

An Investigation of Multi-feature Extraction and Super-resolution with Fast Microphone Arrays

Eric T. Chang^{*1}, Runsheng Wang^{*1}, Peter Ballentine¹, Jingxi Xu¹, Trey Smith², Brian Coltin², Ioannis Kymmissis¹, and Matei Ciocarlie¹

Abstract—In this work, we use MEMS microphones as vibration sensors to simultaneously classify texture and estimate contact position and velocity. Vibration sensors are an important facet of both human and robotic tactile sensing, providing fast detection of contact and onset of slip. Microphones are an attractive option for implementing vibration sensing as they offer a fast response and can be sampled quickly, are affordable, and occupy a very small footprint. Our prototype sensor uses only a sparse array of distributed MEMS microphones (8-9 mm spacing) embedded under an elastomer. We use transformer-based architectures for data analysis, taking advantage of the microphones’ high sampling rate to run our models on time-series data as opposed to individual snapshots. This approach allows us to obtain 77.3% average accuracy on 4-class texture classification (84.2% when excluding the slowest drag velocity), 1.5 mm median error on contact localization, and 4.5 mm/s median error on contact velocity. We show that the learned texture and localization models are robust to varying velocity and generalize to unseen velocities. We also report that our sensor provides fast contact detection, an important advantage of fast transducers. This investigation illustrates the capabilities one can achieve with a MEMS microphone array alone, leaving valuable sensor real estate available for integration with complementary tactile sensing modalities.

I. INTRODUCTION

Designers of robotic tactile sensors can typically choose from many different underlying transduction types, including magnetic, piezoresistive, piezocapacitive, optical, etc. Among these, microphones are appealing for a number of reasons: off-the-shelf microphone integrated circuits offer fast response time in a low cost package with a very small form factor. In particular, due to the frequency response and fast sampling, microphones effectively respond to vibrations and fast changes in contacts, thus enabling applications such as material identification, fast initial touch detection, and slip detection via vibrations [1]–[4].

Microphones typically fall in the category of dynamic sensors (which also includes piezoelectrics, accelerometers, etc.), or sensors which respond to *changes* in contact pressures. In contrast, static transduction methods provide a response to constant pressures and deformations. The inability of microphones to provide such static responses means they are generally not used for determining aspects such as contact location or applied force.

^{*}equal contribution

¹Columbia University, New York, NY 10027, USA. {eric.chang@, rw2967@, peter.ballentine@, jxu@cs., johnkym@ee., matei.ciocarlie@columbia.edu}

²Intelligent Robotics Group, NASA Ames Research Center, Moffett Field, CA 94035, USA {brian.coltin, trey.smith}@nasa.gov

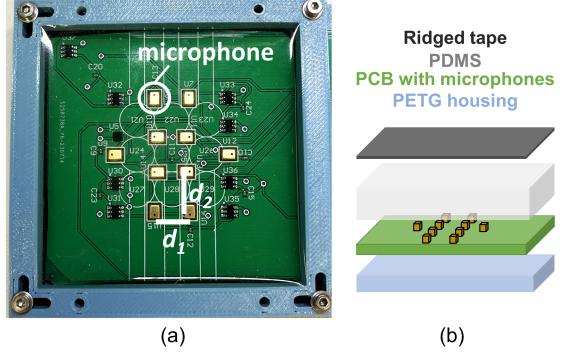


Fig. 1. Photo of the microphone array without the textured tape (a), and an exploded view of the sensor fabrication process (b). The inner square of four microphones are spaced $d_1 = 8$ mm apart. The outer ring of microphones are spaced $d_2 = 9$ mm away from the inner square.

A multimodal tactile array could in theory combine multiple transduction types to get the benefit of all worlds, e.g. use microphones for texture classification and static piezocapacitive transducers for contact localization. However, physical space on a sensor’s surface is always at a premium, and any new modality adds overhead in terms of the complexity of manufacturing for both the sensor and the signal collection and processing.

Thus, in this paper, we investigate the possibility of inferring multiple aspects of contact, including texture, location, and drag velocity, using only a distributed array of MEMS microphones as the underlying transducers (Fig. 1). We aim to combine these abilities with microphones’ natural effectiveness in fast detection of initial contact, due to their native high sensitivity and frequency response. We believe these results will prove useful to designers of tactile systems having to decide which underlying transduction modalities to include, how far apart to space individual sensing elements, etc.

