Continuous Mobile Audio Monitoring for Sleep Apnea Detection

Jing Han, *Senior Member, IEEE*, Tong Xia, and Cecilia Mascolo

***Abstract*— Audio-based sleep apnea detection methods hold great potential to improve access to diagnosis, by providing unattended sleep apnea screening at home via sound collected from mobile sensors during sleep. Our research involved a thorough comparison and evaluation of tracheal and ambient microphone recordings for sleep apnea detection with different granularities. Utilising a variety of acoustic representations and sophisticated deep learning architectures, we performed an extensive analysis on the open PSG-Audio dataset, which encompasses over 850 hours of audio data from 194 subjects. For sleep apnea classification, the most effective model showed a 90.8 % accuracy in detecting sleep apnea, 83.3 % accuracy when hypopneic and apneic events were detected separately, and 75.7 % accuracy when apneic events were further divided into three sub-categories. On overnight recordings, the model achieved a sensitivity of 0.93 and a specificity of 1.0 for moderate sleep apnea screening, and a sensitivity of 0.84 and a specificity of 0.97 for severe sleep apnea screening. This research also provided a unique study to compare and combine respiratory sounds from two different types of sensors for sleep apnea detection. The high performance of our model provides a promising avenue for enabling remote diagnosis and monitoring of sleep apnea.**

***Index Terms*— Sleep apnea detection, apnea-hypopnea index, tracheal sound, ambient sound, acoustic analysis.**

I. Introduction

Sleep disorders are very common and may significantly impact various aspects of daily life. Sleep-related breathing disorders, notably Sleep Apnea Syndromes (SAS), have a high prevalence in the community and are often significantly underdiagnosed and undertreated [1]. Obstructive sleep apnea (OSA), the most important category in SAS, has been estimated to affect nearly one billion adults aged 30–69 years worldwide while approximately 80% of cases remain undiagnosed [2]. SAS is often associated with various symptoms, including loud and habitual snoring [3], disrupted sleep [4], daytime fatigue [5], morning headaches [6], difficulty with memory or concentration [7], mood changes [8], and decreased quality of life [9]. Earlier studies have also shown that this syndrome is a substantial risk factor for severe pathological conditions such as hypertension, stroke, diabetes, cardiovascular diseases, and depression [10]–[14]. As a consequence, it is imperative to increase awareness of SAS and reduce the number of individuals with undiagnosed and untreated SAS. In particular, effective and early SAS detection strategies are needed to not only alleviate the immediate symptoms of SAS but also address the associated long-term health risks [2].

Currently, polysomnography (PSG) is considered the gold standard for SAS diagnosis. It is performed in a hospital or clinic, with multiple sensors continuously monitoring the electrophysiological and cardio-respiratory patterns during a full-night sleep, under the supervision of a specialised physician [15]. Although laboratory-based PSG can provide precise medical diagnoses, its drawbacks, including low cost-effectiveness and limited accessibility pose significant barriers to large-scale and population-level sleep screening.

In the past decade, the rapid advancement of biomedical sensing technologies has significantly transformed healthcare applications, enabling continuous, remote, and personalized monitoring across various clinical and home-based settings [16]–[19]. Particularly, to enhance SAS detection, numerous sensor-based solutions have been proposed by exploiting biomedical sensors to collect user sleep data [20]–[23]. Among these initiatives, approaches leveraging sleep sound analysis have demonstrated promising results [24]–[26]. This is because that SAS is characterised by the recurring partial or complete collapse of the upper airway. Breathing and snoring sounds during sleep can be utilised to detect the presence of these apneic and hypopneic events. Subsequently, SAS can be identified if the frequency of apnea/hypopnea events exceeds a certain threshold. These audio-based approaches hold great promise for facilitating automatic, easy, and convenient SAS detection and monitoring.

In light of recent advancements in deep learning techniques, considerable efforts have been invested in advancing sleep apnea research [27], using different signal sources such as electrocardiogram (ECG) [28], oxygen saturation (SpO2) [29], and respiration signals [30]. Particularly, when it comes to audio-based deep learning models for sleep apnea analysis, researchers have explored the analysis of either tracheal sounds (captured by a contact tracheal microphone on the neck) or ambient sounds (recorded by a non-contact microphone positioned near the bed) [31], [32]. However, a notable limitation across many of these studies is that they reported their performance exclusively on their self-collected datasets [33], [34]. These datasets are typically privately held and cannot be shared for research purposes, making it infeasible to fairly

This work was supported by ERC Project 833296 (EAR).

J. Han is with the Department of Computer Science and Technology, University of Cambridge, and also with the College of Computer Science and Electronic Engineering, Hunan University (e-mail: jh2298@cam.ac.uk).

T. Xia is in Vanke School of Public Health, Tsinghua University (e-mail: tongxia@mail.tsinghua.edu.cn).

C. Mascolo is in the Department of Computer Science and Technology, University of Cambridge (e-mail: cm542@cam.ac.uk).

compare with others. Also, existing research predominantly focuses on either tracheal or ambient sound analysis, with a noticeable absence of comprehensive investigation into the different diagnostic capabilities of the two sound types. Moreover, recent findings have pointed out that long average apnea duration is a factor for morning tiredness and hypertension [35], [36]. Additionally, a potentially important correlation has been observed between the respiratory event duration and the risk of mortality associated with apnea [37]. However, most previous work addresses the binary classification of the disease or the estimation of disease severity, and only a limited number of studies have undertaken event-by-event analyses [38]–[40]. These aforementioned issues underscore the need for more comprehensive investigations on publicly available sleep sound data, exploring both sound types and carrying out fine-grained disease analysis beyond SAS screening and severity estimation.

In this context, the goal of this paper is to explore fine-grained apneic/hypopneic event detection from overnight sleep sounds using advanced deep-learning techniques. The primary objective is to benchmark performance on a large open database, PSG-Audio [41]. The secondary objective is to conduct a detailed examination of the distinct capabilities of tracheal sounds and ambient sounds for this task. To the best of our knowledge, no existing study has yet undertaken a detailed comparison of the two sound types for this task. Additionally, we also compare the obtained performance from our tracheal/ambient-based sleep apnea detection model against other state-of-the-art works on SAS screening, demonstrating superior performance in most cases.

The contributions of this study are as follows:

- *An innovative audio-based sleep event detection pipeline.* We conducted evaluations on an open sleep dataset and presented the efficacy of our approach in detecting apnea/hypopnea events in a segment-by-segment manner.
- *Extensive evaluations on various acoustic feature sets and advanced deep-learning structures.* We demonstrate the potential of utilising a pre-trained model as a feature extractor. The integration of EBranchformer, a novel architecture combining convolution and self-attention, also brings beneficial enhancements to this task.
- *Thorough comparison of two audio sources*. We not only highlighted performance differences between the two sources but also looked into cross-evaluation scenarios. We showcased further performance gains achieved through the combination.
- *Performance Benchmarking.* We compared our work with existing SOTA studies, demonstrating superior performance across multiple metrics, affirming the effectiveness and advancements introduced by our proposed approach.

The remainder of this paper is organised as follows. In Section II, we introduce related studies. Section III describes the overall structure of the audio-based sleep apnea detection pipeline. In Section IV, we describe the experimental setups. Then, we present and discuss the results of the experiments in Section V. Finally, we provide the conclusion in Section VI.

II. Related Work

As aforementioned, the limited accessibility and high cost associated with PSG highlight the need for a more cost-effective and convenient option for SAS screening. Here, we focus on methods based on sleep sound recordings. Several sound-based studies addressing SAS screening have been published [33], [42], [44]. Tasks performed in these studies can be categorised into three tasks based on their diverse objectives: SAS screening, SAS severity prediction, and AHI estimation [33]. SAS screening involves a binary classification task, distinguishing between the presence and absence of the disease. Severity prediction, on the other hand, is a multiclass classification task, predicting the severity degrees using the AHI index, ranging from non-apnea (AHI$<$5) to mild (5$\leq$AHI$<$15), moderate(15$\leq$AHI$<$30), and severe (AHI$\geq$30) apnea. Additionally, AHI estimation tasks a more fine-grained estimation, computing the average number of apneic and hypopneic events per hour during sleep. For instance, Romero *et al.* developed a deep neural network architecture, reporting a sensitivity of 0.79 and a specificity of 0.80 in screening moderate SAS among 103 participants [33]. Likewise, Kim *et al.* employed 132 handcrafted features as acoustic biomarkers for SAS severity prediction, achieving an accuracy of 88.3% across 120 participants [42]. Additionally, Castillo-escario *et al.* analysed the sample entropy of audio signals, demonstrating a remarkable correlation coefficient of 0.99 between the real and estimated AHIs for 13 participants [44].

