The Sound of Water: Inferring Physical Properties from Pouring Liquids

Piyush Bagad
University of Oxford

Makarand Tapaswi
IIIT Hyderabad

Cees G. M. Snoek
University of Amsterdam

Andrew Zisserman
University of Oxford

Project page: bpiyush.github.io/pouring-water-website/

***Abstract*—We study the connection between audio-visual observations and the underlying physics of a mundane yet intriguing everyday activity: *pouring liquids*. Given only the sound of liquid pouring into a container, our objective is to automatically infer physical properties such as the liquid level, the shape and size of the container, the pouring rate, and the time to fill. To this end, we: (i) show in theory that these properties can be determined from the fundamental frequency (pitch); (ii) train a pitch detection model with supervision from simulated data and visual data with a physics-inspired objective; (iii) introduce a new large dataset of real pouring videos for a systematic study; (iv) show that the trained model can indeed infer these physical properties for real data; and finally, (v) we demonstrate strong generalization to various container shapes, other datasets, and in-the-wild YouTube videos. Our work presents a keen understanding of a narrow yet rich problem at the intersection of acoustics, physics, and learning. It opens up applications to enhance multisensory perception in robotic pouring.**

***Index Terms*—Audiovisual, physical estimation, liquid pouring.**

"The blind man of Puisaux judges of his nearness to the fire by the degrees of heat; of the fulness of vessels by the sound made by liquids which pour into them; of the proximity of bodies by the action of the air on his face."

– Denis Diderot, Letter on the Blind (1749)

I. Introduction

What can possibly be scientifically interesting about such a mundane chore as pouring a liquid into a glass? We perform this action all the time but barely realise that we effortlessly learn to infer several useful physical properties in the process. For example, evidence in psychoacoustics suggests that humans can accurately infer the liquid level, the time to fill [1], the size of the container [2], and even the temperature of the liquid [3], merely from the sound of pouring. Such inference (*e.g.*, time to fill) allows us to adaptively control our actions (*e.g.*, stopping pouring to prevent spillage) conforming to the *affordance* theory by Gibson [4]. In this work, we study the physical phenomenon involved in liquid pouring and explore how it can be used to train machines to infer useful physical properties from sound alone as illustrated in Fig. 1.

Despite its mundaneness, liquid pouring has rich physics underpinning it and has been studied for more than a century [5]. The crux of this exploration is summarized well by Berg and Stork [6]: "as the liquid (*e.g.*, water) is filled, a sound consisting of an increasing pitch and some (odd) harmonics superimposed with whooshing, gurgling is observed". This pitch and the harmonics are a function of the physical properties, *e.g.*, trajectory of the pitch depends on the container shape [7], the range of the pitch depends on the container dimensions [1], and slope of pitch depends on the flow rate [1]. Thus, automatically inferring physical properties from the sound of pouring necessitates two stages: (i) detecting pitch from the raw audio signal, and (ii) recovering these physical properties from the pitch. However, there are several challenges in training machines to do this purely from sound.

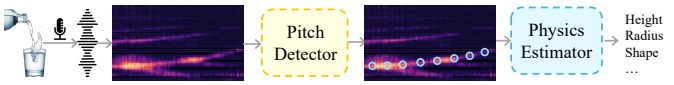

Fig. 1: An overview of the objective and method. We train a pitch detector without any manual supervision and rely on physics to estimate physical properties merely from the sound of water.

First, such a task requires fine-grained time-sensitive audio modeling, while contemporary audio models focus more on coarse tasks like classification [8]. Second, the underlying physics of such a niche activity as pouring is not fully developed for general container-liquid setups, unlike, say, Newtonian mechanics studied analogously in [9]. Third, there is a lack of a large, clean, controlled dataset which is necessary to study physical property estimation. Fourth, supervision either in the form of pitch annotation or the actual physical properties is difficult to obtain and use directly in training.

