average accuracy during cross-validation. The final apneic-epoch classification algorithms, each combining a feature set and a classifier type, are trained on the entire Apnea-ECG training set and tested on the Apnea-ECG test set, the UCD database, and the SOMNIA database.

In order to evaluate the classification performance for different types of events, each epoch in the SOMNIA dataset is labeled with the most frequent occurring event, where each event must last at least three seconds. In case multiple events are present in the epoch, but none of them is longer than three seconds, the epoch is labeled as *unknown*. In case no event is detected, the epoch is labeled as *normal*. The epoch labels are compared with the output of the apneic-epoch classifier with the best performance over the three test sets.

TABLE I

TESTING RESULTS OF THE ALGORITHMS FOR THE DIFFERENT DATABASES (6-FEATURES - 24-FEATURES CLASSIFIERS)

	# epochs (apneic%)	Sensitivity	Specificity	Accuracy	Cohen's kappa	False detection rate
Apnea-ECG test set	17254 (37.9%)	87.2% - 88.3%	87.0% - 88.3%	87.1% - 88.3%	0.73 - 0.75	19.6% - 17.9%
UCD	9843 (27.9%)	50.6% - 33.1%	84.0% - 88.7%	74.7% - 73.2%	0.35 - 0.25	45.0% - 46.9%
SOMNIA	29656 (16.2%)	36.6% - 14.6%	87.8% - 95.0%	79.5% - 82.0%	0.24 - 0.13	63.2% - 63.9%
OSAS∗	15736 (25.8%)	38.0% - 16.0%	86.0% - 96.1%	73.6% - 75.4%	0.26 - 0.16	51.5% - 41.3%
Not OSAS†	13920 (5.4%)	28.9% - 7.4%	89.5% - 94.0%	86.2% - 89.3%	0.12 - 0.01	86.4% - 93.4%

∗SOMNIA subset with OSAS diagnosed patients (30); †SOMNIA subset with patients diagnosed with other or none sleep disorders (27).

III. RESULTS

A. Apnea-ECG database

The cross-validation on the Apnea-ECG training set generated two apneic-epoch classification algorithms. The smallest feature set includes 6 features, namely:

- HRV time domain: IBI RMSSD, SDSD and pNN50.
- HRV frequency domain: normalized max power in the respiration frequency band (0.15:0.40 Hz, HF).
- DFA: scaling exponent proposed by Peng et al. and the one for short time scales by Penzel et al. [18], [19].

The largest feature set comprises 24 features which are, in addition those of the smallest set:

- HRV time domain: IBI 10^{th}, 90^{th} and 95^{th} %ile, and minimum and standard deviation (SD); heart rate 10^{th} and 90^{th} %ile; linearly detrended IBI median and 10^{th} %ile.
- HRV frequency domain: normalized HF band power and the maximum power frequency in HF.
- Hilbert: SD of instantaneous frequency sequence and average amplitude of the IBI Hilbert transform.
- EDR time domain: respiratory frequency (inverse of average breath length) and breath-by-breath correlation.
- EDR frequency domain: band power, maximum power, frequency at the maximum power and its SD (in the two adjacent epochs) of the high frequency (0.15:0.5 Hz).

For both feature sets, the chosen classifier is a logistic regression (L1 penalty). The test set results are reported in Table I in terms of sensitivity, specificity, accuracy, agreement to the reference (Cohen's kappa) and false detection rate (false positive/total positive) on the total number of epochs.

B. UCD and SOMNIA databases

The results obtained for the UCD and SOMNIA databases with the classifiers trained with the two feature sets are reported in Table I. In addition, the table includes also the results obtained for the subjects with and without OSAS diagnosis in the SOMNIA database. Using the classifier trained with the small feature set, the resulting classifications for each epoch type are separately illustrated in Fig. 1.

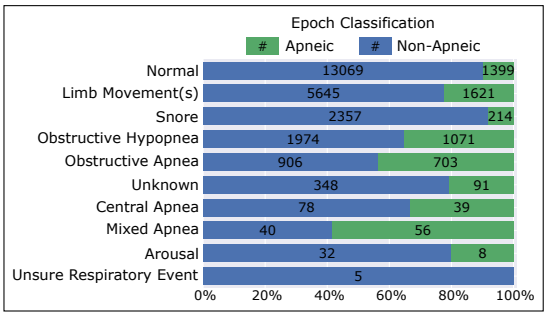

Fig. 1. Overview of the classification of epoch per event type in the SOMNIA database using the 6-features classifier (% over the number of epochs assigned to an event type). Within the bars the number of epochs classified as apneic or non-apneic for each event.

