Analyzing Liquid Pouring Sequences via Audio-Visual Neural Networks

Justin Wilson[1], Auston Sterling[1], and Ming C. Lin[1,2]

***Abstract*— Existing work to estimate the weight of a liquid poured into a target container often require predefined source weights or visual data. We present novel audio-based and audio-augmented techniques, in the form of multimodal convolutional neural networks (CNNs), to estimate poured weight, perform overflow detection, and classify liquid and target container. Our audio-based neural network uses the sound from a pouring sequence—a liquid being poured into a target container. Audio inputs consist of converting raw audio into mel-scaled spectrograms. Our audio-augmented network fuses this audio with its corresponding visual data based on video images. Only a microphone and camera are required, which can be found in any modern smartphone or Microsoft Kinect. Our approach improves classification accuracy for different environments, containers, and contents of the robot pouring task. Our Pouring Sound Neural Networks (PSNN) are trained and tested using the Rethink Robotics Baxter Research Robot. To the best of our knowledge, this is the first use of audio-visual neural networks to analyze liquid pouring sequences by classifying their weight, liquid, and receiving container.**

I. INTRODUCTION

For robots to perform tasks individually or collaboratively, their ability to sense objects and substances in their environment is critical, especially when pouring liquids. Robots are increasingly performing more complicated human tasks, such as household activities, warehouse placements (e.g. Amazon Picking Challenge [10]), and other detection, recognition, and motion-planning tasks. Many methods for performing these robotic tasks use, and often primarily rely on, visual feedback and human interaction.

In this work, we propose using auditory cues to enhance learned feedback for robots in liquid pouring tasks. Audio has been used in robotics for localization of the spatial position of a sound source [38], navigation [17], autonomous systems [30], sensorimotor learning [6], and locomotion control [33], to name a few. Here, we investigate using sound to enhance a robot's ability to estimate poured weights and types of liquids and containers. Humans are able to roughly sense a change in pitch when filling up a container [25], and we demonstrate that robots can learn to do the same. With audio-visual neural networks, we classify weight, pouring contents, and containers for robot pouring tasks.

Until recently, pouring tasks have often used predefined source amounts of a liquid. Now, [9] demonstrates flow and weight estimation from audio-frequency mechanical vibrations of a robot scooping up and pouring granular materials and [41] controls pouring with closed-loop visual feedback.

[1]Department of Computer Science, UNC Chapel Hill, Chapel Hill, NC 27599, USA wilson,austonst,lin@cs.unc.edu

[2]Department of Computer Science, Univ. of Maryland, College Park, MD 20740, USA lin@cs.umd.edu

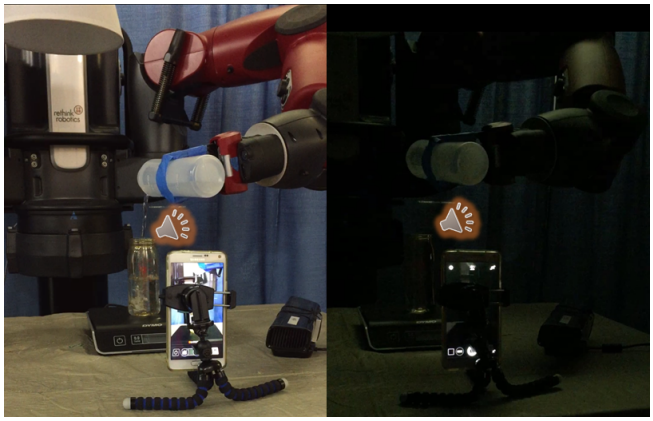

Fig. 1. Our audio-augmented approach performs weight estimation, overflow detection, and content and container classification in bright environments (left) whereas our audio only based approach can be used in dark and occluded environments (right). Pouring sequences are recorded using either a smartphone or Microsoft Kinect's built-in microphone array. Training data is generated by assigning digital scale measurements to discrete audio intervals and tested in experiments using Baxter robot and human experimenter pouring sequences. Various contents (water, rice, soda, and milk) and target containers (glass measuring cup, metal cup, glass bottle, plastic bottle, plastic cup, and square bowl) were evaluated.

Our motivation is to use audio to augment a robot's visual sensing, thereby enabling the use of learned audio-visual feedback. To the best of our knowledge, this is the first use of learned audio-visual feedback to estimate the weight of poured liquids and classify liquid type and container.

The key contribution of this work is a novel, multimodal CNN for weight estimation, overflow detection, and content and container classification for pouring liquids using trained audio-based and audio-augmented neural networks. We demonstrate our approach's ability to compensate for vision-based challenges such as occlusion and transparency by evaluating on target container and liquid pairs trained with hold out pouring sequences for both robot and human experimenter pouring.

1) Training, validation, and test data generated from audio recordings and video images with ground truth measurements from a digital scale;
2) Audio-based convolutional neural network for multi-class weight estimation and binary classification for overflow detection by robotic systems;
3) Audio-augmented neural network enhancing the audio only based method with fused visual inputs for robots pouring contents into various target containers;
4) Pouring content and target container classification for robots, based on pouring sequence audio data.

II. RELATED WORK

In this section, we discuss some of the state-of-the-art audio and video based classification techniques, focusing on temporal classification methods, motion planning, and learned estimation methods for the robot pouring task.

Temporal classification methods: these methods model the dependency, causality, and sequential nature of time series data such as audio. A number of temporal models have been introduced to represent this history and predict the likelihood of consecutive actions. Typical techniques include Hidden Markov Models (HMMs) [37], Conditional Random Fields (CRFs) [24], Recurrent Neural Networks (RNNs) [19], and Long Short-Term Memory (LSTM) [16] networks.

Convolutional filters have also been used for temporal consistency; for example, WaveNet's [32] dilated causal convolutions and Temporary Convolutional Networks' (TCNs) [26] dilated and encoder-decoder implementations. These models have in common the notion of convolution filters across time, computational speedup by updating time steps simultaneously rather than sequentially like recurrent networks, and frame-based classifications as a function of receptive fields (i.e. fixed-length periods of time). These benefits along with state-of-the-art accuracy make TCNs a top choice for audio and visual classification tasks [4].