However, all the above-mentioned studies lack an in-depth examination of individual respiratory events, and only provide a singular diagnosis per participant based on overnight recordings. Recent studies link longer apnea episodes to serious health risks, including hypertension and increased mortality [36], [37]. Given the potential clinical implication of the temporal characteristics of abnormal respiratory events, understanding the timing of the events may offer valuable insights into disease analysis. Therefore, in this work, our focus is on developing a predictive model capable of detecting apnea and hypopnea events. This SAS event detection can subsequently facilitate various downstream tasks, including SAS diagnosis, SAS severity prediction, and AHI estimation. This will be beneficial to the follow-up and treatment of the disease, by providing a more nuanced understanding of the temporal dynamics of the respiratory events.

In addition, several studies have investigated SAS event detection using sleep sounds [38], [40], [45]–[49]. For instance, in [40], a Voice Activity Detection (VAD) algorithm was explored and achieved an accuracy of 73.6% for the detection of all varied apneic events. Likewise, in [38], a real-time epoch-by-epoch apneic event detector was proposed, yielding an accuracy of 88.8% in a three-class classification scenario (apnea, hypopnea, and no-event). In [46], sleep audio features with semantic features and employ XGBoost to classify sleep apneic events and reporting an accuracy of 77.6%. An accuracy of 66.3% was reported in [47] when a pre-trained VGG19 and the long short-term memory (LSTM) fused model was explored to dingtinguish normal snoring and apnea-hypopnea snoring of OSA patients. In another work, a CNN-based model

TABLE I: Summary of related audio-based sleep apnea studies.

	Data	Source		Task			Method	Window
	Availability	Tracheal	Ambient	Event Detection	AHI estimation	SAS Screening	(Feature + Model)	Size
[32]	✗	✓	✗	✓	✗	✗	MFCCs + HMM	60-ms
[33]	✗	✗	✓	✗	✓	✓	FBanks + DNN	30/40-second
[38]	✗	✗	✓	✓	✓	✓	Spectrogram + DNN	30-second
[39]	✗	✗	✓	✓	✓	✓	Spectrogram + DNN	5-minute
[40]	✓	✓	✓	✓	✗	✗	VAD	100-ms
[42]	✗	✗	✓	✗	✗	✓	MFCCs + DNN	5-second
[43]	✗	✓	✗	✓	✓	✓	Spectrogram + DNN	60-second
this work	✓	✓	✓	✓	✓	✓	FBanks&HuBERT + DNN	40-second

MFCCs: Mel Frequency Cepstral Coefcients, HMM: Hidden Markov Model, VAD: Voice Activity Detection,
FBanks: Mel filter banks, HuBERT: features from a pre-trained model.

with novel snore sound features and multi-task learning was proposed [48]. Compared with the studies as mentioned above, our work not only demonstrated comparable performance but, in certain aspects, exhibited better performance.

For the aim of sound-based apnea detection, audio signals can be obtained from various positions. The signals may be collected from a tracheal microphone attached to the trachea, or an ambient microphone positioned near the head of the person. While some prior works have utilised tracheal sound signals [24], [32], [43], [50], most of the others have explored ambient microphone signals [26], [31], [44]. Note that, due to concerns such as patient privacy, confidentiality, and copyright issues, most if not all prior works were evaluated using private datasets. This practice introduces challenges in direct performance comparisons due to diverse and often privately held datasets. To overcome this limitation, our research evaluates both tracheal and ambient sound types simultaneously using the publicly available PSG-audio dataset [41], enabling fair and reproducible comparison under unified experimental settings. Prior studies have compared the two sound sources on the same dataset and reported performance differences between tracheal and ambient recordings [40]. Building upon these comparative analyses, our study provides a systematic benchmark across multiple model architectures and feature representations. This could offer valuable insights for researchers and developers to assess performance differences and make informed decisions when considering other factors such as cost and comfort. In addition, previous studies have overlooked the evaluation across the two sound types; the effect of sound type mismatch for sleep apnea detection has not been investigated. Furthermore, there have been yet no studies integrating the two sound types: tracheal microphone signals and ambient microphone signals. Our study marks the first exploration of the cross-evaluation of tracheal and ambient sounds and the integration of the two for sleep apnea detection.

Earlier studies have primarily focused on leveraging feature engineering techniques to identify distinctive features for sleep apnea detection. Various sets of optimal audio features have been explored as potential acoustic biomarkers for apnea severity estimation in the literature [26], [42], [51]. In more recent developments, researchers have achieved promising results by harnessing deep learning models for this task. Specifically, without manual feature selection, Mel-spectrogram or Mel-filterbank features directly extracted from raw audio can be employed as inputs for deep models. Convolutional neural networks (CNNs) and their variants have been frequently utilised for their efficacy in capturing complex patterns from audio, showcasing enhanced capabilities for sleep apnea detection [33], [38], [39]. Recently, Transformer and its variants have emerged in audio processing, due to their capabilities to capture both spatial and temporal dependencies within sequential signals [52]–[54]. Here, we undertook an exploration of three Transformer architectures, marking the first instance of such investigation. Our research unveiled the exceptional efficacy of EBranchformer in sleep apnea detection, due to the integration of CNN blocks and self-attention mechanisms.

Furthermore, we pioneered the application of transfer learning techniques in sleep apnea detection, leveraging feature extraction from a pre-trained model HuBERT [55]. We demonstrated the effectiveness of this approach by comparing its performance with conventional features, marking a novel contribution to the field. In addition to HuBERT, other pretrained audio foundation models such as BEATs [56], PaSST [57], AST [58], and BYOL-A [59], could also be considered for feature extraction. Unlike HuBERT, which is pretrained on large-scale speech corpora, these models are trained on more diverse general audio datasets. We selected HuBERT because respiratory sounds originate in the human airway and share certain acoustic characteristics with speech, such as airflow-induced turbulence and overlapping frequency ranges.

Table I provides a summary of related studies, highlighting the distinctive contribution of the current research in relation to prior works.

III. Audio-based Sleep Apnea Detection

This study aims to establish a benchmark for audio-based methods in sleep apnea detection, utilising a publicly available sleep sound dataset. Specifically, we focus on evaluating the distinct detection capabilities of tracheal and ambient microphones, as well as the deployment of transfer learning-based features for sleep monitoring.

A. Overview

The proposed pipeline for audio-based sleep apnea detection comprises several key components, including pre-processing,

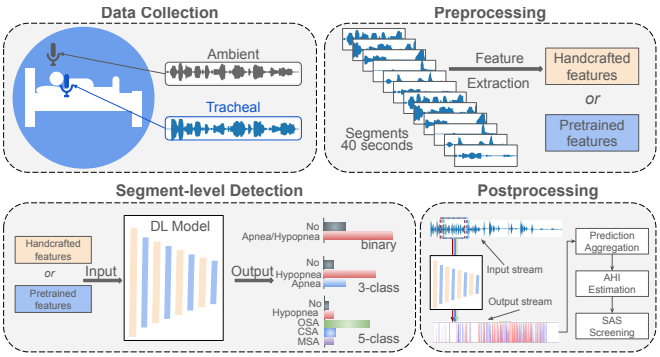


Fig. 1: An overview of our sleep apnea detection framework.

