Nature and Science of Sleep

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

ApneaWhisper: Transformer-Based Audio Segmentation for Fine-Grained Non-Invasive Sleep Apnea Detection

Yunu Kim[1], Myeongbin Kim[1], Jaemyung Shin[2], Minsam Ko[1]

[1]Department of Applied Artificial Intelligence, Hanyang University ERICA, Ansan, Gyeonggi-do, Republic of Korea; [2]DelightRoom, Seoul, Republic of Korea

Correspondence: Minsam Ko, Hanyang University ERICA, Ansan, Gyeonggi-do, Republic of Korea, Email minsam@hanyang.ac.kr

Nature and Science of Sleep downloaded from https://www.dovepress.com/
For personal use only.

Purpose: Sleep apnea is a prevalent sleep disorder with serious health implications. This study introduces ApneaWhisper, a Transformer-based audio segmentation model designed for noninvasive detection of sleep apnea subtypes using PSG-Audio data.

Patients and Methods: We utilized a PSG-Audio dataset from 284 patients. ApneaWhisper leverages a pretrained Whisper encoder to extract 10 ms-resolution frame-level features from 20-second audio clips. A lightweight Transformer decoder with token-based segmentation and a classification head aggregates these features for both frame-level and clip-level predictions. The model was fine-tuned using class-balanced cross-entropy loss to address data imbalance across apnea subtypes.

Results: ApneaWhisper achieved strong performance for sleep apnea detection, with a clip-level F1-score of 0.82 and a frame-level F1-score of 0.70, outperforming conventional baselines including MFCC+DNN, VGGish+bi-LSTM, and VAD-based models. It also showed promising ability in distinguishing between OSA, MSA, CSA, and hypopnea, though with varying success.

Conclusion: The model's fine-grained temporal resolution enables precise apnea event localization, duration estimation, and subtype classification. While ApneaWhisper performs robustly for OSA, challenges remain in distinguishing central (CSA) and mixed (MSA) sleep apnea, due to subtle or ambiguous acoustic patterns. The frame-level segmentation also facilitates accurate apnea–hypopnea index (AHI) estimation, which could reduce dependence on full PSG studies in certain clinical and home-monitoring scenarios. Future improvements may involve multimodal integration (eg, oxygen saturation) and noise-robust training techniques.

Keywords: sleep apnea, sleep breathing, deep learning, audio segmentation, whisper, transformer

Introduction

Sleep apnea is a profoundly impactful sleep-related breathing disorder, globally affecting an estimated 936 million adults aged 30 to 69 years.[1–3] This condition significantly increases the risk of severe comorbidities, including hypertension, stroke, myocardial infarction, and neurocognitive impairment.[2,4–7] OSA, in particular, is characterized by recurrent collapses of the upper airway, manifesting as distinctive acoustic patterns such as loud snoring, gasping, and choking sounds, making the condition particularly amenable to detection via audio signals.[5,8]

Despite its alarming prevalence and severe health consequences, sleep apnea remains severely underdiagnosed, with estimates suggesting that 80–90% of affected individuals in some regions remain undiagnosed.[2,3] The clinical gold standard for diagnosis, overnight polysomnography (PSG), presents significant logistical and financial barriers, costing anywhere from $1000 to over $10,000 in the United States and often requiring long waiting lists and patient discomfort.[9] These limitations have spurred a critical need for noninvasive, more accessible, and cost-effective home-based screening approaches.

Targeted analysis of sleep breathing sounds has emerged as a promising avenue for noninvasive sleep apnea detection, driven by advances in sensor technology and large-scale audio data collection. Notably, breath sounds captured at the trachea closely mirror apneic events, providing a distinctive and sensitive acoustic signature for their identification.[10] Building on this insight, the PSG-Audio dataset[11] offers a unique resource with clinically detailed, time-synchronized annotations of respiratory events paired with high-quality tracheal and room audio recordings. This dual-channel design not only enables precise characterization of apneic episodes but also supports robust comparisons between different acoustic perspectives. By

Received: 16 July 2025
Accepted: 29 August 2025
Published: 4 October 2025

© 2025 Kim et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

combining clinical granularity with audio diversity, PSG-Audio strengthens the bridge between experimental research and real-world clinical deployment, opening the door to scalable, accessible alternatives to traditional PSG.

Early studies relied on handcrafted features extracted from snore or respiratory sounds—such as Mel-frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients (LPCs), spectral flux, zero-crossing rate, pitch, sub-band energy, shimmer, jitter, and harmonics-to-noise ratio (HNR)—and employed traditional machine learning models including Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs) to classify sleep-disordered breathing events or estimate PSG-derived metrics.[5,12,13] However, these approaches often suffered from limited generalizability and high sensitivity to environmental noise in unconstrained home settings. More recently, deep learning–based methods have gained prominence, enabling end-to-end feature learning from raw audio or spectrograms via architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks,[14–20] with hybrid frameworks—for example, VGGish combined with bi-LSTMs—also explored for temporal classification of sleep sounds.[18]

However, most existing deep learning approaches for sleep apnea detection primarily operate on coarse temporal windows (eg, 20–30 second clips), reducing their clinical utility for fine-grained temporal pattern analysis or precise event delineation. Furthermore, many models either classify all apnea types into a single "apnea" category or struggle to differentiate effectively between non-obstructive apnea subtypes, such as central sleep apnea (CSA) and mixed sleep apnea (MSA), which often have subtler or ambiguous acoustic signatures compared to the loud, characteristic sounds of OSA. This difficulty arises because CSA often lacks strong acoustic markers like snoring or gasping,[10,21] as it results from a lack of respiratory effort, while MSA presents an ambiguous mixed acoustic signature combining both central and obstructive features.[22] These limitations reduce the granularity required for accurate Apnea–Hypopnea Index (AHI) calculation and tailored treatment planning.

Recent advances in self-supervised learning (SSL) and Transformer architectures offer unprecedented opportunities to address these gaps. SSL models, such as Wav2Vec 2.0[23] and HuBERT,[24] learn powerful, context-rich representations from vast amounts of unlabeled audio data, making them highly effective for downstream tasks, especially where labeled data is scarce, as is often the case in medical domains. The Whisper model,[25] pretrained on 680,000 hours of diverse labeled audio data for speech recognition, has demonstrated remarkable robustness to noise and generalization across various acoustic environments. Transformer architectures, with their attention mechanisms, excel at modeling long-range dependencies in sequential data, enabling precise temporal localization and more nuanced event classification beyond what traditional CNN-RNN hybrids can achieve. While specific applications of these advanced SSL models directly for audio-only sleep apnea detection are still emerging, their success in related audio processing tasks makes them highly promising. Building on this capability, the WhisperSeg framework[26] adapts Whisper for fine-grained voice activity detection, enabling frame-level temporal localization.

