scientific data

Check for updates

OPEN

DATA DESCRIPTOR

TAME Pain data release: using audio signals to characterize pain

Tu-Quyen Dao[1] ✉, Eike Schneiders[2,3], Jennifer Williams[3], John Robert Bautista[4], Tina Seabrooke[5], Ganesh Vigneswaran[6], Rishik Kolpekwar[7], Ritwik Vashistha[8] & Arya Farahi[8,9] ✉

Accurately assessing pain through speech remains a challenge in medical practice, with profound implications for patient care and patient health outcomes. The TAME Pain dataset addresses this challenge by providing a comprehensive dataset that captures the relationship between induced acute pain and speech in adults. Utilizing the Cold Pressor Task (CPT) method to induce pain, we recorded over 7,000 utterances from 51 participants, correlating their self-reported pain levels with vocal cues. This dataset stands as the largest of its kind to date and includes comprehensive annotations detailing background noise, speech errors, and non-speech vocal features, maximizing its utility for in-depth audio analysis. Our dataset is designed to support the development of reliable, non-invasive pain assessment technologies, particularly in telemedicine and remote healthcare settings. By releasing these data, we aim to facilitate interdisciplinary research in psychology, medical science, and AI, fostering innovations that can enhance pain management practices and improve patient outcomes across diverse clinical environments.

Background & Summary

Pain is often referred to as the "fifth vital sign," yet it remains the only vital sign assessed primarily through subjective patient reports[1]. Accurate characterization of pain's intensity, nature, and location is crucial for diagnostic precision and evaluating therapeutic outcomes across various clinical settings[2]. For instance, distinguishing between the characteristics of pain can be essential in differentiating conditions such as myocardial infarction, heartburn, or aortic dissection[3].

Accurately characterizing pain's intensity, while challenging[4,5], is a fundamental aspect of effective diagnosis and treatment in medical practice and is essential for enhancing healthcare outcomes[6–8]. Despite its importance, understanding the perception of human pain and its expression through speech and vocal cues remains underexplored, particularly within psychology and medical science. Traditional pain assessment methods, which heavily rely on self-reported measures, face significant limitations due to the subjective nature of pain and the influence of various factors, such as individual pain thresholds, cultural differences, and communication abilities[9]. Additionally, reliance solely on patient self-reporting presents challenges, particularly in situations where individuals are unable to communicate their pain effectively. This includes emergency cases involving non-verbal patients, individuals with developmental or neurological impairments, infants lacking developed communication skills, those who do not speak the same language as the medical professional, or those with medical conditions such as learning difficulties, autism, and stroke[10–12]. These limitations often lead to discrepancies in pain management and patient care outcomes[13,14].

Speech production is a physiological process that relies on the coordinated function of neural and muscular systems. Pain, which disrupts normal neural and muscular function, can potentially interfere with the transmission of speech signals by affecting the same physiological pathways involved in speech production[15]. Understanding the vocal expressions of pain (including non-linguistic signals) in adults can provide new

[1]University of Texas at Austin, Department of Molecular Biosciences, Austin, 78712, US. [2]University of Nottingham, School of Computer Science, Nottingham, NG8 1BB, UK. [3]University of Southampton, School of Electronics and Computer Science, Southampton, SO17 1BJ, UK. [4]University of Missouri-Columbia, Sinclair School of Nursing, Columbia, 65211, US. [5]University of Southampton, School of Psychology, Southampton, SO17 1PS, UK. [6]University of Southampton, Faculty of Medicine, Southampton, SO16 6YD, UK. [7]Round Rock High School, Round Rock, 78681, US. [8]University of Texas at Austin, Department of Statistics and Data Sciences, Austin, 78712, US. [9]The NSF-Simons AI Institute for Cosmic Origins, University of Texas at Austin, Austin, TX, 78712, USA. ✉e-mail: tuquyendao@austin.utexas.edu; arya.farahi@austin.utexas.edu

insights into pain perception and management, especially in remote consultations. This is particularly important in telemedicine[16,17], which has seen increased adoption due to the COVID-19 pandemic[18] but has continued post-pandemic. The increasing use of video and telephone consultations in healthcare necessitates the development of reliable, vocal pain assessment technologies that can be effectively utilized in remote settings. By leveraging these technologies to analyze vocal indicators, there is a potential to enhance the accuracy of pain assessment, regardless of a patient's communication abilities, resulting in optimized treatment and improved patient outcomes. Thus, the ability to detect pain in conjunction with other physiological measures, such as voice, is becoming increasingly important.

The ability to judge pain from vocal and paralinguistic (i.e., verbal utterances that do not include language) biomarkers would be a particularly useful low-cost and non-invasive tool for remote communications with patients who cannot verbally articulate their pain. Research with healthy subjects in which pain has been experimentally induced suggests that people experiencing intense pain produce more speech than those experiencing low pain[19]. This suggests that vocal utterances may provide a rich dataset to assess pain levels.

Research indicates that vocal features may provide valuable insights into health in general, and pain more specifically. For example, vocal biomarkers of mental health conditions[20] such as depression[21], schizophrenia[22], and post-traumatic stress disorder[23] has been reported. Duey *et al.*[24] prompted 60 patients with spine disease to self-report their pain levels and provide speech recordings. Using the features identified in the speech signal, the researchers developed a machine learning model that predicted the pain level of the patients (high or low) with a 0.71 accuracy and a 0.73 F1 score. While this is a promising start, it should be noted that there were delays of up to 24 hours between the patient pain ratings and speech recordings. To explore the relationship between speech signals and pain in an uninterrupted manner[25], our team conducted a pilot experiment which indicated a potential correlation between non-vocal pain cues and pain. However, the study was limited to 15, predominantly female, participants. These studies show promise in the detection of pain in speech through an objective signal, but the datasets associated with them are not publicly available, limiting further assessment of the datasets.

The scarcity of reliable, publicly available linguistic, acute pain data complicates the identification of vocal pain-associated features and the development of effective pain assessment technologies. This problem is pronounced by the lack of data collected in controlled environments and insufficient detailed annotation[2]. Previous research has investigated objective pain assessment methods, including physiological indicators like heart rate variability and skin conductance[26]. However, these measures can be invasive, inaccurate, or impractical in many clinical settings, such as telemedicine[27]. Behavioral indicators, including facial expressions and crying, have also been shown useful, particularly in non-verbal populations like infants[28,29]. Most data sets created for the automation of pain assessment focus on facial expressions, electrodermal activity, electrocardiogram, or electromyography to assess pain[30–34]. Recent studies also explored multimodal approaches, combining audio and video data for a more comprehensive analysis of pain expressions[35]. Although the use of audio data is becoming more popular in healthcare, high-quality pain datasets collected in highly controlled environments remain scarce.

Healthcare data in the form of "digital fingerprints"[36] are increasingly being used to provide personalized care. Our study aims to contribute to this growing body of work by releasing a novel and unique dataset that captures the relationship between acute pain and speech and vocal cues in adults[37]. Utilizing the Cold Pressor Task (CPT) with water temperatures ranging from 0°C to 4°C, we created a controlled environment to elicit pain responses while recording speech and the participants' self-reported pain levels. This dataset is designed to provide comprehensive insights into the vocal cues of pain. We augment this dataset by annotating every single audio file, including every sentence spoken by the participant, with details such as background and foreground noise, speech errors, and non-speech vocal features. These annotations enable thorough audio analysis, facilitate pain studies, and aid in identifying both speech and non-speech pain cues.

By focusing on the intersection of pain perception and speech, our dataset aims to provide a unique resource for developing more accurate and objective pain assessment technologies. These advancements have the potential to transform pain management practices, ensuring better care for patients across diverse clinical and remote settings. In addition, our work highlights the importance of interdisciplinary approaches that combine psychology, medical science, and technology to address the complex challenge of pain assessment. Our dataset is expected to spur future research into the development of data-driven pain assessment tools, which could revolutionize how pain is diagnosed and treated in both clinical and remote settings.

