A Predictive Model for Tactile Force Estimation using Audio-Tactile Data

Loïc Niederhauser, Ahalya Prabhakar, Dominic Reber, and Aude Billard

***Abstract*—Robust in-hand manipulation of objects with movable content requires estimation and prediction of the contents' motion with enough anticipation to allow time to compensate for resulting internal torques. The quick estimation of the objects' dynamics can be challenging when the objects' motion properties (e.g., type, amount, dynamics) cannot be observed visually due to robot occlusions or opacity of the container. This can be further complicated by the computational requirements of onboard hardware available for real-time processing and control for robotics. In this work, we develop a simple learning framework that uses echo state networks to predict the torques experienced on the robotic hand with enough anticipation to allow for adaptive controls and sufficient efficiency for real-time prediction without GPU processing. We demonstrate the efficacy of this formulation for tactile force prediction on the Allegro robotic hand with a Tekscan tactile skin using both material-specific and material-agnostic learned models. We show that while both are effective, the material-specific models show an improvement in accuracy due to the difference in inertial properties between the different materials. We also develop a prediction model that uses audio feedback to augment the tactile predictions. We show that adding auditory feedback improves the prediction error, though it significantly increases the computation cost of the model. We validate this formulation for online prediction on the robotic hand moving materials in real-time and adapting grip for slip detection.**

I. Introduction

Safe in-hand manipulation of objects requires the ability to estimate and adapt to the forces experienced by the robot hand from the objects' inertial forces during motion. This is particularly relevant when adaptation requires grip changes that depends on sufficient time to execute. In these situations, force prediction must occur on longer time horizons, allowing for *anticipation* of slip into the future, as opposed to high-speed reactive controls. This can be more challenging when vision is impaired, due to occlusion by the robot or opacity of the container, as objects have different inertial properties that can lead to different force trajectories exerted on the hand during motion. These additional inertial forces can make it difficult for the robotic hand to maintain a stable grasp on the container during manipulation. Most research in this area primarily focuses on either utilizing vision for state estimation and prediction, or short-term slip prediction and detection for reactive torque controllers.

Manuscript received: July 17 2023; Revised: Octobre 13 2023; Accepted: November 10 2023

This paper was recommended for publication by Editor Ashis Banerjee upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported in part by the European Research Council by the CHIST-ERA program through the project CORSMAL and the European Research Council Advanced Grant, project ID: 741945

All authors are affiliated with the LASA lab at École Polytechnique Fédérale de Lausanne (EPFL).

Digital Object Identifier (DOI): see top of this page.

Here, we focus on the problem of tactile prediction on a longer time horizon without the aid of vision. We develop a framework using echo state networks (ESNs) that enable tactile force prediction for a longer-time horizon. We train ESN models that predict the movement and amplitude of forces exerted on the hand during motion . We show that the learned models work over different materials and motion speeds for a given motion type. We demonstrate the differences in tactile forces caused by different materials even during the same motion, and validate the benefits of learning material- and motion-specific models. Furthermore, we show that utilizing auditory feedback in a multimodal prediction model improves the performance of tactile prediction for all materials. Importantly, the use of ESNs as a framework for the proposed approach allows for lower computational requirements, enabling successful execution with low computation resources (i.e., a laptop with only a CPU). Finally, we validate the framework in a real-world experiment, demonstrating both tactile prediction on-line and its use for slip-anticipation for an adaptive grip control in real-time .

II. Related Work

Tactile estimation and prediction research has been of great interest in manipulation applications. Much of the research focuses on slip detection and reactive adaptation, discussed in detail in [1], [2]. Some methods use neural networks for slip detection using tactile data or vision-based tactile sensors for classifying contact or slippage [3]–[7]. Tactile and audio data are sometimes used in neural networks for content estimation, such as weight estimation or liquid height [8], [9]. Other methods learn models of liquid flow of unknown materials by using tactile or force data to estimate physical properties of the contents, including mass, volume and viscosity [10], [11]. These models are used to classify unknown liquids and for content estimation, which could be used to predict flow over time for manipulation. Here, we investigate the use of low-dimensional model learning for general content estimation, without imposing the structure of physical models of liquid flow that the learning here relies upon. Other research uses tactile force estimation and prediction for control and manipulation, discussed in depth in [12]. Many methods use estimation of contact point, force and curvature to drive control for manipulation [6], [13], [14]. Other research methods use neural networks to learn contact force estimation from tactile data [7], [15]. Su et al. [7] uses a neural network to predict the contact forces in the fingertips. The contact force estimation focuses on static grasping (i.e. when the container contents are not moving) to enable slip detection and slip prevention during grasping. [16] also performs force estimation for Biotac

