Let's Grab a Drink: Teacher-Student Learning for Fluid Intake Monitoring using Smart Earphones

Shijia Zhang[1], Yilin Liu[1], and Mahanth Gowda[1]

[1]The Pennsylvania State University, University Park, PA
Email: {scarlettzhang27, yzl470, mahanth.gowda}@psu.edu

***Abstract*—This paper shows the feasibility of fluid intake estimation using earphone sensors, which are gaining in popularity. Fluid consumption estimation has a number of healthcare-related applications in tracking dehydration and overhydration which can be connected to issues in fatigue, irritability, high blood pressure, kidney stones, etc. Therefore, accurate tracking of hydration levels not only has direct benefits to users in preventing such disorders but also offers diagnostic information to healthcare providers. Towards this end, this paper employs a voice pick-up microphone that captures body vibrations during fluid consumption directly from skin contact and body conduction. This results in the extraction of stronger signals while being immune to ambient environmental noise. However, the main challenge for accurate estimation is the lack of availability of large-scale training datasets to train machine learning models (ML). To address the challenge, this paper designs robust ML models based on techniques in data augmentation and semi-supervised learning. Extensive user study with 12 users shows a per-swallow volume estimation accuracy of 3.35 mL ($\approx$ 19.17% error) and a cumulative error of 3.26% over an entire bottle, while being robust to body motion, container type, liquid temperature, sensor position, etc. The ML models are implemented on smartphones with low power consumption and latency.**

I. Introduction

Sufficient hydration is essential for blood circulation, metabolism, temperature regulation, and overall smooth functioning of the human body. However, several surveys have indicated that 50-75% of people can have a net fluid loss (fluid intake is lesser than fluid release), thus leading to chronic dehydration over time [6], [8]. While short-term effects of dehydration include fatigue, foggy memory, irritability, etc., [15], [38] long-term effects due to chronic dehydration can lead to high blood pressure, kidney stones, etc. This can lead to further complications depending on the condition of the body [7]. Proper hydration levels improve cognitive performance and mood [81] while sustaining an overall healthy lifestyle in the long run.

Towards detecting dehydration and alerting users, this paper presents a system called *LiquidMeter*, which shows the feasibility of estimating the volume of fluid intake by exploiting earphone sensors that are gaining in popularity with an expectation to reach a $45.7 billion market by 2026 [35]. At a high level, *LiquidMeter* performs drinking volume estimation by analyzing body sounds during drinking by using bone conduction microphones in the earphones, detailed in Sec. III as Voice-pickup-units (VPU). As the fluid is swallowed, the fluid's motion and the opening and closing of the esophagus (food pipe) for letting the fluid into the stomach will produce acoustic vibrations. These vibrations propagate through the skull, captured in the ears through earphone VPU. In contrast to an ordinary microphone, the VPU measures vibrations directly from a solid surface, thus resulting in a stronger reception and isolation from external noise and interference.

Motivated by the need for monitoring hydration levels, fluid intake monitoring is an active area of research [28]. Vision-based approaches [19], [27] are prevalent in monitoring activities of daily living, including drinking detection. Similarly, smartwatch sensors have also been employed for detecting activities related to drinking vs. eating classification [41], [46], [77]. Smart surfaces that use load cells, pressure sensors, etc., can monitor liquid and food intake when the container is placed on them before and after drinking [61], [76], [83]. Finally, smart containers that utilize capacitive, conductive, pressure, radar sensors, etc., can estimate the drinking volume for fluids consumed with the container [40], [48], [65].

In contrast to prior works, *LiquidMeter* provides the following advantages: (i) Sensing by vision-based approaches can be limited to the camera's view and be susceptible to lighting, resolution, and occlusions. In contrast, *LiquidMeter* uses earphone sensors which can be ubiquitous without limitations on the range of sensing, lighting, or occlusions. (ii) Solutions based on smartwatch devices can mainly detect the action of drinking. The volume estimation is limited to special cases where the user stays *still* while drinking, and the bottle has to be placed on a flat surface before and after drinking. In contrast, *LiquidMeter* works under adhoc and natural conditions, including body and head motion. (iii) Solutions based on smart surfaces or smart containers work under specific settings of having a specialized surface or drinking using a particular container. In contrast, *LiquidMeter*'s solution works on any surface or container since the sensing is done directly on the human body.

Estimation of fluid consumption volume with earphone sensors is challenging for many reasons: (i) Because of the nature of the biological process involved during fluid consumption, sound generation, and its propagation through the face, the relationship between the fluid consumption volume and its acoustic fingerprint can be complex. While machine learning (ML) algorithms can be used to learn this relationship, the training data is limited. Unlike vision and speech domains, there are no large-scale training datasets for the relatively new earphone VPU sensors, or wearable devices [56]. (ii) While *LiquidMeter* designs synthetic training data to address the above challenge, such data is unlabelled (elaborated in Sec. IV-B). (iii) The drinking activity and its acoustic fingerprint

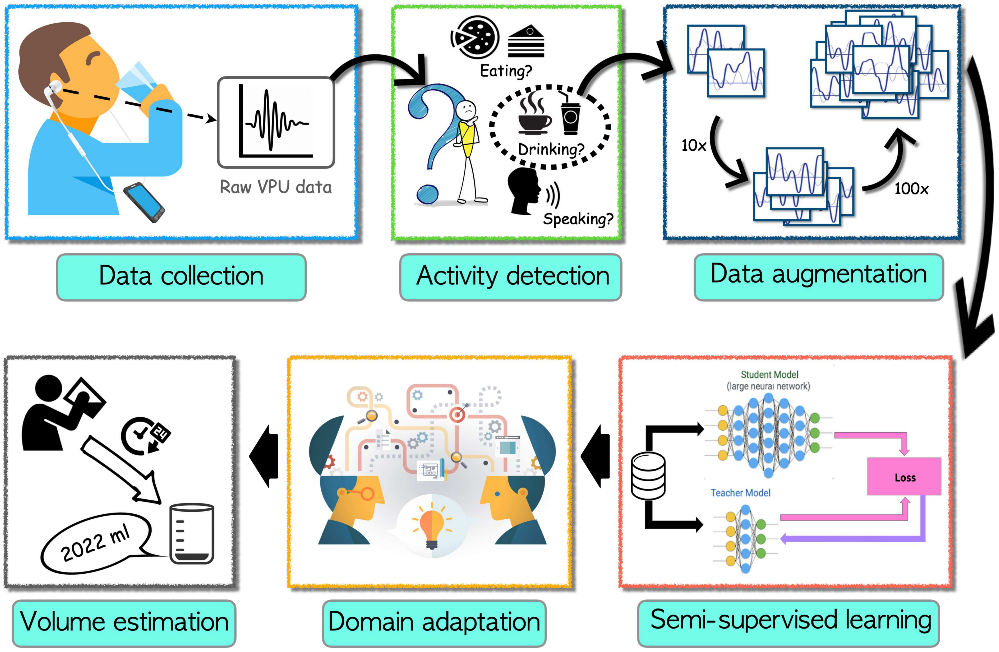

Fig. 1. Overview of *LiquidMeter*: The vibration data from earphones is first used for identifying drinking activity. Drinking related vibration data is combined with data augmentation and semi-supervised learning (teacher-student training) for accurate estimation of the volume of fluid intake with limited training data

can vary across users because of diversity in body sizes, drinking characteristics, etc. (iv) Drinking activities need to be separated from non-drinking activities such as eating, speaking, etc. (v) The choice of earphone sensors in *LiquidMeter* must ensure robustness to external noise and interference.

Enumerated below, *LiquidMeter* exploits a combination of algorithmic and systems-based opportunities to tackle the above challenges. (i) *LiquidMeter* designs data augmentation techniques to transform the limited real data into several samples of synthetic training data by creating variations in pitch, introducing time-shifts, and adding noise. (ii) While the augmented synthetic data may not retain the original labels, *LiquidMeter* designs ML models based on *teacher-student* learning strategies to train efficiently with unlabelled synthetic data. (iii) *LiquidMeter* exploits techniques in domain adaptation for customizing a pre-trained model on other users for a new user. While pretraining helps quickly generate a base model for the new user, domain adaptation will fine-tune the model to customize it to a specific user. (iv) *LiquidMeter* first performs classification of the activity (eating vs. drinking vs. speaking) and triggers the ML model for drinking volume estimation only when a drinking activity is detected. (v) *LiquidMeter* uses bone conduction microphones in contrast to ordinary microphones. This provides isolation from external noise since these microphones have to be in contact with a vibrating surface to detect the sound.

