2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871396

Convolutional Neural Networks for Apnea Detection from Smartphone Audio Signals: Effect of Window Size

Yolanda Castillo-Escario, Lorin Werthen-Brabants, Willemijn Groenendaal, Dirk Deschrijver, *Senior Member, IEEE*, and Raimon Jané, *Senior Member, IEEE*

Abstract— **Although sleep apnea is one of the most prevalent sleep disorders, most patients remain undiagnosed and untreated. The gold standard for sleep apnea diagnosis, polysomnography, has important limitations such as its high cost and complexity. This leads to a growing need for novel cost-effective systems. Mobile health tools and deep learning algorithms are nowadays being proposed as innovative solutions for automatic apnea detection. In this work, a convolutional neural network (CNN) is trained for the identification of apnea events from the spectrograms of audio signals recorded with a smartphone. A systematic comparison of the effect of different window sizes on the model performance is provided. According to the results, the best models are obtained with 60 s windows (sensitivity=0.72, specificity=0.89, AUROC=0.88). For smaller windows, the model performance can be negatively impacted, because the windows become shorter than most apnea events, by which sound reductions can no longer be appreciated. On the other hand, longer windows tend to include multiple or mixed events, that will confound the model. This careful trade-off demonstrates the importance of selecting a proper window size to obtain models with adequate predictive power. This paper shows that CNNs applied to smartphone audio signals can facilitate sleep apnea detection in a realistic setting and is a first step towards an automated method to assist sleep technicians.**

Clinical Relevance— **The results show the effect of the window size on the predictive power of CNNs for apnea detection. Furthermore, the potential of smartphones, audio signals, and deep neural networks for automatic sleep apnea screening is demonstrated.**

I. Introduction

Sleep apnea is one of the most common sleep disorders, affecting 25-50% of the adult population, especially elderly and obese individuals [1]. It is characterized by the occurrence of breathing pauses (apneas) or reductions in the airflow (hypopneas) during the night, leading to intermittent hypoxia and awakenings. Sleep apnea is associated with increased cardiovascular and cerebrovascular morbidity and mortality [2], but also an increased risk of motor vehicle accidents related to sleepiness. However, despite the high prevalence and serious consequences of this disease, most patients remain undiagnosed and untreated.

The gold standard for sleep apnea diagnosis is polysomnography (PSG) in a sleep laboratory. PSG measures multiple physiological signals during the night. Then, trained sleep specialists evaluate these signals to obtain the apnea-hypopnea index (AHI), which is the number of apneas and hypopneas per hour of sleep. Patients can be classified as normal (AHI<5), mild sleep apnea (5≤AHI<15), moderate sleep apnea (15≤AHI<30), and severe sleep apnea (AHI≥30) [3]. Nevertheless, PSG has serious limitations, including its high cost and complexity, the discomfort for the patient, long waiting lists, and the fact that the AHI is measured only for one night, which may not necessarily be representative. For these reasons, there is a need for alternative cost-effective methods for sleep apnea detection.

Much research has been devoted to the development of portable systems for home sleep monitoring [4]. Such devices usually rely on a simplified approach, measuring only one or a few signals that provide a high clinical value. Some of these signals are respiratory flow, thoracic effort, oxygen saturation (SpO_2), or audio signals [4]. The analysis of acoustic breathing and snoring signals has emerged as a promising approach for sleep apnea monitoring since they only require a low-cost sensor (a microphone) and can be used to detect apneas and hypopneas as an absence or reduction in sound [5], [6]. Audio signals can even be recorded with the built-in microphones of smartphones. Indeed, due to their wide availability and their integrated sensors, smartphones can be powerful tools for mobile health (mHealth) applications, including sleep monitoring [7]. Several attempts have been made to screen or monitor sleep apnea at home using smartphones [8]–[11], although they are not yet clinically validated.

