Investigating the Emergent Audio Classification Ability of Whisper

Rao Ma

Speech Group, ALTA institute, University of Cambridge

December 2023

Team Members

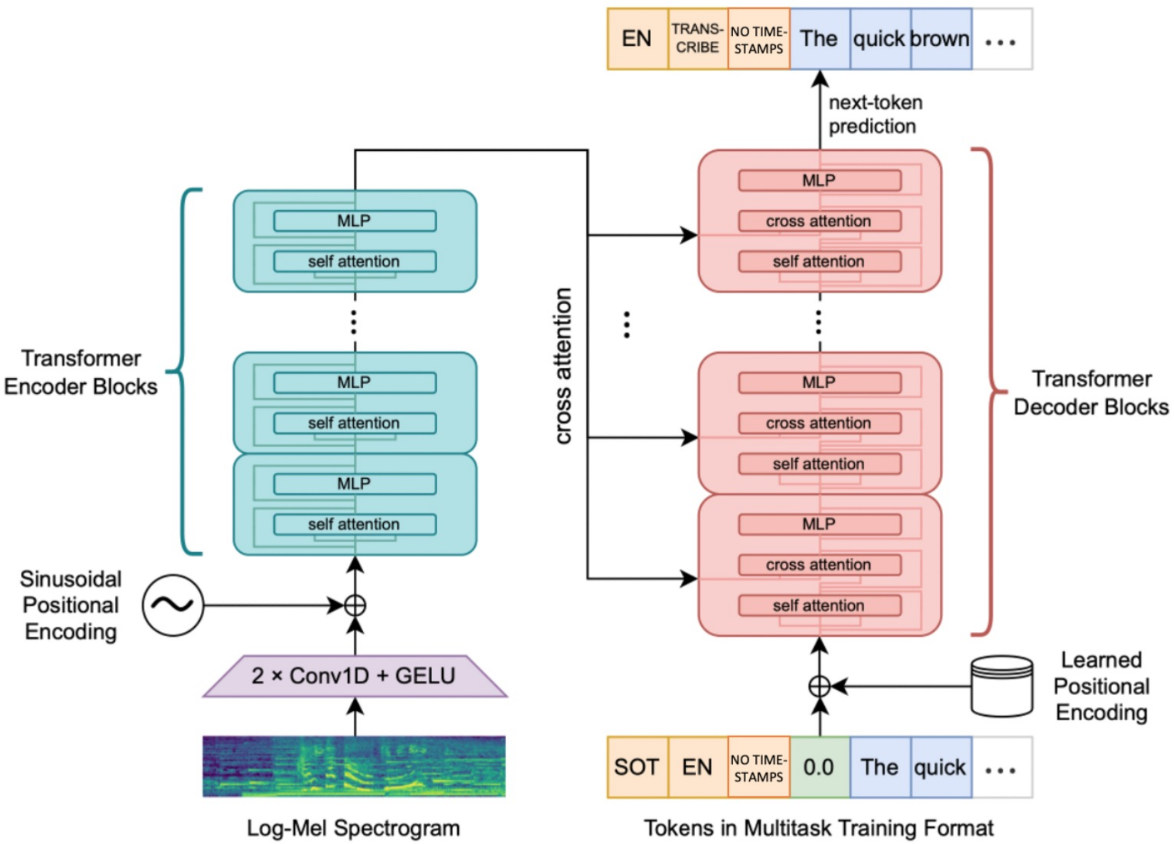

Rao Ma


Adian Liusie


Dr Kate Knill


Prof Mark Gales

Whisper

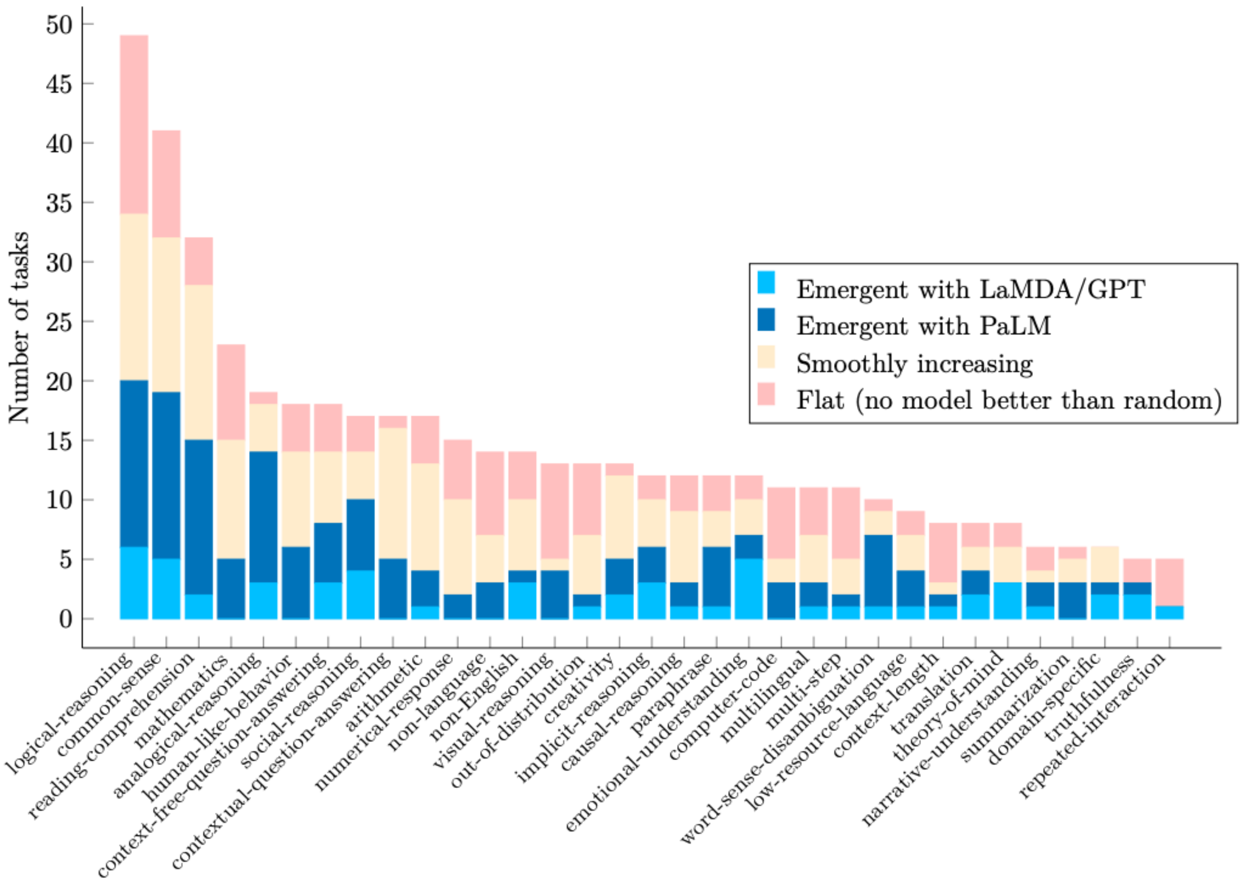


- 680,000 hours of web data
- Multitask training:
  - multilingual ASR
  - speech translation
  - language identification
  - voice activity detection
- Different sizes:
  - tiny – 39M
  - base – 74M
  - small – 244M
  - medium – 769M
  - large – 1550M

Speech Recognition with Whisper

Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

Emergent Ability of Foundation Models



Whisper → new tasks?

Related Work

Table 1: *Summary of our proposed prompts and relative improvement over the default prompts. The differences between our prompt and the default are in* ***bold****. In the AVSR task, CLIP retrie. stands for "CLIP retrieved objects", and <default> stands for <|sot|><|en|><|asr|>, please find detailed description of our prompt for AVSR in section 3. For each task only one case is shown in the table, and similar improvements are shown across different datasets and languages in the main text.*

Task	Language(s)	Default prompt	Our proposed prompt	Improvement
AVSR	En	<|sot|><|en|><|asr|>	**<|sop|>CLIP retrie.**<default>	9%
CS-ASR	Zh+En	<|sot|><|zh|>or<|en|><|asr|>	<|sot|>**<|zh|><|en|>**<|asr|>	19%
ST	En→Ru	<|sot|><|ru|><|st|>	<|sot|><|ru|>**<|asr|>**	45%