A significant development in robotic tactile sensing is the advent of data-driven algorithms that are used on top of the raw sensor signals. For example, learning algorithms have been used with static sensors for object recognition, slip detection, contact localization and super resolution, etc. [5]–[7]. Likewise, using vibration data from dynamic sensors to learn classifiers for material identification is well studied [2], [8]–[12]. However, applying learning algorithms to dynamic sensor data to learn other tactile features is underexplored. Here, we use our distributed microphone sensing platform to

train models from the time domain signals to classify texture, localize contact, and predict drag velocity. In addition, our array also enables fast initial contact detection. Importantly, these are not isolated tasks with carefully selected runtime conditions. Rather, running these models together provides rich, realtime information about the state of sensor contact from solely sparse, dynamic signals. Overall, the main contributions of this paper are as follows:

- To our best knowledge, we are the first to use distributed microphones to simultaneously classify texture, localize contact, and estimate contact velocity, and show that super-resolution techniques can detect aspects of touch previously unreported using transducers of this type.
- We combine the capabilities above with the native abilities of dynamic sensors to provide fast response time and high sensitivity for initial contact onset, resulting in a versatile tactile array despite employing a single transducer type.

II. RELATED WORK

Using vibration-based tactile sensing for material or texture classification has been studied extensively in the robotics literature. Many works use extracted frequency information as input to their classification algorithms [2], [8]–[10]. More recent work is able to use deep learning algorithms to learn texture classifiers that operate on time series input [11]. Since there are texture signatures in the frequency domain that are independent of scanning velocity [13], [14], it has been shown that texture perception can be impressively robust to different scanning velocities [8], [12]. While artificial texture perception has reached impressive performance, these works focus solely on the problem of texture classification. We are interested in what other features can be extracted from a microphone array.

Looking beyond the problem of material classification, there is extensive work on distributed tactile arrays that are sampled at high frequencies and measure vibrations [15]–[17]. These sensors are typically taxel based, and therefore provide spatial information directly. For example, previous work has taken advantage of the spatial acuity of the iCub’s capacitive array [18] to combine both spatial and sliding feedback to improve texture classification [9]. While these works focus on sensors that can sense constant pressures and have dense taxel resolution, our study examines sparse dynamic sensors that do not respond to constant pressures.

There are other works on learning from tactile arrays consisting of only dynamic modalities. The NeuTouch [19] and NUSkin [20] sensors are state-of-the-art tactile sensors using only dynamic piezoresistive transducers to detect and interpret changes in pressure asynchronously. Using various deep learning techniques designed for spike-based inputs, these sensors have been used for container classification, rotational slip detection, [19] and to detect and decode vibrations traveling through objects [20]. However, these works deal with dense dynamic sensing arrays. In this study, we are interested in what features can be learned from a sparser array of microphones.

Lastly, there are a few other works using microphone arrays to learn tactile features. A pair of studies used a tactile skin including sparsely distributed microphones to classify gestures performed by human subjects on the skin via deep learning [21], [22]. While the ability to classify human performed gestures is impressive, the work does not explore localization, velocity, or texture models with the sparse microphone array.

III. MICROPHONE ARRAY DESIGN AND FABRICATION

For this work, we built a planar sensor as a test platform for studying the capabilities of distributed MEMS microphones. Our sensor, shown in Fig. 1, comprises an array of microphones covered with an elastomer layer (PDMS), which acts to distribute strain applied at the surface to multiple underlying transducers, mimicking a finger’s deformable surface.

We designed the array such that the receptive fields are overlapping: one contact event will activate multiple signals. We did this by adjusting the microphone placement, gel/microphone interface, gel stiffness, and gel thickness. Overlapping receptive fields are valuable when using this platform with high capacity machine learning models that can benefit from multiple rich signal inputs. That said, we limit the microphone density in order to cover a large area while leaving room for other sensors, as future designs might combine a vibratory microphone array with complementary modalities.

A. Electronics

This sensor contains a printed circuit board (PCB) with a 24x24 mm sensing area. The PCB contains 10 top-port analog MEMS microphones (Knowles SPU0410HR5H-PB), spaced 8-9 mm apart (Fig. 1). The microphone signals are fed through a multiplexer, followed by a voltage amplifier with a gain of 5.5 and a bias of 1 V. The microphones are sampled at 2500 Hz with the 12-bit analog-to-digital converter on a Teensy 3.6 microcontroller. We use micro-ROS to communicate between a PC and the Teensy over USB [23].