The overall architecture of *LiquidMeter* is depicted in Fig. 1. The VPU data from earphones is first used for isolating drinking activity from other activities such as eating, speaking, etc. When a drinking activity is detected, the ML modules for drinking volume estimation are triggered. These models are trained by exploiting ideas in data augmentation, and semi-supervised learning to handle the challenge of limited training data. Finally, domain adaptation is done on the model thus trained to handle user diversity.

LiquidMeter uses two earphone sensors developed by Sonion. The earphones are embedded with special microphones that can detect vibrations directly from the ear's surface through which the body sound can be captured. The ML models are implemented on smartphones using TensorFlowLite. Evaluated over 6 categories of liquids known to account for $\approx$ 85% of fluid consumption [33], the error in volume estimation during each swallow instance is about 19.17%, whereas the cumulative error over an entire bottle of liquid is around 3.26% (details in Sec. V). Furthermore, our experiments validate robustness to natural variation in earphone wearing positions, body and head motion, ambient acoustic interference, temperature of the fluid, container-type, etc. Therefore, we believe *LiquidMeter* offers a practical solution.

Considering the above possibilities, we summarize *LiquidMeter*'s contributions below: (i) Estimation of liquid intake volume using off-the-shelf earphones under natural and adhoc conditions. (ii) Design of ML models based on data augmentation and student-teacher learning to work with limited training data. (iii) Extensive user study across different liquids, container types, temperature, body-motion, etc to validate the feasibility of the system. (iv) Implementation on embedded devices with low latency and power consumption.

II. Background

We will begin with a brief background on the biological process of swallowing activity and sound generation.

Swallowing food, fluids, and saliva is an essential life-sustaining activity like breathing. Humans swallow 500-700

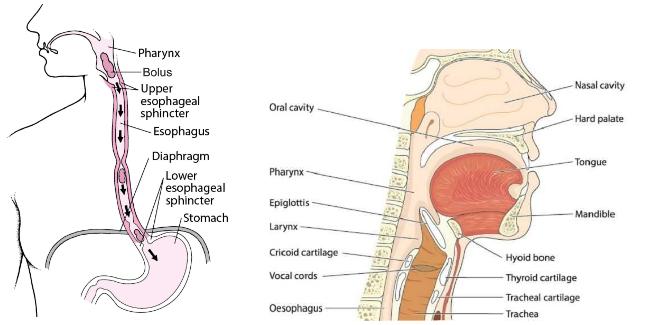

Fig. 2. (a) Flow of contents during human swallowing. Source of image [36] (b) Posterior tongue and hyoid bone can help people to swallow. Source of image [2]

times per day, including 3 times per hour while sleeping and even more while awake [69]. Fig. 2(a) depicts the flow of contents (fluid/food) during swallowing. The three stages involved in the process are: (i) *Oral Stage:* The food is chewed by the posterior tongue, including support from the hyoid bone (tongue bone) depicted in Fig. 2(b). The food is converted into a paste form called *bolus*, whereas, in the case of fluids, chewing may not be needed as it is already in the desired form. (ii) *Pharyngeal stage:* Here, the *larynx* (voice box) moves, and the *epiglottis* closes with the sole aim of preventing the bolus from entering the windpipe. The *hyoid* bone now elevates, and cricopharynx opens to force the bolus into the food pipe. (iii) *Esophageal Stage:* Muscular contractions of the *esophagus* (food pipe) will now propel the bolus into the stomach. Finally, the *larynx* and the *epiglottis* will resume their resting position to prepare for the next swallowing bout [10], [45].

The above three stages generate acoustic vibrations characterized as follows: (i) During the *pharyngeal stage*, the opening of the cricopharynx to force bolus motion into the esophagus creates initial discrete sounds (IDS). (ii) During the *esophageal stage*, a gurgling sound is produced due to the motion of the bolus in the esophagus called bolus transmission sounds (BTS). (iii) Sometimes, a final discrete sound (FDS) might be generated as the bolus reaches the stomach [10]. *LiquidMeter*'s ML models extract features predictive of volume consumed from such sounds. Although the acoustic pattern might vary across people, the typical intensity during swallowing of a fluid varies between 28-62 dB, whereas the frequency varies between 660-1170 Hz [29]. This property is exploited for performing data augmentation (Sec. IV-A).

III. Platform Description

We now discuss the detection of the generated sound on earphones via bone conduction using our platform. We begin by describing the hearing activity, and the role of bone conduction [44]. A spoken sound, will travel through the air and reach the eardrums (depicted as air conduction in Fig. 3). Here, the *cochlea* will convert the acoustic vibrations into electrical impulses to be processed by the brain [18]. In addition to air conduction, the figure also depicts a bone conduction path, through which the sound can reach the cochlea and eventually be converted into electrical impulses reaching the brain [17].

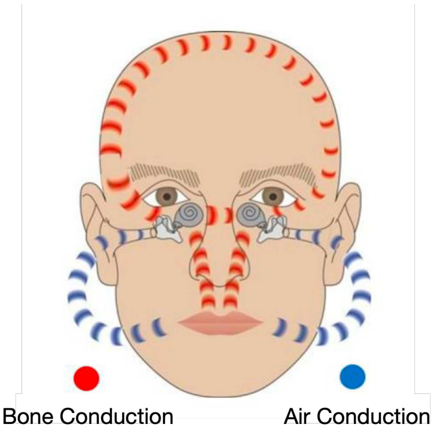

Fig. 3. Bone conduction. Source of image [67]

Ludwig Van Beethoven, a great composer was known to attach a rod between the piano and his head to exploit bone conduction to aid his hearing after he was diagnosed with a hearing loss [59]. We will exploit the bone conduction path for picking up body sounds during fluid intake because of its higher quality and robustness to external noise, especially when fluids are consumed in spaces with high noise such as restaurants, industrial settings, travel (train, plane), etc.

Fig.4 depicts our platform from Sonion [68] which consists of a voice pick-up (VPU) bone sensor. The VPU consists of a microphone (INVN ICS-40619 [3]) in low power mode. A mass-spring is connected to the microphone at its audio port. The role of the mass-spring is to pick up the bone-conducted sound which is much stronger than air-conducted sound. Fig.5 depicts the amplitude of captured vibrations for the utterance of “Let’s grab a drink” with a VPU and an ordinary microphone with a light membrane. Evidently, the VPU can capture stronger vibrations.

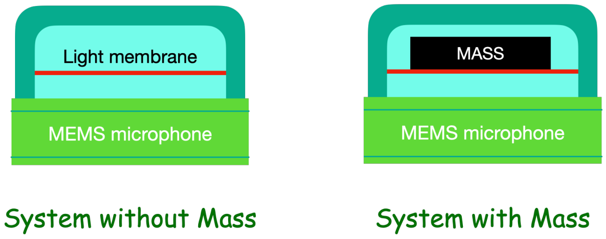

Fig. 4. VPU with and without additional mass-spring

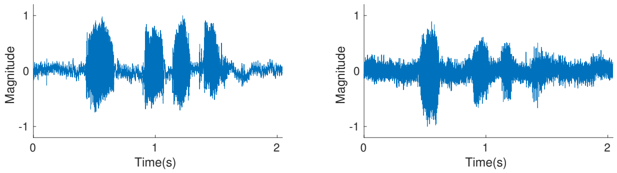

Fig. 5. Noisy Environment: (a) VPU data (b) Microphone data

Fig. 6 depicts an example of raw audio captured by the VPU sensor when a user drank 18.43 mL of water. The patterns, IDS, BTS, and FDS discussed earlier can be clearly seen. *LiquidMeter*'s technical modules discussed next will convert this raw audio into an estimation of the volume of fluid consumed.

