TAME Pain: Trustworthy AssessMEnt of Pain from Speech and Audio for the Empowerment of Patients

Tu-Quyen Dao[1], (PIs) Arya Farahi[1], Eike Schneiders[2], Jennifer Williams[3]
(Collaborators) John Robert Bautista, Rohan Chandra, Liz Dowthwaite, Anna-Maria Piskopani, Tina Seabrooke, and Ganesh Vigneswaran
[1]University of Texas at Austin, [2]University of Nottingham, [3]University of Southampton

GOOD SYSTEMS Ethical AI at UT Austin

UKRI Trustworthy Autonomous Systems Hub

University of Southampton

The University of Nottingham

TEXAS The University of Texas at Austin

Introduction

Traditional methods of evaluating pain in healthcare settings have largely depended on patients' self-reported experiences [1]. This subjective approach to pain assessment is critical for determining accurate diagnoses and treatment plans. Yet, the reliability of this method comes into question for patients with communication challenges. Individuals such as young children, those confronting language barriers, or persons affected by conditions like stroke, learning disabilities, or autism may struggle to express their pain accurately. This communication gap can lead to inadequate care from healthcare professionals. Consequently, there is a pressing need for the development of an objective and non-invasive technique to assess pain levels among these vulnerable groups. Such an innovation would be instrumental in providing them with the precise and empathetic medical attention they require.

Research Goal

We aim to develop and evaluate a proof-of-concept trustworthy algorithm capable of distinguishing between different levels of pain based on bioacoustic markers. Our goal is to therefore empower patients with a communication barrier through an objective pain-detecting method that will ultimately minimize potential medical harm [2].

Methods

This two-phase study design was initiated with a subject study design to collect data for the development of the system, followed by a second part that will assess the level of trust that medical professionals have in the system.

1. Data collection: Participants read aloud provided statements on a computer screen in two separate conditions, a) no-pain condition in which their hand was submerged in warm water, and b) induced pain condition where their hand was submerged in cold water (CPT). They were also prompted to rank their perceived pain level using the Wong-Baker Faces Pain Rating Scale (1-10) [3]. The duration of their participation was audio recorded, and the set-up is shown in Figure 1.

2. Modeling: We will use the self reported pain levels as ground truths for comparison against the model's predicted levels of pain. We will then evaluate medial professionals' trust in the algorithm's predictions by presenting them with the audio recordings of participants experiencing pain coupled with the algorithm's classification.

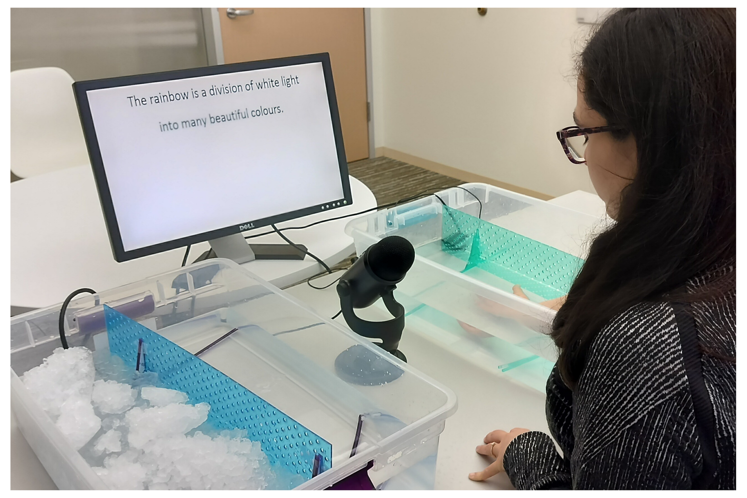

Figure 1: *Set up of data collection process, with the pain induced (cold water) condition on the left, no-pain condition (warm water) on the right, microphone in the middle, and computer screen prompting participants' speech.*

Figures and Results

Data Collection:

Country	Date Completed	Number of Participants	Average Age of Participants	Number of Females	Number of Males	Number of Non-binary	Number of Collected Audio Samples
U.K.	October 2023	15	18.73	12	3	0	1,947
U.S.	February 2024	51	21.33	26	22	3	~7,000

Table 1: *Details of data collection results for the U.K. and the U.S.*

Since the completion of data collection in the U.S. is relatively recent, it has yet to be processed for analysis. The following results are from the U.K.'s data.

Results:

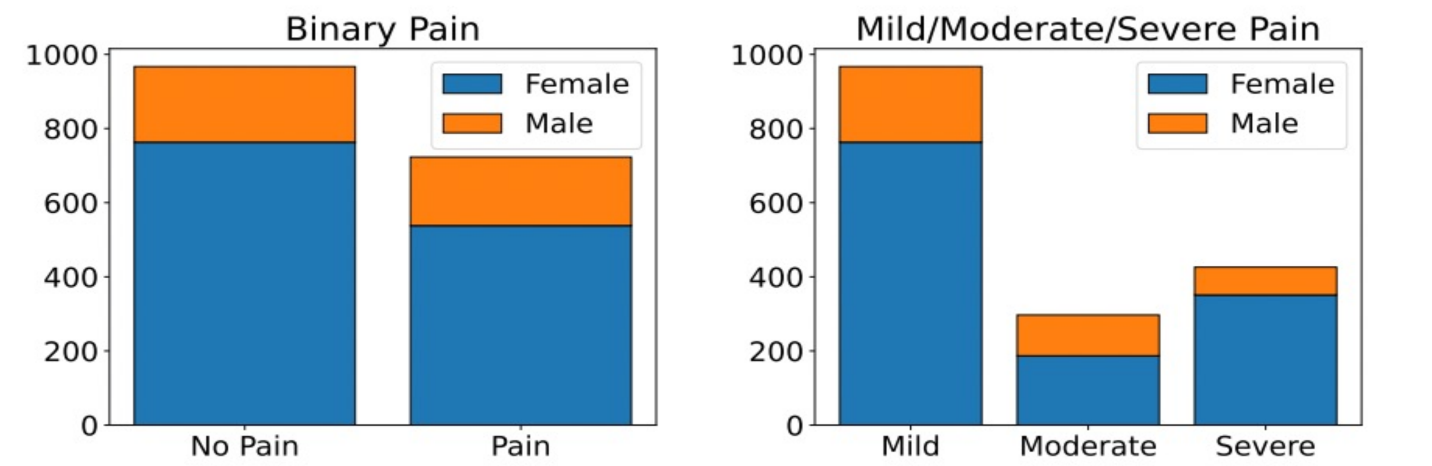

Figure 2: *Distribution of utterances labeled for pain in the binary and three-class discrimination tasks, indicating distribution of male and female utterances*

For binary pain labels, reports of 1-3 was considered as *no pain* and 4-10 as *pain.* For the three-class, 1-3 was treated as *mild*, 4-6 as *moderate*, and 7-10 as *severe* [4]. These conditions are represented in Figure 2.

	Binary		Three-Class	
	Female	Mixed	Female	Mixed
SVM				
- CV	80.0 ± 3.8	80.3 ± 1.7	69.9 ± 4.3	70.9 ± 4.6
- Test	78.0	73.8	70.6	67.3
Logit				
- CV	71.8 ± 3.4	73.8 ± 5.5	66.7 ± 3.5	66.4 ± 2.8
- Test	72.1	70.6	69.6	62.9
MLP				
- CV	79.3 ± 2.5	78.9 ± 3.0	68.3 ± 2.7	68.3 ± 4.9
- Test	78.3	72.8	70.1	68.4
Random	50.0	50.0	33.3	33.3

Table 2: *Classifier performance on two tasks (binary and three-class), reporting normalized F1 scores, for utterances that were female-only as well as mixed male and female. Dark purple is best performance, light purple is second-best performance.*

Several machine learning algorithms (SVM, Logit, MLP) were explored with hyper-parameter tuning, and the performances of the systems are represented by F1 scores in Table 1. All algorithms were able formulate a prediction for both tasks that is consistently greater in accuracy than the baseline Random prediction, with the SVM model performing the best, followed by MLP. However, a separate train/test split uncovered that these models don't generalize well to unseen speakers.

Ethics

Automatic pain discrimination has the potential to be misused outside of the medical domain (e.g., using pain levels for torture or abuse) while it can also provide decision support tools to legitimate health professionals. Our data collection and research activities aligned with a Responsible Research and Innovation (RRI) approach called the AREA (Anticipate, Reflect, Engage, Act) framework to guide our ethics application and to determine whether this work should be undertaken. We followed the ethical guidelines of our institution precisely and we adhere to all applicable regulations regarding data sharing.

The U.S. received IRB approval from the University of Texas at Austin, adhering to the ethical approaches described above (IRB number: STUDY00004954).

Conclusion

The U.K.'s data collection from 15 participants yielded an English speech dataset in which we were able to identify key acoustic features for pain detection and show that our SVM and MLP classifiers performed well on both binary and three-class tasks. However, the small sample size and gender imbalance of participants in the U.K. places limitations on these results.

To address these limitations, the data collection process in the U.S. encompassed a larger dataset from more participants (n=51), featuring a closer gender balance. This will allow for an increased validity and deeper investigation into bioacoustic markers' association with pain detection.

Furthermore, we are interested in developing a more specific pain identification technique that can classify pain levels beyond a three-class task. This will direct us to our ultimate goal of developing a low-cost tool for medical use that can accurately assess pain through speech.

Acknowledgments

This work was supported by the UK EPSRC through the Trustworthy Autonomous Systems (TAS) Hub (EP/V00784X/1), and also by Good Systems, a research grand challenge at the University of Texas at Austin. The authors would like to thank Prof. Gopal Ramchurn, Prof. Joel Fischer, Prof. Sam Baker, and Prof. Sharon Strover, for their helpful feedback and support during this work.

References

[1] Jang, Jin Hee, Won Hee Park, Hyo-In Kim, and Sung Ok Chang. "Ways of reasoning used by nurses in postoperative pain assessment." *Pain Management Nursing* 21, no. 4 (2020): 379-385.
[2] Schneiders, Eike, Jennifer Williams, Arya Farahi, Tina Seabrooke, Ganesh Vigneswaran, John Robert Bautista, Liz Dowthwaite, and Anna-Maria Piskopani. "TAME Pain: Trustworthy AssessMEnt of Pain from Speech and Audio for the Empowerment of Patients." In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, pp. 1-4. 2023.
[3] Wong, Donna L., and Connie M. Baker. "Smiling face as anchor for pain intensity scales." *Pain* 89, no. 2 (2001): 295-297.
[4] Miró, Jordi, Rocío de la Vega, Ester Solé, Mélanie Racine, Mark P. Jensen, Santiago Gálan, and Joyce M. Engel. "Defining mild, moderate, and severe pain in young people with physical disabilities." *Disability and rehabilitation* 39, no. 11 (2017): 1131-1135.