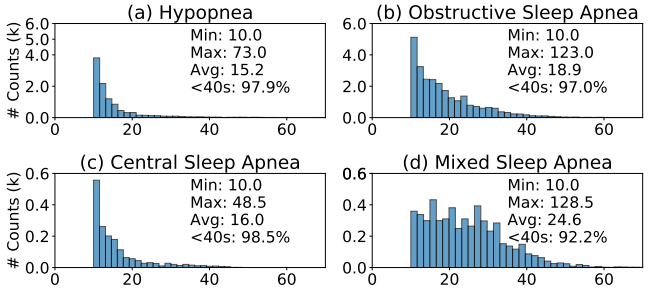


Fig. 2: Duration distribution of respiratory events.

feature extraction, classification, and post-processing stages. In the first stage, we downsampled the original sleep sound recordings and segmented the data into 40-second segments, as well as labelling these segments according to the presence or absence of a certain apnea/hypopnea event. Moving to feature extraction, multi-dimensional segment-level representations were extracted for each segment. We explored two different feature types. In the third stage, the derived representations were fed into varied deep structures for three segment-level classification tasks. Specifically, we first conducted a binary classification task, distinguishing abnormal respiratory events from normal ones. We further delved into finer-grained tasks, including discriminating between apneic and hypopneic events and detecting various types of apneic events. In the last stage of post-processing, we integrated the segment-level predictions to deliver overall estimations of the overnight recording. This final step targeted three tasks, including SAS detection, severe SAS prediction, and AHI estimation. In this way, our proposed framework not only captures the nuances of individual sleep segments but also provides an overall understanding of the overall sleep apnea profile of the participant. Fig. 1 illustrates the whole framework.

B. Signal Preprocessing and Feature Extraction

The original whole-night sleep audio recordings were initially sampled at 48 kHz. We downsampled the audio signals to 8 kHz, as the frequency of respiratory sounds is typically within the frequency range of 50-4000 Hz [60]. According to the Nyquist theorem, an 8 kHz sampling rate sufficiently preserves this frequency range. This also aligns with other established practices in the field (8 kHz used in [34, 42] and 8.82 kHz utilised in [39]). Then, we segmented the long audio recordings into 40-second segments. While an apneic event can last from at least 10 seconds to several minutes [33], [39], [41], a temporal window of 40 seconds is chosen in the present study to effectively capture the temporal patterns of respiratory events. Various window sizes have been utilised in previous studies (see Table I), and research in [33] demonstrated that a longer window size (i.e., 40 seconds) improved detection performance compared to shorter window sizes (i.e., 30 seconds). Our preliminary study aligns with these findings, indicating that a 40-second window size outperforms shorter window sizes (e.g., 10 or 30 seconds). However, increasing the window size beyond 40 seconds did not yield additional performance improvements. Consequently, we have fixed the window size at 40 seconds for the remainder of our study. The duration distributions of four distinct respiratory events on the PSG-Audio dataset are demonstrated in Fig. 2. It can be observed that, while more than 92% of the respiratory events are shorter than 40 seconds and can be covered by a single segment, there are a few cases where the events are longer and cannot be completely covered by one segment.

Following audio segmentation, we apply feature extraction to these segments. In this paper, we focus on two distinct feature types. First, we employ log Mel-filterbank features (FBank) for sleep apnea detection, which is classic and also used in many prior studies [33]. Specifically, a spectrogram is first computed from 100 ms windows with a hop size of 60 ms by applying short-time Fourier transform (STFT). These specific configurations were identified through empirical exploration and were found to produce optimal results for the task at hand. Subsequently, a bank of 80 triangular Mel-filters is applied to the raw power spectrogram, in accordance with human cochlea. Furthermore, a log transform is applied, leading to 80-channel log Mel filterbank features. Consequently, for every 40-second segment, the final obtained feature is a two-dimensional array of size 667*80.

In addition to the classic FBank features, our research also explores the potential of self-supervised representation learning (SSL), by leveraging a cutting-edge pre-trained SSL model, Hidden-unit BERT (HuBERT) as feature extractor [55]. Recently, the effectiveness of pre-trained models has been assessed and shown excellent performance across various tasks. For our study, we chose HuBERT as it has been demonstrated to be the top-performing SSL model compared to various others and exhibited superiority in multiple downstream tasks [55]. Specifically, the HuBERT Base model has a seven-layer CNN and a twelve-layer transformer encoder and has been trained with the 960-hour LibriSpeech dataset [61]. Despite being pretrained on speech data, the model has demonstrated impressive representation capabilities across various non-speech audio datasets [62], [63]. Here parameters of the pre-trained HuBERT model are frozen, and the model is utilised as a feature extractor, to generate representations for the sleep apnea detection. Operating at a frame rate of 20 ms, the features produced by the HuBERT model for each 40-second segment is of size 2000*768.

C. Deep Structures

Features extracted from audio segments are fed into deep structures for sleep apnea recognition. Here, we explore four deep learning architectures to examine which method is optimal for sleep apnea detection. In particular, our exploration encompasses both conventional and cutting-edge deep models designed for audio signal analysis. These models include a recurrent neural network (RNN), Conformer [52], SpeechFormer [54], and E-branchformer [53]. Note that, the latter three structures were initially designed for solving speech recognition; however, previous studies have demonstrated their outstanding performance when applied to various non-speech audio tasks, such as sound event classification and human action classification [64], [65]. Descriptions for each model are provided below, while architecture hyper-parameters of the latter three Transformer-based models are given in Table II.

The first structure we examine is an RNN. This network features a two-layer structure, using 512 bidirectional long short-term memory (BLSTM) cells per layer. Such a structure is effective in learning long-term dependencies within audio inputs in both forward and backward directions, offering a better understanding of the context information. Furthermore, to improve the network's capability for generalisation, a dropout rate of 0.2 is applied to each RNN layer.

The second architecture we investigate is the Convolution-augmented Transformer (Conformer) [52]. This model has been widely used and achieved good success in a wide range of different speech and non-speech audio processing tasks/applications [64], [66]. For our specific task, we utilise only the Conformer encoder, followed by an additional linear layer and activation function to predict the presence of respiratory events. Specifically, the encoder is composed of multiple stacked identical blocks, and each block is composed of four modules stacked together. Within each Conformer block, a multi-head self-attention (MHSA) module and a convolution (CONV) module are sandwiched between two feed-forward (FFN) modules. The MHSA module can learn long-range global context, surpassing the capabilities of traditional RNNs in this respect. Meanwhile, CONV module can capture fine-grained local patterns synchronously. The integration of these two modules enables a comprehensive analysis of both global interaction and local correlations in sleep sound, which is crucial for respiratory event detection in our study.

The third architecture we explore is the Hierarchical Transformer (Hierformer). This model is a purely Transformer-based architecture without convolutions. In a hierarchical setup, the model processes data at multiple levels. The specific architecture we apply here is similar to the one detailed in [54], which consists of four Transformer blocks and three merging blocks. Each Transformer block employs a multi-head attention mechanism, focusing on the part of input dependent upon its importance; the merging block is used after two successive Transformer blocks and refine redundant features by averaging pooling. Intermediate representations are generated by the four Transformer blocks at different granularities: the lower blocks learn fine details from smaller analysis windows, and the higher blocks process signals from larger windows to derive broader contextual information. This hierarchical structure enhances the model's capability to understand and represent sleep sound data at various scales, and thus more manageable at handling long segments of sleep sound than traditional Transformer models.

TABLE II: Model hyper-parameters for Conformer, Hierformer, and EBranchformer.

Model	Conformer	Hierformer	EBranchformer
Num Params (M)	27.3	1.05	27.8
Encoder Layers	8	2-2-2-4 (4 blocks)	12
Encoder Dim	256	256	256
Attention Heads	8	8	8
Linear Units	1024	1024	1024

The final architecture we evaluate is the Enhanced Branchformer (EBranchformer), as proposed in [53]. As mentioned earlier, convolution and self-attention can capture local patterns and global context, respectively. In the aforementioned Conformer model, the strengths of the two are combined sequentially. In contrast, Branchformer applies two distinct branches in parallel, later merging their outputs by concatenation [67]. The EBranchformer further introduces depth-wise convolution in the merging block to combine local and global information both sequentially and in parallel. Although originally developed for automatic speech recognition, EBranchformer has demonstrated its superior performance in various speech-related tasks, often outperforming both Conformer and Branchformer [68]. This suggests its potential in processing non-speech audio signals. In this study, we aim to investigate the effectiveness of EBranchformer in analysing sleep sound.

D. Postprocessing

The outputs of the aforementioned deep model are probabilities indicating the presence of sleep apnea within a provided 40-second segment. For instance, when the deep model is tasked with classifying one segment into one of three categories: no-event, hypopnea, or apnea, and if the output probability corresponding to the apneic event is larger than the other two (no-event and hypopnea), the specific segment then is categorised as one containing apnea. Moreover, when the model is applied to analyse night-long audio recordings spanning several hours, we evaluate the sleep sound data in streaming mode. That is, the long recording is first segmented into a series of 40-second segments (with a hop size of 1 second to facilitate fine-grained temporal resolution during inference), which are then sequentially fed into the model. This yields sequential outputs, indicating the presence of sleep apnea/hypopnea per segment over time.

