International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00407-7

REGULAR PAPER

Sound-based sleep assessment with controllable subject-dependent embedding using Variational Domain Adversarial Neural Network

Ken-ichi Fukui[1] · **Shunya Ishimaru**[2] · **Takafumi Kato**[3] · **Masayuki Numao**[1]

Received: 8 August 2022 / Accepted: 13 June 2023

Abstract
Sleep quality assessment as an indicator of daily health care plays an important role in our physiological and mental activity. Sound during sleep contains rich information on biological activities, such as body movement, snoring, and sleep bruxism. However, sound features differ depending on individual and environmental differences. In order to develop a wide-rage applicable daily sleep assessment, this paper utilizes deep learning to ease individual and environmental differences of sound features. Firstly, by Variational Domain Adversarial Neural Network (VDANN) encodes sound events into latent representation, simultaneously eliminates subject-dependent features. Then, sleep pattern in the obtained latent space is trained by Long Short-Term Memory (LSTM) with associated sleep assessment of one night. We performed age group estimation from normal sleep as an objective indicator of sleep comparing to their age group. The experiment with more than 100 subjects showed that VDANN is able to extract subject independent features, and the proposed method outperforms the conventional method for age group estimation from sleep sound even for new subjects. In addition, our model is able to personalize by controlling subject-dependent embedding when after data accumulation of the subject.

Keywords Sleep quality · Sound · Variational AutoEncoder · Domain adaptation

1 Introduction

Sleep is a physiological phenomenon that affects physical and psychological functions during waking hours. Assessment of the sleep state is an important parameter for health management. Polysomnography (PSG) is used to comprehensively measure sleep; however, due to high PSG-related patient burden, this measurement is only available in specialized facilities. At the same time, the downsizing of measurement devices and advancements in computers and networks in recent years have promoted the research and development of systems using sensor devices for easier sleep assessment, such as single channel EEG or Actigraphy [1, 2], also consumer wrist-worn device measuring acceleration and heart rate [3], and some more devices can be found in [4]. Recently, data-driven machine learning-based sleep assessment, which automate sleep staging, especially with development of deep learning, is evolving [5, 6].

The development of a sleep assessment system that can be used for daily sleep assessment requires the following: (1) convenient noncontact measurement and (2) objective indicators of sleep quality; lastly, machine learning-based estimation requires (3) improved generalizability that takes individual and environmental differences into consideration. This study focused on sleep environmental sound that can be conveniently collected through noncontact recording. Sleep environmental sound is a generic term for sound occurring during sleep. Sound-based sleep assessment has mainly focused on sleep apnea syndrome by detecting snoring [7–9], whereas sleep quality assessment is still uncommon.

This study focused on detectable sound events at the individual (e.g., body movements, sleep bruxism, snoring, sleep talking, etc.) and environmental (e.g., the sound of air con-

Ken-ichi Fukui and Shunya Ishimaru have contributed equally to this work.

Shunya Ishimaru is currently working at Toshiba Digital Solutions Corporation.

✉ Ken-ichi Fukui
fukui@ai.sanken.osaka-u.ac.jp

[1] SANKEN (The Institute of Scientific and Industrial Research), Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

[2] Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan

[3] Graduate School of Dentistry, Osaka University, 1-8 Yamadaoka, Suita, Osaka 565-0871, Japan

Published online: 04 July 2023

ditioners, outdoor noise, etc.) levels. Body movements [10], sleep bruxism [11], and snoring [12] are biological activities that are frequently observed during sleep that reflect individuals' physiological function during sleep.

In addition, using a sleep model that corresponds to individual and environmental differences is preferable during the initial stage of measurement. The machine learning model should identify the same type of events even for unknown subjects who are not included in the training data, thus subject-dependent features should be eliminated. After data accumulation, the use of a personalized sleep model is recommended. This aimed to develop a sleep model based on deep learning that reduces individual differences of sound features. Our model utilizes Variational Domain Adversarial Neural Network (VDANN) [13] to encode sound features, afterward, Long Short-Term Memory (LSTM) [14] was applied to model sleep pattern of a whole night and associate with a target sleep assessment.

Here, a significant correlation exists between age and sleep quality [15, 16]. This may be a useful objective sleep assessment if age estimation based on sleep characteristics is possible. Therefore, this study performed age group estimation using sleep environmental sound. If we could obtain *normal* sleep model associated with real age, then the model can evaluate one's sleep comparing to average sleep in their age.

In this experiment, we collected data on sleep environmental sound in the home environments of more than 100 subjects from a broad range of age groups, and trained latent representation of sleep events. Subsequently, we confirmed VDANN is able to extract subject independent feature and improved estimation capability against unseen subjects which is not included in the training data. In addition, our method is capable of personalization, which is for after data accumulation of the subject, by changing sign of the hyper-parameter in VDANN's loss function.

The contributions of this work can be summarized as follows:

- We proposed daily life sleep assessment model based on deep learning from environmental sound for non-contact and sleep-related body activity measurement.
- The proposed sleep model utilizes VDANN to encode sound event, and simultaneously eliminate subject dependent sound features to enhance generalization to new subjects.
- Also, we proposed to change sign of the hyperparameter in VDANN's loss function for personalization when certain subject's data is accumulated.
- We modeled normal sleep and performed age group estimation from sound events for objective evaluation.
- The performance of the proposed model was validated with more than 100 subjects' data at their home in wide rage of ages.

2 Related studies

2.1 Sleep assessment indicators

Sleep stages comprise rapid eye movement (REM) sleep and nonREM sleep [17]. Sleep stages are used for the physiological assessment of sleep architecture and patterns. In medicine, sleep stages are used for the evaluation of sleep characteristics and treatment-related changes in patients with sleep disorders. Therefore, the most of studies and systems related to conventional sleep assessment using machine learning are included in the estimation of sleep stages [1, 3, 5, 6].

However, various discussions exist on indicators of sleep quality [18]. Two typical sleep indicators related to sleep quality are wake time after sleep onset (WASO) and sleep efficiency (SE). WASO is calculated as the total wake time during sleep. SE is calculated as the ratio of total sleep time to time in bed. The work [19] estimated WASO and SE using machine learning based on the use of actigraphs.

Meanwhile, sleep characteristics are known to vary with age [15, 20]. The incidence of normal and deep nonREM sleep (slow wave sleep) is high during childhood. However, slow wave sleep begins to decrease during the late 30 s. In older adults, slow wave sleep and REM sleep further decrease, whereas WASO increases, preventing older adults from having a good night's sleep [21, 22]. We validated age group estimation performance for normal sleep from sleep sound events, based on the fact upon relationship between sleep characteristics and age.

