Remote Respiration Monitoring of Moving Person Using Radio Signals
(Supplementary Material)

Jae-Ho Choi[1], Ki-Bong Kang[1,2], and Kyung-Tae Kim[1]

[1] POSTECH, Republic of Korea
[2] Samsung Electronics, Republic of Korea
{jhchoi93,kkb131,kkt}@postech.ac.kr

In this supplementary material, we provide more detailed descriptions for radio frequency (RF) signal pre-processing and network architecture of the proposed RF-vital model. We also provide detailed explanations on data acquisition and error metrics, and present additional experimental results.

A1 Details for RF Signal and Pre-processing Pipeline

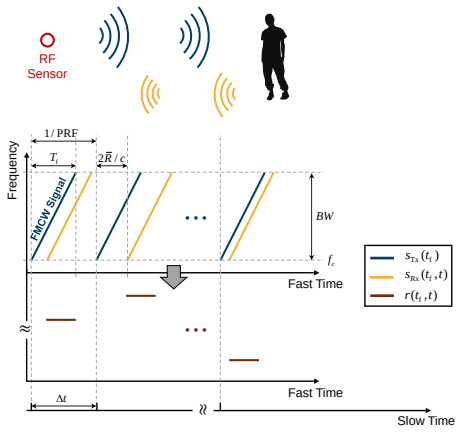


Fig. 1. Reflected FMCW sequences for radar sensor.

Our RF-vital model estimates human respiration based on radio reflections. In this section, we introduce the basic principles of RF signals and our pre-processing pipelines in detail.

Signal Modeling Our method is applicable to any signal data based on RF transmission/reception, but in this work, we utilize a radar sensor whose transmitter and receivers are localized at approximately the same point. The radar periodically transmits a modulated radio signal, then receives the time-delayed

reflections from its surroundings. We adopt a frequency-modulated continuous-wave (FMCW) technique for signal modulation (Fig. 1). The transmitted and received signals for FMCW radar can be expressed as [2,3]:

$$s_{\mathrm{Tx}}(t_{\mathrm{f}}) = \exp\left(j\left(2\pi f_c t_{\mathrm{f}} + \pi\frac{BW}{T_{\mathrm{f}}}t_{\mathrm{f}}^2\right)\right), \tag{1}$$

$$s_{\mathrm{Rx}}(t_{\mathrm{f}}, t) = \sum_i \alpha_i s_{\mathrm{Tx}}\left(t_{\mathrm{f}} - \frac{2\bar{R}_i(t)}{c}\right), \tag{2}$$

where s_{Tx} denotes a transmitted FMCW signal (i.e., waveform with a linearly increasing/decreasing frequency with respect to t_{f}), and s_{Rx} denotes a received signal represented by the linear combination form of time-delayed $s_{\mathrm{Tx}}(t_{\mathrm{f}})$. t_{f} and t refer to the fast time and slow time (Fig. 1), each of which indicates the spatial depth information with respect to the round-trip time-of-flight intervals (i.e., spatial dimension) and sampling time information with respect to the pulse repetition frequency (PRF) of the sensor (i.e., temporal dimension), respectively. f_c, BW, T_{f} and c represent the operating frequency, bandwidth, pulse width, and the speed of light, respectively; α_i is the reflection coefficient of the i-th scatter and $\bar{R}_i$ is the corresponding radial depth. Removing the carrier component through a frequency mixer, the received signal $s_{\mathrm{Rx}}(t_{\mathrm{f}})$ becomes

$$\begin{aligned}
r(t_{\mathrm{f}}, t) &= s_{\mathrm{Tx}}(t_{\mathrm{f}}) s_{\mathrm{Rx}}^*(t_{\mathrm{f}}) \\
&= \sum_i \alpha_i \exp\left(j\left(2\pi f_c t_{\mathrm{f}} + \pi\frac{BW}{T_{\mathrm{f}}}t_{\mathrm{f}}^2\right)\right) \\
&\cdot \exp\left(-j\left(2\pi f_c\left(t_{\mathrm{f}} - \frac{2\bar{R}_i(t)}{c}\right) + \pi\frac{BW}{T_{\mathrm{f}}}\left(t_{\mathrm{f}} - \frac{2\bar{R}_i(t)}{c}\right)^2\right)\right) \\
&= \sum_i \alpha_i \exp\left(j4\pi\frac{\bar{R}_i(t)}{c}\left(f_c + \frac{BW}{T_{\mathrm{f}}}t_{\mathrm{f}}\right) - j4\pi\frac{BW}{T_{\mathrm{f}}}\frac{\bar{R}_i^2(t)}{c^2}\right) \\
&\approx \sum_i \alpha_i \exp\left(j4\pi\frac{\bar{R}_i(t)}{\lambda}\right)\exp\left(j4\pi\frac{BW}{T_{\mathrm{f}}}\frac{\bar{R}_i(t)}{c}t_{\mathrm{f}}\right),
\end{aligned} \tag{3}$$