Motion planning for pouring: while our work currently assumes specific robot and container placements, motion planning for pouring liquids focuses on motion going from start to end targets [34]. Sensory inputs from a chest-mounted camera and a wrist-mounted IMU sensor have also been used to monitor pouring motion [45]. Related work has also shown that size and function, for example, whether a container is fillable, can be determined by using state sequences and a hierarchical spectral clustering algorithm [28] to categorize objects based on size, material, and other features [15]. This work also showed that combining two modalities-sound and proprioception-improved categorization accuracy.

Learning based methods for pouring liquids: [9] is an audio based method that estimates the weight of granular material scooped. The technique is also used for pouring a desired material amount. The approach uses a recurrent neural network with convolutional layers and audio spectrogram input. A benefit of our multimodal approach is that the audio augments the visual data and sample intervals of the pouring sequence are evaluated independently (Table II for baseline comparisons). Analyzing the marginal benefit of recurrent layers in our neural networks is future work. Other learning based methods are based on human demonstrations [46], [47]. These methods model a variety of pouring motions involving shaking and using both robot arms.

Visual control for pouring liquids: [41] is a vision based approach that detects liquid levels using a convolutional network to identify whether a pixel contains liquid and a second stage neural network with a recurrent CNN-LSTM to estimate liquid volume. A Probabilistic Approach to Liquid Level Detection in Cups Using an RGB-D Camera [12] is another liquid detection method. The visual control uses a thermal camera to obtain ground truth labels. It estimates and controls an amount poured without simply pouring the entire contents of the source container or using specialized sensors such as force sensors [40]. This allows for the source container to carry amounts greater than that which the target container can receive.

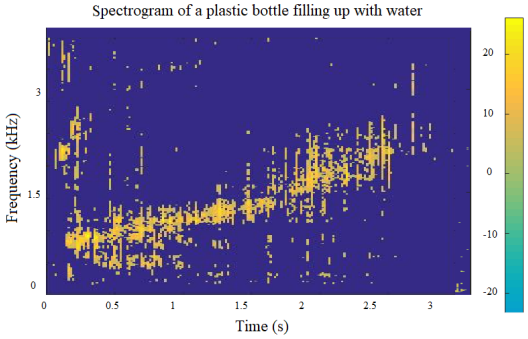

Fig. 2. Spectrogram from a recorded pouring sequence. The frequency of a container filling up can be modeled based on its Helmholtz resonance (also referred to as a resonant cavity) [44]. This resonant frequency increases over time as an object fills up with water as its cavity volume V_c decreases, supporting the use of an audio-based feature for the robot pouring task.

III. TECHNICAL APPROACH

Our neural networks use audio and image data for weight estimation, overflow detection, and poured content and container classification, enhancing learning with sound alone or in conjunction with visual data. By augmenting visual data with sound, we can enhance a robot's ability to detect and perform tasks with transparent or highly reflective containers and liquids in challenging and cluttered environments. To the best of our knowledge, this is the first use of an audio-augmented neural network to analyze liquid pouring sequences in robotics by estimating the weight of a pouring task and classifying poured contents and containers.

Our method allows for a source container to contain amounts greater than the capacity of the target container, as our Pouring Sound Neural Networks (PSNNs) perform multiclass liquid, container, and weight classification and binary classification for overflow detection. Our audio-based approach uses a microphone for input, which can be found in any modern smartphone or Microsoft Kinect. Intervals of recorded audio are assigned a discrete weight class based on digital scale measurements for ground truth labeling. Training is performed offline, while classifications and overflow detection are the results of our neural network predictions.

A. Task Overview

Our task is to use a mel-scaled[1] spectrogram of sound and video image from the target container to predict the poured weight at a point in time during a pouring sequence. A spectrogram is a two-dimensional representation of acoustic energy over frequency and time. Once the target weight is reached or overflow detected, the robot can be signaled to stop pouring and return to its initial position. This task is more difficult than previous work in that it pours a specific

[1]The mel scale is a perceptually linear scale of pitch.

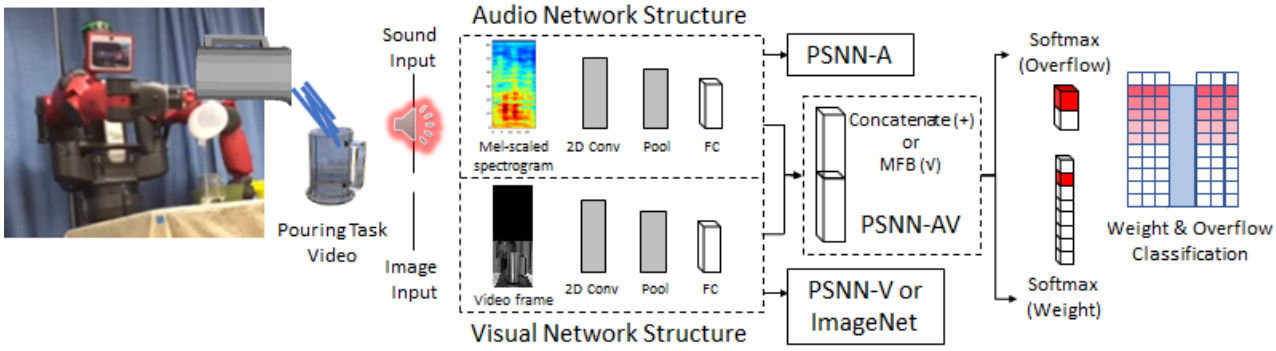

Fig. 3. As the Baxter robot pours liquid from source to target container, a microphone records the audio of the target object filling up with liquid and a camera captures video images. The audio is split into 0.2 second intervals to match the digital scale sampling rate. These audio intervals are converted into mel-scaled spectrograms and passed through a multimodal CNN Pouring Sound Neural Network (we refer to as PSNN) comprised of 2D convolutional, max pooling, fully connected, and softmax layers similar to the Impact Sound Neural Network (ISNN) [42]. Multi-class classification is used for discrete weight estimation (classes of 0.2 oz increments) and liquid and container prediction while binary classification is used for overflow detection. The network's output may be used as a very simple stop command for the robot pouring task. Our method is trained on specific target container and content pairs.

amount rather than simply pouring the entire contents of the source. Moreover, our networks utilize audio information to augment a robot's visual data. The use of audio features are reinforced by the change in audible frequency during a pouring sequence, known as the Helmholtz resonance.

