Assessment of Different Methodologies to Include Temporal Information in Classifying Episodes of Sleep Apnea Based on Single-Lead Electrocardiogram

Tim Willemen[1,2,3], Carolina Varon[2,3], Bart Haex[1,4], Jos Vander Sloten[1], Sabine Van Huffel[2,3]

[1]Department of Mechanical Engineering (Biomechanics section), KU Leuven, 3001 Leuven, Belgium
[2]Department of Electrical Engineering (STADIUS), KU Leuven, 3001 Leuven, Belgium
[3]iMinds Medical IT, 3001 Leuven, Belgium
[4]Imec, 3001 Leuven, Belgium

Abstract

Automated analysis of sleep apnea based on single-lead electrocardiogram would make screening and diagnosis much more accessible. Over the years, several algorithms have been proposed in the literature. In most of them, one or several temporal averaging techniques are used to improve classifier performance. A comprehensive comparison between those techniques however has never been published.

Four different temporal averaging techniques, as well as overlapping of segments, were independently assessed using a database of 70 night-time recordings, originally released for the Computers in Cardiology challenge in 2000. Classification was performed with an LDA classifier. Multiple problem-specific feature sets of 10 features were selected out of a complete set of 304 using a two-step approach.

Averaging classifier input features over neighboring segments led to the highest agreement values on the test set, outperforming the best automatic entry during the original competition (90.4% vs 89.4%). When combining classifier output values, an odd amount of segments should be used. Calculating features on larger segments (> 1-min) led to the worst results, possibly explained by its higher susceptibility to noise. Overlapping of segments improved overall agreement by about 1%.

1. Introduction

Sleep apnea is an under-diagnosed sleep-related breathing disorder which has an estimated prevalence of 4% in men and 2% in women [1]. Clinical diagnosis is based on a presence of five or more apneic events per hour in combination with excessive day time sleepiness that cannot better be explained by other factors. An apneic event is currently defined as a clear decrease from baseline in breathing volume of at least 10 seconds. This decrease can either be more than 50%, or it has to be in combination with at least 3% oxygen desaturation and/or an arousal from sleep. Apneic events can be obstructive (complete or partial obstruction of upper airway with surrounding soft tissue) or central (reduced or absent breathing effort). Mixed events are a combination of both, usually starting as a central event, but continuing as an obstructive event once breathing effort is reinitiated [2].

Current diagnosis is made using polysomnography or polygraphy, requiring the use of an extensive amount of sensors [3]. Annotation of apneic events is performed manually, resulting in large inter- and intra-observer variability [4]. Automated analysis based on less obtrusive and expensive sensors would reduce costs, making diagnosis much more accessible and allowing for cost-effective population-wide screening.

This paper discusses automated analysis of sleep apnea based on single-lead electrocardiogram (ECG). More specifically, it focuses on the assessment of different methodologies to include temporal information, i.e. looking both backward and forward in time in order to improve classification performance. Many authors have used one or several of these techniques to improve their classifier's performance, but a comprehensive comparison between them has never been published.

2. Methodology

The database used in this study was released online at PhysioNet on the occasion of a scientific competition held during the Computers in Cardiology conference in 2000 [5]. It contains 70 night-time recordings of single-lead ECG, simultaneously recorded with full polysomnography to provide expert annotations on the presence of apneic events according to clinical standards. These annotations were transformed by the competition organizers to reflect the presence of sleep apnea in 1-minute segments, leading to the final dataset comprising 34313 minutes of annotated data, split up in 35 training and 35 test nights.

First, RR interval signals were computed from the

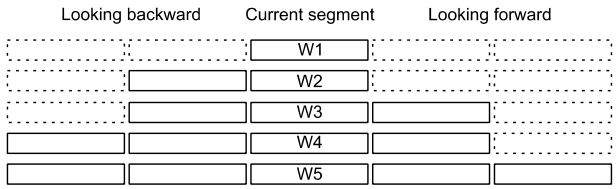

Figure 1. The five combinations of neighboring segments assessed for every temporal information methodology.

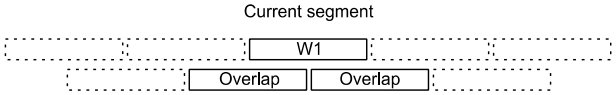

Figure 2. Overlapping segments with 30-second shifts.