2.2 Sleep assessment based on sleep sounds

Many of the studies on sleep-disordered breathing use sleep sounds. For example, Ren et al. [7] recorded sleep sounds using smartphones and earphones with microphones and performed the estimation of the apnea hypopnea index by detecting body movements, snoring, coughing, etc., using a Support Vector Machine and estimating respiratory rates.

Meanwhile, other sleep quality assessment studies estimate SE and WASO. Dafna et al. [23] proposed a method to estimate sleep time, SE, and WASO by calculating the sleep-wake likelihood from sound-based breathing and snoring sounds recorded by a microphone using the AdaBoost algorithm. They also performed a study to estimate SE and WASO based on sleep stages estimated from microphone sounds using a neural network [24]. Chang et al. [25] estimated SE by detecting sleep events (snoring, sleep bruxism, etc.) from smartphones using a decision tree. Zhang et al. [26] detected sleep stages using sounds recorded by a general-purpose microphone and Convolutional Neural Network (CNN) with a spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) as inputs. Unlike the above-mentioned

studies, the purpose of this study was to perform comprehensive assessments of sleep characteristics including snoring, bruxism, coughing, and body movement, not only detecting specific type of event.

2.3 Domain adaptation to reduce individual and environmental differences

Even the same type of sleep events may have different data distributions in the feature space, depending on subject personality traits and environments (collectively called as *domain*). Domain adaptation [27] is a machine learning method that minimizes possible differences in data distribution due to domain. Domain adaptation is a kind of transfer learning that covers change in the data distribution, while the task of source and target domain are the same. The common approach to deal with individual difference is Domain Adversarial Neural Network (DANN) which minimizes the label predictor loss while maximizing the domain classifier loss [28]. The domain classifier loss is maximized to obtain domain irrelevant representation. Jia et al. [29, 30] utilized DANN to obtain subject irrelevant feature of electroencephalogram (EEG) and electrooculogram (EOG) for sleep staging.

Meanwhile, Tu et al. [13] proposed a speaker recognition model that considers dataset differences comprising the adversarial network structure and Variational AutoEncoder (VAE), called Variational Domain Adversarial Neural Network (VDANN). Here, VAE [31] is a neural network to reconstruct inputs with some regularizer in the latent space, and can be used as a feature encoder. VDANN comprises an encoder, a decoder, a speaker classifier, and a domain classifier. Sound features are used as an input, and domains correspond to individual and environmental differences. VDANN acquires domain independent latent representations by training for high accuracy of the speaker classifier, low accuracy of the domain classifier, and low reconstruction error of the decoder. This paper also utilizes VDANN to obtain sound features mitigating individual and environmental differences. In addition, we propose to change a sign of the hyper-parameter in the loss function when target subject's data is enough accumulated, which is for personalization.

3 Proposed method

3.1 Overview

The proposed method uses the following five steps for sleep assessment (pseudo codes are shown in Algorithm 1).

1. Sleep events are extracted using a burst extraction method [32] from continuously recorded audio in a whole night (Algorithm 1 *l.* 1∼2)
2. After converting each sleep event into the frequency domain, a discrete power spectrum is used as an input vector (Algorithm 1 *l.* 3∼4)
3. The sleep events are encoded to the latent variables by VDANN (Algorithm 1 *l.* 5)
4. The sleep assessment estimator is trained by LSTM with a series of the latent variables of sleep events from a whole night (Algorithm 1 *l.* 6∼9)
5. During the reasoning, the same processing in steps (1) and (2) are performed. Then, sleep events are encoded by using the encoder of trained VDANN in step (3). Finally, sleep assessment is performed using the trained LSTM in step (4) with series of the encoded sleep events of a whole night.

Algorithm 1 Training of VDANN+LSTM for sleep assessment

Require:
Data tuple of i^{th} subject of j^{th} night:
$<X_{i,j},\ t_i,\ s_i>$,
where whole night sound signals:
$\mathbf{X} = \{X_{i,j} \mid i = 1, \ldots, ns,\ j = 1, \ldots, nn_i\}$,
the number of subjects: ns,
the number of nights in a subject: nn_i $(i = 1, \ldots, ns)$,
target sleep assessment labels: $\mathbf{t} = \{t_1, \ldots, t_{ns}\}$,
subject IDs: $\mathbf{s} = \{s_1, \ldots, s_{ns}\}$,
VDANN hyper parameters: α, β

Ensure: VDANN Encoder E, LSTM parameters θ_{lstm}

{(1) Extraction of sleep events}
1: **for all** i, j **do**
2: $\quad \{e_{i,j,k}\}_{k=1}^{ne_{i,j}} = Burst\_Extraction(X_{i,j})$
3: **end for**
{(2) Normalize frequency power spectrum for each event}
4: **for all** i, j, k **do**
5: $\quad x_{i,j,k} = FFT\_PowerSpec\_Normalized(e_{i,j,k})$
6: **end for**
{(3) Training of latent representation of sleep events using VDANN}
7: $E = Train\_VDANN(\mathbf{x}, \mathbf{t}, \mathbf{s}, \alpha, \beta)$
{(4) Encoding of latent variables and the training of LSTM}
8: **for all** i, j, k **do**
9: $\quad z_{i,j,k} = Encode\_by\_VDANN(E,\ x_{i,j,k})$
10: **end for**
11: $\theta_{lstm} = Train\_LSTM(\mathbf{z}, \mathbf{t})$
12: **return** E, θ_{lstm}

3.2 Preprocessing

First, sleep events are extracted from the sound continuously recorded during one night. In this study, we extracted sleep events by applying Kleinberg's burst extraction method [32], as done in previous study by Wu et al. [33]. In this method, the events are extracted as parts of the recordings that are

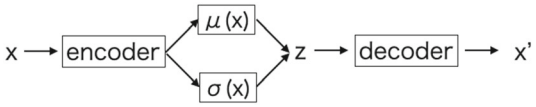

Fig. 1 Network configuration of VAE

estimated to be continuously generated from a normal distribution with a greater variance when compared with stationary noise, assuming that the amplitude of the measured voltage of the microphone has a normal distribution.

Next, Fast Fourier Transformation (FFT) is applied to the extracted sleep events. Here, the sampling rate of the microphone used is 48,000 Hz, and we discretized frequencies between 10 Hz and 24,000 Hz at 10 Hz intervals, thus, an input is a 2400-dimensional vector. Then, each vector was standardized so that the sum equaled one, for using Kullback–Leibler divergence in VAE. The intervals between discrete points were set at 10 Hz considering the balance between resolution with a clear peak and the computational cost after the visual inspection of the power spectrum.