To accurately locate each event, these outputs undergo further processing. Particularly, one event can span multiple segments. In our analysis, both hypopnea and apnea are considered abnormal and are thus grouped into one single category, thus the task is transformed as binary classification. We take three steps to process the sequential output probability of abnormal events. First, we apply a moving average to smooth the probabilities over every ten successive outputs (this was chosen via grid search and consistently yielded better

performance), and this can effectively mitigate single misclassification. Note that smoothing is only applied for night-long audio analysis; no post-processing steps are applied to the segment-level classification evaluation. Subsequently, these smoothened probabilities are converted into binary values, where 0 denotes the absence of an event, and 1 indicates the presence of hypopnea or apnea. Next, we refine the binary outputs by identifying clusters of 1s and reducing each cluster to a single 1 positioned at the cluster's midpoint. This consolidation method ensures that multiple detections of the same event are treated as a single detection within each event cluster. Consequently, we can calculate the total number of apnea/ hypopnea events across long full-night recordings. By dividing this count by the total recording time, we derive the Apnea-Hypopnea Index (AHI). The AHI serves various purposes, including AHI estimation, sleep apnea screening, and assessing sleep apnea severity.

E. Evaluation

In this study, we conduct a comprehensive evaluation and comparison of tracheal and ambient microphone sounds when being applied for audio-based sleep apnea analysis. The original sleep sound dataset is partitioned into three participant-independent splits, namely training, development, and test sets. Throughout the training phase, the aforementioned deep architectures are optimised based on the minimisation of the cross-entropy loss. In this manner, we provide a robust measure of each model's effectiveness and reliability in sleep apnea detection. Our evaluation encompasses three key aspects: respiratory event classification on a segment-by-segment basis, AHI estimation from full-night recordings, and (severe) sleep apnea screening using AHI-based criteria.

Segment-based sleep apnea classification. We divide this analysis into three distinct sub-tasks: (1) binary classification: each segment is categorised as either normal or abnormal; (2) three-class classification: segments are classified into one of three categories: normal, hypopnea, or apnea, (3) five-class classification: this involves a more detailed classification where segments are identified as no-event, hypopnea, obstructive sleep apnea (OSA), central sleep apnea (CSA), or mixed sleep apnea (MSA). For each sub-task, we conduct a thorough comparison between tracheal and ambient sounds in terms of their effectiveness, reporting the performance in terms of accuracy and macro averaged F1 score. Another part of our analysis is the cross-evaluation of models: we investigate how models trained on one type of sound (either tracheal or ambient) perform when tested on the other. This cross-evaluation helps in understanding the generalizability of the models when there is a sound type mismatch. Moreover, we explore the potential of combining tracheal and ambient sounds by aggregating their outputs. This investigation aims to reveal any complementary information that the integration of both sound types might offer. Note that, no post-processing steps are performed on this segment-level classification task.

AHI estimation. After applying postprocessing steps, audio-based AHI estimations can be computed over the whole night recordings. This is achieved by analysing the recordings in a streaming mode using a binary classifier for respiratory event detection. Correlation plots are generated to illustrate the agreement and association between the audio-based AHI estimation and the PSG-based AHI reference. Additionally, We assess the correlation by reporting Pearson correlation coefficient between the two.

Sleep apnea screening. To evaluate the effectiveness of our method in screening for sleep apnea, we report the performance in terms of sensitivity, specificity, precision, and F1 according to various AHI diagnostic cutoffs. We focus on three commonly used AHI cut-offs: 5, 10, and 15. Additionally, for the detection of severe sleep apnea detection, we also examine the performance at a higher cut-off value of 30.

IV. Experiments

A. Data

We evaluated the proposed method using the open-access PSG-Audio dataset [41], which contains high-quality sound recordings collected during sleep along with PSG-based apnea/hypopnea event annotations. This data was collected during a full-night PSG study involving 212 participants by the Sleep Study Unit of the Sismanoglio-Amalia Fleming General Hospital of Athens, including 56 female subjects and 156 male participants, aged between 23 and 85 years old. Audio signals from two microphones – one positioned on the participant's trachea and the other situated approximately one meter above the participant's bed – were recorded simultaneously with the PSG study. Both the PSG data and the supplementary audio data were saved in one European Data Format (EDF) file for each participant. The labelling process of sleep stages and respiratory events was conducted by two health specialists. In particular, four specific respiratory events were annotated, including obstructive apnea, central apnea, mixed apnea, and hypopnea. For each event, the starting time and duration of the event were provided. These annotations were stored in Redline Markup Language (RML) files. Note that, the total number of participants explored in this study is 194 (50 female subjects and 144 male participants), less than the initially stated 212 from the original dataset due to the inaccessibility of some annotation files. In addition, the birth date of each participant was encrypted, making it impossible to retrieve age information from the dataset provided. Ethical clearance approvals were obtained from the Local ethics committee of Sismanoglio Hospital. All participants provided signed consent, permitting the use of their anonymised recorded data for research purposes.

In this study, we explored the tracheal and ambient microphone sound recordings and compared their performance for sleep apnea detection. The original tracheal recordings stored in the EDF files were sampled at 48 kHz, and we reduced the sampling rate to 8 kHz for our analysis. The down-sampled audio was then divided into 40-second segments for further analysis. In particular, positive segments (indicating apnea/ hypopnea events) were extracted according to the annotations of the respiratory events. The cut-off time for a positive segment was set to five seconds after the end of a respiratory event to include subsequent breathing or snoring

TABLE III: Data distribution of the PSG-Audio dataset over the subjects, duration hours, five categories (normal, hypopnea, obstructive sleep apnea [OSA], central sleep apnea [CSA], and mixed sleep apnea [MSA]) in training, development, and test sets. The apnea contains OSA, CSA, and MSA; the abnormal includes OSA, CSA, MSA, and hypopnea.

Data	# subjects (#f:#m)	hours	# normal	# hypopnea	# OSA	# CSA	# MSA	∑apnea	∑abnormal	∑
Train	162 (42:120)	730.3	33 426	8 295	22 381	1 519	4 463	28 363	36 658	70 084
Dev	8 (3:5)	37.2	1 972	613	937	85	350	1 372	1 985	3 957
Test	24 (5:19)	102.2	5 382	1 468	2 860	114	608	3 582	5 050	10 432
All	**194** (50:144)	**869.7**	40 780	10 376	26 178	1 718	5 421	33 317	43 693	84 473

sounds. The choice of five seconds was based on our initial investigation, although a different cut-off time could be used for the same purpose. This ensures the segment contains a complete end of apneic/hypopneic event, preventing the model from confusing events with silent segments and misclassifying entirely silent segments as SAS. From this, we obtained 33 317 apnea segments and 10 376 hypopnea segments. To extract negative segments (segments without any respiratory events), the entire night recording of each participant was divided into 40-second segments with a step size of 10 seconds. The step size of 10 was chosen to balance the number of negative segments with the number of segments containing abnormal respiratory events. Subsequently, segments that had no overlaps with any respiratory events were retained, resulting in 40 780 normal segments. As a result, a total of 84 473 segments were split into training, validation, and test sets in a subject-independent manner, as detailed in Table III.

B. Implementation Details

To assess diverse deep models for audio-based sleep apnea detection, we used the open-sourced S3PRL toolkit [69]. We built aforementioned deep models, designed the data pre- and post-process, and defined the performance evaluation metrics (see Sec. III). The use of open-source data and libraries ensures that our results can be easily reproduced, facilitating benchmark comparisons by others in future studies. Moreover, all code used in the study will be available to public for purposes of reproducing or extending the analysis.

For each specific combination of sub-task (binary, three-class, or five-class), feature type (FBank or HuBERT), model structure (RNN, Conformer, Hierformer, or EBranchformer), and sound source (Tracheal or ambient), a dedicated model was trained. Particularly, each model was trained for 15k steps on the training set, with a batch size of 32. We employed the Adam optimiser, setting the learning rate as 0.0001. Note that an exhaustive grid search over all possible hyperparameter combinations during training was computationally infeasible due to the complexity of our model, which involves multiple hyperparameters as well as different training tasks, model architectures, and feature types. We narrowed the searching space via heuristic selection, and this allowed us to maintain strong performance while keeping computational requirements manageable. This is evidenced by our results that are comparable to SOTA performance in other studies (cf Section V).