B. Audio Feature Analysis: Helmholtz Resonance Frequency

As depicted in Fig. 2, the audio frequency increases as a container fills up with liquid, forming the basis of an audio-based feature for weight estimation and overflow detection. This increase in frequency can be modeled based on the Helmholtz resonance (also referred to as a resonant cavity) [44]. This resonant frequency, f_{res} is calculated as:

$$f_{res} = \frac{c}{2\pi}\sqrt{\frac{s_p}{V_c l_p}}, \tag{1}$$

where f_{res} is proportional to the speed of sound in a gas c and square root of the cross section area s_p of the container neck, divided by cavity volume V_c and neck length l_p. When an object or liquid of volume V_p is placed/poured into the container, the cavity volume V_c decreases by that amount. By substituting $V_c - V_p$ for V_c, then we can solve for poured volume V_p given V_c, f_{res}, and corrected port $l_p^{'}$ [39].

$$V_p = V_c - \frac{s_p}{l_p^{'}(\frac{2\pi f_{res}}{c})^2} \tag{2}$$

While the resonant frequency adds justification for an audio-based network feature, it assumes the container itself will be symmetric, uniform width, and of a similar shape. Therefore, we implement neural network based classifications that are trained on specific container and liquid pairs with holdout pouring sequences to relax some of these constraints.

C. Dataset Generation

We recorded 500 pouring sequences in total, for six target containers of varying material and geometry, each with three liquids and rice. Each container-liquid combination consisted of 20 pouring sequences. 3 hours of audio and video was captured to use 22,239 samples of 0.2 sec. Data was captured using an iPhone or Android mobile device, as well as a Microsoft Kinect built-in microphone array for comparison. Robot and human experimenter pouring was performed.

Example pouring sequence

	Weight Est		Overflow	
Audio	Truth	Pred	Truth	Pred
0.2s	0.0	0.0	NotFull	NotFull
0.4s	0.0	0.0	NotFull	NotFull
0.6s	0.0	0.0	NotFull	NotFull
0.8s	0.0	0.0	NotFull	NotFull
1.0s	0.1	0.0	NotFull	NotFull
1.2s	0.4	0.2	NotFull	NotFull
1.4s	1.0	0.8	NotFull	NotFull
1.6s	1.5	1.6	NotFull	NotFull
1.8s	2.7	2.4	NotFull	NotFull
2.0s	4.2	4.2	NotFull	NotFull
2.2s	5.8	6.4	NotFull	NotFull
2.4s	7.0	7.2	NotFull	Full
2.6s	9.0	8.6	NotFull	Full
2.8s	11.0	10.6	Full	Full
3.0s	11.8	11.4	Full	Full
3.2s	11.8	11.8	Full	Full

TABLE I

GROUND TRUTH AND PREDICTED LABELS FOR AUDIO INTERVALS FROM THE POURING SEQUENCE OF A TARGET CONTAINER. INTERVALS OF 0.2, 0.5, AND 1 SECOND WERE EVALUATED. 0.2 SEC IS ILLUSTRATED HERE AND USED THROUGHOUT THE PAPER.

For poured weight estimation, digital scale measurements were captured at a rate of 5 readings per second and synchronized to the audio and video recordings. The audio sampling rate was 256 kb/s and the video frame rate was 30 per second. Digital scale readings were visible in the video and used for ground truth verification. However, since these video images were also an input into our audio-augmented network, they were cropped to remove the digital scale display and robot arm as to not influence training. For overflow detection, pouring sequences used for training were stopped at the time of overflow so that full labels could be assigned to the last few seconds of audio while the remaining intervals were labeled as not full. For both weight and overflow prediction, ground truth labels were assigned to discrete 0.2 sec intervals (or frames) for audio and visual data. Fig. 3 describes our

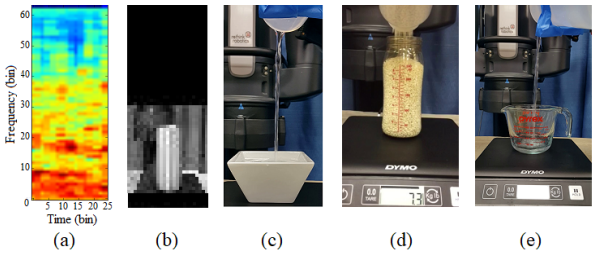

Fig. 4. Audio-visual inputs 2D mel-scaled spectrogram (a) and cropped grayscale image (b). For opaque objects (c), visual information may be occluded. In these cases, PSNN-A outperforms PSNN-V and PSNN-AV. For transparent containers (d-e), our PSNN-V and PSNN-AV networks are able to detect visual deviations for both opaque (d) and transparent (e) pouring contents. The robot arm and digital scale LED are cropped out of images as to not influence network learning (b).

neural network structure and Table I shows an example pouring sequence classified using our frame-based model.

D. Neural Network Architecture of Audio-based Method

Our audio-based neural network model, also referred to as Pouring Sound Neural Network (PSNN-A) shown in Fig. 3, is trained on mel-scaled spectrograms for audio intervals at the digital scale sampling rate of 0.2 seconds. A single convolutional layer followed by two dense layers with feature normalization performs optimally on our classification tasks (Table II). We use consecutive full classification labels to indicate when to stop pouring for overflow detection. Section IV covers our experiments and results against baseline methods. Section V offers analysis and insights into our audio-based (PSNN-A) and audio-augmented (PSNN-AV) convolutional neural networks.

