Multi-frequency Federated Learning for Human Activity Recognition using Head-worn Sensors

Dario Fenoglio[†*], Mohan Li[†*], Davide Casnici[†], [†],
Shkurta Gashi[‡], Silvia Santini[†], Martin Gjoreski[†], Marc Langheinrich[†]
[†] Università della Svizzera italiana, Switzerland
{dario.fenoglio, mohan.li}@usi.ch
[‡] ETH Zurich, Switzerland

Abstract—**Human Activity Recognition (HAR) benefits various application domains, including health and elderly care. Traditional HAR involves Machine Learning (ML) pipelines developed on user data, a process that can be privacy-sensitive in the case of centralized machine learning. This work proposes multi-frequency Federated Learning (FL) to enable: (1) privacy-aware ML; (2) joint ML model learning across devices with varying sampling frequency. We focus on head-worn devices (e.g., earbuds and smart glasses), which are relatively unexplored domains compared to traditional smartwatch or smartphone-based HAR. The results have shown improvements compared to frequency-specific approaches on two datasets, indicating a promising future trend in the multi-frequency FL-HAR task.**

Index Terms—**Federated Learning, Human Activity Recognition (HAR), Head-worn sensors, Earables, Glasses**

I. Introduction

Human Activity Recognition (HAR) refers to the process of identifying and categorizing the specific activities performed by an individual through the analysis of various sensor data. This technology has rapidly emerged as an essential tool with wide-reaching applications across numerous domains in recent decades. The importance of HAR lies in its ability to enable the provision of helpful context information that can be applied in various fields. Among these applications are the management of chronic diseases, where HAR can track and monitor patients' physical activities to provide tailored healthcare solutions and interventions [1]. Similarly, in healthcare settings, detecting abnormal patient behavior can be automated through HAR, thus enhancing the efficiency of care and possibly even saving lives [2]. Furthermore, HAR offers insights into individual habits and routines, from personalized fitness tracking to occupational health and safety, allowing for the design of personalized programs to enhance overall well-being [3]. In summary, HAR not only adds a technological edge to many industries but also brings a humancentric approach to monitoring and understanding behavior. Its applications are expanding, and its impact on daily life is profound, shaping a new era of personalized, context-aware services and care.

Being a crucial component in the realm of ubiquitous computing, HAR demands continuous and seamless integration of computing into everyday life as the main goal. Different approaches have been used in HAR, namely vision-based and sensor-based. Vision-based approaches utilize external sensors like cameras that provide a powerful way to recognize and analyze human activities by capturing visual data. However, they suffer from significant drawbacks. When users are out of the sensors' sensing field, the efficacy of these tools is compromised. Privacy invasion is an even more critical concern, making them potentially unsuitable in personal or sensitive environments [4]. These challenges underscore the limitations of vision-based HAR, particularly in the context of ubiquitous computing, where seamless and unobtrusive integration is a key consideration.

In contrast, sensor-based HAR offers a more flexible and privacy-respecting solution. Both wearable and non-wearable sensors have been deployed, but wearable sensors have become prominent for their ability to overcome the limitations of external sensors, such as their poor portability. Wearable sensors, including magnetometers, gyroscopes, and accelerometers — integral components of inertial measurement units — play a vital role in HAR. Being compact and easily integrated into everyday devices like earbuds or glasses, these sensors align seamlessly with the principles of ubiquitous computing. Their unobtrusive nature allows for continuous monitoring and data collection, permitting a more intuitive and user-centered approach to activity recognition [5].

Unfortunately, the introduction of machine learning, specifically deep learning (DL) algorithms, still complicates privacy preservation as these algorithms require data for training. Even when data are collected in a more privacy-preserving manner, transmitting it to a centralized server for model training cancels users' exclusive ownership of the latter, opening up possibilities for data misuse and inadvertent exposure.

In response to these challenges, Google introduced Federated Learning (FL) [6] in 2016, an innovative ML paradigm that enables neural network models to be trained across multiple decentralized devices or servers, each possessing its local data samples. This approach significantly enhances data privacy and security by ensuring user data remains on their device, presenting a promising equilibrium between advancing HAR capabilities and maintaining robust user privacy protections.

An interesting aspect in FL is the device heterogeneity, i.e. heterogeneous (different) devices can collaboratively train a model with a common goal (e.g., HAR). Recent HAR studies

[*] Equal contribution

have explored multi-device FL in synchronized setups, i.e., if a user wears earbuds and smart glasses simultaneously, FLAME could be used to train joint models across the synchronized devices [7].

Differently from the existing work, this study explores multi-frequency FL in asynchronous setups, i.e., different users can have different active sensors (e.g., combinations of magnetometer, gyroscope, and accelerometer); these sensors can sample at different frequencies (e.g., some devices may use 5Hz and others may use 40Hz); and, these devices — despite utilizing varying sensors and sensor frequencies — can collaboratively build a joint HAR model. Furthermore, we focus on FL for head-worn wearable devices, a relatively unexplored domain compared to traditional smartwatch or smartphone-based HAR.

To this end, this work has made the following contribution: *We propose a novel multi-frequency FL method for HAR.* The method is based on an existing end-to-end learning approach, Spectro-Temporal Residual Network (STResNet) [8], that we adapted to work in a federated and multi-frequency setup. We compared the novel method on two datasets against centralized and frequency-specific models. The results show that our multi-frequency model allows the exploitation of all available data (i.e., all clients and sensors), thus outperforming frequency-specific models. In addition, our model demonstrated high flexibility and robustness, maintaining high performance while accepting a variable number of input sensors.

The rest of the paper is organized as follows: We introduce related works in FL and HAR in Section II. Section III gives details of the two datasets we used. Sections IV and V present our methods and accordingly the experiments with results. In Sections VI and VII, we give a discussion and conclusion on the results with expectations for future research.

