
Proceedings Paper:
Xu, X., Brown, G.J. orcid.org/0000-0001-8565-5476 and Ma, N. (Accepted: 2025) Sound-based sleep staging using pretrained speech foundation models. In: Proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2025), 14-17 Jul 2025, Copenhagen, Denmark. Institute of Electrical and Electronics Engineers (IEEE) (In Press)

Sound-Based Sleep Staging using Pretrained Speech Foundation Models

Xiaolei Xu, Guy J. Brown and Ning Ma[1]

Abstract— **Sleep staging is essential for diagnosing sleep disorders and understanding sleep physiology, yet traditional polysomnography (PSG) is costly, intrusive, and impractical for large-scale or home-based monitoring. Wearable devices provide alternatives but face limitations such as motion artifacts and inconsistent skin contact. In this study, we propose a non-contact sleep staging approach using sound-based analysis. To address data scarcity, we employ transfer learning with pretrained speech foundation models, originally developed for automatic speech recognition but capable of capturing rich acoustic representations beyond linguistic content. By analysing temporal attention weights in the extracted HuBERT embeddings, we demonstrate that these models effectively capture respiratory patterns. Our findings suggest that repurposing speech foundation models for sleep staging provides a scalable, non-contact alternative to PSG, with promising applications in home-based and clinical settings.**

Clinical Relevance—**This demonstrates the feasibility of a non-contact, sound-based approach to sleep staging, potentially improving access and adherence in clinical and home settings.**

I. Introduction

Human sleep is a cyclical process governed by distinct neurophysiological states, broadly classified into wakefulness, non-rapid eye movement (NREM) sleep (N1, N2, and N3), and rapid eye movement (REM) sleep. These sleep stages, as defined by the American Academy of Sleep Medicine (AASM), are conventionally identified through patterns in electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG), and reflect critical variations in brain activity, muscle tone, and autonomic function. Accurate sleep staging is essential for diagnosing sleep disorders such as sleep apnoea, insomnia, and narcolepsy, as well as advancing our understanding of sleep physiology and optimising therapeutic interventions [1].

Polysomnography (PSG) remains the gold standard for sleep staging. Typically this is conducted in a hospital sleep laboratory, and sleep stages are manually scored based on inspection of EEG, EMG and EOG signals. While deep learning models have shown promising results in automating this process with accuracy comparable to human scorers [2], [3], [4], its reliance on multi-channel electrophysiological recordings, clinical expertise, and laboratory-based assessments makes it expensive, time consuming, and impractical for large-scale or home-based monitoring.

To address these limitations, wearable devices such as actigraphy-based wristbands, EEG headbands, and photoplethysmography (PPG)-enabled smartwatches have been explored as alternatives for sleep staging. Actigraphy, which infers sleep-wake patterns from motion data, has been widely adopted due to its ease of use, but it lacks the granularity needed for detailed sleep staging [5], [6]. Wearables equipped with PPG sensors estimate heart rate variability, which has been correlated with sleep stages [7]; however, PPG-based methods are susceptible to motion artifacts and skin-contact inconsistencies. EEG-based wearables offer greater accuracy by directly capturing brain activity, yet they still require adherence to the scalp, which can be intrusive and uncomfortable for long-term monitoring. Moreover, the proprietary algorithms used in many commercial devices limit transparency and validation against gold-standard PSG.

Given these constraints, there is growing interest in passive, non-contact sleep monitoring techniques. Sound-based sleep staging is a promising alternative due to its accessibility; smartphones are ubiquitous and are commonly placed near the bedside while charging overnight, enabling an opportunity for long-term sleep monitoring. Additionally, smartphones can record audio at high sampling rates, capturing rich physiological information through respiration patterns, snoring spectral profiles, vocalisations, and movement-induced sounds, all of which exhibit stage-dependent variations [8], [9], [10]. For example, N3 sleep is characterised by deep, regular breathing, while REM sleep often features irregular respiration and occasional vocalisations [11]. Recent studies have demonstrated the feasibility of using deep learning models to analyse sleep sounds, yet most approaches rely on handcrafted features or task-specific architectures, which may not generalise well across diverse populations and recording environments [8]. Training deep learning models from scratch is also hindered by the scarcity of large-scale labelled datasets, as synchronising high-fidelity audio with PSG ground truth is labour-intensive [10].

