Audio-to-Text Conversion - Data layer

Datasets — Benchmarks & Sources

LibriSpeech (ASR)

- **What it is:** ~1,000 hours of 16 kHz read English speech (audiobooks) with clean/other splits.
- **Why it matters:** Classic ASR benchmark; ubiquitous for pretraining/finetuning and reporting WER.
- **Quirks:** Read speech (not conversational); relatively clean; long utterances; domain mismatch vs meetings/calls.
- **Where:** Hugging Face (`librispeech_asr`).

Common Voice (v11–v17)

- **What it is:** Crowd-sourced, multilingual read speech (dozens of languages, hours vary by locale).
- **Why it matters:** Diversity in accents/devices; permits multilingual/low-resource experiments.
- **Quirks:** Quality varies; label noise; per-language class imbalance (few hours in some locales).
- **Where:** Hugging Face (`mozilla-foundation/common_voice_17_0` and earlier versions).

Multilingual LibriSpeech (MLS)

- **What it is:** ~50k hours across 8 European languages from LibriVox audiobooks (16 kHz).
- **Why it matters:** Scale + multilingual; excellent for cross-lingual transfer and pretraining.
- **Quirks:** Read speech; language imbalance (English/German dominant).
- **Where:** Hugging Face (`facebook/multilingual_librispeech` or `mls` community loaders).

TED-LIUM 3

- **What it is:** ~452 hours of English TED talks with aligned transcripts.
- **Why it matters:** Semi-spontaneous, microphone + auditorium acoustics; popular for domain adaptation beyond read speech.
- **Quirks:** Applause/music segments; varying mic quality; longer talks → need chunking/VAD.
- **Where:** Hugging Face (`tedlium`).

AMI Meeting Corpus

- **What it is:** ~100 hours of multi-speaker, multi-mic English meetings (headsets + room mics).
- **Why it matters:** Meeting/diarization-heavy ASR; far-field and overlapping speech.
- **Quirks:** Crosstalk/overlap; requires beamforming or channel selection; segment boundaries matter.
- **Where:** Hugging Face (`ami`).

VoxPopuli (ASR)

- **What it is:** EU Parliament recordings in 23 languages; thousands of hours, transcribed subsets.
- **Why it matters:** Large-scale multilingual political speech; good for robust, cross-lingual ASR.
- **Quirks:** Formal register; long sessions → chunking; language imbalance.
- **Where:** Hugging Face (`facebook/voxpopuli`).

GigaSpeech

- **What it is:** ~10k hours English from diverse sources (Podcasts, Audiobooks, YouTube, etc.) with normalized transcripts.
- **Why it matters:** Size + domain diversity; modern large-scale pretraining fine-tune set.
- **Quirks:** License tiers (XS–XL); some segments noisy; careful with text normalization.
- **Where:** Hugging Face (`speechcolab/gigaspeech`).

AISHELL-1 / AISHELL-2

- **What it is:** 178 h (A1) / 1k h (A2) Mandarin read speech at 16 kHz with Pinyin/Chinese transcripts.
- **Why it matters:** Standard Mandarin ASR benchmarks; strong for tonal language experiments.
- **Quirks:** Read, relatively clean; character vs pinyin targets → pick one consistently.
- **Where:** Hugging Face (`aishell`, `aishell2` community loaders).

Switchboard + Fisher (LDC)

- **What it is:** ~2.4k hours of English telephone conversational speech (2-party calls).
- **Why it matters:** Conversational ASR staple; great for spontaneous speech and disfluencies.
- **Quirks:** Licensed (LDC); 8 kHz telephony; strong domain shift vs studio audio.
- **Where:** (Licensed) HF community loaders exist but require local data; otherwise via LDC.

WSJ (Wall Street Journal)

- **What it is:** ~80 h read newspaper text (dictation-style) at 16 kHz.
- **Why it matters:** Longstanding benchmark; good for small-scale experiments and decoding research.

- **Quirks:** Small; clean; not representative of conversational speech.
- **Where:** Hugging Face (`wsj` community loaders; often requires local copies).

CHiME-5 / CHiME-6

- **What it is:** Far-field, multi-mic conversational speech in real homes (dinner parties).
- **Why it matters:** Robust ASR under real, noisy, overlapping conditions; multichannel enhancement research.
- **Quirks:** Heavy overlap; needs diarization/VAD; beamforming recommended.
- **Where:** Hugging Face (`chime5`, `chime6` community loaders).

Earnings-22 (and Earnings-21)

- **What it is:** ~100 h+ of English earnings calls from public companies with transcripts.
- **Why it matters:** Domain-specific ASR for finance; long-form, jargon, numbers.
- **Quirks:** Long utterances; many numerals/tickers; requires careful text normalization.
- **Where:** Hugging Face (`mozilla-foundation/earnings22`).