B. Fabrication

To fabricate the sensor, we place the PCB in a Prusament polyethylene terephthalate glycol (PETG) housing. We mix PDMS in a 30:1 weight ratio and degas the PDMS in a vacuum chamber to remove all bubbles. We then place the sensor in a vacuum chamber (-0.75 bar) and pour 17 grams PDMS onto the PCB (6 mm thickness). We pull a vacuum prior to pouring the PDMS to draw the PDMS into the acoustic port, resulting in less hysteresis and increased sensitivity. We degas any remaining bubbles formed during pouring by blowing the sample with a heat gun or leaving the sample under vacuum. We cure the sample at 75°C for 8 hours and allow the sample to cool in the oven.

As shown in other work [24], we find implementing a ridged sensor surface important to achieve texture classification. Imitating the papillary ridges (i.e., fingerprints) of

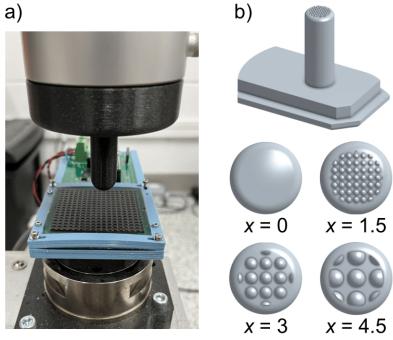


Fig. 2. Our tactile sensor on an F/T sensor (a) and the 4 textures used in data collection (b). The four textures (“a,” “b,” “c,” “d”) follow $x = 0 - 4.5$ mm bump spacing and are printed on 15 mm diameter indenters. The bump diameter is determined by the formula $x\sqrt{\frac{2}{\pi}}$ [8]. The PLA indenter shown in (a) was used for the response time data (Sec. VI)

the human fingers, the ridges serve to amplify and create vibrations during interaction with other textures. To achieve this, we place a single square of patterned high density polyethylene strips (“Non-Slip Grip Strips” by CatTongue) onto the sensor surface. For consistency, we always use the sensor with the tape applied and replace the tape when the ridges begin to fade. We do not observe any issues of adhesion between the tape and the PDMS. Using a removable tape provides standardized, repeatable friction properties, is more durable than an exposed PDMS surface, and is easily replaceable when it wears down or becomes damaged.

IV. DATA-DRIVEN MULTI-FEATURE EXTRACTION

In this study, we leverage machine learning methods to characterize different aspects of contact on our sparse microphone array. Specifically, we use transformer-based architectures to classify texture, localize contact, and estimate contact velocity. The ultimate aim is to investigate the extent to which a microphone array can extract diverse features of a contact episode.

These three tasks are all important elements of tactile perception. Texture discrimination is critical for material classification and tactile exploration. Localization is important for having spatial understanding of a contact. Perceiving contact velocity is an important problem for tasks such as estimating slip velocity and perception during dynamic manipulation scenarios. For slip detection, it is important to estimate velocity over a short time window, since introducing delays removes the ability to use velocity as realtime feedback.

A. Training Data

Using machine learning models requires a large amount of training data to be collected. In this proof of concept study, we do not move the sensor around to interact with the environment as one might do with a tactile finger. Rather, we imitate contact with various surfaces by dragging different indenters across the sensor. Our collected dataset contains isolated drag episodes with uniquely textured indenters, varying velocities, and varying z forces.

The data collection setup consists of our microphone array mounted onto an ATI Industrial Automation Force/Torque (F/T) sensor (Gamma series, calibration SI-130-10) which provides six signals: force and torque in three axes. This sits below a Universal Robots UR5 robot arm. Our microphone array updates at 2500 Hz, the F/T sensor updates at 2400 Hz, and the robot position and velocity data updates at around 130 Hz. We read the 10 microphones, the 6 F/T signals, and the robot pose and velocity at 2000 Hz.

We use four (“a,” “b,” “c,” “d”) different 15 mm diameter textured indenters (Formlabs photopolymer resin), which are mounted onto the UR5 arm. The textures contain hemispherical bumps and are chosen to represent a range of bump densities. The textures are a subset of those used in previous work [8]. The bump diameter and spacing is determined in such a way that preserves contact area across the textures (excluding the flat texture), to avoid introducing spurious features unrelated to the texture (Fig. 2).

We follow one data collection procedure for all of the training data. We collect drag data at velocities ranging from 20 mm/s to 60 mm/s in 5 mm/s increments using fixed accelerations of 0.3-0.4 m/s². The z height and z rotation of the indenter are randomized for every drag to be robust to these factors. The data collection procedure is detailed in Alg. 1. In total, we collected approximately 200 drags per velocity per texture, totalling approximately 7200 drags.