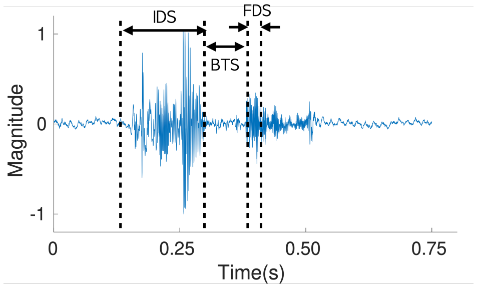

Fig. 6. Raw drinking data showing swallowing stages

IV. Technical Modules

In this section, we discuss ideas for estimating the volume of the liquid consumed based on the sound pattern captured by the earphones. Because of the nature of the biological process involved during fluid consumption, sound generation, and its propagation through the face, the relationship between the fluid consumption volume and its acoustic fingerprint can be complex. Therefore, we design ML models to automatically extract the relationship. However, the success of ML models depends on the availability of large quantities of training datasets. Unlike vision and speech domains, there are no large-scale training datasets for the relatively new earphone VPU sensors, or wearable devices [56]. Given the lack of availability of such training datasets, we exploit opportunities in *data augmentation* and *semi-supervised learning* in achieving a sweet spot in the trade-off between accuracy and training overhead. We expand on various modules in the high-level architecture in Fig. 1.

A. Data Augmentation

Towards handling the challenge of limited training data, we generate synthetic training data by designing a number of transformations to small-scale real training data. The specific *data augmentation* techniques for performing such transformations are discussed below.

Temporal Shift: The ML model accepts 0.75 seconds of audio from the earphone as the input. We create additional input instances by shifting the audio randomly by 0-0.15 seconds towards the right (*fast forward*) or left (*rewind*). With *fast forward*, we add a few seconds of noise at the beginning of the input instance, whereas with *rewind*, we add a few seconds of noise at the end. Alternatively, the noise in the original recording will also do as well as explicitly adding noise at the beginning or end. This creates alternative instances of the input where the swallow event happens at a slightly different instant of time (since the drink was taken into the mouth) than the original input instance. Fig. 7(b) shows an augmented version of the signal in Fig. 7(a) based on time-shifting.

Changing Pitch: The pitch of a sound is related to the perception of the frequency of the sound by the human ear. The pitch of the sound can vary across, gender, age, and person. Given an instance of drinking sound, we change the pitch [4] randomly so as to emulate the generation of a similar drinking activity by a different person, thereby creating more training examples. Fig. 7(c) shows an augmented version of the signal in Fig. 7(a) based on changing the pitch.

Noise Augmentation: We add white gaussian noise to audio samples to create augmented versions of the data. The noise variance is chosen so that the signal-to-noise ratio (SNR) of the original audio degrades from about 40 to 35 dB. Fig. 7(d) shows an augmented version of the signal in Fig. 7(a) based on noise addition. The semi-supervised learning strategies discussed next will build on these data augmentation techniques.

B. Semi-Supervised Learning with Teacher Student Training

Building on the data augmentation techniques discussed above, *LiquidMeter* designs ML models based on *teacher-student* learning to efficiently train ML models with limited training data. Teacher-Student learning is an active area with a number of applications in improving the efficiency of training. Heavy ML models can be compressed to run efficiently on embedded devices like smartphones through knowledge distillation [58], [74]. Image recognition performance on ImageNet and COCO datasets can be enhanced [78], [85]. Distant speech recognition and beamforming techniques also reap benefits [75]. At a high level, *LiquidMeter*'s ML models are inspired by these works. However, the feature representations, network architectures, data augmentation techniques, etc, have been designed carefully to suit our problem domain.

The high-level architecture is depicted in Fig. 8 (i) We first collect small-scale real data by conducting a user study. Using this, we train a smaller ML model (teacher, *Model 1* in Fig. 8) with fewer parameters. While the *teacher* can learn from small training data, the accuracy of such a model can be very limited. (ii) Next, we use data augmentation techniques discussed in Sec. IV-A to expand the real data into synthetic data which is 10 times more than the real data. However, the data augmentation techniques do not necessarily preserve the original labels (particularly with changes to pitch and noise addition) from which a corresponding synthetic data was created. Therefore, the synthetic data thus generated is treated as unlabelled data. (iii) We generated pseudo labels for the unlabelled synthetic data using the *teacher* model. (iv) We now train a larger ML model (student, *Model 2* in Fig. 8) by combining the synthetic data (with pseudo labels) and the real data with original labels. While the pseudo labels might be noisy, the student network is expected to have higher accuracy than the teacher for many reasons: (a) Higher network capacity (b) The training includes a combination of original labels and the pseudo labels, thus providing an overall larger dataset for training. Despite being noisy, our empirical observation is that the pseudo labels are still informative enough to improve the overall learning process. (c) The pseudo labels with low confidence are discarded. (v) The *student* network developed above can be used as a *teacher* network for performing another iteration of teacher-student learning so as to achieve a

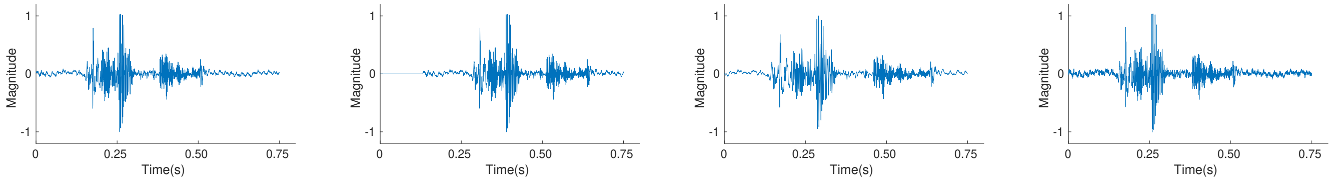

Fig. 7. Data Augmentation: (a) Original Signal (b) Temporal Shift (c) Change of Pitch (d) Noise Addition

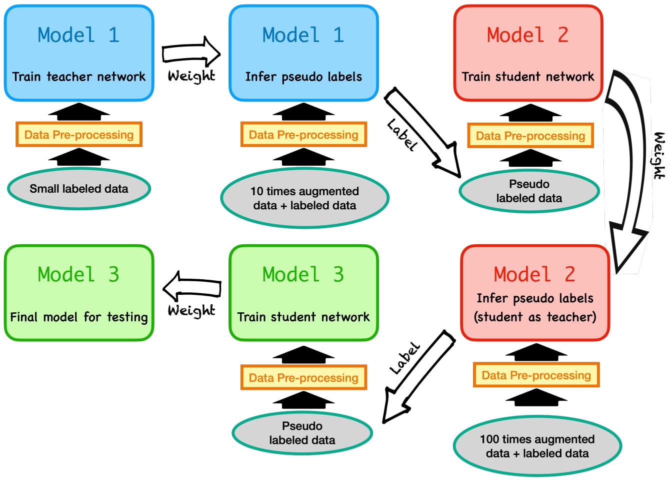

Fig. 8. High-level architecture of teacher-student-based semi-supervised learning

better accuracy (to generate *Model 3* in Fig. 8). *LiquidMeter* performs two iterations of the *teacher-student learning* as discussed above since more iterations beyond this resulted in only marginal gains. We now elaborate on the details of the teacher and student models.

Teacher (compact model): Fig. 9 depicts the architecture of a teacher model (*Model 1*). The model takes time domain and frequency domain features as inputs, computed from T audio samples. With a sampling rate of 4000 Hz and an input audio size of 0.75 s, $T = 3000$. The raw time series data is used as the time-domain input. Therefore, the size of the time domain input is 3000×2 since we have two channels of data from the two ears. The frequency-domain features consist of the Mel-Frequency Cepstral Coefficients (MFCC) features. The MFCC features are known to capture a highly compressed representation of information in the frequency domain, thus popular in many speech processing applications [53]. We compute 13 MFCC co-coefficients from $25ms$ *frames* of audio. After computing the MFCC features for the current $25ms$ *frame*, we move by $10ms$ to capture the next $25ms$ *frame*, whose MFCC features are computed next. This creates an overlap of 15ms between successive *frames*, necessary to minimize information loss due to windowing and other transformations performed while computing MFCC features. Therefore, the size of our frequency-domain input for a 0.75s segment of input audio from both earphones would be $13 \times 74 \times 2$.

