Robust Robotic Pouring using Audition and Haptics

Hongzhuo Liang[1*], Chuangchuang Zhou[1,2], Shuang Li[1], Xiaojian Ma[3],
Norman Hendrich[1], Timo Gerkmann[4], Fuchun Sun[5], Marcus Stoffel[2], and Jianwei Zhang[1]

Abstract— Robust and accurate estimation of liquid height lies as an essential part of pouring tasks for service robots. However, vision-based methods often fail in occluded conditions while audio-based methods cannot work well in a noisy environment. We instead propose a multimodal pouring network (MP-Net) that is able to robustly predict liquid height by conditioning on both audition and haptics input. MP-Net is trained on a self-collected multimodal pouring dataset. This dataset contains 300 robot pouring recordings with audio and force/torque measurements for three types of target containers. We also augment the audio data by inserting robot noise. We evaluated MP-Net on our collected dataset and a wide variety of robot experiments. Both network training results and robot experiments demonstrate that MP-Net is robust against noise and changes to the task and environment. Moreover, we further combine the predicted height and force data to estimate the shape of the target container.

I. INTRODUCTION

Pouring a specific amount of liquid into a container is an important manipulation skill for service robots in applications such as bartending or housekeeping and also in industrial environments. Compared to humans who can effortlessly pour the liquid into a container without spilling, predicting the liquid height in the target container is still challenging for a robot. For the robust and accurate perception of robot pouring, recent approaches mainly rely on vision [1]–[3]. They take the RGB image as input, then infer the liquid height, liquid amount, even pouring trajectories. However, vision fails under occlusions or in the dark. Using audio [4]–[6] is a second option for robot pouring perception, but the performance of this method will degrade in a noisy environment. Besides, if the robot is equipped with force/torque sensing, the amount of liquid poured out from the source container can be measured directly [7], but the liquid height in a target container can only be predicted if the initial fill level and the shape of the container are known. These drawbacks in existing perception methods suggest that to estimate the liquid height precisely, it is necessary to find a robust approach that still works in conditions where one modality becomes incomplete or deteriorated.

*Corresponding author, email: liang@informatik.uni-hamburg.de
[1]Group TAMS, Dept. of Informatics, Universität Hamburg.
[2]Institute of General Mechanics, RWTH Aachen University.
[3]Group VCLA, Dept. of Statistics, University of California, Los Angeles.
[4]Signal Processing (SP), Dept. of Informatics, Universität Hamburg.
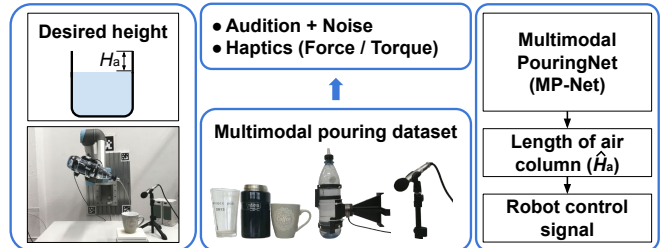[5]Dept. of Computer Science and Technology, Tsinghua University.

Fig. 1. Our multimodal pouring pipeline. (Left) The goal of our pouring task is to pour the desired amount of liquid into a target container. The generated sounds depend on the height of the remaining air column (H_a) in the container. (Center) We collected a multimodal robotic pouring dataset, which contains 300 pouring sequences with audio and force/torque sensor data for training our network. A scale recorded the weight of the poured liquid. We then calculated H_a to serve as ground truth. (Right) By giving the audio and haptic as input, our Multimodal PouringNet (MP-Net) predicts $\hat{H}_a$ as output. This output is then used to control the robotic pouring.

Therefore, we propose to tackle the issues of robust robotic pouring by a multimodal perception method. This multimodal system should be robust in a wild environment, for example, against different levels of noise, types of noise, or light conditions. The perception system should be able to generalize to different target containers, liquid types, and initial liquid height in the target containers.

In this paper, we develop a novel multimodal neural network called MP-Net that takes preprocessed audio spectrograms and force/torque data as the input and estimates the length of the air column in the target container. This idea is inspired by human experiences, as humans can infer whether the target container is almost full from the pouring sound and use their proprioceptive haptic feedback to estimate how much liquid is poured out from the source container [8], [9]. Fig. 1 illustrates the proposed perception pipeline.

The main structure of MP-Net is a recurrent neural network that implicitly integrates the prior knowledge that the liquid level rises monotonically during the pouring. A multimodal pouring dataset with audio and haptics was built up to train this model. To further improve the adaptability to the noisy environment of MP-Net, we augmented the audio data by adding different noise levels of robot noise. The benefits of force/torque data and audio data augmentation are verified in both our dataset and a wide range of robotic experiments, respectively.

To sum up, our main contributions are:

1) We propose a multimodal neural network MP-Net to estimate the liquid height robustly and to enhance the precision and generalization ability of robotic pouring. Not only the haptic input but also the audio data

augmentation facilitate to produce a robust model that can generalize broadly.

2) System assessment and comparison across eight robotic experiments, *e.g.*, pouring water into different containers, in various noisy environments with different noise levels and noise sources, revealed that high accuracy could be achieved by our proposed method.

3) The shape of the target container is generated by combining the force data and realtime height prediction.

II. RELATED WORK

Visual sensing for robotic pouring. Vision remains the most popular modality in robot pouring perception, as humans also highly rely on vision to perform the same task. Pithadiya *et al*. [10] compared several edge detection techniques for the filling height inspection of convex target containers. Schenck *et al*. [11] used a thermal camera to generate pixel-level ground truth data of heated water. The RGB camera is then used to determine the water volume using both a model-based method and a neural network method. This method suffers from poor estimation accuracy due to the varied liquid types and the complex shapes of the water flow. To predict the absolute height of the liquid, Dong *et al*. [12] used a point cloud to model the target container and a proportional-derivative controller to perform the pouring action. Do *et al*. [13] utilized a Kalman filter dealing with opaque and transparent liquids based on an RGB-D camera. However, the mean height errors for ten pourings of three transparent liquids were larger than 4 mm. Recently, [14], [15] deployed reinforcement learning methods for generating motion trajectories during the pouring process.