Most of the approaches presented in the literature for automatic apnea detection use interpretable rule-based algorithms. Instead, other approaches employ machine learning or deep learning techniques, due to their increased learning capabilities when facing complex patterns. Specifically, the advances in deep learning have led to the possibility of building models without human-engineered features. Deep learning is gaining higher interest for sleep

*Research supported by "La Caixa" Foundation (ID 100010434) under fellowship code LCF/BQ/DE18/11670019, the Government of Catalonia (CERCA Program), the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya (GRC 2017 SGR 01770), the Spanish Ministry of Science, Innovation and Universities (RTI2018-098472-B-I00 MCIU/AEI/FEDER), and the Flemish Government (Flanders AI research programme).

Y. Castillo-Escario and R. Jané are with the Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Institute for Bioengineering of Catalonia (IBEC), the Barcelona Institute of Science and Technology (BIST), and Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain (e-mail: {ycastillo, rjane}@ibecbarcelona.eu).

L. Werthen-Brabants and D. Deschrijver are with Ghent University - IMEC, IDLab Technologiepark-Zwijnaarde 126, B-9052 Ghent, Belgium (e-mail: {lorin.werthenbrabants, dirk.deschrijver}@ugent.be)

W. Groenendaal is with IMEC The Netherlands/Holst Centre, Eindhoven 5656AE, The Netherlands (e-mail: willemijn.groenendaal@imec.nl)

apnea detection, and multiple examples can be found in the literature, mostly using convolutional neural networks (CNN), long short-term memory (LSTM) networks, or recurrent neural networks (RNN) from different signals: respiratory signals [12], electrocardiogram (ECG) [13], SpO_2 [14], chest bioimpedance [15], or tracheal sound [16].

In a previous work, a rule-based algorithm was presented, based on entropy analysis of acoustic signals recorded with a smartphone for home sleep apnea diagnosis [10]. The aim of this paper is to test the potential of CNNs for apnea detection from smartphone audio signals. The main reason for choosing CNNs is that there is no need for feature extraction, as they can take images as inputs, and thus directly extract information from the time-frequency maps of audio signals. On the other hand, a key factor in machine and deep learning models, and especially for sleep apnea detection, is the size of the window to segment the recordings, which usually depends on the signal type and the application. For these reasons, the objectives are: 1) to build a CNN model for the identification of apnea events in audio signals recorded with a smartphone, and 2) to analyze the effect of different window sizes in the model performance.

II. Materials and Methods

A. Dataset

This study uses real-life data gathered from 25 patients (14 men, 11 women, mean±std age: 56±16) enrolled for an overnight sleep test. The study was approved by the Ethics Committee of Hospital Clínic de Barcelona (protocol code HCB/2017/0106). Informed consent was obtained from all participants. 16 patients underwent in-lab PSG, while 9 patients underwent a home sleep apnea test with ResMed ApneaLink Air^{TM}. PSG recordings contained multiple channels, including respiratory signals at a sampling rate (fs) of 32 Hz (nasal cannula, thermistor, and thoracic and abdominal effort), single-lead electrocardiogram (fs=256 Hz), and SpO_2 (fs=1 Hz). ApneaLink measured respiratory flow through a nasal cannula (fs=100 Hz), thoracic movement (fs=10 Hz), and SpO_2 (fs=1 Hz). Simultaneously, overnight audio recordings were acquired (fs=48 kHz) with the built-in microphone of a smartphone (Samsung Galaxy S5) that was placed over the subjects' thorax using an elastic band. This configuration had been successfully tested in previous studies [10], [17]. The mean recording length was 6.5±1.4 hours.

The smartphone and the reference system (either PSG or ApneaLink) were synchronized based on timestamps. Data from the reference system were annotated by trained sleep specialists. Following the AASM guidelines [3], 2 subjects had normal AHI, 3 mild AHI, 13 moderate AHI, and 7 severe AHI.

B. Signal Processing and Representation

Audio signals were downsampled to 5 kHz, applying an anti-aliasing low-pass filter. To reflect a realistic measurement and check the robustness of the model when working with raw data, the full-night audio signals were used, and no artifact detection or signal cleaning was performed.