where $*$ denotes a conjugate operator and λ is the wavelength of the transmitted signal ($\lambda = c/f_c$). Note that $r(t_{\mathrm{f}}, t)$ is a linear summation of monotone signals with fundamental frequencies of $2BW\bar{R}_i(t)/(T_{\mathrm{f}}c)$, where BW, T_{f}, and c are constant over time. Therefore, the radial distance of each object can be estimated through the frequency analysis of $r(t_{\mathrm{f}}, t)$ across the fast time domain t_{f}. Using a fast Fourier transform (FFT) algorithm, $r(t_{\mathrm{f}}, t)$ can be transformed to:

$$\begin{aligned}
h(R, t) &= \mathcal{F}_{t_{\mathrm{f}}}\left\{r(t_{\mathrm{f}}, t)\right\} \\
&= \sum_i \alpha_i \delta\left(f - \frac{2BW}{T_{\mathrm{f}}}\frac{\bar{R}_i(t)}{c}\right)\exp\left(j4\pi\frac{\bar{R}_i(t)}{\lambda}\right) \\
&= \sum_i \alpha_i \delta\left(R - \bar{R}_i(t)\right)\exp\left(j4\pi\frac{\bar{R}_i(t)}{\lambda}\right), \ \left(R = \frac{cT_{\mathrm{f}}}{2BW}f\right)
\end{aligned} \tag{4}$$

where $\mathcal{F}_{t_\mathrm{f}}\{\cdot\}$ denotes the FFT operator along the fast time domain and δ indicates an impulse-like signal envelope. R is a radial depth information from the transmitter that can be estimated based on frequency analysis of $r(t_\mathrm{f}, t)$, and has the resolvability of $\Delta R = c/(2BW)$ by $\Delta f = 1/T_\mathrm{f}$. It should be noted that $h(R, t)$ is represented as a linear combination of peak-like signals dependent upon the radial depth of the surrounding objects, forming a 2D range-time RF heatmap.

Doppler Effect The complex nature of the RF signal allows it to measure not only the radial distance information for the targets of interest, but also the instantaneous changes in the radial distance (i.e., radial velocity) based on the Doppler characteristics. Following Eq. (4), the phase difference between the consecutive RF signals can be represented as

$$\Delta\theta\,(t) = \frac{4\pi\Delta\bar{R}}{\lambda} = \frac{4\pi v \Delta t}{\lambda}, \tag{5}$$

where v refers to the relative velocity in the radial direction between the sensor and each body part of the individual. Namely, the sequential RF pulses in a short time window contain some components that increase or decrease in coincidence with v, and such periodicity converges to a certain Doppler shift $f_\mathrm{Doppler} = 2v/\lambda$ in the frequency domain. Therefore, it is possible to estimate the dominant Doppler shift components by applying a frequency analysis technique such as the FFT or short-time Fourier transform (STFT) on the time-windowed RF sequences, which can directly reflect the instantaneous change in radial distance of each body part. In our model, the received continuous RF signals were transformed into radio joint time-frequency (RJTF) maps using STFT based on a Hann window of 300 ms duration, hop length of 60 ms, and FFT size of 256.

Clutter Suppression Meanwhile, $h(R, t)$ involves not only the reflections from the desired sources (i.e., reflections from human), but also the reflections from clutters such as walls, ceilings, and furniture. Such clutter components do not provide any favorable information for human vital signs while maintaining significantly high electromagnetic reflectance, thereby obscuring large portions of human-induced components. Considering that clutter objects remain stationary with respect to slow time whereas a person tends to have a larger variance, it is possible to suppress the clutter reflections through simple high-pass filtering in the slow time direction [1]. We leverage a mean filter to obtain the final $x(R, t)$ as:

$$x(R, t) = h(R, t) - \sum_{t=t-T_\mathrm{s}+1}^{t} h(R, t), \tag{6}$$

where T_s is the slow-time window length for the mean filter, which is set to 10 s in our experiments.

