Dataset Information

# Multi-Accent English ASR Dataset

| | |
|---|---|
| **Name** | Multi-Accent English ASR Dataset |
| **Version** | v1.0 — April 2026 |
| **Tier** | Open Source |
| **Modality** | Audio + text (speech-to-text / ASR) |
| **Spoken language** | English |
| **L1 backgrounds** | 11 — Chinese, Vietnamese, Thai, Japanese, Russian, Polish, General East European, Indonesian, French, German, South Korean |
| **Total recordings** | 7,377 utterances |
| **Total duration** | 10.25 hours |
| **Speaking style** | Read speech — Harvard Sentences corpus (public domain) |
| **Sample rate** | 16 kHz, narrowband, mono |
| **Transcripts** | Verbatim · lowercase a–z + spaces only · disfluencies included · no punctuation, no capitalisation, no non-speech tags |
| **Speaker demographics** | Self-reported by the contributor: ethnicity, country, city, timezone, gender — gender-balanced within every L1 group |
| **Provenance** | 100% human-produced by native L1 speakers on Workbolt in their home country; no synthetic, no scraped web audio |

| | |
|---|---|
| **Recording conditions** | Contributors' own hardware (laptop or phone microphones) in quiet rooms; SNR floor and clipping checks enforced per file |
| **QA pipeline** | Automated acoustic QA (clipping, SNR, completeness) plus human review against the rubric in the Ocular Data Foundry |
| **License** | ODC-By v1.0 (https://opendatacommons.org/licenses/by/1-0/) — permissive, attribution-only; suitable for research and commercial training |
| **Access** | Sample audio and spec public on the Data Studio; full bundles via the contact form at useocular.com/contact |
| **Suggested citation** | Contains information from the Multi-Accent English ASR Dataset by Ocular AI, made available under ODC-By v1.0. |

*Properties listed here describe the v1.0 release shipped April 2026. Future releases will be documented as new versions with their own datasheets.*