The experimental setup was chosen to assess the feasibility of a CNN trained with time-frequency representations of audio signals for the prediction of apnea events. Recordings were split into non-overlapping sections of different lengths (10, 20, 30, 45, 60, 90 and 120 s), to investigate the effect of a varying window size. Each section was labeled as 'apnea' if the central point of that section was inside one of the apnea events annotated by the sleep experts, and as 'non-apnea' otherwise. For this preliminary study, all apneas (obstructive, central, or mixed) were considered, whereas hypopneas were excluded from the analysis. The spectrogram of each section was calculated, with a window length of 0.1 s, 25% overlap, and 512 points for the fast Fourier transform computation. The spectrogram magnitude was log-transformed for a better visualization because breath and snore sounds have more components in the low-frequency bands [5]. Then, the spectrograms were standardized to have zero mean and unit variance, in order to ensure that all characteristics in the sound signal were visible regardless of signal amplitude. An example of the proposed time-frequency representation for an apnea and a non-apnea section can be seen in Fig. 1.

C. Modeling

A custom CNN model was built in Python, using Tensorflow and Keras libraries, to automatically extract time-frequency domain features from the audio spectrograms and classify them into apnea or non-apnea events. The network architecture and model hyperparameters were set after some preliminary testing and based on previous experience. As shown in Fig. 2, the proposed deep neural network architecture consists of five consecutive convolutional blocks followed by a flattening layer, a fully connected layer with 50 neurons and *ReLU* activation, a dropout layer with a dropout rate of 0.4 to reduce overfitting, and finally a fully connected layer with a *softmax* activation function, whose output can be interpreted as the probabilities of 'apnea' and 'non-apnea'. Each convolutional block is composed of a 2D convolutional layer with 100 output filters, a kernel size of 3x3, and *ReLU* as activation function, followed by a 2D max-pooling layer with pool size of 2x2, and a dropout layer with a dropout rate of 0.4.

The networks were trained with a batch size of 16 using *Adam* optimization algorithm with a learning rate of 0.001 and categorical cross entropy as loss function. Early stopping was employed to determine the number of training epochs and prevent overfitting, by saving the model with the best area under the precision recall (PR) curve (AUPRC) in the validation set, with a patience of 50 epochs.

D. Evaluation

The 25 subjects were divided into training and validation sets. Data from 17 subjects were used for training, and from the remaining 8 subjects for validation. It was ensured that data from the same subject were not simultaneously present in both groups and thus the performance of the models was evaluated on data from new, previously unseen patients. As the number of apnea events was much lower than that of non-apnea, the dataset was highly imbalanced. If the models were trained with

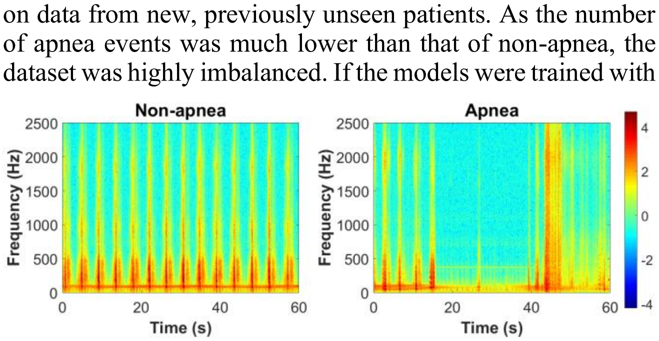

Figure 1. Example of spectrograms of non-apnea and apnea events.