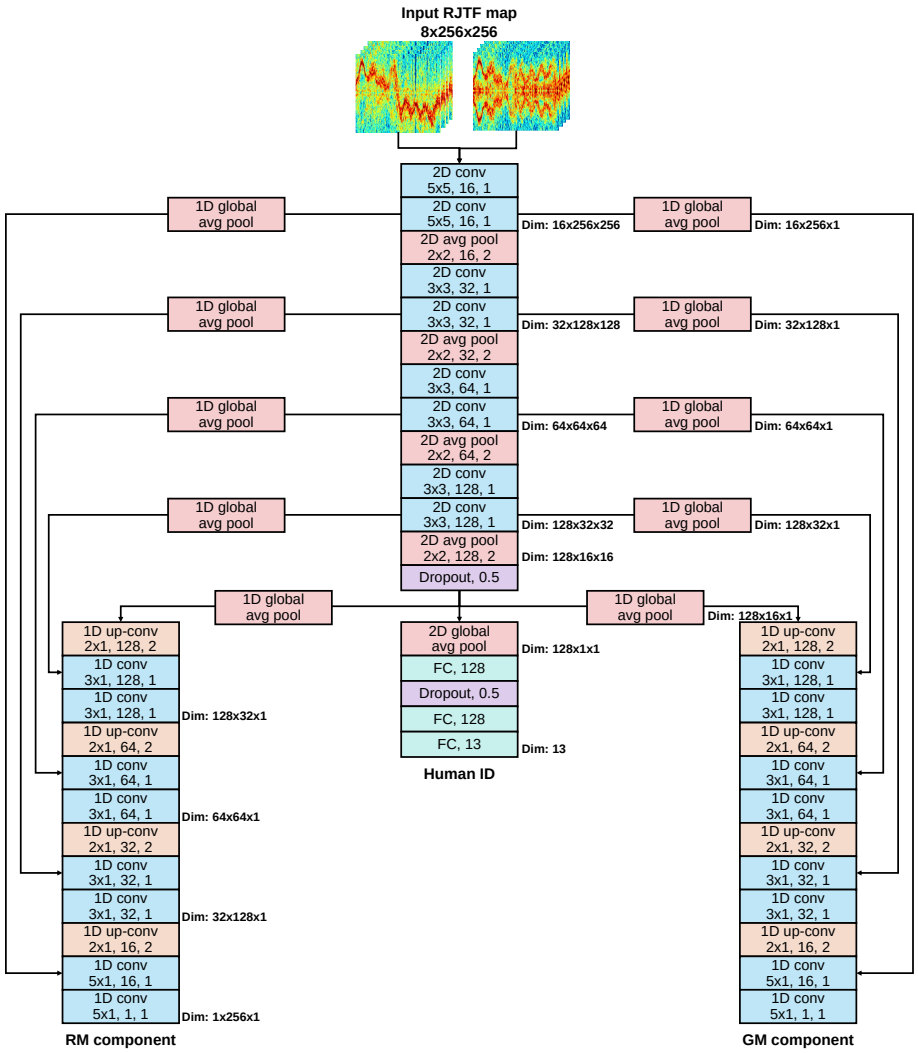


Fig. 2. Detailed network architecture. Each block represents the module type, kernel size, number of output channels, and stride, respectively.

A2 Implementation Details on Network

RF-vital model predicts human respiration based on a 10-s RJTF map formed from the STFT on $\boldsymbol{\alpha}(t) = \{\alpha_m(t)\}_{m=1}^{4}$ and $\exp(j\boldsymbol{\theta}(t)) = \{\exp(j\theta_m(t))\}_{m=1}^{4}$, where $\alpha(t)$ and $\theta(t)$ are the magnitude and phase components of the radio-projected profiles, respectively (Section 4.1). In the experiments, we crop the RJTF map around the frequency band corresponding to the respiratory motion (RM) and global motion (GM) of a person, and then resize all images to 256×256,

resulting in the final network input with a size of 8×256×256. Regarding ground-truth of the network, we utilize respiration signals recorded from the contact chest belt, participant identification (ID), and GM signals obtained from coarse range detections (i.e., magnitude thresholding) on $x(R,t)$ (Section 4.2). The RM and GM ground-truth signals are also resized to 256 length to match the temporal dimension of the final RJTF map, then become normalized between -1 and 1.

Our RF-vital network adopts the U-Net style backbone architecture [6], which is configurated to take the RJTF map of 8×256×256 dimensions as its input and separately predict the 256-length RM and GM signals, and human ID. We leverage 2D and 1D convolutional modules for image encoding and 1D decoding, respectively, and use rectified linear units (ReLU) as layer activation functions. Full details of the RF-vital network are illustrated in Fig. 2.

