HEART RATE AND OXYGEN SATURATION ESTIMATION FROM FACIAL VIDEO WITH MULTIMODAL PHYSIOLOGICAL DATA GENERATION

Yusuke Akamatsu[†] *Yoshifumi Onishi* *Hitoshi Imaoka*

Biometrics Research Laboratories, NEC Corporation, Japan

ABSTRACT

Efforts to estimate multiple physiological parameters such as heart rate and oxygen saturation from facial videos have been made. However, training robust machine learning models for the estimation is challenging without large multimodal physiological datasets containing multiple physiological parameters and facial videos. In this paper, we propose a method to estimate heart rate and oxygen saturation from facial videos with multimodal physiological data generation. To collect sufficient datasets, the proposed method generates multimodal physiological datasets from several datasets containing a part of physiological modalities. Furthermore, to accurately estimate physiological parameters for unseen subjects, $i.e.$, not included in the training data, we generate a multimodal physiological dataset for unseen subjects by using short facial videos of unseen subjects. Experimental results using three public datasets show the effectiveness of our multimodal physiological data generation.

***Index Terms*—** Telemedicine, remote photoplethysmography (rPPG), heart rate, oxygen saturation, multimodal generative model

1. INTRODUCTION

With the pandemic of a novel Coronavirus disease COVID-19, telemedicine has become increasingly important. Telemedicine allows patients to receive medical examinations while staying at home and it is not necessary to go to hospitals. In medical examinations, monitoring multiple physiological parameters such as heart rate (HR) and blood oxygen saturation (SpO2) is essential for the early detection of many diseases. Hence, the development of remote monitoring technologies for these physiological parameters brings significant benefits for telemedicine.

Over the last decade, video-based non-contact physiological measurement has attempted to estimate HR [1–5] and SpO2 [5, 6] from facial videos. This facial video-based measurement allows the patients to monitor their physiological parameters by using cameras on personal mobile devices ($e.g.$, tablet and smartphone) without preparing medical devices ($e.g.$, pulse oximeter). Previous HR and SpO2 estimation methods [1–6] aimed to extract cardiac activity-induced color variations on the face, and achieved good estimation performance under constrained situations. However, noises from several sources, $e.g.$, head motions and illumination changes can easily break prior assumptions of these handcrafted signal processing pipelines. Therefore, the performance is decreased in the practical application scenarios.

To overcome the problem of signal processing approaches, several machine learning (ML)-based approaches [7–11] have been proposed and achieved good performance under unconstrained scenarios. For example, a previous method [8] constructed convolutional neural networks (CNN) to estimate HR from facial color signals. By incorporating facial videos in different head motion scenarios acquired from three cameras into the training data, the method [8] can estimate HR stably with head motions and multiple cameras. Although ML-based approaches may perform well in several practical scenarios, the performance heavily depends on the quantity of training data for ML models. However, it is difficult to collect a large dataset containing facial videos and multiple physiological parameters due to the high cost of data collection. Specifically, a fully public Face+HR+SpO2 dataset [1] is extremely limited. To the best of our knowledge, we know only rPPG dataset [8] consisting of eight subjects. Furthermore, since subjects have large individual differences in appearance and physiology ($e.g$, skin colors and pulse dynamics), ML-based approaches struggle to estimate physiological parameters for unseen subjects, $i.e.$, not included in the training data [12].

In this paper, we propose a method to estimate HR and SpO2 from facial videos with multimodal physiological data generation. We solve the problem about the quantity of training data and the difficulty to estimate physiological parameters for unseen subjects by the following two approaches, respectively:

Approach (i) : To increase the quantity of the training data for ML models, we generate multimodal datasets ($i.e.$, Face+HR+SpO2 dataset) from several datasets containing a part of modalities ($e.g.$, Face+HR or HR+SpO2), which can be collected relatively easily.

Approach (ii) : We use pseudo HR that is estimated from unseen subjects' short facial videos via an unsupervised signal processing approach. Then we generate a multimodal dataset from short facial videos and pseudo HR as training data for unseen subjects, and adapt ML models to unseen subjects. This approach does not require the ground truth HR and SpO2 for unseen subjects.