Methods

The Cold Pressor Task (CPT) was employed as the method to induce pain under controlled laboratory conditions[38]. In the CPT, participants submerge their hand into cold water maintained at a temperature of approximately 0°C to 4°C for a specified duration or until pain tolerance is reached. This method is a common technique in pain research due to its ability to induce pain reliably in a controlled and measurable manner while presenting a low risk of harm to participants. Other techniques used in pain studies include thermal stimulation[39], pressure application[40], and electrical stimulation[41]. For this study, CPT was chosen for its robustness, consistency in pain induction, and minimal risk to participants. The data collection protocol was approved by the University of Texas at Austin's Institutional Review Board (IRB number: STUDY00004954). In this section, we detail our participant selection process, text selection criteria, experimental design, and the procedures used to collect data. Additionally, we outline the steps taken for data cleaning and annotation.

Participants. We recruited 51 participants (26 female, 22 male, and 3 non-binary; 5 Hispanic/Latino, 27 Asian, 1 Black or African American, 14 White, 4 Two or More Races; average age: 21.33, SD: 4.18). To qualify for the study, the participants had to be between 18 and 35 years old, fluent in English, and have health insurance. During the initial screening, we excluded six additional participants who self-reported any of the following medical conditions to reduce the risk[42] of adverse events during the CPT procedure: high blood pressure, heart or

Fig. 1 Experimental set-up for data collection.

circulation problems, dysthymia, cardiovascular disorders, or a history of Raynaud's syndrome, fainting, seizures, or frostbite. Additionally, those with an open cut, sore, or bone fracture on or near either hand, neurological disorders, diabetes, epilepsy, or pregnancy were not eligible[43]. After the eligibility screening phase, the next page of the screening form asked prospective participants to provide their contact email and demographic information, including age, gender, and race/ethnicity. Subsequently, a team member contacted them to schedule an appointment to participate in the study. All participants were recruited using convenience (email lists and flyers), and snowball sampling (word-of-mouth). Participants were given a USD 25 gift card as an incentive.

Prior to data collection, each participant's blood pressure was taken to minimize risk[44]. Following our approved protocol, participants were only allowed to proceed with the study if their blood pressure did not surpass 130 mmHg systolic and 80 mmHg diastolic readings before data collection. All participants were informed about the study aims, procedure, and individual rights, including the right to withdraw from the study or to terminate the experiment without prior notice or reasoning. Verbal informed consent was also obtained before signing a written consent form for the publication of anonymized data and audio recordings. A copy of the consent form was provided to the participants.

Text Selection. Participants were asked to read aloud sentences selected from a randomized list of Harvard sentences[45,46]. The Harvard Sentences are a set of phonetically balanced sentences designed to cover a wide range of English phonemes, ensuring a representative sampling of the language's sounds. This phonetically balanced nature makes them particularly suitable for speech experiments, as they provide consistent and comparable data across different speakers and conditions. The full list of sentences used in this study is provided in the Supplementary Materials file.

Additionally, we incorporated a pain assessment sentence, "On a scale from 1 to 10, the pain I feel right now is —". This regular pain assessment allowed for continuous monitoring of participants' pain levels throughout the experiment, directly linking each batch of Harvard Sentences with a corresponding self-reported pain level. This self-reported pain assessment sentence occurred once for every five Harvard sentences.

Pain Inducement: The Cold Pressor Task. *Experimental Setup.* The lab space, approximately 10 meter^2, was kept consistent for all participants. A custom device was built following[47]. The setup included two plastic containers placed side by side on a desk to minimize movement around the room (see Fig. 1). Each container had a detachable separator reserving the inner side of the container for the participant's hand, while the outer side was reserved for ice (only in cold water conditions) to ensure that participants' skin would not come into direct contact with the ice. The left container, used to induce pain, contained cold water (0-4° C). This pain stimulus, known as a CPT, is a commonly used task for inducing discomfort to mild pain[48,49]. The right container, used as a control condition, contained warm water (34-37° C). By supplementing ice cubes and hot water for the two containers, water temperatures were maintained consistent for all participants. A digital water thermometer was utilized to monitor the temperature (precision ± 0.1° C) in each container and ensure that the water temperature remained within the temperature intervals. To prevent the buildup of microclimate on participants' skin due to lack of water circulation, we ensured water circulation at 5.8 liters per hour using a water pump in each container. Furthermore, water levels remained consistent, at approximately half the container's capacity, allowing participants to fully submerge their hands. Apart from the water temperature, the conditions were identical.

Experimental procedure. Participants were randomly assigned to one of the four groups. Apart from the order in which hands, left (L) or right (R), were placed in cold (C) or warm (W) water, the four groups were identical. Experimental groups were (1) LC-LW-RC-RW, (2) LW-LC-RW-RC, (3) RC-RW-LC-LW, and (4) RW-RC-LW-LC. Participants were asked to keep their hands as still as possible throughout each trial, minimizing external disturbances in the audio recording. Prior to data collection, participants' hand temperatures were taken using an infrared thermometer. This provided a baseline temperature for each hand, allowing the experimenter to bring the participants' hand temperatures back to their personal baseline following the final trial. Next, we equipped the Røde Wireless PRO close-talking lapel mic onto a lanyard to be worn by the participant, at a distance of approximately 10 inches from the participants' chin. This was the primary microphone for the recordings. Additionally, we used the Blue Yeti desktop microphone, which was placed approximately 20 inches from the speaker, and participants read sentences from the monitor, which was placed approximately one meter in front of them. See Fig. 1 for experimental setup.

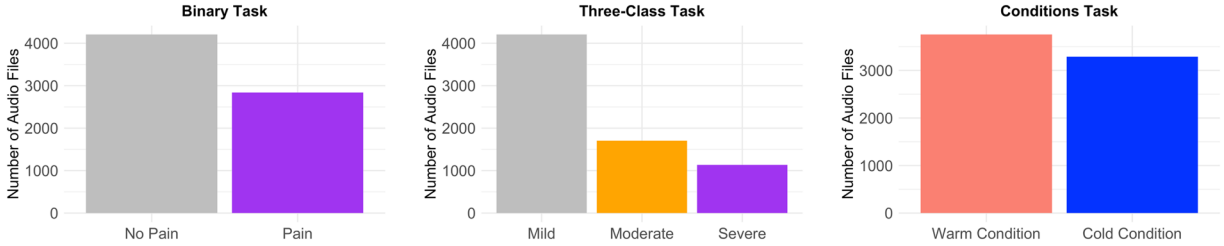

Fig. 2 Left Panel: Distribution of all audio files for the Binary Task. Middle Panel: Distribution of all audio files for the Three-Class Task. Right Panel: Distribution of all audio files for the Conditions Task.

After attaching the primary microphone, we conducted an audio test to verify that the recording device worked as intended. Then, each participant was shown two example sentences. Specifically, from the Harvard sentences, "The rainbow is a division of white light into many beautiful colours" and, for the pain assessment, "On a scale from 1 to 10, the pain I feel right now is —." Once participants had confirmed that they understood and felt comfortable with the task, we proceeded with the actual data collection.

During the experiment, participants were asked to submerge their left or right hand in cold or warm water. The order was dependent on the group. The participants submerged their entire hand in the water, with the palm facing upwards, while performing the reading task. To minimize risk to the participants, and in accordance with the IRB, participants submerged their hand for a maximum duration of up to three minutes—or until they voluntarily withdrew their hand from the water[48]. Following each instance of cold water exposure, participants placed that same hand into warm water to bring their hand up to baseline temperature to minimize discomfort and ensure that each trial for all participants always started at their baseline temperature. We then proceeded with the next condition.

During data collection, each experimental condition started with one batch of six sentences followed by batches of five sentences. Each batch began with a randomized selection of Harvard Sentences, followed by a pain assessment sentence, "On a scale from 1 to 10, the pain I feel right now is —." The order of sentences was randomized for each participant and manually advanced by the research team. Participants continued reading sentences until they either withdrew their hand from the water or reached a three-minute limit. This procedure allowed participants to read at their own pace, leading to variability in the number of sentences read by each participant. Each sentence is saved as one utterance. We collected a total of 7,044 utterances from the 51 participants. The collected audio files are saved as one utterance per *.wav* file, as 16-bit mono PCM 16 kHz. We release the audio of participants' recorded utterances originating from the primary (Røde) lapel mic.