In this study, we introduce ApneaWhisper, a fine-grained audio segmentation model specifically designed for sleep apnea detection. ApneaWhisper builds upon the Whisper architecture, inheriting its strengths while being explicitly configured to identify and temporally localize four major respiratory event types—obstructive apnea, central apnea, mixed apnea, and hypopnea—at a 10 ms resolution. Owing to Whisper's proven robustness against real-world acoustic variability—including background noise, overlapping sounds, and non-verbal audio—ApneaWhisper is particularly well-suited for analyzing sleep breathing sounds in uncontrolled home environments. Leveraging the clinically annotated PSG-Audio dataset, we evaluate the model on two tasks:

1. Clip-level classification: Predict the presence of disordered breathing in 20-second audio segments.
2. Frame-level segmentation: Delineate the onset and offset of apnea and hypopnea events with 10 ms resolution.

We hypothesize that this Transformer-based architecture, leveraging a pretrained Whisper encoder and class-balanced loss, will outperform existing approaches including MFCC+DNN, VGGish+bi-LSTM, and VAD-based methods in both detection accuracy and temporal precision. Our goal is to demonstrate the model's utility as a noninvasive, clinically grounded tool for automated sleep apnea screening and subtype differentiation. This technology is particularly well-suited for integration into modern consumer electronics, such as mobile applications on smartphones or smart home

devices (eg, smart speakers), offering a seamless and accessible method for at-home monitoring. For example, a dedicated sleep monitoring app could utilize ApneaWhisper to analyze overnight audio, providing users with a detailed report on the frequency and type of respiratory events, which could then be shared with a physician for further diagnostic evaluation. This approach would significantly lower the barrier to initial screening, potentially leading to earlier diagnosis and treatment for millions of undiagnosed individuals.

Materials and Methods

Dataset

We utilized the publicly available PSG-Audio dataset,[11] comprising multichannel PSG signals and synchronized overnight audio recordings from 284 individuals. The dataset was collected at the Center for Sleep Disorders, Ioannina, Greece, between 2008 and 2011. Each participant underwent full nocturnal PSG, which included electroencephalography (EEG), electro-oculography (EOG), electromyography (EMG), electrocardiography (ECG), oxygen saturation (SpO_2), respiratory effort (thoracic and abdominal belts), nasal airflow, and body position. Simultaneously, tracheal and room audio recordings were captured at a sampling rate of 48 kHz. Our analysis focused specifically on the room audio recordings and four clinically annotated respiratory event types by a certified sleep physician: OSA, CSA, MSA, and hypopnea. Audio segments lacking any labeled event were categorized as normal breathing. Frame-level annotations were provided at a temporal resolution of approximately 10 milliseconds (100 frames per second), enabling fine-grained event localization. Unlike the prior work that had to subsample the PSG-Audio dataset due to computational constraints,[27] our transformer-based approach leverages the full dataset with fine-grained resolution.

To prepare the data for training and evaluation, the continuous overnight recordings were divided into non-overlapping 20-second audio clips. This clip length was chosen to balance capturing sufficient context for event identification while maintaining manageable computational load. This process resulted in a total of 228,880 clips. Clip-level labels were assigned based on the presence of apnea events: if one or more events (OSA, CSA, MSA, or hypopnea) occurred within a clip, the most dominant apnea type by duration was assigned as the clip's label; otherwise, the clip was labeled as normal. These labels were used for both clip-level classification and frame-level segmentation tasks.

The class distribution within the dataset was inherently imbalanced, reflecting real-world clinical prevalence: normal breathing accounted for 48.6% of clips, followed by OSA (30.0%), hypopnea (11.4%), MSA (8.1%), and CSA (1.8%). To ensure robust generalization and prevent subject-specific biases, we employed a strict subject-level split, allocating 90% of subjects for the training set and 10% for the test set. This split was stratified across all classes to maintain approximate class proportions in both sets, further mitigating the impact of class imbalance during evaluation.

ApneaWhisper Architecture

The proposed Apnea Whisper model extends the WhisperSeg framework[26] for domain-specific analysis of sleep-disordered breathing. It leverages the robust acoustic features learned by the pretrained Whisper encoder,[25] originally designed for large-scale speech recognition, and adapts it to precisely identify and temporally localize diverse respiratory events in sleep audio recordings.

Figure 1 illustrates the architecture of ApneaWhisper. The input raw audio is initially downsampled from its native 48 kHz to 16 kHz to match the sampling rate expected by the pretrained Whisper encoder. These downsampled continuous recordings are then segmented into non-overlapping 20-second clips. Each clip undergoes a feature extraction step, where it is transformed into an 80-bin log-Mel spectrogram. This is achieved using a 25 ms window length and a 10 ms hop length, resulting in an 80×2000 feature map (80 mel-frequency bins over 2000 frames, given 20 seconds at 100 frames/second) per clip. This representation effectively captures the frequency and temporal characteristics of the audio.

The spectrogram is then fed into a lightweight convolutional frontend. This frontend consists of two 1D convolutional layers, each followed by a GELU (Gaussian Error Linear Unit) activation function. These convolutional layers serve to perform local feature extraction and temporal compression, effectively reducing the sequence length while preserving important acoustic patterns. Specifically, they act as a "bottleneck" to downsample the high-resolution spectrogram, making it computationally feasible for the subsequent Transformer encoder.

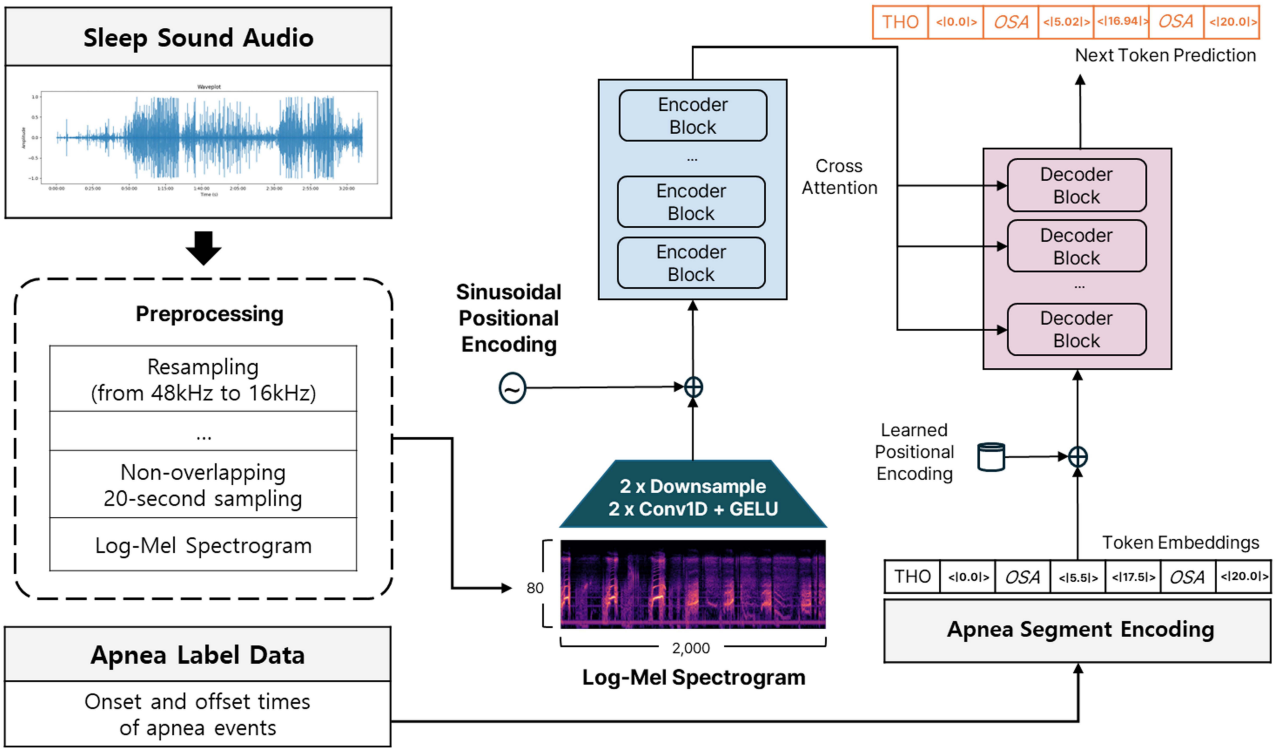

Figure 1 The Architecture of ApneaWhisper.

