Obstructive Sleep Apnea Detection using Pre-trained Speech Representations

Kaibo Zhang[1], *Lili Cao*[2], *Yiming Ding*[2], *Yanru Li*[2], *Chao Zhang*[1], *Ji Wu*[1], *Demin Han*[2]

[1]Dept. of Electronic Engineering, Tsinghua University
[2]Dept. of Otolaryngology, Beijing Tongren Hospital, Capital Medical University

zkb21@mails.tsinghua.edu.cn, Lily_Cao94@126.com, dyment@126.com, liyanruru@aliyun.com, cz277@tsinghua.edu.cn, wuji_ee@tsinghua.edu.cn, deminhan_ent@hotmail.com

Abstract

Obstructive sleep apnea (OSA) is a condition commonly affecting middle-aged men that can disturb sleep, cause daytime tiredness, and increase the risk of heart disease. Speech can serve as a valuable biomarker for identifying and predicting the severity of OSA due to its connection with changes in throat structure. This study proposes a new deep-learning-based method for detecting OSA by analyzing speech recordings of participants in sitting and lying positions. The method utilizes a Siamese structure that employs a pre-trained XLSR model to encode ten utterances for each position, reducing the amount of necessary training data and enabling comparison of throat structure changes between the two positions through voice analysis. The study also explores the use of patient characteristic features. Results show this approach achieves an F1 value of 0.725 on our in-house dataset, proving the feasibility of end-to-end speech OSA detection with foundation models.

Index Terms: Obstructive sleep apnea, Speech foundation models, Speech signal analysis, Siamese network

1. Introduction

Obstructive sleep apnea (OSA) is a common sleep disorder where the upper airway repeatedly collapses during sleep, leading to hypopnea or apnea along with hypoxia, hypercapnia, and cortical arousal. Studies estimate that approximately 17.4% of women and 33.9% of men aged 30 to 70 in the US have OSA [1]. Severe OSA can cause disruptions in sleep structure and result in daytime sleepiness, impaired work performance, and traffic accidents [2]. Additionally, OSA has been linked to other diseases such as hypertension, coronary artery diseases [3], and diabetes [4]. Currently, laboratory-based polysomnography (PSG) is the standard diagnostic test for OSA, using the apnea-hypopnea index (AHI) to determine the severity of the condition. An AHI of 5 or more events per hour indicates OSA, while an AHI of 30 or more events per hour indicates severe OSA [5]. However, PSG is a time-consuming and labor-intensive process, as it involves placing multiple measures on the patient, and it can be challenging to schedule appointments with doctors. Furthermore, the laboratory setting and equipment can interfere with the patient's normal sleep, potentially leading to inaccurate diagnoses. PSG is also inconvenient and expensive for patients, leading to an estimated 24 million undiagnosed OSA cases in the US [6].

The speech signal has been proposed as a viable method for early detection and severity evaluation of OSA [7]. Firstly, speech acquisition in humans is linked to the appearance of OSA from an anatomical perspective [8]. Secondly, the structure and function of the upper airway are key factors in OSA pathophysiology. The intermittent hypoxemia, long-term snoring, and upper airway collapse associated with OSA can affect the structure and function of the upper airway [9], leading to abnormal speech in OSA patients compared to healthy individuals. This can cause articulation, phonation, and resonance anomalies. Theoretically, speech analysis is a faster, less expensive, and more reliable means of predicting OSA.

Several research studies have indicated speech signals can be an efficient and speedy method for diagnosing OSA. However, there were several limitations observed in these studies. For example, [10] employed Chinese vowels and nasal sounds as audio data, and they used a supported vector machine for OSA detection. Similarly, [11] used Chinese syllables and decision trees for classification, but the process also required extracting various types of hand-crafted acoustic features (*e.g.* filter banks) before inputting them into the classifier.