To best capture the spectro-temporal relationships, we use both time and frequency domain features as discussed above as inputs to the ML model. While using Short-time Fourier transforms (STFT) might be one idea to capture spectro-temporal features, our choice of network input design is inspired by recent works which show that using separate time and frequency domain inputs or even designing the network with multiple resolutions of STFTs as input can offer greater flexibility in spectro-temporal feature extraction than conventional STFT-based design [79].

The input passes through a series of convolutional layers with the input downsized at each layer with maxpool operation. The model attempts to capture a compact representation of the input to be used for drinking volume estimation. Batch normalization is used at each layer for accelerating convergence by controlling variation in the input distribution at each layer. The overall size of the model is chosen to be smaller so that it can learn with small-scale training data. While the accuracy of such a model might be low, the student model discussed next will expand on the teacher model for higher accuracy and robustness.

Student (expanded model): Fig. 9 depicts the architecture of the student models (*Model-2* and *Model-3*) at both iterations of the semi-supervised learning strategy. As discussed earlier, the student model for the first iteration serves as the teacher for the second iteration of semi-supervised learning. While the format of the input for the student models is similar to the teacher model, the depth of the networks can larger with an overall higher number of parameters to facilitate better learning. Another key difference between the student and teacher network is the introduction of residual connections (in the second student model, *Model 3*). Residual connections are known to accelerate training of deeper networks while providing a sweet spot between stronger feature representation and convergence of the model [43].

Loss Function: The loss function is the Mean Absolute Error (MAE) for volume estimation as depicted in the simple equation below, where V_{pred} and V_{truth} are predicted and ground truth values of volume estimates.

$$MAE = \sum_{i}^{N} |V_{pred} - V_{truth}|/N \tag{1}$$

C. Domain Adaptation to Handle User Diversity

The vibration pattern of earphone signals might vary across users due to differences in body shape, gender, drinking pattern, etc. While *LiquidMeter* designs techniques based on semi-supervised learning to decrease the training overhead (Sec. IV-B), we also explore domain adaptation techniques to further reduce the overhead of training across multiple users.

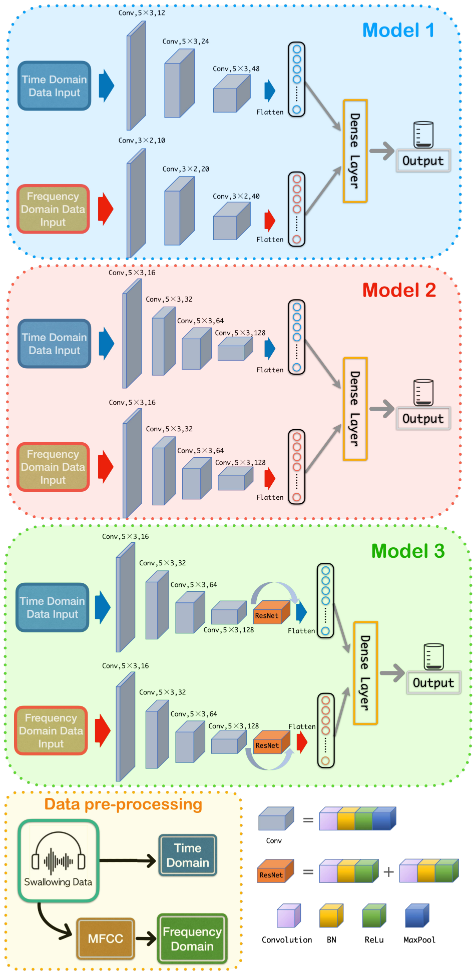

Fig. 9. ML model architectures used in teacher-student learning.

Despite differences across users, the overall biological process of drinking has a lot of similarities. Therefore, *LiquidMeter* designs domain adaptation techniques to adapt a *pretrained* model on one user to perform inferences on a new user with minimal training overhead.

Transfer-learning-based domain adaptation is popular in vision and speech processing. For example, AlexNet model [55] pretrained on ImageNet database [30] was fine-tuned for classifying images in medical domain [84], remote-sensing [42] and breast-cancer [63]. Similarly, a pre-trained BERT language model [31] was fine-tuned for tasks such as text-summarizing [80], question answering, [66] etc. This significantly reduces the burden of training for a new task. In a similar spirit, we use pretrained model from one user and *fine-tune* it for a different user to significantly decrease the training overhead (Fig. 17(b)) without losing much accuracy.

Domain adaptation is performed as enumerated below: (i) We generate a model for one user by first training the model with labeled data from that user – known as the *pretrain*ed model. (ii) We collect small training data from the new (*target*) user. Instead of developing the model for the *target* user from scratch, we initialize the model weights to be the same as the *pretrain*ed model. (iii) We make all layers untrainable except the (*BN*) layers. Using small-scale training data from the *target* user, we update the BN layers to minimize the loss function. This is called *fine tuning*. (iv) Unlike supervised learning, *LiquidMeter* designs ML models based on semi-supervised learning and teacher-student interactions. Therefore, the process of fine-tuning goes through exactly identical iterations of teacher-student training described in Sec. IV-B. However, in contrast to updating the weights of the entire network, only the BN layers are updated. The model thus generated will be used for making inferences on the *target* user.

Finetuning the BN layers help with domain adaptation because of their ability to contain wide oscillations in the distributions of input fed from one layer to the next. Given the sufficient success in BN layers (with only a few parameters) for accelerating convergence by minimizing *covariate shift* [49], we exploit them towards domain adaptation as well. The success of this approach has already been shown in other domains such as computer vision [23], [60]. The BN layers will learn to sufficiently transform the distribution from *target* user to a distribution of the *source* user on which the model is *pretrained* on. If successful, the *pre-trained* model from the *source* user can be used for performing inferences on the target user with the *finetuning* steps discussed here. As discussed in Sec. V, this results in a reduction of training overhead on the *target* user significantly.

D. Activity Detection: Drinking vs Eating vs Speaking vs Null

LiquidMeter needs to identify the activity of the user before performing volume estimation. Therefore, we design a classifier to detect and eliminate other activities such as *eating*, *drinking*, or the *null activity*. Our definition of *null activity* includes other normal activities involving body motion such as walking as long as the user is not drinking, eating, or speaking during that time. Whenever a drinking activity is detected, the corresponding data is analyzed further for volume estimation. Our architecture for the classification network is similar to the architectures in Fig. 9 and Fig. 8 that exploit semi-supervised learning for minimizing the training overhead. The last layers and loss functions are different since the classification network needs to label 4 classes (*Eating*, *Drinking*, *Speaking*, *Null*). In contrast to volume estimation which includes ReLU activation and *MAE* loss function, the last layers include SoftMax activation and *cross-entropy* loss function in the classification network. In the interest of space, we skip detailing out the entire architecture in a separate figure. The evaluation results are discussed next.

V. Performance Evaluation

A. User Study

We conduct a study with 12 users (8 males, 4 females). The users are aged between 20-52, and weigh between 47-96 kgs.

Data Collection Methodology: Our study was approved by the IRB committee. The users wear the smart earphones (Sec. III) as shown in Fig.10 on both ears. The users were then

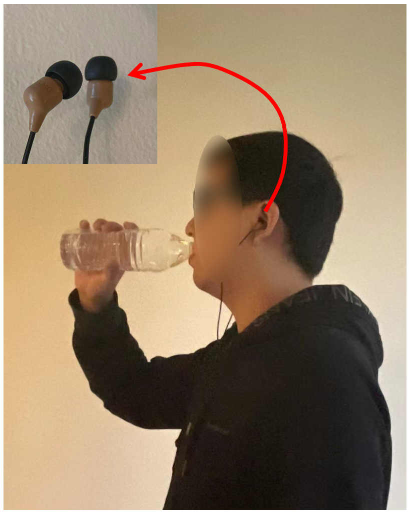

Fig. 10. User wearing earphone sensors while drinking water

instructed to drink six liquids: Water, Milk, Coffee, Tea, Soda, and Juice (Orange). Our chosen category of liquids is known to account for $\approx 85\%$ of fluid intake under daily living conditions [33]. The rest 15% includes special drinks or alcohol, not a part of this study due to IRB restrictions. We allowed the user to drink naturally which includes various speeds depending on their levels of comfort. We decided not to explicitly instruct the users to drink fast or slow since we wanted to be careful with issues such as water getting into the windpipe, therefore we simply instructed the users to drink naturally. Three different containers were used – Bottle, Cup, and Straw – for consuming the fluids. In addition to drinking, the users also performed other activities such as eating, speaking, or the null class where they could walk, or move randomly, or be idle, but they did not speak, eat, or drink. The sensor data was streamed to a smartphone over the audio jack, with a sampling rate of 4000 Hz since higher frequencies are heavily attenuated by the human body [70].