The output from the convolutional frontend is then combined with sinusoidal positional encodings. These encodings are crucial for Transformer architectures, as they inject information about the relative or absolute position of frames within the sequence, allowing the model to leverage the temporal order of acoustic events, which is inherently lost in the self-attention mechanism alone.

The combined representation is subsequently passed through a stack of Transformer encoder blocks. These blocks, comprising multi-head self-attention mechanisms and feed-forward networks, are adept at capturing long-range temporal dependencies and complex contextual relationships within the audio features, far beyond the capabilities of traditional RNNs or fixed-size convolutional filters. The pretrained weights of the Whisper encoder provide a powerful initialization, enabling effective transfer learning from a vast general audio domain to the specialized domain of sleep breathing sounds.

The encoded representation from the Transformer encoder is then fed into a Transformer decoder. This decoder operates autoregressively, using cross-attention to attend to the encoder's output, and self-attention to attend to its own previously generated tokens. The decoder's task is to generate a sequence of tokens that precisely represent event onsets and offsets, each associated with a specific class label (OSA, CSA, MSA, hypopnea, or normal). This token-based formulation, as adopted from WhisperSeg, is key to enabling fine-grained, frame-level segmentation while inherently maintaining temporal continuity. The decoder learns to predict a vocabulary of tokens where each token signifies the start or end of an event of a particular type (eg, START_OSA, END_OSA, START_HYPOPNEA, END_HYPOPNEA).

Training is performed using the token-based supervision strategy of WhisperSeg, where the ground truth is represented as a sequence of event-specific tokens. A critical aspect of our training strategy is the application of a class-balanced cross-entropy loss. Given the inherent data imbalance in the PSG-Audio dataset, where some apnea subtypes (eg, CSA) are significantly underrepresented, a standard cross-entropy loss would bias the model towards over-predicting dominant classes. The class-balanced cross-entropy loss[28] assigns higher weights to less frequent classes and lower weights to more frequent ones during loss calculation, thereby mitigating the impact of class imbalance and ensuring that the model learns robust representations for all event types. This is typically achieved by setting the "weight" parameter in PyTorch's "Cross Entropy Loss" module, where weights are inversely proportional to class frequencies.

The model was implemented in PyTorch and trained using the AdamW optimizer. We used a fixed learning rate of 3e-6, linear scheduling with 100 warm-up steps, weight decay of 0.01, batch size of 16, and trained for 4 epochs on two NVIDIA RTX A6000 GPUs (CUDA 12.1). Audio was sampled at 16 kHz, with spectrogram time step 10 ms, minimum segment length of 0.1 s, and energy threshold $\varepsilon = 0.02$. We adopted these hyperparameters from WhisperSeg and verified their robustness through 10-fold cross-validation on the 5-class clip-level classification task, where performance was stable across folds (mean macro F1 = 0.414, std = 0.038). Given this stability, we applied the same configuration across all experiments.

Evaluation Settings

To systematically assess the performance of ApneaWhisper, we conducted experiments under varying levels of classification granularity based on apnea event types. Specifically, we defined four distinct label configurations to capture different degrees of clinical specificity:

- 3-class: OSA vs Other apnea types (CSA, MSA, Hypopnea) vs Normal breathing. This task focuses on differentiating the most common apnea type.
- 4-class: OSA vs MSA vs Other apnea types (CSA, Hypopnea) vs Normal breathing. This configuration introduces explicit differentiation of MSA.
- 5-class: OSA vs CSA vs MSA vs Hypopnea vs Normal breathing. This represents the most challenging and clinically detailed task, requiring differentiation of all major event types.

For each setup, we trained a separate instance of ApneaWhisper using token-based supervision with a label vocabulary specifically aligned to the respective task. This allowed us to evaluate the model's robustness and discriminative power as the number and specificity of target classes increased.

To benchmark our model, we compared it against three representative baselines from prior work on audio-based sleep apnea detection:

- MFCC-DNN: A conventional machine learning classifier that extracts handcrafted MFCCs from snore sounds. These features are then fed into a fully connected DNN for classification.[12] This baseline represents a traditional feature engineering approach.
- VGGish-biLSTM: A deep learning model that utilizes pretrained audio embeddings from the VGGish network, which provides robust general-purpose audio features. These embeddings are then processed by a bidirectional Long Short-Term Memory (bi-LSTM) network for temporal classification.[18] This baseline represents a common deep learning approach leveraging pretrained embeddings and recurrent architectures.
- VAD-HMM: A rule-based segmentation method that leverages voice activity detection (VAD) principles and a respiratory-probability-driven Hidden Markov Model (HMM).[13] This approach approximates disordered breathing based on silence and energy changes in audio signals. Due to its design, this method supports only binary segmentation (apnea vs normal) and cannot distinguish among different apnea subtypes.

We evaluated model performance at two distinct levels of temporal granularity, aligning with our objectives:

- Frame-level segmentation: Accuracy and F1-score (weighted and macro-averaged) were computed at a high resolution of 10 ms to assess the model's ability to accurately identify event boundaries (onset and offset). This metric directly reflects the model's fine-grained temporal localization capability.
- Clip-level classification: Accuracy and F1-score (weighted and macro-averaged) were used to evaluate the correctness of the predicted label for each 20-second audio clip. Macro-averaging ensures that the performance on minority classes is not overshadowed by that of majority classes, providing a more balanced assessment in the context of imbalanced datasets.

It is important to note that the MFCC-DNN and VGGish+biLSTM models, by design, are clip-level classifiers and therefore were evaluated across all four classification tasks but were not applicable for frame-level segmentation, as they do not generate time-resolved predictions. Conversely, the VAD-HMM baseline produces frame-level outputs and was included in the segmentation evaluation, but due to its binary classification design, it was applied only to Task 1 (Apnea vs Normal) for segmentation.