To generate multimodal datasets in **Approaches (i) and (ii)**, we construct a multimodal physiological generative model called *MultiPhys*. MultiPhys can generate physiological data in arbitrary directions, $e.g.$, Face+HR $\rightarrow$ SpO2, HR+SpO2 $\rightarrow$ Face, and Face $\rightarrow$ HR+SpO2. The experimental results using three public datasets, rPPG dataset [8] (Face+HR+SpO2), UBFC-Phys dataset [13] (Face+HR), and BIDMC dataset [14] (HR+SpO2), show the effectiveness of our multimodal physiological data generation.

2. METHODS

In this section, we explain MultiPhys (see **2.1**) and multimodal physiological data generation (see **2.2**). The procedures of the proposed method are as follows: We first train MultiPhys by a small multimodal dataset, Face+HR+SpO2 dataset. Then, as **Approach (i)**, Face+HR+SpO2 datasets are generated from Face+HR and HR+SpO2 datasets by estimating missing modalities from partially observed modalities via MultiPhys. Furthermore, as **Approach (ii)**, MultiPhys generates a Face+HR+SpO2 dataset for un-

[†]Contact author : yusuke-akamatsu@nec.com

[1]This means the dataset containing facial videos, HR, and SpO2.

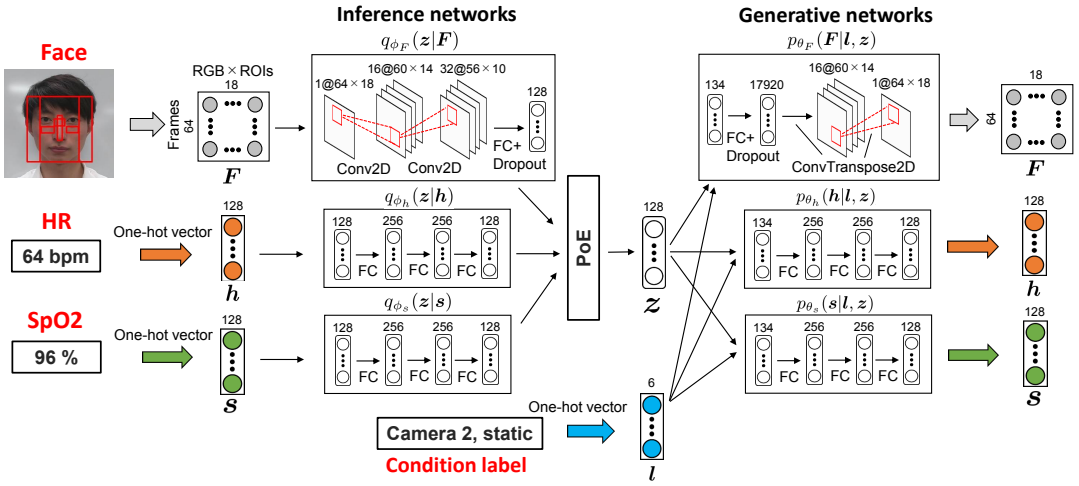

Fig. 1. The architecture of MultiPhys. Facial signals $\boldsymbol{F}$, HR class $\boldsymbol{h}$, and SpO2 class $\boldsymbol{s}$ are converted into the latent variable $\boldsymbol{z}$ via a product-of-experts (PoE) inference network. Then $\boldsymbol{F}$, $\boldsymbol{h}$, and $\boldsymbol{s}$ are reconstructed from $\boldsymbol{z}$ and condition label $\boldsymbol{l}$ via three generative networks. Each network consists of fully connected (FC), convolution (Conv2D), or transposed convolution (ConvTranspose2D) layers.

seen subjects from short facial videos of unseen subjects and pseudo HR. Finally, ML-based estimation models are trained by the original and generated Face+HR+SpO2 datasets; then HR and SpO2 are estimated from facial videos. Note that MultiPhys also can be used to estimate HR and SpO2 from facial videos. In this case, MultiPhys is fine-tuned by the original and generated Face+HR+SpO2 datasets before the estimation. We use MultiPhys and CNN model [8] as the HR and SpO2 estimation methods in our experiment.

While multimodal datasets ($e.g.$, Face+HR+SpO2 dataset) are very insufficient, a single physiological parameter with facial videos or physiological parameters without facial videos can be collected at a relatively low cost. For example, with the recent growth of remote photoplethysmography (rPPG), public Face+HR datasets are available ($e.g.$, LGI-PPGI [15] and UBFC-Phys [13] datasets). We can also utilize public HR+SpO2 datasets ($e.g.$, BIDMC dataset [14]). Therefore, we utilize public Face+HR and HR+SpO2 datasets to generate synthetic Face+HR+SpO2 datasets.