This paper proposes utilizing a pre-trained XLSR model for detecting OSA in spoken sentences [12]. To capture the difference in voices caused by a throat structure change between sitting and lying positions, ten utterances are recorded for each patient in both positions with identical scripts. A Siamese structure [13] with two identical encoders is used to encode the speech of each position, revealing the difference. This approach has several advantages. Firstly, it utilizes complete sentences instead of individual phonemes and syllables, which is a more natural and informative way of taking inputs. Secondly, our approach is end-to-end, reducing error propagation between different components in a modularized system and making it more practical for deployment. Our model also uses raw waveform as input features, allowing the model to learn features more suitable for OSA detection than hand-crafted acoustic features. Thirdly, pre-trained XLSR models are employed to leverage the knowledge embedded in their representations obtained from thousands of hours of data due to the limited training samples available for OSA detection. Lastly, due to the unconstrained feature setting of deep learning models, patient characteristic features, such as age and body mass index (BMI), can easily be incorporated into the model to leverage complementary information and enhance the model's performance. Experimental results on our in-house dataset validate the usefulness of the pre-trained encoder, Siamese structure, and patient characteristic features.

The rest of the paper is organised as follows. Sec. 2 reviews foundation models and their medical uses. Sec. 3 presents our proposed method. Secs. 4 and 5 are the experimental setup and results. We conclude in Sec. 6.

2. Foundation models in Medical Practice

In medical diagnosis, pathological characteristics of patients are used to determine if they have a disease. Similarly, super-

10.21437/Interspeech.2023-278

vised learning models learn patterns between patient features and disease labels to acquire medical knowledge. However, machine learning models, especially deep learning models, require a large number of feature-label data pairs, which is challenging to obtain in the medical field. In the case of OSA detection, polysomnography is needed to obtain the precise AHI value of a patient, and it takes an entire night to complete. Due to the high cost and time consumption, acquiring sufficient paired data for model training is unfeasible.

Self-supervised learning has been a topic of interest in recent years, with significant advances in various fields, including speech and language processing. Self-supervised learning models in speech processing have the advantage of utilizing large amounts of unlabelled speech data, which is abundant and easily accessible. One of the most promising areas of research in self-supervised learning is the development of speech foundation models. These models are pre-trained on large amounts of unlabelled data and are then fine-tuned on specific speech tasks. Speech foundation models can be used for various downstream tasks, including speech recognition, speaker recognition, and speech synthesis.

XLSR, introduced in 2020, is among the most widely used speech foundation models [12, 14, 15, 16], and released a multilingual version pre-trained using 53 languages including Mandarin. The model employs a contrastive learning approach, where it is trained to distinguish between similar and dissimilar speech data pairs. This approach has been highly effective, and XLSR has achieved state-of-the-art results on numerous speech recognition benchmarks [17]. Speech foundation models have considerable value for various medical tasks related to speech, including disease prediction and diagnosis. In the case of OSA prediction, obtaining audio-AHI data pairs can be challenging. Pre-trained speech models have already learned general speech representations, and after fine-tuning with a small amount of data, they can perform well.

3. Method

To predict the severity of a patient's OSA symptoms, the same script is read by a patient in both sitting and lying positions. This is to leverage the clinical finding that the OSA patient's vocal tract shape changes more in different speaking positions. Such a change influence the spectrograms of the whole utterances as shown in Figure 1, and therefore requires an utterance-level model, such as the XLSR. The OSA detection task is designed to classify a subject into two categories based on a pre-specified AHI threshold.

In order to handle the data sparsity issue caused by the small amount of supervised OSA data when training a large deep-learning model, a pre-trained XLSR model is used to leverage a large amount of unsupervised data. The supervised OSA data is then used in the fine-tuning stage. The raw audio waveform is fed into the feature extractor as the input, which comprises seven convolutional neural network (CNN) layers within the pre-trained XLSR model. This enables the model to extract features that are most suitable for OSA detection through joint training. The features are then fed into the Transformer blocks [18]. More details about the model specification can be found in Section 4.2. A statistical pooling layer is used to aggregate the sequence of XLSR-53 representations into a single vector. A classification block is used following the XLSR model to classify the subject into one of the two classes. The cross-entropy function is used as the training loss between the prediction and the actual one-hot label.

