In this document, we provide responses to the questions posed in the Augmented Datasheets for Speech Datasets framework proposed by Papakyriakopoulos et al. (FAccT 2023). Their framework aims to improve transparency in speech dataset documentation.

3.1.1 What is the speech dataset name, and does the name accurately describe the contents of the dataset?

- The pipeline which can be used to download the dataset is hosted at https://github.com/97jamie/public-police-footage, which refers to it as public-police-footage. While not a conventional dataset name, it does accurately reflect the contents: publicly released police body-worn camera footage processed into speech data. The name is descriptive but could benefit from clarification that it includes transcriptions and is built for ASR research.

3.1.2 Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?

- The dataset can be used to draw conclusions on spontaneous speech only.

3.1.3 Describe the process used to determine which linguistic subpopulations are the focus of the dataset

- The dataset was built from body-worn camera (BWC) footage publicly released by 13 police departments across 6 U.S. states (primarily California). Videos were identified through YouTube searches and department websites using terms like “critical incident” and “OIS.” There was no filtering based on speaker demographics, dialect, or other linguistic variables. As a result, the linguistic subpopulations represented in the dataset are determined by geography (i.e., state-level location) and speaker role (e.g., officer, community member, or radio). However, there is no demographic labeling (e.g., race, gender, age, etc.) of the speakers.

3.2.1 How many hours of speech were collected in total (of each type, if appropriate), including speech that is not in the Augmented Datasheets for Speech Datasets and Ethical Decision-Making FAccT ’23, June 12–15, 2023, Chicago, IL, USA dataset? If there was a difference between collected and included speech, why? E.g., if the speech data are from an interview and the dataset contains only the interviewee’s responses, how many hours of speech were collected in interviews from both interviewer and interviewee?

- Hand-cleaned (test + validation): 57.7 minutes = 0.96 hours (From Table I)
- Automatically cleaned (training): 125.8 minutes = 2.10 hours (From Table II)
- Total included in dataset release: ~3.06 hours of speech.

3.2.2 How many hours of speech and number of speakers & words are in the dataset (by each type, if appropriate)?

- See answer to 3.2.1.

3.2.3 Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?

- The dataset does not use any standardized definitions of linguistic subpopulations. The only categorization applied is by speaker role (officer, community member, or radio), which is annotated in the hand-cleaned portion of the data. These roles are not based on linguistic features and are not linked to demographic or standardized population labels. No further metadata is provided regarding speaker identity, language variety, or linguistic subpopulation.

3.2.4 For any linguistic subpopulations identified in the dataset, please provide a description of their respective distributions within the dataset.

- No linguistic subpopulations are identified in the dataset, so no distribution information is available.

3.2.5 How much of the speech data have corresponding transcriptions in the dataset?

- All of the speech data included in the dataset—approximately 3 hours—has corresponding transcriptions. Speech segments without captions or with poor OCR/alignment quality were excluded during processing, so the final dataset consists entirely of transcribed audio.

3.2.6 Does the dataset contain non-speech mediums (e.g. images or video)?

- The original data comes from videos, but the resulting dataset contains no non-speech media. It consists only of transcriptions, metadata (like timestamps and frame numbers), and code to recollect the data. Researchers who want video or images have to download and process the footage themselves using the pipeline.

3.2.7 Do speakers code switch or speak multiple languages, and if so, how is this identified in the data?

- Since the dataset consists of spontaneous police encounters in U.S. cities, some instances of code-switching or non-English speech may occur naturally—especially in community member speech—but the dataset does not explicitly mark or track these.

3.2.8 Does the speech dataset focus on a specific topic or set of topics?

- Yes, the dataset focuses on a specific domain: police body-worn camera footage, particularly from critical incidents such as officer-involved shootings and use-of-force events.

3.2.9 Does the dataset include sensitive content that can induce different emotions (e.g., anger, sadness) that can cause the speakers to produce unusual pitch or tone deviating from plain speech?

- Yes, the dataset includes sensitive and emotionally charged content, as it is constructed from police body-worn camera footage released in connection with critical incidents such as shootings, arrests, and use-of-force events.
- These situations often involve high emotional intensity, including fear, stress, anger, or distress—especially in the speech of community members and officers. As a result,

speakers may produce speech that deviates from neutral patterns in pitch, tone, prosody, or fluency.

- This makes the dataset particularly valuable for developing or evaluating speech models in non-neutral, high-stakes, real-world conditions, but it also introduces acoustic challenges like shouting or overlapping speech.

3.2.10 Does the dataset contain content that complies to the users' needs, or does it result in symbolic violence (the imposition of religious values, political values, cultural values, etc.)?