IV. DISCUSSION

Both classifiers show an apneic-epoch classification performance on the Apnea-ECG database with an accuracy above 87% and a good balance between sensitivity and specificity. The results are in line with most literature for the Apnea-ECG database [4]–[7], although some studies obtained an accuracy above 90% by employing a larger number of features or more complex solutions [5]. Since the goal is to test the generalization power of algorithms trained on a specific database, the results can be considered representative of most results obtained on the topic.

Both classifiers have an overall accuracy drop of almost 15% when tested on the UCD and SOMNIA databases. The decrease is not substantial because of the different amount of apneic-epochs present in the databases (UCD and SOMNIA databases having almost half of the percentage of apneic-epochs as the Apnea-ECG database). The real number of epochs belonging to the non-apneic class is principally contributing to the accuracy in the UCD and SOMNIA databases, leading to a diminished descriptive power of the accuracy metric. In fact, as an example, even if the apneic-epoch classifier with 6 features has an accuracy close to 80%, it tends to misclassify approximately 60% of the epochs containing apneic events, regardless of their type (Fig. 1).

The classifier based on the larger feature set achieves better performance than the classifier with fewer features on the Apnea-ECG database (Table I). This is due to the additional features being included, such as the EDR related features. However, this increase in performance is not seen when the classifiers are tested in the UCD and the SOMNIA databases. Both sensitivity and Cohen's kappa of the classifier trained with the larger feature set are lower than the one with the smaller feature set, indicating that the inclusion of new features, among which some more related to respiratory characteristics (i.e. EDR features), plays a detrimental role in the discrimination between apneic and non-apneic epoch in databases other than the Apnea-ECG database. This effect is presumably caused by the overfitting of the algorithm to a specific topology of the data in the Apnea-ECG database, and this effect is magnified by the increased amount of features. In addition to the over-characterization of the apneic-epochs specific to the Apnea-ECG database, the additional features do not bring any new non-apneic discriminative information

since the false detection rate remains on average unchanged.

Also the classifier based on 6 features shows low performance when tested on databases other than the Apnea-ECG database. The increase in sensitivity and Cohen's kappa of the 6-features classifier, in comparison to the 24-features classifier, comes together with an almost unchanged false detection rate. This suggests that, while decreasing the feature set size improves the generalizability of apneic-epoch classifier, at the same time it decreases the ability of the classifier to capture the differences between apneic and non-apneic phenomena. Therefore, the presence of non-apneic events in the training set prior to feature selection becomes paramount in order to decrease the number of false positives. Interestingly, the sensitivity and Cohen's kappa obtained on the UCD database are better than those obtained on the SOMNIA database. This can be explained by the difference in complexity of the two databases: the SOMNIA database is likely to include a broader spectrum of sleep pathologies when compared with the UCD database, since the first encompasses data from the general sleep center population, while the second includes only data from subjects with suspicion of sleep breathing disorders.

The increase in false detection rate can also be explained by the difference in the characteristics of the patients in the data sets. The more non-apneic events occur which have similar feature characteristics as apneic events, the higher the number of false positives will be. The SOMNIA database is likely to include more non-apneic events, due to the higher number of pathologies included (e.g. limb movement events due to sleep movement disorders, etc.). Consequently, the false detection rate for the SOMNIA is substantially higher than for the UCD database (using the 6-features classifier). The false detection rate does not match the one obtained on the UCD database even in case only the subjects with OSAS are considered, regardless of the improved balance in apneic-epochs percentage. This result can be attributed to other events occurring during sleep, such as limb movements, or the occurrence of non-respiratory arousals since they can influence the features in a similar manner to OSAS [9]. For instance, over 20% of the epochs containing limb movements are classified as apneic and, since they represent almost twice the epochs than those with apneic related events, they play a significant role in the presence of false positives (Fig. 1).

