NIST Pilot Too Close for Too Long (TC4TL) Challenge Evaluation Plan

July 1, 2020
v1.3

1 Introduction

One of the keys to managing the current (and future) epidemic(s) is notifying people of possible virus exposure so they can isolate and seek treatment to limit further spread of the disease. While manual contact tracing is effective for notifying those who may have been exposed, it is believed that automated exposure notification will be a necessary addition as societies open up. Current approaches to automated exposure notification rely on using Bluetooth Low Energy (BLE) signals (or chirps) from smartphones to detect if a person has been too close for too long (TC4TL) to an infected individual. However, the received signal strength indicator (RSSI[1]) value of Bluetooth chirps sent between phones is a very noisy estimator of the actual distance between the phones and can be dramatically affected in real-world conditions by i) where the phones are carried, ii) body positions, ii) physical barriers, and iv) multi-path environments, to mention a few. To better characterize the effectiveness of range and time estimation using the BLE signal, many research organizations around the world are collecting Bluetooth chirp data as well as other phone sensor data (e.g., accelerometer and gyroscope) between various types of phones with simulated real-world variability. The best hope for a solution to this difficult and important problem is to leverage the worldwide research community with common tasks, data, and success metrics that allow for the exchange of and building on collective ideas and approaches.

The National Institute of Standards and Technology (NIST), in coordination with the MIT Private Automated Contact Tracing (PACT) consortium[2], is organizing a TC4TL detector evaluation to facilitate this research effort. The TC4TL challenge presents a new task and domain in the set of technology evaluations conducted by NIST since the 1980s. The evaluation serves the following objectives:

- to provide a common test bed that enables the research community to explore promising new ideas in TC4TL detection using BLE signals,
- to support the community in their development of advanced technology incorporating these ideas,
- to effectively measure system-calibrated performance of the state-of-the-art TC4TL detectors.

The evaluation is intended to be of interest to all researchers in the machine learning community interested in the TC4TL detection problem using BLE signals. To this end the evaluation is designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

TC4TL will be organized in a manner similar to several other NIST evaluations, where NIST will define the evaluation rules and scoring metric, provide development data with ground truth, and then host a leaderboard style evaluation platform where researchers download evaluation data and upload system outputs to be scored and posted.

[1] The RSSI is the received signal strength rounded to the nearest integer. The unit for RSSI is dBm.

[2] `https://pact.mit.edu/`

The data used for the TC4TL evaluation will be derived from data sets being collected by organizations around the world studying this problem. The conditions and available metadata in these data sets may vary and NIST will work to normalize data into a consistent format across data sets. Because this is a new area of data collection, limited amounts of training data will be provided by NIST but we encourage participants to identify and share sources of training data as well as novel ways to create training data. Some of the identified factors affecting distance estimation from RSSI values are (1) the number and time spread of observed chirps, (2) the carriage position of the phones (i.e., hand, front pocket, back pocket, etc), (3) bodies and barriers between phones, and (4) multi-path signals from surfaces (e.g., indoor vs outdoor). Other factors may arise as the phenomenology is further studied. NIST will design evaluation trials to encompass these factors to the extent they are part of data sets used.

Participation in TC4TL challenge is open to all who find the evaluation of interest and are able to comply with the evaluation rules set forth in this plan. There is no cost to participate, but participating teams should be willing to openly share their approaches and training data in the spirit of scientific discovery. We will provide a Google Group forum where participants can hold discussions and we will ask participants to submit a report describing their systems. We also plan to host a virtual meeting at the end of the initial evaluation period, where results can be discussed and sites can provide presentations on their work. Information about evaluation registration can be found on the TC4TL challenge website[3].

Before concluding this section, it is worth making a few remarks regarding the known limitations of this challenge. Note that the TC4TL evaluations are being developed based on newly collected data sets to help coalesce the research community on common tasks, data, and metrics. It is known that initial data sets will have limitations of size and available real-world conditions, but we will design the evaluation to produce meaningful results and grow our collective knowledge. Due to the evolving nature of understanding Bluetooth phenomenology and data collections, we will also be evolving the TC4TL evaluation as necessary.