The model yielding the highest F1-score on the development set was stored. Then we evaluated the performance of the saved model on another independent test set. Note that for the three Transformer variants we tested, we retained the original encoder structure, including the same number of heads and layers as initially proposed. Specifically, these deep models embedded the input features per segment into a 256-dimensional feature, which was subsequently fed into a fully connected layer and softmax for each classification task.

While the segment-level classification evaluation followed the partitions outlined in Table III, the evaluation of overnight recordings was conducted using a three-fold subject-independent cross-validation, as the original test set consisted of only 24 subjects. This approach resulted in a total of 194 AHI estimations, corresponding to each individual subject. For this experiment, a consistent ratio was maintained across three folds regarding gender distribution.

V. Results and Discussion

A. Segment-based Sleep Apnea Classification

We first examined three classification tasks, distinguishing sleep apnea from coarse to fine granularity on segment level. Results are presented in Table IV.

Our results demonstrate that tracheal sounds possess enhanced discriminatory capabilities for sleep apnea detection when compared to ambient microphone sounds. This observation is in line with our expectations, considering the direct capture of respiratory sounds at the trachea. Notably, in the binary classification task, the best tracheal sound-based model obtains an accuracy of 90.8%, with a macro F1 score of 90.7%. This performance surpasses that of the best ambient sound-based model, which gains an accuracy of 89.2% and F1 score of 89.2%. The distinction becomes more pronounced in more complex classification scenarios. For instance, in the three-class case, the top-performing tracheal sound model detects hypopnea and apnea with an 83.3% accuracy and a 76.0% F1 score, exceeding the 80.1% accuracy and 71.4% F1 score achieved using ambient sounds. Likewise, in the five-class classification, the best tracheal sound-based model maintains a 75.7% accuracy and 57.9% F1 score, compared to 71.8% accuracy and 47.6% F1 score for ambient sounds. These findings highlight the robustness of tracheal sound-based models for sleep apnea detection across various scales.

Confusion matrices of the most effective models are depicted in Fig. 3, for tracheal and ambient sounds, respectively. The top three matrices represent results from tracheal sounds. In the binary classification, the tracheal sound-based model attained an accuracy of 96% for normal and 85% for abnormal. In the three-class setting, while the model performs well in distinguishing normal and apnea events, hypopnea remains the

TABLE IV: Performance of segment-level sleep apnea detection through the tracheal and ambient microphone audio recordings via four context-aware audio models with manual-crafted features (FBanks) or high-level representations extracted from a pre-trained model (HuBERT). We examined binary-class, three-class, and five-class cases, respectively.

features	models	ambient						tracheal					
		binary-class		three-class		five-class		binary-class		three-class		five-class	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
FBanks	BLSTM-RNNs	.851	.852	.643	.749	.387	.694	.898	.898	.734	.812	.481	.759
	Conformer	.866	.867	.677	.783	.471	.682	.894	.895	.725	.802	.524	.731
	Hierformer	.866	.867	.687	.749	.389	.666	**.907**	**.908**	.741	.823	.576	.764
	EBranchformer	.879	.880	.694	.782	.469	.688	.900	.900	**.760**	**.833**	**.579**	**.757**
HuBERT	BLSTM-RNNs	.858	.859	.677	.746	.371	.651	.878	.879	.720	.798	.382	.703
	Conformer	.875	.875	.711	.791	.459	.710	.887	.888	.729	.809	.514	.716
	Hierformer	.881	.882	.713	.796	.451	.651	.903	.904	.735	.819	.515	.729
	EBranchformer	**.892**	**.892**	**.714**	**.801**	**.476**	**.718**	.903	.903	.746	.819	.546	.749

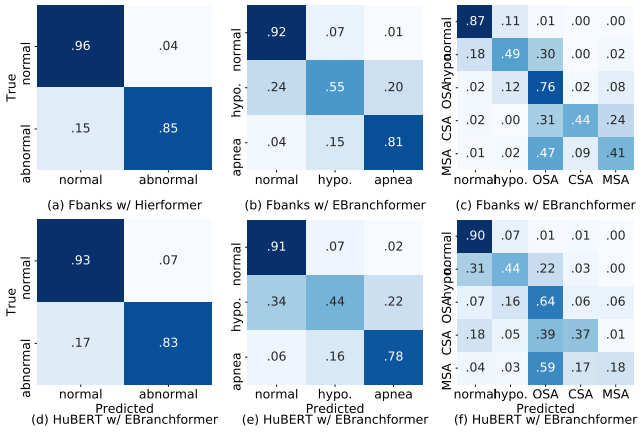


Fig. 3: Confusion matrices of best-performed models for binary-class, three-class, and five-class sleep apnea detection, for tracheal (a)-(c) and ambient (d)-(f) microphone, separately.

most challenging class. Specifically, 24% of hypopnea events are misclassified as normal and 20% as apnea. This may be attributed to the relatively subtle acoustic characteristics of hypopnea events, which often exhibit weaker airflow reduction and less pronounced sound patterns compared to apnea, making them acoustically less distinctive. In the five-class setting, when apnea is further divided into OSA, CSA, and MSA, CSA and MSA are frequently misclassified as OSA (31% of CSA and 47% of MSA predicted as OSA). This confusion likely arises because the acoustic manifestations of different apnea subtypes share overlapping respiratory sound patterns, while their physiological differences (e. g., respiratory effort) are not directly observable from sound alone. Similar patterns are observed in ambient sound-based models, as illustrated in the lower three confusion matrices of Fig. 3. These findings highlight the intrinsic limitation of sound-only approaches in distinguishing fine-grained apnea subtypes. In future work, integrating complementary physiological signals (e. g., respiratory effort or oxygen saturation) may help mitigate these confusions and improve subtype classification performance.

Comparing the four deep structures, our results also indicate the dominance of the EBranchformer architecture in sleep apnea detection, surpassing the performance of the other three architectures studied. Remarkably, in 10 out of all 12 cases analysed (spanning two sound sources, two feature types, and three classification tasks), the EBranchformer consistently emerged as the best model. This underscores the architecture's efficacy in aggregating both local and global information both sequentially and in parallel. Additionally, our analysis reveals unique responses of the two sound types to different feature extraction methods. While HuBERT features demonstrate advantages over FBank features in models based on ambient recordings, a slight preference for FBank features is observed in tracheal recordings. A possible explanation lies in the nature of HuBERT's pretraining data. HuBERT is trained on speech corpora, whose acoustic characteristics could be more similar to ambient recordings. In contrast, tracheal sounds are captured via contact microphones placed on the neck and exhibit distinct spectral and physiological characteristics, including lower-frequency dominance and reduced resemblance to speech. As a result, representations learned from speech data may transfer more effectively to ambient recordings than to tracheal ones. On average, the combination of FBank features and the EBranchformer structure yielded the best performance over the other combinations. Consequently, this combination was chosen for use in subsequent analyses.

We further analysed the impact of varying the offset length (0 to 10 seconds) of the endpoint in sleep apneic event windows on classification accuracy and F1 score, as shown in Figure 4. The results indicate that, across all tasks and sound sources, classification accuracy generally increases with larger offset lengths, peaking around five seconds, before declining as the offset length extends further. A similar trend is observed for the F1 scores. These findings suggest that extending the window endpoint improves model performance, potentially by capturing more relevant contextual information beyond the immediate sleep apneic event.

B. More about Tracheal and Ambient Sounds

We further compared the potential of tracheal and ambient sounds in the context of sleep apnea detection, conducting a statistical comparison between the two sound sources. Specifically, we categorised all segments into four groups: segments accurately predicted by both sound sources, segments identified only by the tracheal sound, segments detected only

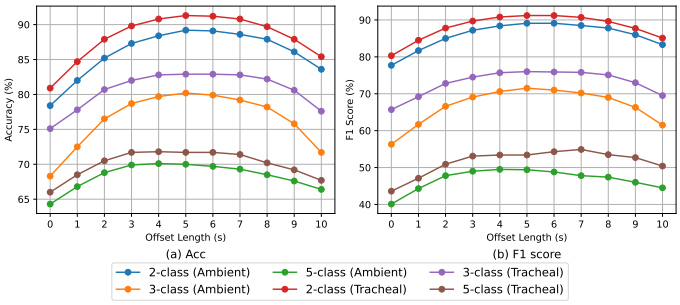


Fig. 4: Effect of varying the offset length (0 to 10 seconds) of the endpoint of sleep apneic event windows on classification accuracy (a) and F1 score (b).

