Open-vocabulary Keyword-spotting with Adaptive Instance Normalization

A. Navon A. Shamsian N. Glazer G. Hetz J. Keshet

aiOla Research

Background

Keyword spotting (KWS):

- **Goal:** Identify specific keywords within an audio utterance.
- Keywords are predefined.

Background

Keyword spotting (KWS):

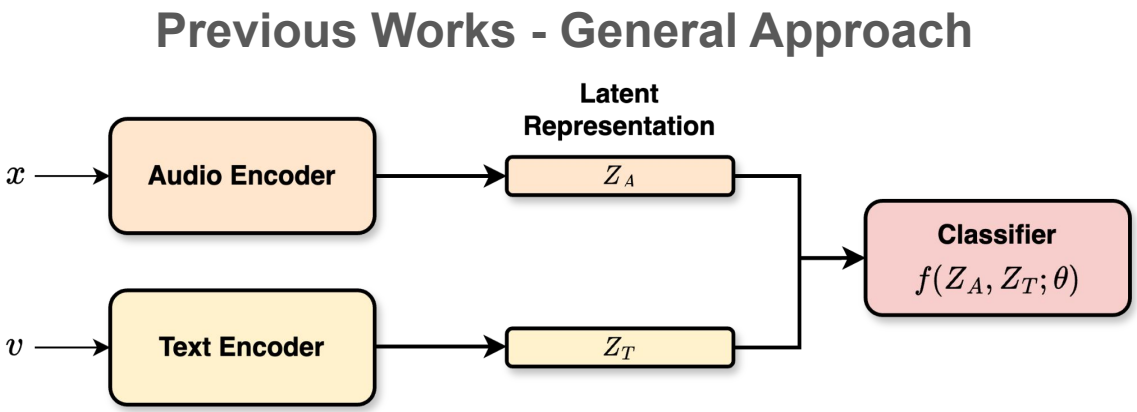
- **Goal:** Identify specific keywords within an audio utterance.
- Keywords are predefined.

Open-vocabulary KWS:

- **Goal:** Detect arbitrary keywords at inference.
- **Motivation:**
  - Personalization, user-defined keywords.
  - No need for additional training to incorporate novel keywords.
  - Low resource languages, specialized domains and jargon.

Overview

- Embed textual and auditory input into a joint space.
- Detect keywords based on audio-text similarity.

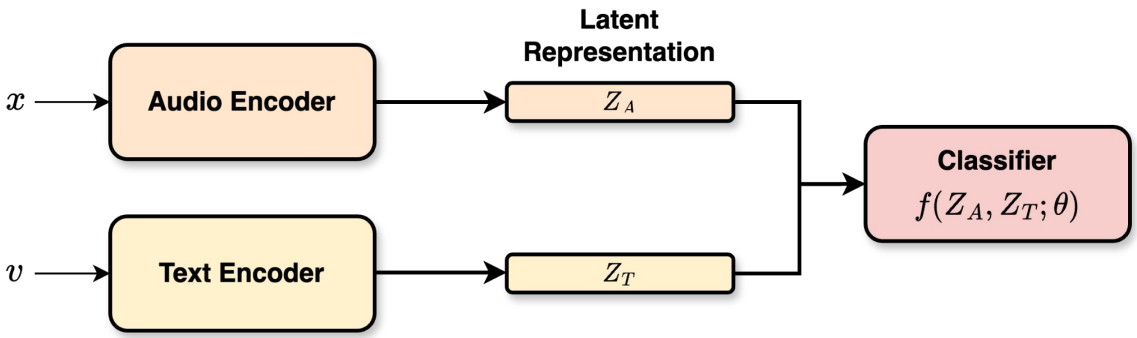


Overview

Key Limitation

Using two encoders for representing heterogeneous modalities in a joint space may cause significant misalignment between audio and text embeddings.

