Norwegian Government Press Conference Speech Corpus

About the corpus

The Norwegian Government Press Conference Speech Corpus (NorGovPCC) consists of more than 138 hours of speech generated from audio with aligned subtitles from press conferences hosted by the Norwegian Government Security and Service Organisation (DSS). The dataset consists of speech segments up to 30 seconds in length with corresponding transcriptions. The transcriptions are extracted from the official subtitles in either Norwegian Bokmål or Norwegian Nynorsk. NorGovPCC is distributed with the NLOD-2.0 open data license.

Point of contact

The Norwegian Language Bank

Languages

The transcriptions in the dataset are in Norwegian Bokmål (nob) or Norwegian Nynorsk (nno).

Dataset Structure

```
dataset_name/
├─ metadata.csv
├─ data/
│  ├─ train/
│  │  ├─ train1.mp3
│  │  ├─ train2.mp3
│  ├─ test/
│  │  ├─ test1.mp3
│  │  ├─ test2.mp3
```

The metadata.csv file describes each audio file in the train and test splits.

This dataset can be loaded as an audio dataset with the datasets library (at the time of writing version V2.19), see the documentation.

Data instances and data fields

This dataset consists of short .mp3 files and their transcriptions. The metadata.csv file contains a row for each .mp3 file with the transcription and other metadata.

Columns in metadata.csv:

- file_name (str): relative path to the .mp3 file
- transcription (str): transcription, the words uttered in the .mp3 file
- duration (float): length of .mp3 file in seconds
- transcription_language (str): nno or nob
- detected_spoken_language (str): no
- language_identification_model (str): model used to detect spoken language
- language_identification_model_certainty (float): model certainty

Data Splits

The data is split into test and train splits of approximately 10% and 90% of the total audio duration. The distribution of nynorsk and bokmål transcriptions in each split is similar to that of the total dataset.

Dataset Statistics

Train

- Number of segments: 17912
- Total duration in hours: 124.95
- Transcript language distribution:
  - **nno**: 7277 segments (40.63%)
  - **nob**: 10441 segments (58.29%)

Test

- Number of segments: 1961
- Total duration in hours: 13.59
- Transcript language distribution:
  - **nno**: 803 segments (40.95%)
  - **nob**: 1158 segments (59.05%)

Dataset Creation

Curation Rationale

The motivation to construct this dataset is the continued need for Norwegian speech datasets with Norwegian transcriptions to develop Norwegian ASR models.

Data Collection and Processing

The dataset was collected and processed using the following steps:

- **Gathering raw audio data:**
  - **Scraping data:** Videos and corresponding subtitles were scraped from webcasts hosted by the Norwegian Government in the period 2022-05-05 – 2024-04-12: Webcast - regjeringen.no.
  - **Extracting audio:** Audio was extracted from the video data using FFmpeg version 4.4.2 [1]
- **Generating base data set:**
  - **Segment alignment:** Subtitle segments were aligned at sentence level using WhisperX [2, 3] and the Wav2Vec2 [4] alignment models NbAiLab/nb-wav2vec2-1b-bokmaal NbAiLab/nb-wav2vec2-1b-nynorsk [5] for Norwegian Bokmål and Norwegian Nynorsk, respectively.
    * We observed that increasing the segment length improved the performance of the Wav2Vec2 model, but also markedly increased the time required by WhisperX to perform the alignment. Therefore, to balance accuarcy and computational time, subtitle segments from scraped VTT files were first combined to be as short as possible while being over 30 seconds and shorter than 120 seconds. If the shortest duration greater than 30 seconds is greater than 120 seconds, then joined segments are kept shorter than 30 seconds. Subtitle segments longer than 120 seconds are discarded.

        * Next, audio- and text segments are aligned using the whisperx.alignment-module.
    - **Language identification:** Each sentence-segment was run through the multilingual Whisper large-v3 model [6] for language indentification The language tokens from the multilingual Whisper model were used to predict the spoken language of each sentence. However, the Whisper large v3 model requires audio segments of exactly 30 seconds in duration and most sentences last for a shorter amount of time than that. A rolling-window approach was chosen to circumvent this issue, in which each sentence started at the start-time of the sentence and ended 30 seconds after the start, independent of the actual duration of the sentence. The final sentence(s) for each video file were padded with zero-values to obtain 30-second audio sequences.
        * ***Note:*** We have observed that automatic language identification will sometimes incorrectly classify English audio clips as Norwegian if the speaker has a strong Norwegian accent. Consequently, the NorGovPCC may contain a small number of clips with English speech and Norwegian transcriptions.
- **Assembling NorGovPCC**
    - The sentence alignment and language identification resulted in a sentence-level base dataset. To construct the NorGovPCC, sentences spoken in languages other than Norwegian (as predicted by the Whisper model) were removed and subsequent sentences were combined into multi-sentence segments that lasted for at most 30 seconds.

Source Language Producers

The NorGovPCC mainly consists of speech from Norwegian politicians and people with high ranking roles in the Norwegian public sector. The transcriptions are from the official subtitles made by DSS.

Personal and Sensitive Information

Both the audio data and subtitles are openly distributed by DSS and are assumed not to include any information exempt from public disclosure.

Additional Information

Dataset Curators

- Tita Enstad (National Library of Norway)
- Marie Roald (National Library of Norway)

Licensing Information

Norwegian Licence for Open Government Data (NLOD) 2.0

References

[1] FFMPEG [computer software]. Version 4.4.2; 2024. Available from: http://ffmpeg.org/

[2] Bain, M, Huh, J, Han, T, Zisserman, A. “WhisperX: Time-Accurate Speech Transcription of Long-Form Audio”. INTERSPEECH 2023 2023.

[3] WhisperX [computer software]. Revision `f2da2f8`; Available from: https://github.com/m-bain/whisperX

[4] Baevski, A, Zhou, Y, Mohamed, A, Auli, M. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. Advances in neural information processing systems 2020; 33:12449–12460.

[5] De La Rosa, J, Braaten, RA, Kummervold, P, Wetjen, FBoosting Norwegian Automatic Speech Recognition. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa) 2023 (pp. 555–564). University of Tartu Library.

[6] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. InProceedings of the International Conference on Machine Learning 2023 Jul 3 (pp. 28492-28518). PMLR.