To enhance the statistical rigor of our evaluation, we employed Stratified Bootstrapping (1000 resamples) to compute 95% confidence intervals (CIs) for each evaluation metric, ensuring stable performance estimation while preserving class distributions in each resample. Furthermore, we conducted Paired Bootstrap Significance Testing between ApneaWhisper and each baseline model. For each bootstrap sample, we computed the metric differences between paired model predictions. A performance difference was considered statistically significant if the 95% CI of the difference excluded zero, indicating that the observed gains were unlikely due to chance.

All the evaluations used stratified sampling to maintain consistent class distributions across splits. While we did not perform external validation due to limited availability of independent datasets, we recognize this as a limitation. Moreover, while our class-balanced loss and macro-averaged metrics help mitigate label imbalance, we acknowledge that additional techniques (eg, focal loss, data augmentation) may further improve performance and generalizability.

Results

Frame-Level Segmentation Performance

Table 1 summarizes the binary frame-level segmentation results (Apnea vs Normal) in terms of accuracy, weighted F1-score, and macro-averaged F1-score, each with 95% bootstrap confidence intervals (N = 1000). ApneaWhisper markedly outperformed the VAD baseline in all metrics. Specifically, ApneaWhisper achieved an accuracy of 0.6922 (95% CI: 0.6919–0.6925), a weighted F1-score of 0.6892 (95% CI: 0.6889–0.6895), and a macro F1-score of 0.6809 (95% CI: 0.6806–0.6811). In contrast, the VAD baseline showed lower accuracy (0.4610) and a substantially reduced weighted F1-score (0.4712).

As shown in Table 2, paired stratified bootstrap testing confirmed that ApneaWhisper significantly outperformed the VAD baseline, with a mean F1-score gain of +0.1624 (95% CI: 0.1617–0.1629, $p < 0.001$). This gain reflects ApneaWhisper's ability to reduce false positives without sacrificing sensitivity, yielding a more balanced precision–recall trade-off. Clinically, this improvement suggests fewer unnecessary apnea flags compared to VAD while maintaining accurate detection of true events.

Figure 2 shows the multi-class frame-level confusion matrices for the 3-class, 4-class, and 5-class settings. OSA events were consistently detected with high recall (>65%) across all configurations. In contrast, CSA segments were often misclassified as Normal breathing, reflecting their weak or absent acoustic markers. MSA confusion patterns depended on label granularity: in the

Table 1 Frame-Level Segmentation Performance for the Binary Classification

Method	Accuracy		F1-Score (Weighted)		F1-Score (Macro)	
	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI
VAD	0.4610	[0.4606, 0.4614]	0.4712	[0.4707, 0.4716]	0.4598	[0.4593, 0.4602]
ApneaWhisper	**0.6922**	[0.6919, 0.6925]	**0.6892**	[0.6889, 0.6895]	**0.6809**	[0.6806, 0.6811]

Note: Bold values indicate the best performance for each metric.
Abbreviations: Avg, Average; 95% CI, 95% Confidence Interval; VAD, Voice Activity Detection.

Table 2 Frame-Level F1-Score Differences and Significance Based on Paired Stratified Bootstrapping (N = 1000)

Comparison	ΔF1	95% CI	*p*-value
ApneaWhisper vs VAD	0.1624	[0.1617, 0.1629]	<0.0001

Abbreviations: ΔF1, Difference in Mean F1 scores; 95% CI, 95% Confidence Interval; VAD, Voice Activity Detection.

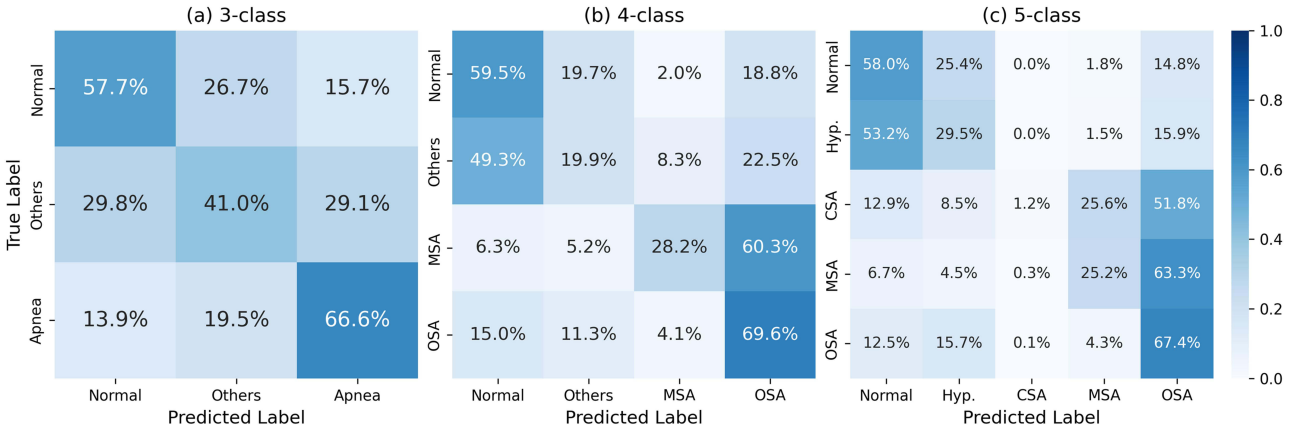

Figure 2 Frame-level confusion matrices across methods: Frame-level results under (**a**) 3-class, (**b**) 4-class, and (**c**) 5-class settings. Rows indicate true labels and columns predicted labels. OSA is consistently detected with high recall, while CSA is often misclassified as Normal and MSA shows mixed errors between OSA and CSA. Hypopnea is frequently confused with Normal due to subtle acoustic cues.
Abbreviations: OSA, Obstructive Sleep Apnea; CSA, Central Sleep Apnea; MSA, Mixed Sleep Apnea; Hyp., Hypopnea.

3-class setting, MSA tended to merge into the "Others" category, while in higher-class configurations it was split between OSA and CSA. Hypopnea events frequently overlapped with Normal predictions, likely due to their lower amplitude and shorter duration. Additional qualitative inspection further revealed that some short OSA events were fragmented into multiple detections at the 10 ms resolution, slightly reducing precision, and that faint breathing noises in noisy segments occasionally triggered false positives for apnea.

Clip-Level Classification Performance

Table 3 presents clip-level classification results for the 2-class, 3-class, 4-class, and 5-class settings. Across all tasks, ApneaWhisper demonstrated superior accuracy and F1-scores compared to MFCC+DNN, VGGish+biLSTM, and VAD. For example, in the 2-class task, ApneaWhisper achieved an accuracy of 0.8234 (95% CI: 0.8185–0.8287), exceeding the VGGish+biLSTM baseline by +0.0939.

As detailed in Table 4, ApneaWhisper consistently outperformed baseline methods with statistically significant gains ($p < 0.001$) across all class configurations. Improvements were most pronounced in 3-class and 4-class settings (+0.2374 and +0.2211 over MFCC+DNN, respectively), highlighting the model's strength in discriminating fine-grained subtypes. In the more challenging 5-class task, the margin over baselines decreased, though ApneaWhisper still maintained a clear advantage. This pattern suggests that while ApneaWhisper effectively captures nuanced acoustic representations, the added complexity and class imbalance in 5-class classification partially reduce the relative effect size.