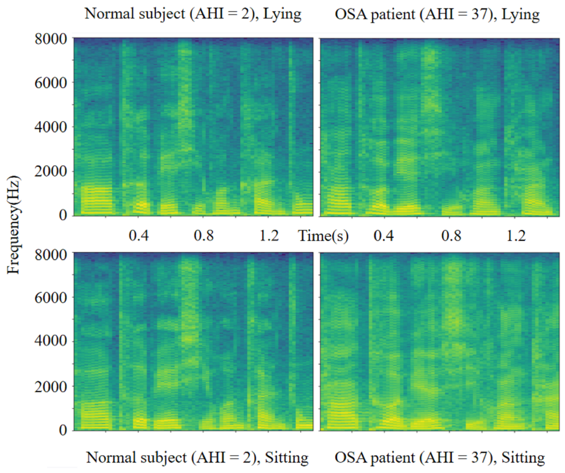

Figure 1: *Spectrograms of lying (upper) and sitting (lower) audios spoken by a normal subject with AHI=2 (left) and an OSA patient with AHI=37 (right). The bottom-right picture has more energy in the higher frequency bands (sitting is yellower on top than lying) since the patient suffers from a more obvious vocal tract change when changing the speaking position.*

3.1. Single network

In the simplest scenario, audios recorded in different positions are not differentiated. The audio files associated with the same subject are first concatenated, and a 6-second long audio segment is randomly selected from a concatenation of all utterances recorded in each position during training. This model is able to leverage any information within each 6-second segment but difficult to leverage the difference between sitting and lying speech segments due to the constrained input audio length. Increasing the input audio length does not result in better performance due to the lack of training data.

3.2. Siamese network

A major issue of the single network method is the neglect of the recordings taken in different positions. A medical study [19] points out that OSA exhibits positional dependence, and changes in body position have important implications for the pathogenesis of OSA. As shown in Figure 1, compared to the normal subject, the OSA patient has a more obvious energy difference in the higher-frequency bands between the sitting and lying audio spectrograms. In order to highlight such a difference, the Siamese network differentiates between the audio recordings taken in the two positions.

As shown in Figure 2, our Siamese network structure has two identical XLSR encoders. For the audio pieces recorded while the patient is sitting and lying, one side of the Siamese network receives the sitting audio while the other receives the lying audio. A 6-second audio segment is randomly selected from a concatenation of all utterances recorded in each position, and the random cutting positions for both sitting and lying audios are identical to facilitate the differentiation of the two audio segments. Afterwards, the two resulting representations are concatenated. A classification block takes the concatenated vector as input and transforms it into two fully connected layers with ReLU activation functions. The final output vector is a 2-dimensional distribution normalised using a softmax function.

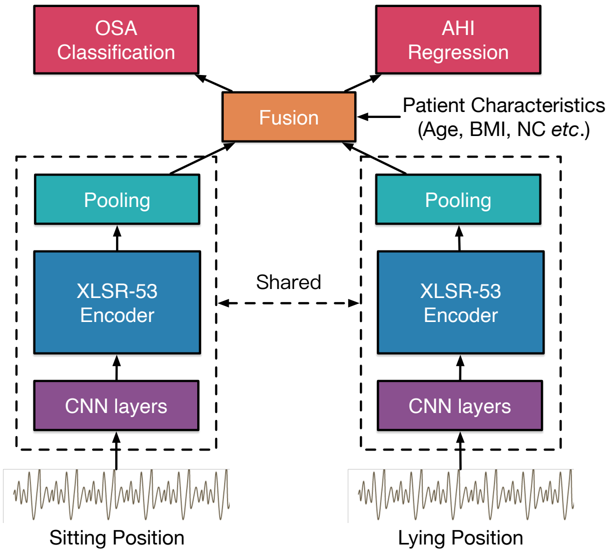

Figure 2: *The Siamese network structure, where "Pooling" refers to the temporal statistics pooling that converts a sequence into its mean and standard deviation; "Fusion" refers to concatenation and a fully-connected layer.*

3.3. Additional Patient Character Features

In addition to the speech recordings, the characteristics of the patients are also collected, including age, BMI, and neck circumference (NC). Each type of such feature is normalized to have a zero mean and unit variance across all patients. During training, a linear layer is utilized to transform each 1-dimensional feature into a 16-dimensional one. As shown in Figure 2, such additional features can be concatenated with the audio features in the fusion block.

