Audio Watermarking Training Dataset Table

Model Comparison

Category	MVP Baseline	Commercial-Grade Upgrade
Training Steps	~70,000	~1,260,000
Dataset Size	~5,000 hours (~5TB)	~100,000 hours (~100TB)
Audio Sampling Rate	16 kHz	44.1 kHz++
Limitations	Muted audio issues, low payload efficiency, no real-time processing	Targeting real-time support, higher payload, and enhanced robustness
Hardware Requirements	Moderate	High-performance GPUs/TPUs
Key Goals	Embed and recover watermarks under limited attacks	Robustness to complex distortions, real-time processing, higher fidelity

Dataset Details (MVP Baseline)

Dataset	Description	Hours Used	Total Hours	Sampling	Purpose
LibriSpeech	English speech from LibriVox	1000	1000	360 training, 40 validation, 40 testing	Speech watermarking
Common Voice	Multilingual speech (10 languages)	1700	1700	1339 training, 171 validation, 200 testing	Multilingual speech watermarking
Audio Set	Generic audio events (subset)	1337	5790	997 training, 140 validation, 200 testing	Audio event watermarking (specific events)
Free Music Archive	High-quality music	888	888	404 training, 146 validation, 248 testing	Music watermarking
Total		5025	9378	2981 training, 456 validation, 608 testing	Diverse audio watermarking

Dataset Expansion for Commercial-Grade Model

- **Increase dataset size:** Source additional open-source and proprietary datasets (music, speech, ambient/environmental sounds).
- **Explore datasets:** Million Song Dataset, TUT Acoustic Scenes, CHiME.
- **Data augmentation:** Implement pitch-shifting, noise injection, echo, and complex attack combinations.
- **Validation & Testing Data:** Allocate a consistent proportion (10-20%) of the final dataset.

Additional Notes

- The total training data for the MVP Baseline is approximately 5,000 hours.
- Data sampling for validation and testing sets is consistent across all datasets.
- The Audio Set dataset is a subset, focusing on specific audio events relevant to watermarking.