2.1. Multimodal Physiological Generative Model

Multimodal physiological generative model (MultiPhys) is shown in Fig.1. MultiPhys is constructed based on the structure of multimodal variational autoencoder (MVAE) [16]. MVAE assumes that multimodal data are converted into latent variables via inference networks, and multimodal data are reconstructed from latent variables via generative networks. Particularly, MVAE can generate arbitrary modalities from one or a subset of modalities. In MultiPhys, we additionally introduce a condition label representing camera type and states of head motions in the generative process of MVAE. Although a condition label is incorporated into MVAE in the previous study [17], the method requires a full set of modalities as input data. On the other hand, MultiPhys requires only one or a subset of modalities as input data by adopting a product-of-experts inference network [16].

Here, suppose that $\boldsymbol{F} \in \mathbb{R}^{D_{\text{frame}} \times D_F}$ represents 2D color signals from facial video. The row of $\boldsymbol{F}$ consists of red, blue, and green (RGB) signals averaged within six regions-of-interest (ROIs) from facial areas ($i.e.$, $D_F = 3 \times 6$). In the same manner as the previous method [8], we use six ROIs as shown in Fig.1 (red rectangles in Face). The column of $\boldsymbol{F}$ consists of multiple frames ($D_{\text{frame}} = 64$ in our experiment). Next, HR and SpO2 are represented as one-hot vectors $\boldsymbol{h} \in \mathbb{R}^{D_c}$ and $\boldsymbol{s} \in \mathbb{R}^{D_c}$, respectively. Note that D_c is the number of classes (128 classes in our experiment). Specifically, the classes are made by splitting the range of admissible HR (40-125bpm) and SpO2 (90-100%) values into D_c segments of equal size. Since the performance of the classification task was higher than that of the regression task for HR estimation in the previous study [8], we employ the classification task for HR and SpO2 estimation. In addition, the condition label is represented as the one-hot vector $\boldsymbol{l} \in \mathbb{R}^{D_l}$, where D_l is the number of camera types and states of head motions (static or motion). For example, facial video recorded by the second camera with the static state is represented as $\boldsymbol{l} = [0, 1, 0, 0, 1, 0]$, where the first four values represent camera type and the last two values represent motion state (the number of all cameras is four in this case). By introducing the condition label, MultiPhys generates multimodal physiological data while considering camera type and motion state.

MultiPhys assumes the latent variable $\boldsymbol{z} \in \mathbb{R}^{D_z}$, where D_z is the dimension of the latent variable. The prior distribution of the latent variable is assumed as the centered isotropic multivariate Gaussian $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$. Then the generative process of facial color signals $\boldsymbol{F}$, the HR class $\boldsymbol{h}$, and the SpO2 class $\boldsymbol{s}$ is denoted as:

$$\begin{aligned} \boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s} &\sim p_\theta(\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s}|\boldsymbol{l}, \boldsymbol{z}) \\ &= p_{\theta_F}(\boldsymbol{F}|\boldsymbol{l}, \boldsymbol{z}) p_{\theta_h}(\boldsymbol{h}|\boldsymbol{l}, \boldsymbol{z}) p_{\theta_s}(\boldsymbol{s}|\boldsymbol{l}, \boldsymbol{z}), \end{aligned} \tag{1}$$

where θ, θ_F, θ_h, and θ_s are the parameters of generative networks. MultiPhys is trained by maximizing the marginal log-likelihood. However, since the calculation of the marginal log-likelihood is intractable, the evidence lower bound (ELBO) is instead maximized. The ELBO is defined via an approximate posterior distribution (an inference network), $i.e.$, $q_\phi(\boldsymbol{z}|\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s})$ as follows:

$$\begin{aligned} \text{ELBO}(\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{l}) =& \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s})}[p_\theta(\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s}|\boldsymbol{l}, \boldsymbol{z})] \\ &- \text{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s})||p(\boldsymbol{z})], \end{aligned} \tag{2}$$