In order to avoid overfitting into these features, a feature-level dropout with a rate of 0.5 is introduced, meaning that for each model update, there is a 50% chance that the expanded patient character feature is set as a 0 vector. Subsequently, a classification block takes the concatenated feature vector as input and returns a 2-dimensional vector.

3.4. Additional regression task

In previous sections, OSA detection is considered only as a binary classification task. This category is determined based on the severity of the symptom of a patient. However, this ignores the difference in AHI values of subjects in the same category, which can be larger than those of some subjects belonging to different categories. Therefore, in order to leverage the information of the absolute AHI values, an additional regression task can be included in the training, which helps the model to learn the difference between patients belonging to the same category but with different OSA severities. To prevent the regression loss from becoming too large, the original AHI values ranging from 0 to 103 are linearly scaled into a range of [0, 1]. Apart from the classification block, we build another regression block. Their structures are almost the same, except the regression block has output dimension 1 with sigmoid function for normalization at the end. The total loss is the sum of cross entropy loss for classification and mean-square error (MSE) loss for regression.

3.5. Test method

During the test, a sliding-window-based method is used, where the patient's audio is divided into multiple segments using a time window of 6 seconds and a moving length of 1 second. For instance, the first segment is from 0 to 6 seconds, the second is from 1 to 7 seconds, and so on. Each segment is input into the model, which outputs the model's judgement of the patient having OSA or not. If the portion of positive judgement exceeds a certain threshold (set at 0.6 in this study), the patient is predicted to have OSA.

4. Experimental Setup

4.1. Dataset

Obtaining PSG (polysomnography) data for Chinese citizens is a challenging task due to its scarcity and sensitivity, making the availability of a public dataset on this topic unfeasible. Thus, in this study, we collaborated with Beijing Tongren Hospital to collect data from 254 patients who were hospitalized due to sleep snoring. All patients were Chinese males with an average age of 39.6 years. For OSA (obstructive sleep apnea) diagnosis, the AHI (apnea-hypopnea index) served as the gold standard, with a threshold AHI = 30 events/h used to separate participants into groups. The objective of this study was to classify patients based on their characteristics and predict whether their AHI values were above or below 30. It is worth mentioning that this study was conducted following the approval of the ethics committee of Beijing Tongren Hospital.

To collect speech data, all patients were asked to read a piece of Chinese text consisting of 10 sentences with decreasing numbers of Chinese characters, while lying down and sitting up. The speech signals were recorded at a sample rate of 16kHz. Additionally, other patient characteristics, such as age, BMI, and NC (neck circumference), were also collected.

The patient cohort was divided into training, validation, and test sets, with corresponding sizes of 149, 30, and 75 patients, respectively. To increase the size of the dataset, the training data was augmented 16-fold and subjected to random cut method during training. The model was trained for 50 epochs, and the hyperparameters were optimized based on performance on the validation set. The prediction results shown in this paper were evaluated on the test set. All experiments were performed using NVIDIA GTX 3090Ti GPUs.

4.2. Model details

For the task of OSA detection, we employ the use of the pre-trained XLSR model [14], which is a multilingual version of the Wav2vec 2.0 model [12]. The XLSR model was trained on the November 2019 release of CommonVoice [20], which is a large multilingual corpus of read speech, covering more than 2000 hours of audio in 38 languages, including Chinese. It should be noted that this base model is only pre-trained on unlabelled speech data, and thus, audio-AHI paired data will be used for fine-tuning.