Figure 3 presents the clip-level confusion matrices. ApneaWhisper maintained a low false-positive rate for Normal breathing even under noisy recording conditions, whereas MFCC+DNN and VGGish+biLSTM frequently misclassified mild OSA as Normal. OSA detection by ApneaWhisper remained robust, with recall exceeding 67% in the 5-class setting. MSA remained challenging for all models; however, ApneaWhisper reduced confusion with OSA and Normal, suggesting improved capture of its mixed obstructive–central acoustic signature. CSA detection performance was limited across all methods (recall <30%), but ApneaWhisper produced fewer CSA→OSA misclassifications compared to the recurrent baseline. Hypopnea events were most often confused with Normal breathing, particularly in clips with low signal-to-noise ratios or subtle respiratory effort changes. Further error analysis revealed that many misclassifications occurred during transitional breathing phases with weak or fragmented acoustic cues, underscoring the potential benefit of integrating complementary physiological signals such as oxygen saturation or respiratory effort.

https://doi.org/10.2147/NSS.S553774

Table 3 Clip-Level Classification Performance Across Different Label Granularities

Method	Accuracy		F1-Score (Weighted)		F1-Score (Macro)	
	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI
2-class (All Apnea Types vs Normal)						
VAD	0.5726	[0.5665, 0.5785]	0.5521	[0.5456, 0.5584]	0.5510	[0.5444, 0.5573]
MFCC+DNN	0.6881	[0.6817, 0.6942]	0.6881	[0.6817, 0.6942]	0.6881	[0.6817, 0.6942]
VGGish+biLSTM	0.7295	[0.7239, 0.7353]	0.7284	[0.7228, 0.7343]	0.7286	[0.7230, 0.7345]
ApneaWhisper	**0.8234**	[0.8185, 0.8287]	**0.8234**	[0.8185, 0.8287]	**0.8234**	[0.8185, 0.8286]
3-class (OSA vs Others vs Normal)						
MFCC+DNN	0.4919	[0.4866, 0.4976]	0.4648	[0.4588, 0.4693]	0.3899	[0.3842, 0.3956]
VGGish+biLSTM	0.5419	[0.5372, 0.5466]	0.5049	[0.4999, 0.5096]	0.4168	[0.4112, 0.4221]
ApneaWhisper	**0.6794**	[0.6736, 0.6853]	**0.6746**	[0.6689, 0.6805]	**0.6273**	[0.6208, 0.6340]
4-class (OSA vs MSA vs Others vs Normal)						
MFCC+DNN	0.3977	[0.3918, 0.4034]	0.3070	[0.3012, 0.3126]	0.4157	[0.4101, 0.4209]
VGGish+biLSTM	0.5032	[0.4995, 0.5070]	0.3070	[0.3005, 0.3135]	0.4475	[0.4433, 0.4518]
ApneaWhisper	**0.6624**	[0.6568, 0.6678]	**0.6501**	[0.6448, 0.6554]	**0.5282**	[0.5208, 0.5356]
5-class (OSA vs CSA vs MSA vs Hypopnea vs Normal)						
MFCC+DNN	0.4242	[0.4183, 0.4300]	0.4588	[0.4531, 0.4647]	0.3007	[0.2948, 0.3065]
VGGish+biLSTM	0.6218	[0.6172, 0.6266]	0.5693	[0.5647, 0.5740]	0.3397	[0.3332, 0.3465]
ApneaWhisper	**0.6306**	[0.6256, 0.6358]	**0.6131**	[0.6086, 0.6178]	**0.3901**	[0.3839, 0.3963]

Note: Bold values indicate the best performance for each metric.
Abbreviations: Avg, Average; 95% CI, 95% Confidence Interval; OSA, Obstructive Sleep Apnea; CSA, Central Sleep Apnea; MSA, Mixed Sleep Apnea; VAD, Voice Activity Detection; MFCC, Mel-frequency Cepstral Coefficients; DNN, Deep Neural Network; VGGish, Google's audio embedding model based on VGG (Visual Geometry Group) architecture; biLSTM, bidirectional Long Short-Term Memory.

Table 4 Pairwise Clip-Level F1-Score Differences and Significance Based on Paired Stratified Bootstrapping (N = 1000)

Comparison	ΔF1	95% CI	*p*-value
2-class (All Apnea Types vs Normal)			
ApneaWhisper vs VAD	0.2724	[0.2667, 0.2783]	<0.0001
ApneaWhisper vs MFCC+DNN	0.1354	[0.1308, 0.1401]	<0.0001
ApneaWhisper vs VGGish+biLSTM	0.0947	[0.0908, 0.0987]	<0.0001
VAD vs MFCC+DNN	−0.1370	[−.1432, −0.1312]	<0.0001
VAD vs VGGish+biLSTM	−0.1777	[−.1841, −0.1720]	<0.0001
MFCC+DNN vs VGGish+biLSTM	−0.0407	[−.0439, −0.0375]	<0.0001

(*Continued*)

Table 4 (Continued).

Comparison	ΔF1	95% CI	*p*-value
3-class (OSA vs Others vs Normal)			
ApneaWhisper vs MFCC+DNN	0.2374	[0.2307, 0.2444]	<0.0001
ApneaWhisper vs VGGish+biLSTM	0.2105	[0.2042, 0.2167]	<0.0001
MFCC+DNN vs VGGish+biLSTM	−0.0269	[−.0299, −0.0239]	<0.0001
4-class (OSA vs MSA vs Others vs Normal)			
ApneaWhisper vs MFCC+DNN	0.2211	[0.2122, 0.2295]	<0.0001
ApneaWhisper vs VGGish+biLSTM	0.2213	[0.2102, 0.2319]	<0.0001
MFCC+DNN vs VGGish+biLSTM	0.0002	[−.0077, 0.0081]	0.9920
5-class (OSA vs CSA vs MSA vs Hypopnea vs Normal)			
ApneaWhisper vs MFCC+DNN	0.0895	[0.0829, 0.0958]	<0.0001
ApneaWhisper vs VGGish+biLSTM	0.0503	[0.0453, 0.0557]	<0.0001
MFCC+DNN vs VGGish+biLSTM	−0.0392	[−.0458, −0.0322]	<0.0001

Abbreviations: ΔF1, Mean F1 score change; 95% CI, 95% Confidence Interval; VAD, Voice Activity Detection; MFCC, Mel-frequency Cepstral Coefficients; DNN, Deep Neural Network; VGGish, Visual Geometry Group; biLSTM, bidirectional Long Short-Term Memory.

Discussion

Key Advantages of ApneaWhisper

This study highlights two pivotal advantages of ApneaWhisper in the context of audio-based sleep apnea detection. First, unlike most prior works that treat apnea detection as a coarse clip-level classification task, ApneaWhisper adopts a segmentation-oriented approach operating at a fine 10 ms temporal resolution. This enables precise localization of respiratory events within 20-second audio clips, accurately delineating their onset and offset. Such granularity supports clinically meaningful applications, including precise calculation of the Apnea–Hypopnea Index (AHI), which depends on the frequency and duration of individual events. It also facilitates the identification of subtle or brief events that may be missed by coarser methods, and lays the groundwork for real-time feedback systems. As evidenced by its consistently higher F1-scores in segmentation compared to the rule-based VAD-HMM baseline, ApneaWhisper's temporal precision enhances its diagnostic potential.