where $\text{KL}[q||p]$ denotes the Kullback-Leibler divergence between distributions q and p. In Eq.(2), the first term is negative reconstruction errors of $\boldsymbol{F}$, $\boldsymbol{h}$, and $\boldsymbol{s}$, and the second term forces $q_\phi(\boldsymbol{z}|\boldsymbol{F}, \boldsymbol{h}, \boldsymbol{s})$ to be similar to $p(\boldsymbol{z})$. As we assume that the joint posterior can be a product of individual posteriors via a product-of-experts (PoE) [16],

the inference network $q_\phi(\boldsymbol{z}|\boldsymbol{F},\boldsymbol{h},\boldsymbol{s})$ is approximated as:

$$q_\phi(\boldsymbol{z}|\boldsymbol{F},\boldsymbol{h},\boldsymbol{s}) \propto p(\boldsymbol{z})\tilde{q}_{\phi_F}(\boldsymbol{z}|\boldsymbol{F})\tilde{q}_{\phi_h}(\boldsymbol{z}|\boldsymbol{h})\tilde{q}_{\phi_s}(\boldsymbol{z}|\boldsymbol{s}), \quad (3)$$

where ϕ, ϕ_F, ϕ_h, and ϕ_s are the parameters of inference networks. In Eq.(3), $\tilde{q}_{\phi_F}(\boldsymbol{z}|\boldsymbol{F})$ is the underlying inference network described as $q_{\phi_F}(\boldsymbol{z}|\boldsymbol{F}) \equiv \tilde{q}_{\phi_F}(\boldsymbol{z}|\boldsymbol{F})p(\boldsymbol{z})$. To effectively train the individual inference networks, we add ELBO terms of single modalities $\boldsymbol{F}$, $\boldsymbol{h}$, and $\boldsymbol{s}$ to Eq.(2). Therefore, the ELBO can be written as:

$$\lambda_{\text{all}} \cdot \text{ELBO}(\boldsymbol{F},\boldsymbol{h},\boldsymbol{s},\boldsymbol{l}) + \sum_{\boldsymbol{x}\in\{\boldsymbol{F},\boldsymbol{h},\boldsymbol{s}\}} \lambda_x \cdot \text{ELBO}(\boldsymbol{x},\boldsymbol{l}), \quad (4)$$

where

$$\text{ELBO}(\boldsymbol{x},\boldsymbol{l}) = \mathbb{E}_{q_{\phi_x}(\boldsymbol{z}|\boldsymbol{x})}[p_\theta(\boldsymbol{x}|\boldsymbol{l},\boldsymbol{z})] - \text{KL}[q_{\phi_x}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})].$$

In Eq.(4), λ_{all} and λ_x are weight parameters of each ELBO term. By maximizing the ELBO in Eq.(4), we can train both joint-modality and single-modality inference networks.

2.2. Multimodal Physiological Data Generation

The proposed method generates Face+HR+SpO2 datasets from Face+HR and HR+SpO2 datasets in **Approach (i)**. Although generating synthetic facial videos from HR information is carried out recently [18], we newly generate physiological modalities in two directions, $i.e.$, Face+HR $\rightarrow$ SpO2 and HR+SpO2 $\rightarrow$ Face.

We first train MultiPhys by using a small multimodal dataset, Face+HR+SpO2 dataset $\mathcal{D}_0$. After the training, the SpO2 class $\boldsymbol{s}_1$ is estimated from facial color signals $\boldsymbol{F}_1$, the HR class $\boldsymbol{h}_1$, and the condition label $\boldsymbol{l}_1$ in a Face+HR dataset via MultiPhys as follows:

$$\begin{aligned}&\boldsymbol{z}_1 \sim q_{\phi_{Fh}}(\boldsymbol{z}|\boldsymbol{F}_1,\boldsymbol{h}_1) \propto p(\boldsymbol{z})\tilde{q}_{\phi_F}(\boldsymbol{z}|\boldsymbol{F}_1)\tilde{q}_{\phi_h}(\boldsymbol{z}|\boldsymbol{h}_1),\\ &\boldsymbol{s}_1 \sim p_{\theta_s}(\boldsymbol{s}_1|\boldsymbol{l}_1,\boldsymbol{z}_1),\end{aligned} \quad (5)$$

where $q_{\phi_{Fh}}(\boldsymbol{z}|\boldsymbol{F}_1,\boldsymbol{h}_1)$ is the joint posterior that is approximated as a product of individual posteriors via PoE. Furthermore, facial color signals $\boldsymbol{F}_2$ are estimated from the HR class $\boldsymbol{h}_2$, the SpO2 class $\boldsymbol{s}_2$, and the condition label $\boldsymbol{l}_2$ (all values of camera types are zero since cameras are not available) in a HR+SpO2 dataset as follows:

