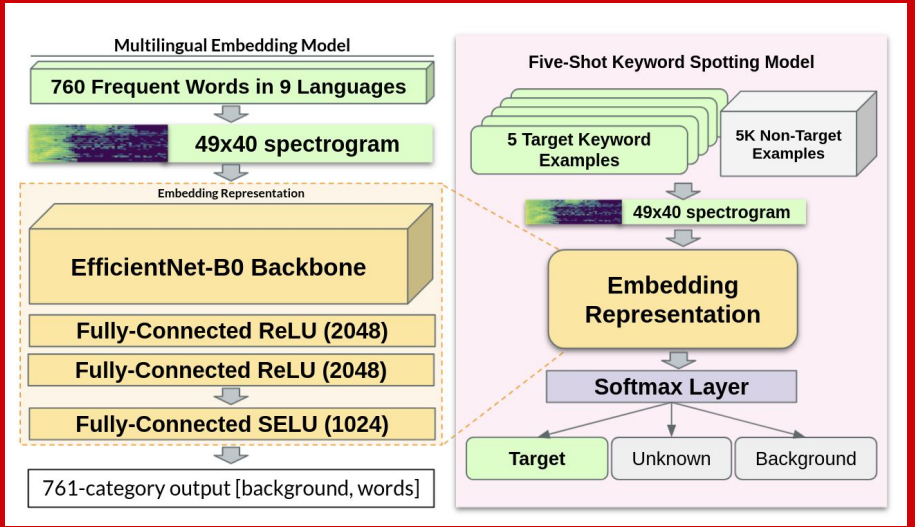
Few Shot Keyword Spotting in Any Language

Mark Mazumder
markmazumder@g.harvard.edu

Harvard John A. Paulson School of Engineering and Applied Sciences



Thanks + Acknowledgements

- Prof. Vijay Janapa Reddi
- Keith Achorn
- Colby Banbury
- Sharad Chitlangia
- Juan Ciro
- Greg Diamos
- Daniel Galvez
- Yiping Kang
- David Kanter
- Peter Mattson
- Josh Meyer
- Mark Sabini
- Pete Warden
- + many others...

Background: Keyword Spotting (KWS) /Wake words/Hotwords

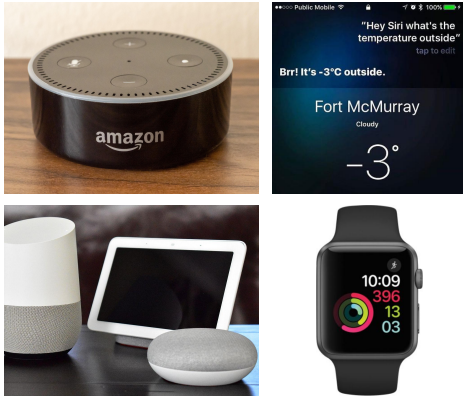


Image credits: Amazon, Apple, Google

- **Always-on voice assistants**: “OK Google”, “Hey Siri”, “Alexa,” …
- Ubiquitous but **limited vocabularies**

Background: Keyword Spotting (KWS) /Wake words/Hotwords

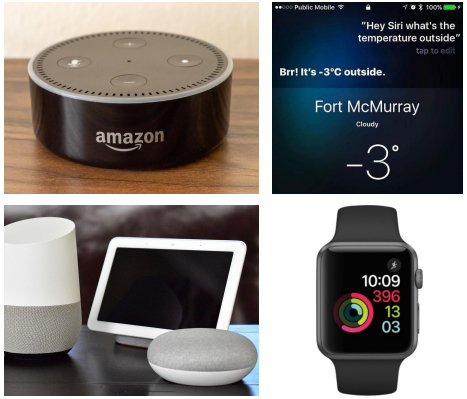


Image credits: Amazon, Apple, Google

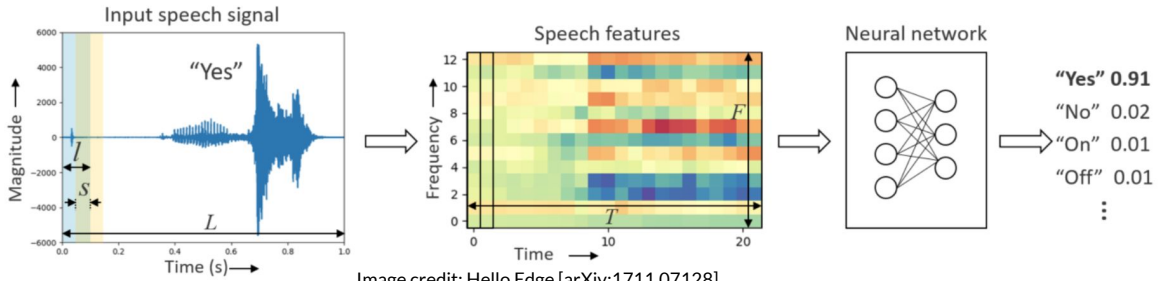


Image credit: Hello Edge [arXiv:1711.07128]

- **Always-on voice assistants**: "OK Google", "Hey Siri", "Alexa," …
- Ubiquitous but **limited vocabularies**

Traditional approach for KWS:

- **Needs thousands of training examples per keyword**

Background: Keyword Spotting (KWS) /Wake words/Hotwords

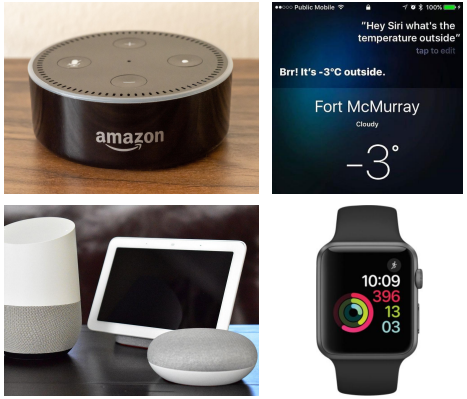


Image credits: Amazon, Apple, Google

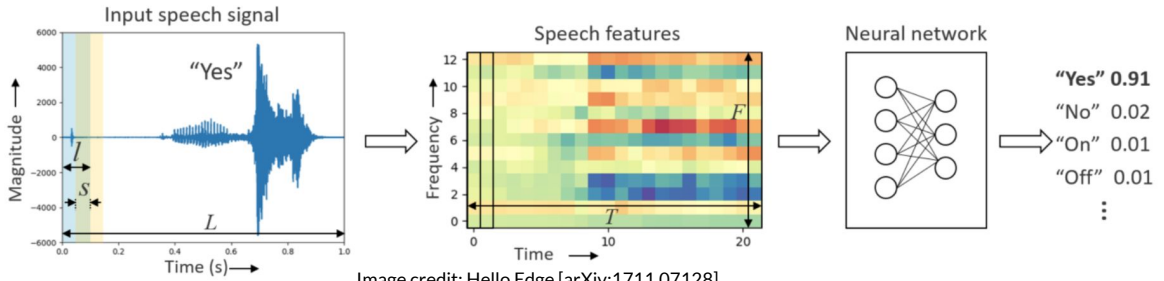


Image credit: Hello Edge [arXiv:1711.07128]

- **Always-on voice assistants**: "OK Google", "Hey Siri", "Alexa," …
- Ubiquitous but **limited vocabularies**

Traditional approach for KWS:

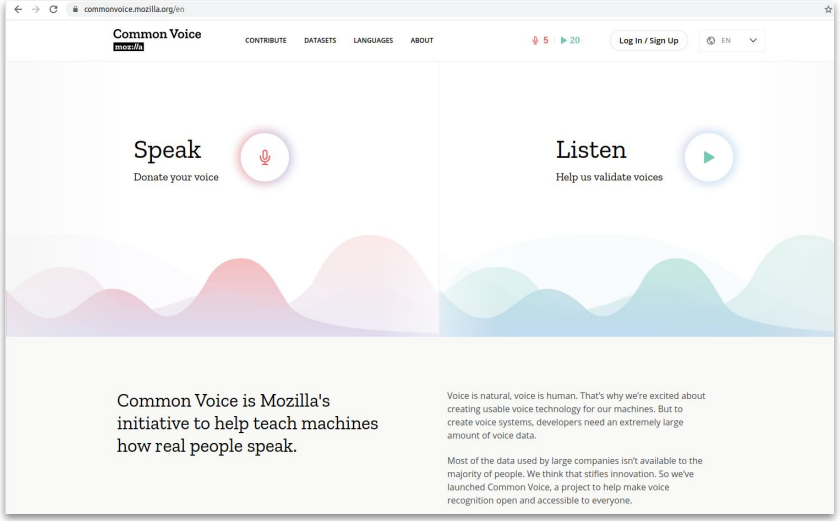
- **Needs thousands of training examples per keyword**

Goal: support **any** keyword in **any** language with just **five** examples

Mozilla Common Voice

commonvoice.mozilla.org [arXiv:1912.06670]

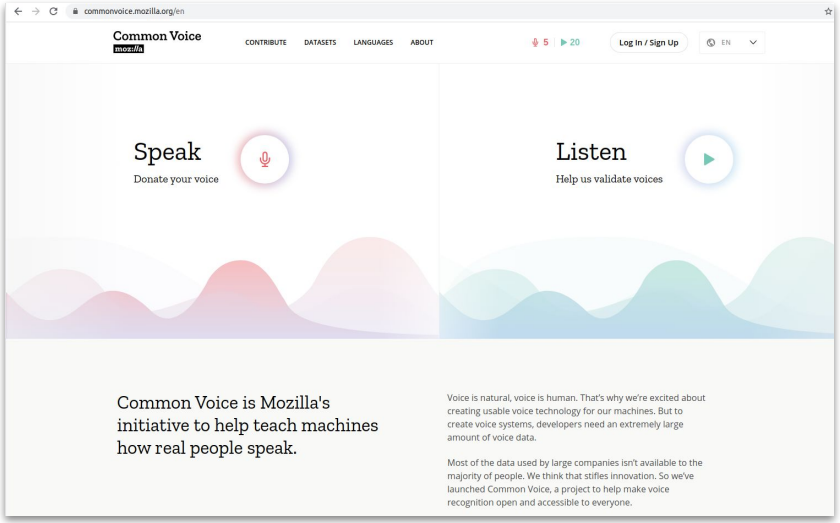
- **60+** languages
- Sentence audio, text transcription
- Crowdsourced



Mozilla Common Voice

commonvoice.mozilla.org [arXiv:1912.06670]

- **60+** languages
- Sentence audio, text transcription
- Crowdsourced

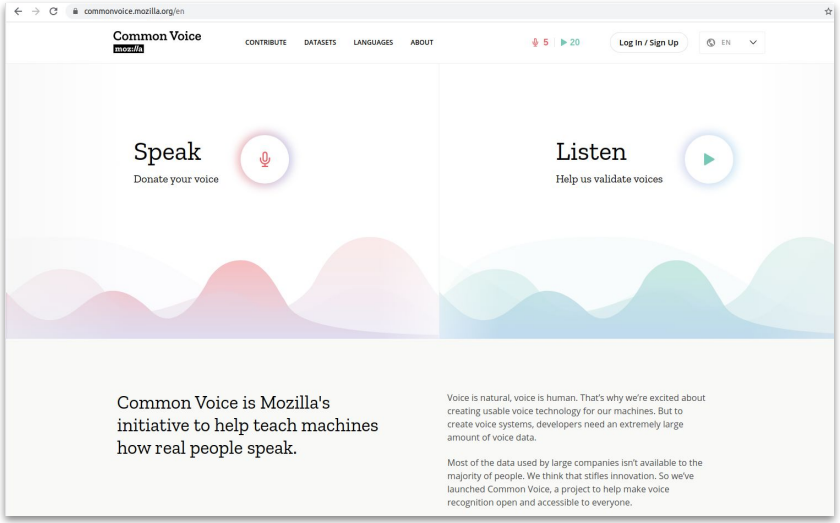


- **Extract** keywords from Common Voice sentences via **forced alignment**
  - **4.3M** examples
  - 3,126 keywords
  - **22 languages**

Mozilla Common Voice

commonvoice.mozilla.org [arXiv:1912.06670]

- **60+** languages
- Sentence audio, text transcription
- Crowdsourced



- **Extract** keywords from Common Voice sentences via **forced alignment**
  - **4.3M** examples
  - 3,126 keywords
  - **22 languages**
- Train a **multilingual embedding model** to represent keywords as speaker-independent **vectors**

Automatic Keyword Dataset Generation

Frequent Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

Automatic Keyword Dataset Generation

Frequent Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

Search CV Sentences

en/validated.tsv

He gazed **up** the steep bank.	clip_29132.mp3
"**Yes**," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
...	...