Second, ApneaWhisper demonstrates superior discriminative ability across various apnea subtypes—including MSA, CSA, and hypopnea—beyond the highly prevalent OSA. Previous approaches have typically collapsed all apnea events into a single "apnea" class or struggled with subtype separation due to subtle acoustic differences and severe data imbalance. By contrast, our model benefits from the pretrained Whisper encoder's general-purpose acoustic representations, which, combined with Transformer-based long-range modeling, allow it to capture nuanced distinctions among apnea types. This leads to improved separation in both frame- and clip-level tasks. As shown in the confusion matrices, traditional classifiers frequently misclassify acoustically ambiguous subtypes (eg, MSA, CSA) as OSA or normal breathing, whereas ApneaWhisper provides comparatively improved differentiation—an important step toward more precise and personalized diagnosis.

Clinical Implications

One of the most clinically meaningful aspects of ApneaWhisper is its fine-grained 10 ms frame-level segmentation, which enables precise delineation of apnea and hypopnea onset/offset times. Such temporal precision directly improves Apnea–Hypopnea Index (AHI) estimation because small changes in event duration or counting criteria can shift severity classification and downstream management decisions. This has been demonstrated for hypopnea scoring rules, where alternative AASM definitions substantially change AHI and severity assignment.[29]

Figure 3 Clip-level confusion matrices across methods: Clip-level classification results for (**a**) 3-class, (**b**) 4-class, and (**c**) 5-class tasks, comparing MFCC+DNN, VGGish +biLSTM, and ApneaWhisper. ApneaWhisper achieves higher recall for OSA, lower false positives for Normal, and improved subtype differentiation compared to baselines, though CSA and Hypopnea remain challenging.
Abbreviations: OSA, Obstructive Sleep Apnea; CSA, Central Sleep Apnea; MSA, Mixed Sleep Apnea; Hyp., Hypopnea.

Moreover, beyond AHI alone, event duration itself carries prognostic information: shorter respiratory events were associated with higher all-cause mortality in the Sleep Heart Health Study, highlighting that accurate temporal measurement adds clinically relevant risk stratification beyond simple event counts.[30]

Fine-scale temporal detection can also support real-time or near-real-time behavioral feedback—most notably positional therapy—which reduces supine time and lowers AHI in positional OSA. Randomized and controlled studies of vibrotactile "sleep position trainer" devices and recent meta-analyses/reviews show meaningful AHI reductions and acceptable adherence in appropriate phenotypes, underscoring the translational potential of frame-level monitoring for just-in-time interventions.[31]

In addition, improved differentiation between apnea subtypes carries significant clinical implications. For example, while OSA is typically managed with continuous positive airway pressure (CPAP) therapy, CSA may require alternative approaches such as adaptive servo-ventilation (ASV), supplemental oxygen, or addressing underlying cardiovascular or neurological conditions.[21,32] MSA often necessitates tailored interventions that combine elements of OSA and CSA management, as standard CPAP therapy may fail to resolve the central component. Accurate hypopnea detection also informs decisions about oxygen desaturation monitoring and airway management. Consequently, the ability to reliably

distinguish between these subtypes can support more individualized treatment plans, improve therapy adherence, and potentially enhance long-term patient outcomes.[21,32]

Technical and Evaluation Limitations

Despite these advantages, several challenges remain. Detection performance for CSA and hypopnea, while improved, remains lower than for OSA. This is consistent with prior reports, such as Le et al[33] which showed reduced accuracy for hypopnea in real-time OSA detection systems. These limitations reflect the intrinsic difficulty of detecting certain events using audio alone. CSA, in particular, lacks distinct acoustic features such as snoring or gasping, making it acoustically similar to normal breathing or silence. MSA, which blends central and obstructive features, also presents ambiguous audio patterns. Reliable classification of these events may require the integration of physiological signals such as respiratory effort or oxygen saturation. Furthermore, the high-resolution segmentation offered by ApneaWhisper introduces computational overhead. Transformer-based architectures like Whisper have substantial memory and compute requirements, which may hinder real-time deployment on edge devices. While Whisper's pretrained features are robust to noise, unpredictable artifacts in unconstrained home environments may still degrade performance.

In addition to these technical challenges, several evaluation-related limitations warrant mention. First, our dataset was collected from a single sleep center in Greece, and we did not conduct external validation on independent populations or recording setups. While we performed 10-fold cross-validation during hyperparameter selection on the 5-class clip-level task and observed consistent performance across folds, broader evaluation across institutions, patient demographics, and devices—including home-based recordings—is necessary to assess generalizability. Second, although class-balanced loss and macro-averaged metrics were employed to mitigate imbalance, rare subtypes such as CSA (~1.8%) and MSA (~8.1%) remain difficult to evaluate robustly. Their scarcity can bias both training and performance metrics. Future work could incorporate sampling techniques (eg, SMOTE), targeted ablation studies, or synthetic data augmentation to better account for rare-event subtypes. Third, we did not perform systematic hyperparameter tuning (eg, grid search or validation-based selection), instead adopting settings from WhisperSeg and verifying them through pilot runs. Although this approach yielded stable performance, more exhaustive tuning could further improve generalization and efficiency. Additionally, a formal ablation study to disentangle the relative contributions of Whisper pretraining and the Transformer architecture was beyond the scope of this work, but represents an important avenue for future research to more precisely characterize the sources of ApneaWhisper's performance gains.

Future Research Directions

Future work could proceed along multiple complementary directions. First, model accuracy could be further improved by addressing the limitations identified in this study. Rare-event handling may be enhanced through targeted sampling techniques (eg, SMOTE), synthetic data augmentation, or dedicated ablation studies to improve the detection of underrepresented subtypes such as CSA and MSA. Broader external evaluation—across multiple institutions, patient demographics, languages, and recording conditions—will be essential to assess and improve generalizability beyond the single-site dataset used in this work. These efforts could be complemented by systematic hyperparameter optimization (eg, grid search, Bayesian optimization) and exploration of alternative self-supervised audio encoders or domain-specific pretraining using large-scale medical or respiratory sound datasets. However, domain-specific pretraining may underperform if the available corpora are not sufficiently large and diverse.[21]

Second, improving model efficiency is crucial for real-world application. Architectural optimizations—such as model compression techniques (eg, quantization, pruning, knowledge distillation[34]) and exploration of more efficient Transformer variants[35]—may enable real-time, on-device inference without sacrificing accuracy. Lightweight attention mechanisms and other efficiency-focused designs could be integrated to further reduce computational overhead while preserving performance.

Third, deployment strategies should be designed to balance performance, privacy, and scalability. For example, a hybrid client–server framework—where the client device performs lightweight preprocessing (eg, denoising, downsampling, or feature extraction) before transmitting data to a secure server for full inference—could enable low-latency applications without requiring full model execution on edge hardware. Such designs can also enhance data privacy by avoiding raw audio transmission and ensuring compliance with health data regulations. Furthermore, incorporating

explainability tools, such as interpretable attention maps or feature attribution methods, may help clinicians understand the basis for predictions and increase trust in the system's outputs.