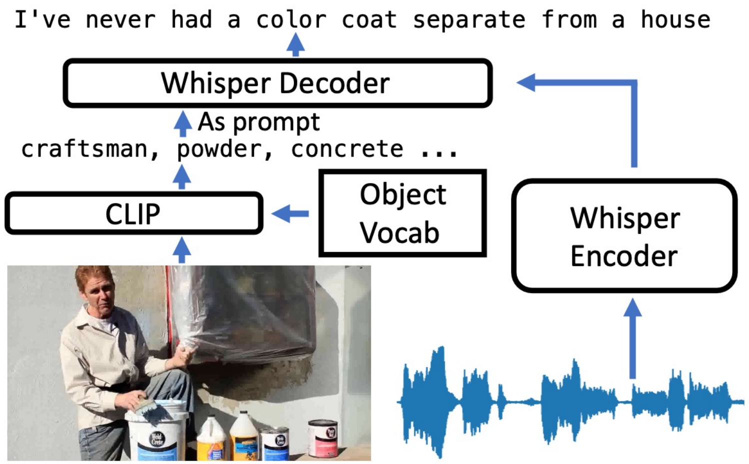
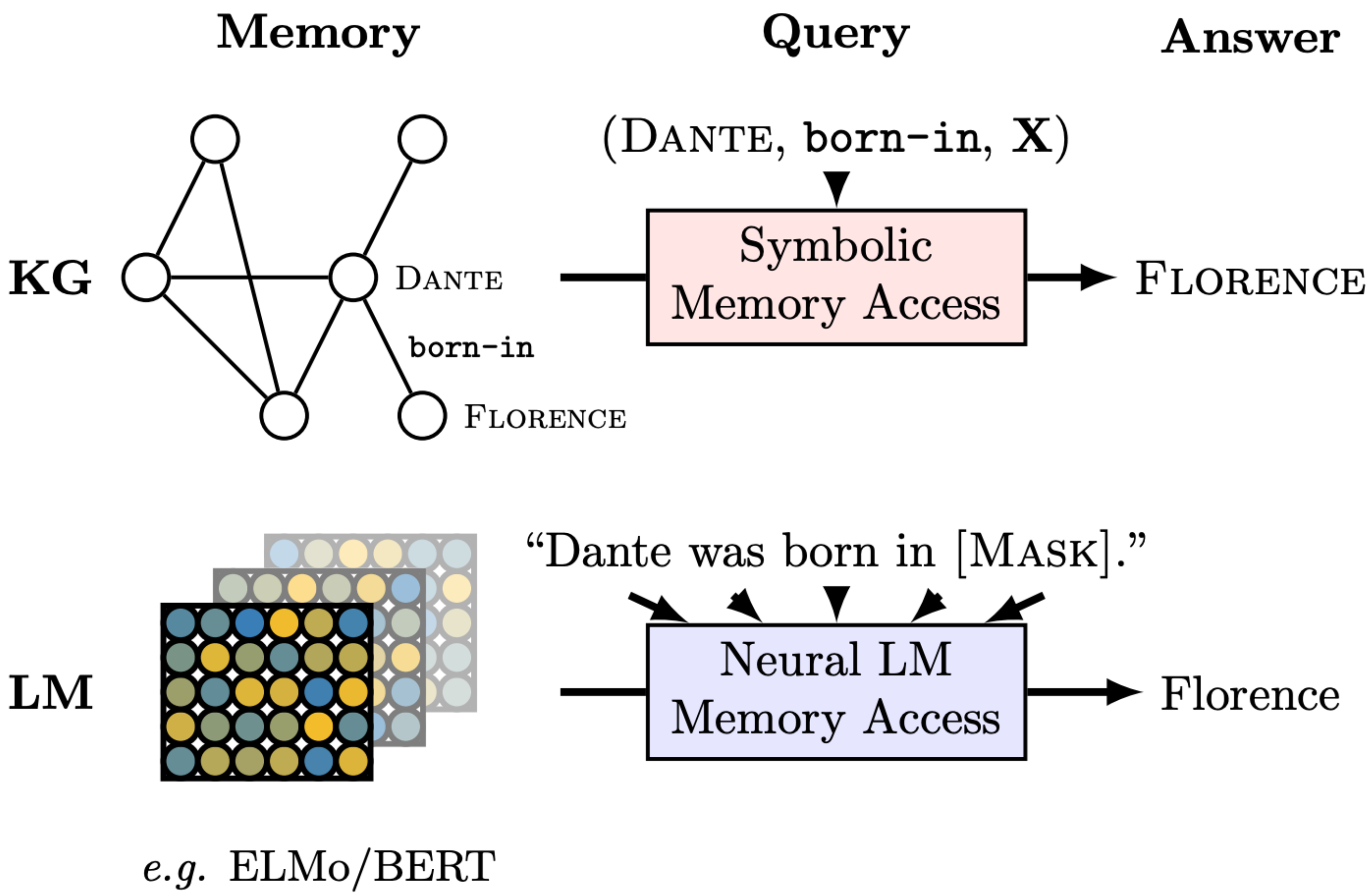


Figure 1: *Framework for visually prompting Whisper. The external object vocab is dataset agnostic.*

Related Work



Audio Classification Tasks

Sound Event Classification

siren / wind / dog / ...

Acoustic Scene Classification

home / tram / office / ...

Vocal Sound Classification

cough / sniff / laugh / ...