All sentences containing keywords

Automatic Keyword Dataset Generation

Frequent Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

Search CV Sentences

en/validated.tsv

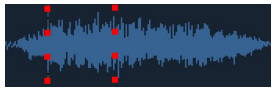
He gazed **up** the steep bank.	clip_29132.mp3
"**Yes**," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
...	...

All sentences containing keywords

Extract Keyword Utterances

Input 5-10 second mp3

*"... gazed **up** the steep..."*



up

Output

1 second wav, silence-padded

Automatic Keyword Dataset Generation

Frequent Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

Search CV Sentences

en/validated.tsv

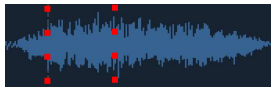
He gazed **up** the steep bank.	clip_29132.mp3
"Yes," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
…	…

All sentences containing keywords

Extract Keyword Utterances

Input 5-10 second mp3

*"… gazed **up** the steep…"*



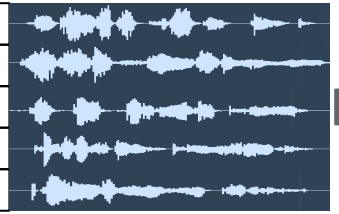
up

Output

1 second wav, silence-padded

- **Forced Alignment** estimates timings from <Audio, Text>
- Well-established technique
- Alignments trained from a *flat start* (no prior acoustic model)

He gazed **up** the steep bank.
"Yes," said Harry sullenly.
Get that ladder **up** here.
Pencils **down**, time is over.
There is **no** help for it.



Estimate Per-Word Timing

Extract Keywords

Large Keyword Dataset

4.3M Utterances
3,126 Keywords
22 Languages

Add

Automatic Keyword Dataset Generation

Frequent Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

Search CV Sentences

en/validated.tsv

He gazed **up** the steep bank.	clip_29132.mp3
"Yes," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
...	...

All sentences containing keywords

Extract Keyword Utterances

Input 5-10 second mp3

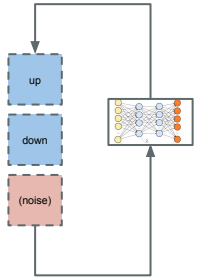
*"... gazed **up** the steep..."*

up

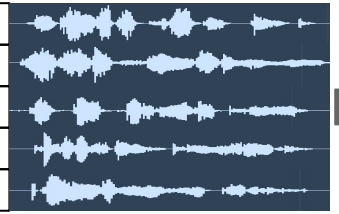
Output

1 second wav, silence-padded

Train KWS Model



He gazed **up** the steep bank.
"Yes," said Harry sullenly.
Get that ladder **up** here.
Pencils **down**, time is over.
There is **no** help for it.



Estimate Per-Word Timing

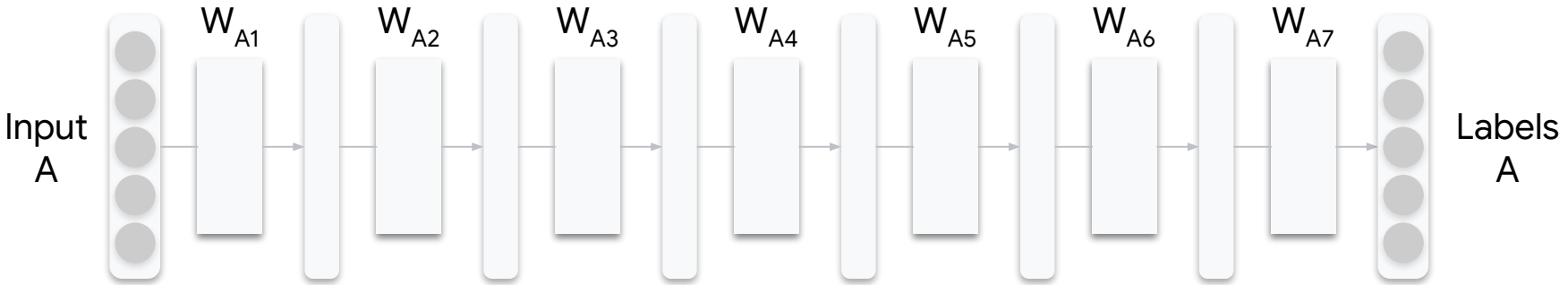
Extract Keywords

Large Keyword Dataset

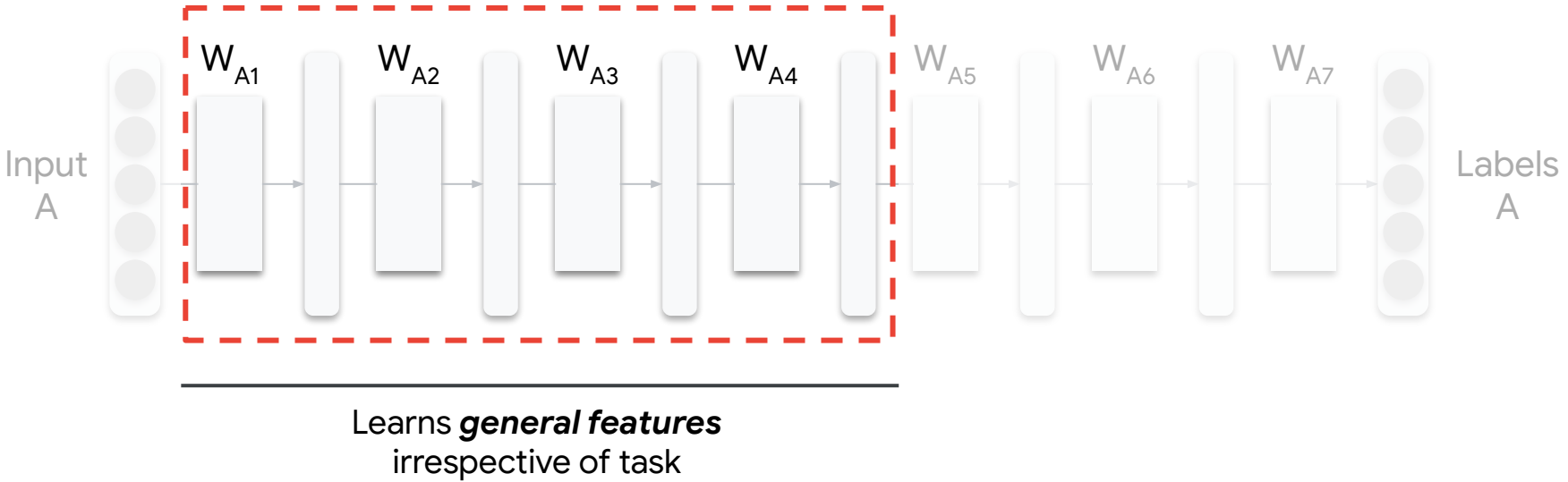
4.3M Utterances
3,126 Keywords
22 Languages