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2023.3340614

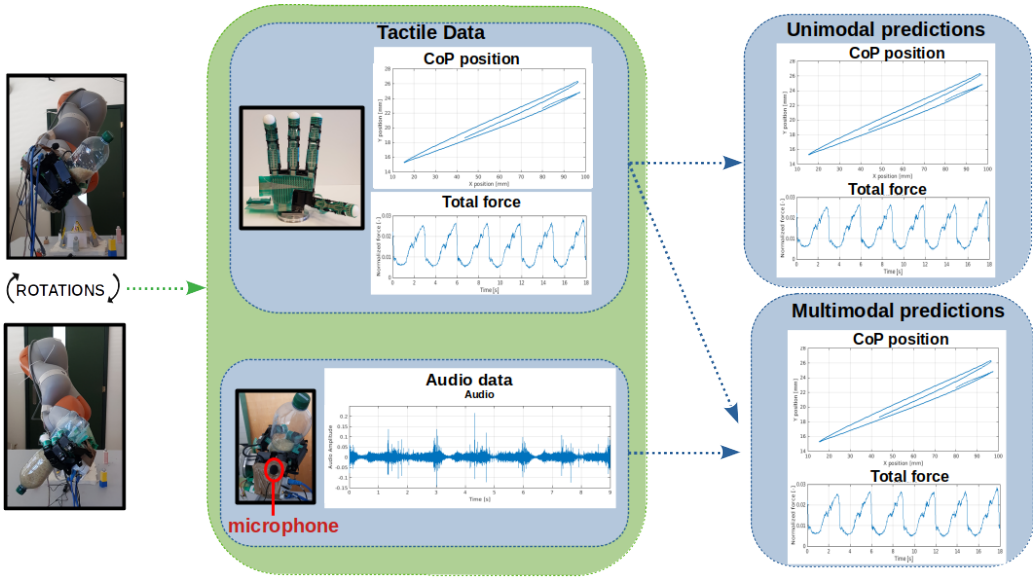


Fig. 1: Description of Experimental Setup. It consists of a Kuka IIWA robot arm mounted with an Allegro hand and a microphone fixed on the robotic hand close to the container. The hand is covered with a Tekscan tactile skin sensor. The hand holds a bottle that is filled with unknown content. In the experiments, content may consist of water, a high-viscosity slurry, rice or gummies. The robot rotates the bottle from left to right such that the content slides inside the bottle. Changes in distribution of mass within the object is revealed through change in pressure exerted on the palm. From the Tekscan data, the position of the center of pressure (cop) and the total force on each surface of the of the hand links is computed. As the content slides, it also generates noise that can be picked up by the microphone. Tactile and audio data are captured and used to refine prediction of tactile response.

sensors located in the fingertips. They use convolutional neural networks to estimate forces during contact, which could be used in a grasp controller. [17] and [18] use graph neural networks and convolutional LSTMs respectively to predict contact forces in the fingertips during grasping. Both use the prediction models to assess grasp stability during lifting. Other methods use deep learning or reinforcement learning to drive touch-based manipulation through unsupervised learning [19], [20]. These methods require large amounts of data (3000-5000 trajectories) and prediction is done on short time horizons (15-18 time steps) due to the complexity of the prediction model. The approach proposed in this paper however focuses developing prediction models that have faster training times and lower complexity for longer time horizon predictions (up to 3 seconds into the future) to accommodate reactive controllers that require grasp adaptation.

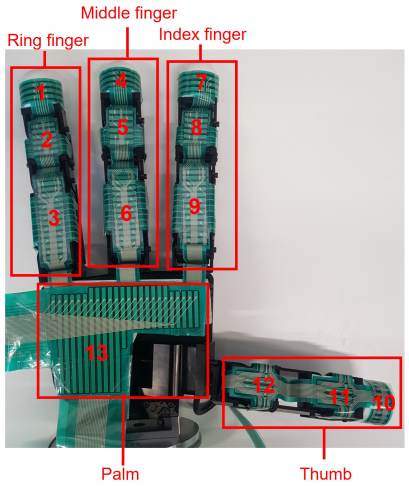
Echo state networks [21], [22] are a type of recurrent neural network with a random dynamical resevoir (i.e., randomly connected neurons within the reservoir). These networks have been shown to be faster to train, requiring fewer parameters to be optimized and not suffering from gradient issues found in methods that use gradient-based optimization (i.e., backpropogation). [23] uses echo state networks to predict the tactile signature of an antenna from action. They use the ESN for the 1-dimensional tactile prediction to discriminate when contact is occurring, rather than for grip control. Furthermore, their work uses echo state networks to predict one time step into the future, rather than for a longer time horizon (e.g., 300 steps into the future) as performed in this work. Furthermore, the other work only use tactile sensors for prediction and estimation. Here, we use auditory feedback to augment the tactile prediction. While there has been research in using auditory feedback for tactile applications [24]–[27], most work focuses on using the multimodal perception for classification and object property identification, rather than real-time tactile prediction.

III. Methods

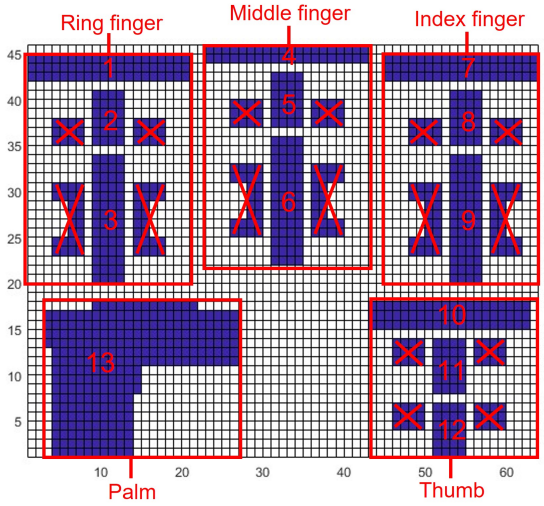
We develop a framework for learning and predicting the tactile pressure distribution experienced in a robotic hand. Specifically, we use echo state networks (ESNs) [21] to learn a predictive model of the magnitude and center of pressure on the hand caused by the objects' movement during manipulation. We describe below the experimental setup and the formulation for the model learning and tactile force prediction.

