Augmented datasets review template

Dataset list

Dataset	Paper	Link
VoxPopuli	VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation - ACL Anthology	facebookresearch/voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation (github.com)

Motivation

What is the speech dataset name, and does the name accurately describe the contents of the dataset?

'VoxPopuli'. The dataset contains voice recordings (transcribed and untranscribed) in multiple languages, all taken from European Parliament recordings. 'Vox' (meaning 'voice') and 'Populi' (meaning 'voice of the people') is usually used to refer to the opinion of a 'person on the street' in English. The name does not very accurately describe the content of the dataset, but does imply it is related to voice.

Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?

In my understanding it should be possible to draw conclusions about both, given that there is both spontaneous and read speech in the dataset, but I cannot evaluate if there is sufficient quality and volume of content to draw conclusions.

Describe the process used to determine which linguistic subpopulations are the focus of the dataset.

The process is not described (though the linguistic subpopulations are listed), the creators of the dataset should have been able to describe the process. The selection of the languages represented appears to be based on the motivation (creation of dataset in multiple languages) and availability of the EU data.

Composition

How many hours of speech were collected in total (of each type, if appropriate), including speech that is not in the dataset? If there was a difference between collected and included, why? E.g., if the speech data are from an interview and the dataset contains only the interviewee's responses, how many hours of speech were collected in interviews from both interviewer and interviewee?

400K hours of unlabelled speech data for 23 languages
1.8K hours of transcribed speech data for 16 languages
17.3K hours of speech-to-speech interpretation data for 15x15 directions
29 hours of transcribed speech data of non-native English intended for research in ASR for accented speech (15 L2 accents)

It is not clear if there was data that was collected but not used, since the data is taken from European Parliament event recordings rather than being specifically collected to create this dataset. The page states that “The raw data is collected from 2009-2020 European Parliament event recordings”, but it is not clear if all voice recordings from all events were included or not.

How many hours of speech and number of tokens are in the dataset (by each type, if appropriate)?

400K hours of unlabelled speech data for 23 languages
1.8K hours of transcribed speech data for 16 languages
17.3K hours of speech-to-speech interpretation data for 15x15 directions
29 hours of transcribed speech data of non-native English intended for research in ASR for accented speech (15 L2 accents)

Unlabelled and transcribed data = 15m tokens, 467m LM tokens. Tokens for speech-to-speech interpretation data and accented speech transcribed data not listed. I am not sure if the creators of the dataset would have been able to provide this information or not.

Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?

Languages, and language pairs in the case of speech-to-speech interpretation data are listed. For accented English speech data, country names are provided for the accents.

For any linguistic subpopulations identified in the dataset, please provide a description of their respective distributions within the dataset.

Unlabelled and transcribed data (LM tokens)
English
60.1M
German
50.0M
French
58.6M
Spanish
57.4M
Polish
13.6M
Italian
52.1M
Romanian
10.3M
Hungarian
13.0M
Czech
13.5M
Dutch
54.6M
Finnish
34.5M
Croatian
285K
Slovak
13.3M
Slovene
12.6M
Estonian
11.3M
Lithuanian
11.5M

Speech-to-speech interpretation data (total target for source, unit not specified)

En
6.0K
De
2.8K
Fr
2.3K
Es

1.6K
Pl
775
It
961
Ro
688
Hu
378
Cs
434
Nl
401
Fi
182
Hr
384
Sk
239
Sl
68
Lt
13

Accented speech transcribed data (transcribed hours per accent)
Dutch
3.52
German
3.52
Czech
3.30
Polish
3.23
French
2.56
Hungarian
2.33
Finnish
2.18
Romanian
1.85
Slovak
1.46
Spanish
1.42
Italian

1.11
Estonian
1.08
Lithuanian
0.65
Croatian
0.42
Slovene
0.25

How much of the speech data have corresponding transcriptions in the dataset?

1.8K hours of transcribed speech data for 16 languages + 29 hours of transcribed speech data of non-native English intended for research in ASR for accented speech

Does the dataset contain non-speech mediums (e.g. images or video)?

No

Is it possible for one speaker to code switch or speak multiple languages, and if so, how is this identified in the data?

This is unclear, the dataset creators should have been able to answer this question. For interpreted speech, source and target are tagged

Does the speech dataset focus on a specific topic or set of topics?

No but all taken from European Parliament recordings, which would probably limit the topics

Does the dataset include sensitive content that can induce different emotions (e.g., anger, sadness) that can cause the speakers to produce unusual pitch or tone deviating from plain speech?

Unknown, creators of the dataset would probably also not be able to answer as they probably did not manually review all content

Does the dataset contain content that complies to the users' needs, or does it result in symbolic violence? E.g. does the speech reproduce values that contradict the values of the lingual groups existing in the dataset?

Unknown, cannot say if creators would have known or not

Collection Process

What mechanisms or procedures were used to collect the speech data, e.g.: is the data a new recording of read speech or an interview? Or is it downloaded speech data from public speeches, lectures, YouTube videos or movies, etc.?

Downloaded from recordings of public meetings

Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?

Unknown, creators of dataset are also unlikely to know

Is there presence of background noise?

Not clear from dataset description or paper

For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are ``fair and neutral"?

Not applicable

If speech data are collected from consenting individuals, do the metadata include all attributes that they have consented to disclosing? E.g., does the metadata include sensitive personal information such as impaired speech due to disabilities impacting speech patterns?

Not possible to answer as consent process is not discussed

Preprocessing/cleaning/labeling

When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?

Not discussed in paper or dataset webpage, only mention silence removal as part of audio processing. Creators of the dataset should have been able to provide this information.

Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?

Used human transcribers for 400 hours of English target speech, training or their familiarity with corpus materials, vocabulary, and linguistic characteristics not discussed. Creators of the dataset should have been able to provide this information.

If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?

Only one method of transcription appears to have been used for human transcribers.

If the speech data include transcriptions, what software was used to generate the transcriptions (including, e.g., software used by human transcribers)? Are timestamps included in transcriptions? Are the alignments provided with the transcripts?

“The VoxPopuli transcribed set comes from aligning the full-event source speech audio with the transcripts for plenary sessions. Official timestamps are available for locating speeches by speaker in the full session, but they are frequently inaccurate, resulting in truncation of the speech or mixture of fragments from the preceding or the succeeding speeches. To calibrate the original timestamps, we perform speaker diarization (SD) on the full session audio using pyannote.audio (Bredin et al., 2020) and adopt the nearest SD timestamps (by L1 distance to the original ones) instead for segmentation.”

Were transcription conventions (such as tagging scheme, treatment of hate speech or swear words, etc.) disclosed along with the corpus?

Not disclosed, creators of the dataset would not have been able to disclose transcription conventions for the original recordings, but should have been able to provide for the 400 hours that was conducted specifically for the dataset.

Extra coding done? (coding is a development from transcription)

As far as I can understand, no.

Uses / Distribution / Maintenance

How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?

Not clear how/if redactions are performed (either for original EU parliamentary recordings or in the audio taken for this dataset), not clear what is done in audio or transcription for PII or sensitive information.

Is there any part of this dataset that is privately held but can be requested for research purposes?

Does not appear to be as they stated they used all audio except Irish language data(due to small quantity of data)

Is there a sample dataset distributed? If so, how well does the sample represent the actual dataset? Do they include all forms of speech included in the dataset? How big is the sample?

Does not appear to be a sample available

Does a documentation exist, where the composition of the dataset as well as details of how the data was collected were laid out and justified?

Composition is laid out on the Github page, collection methodology can be found in the paper. No justification for why the methodology (taking from existing recordings) was used.