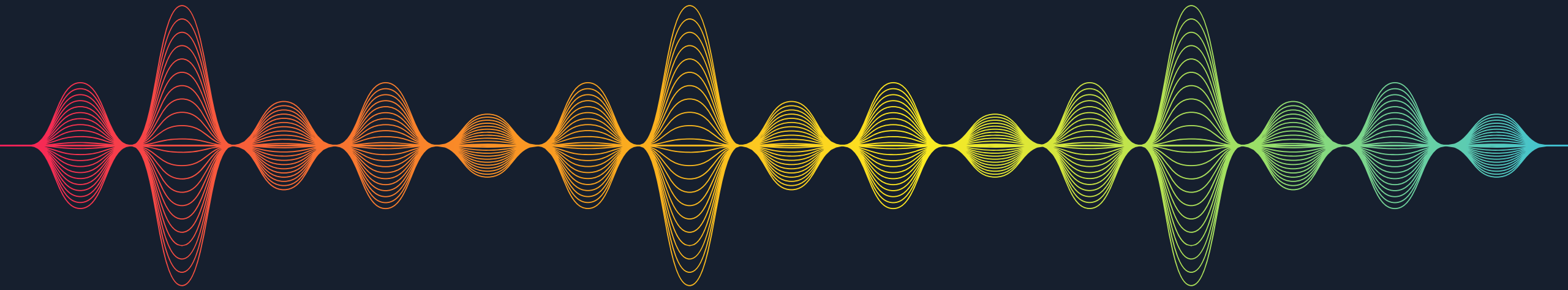
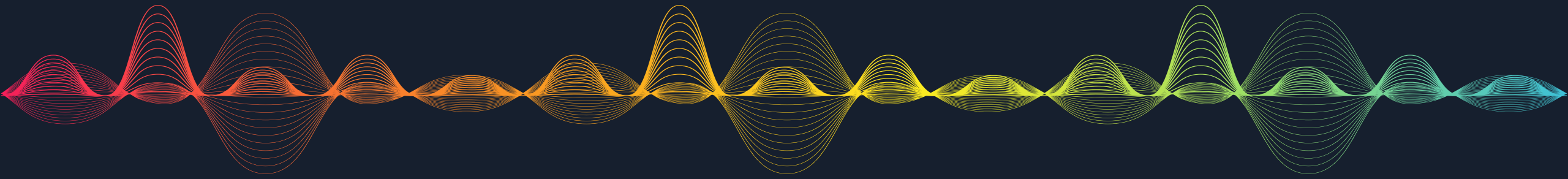
The Wild Bunch

Multilingual Keyword Spotting

... in any language!

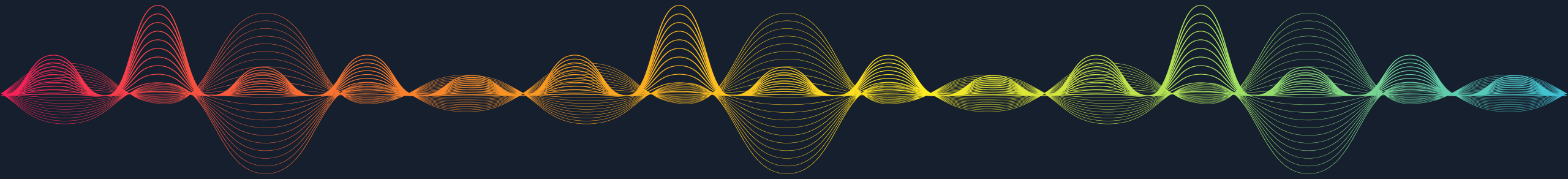


What is keyword spotting?



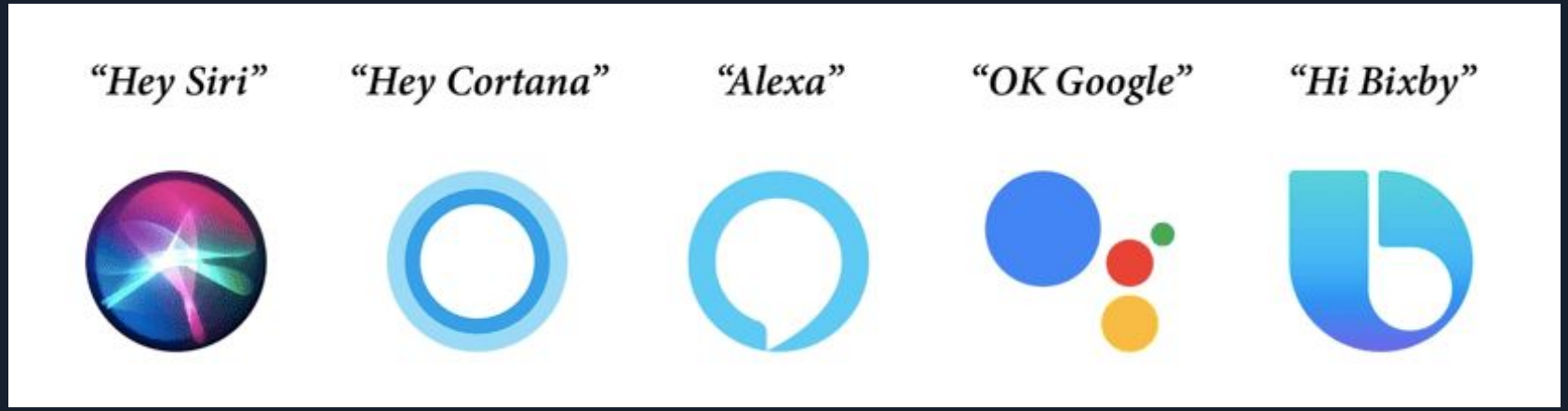
What is keyword spotting?