Previous Works - General Approach



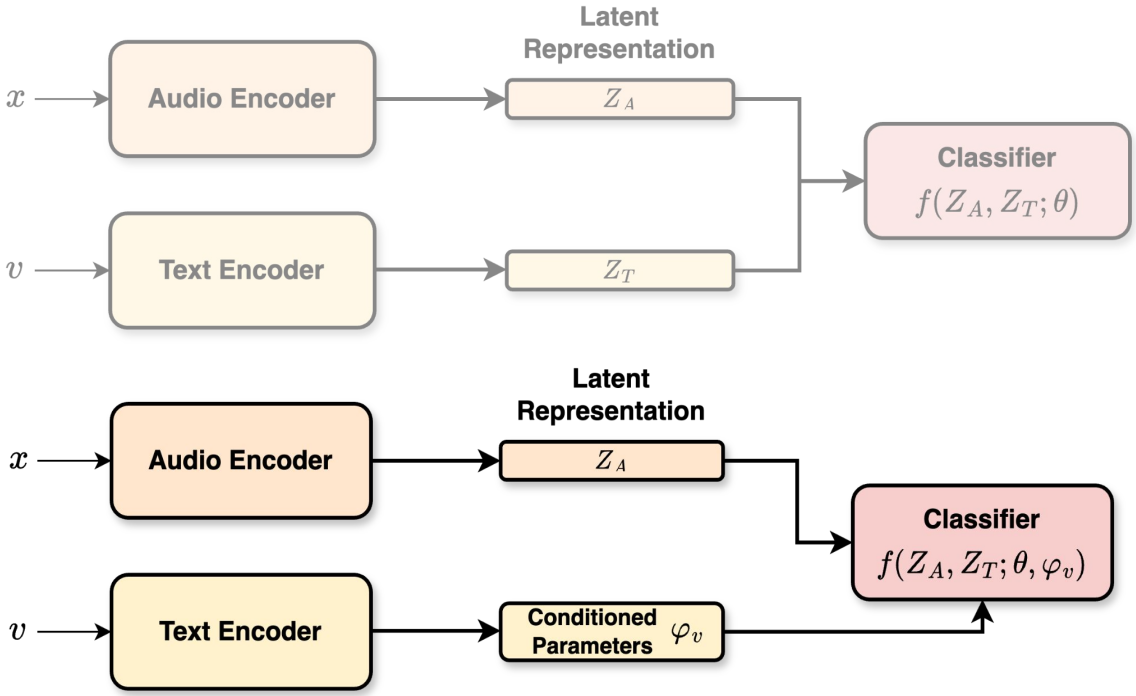
Overview

Key Limitation

Using two encoders for representing heterogeneous modalities in a joint space may cause significant misalignment between audio and text embeddings.

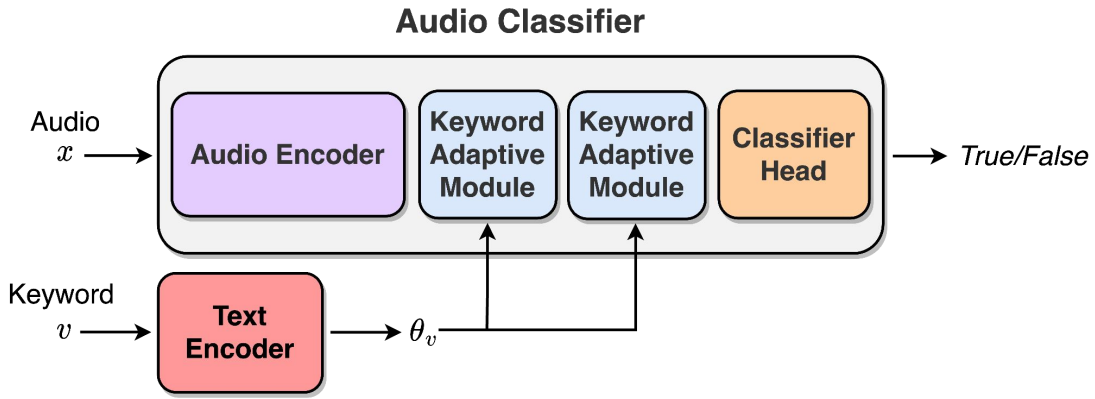
We address this limitation by adaptively condition parameters on query keywords.