ECG signals using Pan-Tompkins [6], while ECG-Derived Respiration signals were calculated from the 200 ms median filtered ECG signal by integrating the enclosed area underneath the QRS complexes, changing under the influence of respiration. Next, for every 1-minute segment of RR and EDR data, both time and frequency domain features were calculated, similarly as in [7]. Calculating features on detrended RR and EDR signals, as well as on the original traces, and adding optional logarithmic transformation and night-specific normalisation, lead to a total feature set of 304 features per 1-minute segment.

Classification was performed using Linear Discriminant Analysis [8], applying gridsearch in order to train its parameters (prior class probability and covariance matrix regularization), both having possible values between 0 and 1.

In order to reduce computational cost, complexity and noise on the classification result, unique subsets of features were selected on the training set as being most descriptive for their respective targeted classification problems (i.e. the different temporal information methodologies). Feature selection was performed in two steps. In step one, features having a single-feature classification accuracy of Cohen's kappa [9] smaller than 0.20 were disregarded from the feature set. In step two, a greedy forward selection algorithm extracted 10 features based on a maximal Cohen's kappa criterion.

To prevent overfitting on the training data, a double cross-validation scheme was introduced. The first cross-validation layer allowed for a robust selection of features, splitting up the training set 10 times at random into 2/3 training data and 1/3 validation data, averaging Cohen's kappa values over these 10 folds. A second similar cross-validation layer, within the training data from each of the 10 first cross validation splits, allowed for a robust gridsearch optimization of the LDA classifier parameters.

Four different temporal information methodologies (i.e. looking both backward and forward in time in order to improve classification performance) were investigated. First, feature values were calculated over wider segments (> 1-minute) around every 1-minute segment, shifting these wider segments with 1-minute steps (WE = window extension). Second, feature values were averaged over neighbouring segments (FA = feature averaging). Third, classifier output values (+1 and -1) were averaged over neighbouring segments using majority vote, prioritizing +1 (apnea) over -1 (normal) in case of tied results (RA = result averaging). Fourth, feature values of both current segment as well as neighbouring segments were used as input for the classifier (FE = feature extension). The amount of neighbouring segments used (looking backward and forward), varied between 0 and 4. Five different combinations were assessed for every temporal information methodology, as visualized in Figure 1.

Finally, apart from the other four temporal information methodologies, the effect of overlapping 1-minute segments was investigated by shifting 1-minute segments with 30-sec steps instead of 1-minute steps, as visualized in Figure 2. This to prevent misdetection of apneic events situated at the boundary between two segments. Classifier output values of these three overlapping segments were averaged by majority vote.

Table 1. List of the 10 greedy forward selected features for standard W1 segments.

#	**W1(all methods)**
1	log MAD EDR; mean(abs(x-mean))
2	EDR relative VLF energy (0.003-0.04 Hz)
3	RR serial correlation coefficient for k = 3
4	log MAD RR; mean(abs(x-mean))
5	norm 10 percentile of 600 s detrended RR
6	log EDR relative VLF energy (0.003-0.04 Hz)
7	RR Fractal Alan factor for k = 10
8	norm log RR absolute VLF energy (0.003-0.04 Hz)
9	log RR absolute LF energy (0.04-0.15 Hz)
10	RR relative VLF energy (0.003-0.04 Hz)

log → ln(1+x); norm → (x-median)/mad

3. Results

Box plots of the 10 fold classification agreement values on the training set (left) and agreement values on the test set (right) are displayed in Figure 3, for each of the four temporal information methodologies and for each of the five neighboring segment combinations. Looking at the training set results, an overall increase in agreement can be seen for all temporal information methodologies when the amount of incorporated neighboring segments is increased ($p < 0.01$ on both student t-test and Mann-Whitney u-test for all methodologies). For the window extension method, this increase is however saturated from W3 on, while for both the window extension method and the result averaging method, an even amount of segments (W2, W4) causes a drop in agreement compared to an odd amount of segments. Agreement values for the feature