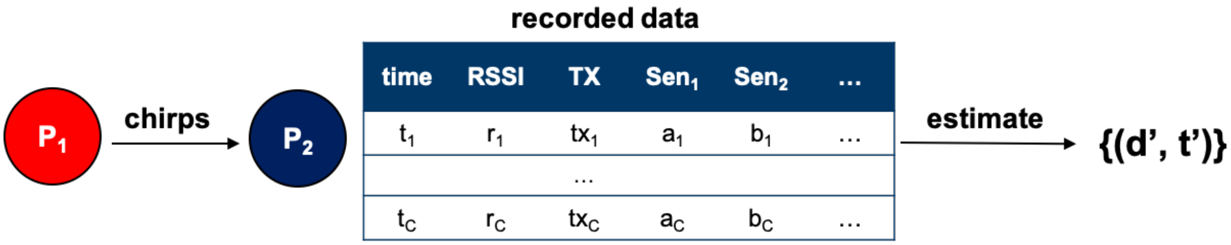
2 Task Description

2.1 Task Definition

The basic task in the NIST TC4TL Challenge is estimating the distance and time between two phones given a series of RSSI values along with other phone sensor data. These distance and time estimates will then be used to solve a two-class hypothesis testing problem, with the null and alternative hypotheses, H_0 and H_1, defined as

$$\begin{cases} H_0: & \text{a TC4TL event} \\ H_1: & \text{not a TC4TL event} \end{cases} \tag{1}$$

A TC4TL event is defined using two parameters, namely distance (D) and time (T).



time	RSSI	TX	Sen_1	Sen_2	…
t_1	r_1	tx_1	a_1	b_1	…
		…			
t_C	r_C	tx_C	a_C	b_C	…

Figure 1: Block diagram demonstrating the basic task in the TC4TL Challenge.

More specifically, as depicted in Figure 1, suppose that phone P_1 (transmitter or beacon) transmits a sequence of BLE chirps for a period of time C, while phone P_2 (receiver or tester) records a sequence of RSSI

[3] `https://tc4tlchallenge.nist.gov/`

values and their corresponding transmit (TX) power[4] from P_1 chirps, along with a sequence values recorded from its inertial measurement units (IMU) such as the accelerometer and gyroscope (denoted as Sen_1, Sen_2, ... in the figure). The task is then to automatically estimate a set of distance and time measurements (i.e., $\{(d',t')\}$) for the contact between P_1 and P_2.

Reference TC4TL labels are generated using true distance and time[5] from contact events. Hypothesized labels are compared to reference labels and probability of false negative (P_{miss}) and probability of false positive (P_{fa}) are calculated. A normalized decision cost function (nDCF) combines these two errors into a single value using weights reflecting the relative cost of each type of error.

2.2 Test Conditions

The evaluation will consist of a set of contact events comprised of series of chirp RSSI values and other sensor data recorded between two phones at varying unknown distances and angular positions. These contact events will be designed to cover some of the known factors affecting distance estimation from RSSI values such as the number and time spread of observed chirps as well as the phone carriage state (i.e., hand or pocket).

More specifically, provided a log of recorded RSSI and phone sensor values for each receiver-transmitter (RX-TX) carriage state configuration (i.e., hand-hand, pocket-hand, pocket-pocket) at a given range (e.g., 1.8 meters), test samples will be extracted using 4-second long windows (aka looks) with variable increments (or skip-sizes) ranging from 10 seconds to 150 seconds. For each test sample, the RX-TX carriage configuration will be either known, or unknown.