$$\begin{aligned}&\boldsymbol{z}_2 \sim q_{\phi_{hs}}(\boldsymbol{z}|\boldsymbol{h}_2,\boldsymbol{s}_2) \propto p(\boldsymbol{z})\tilde{q}_{\phi_h}(\boldsymbol{z}|\boldsymbol{h}_2)\tilde{q}_{\phi_s}(\boldsymbol{z}|\boldsymbol{s}_2),\\ &\boldsymbol{F}_2 \sim p_{\theta_F}(\boldsymbol{F}_2|\boldsymbol{l}_2,\boldsymbol{z}_2).\end{aligned} \quad (6)$$

From the above Eqs.(5) and (6), we can collect the generated multimodal datasets $\mathcal{D}_1 = \{\boldsymbol{F}_1,\boldsymbol{h}_1,\boldsymbol{s}_1,\boldsymbol{l}_1\}$ and $\mathcal{D}_2 = \{\boldsymbol{F}_2,\boldsymbol{h}_2,\boldsymbol{s}_2,\boldsymbol{l}_2\}$ from Face+HR and HR+SpO2 datasets, respectively.

To adapt ML models to unseen subjects, we leverage pseudo HR that is estimated from unseen subjects' facial videos by an unsupervised signal processing method in **Approach (ii)**. We require only short facial videos (10 seconds for each subject in our experiment). To estimate pseudo HR, we employ one of the most accurate unsupervised HR estimation methods, CHROM [4]. By using CHROM, we obtain the pseudo HR class $\boldsymbol{h}_3$ from facial color signals using a single ROI (full bounding box in ref. [8]) collected from unseen subjects' videos. Then the SpO2 class $\boldsymbol{s}_3$ is estimated from facial color signals of unseen subjects $\boldsymbol{F}_3$, the pseudo HR class $\boldsymbol{h}_3$, and the condition label $\boldsymbol{l}_3$ via MultiPhys in the same manner as Eq.(5). Thus, we can obtain the generated multimodal dataset for unseen subjects $\mathcal{D}_3 = \{\boldsymbol{F}_3,\boldsymbol{h}_3,\boldsymbol{s}_3,\boldsymbol{l}_3\}$. The recent ML-based HR estimation method [12] also uses an unsupervised signal processing method to adapt to unseen subjects in a meta-learning framework. Although the proposed method is related to ref. [12], our work mainly aims to generate multimodal data for the training data of ML models.

After multimodal data generation, the pre-trained MultiPhys is fine-tuned or other ML models are trained by the original and generated multimodal datasets $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$. Finally, the HR class $\boldsymbol{h}_{\text{test}}$ and the SpO2 class $\boldsymbol{s}_{\text{test}}$ are estimated from test facial color signals $\boldsymbol{F}_{\text{test}}$. In the case of MultiPhys, the HR and SpO2 classes are estimated as follows:

$$\begin{aligned}&\boldsymbol{z}_{\text{test}} \sim q_{\phi_F}(\boldsymbol{z}|\boldsymbol{F}_{\text{test}}),\\ &\boldsymbol{h}_{\text{test}} \sim p_{\theta_h}(\boldsymbol{h}_{\text{test}}|\boldsymbol{l}_{\text{test}},\boldsymbol{z}_{\text{test}}), \boldsymbol{s}_{\text{test}} \sim p_{\theta_s}(\boldsymbol{s}_{\text{test}}|\boldsymbol{l}_{\text{test}},\boldsymbol{z}_{\text{test}}),\end{aligned} \quad (7)$$

where $\boldsymbol{l}_{\text{test}}$ is the condition label for test data.

3. EXPERIMENTS

3.1. Datasets

rPPG dataset [8] (Face+HR+SpO2): 52 facial videos of eight subjects were recorded on three cameras in different motion scenarios (static or motion). All video sequences were recorded at 15 frames per second (fps) during 60-80 seconds. Simultaneously, HR and SpO2 values were obtained by a pulse oximeter. In the proposed method, ROIs and the sample shape are defined in the same manner as the method [8]. Each sample is 64-frames fragments of facial color signals ($\approx$ 4.3 seconds per fragment), which are obtained by splitting color signals into overlapping segments with a step of 10 frames. As ground truth data, HR and SpO2 values are given for each sample by averaging the values during 64-frames. We set the first 60%, the next 10%, and the last 30% samples from each video sequence to the training, validation, and test data, respectively.

