Sleep Sound Classification Using ANC-Enabled Earbuds

Kenneth Christofferson[1,4], Xuyang Chen[3,5], Zeyu Wang[2,5], Alex Mariakakis[1,4], Yuntao Wang[2,5]

University of Toronto[1], Tsinghua University[2], Beijing University of Posts and Telecommunications[3], Department of Computer Science[4], Department of Computer Science and Technology[5]
kenc@cs.utoronto.edu, cyx2019@bupt.edu.cn, wangzeyu19@mails.tsinghua.edu.cn
yuntaowang@tsinghua.edu.cn, mariakakis@cs.utoronto.edu

2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 978-1-6654-1647-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/PerComWorkshops53856.2022.9767394

***Abstract*—Standard sleep quality assessment methods require custom hardware and professional observation, limiting the diagnosis of sleep disorders to specialized sleep clinics. In this work, we leverage the internal and external microphones present in active noise-cancelling earbuds to distinguish sounds associated with poor or disordered sleep, thereby enabling at-home continuous sleep sound monitoring. The sleep sounds our system is able to recognize include, but are not limited to, snoring, teeth grinding, and restless movement. We analyze the resulting dual-channel audio using a lightweight deep learning model built around a variation of the temporal shift module that has been optimized for audio. The model was designed to have a low memory and computational footprint, making it suitable to be run on a smartphone or the earbuds themselves. We evaluate our approach on a dataset of 8 sound categories generated from 20 participants. We achieve a classification accuracy of 91.0% and an F1-score of 0.845.**

***Index Terms*—earbuds, audio event detection, temporal shift module, health, sleep quality**

I. Introduction

Sleep disorders are often difficult to self-diagnose because it is impossible to recall the severity, frequency, and nature of symptoms that occur while unconscious. This is not only true for respiratory disorders like obstructive sleep apnea and mouth breathing, but also for conditions like sleep bruxism (i.e., teeth grinding). Further, these conditions often do not manifest acute symptoms until damage has already been done. The gold-standard diagnostic technique for many sleep disorders is polysomnography [1], which typically requires that a patient spend at least one night in a specialized sleep clinic. During that time, the patient is instrumented with a battery of sensors ranging from airflow and oxygen saturation sensors for measuring breathing to EMG sensors and movement belts for measuring motion [2]. These studies can be onerous and expensive, meaning that sleep disorders can go undiagnosed until symptoms become severe.

Prior work has shown that it is possible to automatically detect sleep disorder symptoms using the sensors embedded in commodity devices like smartphones [3], smartwatches [4], and wireless routers [5]. However, these approaches are limited in a couple of ways: (1) they are only sensitive to overt symptoms like body movement and snoring, and (2) they can have trouble determining who is presenting the symptom when there are multiple people in the same bed.

Wireless earbuds are quickly becoming a pervasive commodity device, with global sales reaching 129 million units in 2020 [6]. In fact, commercial earbuds like the Bose Sleepbuds II[1] are designed to be worn at night so that people can listen to music and audiobooks as they fall asleep. Not only are earbuds becoming more pervasive, but they are also becoming more sophisticated and sensor-laden. Most notably, many earbuds provide active noise cancellation (ANC), which requires both an interior and exterior microphone to function. The inner microphone's physical isolation and placement in the occluded ear canal allow it to perceive internal body sounds that microphones in other ubiquitous devices cannot. This configuration of microphones also makes it trivial to identify the person producing those sounds since the internal microphone is far more sensitive to the wearer than others in the room.

In this work, we demonstrate that audio collected from ANC-enabled earbuds can be analyzed to identify body sounds associated with sleep disorders. We first modified a set of commodity earbuds so that we could record voluntarily-produced sleep sounds (e.g., grinding, snoring, breathing movement) from 20 participants. We then created a deep learning model that is able to classify those sounds. The model is based on a traditional convolutional neural network (CNN) architecture but is made more computationally efficient by using a modified version of the temporal shift module [7] that was optimized for audio. The model's parameters total 236 kB, making it compact enough to load onto a smartphone or even the earbuds themselves. In our evaluation, we found that our model yields an accuracy of 91.0% and an F1-score of 0.845 across all participants without the need for subject-specific training data.