Add

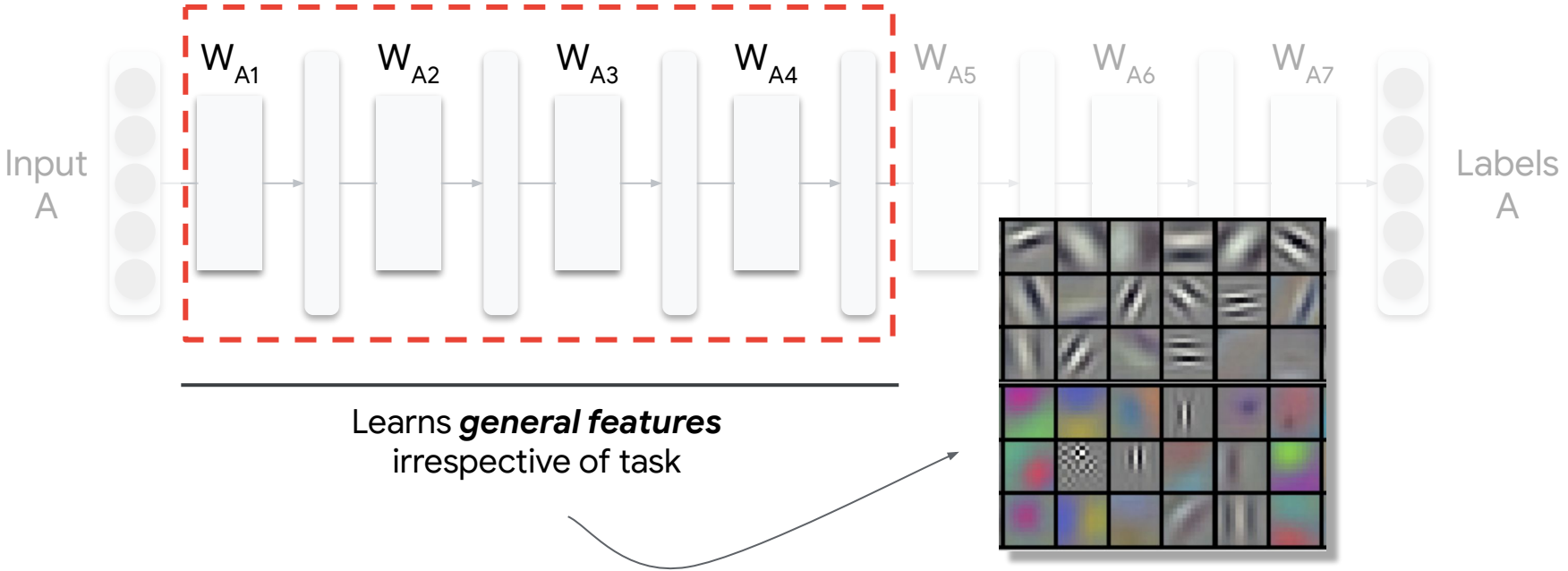
Feature extraction



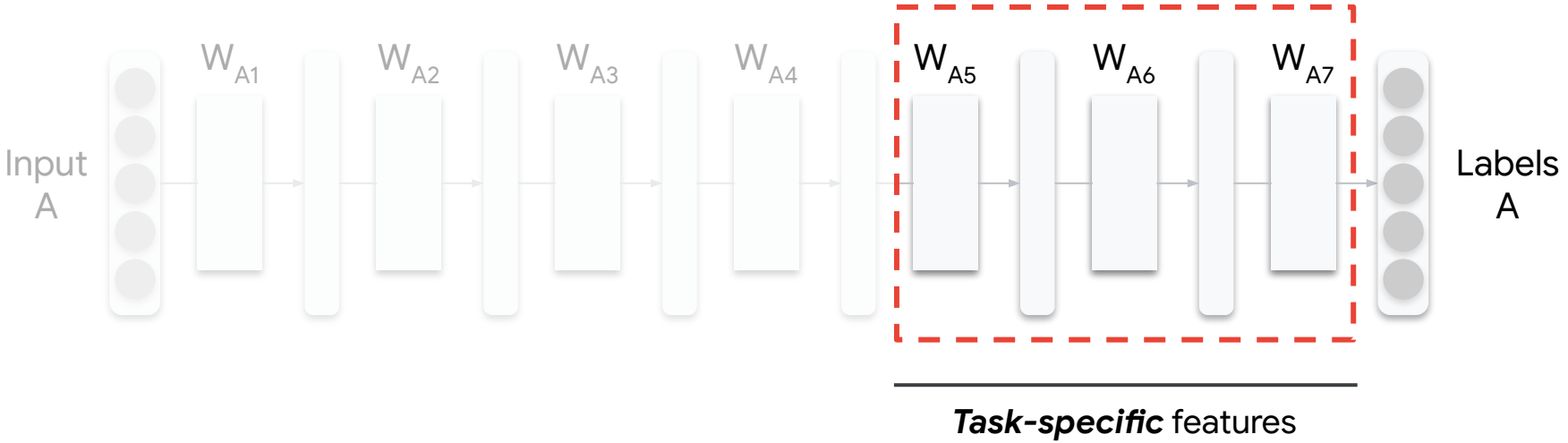
Feature extraction



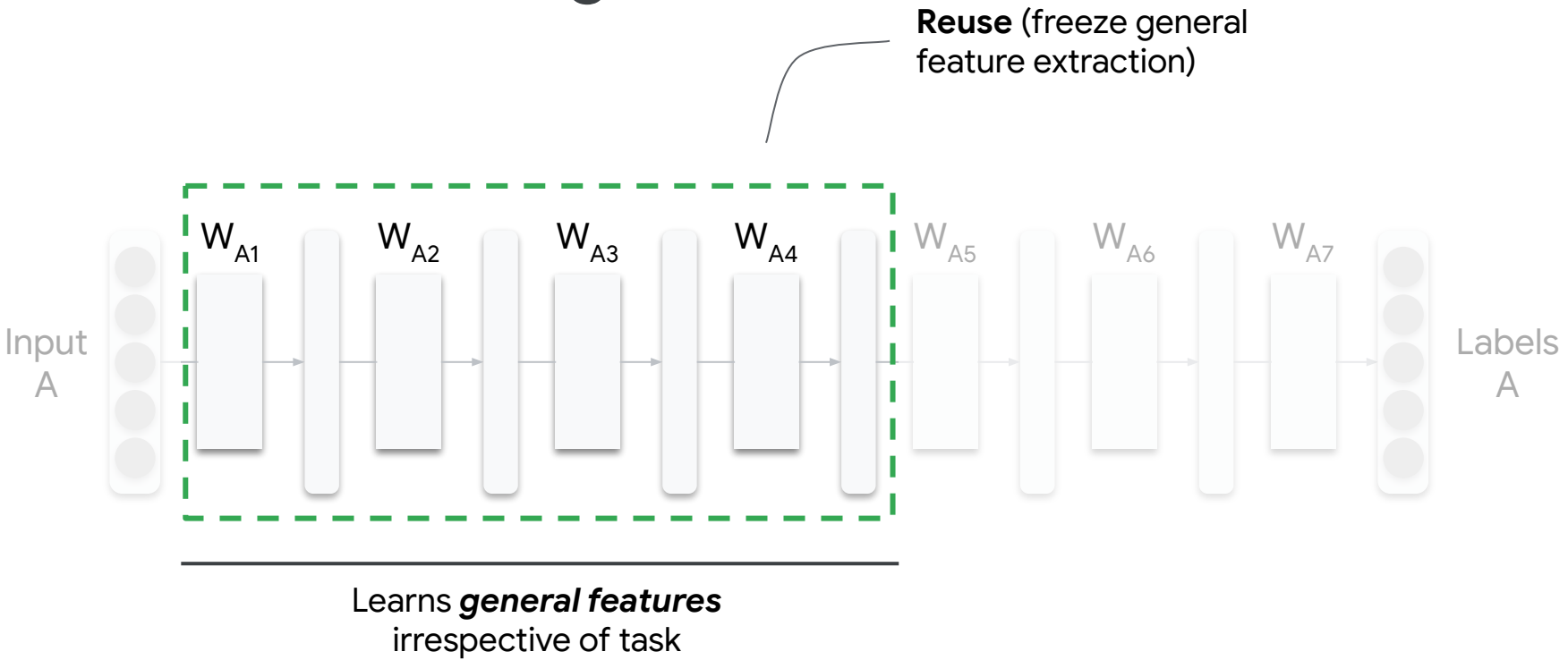
Feature extraction



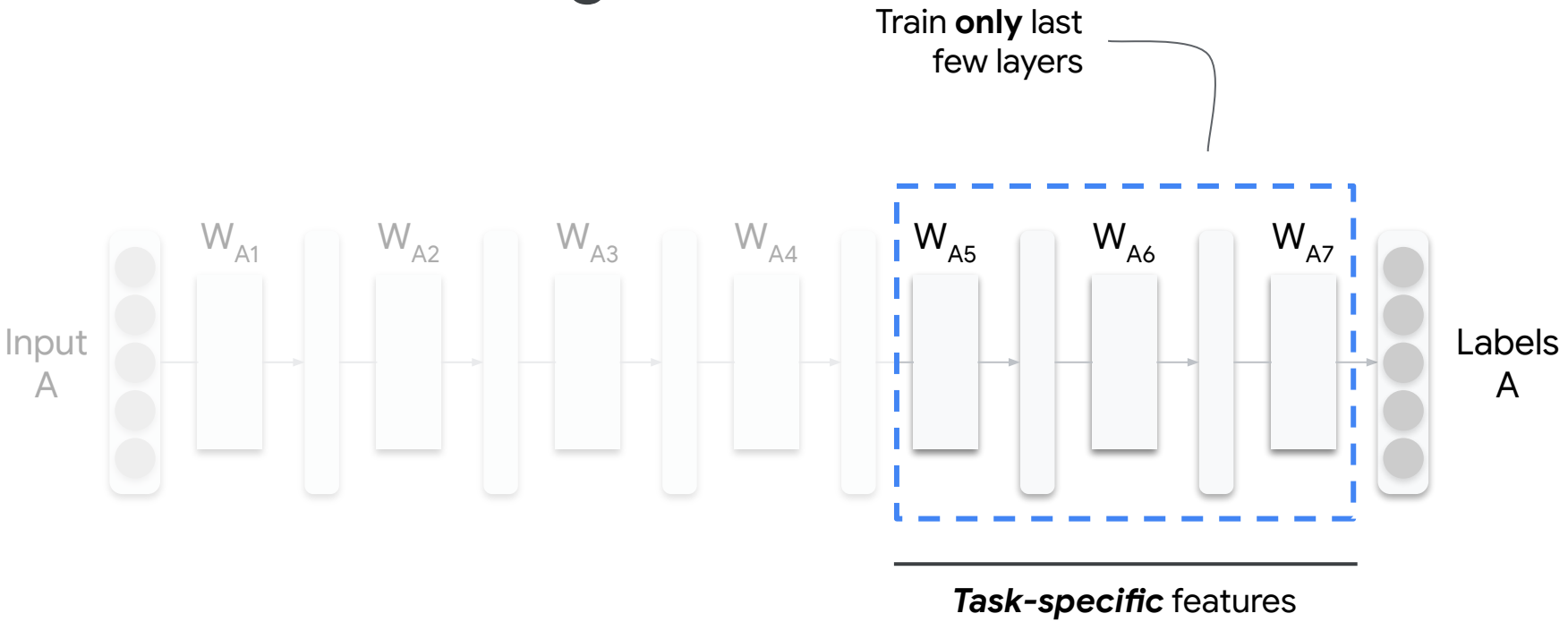
Feature extraction



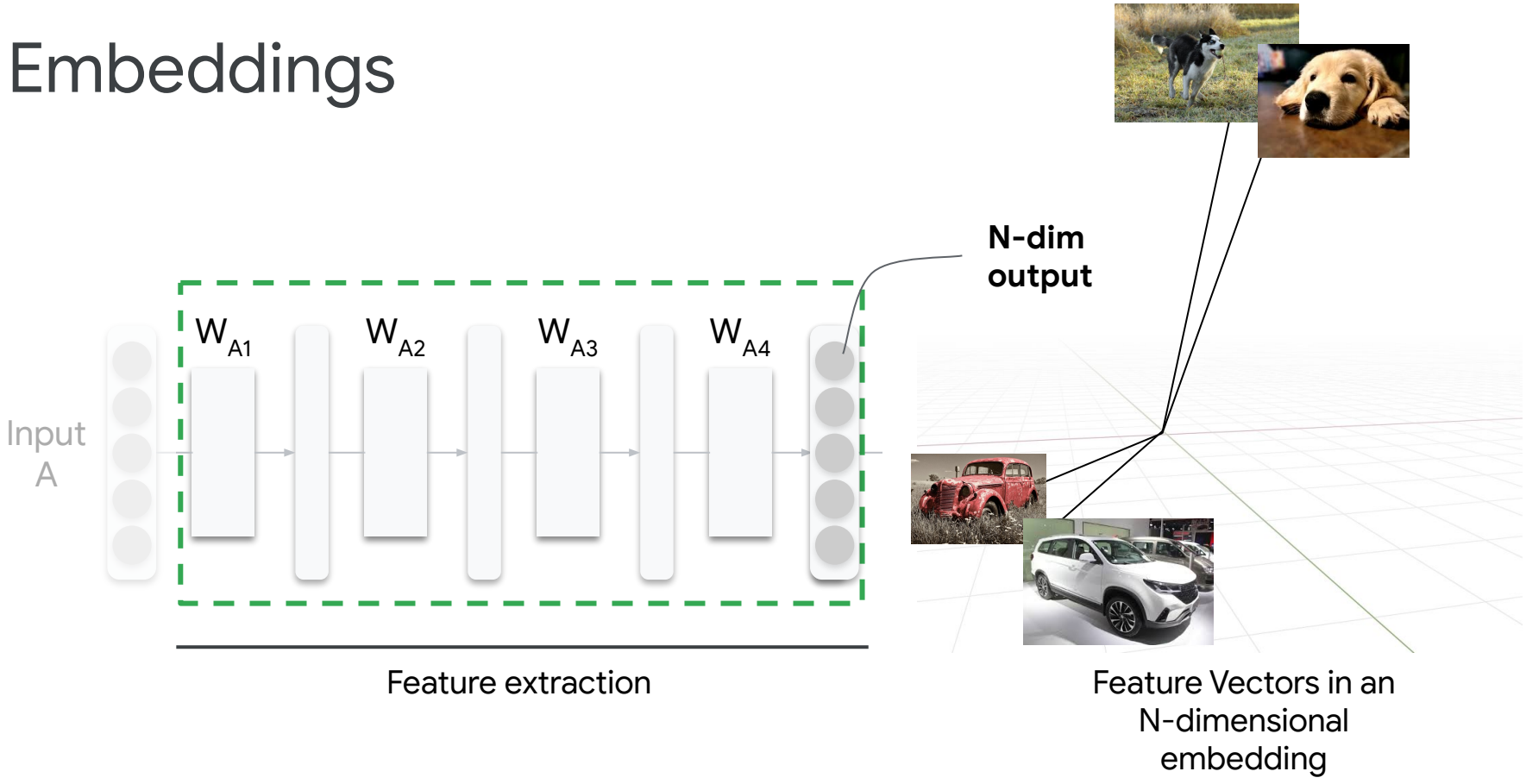
Transfer Learning



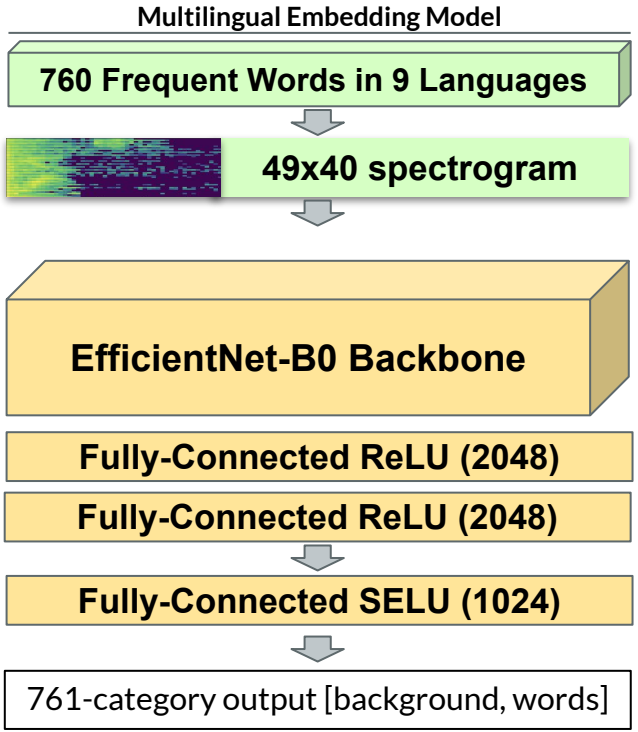
Transfer Learning



Embeddings

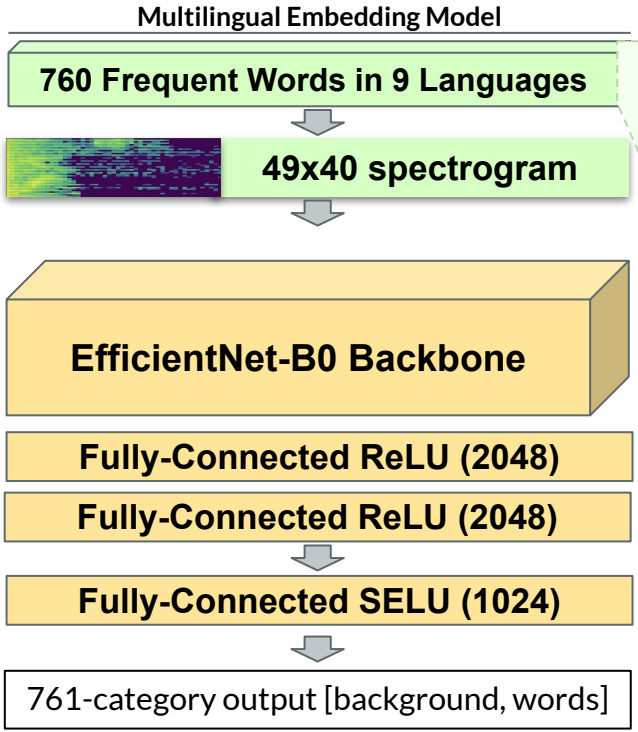

N-dim output
Input A
W_{A1}
W_{A2}
W_{A3}
W_{A4}
Feature extraction
Feature Vectors in an N-dimensional embedding