II. Related Work

This section provides an overview of the related areas. Two main differences between the existing work and ours are: (1) we focus on FL for head-worn wearable devices, a relatively unexplored domain compared to traditional smartwatch or smartphone-based HAR; (2) to the best of our knowledge, this is the first study that explores multi-frequency FL method for head-worn HAR in an asynchronous setup. The closest method to ours is FLAME, which utilizes the synchronization (time alignment) across devices from the same user. Thus, FLAME is useful for scenarios where users simultaneously use multiple devices. On the other hand, in our proposed method, users need at least one of the multiple devices (or sensors) to participate in the FL process.

A. Federated Learning

The FL community has been growing fast since it was introduced. As concluded from recent surveys [9]–[12], most of the research advances focus on core challenges, including reducing computing costs, tackling system or statistic heterogeneity, and enhancing privacy protection. Optimized communication and aggregation strategies [13]–[16] have been proposed to relieve the computational burden without hurting the overall performance. Considering the participation of heterogeneous hardware, efforts have been made with adaptive task-assigning based on device capability, dropout with incapable devices, or tolerance as a more friendly approach [10]. To handle the typical non-independent and non-identically distributed (non-IID) and unbalanced local data, we may resort to personalization through clustering [17], model adaptation [18]–[20], or just data sharing [21]. Even though the data are locally preserved, the vanilla model sharing and aggregation are exposed to malicious attacks such as data poisoning. Many related advanced works have followed Secure Aggregation [22] and Differential Privacy [23] as reliable solutions. Besides HAR, FL has been proven successful in many other fields such as the Internet of Things [24], [25], healthcare [20], [26], [27], vehicular systems [28], and recommender system [29].

B. Human Activity Recognition

Surveys [30], [31] have well captured recent advances in the HAR task with various sensing modalities. Frequently used signals include inertia, electrocardiogram (ECG), and vital signs such as inspiration and temperature. In addition to general HAR task [32]–[34], other targets such as hand gesture recognition [35] or fall detection [36] are also highly related. As indicated from [37], the unexploited unlabeled data have gained much recent attention. A large amount of data remains at the edges and is not applied for model training because labeling them is an overwhelming and knowledge-demanding task. More researchers now try to incorporate them in a semi-supervised or unsupervised manner to achieve better performance [38], [39].

C. Federated Learning for Human Activity Recognition

Even though current smart devices can collect billions of sensor data every day with great potential to improve HAR performance, the cost of data transmission and violation of individual privacy are barely possible to handle. Konstantin et al. [40] have made one of the earliest contributions that deploy the HAR learning task with the FL framework to tackle privacy issues. Tu et al. [41] proposed FedDL, where the center HAR model merges local updates based on a dynamic sharing scheme to speed up the convergence while maintaining high accuracy. Ouyang et al. [42] introduced ClusterFL, a similarity-aware FL approach to cluster different clients in a multitasking manner to achieve high model accuracy and low communication overhead for HAR applications. Xiao et al. [43] develop advanced feature extraction approaches from sensor data to improve the overall performance. Unsupervised learning and personalization have also been proven powerful as future directions [44], [45].

Despite the great success of FL on HAR, few of them have explored wearable device data collected with earbuds and glasses, or in a multi-frequency setup. A recent-to-date work [46] practiced leveraging wearable smart glasses data to achieve personalized treatments and interventions for enhanced healthcare outcomes. Following their promising results, we explore multi-frequency FL for HAR.

III. Datasets

Our head-worn dataset consists of the *USI-HEAR Dataset* [47] and the *OCOsense Smart Glasses HAR Dataset* [46]. An overview is summarized in Table I.

TABLE I
Summary of Datasets.

Dataset	***USI-HEAR***	***OCOsense***
Participations	30	24
Device	eSense earbuds	OCOsense Smart Glasses
Sensors	3-axis accelerometer 3-axis gyroscope	3-axis accelerometer 3-axis gyroscope 3-axis magnetometer pressure sensor 3-axis Euler virtual sensor
Activities	Speak and Walk Head Shaking Speaking Nodding Eating Walking Staying	Sitting Standing Laying Walking Transition Jogging Stair Climbing

A. USI-HEAR Dataset

The USI-HEAR Dataset was collected with the eSense earbuds developed by Nokia Bell Labs [48]. These earbuds consist of two Bluetooth-enabled units, each equipped with a microphone, and the left unit further houses a 6-axis Inertial Measurement Unit (IMU) sensor, comprising a 3-axis accelerometer and a 3-axis gyroscope.

30 participants were provided with the left earbud containing the IMU and performed seven scripted activities, each lasting 3 minutes, with the data subsequently transferred to the experimenter's laptop for verification. The experiment involved seven distinct activities, each carefully selected to represent a range of non-interacting and interacting behaviors. These activities were:

Speak and Walk: Participants combined walking and speaking, illustrating the complexity of simultaneous activities.

Head Shaking: Participants moved their heads horizontally, again with different intensities and intervals, representing a gesture of disagreement or denial.

Speaking: Participants engaged in verbal communication with the experimenters, reflecting natural speech patterns and intonations.

Nodding: Participants were instructed to nod their heads with different intensities and intervals, simulating a common gesture of agreement or acknowledgment.

Eating: Participants consumed food, allowing for the observation of jaw movements and related motions.

Walking: Participants walked at a comfortable pace, capturing the dynamics of regular locomotion.

Staying: Participants remained still or seated, providing a baseline for motion detection.

In total, this dataset comprehends more than 10 hours of streaming data for each single channel of both the gyroscope and accelerometer, with a universal down-sampled frequency at 50Hz.