To summarize, our main contributions are as follows:

1) A lightweight deep learning audio event classification model leveraging a new variant of the temporal shift module [7] for audio data,

[1]https://www.bose.com/en_us/products/wellness/noise_masking_sleepbuds/noise-masking-sleepbuds-ii.html

2) The application of that model on sleep sound events recorded from the microphones embedded in ANC-enabled earbuds, and
3) An evaluation of our approach on a dataset of 8 sound categories generated from 20 participants.

II. Methods

We first describe our data collection process, including the hardware we used to record sleep sounds and the procedure that participants were asked to follow. We then outline the steps we used to process the resulting audio, including pre-processing, data augmentation, and sleep sound classification.

A. Hardware

We modified a Sony WF-1000MX3 ANC-enabled earbud[2] for data collection. Both the interior and exterior microphones were wired out and connected to a recorder for audio capture. The microphones recorded audio at 48 kHz; however, our early experiments showed that most of the frequency content for our target sleep sounds was below 2 kHz. As a result, we downsampled the audio to 4 kHz to enable a lightweight classifier that requires less memory.

B. Data Collection

Our dataset comprises eight acoustic event classes that prior literature has shown to be correlated with poor sleep or might otherwise be produced at night [8]–[10]. It also includes sounds associated with conditions like bruxism, COPD, and asthma that become less regulated during sleep [11]–[14]. Since many of these sounds can be produced in multiple ways (e.g., mouth- and nose-breathing), we further divided our classes into subclasses. This resulted in 20 sound types that were recorded during our experiments. The complete list of sounds can be found in Table I.

We would have preferred to record our dataset during people's natural sleep; however, some sounds and behaviors are naturally produced more frequently than others, which would have led to a highly imbalanced dataset and the possibility of missing classes. Reliably identifying and labelling sleep sounds from continuous sleep also requires a controlled sleep environment and additional instrumentation, which can impact people's natural sleep behaviors [15]. By recording voluntarily produced sounds, we were able to include more types of sounds in our dataset than would have been possible had we recorded participants while they were asleep. Other audio analysis tasks have used voluntary sounds to validate their methods, most notably cough detection [16], [17].

We recruited 20 participants (13 male, 7 female) with an average age of 31.5 years ($s.d. = 8.2$). Because of the COVID-19 pandemic, the study was executed in participants' own homes and instructions were delivered remotely by researchers to maintain social distancing. Hardware was disinfected before and after being sent to each participant. The remote nature of the study introduced real-world noise into our dataset, as nearby appliances and cohabitants contributed extraneous noise to recordings. The uncontrolled nature of our data collection procedure introduced significant variability into our dataset, which simultaneously increased the realism of our dataset and presented challenges to accurate audio classification.

[2]https://www.sony.ca/en/electronics/truly-wireless/wf-1000xm3

TABLE I
THE NUMBER OF ONE-SECOND SEGMENTS FOR EACH SOUND IN OUR SLEEP SOUND DATASET.

Class	Sub-Class	Recorded	Post-Augmentation
Environment	Background sound	813	1,626
Grinding Teeth	Front-to-back quietly	1,003	1,003
	Side-to-side quietly	1,687	1,687
	Front-to-back normally	1,821	1,821
	Side-to-side normally	1,748	1,748
	Clenching teeth	984	984
Swallowing	Swallowing	689	2,058
Speaking	Speaking normally	1,679	1,679
	Murmuring	1,480	1,480
Breathing	Through nose normally	1,523	1,523
	Through mouth normally	1,055	1,055
	Through nose quietly	1,476	1,476
	Through mouth quietly	1,785	1,785
Throat Sounds	Coughing	1,140	1,140
	Clearing throat	1,821	1,821
Snoring	Snoring	505	1,497
Body Motion	Moving legs	177	354
	Moving arms	366	732
	Flipping over	653	1,296
	Re-positioning	634	1,268