To enable a systematic study, we collect a clean, large dataset of 805 videos of pouring across 50 diverse containers. For training, we select a subset of containers shaped like cylinders such that we can approximate the underlying physics with that of a cylinder [1]. We design an audio network for pitch detection based on `wav2vec2` [10] pre-trained on speech data that has characteristic pitch dynamics [11]. As pitch annotations are hard and ambiguous to obtain at scale, we use supervision from simulated data and visual data. We pre-train the network on simulated sounds of liquid pouring. On real data, we fine-tune the network by visual co-supervision with a physics-inspired objective. We demonstrate that the co-supervised audio model is able to predict pitch, and hence estimate physical properties, with a performance far exceeding that of multiple previous methods.

Why is this important, though? First, as far as we know, this is the first work to demonstrate human-like capabilities (or better) in predicting physical properties from sound alone; in fact, we achieve an accuracy of $\pm 0.60\,\mathrm{cm}$ in predicting the air column height for cylinders. Second, although the model is trained on cylinders, we show that the pitch estimation (and the model in general) is applicable beyond cylinders, *e.g.*, it can be used to predict container shape with convincing accuracy. Third, the model generalizes well to videos from other datasets and to in-the-wild YouTube videos. Our dataset, code and models are released on the Project page.

This research was funded by EPSRC Programme Grant VisualAI EP/T028572/1, and a Royal Society Research Professorship.

II. Related Work

Pouring in the literature. Pouring occurs surprisingly often in the literature. It has been studied by roboticists to train robots to pour [12–15]. In computer vision, there has been work on visually perceiving liquids either in a static setting [16–18], or during pouring [19, 20]. For example, [16, 21] detect the amount of liquid in a container and [19, 20] track the stream of pouring liquid. Likewise, there has been work on estimating the dynamic states (*e.g.*, height or mass of liquid at a given time) from multi-modal (vision, audition, haptics, etc.) inputs [22, 23]. Most of these use additional sensory data (*e.g.*, force, torque, hand trajectory, inertia) with vision or audio or both. Such measurements require sophisticated recording equipment. In contrast, our aim is to predict physical properties from sound recorded using regular smartphones instead of bespoke equipment. Closest to our work is that of Wilson et al. [22] where a CNN is supervised to predict the mass of liquid poured at a given time, given instantaneous video and audio clips. Methodologically, our work differs from [22] by incorporating the underlying physics directly in the learning process. By design, our method can estimate several physical properties without supervision and not just liquid mass. We also evaluate our model by linear probing of the co-supervised features on the dataset of [22] and report superior performance.

Audio-visual learning. The natural audio-visual correspondence in videos coupled with large-scale video datasets [8, 24] has led to a variety of work on self-supervised representation learning. These approaches can be broadly categorized as contrastive [25, 26], generative [27, 28], paired sample discriminative [29, 30], clustering [31, 32], and distillation-based [33, 34]. These approaches and tasks ignore fine time-dependent structures in sounds and rely on short and coarse correspondences. In contrast, liquid pouring requires modeling of fine-grained characteristics over time (*e.g.*, pitch).

III. The Physics of Liquid Pouring

As an example, consider a simple cylindrical vessel of radius R, height H as shown in Fig. 2 (left). At time t, suppose that the vessel is filled such that the length of the air column is $l(t)$. We hear a mix of pitch and odd harmonics that correlate with the length of the air column at a given time. We term this resonance as *axial* resonance. This is visible as the blue curve on the spectrogram in Fig. 2 (right).

Pitch in axial resonance. As the liquid fills up, it pushes air out of the air column creating a frequency pattern that resembles blowing air in an organ pipe closed at one end. As the water level increases, the vacant space for air molecules to vibrate reduces and hence the frequency increases. At time t, the fundamental frequency $f(t)$ is given by $f(t) = c/(4l(t))$, where c is the speed of sound in air. This expression arises from a standing wave of wavelength $\lambda(t) = 4l(t)$ where the amplitude is zero at the water surface and maximum at the top of the vessel. Rayleigh [5] and others studied this and found an experimental end-correction that depends on the radius: $f(t) = c/4.\left(l(t) + \beta R\right)$, where β is the end-correction factor generally agreed to be 0.62 [35, 36]. A spectrogram of the sound of pouring in a sample container is shown in Fig. 2 (right) with pitch $f(t)$ (blue circles) and first harmonic (green crosses) marked. To avoid working with an inverse relation, we look at this equation in terms of wavelength $\lambda(t)$,

$$\lambda(t) = \frac{c}{f(t)} = 4\left(l(t) + \beta R\right). \tag{1}$$

All these quantities are in metric units. The LHS is observable from audio while the RHS is observable from the video of liquid pouring (up to a scale factor). This implies that the audio is effectively a *metric ruler* for objects in the video.

