

Plug-and-Play Multilingual Few-shot Spoken Words Recognition

Aaqib Saeed · Vasileios Tsouvalas
Eindhoven University of Technology, The Netherlands
arXiv:2305.03058 · 2023

01 Motivation

The problem

Fixed-vocabulary keyword spotting (KWS) requires new training when users add words or languages. Low-resource languages and personal vocabularies need recognition from only a few recordings.

Contributions

- **Multilingual embeddings** learn from over 12 million spoken-word clips across 20 languages.
- **Unseen-class evaluation** tests new words, five new languages and language-identification transfer.
- **Edge measurements** compare small and base encoders across three devices.

Few-shot episodes

An episode contains N candidate words and K labeled support clips per word. Queries are separate clips from those same classes; training and test vocabularies are disjoint.

Closed-set recognition

Each query belongs to one candidate word in the support set. Experiments use isolated clips; continuous listening and unknown-word rejection are outside this evaluation.

Scope

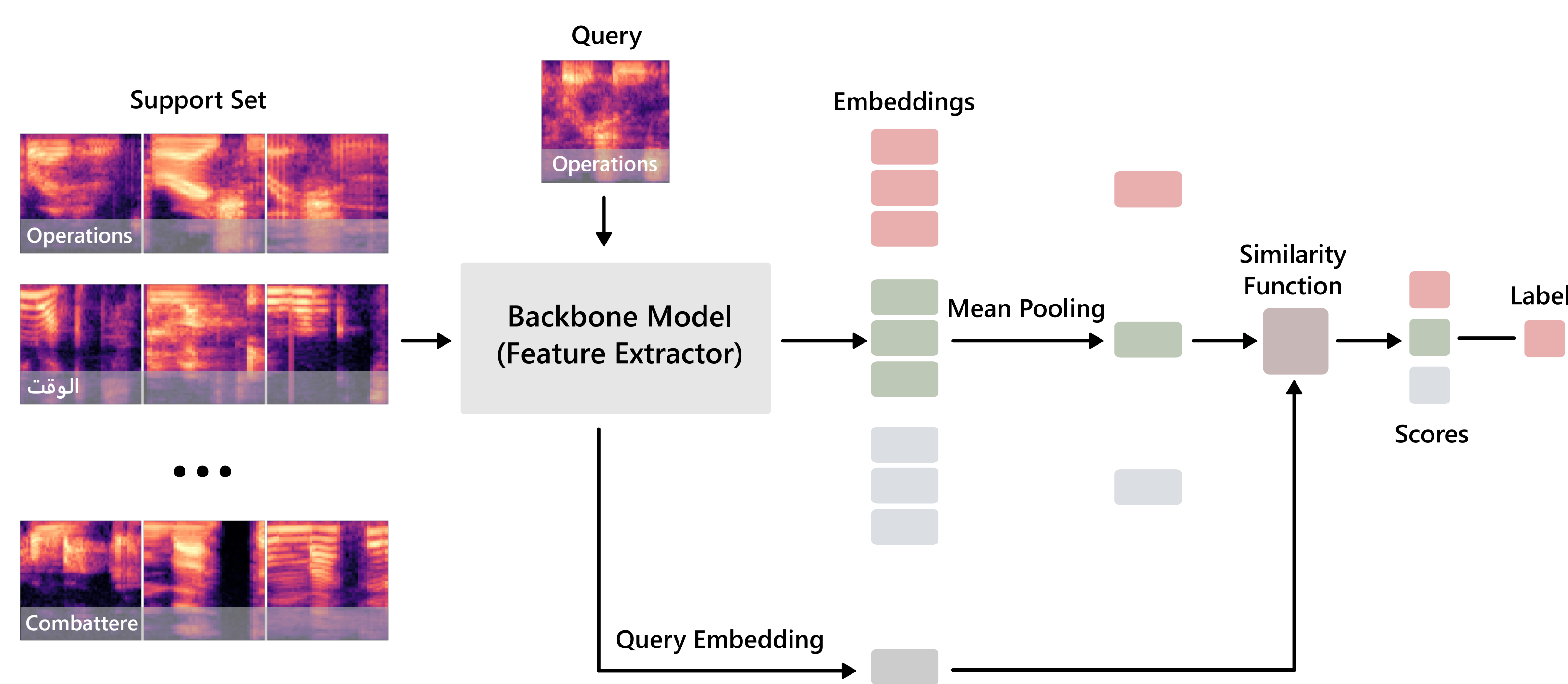
Encoder latency excludes preprocessing and classification. Recognition accuracy after quantization is not reported.

Notation

Multilingual Spoken Words Corpus (MSWC) · few-shot learning (FSL)

02 PLiX

New examples add vocabulary without retraining the encoder



Technical core

$$\mathbf{p}_i = \frac{1}{|S_i|} \sum_{(x, y) \in S_i} f_{\theta}(x)$$

S_i contains support examples of word i ; f_{θ} encodes each clip. Their mean $\mathbf{p}_i$ defines that word without updating θ .

$$\hat{y} = \arg \min_i \|f_{\theta}(x_q) - \mathbf{p}_i\|_2^2$$

Encode query x_q separately and choose the closest word prototype. Query embeddings are never averaged into $\mathbf{p}_i$.



Paper



Code

03 Evaluation

95%

80-word recognition

English-only base; five supports/word

5.12 ms

small encoder latency

Quantized TinyNet-E; Pixel 7; encoder only

More support handles larger vocabularies

English test vocabulary	Supports per word	Accuracy (%) (Ours)
40 words	1	90
80 words	1	86
60 words	5	96
80 words	5	95

English-only base model, held-out MSWC words; 1,000 episodes. Rounded mean accuracies.

Quantized encoder latency

Device	Small (ms)	Base (ms)
Pixel 7	5.12	78.07
Raspberry Pi 4	6.92	107.13
Jetson Nano	7.32	101.37

Quantized TinyNet-E / EfficientNet-v2 Medium; 200-run mean encoder latency. Excludes log-mel preprocessing and prototype classification.

Takeaway

New support examples add vocabulary without encoder retraining, with accuracy depending on language, vocabulary size and support count. The **5.12 ms** Pixel 7 result measures only the separate small encoder, excluding preprocessing and classification.