Participants were asked to lay in their own bed to better recreate how the sounds would be made during sleep. When applicable, participants were provided with instructions detailing how some of the sounds should be performed. For example, participants were asked to read a standard paragraph for all the recordings related to speech. Some of the sounds in Table I are discrete events (e.g., cough), while others are more continuous (e.g., speech and breathing). This nuance made it challenging to generate a dataset that was balanced in terms of instances and duration. As a result, we asked participants to record themselves making each sound for a fixed duration between 15–60 seconds depending on a variety of factors. For example, participants were asked to speak for 60 seconds and to cough for 30 seconds — the latter being shorter since it is difficult to cough repeatedly. Participants were encouraged to take breaks between recordings and were allowed to skip any activities that led to discomfort.

C. Pre-Processing

Mel-frequency cepstrum coefficients (MFCCs) and the short-time Fourier transform (STFT) are the most common audio feature-extraction techniques used in deep learning audio classification systems. MFCCs are better suited for speech data [18], which only comprises a subset of our dataset. As a result, we extracted STFT features and normalized them with linear scaling. Since our audio events did not have uniform length, we generated our dataset by splitting our

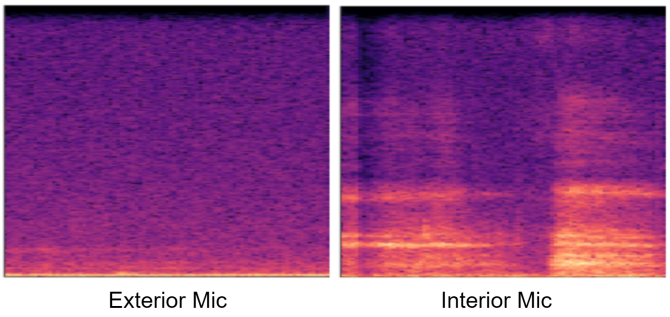

Fig. 1. Spectrograms showing a single respiratory cycle (inspiration and exhalation) as recorded by the (left) exterior and (right) interior microphones.

audio recordings into one-second sliding windows with 16.7% overlap ($1/6^{th}$ of a second).

Despite our efforts to collect an equal number of one-second windows for each sound type, we ended up with an imbalanced dataset. We ameliorated this issue using data augmentation [19], [20]. We synthetically generated new audio examples of underrepresented classes using the following audio data augmentation techniques [21]:

- **Time stretching:** Adjusting the length of an audio sample while maintaining pitch
- **Pitch shifting:** Adjusting the pitch of an audio sample without changing its length
- **Background noise:** Adding background noise, or mixing generated noise with a sample

To mitigate overreliance on data augmentation, we also employed random undersampling on the over-represented classes to balance the training set. In the end, each training fold of our models had no more than 1,750 samples of each sound.

D. Classification Model

Although standard 2D CNNs and recurrent network architectures have been used to great effect in processing spectrogram data [22], they can be computationally intensive and thus of limited utility in mobile and embedded systems. We propose a CNN-variant that leverages a modified temporal shift module (TSM) [7], which was designed by Lin et al. to reduce the computational intensity of video processing networks that rely on 3D convolutional layers. TSM layers replace 3D convolutions with 2D convolutions and shift data along the filter dimension in order to make data from adjacent timesteps consumable by 2D convolution (Fig. 2). Because TSM layers only shift the location of data according to a hyperparameter, they do not require any additional learnable parameters and therefore significantly decrease computational overhead. That being said, TSM layers do not provide the same exchange of temporal information as a higher dimensional convolution.

The modified TSM layers we use for audio data, originally applied to audio super resolution [23], operate on the output from an initial 2D convolution and return a 3D tensor with frequency bin, time point, and convolutional filter dimensions.