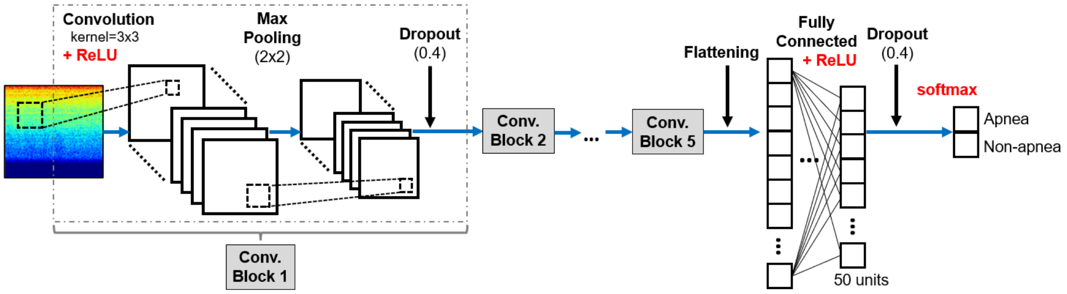

Figure 2. The proposed convolutional neural network architecture.

the original data distribution, they would be biased towards predicting non-apnea events. However, to avoid this problem, the majority class was randomly subsampled in order to obtain a training set that is balanced (equal number of apnea and non-apnea events). For evaluation purposes, though, it is important to keep the original imbalanced data to reflect a realistic measurement. When evaluating the performance of the model, all possible data from the patients in the validation set were retained, in this case using overlapping windows 1 second apart (1 s stride), so that a prediction for each second of the recording can be made.

For each window size (ranging from 10 s to 120 s), audio spectrograms were calculated, a CNN model was trained, and then it was used to predict the labels of the validation set. A decision threshold of 0.5 was used to determine the apnea/non-apnea classifications. The metrics that were used to assess the performance of the models in the validation set included the sensitivity (Se, also known as recall), specificity (Sp), positive predictive value (PPV, also known as precision), negative predictive value (NPV), and accuracy (Acc). The combinations of Se and Sp for different decision thresholds can be summarized in a receiver operating characteristic curve (ROC). The area under this curve (AUROC), and the AUPRC, which is often preferred for imbalanced classification, were also calculated.

TABLE I. Performance of models for varying window size

Window	Se	Sp	PPV	NPV	Acc	AUROC	AUPRC
10 s	0.667	0.514	0.050	0.976	0.520	0.626	0.058
20 s	0.540	0.532	0.042	0.968	0.533	0.588	0.063
30 s	0.572	0.859	0.133	0.982	0.849	0.754	0.186
45 s	0.570	0.917	0.207	0.983	0.905	0.802	0.283
60 s	0.718	0.891	0.201	0.988	0.885	0.877	0.522
90 s	0.668	0.913	0.225	0.986	0.904	0.846	0.286
120 s	0.593	0.931	0.246	0.984	0.918	0.813	0.243

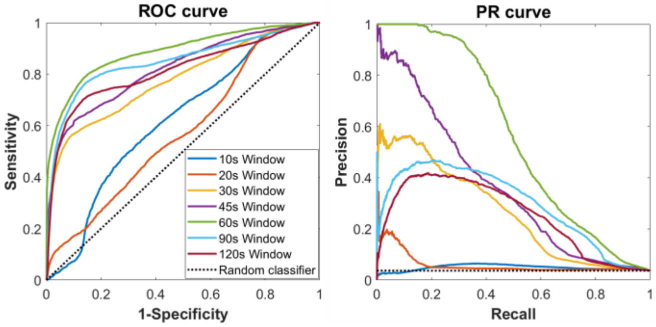

Figure 3. ROC and PR curves for each window size. The dotted lines indicate the performance of a model with random predictions based on the imbalanced data, clearly outperformed by the CNN models.

III. Results and Discussion

The performance of the models for each window size is summarized in Table I and in the ROC and PR curves in Fig. 3. The results suggest that the best window size for the detection of apnea events with this CNN model is 60 s, as indicated by the highest AUPRC, AUROC, and Se, while having good values of Sp, NPV, and Acc. With small window sizes (10 s and 20 s), results are much worse, presumably because the windows are shorter than most apnea events and sound reductions cannot be captured. The model performance rises for windows longer than 30 s, reaching the best AUPRC at 60 s. For windows of 90 and 120 s, the predictive power of the model drops because these windows are too long and may contain multiple respiratory events of different origin.