Audio sensing for robotic pouring. The auditory signal generated when liquid or granular materials interact with other objects also comprises reliable clues, such as resonance frequency and vibration behavior of the liquids and the air. Griffith *et al*. [16] indicated that auditory and proprioceptive data enhance the classification tasks for the interaction between objects and water. Clarke *et al*. [4] used audio-frequency vibration generated by shaking a granular material to evaluate the weight that is poured out. However, the weight of the poured out granular materials does not allow us to estimate the filling height of the target container.

None of the above work explored the fact that the audio vibration of the air in the target container can be analyzed to solve the liquid height regression in robotic pouring. In our previous work [6], we used audio as input and trained an RNN based network to directly predict the length of the air column, including generalization to different target containers. However, the network failed in a noisy environment or when pouring liquids with high viscosity as the audio signal is weak in these conditions.

Haptic sensing for robotic pouring. Haptic sensing, especially force and torque sensing, is also an essential signal of robotic pouring. The goal of learning robust robot pouring from human demonstrations including hand trajectories and force/torque data is explicitly discussed by Huang and Sun [17], as part of their multi-task dataset on human daily
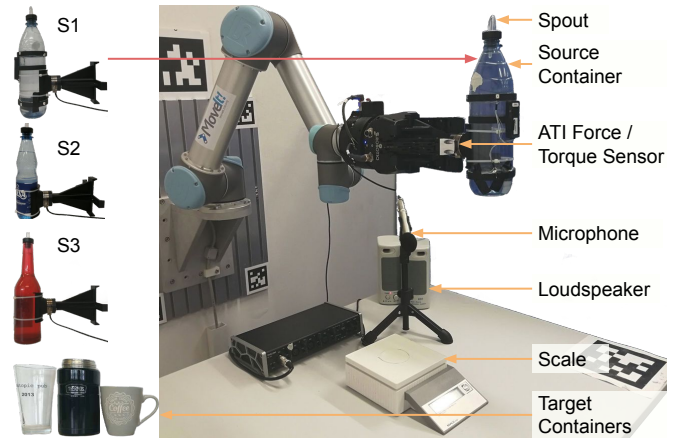

Fig. 2. Pouring setup used to collect our multimodal dataset and to perform the robot experiment. The three target containers used to collect the dataset are shown in the bottom left corner. We put these containers on a calibrated scale to get the realtime label of the liquid height. A force/torque sensor is mounted on the source container to collect force and torque information while pouring, and a microphone was set on the table to collect the pouring audio. Loudspeakers were used to create ambient noise for robot experiments. S1, S2, S3 in the left represent three different source containers where S1 is the container used to collect dataset, S2, and S3 are novel source containers.

interactive manipulation. Rozo *et al*. [18] used a parametric hidden Markov model to retrieve joint-level commands given the force/torque inputs from the human demonstration. Saal *et al*. [19] examined the viscosity estimation of various liquids inside a bottle from tactile sensory data. Matl *et al*. [20] utilized a force sensor mounted at the end of a robotic arm with a container grasped by a gripper to get the change of wrenches, and they used a physics-based model to estimate the mass and volume of the liquid. In the above three papers, perception happens before the pouring action begins. Although the force signal from the pouring container can explicitly represent the volume of the poured-out liquid, it cannot measure the liquid height in an unseen target container.

Multimodal fusion for robotic pouring. Recent studies have shown that multimodal sources represent the environmental features better than single modality in pouring tasks [21] as weel as other tasks [22]. Wu *et al*. [23] presented a hierarchical long short-term memory [24] model, which could detect if a pouring sequence was successful or failed. This model is based on a pouring dataset, including visual sequences and IMU data. However, this work only classifies whether a pouring action was successful but cannot predict the precise height of the poured-out liquid. Wilson *et al*. [5] implemented a multimodal convolutional neural network to fuse audio and visual data to predict the weight of the poured liquid, detect overflow and to classify the liquid and the target container. Park *et al*. [25] proposed another interesting robot application for anomaly detection during robot manipulation using haptic, visual, auditory, and kinematic sensing.

Based on how humans rely on the correlations of haptic and audio while pouring, in this paper, we take advantage of audio and force/torque data as the input to promote the robustness of robotic pouring.

III. MULTIMODAL POURING DATASET

A. Dataset Collection Setup

To create the multimodal dataset for the height estimation task, we designed a robot pouring setup, as shown in Fig. 2. It consists of a UR5 robot arm with a Robotiq 3-finger hand, and a custom 3D-printed bottle holder with an embedded ATI force/torque sensor. A standard plastic bottle (S1) is used as the source container, with an optional bottle spout to limit liquid flow. The target container is placed on a MAULlogic digital scale which measures the combined weight of the container and the liquid inside. Audio is recorded with a microphone in an environment with only UR5 robot ego-noise. The dataset setup contains three different target containers with three different materials and heights. The properties of these containers are shown in Table I with ID 1-3.

Unlike our previous work that used human pouring sequences as training data [6], we use the robot to perform the pouring task. The main reason is that the haptic data coming from human pouring is massively different from the robot pouring data, which makes it hard to transform human pouring to robotic pouring.

We set a fixed pouring trajectory and place the mouth center of the source container 310 mm above the scale plane. In this manner, we collected 100 trials each for three different target containers. For each trial, we recorded both audio and force/torque data starting just before the scale reading begins to change, and stopping after the reading of the scale becomes stable again. The lengths of one pouring recording varied from 24-40 seconds according to the capacity of the different target containers.

B. Data Analysis

Audio-frequency data. To better extract audio information from the microphone, we resampled all audio data from 44.1 kHz to 16 kHz and computed spectrograms with a window length of 32 ms and 50% overlap without zero-padding. For network training, we randomly chose 4 seconds of audio clips from one complete pouring audio sequence. Using fixed length training samples allows for batch processing for better GPU performance, and random starts correspond to different initial liquid amounts, so that the network can generalize to partially filled containers.