A. Experimental Setup

The experimental setup, shown in Figure 1, uses a Kuka IIWA 7 with an attached Allegro hand for robotic manipulation. A Tekscan tactile sensor consisting of 1050 tactile pixels of 4.6 x 4.6 [mm] covering palm and fingers provides pressure data across the Allegro hand and is sampled at a rate of approximately 100 Hz. The tactile data values are calibrated and normalized such that they range from 0 to 255 values (similar to values of 8-bits gray scale images) and then converted to Newtons by multiplying the value by a factor of 0.007. We represent the tactile sensors on the hand as



(a) Tekscan tactile skin attached on the Allegro hand



(b) Diagram of the flattened Tekscan sensor with each section numbered. The sensor has sensing cells on the sides of the fingers (marked with red crosses), which are not taken into consideration as they do not come in contact with the container.

Fig. 2: Diagram of the Tekscan sensor showing the cells and the numbered sections covering the hand.

13 sections covering the palm and the fingers, as shown in Figure 2. From each section of the hand, we calculate the center of pressure (CoP) and amplitude of the activated sensors over the region. The motion of the CoP and its amplitude over the different sections reflects the torques exerted on the hand during the motions by the contents. We train our model (discussed in detail below) to predict the CoP position and amplitude in each section. An audio microphone is attached to the Allegro hand (shown in Figure 1), capturing the sounds inside the container during manipulation with a sampling rate of 16 kHz.

B. Data Collection

To generate data, we develop a robot controller that rotates the container a desired amount in a specified period. The rotation motion uses a position-based dynamical system (DS) controller that combines a DS control to maintain the desired position, while changing the orientation a specified range in a sinusoidal pattern to generate the motion. Throughout the motion, the Allegro hand maintains a envelop grasp on the container with a base torque of 1.5 Nm.

We train prediction models for 4 classes of material contents with different inertial properties: water, a high-viscosity slurry, rice and gummies. The total weight of the content is 700 [g], 750 [g], 500 [g] and 450 [g] respectively. We collect demonstrations of 4 trials of 10 rotations for each material with orientation range of 120° over a period of 3 seconds. Between each trial, the bottle is replaced in the hand and the grasp is reset leading to variance in grasp between trials.

Parameter	Lower bound	Upper bound
Spectral radius	0.01	1
Input scaling	0.0001	10
Noise level	0.0	0.3
Time constants	0	1
Leakage	0	1

TABLE I: Particle Swarm Optimization Ranges for Hyper-parameters for Echo State Network Models. The parameters were chosen by performing particle swarm optimization over the parameters to optimizing the average training error over time.

C. Model Learning

Echo state networks are used for tactile pressure prediction. Echo state networks (ESNs) [21], [22] consist of a recurrent neural network (RNN) with randomly assigned connectivity and weights for the hidden layer. While RNNs are trained with backpropagation, echo-state networks only optimize the weights of the output neurons , enabling computationally efficient model learning which effectively captures nonlinear time series behavior. The implementation used in this work follows the method introduced in [28], which introduces additional parameters for improving performance (e.g., leakage, skipped connections). We optimize the parameters using particle swarm optimization, within the bounds listed in Table I. For testing, all ESN models were run on a computer with an Intel Core i7-6700 CPU@3.40GHz and 16GB RAM, with no GPU.[1]

1) Tactile Pressure Prediction: Tactile pressure prediction consists of an echo state network model predicting the pressure experienced over the hand throughout the motion. The echo state network input and output consists of the CoP position (x and y) and total force for each of the 13 sections over the hand.

[1]We optimize the following parameters using particle swarm optimization (PSO). The parameters are: spectral radius, scaling of the input, noise level in neurons during training, time constants of neurons (all neurons are equally split between a small number of time constant in order to reduce the degrees of freedom of the optimization), leakage of neurons and connectivity of the reservoir. The parameters are given sensible bounds (listed in Table I), the initial swarm and inertia matrices are initialized by randomly sampling particles and inertia from a bounded uniform distribution.

Parameter	Value ranges
Number of Audio Time Stamp Predictions	10
Mel Spectrogram: Window Length	[0.1 - 0.9] * (1.5 [s])
Mel Spectrogram: Window Overlap	0.1 - 0.9
ESN Spectral Radius	0 - 1
4 Input Scalings (for groups of Mel bands: 1-8, 9-16, 17-24, 25-32)	0 - 1
ESN Leakage	0 - 1
Weights for weighted predictions (30 ratio constants for 10 timestamps x 3dim)	0 - 1

TABLE II: Particle Swarm Optimization Ranges for Hyper-parameters for Audio-Based Echo State Network Models. The parameters were chosen by performing particle swarm optimization over the parameters to optimizing the average training error over time.