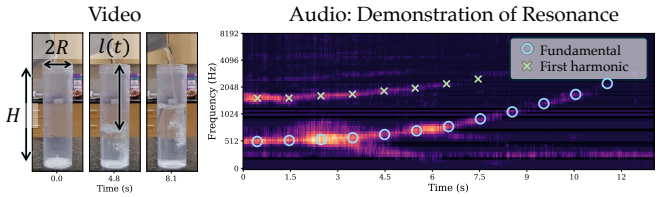

Fig. 2: **Demonstration of axial resonance in liquid pouring.** As liquid is poured in the container (left), theoretical estimates of the resonant frequencies are shown (right) overlaid on a MelSpectrogram. Blue circles show the pitch (fundamental frequency) and green crosses show the first harmonic. Note, the fainter curve starting at around 3s due to *radial resonance*, a different kind of resonance caused by the radial vibration of the container rim which decreases over time. We defer its study to future work.

Physical properties from pitch. We categorize the physical properties of the container and liquid in two sets. (i) *Static*: these are inherent to the container-liquid system (*e.g.*, container size) and do not vary over time. (ii) *Dynamic*: these vary over time (*e.g.*, air column length). We first derive the air column length and later compute other properties from that.

1) **Length of air column**: We want to estimate $l(t)$ given $\lambda(t)$ at a given time t. Using the boundary condition $l(T) = 0$ in Eq. (1), where T is the total pouring duration, we get:
$$l(t) = \frac{1}{4}\left[\lambda(t) - \lambda(T)\right], \forall t \in [0, T]. \tag{2}$$
2) **Container size**: Container height and radius are directly obtained from the boundary conditions: $H = l(0) = (\lambda(0) - \lambda(T))/4$; and $R = \lambda(T)/4\beta$.
3) **Volume flow rate**: For volume flow rate $Q(t)$, suppose the volume at time t is $V(t) = \pi R^2(H - l(t))$. Then, $Q(t) = \frac{\mathrm{d}V}{\mathrm{d}t} = -\frac{1}{4}\pi R^2 \frac{\mathrm{d}\lambda}{\mathrm{d}t}$, where the derivative can be numerically approximated using the estimated $\lambda(t)$.
4) **Time to fill**: Here, we assume a constant flow rate (since otherwise, one could pause pouring midway leading to an ill-defined time to fill). Also, we do not know the true duration T and are only given a partial audio, *i.e.*, the first t seconds. Here, following Cabe and Pittenger [1], we make an additional assumption that the end-correction term βR is small at the start of pouring ($\beta R \ll H$). Thus, in a short interval at the start of pouring $t' \in (0, \delta)$, we have
$$\tau(t') = \left[\frac{l(t')}{-\frac{\mathrm{d}l}{\mathrm{d}t}}\right] = \left[\frac{\lambda(t')/4 - \beta R}{-\frac{1}{4}\frac{\mathrm{d}\lambda}{\mathrm{d}t}}\right] \approx -\frac{\lambda(t')}{\frac{\mathrm{d}\lambda}{\mathrm{d}t}}, \tag{3}$$

Then, we can use the property of τ to get time to fill: $\tau(t) = T - t = (T - t') + t' - t = \tau(t') - (t - t')$, for some $t' \in (0, \delta), \delta \ll t$. Note that this needs reliable estimates of λ and its derivative at the start of the audio.

IV. Network and Training

Our objective is to predict physical properties (*e.g.*, length of the air column) from the sound of pouring. Our approach is formulated in two stages: (i) detect the pitch from the raw audio signal, and (ii) recover physical properties from the pitch as previously described. To detect pitch, we train an audio network, first with supervision from synthetic audio of pouring water; and then with real videos using the visual stream to provide the supervision on pitch.