Additionally, the test samples will be divided into two subsets, namely fine-grain and coarse-grain distance sets. In both subsets, the distance between phones, d, can vary within a single contact event file. In the fine-grain distance and course-grain distance subsets, d can vary within a single contact event file by up to 0.9 m and 2.1 m, respectively. In the fine-grain subset, variable values of D (see section 3) will be used to define contact events and compute performance (i.e., multiple TC events will be evaluated by using different values of $D \in \{1.2\text{ m}, 1.8\text{ m}, 3.0\text{ m}\}$), while in the coarse-grain subset performance will only be computed for a single $D = 1.8$ m (i.e., a single TC event is evaluated).

3 Performance Measurement

For the initial data set available, we will not be assessing the time component of the TC4TL detector (i.e., we will set $T = 0$), and we will be focusing only on the distance component (D). Note that the definition of a TC4TL event type (D in this case) is a parameter of the metric and will be varied to explore the sensitivity of the estimators to potential changes in TC4TL definitions by Public Health Authorities.

For a contact event, the system will output a single, real-valued distance estimate d' between the two phones. Estimated distances may be quantized to any desired precision (suggestion is for 0.3 meter increments). The estimated distance is converted to a hypothesized event type as follows

$$\begin{aligned} d' \leq D &: \text{hypothesized = TC4TL event} \\ d' > D &: \text{hypothesized = not-TC4TL event} \end{aligned} \tag{2}$$

The true distance (d) for the contact event is also converted to a reference event type as follows

$$\begin{aligned} d \leq D &: \text{reference = TC4TL event} \\ d > D &: \text{reference = not-TC4TL event} \end{aligned} \tag{3}$$

[4] The transmit power or measured power is a pre-calibrated constant indicating the expected RSSI at a distance of 1 meter for iBeacon and 0 meter for Eddystone from the beacon. For this challenge, the path loss (or attenuation) can be computed using the following formula: $P_L = P_t - 41 - P_r$, where P_L, P_t, and P_r, denote the path loss, the transmit power (TXPower), and the received RSSI, respectively.

[5] Note that for this pilot challenge only the too-close (TC) aspect of contact events will be considered. We may incorporate the too-long (TL) aspect in future evaluations.

The reference and hypothesized event types are compared over a set of contact events and the probability of miss (P_{miss}) and probability of false alarm (P_{fa}) values are calculated

$$\begin{aligned} P_{miss} &= \frac{\text{\# (ref = TC4TL and hyp = not-TC4TL)}}{\text{\# ref = TC4TL}} \\ P_{fa} &= \frac{\text{\# (ref = not-TC4TL and hyp = TC4TL)}}{\text{\# ref = not-TC4TL}} \end{aligned} \tag{4}$$

These probabilities will be combined using a normalized decision cost function

$$\text{nDCF} = \frac{w_{miss} P_{miss} + w_{fa} P_{fa}}{\min(w_{miss}, w_{fa})} \tag{5}$$

where w_{miss} and w_{fa} are costs associated with missed and spurious detections, respectively. For this challenge, the weights will be set to $w_{miss} = 1$ and $w_{fa} = 1$.

4 Data Description

4.1 Data Organization

The development and test sets follow a similar directory structure:
<base_directory>/

README.txt
data/
docs/

The data/ directory will contain a set of *test* contact event files with names [a-z]_tc4tl20.csv, where [a-z] is a random 8-character alphabetical string (all lower case). These files will be in ASCII format containing comma separated values (CSV) with the following fields and values

– TXDevice,<device_type>

– TXPower,<txpower_value>

– RXDevice,<device_type>

– TXCarry,<TX_device_carriage_state>

– RXCarry,<RX_device_carriage_state>

– RXPose,<RX_user_pose>

– TXPose,<TX_user_pose>

– <timestamp>,Bluetooth, <RSSI_value>

– <timestamp>,Accelerometer, <a_x >,<a_y >,<a_z >

– <timestamp>,Gyroscope, <g_x >,<g_y >,<g_z >

– <timestamp>,Attitude, <pitch >,<roll >,<yaw >

– <timestamp>,Gravity, <x >,<y >,<z >

– <timestamp>,Magnetic-field, <x >,<y >,<z >,<accuracy∈{uncalibrated, low, medium, high}>