3.3 Variational AutoEncoder (VAE)

VAE is a deep generative model proposed by Kingma and Welling [31]. As shown in Fig. 1, the network configuration of VAE comprises an encoder E and a decoder G. The network parameter of encoder E and decoder G is defined as ϕ_e and θ_g, respectively, whereas q_ϕ represents the stochastic encoder, and p_θ represents the stochastic decoder. The VAE training is performed by maximizing the evidence lower bound (ELBO); $L_{\mathrm{VAE}}(x;\theta_g,\phi_e)$ in below.

$$L_{\mathrm{VAE}}(x;\theta_g,\phi_e) = \mathcal{L}_{\mathrm{reconst}}(x;\ \theta_g,\phi_e) - KL[q_\phi(z|x) \parallel p_\theta(z)] \tag{1}$$

The first term is reconstruction error between input and the output from decoder G. The second term represents the Kullback–Leibler (KL) divergence of the latent variable z between the encoder and the decoder. The reconstruction error term can be obtained by assuming the distribution from decoder $p_\theta(x)$. As in Eq. (2), the terms for the KL divergence can be obtained by assuming normal distributions of $\mathcal{N}(z;\mathbf{0},\mathbf{I})$ and $\mathcal{N}(z;\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ for $p_\theta(z)$ and $q_\phi(z|x)$, respectively, where L represents the number of dimensions of the latent space.

$$-KL[q_\phi(z|x) \parallel p_\theta(z)] = \frac{1}{2}\sum_{l=1}^{L}\left(1+\log\sigma_l^2-\mu_l^2-\sigma_l^2\right) \tag{2}$$

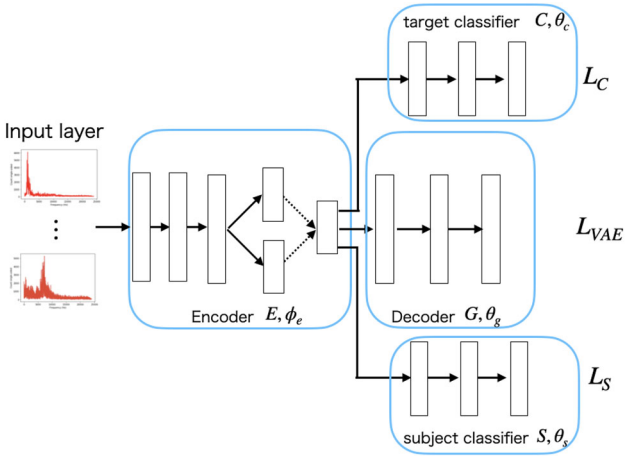

Fig. 2 Outline of the VDANN structure

3.4 Training of VDANN using sleep event data

As shown in Fig. 2, VDANN comprises encoder E, decoder G, target classifier C, and the subject (domain) classifier S.

The subject classifier is for obtaining subject independent feature by maximizing the classification error. VDANN can be expressed as the following equation. $x \in R^d$ represents the d-dimensional input vector, whereas $z \in R^L$ represents the L-dimensional latent variable. t represents the K-dimensional one-hot vector of the target label, whereas y represents the J-dimensional one-hot vector of subject labels. Also, $\phi_e, \theta_g, \theta_c, \theta_s$ represent the network parameters of encoder E, decoder G, target classifier C, and the subject classifier S, respectively.

In this study, supposing a power spectrum is a probability distribution (normalized as $\sum_i x_i = 1$), the reconstruction loss is defined by KL divergence between input x and the output of decoder G; $p(z)$, as follows:

$$\mathcal{L}_{reconst}(x;\ \theta_g,\phi_e) = d\cdot\mathbb{E}_{p(x)}[KL[x \parallel p(z)]] \tag{3}$$

Note that we found the KL divergence reconstruction loss empirically achieved better reconstruction compared to common MSE (Mean Squared Error) loss in our data (shown in Sect. 4.3).

In addition, as with the normal VAE, z was sampled from the output of the encoder using the reparameterization trick.

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon\sigma} \quad (\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})) \tag{4}$$

The loss functions of target classifier C and the subject classifier S were, respectively, defined as cross-entropy loss as in Eqs. (5) and (6).

$$\mathcal{L}_C(x;\theta_c,\phi_e) = \mathbb{E}_{p(x)}\left[-\sum_{k=1}^{K} t_k \log C(z)_k\right] \tag{5}$$

$$\mathcal{L}_S(x;\theta_s,\phi_e) = \mathbb{E}_{p(x)}\left[-\sum_{j=1}^{J} y_j \log S(z)_j\right] \tag{6}$$

The softmax function was used as the activation function of the output layer of the target classifier C and the subject classifier S. Based on the above, the VDANN loss function is defined as follows.

$$\begin{aligned}&\mathcal{L}_{\mathrm{VDANN}}(x;\phi_e,\theta_g,\theta_c,\theta_s)\\ &\quad= \mathcal{L}_{\mathrm{VAE}}(x;\phi_e,\theta_g) + \alpha\mathcal{L}_C(x;\theta_c,\phi_e) + \beta\mathcal{L}_S(x;\theta_s,\theta_g)\end{aligned} \tag{7}$$

where α and β are hyperparameters to adjust the weight of $\mathcal{L}_{\mathrm{VAE}}$, $\mathcal{L}_C$ and $\mathcal{L}_S$. Then, the training of VDANN is performed by the following optimization.

$$\min_{\phi_e,\theta_g,\theta_c,\theta_s} \sum_x \mathcal{L}_{\mathrm{VDANN}}(x;\phi_e,\theta_g,\theta_c,\theta_s) \tag{8}$$

As with the VDANN of Tu et al. [13], this is a model with lower individual differences when $\beta < 0$, by maximizing the classification error. In addition, unlike study of Tu et al., we show better performance by setting $\beta > 0$, when estimating unknown night of known subject when using age group estimation as the target.

When $\beta < 0$, the following adversarial training is performed. First, the optimization of the subject classifier S is performed by maximizing $\mathcal{L}_S$. Next, the optimization of VDANN is performed by minimizing target classification loss $\mathcal{L}_C$ and reconstruction error $\mathcal{L}_{\mathrm{VAE}}$ with a fixed parameter S. This alternate optimization is performed until convergence.

$$\hat{\theta}_s = \arg\max_{\theta_s} \sum_x \mathcal{L}_S(x;\theta_s|\,\hat{\phi}_e,\hat{\theta}_g,\hat{\theta}_c) \tag{9}$$

$$(\hat{\phi}_e,\hat{\theta}_g,\hat{\theta}_c) = \arg\min_{\phi_e,\theta_g,\theta_c} \sum_x \mathcal{L}_{\mathrm{VDANN}}(x;\phi_e,\theta_g,\theta_c,\hat{\theta}_s) \tag{10}$$

3.5 Effect of β coefficient

In this section, we show the effect of β coefficient, which controls individual differences of sound features in the latent space. Figure 3 shows the differences in the distribution of sleep events in the latent space obtained by VDANN with changing sign of β, plotted by age group and by different subjects.