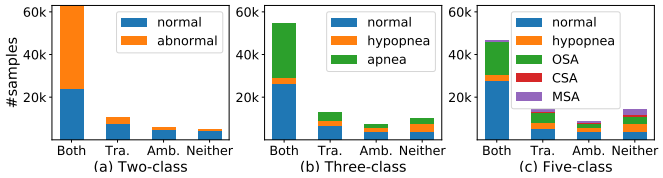


Fig. 5: Detailed performance in terms of the number of correctly predicted samples for the binary (a), three (b), and five (c) classes, respectively. We split samples into four types: predicted correctly by both modalities, predicted correctly only by tracheal microphone, predicted correctly only by ambient microphone, and predicted correctly by neither of the two.

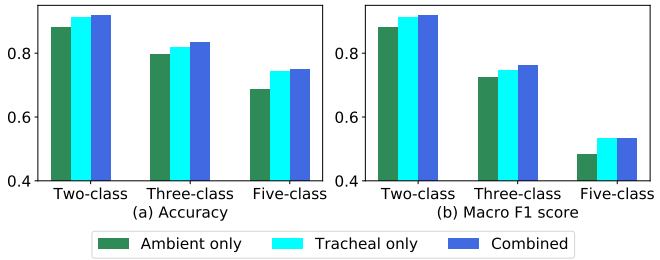


Fig. 6: Performance comparison in terms of accuracy (a) and F1 (b) among different model settings, i. e., ambient only, tracheal only, and the combination of the two (Combined).

by the ambient sound, and segments incorrectly predicted by both. As illustrated in Fig. 5, the majority of segments across all three classification levels were correctly predicted by both sources, indicating the overall effectiveness of both sound types. In addition, though tracheal sounds generally outperformed ambient sounds, a notable portion of segments were detected exclusively by ambient sound. We also note that the distribution of classes detected by each source was similar. This suggests that there are no significant differences in the responses of the two sound types to varied sleep apneic events.

As shown in Fig. 5, certain events were accurately predicted by only one sound source. Considering that the model outputs a probability value, which can be interpreted as a confidence level, we explored a joint prediction approach. This involved comparing the confidence levels from both sources when they disagreed and selecting the prediction with higher confidence. The results of this combined approach, as shown in Fig. 6, indicate enhanced accuracy across all levels of sleep apnea detection. Furthermore, the combination also improved the F1 score in binary and three-class classifications, highlighting the potential benefits of integrating tracheal and ambient sounds for sleep apnea detection during sleep.

Additionally, we investigated the scenarios where the training and test sets came from different sound sources. This cross-source evaluation assesses the models' robustness and generalisation capabilities, enhancing our understanding of the different sources. A model that performs well on both tracheal and ambient recordings is likely more robust. Additionally, it helps detect if the model overfits to any source-specific noise rather than learning the relevant signals for SAS detection. The results, presented in Table V, reveal a decline in performance due to source mismatch. Models trained on tracheal sounds exhibited a more significant performance drop when tested on ambient sounds, likely due to a lack of noise tolerance. Conversely, models trained with ambient sounds showed reasonable performance on tracheal sounds, though slightly inferior to their performance with ambient sounds. This suggests that while tracheal sound-based models excel with matched data, they require further adaptation or preliminary data denoising for effective application to ambient sound recordings. Hence, if a model is exclusively trained on tracheal sounds, it requests additional adaptation to suit ambient sleep sound recordings. In the future, it would be interesting to investigate augmenting the training data with a mix of both tracheal and ambient recordings to improve the model's robustness.

C. AHI Estimation

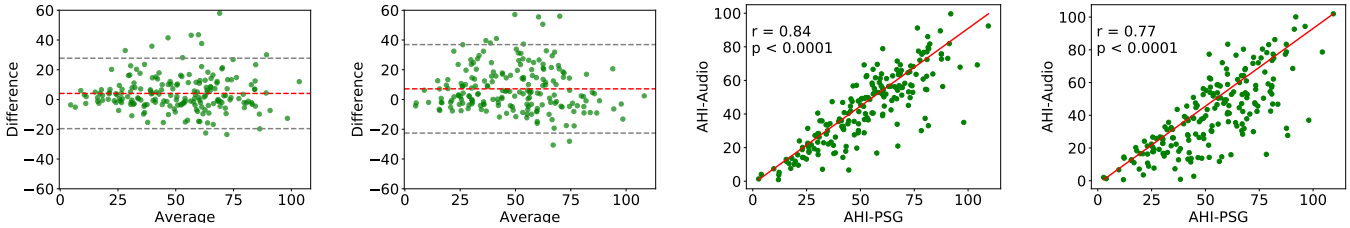
Upon processing overnight recordings through a trained model, the sequential outputs were aggregated and utilised to calculate corresponding AHI estimations, as detailed in Section III-D. As a result, Fig. 7 presents the Bland-Altman plots and correlation plots between the PSG-based AHI labels and the AHI estimations derived from sleep sounds, for tracheal and ambient microphones, respectively. Overall, our sound-based predictions match the reference AHI patterns. In particular, a substantial correlation is shown between tracheal sound-based AHI estimations and the PSG-based AHI, with a correlation coefficient of 0.84. In comparison, the ambient sound-based AHI estimations exhibited a slightly lower correlation coefficient of 0.77 with the PSG-based AHI. These findings indicate the feasibility of our proposed methods in estimating the AHI from sounds.

D. Sleep Apnea Screening

Table VI presents the performance of our method in diagnosing sleep apnea across various AHI cut-off values. We also demonstrate a comparison to other state-of-the-art (SOTA) studies. Our approach in most cases surpassed these SOTA methods in effectiveness. However, it is important to note that these comparisons are based on results from different datasets, making the comparison somewhat indirect. Despite this limitation, the comparison still offers valuable insights, particularly considering that the number of subjects evaluated in our study is comparable to those in other studies. Most

TABLE V: Performance in terms of F1 score and accuracy (Acc) when training and evaluating different sound sources.

Training \ *Test*	ambient						tracheal					
	binary-class		three-class		five-class		binary-class		three-class		five-class	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
ambient	.879	.880	.694	.782	.469	.688	.839	.842	.677	.743	.420	.655
tracheal	.756	.759	.499	.618	.361	.565	.900	.900	.760	.833	.579	.757



(a) Tracheal AHI estimations (b) Ambient AHI estimations (c) Tracheal AHI estimations (d) Ambient AHI estimations

Fig. 7: Agreement between AHI-PSG and AHI-Audio, for tracheal sounds and ambient sounds, respectively. In Bland-Altman plots (a) and (b), the red and grey dashed lines indicate the mean difference and 1.96 times the standard deviation of the differences, respectively. In scatter plots (c) and (d), the red diagonal line indicates when two observations correlate perfectly.

TABLE VI: Performance comparison between our methods and other state-of-the-art approaches. Results are reported in terms of sensitivity (Se), specificity (Sp), precision (Pr), and F1 score. Three AHI cut-off points (5, 10, and 15) are explored for sleep apnea screening, and one AHI cut-off value (30) is applied for severe sleep apnea screening.

Method (#Subjects, Sound Source)	AHI Cut-off=5				AHI Cut-off=10				AHI Cut-off=15				AHI Cut-off=30			
	Se	Sp	Pr	F1	Se	Sp	Pr	F1	Se	Sp	Pr	F1	Se	Sp	Pr	F1
Ours (194, Tracheal)	.99	1.0	1.0	.83	.96	1.0	1.0	.72	.93	1.0	1.0	.71	.84	.97	.99	.80
Ours (194, Ambient)	.98	1.0	1.0	.74	.96	1.0	1.0	.70	.89	1.0	1.0	.66	.75	1.0	1.0	.73
SoundSleepNet [38] (150, Ambient)	.97	.89	-	-	-	-	-	-	.85	.84	-	-	.96	.91	-	-
DNN [33] (103, Ambient)	.87	.71	-	-	.73	.70	-	-	.77	.80	-	-	.73	.95	-	-
OSAnet [39] (59, Ambient)	.94	.83	.96	-	.83	.79	.89	-	.89	.96	.97	-	.96	.92	.88	-
Firefly [34] (120, Ambient)	-	-	-	-	-	-	-	-	.88	.80	.82	-	-	-	-	-
TS-DNN [43] (304, Tracheal)	.98	.76	-	-	-	-	-	-	.97	.90	-	-	.92	.94	-	-

SOTA methods focus solely on OSA detection. In particular, SoundSleepNet [38] employed a two-step training strategy to sequentially learn a CNN+RNN+Transformer structure, enabling a three-class OSA event detection task. Various CNN structures were explored in [33], [34], [39], [43] for OSA. Moreover, [34] further utilised active sonar signal processing, which improved the detection of CSA, even in the absence of passive snoring or recovery breaths that typically accompany obstructive events and not central ones. In contrast, our best-performing EBranchformer model exclusively exploited passive acoustic signals and aggregated relevant information both sequentially and in parallel. Also, we investigated SAS in general, considering both OSA and CSA.