For each material, the data is split in training/testing sets with a $50\% - 50\%$ ratio. Each set is then further split into 4000-4500 time windows of 150 samples, depending on the length of the trial. The ESN is primed with 150 samples and trained to predict the following 300 samples in open loop.

Accuracy of the network's prediction is assessed using the testing set, which is composed of an equal number of time windows. In addition, the capacity of the ESN to predict position on unseen material is assessed, as well as it's capacity to work with different movement speed by varying the movement speed for a single material. We also assess the results of the prediction while the robot is running in a real-time scenario, described below in Section III-E.

D. Audio-Based Tactile Prediction

We also develop a learning model that takes in auditory feedback to improve the tactile force predictions. We train an ESN to predict the tactile data for 10 time steps in the future based on 1.5 seconds of audio data. The audio data is broken into smaller time windows of 0.1 seconds, from which Mel spectrograms are generated. These spectrograms are fed into the ESN sequentially and the next 10 time steps of tactile data is predicted in one shot. To integrate this prediction with the tactile prediction, we implement a weighted combination of the predictions from the audio and tactile ESNs and feed the result back into the tactile ESN through the open-loop routine of the prediction. We use particle swarm optimization to optimize the parameters of the audio-based ESN described in Table II.

E. Robotics implementation

We validate the learned models in a robotic experiment on the Kuka IIWA with the attached Allegro hand for real-time prediction and control. We perform a slip prediction experiment that utilizes the trained models to perform online tactile prediction to anticipate when slip will occur as illustrated on figure 3. We place the container on the robotic hand with the hand in the open position. The hand performs a small rotation in one direction (10 degrees) in which the bottle remains stable on the hand followed by a wider rotation in the opposite direction (45 degrees), during which the container will slip out of the hand. For training, we collect data with the hand in the open configuration and the bottle attached securely to the palm to prevent the bottle from slipping out of the hand. We collect 20 trials of data with a bottle filled with water using a rotation motions of 5 seconds. For the experimental study, as the hand executes the rotation motion, we use the tactile predictions to anticipate when slip will occur, and accordingly execute a close motion on the hand to grasp it. We define the slip threshold as when the force amplitude is close to 0 (i.e under 1 N) or when the CoP measurements are close to the side of the palm (i.e closer to 5 mm from the side of the palm). We perform the experiment in both directions, five times by starting the rotation towards the thumb and five times by starting the rotation towards the ring finger. To evaluate performance, we analyze the closing time of the hand over the container. Early and on time closings are both considered successful, while late closings are classified as failures. Early closing is when the hand closes when the bottle doesn't move or is still stable (i.e, during the smaller first rotation thumb side). On time closing refers to the hand closing when the bottle is moving but has not yet slipped out of the hand during the larger rotation. Late closing occurs when the hand begins to close after the bottle has fallen.

IV. RESULTS

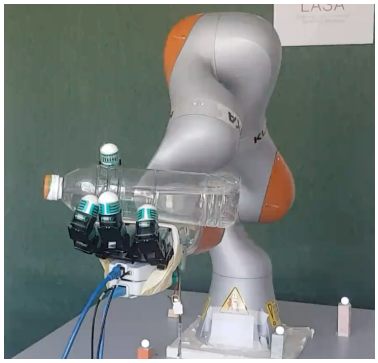
In this section, we discuss the results of the tactile and audio-tactile model learning and the experimental studies conducted. Figure 4 shows the data of the CoP and total force data over a rotation for each material. While the overall temporal patterns caused by the rotation of the contents on the hand resemble each other, the inertial dynamics of each material has a significant impact on both the CoP location and total force exerted on the hand, highlighting the importance of taking material content into account for force prediction.

Figure 5 shows the FFT decomposition of the signal for different material. In this case, the robot noise induces some similarity in the frequency decomposition for different material. However, differences in shape and intensities are still observed for different materials.

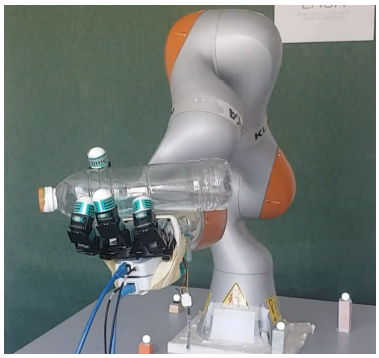
Figure 6 shows the results for open-loop tactile prediction using the material-specific ESN of the CoP and force for water for 300 timesteps (or one rotation). The motion of the center of pressure caused by the contents moving back and forth over the palm is accurately predicted for the full rotation by the ESN. Figures 6a-6d shows the results for a single trial. 6e-6f shows the average error over all the trials. Overall, the ESN is able to accurately predict the center of pressure and amplitude over time for all trials with water.

Table III compares the performances and complexity [2] of different prediction models. The MLP model was implemented by giving the full primer as input (0.1 [s] of signal) and predicting 3 [s] of signal which means that the minimal number of weights it can have is 2101. The LSTM is implemented in the

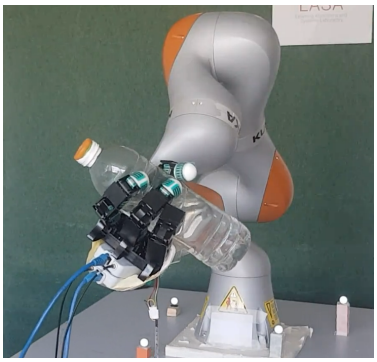
[2]The time of prediction is not reported since the Matlab deep learning library uses low level optimization that could not be implemented for ESNs. This a relevant metric of the complexity of almost all algorithms. Only the LSTM's has non-linear activation functions bringing its computational complexity above the other algorithms.