B. OCOsense Smart Glasses HAR Dataset

The dataset was collected in 2022 by Emteq Labs using their *OCOsense Smart Glasses* [49]. The device is equipped with one 3-axis accelerometer, one 3-axis gyroscope, one 3-axis magnetometer, one pressure sensor (barometer), and one 3-axis Euler virtual sensor to combine data from the accelerometer and gyroscope to provide the orientation of the glasses in three dimensions (yaw, pitch, roll).

24 participants were asked to perform the following activities while wearing the smart glasses:

Sitting (39.3%)—Includes *Sitting*, *Sitting Still*, *Sitting-looking around*, *Sitting-using a PC*, and *Sitting-using a phone*.

Standing (27.3%)—Includes *Standing*, *Standing Still*, *Standing-looking around*, and *Standing-using a phone*.

Laying (18.3%)—Includes *On the back*, *On the left side*, *On the right side*, and *On the stomach*.

Walking (9.1%)—Includes *Walking*, *Walking-looking around*, and *Walking-using a phone*.

Transition (2.2%)—Includes *Sitting down* and *Standing up*.

Jogging (1.7%)—Includes *Jogging*.

Stair climbing (1.7%)—Includes *Stair climbing*.

We have 1.7M samples (9.5 hours duration in total) approximately equally distributed among all participants labeled by themselves. The sampling frequency coordinates with the earbuds dataset at 50Hz.

IV. Methods

In this section, we describe our centralized training comparison between a few pipelines to choose the final model for FL, in addition to the setup of FL and multi-frequency network.

A. Centralized Machine Learning

The DL pipeline employed in this study encompasses five distinct neural network architectures. Among these, four are 1D convolutional neural networks (ConvNets), and the remaining model is a deep multimodal spectro-temporal residual neural network (STResNet) [50].

TABLE II
Main DL 1D Convolutional Neural Networks Architecture's Features.

Model	**# Conv.**	**# Dense**	**Act. functions**	**# params.**
ConvNet1	3	4	LeakyReLU	78,680,443
ConvNet2	3	3	ReLU	5,962,453
ConvNet3	2	2	ReLU	1,937,575
ConvNet4	2	2	PReLU/ReLU	1,909,031

The four 1D-convolutional deep models are characterized by different configurations of convolutional and dense layers, along with specific activation functions. The architectures are summarized in Table II, and they share common features such as the softmax function as the final activation function and sparse categorical cross-entropy as the loss function. To mitigate overfitting, L2 regularization (rate = 0.0001) and dropout (rate = 0.5) were employed. Additionally, early stopping was implemented using the validation loss as the stopping criterion, further safeguarding against overfitting.

The STResNet model builds upon the concept of end-to-end unimodal time-series classification using residual networks. It incorporates multimodal and spectro-temporal information fusion, essential components for a successful HAR system. STResNet extracts channel-specific spectro-temporal information for each sensor channel. The spectral information is obtained by calculating the logarithm of the amplitude spectrogram in decibels for each input window. The temporal representation is extracted by residual blocks containing CNN layers with 1-dimensional (1D) filters. The shortcut connections in the residual blocks combat the gradient vanishing problem, making training more tractable. L2 regularization and dropout were applied to the dense layers, and the final output is provided by a softmax layer, representing class probability for each of the seven activities. Among the models finally employed in this study, STResNet stands out as the second most computationally demanding model, with a total of 14,005,415 trainable parameters. This substantial complexity is indicative of the model's capacity to capture intricate patterns and relationships within the data. However, it is second only to ConvNet1 in terms of computational demand, reflecting a careful balance between model complexity and computational efficiency within the overall DL pipeline.

B. Federated Learning

In our study, we implemented the Weighted Federated Averaging algorithm via the Flower library [51]. Each participating client performed one training epoch on their local dataset and then forwarded their model weights and the count of their training samples to the server. To balance the contributions from all clients, the server applied a weighted average to these weights. We chose this approach due to the inhomogeneous distribution of data across our datasets, resulting in varying numbers of samples per client. The model training was conducted using a sparse categorical cross-entropy loss function, combined with an Adam optimizer set at a learning rate of 0.0001. Considering the average number of samples per client, we standardized the batch size at 32 samples for both datasets. The final model was chosen based on its highest accuracy on the validation set during training.

For our FL environment, we exclusively employed the STResNet model, owing to its superior performance in predicting activity on the USI-HEAR dataset. To ensure a precise comparison between centralized and federated learning, STResNet was implemented under both these settings across the OCOsense and USI-HEAR datasets. Additionally, to evaluate the robustness of FL under different conditions, we tested it initially using all available clients in the datasets, followed by a gradual reduction in the size of the training set. This allowed us to assess the impact of dataset size on the performance and effectiveness of the FL.

Specifically, to validate both centralized and federated using the same test set clients, we employed a person-independent 5-fold cross-validation. This involved dividing the dataset into five distinct groups of clients, ensuring each client's data was excluded from the training process once. In each fold, clients not involved in the training were evenly split into validation and test sets. This strategy ensured a comprehensive and unbiased evaluation of both centralized and federated learning approaches.

C. Multi-Frequency STResNet

To address the challenge of clients equipped with sensors operating at varying frequencies, we developed a novel Multi-frequency STResNet model. This model processes input signals from sensors with different sampling frequencies. For instance, we simulated a scenario in which half of the clients' sensors operated in low battery mode at 5Hz, while the other half functioned at 40Hz. Instead of employing separate models for each battery mode—which would reduce the dataset size—we proposed a versatile model that expands the potential data and client pool. Our approach involves creating distinct spectral-temporal encoders for each sensor (or channel), tailoring both temporal and spectral feature extraction to the sensor's sampling frequency.

As depicted in Figure 1, our encoder receives a raw sensor signal as input and processes it through parallel pathways. In the temporal pathway, after an initial batch normalization step, the signal undergoes through four residual blocks. Each block consists of two 1D convolutions and two Leaky ReLU activations, followed by a single 1D max pooling operation to capture the temporal dynamics effectively. Concurrently, in the spectral pathway, the signal is transformed into a spectrogram and then processed through three blocks, each containing a 2D convolution paired with a Leaky ReLU activation. To integrate the multidimensional spectral features, a dense layer and dropout regularization are employed.