When $\beta < 0$, the adversarial training is performed as described in Eqs. (9), (10), resulting subject independent embedding can be obtained. In other words, the latent space is trained to make it impossible to identify individuals in the latent space (the right bottom in Fig. 3), so that sleep events for the same age group are close in the latent space (the left bottom in Fig. 3). This increases the accuracy of age group estimation in new subjects because the latent space can be useful for other subjects.

In contrast, when $\beta > 0$, $\mathcal{L}_{\mathrm{VAE}}$, $\mathcal{L}_{\mathrm{C}}$, $\mathcal{L}_{\mathrm{S}}$ are all minimized. The subject classifier $\mathcal{L}_{\mathrm{S}}$ and the target classifier $\mathcal{L}_{\mathrm{C}}$ are both trained to reduce misclassification and acquire a feature encoder that identifies subjects (the right top in Fig. 3). This encoder has benefit for the target such as age group, because age group as a estimation target remains the same in the same subjects within a short period of time. However, the location of the sound events from new subjects in the latent space may not predictable, thus the accuracy of age group estimation for new subjects will be low when $\beta > 0$.

Table 1 Basic statistics of dataset by age group

Age group	#subject	#total nights	Mean age (SD)
20 s	57	770	23.0 (2.0)
30 s	11	142	34.9 (3.1)
40 s	27	439	45.0 (2.6)
50 s	16	221	54.2 (2.3)

3.6 Training of LSTM for sleep assessment

Lastly, sequence-to-one (label) training is performed using a series of latent variables $\{z_1, z_2, \ldots, z_n\}$ as an input of LSTM after encoding a set of sleep events during one night $\{x_1, x_2, \ldots, x_n\}$ using the VDANN encoder E. Sequence-to-one (label) training is a type of training that generates one class label as an output, i.e., age group. The softmax function is used as the activation function of the output layer. The variable t represents the K-dimensional one-hot vector of the target label, whereas y represents the output of LSTM where the loss function of LSTM is defined by the cross-entropy error as in Eq. (11).

$$\mathcal{L}_{\mathrm{LSTM}} = -\sum_{k=1}^{K} t_k \log y_k \tag{11}$$

4 Experiment and results

4.1 Dataset

In this study, data were collected from Japanese subjects recruited from a broad range of age groups. This experiment was performed with the approval of the Institutional Ethics Committee in SANKEN, Osaka University, Japan (approval ID: H29-05). The subjects were asked to record sleep environmental sound by a smartphone (ASUS/ZenFone 3 Laser) in a single bedroom for 28 days. Before and after bedtime,

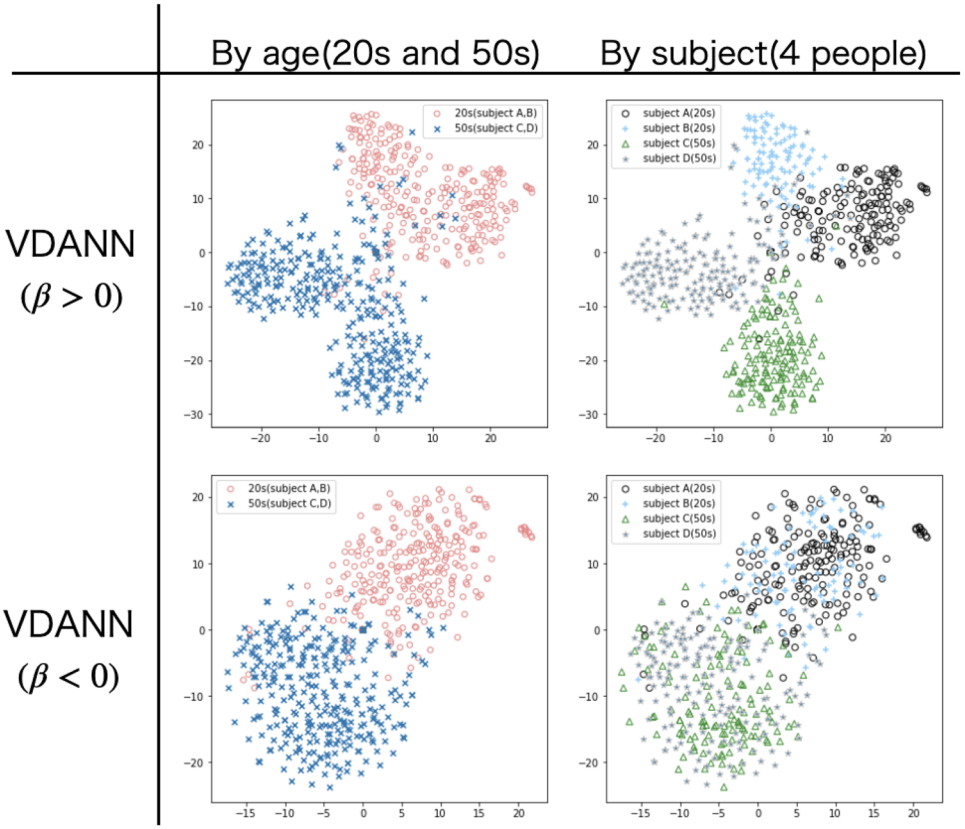

Fig. 3 Visualization example of sleep events in the latent space. The distributions of sleep events in two subjects in their 20 s and two subjects in their 50 s are plotted by t-SNE [34]

they were asked to fill out a questionnaire on mental and physical fatigue, sleep satisfaction, sleep quality (5 levels), and room environment during sleep also by a smartphone. As of June 2020, a total of about 7500 days of data were collected from 267 subjects. The large sample size of this study allowed us to estimate age groups by modeling sleep patterns from sounds.

As for the age group estimation, questionnaire data completed on days without air conditioning were used to develop a model of normal sleep patterns using low-noise sleep environmental sound. In addition, in order to develop *normal* sleep model for age group estimation, we selected subjects with more than seven days were better-than-usual sleep satisfaction (above 3 out of 5 by questionnaire) and who had the flu or injuries were excluded from analysis. A total of 1572 days of data from 111 subjects were identified in the above sampling. Table 1 shows the number of subjects and the number of days for each age group.