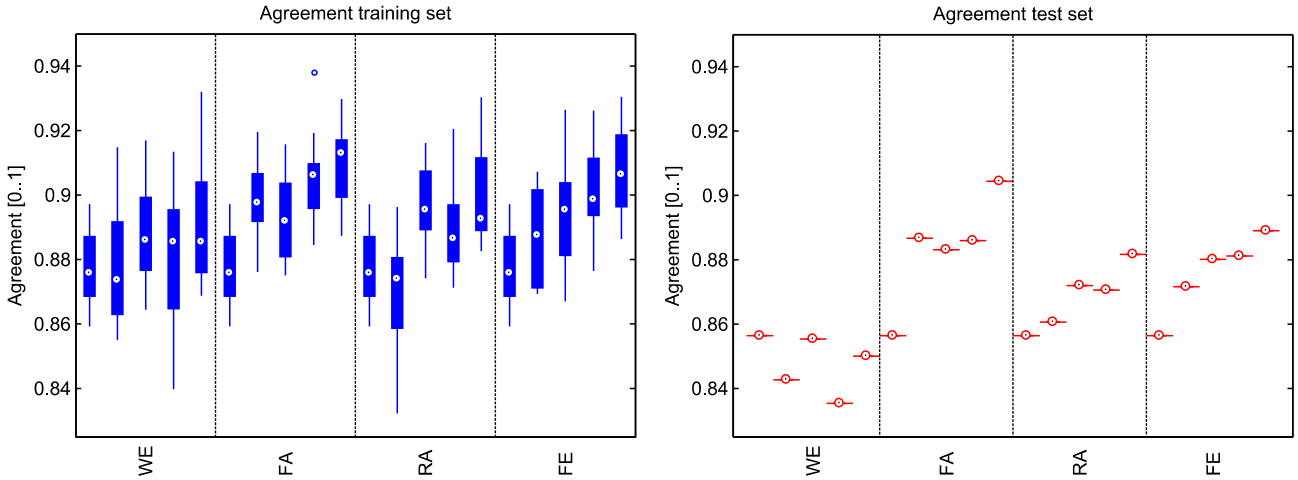

Figure 3. Box plots of the 10 fold classification agreement values on the training set (left) and agreement values on the test set (right) for each of the four temporal information methodologies (Window Extension, Feature Averaging, Result Averaging, Feature Extension). The five combinations of neighboring segments assessed (W1 through W5) are plotted from left to right.

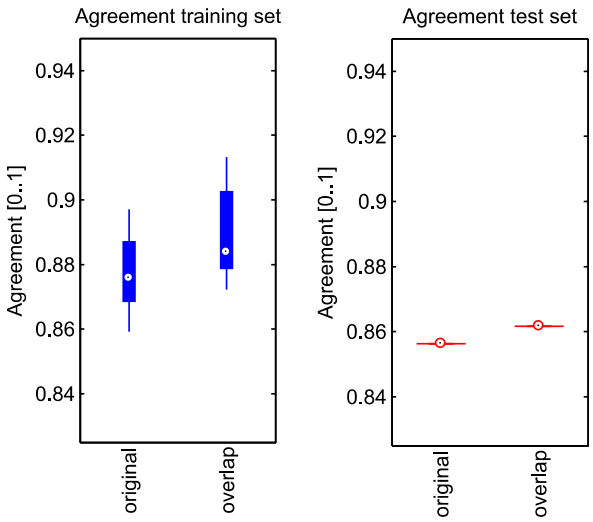

Figure 4. Box plots of the 10-fold classification agreement values on the training set (left) and agreement values on the test set (right) for standard W1 segments (original) and overlapping 1-minute segments (overlap).

averaging and feature extension methods are the highest, with slightly higher agreement for the feature averaging method.

Looking at the test set, results are similar as on the training set, except for the window extension method where a drop in agreement can be seen, especially for an even amount of segments. Agreement values are again the highest for the feature averaging method, reaching as high as 90.44% with W5.

Overall, single-feature classification accuracies (not shown) improved the most for result averaging, especially for an odd number of segments (with W5, 255 of 304 features improve), followed by feature averaging (163 of 304 features), feature extension (149 of 304 features) and window extension (137 of 304 features). For the latter three, mainly more robust features improved (percentiles, frequency content location values, logarithmic transform), while others degraded.

Table 1 lists the 10 selected features for W1. The feature sets for W5 are similar in all temporal information methodologies, except for the absence of VLF energy features in the window extension method, which are replaced by more robust frequency location values of energy content. Classification agreement values on training and test set, with all methods using the same feature set W1 instead of their method-specific trained versions, show similar trends to those shown in Figure 3. Saturation of increasing agreement with amount of incorporated neighboring segments is now present however from W3 on for all temporal information methodologies, leading to slightly lower agreement values (about 1% for W3, about 2% for W5).

The result of the overlapping 1-minute segment methodology compared to standard W1 segments is shown in figure 4; left the boxplots of the 10 fold classification agreement values on the training set, right the agreement values on the test set. Overlapping of segments increases agreement in both training and test set by 1%, although the difference between both is not significant ($p = 0.07$ on a student t-test, $p = 0.12$ on a Mann-Whitney u-test).