(a) Initial position for the robotics implementation test



(b) First, the hand rotates 10 degrees on one side. While the bottle remains stable, the hand does not need to close.



(c) The hand rotates 45 degrees in the opposite direction. The hand should close and catch the bottle to avoid it falling.

Fig. 3: Example of Robotic Experiment for Slip Prediction and Recovery. Initially, the container is placed on the robotic palm with the fingers open. As the hand rotates, the model predicts the tactile forces during the motion and anticipates when slip will occur. Upon slip prediction, the robotic hand changes the grip in to a closed envelop grasp over the container to maintain stability.

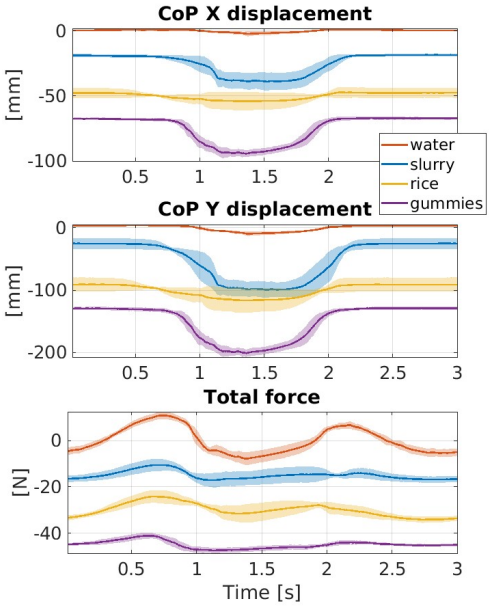


Fig. 4: Center of Pressure and Amplitude data over time for each material (water, slurry, rice, gummies). Mean (line) and variance (shading) is shown for all trials of each material. The effect of different dynamics of each material on both CoP location and amplitude can be seen over the rotation.

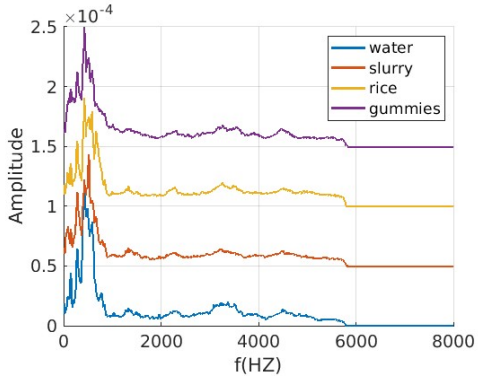


Fig. 5: FFT of the audio signal for different material. Across materials it shows similarities in shape due to the noise of the robot being similar across different material. Differences can be seen in intensity and shape of the peaks

same sequential manner as the ESN. The result show that for a long time horizon, a simple linear interpolation doesn't yield good result. MLP needs about 7 times more complexity than the ESN to match its performance; however, vastly increasing the complexity of the model can yield better results than the ESN. The LSTM network can match ESN performances, and even improve on it, but also requires a lot more complexity to improve on the performances. In all those methods ESN's are showing better performance at the lowest complexity.

	CoP error [mm]	Amplitude error [N]	Num. of weights
Linear predictor (10 [ms])	2.23 ± 1.62	6.20 ± 2.42	2
Linear predictor (3000 [ms])	44. ± 4.01	17.7 ± 9.74	2
ESN (30 hidden units 0.1 connectivity)	3.03±1.67	2.57 ± 1.29	333
MLP (1 hidden unit)	3.89 ± 2.00	5.43 ± 2.83	2101
MLP (30 hidden unit)	1.25 ± 0.73	1.87 ± 1.15	36930
LSTM (30 lstm unit)	3.33 ± 2.19	5.23 ± 2.67	153
LSTM (200 lstm unit)	1.14 ± 0.64	1.05 ± 0.76	333

TABLE III: Comparison of performances for different algorithms

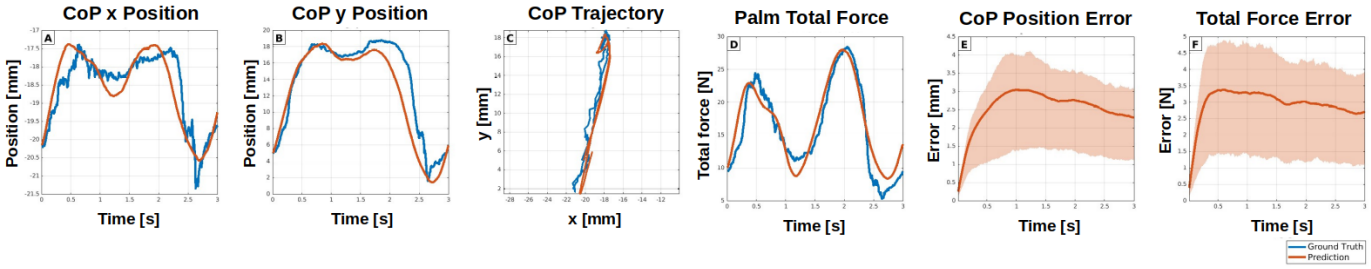


Fig. 6: Prediction of center of pressure over palm for water using the material specific model. (a) and (b) shows the ground truth (blue) and predicted trajectories (red) in the x-domain and y-domain respectively over time for a single trial. (c) shows the ground truth data and predicted trajectories of the center of pressure over the palm in the x-y domain. (d) shows the ground truth and predicted amplitude of pressure on the palm over time. (e) and (f) show the errors over time for the CoP position and amplitude respectively for all trials. The mean (line) and variance (shading) are shown for all trials.