Force/Torque data. The force/torque sensor we use is ATI Mini45 (500 Hz). As the weight of the source container is far below the maximum payload of this sensor (580 N in Fx/Fy), the raw data is rather noisy but becomes usable after simple low-pass filtering. See Fig. 3 for typical raw force/torque data (in N/Nm) during a sample pouring sequence. In each sub-figure, we plot the noisy raw signal from the sensor shown as a light color together with a low-pass filtered signal shown as a dark color. The change of the bottle weight during pouring can clearly be seen in the force/torque readings for the x- and y-axis, while the z-axis measurements show only little change. This is due to the specific orientation of the sensor in our 3D-printed holder, dictated by the cable routing. While the sensor z-axis is almost horizontal (orthogonal to gravity),
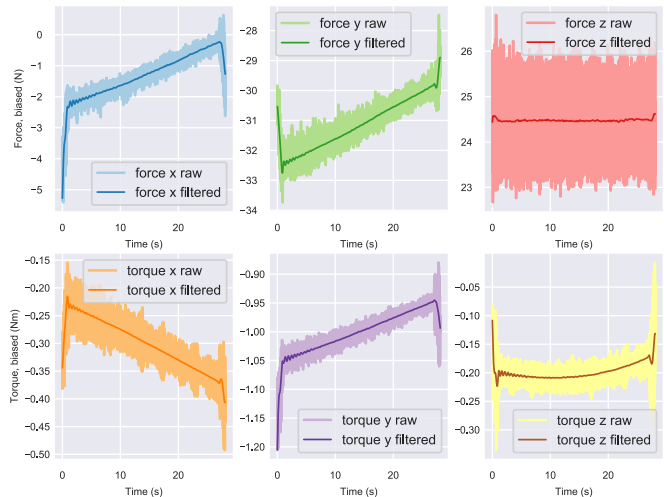

Fig. 3. A sample of force/torque data (N/Nm) in our pouring dataset. In each sub-figure, we show the raw sensor data (light line) and the filtered data using a Butterworth low pass filter (dark line). See the text for details.
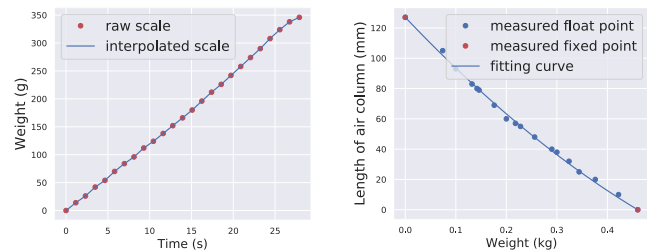

Fig. 4. (a, left) Sample scale readings of a pouring sequence in our dataset from bottle ID 1. Each red dot is one weight measurement from the scale. The blue line is the interpolated curve. (b, right) Converting liquid weights (kg) into length of the air column (mm) for bottle ID 1. The dotted points are the manually measured data, where the red points indicate the empty and completely filled container. The blue line is the fitting curve that calculates from the scale reading to the length of the air column.

the x- and y-axes of the sensor are pointing diagonally upwards, so that the raw sensor readings actually increase when pouring from the source container. The pouring motion from the robot slightly rotates the bottle around the z-axis.

Scale data. To get the ground truth of the pouring perception problem, we used a MAULlogic digital scale to measure the weight of the target container (accuracy ± 2 g). Because the publishing frequency of the scale is only 1 Hz, we deployed a linear interpolation (Fig. 4(a)) to get realtime scale readings. To convert the measured weights into the heights H_a of the air column needed for network training, we sampled 10-15 random amounts of liquid for each target container, up to full container capacity, and manually measured the length of the air column. A polynomial curve fitting is utilized using this data to calculate the approximate air column length for different amounts of liquid and corresponding measured weights (Fig. 4(b)).

C. Audio Data Augmentation

Considering that the variability of the audio data is significant enough, models trained on it will better generalize in different audio conditions. Therefore, we augmented the audio data in our dataset by adding noise. We recorded an ego noise from a humanoid PR2 robot, which is commonly
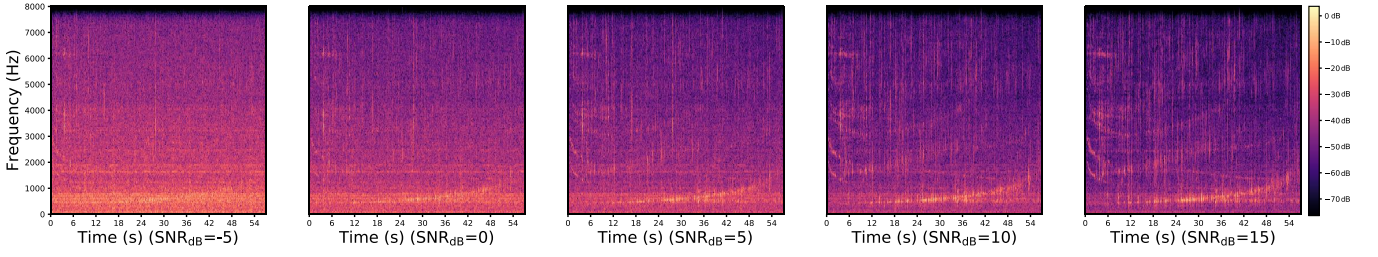
Fig. 5. Examples of audio spectrograms that add different noise levels of the same audio signal in our dataset. While the $\mathrm{SNR_{dB}}$ is high (*e.g.*, $\mathrm{SNR_{dB}} = 15$, right), which indicates the audio signal is rather clear, we can clearly see rising frequency curves between 256 Hz-2048 Hz. But when the $\mathrm{SNR_{dB}}$ is low (*e.g.*, $\mathrm{SNR_{dB}} = -5$, left), there is no meaningful structure of the resonance frequency.

used for research in household environments. In detail, we employed the signal-to-noise ratio in decibels ($\mathrm{SNR_{dB}}$) as our reference on how much noise was added.

$$\mathrm{SNR_{dB}} = 10 \log \left(\frac{A_{\mathrm{signal}}}{\alpha_{\mathrm{nf}} \times A_{\mathrm{noise}}} \right)^2 \qquad (1)$$

$$\alpha_{\mathrm{nf}} = \frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}} \times 10^{\frac{\mathrm{SNR_{dB}}}{-20}} \qquad (2)$$

$$\mathrm{audio} = \mathrm{signal} + \mathrm{noise} \times \alpha_{\mathrm{nf}} \qquad (3)$$

Equations (1) - (3) illustrate how noise was added, where A_{signal} and A_{noise} are the root mean square (RMS) amplitude for the recorded pouring audio and noise, respectively. α_{nf} is the ratio of the input noise.