UBFC-Phys dataset [13] (Face+HR): 56 facial videos of 56 subjects were recorded on a single camera in rest, speech, and arithmetic tasks. Since all video sequences were recorded at 35 fps during 180 seconds, we converted the fps into 15 in order to make the same condition as the rPPG dataset. As ground truth data, we transformed the blood volume pulse recorded in this dataset into HR by using Python framework for Virtual Heart Rate (pyVHR) [19]. The sample shape is defined in the same manner as the rPPG dataset. After randomly selecting videos in the rest task, we constructed two datasets, $i.e.$, **seen-subjects dataset** and **unseen-subjects dataset**. The seen-subjects dataset consists of 20 subjects, where the first 70%, the next 10%, and the last 20% from each video sequence are set to the training, validation, and test data, respectively. The unseen-subjects dataset consists of 10 subjects who are not used in the seen-subjects dataset, where the first 10 seconds and the remaining periods from each video sequence are set to training ($i.e.$, short videos) and test data, respectively. Note that, since subjects' identification numbers are not provided from the rPPG dataset, we construct the unseen-subjects dataset by using the UBFC-Phys dataset. Therefore, the effectiveness of the adaptation to unseen subjects, $i.e.$, **Approach (ii)** is verified for only HR estimation.

BIDMC dataset [14] (HR+SpO2): HR and SpO2 were acquired from 53 subjects during 8 minutes. We used HR and SpO2 of randomly selected 20 subjects, which is the same number of subjects as the seen-subjects dataset. Then we extracted HR and SpO2 for 100 seconds from each subject.

We summarize the rPPG, seen-subjects, unseen-subjects, and BIDMC datasets in Table 1. The samples in the unseen-subjects dataset were augmented five times ($i.e.$, 80$\times$5) since the sample size is very small. The total number of training data $\mathcal{D}_1$+$\mathcal{D}_2$+$\mathcal{D}_3$ ($i.e.$, 6900) on the generated multimodal datasets is more than the number of training data $\mathcal{D}_0$ ($i.e.$, 3834) on the original multimodal dataset.

Table 1. Datasets, modalities, and the numbers of samples in the training (index), validation, and test data.

Dataset	Modality	Train (index) / Val / Test
rPPG	Face+HR+SpO2	3834 ($\mathcal{D}_0$) / 298 / 1956
Seen-subjects	Face+HR	3640 ($\mathcal{D}_1$) / 400 / 940
Unseen-subjects	Face	400 ($\mathcal{D}_3$) / - / 2480
BIDMC	HR+SpO2	2860 ($\mathcal{D}_2$) / - / -

Table 2. HR and SpO2 estimation performance from facial videos on the rPPG, seen-subjects, and unseen-subjects datasets. The best values for each evaluation metric are represented in bold. The performance denoted by "-" is not applicable to the evaluation.

		rPPG dataset						Seen-subjects dataset			Unseen-subjects dataset		
		HR (bpm)			SpO2 (%)			HR (bpm)			HR (bpm)		
Method	Training dataset	MAE	RMSE	ρ	MAE	RMSE	ρ	MAE	RMSE	ρ	MAE	RMSE	ρ
ICA [2]	None	13.0	19.2	0.252	-	-	-	11.8	16.9	0.306	13.9	20.1	0.317
CHROM [4]	None	13.5	20.3	0.256	-	-	-	8.76	15.0	0.421	10.1	16.2	0.567
SpO2 [5, 6]	$\mathcal{D}_0$	-	-	-	0.647	0.901	0.079	-	-	-	-	-	-
CNN [8]	$\mathcal{D}_0$	4.87	7.98	0.849	-	-	-	-	-	-	-	-	-
CNN [8]+**Gen.**	$\mathcal{D}_0$+$\mathcal{D}_1$+$\mathcal{D}_2$	4.57	7.52	0.865	-	-	-	6.45	9.55	0.653	18.4	21.0	-0.211
CNN [8]+**Gen.+Adap.**	$\mathcal{D}_0$+$\mathcal{D}_1$+$\mathcal{D}_2$+$\mathcal{D}_3$	**4.47**	**7.36**	**0.871**	-	-	-	6.11	**9.02**	0.672	8.75	13.3	0.713
MultiPhys	$\mathcal{D}_0$	4.99	8.49	0.828	0.587	0.932	0.326	-	-	-	-	-	-
MultiPhys+Gen.	$\mathcal{D}_0$+$\mathcal{D}_1$+$\mathcal{D}_2$	4.58	7.77	0.853	**0.551**	**0.879**	**0.450**	**5.94**	9.18	**0.677**	15.5	18.5	-0.023
MultiPhys+Gen.+Adap.	$\mathcal{D}_0$+$\mathcal{D}_1$+$\mathcal{D}_2$+$\mathcal{D}_3$	4.51	7.63	0.859	0.552	0.881	0.438	6.18	9.57	0.646	**8.19**	**12.5**	**0.727**