Algorithm 1 Dragging Data Collection

- 1: Calibration: set the Z height bounds (Z_{max} , Z_{min}) to positions corresponding to 1N and 5N
 - 2: Move to 5 mm above a randomly sampled starting point along the border of the sensing area
 - 3: **while** collecting data **do**
 - 4: Move to a Z height sampled from $[Z_{min}, Z_{max}]$
 - 5: Rotate to a Z -axis angle sampled from $[0, 45]$ degrees
 - 6: Move (while dragging along the sensor) to a new randomly sampled point along the edge of the sensing area that is ≥ 15 mm from the current point
 - 7: **end while**
-

B. Preprocessing

Before training, we do a number of preprocessing steps on the raw data. First, we center the microphone signals around 0 by subtracting a baseline value independently for each microphone. To extract segments of data corresponding to drags, we ignore any data regions where the robot arm was not moving with significant velocity, yielding a collection of segments, each representing a single, complete drag. The dataset is then augmented by sliding a window across each complete drag segment, with overlaps between adjacent windows.

We adapt window sizes according to the specific task at hand. We use a window size of 100 timesteps (0.05 seconds) for position prediction, 200 timesteps (0.1 seconds) for velocity estimation, and 500 timesteps (0.5 seconds) for

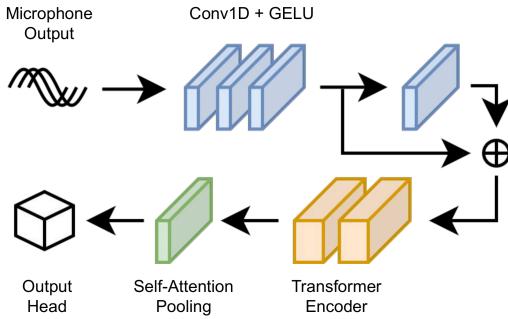


Fig. 3. An illustration of our network architecture.

texture classification. Across all tasks, we maintain a uniform window offset of 50 timesteps to allow for overlaps. To eliminate the DC components in each window, we apply a third-order bidirectional high-pass filter with a cutoff frequency at 3 Hz. Each resulting sliding window is considered a sample, containing 10-channel sensor readings over n timesteps at a sampling frequency of 2000 Hz. Using this approach, we were able to extract over 110000 samples from the raw data.

For each sample in our dataset, we extract the texture labels directly from the associated data files. Position labels are acquired by capturing the robot arm’s encoder-reported position at the last timestep within the established sliding window. To obtain velocity labels, we calculate the median velocity within the sliding window from the UR5 velocity encoders, a measure chosen for its robustness to minor fluctuations.

C. Learning Model Architecture

Our machine learning model consists of a neural network comprised of two primary modules: a convolution module and a transformer module. The convolution module employs three layers of 1-D strided convolution along the temporal dimension for feature encoding. We apply GELU activation between each pair of convolution layers. The convolutional module outputs a compressed 10-channel latent representation.

Before passing the latent representation into the transformer, we apply an additional convolution to the latent representation, pass the output through a GELU activation function, and then sum it with the original latent representation. The operation was introduced to capture a more expressive set of attributes at two different levels of abstraction.

We pass the enriched latent representation into a 2-layer transformer encoder block. To connect the transformer output with the subsequent classification and regression layers, we use self-attention pooling [25]. The pooled output is directed either to a classification head for texture discrimination or to a regression head for predicting position and velocity. Collectively, the convolution and transformer modules serve as a unified signal encoder.

D. Training

We employ task-specific methodologies when partitioning training, validation, and test sets. To evaluate the generaliza-

tion capabilities of our model, we adopt a held-out-velocity strategy for texture and position tasks [8]. The model is trained on discrete drag velocity profiles within a predefined range and evaluated on a distinct velocity profile that, while within the range, was not part of the training set. We conduct cross-validation using five different velocity profiles as the held-out set. For velocity prediction, we execute ten rounds of cross-validation, each utilizing a unique held-out test set containing samples from a diverse range of velocities. Whenever sampling from the datasets, we sample individual drags before dividing into windows to ensure that adjacent windows from the same drag do not end up in different training/validation/test sets. On all tasks, we select the best-performing model using validation loss, and report test set performances using the selected model.

For texture classification and both regression tasks, we use cross-entropy and mean squared error (MSE) loss, respectively. We train on an RTX 3090 Ti GPU with the AdamW optimizer at 0.0001 learning rate and a batch size of 64. The texture model converges within 20 epochs, whereas velocity model and localization model converges in 30 and 60 epochs, respectively. We use random seed 42 across all tasks.