A3 Details for Data Acquisition

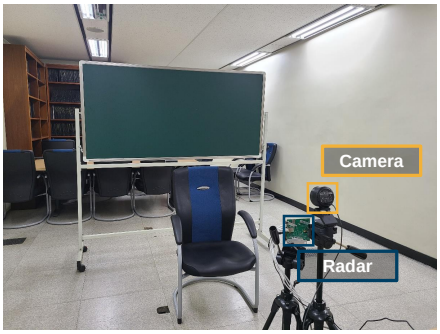


Fig. 3. Experimental environment for RRM-static dataset where a person looks forward while sitting in a chair.

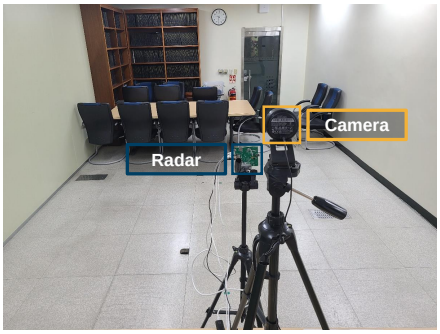


Fig. 4. Experimental environment for RRM-moving dataset where a person is allowed to move around freely.

Because there exists no public dataset for the RF-based non-contact respiration rate measurement (nRRM) task (especially in large motion scenarios), we collected our own dataset in two different conditions, which we refer to as RRM-static and RRM-moving for stationary and moving conditions, respectively.

Specifically, the RRM-static and RRM-moving datasets consist of synchronized FMCW radar echoes (i.e., RF signals), RGB videos, and ground-truth respiration signals (Table 1), each of which was collected in a situation where a person faces forward while sitting in a chair or freely walks around the interior room, respectively. For static cases, we placed a chair about 70 cm away from the radar and camera, and then requested each participant to look straight ahead while holding her/his breath intermittently (Fig. 3). For moving cases,

Sensor	Description
RF sensor	Texas Instruments Inc. IWR1443BOOST radar: FMCW modulation, 77 GHz operating frequency, 1.5 GHz bandwidth, 25 μs pulse width, 1 kHz PRF
Camera	Razer Kiyo Pro webcam: 1280×720 resolution, uncompressed YUY2 format, 30 Hz FPS
Contact sensor	Vernier GDX-RB respiration belt, wireless connection via Bluetooth technology, 10 Hz FPS

Table 1. Specifications of each measurement sensor.

as shown in Fig. 4, we obtained FMCW radar reflections in moving scenarios where each person was allowed to walk around freely except for irregular movements such as a person running or falling, within a space of about 4 m × 5 m (in this case, RGB videos from the camera were leveraged as visual reference). Moreover, some additional samples were collected in more challenging settings such as measurements under no lighting, a person wearing a mask, and a person bowing her/his head, all of which were used only for qualitative evaluation of our RF-vital model.

A4 Error Metrics

To evaluate the nRRM performance of our RF-vital model, we estimate the respiration rates (RRs) of each individual by post-processing the output signals of the network through a band pass filter with a [0.08 Hz, 0.6 Hz] passband range. The predicted RRs are then compared with the ground-truth RR measurements for each 10-s time window. Following [4], we adopt widely utilized error metrics, i.e., mean absolute error (MAE), root mean squared error (RMSE), Pearson's correlation (ρ), and standard deviation (Std), which are described below.

Mean absolute error (MAE):

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\left|RR_i - RR_i'\right|, \tag{7}$$

Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(RR_i - RR_i'\right)^2}, \tag{8}$$

where M is the total number of window samples, RR_i is the ground-truth RR obtained from contact signal, and RR_i' is the estimated RR.

Pearson's correlation coefficient (ρ): can be obtained by computing the normalized covariance between the real RRs for each time instant, i.e., $RR = [RR_1, RR_2, \ldots, RR_M]$, and the estimations $RR' = [RR'_1, RR'_2, \ldots, RR'_M]$.
Standard deviation (Std): is calculated based on the Std of the errors between the real and estimated RRs [5].