Pain Annotation. The pain statements were used to label the utterances by manually extrapolating the reported pain level backward for the previous unlabeled utterances, allowing us to label every utterance with the subject's self-reported pain level. If a subject reported a pain level of 0, this was re-labeled as 1 since we used a pain scale of 1-10. This adjustment was documented as a revised pain level. If the pain statement was not available for a batch of sentences, we copied forward, rather than backward, from the preceding pain rating, if available. If a preceding pain rating was unavailable, we copied backward from the following pain rating. For cases with no adjacent pain statement for that condition, we removed the utterances that remained unlabeled; this was applicable for five utterances.

Out of 7,044 utterances, five utterances were not labeled with a pain level due to the absence of a pain statement from that participant's condition task. Notes regarding audio disturbances and labeling technicalities were made for these five utterances, but since they don't have an associated pain rating, we excluded them from all figures and tables. Our working dataset contains 7,039 utterances and 311.24 minutes of data (average duration: 2.65 seconds, SD: 0.57 seconds, minimum: 0.33 seconds, maximum: 5.88 seconds).

We adjusted the revised pain levels in our data set to align with three discriminative tasks: presence or absence of pain ("Binary task"), mild, moderate, and severe pain ("Three-Class Task"), and cold/warm conditions ("Condition Task"). For binary pain labels, we labeled 1-3 as *No pain* and 4-10 as *Pain*. For the three-class problem, we treated 1-3 as *Mild*, 4-6 as *Moderate*, and 7-10 as *Severe*[50]. For the condition task, the abbreviations LW and RW were used to indicate *Warm Condition* and LC and RC as *Cold Condition*. Figure 2 shows the number of audio files for each classification.

Speech Data Pre-processing and Annotations. All recordings were trimmed using voice activity detection (VAD) to remove any leading and trailing silence using the Python *webrtcvad* toolkit (https://github.com/wiseman/py-webrtcvad/) with the lowest aggressiveness setting. Then, TD listened to all audio files using Sony WH-1000XM5 Wireless Noise Canceling Headphones for the manual annotation process. 7,039 utterances were manually labeled with a pain level and a revised pain level in the `PAIN LEVEL` and `REVISED PAIN` columns of the `meta_audio.csv`. When an audio file demonstrated the presence of an audible audio feature or labeling technicality, annotations were added to the `NOTES` column of the `meta_audio.csv`. Of 7,044 utterances, 2,869 utterances contained annotations in the `NOTES` column. These audio files with annotations in the `NOTES` column were then compiled into seven distinct annotation classes, organized in folder *Annotations*. These seven classes are (1) an external disturbance was present, (2) a speech error and/or disturbance occurred, (3) the audio

was cut out, (4) an audible breath could be heard, (5) there was no pain rating reported, so we copied from an adjacent pain rating, (6) the assigned sentence was not spoken, or (7) there was no pain rating at all.

The categories described below are organized in separate datasets within folder *Annotations*. The following describe how each annotation category was made. These categories were treated as non-mutually exclusive, i.e., an utterance can be assigned to multiple categories.

1. `External_Disturbances.csv`: (1,852 utterances) Includes any external noise unrelated to a participants' vocalization. The annotations were made by (1) describing the disturbance's intensity, (2) defining the noise type, and (3) describing the location of the disturbance in the audio. (1) The intensity of the disturbance was indicated by an adjective preceding the noise type (e.g., "slight" indicated a low-intensity disturbance, and "loud" indicated a high-intensity disturbance), while no preceding adjective indicated moderate intensity. (2) The noise type was defined by a single word to best phonetically imitate the disturbance heard in the audio (e.g., beep, click, creak, shuffling, slap, static, etc.). (3) This was followed by a description of the disturbance's general location in the audio file (e.g., beginning, middle, end). However, a location descriptor was omitted for instances where the disturbance was persistent through the audio, replaced by noting that the disturbance was "constant" or occurred "throughout". To distinguish the general source of the noise, whether it was related to the participant's movements (e.g., the movement of water heard from the experimental conditions, which was likely caused by the movement of the participant's hand) or unrelated to the participant (e.g., a beeping truck outside). We labeled each row with a "foreground" disturbance (likely related to the speaker), a "background" disturbance (likely unrelated to the speaker), or both a "foreground and background" disturbance (both "foreground" and "background" annotations were made for the same audio file). This distinction was made in accordance to the external disturbances annotations. A disturbance was classified as having a "foreground" relation if the annotation noise type consisted of an "airy", "creak", "scrape", "shuffle", "squeak", "static", "tap", or "water" noise. A disturbance was classified as having a "background" relation if the annotation noise type consisted of a "background voice", "beep", "buzz", "clank", "clash", "clatter", "click", "crinkle", "crunch", "hum", "knock", "ring", "slap", "tink" or "zip" noise. Contains one audio file without a pain rating.
2. `Speech_Errors_and_Disturbances.csv`: (495 utterances) Includes speech errors (mispronunciations, reading errors, stutters, and added/deleted words/phrases), indicated by describing the verbal mistake and the word/phrase subject to the error. This also includes speech disturbances made by the participant (laugh, sniff, throat clearing "um", "uh"), described by identifying the verbal disturbance and its location in the audio. Contains one audio file without a pain rating.
3. `Audio_Cut_Out.csv`: (525 utterances) Includes instances where the audio was cut out, which resulted in loss of parts of the assigned sentence. The annotation was made by first stating the general location or quantity of the audio cut, followed by the phrase/word/letter(s) that was cut out, in parenthesis. Contains one audio file without a pain rating.
4. `Audible_Breath.csv`: (140 utterances) Includes audible inhales/exhales made by a participant, followed by a description of its general location (e.g., beginning, middle, end) in the audio.
5. `No_Pain_Rating_Given_So_Copied.csv`: (503 utterances) Includes audio files without a pain rating as a result of an audio cut at the end of a pain statement or in the case that a condition task doesn't end on a pain statement. For instances where a pain statement from the same condition was available, we extrapolated an adjacent reported pain level forward, instead of backward, for the subsequent utterances without a pain rating. For other instances where the first pain rating of a condition was not available (on 6th audio file), we copied backward from the adjacent reported pain level (11th audio file) and reported the file identification number of the file we copied from.
6. `No Assigned Sentence.csv`: (13 utterances) Includes audio files where the assigned sentence was not spoken at all (e.g., a participant asks to take their hand out). Any extra dialogue was reported if present.
7. `No_Pain_Rating.csv`: (5 utterances) Includes audio files where no pain rating was reported and there was nothing from that condition to copy from (only happened for five audio files identified as `p61395.LC`).

For all audio recordings, there is a uniform background fan noise. This noise is the loudest for participants `p15965`, `p37540`, `p59520`, `p60145`, `p71740`, `p72315`, and `p93975` (loudest). The fan noise is not included in our annotation.

Action Labels. Based on the annotations, an action label with a pain rating was assigned to each utterance. This label can be found in the `ACTION LABEL` column in the `meta_audio.csv` file. This excludes utterances classified as "No Pain Rating". From an integer scale of 0 to 4, this label broadly indicates the quality of an audio file, with 0 being the highest quality and 4 being the lowest quality. We define low-quality audio as containing an audible feature that is likely to confound processed speech data. If multiple annotations were made for a single utterance, the highest action label was assigned to that utterance.

The parameters of each label are defined below.

0. (4,658 utterances) Clean audio files free from disturbances or errors. This label includes audio files that featured: no annotations made in the `NOTES` column, no external disturbances, no speech errors/disturbances, minor added/deleted words ("a", "the", "this", "to", "of"), audible breaths, "No Pain Rating So Copied" annotations, or "Audio Cut Out" annotations featuring $\leq$1 word audio cuts.
1. (408 utterances) Includes audio files that featured no external disturbances, speech errors

	Action Label				
	0	1	2	3	4
External Disturbances	0	0	604	1173	74
Speech Errors+Disturbances	25	294	40	108	27
Audio Cut Out	142	142	134	102	4
Audible Breath	84	7	17	26	6
No Pain Rating Given So Copied	278	65	57	81	22
No Assigned Sentence	0	0	0	0	13

Table 1. Frequency of utterances for action labels observed in each annotation category with a pain label.

		Action Label				
		0	1	2	3	4
Binary Task	No Pain	2,905	246	388	631	32
	Pain	1,753	162	310	544	68
Three-Class Task	Mild	2,905	246	388	631	32
	Moderate	1,067	86	190	326	34
	Severe	686	76	120	218	34
Conditions Task	Warm Condition	2,585	240	345	558	25
	Cold Condition	2,073	168	353	617	75

Table 2. Frequency of utterances for action labels observed in each classification task.