Audio network. The network takes in raw audio samples and outputs wavelength (pitch) estimates at each time step. The architecture is based on `wav2vec2` [10] adapted for pitch detection on pouring sounds. Note that we predict the fundamental wavelength as opposed to frequency because wavelength varies linearly with length of air column and we want to bake in this linearity in the learned features. The input waveform is resampled at a rate of 16 kHz. First, the waveform is tokenized using a 1D CNN encoder that takes in windows of 25 ms with a hop length of 20 ms. Moreover, we add sinusoidal position embeddings to the tokens to enhance temporal information. These are then passed through a Transformer network with 12 blocks (d_{model}=768, n_{heads}=8). This is followed by a prediction head, a linear regressor that maps from $\mathbb{R}^{768} \rightarrow \mathbb{R}^{K}$, where K=64 is the number of wavelength bins followed by a softmax to obtain a distribution.

Pre-training with synthetic data. Since it is hard to obtain pitch annotations on real samples, we first pre-train the network on synthetic samples. To generate synthetic data, we train a Differentiable Digital Signal Processing (DDSP) [37] autoencoder with independent control over pitch and amplitude. It is an encoder-decoder model capable of generating pouring sounds given a specific pitch and loudness profile. The encoder consists of three modules: a loudness encoder, a pitch encoder, and a residual latent encoder. The residual encodes background noise and room reverberation. The decoder is composed of synthesizers based on classical signal processing techniques. Loudness and residual are conditioned on a real audio sample while the pitch can be arbitrary. This enables us to provide arbitrary pitch $f(t), \forall t$ and the model generates a waveform with this pitch. To generate a sample, we randomly sample radius R, height H, compute $l(t)$=$(-H/T)t + H$ and $f(t)$ from Eq. (1) and pass it to the decoder. In total, we generate 10K samples and pre-train the pitch detection part of our network using the KL divergence loss. We train only the penultimate K=8 layers of the Transformer and the prediction head, keeping the rest of the network frozen. The 1D CNN and Transformer are initialized from pre-trained `wav2vec2`.

Fine-tuning with visual co-supervision. We fine-tune the network on a small number of real samples to overcome the *sim2real* gap [38]. Since it is hard to annotate pitch precisely on real data, we use video as a source of weak supervision.

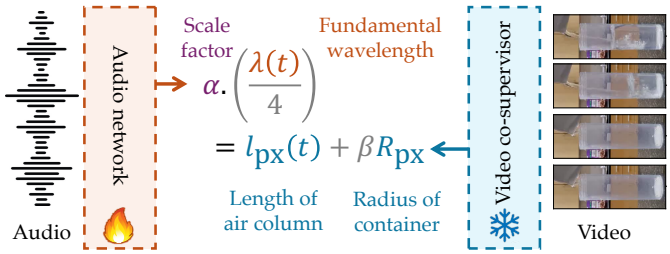

Fig. 3: **Visual co-supervision for pitch detection.** Video co-supervisor provides air column length and radius (in px) which supervises the audio network that predicts wavelength (in cm).

To use visual co-supervision, we train a visual network to estimate the length of air column $l(t)$ and container radius R (in pixels) to compute the RHS in Eq. (1). These then supervise the audio network to predict $\lambda(t)$, the LHS in Eq. (1). A schematic diagram is presented in Fig. 3. The visual network takes in the video as input and outputs container radius R, and $l(t), \forall t$. Estimating R is trivial using a segmentation mask obtained by SAM [39]. To estimate $l(t)$, we design a network based on DINO [40]. DINO's dense spatial feature maps for a frame sequence are passed through a Transformer to model temporal dependencies. This is followed by a prediction head that regresses a 1D bounding box spanning the air column. We train this network with MSE loss using pseudo-labels obtained from temporal difference of frames. We use a temporal context of N=20 frames (4 FPS), the Transformer has one block (d=512, n_{heads}=4). To account for the unknown scale factor between the metric audio outputs to pixel video outputs, we estimate the scale factor for each video by simply computing the ratio of wavelength from audio to the pixel lengths from video. Then, we fix the scale factors and the video network and fine-tune the audio network to improve predictions of λ.