A5 Additional Results

A5.1 Qualitative results for different combinations of decoders

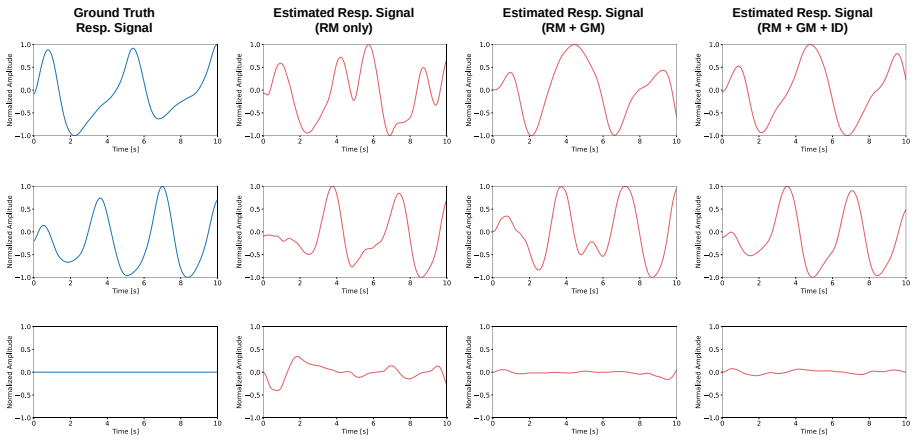


Fig. 5. Qualitative results of the RF-vital network using different decoder combinations. The first column shows the ground-truth respiration signals recorded from the contact sensor. The second and third columns show the predicted respiration signals using only the RM decoder or using RM and GM decoders together, respectively. The fourth column shows the outputs of our RF-vital model leveraging all decoding branches (i.e., RM decoder, GM decoder, and ID discriminator).

Fig. 5 shows the qualitative results under moving conditions for several combinations of decoders in the RF-vital model. We can see that the predicted respiration signals are likely to be contaminated by large GMs when the model is trained without a GM decoder, demonstrating the effectiveness of the adversarial GM decoder in our RF-vital model. Also, the addition of an ID discriminator can further contribute to stable RR predictions.

A5.2 Performance across different STFT windows

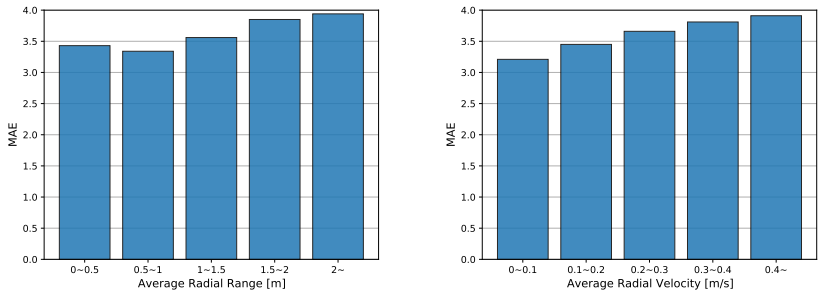
The RF spectrogram provides information of different weights (on time or frequency dimension) depending upon the duration of STFT window. As shown in

STFT Window	MAE↓	RMSE↓	ρ↑	Std↓
100 ms	3.73	6.89	0.32	6.21
300 ms	3.67	7.02	0.32	6.39
500 ms	4.12	7.33	0.29	7.04
700 ms	4.98	8.06	0.24	7.23

Table 2. Performance comparison across different STFT windows.

Table 2, we compared the numerical nRRM performance of our RF-vital model for input RJTF maps with different window durations (hop size was set to 20% of each window size). It can be observed that the smaller window size tends to achieve better performance owing to enhanced resolvability in the temporal dimension. However, significantly reduced window size may limit or rather degrade the performance due to decreased frequency resolution.

A5.3 Detailed Analysis across Radial Distance and Velocity



(a) (b)

Fig. 6. Performance comparison of RF-vital model with respect to the radial range or velocity of each individual.

We further broke down the test results with respect to the radial range or velocity of each individual (Fig. 6). Predictably, as the distance or velocity (instantaneous movement) from the RF sensor increases, the estimation performance tends to deteriorate. Nevertheless, even under some adverse conditions (i.e., low SNR reflections due to far distances or contaminations due to rapid motions), the average MAE loss remains below 4.

References

1. Choi, J.H., Kim, J.E., Kim, K.T.: People counting using IR-UWB radar sensor in a wide area. IEEE Internet Things J. **8**(7), 5806–5821 (2021)
2. Jiang, C., Guo, J., He, Y., Jin, M., Li, S., Liu, Y.: mmVib: Micrometer-level vibration measurement with mmWave radar. In: ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom). pp. 1–13 (2020)
3. Li, J., Stoica, P.: MIMO radar signal processing. John Wiley & Sons, Hoboken, New Jersey, USA (2008)
4. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. In: Adv. Neural Inform. Process. Syst. (NIPS). pp. 1–23 (2020)
5. Niu, X., Han, H., Shan, S., Chen, X.: VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In: Asian Conf. Comput. Vis. (ACCV). pp. 562–576 (2018)
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Int. Conf. Med. Image Comput. Computer-Assist. Interven. (MICCAI). pp. 234–241 (2015)