We mixed the original audio sequences with noise of different levels ranging from -20-20 $\mathrm{SNR_{dB}}$ with a step size of 5 dB. The number of generated audio clips, of 4 seconds length each, is proportional to the length of a pouring trial. Fig. 5 shows examples of mixed audio spectrograms for different $\mathrm{SNR_{dB}}$. When $\mathrm{SNR_{dB}} > 0$, one high-energy and rising curve between 256 Hz and 2048 Hz is clearly visible during pouring. This curve represents the resonance frequency of the air. But when $\mathrm{SNR_{dB}} \leq 0$, the spectrogram has little structure of the resonance frequency, which indicates this audio sequence contains more noise than signal.

IV. MULTIMODAL POURING NETWORK

Our goal is to design a robust network architecture to acquire the liquid filling height by making sense of audition and haptics together. Audio vibration results in the changes in the resonance frequency of the air in the target container. Force/torque changes result from the liquid poured out of the source container, which yields additional guidance about the liquid in the target container, especially when the audio signal is deteriorated.

The pouring process is a typical sequential problem and the predicted length has a temporal relationship. Intuitively, we choose a recurrent network [26] as our model architecture. The multimodal network architecture MP-Net is shown in Fig. 6. In detail, we utilize the long short-term memory (LSTM) unit to process the multimodal inputs. We mix audio clips with noise using Equation (3) and transform it into a $257 \times n$ matrix using Short-Time Fourier Transform (STFT). The haptic data is processed to a $6 \times 8 \times n$ matrix (6 F/T sensor channels, and 8 F/T samples arrive during each 16 msec STFT interval) and then concatenated with the audio
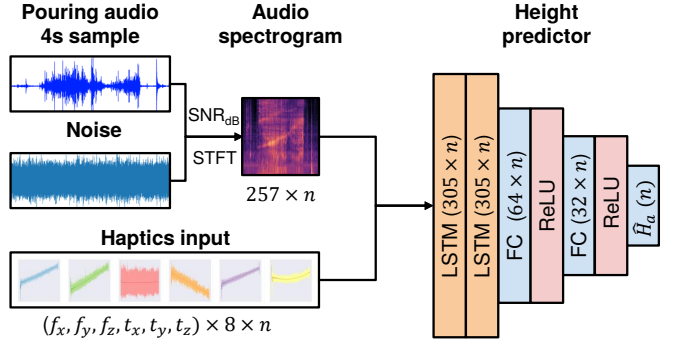

Fig. 6. MP-Net architecture. The raw audio data is transformed into a spectrogram with 257 descriptors. Then the encoder module (a recurrent neural unit) is progressively fed each time slice of the audio-frequency spectrogram and the corresponding haptic data. Finally, the height predictor module produces the 1D length of the air column of the target containers. The blue rectangle denotes a fully-connected layer followed by a batch normalization layer and a rectified linear unit. $n = 251$ for a 4 s pouring sample.

data to a $305 \times n$ matrix for input. Here, n is the number of the time slice, and we get $n = 251$ when using audio clips of 4 seconds duration from our dataset. Then each time slice of the fusion data is progressively fed into the encoder module (2-layer LSTM unit) to a layer of 56 recurrent features A_h.

Besides, it is crucial to find a well-suited ground truth in supervised learning. Due to the fact that when the length of the air column gets shorter in an organ pipe, the air vibrates faster and the resonance frequency of the air increases [27], [28], it is more indicative to choose the length of the air column H_a as the ground truth of our model instead of the liquid height. Thus, the height predictor (a 2-layer multilayer perceptron) in MP-Net takes the recurrent vector A_h as input and performs a regression of the length of the air column H_a. The height predictor is supervised with a mean squared error (MSE) loss $\mathcal{L}_{height}$

$$\mathcal{L}_{height} = \|\hat{H}_a - H_a\|^2. \qquad (4)$$

In addition, an auxiliary $\mathcal{L}_{mono}$ is introduced to enforce the decrease of the estimated length of the air column over time t

$$\mathcal{L}_{mono} = \sum_t [\max(0, (\hat{H}_{a_{t+1}} - \hat{H}_{a_t}))]. \qquad (5)$$

Overall loss. Combining with $\mathcal{L}_{height}$ and $\mathcal{L}_{mono}$, the complete training objective for MP-Net is defined by $\mathcal{L}_{mp}$

$$\mathcal{L}_{mp}(\theta) = \mathcal{L}_{height} + \alpha \cdot \mathcal{L}_{mono}, \qquad (6)$$

where α is a hyperparameter for balancing these two loss functions. In our implementation, we set it to 0.01 for the best performance via some preliminary experiments.

V. EXPERIMENT

A. MP-Net Evaluation

We examined our proposed multimodal network MP-Net against the following baseline methods: 1) MP-Net*: the same network architecture as MP-Net but trained without adding noise to the audio data, 2) AP-Net: only audio branch of MP-Net and training on augmented audio data with noise, 3) AP-Net*: same network architecture as AP-Net without audio augmentation [6], 4) FT-Net: only force/torque branch of MP-Net that only takes force/torque data as input. All the models mentioned above were trained on our whole dataset which contains pouring samples collected from three different target containers.

First, we evaluated MP-Net and four baselines using the fraction of the input sequences whose length prediction error $|\hat{H}_a - H_a|$ is below a threshold e. We trained MP-Net and AP-Net on noisy audio, with noise levels from the SNR_{dB} set $[0, 5, 10, 15, 20]$. Except for FT-Net, MP-Net and the other three baseline models were then tested on audio data with $\text{SNR}_{\text{dB}} = 5$ noise level. Fig. 7(a) shows that MP-Net has a stable advantage of about 8% over AP-Net over the full range of tested length error thresholds, which indicates that the force data provide a positive gain to the training result. The performances of MP-Net* and AP-Net* are both poor compared to MP-Net and AP-Net, respectively, which indicates that augmenting audio data also contributes a lot to the networks. FT-Net was not trained on audio and has learned to predict the air column length averaged over all containers.