Previous Works - General Approach



The AdaKWS Model

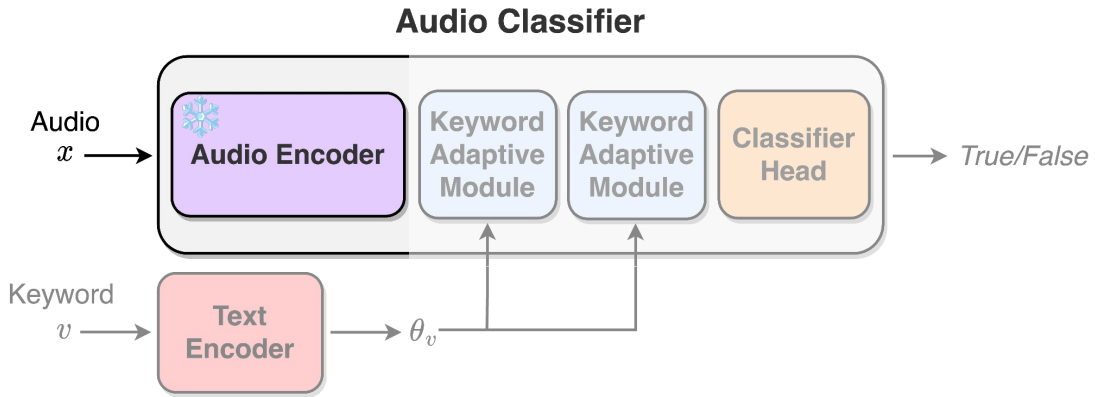
The audio is processed using the classifier which is conditioned on the keyword through the keyword-adaptive modules.



The AdaKWS Model

The audio is processed using the classifier which is conditioned on the keyword through the keyword-adaptive modules.

The **Audio Encoder** is a pre-trained Whisper encoder.

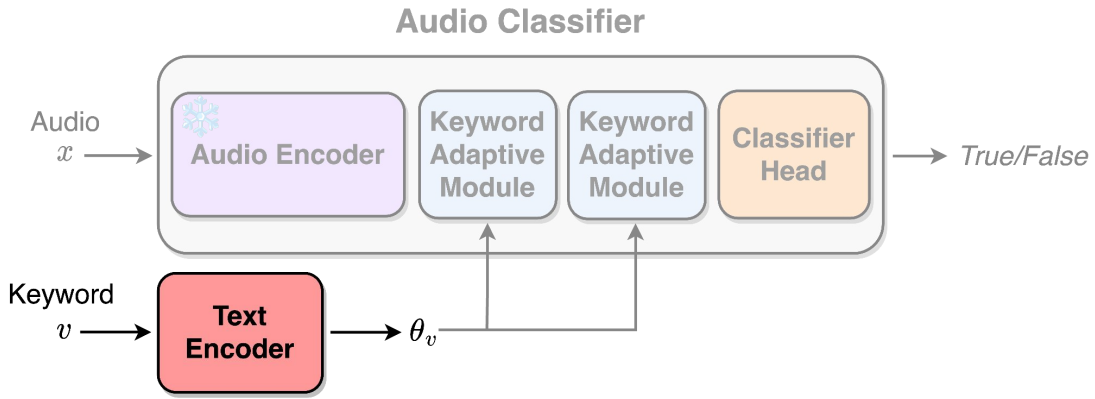


Robust Speech Recognition via Large-Scale Weak Supervision, Radford et al., 2022.

The AdaKWS Model

The audio is processed using the classifier which is conditioned on the keyword through the keyword-adaptive modules.

The **Text Encoder** outputs a set of keyword-conditioned normalization parameters.