3.2. Settings

The networks of MultiPhys were implemented as shown in Fig.1. All hyperparameters of MultiPhys were selected by using training and validation data. The weight parameters of ELBO terms in Eq.(4) were set to $\lambda_{\text{all}} = 0.5$, $\lambda_F = 0.1$, $\lambda_h = 5.0$, and $\lambda_s = 5.0$. Each layer is followed by ReLU activation [20]. We used an Adam optimizer [21], and both numbers of epochs for the training using $\mathcal{D}_0$ and the following fine-tuning were 100.

We compared our HR estimation performance with signal processing approaches, ICA [2] and CHROM [4] [2]. Our SpO2 estimation performance was also compared with the methods [5, 6] [3]. In the comparative methods [2, 4–6], we used facial color signals from a single ROI (full bounding box in ref. [8]), where the highest estimation performance was achieved. Furthermore, we applied the generated datasets to a CNN model [8] [4] to verify the effectiveness of our approaches. Since the CNN model is used to estimate only HR, we applied the generated Face+HR datasets. We generated the multimodal datasets four times via MultiPhys whose networks were trained with four random seeds, and the performance was averaged across all adaptation of the generated datasets to MultiPhys and the CNN model [8]. The HR and SpO2 estimation performance was evaluated using the mean absolute error (MAE), root mean squared error (RMSE), and correlation coefficients ρ between the estimated values and the corresponding ground truth data.

3.3. Results and Discussion

We show the HR and SpO2 estimation performance in Table 2 and discuss the effectiveness of **Approaches (i) and (ii)**. In the table, the method that uses the generated multimodal physiological data $\mathcal{D}_1$ and $\mathcal{D}_2$ is represented as *Gen.*, and the method that adapts to unseen subjects using $\mathcal{D}_3$ is represented as *Adap.*.

First, by comparing *MultiPhys+Gen.* with *MultiPhys* on the rPPG dataset, we confirm that the multimodal physiological data generation, *i.e.*, **Approach (i)** improves the HR and SpO2 estimation performance. This result suggests that augmenting training data from other datasets is effective for the estimation. In more detail, Fig.2 shows the HR and SpO2 estimation performance on the rPPG dataset when using $\mathcal{D}_1$ or $\mathcal{D}_2$ that is generated from Face+HR or HR+SpO2 datasets, respectively. Compared with the method using only $\mathcal{D}_0$, we can see that both Face+HR and HR+SpO2 datasets are effective for the improvement. Importantly, we observe the effectiveness even when using the dataset without facial videos, *i.e.*, HR+SpO2 dataset. Furthermore, the comparison results of *CNN+Gen.* against *CNN* on the rPPG dataset show the effectiveness of the physiological data generation for other ML models.

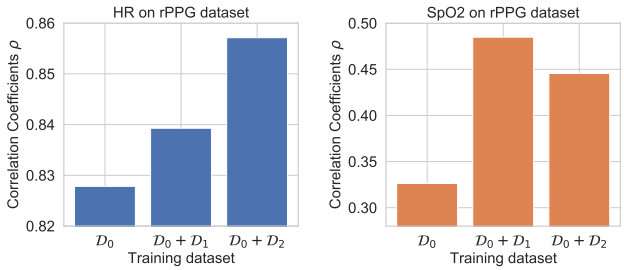

Fig. 2. Correlation coefficients ρ with several training datasets.