(mispronunciations, reading errors, stutters, all other added/deleted words/phrases), or "Audio Cut Out" annotations featuring > 1 word audio cuts.
2. (698 utterances) Includes audio files that featured low severity external disturbances (indicated in "External Disturbances" annotations with an adjective "slight", "faint", "short", "low", or "small"), or "Audio Cut Out" annotations that indicate the "majority" of the audio was cut out.
3. (1,175 utterances) Includes files that featured moderate severity external disturbances (indicated in "External Disturbances" annotations with no adjective or if a low severity disturbance was annotated to be "throughout" or "constant"), or the word "ouch" was added.
4. (100 utterances) The lowest audio quality with a high potential to confound with speech data processing, this label includes files with high background contamination (indicated in "External Disturbances" annotations with an adjective "loud", "obvious", or "distinct"), files residing in the annotation category "No Assigned Sentence", speech disturbances (laughter, sniffs, throat clearing), or if the previous sentence was read by mistake.

Table 1 shows that higher action labels were designated for annotated utterances that have a higher potential to confound or corrupt the speech data, as seen in "External Disturbances" and "No Assigned Sentence." Table 2 shows that the baseline classifiers ("No Pain" in Binary Task, "Mild" in Three-Class Task, and "Warm Condition" in Conditions Task) have a higher frequency of lower action labels (0, 1, 2) than their task's corresponding pain-stimulated classifiers ("Pain" in Binary Task, "Moderate" and "Severe" in Three-Class Task, and "Cold Condition" in Conditions Task). Moreover, a higher frequency of the highest action label (4) in all pain-stimulated classifiers is observed.

Data Records

The TAME Pain data are available on the PhysioNet data platform[51]. It consists of three files and one folder described in the following subsections: (1) audio recordings, (2) audio metadata file, (3) participant data, and (4) a folder that includes annotations of audio file data.

Audio Recordings. `mic1_trim_v1.zip` contains 51 subfolders, each identified with a participant ID (PID) that corresponds to 51 participants. The contents of each subfolder consist of that participant's audio recordings saved in *.wav* format, after being trimmed using VAD. Each audio file is named according to the `PID.COND.UTTNUM.UTTID.wav` format. The identifying labels are described below.

- `PID` (Participant Identification): Begins with the letter p followed by a randomly generated five-digit number. 51 unique PIDs are used to deidentify the 51 participants. This label corresponds to the title of each folder.
- `COND` (Condition): Has four unique variables: LC (Left Cold), LW (Left Warm), RC (Right Cold), RW (Right Warm). This label corresponds to the different experimental conditions.
- `UTTNUM` (Utterance Number): Utterances are the speech data collected in the audio files, encompassing Harvard Sentences or pain statements. Each audio file consists of a single utterance. This label numbers utterances of each condition in numerical order, starting at 1 for the first utterance of each condition. The first pain

statement occurs at the sixth Utterance Number and reoccurs every fifth Utterance Number thereafter (e.g., 6, 11, 16, 21, etc).
- `UTTID` (Utterance ID): Identifies the assigned sentence of an utterance taken from Harvard Sentences. All the sentences used are found in the Supplementary Materials file, and the Utterance ID can be identified by the sentence's corresponding list number in the Appendix. All the pain statements are assigned an Utterance ID of '99999' for distinction from non-pain statements.

These files are sorted according to the numerical order of `PID` first, alphabetical order of `COND` (i.e., LC, LW, RC, RW) next, and numerical order of `UTTNUM` last.

Audio File and Participant Data. There are two metadata files, `meta_audio.csv` and `meta_participant.csv`, that consist of Audio file metadata and participant data, respectively.

Audio metadata (`meta_audio.csv`). Each row represents a single audio file and is sorted in the same order described for the audio recordings. The first four columns, `PID`, `COND`, `UTTNUM`, and `UTTID` maintain the same definitions as above. The other columns are defined below.

- `PAIN LEVEL`: Raw self-reported pain levels extracted from the audio data.
- `REVISED PAIN`: Self-reported pain levels modified to fit with our scale definition.
- `DURATION`: The length of the audio file in seconds.
- `ACTION LABEL`: A discrete scale, from 0 to 4, that labels the quality of the audio, with 0 being the highest quality and 4 being the lowest quality.
- `NOTES`: Manual annotations made by the authors. Multiple annotations for a single file are separated by semicolons. Annotations are made in the order that they occur in the audio.

Participant Data (`meta_participant.csv`). Each row represents a single participant, sorted in numerical order according to `PID`, present in the first column. The other columns are defined below.

- `GENDER`: Self-reported gender in the screening survey, which included a multiple-choice question with the following options: Man, Woman, Non-Binary, and Prefer to self-describe. While a text box was available for those choosing to self-describe, it was not utilized by any respondents. Participants include 26 female, 22 male, and 3 non-binary.
- `AGE`: Self-reported age from the screening survey (average age: 21.33 years, SD: 4.18 years).
- `RACE/ETHNICITY`: Self-reported race/ethnicity from the screening survey. It was presented as a multiple-choice question with the options including Hispanic/Latino, American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Two or More Races. Participants include 5 Hispanic/Latino, 27 Asian, 1 Black or African American, 14 White, and 4 Two or More Races.
- `FOLDER SIZE`: Digital storage size of all audio files for that participant, in megabytes (average: 11.72 MB, SD: 2.14 MB).
- `NUMBER OF FILES`: Count of audio files for that participant. (average: 138.12 audio files, SD: 27.62 audio files).
- `TOTAL DURATION`: Sum of lengths of all audio files for that participant, in seconds. (average: 365.90 seconds, SD: 67.12 seconds).
- `LC`: Abbreviation for "Left Cold", a "1" indicates that the participant completed the condition (hand was submerged for the maximum duration of three minutes) while a "0" indicates an incomplete condition (withdrew the hand from the water before three minutes or did not attempt condition). Includes 37 completed conditions and 14 incomplete conditions.
- `LW`: Abbreviation for "Left Warm", a "1" indicates that the participant completed the condition (hand was submerged for the maximum duration of three minutes) while a "0" indicates an incomplete condition (withdrew the hand from the water before three minutes or did not attempt condition). Includes 51 completed conditions and zero incomplete conditions.
- `RC`: Abbreviation for "Right Cold", a "1" indicates that the participant completed the condition (hand was submerged for the maximum duration of three minutes) while a "0" indicates an incomplete condition (withdrew the hand from the water before three minutes or did not attempt condition). Includes 42 completed conditions and nine incomplete conditions.
- `RW`: Abbreviation for "Right Warm", a "1" indicates that the participant completed the condition (hand was submerged for the maximum duration of three minutes) while a "0" indicates an incomplete condition (withdrew the hand from the water before three minutes or did not attempt condition). Includes 49 completed conditions and two incomplete conditions.

Annotation of File Data. Folder *Annotations* consists of seven `.csv` data files

1. `External_Disturbances.csv`
2. `Speech_Errors_and_Disturbances.csv`
3. `Audio_Cut_Out.csv`
4. `Audible_Breath.csv`
5. `No_Pain_Rating_So_Copied.csv`
6. `No_Assigned_Sentence.csv`
7. `No_Pain_Rating.csv`

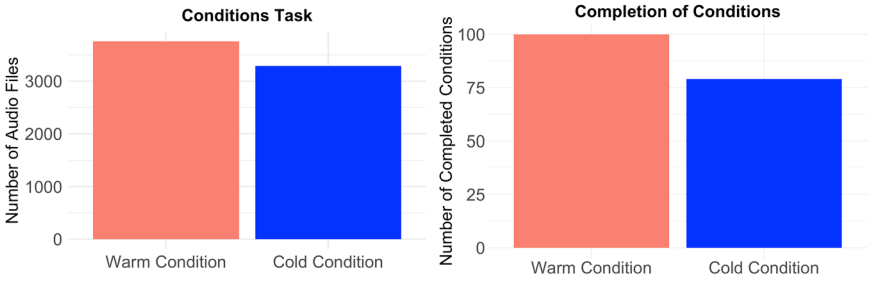

Fig. 3 Left Panel: Frequency of audio files for the Conditions Task. Right Panel: Frequency of completed conditions for all participants.