Fourth, multimodal integration with physiological signals—such as oxygen saturation, respiratory effort, or heart rate—offers a promising path to improving the detection of acoustically ambiguous events like CSA and hypopnea.[36,37] Such signals can now be obtained noninvasively through consumer-grade wearable devices[38] or home-based monitoring systems, making them more practical for large-scale screening compared to full polysomnography (PSG). Combining audio-based respiratory analysis with complementary physiological data can provide a more comprehensive and robust representation of respiratory events, while maintaining the accessibility of noninvasive screening. Different fusion strategies (eg, early vs late fusion) and domain-specific self-supervised pretraining[30] may further enhance robustness and clinical reliability.

Conclusion

In this study, we presented ApneaWhisper, a novel Transformer-based audio segmentation framework designed for noninvasive and fine-grained detection of sleep-disordered breathing events. By leveraging a token-based modeling approach adapted from WhisperSeg and operating at a high temporal resolution of 10 ms, ApneaWhisper enables precise localization and detailed classification of respiratory events including OSA, MSA, CSA, and hypopnea directly from raw sleep audio. Our model consistently and significantly outperformed conventional machine learning baselines (MFCC +DNN, VGGish+bi-LSTM) and rule-based methods (VAD-HMM) in both frame-level segmentation and clip-level classification, demonstrating its superior utility for scalable and automated sleep apnea screening.

Beyond methodological contributions, ApneaWhisper's ability to distinguish among apnea subtypes has direct clinical relevance. Accurate subtype differentiation can inform tailored treatment strategies—such as CPAP for OSA, adaptive servo-ventilation or supplemental oxygen for CSA, and combination approaches for MSA—enabling more personalized care pathways. Its fine-grained temporal precision also supports accurate Apnea–Hypopnea Index estimation and event duration analysis, both of which influence severity grading and long-term risk stratification.

From a translational perspective, ApneaWhisper could be integrated into real-world workflows through mobile applications, home monitoring devices, or hybrid client–server systems for clinical triage. Such deployment strategies may facilitate early detection and ongoing monitoring outside the sleep laboratory, reducing diagnostic delays and broadening access. To ensure safe and equitable implementation, future work should address cross-population validation, robustness to diverse recording conditions, and compliance with privacy regulations. Incorporating explainability mechanisms could further support clinician trust and patient engagement in regulated clinical environments.

Despite the remaining challenges—particularly in detecting acoustically subtle or underrepresented subtypes—ApneaWhisper provides a strong foundation for next-generation, noninvasive diagnostic systems that combine high-resolution analysis, clinical interpretability, and deployment feasibility in both home and clinical settings.

Abbreviations

AHI, Apnea–Hypopnea Index; bi-LSTM, Bidirectional Long Short-Term Memory; CNN, Convolutional Neural Network; CSA, Central Sleep Apnea; DNN, Deep Neural Network; ECG, Electrocardiography; EEG, Electroencephalography; EMG, Electromyography; EOG, Electrooculography; GELU, Gaussian Error Linear Unit; GMM, Gaussian Mixture Model; HMM, Hidden Markov Model; HNR, Harmonics-to-Noise Ratio; LPC, Linear Prediction Coefficient; LSTM, Long Short-Term Memory; MFCC, Mel-frequency Cepstral Coefficient; MSA, Mixed Sleep Apnea; OSA, Obstructive Sleep Apnea; PSG, Polysomnography; RNN, Recurrent Neural Network; SpO_2, Oxygen Saturation; SSL, Self-Supervised Learning; SVM, Support Vector Machine; VAD, Voice Activity Detection.

Data Sharing Statement

The dataset supporting the results reported in this paper, PSG-Audio, is a publicly available dataset. It can be accessed and downloaded from https://www.scidb.cn/en/detail?dataSetId=778740145531650048. No additional unpublished data from this study are available for sharing, as all relevant data are either part of the publicly accessible dataset or are derived results presented within the paper.

Ethics Approval and Informed Consent

This study complies with the principles of the Declaration of Helsinki. It utilized the publicly available PSG-Audio dataset, originally collected and annotated by the Sleep Study Unit of Sismanoglio – Amalia Fleming General Hospital of Athens, and openly accessible via https://doi.org/10.11922/sciencedb.00345 under a CC BY 4.0 license. According to the dataset descriptor, all recordings were conducted with written informed consent from participants, and ethical approval was obtained from the Local Ethics Committee of Sismanoglio Hospital, in compliance with the European Regulation for Personal Data Protection (GDPR). All personal identifiers were removed, and data were fully anonymized prior to public release. Our study involved only secondary analysis of fully de-identified, pre-existing public data, with no new data collection, no direct contact with human subjects, and no manipulation of clinical procedures. In accordance with institutional policies, the Hanyang University Institutional Review Board (HYU-IRB) reviewed the study protocol and determined that it was exempt from further ethical review.

Author Contributions

Y.K., Data curation, Software, Validation, Formal analysis, Writing—original draft. M. Kim, Software, Validation, Visualization, Writing—review and editing. J. Shin, Conceptualization, Formal analysis, Supervision, Writing—review and editing. M. Ko, Conceptualization, Funding acquisition, Data curation, Formal analysis, Writing—review and editing, Supervision. All authors agreed on the choice of journal; reviewed and approved all versions of the article including the final version and any significant changes during the proofing stage; and accept full responsibility for the integrity of the work.

Funding

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (RS-2023-00224524). It was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (IITP-2024-RS-2024-00423071) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Young T, Evans L, Finn L, Palta M. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. *Sleep*. 1997;20(9):705–706. doi:10.1093/sleep/20.9.705
2. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;7(8):687–698. doi:10.1016/S2213-2600(19)30198-5
3. Ghavami T, Kazeminia M, Ahmadi N, et al. Global prevalence of obstructive sleep apnea in the elderly and related factors: a systematic review and meta-analysis study. *J Perianesth Nurs*. 2023;38(6):865–875. doi:10.1016/j.jopan.2023.01.018
4. Tregear S, Reston J, Schoelles K, Phillips B. Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis. *J Clin Sleep Med*. 2009;5(6):573–581. doi:10.5664/jcsm.27662
5. Abeyratne U, De Silva S, Hukins C, Duce B. Obstructive sleep apnea screening by integrating snore feature classes. *Physiol Meas*. 2013;34 (2):99–121. doi:10.1088/0967-3334/34/2/99
6. Yaggi HK, Concato J, Kernan WN, et al. Obstructive sleep apnea as a risk factor for stroke and death. *N Engl J Med*. 2005;353(19):2034–2041. doi:10.1056/NEJMoa043104
7. Ho M, Brass S. Obstructive sleep apnea. *Neurol Int*. 2011;3(3):15. doi:10.4081/ni.2011.e15
8. Nakano H, Hirayama K, Sadamitsu Y, et al. Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept. *J Clin Sleep Med*. 2014;10(1):73–78. doi:10.5664/jcsm.3364
9. The Sleep Foundation. How much does a sleep study cost? SleepFoundation.org. 2024. Available from: https://www.sleepfoundation.org/sleep-apnea/how-much-does-a-sleep-study-cost. Accessed July 12, 2024.
10. Penzel T, Sabil A. The use of tracheal sounds for the diagnosis of sleep apnoea. *Breathe*. 2017;13(2):37–45. doi:10.1183/20734735.008817
11. Korompili G, Amfilochiou A, Kokkalas L, et al. PSG-Audio: a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Sci Data*. 2021;8(1):197. doi:10.1038/s41597-021-00977-w
12. Sillaparaya A, Bhatranand A, Sudthongkong C, et al. Obstructive sleep apnea classification using snore sounds based on deep learning. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE; 2022:1152–1155. doi:10.23919/APSIPAASC55919.2022.9980194.
13. Korompili G, Kokkalas L, Mitilineos SA, et al. Detecting apnea/hypopnea events time location from sound recordings for patients with severe or moderate sleep apnea syndrome. *Appl Sci*. 2021;11(15):6888. doi:10.3390/app11156888
14. Mostafa SS, Mendonça F, Ravelo-García AG, Morgado-Dias F. A systematic review of detecting sleep apnea using deep learning. *Sensors*. 2019;19 (22):4934. doi:10.3390/s19224934