Audio input: two audio input forms were considered – they are a 1D raw audio data and a 2D mel-scaled spectrogram. Using spectrograms as audio input has been shown to reduce over-fitting and improve accuracy [18]. They are computed using a short-time Fourier transform with a Hann window of 2048 samples and an overlap of 25%. Frequency and time axes are downsampled and mapped into 64 mel-scaled frequency bins and 25 time bins to match the logarithmic perception of frequency [42]. We downsample the mel-spectrogram audio input and use a convolution kernel with an increased frequency resolution to reduce over-fitting.

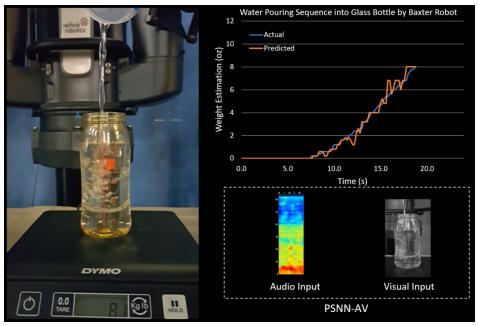

Fig. 5. Demo video of liquid weights predicted by our PSNN neural network for a robot pouring sequence. (Left) video. (Top Right): actual versus predicted weights over time. (Bottom Right): audio and visual neural network inputs. Supplemental materials available at http://gamma.cs.unc.edu/PSNN/

E. Neural Network Architecture of Audio-Visual Method

The input size for audio and visual data have equivalent sizes (25 by 64 pixels). The inputs were designed this way to highlight the importance of estimating weight by changing vertical dimensions of frequency for audio and height for images respectively. Visualizations of inputs that maximize activation illustrate these distinguishing features (Fig. 9). Equivalence by concatenating inputs or fusing based on a bilinear model [48] also allows the network to appropriately weight audio, visual, and audio-visual, given transparent or opaque target containers and contents.

Visual input: for our visual and audio-augmented networks, video images from a mobile device were assigned to corresponding audio intervals and digital scale recordings. To improve training and the effectiveness of our classification, visual data was augmented using techniques discussed in [35] such as cropping. Correctly aligning the multimodal inputs with different sampling rates was also important as to not degrade neural network performance.

F. Implementation Details

All models were implemented with Tensorflow [1] and Keras [8]. Parameters were learned using categorical cross entropy loss with Stochastic Gradient Descent. Training was performed using ADAM [22] and run with a batch size of 64, with remaining hyperparameters tuned manually based on a separate validation set before final test set evaluation. Only audio-based methods were evaluated for overflow detection as incorporating visual information oversimplifies the task. Since there are fewer Full examples in a pouring sequence, audio data was balanced by randomly selecting an equal number of Full/Not Full audio intervals. Our datasets are available to aid future research in this area.

IV. EXPERIMENTS AND RESULTS

We compared our method against baselines by conducting quantitative experiments on a variety of target containers, liquids, and rice. All baselines (including KNN and Linear SVM) are trained on the same input data in order to provide a fair comparison. Pouring sequences were randomly divided into 80% training, 10% validation, and 10% test sets. All target containers and pouring contents were included in training. Test data was based on hold out pouring sequences.

A. Data Capture and Training

Video was recorded using a Samsung Galaxy Note 4 running Android 6.0.1, iPhone 6, and Microsoft Xbox 360 Kinect Sensor. Training was performed using a TITAN X GPU running on Ubuntu 16.04.5 LTS.

B. Pouring Sequence Experiments

Our experiments contained both human experimenter and robot pouring sequences. While robot pouring was varied by adjusting source container volume, experimenter pouring sequences offered additional variability, e.g. unfixed starting positions. All of our robot experiments were performed on a Rethink Robotics Baxter Research Robot, shown in Fig. 1.

Weight Estimation and Overflow Detection Accuracy by Method for Robot and Human Experimenter Water Pouring Sequences

		Glass Bottle, Robot Pour, N=20			Glass Bottle, Human Pour, N=20			Combined Container Dataset, N=40		
Method	Input	+/- 0.4 oz	Ave Err	Overflow	+/- 0.4 oz	Ave Err	Overflow	+/- 0.4 oz	Ave Err	Overflow
kNN [11]	A	66.4%	1.9 oz	71.9%	54.2%	2.7 oz	62.5%	58.8%	2.4 oz	77.1%
Linear SVM [5]	A	4.6%	3.8 oz	50.0%	13.6%	4.3 oz	50.0%	12.7%	4.0 oz	60.4%
SoundNet5 [3]	A	46.0%	1.9 oz	50.0%	42.4%	3.6 oz	50.0%	21.2%	3.3 oz	50.0%
SoundNet8 [3]	A	11.2%	3.3 oz	50.0%	29.2%	4.7 oz	50.0%	35.4%	4.4 oz	50.0%
TCN [26]	A	78.4%	0.9 oz	50.0%	40.1%	3.7 oz	50.0%	49.6%	2.6 oz	50.0%
PSNN-A (Ours)	A	**88.0%**	**0.5 oz**	**78.1%**	**75.8%**	**1.9 oz**	**64.3%**	**80.8%**	**1.3 oz**	**83.3%**
ImageNet [23]	V	**83.8%**	**0.3 oz**	–*	**71.2%**	**0.4 oz**	–*	68.1%	1.1 oz	–*
PSNN-V (Ours)	V	79.9%	0.6 oz	–*	66.5%	0.6 oz	–*	**78.0%**	**0.4 oz**	–*
PSNN-AV Cat (Ours)	AV	**91.5%**	**0.2 oz**	–*	**86.4%**	**0.2 oz**	–*	82.0%	0.3 oz	–*
PSNN-AV MFB (Ours)	AV	88.8%	**0.2 oz**	–*	71.2%	2.1 oz	–*	**86.7%**	**0.2 oz**	–*

TABLE II

MULTIPLE NETWORK MODELS (OURS IS PSNN) AND BASELINES WERE EVALUATED TO DETERMINE THE BEST NEURAL NETWORK STRUCTURE FOR AUDIO AND AUDIO-VISUAL BASED LIQUID POURING ANALYSIS. AUDIO INPUT ARE MEL-SCALED SPECTROGRAMS; VISUAL INPUT, GRAYSCALE IMAGES. * ONLY AUDIO-BASED NEURAL NETWORKS WERE EVALUATED FOR OVERFLOW AS VISUAL INFORMATION OVERSIMPLIFIED THE TASK.