If multiple annotations were made for a single file, then that file is found across multiple categories pertaining to each annotation made. All data files contain the columns: `PID`, `COND`, `UTTNUM`, `UTTID`, `PAIN LEVEL`, `REVISED PAIN`, and `NOTES`. An `ACTION LABEL` column is also present in every data file except in `No_Pain_Rating.csv` because it contains files without a pain rating. Only `External_Disturbances.csv` contains a `NOISE RELATION` column to distinguish the broad source of the disturbance relative to the speaker.

Technical Validation

The validation process was carried out in two key stages: technical quality and usefulness. This section outlines the experiments and analyses conducted to ensure the technical quality and reliability of the TAME Pain dataset.

Technical Quality. In validating our dataset with the experimental conditions, we observe that this dataset contains more audio files for the warm water condition. Similarly observed in the experimental conditions, there is a higher frequency of incomplete cold water condition tasks, mostly due to voluntary hand-withdrawals prior to the full task duration. Figure 3 compares the conditions of all audio files (left) to the completed experimental conditions of all 51 participants (right). For the right panel of Fig. 3, completed conditions imply that a participant's hand was submerged for the full duration of three minutes. Incomplete conditions, not included in the graph, imply a premature hand withdrawal (14 premature LC hand-withdrawals, 8 premature RC hand-withdrawals) or an unattempted condition (1 unattempted RC, 2 unattempted RW).

The accuracy and comprehensiveness of the labeling was ensured through a collaborative process. The initial set of labels was proposed by JW, the team's audio processing expert, outlining the key features to be annotated. This was followed by a round of discussions among TD, AF, ES, and JW, where the labels were refined, and sample cases were reviewed to ensure a shared understanding and accuracy of the labeling criteria. TD then carried out the primary labeling task, applying the agreed-upon labels across the entire dataset. This was followed by a secondary categorization task, suggested by JW and performed by TD, to broadly define the labels. The results were subsequently presented to the entire research team for feedback. This collaborative review led to further refinements, ensuring that the labels accurately captured the relevant features. Example files for each labeling assignment were shared with the team to maintain consistency. To validate the accuracy of the labeling, RK independently reviewed a random selection of 100 utterances. No objective inconsistencies were found between RK's and TD's labels, confirming the reliability of the labeling process.

Although we attempted to create a highly controlled environment by instructing participants to avoid making noise and conducting the experiment in a relatively quiet room, external disturbances were inevitable. To address this, we developed a labeling system to differentiate data based on audio quality. Table 2 reveals a pattern where higher quality audio (action label 0) is more frequent in baseline conditions and lower quality audio is more prevalent in pain-stimulated conditions. This observation suggests a potential link between more pronounced disturbances in audio and pain detection. Beyond focusing solely on speech data, considering all audio signals-including non-speech elements like co-speech gestures[19] that may translate into audio – could enhance the practical applications of this dataset. This broader approach can be useful particularly in clinical settings, where environmental control is often limited, and understanding the full range of audio signals can contribute to more effective pain assessment tools.

Usefulness. This dataset contains a total of 311.24 minutes of audio data, with individual utterances averaging 2.65 seconds in duration (SD: 0.57 seconds, ranging from 0.33 to 5.88 seconds). Figure 4 illustrates the distribution of utterance durations. Figure 5 shows the average duration and the standard error of the mean for each class. The p-values are computed using a paired t-test. It is observed that the only significant difference is found between the "Moderate" and "Severe" classifiers of the Three-Class Task ($p = 0.011$) when $\alpha = 0.05$.

We demonstrate the utility of our annotations through a stratified analysis of the Binary, Three-Class, and Conditions classifications. By comparing the overall patterns of these classifications (Fig. 2) with those observed within specific subgroups – such as External Disturbances, Speech Errors+Disturbances, and Audible Breath – we observe notable differences in distribution (Fig. 6). The three rightmost graphs in each row of Fig. 6 show a higher frequency of audio files associated with pain-stimulated classifications compared to the reference graphs. This is particularly evident in the Audible Breath category, which suggests a potential correlation between audible inhales/exhales and pain detection. Furthermore, this correlation supports past findings that show pain is a

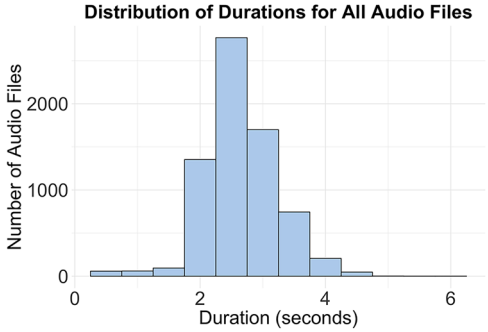

Fig. 4 Histogram of durations for the working dataset of 7,039 audio files.

known simulator for changes in respiratory pattern[52,53]. These findings highlight the importance of considering non-speech vocal cues in the analysis, as they may provide additional insights into the physiology, detection, and assessment of pain.

In Fig. 7, we further stratify the External Disturbances category into Foreground and Background subgroups. As expected, the Background Disturbance distributions (middle column) show no statistically significant deviation from the reference graphs (left column), indicating that background noises, which are unrelated to the speaker, do not significantly differ between baseline and pain-stimulated conditions. In contrast, the Foreground Disturbance distributions (right column) exhibit a notable deviation, with a higher proportion of audio files in the pain-stimulated classifiers compared to the reference graphs. This suggests that foreground disturbances, which are more likely to be related to the speaker's actions (e.g., chair squeaks or other movements), may correlate with pain expression. These findings imply that speaker-induced noises might be an important non-verbal cue in pain detection, highlighting the potential significance of monitoring and analyzing such disturbances in pain assessment studies.

These examples demonstrate that the provided labels are not merely tools for tracking data quality; they also offer potentially useful insights that can be leveraged to enhance our understanding of pain and to develop more accurate pain assessment tools. Further analysis of these labels in various contexts can reveal patterns and correlations that contribute to more sophisticated and effective methods for assessing pain.

Usage Notes

Pain Detection. Previous studies have demonstrated that audio signals, such as cries in infants, can be used to distinguish between pain and discomfort with a high accuracy[54–56]. These findings suggest potential for detecting pain through adult speech, as cries are an early form of speech and involve similar cognitive and muscular mechanisms[57]. Based on this premise, recent studies have begun to explore adult speech in relation to pain reports, whether from induced acute pain stimuli[58] or diagnosed chronic pain[59]. However, these studies have been limited by the small size of their datasets and the lack of data availability, restricting further research and validation.

TAME Pain performed the first pilot of data collection in the UK using the same data collection protocol[25,37]; however, the UK study recruited only 15 participants, mostly female, and the annotation was done by a local team at the University of Southampton and is less comprehensive. The results of the UK study suggested that there are non-vocal pain cues, which became a key motivation behind conducting the US study. Our data release does not include the UK data which is limited for distribution due to UK GDPR laws.

This study addresses the limitations of the UK's small-scale and private dataset by providing the largest publicly released dataset to date that links adult speech with self-reported pain levels. Table 3 provides a comparative overview of our dataset alongside other relevant datasets in this domain[25,58,59]. TAME Pain is distinctive as the only publicly accessible dataset, facilitating further research and development in pain assessment. While there are other growing voice datasets, such as the Bridge2AI Voice[60], these do not include pain-related information. Establishing connections between our dataset and these broader voice datasets could provide a promising avenue for future research. Such efforts would enable the exploration of synergies between pain assessment and other vocal characteristics, which could potentially advance our understanding of speech-based analysis in clinical contexts.

Quality Control Scores. The quality control scores, represented by the action labels ranging from 0 to 4, were designed to categorize the audio files based on their overall quality and the presence of confounding features for pain studies. A score of 0 denotes the highest quality, where the audio is virtually free of disturbances, while a score of 4 represents the lowest quality, with significant background noise or speech disturbances that could interfere with data processing. These labels help ensure that researchers can filter and select audio files that meet the desired quality standards for their specific analyses.

However, researchers should consider adapting the quality thresholds based on the specific requirements of their scientific application. Depending on the sensitivity of the analysis to audio quality, different quality cut-offs may be more appropriate. For instance, some studies might tolerate minor disturbances (e.g., action labels of 1 or 2) without compromising the validity of the results, while others might require the strictest criteria (only using files labeled as 0) to avoid any potential confounding factors.