Labels for Training and Testing: The VPU sensor data from both earphones are collected as the user drinks the fluid for predicting the volume of consumption. We use a high-precision weighing scale [1] to measure the ground truth. The user is instructed to place the container on the scale before and after each drinking bout, and the difference in weight as measured by the scale is noted. The weight is later divided by the density of the liquid to obtain the ground truth of volume. We note that the weighing scale is only used for collecting ground truth, it is not a part of our system. The labels for activity classification (eating, speaking, drinking, null class) were derived manually.

Training Data: Towards keeping the volume of fluids consumed per session within natural and comfortable limits, we spread the user study across six different days, with over four sessions per day at different times. Over the period of six days, each user consumed about 2100 mL of water, and 600 mL of each of the 5 other types of liquids: Milk, Coffee, Tea, Soda, and Orange Juice. An equal amount of liquid was consumed each day. We used bottled mineral water at room temperature for "water" studies above. Half of Milk, Coffee, and Tea was tested under both "hot" and "room temperature" conditions, whereas, Soda and Orange Juice were tested under "room temperature", and "cold" conditions. Given ideas in semi-supervised learning discussed in Sec. IV-B, our evaluation depicts that the data thus collected is sufficient for generating robust ML models. The initial data collection was further augmented as elaborated later in this section to test special cases of usability. For activity classification, approximately 20 minutes of data per day was collected for each of the non-drinking activities (eating, speaking, null class). Three kinds of models were developed using the training data: (i) **User-dependent model:** A model for each user that requires training data from the same user. (ii) **Model with domain adaptation:** A model for each user where a pre-trained model from a different user is taken and fine-tuned using techniques in Section IV-C such that only a small fraction of user-specific training data is used for developing a model for the user. (iii) **Multi-user model:** This is a user-independent model. Here, we train a model based on training data from multiple users. The trained model is directly used for inferences on a new user without any training data from the new user.

Test Data: Because of data augmentation and semi-supervised learning strategies, the user-dependent model converged with only a small amount of training data – 65 instances of drinking events (swallowing the liquid) – which is approximately 1500 mL of liquid, and the testing was done with the rest of the data in a randomized cross-validation manner such that the training and the test data are not taken from the same day. With domain adaptation, even lesser training data is needed (15 swallow instances, details in Fig. 17(b)). We only used water for training the model, which generalizes well to other fluids in the study (details in Fig. 12). Given that training and testing data are sampled from different days, the earphone sensor has to be removed and remounted between training and testing, thus providing a benchmark for validating robustness to natural variation in sensor positions. In addition, various other test cases including drinking while walking, drinking with head movements, ambient acoustic interference were considered as described in detail in appropriate subsections.

Metrics of Evaluation: In addition to MAE from Equation 1, we also use the *mean absolute percentage error (MAPE)*, and *mean percentage error (MPE)*, depicted in Equation 2. These are popular metrics for validating fluid intake accuracy [28]. While MAPE computes the absolute error, MPE computes the cumulative sum of errors over time where the positive and negative errors may cancel out. MPE might be more relevant in the context of longer intervals of tracking (over an entire bottle of drink). However, MAPE computes average error across instances of swallowing a drink, providing an estimate of worst-case errors.

$$MAPE = \frac{100}{N}\sum_{i}^{N}\frac{|V(Predicted) - V(GroundTruth)|}{V(GroundTruth)}$$

$$MPE = \frac{100}{N}\sum_{i}^{N}\frac{V(Predicted) - V(GroundTruth)}{V(GroundTruth)} \tag{2}$$

B. Implementation

LiquidMeter is implemented on a combination of desktop and smartphone devices. The ML model is implemented with TensorFlow [9] packages and the training is performed on a desktop with Intel i7-8700K CPU, 16GB RAM memory, and Nvidia GTX 1080 GPU. We use the Adam optimizer [52] with a learning rate of 0.1, β_1 of 0.9 and β_2 of 0.999. To avoid over-fitting issues that may happen in the training process, we apply the L2 regularization [16] on each CONV layer with a parameter of 0.01 and also add dropouts [72] with a parameter of 0.1 following each RELU activations. Once a model is generated from training, the inference is done entirely on a smartphone device using TensorFlowLite [39] on Samsung S20, and Oneplus 9 Pro smartphones.

C. Performance Results

While we mainly provide MAPE and MPE errors in the graphs for brevity, the MAE errors are discussed while evaluating the training overhead of the ML models. If not stated otherwise, the general reported results are from the *model with domain adaptation*. The results from other models (*user dependent*, *multi-user*) are discussed separately.

Activity Classification: Eating vs Speaking vs Drinking vs Null: A user could be performing any activity such as being idle, walking, speaking, or eating in addition to drinking. We first identify the activity. If a drinking activity is detected, the ML module for volume estimation is triggered. We did not have a single instance of missed classification and the accuracy was 100%. Because of the vast differences in generated sounds across activities, they are easily distinguishable.

Accuracy vs Users: Fig.11(a) shows the breakup of accuracy across users. Although the direct use of a model trained from 11 users (multi-user model) and tested on a new user (without domain adaptation) provides a decent accuracy (MAPE = 34.89%, MPE = 7.11%), domain adaptation in *LiquidMeter* can significantly cut down the errors (MAPE = 19.17%, MPE = 3.26%), the performance is close to user-dependent model with less training overhead. The overall accuracy is robust with diversity in users, body mass indices, gender, etc.

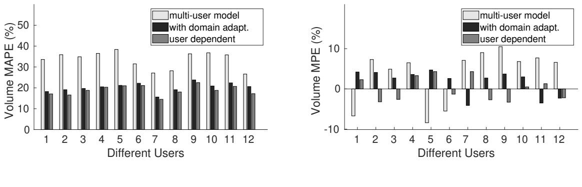

Fig. 11. Accuracy vs users (a) *MAPE* (b) *MPE*

Accuracy vs Type of Drink: Fig. 12 depicts the overall accuracy as a function of the liquid. Evidently, the model trained with water continues to hold well for other liquids, with only small differences. Because of similarity in sound generated across liquids, reliable classification of the type of drink from the VPU data was not possible. However, we believe that because of the same reason, the model trained with water continues to hold for other liquids as well. The accuracy is consistent across liquids, indicating promise in seamless monitoring of fluid intake for commonly used liquids.

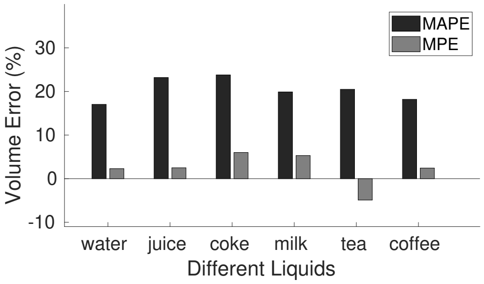

Fig. 12. Accuracy vs type of liquid

Accuracy vs Volume Consumed: Fig. 13(a) depicts the MAE as a function of the amount of liquid consumed. The accuracy is slightly higher within the 10-30 mL regime. Given that the natural distribution of volumes of liquid consumed per swallowing instance (in Fig. 13(b)) tends to be more within 10-30 mL, the ML models are trained with more data within this domain, thus resulting in higher accuracy. Nevertheless, the overall MPE is less than 5%, thus showing promise in accurately tracking the volume over time.

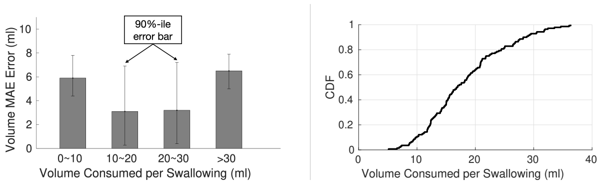

Fig. 13. (a) Accuracy vs volume of liquid consumed per swallow (b) Distribution of volume intake per swallow event

Accuracy vs Type of Container Fig. 14(a) depicts the accuracy vs popularly used containers for drinking the liquid. Evidently, the accuracy is consistent across all containers. This is because the container only influences how the liquid is ingested into the mouth. On the other hand, the volume estimation depends on the sound produced when the liquid passes through the food pipe from the mouth, which is independent of the type of container. Our experiments are in agreement with this hypothesis.

