India Speaks – Preprocessing Pipeline: Design & Limitations

Objective

To design and implement a scalable multilingual audio-text preprocessing pipeline that adheres to the India Speaks standards. The pipeline prepares real-world, crowd-sourced speech data for downstream ASR and TTS model training.

Key Design Features

1. Audio Validation (Stubbed)

- File path pattern check: must begin with `s3://`
- Duration validation: reject rows > 15s
- Sample rate/corruption: not tested (no audio access); simulated using stub logic as per task instruction.

2. Text Normalization

- Unicode normalization (NFC)
- Hindi digit expansion using standard mapping (e.g., `2` → "दो")
- Allowed punctuation only (as per PDF): `. , ? ! ' - : ;`
- English text is lowercased
- Non-verbal tokens (e.g., `[laugh]`, `[breath]`) are preserved using square brackets

3. Language Mismatch Detection

- Used `fastText` (`lid.176.bin`) for robust multilingual language detection
- Rows where predicted language ≠ labeled language are flagged (not rejected)
- Added `lang_warning` column to track mismatches

4. Rejection Reasons

Rows are rejected only for:

- Invalid audio path
- Duration > 15s
- Missing transcription

Each rejected row is saved in `rejected.csv` with a specific `reason`.

Outputs

- `train_ready.csv`: Clean rows (includes `lang_warning` if applicable)
- `rejected.csv`: Problematic rows with `reason`

Known Limitations

1. WER Mismatch Not Implemented
   - The dataset provides only one transcription (`transcription_raw`)

- WER requires a reference transcription (ground truth)
- Thus, WER-based rejection was skipped and documented

2. Digit Expansion Only for Hindi
- Due to lack of digit maps for all languages
- Matches PDF policy; other languages untouched

3. CLI Only Implemented in Script
- Not fully packaged as installable module (pip)
- However, script exposes CLI arguments (`--input_csv`, `--output_dir`) with help flags

Conclusion

The solution delivers a clean, efficient, and extensible preprocessing pipeline fully aligned with the India Speaks data standards. It demonstrates robust text normalization, language detection, and audio validation while being adaptable for future model training pipelines.

```
Clean rows: 1886
Rejected rows: 114
Sample rejections:
  utterance_id                reason
0     utt_0019  MissingTranscription
1     utt_0030  MissingTranscription
2     utt_0072  MissingTranscription
3     utt_0093  MissingTranscription
4     utt_0147  MissingTranscription
```