Additionally, we introduced a context vector to mask activations from sensors that are not present (Figure 2). This vector assigns a value of 1 where sensor input is available and 0 otherwise. Correspondingly, input channels are set to zero in the absence of sensor data. Prior to the fully connected layers, the context vector is utilized to zero out activations from absent sensors, thereby preventing consideration of non-zero values due to the bias values involved in the training of the neural networks.

Figure 2 illustrates how our multi-frequency STResNet accommodates users with sensors in both low and full battery modes. We evaluated our model in a centralized environment using person-independent 10-fold Monte Carlo cross-validation, ensuring consistent training and testing on the same client groups for each iteration. We benchmarked our model against those trained exclusively on clients with 5Hz or 40Hz signals, a model trained on all clients at 5Hz (including downsampling those at 40Hz), and an ideal scenario where all

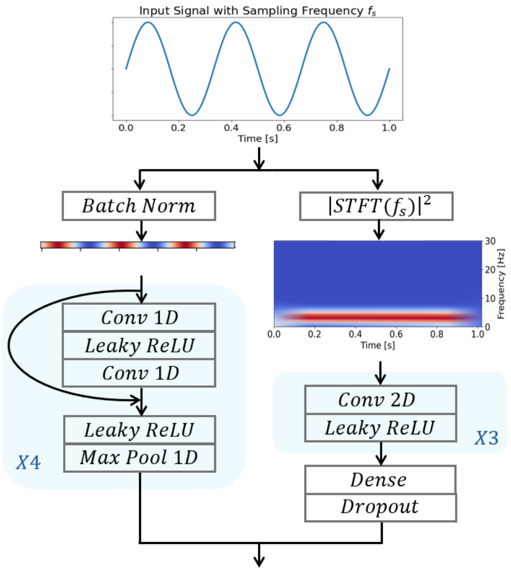

Fig. 1. Encoder Architecture for a Single Input Channel.

clients operate at 40Hz (i.e., no low battery mode). Furthermore, as our model is capable of being evaluated on test clients at both 5Hz and 40Hz, assessments were conducted under each condition to facilitate a fair comparison with frequency-specific models.

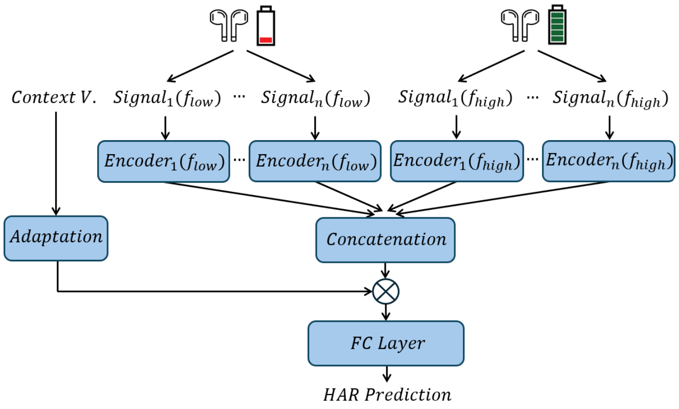
Context V.
$Signal_1(f_{low})$ ··· $Signal_n(f_{low})$
$Signal_1(f_{high})$ ··· $Signal_n(f_{high})$
$Encoder_1(f_{low})$ ··· $Encoder_n(f_{low})$
$Encoder_1(f_{high})$ ··· $Encoder_n(f_{high})$
Adaptation
Concatenation
FC Layer
HAR Prediction

Fig. 2. Multi-Frequency Model Architecture to Handle Low and Full Battery Mode Devices.

V. Experiments and Results

This section outlines the experiments conducted and the corresponding results. In Subsection V-A, we evaluated five DL models across various sensor streams, identifying the STResNet model as the most accurate. Subsection V-B illustrates the comparable accuracy of federated and centralized learning, emphasizing FL's scalability and adaptability, even with varying numbers of training clients, while still ensuring user privacy. Finally, Subsection V-C presents the results for the novel approach, demonstrating the multi-frequency STResNet model's efficiency in handling different sensor frequencies, outperforming frequency-specific setups. It should be noted that all models in our experiments were person-independent, ensuring non-overlapping training and testing client groups, which is crucial for the generalizability and applicability of our findings in real-world scenarios.

A. Model Selection with Centralized Training

We first carry out a centralized training with all five DL models to select the one with the best performance for the FL setup. We use different sensor streams for comprehensive validation: accelerometer (ACC), gyroscope (GYR), magnitude (MAG), first-order derivatives (DER), and all combined (ALL). The results are shown in Table III with accuracy scores expressed as percentages, and standard deviations as percentage points.

TABLE III
Training F1-Scores of DL Models with Standard Deviation

Model	ACC	GYR	MAG	DER	ALL
ConvNet1	38.95 (17.29)	50.77 (15.53)	52.89 (13.97)	61.74 (17.52)	57.12 (13.50)
ConvNet2	**42.91** **(14.94)**	56.02 (14.54)	55.12 (13.83)	62.23 (17.17)	65.49 (12.11)
ConvNet3	34.86 (14.05)	45.64 (17.74)	**58.32** **(13.05)**	56.03 (15.32)	61.78 (11.43)
ConvNet4	20.01 (5.63)	43.16 (18.75)	36.23 (10.07)	38.33 (16.31)	40.55 (13.25)
STResNet	38.86 (16.31)	**57.13** **(13.31)**	57.12 (14.74)	**62.55** **(11.91)**	**69.22** **(11.78)**

The results elucidate that among the original sensor streams, the gyroscope emerges as the most pertinent. When considering virtual sensor streams, the derivatives stand out as a better input for most models compared to the magnitudes. The highest performance is attained with the STResNet model, taking advantage of the entirety of the sensor streams and exploiting the spectral information of the signals, underscoring the importance of a comprehensive approach in sensor data analysis. This consistent performance across different scenarios justifies our choice of STResNet for the rest of the experiment setup.