V. Experiments

In this section, we present various experiments to demonstrate physical understanding from sounds of liquid pouring.

LiquidPouring50 Dataset. Our data consists of videos that show a human hand pouring liquid in a container with a fixed camera facing the container. Across videos, we randomly vary the flow rate but keep it constant within a single video. In total, we collect 805 videos across 50 containers (4 shapes, 5 materials) and 2 liquids (hot and normal water). The shapes are cylindrical, semiconical, bottleneck and hemispherical. The materials include glass, plastics, ceramics, steel and cardboard. Each container is annotated with its shape, material and basic measurements. The mean container height is 11.5 cm, (base) radius is 3.1 cm and video duration is 10.5 s. We carefully create splits with multiple test sets described in Table I.

Comparison with baselines. We compare our models with standard pitch detection baselines [41–43] in estimating $l(t)$. We obtain ground-truth assuming constant flow rate as $l(t) = (-H/T).t + H$. The models predict pitch $\lambda(t)$, and $l(t)$ is computed using Eq. (2). As reported in Table II, our models comprehensively beat all baselines with the best model achieving an MAE of 0.6 cm.

Split	Opacity		Container shapes			Containers	Videos
	Transparent	Opaque	Cyl.	Sem.	Bot.		
Train	✓	✗	✓	✓	✗	18	195
Test I	✓	✗	✓	✓	✗	13	54
Test II	✗	✓	✓	✓	✗	19	327
Test III	✓	✓	✓	✓	✓	25	434

TABLE I: **Splits in LiquidPouring50.** Train, Test I and II all have cylinder-like containers and are disjoint in terms of videos. The containers in Test I are a subset of those in Train, while Test II has novel containers. Test III is used for shape classification and overlaps only with Test II in terms of containers. This adds up to $18+25{=}43$ containers, and $195 + 54 + 434{=}683$ videos. The remaining 122 videos (out of a total of 805) are of hemispherical/freeform containers only used for qualitative analysis. Here, Cyl., Sem., and Bot. denote cylindrical, semiconical and bottleneck respectively.

Method	Test set I seen containers ↓	Test set II unseen containers ↓
Baselines		
Yin [41]	30.80	27.30
PESTO [42]	11.70	10.60
CREPE [43]	7.61	9.40
Argmax on spectrogram	4.60	5.11
Ours		
Audio-only	0.78	0.82
Co-supervised	**0.60**	**0.71**

TABLE II: **Comparison with baselines in estimating length of air column.** Mean absolute error (in cm) in estimating $l(t)$ on the two test sets (cylinder-like containers). Our models comfortably beat all the pitch detection baselines. Generally, performance on Test set II is poorer as it consists of containers not seen during training.

Property	Units	Test set I		Test set II	
		Synthetic ↓	*Co-supervised* ↓	*Synthetic* ↓	*Co-supervised* ↓
Static properties					
Height	cm	2.23	2.27	2.77	2.85
Radius	cm	1.62	1.39	2.24	1.88
Dynamic properties					
Flow rate	ml/s	25.2	22.5	45.7	40.4
	s	3.96	4.16	4.39	4.10
Time to fill	s	1.62	1.49	3.44	2.99
	s	1.53	1.07	2.66	2.21

TABLE III: **Co-supervision improves physical property estimation.** Mean absolute error in estimating various physical properties. Visually co-supervised model generally improves over the synthetic-trained model in estimating physical properties from pitch. We observe noticeable improvements in estimating radius and flow rate. This suggests co-supervision particularly improves estimation of pitch towards the end of the audio as well as the slope of pitch generally.