The audio model we adopt is composed of four main parts: feature extraction, encoding, pooling and classification. The feature extraction part splits the audio waveform into smaller segments and employs convolutional neural networks (CNN) to predict future frames for each segment. The encoder component employs a transformer-based architecture. Our model follows the structure of the large version of XLSR [14], which comprises 24 transformer blocks. For pooling, both mean pooling and standard deviation pooling are utilized to condense features from various audio segments into a single feature representation. The classification block comprises of two linear layers with a ReLU activation function in the middle.

5. Results

5.1. Pre-trained model vs randomly initialised model

We investigate the impact of inheriting pre-trained model parameters on the performance of the model in comparison to randomly initialized model parameters, when training both the encoder part and the classification block. Our experimental results, as summarized in Table 1, demonstrate that the F1 score is 0.700 for the model with inherited parameters, while it is 0.667 for the model with randomly initialized parameters. These findings suggest that the pre-trained model parameters provide significant advantages in terms of OSA detection, indicating that the general speech knowledge learned by the pre-trained model can greatly assist in this task.

5.2. Training the encoder or not

We aim to examine the influence of fine-tuning pre-trained model parameters on OSA prediction, while considering different training scenarios involving solely training the classification block and training both the encoder part and the classification block. Specifically, in the former scenario, we choose the block from the encoder part with the best validation F1 score, which in our case is the 14th block, as the feature output block. Our results, as presented in Table 1, reveal that the F1 score is 0.700 for the encoder being trained, whereas it is 0.647 for the encoder not being trained. These findings indicate that fine-tuning the pre-trained model can yield significant benefits in terms of OSA prediction.

5.3. Using the single or Siamese network structure

We analyze the effect of the encoder network structure on OSA prediction, by comparing the performance of Siamese and single network structures when handling audio samples of sitting and lying positions. Specifically, in the Siamese network structure, the audio samples of each patient's sitting and lying positions are fed into separate halves of the network, whereas in the single network structure, the distinction between sitting and lying audio is not made and both samples are sent to the network. Our results, as presented in Table 1, demonstrate that the F1 score on the test set for the siamese network is superior to that for the single network. These findings suggest that distinguishing between sitting and lying speech can be helpful for accurate OSA prediction.

Table 1: *Classification results (F1 score) with audio features, where "No fine-tuning" means not training the encoder part with OSA labelled data; "No pre-training" means not using parameters from pre-trained model. For both "Single network" and "Siamese network", we do fine-tuning and pre-training.*

No fine-tuning	No pre-training
0.647	0.667
Single network	**Siamese network**
0.679	0.700

5.4. Additional patient characteristic features

Various patient characteristics can have varying impacts on the detection of OSA. As evidenced by the data presented in Table 2, Neck Circumference (NC) plays a beneficial role in detecting OSA. Moreover, by integrating all characteristics jointly, the model learns to discriminate among characteristics, leading to a performance that is deemed satisfactory.

Table 2: *Classification results (F1 score) with different additional characteristics of patients. For instance, "Age" indicates combining audio and age characteristics to detect OSA.*

None	Age	BMI	NC	All
0.700	0.674	0.660	0.706	0.707

5.5. Additional regression task

We assess whether the inclusion of a regression task to predict the AHI value would aid in the classification of OSA in the test set. To improve the performance of the model, we divided the MSE loss for the regression task by a factor of 2. The results presented in Table 3 demonstrate that incorporating the regression task during training of the Siamese network led to a notable improvement in the F1 score, from 0.700 to 0.708.

5.6. Combination of characteristics and regression

As detailed in Section 5.4 and Section 5.5, the inclusion of patient characteristics and a regression task has proven to enhance the performance of the model. It is therefore reasonable to explore the integration of these two methods. Based on the findings presented in Table 2, we selected NC as the patient characteristic to incorporate. The results reported in Table 3 indicate that the model utilizing both NC and regression yielded an F1 score of 0.725, surpassing the performance of all other models.