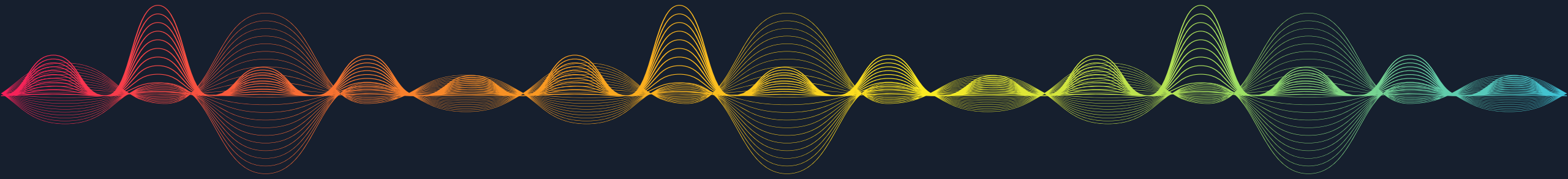
Automatically detecting specific words from a continuous audio



What is keyword spotting?

Automatically detecting specific words from a continuous audio





Few-Shot Keyword Spotting in Any Language

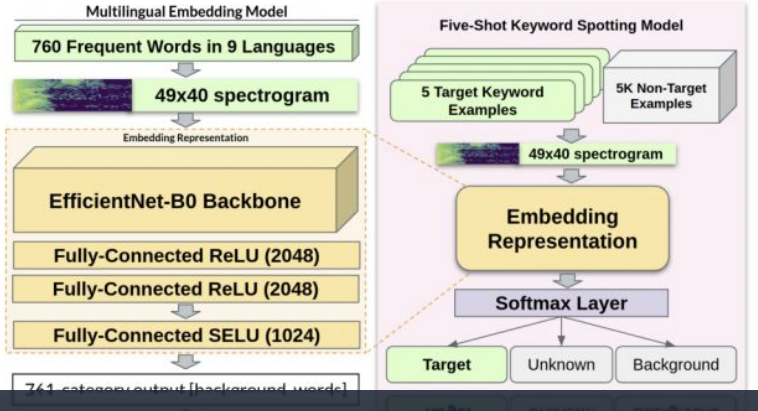
Mark Mazumder[1], *Colby Banbury*[1], *Josh Meyer*[2*], *Pete Warden*[3], *Vijay Janapa Reddi*[1]

[1]Harvard University, USA
[2]Coqui, Germany
[3]Google, USA

{markmazumder,cbanbury}@g.harvard.edu, josh@coqui.ai, petewarden@google.com, vj@eecs.harvard.edu

Abstract

We introduce a few-shot transfer learning method for keyword spotting in any language. Leveraging open speech corpora in nine languages, we automate the extraction of a large multilingual keyword bank and use it to train an embedding model. With just five training examples, we fine-tune the embedding model for keyword spotting and achieve an average F_1 score of 0.75 on keyword classification for 180 new keywords unseen by the embedding model in these nine languages. This embedding model also generalizes to new languages. We achieve an aver-



Keyword Spotting Pipeline

Spoken Words

Background Noise

Embedding Representation

Keyword Spotting (Fine-tuning)

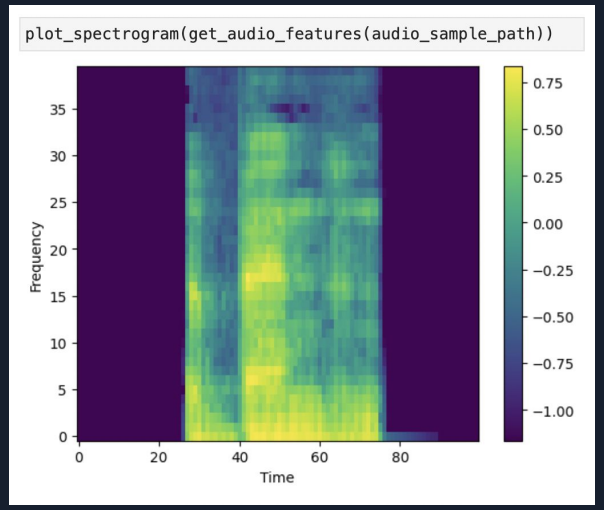
Target Examples

Negative Examples

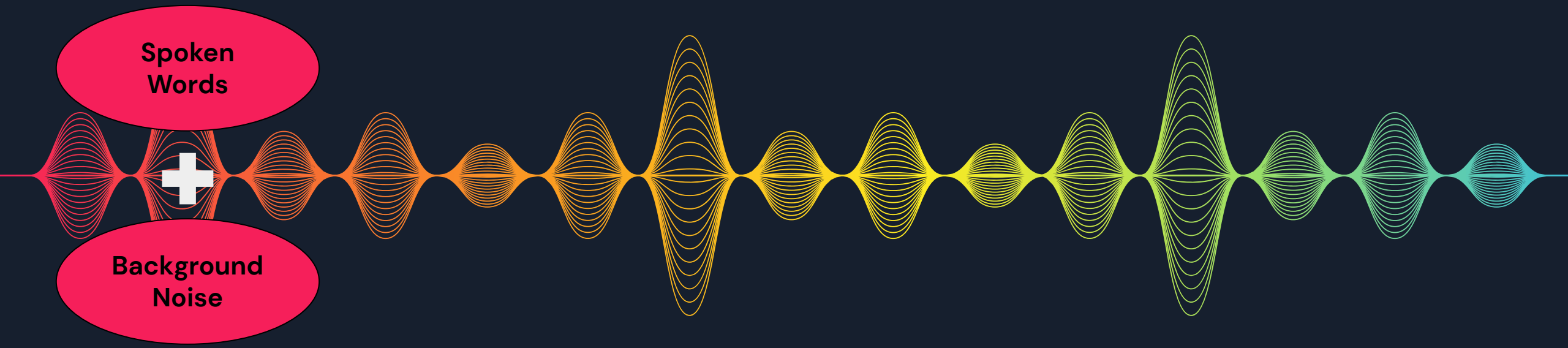
Non-Target

Target

Background



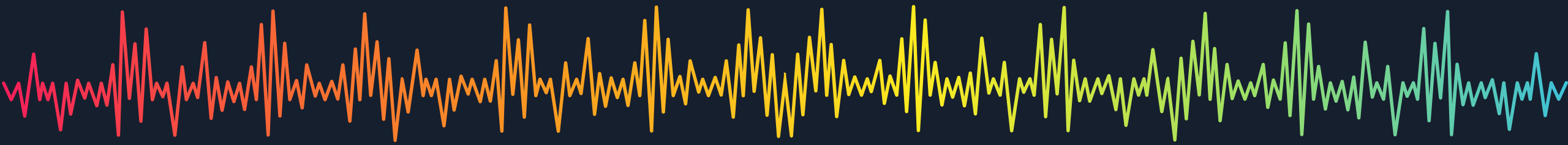
Keyword Spotting Pipeline



Which corpora did we use?