Figures 7 and 8 show the average CoP and force prediction errors for 300 steps into the future respectively over all trials for each material using the material-specific ESNs. The material-specific models accurately predict the CoP position for 300 timesteps into the future for all materials. For both the slurry and gummies, the mean and variance in error is higher, as the motion of the materials inside the container varied depending on how the material stuck to the container and moved during rotation. For both rice and water, which were more predictable in the flow, the overall error during the prediction was lower.

Figure 9 compares the results for generalized and specific model learning. Figure 9 shows average prediction error for CoP position and total force for each material for both the generalized and material-specific models. Figure 9 shows that material-specific models results in lower prediction error compared to the generalized model; however, the generalized model performed better for materials were more predictable flow (i.e., water, slurry and rice) compared to those with more stochastic flows. The generalized model performed comparably well compared to the material-specific models for total force prediction for all materials.

In addition, ESNs trained on only a subset of materials generally result in better performances on the materials within the training set (water and gummies). For unseen materials, the model performs with similar performances as the material-agnostic model but with generally higher variance.

Figure 10 shows that frequency-specific models result in significantly better performance compared to the generalized model, both in CoP and amplitude prediction. In addition, the frequency-agnostic model trained on only 3 of the frequencies show the same performances as one trained on all 5 frequencies, showing the model's ability to generalize to unseen frequencies in the vicinity of the ones it was trained on.

Figure 11 compares the results of using the audio-tactile ESN model compared with the tactile-only ESN. Figure 11a shows that incorporating the audio data into the prediction significantly improves the prediction of the CoP location throughout the trajectory. On the other hand, Figure 11b shows that the amplitude prediction is not significantly improved with audio-data augmentation. This is possibly due to the audio data does not correlating with amount of force the material was exerting on the hand. Figure 12 showed that for all materials, the audio-tactile models resulted in better performance. However, for slurry, the improvement was not significant.

Rotation start on thumb side rotation			
Prediction length	Early Closing	On time	Late closing
10 [ms]	0	1	4
30 [ms]	1	4	0
50 [ms]	5	0	0
Rotation start on pinky side rotation			
10 [ms]	0	0	4
30 [ms]	0	4	1
50 [ms]	4	1	0

TABLE IV: Result of the hand rotation for 5 trials with different prediction lengths for rotations starting either from thumb or pinky side. As prediction horizon increases, the number of late closings decrease and the robotic hand is successfully able to prevent the container from slipping.

Tables IV shows the results of the experimental study conducted using online tactile prediction and control for grip closure. As described above in Section III-E, the container is placed in the open palm, and online tactile prediction is used to anticipate when slip will occur and enact the grip closure command accordingly. As the prediction is done for a shorter time frame, i.e. before the first time step for use of audio data, audio is not used in this setting. For both directions of rotation, as the prediction horizon increases, the number of failed (i.e., late closures) that occur decrease to 0. Importantly, the 10-ms prediction time is insufficient to both recognize and enact the grip closure command in time to prevent the container slipping. For both the 30 and 50 ms prediction horizons, the longer prediction horizons allow for greater anticipation, enabling the robot to complete grip closure with additional time (i.e., early closing). It is of note that by predicting for shorter time frame the ESN has lower variance in the result as can be seen in figure 7 making the experiment more repeatable While this additional time may not be necessary for this scenario, the necessity of long-time horizon predictions is highlighted when reactive torque changes are not sufficient, but grip changes need to occur for stable manipulation.

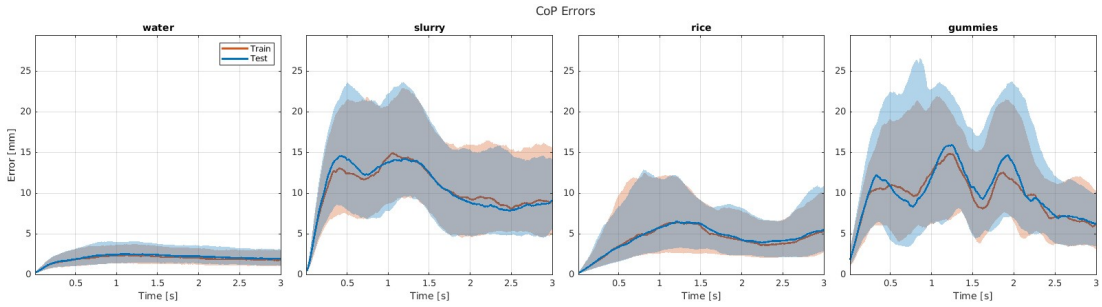


Fig. 7: Average CoP error for training set (shown in blue) and test set (shown in red) during forward prediction of the tactile data values during motion using the material-specific echo state network.