V. CONTACT CHARACTERIZATION PERFORMANCE

A. Texture Classification

Upon training a texture classifier, we evaluate the model’s ability to distinguish between the four textures while generalizing to velocities not seen in training. We plot confusion matrices across five different test sets, where each matrix is the performance on a velocity that was held out during training (Fig. 4). The average accuracy across the 5-fold cross validation is 77.3%. The model’s robustness to varying velocities and strong performance on unseen velocities suggests that the model learns to classify textures based on signal features that are independent of velocity.

We note that the 20 mm/s velocity has significantly worse performance than the other velocities. This is likely due to a slower drag causing less prominent signals, and therefore fewer distinguishing characteristics between the textures. Poor classification performance on this velocity relative to higher velocities is consistent with previous work [8]. Excluding this velocity, the average classification performance among the other test sets is 84.2%. We also observe from the confusion matrices that when samples are misclassified, the models tend to predict a neighboring texture.

B. Localization

The learned localization regressor estimates the current contact position from a short history of data (100 timesteps, 0.05 seconds). As with the texture model, we examine the localizer’s performance on five test sets containing unseen velocities. The average mean/median error over the five test sets is 1.8/1.5 mm.

We find this learned position model to outperform a heuristic-based method, which assigns the position of the microphone with the highest signal-to-noise ratio (SNR) over the history window as the estimated location. The mean /

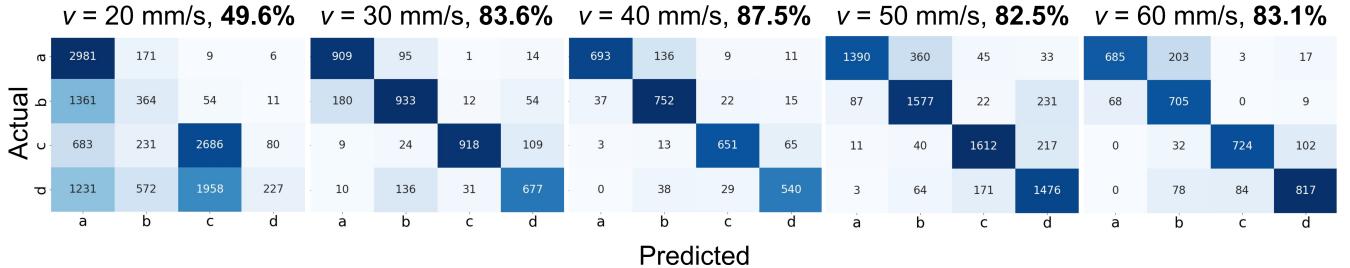


Fig. 4. Confusion matrices on 4 class texture classification. Each matrix is for a different test set, each of which correspond to the held-out velocities.

median localization error for the baseline is 5.5 / 5.3 mm. We note that using SNR over the history window means that this baseline predicts the average position over the entire window of history, whereas the learned model is trained to predict the most recent position. That said, the delay introduced by such a short window (0.05 seconds) is small.

C. Direct Velocity Regression

We analyze velocity estimation results when the model has access to different windows of history: 0.05, 0.10, and 0.25 seconds. We achieve a median absolute velocity error of 5.2, 4.5, and 5.0 mm/s for 0.05, 0.10, and 0.25 second history windows, respectively, or 13.0%, 11.3%, and 12.5% of the 20 - 60 mm/s range of velocities explored (Tab. I). We find that a window larger than 0.05 seconds of history is necessary for learning meaningful velocity estimation; however, results do not noticeably improve for the 0.25 s window (Tab. I).

We hypothesize that directly estimating velocity with a designated model will be more accurate than using the derivative of a position measurement. To test this hypothesis, we compare our results to inferring velocity using the learned position model. We make this velocity estimate by using the estimated start and end positions over a 200 timestep (0.10 s) time window, resulting in a median error of 9.7 mm/s. It is not surprising that the learned model outperforms this baseline, as errors in the position estimation will be amplified when computing velocity over a short time window. However, for the conditions investigated, this result suggests that there is value to making direct velocity estimates.

VI. RESPONSE TIME ANALYSIS

While being able to extract diverse features from a microphone array is useful, the primary benefit of using vibration modalities in practice is for their fast response time and sensitive contact detection. Therefore, to complement our machine learning study, we also provide an analysis of this sensor's response time to illustrate this benefit in a MEMS microphone array.

To achieve this, we measure the relative response time between a single microphone and an F/T sensor used as a baseline – i.e., the difference between the time the F/T sensor and the microphone detected contact. While an F/T sensor is not a ground truth measurement of the start of contact, we treat it as a baseline as follows: we use the F/T data offline to estimate when the true contact began (i.e.,

we look ahead in the data to assign contact labels), resulting in contact identification while the F/T signal is still in the sensor's noise range, and thus earlier than the F/T sensor could have detected it in realtime. We detail this procedure below.