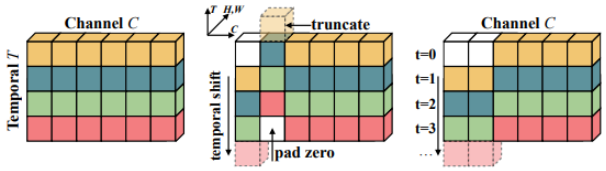

Fig. 2. A visualization illustrating the use of temporal shifting in the TSM module (visualization taken from Lin et al. [7]).

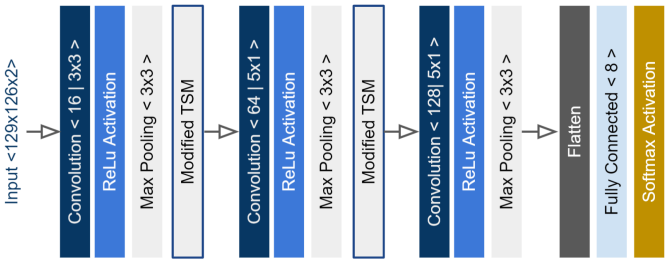

Fig. 3. The network architecture we use for our final evaluation. There are three 2D convolutional layers followed by max pooling. The first two layers are shifted by a modified TSM layer. The network is completed with a fully connected layer.

The TSM iterates over the filter dimension and shifts some filters along the time dimension, effectively aligning frequency content from adjacent timesteps across filters. In this way, 1D convolutions oriented along the frequency dimension are able to consume information from different time steps, which in turn enables them to embed temporal information. Fig. 3 shows the CNN architecture we use in our final evaluation, including the modified TSM layers that follow some of the convolutional layers.

III. Results

In this section, we compare our lightweight classifier to a set of baseline CNNs with varied depth and structure to investigate the tradeoff between accuracy and computational efficiency. The set of baseline models include the following four configurations:

- **Tiny CNN:** 3 convolution layers, 32–64 filters per layer
- **Very Small CNN:** 4 convolutional layers, 32–64 filters per layer
- **Small CNN:** 5 convolutional layers, 64 filters per layer
- **Medium CNN:** 7 convolutional layers, 64 filters per layer

All models were trained with the Adam optimizer to minimize categorical cross-entropy loss. The models were trained for 300 epochs at a batch size of 128 using TensorFlow 2.4.1 and an NVIDIA GeForce GTX 980Ti GPU. Random selection search was employed to tune hyperparameters. To demonstrate that our approach generalizes across participants, we trained and tested our models using 5-fold cross-validation such that all of a given participant's data was kept in the same fold; no synthetic data generated using data augmentation was included in any of the test sets.

TABLE II
EXPERIMENTAL RESULTS FOR CLASSIFIER ACCURACY AND COMPUTATIONAL EFFICIENCY

Type	Acc.	F1	Params	Size	FLOPs
Tiny Modified TSM	88.1%	0.815	19.9k	83 kB	10.8M
Modified TSM	**91.0%**	**0.845**	54.8k	236 kB	30.8M
Tiny CNN	87.3%	0.816	55.0k	214 kB	69.2M
Very Small CNN	89.4%	0.836	187.1k	731 kB	921.9M
Small CNN	89.7%	0.843	411.5k	1607 kB	4.7B
Medium CNN	88.3%	0.831	845.6k	3338 kB	9.7B

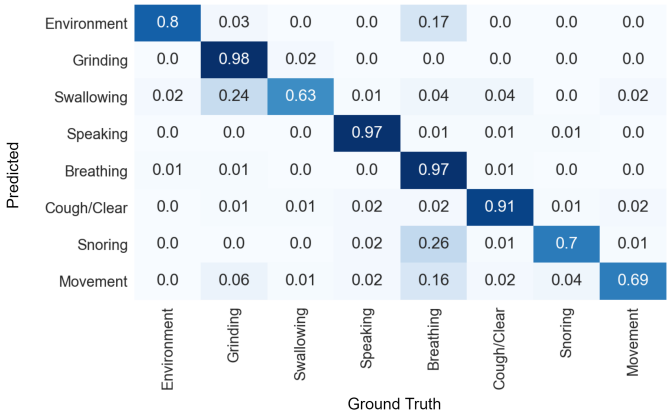

Fig. 4. A confusion matrix showing the modified TSM classifier's accuracy across classes.

