P6-18-SLR

Measuring Sound Symbolism in Audio-visual Models

Wei-Cheng Tseng, Yi-Jen Shih*, David Harwath and Raymond Mooney*

Department of Computer Science, The University of Texas at Austin, USA

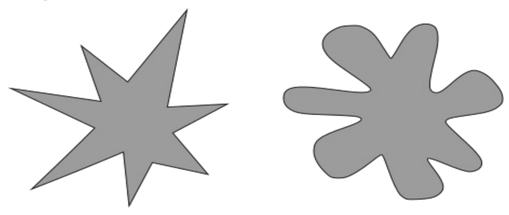
TL;DR

This work aims to bridge the gap between cognitive science and speech technologies. We investigate the existence of a linguistic phenomenon—**sound symbolism**—in pre-trained audio-visual models. Experimental results show that some models can capture sound-meaning associations akin to those observed in humans, influenced by the pre-training methods and datasets.

Background & Motivation

- Audio-visual representation learning has become an important research topic
- Pre-trained audio-visual models emerges intriguing language processing capabilities without direct supervision:
  - *AV-HuBERT* [1] → better acoustic unit discovery
  - *VG-HuBERT* [2] → word discovery & segmentation
  - *AV-NSL* [3] → syntax acquisition
- **What about more abstract linguistic phenomena?**
- We look at sound symbolism. Specifically, "**Kiki-bouba Effect**"
  - Independent of cultural and linguistic background
  - Important for early language acquisition

Dataset Collection



Audio:

- Synthesize sounds from predefined sets [4] of *sharp* or *round* phones w/ certain template:

  C_{SHARP} = {*k t p ʧ ʤ z*} V_{SHARP} = {*ɛ iː*}

  C_{ROUND} = {*m n l b d g*} V_{ROUND} = {*ɔː uː*}

  $C_{NEUTRAL}$ = {*f s v*} $V_{NEUTRAL}$ = {*aː*}

- $(CV)_1(CV)_2(CV)_1$
- Can't draw from opposite groups at same time

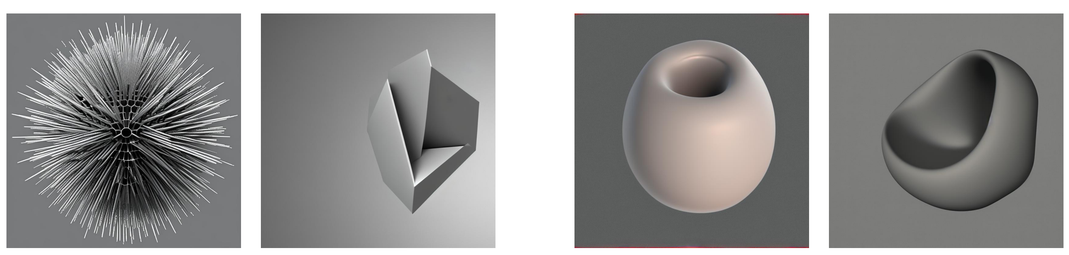
Image:

- Define sets of *sharp* or *round* adjectives:

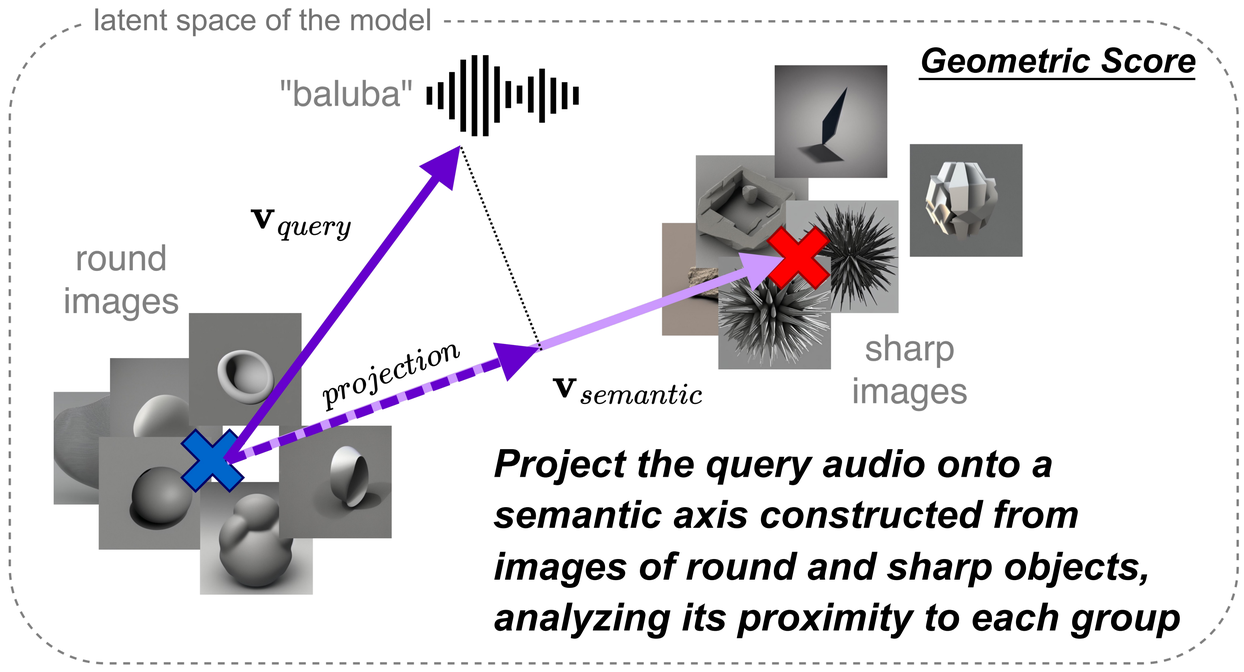
  W_{SHARP} = {*sharp spiky angular … pointed rugged*}

