

Continuous Training: Data Processing Engine



MLOps: Continuous Training



The MLOps Personas



ML
Engineer



ML
Researcher



Data
Scientist



Data
Engineer



Software
Engineer

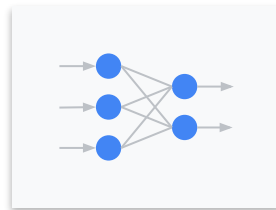
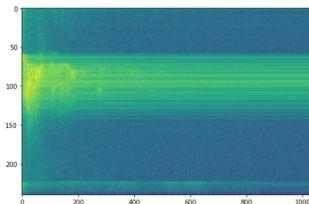
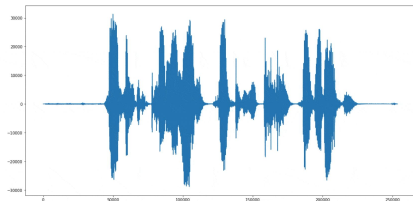


DevOps



Business
Analyst

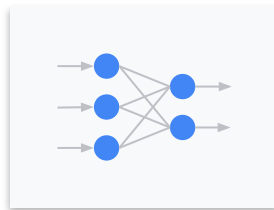
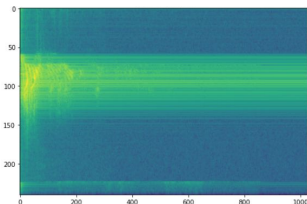
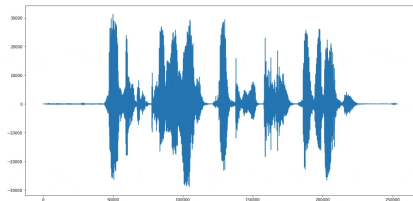
Keyword Spotting



“Yes” - 0.91
“No” - 0.09

- Voice assistants are **ubiquitous**

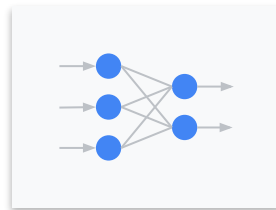
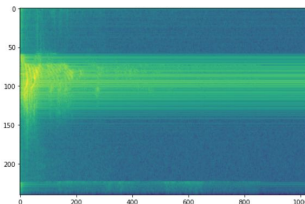
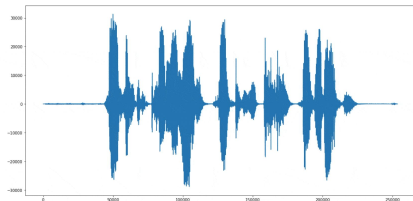
Keyword Spotting



“Yes” - 0.91
“No” - 0.09

- Voice assistants are **ubiquitous**
- **Limited** vocabulary and languages

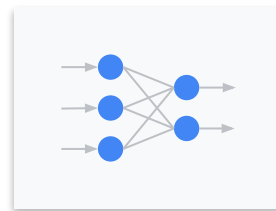
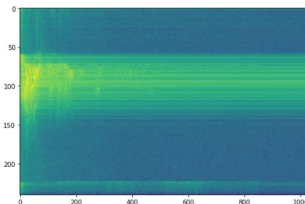
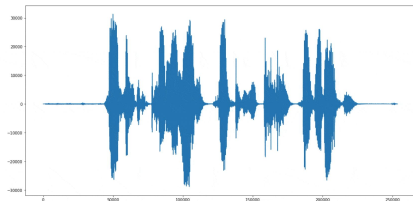
Keyword Spotting



“Yes” - 0.91
“No” - 0.09

- Voice assistants are **ubiquitous**
- **Limited** vocabulary and languages
- Datasets require **thousands** of training examples *per keyword*