4.2 Sleep event encoding

In this experiment, we compared the encoding of sleep events by AutoEncoder (AE) and VAE. Each network structure of both the encoder and decoder comprised two fully-connected hidden layers. The encoder had the Batch Normalization [35] after the input layer. The number of dimensions of latent space were tested though $L \in \{2, 4, 8, 16, 32, 64, 128\}$, L was set at 64 because F-score of the age group classification were not improved after $L = 64$.

Table 2 Average reconstruction error (MSE) and standard deviation of AE and VAE

	Training data	Test data
AE	20.82 ± 0.63	22.25 ± 0.26
VAE	18.72 ± 0.58	19.92 ± 0.23

Table 2 shows the average reconstruction error (MSE) and the standard deviation of the training and test data, comparing AE and VAE using MSE as the reconstruction loss. This result quantitatively shows the superiority of VAE compared with AE.

After the training of VDANN using sleep event data, reconstructed power spectrum of sleep events using test data were confirmed. Figure 4 shows the comparison of the results of the reconstruction after the adversarial training with $\alpha = 1$ and $\beta = -1000, -2000$ when the number of dimensions of the latent space is $L = 64$. It is evident that the greater the value of β is in the negative direction, the greater the reduction in inter-individual differences in sleep events in the training set, preventing the acquisition of appropriate latent representation. Therefore, it is necessary to adjust β appro-

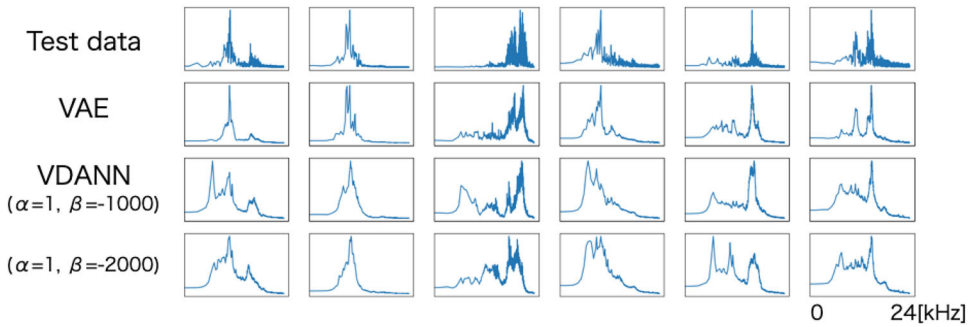

Fig. 4 Comparison of the reconstruction by VAE and VDANN (the horizontal axis is a logarithmic scale)

priately based on the target (sleep assessment) classification accuracy.

4.3 MSE vs. KLD reconstruction loss

In this section, we compared MSE and KLD reconstruction loss in AE and VAE. Since MSE and KLD are different reconstruction evaluation criteria, we visually inspected by plotting data distribution in the latent space by t-SNE as shown in Fig. 5. The event labeling was made by actually listening to some sounds during typical sleep events, body movement, snoring, cough, etc.

Comparing AE with MSE (left top in Fig. 5) and AE with KLD (left bottom), KLD is more dense in snoring events than using MSE, which means KLD is suitable to capture the similarity of snoring events. While comparing VAE with MSE (right top) and VAE with KLD (right bottom), though MSE is almost random distribution, by using KLD each event type is closely distributed. When comparing AE with KLD (left bottom) and VAE with KLD (right bottom), VAE with KLD is more dense in snoring, cough, and TV noise events than AE with KLD. In conclusion, VAE with KLD is the best combination for sleep event encoding.

4.4 Effect of VDANN encoder in event type distribution

The differences in the distribution of sleep events in the latent space in two subjects between VAE and VDANN are shown in Fig. 6. In the case of VAE, coughing events and snoring events of subject A (red circle in Fig. 6a) and subject B (blue circle in Fig. 6a) are close to each other in the latent space because the frequency power spectrum distributions of sleep events are similar in the same subject. In contrast, sleep events in different subjects are dispersed in the latent space due to different frequency power spectrum distributions. For example, "snoring" events in subjects A and B are dispersed. Meanwhile, it is clear that VDANN is trained so that coughing and snoring events in subjects A and B are, respectively, close in the latent space. The coughing events of subjects A and B are both closely distributed in Fig. 6b (events 1–3), and the snoring events (events 4–6) as well. This enables the model to generalize for unseen subjects.

4.5 Age group estimation

4.5.1 Experimental setup

This experiment compared the results of age group estimation based on sleep environmental sound using the same dataset (111 subjects, a total of 1572 days) as that used in the previous work [36]. Four age groups (20 s, 30 s, 40 s, and 50 s) were used for age estimation. This study used 10-year age groups because such age differences usually exhibit general differences in sleep characteristics [37, 38]. The work [36] performed age groups estimation by converting sleep environmental sound during a whole night into a spectrogram without the event extraction, and using a deep learning model comprising CNN and transfer learning. To the best of our knowledge, no other work performed age group estimation from sound.

The number of sleep events that can be identified during one night is varied about 100–14,000 in this dataset. To minimize training time and bias in the number of sleep events, sleep events of < 100 ms were excluded from this experiment. If the number of sleep events during one night exceeded 2000, sampling of sleep events was performed after setting the time interval threshold. The time interval threshold was set between 1 and 10 s so that the number of sleep events was ≤ 2000/day.

We performed two types of cross-validation; subject-level and night-level. At the subject-level, sound data of subjects that are not included in the training data was used as test data. Hence, the subject-level is for testing completely new subjects. While, at the night-level, different nights of sound data from the same subject are included in the training and test data. At the night-level, we examined whether it is possible to perform age group estimation from sound feature, and also for when the target subject's data is accumulated.