We measured the classification performance of the models using both accuracy and F1 score to account for the uneven distribution of our dataset. We also measured the computational efficiency of the models according to the number of learned parameters, memory footprint, and number of floating-point operations (FLOPs) they required. The overall results of these experiments are shown in Table II.

A. Performance

All the models achieved a cross-validated accuracy between 87.3% and 91.0%. Our proposed classifier achieved the highest accuracy (91.0%) and F1 score (0.845). Comparisons with previous literature are difficult because of differences in sensors and classification targets; nevertheless, the accuracy numbers seen in those works rarely exceed 90% [24], [25]. The discrepancy between accuracy and F1 score for all the models we tested suggests that many of the classification errors were concentrated in underrepresented classes. The proposed model's confusion matrix, shown in Fig. 4, confirms this observation. It shows that classification errors were concentrated in four classes: environment (no activity), snoring, movement, and swallowing. Environmental noise was often confused with breathing. Although participants were asked to breathe particularly lightly while gathering the environment recordings, this result was not unexpected since some of them were still quiet enough to pick up on unconscious breathing sounds.

Snoring was sometimes confused with breathing, which can be attributed to the fact that subtle snoring sounds similar to regular breathing. Movement was also confused with movement in cases when people breathed between or over movement noises. Swallowing was correctly classified only 63% of the time, with those errors almost entirely being classified as teeth grinding. Upon reviewing the audio, we found that swallowing events were often preceded by a gathering of saliva in the mouth, manifesting as a series of rapid "pops" and "clicks" that resemble teeth grinding noises.

We hypothesize that the high error rate in the swallowing class can be addressed in one of two ways. First, gathering more examples of swallowing may improve our model's accuracy since it was one of the most underrepresented classes in our dataset. Second, we hypothesize that the gathering of saliva before swallowing may have been somewhat exaggerated since participants were asked to swallow on command; providing participants with liquid to moisten their mouths may make the swallowing sounds more natural.

B. Computational Efficiency

Edge devices like smartphones and earbuds are becoming increasingly capable of running deep learning models locally. The Qualcomm QCC5144[3] is a chip comparable to the ones embedded in ANC-enabled earbuds. It includes 448 kb of data memory and external flash storage (Q-SPI), which is sufficient for both our modified TSM network and spectral input data. Although smaller neural networks tend to perform worse than larger ones in classification tasks [26], being able to perform prediction on a chip like the QCC5144 can ameliorate some of the privacy concerns associated with transmitting private audio to a central service.

The difference in F1 score between our largest and smallest CNN baselines was about 0.03, yet the smallest model had 6.5% of the parameters and required less than 1% of the FLOPs needed for the largest model. Our proposed model outperformed the best performing baseline CNN while only requiring 13% of the parameters and 0.7% of the FLOPs.

IV. RELATED WORK

A. Sleep Monitoring

The literature has explored various ubiquitous sensing modalities, form factors, and target symptoms for assessing sleep quality. One of the most basic modalities that people have leveraged for sleep detection is actigraphy, which relies on body movement that is detectable by an accelerometer. For example, Natale et al. [3] demonstrated that a smartphone placed near a person's pillow was just as good at estimating total sleep duration as a wrist-worn actigraph. Breathing is another characteristic of sleep that has been monitored via ubiquitous sensing. While sleep clinics rely on uncomfortable chest-straps to directly measure the expansion and contraction of the chest, researchers have shown that it is possible to measure breathing rate non-invasively by detecting perturbations in wireless signals [5], [27]. By tracking the rate and consistency in a person's breathing, many of these works also attempt to

[3]https://www.qualcomm.com/products/qcc5144

estimate sleep stages and sleep quality. Researchers have even looked at behavioral predictors of a person's sleep quality. Min et al. [28] and Chen et al. [29] both leverage smartphone app usage data and battery consumption among other data sources (e.g., ambient light and sound levels) to predict characteristics of sleep.