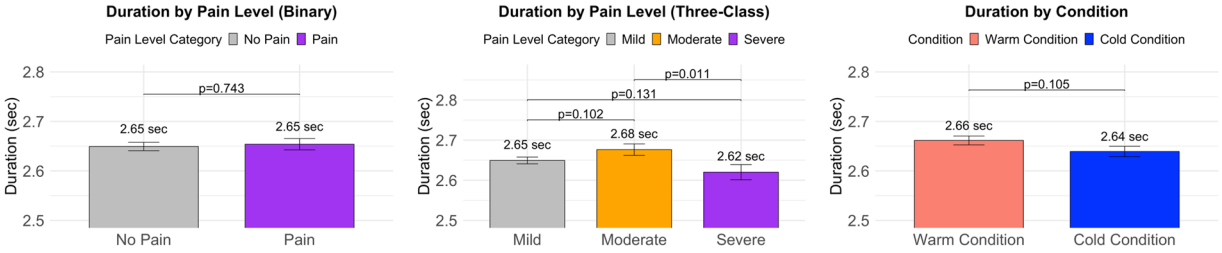

Fig. 5 Left Panel: The average utterance duration for each in the binary pain – no pain Task. The No Pain class serves as the baseline and the Pain class serves as the pain-stimulated condition. Middle Panel: The average utterance duration for the three-class pain. The Mild class serves as the baseline, and the Moderate and Severe class serve as the pain-stimulated condition. Right Panel: The average utterance duration for warm/cold condition. The error bars are the standard error of the mean. The *p*-values are computed using a paired t-test, compares.

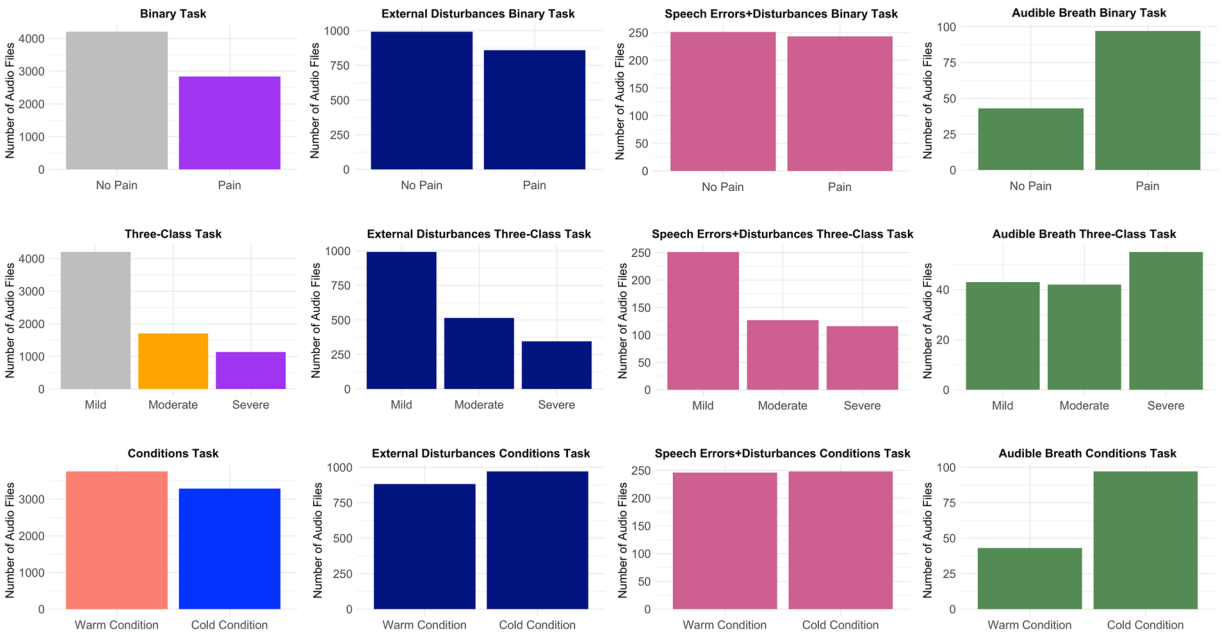

Fig. 6 First (Leftmost) Column: The reference distribution for (from top to bottom) Binary, Three-Class, and Conditions classification. Second Column: Distribution of utterances with label External Disturbances. Third Column: Distribution of utterances with label Speech Errors+Disturbances. Last (Rightmost) Column: Distribution of utterances with label Audible Breath. All categories shows a deviation from the reference distribution.

Application in Real-world Environment. Conducting our experiment in a controlled, quiet environment reduces the influence of external factors that could confound or obscure the pain signal, yielding a robust dataset for studying pain biomarkers and indicators. While this controlled setting does not replicate the acoustic complexities of real-world conditions, the collected data can be augmented with noise profiles from real-world environments or paired with simulated acoustic features to mimic practical scenarios. This flexibility allows the dataset to be tailored for diverse applications by introducing noise systematically, all while maintaining the integrity of the original data.

Ethical Considerations. Although the dataset is intended to help accelerate the development of pain detection and classification models in decision support tools to aid optimal pain management, we recognize potential ethical issues and unintended consequences. For instance, given the relatively small number of participants leading to a less racial and ethnically diverse dataset, caution should be exercised in interpreting outputs from models developed based on our dataset to avoid widening health disparities caused by biased automated systems. We also acknowledge the possibility of dataset misuse that could lead to the development of harmful tools. Given the open-access availability of the dataset, we request that people adhere to ethical principles of non-maleficence (i.e., first, do no harm) and beneficence (i.e., to do good).

To enhance our awareness of such issues, we ensured that our data collection and research activities aligned with a Responsible Research and Innovation (RRI) approach[61,62] based on the AREA (Anticipate, Reflect,

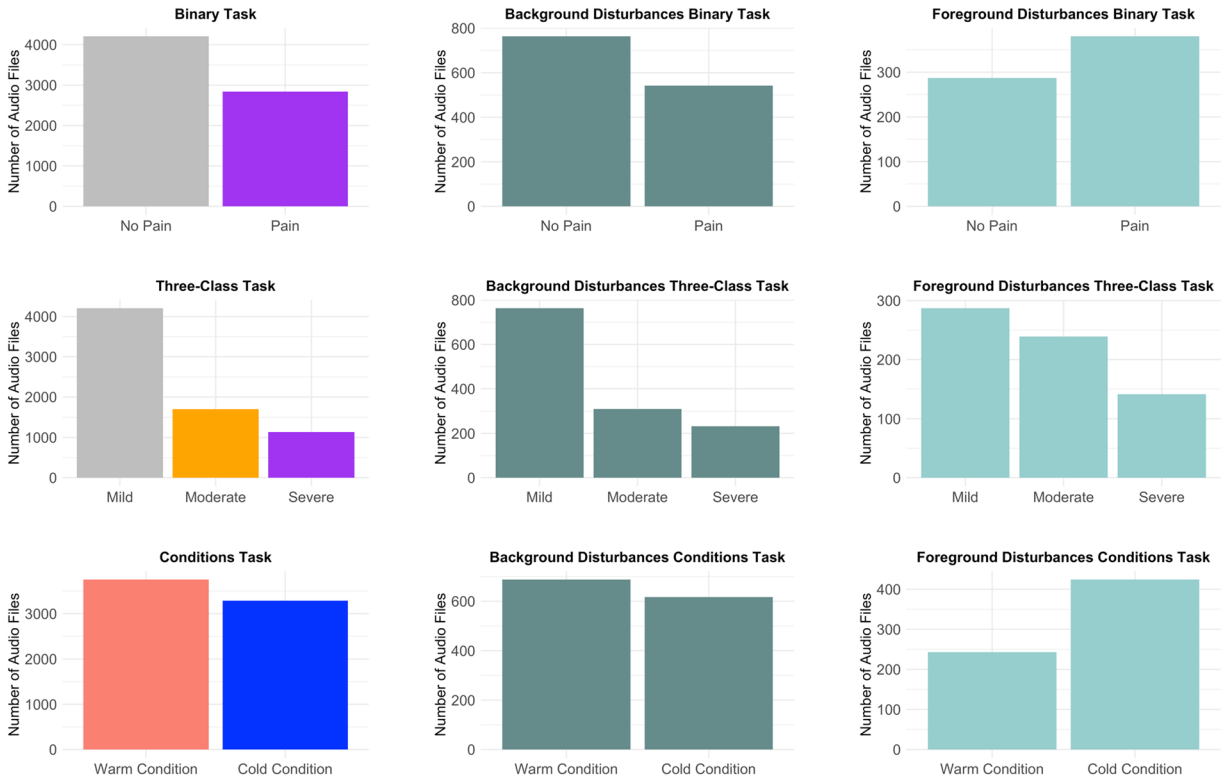

Fig. 7 Left Column: The reference distribution for (from top to bottom) Binary, Three-Class, and Conditions classification. Middle Column: Distribution of utterances with label background disturbance. Right Column: Distribution of utterances with label foreground disturbance. While the foreground disturbance shows a deviation from the reference distribution, the background disturbance is consistent with the reference distribution.