Multilingual Embedding Model



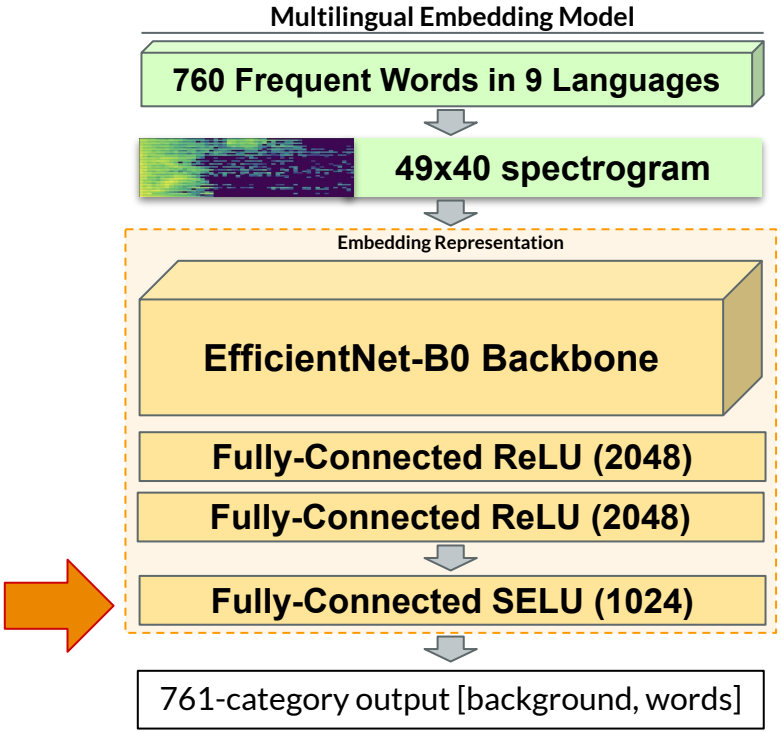
Simple classifier for 760 frequent words in 9 languages

Multilingual Embedding Model



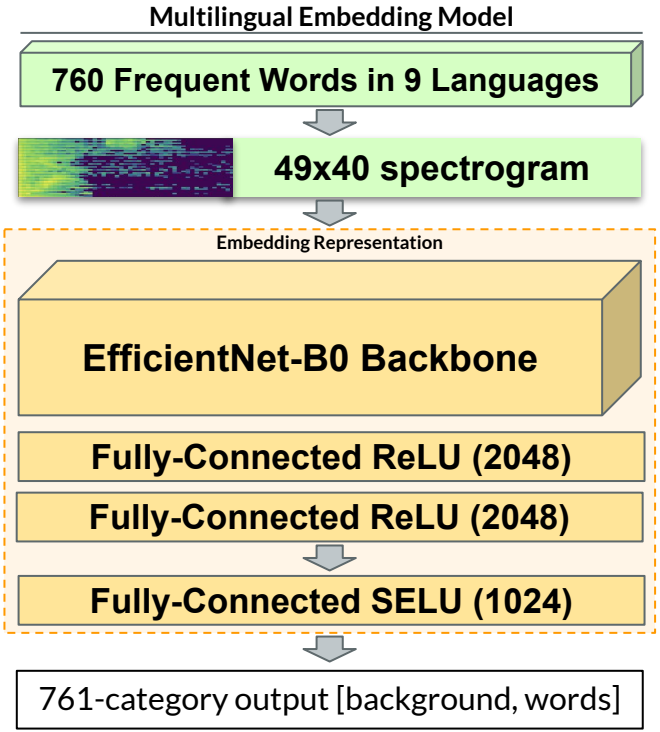
Language	# words	# train
English	265	518760
German	152	287100
French	105	205920
Kinyarwanda	68	134640
Catalan	80	132660
Persian	35	69300
Spanish	31	61380
Italian	17	31680
Dutch	7	13860
Model	760	1455300

Few-Shot Keyword Spotting

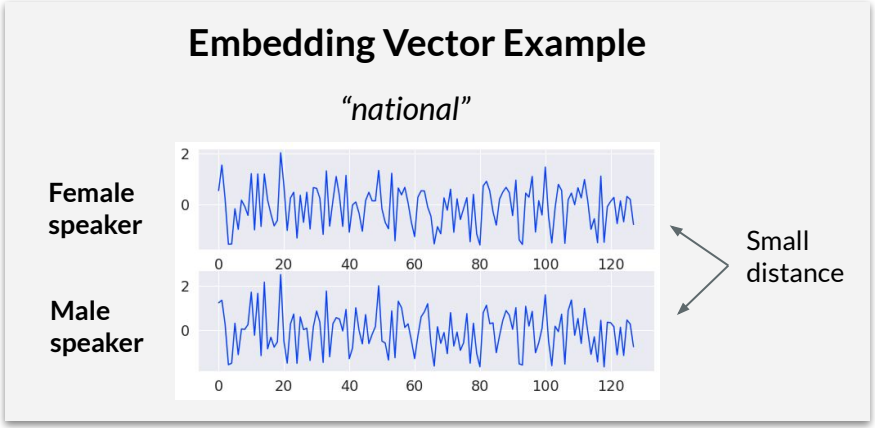


Use **penultimate layer** for embedding representation in **keyword-spotting model**