Table 3: *Classification results (F1 score) with/without the regression task or NC characteristic. "Baseline" means model without regression or NC.*

Baseline	Regression	NC	Regression and NC
0.700	0.708	0.706	0.725

6. Conclusions

In this paper, we present a novel method for OSA detection, which is a binary classifier built on XLSR, a pre-trained speech foundation model. The end-to-end nature of our approach makes it easy to scale up and deploy in real-world applications. Our experimental results verify the effectiveness of utilizing pre-trained representations for OSA detection. It is demonstrated that the Siamese network structure outperforms the single network structure by leveraging the differences between audio spoken in sitting and lying positions. Moreover, combining patient characteristic features with audio features and training with an additional AHI regression task can further enhance the model's performance. This preliminary study demonstrates the potential for AI-assisted OSA diagnosis. As further progress can be anticipated by collecting more data and employing more advanced speech foundation models. Our future work includes achieving multimodal diagnosis using visual features such as oral cavity images.

7. References

[1] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American journal of epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[2] C. F. George, "Reduction in motor vehicle collisions following treatment of sleep apnoea with nasal cpap," *Thorax*, vol. 56, no. 7, pp. 508–512, 2001.

[3] V. K. Somers, D. P. White, R. Amin, W. T. Abraham, F. Costa, A. Culebras, S. Daniels, J. S. Floras, C. E. Hunt, L. J. Olson *et al.*, "Sleep apnea and cardiovascular disease: An american heart association/american college of cardiology foundation scientific statement from the american heart association council for high blood pressure research professional education committee, council on clinical cardiology, stroke council, and council on cardiovascular nursing in collaboration with the national heart, lung, and blood institute national center on sleep disorders research (national institutes of health)," *Journal of the American College of Cardiology*, vol. 52, no. 8, pp. 686–717, 2008.

[4] A. Briançon-Marjollet, M. Weiszenstein, M. Henri, A. Thomas, D. Godin-Ribuot, and J. Polak, "The impact of sleep disorders on glucose metabolism: endocrine and molecular mechanisms," *Diabetology & metabolic syndrome*, vol. 7, no. 1, pp. 1–16, 2015.

[5] W. R. Ruehland, P. D. Rochford, F. J. O'Donoghue, R. J. Pierce, P. Singh, and A. T. Thornton, "The new aasm criteria for scoring hypopneas: impact on the apnea hypopnea index," *sleep*, vol. 32, no. 2, pp. 150–157, 2009.

[6] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. Hla, "Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study," *WMJ: official publication of the State Medical Society of Wisconsin*, vol. 108, no. 5, p. 246, 2009.

[7] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, "Speech signal and facial image processing for obstructive sleep apnea assessment," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[8] T. M. Davidson, "The great leap forward: the anatomic basis for the acquisition of speech and obstructive sleep apnea," *Sleep medicine*, vol. 4, no. 3, pp. 185–194, 2003.

[9] H. Zhang, J.-Y. Ye, L. Hua, Z.-H. Chen, L. Ling, Q. Zhu, L.-M. Wang, L. Zheng, and Y.-H. Zhang, "Inhomogeneous neuromuscular injury of the genioglossus muscle in subjects with obstructive sleep apnea," *Sleep and Breathing*, vol. 19, no. 2, pp. 539–545, 2015.

[10] Y. Ding, J. Wang, J. Gao, Q. Fang, Y. Li, W. Xu, J. Wu, and D. Han, "Severity evaluation of obstructive sleep apnea based on speech features," *Sleep and Breathing*, vol. 25, no. 2, pp. 787–795, 2021.

[11] Y. Ding, Y. Sun, Y. Li, H. Wang, Q. Fang, W. Xu, X. Chen, J. Wu, J. Gao, and D. Han, "Selection of osa-specific pronunciations and assessment of disease severity assisted by machine learning," *Journal of Clinical Sleep Medicine*, pp. jcsm–9798, 2021.

[12] 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.

[13] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015.

[14] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[15] 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.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] H. Nakano, T. Ikeda, M. Hayashi, E. Ohshima, and A. Onizuka, "Effects of body position on snoring in apneic and nonapneic snorers," *Sleep*, vol. 26, no. 2, pp. 169–172, 2003.

[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.