	Speech Prosody[58]	Duesseldorf[59]	TAME Pain (UK)[25]	TAME Pain
# Audio Files	400	844	1,690	7,044
# Participants	27	80	15	51
Total Duration (min.)	—	≈180	76.88	311.24
Ave. Audio File Duration (sec.)	0.93	12.8	2.72	2.65
Language	English	German	English	English
Data Released	No	No	No	Yes

Table 3. Comparison of datasets exploring the relationship between adult speech and pain. TAME Pain (UK) is not included in our data release.

Engage, Act) framework[63]. Liz Dowthwaite conducted RRI workshops among the research team to guide our ethics application and project objectives. This is in addition to securing ethical approval from the Institutional Review Board of the University of Texas at Austin (IRB number: STUDY00004954), have been in compliance with guidelines set by each researcher's institution, and have received a separate ethical approval from the University of Southampton, UK (reference ERGO approval 80074.A1).

Limitations. This data set can be used to extract features for analysis and develop a greater understanding of how pain affects speech. However, it should be noted that the data collection room, although enclosed, was not a controlled audio environment. Therefore, some files contain audio disruptions unrelated to the speaker, which were carefully annotated. Furthermore, the audio files in `mic1_trim_v1.zip` is a raw data set and was not cleaned for disruptions that may confound with feature extraction. We advise removing extreme disturbances, indicated with an Action Label equal to 4, prior to utilizing the files for training or analysis.

The provided manual annotations are useful for model training and validation, especially since hand-labeling is often expensive and time consuming. However, these annotations pose limitations of being subject to human error since they were manually reviewed by a single author, TD. Furthermore, TD was not blinded to the pain levels or conditions reported in each audio file during the annotation process. Additionally, while participants were required to be fluent in English, factors that might independently influence speech patterns, such as developmental learning disorders (e.g., dyslexia), language dominance, or linguistic competence, were not explicitly

monitored. While only healthy subjects participated in our study, conditions such as post-stroke states, depression, or schizophrenia, which could also affect speech, were not formally excluded.

The manual annotations are objective in recognizing a disturbance, but the intensity and single-word descriptor of the noise is subjective, and may be open to interpretation. Similarly, speech errors such as misreadings are an objective identification, while other speech errors like mispronunciations are subjective and were determined errors based on being the minority pronunciation, relative to the dataset.

We have focused on self-reported pain from participants as the ground truth in our dataset. We acknowledge that medical professionals and clinicians may interpret pain signals differently from those reported by the participants. We also acknowledge that our use of Harvard sentences read aloud may not align fully with the breadth of utterances, vocalisations, and non-verbal cues that a person experiencing pain may exhibit when presenting to medical professionals in times of distress.

Data Access. Users must be registered on the PhysioNet data platform and sign a specified data use agreement before accessing the TAME Pain dataset files.

Code availability

For data processing, we used "webrtcvad" (https://github.com/wiseman/py-webrtcvad) to trim the audio data. The following R libraries were used to create the visualizations: dplyr, ggpubr, ggplot2, and tibble. No additional custom code was used to process the data, as all annotations were performed manually.

Received: 11 October 2024; Accepted: 28 February 2025;
Published online: 10 April 2025

References

1. Merboth, M. K. & Barnason, S. Managing pain: the fifth vital sign. *Nursing Clinics of North America* **35**, 375–383 (2000).
2. Breivik, H. *et al*. Assessment of pain. *British journal of anaesthesia* **101**, 17–24 (2008).
3. Karnath, B., Holden, M. D. & Hussain, N. Chest pain: differentiating cardiac from noncardiac causes. *Hospital Physician* **38**, 24–27 (2004).
4. Ruben, M. A., van Osch, M. & Blanch-Hartigan, D. Healthcare providers' accuracy in assessing patients' pain: A systematic review. *Patient education and counseling* **98**, 1197–1206 (2015).
5. Ruben, M. A., Blanch-Hartigan, D. & Shipherd, J. C. To know another's pain: A meta-analysis of caregivers' and healthcare providers' pain assessment accuracy. *Annals of Behavioral Medicine* **52**, 662–685 (2018).
6. Glowacki, D. Effective pain management and improvements in patients' outcomes and satisfaction. *Critical care nurse* **35**, 33–41 (2015).
7. Small, C. & Laycock, H. Acute postoperative pain management. *Journal of British Surgery* **107**, e70–e80 (2020).
8. Loued-Khenissi, L., Martin-Brevet, S., Schumacher, L. & Corradi-Dell'Acqua, C. The effect of uncertainty on pain decisions for self and others. *European Journal of Pain* **26**, 1163–1175 (2022).
9. Karcioglu, O., Topacoglu, H., Dikme, O. & Dikme, O. A systematic review of the pain scales in adults: which to use? *The American journal of emergency medicine* **36**, 707–714 (2018).
10. Nesbitt, J., Moxham, S. & Williams, L. *et al*. Improving pain assessment and managment in stroke patients. *BMJ Open Quality* **4**, u203375–w3105 (2015).
11. Cook, A. K., Niven, C. A. & Downs, M. G. Assessing the pain of people with cognitive impairment. *International Journal of Geriatric Psychiatry* **14**, 421–425 (1999).
12. Liu, J. *et al*. Challenges in the diagnosis and management of pain in individuals with autism spectrum disorder. *Review Journal of Autism and Developmental Disorders* **7**, 352–363 (2020).
13. McNeill, J. A., Sherwood, G. D. & Starck, P. L. The hidden error of mismanaged pain: a systems approach. *Journal of pain and symptom management* **28**, 47–58 (2004).
14. McNeill, J. A., Sherwood, G. D., Starck, P. L. & Thompson, C. J. Assessing clinical outcomes: patient satisfaction with pain management. *Journal of pain and symptom management* **16**, 29–40 (1998).
15. Roy, N., Volinn, E., Merrill, R. M. & Chapman, C. R. Speech motor control and chronic back pain: a preliminary investigation. *Pain Medicine* **10**, 164–171 (2009).
16. Kichloo, A. *et al*. Telemedicine, the current covid-19 pandemic and the future: a narrative review and perspectives moving forward in the usa. *Family medicine and community health* **8**, e000530 (2020).
17. Patel, S. Y. *et al*. Trends in outpatient care delivery and telemedicine during the covid-19 pandemic in the us. *JAMA internal medicine* **181**, 388–391 (2021).
18. Colbert, G. B., Venegas-Vera, A. V. & Lerma, E. V. Utility of telemedicine in the covid-19 era. *Reviews in cardiovascular medicine* **21**, 583–587 (2020).
19. Rowbotham, S., Wardy, A. J., Lloyd, D. M., Wearden, A. & Holler, J. Increased pain intensity is associated with greater verbal communication difficulty and increased production of speech and co-speech gestures. *Plos one* **9**, e110779 (2014).
20. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology* **5**, 96–116 (2020).
21. Koops, S. *et al*. Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)* **22**, 152–160 (2023).
22. De Boer, J. *et al*. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychological medicine* **53**, 1302–1312 (2023).
23. Marmar, C. R. *et al*. Speech-based markers for posttraumatic stress disorder in us veterans. *Depression and anxiety* **36**, 607–616 (2019).
24. Duey, A. H. *et al*. Daily pain prediction using smartphone speech recordings of patients with spine disease. *Neurosurgery* **93**, 670–677 (2023).
25. Williams, J. *et al*. Predicting acute pain levels implicitly from vocal features. In *Proc. Interspeech 2024*, 1460–1464, https://doi.org/10.21437/Interspeech.2024-15 (2024).
26. Tracy, L. M. *et al*. Meta-analytic evidence for decreased heart rate variability in chronic pain implicating parasympathetic nervous system dysregulation. *Pain* **157**, 7–29 (2016).
27. Lawrence, K. *et al*. Building telemedicine capacity for trainees during the novel coronavirus outbreak: a case study and lessons learned. *Journal of General Internal Medicine* **35**, 2675–2679 (2020).
28. Johnston, C. C. & Strada, M. E. Acute pain response in infants: a multidimensional description. *Pain* **24**, 373–382 (1986).