This study proposes a novel approach that leverages speech foundation models for sound-based sleep staging. Speech foundation models have been shown to be effective in speech-related tasks with limited data, including speech recognition in low-resource languages [12], speech emotion recognition [13], and speech depression detection [14]. Trained on large-scale speech corpora, these models capture rich hierarchical representations of acoustic signals beyond linguistic content [15], [16]. Given the limited availability of labelled sleep audio data and the inherent class imbalance in sleep stages, we hypothesise that latent features from speech foundation models can effectively capture subtle biomarkers (e.g., breathing rate, snore characteristics) that discriminate sleep stages. Additionally, we introduce an effective pipeline to address class imbalance that arises from the predominance

All authors are with School of Computer Science, University of Sheffield, Sheffield S1 4DP, UK, {xxu97, g.j.brown, n.ma}@sheffield.ac.uk

[1]Ning Ma is also with Insigneo Institute, University of Sheffield, UK

of NREM in sleep staging. Our proof-of-concept study suggests that repurposing speech foundation models for sleep staging can offer a scalable, non-contact alternative to PSG, with potential applications in home-based monitoring.

II. Data

This study used the PSG-Audio dataset [17], an open-access database containing PSG recordings with synchronised audio from 212 participants (56 female, 156 male) aged 34 to 76 years. Sleep stages were manually annotated into wake, REM, NREM1, NREM2, and NREM3 using non-overlapping 30-sec EEG epochs [17].

The recordings were collected in a hospital sleep study unit, where 88.7% of participants were diagnosed with severe obstructive sleep apnoea, defined by an apnoea-hypopnoea index (AHI) of ≥ 30. As a result, the sleep stage distribution in this dataset differs from that of a healthy population. In healthy adults, REM sleep typically accounts for 20–25% of total sleep time, while NREM3 (deep sleep) comprises 15–25%. However, in individuals with severe OSA, REM and NREM3 stages are often significantly reduced due to frequent arousals and sleep fragmentation. This pattern is evident in the dataset, as shown in Table I, where the prevalence of severe OSA has contributed to a notable class imbalance. Although such imbalance poses challenges for the development of robust deep learning models, this dataset remains the largest public resource that provides synchronised audio recordings and expert-labelled sleep stage annotations.

TABLE I: Data distribution across classes in PSG-Audio.

Class	Wake	REM	NREM1	NREM2	NREM3
Samples	8,774	3,123	9,714	76,882	4,540
Proportion	8.5%	3.0%	9.4%	74.7%	4.4%

Two synchronised audio recordings were collected [17]: one from an ambient microphone positioned one metre above the participant's head and another from a tracheal microphone attached to the participant's neck. Both recordings were sampled at 48 kHz with 24-bit depth in uncompressed WAV format. The "snore" channel from PSG (a contact microphone sampled at 500 Hz) was used by [17] to synchronise audio recordings with PSG signals. The time difference between the audio recordings and the PSG system was estimated by maximising the cross-correlation between the tracheal microphone and the snore channel.

We identified 21 participants in the dataset with missing audio or annotation errors. After excluding these recordings, we used the "clean" subset of the PSG-Audio dataset, consisting of 191 participants, as the final study cohort.

III. System

This study aims to estimate sleep stages from ambient sound recordings by leveraging the latent patterns encoded in acoustic features using speech foundation models. Additionally, we address the data imbalance issue (see Section II) to improve the robustness of the proposed model.

A. Data Preprocessing

All audio recordings were downsampled to a rate of 16 kHz to reduce computational load. We first checked for empty audio segments across the dataset. Any 30-second epoch with a root mean square (RMS) amplitude of zero for more than 10 seconds was considered invalid. This procedure identified and removed approximately 19.5 hours of invalid data in the dataset (about 1%). These empty segments predominantly occurred in the later stages of the night, though some recordings exhibited substantial gaps of up to one hour in the middle of the night. Subsequently, we applied a biquad high-pass filter with a 50 Hz cutoff frequency to remove stationary noise commonly present in hospital environments. Finally, a pre-emphasis filter was used to enhance high-frequency breathing and snore sounds while attenuating low-frequency noise.

