

# Weekly Updates 1





April 14<sup>th</sup> – 20<sup>th</sup>

**Cevin Samuel**

GCF Intern

# Recap

## Last week TODOS:

- ASR Data Processing Pipeline (in progress ) -> some FINDINGS in next slide.
- Set up data pipelines (Python, HuggingFace):
  - Connect or load dataset, `espnet/yodas` (done ) .
  - Normalize transcripts text data (hasn't started ):
    - How to handle (?):
      - S Substitutions.
      - I Insertions.
      - D Deletions.
  - Connect to Whisper model for transcription (done ) .

# Data Audit `espnet/yodas/Indonesian` [shard id000 (1/3)]

URL: <https://huggingface.co/datasets/espnet/yodas/tree/main/data/id000>

- Some audios are in wrong code (not even code-mixed) BUT THEY HAVE CORRECT INDONESIAN TRANSCRIPT:
  - 310+ utters are non-speech (e.g., only music, noise, etc): YlwOo8o1Cq4, YXht-gEBSBI, YxtjikMAYH0.
  - 761 utters in 100% Russian: 1OI6crYhlvA.
  - 541 utters in 100% Japanese: Y3sWTnNpgl4.
  - 185+ utters in 100% English: 28ZAZ7a\_3Tk, YL1bylpzYKk, YW4oxbMnBxA.
  - 120 utters in 100% Malay.
  - 50+ utters in 100 Javanese.
  - 50+ utters in 100% Sundanese: YiL\_ACfNO58.
- TOTAL:
  - 2017+ bad utters out of 5940  $\approx$  34% (i.e. 66% is usable).

**This may indicate that, across the whole `espnet/yodas` multilingual dataset have this kind of data (around 2/3<sup>rd</sup> is usable).**

**This CLEARLY indicates that, the team behind `yodas` dataset retrieved videos by looking at their subtitle, not their actual audio.**