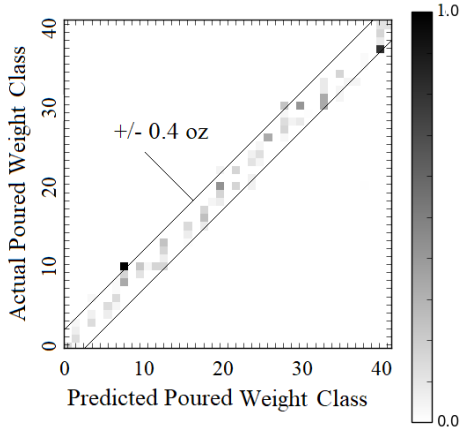

Fig. 6. ***PSNN-AV***: confusion matrix comparing actual to predicted poured amounts by classes of 0.2 oz (about 6 ml) weight increments. Class 0 represents empty; Class 1, 0.2 oz; and so on. Using audio and visual improves accuracy, especially at the beginning and end of the pouring sequence. Our system achieves up to **91.5%** (Table II) and **91.2%** (Table IV) classification accuracy to within +/- 0.4 oz using PSNN-AV.

Pouring consisted of experimenters using both hands to hold the source container for human pouring and the Baxter robot's 7 DOF left arm for robot pouring sequences. We used the Dymo Digital USB Postal Scale for ground truth weight estimates and a Samsung Galaxy Note 4 for video recording.[2]

For robot experiments, the target container rests on a tabletop, positioned slightly to the side and below the source container. The source container is fixed to the robot gripper and is pre-filled with an amount not known to the robot but greater than the amount required to fill the target container.

After a pouring sequence is initiated, audio from the target container filling up is recorded with a smartphone. Each audio interval is transformed into a mel-scaled spectrogram and input into our neural network model for weight and overflow classification. Once the desired pour amount is classified or overflow is detected, the robot can be signaled to stop the pouring sequence and return to its initial position.

[2]Audio and video was also captured using an iPhone 6 and Microsoft Xbox 360 Kinect Sensor with built-in microphone array for comparison.

C. Our Results vs. Baseline

As illustrated in Table II and Fig. 6, up to 91.5% of the audio intervals for the robot pouring sequence into a glass bottle were classified to a weight class within 0.4 oz using our audio-augmented convolutional neural network (PSNN-AV); likewise, 86.4% of the human pouring sequence. This resulted in an average error of 0.2 oz and 0.2 oz respectively. We also performed an evaluation on a combined pouring dataset containing both robot and human pouring sequences to explore the opportunity for transfer learning. More details in Analysis Section V.

Tables III and IV demonstrate our method's ability to be trained on different liquids (Fig. 7) and types of containers, including asymmetric objects. First, our audio-based PSNN-A network outperforms all baseline methods for audio only input. Second, when pouring content is visible, audio-augmented (PSNN-AV) outperforms audio-based (PSNN-A). This is especially true for more viscous liquids, such as milk, which make less noise during a pouring sequence.

Classification Accuracy for Plastic Bottle Weight Estimates via PSNN-AV and Human Pouring

Pl. Bottle	+/- 0.2 oz	+/- 0.4 oz	+/- 0.6 oz
Milk	57.8%	63.9%	68.4%
Rice	49.1%	64.4%	73.0%
Soda	73.0%	82.9%	88.4%
Water	69.6%	77.2%	84.0%

TABLE III

VARIOUS POURING CONTENTS WERE EVALUATED USING PSNN-AV TO ESTIMATE WEIGHT GIVEN EXPERIMENTER POURING SEQUENCES.

We should note, however, that due to the relatively small size of the training set, our neural networks work well for target container and pouring content pairs that are described in this paper. Since all liquid-container pairs are included in training with hold out pouring sequences, future work is needed for generalization to unseen and untrained target containers or pouring contents.

D. Pouring content and target container classification

Table V highlights our networks' ability to classify the pouring content and target container (Fig. 8) from pouring

Classification Accuracy and Average Error by Method, Input, and Target Container for Robot Pouring Sequences