4. Discussion

This paper discussed different methodologies to

include temporal information in a classification procedure in order to improve overall performance of the classifier. Results on the training and test data show that the feature averaging method led to the highest agreement values, while the window extension methods led to the worst. Agreement values are similar as those reached by submitted entries during the scientific Computing in Cardiology competition, as reported in [10]. The test set result of the feature averaging method (with W5) even outperformed the highest results of all automatic classifiers submitted during the competition (90.44% vs 89.4%). While the amount of neighboring segments used was limited to four, it seems worthwhile to investigate even higher amounts, since in [11] a maximal accuracy was reached with a total window size of 7 segments using a similar feature averaging approach.

The worst performance of the window extension method can possibly be explained by a higher susceptibility to noise. While other methods average out noisy features or classifier output over possibly noise-free neighboring segments, in the window extension method, features of neighboring noise-free segments will also be affected. It explains why more robust versions of features are preferred when window width increased (up to W5). The window extension method also shows the largest difference between training and test set agreement, indicating higher overfitting and again higher susceptibility to noise present in the test set.

The result averaging and window extension method showed drops in agreement when an even amount of segments was used. This can be explained by the possible ties in output class (+1 or -1) when averaging the classifier output over the neighboring segment outputs. Since priority is given to +1 (apnea) in these cases, it leads to an increased amount of false positives. Given priority to -1 (no apnea) would in the same way lead to an increased amount of false negatives. For this reason, the use of an odd amount of segments should be preferred.

While overlapping of segments did improve overall agreement slightly, it suffers from the same problem as described in the paragraph above. The final segment output needs to be calculated by averaging over three classifier output values (one from the segment itself, and two from the left and right overlapping segments). This can again lead to the introduction of false positives when neighboring segments both contain apnea but the middle one does not. However, since apneic events are mostly occurring in a repetitive pattern [2], the amount of false positives introduced this way is limited.

Future work will consist of improving the calculation of the RR interval and EDR signals by applying more robust methodologies (e.g. kPCA for EDR calculation [12]), using the LDA-based extracted feature sets in a more versatile LS-SVM classifier [13], and increasing the amount of neighboring segments as mentioned earlier.

Acknowledgements

Research supported by Research Council KUL: GOA MaNet, PFV/10/002 (OPTEC), several PhD/postdoc & fellow grants; iMinds: SBO dotatie 2013, ICON: NXT_Sleep, Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, 2012-2017).

References

[1] Young T, Palta M, Dempsey J, Skatrud J, Weber S, Bader S. The occurrence of sleep disordered breathing in middle-aged adults. New Engl J Med 1993;328:1230-5.

[2] Task Force of the American Academy of Sleep Medicine. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. Sleep 1999;22:667-89.

[3] Iber C, Ancoli-Israel S, Chesson AL, Quan SF. The AASM manual for the scoring of sleep and associated events. Whestchester, IL: AASM, 2007.

[4] Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, Surovec S, Nieto FJ. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21:749-57.

[5] Penzel T, Moody GB, Mark RG, Goldberger AL, Peter JH. The apnea-ECG database. Computers in Cardiology 2000;27:255-8.

[6] Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE Trans on Biomed Eng 1985;32:230-6.

[7] Bsoul M, Minn H, Tamil L. Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG. IEEE Trans Inf Technol Biomed 2011;15:416-27.

[8] Ripley BD. Pattern recognition and neural networks. Cambridge, UK: Cambridge Univ Press, 1996.

[9] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

[10] Penzel T, McNames J, Murray A, de Chazal P, Moody G, Raymond B. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. Med Biol Eng Comput 2002;40:402-7.

[11] Mendez O, Corthout J, Van Huffel S, Matteucci M, Penzel T, Cerutti S, Bianchi AM. Automatic screening of obstructive sleep apnea from the ECG based on empirical mode decomposition and wavelet analysis. Physiol Meas 2010;31:273-89.

[12] Widjaja D, Varon C, Caicedo A, Suykens J, Van Huffel S. Appication of kernel principal component analysis for single lead ECG-Derived Respiration. IEEE Trans Biomed Eng 2012;59:1169-76.

[13] De Brabanter K, Karsmakers P, Ojeda F, Alzate C, De Brabanter J, Pelckmans K, De Moor B, Vandewalle J, Suykens J. LS-SVMlab toolbox user's guide version 1.8. Internal report 2010:10-146.

Address for correspondence.

Tim Willemen.
Celestijnenlaan 200C PB 2419, B-3001 Leuven, Belgium.
tim.willemen@kuleuven.be