– <timestamp>,Altitude, <x >,<y >

- <timestamp>,Activity, <start_time >,<activity_type >,<confidence_level$\in \{0, 1, 2\}$ >
- <timestamp>,Heading, <true_heading >,<magnetic_heading >,<heading_accuracy>,<x>,<y>,<z>

For example

```
TXDevice,phone1
TXPower,12
RXDevice,phone2
TXCarry,pocket
RXCarry,hand
RXPose,standing
TXPose,sitting
0.000,Bluetooth,-58
0.001,Bluetooth,-67
0.034,Accelerometer,0.0939483642578125,-0.82562255859375,-0.7227325439453125
0.035,Gyroscope,0.7683576345443726,-0.2502034604549408,0.1180611252784729
0.036,Attitude,-0.47507938022821944,-1.5496823294006532,-2.8489995770785375
0.037,Gravity,-0.8890581727027893,0.4574090242385864,-0.01877436228096485
0.037,Magnetic-field,-40.492794036865234,27.20453643798828,20.183868408203125,high
0.035,Heading,28.372474670410156,42.55014419555664,15.139200210571289,...
0.110,Activity,412815.77871700004,walking,2
```

The timestamps denote the recording time in seconds. Note that the contact event files will have varying number of Bluetooth chirps and sensor values. Detailed descriptions of the various motion sensors (e.g., accelerometer and gyroscope) and the information derived based off these sensors (e.g., activity) are beyond the scope of this evaluation plan. We refer the interested reader to the Apple[6] developer documentation for Core Motion[7] and Core Location[8] services.

Any associated metadata for the event files, including phone carriage state, step size (in seconds) for the 4 s windows, and the granularity of distance subsets (i.e., coarse-grain vs fine-grain), will be located under the `docs/` directory.

4.2 Development Set

A small development set of contact events along with their corresponding ground truths and metadata will be provided for system training and hyperparameter tuning. The metadata will include phone carriage states and positions (e.g., in hand or pocket, sitting or standing) during recording, as well as the step-sizes used for extracting the 4 s windows.

4.3 Training Set

Participants may use any data, except for the *MIT Matrix Data* available on the PACT data repository site[9], for system training and hyper-parameter tuning purposes. Some example data sets are available publicly from the PACT data repository site. We ask that participants, to the extent possible, share their training data via the PACT data repository. All personally identifiable information (PII) should be removed or anonymized. No PII is available in NIST data.

[6] See Disclaimer in Section 8.

[7] https://developer.apple.com/documentation/coremotion

[8] https://developer.apple.com/documentation/corelocation

[9] https://mitll.github.io/PACT/datasets.html

5 Evaluation Rules and Requirements

The TC4TL challenge is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to NIST for scoring. As such, the participants have agreed to process the data in accordance with the following rules to maintain the scientific integrity of the evaluation:

- Each test contact event should be processed independently to produce a distance estimate. Processing the entire test set as a whole or using any portion of the test set to train/tune a system is not allowed.
- Participants should not use the *MIT Matrix Data* or the *test* set designated by NIST, for system training, development, and hyper-parameter tuning. Datasets available from the PACT data repository site[9] (except for the *MIT Matrix Data*) may be used for system training and development purposes.

In addition to the above data processing rules, participants agree to comply with the following general requirements:

- While participants may report their own results, participants may not make advertising claims about their standing in the evaluation, regardless of rank, or winning the evaluation, or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113) shall be respected[10]: *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
- At the conclusion of the evaluation NIST generates a report summarizing the system results for conditions of interest, and these results/charts will contain the names of the participating teams involved with their consent. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.
- The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

6 Evaluation Protocol

To facilitate information exchange between the participants and NIST, all evaluation activities are conducted over a web-interface.