A. Data Collection and Processing

To collect data for this experiment, we use the same data collection setup as for the training data, but with a hemispherical, polylactic acid (PLA) indenter on the UR5 arm, and with a downward motion instead of a dragging one (Fig. 2). We also increase the rate at which the F/T sensor updates the ethernet port to 5000 Hz. However, we note that the internal F/T strain gauge measurements always occur at 7000 Hz [26]. We also increase the rate at which data is synchronized to a file to 2300 Hz. We selected a single microphone with a representative sensitivity to use for the analysis.

The data collection process consisted of an indentation procedure repeated over 45 episodes. The procedure consisted of: 1) move the indenter above the sensor such that a constant velocity can be reached before contact, 2) move the indenter downwards at a preset velocity (acceleration = 4.0 m/s²) onto a preset location at a depth correlated with 2-3 N, and remain on the sensor for 1 second.

We select the contact points to be 0, 2, 4, and 6 mm away from the microphone of interest, and we set the incoming velocities to be 10, 55, and 100 mm/s. Empirically, we find the microphone receptive fields to significantly decay around 4 mm away from the microphone, so we include 6 mm in our analysis to look at the response beyond the nominal receptive field. In practice, none of the microphones in the majority of the sensing area are spaced more than 8-9 mm apart (Fig. 1), so 6 mm away from a single microphone is always well into the receptive field of another.

Before analyzing where contact starts, we first use a section of data at the beginning of each episode to individually bias (i.e., tare) the episode to offset drift in the F/T readings. We also remove any episodes in which the contact data was affected by the F/T sensor momentarily flat-lining and failing to update. After these steps, we are left with at least 30 episodes per location and velocity condition.

Identifying F/T Contact Time. To identify the time at which the F/T sensor detects contact for each episode, we use an offline heuristic to identify the start of contact as

TABLE I

LOCALIZATION AND VELOCITY ESTIMATION ERROR. LOCALIZATION PERFORMANCE IS REPORTED ON VELOCITIES NOT SEEN IN TRAINING (MEDIAN ERROR OVER 5 TEST SETS IS 1.6 MM). VELOCITY ESTIMATION IS TRAINED ON ALL VELOCITIES AND REPORTED IN 5MM/S VELOCITY BINS (MEDIAN VELOCITY ERROR IS 5.1 / 4.5 / 5.3 MM/S FOR A 0.05 / 0.10 / 0.25 S HISTORY WINDOW.

| Task | History (s) | Drag Speed (mm/s) | | | | | | | | | |
|------------------------|-------------|-------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--|
| | | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | |
| M/Mdn Loc. Err. (mm) | 0.05 | 2.0 / 1.6 | - | 1.9 / 1.6 | - | 1.6 / 1.3 | - | 1.8 / 1.6 | - | 1.8 / 1.5 | |
| M/Mdn Vel. Err. (mm/s) | 0.05 | 7.9 / 6.4 | 5.5 / 4.5 | 4.3 / 2.8 | 5.5 / 5.1 | 6.0 / 5.6 | 6.3 / 5.2 | 6.3 / 4.5 | 7.0 / 4.8 | 9.7 / 7.8 | |
| M/Mdn Vel. Err. (mm/s) | 0.10 | 6.3 / 5.5 | 5.0 / 4.3 | 4.2 / 2.9 | 5.0 / 4.5 | 5.1 / 4.4 | 5.1 / 4.0 | 5.2 / 3.6 | 6.1 / 4.4 | 8.8 / 7.3 | |
| M/Mdn Vel. Err. (mm/s) | 0.25 | 9.7 / 9.5 | 8.0 / 8.0 | 5.1 / 4.3 | 4.8 / 3.8 | 5.4 / 5.0 | 4.4 / 3.7 | 3.7 / 2.7 | 3.9 / 2.7 | 6.6 / 5.7 | |

TABLE II

MEAN (STANDARD DEVIATION) RELATIVE RESPONSE TIME (IN MS) BETWEEN THE MICROPHONE AND THE F/T SENSOR.

| | 0 mm | 2 mm | 4 mm | 6 mm |
|----------|-----------|-----------|-----------|-----------|
| 10 mm/s | 3.0 (3.1) | 4.2 (2.5) | 5.1 (1.4) | - |
| 55 mm/s | 4.8 (2.5) | 3.1 (1.4) | 3.0 (1.3) | 4.0 (1.6) |
| 100 mm/s | 3.0 (1.3) | 2.6 (1.7) | 3.1 (1.6) | 3.8 (2.3) |

early as possible. The heuristic would not work in realtime because it looks both forward and backward in the data. By sliding a window function across each episode, we found this combination of heuristics to identify contact consistently: 1) at least 1 point in the window is less than -1 N; 2) all future points in the window are less than the current point; 3) the curve starts decreasing within the next 4 timesteps; and 4) the curve decreases significantly 25 timesteps later. We verify that each episode was processed correctly, noting that the heuristic is not perfect and occasionally identifies contact too early.