Accuracy vs Temperature: Fig. 14(b) depicts the accuracy variation with temperature. While beverages are typically served hot, the safe temperature for a hot beverage is known to be below 56° [20], and we maintain the temperatures of our hot beverage below this value. Similarly, we maintain the temperature of our cold beverage above 4° using a refrigerator.

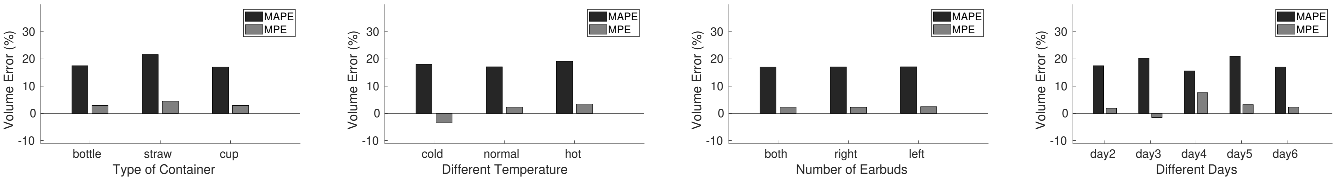

Fig. 14. Accuracy variation across (a) Type of container (b) Temperature (c) Earbuds (d) Sensor position over days

Under these conditions, we observe that the acoustic vibrations generated in the body are similar to those when the drink is taken at room temperature. Accordingly, the volume estimation accuracy is consistent across all temperatures.

Accuracy vs Number of Earbuds: Fig. 14(c) depicts the accuracy for individual earphones as well as the overall accuracy when both are used. The accuracy is uniform across all cases. Inspection of raw recordings reveals that there is substantial redundancy in information captured across both earphones. Thus, we believe it is sufficient to put a VPU sensor in one earphone.

Robustness to Sensor Position Variation: Fig. 14(d) depicts the accuracy over different days of the user study. Although the earphones can fit snugly, there might be small variations in sensor position across days. The training and test data sets were sampled across completely different days to validate robustness to sensor positions. Evidently, the accuracy is consistent across days. The ML models have been trained with noisy data augmentation techniques, we believe this makes the model robust.

Accuracy vs Hands: Fig. 15 depicts the accuracy as a function of the hand used for drinking. Given that the volume estimation depends on sounds produced when the liquid is passing through the food pipe, which is independent of the hand, we did not notice a difference in accuracy due to the hand used for drinking.

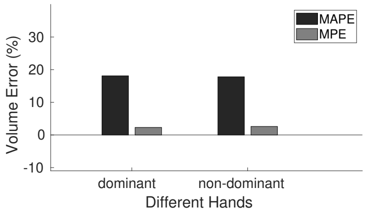

Fig. 15. Accuracy vs Hands

Robustness to Ambient Noise: Since the VPU sensor used in *LiquidMeter* picks up vibrations directly from bone conduction, it is immune to interference from ambient acoustic noise. To validate the hypothesis, we simulate a noise environment by playing airport noise in the ambiance. The volume of the noise was set to 60 dB which is at similar levels to actual ambient noise in an airport. Our results in Fig. 16(a) indicate that the ambient noise levels do not affect the accuracy in comparison to a clean environment.

Robustness to Head and Body Motion: In a natural setting, a user may consume drinks while walking (in a party), or moving the head while talking to others. To evaluate robustness under such conditions, we test the accuracy under the following conditions: (i) Turning head to the right (ii) Turning head to the left (iii) Walking (iv) Static. Depicted in Fig. 16(b), our results indicate that the accuracy is stable across conditions of mobility, close to static conditions.

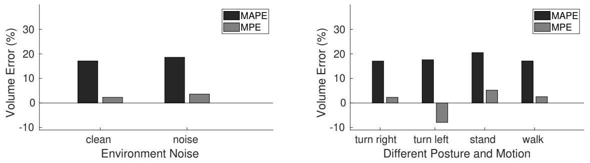

Fig. 16. Robustness to (a) Ambient Noise (b) Body Motion

Effectiveness of Semi-Supervised Learning: Fig. 17(a) depicts median and 90^{th} percentile MAE for for the following three cases: (i) without data augmentation or semi-supervision *(base)* (ii) with one iteration of teacher-student training *(Iteration-1)* (iii) with two iterations of teacher-student training *(Iteration-2)*. Evidently, the semi-supervision in *LiquidMeter* creates robust ML models bringing down the error substantially from 10.4 mL to 3.91 mL and 3.35 mL after the first and second iterations respectively. Similarly, the 90^{th} percentile errors are cut down from 21.19 mL to 7.54 mL and 6.50 mL with the two iterations. Additional iterations beyond this result in only marginal gains.

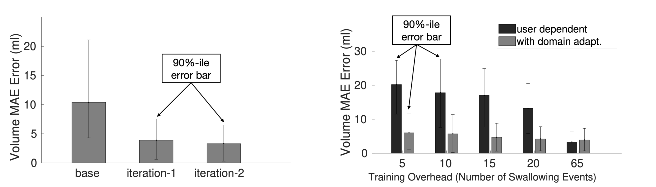

Fig. 17. Accuracy Variation with (a) Iterations of Semi-Supervision (b) Size of Training Data

Training Overhead: Fig.17(b) shows the median and 90^{th} percentile MAE as a function of the size of training data. Because of data augmentation and semi-supervised learning techniques (Sec. IV) incorporated in *LiquidMeter*, even with the *user dependent* model, we only need 65 instances of drinking (swallowing) events to converge and achieve an MAE of 3.12 mL. While we believe, the *user dependent* model does not have a big training overhead, *domain adaptation* can further cut down the number of training instances (swallow events) to 15 with MAE levels (3.35 mL) close to the *user dependent* model.

Latency and Power Consumption: Fig. 18(a) depicts the latency of executing ML models in *LiquidMeter* on smartphone devices. Evidently, the latency for both activity classification and volume estimation is very low. For profiling the energy of the TensorflowLite model, we use Batterystats and Battery Historian [5] tools. We compare the difference in power between two states: (i) The device is idle with the screen on. (ii) The device is making inferences using the TensorflowLite model. Fig. 18(b) depicts a low power consumption profile for both activity detection and volume estimation modules. The earphone sensor draws 55uA at 1.8V, thus consuming a small power [67].

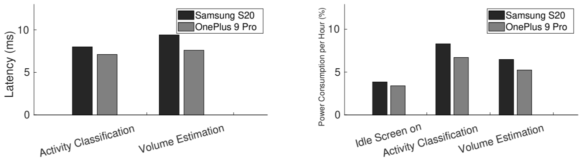

Fig. 18. System evaluation (a) Latency (b) Power Consumption

VI. Related Work

Vision: Work in [27] uses a depth camera installed on the ceiling to detect activities such as eating and drinking by using Artificial neural Networks (ANN) on the depth + RGB images. Work in [22] uses Kinect cameras on a mobile robot to detect various activities in daily life including drinking. The information from depth, color, and optical flow, etc., is integrated into a 3D CNN algorithm for performing activity detection. Work in [50] performs eating and drinking activity classification by exploiting both spatial as well as temporal variation of the human body pose detected by a camera. Linear Discriminant Analysis (LDA) is used to map the input to a compact feature representation space so as to facilitate classification based on clustering algorithms in the feature space. Similarly, the work in [19] uses skeleton structures extracted from two Kinect sensors to detect eating and drinking activities in a meeting room. While vision based sensing can be passive, they can be susceptible to occlusion and lighting conditions. Also, all of the above works only perform detection of drinking but do not estimate the actual volume of the content consumed. In contrast, *LiquidMeter*'s solution is more ubiquitous while being able to continuously estimate the volume of fluid consumption.