Emotion Recognition

angry / happy / sad / ...

Music Genre Classification

blues / jazz / pop / ...

Speaker Counting

0 / 1 / 2 / 3 / 4 / 5 / ...

Task	Dataset	Utts	Avg. Dur.	K
SEC	ESC50	2,000	5.0	50
	UrbanSound8K	8,732	3.6	10
ASC	TUT2017	1,620	10	15
VSC	Vocal Sound	3,594	5.0	6
ER	RAVDESS	1,440	3.7	8
	CREMA-D	7,442	5.0	6
MGC	GTZAN	1,000	30	10
SC	LibriCount	5,720	5.0	11

Prompting Whisper for Audio Classification

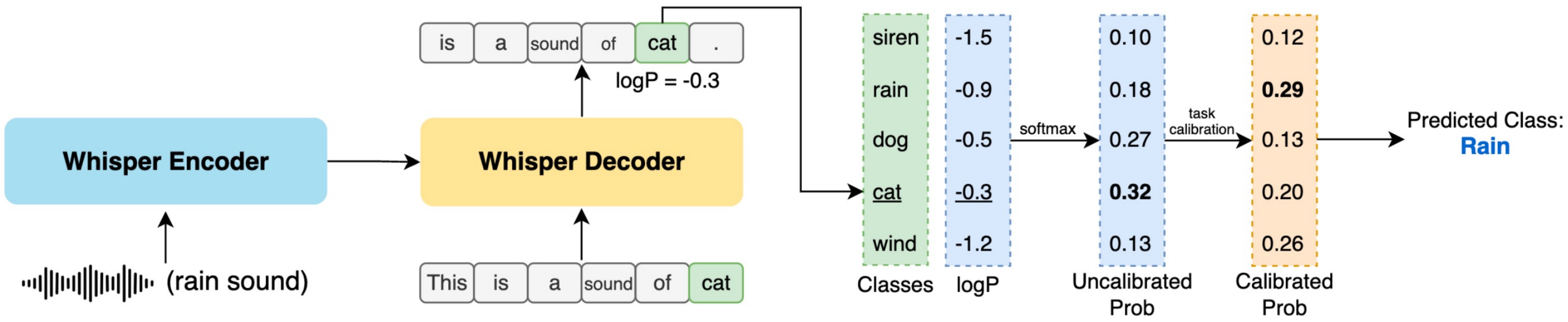


Figure 2: ASR foundation models are leveraged for zero-shot audio classification by prompting the decoder to calculate the log-likelihood of label sequences associated with each class. The log-likelihood for each class is converted to probabilities and post-processed to a predicted class. This process is displayed for Whisper.

Task Calibration: Prior-matching

$$\tilde{P}_\theta(y_k|x) = \frac{P_\theta(t(y_k)|x)}{\sum_{y_j} P_\theta(t(y_j)|x)} \tag{1}$$

$$\hat{P}_\theta(y_k|x, \alpha_{1:K}) = \frac{\alpha_k \tilde{P}_\theta(y_k|x)}{\sum_i \alpha_i \tilde{P}_\theta(y_i|x)} \tag{2}$$

$$\hat{P}_\theta(y_k|\alpha_{1:K}) = \mathbb{E}_x\{\hat{P}_\theta(y_k|x, \alpha_{1:K})\} \tag{3}$$

$$\bar{\alpha}_{1:K} = \underset{\alpha_{1:K}}{\text{argmin}} \sum_{\forall y_k} |\hat{P}_\theta(y_k|x, \alpha_{1:K}) - \frac{1}{K}| \tag{4}$$