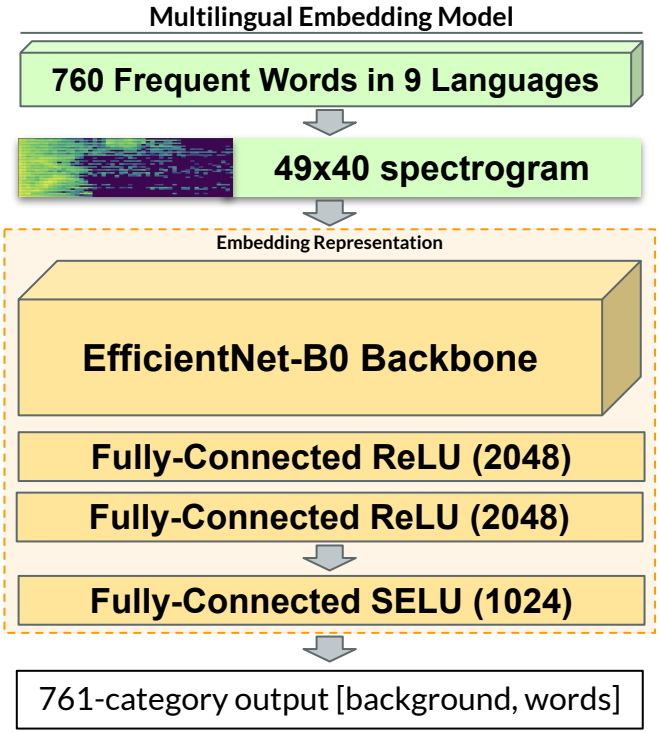
Multilingual Embedding Model



- **Penultimate layer output** as embedding vector

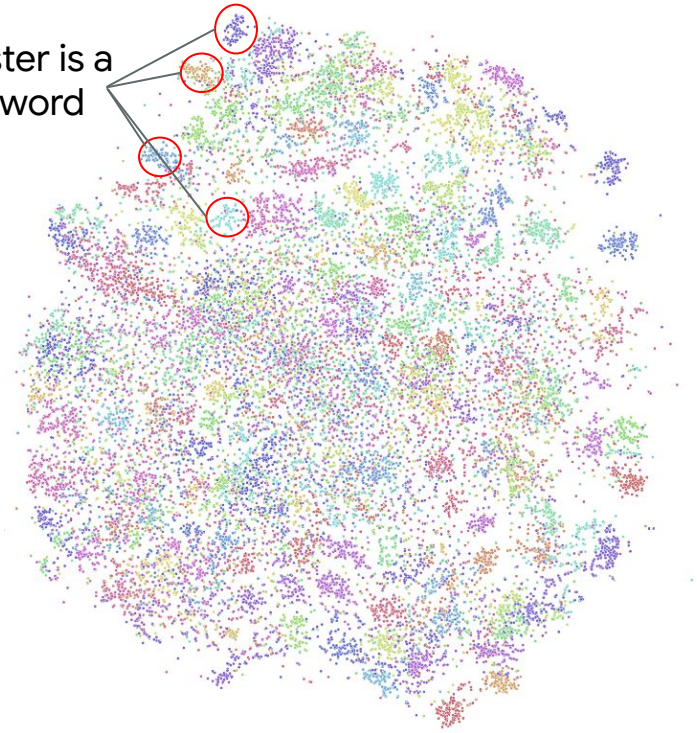


Multilingual Embedding Model

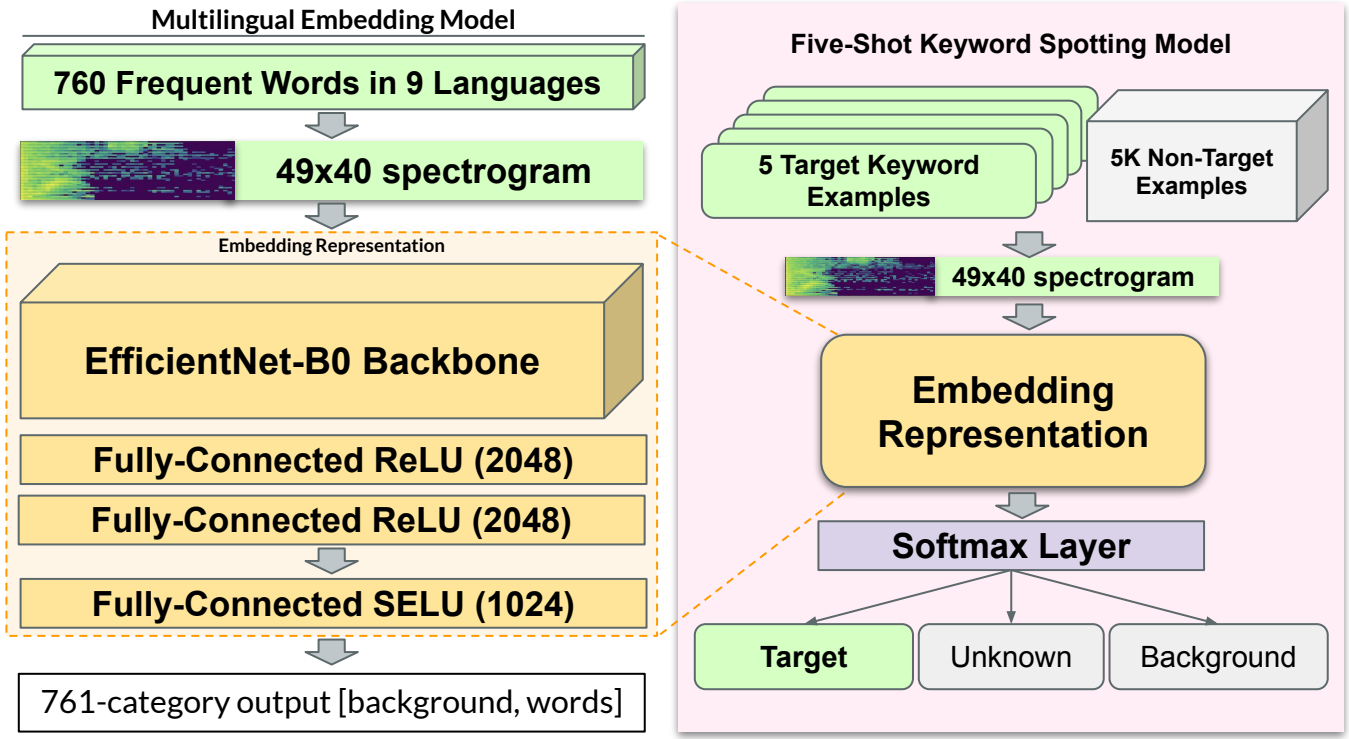


t-SNE view of 165 novel words (not used to train the classifier)