To compare the robustness of MP-Net, AP-Net, FT-Net with regard to audio noise, we evaluate the percentage of input sequences whose length prediction error stays below $5\,\text{mm}$, when both the training and the testing data were augmented with the same noise levels. Each model was trained and tested on audio data augmented with 9 different SNR_{dB} levels from the set $[-20, -15, -10, -5, 0, 5, 10, 15, 20]$ respectively. Fig. 7(b) illustrates that the accuracy of MP-Net and AP-Net increases with SNR_{dB}. The accuracy curve of FT-Net is straight because force/torque data is independent of audio noise. We can see that MP-Net has at least 27% higher accuracy than AP-Net when $\text{SNR}_{\text{dB}} \leq -15$, and has 8% higher accuracy when $\text{SNR}_{\text{dB}} = 0$. Note that the performance of AP-Net degrades quickly for $\text{SNR}_{\text{dB}} \leq -15$, which indicates that the noise interferes with the audio and the network cannot gain useful information from it anymore. Nevertheless, with the help of force/torque data, MP-Net can still perform comparably to FT-Net in such a noisy situation. This phenomenon proves that MP-Net has learned to predict liquid height by conditioning on both audio and haptic input.

B. Robotic Experiments

To verify and compare the reliability and robustness of the proposed MP-Net in robotic pouring tasks, we carried
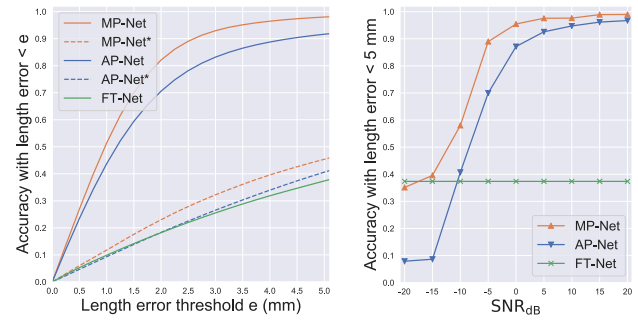

Fig. 7. Network evaluation results of MP-Net and four baselines. (a, left) The fraction of the input sequence whose length prediction error is below a threshold between five models. (b, right) The fraction of the input sequence whose length prediction error is below 5 mm between MP-Net, AP-Net, FT-Net. MP-Net and AP-Net trained and tested on nine datasets, where the only difference between each dataset is the noise level of audio data.

out eight evaluation experiments on different target containers, different pouring heights, different source containers, different noise levels, varying positions of the noise source, different initial liquid heights, different types of liquid, and different noise sources.

All experiments used both MP-Net and AP-Net to compare their real-world performance, except for the different source containers experiments. To adapt to different noisy environments, MP-Net and AP-Net were trained on noisy audio data with noise levels SNR_{dB} from 0-20. The experiment setup was the same as in our dataset collection setup and the initial water level of the source container was random. MP-Net took the latest $4\,\text{s}$ sequence of audio and haptic data. While the sequence length of the input signal was less than $4\,\text{s}$, all the input signals were fed into the network. Once the estimated length of the air column was smaller than the desired one, the robot immediately stopped pouring. To compensate for the liquid poured out during the stopping motion, a corresponding correction was applied.

1) Evaluation of Different Target Containers: We used nine different target containers for robot experiments. The first five rows in Table I illustrate the properties of these nine containers, where containers 1-3 are included in the training dataset while 4-9 are only used for testing. We kept the distance between the target containers and the microphone the same as in our original dataset. Then we set the desired length of the air column to $60\,\text{mm}$. We synthesized a $\text{SNR}_{\text{dB}} = 5$ audio signal by playing the PR2 robot noise from a pair of loudspeakers located at positions 1&1 as shown in Fig. 10(a). Here, 1&1 means that both loudspeakers were placed at position 1. We repeated the experiments five times for each container.

Fig. 8(a) displays that the absolute mean errors of the liquid height are below $8\,\text{mm}$ and the standard deviations are below $4\,\text{mm}$ for both MP-Net and AP-Net among the known target containers. Container 3 performs best for both networks due to the stainless steel material which makes the crispest sound. Furthermore, we converted the height error of each cup to a weight error, shown in rows seven and eight of Table I. We can see that AP-Net performs well on known containers and even outperforms MP-Net a

THE PROPERTIES OF NINE TARGET CONTAINERS AND CONTROLLING RESULTS OF POURING WATER INTO A DESIRED LENGTH OF THE AIR COLUMN.

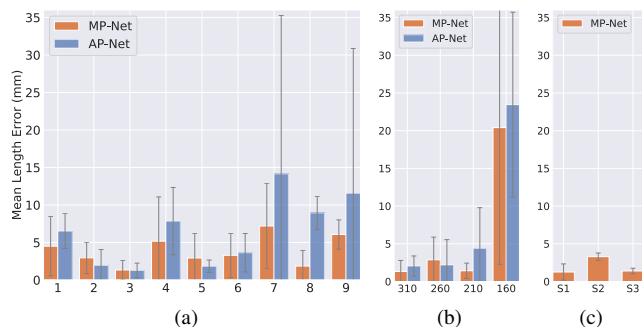| ID	1	2	3	4	5	6	7	8	9
Container									
Material	Glass	Ceramics	Steel	Ceramics	Ceramics	Plastic	Plastic	Enamel	Glass
Height (mm)	127	99	150	97	94	103	115	78	135
Volume (ml)	460	428	766	310	298	416	408	382	358
Error in volume (ml) with $\mathrm{SNR_{dB}} = 5$, $H_a = 60\,\mathrm{mm}$									
MP-Net	15.8 ± 14.0	11.5 ± 8.1	6.3 ± 6.1	16.6 ± 19.0	8.2 ± 9.3	12.8 ± 11.5	23.7 ± 18.7	8.5 ± 9.6	13.9 ± 4.4
AP-Net	23.1 ± 8.2	7.5 ± 8.2	6.0 ± 4.7	25.3 ± 14.5	5.2 ± 2.3	14.3 ± 10	46.6 ± 70.6	41.6 ± 10.4	26.6 ± 45.4

Fig. 8. In the following robot experiments, the $\mathrm{SNR_{dB}}$ were set to 5 dB. Robot experiment result (a) on pouring water into different target containers ($H_a = 60\,\mathrm{mm}$, ID 1-3 are in the training dataset, ID 4-9 are novel), (b) on four different pouring heights (pouring height 310 mm was the height for dataset collection, $H_a = 40\,\mathrm{mm}$), (c) on one known source container S1 and two novel source containers S2, S3 ($H_a = 40\,\mathrm{mm}$).