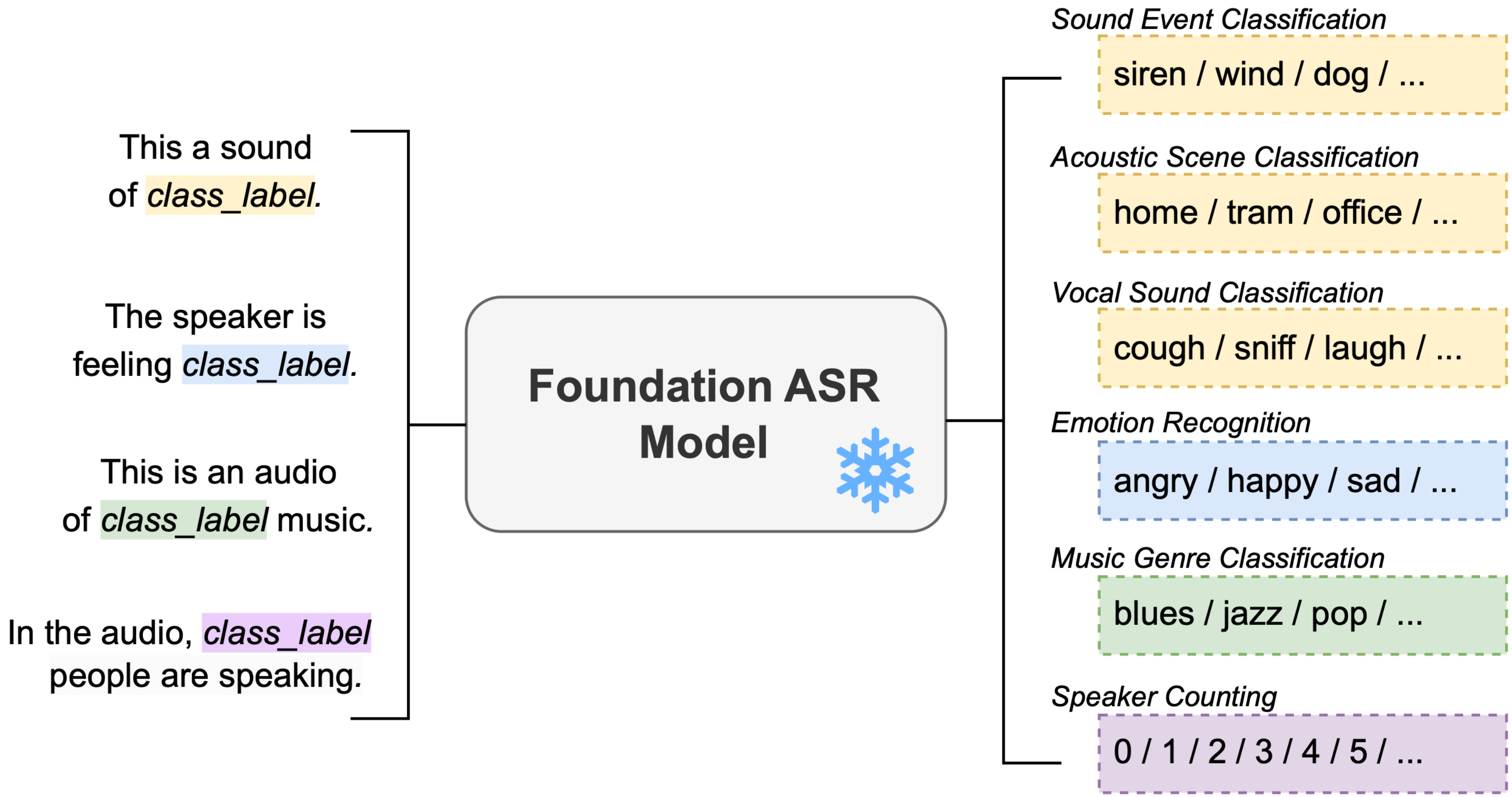
Task Calibration: Null-input

$$\tilde{P}_\theta(y_k|x) = \frac{P_\theta(t(y_k)|x)}{\sum_{y_j} P_\theta(t(y_j)|x)} \quad (1)$$

$$\hat{P}_\theta(y_k|x, \alpha_{1:K}) = \frac{\alpha_k \tilde{P}_\theta(y_k|x)}{\sum_i \alpha_i \tilde{P}_\theta(y_i|x)} \quad (2)$$

$$\bar{\alpha}_k \approx \frac{1}{\mathbb{E}_x\{P_\theta(y_k|x)\}} \approx \frac{1}{P_\theta(y_k|\phi)} \quad (5)$$

Prompts


This a sound
of class_label.
The speaker is
feeling class_label.
This is an audio
of class_label music.
In the audio, class_label
people are speaking.
Foundation ASR
Model
Sound Event Classification
siren / wind / dog / ...
Acoustic Scene Classification
home / tram / office / ...
Vocal Sound Classification
cough / sniff / laugh / ...
Emotion Recognition
angry / happy / sad / ...
Music Genre Classification
blues / jazz / pop / ...
Speaker Counting
0 / 1 / 2 / 3 / 4 / 5 / ...

