Data Sheet for Multilingual Speech Intent Recognition Dataset for Home Automation: Luganda and Runyankore

We present the Multilingual Speech Intent Recognition Dataset for Luganda and Runyankore data sheet created by researchers at Marconi Research and Innovation Lab in Makerere University in Uganda. We follow the datasheet for the dataset framework created by (Gebru et al. 2021).

Motivation	
For what purpose was the dataset created?	The dataset was created to enhance multilingual speech intent recognition, specifically for voice control and smart home applications in low-resource languages, namely Luganda and Runyankore.
Was there a specific task in mind?	The specific task was to develop speech intent recognition systems that can understand, and process voice commands in Luganda and Runyankore for IoT applications.
Who created the dataset?	The dataset was created by researchers at the Marconi Research and Innovation Lab at Makerere University, including John Trevor Kasule, Sudi Murindanyi, and Andrew Katumba, with the support of volunteers.
Who funded the creation of the dataset?	The creation of the dataset was funded by the Marconi Research and Innovations Laboratory at Makerere University.

Composition	
What do the instances that comprise the dataset represent?	The instances represent audio recordings of voice commands associated with specific intents for controlling smart home devices.
How many instances are there in total (of each type, if appropriate)?	1. **Luganda:** 10,200 audio files from 81 speakers(43 Male, 38 Female)

	2. **Runyankore:** 6,752 audio files from 52 speakers (28 Male, 24 Female)
Does the dataset contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set?	The dataset is not a sample from a larger set; it comprises all collected instances for the defined intents.
What data does each instance consist of? “Raw” data or features?	Each instance consists of raw audio data recorded as voice commands
Is there a label or target associated with each instance? If so, please provide a description.	Yes, each instance is within a folder labelled with an intent that describes the action corresponding to the voice command (e.g., "turn on lights").
Is any information missing from individual instances?	No information is missing from individual instances.
Are relationships between individual instances made explicit?	No explicit relationships between individual instances are defined.
Are there recommended data splits (for example, training, development/validation, testing)?	Yes, recommended splits include: • Luganda: Training (8,369), Testing (1,068), Validation (942), which corresponds to 80% training, 10% testing and 10% validation. • Runyankore: Training (5,628), Testing (612), Validation (512), which corresponds to 80% training, 10% testing and 10% validation. Which is also represented in the shared dataset.
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.	None
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	The dataset is self-contained and does not rely on external resources.
Does the dataset contain data that might be considered confidential?	No

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

Collection Process	
How was the data associated with each instance acquired?	The data was acquired through recordings of voice commands spoken by participants using different smartphones.
What mechanisms or procedures were used to collect the data?	Volunteers recorded their voices following a predefined list of commands related to smart home actions.
What was the sampling strategy if the dataset is a sample from a larger set?	The dataset is not from a larger set.
Who was involved in the data collection process?	A group of University students aged between 18 and 25 volunteered to participate in data collection.
Over what timeframe was the data collected?	The dataset was collected over a period of one year.
Were any ethical review processes conducted (for example, by an institutional review board)?	No

Preprocessing, cleaning, and labelling	
Was any preprocessing/cleaning/labelling of the data done (for example, discretisation or bucketing, tokenisation, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	Yes, audio files with high noise levels were cleaned using the Audacity Software Tool. The audio files were converted to .wav format at a sampling rate of 16KHz.
Was the “raw” data saved besides the preprocessed/cleaned/labelled data (for example, to support unanticipated future uses)? If so, please provide a link or other	No

access point to the “raw” data.	
Is the software used to preprocess/clean/label the data available? If so, please provide a link or other access point.	Audacity Software Tool

Uses	
Has the dataset been used for any tasks already? If so, please provide a description.	We have used the dataset to develop speech intent classification models and deployed them on resource-constrained IoT devices to demonstrate home automation.
Is there a repository that links to any or all papers or systems that use the dataset?	Yes, one conference paper was written and presented in AfricaNLP 2024. Here is the link “https://openreview.net/forum?id=GBq2FsT2Us”
What (other) tasks could the dataset be used for?	The dataset can be used to improve voice control, develop Spoken Language Understanding models, enhance Speech Intent Recognition models, serve as a baseline for collecting similar data, etc.
Is there anything about the dataset's composition or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?	No

Distribution	
Will the dataset be distributed to third parties outside of the entity (for example, a company, institution, or organisation) on behalf of which the dataset was created? If so, please provide a description.	Yes, the dataset will be made publicly available.
How will the dataset be distributed (for	The dataset and associated metadata are

example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?	stored on the Harvard Dataverse, an open-source data repository. The dataset is assigned a Digital Object Identifier (DOI): ***https://doi.org/10.7910/DVN/JWQAZZ***
When will the dataset be distributed?	The dataset is available under the specified DOI.
Will the dataset be distributed under a copyright or other intellectual property (IP) license and/or under applicable terms of use (ToU)?	The dataset is licensed under the CC BY license, which allows users to share and adapt the dataset as long as they give credit to data set creators.
Have any third parties imposed IP-based or other restrictions on the data associated with the instances?	No
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	No

Maintenance	
Who will be supporting/hosting/maintaining the dataset?	The research team at the Marconi ab will support, host, and maintain the dataset. The dataset is hosted on the dataverse platform.
How can the owner/curator/ manager of the dataset be contacted (for example, email address)?	The dataset managers can be contacted via email. Trevor: kasulejohntrevor@gmail.com Sudi: murindanyi@gmail.com Andrew: andrew.katumba@mak.ac.ug
Is there an erratum?	No
Will the dataset be updated (for example, to correct labelling errors, add new instances, delete instances)?	All updates to the dataset will be documented and communicated through the Marconi Lab dataverse, as well as Hugging Face.
Will older versions of the dataset continue to be supported/hosted/ maintained? If so, please describe how.	Yes, dataset natively supports versioning. Copies of older versions will be stored locally on data storage servers in the Marcon lab and the Marconi Lab Hugging Face organisation.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	Interested researchers can email managers to discuss dataset extension and contribution.

Reference

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.