Datasheet for the Public Airwaves Dataset

Generated by Automated Data Collection

March 10, 2025

1 Motivation

The **Public Airwaves** dataset consists of recorded audio clips from various online radio stations, specifically focusing on distinguishing between human speech and music. The dataset can serve as a foundational resource for building lightweight models to classify live audio streams into categories such as *news* or *music*.

2 Composition

2.1 Data Collection Process

The dataset was collected by streaming audio from multiple publicly available radio stations using `ffmpeg`. Each recording was stored as an `.mp3` file with a duration of 30 seconds.

- **Radio Stations:** The following online stations were included:
  - **BBC Radio 1:** `http://bbcmedia.ic.llnwd.net/stream/bbcmedia_radio1_mf_p`
  - **NPR Live:** `http://npr-ice.streamguys1.com/live.mp3`
  - **KEXP Radio:** `http://live-aacplus-64.kexp.org/kexp64.aac`
- **Total Clips:** 30 recordings
- **Format:** MP3 audio files
- **Metadata:** Each file is associated with metadata stored in a CSV file, containing:
  - Station Name
  - Filename
  - Timestamp of Recording
  - Duration (30 seconds per clip)

2.2 Dataset Structure

The dataset is organized into a folder named `Public_Airwaves`, containing:

- MP3 audio recordings labeled as: `[Station]_[Timestamp]_[ID].mp3`
- A metadata file (`metadata.csv`) storing structured information about each clip.

3 Uses and Applications

- Training machine learning models to classify audio streams into **news** or **music**.
- Analyzing the acoustic features of speech vs. instrumental audio.
- Developing real-time monitoring systems for media content categorization.

4 Limitations

- No manual labels: Clustering was used to infer audio categories.
- Data limited to specific radio stations and may not generalize globally.
- No speech-to-text transcripts; focus is on acoustic features.

5 Ethical Considerations

- The dataset only contains publicly available broadcasted content.
- No private or user-generated content was recorded.
- Any commercial use should ensure compliance with broadcasting regulations.

6 Future Work

- Expanding dataset diversity with additional radio stations.
- Incorporating speech recognition for further classification.
- Adding more granular categories (e.g., advertisements, interviews, talk shows).

7 Acknowledgments

This dataset was generated using automated audio recording scripts via `ffmpeg`. The radio station streams are publicly accessible and belong to their respective broadcasters.