In this experiment, the encoder part of VDANN comprised the single Batch Normalization layer and three fully-connected layers (the number of units are 256, 128, and

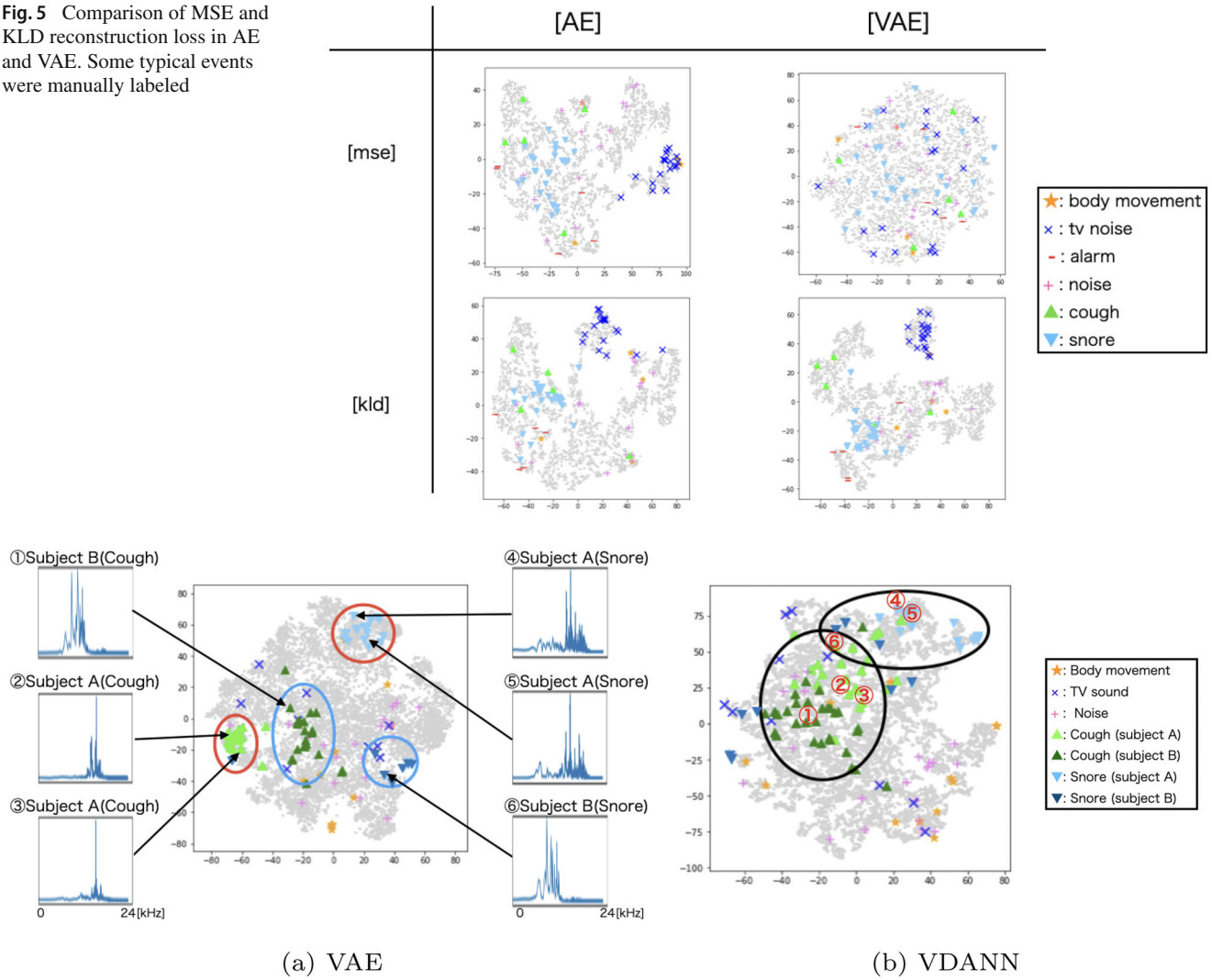

Fig. 5 Comparison of MSE and KLD reconstruction loss in AE and VAE. Some typical events were manually labeled

Fig. 6 Example of the relationship between the distribution of sleep events in the latent space and the frequency power spectrum (the horizontal axis of the plot is in a log scale), with $\alpha = 0$, $\beta = -2000$ in VDANN (color figure online)

64), whereas the decoder part comprised two fully-connected layers (the number of units are 128 and 256). The subject classifier and target classifier comprised a fully-connected layer (number of units is 128) and the dropout layer (dropout rate of 20%). These hyperparameters were adjusted to reduce or minimize the reconstruction error in the training with sleep events. The ReLU function was used as the activation function for the hidden layers. The softmax function was used as the activation function for the output layer. In addition, a single hidden layer was used for LSTM.

In this experiment, we performed stratified fivefold cross-validation. By changing hyperparameters α and β, the number of dimensions L of latent space in VDANN, and the number of units of the hidden layer in LSTM, the final hyperparameters were as follows: $\alpha = 2000$, $\beta = \pm 200$, $L = 64$, and the number of units in LSTM is 128.

4.5.2 Experimental results

Tables 3 and 4 show the results of age group estimation at the subject-level and the night-level for each model. When $\beta = 0$ which is without domain adaptation, the accuracy of age group estimation was slightly lower at the subject-level and significantly lower at the night-level than that in the previous work [36]. However, when $\beta < 0$ which leads subject-independent feature, the proposed model had higher performance than the conventional models by capturing the characteristics of each age group to reduce individual differences. Here, F-score of subject-level estimation is low, however, it is above the chance level, as the F-score when estimating by the majority age group (20 s) was 0.1635.

While, when $\beta > 0$ which leads subject-dependent feature, F-score at the night-level was higher than that in the conventional models, which do not use subject labels. The

Table 3 Subject-level cross-validation results of age group estimation (average and standard deviation)

	Precision	Recall	F-score
VAE + LSTM	0.296 ± 0.039	0.301 ± 0.047	0.279 ± 0.028
VDANN+LSTM ($\beta > 0$)	0.311 ± 0.020	0.303 ± 0.022	0.288 ± 0.028
VDANN+LSTM ($\beta = 0$)	0.290 ± 0.043	0.300 ± 0.054	0.291 ± 0.046
VDANN+LSTM ($\beta < 0$)	$\mathbf{0.328 \pm 0.031}$	$\mathbf{0.324 \pm 0.023}$	$\mathbf{0.319 \pm 0.028}$
CNN [36]	0.305 ± 0.050	0.304 ± 0.032	0.304 ± 0.041
VGG16 [36]	0.122 ± 0.011	0.250 ± 0.000	0.164 ± 0.010
InceptionV3 [36]	0.218 ± 0.071	0.251 ± 0.019	0.230 ± 0.044
Xception [36]	0.312 ± 0.023	0.291 ± 0.019	0.300 ± 0.018

Boldface indicates our proposed method and the highest score in each metric

Table 4 Night-level cross-validation results of age group estimation (average and standard deviation)

	Precision	Recall	F-score
VAE + LSTM	0.686 ± 0.022	0.678 ± 0.025	0.678 ± 0.023
VDANN+LSTM ($\beta > 0$)	0.848 ± 0.024	$\mathbf{0.825 \pm 0.033}$	$\mathbf{0.830 \pm 0.022}$
VDANN+LSTM ($\beta = 0$)	0.791 ± 0.042	0.774 ± 0.064	0.764 ± 0.055
VDANN+LSTM ($\beta < 0$)	0.821 ± 0.028	0.779 ± 0.023	0.785 ± 0.028
CNN [36]	0.752 ± 0.035	0.748 ± 0.049	0.750 ± 0.050
VGG16 [36]	0.123 ± 0.000	0.250 ± 0.000	0.164 ± 0.000
InceptionV3 [36]	0.592 ± 0.119	0.551 ± 0.090	0.568 ± 0.095
Xception [36]	$\mathbf{0.851 \pm 0.027}$	0.803 ± 0.050	0.826 ± 0.031

Boldface indicates our proposed method and the highest score in each metric

proposed method captures sound feature of each subject together with age group, and leading a higher accuracy of the estimation. This model has a benefit when after accumulating data from target subject.