B. Addressing Class Imbalance

The PSG-Audio dataset exhibits a highly imbalanced distribution of sleep stages, with NREM stages comprising 88.5% of the samples, while REM and wake stages account for only 3% and 8.5%, respectively (see Table I). To reduce model overfitting to the majority class, we employed an upsampling strategy for minority classes by generating overlapping segments. Specifically, if two consecutive non-overlapping epochs shared the same sleep stage, we created an additional sample by shifting the window by 15 seconds, resulting in a 50% overlap. This approach not only preserved transient sound events that occur at epoch boundaries, such as body movements and coughs, but also increased the number of minority samples with minimal redundancy. Using this method approximately doubled the size of the wake and REM classes.

C. Sleep Staging using Speech Foundation Models

HuBERT is a self-supervised speech foundation model trained on the 960-hour LibriSpeech dataset [18]. Recent studies have shown promising results of using HuBERT for non-speech audio tasks such as acoustic event detection and music analysis [19]. Further, [20] showed that HuBERT could reconstruct respiratory effort signals from speech audio, suggesting that its representations inherently capture breathing-related features.

We employ HuBERT-Base, the smallest model in the HuBERT family, which consists of convolutional neural network (CNN) encoders followed by transformer layers. The CNN encoders function as filters and extract local features from the raw waveform, while the transformer layers capture temporal dependencies. The model outputs a sequence of 768-dimensional feature vectors for each time frame.

As shown in Fig. 1, we froze the pretrained HuBERT model to extract feature vectors over time (HuBERT embeddings). These embedding vectors were then processed through a self-attention layer to determine element-wise importance weights. A softmax function normalised the weights to obtain the importance score for each frame. The final 768-dimensional representation was obtained via

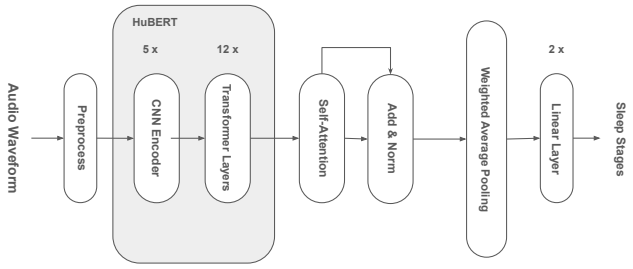


Fig. 1: Model architecture

a weighted average of the HuBERT embeddings. To classify sleep stages, we employed two fully-connected layers: the first reduced the dimensionality by half, while the second projected the output into a three-class space corresponding to three sleep stages: wake, NREM, and REM.

D. Training Setup

We explored two model configurations: one with 30-sec inputs and another with 90-sec inputs. The 30-sec model predicts sleep stages from a single 30-sec segment, whereas the 90-sec system incorporates the adjacent segments to predict the middle segment. Before training, local mean-variance normalisation was applied to enhance the contrast of breathing and snoring events.

The model consisted of 94.7 million frozen parameters and 2.1 million trainable parameters. Training used a weighted multi-class cross-entropy loss, where class weights were set as the inverse of class frequencies to address data imbalance. An initial learning rate of 5×10^{-5} was used, which decayed linearly after 6,000 steps to a minimum of 2.5×10^{-6} over five training epochs.

E. Visualisation of Attention Weights

To examine the effect of the extracted HuBERT embeddings, we visualised the attention weights from the self-attention layer to identify the regions the model focused on within an audio segment. Fig. 2 shows the attention weights for a 30-sec NREM3 sample. In the top Mel-spectrogram, brighter areas indicate snores, while the bottom panel highlights regions where the model assigns higher attention, with higher attention weights. Overall, the attention weights tend to align with the inspiration and expiration phases of breathing in the audio segment, albeit with a slight delay. Notably, the model assigns greater weights to expiration sounds, even when they are very quiet. These findings support our hypothesis that speech foundation models can capture subtle breathing patterns from audio.