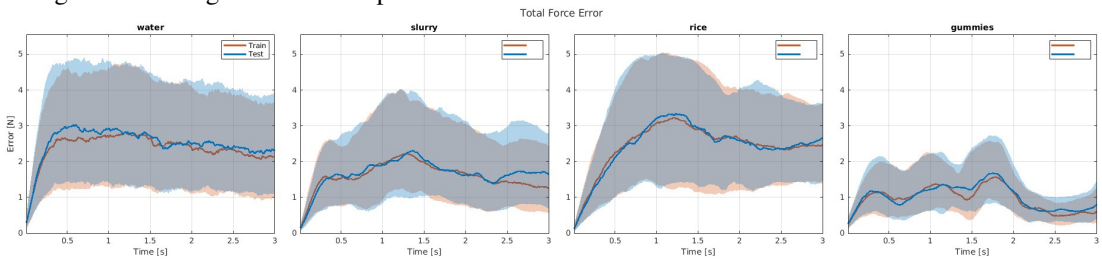


Fig. 8: Average total force error for training set (shown in blue) and test set (shown in red) during forward prediction of the tactile data values during motion using the material-specific echo state network.

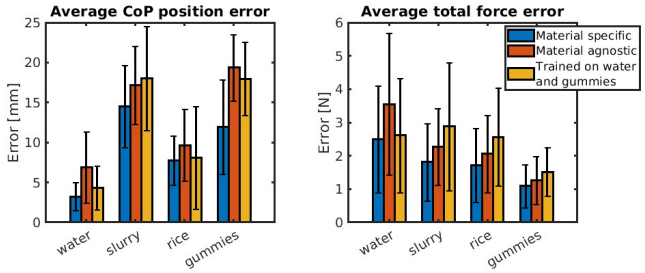


Fig. 9: Comparison of CoP and amplitude prediction errors using generalized and material-specific models as well as model trained on only a subset of material to assess generalization properties.

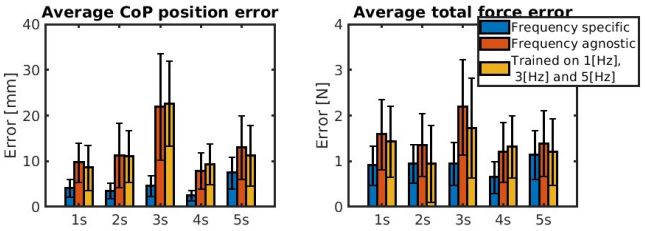
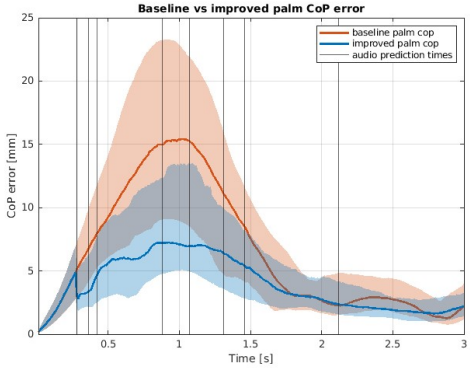


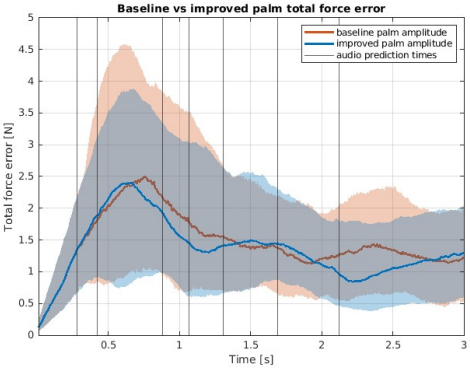
Fig. 10: Comparison of CoP and amplitude prediction errors using generalized, frequency-specific ESN, as well as generalized ESNs tested on frequencies unseen in training.

V. Conclusion

In this paper, we introduced a novel formulation for enabling long-time horizon tactile pressure prediction. We show that echo state networks are capable of accurately predicting tactile pressure trajectory during motion within real-time time bounds and on experimental data collected using the Allegro hand mounted on a Kuka IIWA robot. We show that material-specific echo state networks are effective for capturing and predicting the tactile pressure flow and the additional benefits of audio-data augmentation. We compare the ESN's to other



(a)



(b)

Fig. 11: (a) CoP and (b) Amplitude prediction error over palm over time in water trial using the tactile-only and audio-tactile ESN models. The tactile-only baseline (in red) and audio-tactile model (in blue) are shown with the mean and variance over all trials. Times when the audio prediction is incorporated into the prediction for the audio-tactile model are indicated with the black vertical lines.