Method	In	Transparent Plastic Cup Water +/-0.4 oz/AveErr	Transparent Glass Meas. Cup Water +/-0.4 oz/AveErr	Opaque Porcelain Bowl Water +/-0.4 oz/AveErr	Opaque Metal Cup Water +/-0.4 oz/AveErr	Transparent Glass Bottle Milk +/-0.4 oz/AveErr	Transparent Glass Bottle Rice +/-0.4 oz/AveErr
kNN	A	34.7% / 3.4 oz	25.9% / 3.6 oz	48.1% / 2.2 oz	41.0% / 2.5 oz	38.2% / 2.7 oz	48.4% / 1.7 oz
Linear SVM	A	5.4% / 3.4 oz	8.0% / 4.8 oz	8.9% / 3.3 oz	7.0% / 4.1 oz	33.2% / 3.5 oz	12.8% / 2.3 oz
SoundNet5	A	14.0% / 3.4 oz	5.3% / 4.2 oz	6.4% / 4.4 oz	4.4% / 4.7 oz	9.7% / 3.0 oz	9.6% / 2.4 oz
SoundNet8	A	11.6% / 3.2 oz	20.5% / 6.1 oz	9.4% / 3.5 oz	13.1% / 4.2 oz	13.4% / 5.8 oz	8.8% / 3.4 oz
TCN	A	50.0% / 1.5 oz	39.5% / 1.9 oz	43.0% / 2.0 oz	51.5% / 1.7 oz	34.0% / 3.9 oz	52.7% / 1.7 oz
PSNN-A (Ours)	A	**59.1% / 1.2 oz**	**46.8% / 1.2 oz**	**60.9% / 1.3 oz**	**65.9% / 0.7 oz**	**45.0% / 1.8 oz**	**74.1% / 1.0 oz**
ImageNet	V	64.5% / 0.6 oz	51.7% / 1.2 oz	29.4% / 3.9 oz	20.0% / 6.1 oz	65.1% / 0.4 oz	77.0% / **0.4 oz**
PSNN-V (Ours)	V	**79.8% / 0.3 oz**	**63.9% / 0.5 oz**	**36.2% / 2.7 oz**	**25.3% / 4.6 oz**	**68.9% / 0.4 oz**	**83.7% / 0.4 oz**
PSNN-AV Cat (Ours)	AV	**79.0% / 0.3 oz**	**70.0% / 0.4 oz**	40.0% / 3.4 oz	48.5% / 1.9 oz	71.8% / 0.4 oz	**91.2% / 0.2 oz**
PSNN-AV MFB (Ours)	AV	69.2% / 0.4 oz	44.9% / 1.7 oz	**42.6% / 2.6 oz**	**65.5% / 1.2 oz**	**82.4% / 0.2 oz**	81.8% / 0.3 oz

TABLE IV

MULTIPLE NETWORK MODELS AND BASELINES WERE EVALUATED. OURS IS PSNN. TOP PERFORMING METHOD FOR EACH INPUT HIGHLIGHTED IN BOLD. HEADINGS INDICATE DISTINGUISHING PROPERTIES BEING EVALUATED. PSNN OUTPERFORM OTHER BASELINE METHODS.

Fig. 7. Various pouring contents were evaluated using PSNN-AV. This graph displays the percentage of classified weights within +/- 0.2 oz (blue), 0.4 oz (orange), and 0.6 oz (gray) of ground truth. For instance, soda and water were easier to estimate pour weights into a plastic bottle than rice and milk. Rice was most difficult to precisely predict within +/- 0.2 oz.

Classification Accuracy for Pouring Content via Human Pouring and Target Container via Robot Pouring

Pl. Bottle	Content %	Water	Container %
Milk	86.5%	Pl. Bottle (0)	99.6%
Rice	79.6%	Metal Cup (1)	88.4%
Soda	72.4%	Gl. Bottle (2)	69.2%
Water	97.9%	Gl. Measuring Cup (3)	64.2%
		Porcelain Square Bowl (4)	61.3%
		Pl. Cup (5)	78.5%

TABLE V

PSNN-A TO PREDICT POURING CONTENT AND TARGET CONTAINER FROM POURING SEQUENCE AUDIO. PLASTIC (PL.) AND GLASS (GL.) PERCENTAGES ARE 0.2 SEC INTERVALS CLASSIFIED CORRECTLY.

sequence data. Viscous liquids with greater damping (e.g. milk) and carbonated beverages (e.g. soda) offer distinguishing features. For future work, we plan to investigate if accuracy varies over time. For instance, is content classification accuracy higher in the beginning of a pouring sequence?

We concluded our testing with an ablative analysis for hyper-parameter optimization (e.g. training epochs, interval length, etc.). Our pouring sequence dataset with audio and visual data is made available to support future research and evaluation in this area of robotics.

V. ANALYSIS

In this work, we implement multimodal neural networks based on audio and visual data to the robotic task of weight

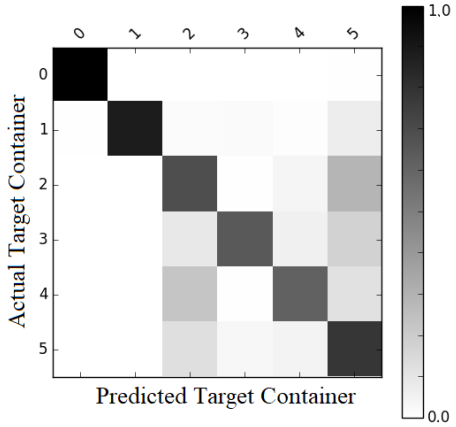

Fig. 8. Confusion matrix of actual to predicted target container classifications based on audio from pouring sequences. Higher accuracy is achieved when we exclude before and after pouring, i.e. exclude intervals when audio is not present during the pouring sequence. Labels (0-5) in Table V.

estimation for pouring a liquid, overflow detection, and liquid and container classification. Our PSNN neural networks outperform existing methods in the experiments that we have performed. Our contributions include new audio-visual datasets and multimodal neural network architectures designed for the robot pouring task. In this section, we analyze the improved performance of using our methods.

A. Activation Maximization Visualizations

We analyzed activation maximizations to visualize the spectrogram audio and visual input which would produce the highest activation for a given volume class. Fig. 9 shows activation maximization for the audio-based PSNN-A network as additional volume is poured (a-b) and the visual-based PSNN-V network (c-d). Both highlight the importance of vertical dimensions for audio and visual when estimating poured volume based on frequency and height respectively.