Identifying Microphone Contact Time. Whereas an offline heuristic is appropriate for the F/T data, identifying the microphone contact time requires a realtime heuristic so that the measured response time is relevant in practice. We chose a heuristic that identifies contact when the microphone signal is at least 18 ADC counts higher than the median of the last 20 signals. Intuitively, this examines if the signal has sufficiently exceeded the noise floor. We found this heuristic to work in realtime with almost no false positives.

B. Response Time Analysis Results

We find the average response time of the sensor to range from 2.6 - 5.1 ms across the conditions studied (Tab. II). The response shows intuitive trends as we vary velocity and distance. Response time decreases with increasing velocity, especially going from 10 to 55 mm/s. Likewise, as the tap moves further away from the microphone, response time also decreases. At the 6 mm and 10 mm/s condition, the contacts were not detected using our microphone contact heuristic. That said, these responses have signal changes noticeable to the naked eye, but they are small relative to the other responses and come as far as 200 ms after initial contact, making them impractical to use in realtime.

A 3 ms response correlates to a response above 330 Hz – faster than many static sensor arrays. However, the microphone has a reported sensitivity limit of 10 kHz and we sample the system at 2300 Hz, putting the Nyquist frequency at 1150 Hz – far higher than our measured response time. We therefore attribute the limiting factor of the response time to the latency introduced by the elastomer.

Limitations of this Analysis. Although we denote the start of contact to be just before the F/T sensor senses the contact, this is not a precise ground truth measurement for the beginning of contact as the F/T sensor has its own response characteristics. According to its datasheet [26], the -3dB cutoff of the frequency response is approximately 1800 Hz, and it can deliver data to the ethernet port with a delay of 0.286 ms. In addition, we do not quantify delays of reading from the ethernet port, serial communication with the Teensy, or with ROS2 publishers and subscribers. Ultimately, this experiment provides an estimate of the response time relative to a fast or quasi-real-time baseline.

VII. CONCLUSIONS

This study has provided an investigation into the ability of distributed MEMS microphone arrays to provide information about multiple aspects of contact including simultaneous texture classification, contact localization, and contact velocity determination. In particular, contact localization and velocity determination are previously unreported using this transduction modality. We believe that the ability to determine such features stems from data-driven techniques that make use of a short window of data as opposed to a single snapshot, processed by a self-attentive transformer architecture that is known to be highly effective at analyzing time-series data. We complement these results with the ability of MEMS microphones to perform fast initial contact detection using simple heuristics.

This study showcases that a diverse set of tactile features can be extracted from *only* a sparse microphone array. Thus, using microphones in this way can be a valuable method of implementing fast, low cost, and space-efficient vibration sensing. In the pursuit of information rich tactile sensors that mimic both the static and vibration receptors in human hands, we hope these results will provide insight into the capabilities of MEMS microphones as a tactile modality.

ACKNOWLEDGMENT

We thank Kevin Kam for help with PCB design.