In addition, the number of parameters used for the training in the proposed method is 1/100 or less than that of [36] with a significantly-reduced amount of spatial complexity in the model.

4.6 Effects of changes in α and β on F-scores

Figures 7 and 8 show the mean F-scores when α is changed in $\{0, 200, 2000, 20000\}$ and β is changed in $\{-1600, -800, -400, -200, 0, 200, 400, 800, 1600\}$, respectively, where $\beta = 0$ represents the case in which the subject (domain) classifier is not used. At the subject-level (Fig. 7), the F-score peaked at $\alpha = 2000$ and $\beta = -200$, showing the best score when $\beta < 0$. The F-score generally tended to be higher at $\beta < 0$ than at $\beta > 0$, showing the effects of β (regardless of α) on F-score.

In contrast, at the night-level (Fig. 8), the F-score peaked when $\beta > 0$; $\alpha = 2000$ and $\beta = 200$. Here, a pronounced saturation effect for large F-scores due to significant changes in α and β was shown at $\alpha \geq 2000$. When $\alpha = 2000$ and $\beta = 0$, the mean F-score was approximately 0.76, whereas when $\alpha = 2000$ and $\beta = 200$, the mean F-score was approximately 0.8, showing improved accuracy of age group estimation using a subject classifier.

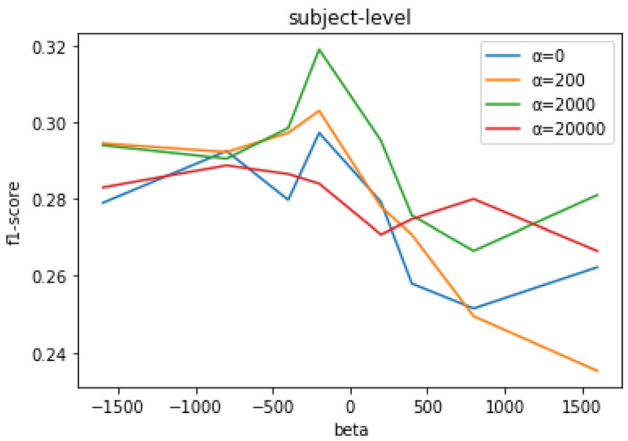

Fig. 7 Effect of α and β in subject-level evaluation

4.7 Discussion

The purpose of this experiment was to verify the accuracy of the method of age group estimation using normal sleep. Since this experiment cannot distinguish classification error and poor sleeper, future studies should include subjects with low sleep quality or disease, or comparison to sleep assessment by PSG, in order to validate the classification accuracy for poor sleepers. The other direction is that the proposed method should be validated in various conditions such as other recording devices, regions, or countries.

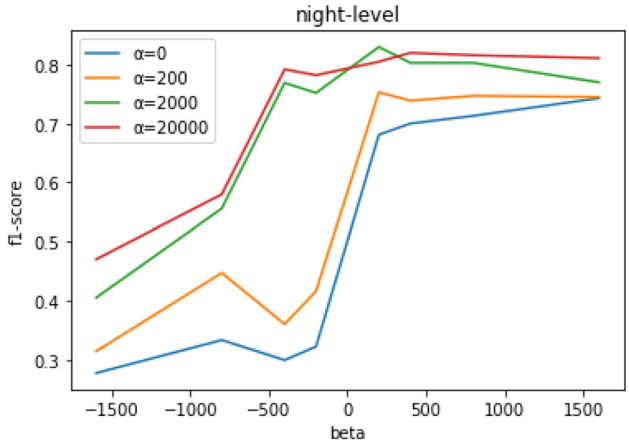

Fig. 8 Effect of α and β in night-level evaluation

As for technical improvement of deep learning, Time-Aware LSTM (T-LSTM) [39] with irregular intervals of events or Transformer model [40] with attention mechanism may improve time series modeling. Also different input expressions, such as the (log) power spectrogram and (log) Mel spectrogram, may be an option. Furthermore, automatic optimization of hyperparameters by Bayesian optimization and data augmentation may improve the accuracy of estimation.

Moreover, in this study, we attempted to estimate age groups based on biological activities during sleep from sounds. However, various other factors affect sleep, such as life factors (sleep and wake patterns, physical activity during day, etc.), physiological factors (metabolic function, etc.), physical factors (BMI, etc.), psychological factors (stress, anxiety, etc.), disease, and environmental factors (room temperature and illuminance). Our current model does not adequately consider these factors. The currently-collected dataset includes information on individual profiles, such as gender, height, weight, physical activity during the day, heart rate, and room sensors. Therefore, the development of a sleep model that considers this information in addition to sleep environmental sound may contribute to improved sleep assessment.

5 Conclusion

In this study, we developed a sleep sound event encoder that controls individual differences of sound features by utilizing Variational Domain Adversarial Neural Network (VDANN), then age group estimation is performed using LSTM with a series of the encoded latent variables as inputs.

In the training of VDANN, we improved the estimation accuracy before-and-after data accumulation in new subjects by switching the positive and negative signs of hyperparameter β. The experiment using sleep sound data at home environment with more than 100 subjects shows that the effectiveness of the proposed method compared with conventional methods. Our method can improve the classification accuracy even for unseen subjects by mitigating individual differences of sound features.

Acknowledgements This research was supported by JSPS KAKENHI Grant Numbers JP22K19832, the Center of Innovation Program from Japan Science and Technology Agency (JST), and by a Project, JPNP06046, subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

Funding Open access funding provided by Osaka University.