IV. Evaluation

Experiments were conducted on the "clean" subset of the PSG-Audio dataset (see Section II) using ambient audio recordings from 191 participants. We allocated 80% of the dataset (153 participants) for training, while the remaining 20% (38 participants) were reserved for evaluation.

To evaluate the benefits of leveraging speech foundation models and incorporating adjacent context, we compared two proposed models: (1) a 30-sec HuBERT-based model using a single 30-sec audio segment and (2) a 90-sec HuBERT-based model aggregating features from 90-sec windows to incorporate adjacent context.

Additionally, we included SoundSleepNet [10], a state-of-the-art sound-based sleep staging model, as a baseline. SoundSleepNet employs a CNN-Transformer architecture with Mel-spectrograms as input. Its feature extractor consisted of CNN encoders, two BiLSTM layers, and two Transformer layers. Since its source code was unavailable, we implemented the model ourselves. The reimplemented model contained 29.7 million trainable parameters and was trained with a learning rate of 5×10^{-4}, using the same scheduler and loss function as our proposed systems.

All models were trained to classify three sleep stages: wake, REM and and NREM. Model performance was evaluated using confusion matrices, along with class-wise sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV).

V. Results and Discussion

Fig. 3 presents the confusion matrices for each model in differentiating the three sleep stages. Class-wise performance of the three models is listed in Table II. For the "Wake" stage, all the three models shows strong sensitivity and specificity, indicating they correctly identify a high proportion of true wake stages and avoid misclassifying non-wake stages. The proposed 90-sec HuBERT improves slightly over the 30-sec HuBERT in sensitivity, specificity, and PPV, suggesting that longer audio segments may help reduce false positives while maintaining high true positive rates. All models struggle with precision (low PPV), likely due to class imbalance or overlapping acoustic features with other stages. Additionally, we observed that unfreezing the HuBERT model leads to overfitting, resulting in poor generalisation performance on the test set. Therefore, we only used the HuBERT model as the feature extractor.

For the "REM" stage, the SoundSleepNet model has very low sensitivity (0.21) and PPV (0.07), indicating it misses most REM stages and has a high false positive rate. The proposed 30-sec HuBERT improves sensitivity but still has low PPV. The 90-sec HuBERT, with a longer audio context, shows the best performance for REM stages, with higher sensitivity and specificity, though PPV remains low.

The "NREM" stage is generally well-detected by all models, with high precision (PPV $>$ 0.93). SoundSleepNet has higher sensitivity, while the two HuBERT models trade sensitivity for specificity. However, the low NPV ($<$0.25) across all models indicates that non-NREM predictions are less reliable. This issue, along with very low PPVs for wake and REM, is likely due to the class imbalance in the dataset – NREM comprises 88% of the samples – which stems from the dataset's bias towards severe OSA patients. Despite efforts to address the class imbalance issue, the model still tend to favour NREM predictions, reducing their ability to distinguish minority classes.