B. Comparison between Centralized and Federated

Table IV presents the accuracy, F1-score, and cross-entropy loss for both centralized and federated learning across the USI-HEAR and OCOsense datasets. These results underscore the efficacy of FL in handling diverse HAR datasets. Notably, the performance metrics in the FL setup were comparable to those in the centralized setup. This equivalence highlights FL's ability to effectively train a global model without the need for direct data sharing from clients.

Furthermore, Table V demonstrates the robustness of FL when the number of training clients is reduced. The similarity in results between centralized and federated methods, even under these conditions, indicates a high degree of scalability and adaptability in the FL approach. This suggests that FL can maintain reliable performance despite variations in the size of training data and the number of clients, crucial in real-world applications where client availability can vary.

TABLE IV
COMPARISON BETWEEN CENTRALIZED AND FEDERATED LEARNING

	Centralized		Federated	
	USI-HEAR	**OCOsense**	**USI-HEAR**	**OCOsense**
Accur.	**70.0 ± 4.7**	84.9 ± 2.7	69.43 ± 4.14	**85.19 ± 1.99**
F1	**70.2 ± 4.8**	**87.8 ± 1.7**	67.26 ± 4.02	87.72 ± 1.72
CE loss	**1.052 ± 0.165**	**0.389 ± 0.083**	1.105 ± 0.242	0.444 ± 0.085

TABLE V
COMPARISON OF ACCURACY BETWEEN CENTRALIZED AND FEDERATED LEARNING UNDER VARIOUS PARTICIPATION SIZE

#User	Centralized		Federated	
	USI-HEAR	**OCOsense**	**USI-HEAR**	**OCOsense**
1	**50.68 ± 7.53**	**70.22 ± 2.90**	50.68 ± 7.53	70.22 ± 2.90
2	**55.87 ± 1.37**	**73.72 ± 2.39**	55.60 ± 2.43	73.06 ± 2.73
3	**60.30 ± 6.56**	76.73 ± 3.61	58.78 ± 4.04	**77.00 ± 3.21**
4	59.69 ± 4.83	78.37 ± 2.04	**60.12 ± 5.31**	**78.64 ± 1.95**
6	60.04 ± 7.16	80.92 ± 2.12	**62.09 ± 5.61**	**82.13 ± 2.27**
8	63.02 ± 4.91	81.75 ± 2.63	**63.56 ± 6.32**	**83.41 ± 1.83**

C. Multi-frequency Model

Table VI provides a comprehensive comparison of F1-scores and the number of training users for the USI-HEAR and OCOsense datasets across various frequency settings. This table delineates the performance of our multi-frequency STResNet model against various configurations: exclusively 5Hz clients, all clients down-sampled to 5Hz (Down-5Hz), exclusively 40Hz clients, and an ideal scenario with all clients at 40Hz (Ideal 40Hz). To ensure a fair comparison, our multi-frequency model, capable of processing both 5Hz and 40Hz frequencies, was tested under both these conditions (Multi-5Hz and Multi-40Hz) on the same test clients. The column ”Users” presents the number of training users (clients) available for the specific scenario.

Notably, our multi-frequency model exhibited superior F1-scores in both Multi-5Hz and Multi-40Hz configurations across both datasets, demonstrating its effectiveness over single-frequency settings (5Hz and 40Hz clients). This improvement is mainly attributed to the model’s ability to utilize all available original data for training a unified model. However, it should also be noted that the ”Down-5Hz” model, i.e., an approach that first downsamples all the data to the lowest joint frequency (5Hz in this case) and then trains a model, slightly outperforms the multi-frequency approach.

Precisely, in the USI-HEAR dataset, the multi-frequency model achieved F1-scores of 63.26% ± 3.08% and 65.53% ± 3.61% for 5Hz and 40Hz, respectively. These scores significantly surpass the single-frequency models’ scores of 59.30% ± 3.33% (5Hz) and 62.57% ± 4.49% (40Hz), and they closely align with the ideal outcome where all clients operate at a sampling frequency of 40Hz.

In the case of the OCOsense dataset, a similar pattern emerges. Notably, as shown in Table VI, down-sampling to 5Hz resulted in a higher F1-score (86.17% ± 2.29%) than the ideal scenario of all 40Hz clients (85.77% ± 2.17%), suggesting that a 5Hz sampling rate is sufficient for accurate HAR tasks. Our multi-frequency model also performed better on the same test clients when down-sampled to 5Hz compared to 40Hz. Once again, the multi-frequency model outperformed single-frequency setups, underscoring the versatility of our model in effectively handling diverse sensor frequencies.

TABLE VI
COMPARISON OF F1-SCORE AND TRAINING USERS FOR USI-HEAR AND OCOSENSE

Freq.	USI-HEAR		OCOsense	
	F1-score	**Users**	**F1-score**	**Users**
5Hz	59.30 ± 3.33	7 (5Hz)	72.81 ± 14.17	5 (5Hz)
Down-5Hz	65.38 ± 2.39	14 (5Hz)	86.17 ± 2.29	10 (5Hz)
40Hz	62.57 ± 4.49	7 (40Hz)	79.74 ± 2.66	5 (40Hz)

Multi-5Hz	63.26 ± 3.08	7(5Hz)
		7(40Hz)
			85.38 ± 2.52	5(5Hz)
				5(40Hz)

Multi-40Hz	65.53 ± 3.61	7(5Hz)
		7(40Hz)
			83.45 ± 1.99	5(5Hz)
				5(40Hz)

Ideal 40Hz	69.14 ± 2.96	14 (40Hz)	85.77 ± 2.17	10 (40Hz)

VI. DISCUSSION

We discuss the experiment results and our findings in this section, respectively, on the innovations in HAR modalities (Subsection VI-A), insights into multi-frequency in HAR tasks (Subsection VI-B), and expectations for future research (Subsection VI-C).