Smart Surface: Work in [21] designs a diet aware dining table where the surface of the table is embedded with RFID and weight detection sensors for tracking food and drink consumption. Work in [83] designs a smart table embedded with a fine grained pressure sensing textile matrix as well as a weight sensitive tablet. Various fine grained activities such as cutting, spooking, poking, etc., can be detected for monitoring eating and drinking related actions. Similarly, work in [61] uses a table with embedded weighing scale to detect bits of eating and drinking. In contrast to above works, a portable sensing mat [76] with embedded sensors is designed to track food and drink consumption. While sensor embedded surfaces can track fine grained eating and drink related activities, the sensing is only restricted to the surface, and the food/drink has to be placed on the smart surface before and after consumption. In contrast, *LiquidMeter* provides a more ubiquitous solution.

Inertial Sensors: Inertial sensors worn on the wrist and the head are explored in [77] for performing drinking activity detection. Work in [13] first uses a waist worn inertial sensor to determine whether a person is sitting, standing, and moving. This information is fused with wrist worn inertial sensors to detect eating and drinking activities. Similarly, work in [14] uses a single wrist watch inertial sensor worn on the dominant hand for eating and drinking detection. The above works do not perform drinking volume estimation. Works in [41], [46] estimates drinking volume using smart watch sensors. However, the subject has to remain relatively still while drinking and has to place the bottle down before beginning another bout of drinking. In contrast, *LiquidMeter*'s solution is more generic with ability to estimate the volume under adhoc conditions including mobility.

Sensors on the Throat and Neck: Capacitive sensors have been used on the throat and neck for classifying activities such as chewing, swallowing, speaking, and sighing [24], [25]. Bioimpedance and pressure sensor have similarly been used for detection of swallow events [82]. EMG sensors in combination with throat microphones have been used for detecting swallow events as well as classifying volume consumption into three categories: low, medium, and high [12]. Electroglottography (EGG) [37] signals have also been used for detecting food intake. In contrast to these works that detect and classify activities at a higher granularity, *LiquidMeter* tracks fluid intake at a finer granularity of milliliters.

Radio Frequency (RF) based Sensing: RFID sensors are attached at the bottom of the liquid container for detecting drinking events [51]. UltraWideBand [32] and RFID sensors [73] have been used to detect liquids placed in a container based on properties of RF reflections. Several RFID tags are attached to the container and the volume level in the container is estimated at a resolution of 35 mL based on properties of signals received from these tags by a RFID reader [54]. While these works are innovative in nature, the range of RF based solutions is limited to the environment in which the infrastructure is installed. In contrast to these works, *LiquidMeter* offers an order of magnitude higher accuracy levels, while being fully ubiquitous.

Smart Container: A number of smart containers have been made commercially available for estimation of fluid volume consumption [34], [40], [47], [48], [65], [71]. HydrateSpark [48] uses capacitive and IMU sensors to estimate the volume. The data syncs with a smartphone app using Bluetooth and the smart container includes an LED based reminder if the user has not consumed enough fluid. On the other hand, H2OPal [40] uses load cells and IMU sensors at the bottom of the device. Any container with similar size can be inserted into the device for tracking. The Thermos Smart Lid [71] includes sensors in the lid to measure liquid levels and the temperature. In addition to tracking volume consumption, the Ozmo smart bottle [65] can also differentiate between coffee and water. While effective in tracking, the user has to use the same container for drinking any fluid. In contrast, *LiquidMeter*'s solution works with any container.

Activity Detection using Earphones: Earphone sensors are gaining in popularity with a number of applications in emotion sensing, dental hygiene detection, smart health, and augmented reality [26]. The feasibility of sensing eog (eye), eeg (brain), and emg (facial muscles) signals at earphone electrodes has been explored in LIBS [64] for applications in healthcare. EarFS [62] and ECTF [11] show the feasibility of detecting facial expressions by detects electrical signals in earphones as well as embedded microphones. Perhaps, closest to our work related to food consumption is [57], where earphone sensors are used for classifying activities such as head nodding, speaking, eating, etc. They do not estimate liquid volume consumption. In contrast to above works, to our best knowledge, *LiquidMeter* is the first work that uses earphone sensors for drinking volume estimation.

VII. Discussion: Limitations and Future Work

Wireless Streaming of Earphone Sensor Data: Our platform developed by *Sonion* currently does not support wireless streaming of sensor data, the earphones need to be connected to the smartphone's audio port. We believe providing wireless streaming of sensor data will improve the usability of the system. While *LiquidMeter* shows the feasibility of sensing, we plan to incorporate this feature in our future work.

Food Classification and Quantification: We plan to extend this work towards the identification and quantification of food intake. Diverse classes of food such as bread, rice, pizza, salad, etc might generate different sound patterns at the earphones. Analysis of such information might offer valuable insights about food intake which is a part of our future investigation.

Earphone Wearability for Whole Day: One of the limitations of *LiquidMeter* is that it can only track drinking volume when the users are wearing earphones. Given the rise in popularity of smart earphones, particularly in the context of healthcare applications such as monitoring of respiration rate, blood pressure, and sleep stages, etc, we believe earphones in the future could be suitable for long-term wearability.

VIII. Conclusion

Ensuring proper hydration levels is critical for the smooth functioning of the human body. Towards this end, this paper presented a system called *LiquidMeter* that shows the feasibility of estimating the volume of fluid intake using smart earphones. The bone conduction sensor in the earphones picks up acoustic vibrations during fluid intake which is analyzed for estimating the volume of intake. While the lack of large-scale training data is a challenge, *LiquidMeter* builds robust ML models by designing techniques based on data augmentation and semi-supervised learning. Extensive measurement based evaluation across diverse users depicts an accuracy of 3.35 mL ($\approx$ 19.17% error) over commonly consumed liquids. Furthermore, the accuracy is robust to sensor mounting positions, body, and head motion, ambient acoustic interference, container type, temperature, etc. While the results are promising, we believe there are additional opportunities to be explored in the space of smart healthcare such as food intake monitoring, nutrient, and calorie estimation, etc.

IX. Acknowledgements

This research was partially supported by NSF grants: CAREER-2046972, CNS-1909479, CNS-2008384. We are also thankful to Sonion for providing us with the earphone hardware.

References

[1] Food network™ precision digital kitchen scale. https://www.kohls.com/product/prd-3758577/food-network-precision-digital-kitchen-scale.jsp.

[2] Hyoid bone. https://biologydictionary.net/hyoid-bone/.

[3] ics40619 datasheet. https://product.tdk.com/system/files/dam/doc/product/sw_piezo/mic/mems-mic/data_sheet/ics-40619-datasheet.pdf.

[4] Pitch and frequency. https://www.physicsclassroom.com/class/sound/Lesson-2/Pitch-and-Frequency.

[5] Profile battery usage with batterystats and battery historian. https://developer.android.com/topic/performance/power/setup-battery-historian.

[6] Study finds inadequate hydration among u.s. children. https://www.hsph.harvard.edu/news/press-releases/study-finds-inadequate-hydration-among-u-s-children/.

[7] What does it mean when dehydration becomes long-term and serious? https://www.healthline.com/health/chronic-dehydration.

[8] Survey of 3003 americans. *Nutrition Information Center, New York Hospital-Cornell Medical Center* (1998).

[9] Abadi, M., et al. Tensorflow: A system for large-scale machine learning. In *OSDI* (2016), pp. 265–283.

[10] Aboofazeli, M., et al. Analysis of temporal pattern of swallowing mechanism. In *IEEE EMBC* (2006).

[11] Amesaka, T., et al. Facial expression recognition using ear canal transfer function. In *ACM ISWC* (2019).

[12] Amft, O., et al. Methods for detection and classification of normal swallowing from muscle activation and sound. In *2006 Pervasive Health Conference and Workshops* (2006), IEEE, pp. 1–10.

[13] Anderez, D. O., et al. A hierarchical approach in food and drink intake recognition using wearable inertial sensors. In *ACM PETRA* (2018).

[14] Anderez, D. O., et al. Eating and drinking gesture spotting and recognition using a novel adaptive segmentation technique and a gesture discrepancy measure. *Expert Systems with Applications* (2020).

[15] Armstrong, L. E., et al. Mild dehydration affects mood in healthy young women. *The Journal of nutrition* (2012).

[16] Bertero, M., et al. The stability of inverse problems. In *Inverse scattering problems in optics*. Springer, 1980, pp. 161–214.