Baseline: CLAP

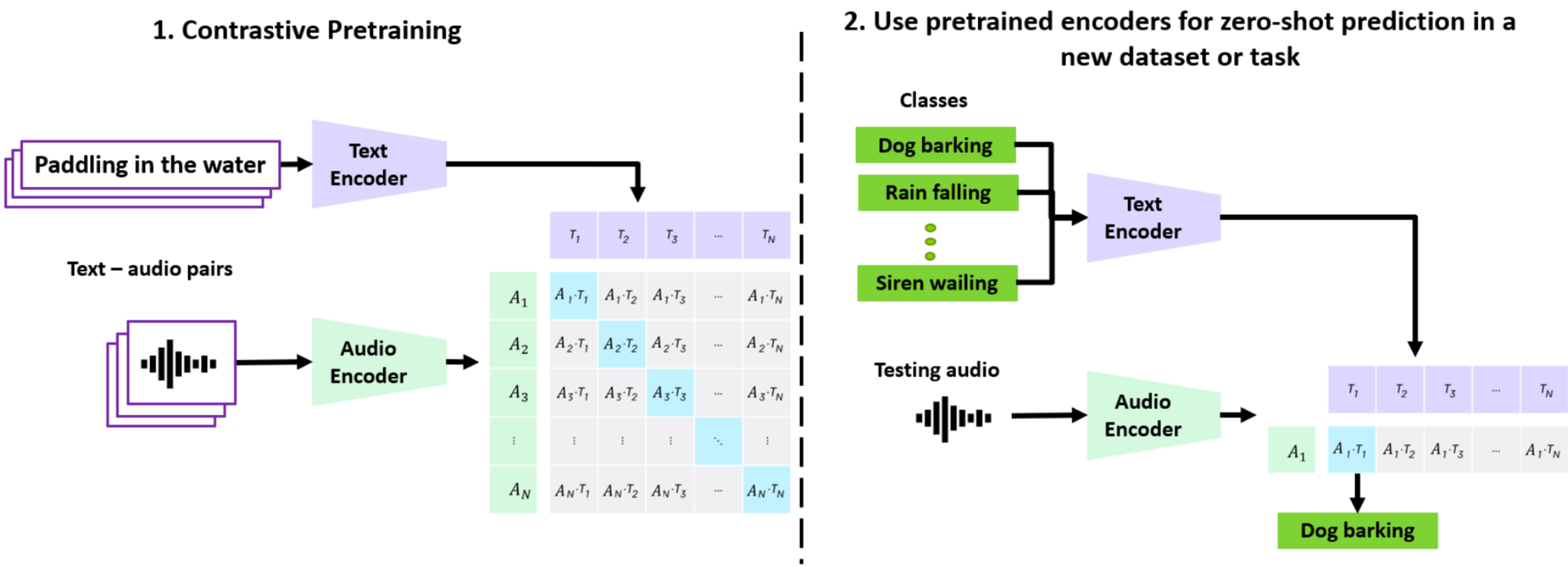


Fig. 1. CLAP jointly trains an audio and a text encoder to learn the (dis)similarity of audio and text pairs in a batch using contrastive learning. At testing time, the pretrained encoders are used to extract audio embeddings from the testing audio and text embeddings from the class labels. Zero-Shot linear classification is achieved by computing cosine similarity between the embeddings.

Main Results

Model	ESC50	US8K	TUT	Vocal	RAVDESS	CREMA-D	GTZAN	LibriCount	**Avg.**
Baselines (§4.3)									
Random	2.0	10.0	6.0	16.7	12.5	16.7	10.0	9.1	10.4
AudioCLIP	69.4	65.3	-	-	-	-	-	-	-
CLAP	82.6	73.2	29.6	49.4	16.0	17.8	25.2	17.9	39.0
Uncalibrated (§3.1)									
MMS large (1B)	1.7	9.6	4.9	14.2	13.5	17.2	8.3	8.4	9.7
Whisper medium.en (769M)	27.9	39.5	7.2	59.0	15.3	20.9	15.2	8.2	24.2
Whisper medium (769M)	29.7	45.8	7.5	44.6	16.7	19.9	28.4	9.4	25.2
Whisper large-v2 (1.6B)	38.9	50.5	7.7	60.1	15.1	20.2	38.2	9.2	30.0
Prior-matched (§3.3)									
MMS large (1B)	2.4	10.9	7.6	11.5	12.2	17.2	10.5	11.5	10.5
Whisper medium.en (769M)	56.2	60.9	18.3	82.8	29.0	22.6	29.7	9.8	38.7
Whisper medium (769M)	57.5	61.6	25.2	82.4	35.0	25.9	48.6	16.3	44.1
Whisper large-v2 (1.6B)	65.4	60.4	26.0	84.9	41.7	28.8	60.9	17.3	**48.2**