Keyword Spotting



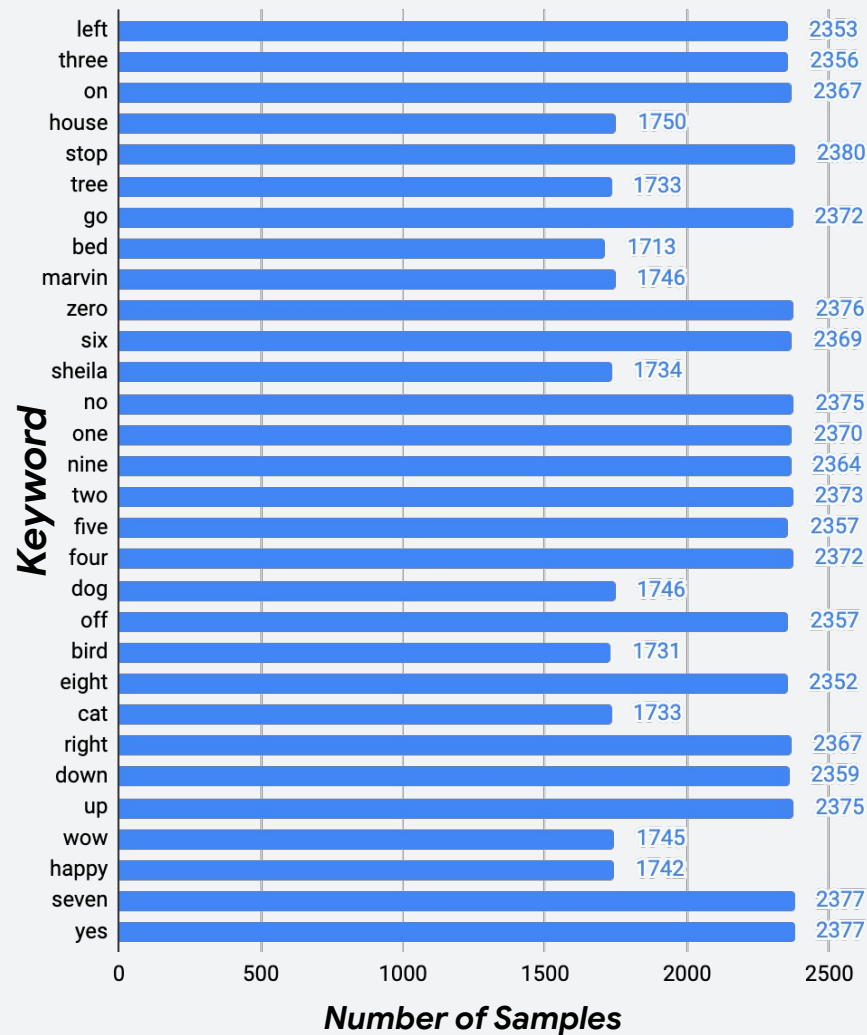
“Yes” - 0.91
“No” - 0.09

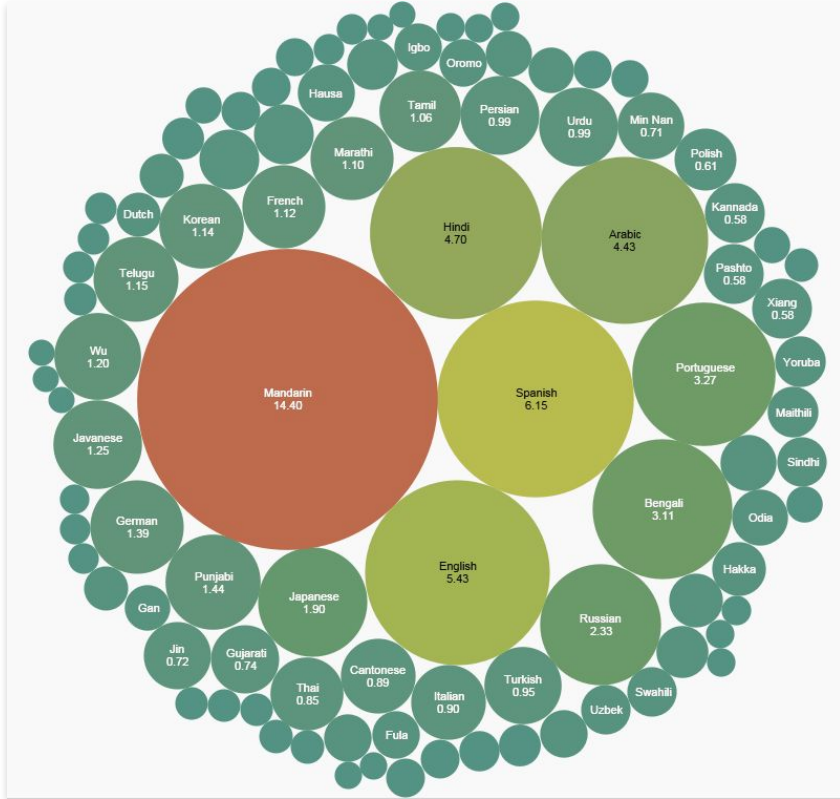
- Voice assistants are **ubiquitous**
- **Limited** vocabulary and languages
- Datasets require **thousands** of training examples *per keyword*