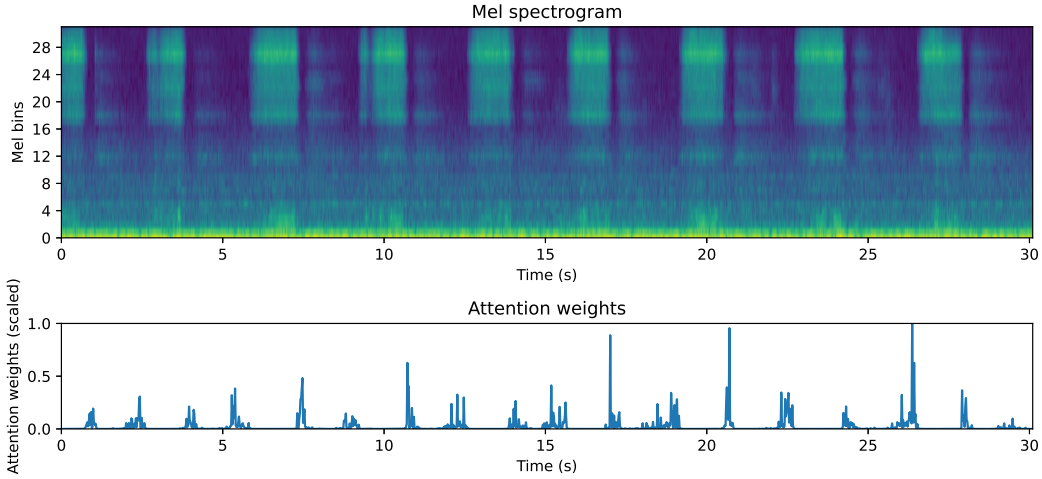
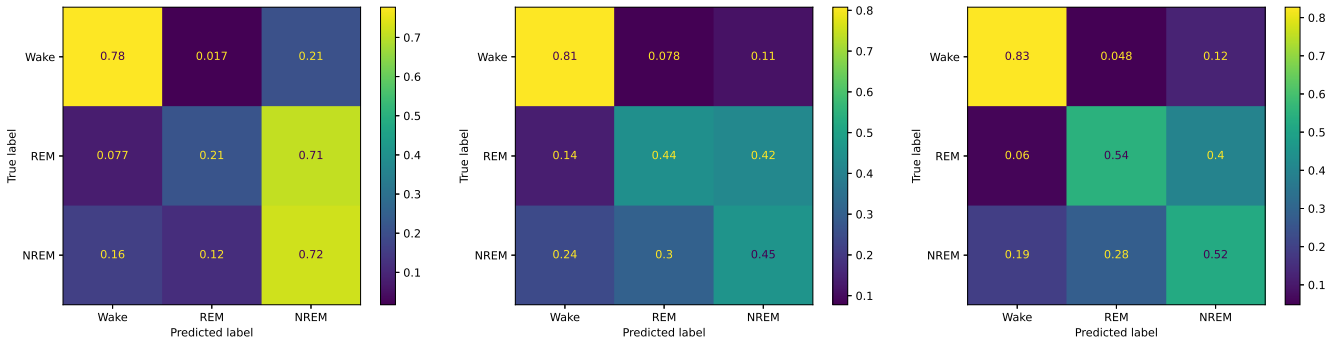


Fig. 2: Visualisation of time-wise importance generated from the self-attention layer after extraction of HuBERT embeddings. The top line shows the Mel-spectrogram of a sleep audio recording during NREM-3, and the bottom line shows the corresponding time-wise attention weights (scaled to [0,1]).



(a) SoundSleepNet (b) 30-sec HuBERT (c) 90-sec HuBERT

Fig. 3: Normalised confusion matrices for the three models evaluated in this study.

The comparison between 30-sec and 90-sec HuBERT shows that a longer context window can improve performance (e.g., REM sensitivity increases from 0.44 to 0.54, and wake sensitivity and specificity also see modest improvements). This suggests that incorporating longer temporal context helps capture more relevant acoustic cues.

The use of an ambient microphone instead of a tracheal microphone for audio recording introduces several considerations that may affect model performance. Ambient microphones capture a broader range of sounds, including environmental noise (e.g., room ventilation, medical equipment noise, or external disturbances), which may obscure subtle respiratory patterns critical for distinguishing sleep stages. The high sensitivity but low PPV for Wake may be partly attributed to ambient noise. Sounds such as rustling sheets or environmental disturbances can be misclassified as Wake, leading to false positives. This effect is particularly pronounced in datasets with severe OSA patients, where frequent movements and OSA events exacerbate noise levels.

TABLE II: Class-wise performance of the three models evaluated in this study.

Model	Class	SE	SP	PPV	NPV
SoundSleepNet	**Wake**	0.78	0.84	0.32	0.98
	REM	0.21	0.89	0.07	0.97
	NREM	0.72	0.64	0.93	0.25
30-sec HuBERT	**Wake**	0.81	0.76	0.24	0.98
	REM	0.44	0.72	0.06	0.97
	NREM	0.45	0.79	0.94	0.17
90-sec HuBERT	**Wake**	0.83	0.81	0.29	0.98
	REM	0.54	0.74	0.08	0.98
	NREM	0.52	0.79	0.95	0.19

VI. Conclusions

This study has demonstrated the potential of pretrained speech foundation models for non-contact sound-based sleep staging. By analysing temporal attention weights in the extracted HuBERT embeddings, we have shown that these

models effectively capture respiratory patterns. The 90-sec HuBERT model performed better than others in detecting wake and REM stages, though it still struggled with low positive predictive value for REM. The use of ambient microphones in this study introduces challenges related to background noise and the capture of quiet breathing sounds, particularly affecting REM detection and overall staging precision. Moreover, using a dataset with cases of severe sleep apnea presents challenges due to atypical respiratory patterns and a more pronounced class imbalance in sleep stages.