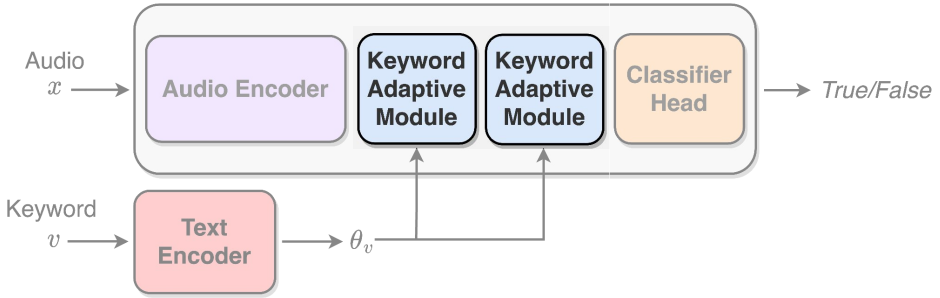


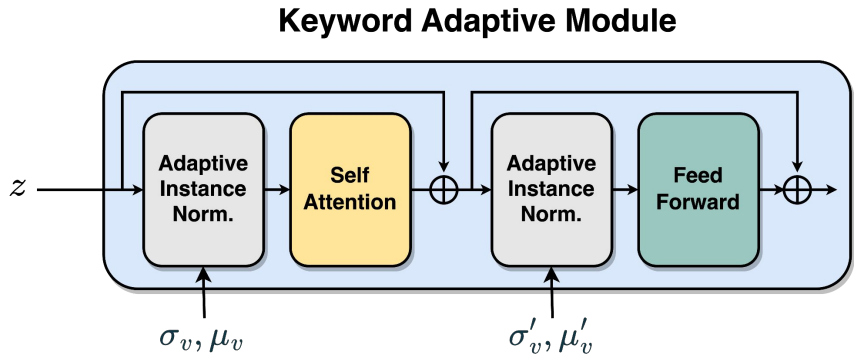
Keyword Adaptive Module

Each module is a standard transformer encoder block in which we swap the Layer Normalization layers with **Adaptive Instance Normalization layers (AdaIN)**. An **AdaIn** is a normalization layer of the form:

$$\text{AdaIN}(z, v) = \sigma_v \left(\frac{z - \mu_z}{\sigma_z} \right) + \mu_v$$

z is the audio representation and
v is the target keyword





aiOla

Online Keyword Sampling

Training examples are pairs of speech utterances and keywords.

Online Keyword Sampling - Positive

Training examples are pairs of speech utterances and keywords.

Positive Examples

Random keyword phrase from the speech transcription.

Online Keyword Sampling - Positive

Training examples are pairs of speech utterances and keywords.

Consider the utterance:

> *“Open vocabulary keyword spotting with adaptive instance normalization”*

Positive Examples

Random keyword phrase from the speech transcription.

Online Keyword Sampling - Positive

Training examples are pairs of speech utterances and keywords.

Consider the utterance:

“Open vocabulary keyword spotting with adaptive instance normalization”

Positive Examples

Random keyword phrase from the speech transcription:

keyword

Online Keyword Sampling - Positive

Training examples are pairs of speech utterances and keywords.

Consider the utterance:

“Open vocabulary keyword spotting with adaptive instance normalization”

Positive Examples

Random keyword phrase from the speech transcription:

instance normalization

Online Keyword Sampling - Negative

“Open vocabulary keyword spotting with adaptive instance normalization”

Random Negative

Random keyword phrase from different speech utterances in the batch:

animals, drink, neighborhood

Online Keyword Sampling - Negative

“Open vocabulary keyword spotting with adaptive instance normalization”

Nearest Keyword

Represent each keyword with last layer’s embedding representation.

Querying for the keyword with the smallest cosine distance within training batch:

instance → *distance*, *keyword* → *backward*

Online Keyword Sampling - Negative

“Open vocabulary keyword spotting with adaptive instance normalization”

Keyword Concatenation

Concatenating a positive keyword with a random negative keyword:

spotting → *spotting animals*, *animals spotting*

Online Keyword Sampling - Negative

“Open vocabulary keyword spotting with adaptive instance normalization”

Character Swapping

Alter a positive keyword by substituting one or more characters.

Character can be chosen randomly, or according to an a priori mapping of acoustically similar characters (“*s*”→“*z*”,“*p*”→“*b*”, etc.).

spotting → *zbotting*, *adaptive* → *adantive*

Experimental Setup

- Datasets:
  - VoxPopuli dataset - multilingual corpus with transcribed utterances from 16 languages.
  - Keyword spotting benchmark: LibriPhrase easy (LE) and hard (LH).
  - Zero-shot evaluation: Fleurs and multilingual-LibriSpeech.
- Several model sizes: AdaKWS-Tiny, AdaKWS-Base, AdaKWS-Small.
- Diverse and challenging evaluation sets, with random, concat and swap negatives.
- ASR-based and KWS baselines.