Each cluster is a different word

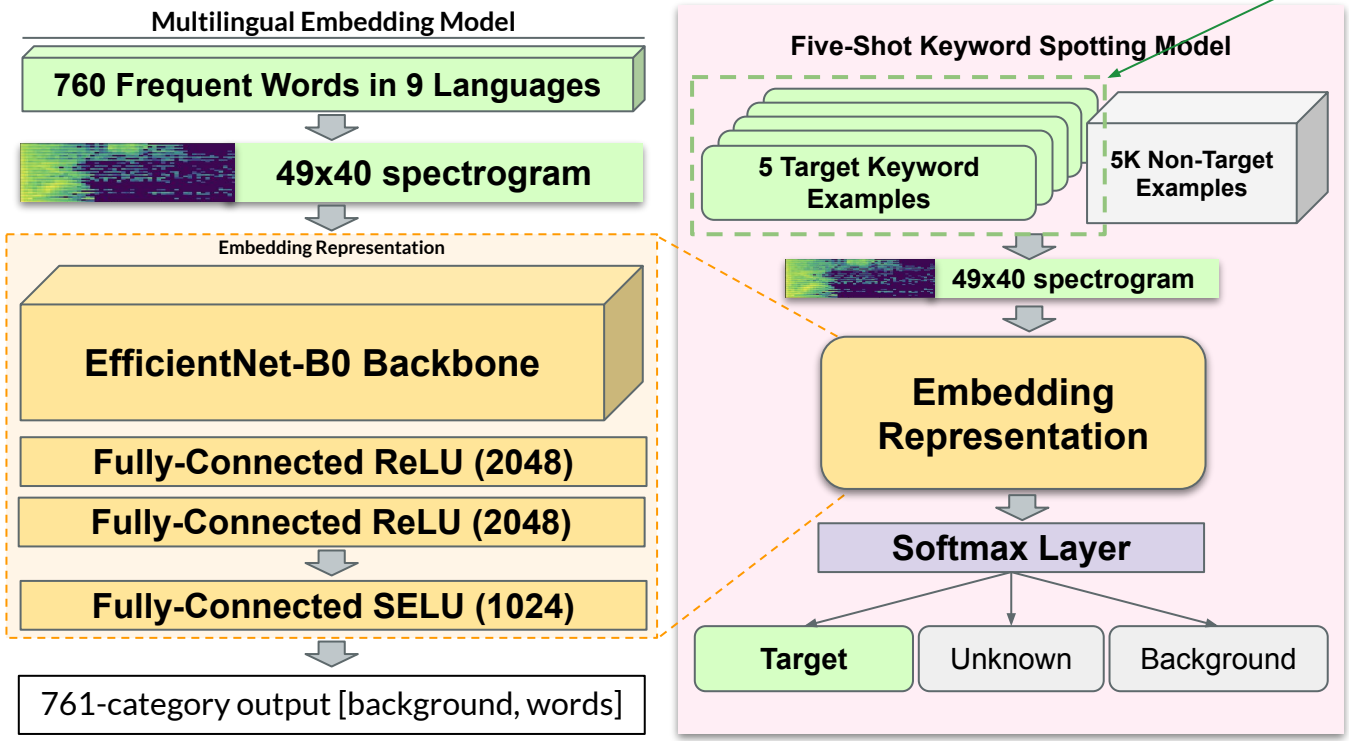


Few-Shot Keyword Spotting



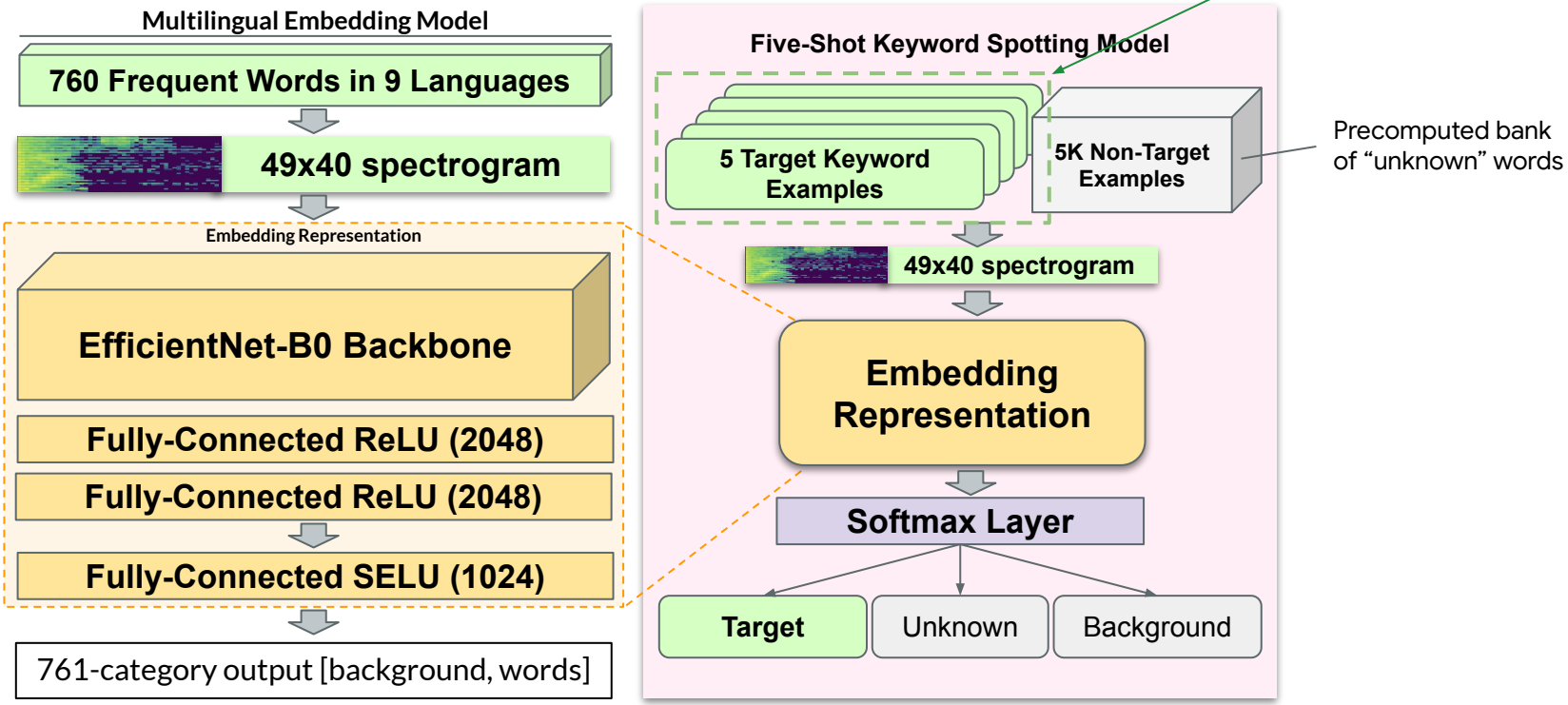
Few-Shot Keyword Spotting

Reduces training examples from **thousands** to just **five**



Few-Shot Keyword Spotting

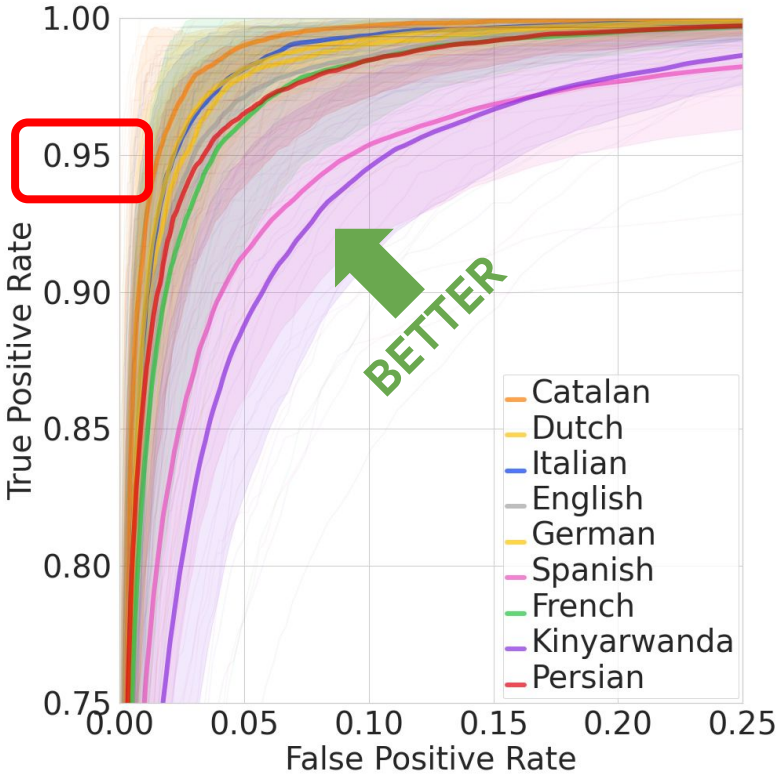
Reduces training examples from **thousands** to just **five**



Evaluation

- Classification and streaming accuracy
- 440 keywords
- 22 Languages
- 5 random training samples per keyword

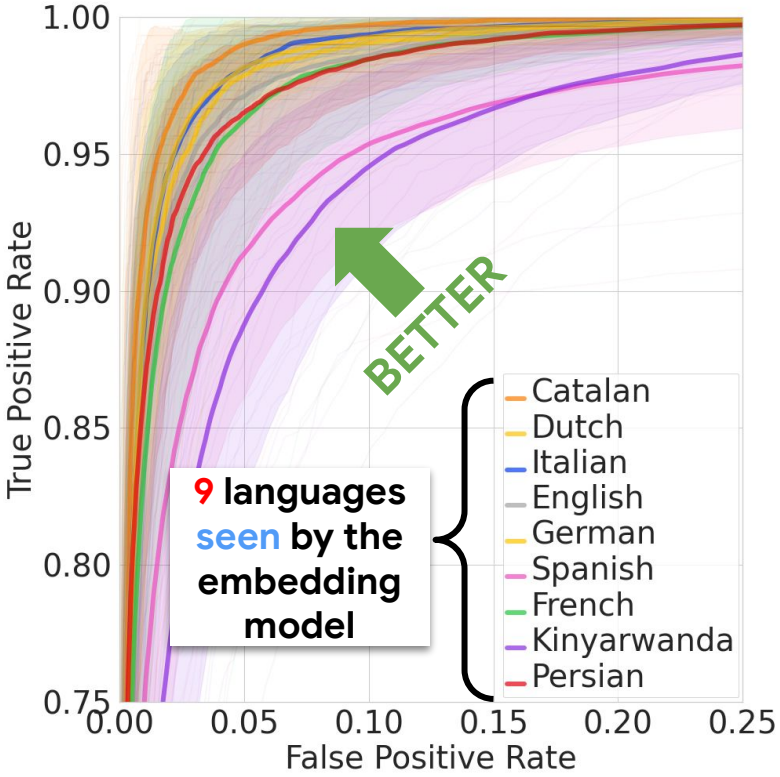
5-shot Keyword Spotting Results



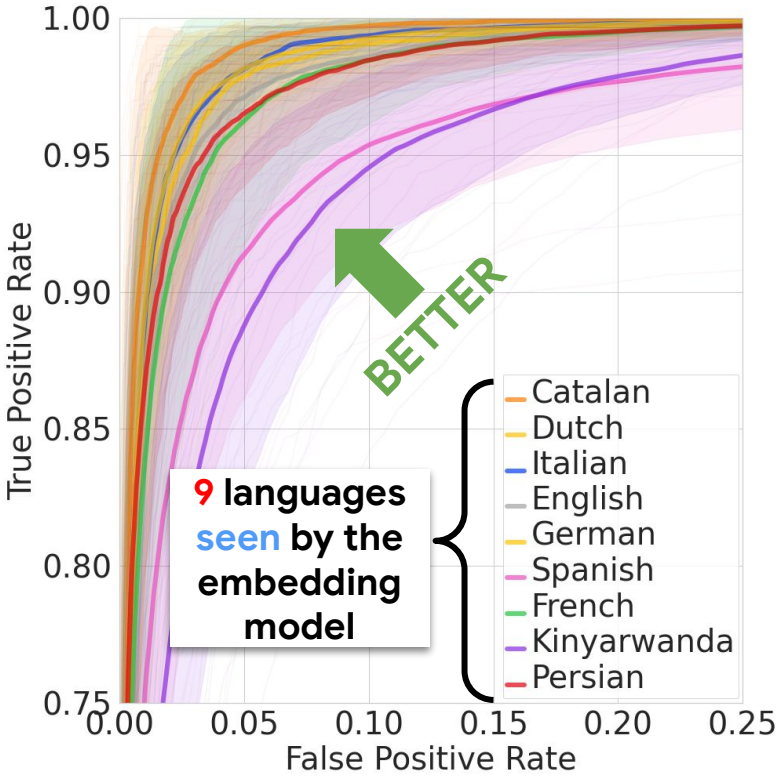
- **Classification performance** shown as ROC curves
- High top-1 accuracy on keywords **unseen** by the embedding model with only **five** training examples
- Avg F_1@threshold 0.8 = 0.75