B. Model Comparisons

For opaque target containers, the audio only PSNN-A performs the best compared to PSNN-V and PSNN-AV due to occlusion. For transparent target containers, multimodal

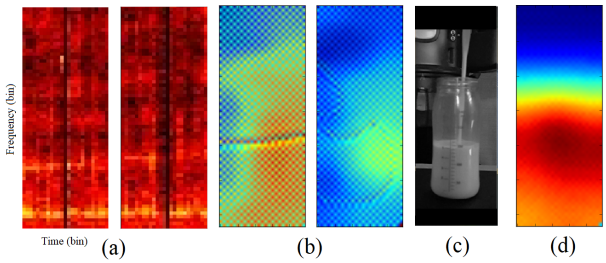

Fig. 9. **Audio activations**: example pouring sequence spectrograms of frequency versus time (a). Audio inputs that would maximize our audio-based neural network activation for a couple of specific weights (b). This demonstrates the PSNN-A neural network's ability to learn changes in frequency to distinguish between weight classes. **Visual activations**: example grayscale, cropped visual input (c). Visual input that would maximize the activation of our visual neural network (d). This shows the PSNN-V neural network's ability to learn visual features for distinguishing between classes for visible pouring contents (Fig. 4).

PSNN-AV provides the maximum classification accuracy and minimum average error. Even for a quiet, viscous liquid like milk, augmenting visual data with audio outperformed audio or visual only with 82.4% accuracy and 0.2 oz average error compared to 45.0% and 68.9% respectively. (Table IV).

1) PSNN-A Normalized: normalizing the features allows for a more symmetric optimization between frequency and time given a mel-scaled spectrogram input. Scaling is important to normalize the differences in feature scale. When feature scaling is not applied, then gradient descent may require a smaller learning rate to ensure that the optimization converges and does not over step the minimum.

2) PSNN-A and Temporal Convolutional Networks (TCN): our methods outperform time distributed baselines because while the pouring task is sequential, it does not rely as heavily on previous inputs since each 0.2 second spectrogram encodes the current state. Furthermore, time distributed methods may overfit and fail to cover more general and inconsistent pouring behavior. PSNN can evaluate inputs independently since each mel-scaled spectrogram already encodes historical information given a frame-based interval.

3) Robot and Human Poured: Robot pouring sequences are more accurate than human poured given an equal number of training examples and epochs (Table II). In other words, robot pouring sequences require less data and training time because of more uniform pouring sequences, producing more consistent audio and visual data for each weight class.

4) Combined Pour Dataset: For TCN and PSNN, the combined dataset of robot and human pouring sequences mostly performs medially as compared to each separately (Table II). For PSNN-V, however, additional visual data of a combined dataset performs better with 0.4 oz average error compared to 0.6 oz for both robot and human pouring. This implies visual data is less affected by pouring consistency than audio, benefiting from additional yet mixed data.

5) Interval Length: Audio sampling intervals of 0.2, 0.5, and 1 second were evaluated. 0.2 is the minimum based on the digital scale sampling rate. Faster intervals performed better, which is to be expected since the interval is assigned a single ground truth weight and smaller time intervals would represent a smaller change in poured amount over that time. As the length increases, the interval likely has a larger variation of frequencies for each training example.

VI. CONCLUSION AND FUTURE WORK

We present novel, audio-based and audio-augmented neural networks to estimate poured weight, perform overflow detection, and classify pouring liquid and target container based on pouring sequence audiovisual data. By recording the sound of the pouring sequence as the target container fills up, an audio-based feature can be applied to different containers and liquids for the robot pouring task. Our method is trained on specific target container and content pairs using both human and robot pouring sequences and is tested on the Baxter robot. We also evaluate our dataset on a combined container dataset and make our audio-visual data available for future research. To the best of our knowledge, this is the first use of audio-visual neural networks to analyze liquid pouring sequences by classifying their weight, liquid, and receiving container.

Future Directions: to increase accuracy beyond current performance, we plan to analyze augmentations of our audio data with environmental, room acoustics, and other alterations. As the task involves temporal data, sequential layers can be introduced into the neural network model, such as recurrent, LSTM, or GRU layers or HMM filtering. This may be especially helpful for audio only PSNN-A classification at the beginning and end of pouring sequences when there are no pouring sounds. In addition, we plan to compare against a lower-dimensional parameterization of the sound such as a set of audio features like spectral centroid, skew, kurtosis, and rolloff. Comparison with model-based approaches where the target container 3D geometry is known may shed new insight as well.

Our current neural networks do not generalize to unseen target containers or pouring contents. We plan to research ways to generalize our approach, which may involve increasing the size of our training set, adding more audio and visual data augmentations, or incorporating synthetic pouring sequences. Using a multiple output neural network rather than separately trained neural networks for poured weight, content, and target container classification may also help as well as using a ratio of volume over the target container volume or a combination of all of the above.

Finally, we will explore if our approach can be applied to other granular materials and liquids in addition to rice and the liquids that we've tested to date. Furthermore, we plan to evaluate if container size and function (e.g. fillable or not) can be determined by using the spectral hierarchical clustering algorithm [28] or PSNN to categorize objects based on size, material, and other features [15].

ACKNOWLEDGMENT

This work is supported in part by the U.S. National Science Foundation and the Elizabeth Stevinson Iribe Chair Professorship.

REFERENCES

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu. Y, and Zheng, X. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. https://www.tensorflow.org/

[2] Adrien, J.-M. 1991. The missing link: Modal synthesis. In Representations of musical signals, pages 269–298. MIT Press

[3] Aytar, Y., Vondrick, C., and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems. pp. 892–900

[4] Bai, S., Kolter, J. Z., Koltun, V. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. CoRR.

[5] Bottou, L. 2010. Large-Scale Machine Learning with Stochastic Gradient Descent. Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)

[6] Boyer, E. 2015. Continuous Auditory Feedback for Sensorimotor Learning. Doctoral dissertation, Universit Pierre et Marie Curie-Paris VI. ED3C. IRCAM University Paris Descartes. ISMES.

[7] Chadwick, J.N. and James, D. L. 2011. Animating fire with sound. In ACM Transactions on Graphics (TOG). Vol. 30. ACM, 84

[8] Chollet, F. and others. 2015. Keras. https://github.com/keras-team/keras

[9] Clarke, S., Rhodes, T., Atkeson, C., and Kroemer, O. 2018. Learning Audio Feedback for Estimating Amount and Flow of Granular Material. Conference on Robot Learning. Proceedings of Machine Learning Research.