Meanwhile, microphones embedded in smartphones, smartwatches, and smartspeakers have been used to capture and process sounds associated with sleep. Beyond using overall sound level as a feature for sleep detection [28], some researchers have analyzed audio to identify breathing and snoring. For example, Ren et al. [30] recorded audio from a bedside smartphone and simultaneously applied signal processing to infer a person's breathing rate and machine learning to detect notable sleep events like snoring and coughing. In contrast to such work, our system leverages ANC-enabled earbuds for audio recording, which has a couple of advantages. First, the internal microphone is able to detect sounds that are internal to the body (e.g., teeth grinding, swallowing) and would be difficult to hear from a distant microphone. Second, earbuds make it trivial to identify who is making the sound when multiple people share the same bed since the internal microphone is directed towards the source and is somewhat shielded from other sounds.

B. Audio Event Classification

Audio events classification has been used to great effect in ubiquitous computing applications. One of the most popular domains for audio recognition entails human activity recognition in the home [31], [32], relying on the fact that activities can generate unique noises depending on their typical locations and the appliances that they require. These approaches have also been generalized to urban areas, relying on sounds like car horns and construction for inference [33], [34].

Audio has also been analyzed to identify both external and internal body sounds relevant to clinical assessments, with respiratory health being one of the more prominent topics. Researchers have used portable microphones for cough detection [35]–[37] and classification [38]–[40]. For example, Laguarta et al. [38] recently demonstrated that it is possible to discriminate normal coughs from coughs produced by patients with COVID-19. It should be noted that such works have examined both naturally produced and forced coughs. Rahman et al. [41] demonstrated that a contact microphone pressed against one's throat can be used to not only capture cough sounds, but also internal body sounds like swallowing and chewing that would enable food journaling applications. As mentioned earlier, the most relevant use of audio event classification to our own work comes from efforts like those of Ren et al. [30], who processed audio to track external sleep sounds like heavy breathing and snoring. Our work expands upon this literature by examining internal body sounds associated with sleep and captured through an underexplored modality — earbuds.

V. Discussion and Conclusion

Our research demonstrates the potential utility of ANC-enabled earbuds for recording and classifying acoustic sleep events. We were able to achieve strong accuracy using a lightweight CNN with a modified TSM, which is notable since our model could be deployed on a paired mobile device or the earbuds themselves. We now briefly describe the limitations of our current work and the potential for future improvements.

First, we did not explore the full gamut of sounds that can be recorded by earbuds during sleep. We conducted our data collection efforts within participants' homes so that the resulting audio would include varied background noises, yet the possibility of concurrent sounds can expand beyond that. Future work could explore whether our classification approach is robust to confounding audio generated by the earbuds themselves, such as sounds from audiobooks or music. Sleep sounds can also overlap (e.g., coughing while moving in bed), so a new classification approach may be needed to separate superimposed audio signals.

Our work is also limited because we evaluated our system on voluntarily produced sounds. This approach has been used to validate other ubiquitous computing applications involving audio event classification [16], [17], [41], yet voluntary and reflex sounds are known to have different characteristics [42]. We plan to expand upon our preliminary findings by collecting more ecologically valid data during clinical polysomnography sessions. This would not only give us access to natural sleep sounds during sleep, but also provide access to additional sensor data streams (e.g., EEG, EKG) that would allow us to automate some components of the annotation process. We then hope to deploy our system in such studies for sleep sound classification, which would provide clinicians with an additional source of information as they diagnose patients' sleep issues. Our ability to integrate our system into clinical practice will be contingent on criteria beyond high classification accuracy. First, we will need to confirm that the voluntary sounds collected during our study are comparable to natural sounds made during sleep. Second, since we will want the earbuds to be wireless and powered for an entire night, we will need to be able to process and classify sounds in real-time to avoid costs associated with data storage or transmission. The modified TSM made it possible for us to significantly reduce the computational intensity and size of our deep learning model, but there are also latency and power costs associated with generating audio spectrograms used as input to that model; future work would need to explore how those computational costs impact real-time processing. Lastly, the earbuds' form factor should be comfortable and not impact how people sleep, especially when they visit a sleep clinic seeking a medical diagnosis.