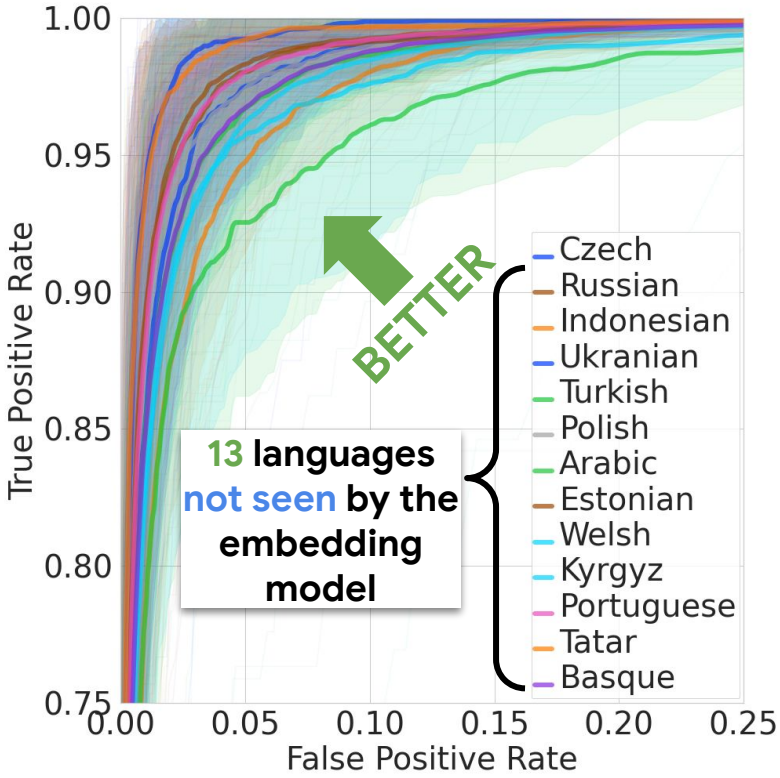
ROC: Receiver Operating Characteristic

Generalizing to **Any** Language

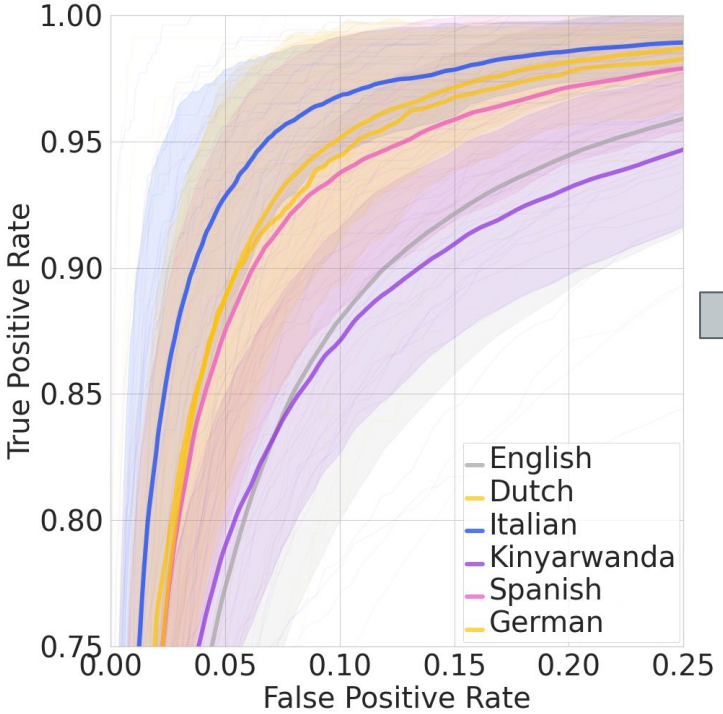


Generalizing to **Any** Language

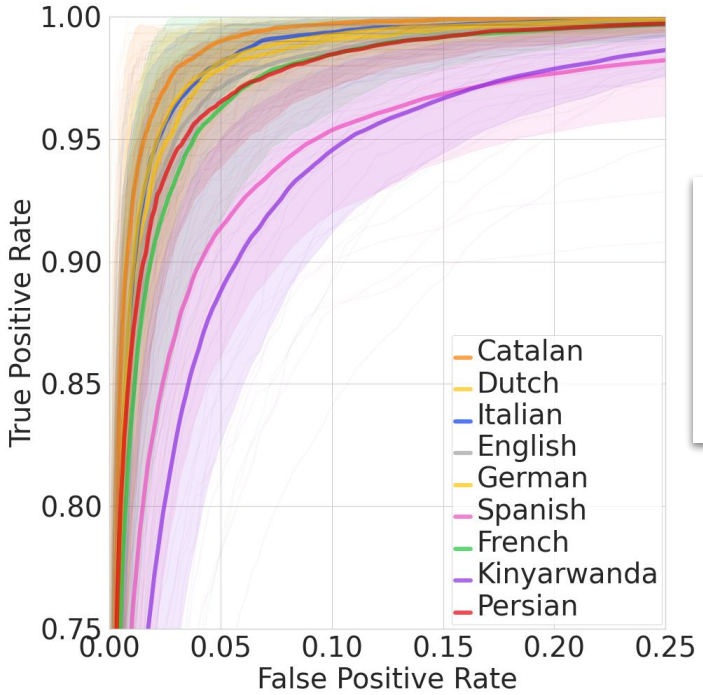




Monolingual vs Multilingual Embedding



Six Monolingual Embedding Models



Multilingual Embedding Model

Performance improves across all languages with no additional data

Streaming Accuracy Scenarios

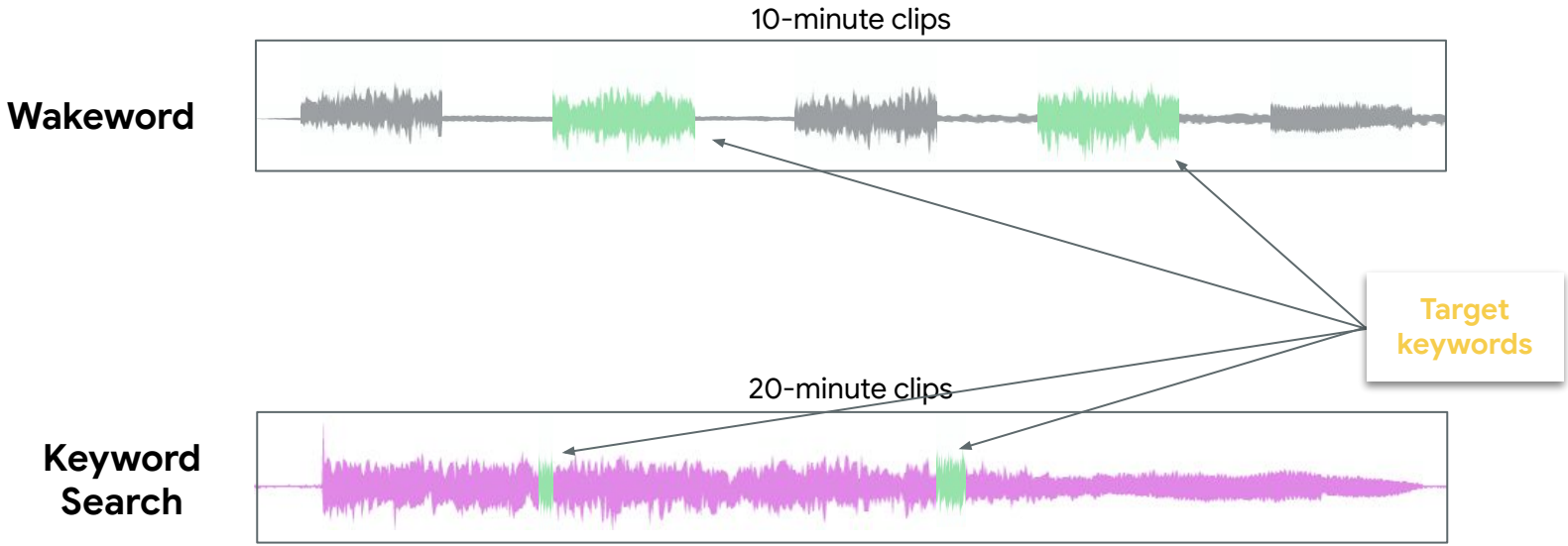
Wakeword

10-minute clips

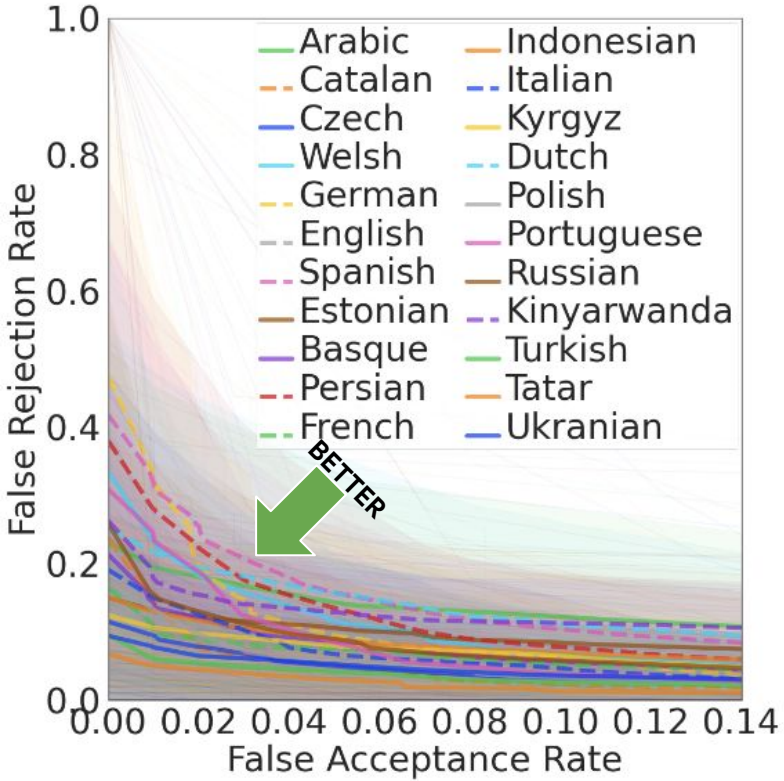
Keyword Search

20-minute clips

Streaming Accuracy Scenarios



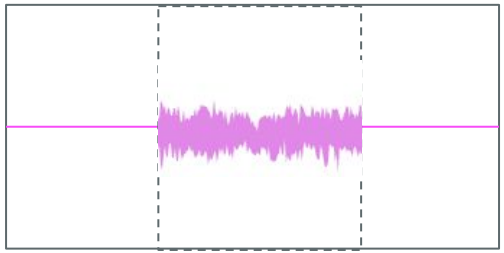
Streaming Accuracy Tests Across 22 Languages



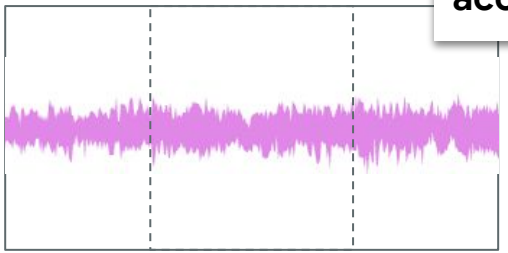
- Wakeword scenario: **Avg TPR 87.4%**

Extracting keywords with audio context

Keyword in Common Voice sentence



1-second **silence-padded** extraction

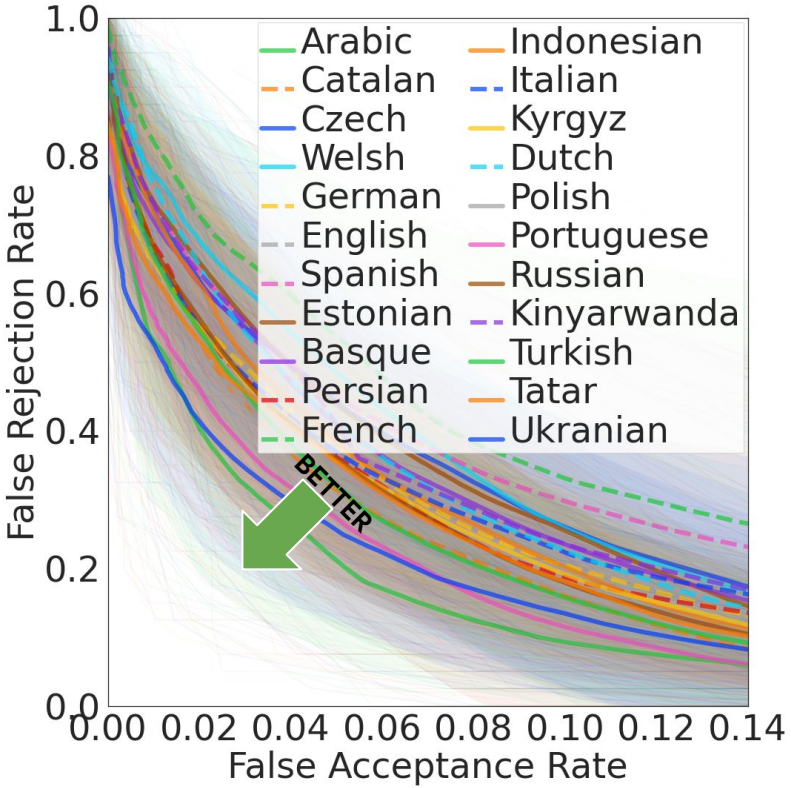


1-second **context-padded** extraction

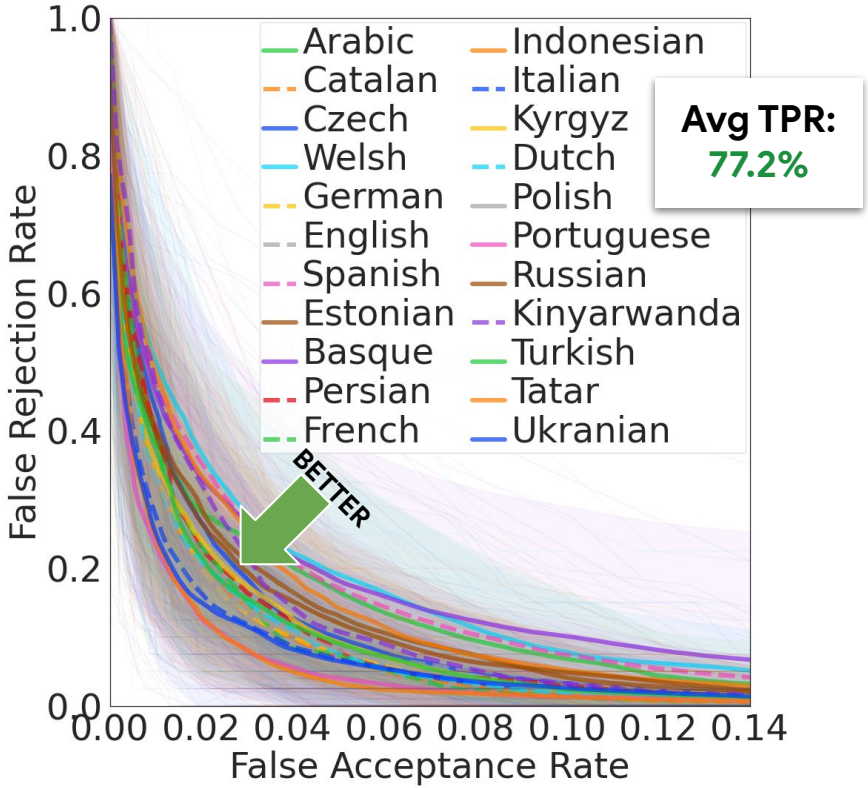
Improves keyword search accuracy

Streaming Accuracy on Keyword Search

Embedding trained **only on silence-padded** keywords



Embedding trained on **silence + context-padded** keywords



Broadcast Radio Monitoring

MAKERERE UNIVERSITY

Makerere AI Lab

- **Problem Description:** Create a Covid-19 keyword spotting system to monitor public radio broadcasts for the Uganda Ministry of Health
- **Impact Goals:** Estimate Covid spread, vaccine sentiment & info, other topics (crop disease, …)
- **Domain experts:**
  - Dr. Joyce Nabende, Jonathan Mukiibi (Makerere AI Lab)
  - Dr. Josh Meyer (Mozilla Foundation Machine Learning Fellow, Coqui.io)