FLEURS

- **What it is:** 102 languages, prompted read sentences; designed for speech recognition and language ID.
- **Why it matters:** Strong coverage for low-resource, multilingual ASR evaluation.
- **Quirks:** Short prompted phrases; limited hours per language.
- **Where:** Hugging Face (`google/fleurs`).

SPGISpeech *(restricted)*

- **What it is:** ~5k hours of English finance/earnings calls with transcripts.
- **Why it matters:** Large domain corpus for enterprise ASR.
- **Quirks:** Access-restricted; variable audio quality; long segments.
- **Where:** Hugging Face (`speechcolab/spgispeech`, gated).

Preprocessing (what to do and why)

Resampling

We convert all audio to a consistent sample rate so models see uniform time scales.

- **Target 16 kHz mono:** `torchaudio.transforms.Resample(orig_sr, 16000) mono` → Matches most wav2vec2/Conformer/Whisper finetunes; lowers compute if source >16 kHz.
- **8 kHz telephony kept at 8 kHz (optional upsample):** `keep_8k or upsample_to_16k` → Avoids artifacts; if model expects 16 kHz, upsample with high-quality resampler.

Loudness / Amplitude Normalization

We standardize loudness to stabilize training across devices and speakers.

- **Peak/ RMS/ LUFS normalize to target (e.g., −23 LUFS):** `loudnorm` → Reduces variance; prevents clipping and vanishing signals.

Voice Activity Detection (VAD) & Chunking

We trim silence and split long recordings to manageable windows.

- **Silero/WebRTC VAD + max_len (e.g., 20–30 s):** `trim + split` → Lowers padding; improves batch utilization; enables streaming decoding.

Feature Extraction

We transform waveforms into model-friendly features.

- **Raw waveform (wav2vec2/Conformer):** `float32 in [-1,1]` → End-to-end models learn features; minimal preprocessing.
- **Log-Mel spectrogram (Whisper/ESPnet):** `80-mel, 25 ms win, 10 ms hop, log(·)` → Matches pretrained expectations for spectrogram-based models.
- **Cepstral Mean/Var Norm (CMVN) for classical pipelines:** `per-speaker/per-utterance` → Stabilizes MFCC/Mel features.

Text Normalization

We standardize transcripts to reduce label entropy and improve WER.

- **Lowercase, strip punctuation (task-dependent):** `normalize_numbers=true` → Consistent targets; decide on numerals (e.g., “twenty-one” vs “21”).
- **Language-specific rules:** `Chinese no spaces; Arabic diacritics` → Prevents tokenization drift; improves cross-locale comparability.

Tokenization / Labeling

We convert text to token IDs for CTC or seq2seq decoders.

- **CTC char/BPE vocab:** `AutoTokenizer or custom SentencePiece` → CTC needs blank token; keep vocab small for low-resource.
- **Seq2seq (Whisper/Transducer) tokenizer:** `WhisperProcessor / SentencePiece` → Handles language/task tokens (e.g., `<|en|><|transcribe|>`).

Data Augmentation

We simulate real-world conditions to improve robustness.

- **SpecAugment:** `Time/freq masking` → Regularizes; strong gains for low-resource.

- **Speed perturb:** `sox tempo 0.9/1.0/1.1` → Speaker/rate diversity; cheap and effective.
- **Additive noise & RIRs:** `MUSAN + simulated room IRs` → Robust to background noise/reverb; essential for far-field.
- **Codec/phone effects:** `Opus/GSM reencode` → Domain-match for telephony or meeting platforms.

Padding & Batching

We pad to common lengths to form efficient batches without OOM.

- **Dynamic padding by longest in batch:** `DataCollator with padding=true` → Minimizes wasted compute; combine with bucketing by duration.
- **Length capping:** `truncate or sliding windows` → Keeps VRAM bounded; use long-form decoding with chunked attention.

Dataloading tips

We prepare the dataset so training is fast, reproducible, and efficient.

- **Streaming datasets:** `load_dataset(..., streaming=True)` → Start training without full download; ideal for 1000h+ corpora.
- **Duration bucketing:** `group_by_length(duration_sec)` → Reduces padding; steadier step times and more stable training.
- **Prefetch & pin memory:** `DataLoader(pin_memory=True, prefetch_factor>1)` → Keeps GPUs busy; fewer host→device stalls.
- **Mixed precision I/O:** store `float32` on disk, cast to `float16` on device → Saves VRAM and speeds up compute with minimal quality loss.
- **Worker init & seeding:** `worker_init_fn=seed_all` → Reproducible augmentations and shuffling.
- **Deterministic validation:** fixed VAD/segmentation + consistent text normalization → Fair, comparable WER across checkpoints.
- **Long-form eval:** stitch chunked hypotheses with timestamps → Accurate WER on TED/meetings/earnings calls without truncation bias.