29. Zamzami, G. *et al*. Pain assessment in infants: Towards spotting pain expression based on infants' facial strain. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 5, 1–5 (IEEE, 2015).
30. Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E. & Matthews, I. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 57–64 (IEEE, 2011).
31. Werner, P. *et al*. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing* **8**, 286–299 (2016).
32. Zamzmi, G. *et al*. A review of automated pain assessment in infants: features, classification tasks, and databases. *IEEE reviews in biomedical engineering* **11**, 77–96 (2017).
33. Werner, P. *et al*. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing* **13**, 530–552 (2019).
34. Gkikas, S. & Tsiknakis, M. Automatic assessment of pain based on deep learning methods: A systematic review. *Computer methods and programs in biomedicine* **231**, 107365 (2023).
35. Salekin, M. S. *et al*. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Computers in biology and medicine* **129**, 104150 (2021).
36. Powell, D. Walk, talk, think, see and feel: harnessing the power of digital biomarkers in healthcare. *NPJ Digital Medicine* **7**, 45 (2024).
37. Schneiders, E. *et al*. 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*, TAS '23 (2023).
38. Mitchell, L. A., MacDonald, R. A. & Brodie, E. E. Temperature and the cold pressor test. *The Journal of Pain* **5**, 233–237 (2004).
39. Brown, J. E., Chatterjee, N., Younger, J. & Mackey, S. Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal stimulation. *PloS one* **6**, e24124 (2011).
40. Barton, N. J. *et al*. Pressure application measurement (pam): a novel behavioural technique for measuring hypersensitivity in a rat model of joint pain. *Journal of neuroscience methods* **163**, 67–75 (2007).
41. Katsarava, Z. *et al*. A novel method of eliciting pain-related potentials by transcutaneous electrical stimulation. *Headache: The Journal of Head and Face Pain* **46**, 1511–1517 (2006).
42. Velasco, M., Gómez, J., Blanco, M. & Rodriguez, I. The cold pressor test: pharmacological and therapeutic aspects. *American journal of therapeutics* **4**, 34–38 (1997).
43. McIntyre, M. H. *et al*. Validity of the cold pressor test and pain sensitivity questionnaire via online self-administration. *PLOS ONE* **15**, 1–16 (2020).
44. Zhao, Q. *et al*. Reproducibility of blood pressure response to the cold pressor test: the gensalt study. *American journal of epidemiology* **176**, S91–S98 (2012).
45. Rothauser, E. IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics* **17**, 225–246 (1969).
46. Kwong, A. Harvard sentences Accessed: 2024-07-24 (2024).
47. Lakhsassi, L., Borg, C., Martusewicz, S., van der Ploeg, K. & de Jong, P. J. The influence of sexual arousal on subjective pain intensity during a cold pressor test in women. *Plos one* **17**, e0274331 (2022).
48. Ghiasi, S., Greco, A., Barbieri, R., Scilingo, E. P. & Valenza, G. Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge. *Scientific Reports* **10**, 5406 (2020).
49. Schwabe, L., Haddad, L. & Schachinger, H. HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology* **33**, 890–895 (2008).
50. Miró, J. *et al*. Defining mild, moderate, and severe pain in young people with physical disabilities. *Disability and rehabilitation* **39**, 1131–1135 (2017).
51. Dao, T.-Q. *et al*. Tame pain: Trustworthy assessment of pain from speech and audio for the empowerment of patients (version 1.0.0). *PhysioNet* https://doi.org/10.13026/20e2-1g10 (2025).
52. Borgbjerg, F. M., Nielsen, K. & Franks, J. Experimental pain stimulates respiration and attenuates morphine-induced respiratory depression: a controlled study in human volunteers. *Pain* **64**, 123–128 (1996).
53. Kato, Y., Kowalski, C. J. & Stohler, C. S. Habituation of the early pain-specific respiratory response in sustained pain. *Pain* **91**, 57–63 (2001).
54. Mittal, V. K. Discriminating features of infant cry acoustic signal for automated detection of cause of crying. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5 (IEEE, 2016).
55. Ntalampiras, S. Audio pattern recognition of baby crying sound events. *Journal of the Audio Engineering Society* **63**, 358–369 (2015).
56. Corvin, S. *et al*. Pain cues override identity cues in baby cries. *Iscience***27** (2024).
57. Lieberman, P. The physiology of cry and speech in relation to linguistic behavior. In *Infant crying: Theoretical and research perspectives*, 29–57 (Springer, 1985).
58. Oshrat, Y. *et al*. Speech prosody as a biosignal for physical pain detection. In *Conf Proc 8th Speech Prosody*, 420–24 (2016).
59. Ren, Z. *et al*. Evaluation of the pain level from speech: Introducing a novel pain database and benchmarks. In *Speech Communication; 13th ITG-Symposium*, 1–5 (VDE, 2018).
60. Bensoussan, Y., Elemento, O. & Rameau, A. Voice as an ai biomarker of health-introducing audiomics. *JAMA Otolaryngology–Head & Neck Surgery* **150**, 283–284 (2024).
61. Owen, R. & Pansera, M.*Responsible innovation and responsible research and innovation* (Edward Elgar Publishing, 2019).
62. Stilgoe, J., Owen, R. & Macnaghten, P. Developing a framework for responsible innovation. In *The Ethics of Nanotechnology, Geoengineering, and Clean Energy*, 347–359 (Routledge, 2020).
63. Stahl, B. C. *et al*. Assessing responsible innovation training. *Journal of Responsible Technology* 100063 (2023).

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/V00784X/1] UKRI Trustworthy Autonomous Systems Hub, Responsible AI UK [grant number EP/Y009800/1], and Good Systems, a research grand challenge at the University of Texas at Austin. The authors would like to thank Prof. Sarvapaili (Gopal) Ramchurn, Prof. Joel Fischer, Prof. Sharon Strover, Dr. Liz Dowthwaite, Dr. Rohan Chandra, and Dr. Anna-Maria Piskopani for their helpful feedback during this work.

Author contributions

A.F., E.S., J.W., T.S., J.R.B., and G.V. conceptualized the project. E.S., J.W., R.B., T.S., and G.V. developed the study protocol. E.S., T.S., and J.W. designed the experimental setup. J.R.B., T.Q.D., and A.F. prepared and acquired the IRB. T.Q.D., R.V., and A.F. obtained consent. E.S., T.Q.D., R.V., A.F., and T.S. were involved in data acquisition. T.Q.D. and A.F. processed the data. T.Q.D. carried out data annotation task. A.F., E.S., J.W., T.S., and R.K. helped in data annotation validation. T.Q.D., R.K., and A.F. validated the data. T.Q.D., R.K., and A.F. performed analysis and created the visualizations. T.Q.D. and A.F. wrote the paper draft. All authors contributed to editing the paper and reviewed the manuscript.

Competing interests

Jennifer Williams reports a relationship with The Alan Turing Institute that includes consulting or advisory. Jennifer Williams reports a relationship with MyVoice AI that includes employment. Jennifer Williams has patent #US20230186896A1 pending to MyVoice AI Ltd. Jennifer Williams has patent #US20220405363A1 pending to MyVoice AI Ltd. In regard to prior employment at MyVoice AI Ltd, Jennifer Williams was previously employed part-time while the work in this manuscript was undertaken (ending in February 2024) and does not have any remaining interactions nor any restrictive covenants, but there are two pending patents on voice-based biometric identification pending where Jennifer Williams is listed as a lead inventor, and MyVoice AI Ltd is the assignee. Those two patents are not topically related to the work in this manuscript. The rest of the authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04733-2.

Correspondence and requests for materials should be addressed to T.-Q.D. or A.F.