Broadcast Radio Monitoring in Luganda

Potential for social impact

- Uganda Ministry of Health can gather real-time updates on health, safety, food security

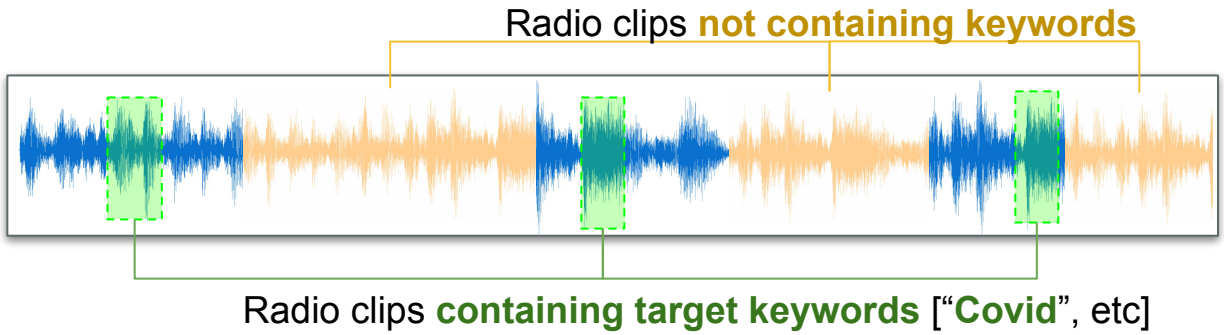
https://radio.unglobalpulse.net/uganda/

In Uganda, internet infrastructure is often poorly developed, precluding the use of social media to gauge sentiment. Instead, community radio phone-in talk shows are used to voice views and concerns. In a project piloted by the United Nations (UN), radio browsing systems have been developed to monitor such radio shows [1, 2]. Currently, these systems are actively and successfully supporting relief and developmental programmes by the organisation. However, the deployed radio browsing systems use automatic speech recognition (ASR) and are therefore highly dependent on the availability of substantial transcribed speech corpora in the target language. This has proved to be a serious impediment when quick intervention is required, since the development of such a corpus is always time-consuming.

Excerpt from *Menon et. al.* Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. INTERSPEECH 2019

Radio Search: Evaluation

- Assembled streaming wavs from transcribed radio data
  - Interspersed with non-target radio clips



Multilingual Spoken Words Corpus

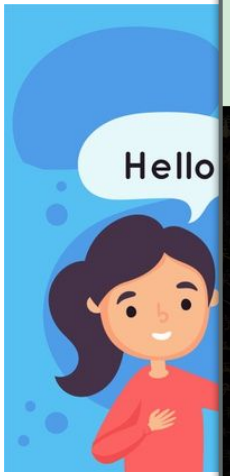
Under review: https://openreview.net/forum?id=c20jiJ5K2H

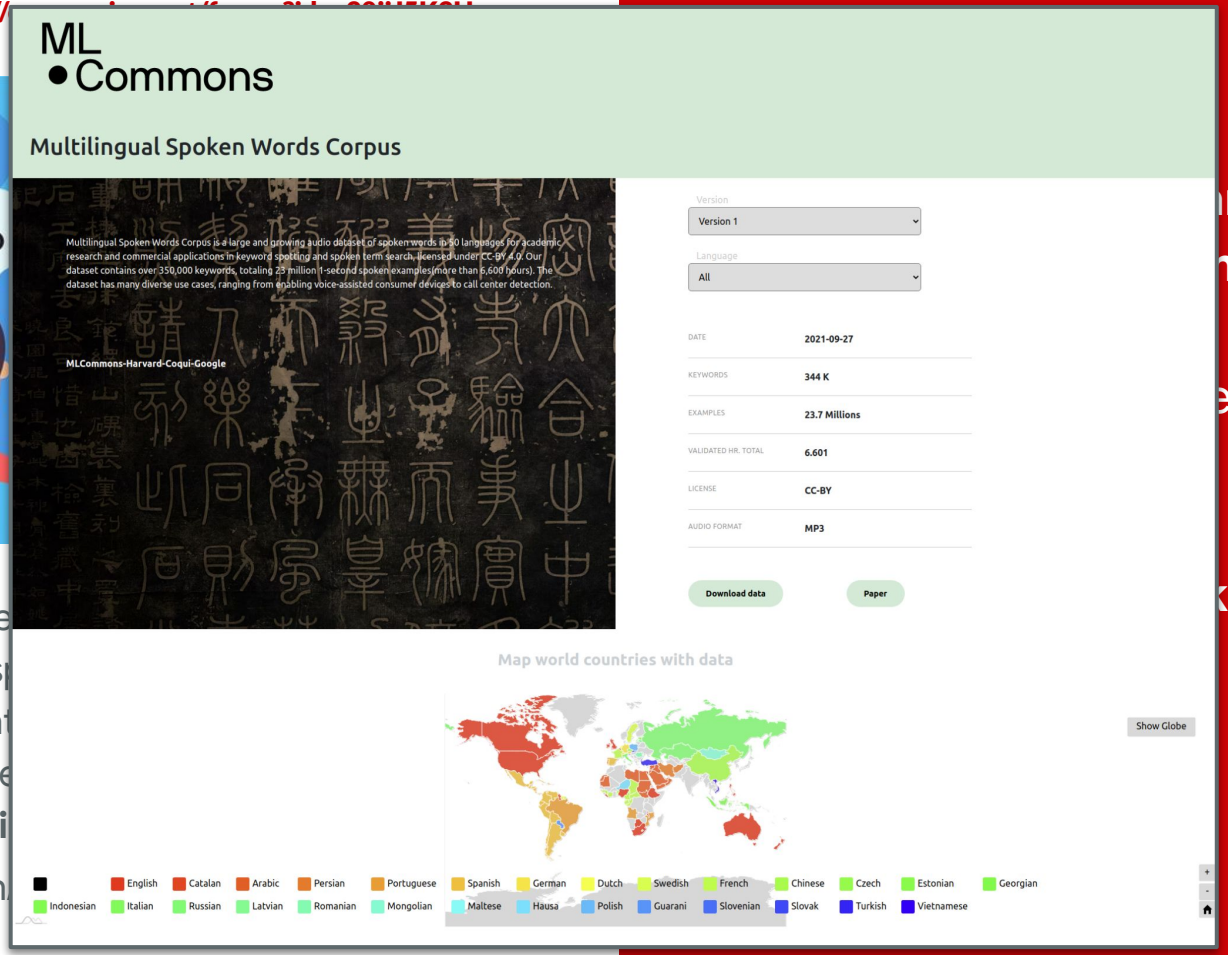
- **50+** languages
- Collectively spoken by over **5 Billion** people
- Regular updates with more data
- Includes forced alignments for **all of Common Voice**
- Includes train/dev/test **splits**

- Speech recordings **of spoken words** in over 50 languages
- Extracted from **Common Voice**
- 340,000+ words
- **23.7 million** one-second recordings
- 6,000+ hours
- Commercial use **ok** (CC-BY)
- Maintained by **MLCommons.org**

Multilingual Spoken Words Corpus

Under review: https://





of spoken

nguages

mmon Voice

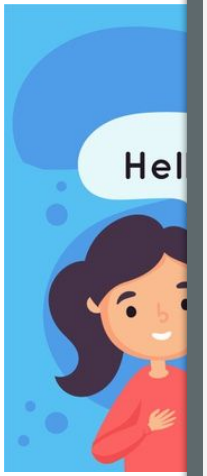
econd

k (CC-BY)

- **50+** language
- Collectively sp
- Regular updat
- Includes force
  Common Voi
- Includes train

Multilingual Spoken Words Corpus

Under review: https://openreview.net/forum?id=c20jiJ5K2H



- **50+** languag
- Collectively
- Regular upd
- Includes for
  Common V
- Includes tra

Paper available on OpenReview (to appear in NeurIPS 2021 Datasets track):

https://openreview.net/forum?id=c20jiJ5K2H

OpenReview.net | Search OpenReview... | Login

← Go to **NeurIPS 2021 Track Datasets and Benchmarks Round2** homepage

Multilingual Spoken Words Corpus

Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, Greg Diamos, Pete Warden, Josh Meyer, Vijay Janapa Reddi

20 Aug 2021 (modified: 30 Sept 2021) NeurIPS 2021 Datasets and Benchmarks Track (Round 2) Readers: Everyone Show Bibtex

Keywords: keyword spotting, speech recognition, low resource languages

TL;DR: Multilingual Spoken Words Corpus is a speech dataset of over 340,000 spoken words in 50 languages, with over 23.7 million examples.

Abstract: Multilingual Spoken Words Corpus is a large and growing audio dataset of spoken words in 50 languages collectively spoken by over 5 billion people, 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. We provide a detailed analysis of the contents of the data and contribute methods for detecting potential outliers. We report baseline accuracy metrics on keyword spotting models trained from our dataset compared to models trained on a manually-recorded keyword dataset. We conclude with our plans for dataset maintenance, updates, and open-sourced code.

Supplementary Material: zip

URL: During review, the dataset is available via a private URL to reviewers through OpenReview. Following the review period, our dataset will be hosted, maintained, and advanced by MLCommons.org

Dataset will be released publicly at **NeurIPS 2021** this December

oken
ges
n Voice

BY)

Conclusions

- **More data always helps**: KWS performance improves using data from **other** languages
- **Context** helps **keyword search** without impacting wakeword performance
- **Crowdsourced** data enables large-scale evaluation (many languages)

Code, models, & colabs:

github.com/harvard-edge/multilingual_kws