The model with 60-s windows demonstrates a remarkable Sp, NPV, Acc, and AUROC. The Se is 72%, due to the epoch-by-epoch comparison, and the fact that the exact start and end times of apneas do not perfectly match the manual annotations. Most apneas in the validation set (94%) are correctly detected by the CNN model, albeit sometimes slightly shifted. This is also because human annotations of the position of apneas are not exact to the second. Additionally, human annotations were based on signals of different origin and using several channels (respiratory flow, thoracic effort, oxygen saturation…), whereas the CNN model is trained only with audio signals. Nonetheless, detecting 94% of the apneas is a slight improvement from the 89% reported in a previous paper based on a rule-based algorithm for sleep apnea diagnosis from smartphone audio signals [10].

Although most of the metrics are reasonably good, the PPV is quite low for all the window sizes, which also affects the AUPRC. This is a common problem when dealing with highly imbalanced data [18]. The models were trained with balanced classes, leading to a very good performance in that scenario (for the 60-s window: Se=0.92, Sp=0.90, PPV=0.91, NPV=0.92, Acc=0.91). However, when moving to the original distribution, the ratio between apneas and non-apneas is almost 1:50, and thus the PPV is seriously affected. Most false positives (FP) occur in healthy subjects or patients with mild sleep apnea, especially in regions where there are sound artifacts or the quality of the audio signal was poor. A pre-processing step for cleaning the signal, discarding artifacts and low-quality regions, could help to improve the performance of the model. Moreover, typically, only events longer than 10 s are annotated, but the models make decisions on a per-second basis and thus can lead to isolated or short FP. A post-processing step can help to remove these FP. The combination of audio with other signals (e.g., oxygen saturation or thoracic movement) could also contribute to increase the model precision and could easily be integrated in the smartphone system.

To test the potential benefits of overlapping windows as a method for data augmentation to improve the generalization capabilities of the model, the best model (60 s windows) was trained when applying a 59 s overlap (1 s stride). The results with that model were: Se=0.880, Sp=0.775, PPV=0.130, NPV=0.994, Acc=0.779, AUROC=0.910, AUPRC=0.456. Using a large window overlap, more data is available to train the model, but this has a higher computational cost. In the experiment, when using overlapping windows, the model predicted more apnea events and, therefore, the Se notably increased, but at the cost of more FP and hence a lower Sp and PPV. Future investigations will focus on the effect of window overlap, to robustly determine the best strategy in this regard.

IV. Conclusion

Sleep apnea is a highly prevalent sleep disorder, but most patients remain undiagnosed and untreated. One of the reasons is the limitations in current diagnostic techniques. Here we propose a CNN for detecting apnea events from time-frequency representations of audio signals recorded with a smartphone. We also provide a systematic comparison of different window sizes, showing the effect of this parameter in the performance of the model.

The best predictive power was obtained when the model was trained with audio spectrograms from 60-s windows, a size that was long enough to capture the sound reductions corresponding to apneas, without including multiple events. These results should be corroborated in future studies with a higher sample size to validate the generalization capabilities of the model with data from new patients and extract more robust conclusions. Other time-frequency methods could be tested, to check if they can improve the predictions. Model complexity could also be increased to include hypopneas and other respiratory events in the predictions, while the combination of audio with other signals could help to further guide the algorithm and improve its accuracy. However, this proof-of-concept shows the potential of CNN from smartphone audio signals as a simple, automatic, and cost-effective tool to facilitate sleep apnea detection in a realistic setting, and thus assist sleep technicians in the monitoring and management of patients.

Acknowledgment

The authors would like to thank Dr. Josep Maria Montserrat and the technicians from the Hospital Clínic sleep laboratory for their help in patient recruitment and data acquisition and annotation.