Fig. 9. Robot experiment results of the performance of MP-Net, AP-Net, and AP-Net* in environments with different levels of noise (measured by $\mathrm{SNR_{dB}}$). We evaluate with five different target liquid heights and the results are demonstrated in five different colors, respectively. The dashed lines represent the desired lengths of the air column, while the solid dots (with error bars) show the actual ones when the pouring terminates.

little on containers 2 and 3. Regarding the unseen containers, especially for container 7 and 9, MP-Net exhibits a stronger generalization ability than AP-Net in a noisy environment.

2) Evaluation of Different Pouring Heights: In this experiment, we placed the source container at four different heights from the set [310, 260, 210, 160] mm respectively. The height of the source container is the vertical distance from the mouth center of the source container to the scale plane. We carried out five robot experiments on each height using target container 2 when $\mathrm{SNR_{dB}} = 5$. Fig. 8(b) indicates that our algorithm performs well for the three higher heights but not at the lowest height. As the pouring height decreases, the volume of the pouring sound also decreases. With the source container at height 160 mm the pouring sound is too low, and haptics only is not sufficient for the network to perceive the height of the air column.

3) Evaluation of Different Source Containers: The source container influences the flow-rate and the initial force/torque data. Therefore, we tested two novel source containers S2 (44 g) and S3 (484 g), which are shown in the left side of Fig. 2, to compare with the container S1 (64 g) used for network training and all other experiments. We kept the pouring height at 310 mm, and the other experimental setup was the same as in the evaluation of the different pouring height. Fig. 8(c) suggests that different source containers hardly affect network performance.

4) Evaluation of Varying Noise Conditions: To further verify the performance of our MP-Net model in different
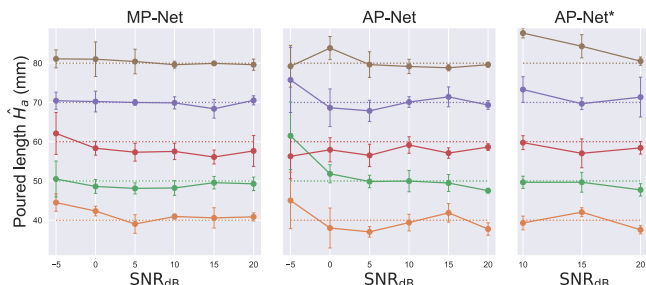
noise conditions, we implemented a set of experiments using target container 2 under six different noise levels of [-5, 0, 5, 10, 15, 20] $\mathrm{SNR_{dB}}$. The loudspeakers were again located at positions 1&1. For each $\mathrm{SNR_{dB}}$ level, we tested five different target lengths of air column, namely [40, 50, 60, 70, 80] mm. We carried out five robot experiments on each audio and target length condition.

As visualized in Fig. 9, MP-Net has a substantial advantage over AP-Net when $\mathrm{SNR_{dB}} = -5$, which further indicates the advantages of multimodal fusion. In this experiment, we also tested AP-Net* under different noise conditions. AP-Net* performed well while $\mathrm{SNR_{dB}} \geq 10$, but when $\mathrm{SNR_{dB}} < 10$, the robot either stopped pouring immediately or overfilled the target containers. Therefore we did not list the experiment results of AP-Net* tested on audio with $\mathrm{SNR_{dB}} < 10$.

5) Evaluation of Varying Positions of Noise Source: To assess whether MP-Net is sensitive to the direction of the noise source, we set up six different position combinations [1&1, 2&2, 3&3, 4&4, 2&4, 1&3] of the two loudspeakers.

The two loudspeakers played a synthetic $\mathrm{SNR_{dB}} = 5$ noise signal at each position. We used the target container 2 and a desired air column length of 40 mm. The other experimental setup was the same as in the evaluations of the different target containers. Then at each combined position of the two loudspeakers, we poured water five times. As shown in Fig. 10(b), when the loudspeakers are at position 1&1, both models perform best as the loudspeakers are behind the microphone. MP-Net generalizes better than AP-Net to
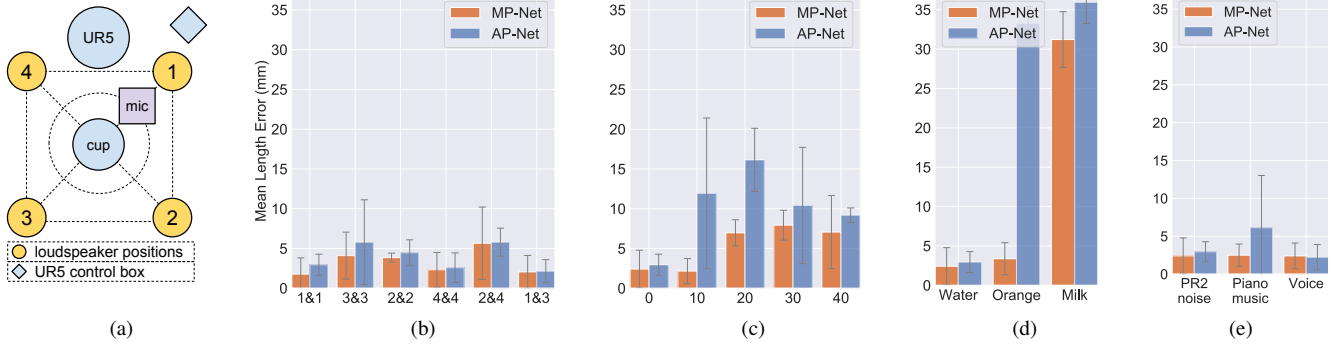
Fig. 10. (a) Schematic diagram of 4 different loudspeaker positions relative to the target containers, the UR5 robot position, and the control box of the UR5 robot. Evaluation results of (b) six combinations of two loudspeakers positions, (c) varying initial heights of the source container, (d) different types of liquids and (e) different types of noise sources. In these experiments, the SNR_{dB} was set to 5 dB. The target height H_a was set to 40 mm.

the different positions of the loudspeakers due to the lower consistent mean height error among all tested positions.