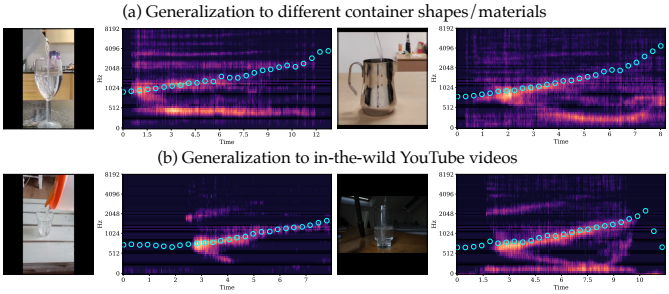

Fig. 4: **Qualitative results.** Predicted pitch is shown in cyan. We show generalization across (a) container shapes, (b) to in-the-wild videos. Despite visual variations in (b), accurate pitch helps infer precise container size.

Improvement by co-supervision. We show benefits of co-supervision in estimating other physical properties. Ground truth for H, R and time to fill are available while flow rate is estimated as the ratio of volume to duration. Predictions are obtained following the steps in Section III. As reported in Table III, visual co-supervision usually improves over the synthetic-trained model. On height estimation, co-supervision suffers a meagre drop (≤ 0.8mm). We attribute this to possibly imprecise start and end annotated timestamps of pouring.

Container shape recognition. While we train on cylinders, we evaluate if our models generalize to other container shapes and to other data sources. On the *Test III* split, we check if container shape (cylindrical, semiconical, or bottleneck) can be inferred from features learned by co-supervision. Concretely, we compute features $z(t), \forall t$ from the last layer of the co-supervised `wav2vec2`. Then, we construct features $h := \text{concat}(\mathbb{E}[z(t)], z(t/4), z(t/2), z(3t/4))$, and train and evaluate a linear probe on these features. This achieves a sample accuracy of 90.91% and a mean class accuracy of 92.47%. In comparison, features without co-supervision achieve 88.63% and 89.44%. This shows: (i) a model trained to detect pitch implicitly encodes container shape, and (ii) co-supervision further improves shape recognition.

Liquid mass estimation. On the dataset by Wilson et al. [22], we evaluate the estimation of the liquid mass from the pouring sounds. Note, in this dataset the pouring conditions, liquid types, and environmental conditions differ from those of our LiquidPouring50 dataset. We attach a regressor to the co-supervised features $z(t)$ to predict the liquid mass $m(t)$. We report results on the same splits as [22]. Averaging across six containers, our model achieves MAE of $1.20\,\text{oz}$ outperforming the best audio model from [22], a fully supervised CNN that achieves MAE of $1.35\,\text{oz}$.

Generalization and failure cases. Qualitatively, we also find that our model generalizes to different container shapes and to in-the-wild YouTube videos (Fig. 4). Finally, we report some failure cases in estimating the pitch in our preprint on arXiv [44]. These are likely due to a limitation of our model: it is trained to pick only the fundamental frequency while ignoring higher harmonics and other kinds of resonance [45].

VI. Conclusion

We have shown early evidence that machines, like humans, can be trained to infer physical properties from pouring sounds. On cylinder-like containers, we demonstrated precise estimation of physical properties such as container size, flow rate and time to fill. Furthermore, we showed our model is more generally applicable beyond cylinders and can be used to estimate container shape and liquid mass. It generalizes well to other shapes and to in-the-wild YouTube videos. Our work strengthens multimodal perception in robotic pouring and opens up possibilities to infer even subtler properties like liquid temperature, viscosity, and container material, merely from pouring sounds. We hope that our work also prompts similar studies for physical understanding from the sound of other mundane activities.

References

[1] Patrick A. Cabe and John B. Pittenger. Human sensitivity to acoustic information from vessel filling. *Journal of experimental psychology. Human perception and performance*, 2000. 1, 2

[2] Hannah Perfecto, Kristin Donnelly, and Clayton R Critcher. Volume estimation through mental simulation. *Psychological science*, 2019. 1

[3] Carlos Velasco, Russ Jones, Scott King, and Charles Spence. The sound of temperature: What information do pouring sounds convey concerning the temperature of a beverage. *Journal of Sensory Studies*, 2013. 1

[4] James J Gibson. *The ecological approach to visual perception*. Psychology press, 2014. 1

[5] John William Strutt Baron Rayleigh. *The theory of sound*. Macmillan, 1896. 1, 2