REFERENCES

- [1] Z. Kappassov, J.-A. Corrales, and V. Perdereau, “Tactile sensing in dexterous robot hands — review,” *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921889015001621>
- [2] J. Fishel and G. Loeb, “Bayesian exploration for intelligent identification of textures,” *Frontiers in Neurorobotics*, vol. 6, 2012. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fnbot.2012.00004>
- [3] J. A. Fishel and G. E. Loeb, “Sensing tactile microvibrations with the biotac — comparison with human sensitivity,” in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 1122–1127.
- [4] S. Teshigawara, T. Tsutsumi, S. Shimizu, Y. Suzuki, A. Ming, M. Ishikawa, and M. Shimojo, “Highly sensitive sensor for detection of initial slip and its application in a multi-fingered robot hand,” in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1097–1102.
- [5] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, “The feeling of success: Does touch sensing help predict grasp outcomes?” 2017.
- [6] J. Lin, R. Calandra, and S. Levine, “Learning to identify object instances by touch: Tactile recognition via multimodal matching,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3644–3650.
- [7] P. Piacenza, K. Behrman, B. Schifferer, I. Kymmissis, and M. Ciocarlie, “A sensorized multicurved robot finger with data-driven touch sensing via overlapping light signals,” *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2416–2427, 2020.
- [8] N. Pestell and N. F. Lepora, “Artificial sa-i, ra-i and ra-ii/vibrotactile afferents for tactile sensing of texture,” *Journal of The Royal Society Interface*, vol. 19, no. 189, p. 20210603, 2022. [Online]. Available: <https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2021.0603>
- [9] T. Taunyazov, H. F. Koh, Y. Wu, C. Cai, and H. Soh, “Towards effective tactile identification of textures using a hybrid touch approach,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4269–4275.
- [10] K. Dai, X. Wang, A. M. Rojas, E. Harber, Y. Tian, N. Paiva, J. Gnehm, E. Schindewolf, H. Choset, V. A. Webster-Wood, and L. Li, “Design of a biomimetic tactile sensor for material classification,” in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10774–10780.
- [11] S. S. Baishya and B. Bäuml, “Robust material classification with a tactile skin using deep learning,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 8–15.
- [12] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, “Multimodal feature-based surface material classification,” *IEEE Transactions on Haptics*, vol. 10, no. 2, pp. 226–239, 2017.
- [13] H. P. Saal, X. Wang, and S. J. Bensmaia, “Importance of spike timing in touch: an analogy with hearing?” *Current Opinion in Neurobiology*, vol. 40, pp. 142–149, 2016, systems neuroscience. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0959438816300988>
- [14] Z. M. Boundy-Singer, H. P. Saal, and S. J. Bensmaia, “Speed invariance of tactile texture perception,” *Journal of Neurophysiology*, vol. 118, no. 4, pp. 2371–2377, 2017, PMID: 28724777. [Online]. Available: <https://doi.org/10.1152/jn.00161.2017>
- [15] W. Lin, D. Zhang, W. W. Lee, X. Li, Y. Hong, Q. Pan, R. Zhang, G. Peng, H. Z. Tan, Z. Zhang, L. Wei, and Z. Yang, “Super-resolution wearable electrotactile rendering system,” *Science Advances*, vol. 8, no. 36, p. eabp8738, 2022. [Online]. Available: <https://www.science.org/doi/abs/10.1126/sciadv.abp8738>
- [16] T. P. Tomo, A. Schmitz, W. K. Wong, H. Kristanto, S. Somlor, J. Hwang, L. Jamone, and S. Sugano, “Covering a robot fingertip with uskin: A soft electronic skin with distributed 3-axis force sensitive elements for robot hands,” *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 124–131, 2018.
- [17] Tekscan, “Pressure mapping sensor 4256e.” [Online]. Available: <https://www.tekscan.com/products-solutions/pressure-mapping-sensors/4256e>
- [18] P. Maiolino, M. Maggiali, G. Cannata, G. Metta, and L. Natale, “A flexible and robust large scale capacitive tactile system for robots,” *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3910–3917, 2013.
- [19] T. Taunyazov, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. Tee, and H. Soh, “Event-driven visual-tactile sensing and learning for robots,” in *Proceedings of Robotics: Science and Systems*, July 2020.
- [20] T. Taunyazov, L. S. Song, E. Lim, H. H. See, D. Lee, B. C. Tee, and H. Soh, “Extended tactile perception: Vibration sensing through tools and grasped objects,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1755–1762.
- [21] M. J. Yang, K. Park, and J. Kim, “A large area robotic skin with sparsely embedded microphones for human-robot tactile communication,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 3248–3254.
- [22] M. J. Yang, J. Cho, H. Chung, K. Park, and J. Kim, “Touch classification on robotic skin using multimodal tactile sensing modules,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9917–9923.
- [23] K. Belsare, A. C. Rodriguez, P. G. Sánchez, J. Hierro, T. Kolcon, R. Lange, I. Lütkebohle, A. Malki, J. M. Losa, F. Melendez, M. M. Rodriguez, A. Nordmann, J. Staschulat, , and J. von Mendel, *Micro-ROS*. Springer, 2023, pp. 3–55.
- [24] J. Scheibert, S. Leurent, A. Prevost, and G. Debrégeas, “The role of fingerprints in the coding of tactile information probed with a biomimetic sensor,” *Science*, vol. 323, no. 5920, pp. 1503–1506, 2009. [Online]. Available: <https://www.science.org/doi/abs/10.1126/science.1166467>
- [25] P. Safari, M. India, and J. Hernando, “Self-attention encoding and pooling for speaker recognition,” 2020.
- [26] *Network Force/Torque Sensor System*, ATI Industrial Automation, 2023.