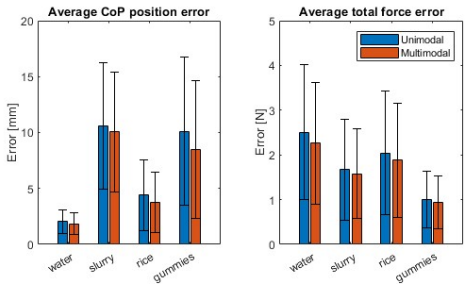


Fig. 12: Comparison of average CoP and amplitude prediction errors for each material using the tactile-only (unimodal) and audio-tactile (multimodal) models. The multimodal models resulted in performance improvement for all materials in both yielded Improvement yielded by the audio prediction for each material

machine learning models, and show that ESN's have the lowest computational complexity while being able to predict tactile data with sufficient accuracy, making them the best choice when looking to minimize the computational impact of tactile prediction on any low capacity hardware. We validate our results for different material contents and online in a real-world experimental setup, where for models with lower computational requirements are required . Finally, we highlight the importance of long-time tactile prediction for manipulation in tasks where grip changes need to occur. Future work seeks to improve the model to accommodate a wider variety of materials and motion types. Furthermore, we seek to use the proposed framework with a more complex grasp planner that would benefit from long-time horizon prediction.

ACKNOWLEDGMENT

We would like to thank Stanislas Furrer, Lorenzo Panchetti, and Maxime Perret for their help in the initial experimental setups for this work.

REFERENCES

[1] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[2] R. A. Romeo and L. Zollo, "Methods and sensors for slip detection in robotics: A survey," *Ieee Access*, vol. 8, pp. 73 027–73 050, 2020.

[3] B. S. Zapata-Impata, P. Gil, and F. Torres, "Learning spatio temporal tactile features with a convlstm for the direction of slip detection," *Sensors*, vol. 19, no. 3, p. 523, 2019.

[4] M. Meier, F. Patzelt, R. Haschke, and H. J. Ritter, "Tactile convolutional networks for online slip and rotation detection," in *International Conference on Artificial Neural Networks*. Springer, 2016, pp. 12–19.

[5] J. Reinecke, A. Dietrich, F. Schmidt, and M. Chalon, "Experimental comparison of slip detection strategies by tactile sensing with the biotac® on the dlr hand arm system," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 2742–2748.

[6] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 241–248.

[7] Z. Su, K. Hausman, Y. Chebotar, A. Molchanov, G. E. Loeb, G. S. Sukhatme, and S. Schaal, "Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 297–303.

[8] S. Clarke, T. Rhodes, C. G. Atkeson, and O. Kroemer, "Learning audio feedback for estimating amount and flow of granular material," in *Proceedings of The 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 87. PMLR, 29–31 Oct 2018, pp. 529–550.

[9] H. Liang, C. Zhou, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, M. Stoffel, and J. Zhang, "Robust robotic pouring using audition and haptics," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 880–10 887.

[10] H. Huang, X. Guo, and W. Yuan, "Understanding dynamic tactile sensing for liquid property estimation," in *Robotics: Science and Systems XVIII, New York City, NY, USA*, 2022.

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

[12] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.

[13] H. Liu, K. C. Nguyen, V. Perdereau, J. Bimbo, J. Back, M. Godden, L. D. Seneviratne, and K. Althoefer, "Finger contact sensing and the application in dexterous hand manipulation," *Autonomous Robots*, vol. 39, no. 1, pp. 25–41, 2015.

[14] Q. Li, C. Schürmann, R. Haschke, and H. J. Ritter, "A control framework for tactile servoing." in *Robotics: Science and systems*. Citeseer, 2013.

[15] N. Wettels, A. Parnandi, J.-H. Moon, G. Loeb, and G. Sukhatme, "Grip control using biomimetic tactile sensing systems," *Mechatronics, IEEE/ASME Transactions on*, vol. 14, pp. 718 – 723, 09 2010.

[16] B. Sundaralingam, A. S. Lambert, A. Handa, B. Boots, T. Hermans, S. Birchfield, N. Ratliff, and D. Fox, "Robust learning of tactile force estimation through robot interaction," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9035–9042.

[17] A. Garcia-Garcia, B. S. Zapata-Impata, S. Orts-Escolano, P. Gil, and J. Garcia-Rodriguez, "Tactilegcn: A graph convolutional network for predicting grasp stability with tactile sensors," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[18] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional lstm network," *arXiv preprint arXiv:1810.02653*, 2018.

[19] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 818–824.

[20] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," *Advances in neural information processing systems*, vol. 29, 2016.

[21] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[22] ——, "Adaptive nonlinear system identification with echo state networks," *Advances in neural information processing systems*, vol. 15, 2002.

[23] N. Harischandra and V. Dürr, "A forward model for an active tactile sensor using echo state networks," in *2012 IEEE International Symposium on Robotic and Sensors Environments Proceedings*, 2012, pp. 103–108.

[24] Y. Jonetzko, N. Fiedler, M. Eppe, and J. Zhang, "Multimodal object analysis with auditory and tactile sensing using recurrent neural networks," in *International Conference on Cognitive Systems and Signal Processing*. Springer, 2020, pp. 253–265.

[25] J. Sinapov, C. Schenck, and A. Stoytchev, "Learning relational object categories using behavioral exploration and multimodal perception," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 5691–5698.

[26] S. Jin, H. Liu, B. Wang, and F. Sun, "Open-environment robotic acoustic perception for object recognition," *Frontiers in neurorobotics*, vol. 13, p. 96, 2019.

[27] Q. Liu, F. Feng, C. Lan, and R. H. Chan, "Va2mass: Towards the fluid filling mass estimation via integration of vision and audio learning," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 451–463.

[28] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 659–686.