V. Conclusion

Our research shows that the performance of apnea detection algorithms is strongly influenced by the choice of the database used to train the classifier. Databases that do not encompass the full complexity and variety of sleeping and sleep pathologies are prone to generate solutions which cannot be easily employed in more complex situations, for instance in case of sleep comorbidities. Although the Apnea-ECG database is of critical importance in the advancement of ECG-based apnea detection research, our work suggests that this database has to be considered as a starting point for research in this area rather than as the definitive database in the field. Therefore, we want to promote a new effort for larger, more comprehensively annotated and multi-center data collections.

References

[1] P. Lévy, M. Kohler, W. T. McNicholas, F. Barbé, R. D. McEvoy, V. K. Somers, L. Lavie, and J.-L. Pepin, "Obstructive sleep apnoea syndrome," *Nature Reviews Disease Primers*, vol. 1, p. 15015, 2015.

[2] T. Penzel, "Technology to assess sleep," *Sleep Medicine Clinics*, vol. 11, no. 4, p. i, 2016. Technology to Assess Sleep.

[3] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ecg database," in *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, pp. 255–258, 2000.

[4] T. Penzel, J. McNames, P. de Chazal, B. Raymond, A. Murray, and G. Moody, "Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings," *Medical and Biological Engineering and Computing*, vol. 40, pp. 402–407, Jul 2002.

[5] P. de Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley, "Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp. 686–696, June 2003.

[6] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. V. Huffel, "A novel algorithm for the automatic detection of sleep apnea from single-lead ecg," *IEEE Transactions on Biomedical Engineering*, vol. 62, pp. 2269–2278, Sept 2015.

[7] S. Martín-González, J. L. Navarro-Mesa, G. Juliá-Serdá, J. F. Kraemer, N. Wessel, and A. G. Ravelo-García, "Heart rate variability feature selection in the presence of sleep apnea: An expert system for the characterization and detection of the disorder," *Computers in Biology and Medicine*, vol. 91, pp. 47 – 58, 2017.

[8] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[9] M. Nano, P. Fonseca, R. Vullings, and R. Aarts, "Measures of cardiovascular autonomic activity in insomnia disorder: A systematic review," *PLOS ONE*, vol. 12, pp. 1–31, 10 2017.

[10] C.-K. Yang, A. S. Jordan, D. P. White, and J. W. Winkelman, "Heart rate response to respiratory events with or without leg movements," *Sleep*, vol. 29, no. 4, pp. 553–556, 2006.

[11] M. J. Lado, X. A. Vila, L. Rodríguez-Liñares, A. J. Méndez, D. N. Olivieri, and P. Félix, "Detecting sleep apnea by heart rate variability analysis: Assessing the validity of databases and algorithms," *Journal of Medical Systems*, vol. 35, pp. 473–481, Aug 2011.

[12] C. Heneghan, "St. Vincent's University Hospital/University College Dublin Sleep Apnea Database," 2011. http://physionet.org/physiobank/database/ucddb/.

[13] C. Iber, "Are we ready to define central hypopneas?," *Sleep*, vol. 36, no. 3, pp. 305–306, 2014.

[14] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ecg and respiratory effort," *Physiological Measurement*, vol. 36, no. 10, p. 2027, 2015.

[15] D. Widjaja, J. Taelman, S. Vandeput, M. A. Braeken, R. A. Otte, B. R. V. den Bergh, and S. V. Huffel, "Ecg-derived respiration: Comparison and new measures for respiratory variability," in *2010 Computing in Cardiology*, pp. 149–152, Sept 2010.

[16] A. J. Camm, M. Malik, J. Bigger, G. Breithardt, S. Cerutti, R. Cohen, P. Coumel, E. Fallen, H. Kennedy, R. Kleiger, *et al.*, "Heart rate variability: standards of measurement, physiological interpretation and clinical use.," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[17] A. Hossen, D. Jaju, B. Al-Ghunaimi, B. Al-Faqeer, T. Al-Yahyai, M. Hassan, and M. Al-Abri, "Classification of sleep apnea using wavelet-based spectral analysis of heart rate variability," *Technology and Health Care*, vol. 21, no. 4, pp. 291–303, 2013.

[18] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of dna nucleotides," *Phys. Rev. E*, vol. 49, pp. 1685–1689, Feb 1994.

[19] T. Penzel, J. W. Kantelhardt, L. Grote, J. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp. 1143–1151, Oct 2003.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, Jan 2002.