Our method demonstrated exceptional screening capabilities for tracheal sound-based sleep apnea diagnosis. On the PSG-Audio dataset, it achieved over 90% sensitivity and 100% specificity across different AHI cut-off points. Similar patterns can be seen for ambient sound-based models. For the detection of severe sleep apnea, the models showed sensitivities of 84% and 75% for tracheal and ambient sounds, respectively. As indicated in Fig. 7, there were instances where the AHI estimations fell short of the reference values, especially when the reference AHI exceeded 40. This resulted in a subset of participants being incorrectly classified as non-severe, consequently reducing the sensitivity in severe sleep apnea screening. This implies that the respiratory event patterns in certain severe sleep apnea patients might exhibit diverse characteristics and reveals a limitation in our current model's ability to detect specific types of sleep apnea in these individuals. Particular, a small percentage of SAS patients, especially those with CSA, do not habitually snore or produce heavy breathing sounds during sleep, making them difficult to identify with our model. We consider to address this issue in future by integrating other wearable devices to leverage other biosignals.

VI. Conclusion

This study comprehensively examined the capability of tracheal and ambient sounds in sleep apnea research, focusing on detailed segment-by-segment sleep apnea detection. Evaluating on PSG-Audio, our research highlighted the significant potential of audio-based analysis in remotely monitoring sleep-rated respiratory anomalies. In particular, the EBranchformer model emerged as the most promising model among the four models tested, underscoring the importance of considering both local and global information in model design. Additionally, the study demonstrated enhanced performance

through the use of a pre-trained model for extracting deep representations from ambient sounds. Moreover, our findings indicate that while tracheal sounds inherently possess superior capabilities for distinguishing sleep apneic events, combining them with ambient sounds can lead to even more accurate detection. In the cross-source validation, a notable decrease in performance was observed, suggesting the need for model adaptation and noise reduction in response to these variations raised by difference sources. Overall, our method yields outstanding performance in overnight AHI estimation and sleep apnea screening across various AHI cut-offs, setting a new benchmark that surpasses existing state-of-the-art methods.

In conclusion, this study offers valuable insights into the development of IoT-enabled, audio-based sleep apnea detection systems. By thoroughly comparing tracheal and ambient sound recordings, we lay a foundation for future innovations in non-invasive, scalable sleep monitoring solutions. Our findings highlight the promise of audio sensing as a viable approach for unobtrusive, at-home screening of sleep apnea, with potential to transform current diagnostic practices.

Yet, several limitations should be acknowledged. The present study was evaluated on PSG-Audio only, which was collected in a single clinical setting with a specific recording configuration. Variations in recording devices, placement, background noise conditions, and patient populations may introduce domain shifts that affect performance. In particular, real-world home settings may involve higher ambient noise levels and less standardised sensor positioning. Moreover, although HuBERT was adopted as a representative pretrained model in this study, other foundation models trained on large-scale general audio datasets may also be suitable for deep feature extraction. As our objective was not to identify the optimal pretrained model, systematically evaluating alternative general-purpose audio representations remains an important direction for future work and may further improve robustness and generalisation in sleep apnea detection.

Looking ahead, we plan to integrate signals from diverse sleep-related sensors and develop noise-robust models, with the goal of advancing real-world deployable solutions that align with the vision of more accurate and efficient sleep health monitoring.

REFERENCES

[1] F. Yang *et al.*, "Internet-of-things-enabled data fusion method for sleep healthcare applications," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15 892–15 905, 2021.

[2] A. V. Benjafield *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *The Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, Aug. 2019.

[3] H. Palomäki, M. Partinen, T. Erkinjuntti, and M. Kaste, "Snoring, sleep apnea syndrome, and stroke." *Neurology*, vol. 42, no. 7, pp. 75–81, 1992.

[4] M. Younes *et al.*, "Contribution of obstructive sleep apnea to disrupted sleep in a large clinical cohort of patients with suspected obstructive sleep apnea," *Sleep*, vol. 46, no. 7, p. zsac321, 2023.

[5] N. Meslier *et al.*, "Prevalence of symptoms of sleep apnea syndrome. study of a french middle-aged population," *Revue Des Maladies Respiratoires*, vol. 24, no. 3, pp. 305–313, 2007.

[6] M. S. Aldrich and J. B. Chauncey, "Are morning headaches part of obstructive sleep apnea syndrome?" *Archives of Internal Medicine*, vol. 150, no. 6, pp. 1265–1267, 1990.

[7] K. Gagnon *et al.*, "Cognitive impairment in obstructive sleep apnea," *Pathologie Biologie*, vol. 62, no. 5, pp. 233–240, 2014.

[8] M. Harris, N. Glozier, R. Ratnavadivel, and R. R. Grunstein, "Obstructive sleep apnea and depression," *Sleep Medicine Reviews*, vol. 13, no. 6, pp. 437–444, 2009.

[9] W. Lee, S.-A. Lee, H. U. Ryu, Y.-S. Chung, and W. S. Kim, "Quality of life in patients with obstructive sleep apnea: Relationship with daytime sleepiness, sleep quality, depression, and apnea severity," *Chronic Respiratory Disease*, vol. 13, no. 1, pp. 33–39, 2016.

[10] A. G. Logan *et al.*, "High prevalence of unrecognized sleep apnoea in drug-resistant hypertension," *Journal of Hypertension*, vol. 19, no. 12, pp. 2271–2277, Dec. 2001.

[11] H. K. Yaggi, J. Concato, W. N. Kernan, J. H. Lichtman, L. M. Brass, and V. Mohsenin, "Obstructive sleep apnea as a risk factor for stroke and death," *New England Journal of Medicine*, vol. 353, no. 19, pp. 2034–2041, Nov. 2005.

[12] D. Einhorn, D. A. Stewart, M. K. Erman, N. Gordon, A. Philis-Tsimikas, and E. Casal, "Prevalence of sleep apnea in a population of adults with type 2 diabetes mellitus," *Endocrine Practice*, vol. 13, no. 4, pp. 355–362, July 2007.

[13] C. L. Jackson, S. Redline, and K. M. Emmons, "Sleep as a potential fundamental contributor to disparities in cardiovascular health," *Annual Review of Public Health*, vol. 36, pp. 417–440, Mar. 2015.

[14] E. Sforza, Z. de Saint Hilaire, A. Pelissolo, T. Rochat, and V. Ibanez, "Personality, anxiety and mood traits in patients with sleep-related breathing disorders: effect of reduced daytime alertness," *Sleep Medicine*, vol. 3, no. 2, pp. 139–145, Mar. 2002.

[15] F. Mendonca, S. S. Mostafa, A. G. Ravelo-Garcia, F. Morgado-Dias, and T. Penzel, "A review of obstructive sleep apnea detection approaches," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 825–837, Mar. 2018.

[16] Y. Guo, X. Gu, and G.-Z. Yang, "MCDCD: Multi-source unsupervised domain adaptation for abnormal human gait detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 4017–4028, 2021.

[17] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 465–474, 2019.

[18] Y. Zhang, T. Xia, A. Saeed, and C. Mascolo, "RespLLM: Unifying Audio and Text with Multimodal LLMs for Generalized Respiratory Health Prediction," in *Proc. Machine Learning for Health (ML4H)*. PMLR, 2025, pp. 1053–1066.

[19] F. Deligianni, Y. Guo, and G.-Z. Yang, "From emotions to mood disorders: A survey on gait analysis methodology," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2302–2316, 2019.