- The dataset does not impose religious, political, or cultural values. It's a tool for studying speech in public police footage and is focused on accountability-related research.

3.3.1 What mechanisms or procedures were used to collect the speech data, e.g.: is the data a new recording of read speech, or an interview? Or is it downloaded speech data from public speeches, lectures, YouTube videos or movies, etc.?

- The speech data was collected by downloading publicly released police body-worn camera videos from YouTube, using links found on police department websites or through keyword searches (e.g., "critical incident," "OIS").

3.3.2 Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?

- No, the data was not collected in a consistent technical setting. All audio comes from real-world police body-worn camera footage, which varies by department, device, environment, and editing practices.

3.3.3 Is there presence of background noise?

- Yes, there is background noise throughout the dataset. The audio comes from real-world police body camera footage, so background sounds like traffic, wind, overlapping speech, and radio transmissions are common. This makes the audio more challenging than studio or lab-recorded speech.

3.3.4 For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are "fair and neutral"?

- Not applicable.

3.3.5 Have data subjects consented to the disclosure of the metadata in the dataset? Also, does the metadata include sensitive personal information such as disability status?

- No, data subjects did not provide consent for disclosure, but the footage was already publicly released by police departments. The dataset includes no sensitive personal information, and the authors specifically exclude segments with names or identifying content.

3.4.1 When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?

- No, background noise was not deleted or adjusted. The dataset preserves the original audio conditions from the publicly released videos, which vary in quality. No normalization was applied to make recordings sound more similar.

3.4.2 Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?

- Yes, human annotators were hired to transcribe and correct a subset of the data. All annotators were fluent English speakers and also co-authors of the paper, meaning they were familiar with the corpus material, vocabulary, and general characteristics of the speech domain. While not formally trained in transcription, they had contextual knowledge of the dataset and task.

3.4.3 If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?

- The dataset used two transcription methods: automated OCR-based caption extraction and manual correction for a subset of the data. In the hand-corrected set, transcriptions were validated by the annotators (who are also the authors) using a consistent annotation procedure. Captions were corrected via rule-based edits, not freeform rewriting, to support reproducibility. Annotators also aligned time boundaries to audio and labeled speaker roles and overlapping speech.

3.4.4 If the speech data include transcriptions, what software was used in the generation of the transcriptions (including, e.g., software used by human transcribers)? Are timestamps included in transcriptions? Are the alignments provided with the transcripts?

- Transcriptions were initially generated using OCR software, primarily PaddleOCR, applied to on-screen captions in the videos. For the hand-corrected set, annotators (i.e., the authors) reviewed and edited the OCR output using spreadsheets and Praat. Timestamps are included for all transcriptions. In the hand-corrected data, these timestamps were manually aligned to the audio. The alignments are released alongside the transcripts and include start/end times for each captioned segment.

3.4.5 Were transcription conventions (such as tagging scheme, treatment of hate speech or swear words, etc.) disclosed Augmented Datasheets for Speech Datasets and Ethical Decision-Making FAccT ’23, June 12–15, 2023, Chicago, IL, USA along with the corpus?

- No, transcription conventions (such as a tagging scheme or treatment of swear words) were not formally disclosed alongside the dataset. In the hand-corrected set, captions were edited for accuracy using rule-based edits, but there was no standardized scheme for things like hate speech or profanity. Swear words were generally left as-is unless already censored in the video.

3.4.6 Is additional coding performed, separate to transcriptions and tagging?

- Yes, some additional coding was done. In the hand-corrected set, we labeled speaker roles, marked overlapping speech, and flagged identifying info for exclusion. That’s it—no linguistic tagging or sentiment coding.

3.5.1 How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?

- Redactions are handled by excluding any segments with identifying information. No audio is censored—if a caption included a name, we just didn't include it. No redistribution of raw audio means no in-line redaction needed.

3.5.2 Is there any part of this dataset that is privately held but can be requested for research purposes?

- No.

3.5.3 Is there a sample dataset distributed? If so, how well does the sample represent the actual dataset? Do they include all forms of speech included in the dataset? How big is the sample?

- Yes, the hand-corrected set of 20 videos functions as a sample dataset. It includes approximately 0.96 hours of speech and is used for validation and testing. The sample is representative in terms of data type (body-worn camera footage with embedded captions) and includes all speaker roles found in the full dataset: officer, community member, and radio. It's also the only part of the dataset with manually verified transcriptions and alignments, making it a higher-quality slice of the full pipeline output.

3.5.4 Aside from this datasheet, is other documentation available about the data collection process (e.g., agreements signed with data subjects and research methodology)?

- Yes, the paper itself provides detailed documentation of the data collection and processing methodology, including video identification, OCR, caption extraction, alignment, and filtering procedures.