References

[1] R. Heinzer *et al.*, “Prevalence of sleep-disordered breathing in the general population: THE HypnoLaus study,” *The Lancet Respiratory Medicine*, vol. 3, no. 4, pp. 310–318, Apr. 2015.

[2] A. Yoshihisa and Y. Takeishi, “Sleep Disordered Breathing and Cardiovascular Diseases.,” *Journal of atherosclerosis and thrombosis*, vol. 26, no. 4, pp. 315–327, Apr. 2019.

[3] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. v Vaughn, *The AASM manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications.*, vol. 176, 2012.

[4] M. Ahmed, N. P. Patel, and I. Rosen, “Portable Monitors in the Diagnosis of Obstructive Sleep Apnea,” *Chest*, vol. 132, no. 5, pp. 1672–1677, Nov. 2007.

[5] D. Pevernagie, R. M. Aarts, and M. de Meyer, *The acoustics of snoring*, vol. 14, no. 2, pp. 131–144, Apr. 2010.

[6] H. Jin *et al.*, “Acoustic Analysis of Snoring in the Diagnosis of Obstructive Sleep Apnea Syndrome: A Call for More Rigorous Studies.,” *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, vol. 11, no. 7, pp. 765–71, Jul. 2015.

[7] E. Fino and M. Mazzetti, *Monitoring healthy and disturbed sleep through smartphone applications: a review of experimental evidence*, vol. 23, no. 1, pp. 13–24, Mar. 2019.

[8] J. Behar *et al.*, “SleepAp: An Automated Obstructive Sleep Apnoea Screening Application for Smartphones,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 325–331, Jan. 2015.

[9] M. Al-Mardini, F. Aloul, A. Sagahyroon, and L. Al-Husseini, “Classifying obstructive sleep apnea using smartphones,” *Journal of Biomedical Informatics*, vol. 52, pp. 251–259, Dec. 2014.

[10] Y. Castillo-Escario, I. Ferrer-Lluis, J. M. Montserrat, and R. Jané, “Entropy analysis of acoustic signals recorded with a smartphone for detecting apneas and hypopneas: A comparison with a commercial system for home sleep apnea diagnosis,” *IEEE Access*, vol. 7, pp. 128224–128241, Sep. 2019.

[11] I. Ferrer-Lluis, Y. Castillo-Escario, J. M. Montserrat, and R. Jané, “SleepPos App: An Automated Smartphone Application for Angle Based High Resolution Sleep Position Monitoring and Treatment,” *Sensors*, vol. 21, no. 13, p. 4531, Jul. 2021.

[12] T. Van Steenkiste, W. Groenendaal, D. Deschrijver, and T. Dhaene, “Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2354–2364, Nov. 2019.

[13] U. Erdenebayar, Y. J. Kim, J. U. Park, E. Y. Joo, and K. J. Lee, “Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram,” *Computer Methods and Programs in Biomedicine*, vol. 180, Jul. 2019.

[14] F. Vaquerizo-Villar *et al.*, “A Convolutional Neural Network Architecture to Enhance Oximetry Ability to Diagnose Pediatric Obstructive Sleep Apnea,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 2906-2916, Aug. 2021.

[15] T. Van Steenkiste *et al.*, “Portable Detection of Apnea and Hypopnea Events Using Bio-Impedance of the Chest and Deep Learning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2589–2598, Sep. 2020.

[16] H. Nakano, T. Furukawa, and T. Tanigawa, “Tracheal sound analysis using a deep neural network to detect sleep apnea,” *Journal of Clinical Sleep Medicine*, vol. 15, no. 8. pp. 1125–1133, Aug. 2019.

[17] H. Nakano *et al.*, “Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: Proof of concept,” *Journal of Clinical Sleep Medicine*, vol. 10, no. 1, pp. 73–78, Jan. 2014.

[18] P. Ranganathan and R. Aggarwal, “Common pitfalls in statistical analysis: Understanding the properties of diagnostic tests – Part 1,” *Perspectives in Clinical Research*, vol. 9, no. 1, p. 40, Jan. 2018