Goal: make voice interfaces available to a **worldwide** audience

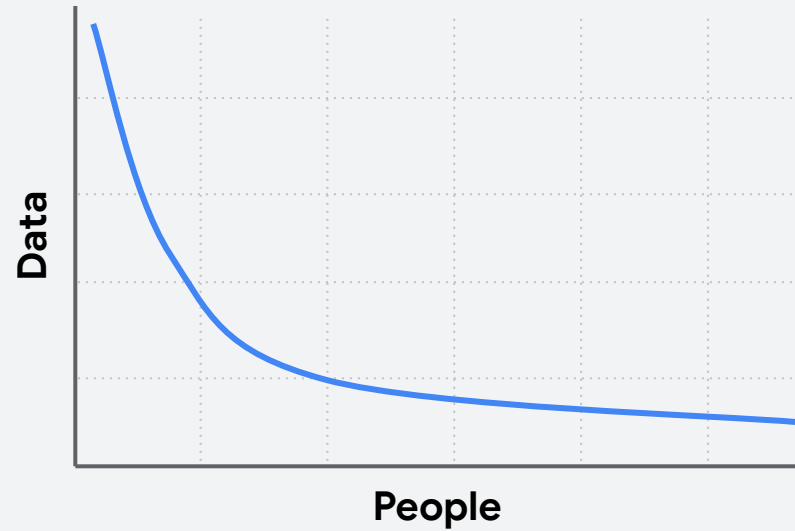
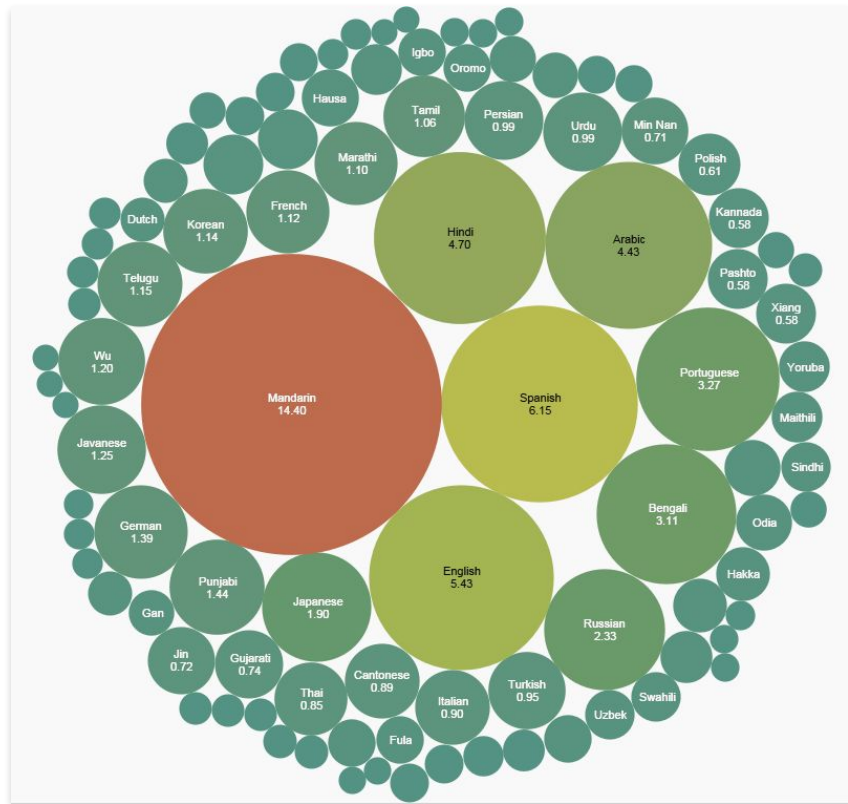
Google Speech Commands (GSC)