15. Li R, Li W, Yue K, Li Y. Convolutional neural network for screening of obstructive sleep apnea using snoring sounds. *Biomed Signal Process Control*. 2023;86:104966. doi:10.1016/j.bspc.2023.104966
16. Cheng S, Wang C, Yue K, et al. Automated sleep apnea detection in snoring signals using long short-term memory neural networks. *Biomed Signal Process Control*. 2022;71:103238. doi:10.1016/j.bspc.2021.103238
17. Wang B, Tang X, Ai H, et al. Obstructive sleep apnea detection based on sleep sounds via deep learning. *Nat Sci Sleep*. 2022;14:2033–2045. doi:10.2147/NSS.S373367
18. Serrano S, Patané L, Serghini O, Scarpa M. Detection and classification of obstructive sleep apnea using audio spectrogram analysis. *Electronics*. 2024;13(13):2567. doi:10.3390/electronics13132567
19. Lillini D, Aironi C, Migliorelli L, Gabrielli, L, Squartini, S, et al. Sicrnn: a siamese approach for sleep apnea identification via tracheal microphone signals. *Sensors*. 2024;24(23):7782. doi:10.3390/s24237782
20. ElMoaqet H, Eid M, Ryalat M, Penzel, T. Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals. *Sensors*. 2020;20(18):5037. doi:10.3390/s20185037
21. Kushida CA, Littner MR, Hirshkowitz M, et al. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. *Sleep*. 2005;28(4):499–521. doi:10.1093/sleep/28.4.499
22. Iber C, Ancoli-Israel S, Chesson A, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Westchester, IL: American Academy of Sleep Medicine; 2007.
23. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. *Adv Neural Inf Process Syst*. 2020;33:12461–12471.
24. Hsu WN, Bolte B, Tsai YHH, et al. Hubert: self-supervised speech pre-training by masked prediction of hidden units. *IEEE/ACM Trans Audio Speech Lang Process*. 2021;29:3451–3460. doi:10.1109/TASLP.2021.3122291
25. Radford A, Kim JW, Xu T, et al. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356. 2022. Available from: https://arxiv.org/abs/2212.04356. Accessed August 14, 2025.
26. Gu N, Lee K, Barsha M, Ram, SK, You, G, Hahnloser, RHR, et al. Positive transfer of the Whisper speech transformer to human and animal voice activity detection. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2024. doi:10.1101/2023.09.30.560270.
27. Ding L, Peng J, Song L, Zhang X. Automatically detecting apnea–hypopnea from snore signals based on a VGG19+LSTM architecture. *Biomed Signal Process Control*. 2023;80:104351. doi:10.1016/j.bspc.2022.104351
28. Cui Y, Jia M, Lin TY, et al. Class-balanced loss based on effective number of samples. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2019:9208–9217. doi:10.1109/CVPR.2019.00943.
29. Ruehland WR, Rochford PD, O'Donoghue FJ, et al. The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index. *Sleep*. 2009;32(2):150–157. doi:10.1093/sleep/32.2.150
30. Butler MP, Emch JT, Rueschman M, et al. Apnea–hypopnea event duration predicts mortality in men and women in the sleep heart health study. *Am J Respir Crit Care Med*. 2019;199(7):903–912. doi:10.1164/rccm.201804-0758OC
31. van Maanen JP, de Vries N, et al. Long-term effectiveness and compliance of positional therapy with the sleep position trainer in the treatment of positional obstructive sleep apnea syndrome. *Sleep*. 2014;37(7):1209–1215. doi:10.5665/sleep.3840
32. Oldenburg O, Lamp B, Faber L, Teschler H, Horstkotte D, Topfer V. Sleep-disordered breathing in patients with symptomatic heart failure: a contemporary study of prevalence in and characteristics of 700 patients. *Eur Respir J*. 2007;30(1):125–131. doi:10.1183/09031936.00108406
33. Le VL, Kim D, Cho E, et al. Real-time detection of sleep apnea based on breathing sounds and prediction reinforcement using home noises: algorithm development and validation. *J Med Internet Res*. 2023;25:e44818. doi:10.2196/44818
34. Li Z, Li H, Meng L. Model compression for deep neural networks: a survey. *Computers*. 2023;12(3):60. doi:10.3390/computers12030060
35. Niizumi D, Takeuchi D, Yasuda M, Nguyen, BT, Ohishi, Y, Harada, N, et al. Towards pre-training an effective respiratory audio foundation model. arXiv preprint arXiv:2505.15307. 2025. Available from: https://arxiv.org/abs/2505.15307. Accessed September 15, 2025.
36. Zhang Y, Zhou L, Zhu S et al, et al. Deep learning for obstructive sleep apnea detection and severity assessment: a multimodal signals fusion multiscale transformer model. *Nat Sci Sleep*. 2025;17:1–15. doi:10.2147/NSS.S492806
37. Romero HE, Ma N, Brown GJ, Johnson S. Obstructive sleep apnea screening with breathing sounds and respiratory effort: a multimodal deep learning approach. *Proc Interspeech*. 2023;5451–5455. doi:10.21437/Interspeech.2023-209
38. Abd-Alrazag A, Aslam H, AlSaad R et al, et al. Detection of sleep apnea using wearable AI: systematic review and meta-analysis. *J Med Internet Res*. 2024;26():e58187. doi:10.2196/58187

Nature and Science of Sleep

Dovepress
Taylor & Francis Group

Publish your work in this journal

Nature and Science of Sleep is an international, peer-reviewed, open access journal covering all aspects of sleep science and sleep medicine, including the neurophysiology and functions of sleep, the genetics of sleep, sleep and society, biological rhythms, dreaming, sleep disorders and therapy, and strategies to optimize healthy sleep. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/nature-and-science-of-sleep-journal