  W_{ROUND} = {*round circular soft … smooth chubby*}

- Synthesize w/ certain template
  - **A 3D-rendering of a ⟨*w*⟩ object**

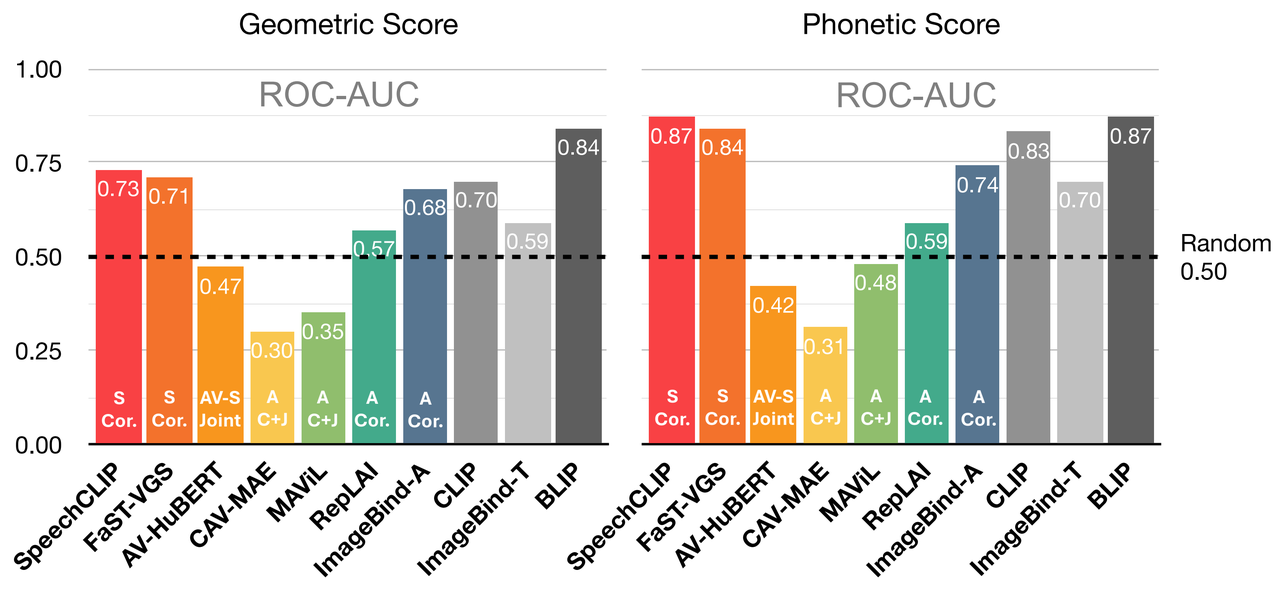


Evaluation: Zero-shot Probing



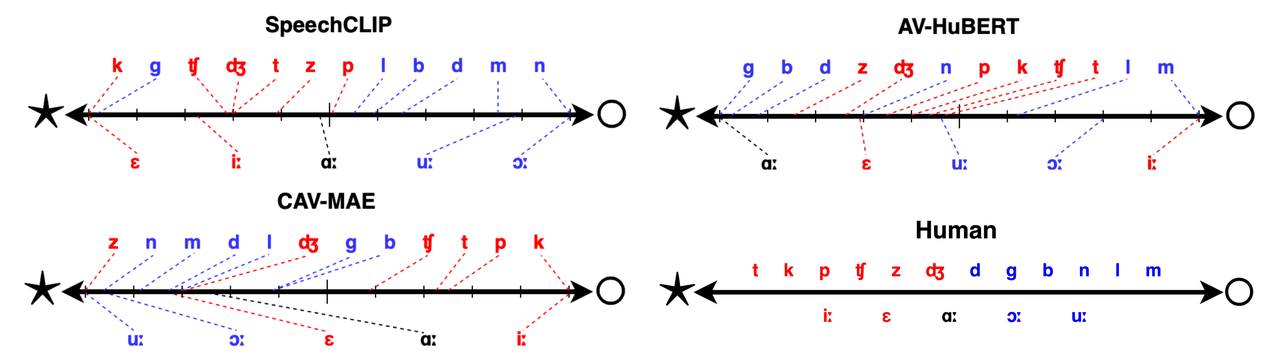
Results

- Apply a threshold to the scores (magnitude of the projected vectors) for binary classification



- Models pretrained on **spoken image captions** show more profound association than those trained on **general audio**
- Pretraining objectives/methodologies also affect

- Ranking of proximity of phone to sharp/round attr.



Contributions & Conlusions

- Investigate presence of *kiki-bouba effect* in pre-trained audio-visual models
- Reveal significant corr. between patterns of sound symbolism and models learned on spoken captions, analogous to human learning process
- Support non-arbitrariness of language and provide insight into audio-visual learning algorithm

Reference

[1] B. Shi et al., "Learning audio-visual speech repre- sentation by masked multimodal cluster prediction," 2022.
[2] P. Peng et al., "Word Discovery in Visually Grounded, Self-Supervised Speech Models," 2022.
[3] C.J. Lai et al., "Audio-visual neural syntax acquisition," 2023.
[4] Kelly McCormick et al., "Sound to meaning mappings in the bouba-kiki effect.," 2015.

Paper link

Email: raytseng@utexas.edu