By comparing *MultiPhys+Gen.+Adap.* with *MultiPhys+Gen.* on the unseen-subjects dataset, we confirm that the adaptation to unseen subjects, *i.e.*, **Approach (ii)** significantly improves the HR estimation performance for unseen subjects. This result suggests that the utilization of short facial videos and pseudo HR are extremely effective for the estimation. The comparison results of *CNN+Gen.+Adap.* against *CNN+Gen.* also show the effectiveness of adaptation to unseen subjects for other ML models. From the estimation performance on the seen-subjects and unseen-subjects datasets, *MultiPhys+Gen.+Adap.* and *CNN+Gen.+Adap.* can estimate HR of unseen subjects with the same level as that of seen subjects. Furthermore, on the rPPG and seen-subjects datasets, *MultiPhys+Gen.+Adap.* and *CNN+Gen.+Adap.* can still estimate HR or SpO2 as accurately as *MultiPhys+Gen.* and *CNN+Gen.*, or rather the HR estimation performance is higher on the rPPG dataset. This result suggests that the methods that adapt to unseen subjects have no negative effects on the estimation performance of seen subjects.

Overall, ML models along with **Approaches (i) and (ii)** outperform the ML models without these approaches and the previous HR and SpO2 estimation methods [2, 4–6]. We believe that the generated multimodal datasets can be applied to any ML model and improve the estimation performance for multiple physiological parameters.

4. CONCLUSION

This paper has proposed the HR and SpO2 estimation method from facial videos with multimodal physiological data generation. We introduce the following two approaches: (i) generating multimodal physiological datasets from a part of physiological modalities and (ii) adapting ML models to unseen subjects. The experimental results show the effectiveness of the two approaches and the robust estimation performance on several public datasets.

[2] Signal processing pipelines including the preprocessing, signal decomposition, and spectral analysis were implemented by using pyVHR [19].

[3] The coefficients A and B were optimized by using training data $\mathcal{D}_0$.

[4] We used CL+F model whose networks were trained with a random seed. The CL+F model achieves the best performance in ref. [8].

5. REFERENCES

[1] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[2] M. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[3] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity," in *federated conf. computer science and information systems (FedCSIS)*, 2011, pp. 405–410.

[4] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[5] H. Rahman, M. U. Ahmed, and S. Begum, "Non-contact physiological parameters extraction using facial video considering illumination, motion, movement and vibration," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 88–98, 2019.

[6] G. Casalino, G. Castellano, and G. Zaza, "A mhealth solution for contact-less self-monitoring of blood oxygen saturation," in *IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1–7.

[7] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–365.

[8] M. Kopeliovich, Y. Mironenko, and M. Petrushan, "Architectural tricks for deep learning in remote photoplethysmography," in *Proc. the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[9] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.

[10] Z. Yu, W. Peng, X. Li, X. Hong, et al., "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement," in *Proc. the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 151–160.

[11] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *arXiv preprint arXiv:2006.03790*, 2020.

[12] X. Liu, Z. Jiang, J. Fromm, X. Xu, S. Patel, and D. McDuff, "Metaphys: few-shot adaptation for non-contact physiological measurement," in *Proc. the Conference on Health, Inference, and Learning*, 2021, pp. 154–163.

[13] R. Meziatisabour, Y. Benezeth, P. De Oliveira, J. Chappe, et al., "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, 2021.

[14] M. AF. Pimentel, A. EW. Johnson, P. H. Charlton, D. Birrenkott, et al., "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2016.

[15] C. S. Pilz, S. Zaunseder, J. Krajewski, and V. Blazek, "Local group invariance for heart rate estimation from face videos in the wild," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1254–1262.

[16] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," *arXiv preprint arXiv:1802.05335*, 2018.

[17] Y. Akamatsu, K. Maeda, T. Ogawa, and M. Haseyama, "Classification of expert-novice level using eye tracking and motion data via conditional multimodal variational autoencoder," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1360–1364.

[18] Y. Tsou, Y. Lee, and C. Hsu, "Multi-task learning for simultaneous video generation and remote photoplethysmography estimation," in *Proc. the Asian Conference on Computer Vision (ACCV)*, 2020.

[19] G. Boccignone, D. Conte, V. Cuculo, A. D'Amelio, et al., "An open framework for remote-ppg methods and their assessment," *IEEE Access*, vol. 8, pp. 216083–216103, 2020.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. International Conference on Machine Learning (ICML)*, 2010.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.