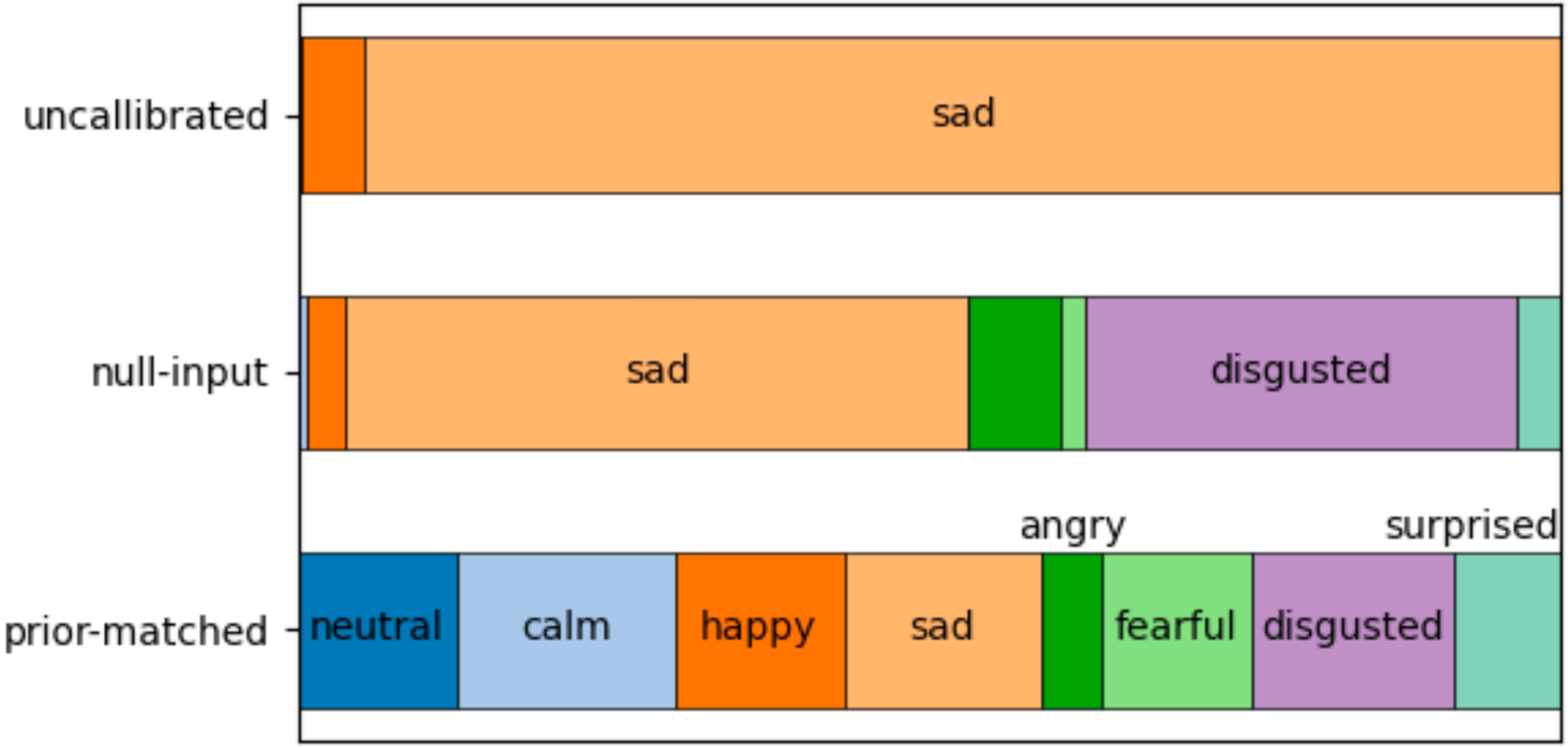
Table 3: Baseline and zero-shot task performance using the default prompts (of Table 2).