Multilingual KWS - Voxpopuli

F1 results for the VoxPopuli test dataset:

	CS	DE	EN	ES	ET	FI	FR	HR	HU
Whisper-Tiny	48.5	77.0	83.4	85.6	39.5	58.4	78.3	48.6	47.0
Whisper-Small	77.9	89.3	84.1	86.3	55.2	81.5	90.5	72.5	72.5
Whisper-Large-V2	91.0	93.2	80.5	87.8	78.5	89.1	93.2	67.9	87.4
AdaKWS-Tiny	91.8	94.4	95.5	95.1	86.2	92.1	94.9	89.3	90.6
AdaKWS-Base	92.3	95.2	95.9	95.1	85.2	92.2	**96.3**	90.4	91.3
AdaKWS-Small	**94.4**	**95.7**	**96.3**	**95.6**	**91.1**	**94.8**	95.9	**93.0**	**92.7**

	IT	LT	NL	PL	RO	SK	SL	Overall	# Params
Whisper-Tiny	72.2	38.4	67.8	74.7	52.0	43.8	46.0	69.3	39M
Whisper-Small	85.4	62.5	85.8	89.7	77.3	68.8	64.3	83.3	244M
Whisper-Large-V2	87.7	76.1	**92.6**	93.8	87.7	84.0	80.0	88.4	1550M
AdaKWS-Tiny	90.8	**85.6**	91.1	93.0	91.9	91.7	85.6	92.8	15M
AdaKWS-Base	92.0	82.1	91.5	94.3	92.9	93.9	88.5	93.7	31M
AdaKWS-Small	**92.8**	78.5	92.2	**95.5**	**94.8**	**94.5**	**89.9**	**94.6**	109M

aiOla

Effect of Negative Sampling Methods

Results for the AdaKWS-tiny model, trained using different negative sampling approaches, on the VoxPopuli test dataset:

	F1 ↑	AUC ↑	EER ↓
Random	79.37	88.09	19.37
Random + NK	81.24	88.79	18.80
Random + NK + Cat.	84.99	91.20	15.91
Random + NK + Cat. + Swap	**92.87**	**97.78**	**7.16**

KWS Benchmark - LibriPhrase

Results for the LibriPhrase easy (LE) and LibriPhrase hard (LH) test dataset:

	AUC (%) ↑		EER (%) ↓	
	LH	LE	LH	LE
Triplet	54.88	63.53	44.36	32.75
Attention	62.65	78.74	41.95	28.74
DONUT	54.88	63.53	44.36	32.75
CMCD	73.58	96.70	32.90	8.42
EMKWS	84.21	97.83	23.36	7.36
CED	92.70	**99.84**	14.40	1.70
AdaKWS-Tiny	93.75	99.80	13.47	1.61
AdaKWS-Base	94.39	99.81	12.60	1.37
AdaKWS-Small	**95.09**	99.82	**11.48**	**1.21**

Zero-shot Performance - Multilingual LibriSpeech

F1 results for the Multilingual LibriSpeech test dataset:

	DE	EN	ES	FR	IT	NL	PL	Overall	# Params	Inference Time (MS) ↓
Whisper-Tiny	77.6	85.8	83.6	73.4	72.4	69.2	76.0	77.9	39M	260
Whisper-Small	90.8	91.3	92.9	88.2	85.8	86.1	91.5	89.6	244M	621
Whisper-Large-V2	**95.0**	**93.8**	**95.1**	93.8	91.7	**92.3**	**95.2**	**93.9**	1550M	1836
AdaKWS-Tiny	92.6	91.4	92.1	91.9	92.3	87.9	92.6	91.3	15M	6
AdaKWS-Base	93.8	92.6	92.2	92.8	92.7	90.0	92.8	92.4	31M	7
AdaKWS-Small	94.6	**93.8**	94.4	**94.1**	**94.2**	91.8	94.4	93.8	109M	11

