DATASHEET:
Multilingual Spoken Words Corpus

We, the authors of this work, will bear all responsibility in case of a violation of rights and we confirm the license for our data is CC-BY 4.0

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The Multilingual Spoken Words Corpus was created to provide word length speech data in numerous languages to be used to study and improve speech recognition systems. Existing word length datasets are limited in size and do not support multiple languages, which has limited the scope of keyword spotting research.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by a team of researchers from Harvard University's Edge Computing Lab, Coqui.ai, Google and MLCommons.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
The creation of this dataset was sponsored in part by the SRC consortium. The dataset will be hosted by MLCommons.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
The dataset comprises of anonymized audio clips submitted by people in the Common Voice Database.

How many instances are there in total (of each type, if appropriate)?
Total number of instances are 23.4 Million audio clips

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset is our best effort to extract and represent as much diversity (in terms of various different languages) from Common Voice as possible. In particular, it comprises of 50 different languages.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance is a 1 second opus audio clip of a particular word in a particular language, organized by filename and parent directory in a compressed tarball.

Is there a label or target associated with each instance? If so, please provide a description.
The label or target associated with each instance is the spoken word in the filename and the language ISO code in the parent directory.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
We have not removed any information manually.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Yes, extractions corresponding to the same keyword are stored and managed together. Associated data includes forced alignments as Praat textgrids and text files containing data splits, all linked through unique filenames and language ISO codes.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
Yes, data splits are provided and a detailed description is provided in Section 4 of the paper.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Sources of noise in the dataset we derive from exist. These are natural sources of noise encountered while users submit recordings on the Common Voice database collection website, which we derive our data from. More explanation is provided in Section 3 of the text. We provide a metric to assess potential errors caused by forced alignment, transcription mismatches, mispronunciations, and saturating background noises in Section 4 of the paper.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is fully self-contained (though derived from Common Voice, an external, maintained source with a CC License)

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No, the dataset is not confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No, our source of audio is Common Voice, which is designed to avoid offensive user-submitted audio content by using an approved list of text sources, and a crowdsourced validation and reporting process for each audio clip.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
Our dataset identifies subpopulations in two forms: gender and language. The source of our data, Common Voice, contains additional optionally volunteered metadata that can be associated externally with our data.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No, it is not possible to identify individuals directly or indirectly from the dataset, which Common Voice takes steps to prevent.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
To our best efforts we have ensured that our dataset does not contain any sensitive information. We do not include or reconstruct any potentially sensitive demographic information in our dataset, and we note that Common Voice also provides their own safeguards and user agreements in an effort to avoid collecting or exposing sensitive information for those who have donated their voices, and Common Voice also provides a publicly-reviewable whitelist of text sources which are restricted to public domain data.

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data was recorded and validated by volunteers and submitted to the Common Voice Dataset. We then used forced alignment to extract audio clips of individual words from the dataset. The alignments are not validated, but word error rates are estimated based on random sampling of English data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
Most of the data is sourced from Common Voice version 3 released in June 2020 but we added some of the new languages added in Common Voice version 7 released in July 2021.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
The original audio was recorded by volunteers with a variety of microphones on the Common Voice web interface. The

data was validated by volunteers on the same platform. We then performed forced alignment with the Montreal forced alignment tool.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint.
Running forced alignment and word extraction is a time consuming and expensive process. We estimate that it cost approximately $2000 USD to create the dataset. We used three compute engines on Google Cloud Platform and a 64 core CPU to obtain alignments, extract words, and run data analysis.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
This dataset was generated from Common Voice. The process is described in the paper. In short, we extracted individual words with at least 3 characters from sentence length data.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data was donated and validated by unpaid volunteers.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No, there was no formal ethical review process.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Yes the dataset relates to people and speech.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data was originally collected by people for the Common Voice Dataset but was downloaded via commonvoice.mozilla.org and then word aligned for the Multilingual Spoken Words Corpus.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
The original speech was donated on commonvoice.mozilla.org and the individuals are notified that their contribution will be included in a public dataset. The individuals were not notified about the creation of this dataset but the usage of this dataset does not differ from Common Voice.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Yes the individuals consented to donate their voice to be included in Common Voice. The website commonvoice.mozilla.org shows how voice donation is presented.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
We will stay up to date with the current iteration of Common Voice including the removal of data that is no longer present in Common Voice. MLCommons (the entity that will host the dataset) will support requests for the removal of data based on the session ID provided in Common Voice.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
A discussion of the datasets impact is included in the Multilingual Spoken Words Corpus paper.

Any other comments?
All voice clips in the dataset are scrubbed of personally identifying information by Common Voice. Please see `commonvoice.mozilla.org/en/faq` for more information on how the data was collected.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
The Multilingual Spoken Words Corpus is produced by extracting individual words using alignments generated from the Montreal Forced Aligner. We select keywords with a minumum character length of three. Features are extracted using the TensorFlow MicroFrontend for spectrograms, and an embedding model provided in our open-source repository for a 1024-dimensional feature vector. We recommend an optional filtering method: filtering based

on nearest-neighbor outlier detection, discussed in Section 4.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
The unextracted data is available in the Common Voice Dataset. The data is provided without applying the optional filtering methods described.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Yes the Montreal Forced Aligner is used to generate word alignments. `https://montreal-forced-aligner.readthedocs.io/en/latest/`

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Yes the dataset was used to create the multilingual speech embedding described in Sections 4 and 5 of the paper and a radio monitoring application in Section 6 in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
We do not actively track the papers and systems that use our dataset as of writing.

What (other) tasks could the dataset be used for?
The dataset can be used for a variety of speech related applications, like keyword spotting systems. The paper describes the applications of the dataset in more detail.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
The dataset contains some noisy, unintelligible, and mislabeled data. The data was extracted using alignments generated by Montreal Forced Aligner. This has the potential to introduce some bias into the data as the aligner may produce worse alignments for specific groups. Furthermore, the optional filtering techniques described in the paper have the potential to introduce additional bias to the dataset and should only be used with an understanding of this potential issue.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The user of the dataset should not attempt to identify any individual who contributed their voice.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes the dataset will be publically available under CC-BY 4.0.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
That dataset will be distributed in the form of tarballs made available on a website. The dataset does not yet have a DOI, but we will add a DOI when we publicly release the dataset.

When will the dataset be distributed?
The dataset will be publicly distributed for usage through MLCommons.org after the review period.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The Dataset is distributed under CC-BY 4.0 License. This allows usage for academic research and commercial uses.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No third parties have imposed IP-based or any other restrictions on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No export controls or other regulatory restrictions apply to the dataset or to any of its individual instances.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is supported, hosted and will be maintained by MLCommons

How can the owner/curator/manager of the dataset

be contacted (e.g., email address)?
Yes the hosting organization (MLCommons) can be contacted at datasets@mlcommons.org.

Is there an erratum? If so, please provide a link or other access point.
No - however, planned future releases will contain fixes to any errors reported or found.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
Yes, the dataset will be constantly updated to add new languages and correct any errors present at the time of submission and unknown to the authors.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
We will follow Common Voice's application of data retention policies.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Yes, all versions of the dataset will be receive continuous support, hosting and maintenance services by MLCommons

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
The Multilingual Spoken Words Corpus is available under the CC-BY 4.0 license which allows adaptation with attribution. To contribute to the dataset you can donate your voice to the Common Voice Dataset (https://commonvoice.mozilla.org)

Any other comments?
We plan to extend the dataset to include data sourced from other sentence length corpora. This process will be documented and this datasheet will be updated to reflect the current state of the Multilingual Spoken Words Corpus.