- **English** dataset
- **35** spoken words
- **105,000** audio samples
- **Individually recorded** words





- Speech commands for the **whole planet?**
- For **more than** just voice assistants



Data Engineering

Data Engineering

Requirements

- Problem definition
- Machine & human usable format
- Permissions & rights

Data Engineering

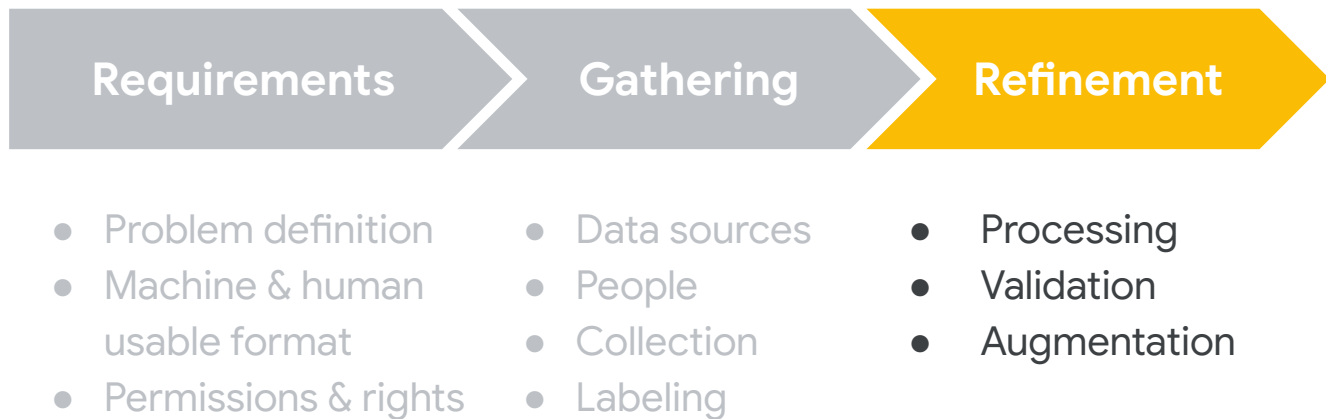
Requirements

- Problem definition
- Machine & human usable format
- Permissions & rights

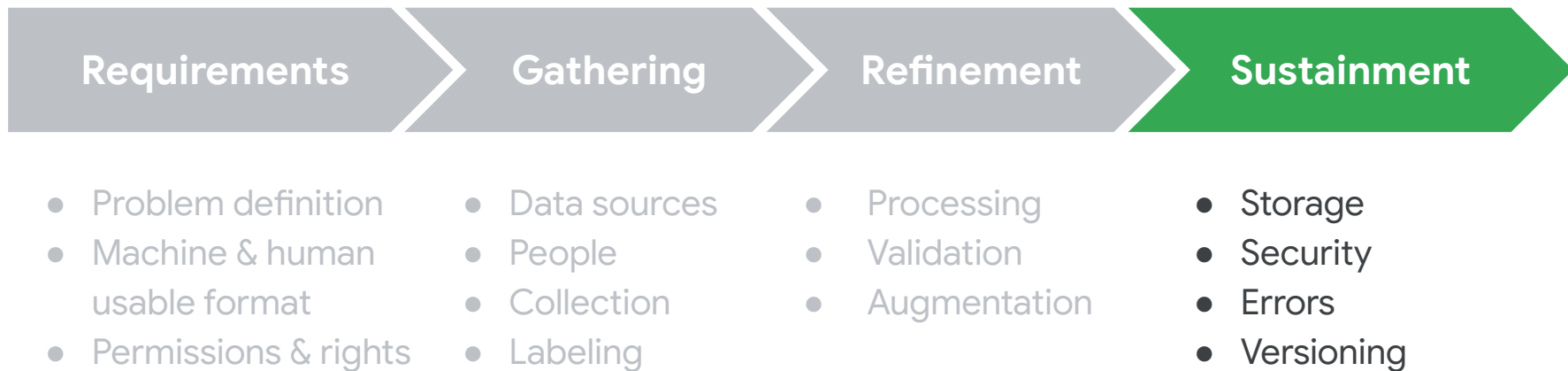
Gathering

- Data sources
- People
- Collection
- Labeling