[17] Bone conduction headsets ("bonephones") research. http://sonify.psych.gatech.edu/research/bonephones/.

[18] Bone conduction: How it works. http://www.goldendance.co.jp/English/boneconduct/01.html.

[19] Brena, R. F., et al. Activity recognition in meetings with one and two kinect sensors. In *Mexican Conference on Pattern Recognition* (2016).

[20] Brown, F., et al. Calculating the optimum temperature for serving hot beverages. *Burns 34*, 5 (2008), 648–654.

[21] Chang, K.-h., et al. The diet-aware dining table: Observing dietary behaviors over a tabletop surface. In *IEEE PerCom* (2006).

[22] Chang, M.-J., et al. A vision-based human action recognition system for moving cameras through deep learning. In *ACM SPML* (2019).

[23] Chang, W.-G., et al. Domain-specific batch normalization for unsupervised domain adaptation. In *IEEE CVPR* (2019).

[24] Cheng, J., et al. Active capacitive sensing: Exploring a new wearable sensing modality for activity recognition. In *IEEE PerCom* (2010).

[25] Cheng, J., et al. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *UbiComp Adjunct* (2013).

[26] Choudhury, R. R. Earable computing: A new area to think about. In *ACM HotMobile* (2021).

[27] Cippitelli, E., et al. Unobtrusive intake actions monitoring through rgb and depth information fusion. In *IEEE ICCP* (2016).

[28] Cohen, R., et al. Fluid intake monitoring systems for the elderly: A review of the literature. *Nutrients* (2021).

[29] de Lima Nunes, E., et al. Swallowing acoustic characteristics of time, intensity, and frequency in healthy adults. *Open Journal of Otolaryngology* (2019).

[30] Deng, J., et al. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR* (2009).

[31] Devlin, J., et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[32] Dhekne, A., et al. Liquid: A wireless liquid identifier. In *ACM MobiSys* (2018).

[33] Drewnowski, A., et al. Water and beverage consumption among adults in the united states: cross-sectional study using data from nhanes 2005–2010. *BMC public health* (2013).

[34] Drinkup smart water bottle review. https://the-gadgeteer.com/2018/05/02/drinkup-smart-water-bottle-review/, 2021.

[35] Global wireless headphones market to reach $45.7 billion by 2026. https://www.globenewswire.com/news-release/2021/07/29/2270984/0/en/Global-Wireless-Headphones-Market-to-Reach-45-7-Billion-by-2026.html, 2021.

[36] Overview of the esophagus. https://www.merckmanuals.com/home/digestive-disorders/esophageal-and-swallowing-disorders/overview-of-the-esophagus.

[37] Farooq, M., et al. A novel approach for food intake detection using electroglottography. *Physiological measurement* (2014).

[38] Ganio, M. S., et al. Mild dehydration impairs cognitive performance and mood of men. *British Journal of Nutrition* (2011).

[39] Google. Deploy machine learning models on mobile and IoT devices. "https://www.tensorflow.org/lite", 2019.

[40] H2opal smart water bottle hydration tracker. https://www.h2opal.com/, 2021.

[41] Hamatani, T., et al. Fluidmeter: Gauging the human daily fluid intake using smartwatches. *ACM IMWUT* (2018).

[42] Han, X., et al. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing* (2017).

[43] He, K., et al. Deep residual learning for image recognition. In *IEEE CVPR* (2016).

[44] Henry, P., et al. Bone conduction: Anatomy, physiology, and communication. Tech. rep., Army research lab aberdeen proving ground md human research and engineering, 2007.

[45] Honda, T., et al. Characterization of swallowing sound: preliminary investigation of normal subjects. *PloS one 11*, 12 (2016), e0168187.

[46] Huang, H.-Y., et al. Fluid intake monitoring system using a wearable inertial sensor for fluid intake management. *Sensors 20*, 22 (2020), 6682.

[47] Hydracoach intelligent water bottle. https://www.walmart.com/ip/HydraCoach-Intelligent-Water-Bottle/24074164, 2021.

[48] Smart water bottle hydratespark bluetooth water. https://hidratespark.com/, 2021.

[49] Ioffe, S., et al. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[50] Iosifidis, A., et al. Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes. In *IEEE ICASSP* (2012).

[51] Jayatilaka, A., et al. Real-time fluid intake gesture recognition based on batteryless uhf rfid technology. *Pervasive and Mobile Computing* (2017).

[52] Kingma, D. P., et al. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[53] Kinnunen, T., et al. Voice activity detection using mfcc features and support vector machine. In *SPECOM 2007*.

[54] Kreutzer, J. F., et al. Radio frequency identification based detection of filling levels for automated monitoring of fluid intake. In *IEEE Robio* (2014).

[55] Krizhevsky, A., et al. Imagenet classification with deep convolutional neural networks. In *NIPS* (2012).

[56] Kwon, H., et al. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *arXiv preprint arXiv:2006.05675* (2020).

[57] Laporte, M., et al. Detecting verbal and non-verbal gestures using earables. In *ACM UbiComp and ISWC* (2021).

[58] Liu, Y., et al. Structured knowledge distillation for semantic segmentation. In *IEEE CVPR* (2019).

[59] Mai, F. M. *Diagnosing genius: The life and death of Beethoven*. McGill-Queen's Press-MQUP, 2007.

[60] Mancini, M., et al. Boosting domain adaptation by discovering latent domains. In *IEEE CVPR* (2018).

[61] Mattfeld, R. S., et al. Measuring the consumption of individual solid and liquid bites using a table-embedded scale during unrestricted eating. *IEEE BHI* (2016).

[62] Matthies, D. J., et al. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *ACM CHI* (2017).

[63] Nawaz, W., et al. Classification of breast cancer histology images using alexnet. In *ICIAR* (2018), Springer.

[64] Nguyen, A., et al. A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *ACM SenSys* (2016).

[65] Smart water bottle that integrates with fitbit: Ozmo. https://www.ozmo.io/, 2021.

[66] Qu, C., et al. Bert with history answer embedding for conversational question answering. In *ACM SIGIR* (2019).

[67] Sonion voice-pick-up-vpu slides. https://invensense.tdk.com/wp-content/uploads/2018/10/Sonion-Voice-Pick-Up-VPU-Sensor-Paul-Clemens.pdf.

[68] Sonion vpu sensor. https://www.sonion.com/vpu-voice-pick-up-sensor/.

[69] Dysphagia. a difficult diagnosis to swallow! https://xavier.org.au/resources/news/2021/march/214/dysphagia_a_difficult_diagnosis_to_swallow, 2021.

[70] Tagliasacchi, M., et al. Seanet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095* (2020).

[71] Thermos connected hydration bottle with smart lid review. https://www.pcmag.com/reviews/thermos-connected-hydration-bottle-with-smart-lid, 2021.

[72] Wager, S., et al. Dropout training as adaptive regularization. In *Advances in neural information processing systems* (2013).

[73] Wang, J., et al. Tagscan: Simultaneous target imaging and material identification with commodity rfid devices. In *ACM MobiCom* (2017).

[74] Wang, J., et al. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE TVCG* (2019).

[75] Watanabe, S., et al. Student-teacher network learning with enhanced features. In *IEEE ICASSP* (2017).

[76] Watanabe, T., et al. A portable sensor sheet for measuring the eating pace in meal assistance care. In *IEEE EMBC* (2019).

[77] Wellnitz, A., et al. Fluid intake recognition using inertial sensors. In *iWOAR* (2019).

[78] Xie, Q., et al. Self-training with noisy student improves imagenet classification. In *IEEE/CVF CVPR* (2020).

[79] Yao, S., et al. Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In *The World Wide Web Conference* (2019).

[80] Zhang, H., et al. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243* (2019).

[81] Zhang, N., et al. Effects of dehydration and rehydration on cognitive performance and mood among male college students in cangzhou, china: A self-controlled trial. *IJERPH* (2019).

[82] Zhang, R., et al. A generic sensor fabric for multi-modal swallowing sensing in regular upper-body shirts. In *ACM ISWC* (2016).

[83] Zhou, B., et al. Smart table surface: A novel approach to pervasive dining monitoring. In *IEEE PerCom)* (2015).

[84] Zhou, Z., et al. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE CVPR* (2017).

[85] Zoph, B., et al. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882* (2020).