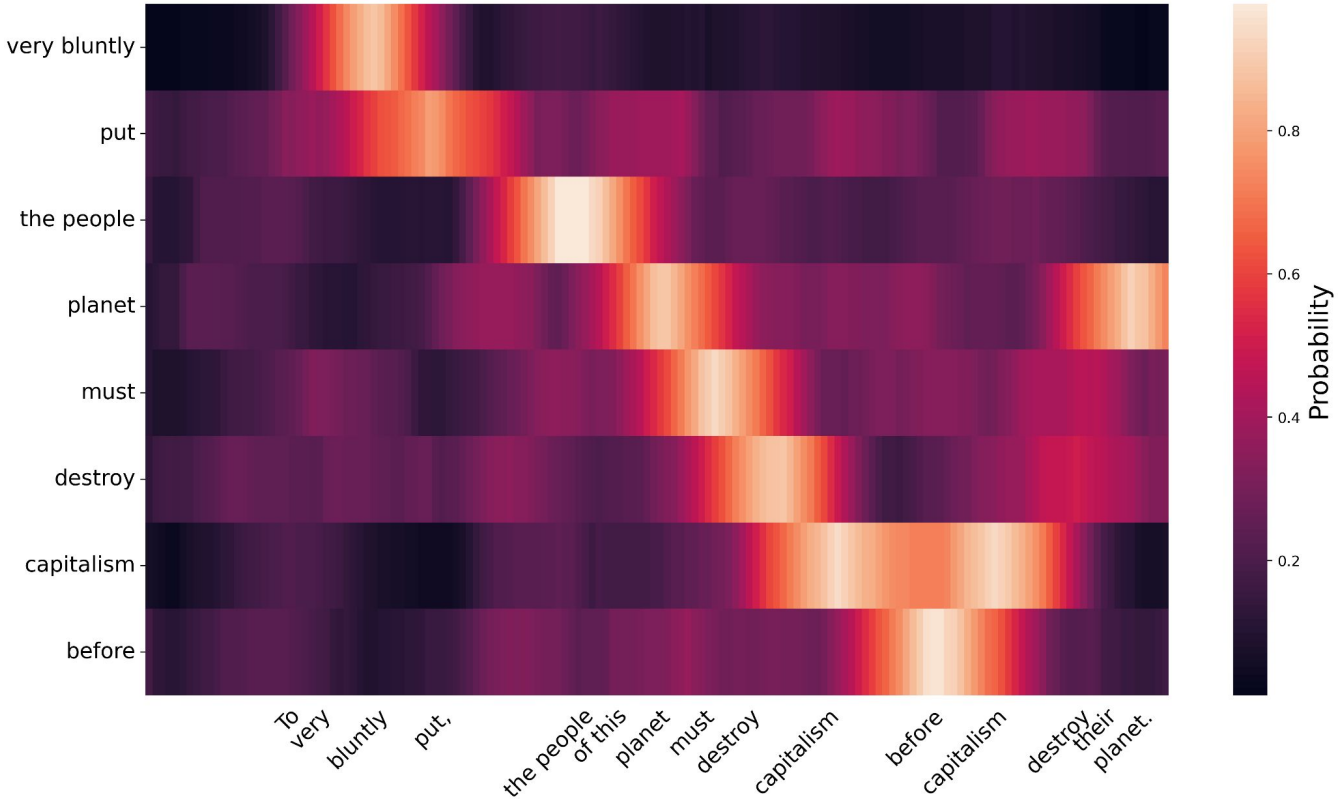
aiOla

Generalization to Novel Languages - Fleurs

Zero-shot performance for low resource, novel languages: F1 results for four languages from the Fleurs test dataset:

	Icelandic	Maltese	Swahili	Uzbek	Overall
Whisper-Tiny	42.3	37.6	43.9	34.2	37.8
Whisper-Small	52.0	34.9	55.6	35.4	40.4
Whisper-Large-V2	69.2	38.6	**74.9**	36.6	48.1
AdaKWS-Tiny	69.2	**76.7**	70.1	**67.5**	**71.9**
AdaKWS-Base	**71.5**	**76.7**	71.6	65.7	71.6
AdaKWS-Small	70.6	76.0	69.4	66.1	70.9

Visualization of AdaKWS Prediction



aiOla

Conclusion

- We present AdaKWS, a novel open-vocabulary KWS model.
- We present a new keyword-adaptive normalization layers.
- We propose several online negative sampling techniques and evaluate their effectiveness.
- We evaluate AdaKWS on a diverse set of multilingual benchmarks, as well as generalization to novel datasets and languages.

Thank you
for listening!

aiOla