6) Evaluation of Varying Initial Liquid Height in Target Containers: In this experiment, we poured water into the target container 2, starting from five different initial liquid heights [0, 10, 20, 30, 40] mm. We tested five times from each initial level. We put two loudspeakers at 1&1 positions and kept the other test setups the same as in the evaluations of the varying direction of noise sources. The results in Fig. 10(c) demonstrate that MP-Net is again more robust than AP-Net. Force and torque data yield a meaningful indication of how much water was poured out.

7) Evaluation of Different Types of Liquid: We conducted pouring experiments with different liquids: pure water, orange juice and 1.8% fat milk. We used the same experimental setting as in the evaluations of different microphone positions and poured each type of liquid for five times. As manifested in Fig. 10(d), MP-Net can generalize to common household liquids like water and orange juice while AP-Net cannot handle the task of pouring orange juice well under $\text{SNR}_{\text{dB}} = 5$. However, similar to [6], due to the high viscosity of milk, both models cannot generate correct height prediction.

8) Evaluation of Different Types of Noise Sources: We also assessed our model with three noise types: PR2 robot noise, human voices and a continuous piece of piano music. The human voice is represented by discrete sounds of a man counting numbers in English. We poured water under each type of noise five times. All experimental settings were the same as in the evaluations of different types of liquid. Fig. 10(e) shows that MP-Net is not affected by different noise types, but the accuracy of AP-Net has a small fluctuation under a musical disturbance.

C. Shape Prediction of Target Containers

In this section, we applied MP-Net to predict the shape of symmetric target containers. In this case, the edge profile is sufficient to describe the shape of the containers, which is determined by the correlation between height and radius [29]. Fig. 11(a) shows a volume profile filled with liquid of density ρ, where $\Delta V, \Delta m, \Delta h$ are the poured liquid volume, and the weight and liquid height differences during a time interval Δt respectively. Assuming that Δt is very small, then Δm

can be calculated by approximating the shape of ΔV as a cylinder,

$$\Delta m = \rho \Delta V = \rho \pi r^2 \Delta h \tag{7}$$

We can determine Δm by the (f_x, f_y, f_z) values from the force/torque sensor and Δh through our neural network output $\hat{H}_a$. In the robot experiments, the frequency of Δm and $\hat{H}_a$ was 500 Hz and 12 Hz, respectively. To get a smooth and accurate estimation of the container shape, we used a quadratic function to fit the scatter points,

$$r(h) \approx \sqrt{\frac{\Delta m}{\Delta h} \frac{1}{\rho \pi}} \tag{8}$$

For target containers 1, 2, 4, 5, 6, 7, we conducted five trials of the experiment in which the robot pours into these target containers. We recorded the realtime estimation of $\hat{H}_a$ and force data into a rosbag. When the target container was filled to about 90% of its total height, we stopped the pouring and the recording. Using the data from these rosbags, we calculated the edge-profiles of the target containers. In Fig. 11(b), the thick black curves are the ground truth profiles, while five colored curves around black curves depict the experimental results. The magenta area in the middle of each target container visualizes the mean error of the radius prediction. As expected, the mean radius estimation error is highest for an empty container, when our recursive network cannot yet rely on its memory but stabilizes as the liquid level rises. Due to the restriction to quadratic functions, the reconstruction works best for containers with low edge curvature (such as containers 1, 7).

VI. CONCLUSION AND FUTURE WORK

In this paper, we motivate the need for combining audio and haptic information for robot pouring tasks. We recorded a robot pouring dataset that includes 300 complete robot pouring sequences with audio and force/torque data. We propose a novel audio-haptic recurrent deep network (MP-Net) trained on this dataset that predicts liquid height in realtime. The multimodal perception system is systematically tested across four baselines and a wide range of robotic pouring experiments in a noisy environment. The results substantiate that MP-Net is quite robust against noise and against changes in different tasks and varying environments.
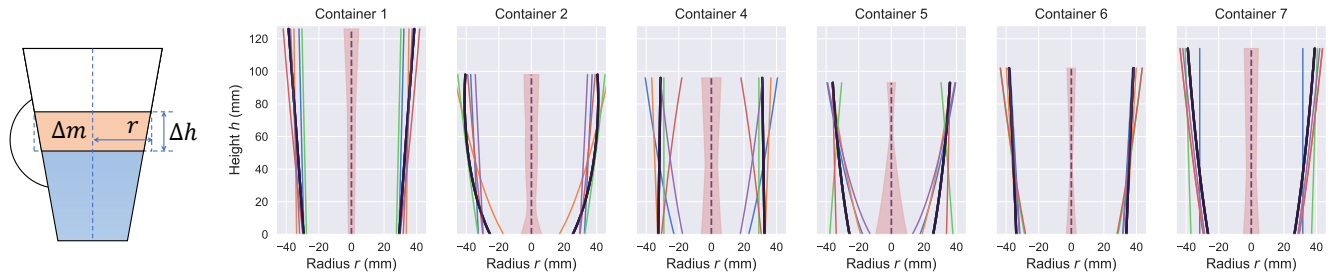
Fig. 11. (a, left) Schematic diagram of a symmetric container. In a specified time interval Δt, the change in mass Δm can be determined by the F/T sensor and the change in height Δh can be derived through MP-Net. Then the radius r at each height can be calculated to form an edge-profile of this container. (b, right) Prediction result of estimating the target container shape. The black curve is the ground truth, while the five different colored curves are the estimated target container shape in five trials. The mean error of the estimated container radius at different heights is plotted in the middle of each subplot (the shaded magenta area).