6.1 Evaluation Account

Participants must sign up for an evaluation account where they can perform various activities such as registering for the evaluation, signing the data license agreement, as well as uploading the submission and system description. To sign up for an evaluation account, go to `https://tc4tlchallenge.nist.gov`. The password must be at least 12 characters long and must contain a mix of upper and lowercase letters, numbers, and symbols. After the evaluation account is confirmed, the participant is asked to join a site or create one if it does not exist. The participant is also asked to associate his site to a team or create one if it does not exist. This allows multiple members with their individual accounts to perform activities on behalf of their site and/or team (e.g., make a submission) in addition to performing their own activities (e.g., requesting workshop invitation letter).

[10] See `http://www.ecfr.gov/cgi-bin/ECFR?page=browse`

- A participant is defined as a member or representative of a site who takes part in the evaluation (e.g., John Doe)
- A site is defined as a single organization (e.g., NIST)
- A team is defined as a group of organizations collaborating on a task (e.g., Team1 consisting of NIST and MIT)

6.2 Evaluation Registration

One participant from a site must formally register his or her site to participate in the evaluation by agreeing to the terms of participation. For more information about the terms of participation, see Section 5.

6.3 Submission Requirements

Each team must make at least one valid submission for the challenge, processing all *test* event files. Submissions with missing *test* event files, or event files appearing in orders different than that of the trial list, will not pass the validation step, and hence will be rejected.

6.3.1 System Output Format

The system output is a text file (with a `.tsv` extension) composed of a header and a set of records where each record contains an event file and a distance estimate[11] output by the system for the file. The order of the event files in the system output file must follow the same order as the trial list provided under `docs/` directory (`tc4tl_test_trials.tsv`). Each record is a single line containing 2 fields separated by the tab (\t) character in the following format:

```
fileid<TAB>distance<NEWLINE>
```

where
`fileid` - The event log file identifier
`distance` - The estimated distance in meters as a floating point value

For example:

```
fileid distance
sakglcba_tc4tl20.csv 2.4
usoglecr_tc4tl20.csv 1.5
cxxvgmom_tc4tl20.csv 3.0
```

6.3.2 Leaderboard Submission

The participants may make multiple challenge submissions (up to 5 per day). A leaderboard will be maintained by NIST indicating the best submission performance results thus far received and processed. The submission file must be a compressed `.zip` or `.tgz` file, with the system output file described in the previous section as the only content (i.e., no directories or sub-directories). In order to mitigate the possibility of overtuning/overfitting to the *test* set, the leaderboard will display the results computed on only a subset of the *test* event files. The official final results will be computed on the full *test* and announced at the end of the challenge.

[11] As noted previously, it is suggested that estimated distances be quantized to a precision level of 0.3 meter increments (e.g., 0.3 m, 0.9 m, 1.2 m, etc.)

6.3.3 System Description

To allow sites to learn from each other and increase the collective scientific knowledge, we ask participants to produce and share clear system descriptions, covering training, tuning and inference, so other researchers could reasonably reproduce their work. System descriptions can be uploaded and shared via the challenge web platform (`https://tc4tlchallenge.nist.gov`).

In order for NIST to receive comparable and informative system descriptions, the following information is recommended to be included:

- Abstract
- Notable highlights (novel aspects)
- Data resources (training and development datasets)
- Algorithmic description
- Experimental results
- Hardware description and timing report (CPU/GPU resources and run-times, memory footage)

The system descriptions should follow the latest IEEE conference proceeding template available at: `https://www.ieee.org/conferences/publishing/templates.html`.

We will also be setting up a TC4TL Challenge Google Group[12] to provide a place to share ideas, find collaborators, as well as facilitate discussions among participants. Additionally we will plan for a virtual meeting to cover results of the evaluation and discuss future plans.

7 Schedule

Milestone	Date
Evaluation plan published	June 19 2020
Registration starts	June 22 2020
Evaluation data available to participants	July 1 2020
Leaderboard open for submissions	July 2020
Leaderboard closed for submissions	July 2020
Post evaluation virtual workshop	July 2020

8 Disclaimer

Certain commercial equipment, instruments, software, services, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software, services, or materials are necessarily the best available for the purpose.

[12] See Disclaimer in Section 8.