References

[1] D. Manfredini, J. Ahlberg, T. Castroflorio, C. Poggio, L. Guarda-Nardini, and F. Lobbezoo, "Diagnostic accuracy of portable instrumental devices to measure sleep bruxism: a systematic literature review of polysomnographic studies," *Journal of oral rehabilitation*, vol. 41, no. 11, pp. 836–842, 2014.

[2] N. Douglas, S. Thomas, and M. Jan, "Clinical value of polysomnography," *The Lancet*, vol. 339, no. 8789, pp. 347–350, 1992. Originally published as Volume 1, Issue 8789.

[3] V. Natale, M. Drejak, A. Erbacci, L. Tonetti, M. Fabbri, and M. Martoni, "Monitoring sleep with a smartphone accelerometer," *Sleep and Biological Rhythms*, vol. 10, no. 4, pp. 287–292, 2012.

[4] L. Chang, J. Lu, J. Wang, X. Chen, D. Fang, Z. Tang, P. Nurmi, and Z. Wang, "Sleepguard: Capturing rich sleep information using smartwatch sensing data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–34, 2018.

[5] F. Zhang, C. Wu, B. Wang, M. Wu, D. Bugos, H. Zhang, and K. R. Liu, "Smars: Sleep monitoring via ambient radio signals," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 217–231, 2019.

[6] C. Research, "True wireless hearables sales to climb to 129 million units globally by 2020," tech. rep., 2019.

[7] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

[8] H. Middelkoop and G. Kerkhof, "Nocturnal wrist motor activity and subjective sleep quality in young and elderly subjects," 1990.

[9] M. T. Hyyppä and E. Kronholm, "Sleep movements and poor sleep in patients with non-specific somatic complaints—ii. affective disorders and sleep quality," *Journal of psychosomatic research*, vol. 31, no. 5, pp. 631–637, 1987.

[10] M. R. Lemke, P. Puhl, and A. Broderick, "Motor activity and perception of sleep in depressed patients," *Journal of psychiatric research*, vol. 33, no. 3, pp. 215–224, 1999.

[11] P. Svensson, T. Arima, G. Lavigne, and E. Castrillon, "Chapter 144 - sleep bruxism: Definition, prevalence, classification, etiology, and consequences," in *Principles and Practice of Sleep Medicine (Sixth Edition)* (M. Kryger, T. Roth, and W. C. Dement, eds.), pp. 1423–1426.e4, Elsevier, sixth edition ed., 2017.

[12] A. M. Alencar, D. G. V. da Silva, C. B. Oliveira, A. P. Vieira, H. T. Moriya, and G. Lorenzi-Filho, "Dynamics of snoring sounds and its connection with obstructive sleep apnea," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 1, pp. 271–277, 2013.

[13] J. Krönig, O. Hildebrandt, A. Weissflog, W. Cassel, V. Gross, K. Sohrabi, P. Fischer, and U. Koehler, "Long-term recording of night-time respiratory symptoms in patients with stable copd ii–iv," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 14, no. 5, pp. 498–503, 2017.

[14] W. T. McNicholas, D. Hansson, S. Schiza, and L. Grote, "Sleep in chronic respiratory disease: Copd and hypoventilation disorders," *European Respiratory Review*, vol. 28, no. 153, 2019.

[15] J. Newell, O. Mairesse, P. Verbanck, and D. Neu, "Is a one-night stay in the lab really enough to conclude? first-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples," *Psychiatry research*, vol. 200, no. 2-3, pp. 795–801, 2012.