[6] Richard E Berg and David G Stork. *The physics of sound*. Pearson Education India, 1982. 1

[7] Mark P Silverman and Elizabeth R Worthy. Musical mastery of a coke™ bottle: physical modeling by analogy. *The Physics Teacher*, 1998. 1

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1, 2

[9] Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016. 1

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020. 1, 3

[11] Masanori Morise et al. Harvest: A high-performance fundamental frequency estimator from speech signals. In *INTERSPEECH*, 2017. 1

[12] Connor Schenck and Dieter Fox. Visual closed-loop control for pouring liquids. In *ICRA*, 2017. 2

[13] Zherong Pan, Chonhyon Park, and Dinesh Manocha. Robot motion planning for pouring liquids. *International Conference on Automated Planning and Scheduling*, 2016.

[14] Yongqiang Huang, Juan Wilches, and Yu Sun. Robot gaining accurate pouring skills through self-supervised learning and generalization. *Robotics and Autonomous Systems*, 2021.

[15] Yongqiang Huang and Yu Sun. Learning to pour. In *IROS*, 2017. 2

[16] Chau Do, Tobias Schubert, and Wolfram Burgard. A probabilistic approach to liquid level detection in cups using an rgb-d camera. In *IROS*, 2016. 2

[17] Roozbeh Mottaghi, Connor Schenck, Dieter Fox, and Ali Farhadi. See the glass half full: Reasoning about liquid containers, their volume and content. *ICCV*, 2017.

[18] Sagi Eppel, Haoping Xu, Yi Ru Wang, and Alan Aspuru-Guzik. Predicting 3d shapes, masks, and properties of materials, liquids, and objects inside transparent containers, using the transproteus cgi dataset, 2021. 2

[19] Connor Schenck and Dieter Fox. Detection and tracking of liquids with fully convolutional networks, 2016. 2

[20] Haitao Lin, Yanwei Fu, and Xiangyang Xue. Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring. In *ICCV*, 2023. 2

[21] Gautham Narayan Narasimhan, Kai Zhang, Ben Eisner, Xingyu Lin, and David Held. Self-supervised transparent liquid segmentation for robotic pouring. *ICRA*, 2022. 2

[22] Justin Wilson, Auston Sterling, and Ming Lin. Analyzing liquid pouring sequences via audio-visual neural networks. In *IROS*, 2019. 2, 4

[23] Hongzhuo Liang, Chuangchuang Zhou, Shuang Li, Xiaojian Ma, Norman Hendrich, et al. Robust robotic pouring using audition and haptics. In *IROS*, 2020. 2

[24] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 2

[25] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv:2210.07839*, 2022. 2

[26] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv:2009.09805*, 2020. 2

[27] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *ICCV*, 2023. 2

[28] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *NeurIPS*, 2024. 2

[29] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 2

[30] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *NeurIPS*, 2018. 2

[31] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Selfsupervised learning by crossmodal audiovideo clustering. *NeurIPS*, 2020. 2

[32] Brian Chen, Andrew Rouditchenko, Kevin Duarte, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *ICCV*, 2021. 2

[33] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *NeurIPS*, 2016. 2

[34] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2

[35] S. Herbert Anderson and Floyd C. Ostensen. Effect of frequency on the end correction of pipes. *Physical Review*, 1928. 2

[36] Arthur Taber Jones. End corrections of organ pipes. *Journal of the Acoustical Society of America*, 1941. 2

[37] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. In *ICLR*, 2020. 3

[38] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, et al. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *NeurIPS*, 2022. 3

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, et al. Segment anything. In *ICCV*, 2023. 3

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[41] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 2002. 3, 4

[42] Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. Pesto: Pitch estimation with self-supervised transposition-equivariant objective. In *ISMIR*, 2023. 4

[43] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *ICASSP*, 2018. 3, 4

[44] Piyush Bagad, Makarand Tapaswi, Cees GM Snoek, and Andrew Zisserman. The sound of water: Inferring physical properties from pouring liquids. *arXiv:2411.11222*, 2024. 4

[45] Anthony P. French. In vino veritas: A study of wineglass acoustics. *American Journal of Physics*, 1983. 4