References

1. Fu, M., Wang, Y., Chen, Z., Li, J., Xu, F., Liu, X., Hou, F.: Deep learning in automatic sleep staging with a single channel electroencephalography. Front. Physiol. (2021). https://doi.org/10.3389/fphys.2021.628502
2. Park, K., Choi, S.H.: Smart technologies toward sleep monitoring at home. Biomed. Eng. Lett. **9**, 73–85 (2019)
3. Walch, O., Huang, Y., Forger, D., Goldstein, C.: Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep **42**(12), zsz180 (2019)
4. Kwon, S., Kim, H., Yeo, W.-H.: Recent advances in wearable sensors and portable electronics for sleep monitoring. iScience **24**(5), 102461 (2021)
5. Supratak, A., Dong, H., Wu, C., Guo, Y.: DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(11), 1998–2008 (2017). https://doi.org/10.1109/TNSRE.2017.2721116
6. Mousavi, S., Afghah, F., Acharya, U.R.: SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. PLoS ONE **14**(5), 1–15 (2019). https://doi.org/10.1371/journal.pone.0216456
7. Ren, Y., Wang, C., Yang, J., Chen, Y.: Fine-grained sleep monitoring: hearing your breathing with smartphones. In: Proceedings of 2015 IEEE Conference on Computer Communications (INFOCOM), pp. 1194–1202 (2015)
8. Jiang, Y., Peng, J., Zhang, X.: Automatic snoring sounds detection from sleep sounds based on deep learning. Phys. Eng. Sci. Med. **43**(2), 679–689 (2020). https://doi.org/10.1007/s13246-020-00876-1
9. Xie, J., Aubert, X., Long, X., van Dijk, J., Arsenali, B., Fonseca, P., Overeem, S.: Audio-based snore detection using deep neural

networks. Comput. Methods Programs Biomed. (2021). https://doi.org/10.1016/j.cmpb.2020.105917
10. Merlino, G., Gigli, G.L.: Sleep-related movement disorders. Neurol. Sci. **33**, 491–513 (2012)
11. Kato, T., Yamaguchi, T., Okura, K., Abe, S., Lavigne, G.J.: Sleep less and bite more: sleep disorders associated with occlusal loads during sleep. J. Prosthodont. Res. **57**, 69–81 (2013)
12. Hall, A.P.: Sleep, sleep studies and sleep-disordered breathing. Curr. Opin. Anaesthesiol. **30**(1), 163–167 (2017)
13. Tu, Y., Mak, M., Chien, J.: Variational domain adversarial learning for speaker verification. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 4315–4319 (2019)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
15. Mander, B., Winer, J., Walker, M.: Sleep and human aging. Neuron **94**, 19–36 (2017)
16. Gadie, A., Shafto, M., Leng, Y., Kievit, R.A.: How are age-related differences in sleep quality associated with health outcomes? An epidemiological investigation in a UK cohort of 2406 adults. BMJ Open **7**(7), e014920 (2017)
17. Berry, R., Brooks, R., Gamaldo, C., Harding, S., Lloyd, R., Quan, S., Troester, M., Vaughn, B.: AASM scoring manual updates for 2017 (version 2.4). J. Clin. Sleep Med. **13**(5), 665–666 (2017)
18. Mendonca, F., Mostafa, S., Morgado-Dias, F., García, A.G., Penzel, T.: A review of approaches for sleep quality analysis. IEEE Access **7**, 24527–24546 (2019)
19. Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., Arora, T., Taheri, S.: Sleep quality prediction from wearable data using deep learning. JMIR Mhealth Uhealth **4**(4), 125 (2016)
20. Crowley, K.: Sleep and sleep disorders in older adults. Neuropsychol. Rev. **21**, 41–53 (2011)
21. Van Cauter, E., Leproult, R., Plat, L.: Age-related changes in slow wave sleep and REM sleep and relationship with growth hormone and cortisol levels in healthy men. JAMA **284**(7), 861–868 (2000)
22. Ohayon, M.M., Carskadon, M.A., Guilleminault, C., Vitiello, M.V.: Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. Sleep **27**(7), 1255–1273 (2004)
23. Dafna, E., Tarasiuk, A., Zigel, Y.: Sleep-wake evaluation from whole-night non-contact audio recordings of breathing sounds. PLoS ONE **10**(2), 1–22 (2015). https://doi.org/10.1371/journal.pone.0117382
24. Dafna, E., Tarasiuk, A., Zigel, Y.: Sleep staging using nocturnal sound analysis. Sci. Rep. (2018). https://doi.org/10.1038/s41598-018-31748-0
25. Chang, X., Peng, C., Xing, G., Hao, T., Zhou, G.: iSleep: a smartphone system for unobtrusive sleep quality monitoring. ACM Trans. Sens. Netw. (2020). https://doi.org/10.1145/3392049
26. Zhang, Y., Chen, Y., Hu, L., Jiang, X., Shen, J.: An effective deep learning approach for unobtrusive sleep stage detection using microphone sensor. In: Proceedings of 2017 International Conference on Tools with Artificial Intelligence (ICTAI), pp. 37–44 (2017)
27. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. CoRR arxiv:2010.03978 (2020)
28. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**, 1–35 (2016)
29. Jia, Z., Lin, Y., Wang, J., Ning, X., He, Y., Zhou, R., Zhou, Y., Lehman, L.-W.H.: Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. IEEE Trans. Neural Syst. Rehabil. Eng. **29**, 1977–1986 (2021). https://doi.org/10.1109/TNSRE.2021.3110665
30. Jia, Z., Cai, X., Jiao, Z.: Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. IEEE Sens. J. **22**(4), 3464–3471 (2022). https://doi.org/10.1109/JSEN.2022.3140383
31. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR) (2014)
32. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1–25 (2002)
33. Wu, H., Kato, T., Numao, M., Fukui, K.: Statistical sleep pattern modelling for sleep quality assessment based on sound events. Health Inf. Sci. Syst. **5**, 1–11 (2017)
34. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008)
35. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning (ICML), pp. 448–456 (2015)
36. Kalintha, W., Kato, T., Fukui, K.: SleepAge: sleep quality assessment from nocturnal sounds in home environment. In: Procedia Computer Science, vol. 176, pp. 898–907 (2020)
37. Van Cauter, E., Leproult, R., Plat, L.: Age-related changes in slow wave sleep and REM sleep and relationship with growth hormone and cortisol levels in healthy men. J. Am. Med. Assoc. **284**(7), 861–868 (2000)
38. Li, L., Nakamura, T., Hayano, J., Yamamoto, Y.: Age and gender differences in objective sleep properties using large-scale body acceleration data in a Japanese population. Sci. Rep. **11**, 9970 (2021)
39. Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient subtyping via time-aware LSTM networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 65–74 (2017)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)