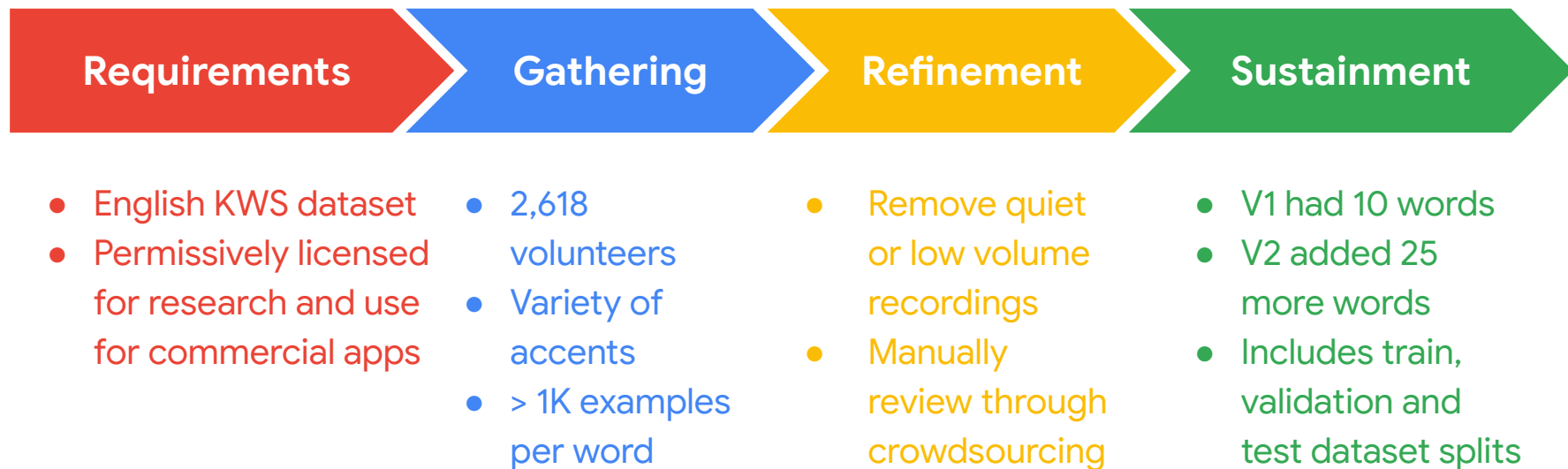
Data Engineering



Data Engineering

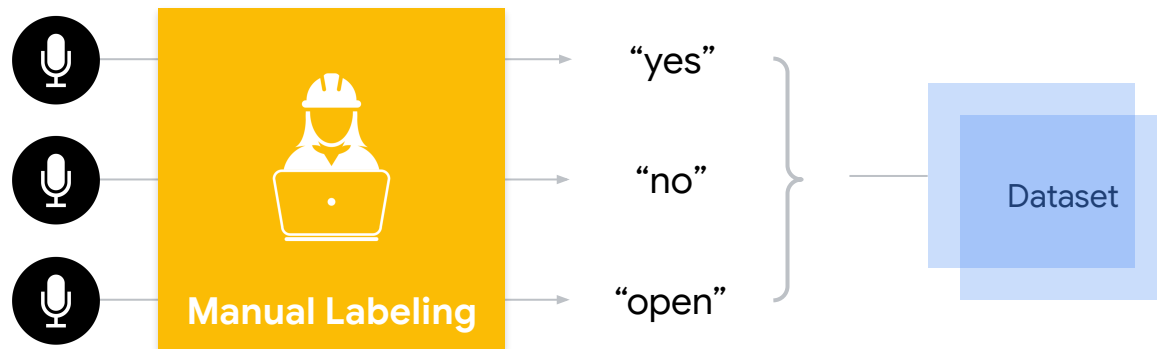


Data Engineering for Speech Commands

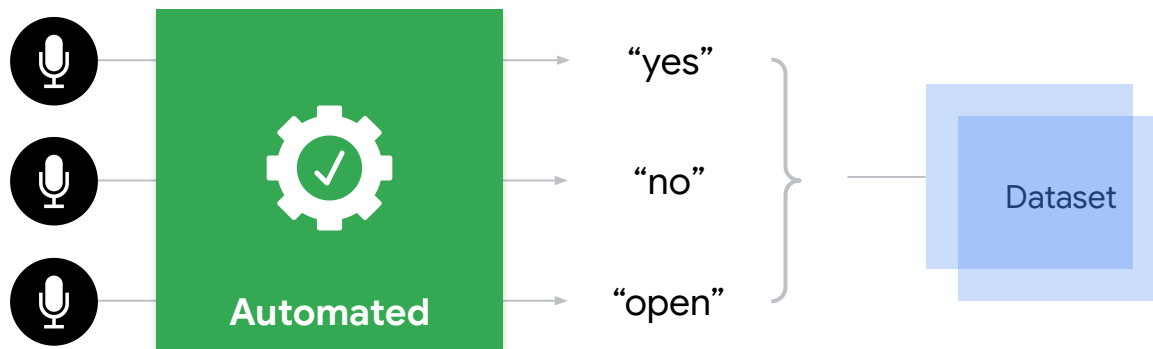


How do we **scale** past this?

Labeling is Expensive



How do we **automate**?



Multilingual Spoken Words Corpus

Spoken Words in **50 Languages**

Spoken Words in 50 Languages

HOURS

CATEGORIZATION

LANGUAGES

<10

Low-Resource

Arabic, Assamese, Breton, Chuvash, Chinese (zh-CN), Dhivehi, Frisian, Georgian, Guarani, Greek, Hakha Chin, Hausa, Interlingua, Irish, Latvian, Lithuanian, Maltese, Oriya, Romanian, Sakha, Slovenian, Slovak, Sursilvan, Tamil, Vallader, Vietnamese

Spoken Words in 50 Languages

HOURS	CATEGORIZATION	LANGUAGES
<10	Low-Resource	Arabic, Assamese, Breton, Chuvash, Chinese (zh-CN), Dhivehi, Frisian, Georgian, Guarani, Greek, Hakha Chin, Hausa, Interlingua, Irish, Latvian, Lithuanian, Maltese, Oriya, Romanian, Sakha, Slovenian, Slovak, Sursilvan, Tamil, Vallader, Vietnamese
10-100	Medium-Resource	Czech, Dutch, Estonian, Esperanto, Indonesian, Kyrgyz, Mongolian, Portuguese, Swedish, Tatar, Turkish, Ukrainian