[16] F. Barata, K. Kipfer, M. Weber, P. Tinschert, E. Fleisch, and T. Kowatsch, "Towards device-agnostic mobile cough detection with convolutional neural networks," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–11, IEEE, 2019.

[17] J. Amoh and K. Odame, "Deepcough: A deep convolutional neural network in a wearable cough detection system," in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, IEEE, 2015.

[18] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE signal processing letters*, vol. 13, no. 1, pp. 52–55, 2005.

[19] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[20] Q. Wen, L. Sun, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," *arXiv preprint arXiv:2002.12478*, 2020.

[21] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[22] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, IEEE, 2017.

[23] Y. Li, Y. Wang, X. Liu, Y. Shi, and S.-f. Shih, "Enabling real-time on-chip audio super resolution for bone conduction microphones," *arXiv preprint arXiv:2112.13156*, 2021.

[24] D. Ma, A. Ferlini, and C. Mascolo, "Oesense: employing occlusion effect for in-ear human sensing," *arXiv preprint arXiv:2106.08607*, 2021.

[25] X. Xu, H. Shi, X. Yi, W. Liu, Y. Yan, Y. Shi, A. Mariakakis, J. Mankoff, and A. K. Dey, "Earbuddy: Enabling on-face interaction via wireless earbuds," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

[26] L. J. Ba and R. Caruana, "Do deep nets really need to be deep?," *arXiv preprint arXiv:1312.6184*, 2013.

[27] T. Rahman, A. T. Adams, R. V. Ravichandran, M. Zhang, S. N. Patel, J. A. Kientz, and T. Choudhury, "Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 39–50, 2015.

[28] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 477–486, 2014.

[29] C.-Y. Chen, S. Vhaduri, and C. Poellabauer, "Estimating sleep duration from temporal factors, daily activities, and smartphone use," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 545–554, IEEE, 2020.

[30] Y. Ren, C. Wang, Y. Chen, J. Yang, and H. Li, "Noninvasive fine-grained sleep monitoring leveraging smartphones," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8248–8261, 2019.

[31] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pp. 509–514, IEEE, 2012.

[32] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 213–224, 2018.

[33] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 171–175, IEEE, 2015.

[34] Z. Huang, C. Liu, H. Fei, W. Li, J. Yu, and Y. Cao, "Urban sound classification based on 2-order dense convolutional network using dual features," *Applied Acoustics*, vol. 164, p. 107243, 2020.

[35] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 375–384, 2011.

[36] S. Birring, T. Fleming, S. Matos, A. Raj, D. Evans, and I. Pavord, "The leicester cough monitor: preliminary validation of an automated cough detection system in chronic cough," *European Respiratory Journal*, vol. 31, no. 5, pp. 1013–1018, 2008.

[37] X. Sun, Z. Lu, W. Hu, and G. Cao, "Symdetector: detecting sound-related respiratory symptoms using smartphones," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 97–108, 2015.

[38] J. Laguarta, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[39] G. Botha, G. Theron, R. Warren, M. Klopper, K. Dheda, P. Van Helden, and T. Niesler, "Detection of tuberculosis by automatic cough sound analysis," *Physiological measurement*, vol. 39, no. 4, p. 045005, 2018.

[40] V. Swarnkar, U. R. Abeyratne, A. B. Chang, Y. A. Amrulloh, A. Setyati, and R. Triasih, "Automatic identification of wet and dry cough in pediatric patients with respiratory diseases," *Annals of biomedical engineering*, vol. 41, no. 5, pp. 1016–1028, 2013.

[41] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "Bodybeat: a mobile system for sensing non-speech body sounds.," in *MobiSys*, vol. 14, pp. 2–594, Citeseer, 2014.

[42] C. Magni, E. Chellini, F. Lavorini, G. A. Fontana, and J. Widdicombe, "Voluntary and reflex cough: similarities and differences," *Pulmonary pharmacology & therapeutics*, vol. 24, no. 3, pp. 308–311, 2011.