A. Novel Modalities in HAR

Compared to traditional visual-based HAR, which demands image or video data from cameras, sensor-based HAR improves device availability and reduces data size with sensor streams while preserving competitive performance. As the main raw data source, inertia information is collected with IMUs, which can be fused with different devices and accessories. In this paper, we looked into datasets from earbuds and glasses, two relatively underexplored domains compared to other works with mobile and wrist-worn devices.

It should be noted that the on-body locations of sensors could potentially introduce divergence in results: Activities with delicate changes on the face or head, such as from speaking to eating, are more distinguishable with sensor data above the neck. Likewise, activities with limb motion, such as running and walking, could be more clearly shown with sensors on the wrist or from the waist down. Predictably, a more systematic and comprehensive HAR framework should have multiple modalities to fully cover the range of human activities, and our work could be a good start to the concentration on more delicate differentiation of facial and head activities.

B. Multi-frequency HAR in The Wild

Data frequency has a direct impact on the size of the data stream and storage. From the perspective of an end-user or client, an acceptable performance with minimum data frequency is desired with a generally faster response, higher device refresh rate, and less running-time memory taken. From the perspective of a server or aggregator, accommodation for a wider range of data frequency signifies a larger size of the training dataset, participation from more diverse users, and thus better performance for all. Furthermore, multi-frequency tolerance in FL introduces great possibilities for communication cost reduction and heterogeneous system collaboration. We hope this work may pave the way for further research on this track.

In terms of multi-frequency in HAR tasks, we noticed that the ”Down-5Hz” model, i.e., an approach that first downsamples all the data to the lowest joint frequency (5Hz in this case) and then trains a model, has competitive results with the multi-frequency setup. This indicates a possible frequency threshold in the specific experimental datasets, where information introduced by higher frequency is redundant. However, we expect that in real-life applications where the activities to be recognized are more dynamic, a 5Hz sampling rate would not be sufficient to achieve good HAR performance. Thus, we hope our work inspires more flexible, lightweight, and energy-friendly frameworks.

C. Challenges and The Future

Besides the future research directions above, there are more opportunities and challenges in the FL-HAR field. For this paper, we hope to address two important aspects in later works. First, we did not leverage the enormous unlabeled data in this task. While millions of data streams appear in various sensor devices, end-users usually do not hold clean and well-labeled data. In our next step, we will explore unsupervised or semi-supervised methods to further improve the model quality and label efficency. Second, our work did not apply any personalization approach, such as local re-training, which could be a solution to address the notorious non-IID problem in data heterogeneity. In the future, we may deploy a clustering or privacy-friendly knowledge-sharing approach to achieve better performance.

VII. Conclusion

This paper has introduced a multi-frequency FL framework for the HAR task, which builds one united model to accommodate sensor data from different frequencies and classify human activities under various device status, such as low battery mode. We tested our framework with simulated 5Hz and 40Hz data streams from one earbuds dataset and one smartglasses dataset, with promising results showing a successful frequency fusion approach with heterogeneous sensors in recognizing human activities. We hope our work may inspire more attention to the data frequency issue in HAR and encourage more research on heterogeneity-friendly FL systems.

Acknowledgment

This study was funded by the projects TRUST-ME (205121L_214991), SmartCHANGE (GA No. 101080965), and XAI-PAC (PZ00P2_216405).

References

[1] Rex Liu, Albara Ah Ramli, Huanle Zhang, Erik Henricson, and Xin Liu. An overview of human activity recognition using wearable sensors: Healthcare and artificial intelligence. In Lecture Notes in Computer Science, volume 12993. Springer, 2022.

[2] F Serpush, MB Menhaj, B Masoumi, and B Karasfi. Wearable sensor-based human activity recognition in the smart healthcare system. Computational Intelligence and Neuroscience, 2022.

[3] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1797–1806. ACM, 2008.

[4] M. S. Ryoo and C. Chen. Privacy protection in video-based human activity recognition: A survey. InWorkshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (AAAI), 2015.

[5] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. Sensors, 15:31314–31338, 2015.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” proceedings.mlr.press, Apr. 10, 2017.

[7] Hyunsung Cho, Akhil Mathur, and Fahim Kawsar. 2022. FLAME: Federated Learning across Multi-device Environments. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6, 3, Article 107 (September 2022).

[8] Gjoreski, Martin, Vito Janko, Gašper Slapničar, Miha Mlakar, Nina Reščič, Jani Bizjak, Vid Drobnič et al. ”Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors.” Information Fusion 62 (2020): 47-62.

[9] Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz et al. ”Advances and open problems in federated learning.” Foundations and Trends® in Machine Learning 14, no. 1–2 (2021): 1-210.

[10] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, May 2020.

[11] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. ACM Trans. Intell. Syst. Technol. 10, 2, Article 12 (March 2019).

[12] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, “A survey on federated learning,” Knowledge-Based Systems, vol. 216, p. 106775, Mar. 2021.

[13] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny et al., “Towards federated learning at scale: System design,” in Proc. Conf. Machine Learning and Systems, 2019.

[14] J. Hamer, M. Mohri, and A. T. Suresh, “FedBoost: A Communication-Efficient Algorithm for Federated Learning,” proceedings.mlr.press, Nov. 21, 2020.

[15] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, “FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization,” proceedings.mlr.press, Jun. 03, 2020.