Results of Null-input Calibration

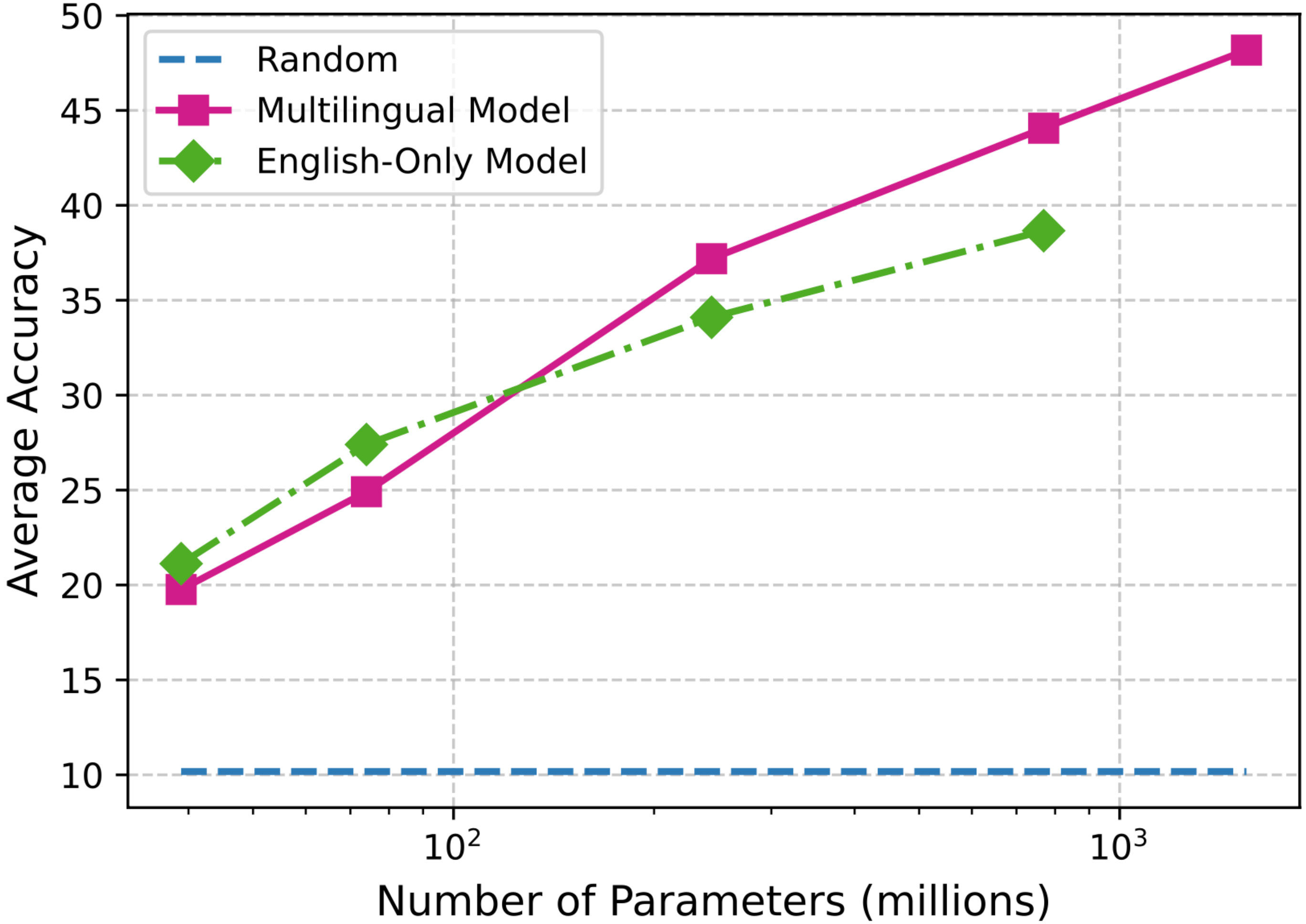
Method	medium.en	medium	large-v2
Uncalibrated	24.2	25.2	30.0
Zero Input	29.8	34.8	34.9
Gaussian Noise	28.5	29.5	35.8

Table 6: Average accuracy of 8 audio classification tasks with null-input calibration.

Distribution of Predictions on RAVDESS



Parameter Size vs Average Accuracy



Ablation of Prompts on RAVDESS

Prompt	Acc
The speaker is feeling *class_label*.	41.7
class_label	20.7
(class_label)	33.1
[class_label]	32.6
The person talking feels *class_label*.	38.5
The speaker is experiencing *class_label* emotions.	20.8
The person speaking is in a *class_label* mood.	29.9
The speaker's emotion is *class_label*.	33.6
The person talking is filled with *class_label* feelings.	39.7
Ensemble of Prompts	44.0

Ablation of Prompts on All Tasks

Dataset	Default	Ensemble
ESC50	65.4	67.1
US8K	60.4	67.6
TUT	26.0	25.2
Vocal	84.9	87.3
RAVDESS	41.7	44.0
CREMA-D	28.8	33.1
GTZAN	60.9	60.0
LibriCount	17.3	22.0
Average	48.2	**50.8**

Conclusions

- The first to examine the emergent ability of foundation ASR models on audio-classification tasks
- Zero-shot prompting of Whisper can yield effective performance
- Calibration methods can be used to readjust the output distribution for better task alignment
- Performance increases with model size, implying that as ASR foundation models scale up, they may exhibit improved zero-shot performance

UNIVERSITY OF CAMBRIDGE

Thank you for listening!