Finally, the multimodal nature of our network lets us reconstruct the shape of the target container. The dataset and associated software are public and are available at https://lianghongzhuo.github.io/MultimodalPouring.

One surprising limitation of our approach is the poor generalization to liquids like milk or fruit juices, which would be considered quite similar to water by many humans, while the pouring noises are actually quite different. Training on different liquid types would improve network performance, but MP-Net will still fail in situations where the auditory signal is too weak. Another issue is our use of raw force/torque data as the network input, which changes significantly for different grasp types and pouring motions. This could be resolved by training on many grasps, or simply by feeding preprocessed weight data into the network.

For future work, using audio and haptic information for dynamic control of robotic pouring would be an exciting research direction.

REFERENCES

[1] C. Schenck and D. Fox, "Perceiving and reasoning about liquids using fully convolutional networks," *The Int. Journal of Robotics Research (IJRR)*, pp. 452–471, 2018.

[2] R. Mottaghi, C. Schenck, D. Fox, and A. Farhadi, "See the glass half full: Reasoning about liquid containers, their volume and content," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 1871–1880.

[3] C. Schenck and D. Fox, "Reasoning about liquids via closed-loop simulation," in *Robotics: Science and Systems (RSS)*, 2017.

[4] S. Clarke, T. Rhodes, C. G. Atkeson, and O. Kroemer, "Learning audio feedback for estimating amount and flow of granular material," in *Proc. of The 2nd Conf. on Robot Learning*, 2018, pp. 529–550.

[5] J. Wilson, A. Sterling, and M. C. Lin, "Analyzing liquid pouring sequences via audio-visual neural networks," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 7696–7703.

[6] H. Liang, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang, "Making sense of audio vibration for liquid height estimation in robotic pouring," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 5333–5339.

[7] Y. Huang and Y. Sun, "Learning to pour," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017, pp. 7005–7010.

[8] S. Ikeno, R. Watanabe, R. Okazaki, T. Hachisu, M. Sato, and H. Kajimoto, "Change in the amount poured as a result of vibration when pouring a liquid," in *Haptic Interaction*. Springer, 2015, pp. 7–11.

[9] C. Weadon, "Pouring in the dark," *Future Reflections*, vol. 10, no. 3, 1991.

[10] K. J. Pithadiya, C. K. Modi, and J. D. Chauhan, "Selecting the most favourable edge detection technique for liquid level inspection in bottles," *Int. Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, pp. 2150–7988, 2011.

[11] C. Schenck and D. Fox, "Visual closed-loop control for pouring liquids," in *IEEE Int. Conf. on Robotics and Automat. (ICRA)*, 2017, pp. 2629–2636.

[12] C. Dong, M. Takizawa, S. Kudoh, and T. Suehiro, "Precision pouring into unknown containers by service robots," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 5875–5882.

[13] C. Do and W. Burgard, "Accurate pouring with an autonomous robot using an RGB-D camera," in *Intelligent Autonomous Systems(IAS)*, 2018, pp. 210–221.

[14] C. Do, C. Gordillo, and W. Burgard, "Learning to pour using deep deterministic policy gradients," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2018, pp. 3074–3079.

[15] P. Sermanet, K. Xu, and S. Levine, "Unsupervised perceptual rewards for imitation learning," *arXiv preprint arXiv:1612.06699*, 2016.

[16] S. Griffith, V. Sukhoy, T. Wegter, and A. Stoytchev, "Object categorization in the sink: Learning behavior–grounded object categories with water," in *ICRA Workshop on Semantic Perception, Mapping and Exploration*, 2012.

[17] Y. Huang and Y. Sun, "A dataset of daily interactive manipulation," *The Int. Journal of Robotics Research (IJRR)*, pp. 879–886, 2019.

[18] L. Rozo, P. Jiménez, and C. Torras, "Force-based robot learning of pouring skills using parametric hidden markov models," in *IEEE Int. Workshop on Robot Motion and Control*, 2013, pp. 227–232.

[19] H. P. Saal, J.-A. Ting, and S. Vijayakumar, "Active estimation of object dynamics parameters with tactile sensors," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2010, pp. 916–921.

[20] C. Matl, R. Matthew, and R. Bajcsy, "Haptic perception of liquids enclosed in containers," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 7142–7149.

[21] R. Sanchez-Matilla, K. Chatzilygeroudis, A. Modas, N. F. Duarte, A. Xompero, P. Frossard, A. Billard, and A. Cavallaro, "Benchmark for human-to-robot handovers of unseen containers with unknown filling," *IEEE Robotics and Automat. Lett.*, vol. 5, no. 2, pp. 1642–1649, 2020.

[22] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *IEEE Int. Conf. on Robotics and Automat. (ICRA)*, 2020.

[23] T. Wu, J. Lin, T. Wang, C. Hu, J. C. Niebles, and M. Sun, "Liquid pouring monitoring via rich sensory inputs," in *European Conf. on Computer Vision (ECCV)*, 2018, pp. 335–351.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] D. Park, Z. Erickson, T. Bhattacharjee, and C. C. Kemp, "Multimodal execution monitoring for anomaly detection during robot manipulation," in *IEEE Int. Conf. on Robotics and Automat. (ICRA)*, 2016.

[26] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014, pp. 103–111.

[27] A. P. French, "In vino veritas: A study of wineglass acoustics," *American Journal of Physics*, pp. 688–694, 1983.

[28] E. S. Webster and C. E. Davies, "The use of helmholtz resonance for measuring the volume of liquids and solids," *Sensors*, vol. 10, no. 12, pp. 10663–10672, 2010.

[29] M. Kennedy, K. Schmeckpeper, D. Thakur, C. Jiang, V. Kumar, and K. Daniilidis, "Autonomous precision pouring from unknown containers," *IEEE Robotics and Automat. Lett.*, vol. 4, no. 3, pp. 2317–2324, 2019.