[16] X. Yao, C. Huang and L. Sun, ”Two-Stream Federated Learning: Reduce the Communication Costs,” 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 2018, pp. 1-4.

[17] F. Sattler, K. -R. Müller and W. Samek, ”Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 8, pp. 3710-3722, Aug. 2021.

[18] Hanzely, Filip, and Peter Richtárik. ”Federated learning of a mixture of global and local models.” arXiv preprint arXiv:2002.05516 (2020).

[19] Arivazhagan, Manoj Ghuhan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. ”Federated learning with personalization layers.” arXiv preprint arXiv:1912.00818 (2019).

[20] Dario Fenoglio, Daniel Josifovski, Alessandro Gobbetti, Mattias Formo, Hristijan Gjoreski, Martin Gjoreski, and Marc Langheinrich. 2023. Federated Learning for Privacy-aware Cognitive Workload Estimation. In Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia (MUM ’23). Association for Computing Machinery, New York, NY, USA, 25–36.

[21] T. Tuor, S. Wang, B. J. Ko, C. Liu and K. K. Leung, "Overcoming Noisy and Irrelevant Data in Federated Learning," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 5020-5027.

[22] Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical secure aggregation for federated learning on user-held data." arXiv preprint arXiv:1611.04482 (2016).

[23] McMahan, H. Brendan, Daniel Ramage, Kunal Talwar, and Li Zhang. "Learning differentially private recurrent language models." arXiv preprint arXiv:1710.06963 (2017).

[24] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," in IEEE Communications Surveys Tutorials, vol. 23, no. 3, pp. 1622-1658, third quarter 2021.

[25] A. Imteaj, U. Thakker, S. Wang, J. Li and M. H. Amini, "A Survey on Federated Learning for Resource-Constrained IoT Devices," in IEEE Internet of Things Journal, vol. 9, no. 1, pp. 1-24, 1 Jan.1, 2022.

[26] Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119 (2020).

[27] Xu, Jie, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. "Federated learning for healthcare informatics." Journal of Healthcare Informatics Research 5 (2021): 1-19.

[28] Z. Du, C. Wu, T. Yoshinaga, K. -L. A. Yau, Y. Ji and J. Li, "Federated Learning for Vehicular Internet of Things: Recent Advances and Open Issues," in IEEE Open Journal of the Computer Society, vol. 1, pp. 45-61, 2020.

[29] Alamgir, Z., Khan, F.K. and Karim, S. Federated recommenders: methods, challenges and future. Cluster Comput 25, 4075–4096 (2022).

[30] Qiu, Sen, Hongkai Zhao, Nan Jiang, Zhelong Wang, Long Liu, Yi An, Hongyu Zhao, Xin Miao, Ruichen Liu, and Giancarlo Fortino. "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges." Information Fusion 80 (2022): 241-265.

[31] Nweke, Henry Friday, Ying Wah Teh, Mohammed Ali Al-Garadi, and Uzoma Rita Alo. "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges." Expert Systems with Applications 105 (2018): 233-261.

[32] M.M. Hassan, M.Z. Uddin, A. Mohamed, A. Almogren, A robust human activity recognition system using smartphone sensors and deep learning, Future Gener. Comput. Syst. 81 (2018) 307–313

[33] R. Mutegeki, D.S. Han, A CNN-LSTM approach to human activity recognition, in: 2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC, IEEE, 2020, pp. 362–366.

[34] S. Mekruksavanich, A. Jitpattanakul, LSTM networks using smartphone data for sensor-based human activity recognition in smart homes, Sensors 21 (5) (2021) 1636.

[35] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, L. Yuan, Hand gesture recognition using deep feature fusion network based on wearable sensors, IEEE Sens. J. 21 (1) (2020) 539–547.

[36] R. Li, H. Li, W. Shi, Human activity recognition based on LPA, Multimedia Tools Appl. 79 (41) (2020) 31069–31086.

[37] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2022. Assessing the State of Self-Supervised Human Activity Recognition Using Wearables. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6, 3, Article 116 (September 2022), 47 pages.

[38] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6, 1, Article 17 (March 2022), 28 pages. https://doi.org/10.1145/3517246

[39] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. 2020. Masked reconstruction based self-supervision for human activity recognition. In Proceedings of the 2020 ACM International Symposium on Wearable Computers (ISWC '20). Association for Computing Machinery, New York, NY, USA, 45–49. https://doi.org/10.1145/3410531.3414306

[40] K. Sozinov, V. Vlassov and S. Girdzijauskas, "Human Activity Recognition Using Federated Learning," 2018 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Ubiquitous Computing and Communications, Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom), Melbourne, VIC, Australia, 2018, pp. 1103-1111.

[41] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. FedDL: Federated Learning via Dynamic Layer Sharing for Human Activity Recognition. In Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys '21). Association for Computing Machinery, New York, NY, USA, 15–28.

[42] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. ClusterFL: a similarity-aware federated learning system for human activity recognition. In Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21). Association for Computing Machinery, New York, NY, USA, 54–66.

[43] Xiao, Zhiwen, Xin Xu, Huanlai Xing, Fuhong Song, Xinhan Wang, and Bowen Zhao. "A federated learning system with enhanced feature extraction for human activity recognition." Knowledge-Based Systems 229 (2021): 107338.

[44] Youpeng Li, Xuyu Wang, and Lingling An. 2023. Hierarchical Clustering-based Personalized Federated Learning for Robust and Fair Human Activity Recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7, 1, Article 20 (March 2023), 38 pages.

[45] Lulu Gao and Shin'ichi Konomi. 2023. Personalized Federated Human Activity Recognition through Semi-supervised Learning and Enhanced Representation. In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing ; the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct). Association for Computing Machinery, New York, NY, USA, 463–468.