01 Multilingual Spoken Words Corpus

1-second recordings of different words, automatically extracted from continuous audio



https://mlcommons.org/en/multilingual-spoken-words/

Multilingual Spoken Word Corpus

50

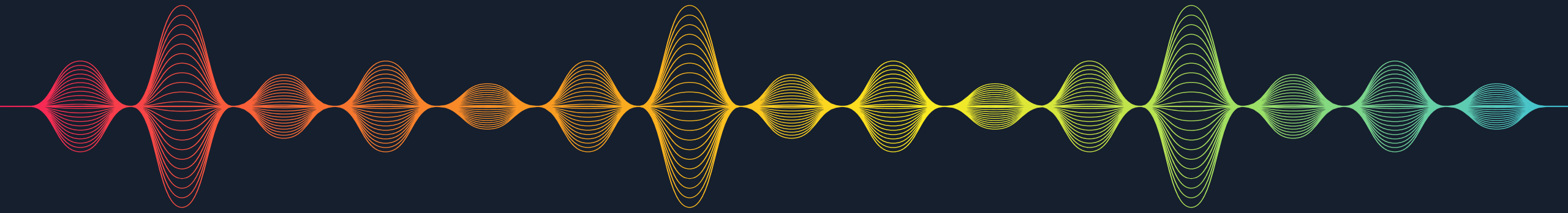
Languages

23,400,000

1-second spoken words

124GB

Full Dataset Size



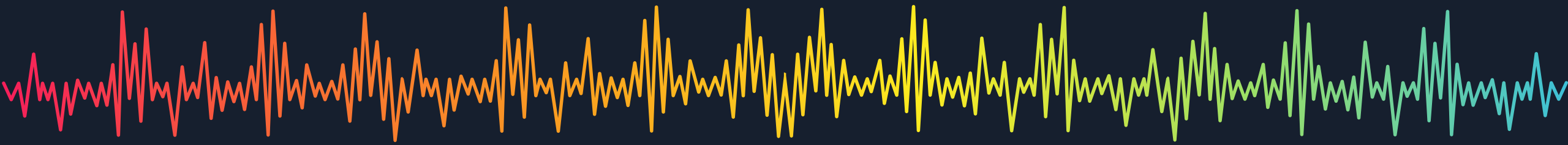
Which corpora did we use?

01 Multilingual Spoken Words Corpus

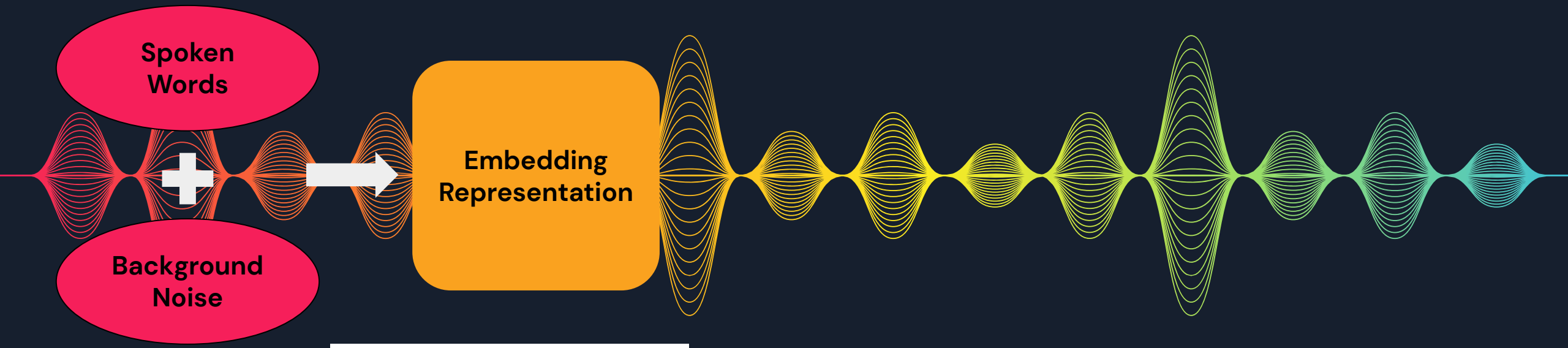
1-second recordings of different words, automatically extracted from continuous audio

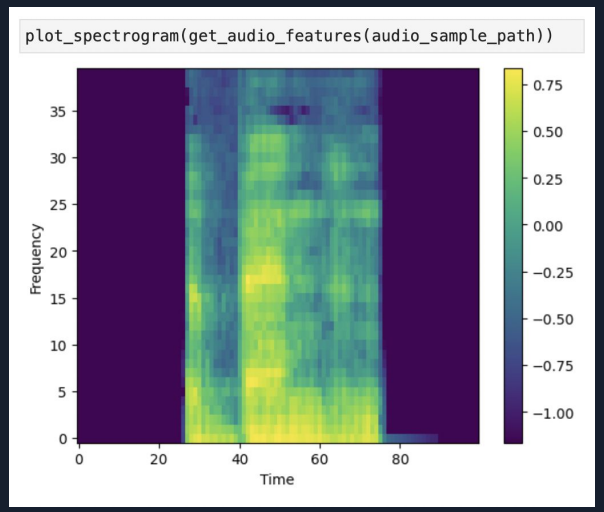
02 Google Speech Commands

Only to extract 1-second audio clips of background noise



Keyword Spotting Pipeline





Embedding Model

30 English Keywords

was	for	not	can
and	have	but	said
you	what	one	with
that	had	there	two
are	they	also	him
his	from	all	about
this	were	boy	out
its	your		