Future work could explore hybrid approaches combining ambient audio with other modalities or investigate the use of tracheal microphones in controlled settings to establish performance benchmarks. Furthermore, these pretrained models could be fine-tuned on explainable tasks (e.g., snore detection, snore rate, and AHI estimation) to encourage the model to prioritise features linked to sleep stages, rather than relying solely on models trained on generic speech or audio event data. These efforts will be essential for advancing sound-based sleep staging as a scalable solution for home-based monitoring and clinical applications.

ACKNOWLEDGMENT

The authors would like to thank Passion For Life Healthcare (UK) Limited for providing access to anonymised data collected via their SoundSleep app, which will support future developments of this research. Ning Ma was partly funded by MRC IAA grant 182731 and GOSH BRC grant 187217.

REFERENCES

[1] M. A. Carskadon, W. C. Dement, *et al.*, "Normal human sleep: An overview," *Principles and Practice of Sleep Medicine*, vol. 4, no. 1, pp. 13–23, 2005.

[2] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[3] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[4] A. N. Olesen, P. Jørgen Jennum, E. Mignot, and H. B. D. Sorensen, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," *Sleep*, vol. 44, no. 1, p. zsaa161, 2021.

[5] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ecg and respiratory effort," *Physiological measurement*, vol. 36, no. 10, p. 2027, 2015.

[6] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, "Sleep stage classification from heart-rate variability using long short-term memory neural networks," *Scientific reports*, vol. 9, no. 1, p. 14149, 2019.

[7] K. Kotzen, P. H. Charlton, S. Salabi, L. Amar, A. Landesberg, and J. A. Behar, "SleepPPG-Net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 924–932, 2022.

[8] E. Dafna, A. Tarasiuk, and Y. Zigel, "Sleep staging using nocturnal sound analysis," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[9] B. Xue, B. Deng, H. Hong, Z. Wang, X. Zhu, and D. D. Feng, "Non-contact sleep stage detection using canonical correlation analysis of respiratory sound," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 614–625, 2019.

[10] J. Hong, H. H. Tran, J. Jung, H. Jang, D. Lee, I.-Y. Yoon, J. K. Hong, and J.-W. Kim, "End-to-end sleep staging using nocturnal sounds from microphone chips for mobile devices," *Nature and Science of Sleep*, pp. 1187–1201, 2022.

[11] R. B. Berry, R. Brooks, C. Gamaldo, S. M. Harding, R. M. Lloyd, S. F. Quan, M. T. Troester, and B. V. Vaughn, "AASM scoring manual updates for 2017 (version 2.4)," pp. 665–666, 2017.

[12] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[13] Y. Wang, A. Boumadane, and A. Heba. A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding.

[14] X. Zhang, X. Zhang, W. Chen, C. Li, and C. Yu, "Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments," *Scientific Reports*, vol. 14, no. 1, p. 9543, 2024.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

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

[17] G. Korompili, A. Amfilochiou, L. Kokkalas, S. A. Mitilineos, N.-A. Tatlas, M. Kouvaras, E. Kastanakis, C. Maniou, and S. M. Potirakis, "PSG-Audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies," *Scientific data*, vol. 8, no. 1, p. 197, 2021.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[19] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, L. Cagliero, P. Garza, and S. M. Siniscalchi, "Benchmarking representations for speech, music, and acoustic events," *arXiv preprint arXiv:2405.00934*, 2024.

[20] V. Mitra, A. Chatterjee, K. Zhai, H. Weng, A. Hill, N. Hay, C. Webb, J. Cheng, and E. Azemi, "Pre-trained foundation model representations to uncover breathing patterns in speech," *arXiv preprint arXiv:2407.13035*, 2024.