Spoken Words in 50 Languages

HOURS	CATEGORIZATION	LANGUAGES
<10	Low-Resource	Arabic, Assamese, Breton, Chuvash, Chinese (zh-CN), Dhivehi, Frisian, Georgian, Guarani, Greek, Hakha Chin, Hausa, Interlingua, Irish, Latvian, Lithuanian, Maltese, Oriya, Romanian, Sakha, Slovenian, Slovak, Sursilvan, Tamil, Vallander, Vietnamese
10-100	Medium-Resource	Czech, Dutch, Estonian, Esperanto, Indonesian, Kyrgyz, Mongolian, Portuguese, Swedish, Tatar, Turkish, Ukrainian
>100	High-Resource	Basque, Catalan, English, French, German, Italian, Kinyarwanda, Persian, Polish, Russian, Spanish, Welsh

Spoken Words in 50 Languages



Keywords per language

LOW-RESOURCE

AVERAGE PER LANGUAGE

552

KEYWORDS

10,431

SAMPLES

MEDIUM-RESOURCE

AVERAGE PER LANGUAGE

4,054

KEYWORDS

109,597

SAMPLES

HIGH-RESOURCE

AVERAGE PER LANGUAGE

23,408

KEYWORDS

1.8M

SAMPLES

Multilingual Spoken Words

Multilingual Spoken Words Corpus is a large and growing audio dataset of spoken words in 50 languages for academic research and commercial applications in keyword spotting and spoken term search, licensed under CC-BY 4.0. The dataset contains more than 340,000 keywords, totaling 23.4 million 1-second spoken examples (over 6,000 hours). The dataset has many use cases, ranging from voice-enabled consumer devices to call center automation. We generate this dataset by applying forced alignment on crowd-sourced sentence-level audio to produce per-word timing estimates for extraction. All alignments are included in the dataset. Please see our paper for a detailed analysis of the contents of the data and methods for detecting potential outliers, along with baseline accuracy metrics on keyword spotting models trained from our dataset compared to models trained on a manually-recorded keyword dataset.

Read our full paper [here](#)

Join the MSWC mailing list [here](#)

Connect with other MSWC users on our [discord server](#)

Get started by trying out our introductory tutorial notebook [here on Google Colab](#)

Kind

Full dataset

LICENSE

CC-BY 4.0

AUDIO FORMAT

Opus

DESCRIPTION

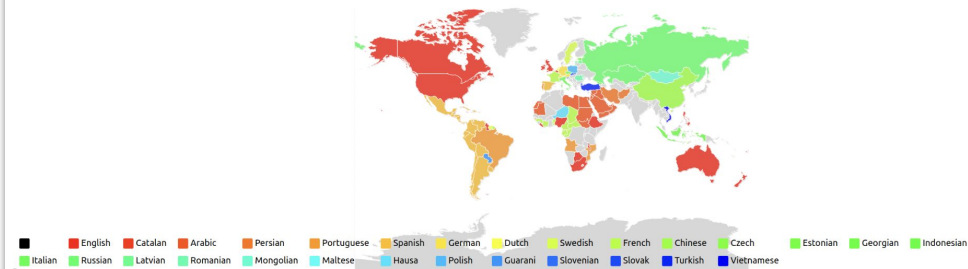
All 50 languages

Download

Google mirror

Primary languages in our dataset by country

This map depicts 28 primary languages which are included in our 50-language dataset, highlighted by country. Our dataset contains keywords in the following 50 languages: Arabic, Assamese, Basque, Breton, Catalan, Chinese, Chuvash, Czech, Dhivehi, Dutch, English, Esperanto, Estonian, French, Frisian, Georgian, German, Greek, Guarani, Hakha Chin, Hausa, Indonesian, Interlingua, Irish, Italian, Kinyarwanda, Kyrgyz, Latvian, Lithuanian, Maltese, Mongolian, Oriya, Persian, Polish, Portuguese, Romanian, Russian, Sakha, Slovak, Slovenian, Spanish, Sursilvan, Swedish, Tamil, Tatar, Turkish, Ukrainian, Vallader, Vietnamese, and Welsh



Show Globe