Embedding Model

Data Preparation

30

English keywords

1,000 samples each from MSWC

1

Background noise

Random extractions from GSC

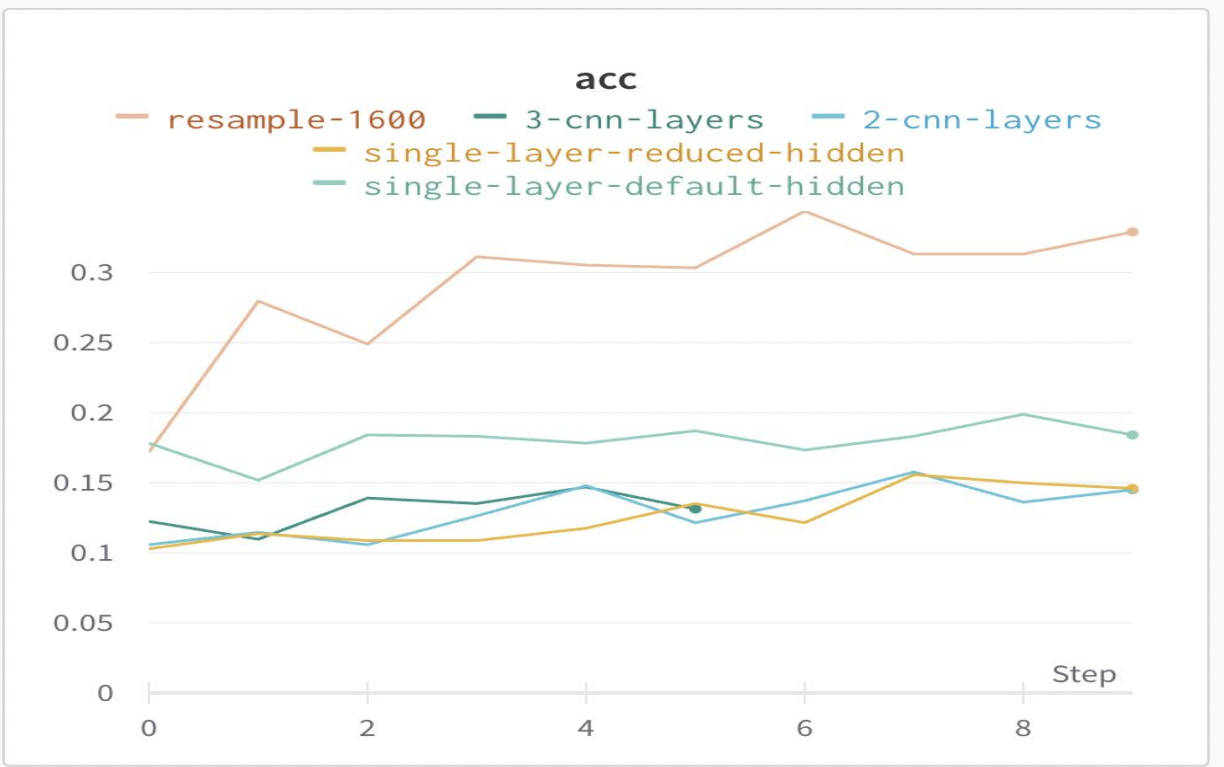
31

Category

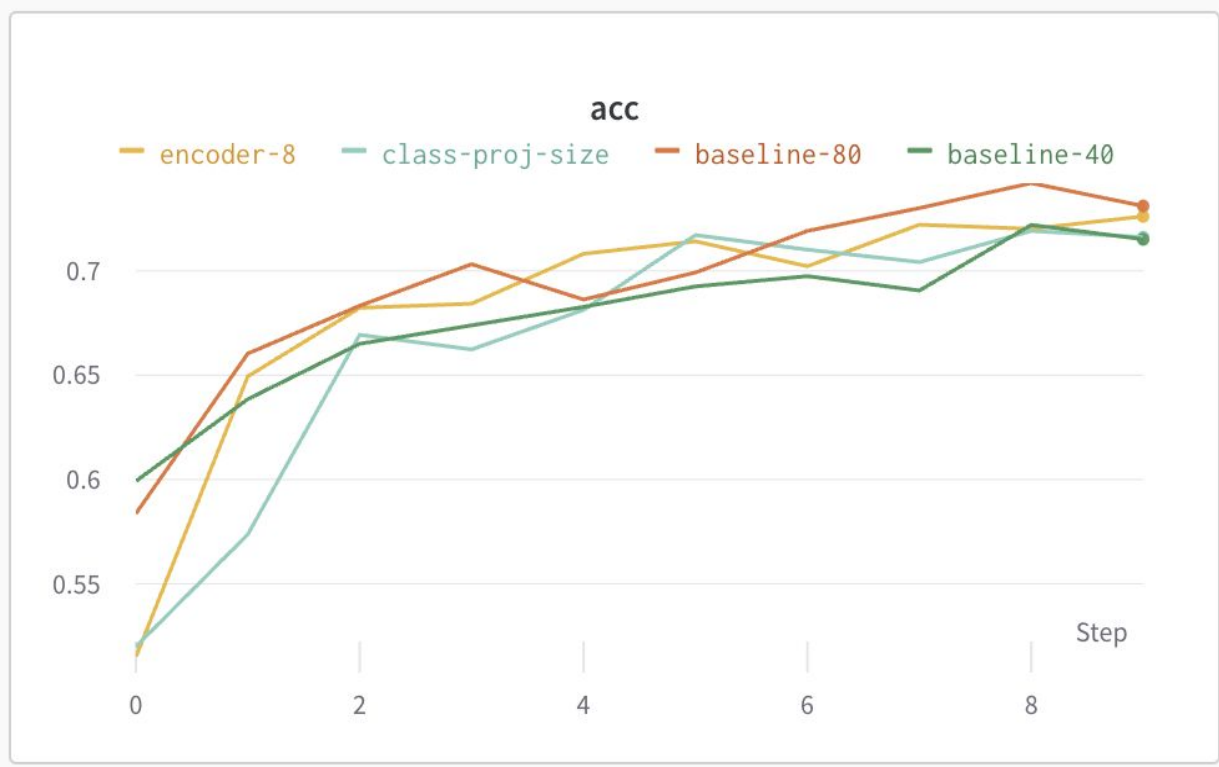
Classification task

Embedding Model

Wav2Vec2