[10] Correll, N., Bekris, K. E., Berenson, D., Brock, O., Causo, A., Hauser, K., Okada, K., Rodriguez, A., Romano, J. M., and Wurman, P. R. 2018. Analysis and Observations From the First Amazon Picking Challenge. IEEE Transactions on Automation Science and Engineering. Vol. 15.

[11] Cover, T. and Hart, P. 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. pp. 21–27

[12] Do, C., Schubert, T., and Burgard, W. 2016. A Probabilistic Approach to Liquid Level Detection in Cups Using an RGB-D Camera. IROS 2016.

[13] van den Doel, K., and Pai, D. K. 1996. The sounds of physical shapes. Presence 7, 382–395

[14] Duch, W., Kacprzyk, J., Oja, E., and Zadrozny, S. 2005. Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part II, volume 3697 of Lecture Notes in Computer Science. Springer

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

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

[17] Huang, J., Supaongprapa, T., Terakura, I., Wang., F., Ohnishi, N., and Sugie, N. 1999. A model-based sound localization system and its application to robot navigation. Robotics and Autonomous Systems. Volume 27, Issue 4, 30 June 1999, pp. 199–209

[18] Huzaifah, M. 2017. Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. CoRR. Volume abs/1706.07156.

[19] Jain, L. C. and Medsker, L. R. 1999. Recurrent Neural Networks: Design and Applications. CRC Press, Inc.

[20] Jaitly, N. and Hinton, G. 2013. Vocal Tract Length Perturbation (VTLP) improves speech recognition

[21] James, D. L., Langlois, T. R., Mehra, R., and Zheng, C. 2016. Physically Based Sound for Computer Animation and Virtual Environments. ACM SIGGRAPH 2016 Course, http://graphics.stanford.edu/courses/sound/

[22] Kingma, D. P. and Ba., J. 2014. Adam: A method for stochastic optimization

[23] Krizhevsky, A., Sutskever, I., and Hinton, G. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. pp. 1097–1105

[24] Lafferty, J. D., McCallum, A. and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289.

[25] Lawson, R. B. Pitch Perception. Technical Note 7-65. Human Engineering Laboratories.

[26] Lea, C., Flynn, M., Vidal, R., Reiter, A., and Hager, G. 2016. Temporal Convolutional Networks for Action Segmentation and Detection. CoRR

[27] Lin, T., RoyChowdhury A., and Maji S. 2015. Bilinear CNN Models for Fine-grained Visual Recognition. International Conference on Computer Vision (ICCV)

[28] von Luxburg, U. 2007. A tutorial on spectral clustering. Statistics and Computing, vol. 17, no. 4, pp. 395-416

[29] Lysenkov, I. and Eruhimov, V., and Bradski, G. 2008. Recognition and Pose Estimation of Rigid Transparent Objects with a Kinect Sensor. Robotics: Science and Systems (RSS)

[30] Martinson, E. and Schultz, A. 2009. Discovery of sound sources by an autonomous mobile robot. Autonomous Robots.

[31] Moss, W., Yeh, H., Hong, J., Lin, M., AND Manocha, D. 2010. Sounding Liquids: Automatic Sound Synthesis from Fluid Simulation. ACM Transactions on Graphics (TOG)

[32] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio

[33] Ozkul, M. C., Saranli, A., Yazicioglu, Y. 2012. Acoustic Surface Perception for Improved Mobility of Legged Robots. ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, 2012.

[34] Pan, Z., Park, C. and Manocha, D. 2016. Robot Motion Planning for Pouring Liquids. International Conference on Automated Planning and Scheduling (ICAPS)

[35] Perez, L and Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. CoRR.

[36] Quinlan, J. R. 1986. Induction of Decision Trees. Journal Machine Learning. pp. 81–106

[37] L. R. Rabiner. 1990. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Readings in Speech Recognition. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. pp. 267–296

[38] Rascon, C. and Meza, I. 2017. Localization of sound sources in robotics: A review. Robotics and Autonomous Systems. Volume 96, October 2017, pp. 184–210

[39] Rayleigh, J.W.S. 1945. The Theory of Sound, Volume Two; Dover Publications, Inc.: New York, NY, USA

[40] Rozo, L., Jimenez, P., and Torras, C. 2013. Force-based robot learning of pouring skills using parametric hidden markov models. RoMoCo, pp. 227-232

[41] Schenck, C. and Fox, D. 2017. Visual Closed-Loop Control for Pouring Liquids. International Conference on Robotics and Automation (ICRA)

[42] Sterling, A., Wilson, J., Lowe, S., and Lin, M. C. 2018. ISNN: Impact Sound Neural Network for Audio-Visual Object Classification. Proceedings of the European Conference on Computer Vision (ECCV)

[43] Tenenbaum, J. and Freeman, W. 2000. Separating Style and Content with Bilinear Models. Neural Comput.

[44] Webster, E. and Davies, C. 2010. The Use of Helmholtz Resonance for Measuring the Volume of Liquids and Solids

[45] Wu, T., Lin, J., Wang, T., Hu, C., Niebles, J., Sun, M. 2018. Liquid Pouring Monitoring via Rich Sensory Inputs

[46] Yamaguchi, A., Atkeson, C., Niekum, S., Ogasawara, T. 2014. Learning pouring skills from demonstration and practice. IEEE RAS International Conference on Humanoid Robots.

[47] Yamaguchi, A., Atkeson, C., Ogasawara, T. 2015. Pouring Skills with Planning and Learning Modeled from Human Demonstrations. International Journal of Humanoid Robotics, Vol. 12, No. 3.

[48] Yu, Z., Yu, J., Fan, J., and Tao, D. 2017. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. IEEE International Conference on Computer Vision (ICCV)

[49] Zheng, C. AND James, D. L. 2010. Rigid-body fracture sound with precomputed soundbanks. ACM Trans. Graph. 29, 69:1–69:13