[20] Q. Shen, X. Yang, L. Zou, K. Wei, C. Wang, and G. Liu, "Multitask residual shrinkage convolutional neural network for sleep apnea detection based on wearable bracelet photoplethysmography," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25 207–25 222, 2022.

[21] J. Levy, D. Álvarez, F. Del Campo, and J. A. Behar, "Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry," *Nature Communications*, vol. 14, no. 1, pp. 1–12, Aug. 2023.

[22] A. Yadollahi, E. Giannouli, and Z. Moussavi, "Sleep apnea monitoring and diagnosis based on pulse oximetery and tracheal sound signals," *Medical & Biological Engineering & Computing*, vol. 48, pp. 1087–1097, Aug. 2010.

[23] T. Choksatchawathi *et al.*, "Apsense: Data-driven algorithm in PPG-based sleep apnea sensing," *IEEE Internet of Things Journal*, vol. 11, no. 20, pp. 33 915–33 926, 2024.

[24] T. Penzel and A. Sabil, "The use of tracheal sounds for the diagnosis of sleep apnoea," *Breathe*, vol. 13, no. 2, pp. 37–45, June 2017.

[25] A. K. Ng, T. San Koh, E. Baey, T. H. Lee, U. R. Abeyratne, and K. Puvanendran, "Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?" *Sleep Medicine*, vol. 9, no. 8, pp. 894–898, Dec. 2008.

[26] E. Dafna, A. Tarasiuk, and Y. Zigel, "Osa severity assessment based on sleep breathing analysis using ambient microphone," in *Proc. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 2044–2047.

[27] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, "A systematic review of detecting sleep apnea using deep learning," *Sensors*, vol. 19, no. 22, pp. 1–26, 2019.

[28] D. Novak, K. Mucha, and T. Al-Ani, "Long short-term memory for apnea detection based on heart rate variability," in *Proc. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2008, pp. 5234–5237.

[29] S. S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, "SpO2 based sleep apnea detection using deep learning," in *Proc. IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, 2017, pp. 91–96.

[30] H. ElMoaqet, M. Eid, M. Glos, M. Ryalat, and T. Penzel, "Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals," *Sensors*, vol. 20, no. 18, pp. 5037–5066, 2020.

[31] J. Hong *et al.*, "End-to-end sleep staging using nocturnal sounds from microphone chips for mobile devices," *Nature and Science of Sleep*, vol. 14, pp. 1187–1201, June 2022.

[32] Y. Liu *et al.*, "Tracheal sound-based apnea detection using hidden markov model in sedated volunteers and post anesthesia care unit patients," *Journal of Clinical Monitoring and Computing*, vol. 37, p. 1061–1070, May 2023.

[33] H. E. Romero, N. Ma, G. J. Brown, and E. A. Hill, "Acoustic screening for obstructive sleep apnea in home environments based on deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 2941–2950, Feb. 2022.

[34] R. Tiron *et al.*, "Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology," *Journal of Thoracic Disease*, vol. 12, no. 8, p. 4476–4495, Aug. 2020.

[35] S. Saraç and G. C. Afsar, "Effect of mean apnea-hypopnea duration in patients with obstructive sleep apnea on clinical and polysomnography parameter," *Sleep and Breathing*, vol. 24, pp. 77–81, June 2020.

[36] A. Malhotra *et al.*, "Metrics of sleep apnea severity: beyond the apnea-hypopnea index," *Sleep*, vol. 44, no. 7, p. zsab030, July 2021.

[37] M. P. Butler *et al.*, "Apnea–hypopnea event duration predicts mortality in men and women in the sleep heart health study," *American Journal of Respiratory and Critical Care Medicine*, vol. 199, no. 7, pp. 903–912, Apr. 2019.

[38] V. L. Le *et al.*, "Real-time detection of sleep apnea based on breathing sounds and prediction reinforcement using home noises: Algorithm development and validation," *Journal of Medical Internet Research*, vol. 25, pp. 1–15, Feb. 2023.

[39] B. Wang *et al.*, "Obstructive sleep apnea detection based on sleep sounds via deep learning," *Nature and Science of Sleep*, pp. 2033–2045, Dec. 2022.

[40] G. Korompili, L. Kokkalas, S. A. Mitilineos, N.-A. Tatlas, and S. M. Potirakis, "Detecting apnea/hypopnea events time location from sound recordings for patients with severe or moderate sleep apnea syndrome," *Applied Sciences*, vol. 11, no. 15, pp. 1–15, July 2021.

[41] G. Korompili *et al.*, "PSG-Audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies," *Scientific Data*, vol. 8, no. 1, pp. 1–13, Aug. 2021.

[42] T. Kim, J.-W. Kim, and K. Lee, "Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques," *Biomedical Engineering Online*, vol. 17, pp. 1–19, Dec. 2018.

[43] 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.

[44] Y. Castillo-Escario, I. Ferrer-Lluis, J. M. Montserrat, and R. Jane, "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. 128 224–128 241, Sep. 2019.

[45] S. Cao, I. Rosenzweig, F. Bilotta, H. Jiang, and M. Xia, "Automatic detection of obstructive sleep apnea based on speech or snoring sounds: a narrative review," *Journal of Thoracic Disease*, vol. 16, no. 4, pp. 2654–2667, 2024.

[46] X. Qiu *et al.*, "An audio-semantic multimodal model for automatic obstructive sleep apnea-hypopnea syndrome classification via multi-feature analysis of snoring sounds," *Frontiers in Neuroscience*, vol. 18, pp. 1–12, 2024.

[47] L. Ding, J. Peng, L. Song, and X. Zhang, "Automatically detecting apnea-hypopnea snoring signal based on VGG19+ LSTM," *Biomedical Signal Processing and Control*, vol. 80, pp. 1–10, 2023.

[48] A. Hu *et al.*, "Snore sound features based on percussive enhancing and positional encoding combined with multi-task learning for osahs detection," in *Proc. 49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 901–905.

[49] Y. Huang, L. Chen, and Q. Huang, "Fine-grained detection of apnea-hypopnea events based on transformer network in audio recordings," in *Proc. 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 2023, pp. 580–585.

[50] J. Liu *et al.*, "Tracheal sounds accurately detect apnea in patients recovering from anesthesia," *Journal of Clinical Monitoring and Computing*, vol. 33, pp. 437–444, June 2019.

[51] T. Rosenwein, E. Dafna, A. Tarasiuk, and Y. Zigel, "Breath-by-breath detection of apneic events for osa severity estimation using non-contact audio recordings," in *Proc. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 7688–7691.

[52] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[53] K. Kim *et al.*, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 84–91.

[54] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "SpeechFormer: A Hierarchical Efficient Framework Incorporating the Characteristics of Speech," in *Proc. INTERSPEECH*, 2022, pp. 346–350.

[55] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, Oct. 2021.

[56] S. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. International Conference on Machine Learning (ICML)*, 2023, pp. 5178–5193.

[57] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. INTERSPEECH*, 2022, pp. 2753–2757.

[58] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. INTERSPEECH*, 2021, pp. 571–575.

[59] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 137–151, 2022.

[60] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clinical Medicine Insights. Circulatory, Respiratory and Pulmonary Medicine*, vol. 2, pp. 45–58, May 2008.

[61] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[62] T.-Y. Wu, T.-Y. Hsu, C.-A. Li, T.-H. Lin, and H.-y. Lee, "The efficacy of self-supervised speech models for audio representations," *Proceedings of Machine Learning Research, HEAR: Holistic Evaluation of Audio Representations*, pp. 90–110, 2021.

[63] M. La Quatra *et al.*, "Benchmarking representations for speech, music, and acoustic events," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 505–509.

[64] S. Srivastava *et al.*, "Conformer-based self-supervised learning for non-speech audio tasks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8862–8866.

[65] Y. Chae, J. Koo, S. Lee, and K. Lee, "Exploiting time-frequency conformers for music audio enhancement," in *Proc. 31st ACM International Conference on Multimedia (MM)*, 2023, pp. 2362–2370.

[66] P. Guo *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5874–5878.

[67] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. International Conference on Machine Learning (ICML)*, 2022, pp. 17 627–17 643.

[68] Y. Peng *et al.*, "A comparative study on E-Branchformer vs Conformer in speech recognition, translation, and understanding tasks," in *Proc. INTERSPEECH*, 2023, pp. 2208–2211.

[69] S. wen Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.