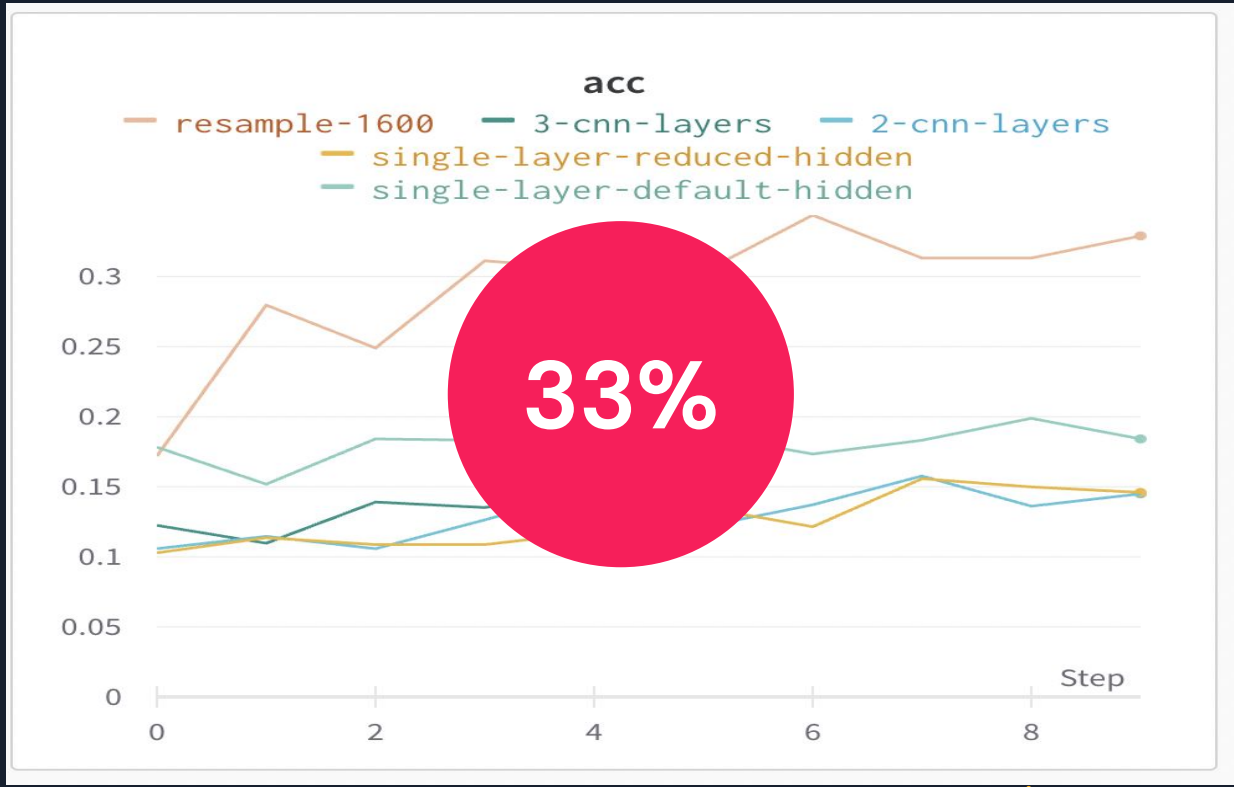


Whisper

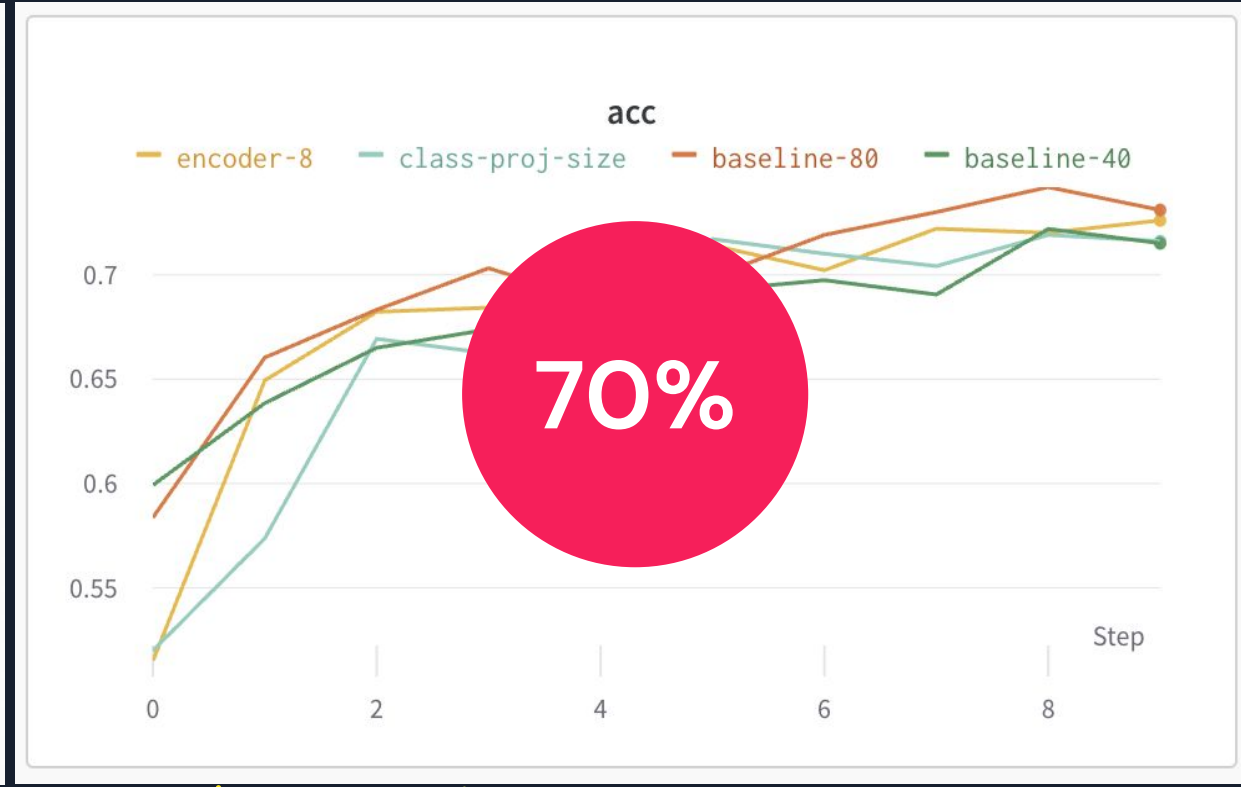


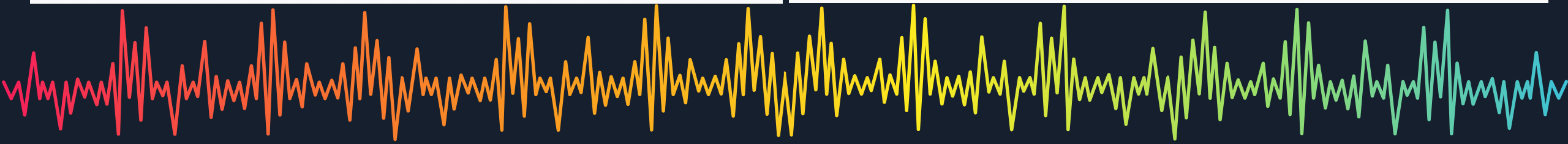
Embedding Model

Wav2Vec2

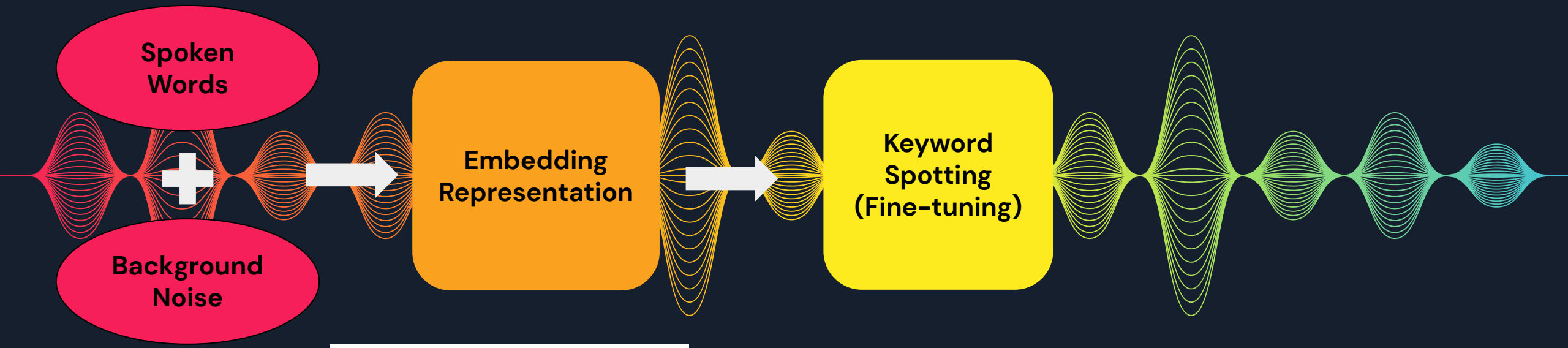


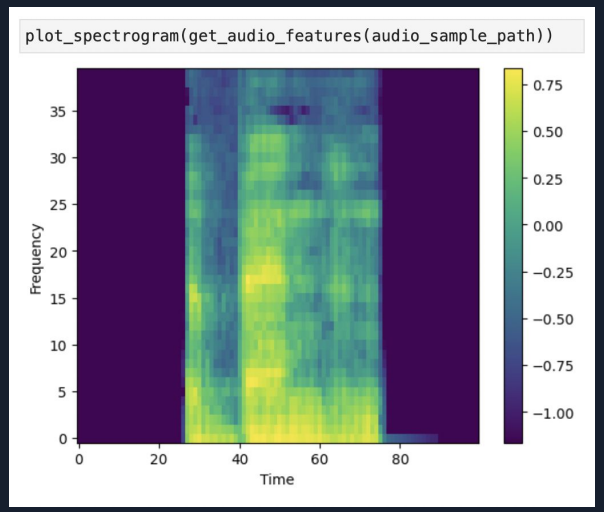
Whisper





Keyword Spotting Pipeline