[46] Borjan Sazdov, Bojan Jakimovski, Simon Stankoski, Ivana Kiprijanovska, Bojan Sofronievski, Martin Gjoreski, Charles Nduka, and Hristijan Gjoreski. 2023. Privacy-aware Human Activity Recognition with Smart Glasses for Digital Therapeutics. In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing; the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct). Association for Computing Machinery, New York, NY, USA, 592–596.

[47] Davide Casnici. 2023. Federated Learning for Human Activity Recognition using Ear-Worn Sensors ; Master's thesis at Università della Svizzera italiana. https://thesis.bul.sbu.usi.ch/theses/2171-2223Casnici/pdf?1696593780

[48] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. ESense: Open Earable Platform for Human Sensing. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems (SenSys '18). Association for Computing Machinery, New York, NY, USA, 371–372.

[49] J. Archer et al., "OCOsenseTM Smart Glasses for Analyzing Facial Expressions Using Optomyographic Sensors," in IEEE Pervasive Computing, vol. 22, no. 3, pp. 18-26, 1 July-Sept. 2023.

[50] Zhang, Junbo, Yu Zheng, and Dekang Qi. "Deep spatio-temporal residual networks for citywide crowd flows prediction." In Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 1. 2017.

[51] Beutel, Daniel J and Topal, Taner and Mathur, Akhil and Qiu, Xinchi and Fernandez-Marques, Javier and Gao, Yan and Sani, Lorenzo and Kwing, Hei Li and Parcollet, Titouan and Gusmão, Pedro PB de and Lane, Nicholas D. "Flower: A Friendly Federated Learning Research Framework." arXiv preprint arXiv:2007.14390 (2020).

VIII. Appendix

This appendix provides extended results and additional metrics that complement the findings presented in the main body of the paper. These results offer deeper insights into the performance of the multi-frequency STResNet model under various conditions and with different metrics.

A. Extended Performance Metrics

Table VII presents a comprehensive set of performance metrics for the multi-frequency STResNet model, also introducing accuracy to the F1-score. This table complements Table VI reported in the main results.

TABLE VII
Comparison of Accuracy, F1-score and Number of Training Users for USI-HEAR and OCOsense (40-5Hz)

	USI-HEAR			OCOsense		
Freq.	**Accuracy**	**F1-score**	**Users**	**Accuracy**	**F1-score**	**Users**
5Hz	60.17 ± 3.40	59.30 ± 3.33	7 (5Hz)	73.06 ± 6.43	72.81 ± 14.17	5 (5Hz)
Down-5Hz	65.74 ± 2.61	65.38 ± 2.39	14 (5Hz)	81.36 ± 2.44	86.17 ± 2.29	10 (5Hz)
40Hz	63.75 ± 4.05	62.57 ± 4.49	7 (40Hz)	76.90 ± 3.20	79.74 ± 2.66	5 (40Hz)

Multi-5Hz	63.69 ± 3.28	63.26 ± 3.08	7(5Hz)
			7(40Hz)
				80.00 ± 1.87	85.38 ± 2.52	5(5Hz)
						5(40Hz)

Multi-40Hz	66.34 ± 3.31	65.53 ± 3.61	7(5Hz)
			7(40Hz)
				78.78 ± 2.51	83.45 ± 1.99	5(5Hz)
						5(40Hz)

Ideal 40Hz	69.73 ± 3.02	69.14 ± 2.96	14 (40Hz)	80.77 ± 2.47	85.77 ± 2.17	10 (40Hz)

B. Additional Experiments

In addition to our primary experiments at a sampling frequency of 5Hz, we extended our investigation to assess the model's performance with signals sampled at a lower frequency of 3Hz. Table VIII presents the accuracy and F1-scores of our multi-frequency STResNet model, comparing its performance with configurations for exclusively 3Hz clients, clients downsampled to 3Hz (Down-3Hz), exclusively 40Hz clients, and an ideal scenario where all clients operate at 40Hz (Ideal 40Hz). Despite the reduced sampling rate, our multi-frequency model preserves its advantages over single-frequency models, registering an F1-score of 80.69% ± 3.64% and 80.49% ± 2.51% for 40Hz and 3Hz respectively, as compared to 78.90% ± 3.01% and 66.26% ± 18.15% for their single-frequency counterparts. Overall, it is observed that lowering the sampling frequency to 3Hz generally leads to a decline in HAR prediction accuracy. Nevertheless, particularly within the OCOsense dataset, down-sampling all clients to 3Hz yields results that are still comparable to using the full 40Hz frequency.

TABLE VIII
Comparison of Accuracy, F1-score and Number of Training Users for USI-HEAR and OCOsense (40-3Hz)

	USI-HEAR			OCOsense		
Freq.	**Accuracy**	**F1-score**	**Users**	**Accuracy**	**F1-score**	**Users**
3Hz	57.48 ± 4.45	57.06 ± 4.24	7 (3Hz)	71.56 ± 6.65	66.26 ± 18.15	5 (3Hz)
Down-3Hz	63.68 ± 2.43	63.09 ± 2.46	14 (3Hz)	80.12 ± 2.21	82.39 ± 2.56	10 (3Hz)
40Hz	62.94 ± 3.65	62.05 ± 3.95	7 (40Hz)	76.61 ± 3.50	78.90 ± 3.01	5 (40Hz)

Multi-3Hz	61.48 ± 3.45	61.13 ± 3.52	7(3Hz)
			7(40Hz)
				78.48 ± 1.95	80.49 ± 2.51	5(3Hz)
						5(40Hz)

Multi-40Hz	64.85 ± 2.20	64.10 ± 2.03	7(3Hz)
			7(40Hz)
				77.44 ± 2.56	80.69 ± 3.64	5(3Hz)
						5(40Hz)

Ideal 40Hz	68.58 ± 3.01	68.08 ± 3.09	14 (40Hz)	78.59 ± 2.59	82.65 ± 2.18	10 (40Hz)