DATASHEET:
Lelapa-X-ASR (Afrikaans)

Lelapa AI
info@lelapa.ai

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most updated version here.

MOTIVATION

For what purpose was the dataset created?
The dataset was created to enable research on Automatic Speech Recognition (ASR) for Afrikaans i.e. given an audio file containing Afrikaans speech, transcribe the speech into the language specific text. The dataset was created intentionally with that task in mind, focusing on South Africa call centre audio where ASR is often required.

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 Lelapa AI.

What support was needed to make this dataset?
Lelapa AI.

Any other comments?
None.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
The instances are audio files taken from a combination of open-source and closed synthetic resources, together with the corresponding transcription of the audio file.

How many instances are there in total (of each type, if appropriate)?
93723 audio files

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
The dataset contains all possible instances and is fully representative, having been collected with the widest range of demographic elements and the most equal split between gender distribution as possible, as validated by the relevant sources.

What data does each instance consist of?
Each instance consists of an audio segment and the corresponding transcription of the audio file. The data was collected, annotated and curated with natural language processing best practice in mind.

Is there a label or target associated with each instance?
No.

Is any information missing from individual instances?
Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
Yes, the speaker ID of each audio segment is explicitly noted.

Are there recommended data splits (e.g., training, development/validation, testing)?
The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in transcription Word Error Rate (WER).

Are there any errors, sources of noise, or redundancies in the dataset?
There are some natural instances of noise in the audio files which is prevalent in the call centre domain. There are certain redundancies where transcripts are the same for different audio files by different speakers

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
The dataset is entirely self-contained.

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)?
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
No.

Does the dataset relate to people?
Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)?
Yes. The sub-populations are identified by gender (male or female) with a 50.55% and 49.45% distribution respectively, and by age-group (18-29 = 55.54%, 30-49 = 30.82%, >49 = 12.29%, Unknown = 1.35%)

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?
No.

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)?
No.

Any other comments?
None.

COLLECTION

How was the data associated with each instance acquired?
The data was acquired from a combination of open-source and closed resources.

Over what timeframe was the data collected?
Unknown to the authors of the datasheet.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
The audio files were collected using electronic hardware such as laptops, smartphones and desktop computers. The transcripts were scripted or obtained from manual human annotation.

What was the resource cost of collecting the data?
Unknown to the authors of the datasheet.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
N/A

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)?
Unknown to the authors of the datasheet

Were any ethical review processes conducted (e.g., by an institutional review board)?
Unknown to the authors of the datasheet.

Does the dataset relate to people?
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Via third parties.

Were the individuals in question notified about the data collection?
Yes the individuals created the synthetic data specifically for the ASR use case.

Did the individuals in question consent to the collection and use of their data?
Yes. The synthetic data was created by the individuals specifically for the ASR use case, therefore the authors were explicitly informed that the recordings would be used in this way.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
Unknown to the authors of the datasheet.

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?
N/A.

Any other comments?
None.

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)?
The speech tags in the transcriptions were normalised to ensure consistency across combined datasets.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
Yes.

Is the software used to preprocess/clean/label the instances available?

Yes.

Any other comments?
None.

USES

Has the dataset been used for any tasks already?
Yes, for research on Automatic Speech Recognition (ASR) for Afrikaans.

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.
N/A.

What (other) tasks could the dataset be used for?
The dataset could be used for anything related to modeling or understanding Afrikaans speech.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
No.

Are there tasks for which the dataset should not be used?
No.

Any other comments?
None.

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?
No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
N/A.

When will the dataset be distributed?
N/A.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
N/A.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
Yes. The proprietary/ closed data is under a Licence Agreement between Lelapa AI and the third party.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
No.

Any other comments?
None.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
Lelapa AI.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Email: info@lelapa.ai

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
Yes. As required for company research and commercial purposes.

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)?
N/A.

Will older versions of the dataset continue to be supported/hosted/maintained?
Older versions will be kept around for consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
No.

Any other comments?
None.

REFERENCES

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.