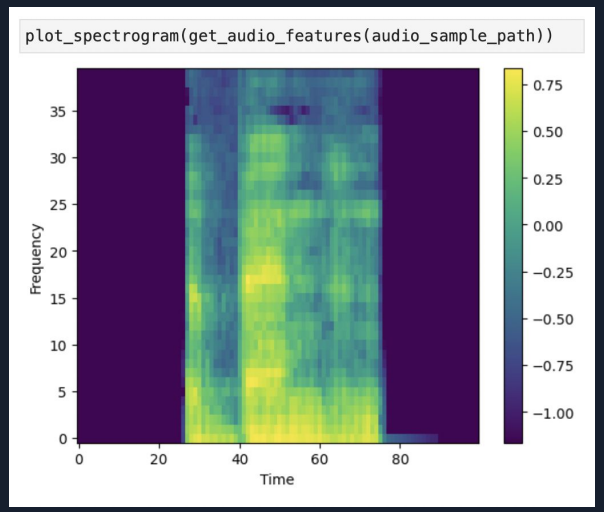
Keyword Spotting Pipeline

Spoken Words

Background Noise

Embedding Representation

Keyword Spotting (Fine-tuning)



Target Examples

Negative Examples

Keyword Spotting Pipeline

Spoken Words

Background Noise

Embedding Representation

Keyword Spotting (Fine-tuning)

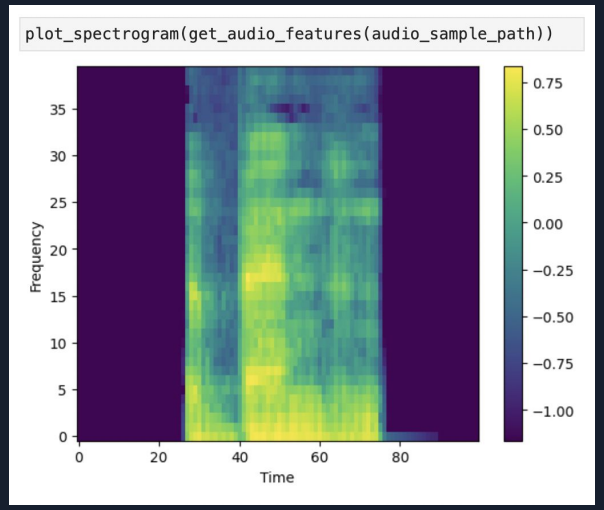
Target Examples

Negative Examples

Non-Target

Target

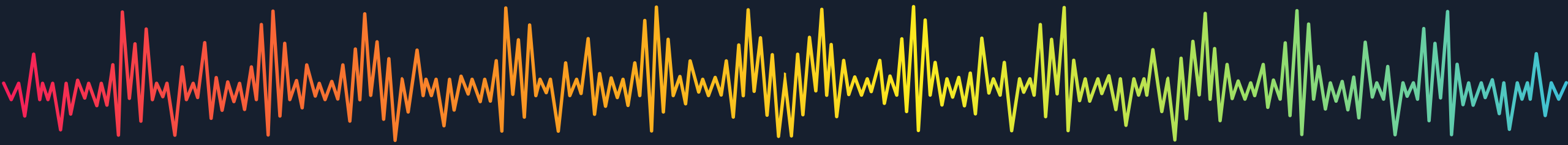
Background



3.2. Few-shot Transfer Learning

For 5-shot transfer learning (Fig. 1b), we use five target samples to fine-tune a 3-class softmax layer (with *target, unknown,* and *background* categories) on the output feature vector of the embedding layers, along with 128 non-target samples drawn from a precomputed bank of 5,000 "unknown" utterances in the nine embedding languages. When training KWS models in languages not seen by the embedding model (e.g., in Welsh), non-target samples are still drawn from this bank, i.e., to train a KWS model in Welsh a user would only need to collect 5 target samples of a Welsh keyword, without also needing to collect non-target examples in Welsh. The weights in the embedding layers are frozen when fine-tuning; we only update the softmax layer. Across 256 total training samples, approximately 45% are in the target category (random augmentations of the five target examples using the same strategy as Sec. 3.1), 45% are negative samples drawn from the precomputed set of non-target words, and 10% are background noise (Sec. 5.1.2).

"random augments of the five target examples" until you have 128 samples…



Keyword Spotting

Data Preparation

45%

Target word examples

45%

Non-target word examples

10%

Background noise

For a total of 256 data points

Keyword Spotting

English

"people"

81%

Keyword Spotting

English

"people"

81%

Chinese

"日本"

75%

Keyword Spotting

English

"people"

81%

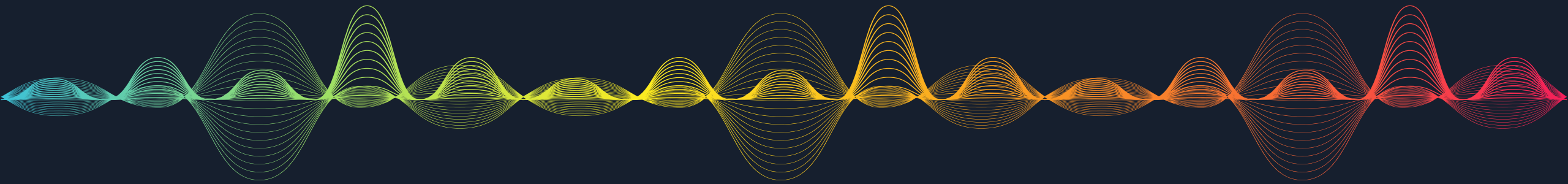
Chinese

"日本"

75%

For a total of 256 data points

Next Steps . . .



Next Steps . . .

01 Train a multilingual embedding model

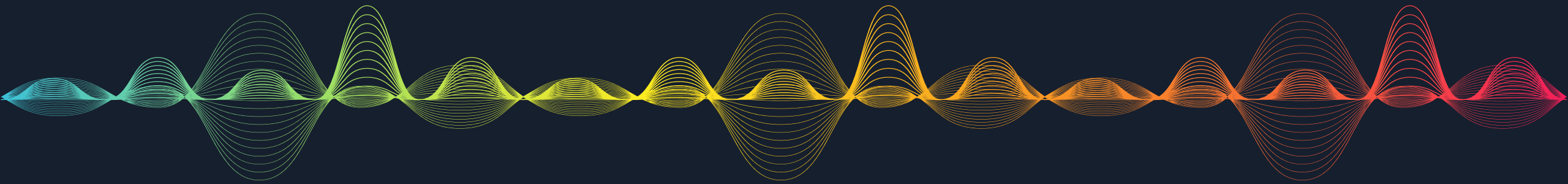
Which performs better KWS – Monolingual or Multilingual embedding?

02 Test keyword spotting on more languages

Original authors claimed this can be

03 Hyperparameter optimization

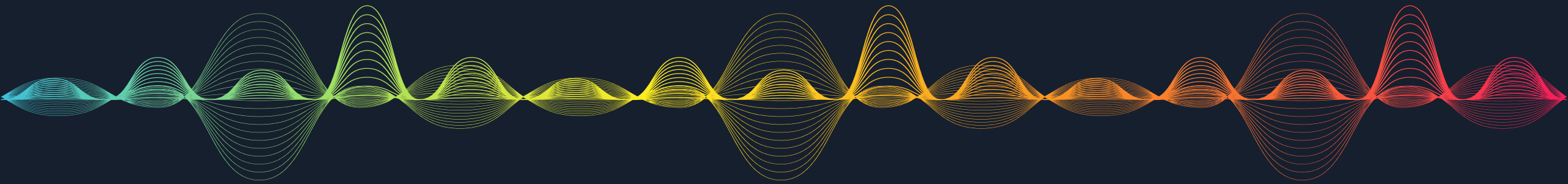
Try to improve embedding accuracy and keyword spotting accuracy



Next Steps . . .

01 Train a multilingual embedding model

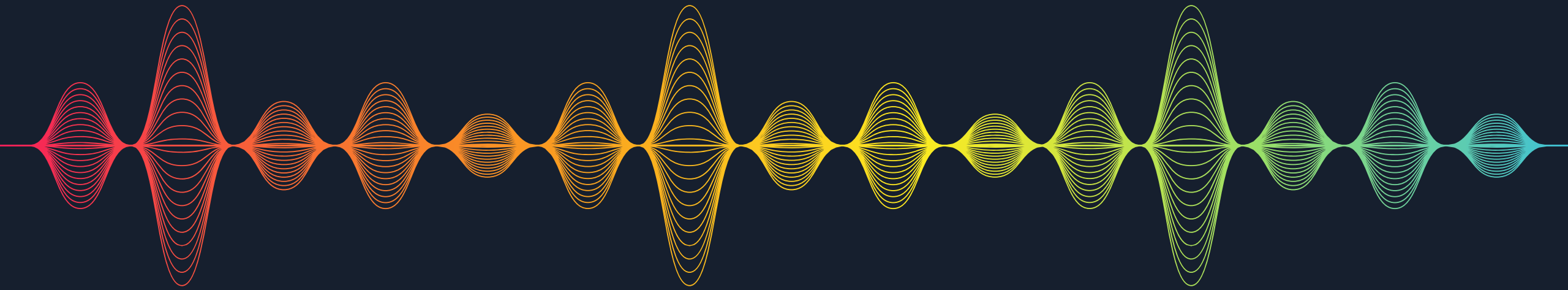
Which performs better KWS – Monolingual or Multilingual embedding?



The Wild Bunch

Multilingual Keyword Spotting

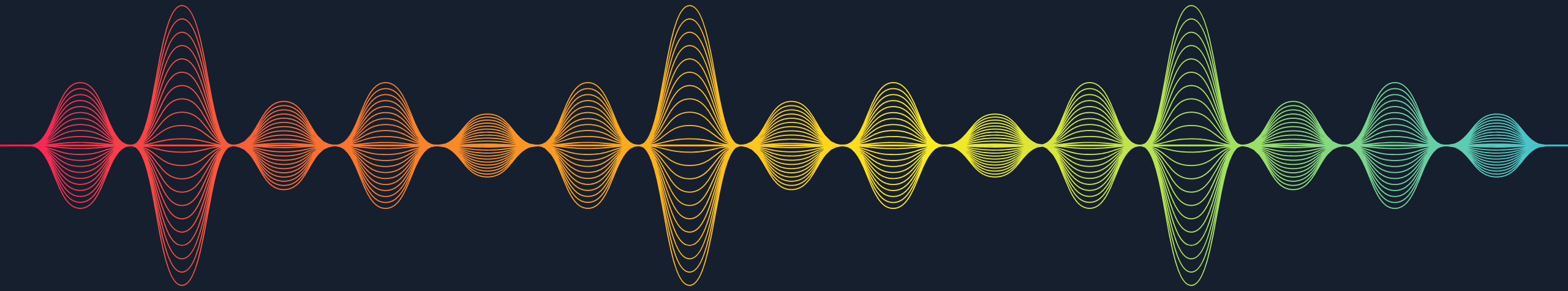
... in any language!



The Wild Bunch

Monolingual Keyword Spotting

... in any language!



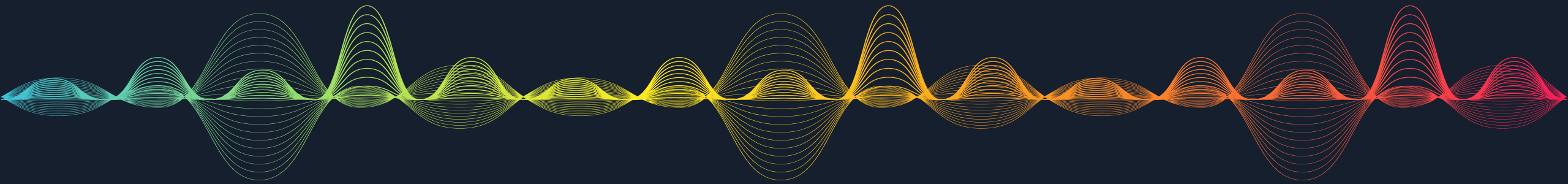
Next Steps . . .

01 Train a multilingual embedding model

Which performs better KWS – Monolingual or Multilingual embedding?

02 Test keyword spotting on more languages

Original authors claimed this model can be used for KWS in any language!



Next Steps . . .

01 Train a multilingual embedding model

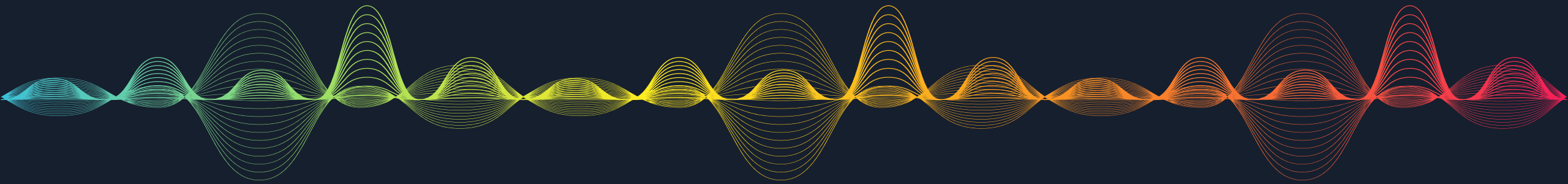
Which performs better KWS – Monolingual or Multilingual embedding?

02 Test keyword spotting on more languages

Original authors claimed this model can be used for KWS in any language!

03 Hyperparameter optimization

Try to improve embedding accuracy and keyword spotting accuracy



Thank you!

Do you have any questions?

Bingyang Hou Jae Ihn Behrooz Qiassi Min Zeng